

Fundamentos de ciencia de datos con R

Gema Fernández-Avilés y José-María Montero

2023-06-16

Índice general

Prefacio	17
¡Hola mundo!	17
¿Por qué este libro?	18
¿A quién va dirigido?	19
El paquete CDR	20
¿Por qué R?	20
Agradecimientos	21
0.1. Este manual está publicado por McGraw Hill. Las copias físicas están disponibles en McGraw Hill. La versión <i>online</i> se puede leer de forma gratuita en https://cdr-book.github.io/ y tiene la licencia de Creative Commons Reconocimiento-NoComercial-SinObraDerivada 4.0 Internacional. Si tiene algún comentario o sugerencia, no dude en contactar con los editores y los autores. ¡Gracias!	22
I Ciencia, datos, software... y científicos	23
1. ¿Es la ciencia de datos una ciencia?	25
1.1. ¿Qué se entiende por ciencia?	25
1.2. ¿Qué es la ciencia de datos?	26
1.3. Lo científico de la ciencia de datos	28
2. Metodología en ciencia de datos	31
2.1. Preliminares	31
2.2. Principales metodologías en ciencia de datos	32
2.3. CRISP-DM para ciencia de datos	33

3. R para ciencia de datos	37
3.1. Introducción	37
3.2. La sesión de R	38
3.3. Instalación de R	39
3.4. Trabajar con proyectos de RStudio	40
3.5. Tratamiento de datos con R	40
3.6. Organización de datos con el <i>tidyverse</i>	45
4. Ética en la ciencia de datos	55
4.1. ¿Qué es la ética?	55
4.2. Los principios éticos	56
4.3. Equidad: la importancia de los sesgos	58
4.4. ¿Es necesaria la explicabilidad?	61
4.5. Recursos en R para trabajar en sesgos y explicabilidad	63
II Bienvenidos a la jungla de datos	65
5. Gestión de bases de datos relacionales	67
5.1. Introducción	67
5.2. Concepto de base de datos	68
5.3. SQL: el lenguaje estructurado de consulta	70
5.4. Acceso y explotación de bases de datos desde R	73
6. Gestión de bases de datos NoSQL	83
6.1. Introducción al big data	83
6.2. Las V's del big data	84
6.3. Tipos de datos en entornos big data	85
6.4. ¿Por qué bases de datos NoSQL?	86
6.5. Bases de datos NoSQL	87
6.6. Integración de bases de datos NoSQL en R	93

Índice general

5

7. Gobierno, gestión y calidad del dato	99
7.1. Introducción	99
7.2. Concepto de gobierno del dato	100
7.3. Marcos y metodologías de gobierno del dato	103
7.4. Gestión de calidad del dato	105
8. Integración y limpieza de datos	115
8.1. Introducción	115
8.2. Integración de datos	115
8.3. Limpieza de datos	119
9. Selección y transformación de variables	133
9.1. Introducción	133
9.2. Selección de variables	134
9.3. Transformación de variables	143
9.4. Reducción de dimensionalidad	148
10. Herramientas para el análisis en ciencia de datos	151
10.1. Introducción	151
10.2. Partición del conjunto de datos	151
10.3. Técnicas para manejar datos no equilibrados	154
10.4. El enfoque de validación	155
10.5. Compensación (<i>trade off</i>) entre sesgo y varianza	158
10.6. Ajuste de hiperparámetros	159
10.7. Evaluación de modelos	161
11. Análisis exploratorio de datos	169
11.1. Introducción	169
11.2. Análisis exploratorio de una variable	173
11.3. Análisis exploratorio de varias variables	181

III Fundamentos de estadística	193
12. Probabilidad	195
12.1. Introducción a la probabilidad	195
12.2. Probabilidad: elementos básicos, definición y teoremas	196
12.3. Variable aleatoria y su distribución de probabilidad	198
12.4. Modelos de distribución de probabilidad	199
12.5. Teorema central del límite	206
12.6. Distribuciones de probabilidad en R	207
13. Inferencia estadística	213
13.1. Introducción	213
13.2. Muestreo aleatorio simple	214
13.3. Estimación puntual	216
13.4. Estimación por intervalos	218
13.5. Contrastes de hipótesis	219
13.6. Inferencia estadística paramétrica sobre poblaciones normales	220
13.7. Inferencia sobre poblaciones normales con R	223
13.8. Inferencia estadística no paramétrica: contrastes de normalidad	225
14. Muestreo y remuestreo	229
14.1. Introducción al muestreo	229
14.2. Muestreo aleatorio simple	230
14.3. Muestreo estratificado	234
14.4. Otros tipos de muestreo probabilístico	236
14.5. Técnicas de remuestreo: Bootstrap	237
IV Modelización estadística	243
15. Modelización lineal	245
15.1. Modelización	245
15.2. Procedimiento de modelización	247
15.3. Procedimiento con R : la función <code>lm()</code>	249

Índice general

7

15.4. Casos prácticos	250
15.5. Comentarios finales	258
16. Modelos lineales generalizados	259
16.1. Introducción	259
16.2. El modelo y sus componentes	260
16.3. Procedimiento con R : la función <code>glm()</code>	261
16.4. Regresión logística	262
16.5. Regresión de Poisson	266
16.6. Casos prácticos	266
17. Modelos aditivos generalizados	279
17.1. Introducción	279
17.2. Splines con penalizaciones	280
17.3. Aspectos metodológicos	281
17.4. Procedimiento con R : la función <code>gam()</code> del paquete <code>mgcv</code>	284
17.5. Casos prácticos	285
18. Modelos mixtos	297
18.1. Conceptos básicos	297
18.2. Formulación del modelo con efectos aleatorios o modelos mixtos	301
18.3. Procedimiento con R para ajustar modelos mixtos	303
18.4. Caso práctico	303
19. Modelos <i>sparse</i> y métodos penalizados de regresión	317
19.1. Introducción	317
19.2. Selección del mejor subconjunto	318
19.3. Métodos <i>shrinkage</i>	324
20. Modelización de series temporales	337
20.1. Conceptos básicos	337
20.2. Modelos ARIMA	339
20.3. Análisis de series temporales con R	340

21. Análisis discriminante	359
21.1. Introducción	359
21.2. Análisis discriminante lineal	361
21.3. Análisis discriminante cuadrático	371
22. Análisis conjunto	375
22.1. Introducción y conceptos clave	375
22.2. Tipos de análisis conjunto	376
22.3. Etapas de la realización del análisis conjunto	377
22.4. Procedimiento con R: la función <code>Conjoint()</code>	382
23. Análisis de tablas de contingencia	387
23.1. Introducción	387
23.2. Contraste de independencia en tablas 2×2	390
23.3. Contraste de independencia en tablas $R \times C$	396
23.4. Medidas de asociación en tablas 2×2	399
23.5. Medidas de asociación en tablas $R \times C$	402
23.6. Contrastos de independencia en tablas multidimensionales	405
V Machine learning supervisado	407
24. Árboles de clasificación y regresión	409
24.1. Introducción	409
24.2. Procedimiento con R: la función <code>rpart()</code>	412
24.3. Árboles de clasificación	412
24.4. Árboles de regresión	427
25. Máquinas de vector soporte	437
25.1. Introducción	437
25.2. Algoritmo SVM para clasificación binaria	439
25.3. ¿Y si tengo más de dos clases?	440
25.4. Truco del <i>kernel</i> : tratando con la no linealidad	440
25.5. Procedimiento con R: la función <code>svm()</code>	442
25.6. Aplicación de un modelo SVM Radial con ajuste automático en R	442

Índice general

9

26. Clasificador k-vecinos más próximos	449
26.1. Introducción	449
26.2. Decisiones a tener en cuenta	450
26.3. Procedimiento con R : la función <code>knn()</code>	452
26.4. Aplicación del modelo KNN en R	452
27. Naive Bayes	457
27.1. Introducción	457
27.2. Teorema de Bayes	457
27.3. El algoritmo <i>naive</i> Bayes	458
27.4. Procedimiento con R : la función <code>naive_bayes()</code>	461
27.5. Clasificación de clientes utilizando el modelo <i>Naive Bayes</i>	462
28. Métodos ensamblados: bagging y random forest	465
28.1. Introducción a los métodos ensamblados	465
28.2. Bagging	465
28.3. Random Forest	471
29. Boosting y el algoritmo XGBoost	483
29.1. Métodos ensamblados: bagging vs boosting	483
29.2. ¿Qué es el boosting?	484
29.3. Gradient Boosting (GB)	484
29.4. eXtreme Gradient Boosting (XGB)	493
VI Machine learning no supervisado	499
30. Análisis cluster: clusterización jerárquica	501
30.1. Introducción	501
30.2. Selección de las variables	502
30.3. Elección de la distancia entre elementos	503
30.4. Técnicas de agrupación jerárquicas	508
30.5. Calidad de la agrupación y número de clusters	519

31. Análisis cluster: clusterización no jerárquica	525
31.1. Métodos de reasignación	525
31.2. Métodos basados en la densidad de elementos	530
31.3. Otros métodos	532
31.4. Nota final	533
32. Análisis de componentes principales	535
32.1. Introducción	535
32.2. Obtención de las componentes principales	536
32.3. Estimación de las componentes principales	540
32.4. Número de componentes a retener	540
32.5. Interpretación de las componentes principales	542
32.6. Reproducción de los datos tipificados y de la matriz de	544
32.7. Limitaciones del análisis de componentes principales	545
33. Análisis factorial	547
33.1. Introducción	547
33.2. Elementos teóricos del análisis factorial	548
33.3. El análisis factorial en la práctica	552
33.4. Relaciones y diferencias entre el AF y el ACP	565
34. Escalamiento multidimensional	567
34.1. Introducción	567
34.2. Medición de distancias y similitudes	568
34.3. Modelo de escalamiento multidimensional	569
34.4. Tipos de escalamiento multidimensional	571
35. Análisis de correspondencias	581
35.1. Introducción	581
35.2. Metodología del análisis de correspondencias	582
35.3. Procedimiento con R: la función ca()	584

Índice general

11

VII Deep learning	591
36. Redes neuronales artificiales	593
36.1. ¿Qué es el <i>deep learning</i> ?	593
36.2. Aplicaciones del <i>deep learning</i>	595
36.3. Redes neuronales	598
36.4. Perceptrón o neurona	598
36.5. Perceptrón multiclasa	600
36.6. Funciones de activación	601
36.7. Perceptrón multicapa	605
36.8. Instalación de librerías de <i>deep learning</i> en R : Tensorflow/Keras	608
36.9. Ejemplo de red para clasificación en R	609
36.10. Ejemplo de red para regresión en R	616
37. Redes neuronales convolucionales	621
37.1. Introducción	621
37.2. Convolución	622
37.3. Neuronas convolucionales	624
37.4. Relleno del borde	625
37.5. Capas de agrupación	626
37.6. Desvanecimiento del gradiente	627
37.7. Sobreajuste	628
37.8. Generación de datos de entrenamiento artificiales	629
37.9. Ejemplo en R para el conjunto de datos CIFAR10	631
VIII Ciencia de datos de texto y redes	639
38. Minería de textos	641
38.1. Introducción	641
38.2. Conceptos y tareas fundamentales	642
38.3. Análisis de sentimientos	645
38.4. Minería de textos en R	646
38.5. Ejemplo de aplicación	647

39. Análisis de grafos y redes sociales	659
39.1. Introducción	659
39.2. Teoría de grafos	660
39.3. Elementos de un grafo	661
39.4. Procedimiento con R: el paquete <i>igraph</i>	664
39.5. Análisis de influencia en un grafo aplicado a RRSS	666
39.6. Entorno social en el universo cinematográfico Marvel	672
IX Ciencia de datos espaciales	677
40. Trabajando con datos espaciales	679
40.1. Introducción	679
40.2. Conceptos clave	680
40.3. Mi primer mapa	688
40.4. ¿Cómo (no) mentir con la visualización?	689
40.5. Mapas espacio-temporales	691
40.6. Mapas interactivos	692
41. Geoestadística	695
41.1. Introducción	695
41.2. Preliminares	696
41.3. Análisis estructural de la dependencia espacial	697
41.4. Kriging	709
42. Modelos econométricos espaciales	713
42.1. La dependencia espacial	713
42.2. Medidas de autocorrelación espacial	717
42.3. Modelos econométricos espaciales de corte transversal	719
43. Procesos de puntos	731
43.1. Introducción	731
43.2. Patrones puntuales espaciales en \mathbb{R}^2	732
43.3. Patrones puntuales espaciales sobre redes lineales	743

Índice general 13

X Comunica y colabora	749
44. Informes reproducibles con R Markdown y Quarto	751
44.1. ¿Por qué informes reproducibles?	751
44.2. Documentos Quarto	756
44.3. Otros formatos	763
45. Creación de aplicaciones web interactivas con Shiny	765
45.1. Introducción	765
45.2. Componentes mínimos de una aplicación Shiny y disposición básica	766
45.3. Diseño de una aplicación Shiny	767
45.4. Elementos para la introducción de datos	770
45.5. Elementos para visualización (salida)	774
45.6. Reactividad	775
45.7. Publicación de la aplicación en la web	777
45.8. Extensiones de Shiny	779
46. Git y GitHub R	781
46.1. ¿Qué es Git y GitHub?	781
46.2. ¿Por qué usar Git y GitHub?	781
46.3. Instalación y/o actualización de R y RStudio	782
46.4. Configuración de Git y GitHub	783
46.5. Conectar Git y GitHub con Rstudio	786
46.6. Flujo de trabajo general de Git y GitHub en RStudio	790
47. Geoprocесamiento en nube	795
47.1. Introducción	795
47.2. Sintaxis de Google Earth Engine	796
47.3. Primeros pasos	796
47.4. Cálculo de anomalías	797

XI Casos de estudio en ciencia de datos	801
48. Análisis de una red criminal	803
48.1. Introducción	803
48.2. El conjunto de datos <i>Oversize</i>	803
48.3. Creación de la red mafiosa	804
48.4. Visualización de la red mafiosa	804
48.5. Importancia de los actores (delincuentes)	806
48.6. Identificación de comunidades de la mafia	807
48.7. Visualización de comunidades de la mafia	807
49. Optimización de inversiones publicitarias	811
49.1. Metodologías para optimizar las inversiones publicitarias	811
49.2. Robyn como alternativa <i>open-source</i> en R	813
50. ¿Cómo twitea Elon Musk?	821
50.1. Introducción	821
50.2. Análisis visual de los <i>tweets</i> de Elon Musk	821
51. Análisis electoral: de Rstudio a su periódico	835
51.1. Motivación	835
51.2. Obtención de los datos	835
51.3. Transformación y primeros gráficos	836
52. Crisis: impacto en el paro de Castilla-La Mancha	843
52.1. Planteamiento	843
52.2. Evolución del paro medio anual en Castilla-La Mancha	844
52.3. Evolución del paro medio anual en función de la edad y el sexo	845
52.4. Evolución del paro medio anual según el tiempo de búsqueda de empleo	848
52.5. Evolución del paro medio anual según sexo, edad y sector de procedencia	849
52.6. Conclusiones	850

Índice general 15

53. Segmentación de clientes en el comercio minorista	851
53.1. Motivación y conceptos clave	851
53.2. El modelo <i>Recency, frequency, monetary</i> tradicional	852
53.3. El modelo <i>Recency, frequency, monetary</i> extendido	852
54. Análisis de datos en medicina	859
54.1. Justificación	859
54.2. Introducción al uso de datos en investigación clínica y ensayos clínicos	859
54.3. Análisis de supervivencia	864
54.4. Regresión de COX	866
54.5. Conclusión	868
55. Messi y Ronaldo: dos ídolos desde la perspectiva de los datos	869
55.1. Motivación	869
55.2. Las estadísticas y el fútbol	869
56. Un dato sobre el cambio climático	879
56.1. Introducción	879
56.2. Consideraciones iniciales	880
56.3. Paquetes	880
56.4. Visualización de mapas “pequeños múltiples”	881
57. Predicción de consumo eléctrico con redes neuronales	889
57.1. Introducción	889
57.2. Datos de entrada	889
57.3. Modelización	890
58. Implementación de un sistema experto en el ámbito pediátrico	897
58.1. Introducción	897
58.2. Marco teórico	897
58.3. Sistema experto para el ámbito pediátrico en atención primaria	902
59. El procesamiento del lenguaje natural para tendencias de moda en textil	911
59.1. Introducción	911
59.2. Análisis de tendencias de moda en textil	911

60. Detección de fraude de tarjetas de crédito	919
60.1. Introducción	919
60.2. Modelización del fraude en la compra con tarjetas de crédito	920
A. Información de la sesión	929

Prefacio

¡Hola mundo!

El siglo XXI está siendo testigo de grandes cambios vertiginosos en el contexto social y tecnológico, entre otros. Los tiempos han cambiado, la sociedad se ha globalizado y “exige” respuestas inmediatas a problemas muy complejos. Vivimos en el mundo de la **información**, de los **datos**, o mejor, de las **bases de datos masivas**, y los ciudadanos y, sobre todo, las empresas y los gobiernos, dirigen su mirada hacia el mundo científico para que les ayude a “**oír las historias**” que cuentan esos datos acerca de la realidad de la que han sido extraídos. Y dado su enorme volumen y sofisticación (en el nuevo mundo las imágenes y los textos, por ejemplo, también son datos), exigen algoritmos de nueva generación en el campo del *machine learning*, o incluso del *deep learning*, para “oír las historias” que cuentan. No parecen mirar al “antiguo” investigador científico, sino al “nuevo” *científico de datos*.

Ello, inevitablemente, se traduce en la necesidad de profesionales con una gran capacidad de adaptación a este nuevo paradigma: los científicos de datos, también llamados por algunos los “nuevos hombres del Renacimiento”, para lo cual las Universidades y demás instituciones educativas especializadas se apresuran a incluir el grado de Ciencia de Datos en su oferta educativa y a ofrecer seminarios de software estadístico de acceso abierto para sus estudiantes de primeros cursos.

Con la emergencia de la nueva sociedad, en la que el manejo de la ingente cantidad de información que genera se hace absolutamente necesario para circular por ella, la **Ciencia de Datos** ha venido para quedarse. Sin embargo, el mundo de la Ciencia de Datos es cualquier cosa menos sencillo. En él, cualquier ayuda, cualquier guía es bienvenida. Por ello, es muy recomendable que la persona que se quiera introducir en él, sea con fines de investigación o con fines profesionales, se agarre de la mano de un guía especializado que le lleve, de una manera amena, comprensible y eficiente, desde el planteamiento de su problema y la captura de la información necesaria para poderle dar una solución, hasta la redacción de las conclusiones finales que ha obtenido con los modernos informes reproducibles colaborativos. Y como en la parte central de ese camino tendrá que luchar con grandes gigantes (en la actualidad denominados técnicas estadísticas y algoritmos), el guía tendrá que explicarle, de manera sencilla y amena, en qué consiste la lucha (las técnicas y los algoritmos) y cómo llegar a la victoria lo más rápido posible, enseñándole a moverse por el mundo del software estadístico, en nuestro caso **R**, que le permitirá realizar los cálculos necesarios para vencer al problema planteado a una velocidad vertiginosa.

En resumen, la información masiva y el moderno tratamiento estadístico de la misma son la “mano invisible” que gobierna la sociedad del siglo XXI, y este manual pretende ser el guía anteriormente mencionado que le llevará de la mano cuando quiera caminar por ella.

¿Por qué este libro?

Lo dicho anteriormente ya justifica por sí solo la aparición de este manual. Afortunadamente, no es el primero en la materia, pues son ya bastantes los materiales de calidad publicados sobre Ciencia de Datos. Sin embargo, quizás, éste pueda ser considerado el más completo. Y ello por varias razones.

La primera es su **completitud**: este manual lleva de la mano al lector desde el planteamiento del problema hasta el informe que contiene la solución al mismo; o desde no saber qué hacer con la información de la que dispone, hasta ser capaz de transformar tales bases de datos masivas, y casi imposibles de manejar, en respuestas a problemas fundamentales de una empresa, institución o cualquier agente social.

La segunda es su **amplitud temática**:

- (I) Parte de las dos primeras preguntas que un neófito se puede hacer sobre esta temática: ¿qué es eso de la Ciencia de Datos que está en boca de todos? Y, ¿qué diablos es **R** y cómo funciona?
- (II) Enseña cómo moverse en la jungla del *Big Data* y de los “nuevos” tipos de datos, siempre bajo el paraguas de la ética de los datos y del buen gobierno de dichos datos.
- (III) Muestra al lector cómo obtener conocimiento de la oscuridad del enorme banco de información a su disposición, que no sabe cómo abordar ni manejar.
- (IV) No deja a nadie atrás, y de forma previa al contenido central del manual (las técnicas de Ciencia de Datos), incluye unas breves, pero magníficas, secciones sobre los rudimentos de la probabilidad, la inferencia estadística y el muestreo, para aquéllos no familiarizados con estas cuestiones.
- (V) Aborda una treintena de técnicas de Ciencia de Datos en el ámbito de la modelización, análisis de datos cualitativos, discriminación, *machine learning* supervisado y no supervisado, con especial incidencia en las tareas de clasificación y clusterización -así como, en el caso no supervisado, de reducción de la dimensionalidad, escalamiento multidimensional y análisis de correspondencias-, *deep learning*, análisis de datos textuales y de redes, y, finalmente, ciencia de datos espaciales (desde las perspectivas de la geoestadística, la econometría espacial y los procesos de punto).
- (VI) Hace especial hincapié en la reproducibilidad en tiempo real (o no) entre los distintos miembros de un equipo (sea universitario, empresarial, o del tipo que sea) y en la difusión de los resultados obtenidos, enseñando al lector cómo generar informes reproducibles mediante RMarkdown y documentos Quarto o en otros modernos formatos.
- (VII) Dedica un capítulo a la creación de aplicaciones web interactivas (con Shiny).

Índice general

19

- (viii) Para aquéllos con pasión por la codificación, y que quieran compartir código y colaborar con otros desarrolladores, este manual aborda la gestión rápida y eficaz de proyectos (del tamaño que sean) mediante Git, un sistema de control de versiones distribuido, gratuito y de código abierto, y GitHub, un servicio de alojamiento de repositorios Git del cual, aquellos no familiarizados con la cuestión de la codificación, o con aversión a ella, podrán tomar el código que necesitan.
- (ix) Muestra al lector los primeros pasos para iniciarse en el geoprocесamiento en la nube.
- (x) Y, finalmente, aborda más de una docena de casos de uso (en medicina, periodismo, economía, criminología, marketing, moda, demanda de electricidad, cambio climático, reconocimiento de patrones en la forma de tuitear...) que ilustran la puesta en práctica de todos los conocimientos anteriormente adquiridos.

La cuarta razón es que todo lo que el lector aprende en este manual lo puede reproducir y poner en práctica inmediatamente con **R**, puesto que el manual está trufado de *chunks* (o trozos de código **R**) que no tiene más que cortar y pegar para reproducir los ejemplos que se muestran en el libro, cuyos datos están en el paquete CDR; o utilizar dichas *chunks* para abordar el problema que le ocupa con los datos que tenga a su disposición. Una buena razón, sin duda. Por consiguiente, el manual es una buena combinación “teoría-práctica-software” que permite abordar cualquier problema que el científico de datos se plante en cualquier disciplina o situación empresarial, médica, periodística...

La quinta es su **variedad de perspectivas**. Son **más de 40 los participantes** en este manual. Algunos de ellos, prestigiosos profesores universitarios; otros, destacados miembros de instituciones públicas; otros, CEOs de empresas en la órbita de la ciencia de datos; otros, *big names* del mundo de **R** software... El manual es, sin duda, un magnífico ejemplo de colaboración Universidad-Empresa para buscar soluciones a los problemas de las sociedades modernas.

¿A quién va dirigido?

Fundamentos de ciencia de datos con R está dirigido a todos aquellos que desean desarrollar las habilidades necesarias para abordar proyectos complejos de Ciencia de Datos y “pensar con datos” (como lo acuñó Diane Lambert, de Google). El deseo de resolver problemas utilizando datos es su piedra angular. Por tanto, como se avanzó anteriormente, este manual no deja a nadie atrás, y lo único que requiere es “el deseo de resolver problemas utilizando datos”. No excluye ninguna disciplina, no excluye a las personas que no tengan un elevado nivel de análisis estadístico de datos, no excluye a nadie. Se ha procurado una combinación de rigor y sencillez, y de teoría y práctica, todo ello con sus correspondientes códigos en **R**, que satisfaga tanto a los más exigentes como a los principiantes.

También está destinado a todos aquellos que quieran sustituir la navegación por la web (la búsqueda del video, publicación de blog o tutorial *online* que solucione su problema –frustración tras frustración por la falta de consistencia, rigor e integridad de dichos materiales, así como por su sesgo hacia paquetes singulares para la implementación de las cuestiones que tratan–), por

una “**biblia de la ciencia de datos**” rigurosa pero sencilla, práctica y de aplicación inmediata sin ser ni un experto estadístico ni un experto informático.

Pero si a alguien está destinado especialmente, es a la comunidad hispano hablante. Este manual es un guiño a dicha comunidad, para que tenga a su disposición, en su lengua nativa, uno de los mejores manuales de Ciencia de Datos de la actualidad.

El paquete CDR



El paquete **CDR** contiene la mayoría de conjuntos de datos utilizados en este libro que no están disponibles en otros paquetes. Para instalarlo use la función `install_github()` del paquete `remotes`.

```
# este comando sólo necesita ser ejecutado una vez
# si el paquete remotes no está instalado, descomentar para instalarlo

# install.packages("remotes")
remotes::install_github("cdr-book/CDR")
```

La lista de todos los conjuntos de datos puede obtenerse haciendo `data()`.

```
library('CDR')
data(package = "CDR")
```

Este paquete ayudará al lector a reproducir todos los ejemplos del libro. De acuerdo con las mejores prácticas en **R**, el paquete **CDR** sólo contiene los datos utilizados en el libro.

¿Por qué R?

R es un lenguaje de código abierto para computación estadística que se ha consolidado entre la comunidad científica internacional, en las últimas dos décadas, como una herramienta de primer

nivel, consolidándose como líder permanente en el ámbito de la implementación de metodologías estadísticas para el análisis de datos. La utilidad de **R** para la Ciencia de Datos deriva de un fantástico ecosistema de paquetes (activo y en crecimiento), así como de un buen elenco de otros excelentes recursos: libros, manuales, *blogs*, foros y *chats* interactivos en las redes sociales, y una gran comunidad dispuesta a colaborar, a orientar y a resolver diferentes cuestiones relacionadas con **R**.

Por otra parte, **R** es el lenguaje estadístico y de análisis de datos más utilizado en la mayoría de los entornos académicos y, cómo no, por una larga lista de importantes empresas, entre las que se cuentan Facebook (análisis de patrones de comportamientos relacionado con actualizaciones de estado e imágenes de perfil), Google (para la efectividad de la publicidad y la previsión económica), Twitter (visualización de datos y agrupación semántica), Microsoft (adquirió la empresa Revolution R), Uber (análisis estadístico), Airbnb (ciencia de datos), IBM (se unió al grupo del consorcio R), New York Times (visualización)...

La comunidad **R** también es particularmente generosa e inclusiva, y hay grupos increíbles, como *R-Ladies* y *Minority R Users*, diseñados para ayudar a garantizar que todos aprendan y usen las capacidades de **R**.

Agradecimientos

No queremos dar por finalizado este prefacio sin agradecer a los 44 autores participantes en esta obra su esfuerzo por condensar, en no más de 20 páginas, la teoría, práctica y tratamiento informático de la parte de la Ciencia de Datos que les fue encargada. Y no sólo eso; el “más difícil todavía” fue que debían dirigirse a un abanico de potenciales lectores tan grande como personas haya con “el deseo de resolver problemas utilizando datos”. Era misión imposible. Sin embargo, a la vista del resultado, ha sido misión cumplida. El esfuerzo mereció la pena.

Además, nos gustaría agradecer el apoyo incondicional recibido por (en orden alfabético): Itzcoatl Bueno, Ismael Caballero, Emilio L. Cano, Diego Henangómez, Ricardo Pérez, Manuel Vargas y Jorge Velasco.

También queremos poner de manifiesto que la edición de este texto ha sido financiada por diversos entes de la Universidad de Castilla-La Mancha. En su mayor parte, por el **Máster en Data Science y Business Analytics (con R software)** (a través de la orgánica: 02040M0280), pero también por la Facultad de Ciencias Jurídicas y Sociales de Toledo (a través de su contrato programa: orgánica 00440710), el Departamento de Economía Aplicada I (mediante sus fondos departamentales, DEAI 00421I126) y el Grupo de Investigación Economía Aplicada y Métodos Cuantitativos (que ha dedicado parte de sus fondos a la edición de esta obra, orgánica 01110G3044-2023-GRIN-34336).

A todos, eternamente agradecidos por ayudarnos en este reto de transformar la oscuridad en conocimiento, de convertir en una ciencia y en un arte la difícil tarea de sacar valor de los datos, el petróleo del futuro. Quizás en este momento no seamos conscientes de que hemos puesto nuestro granito de arena a la ciencia que, a buen seguro, juegue uno de los papeles más importantes de este siglo, caracterizado por el predominio de la información. Una ciencia, la Ciencia de Datos, que combina el análisis estadístico de datos, la algoritmia y el conocimiento del

negocio para sacar valor del bien más abundante de la sociedad en la que vivimos: la información. Una disciplina cuyo dominio caracteriza a los científicos de datos (también denominados los nuevos personajes del Renacimiento), profesión que ya fue calificada hace más de veinte años en la *Harvard Business Review* y en *The New York Times*, entre otros, como la “más sexy del siglo XXI”.

Nota

- 0.1. Este manual está publicado por McGraw Hill. Las copias físicas están disponibles en McGraw Hill. La versión *online* se puede leer de forma gratuita en <https://cdr-book.github.io/> y tiene la licencia de Creative Commons Reconocimiento-NoComercial-SinObraDerivada 4.0 Internacional. Si tiene algún comentario o sugerencia, no dude en contactar con los editores y los autores. ¡Gracias!**

Parte I

Ciencia, datos, software... y científicos

Capítulo 1

¿Es la ciencia de datos una ciencia?

Gema Fernández-Avilés¹

Universidad de Castilla-La Mancha

1.1. ¿Qué se entiende por ciencia?

No es posible determinar si la ciencia de datos es una ciencia sin previamente consensuar la definición de ciencia. La palabra *ciencia* deriva del latín (*scientia*), que deriva a su vez del verbo latino *scire* (saber). En este sentido, Bunge (2004) distingue dos categorías fundamentales de conocimiento: el vulgar y el científico. El conocimiento vulgar se adquiere de manera cotidiana a partir de percepciones y sensaciones individuales, apoyándose en la evidencia (pruebas) y el sentido común. Este tipo de conocimiento constituye la base sobre la que se sustenta el conocimiento científico, que se enriquece al verificar, con la ayuda de un método, la validez de las observaciones realizadas. Se adquiere, por tanto, de forma consciente, deliberada y metódica. Puede ser sometido a prueba y, llegado el caso, ser superado Montero (1997).

Una definición generalmente aceptada de ciencia es la propuesta por Blaug (1980): “ciencia es el cuerpo de proposiciones sintéticas acerca del mundo real que es susceptible, al menos en principio, de falsaciones por medio de la observación empírica, ya que excluye la posibilidad de que ciertos acontecimientos se produzcan. Así pues, la ciencia se caracteriza por su método de formulación de proposiciones contrastables, y no por su contenido, ni por su pretensión de certeza en el conocimiento; si alguna certeza proporciona la ciencia, esta será más bien la certeza de la ignorancia.”

Si bien esta definición está encuadrada dentro del enfoque del falsacionismo de Karl Popper (para el cual las teorías son conjjeturas, hipótesis generales que permiten explicar fenómenos),

¹Quisiera agradecer a Bilal Laouah y a José-María Montero los comentarios realizados durante el desarrollo de este capítulo.

introduce la palabra clave de la discusión planteada, el **método**, y más concretamente, el **método científico**.

La palabra *método* procede del latín *methodus**, y esta del griego $\mu\epsilon\thetao\delta0\varsigma$, que quiere decir “**el camino a seguir para ir más allá**”. Es, pues, un procedimiento para conseguir algo; y como el fin que busca la ciencia es la verdad, el método científico es el camino mediante el cual las ciencias pueden llegar a encontrar sus respectivas verdades. El método científico es, por tanto, el conjunto de procedimientos por medio de los cuales se adquieren conocimientos rigurosos, ciertos y seguros acerca de un objeto. El método científico ha de recorrerse siguiendo cuatro fases ([Cancelo, 1997](#)): (i) inventario previo de los fenómenos o de los hechos significativos no rutinarios; (ii) planteamiento de un tema problemático que hace necesaria una explicación; (iii) ideación de conjeturas tendentes a darla, y (iv) tratamiento de las diversas hipótesis hasta que sólo una se mantenga.

Por tanto, aunque no hay acuerdo a la hora de dar una definición exacta de ciencia, sí lo hay a la hora de aceptar el método científico como el elemento que define la ciencia: “el método científico y la finalidad a la cual se aplica (el conocimiento objetivo del mundo) constituyen la entera diferencia que existe entre la ciencia y la no-ciencia [...]. El método científico es un rasgo característico de la ciencia, tanto de la pura como de la aplicada: donde no hay método científico no hay ciencia”, [Bunge \(2004\)](#). “Una ciencia es una disciplina que utiliza el método científico con la finalidad de hallar estructuras generales (leyes).” [Montero \(1997\)](#).

Dado que la utilización del método científico es la piedra angular alrededor de la cual se articula el concepto de ciencia, a continuación se aborda la cuestión de si, en base a lo anterior, la ciencia de datos es o no una ciencia. No obstante, previamente, se abordará la cuestión de qué se entiende por ciencia de datos”

1.2. ¿Qué es la ciencia de datos?

La Ciencia de datos es una disciplina emergente donde, a diferencia de otros saberes como, por ejemplo, las ciencias matemáticas, el corpus o la acumulación de conocimiento se ha generado en un lapso de tiempo relativamente corto (y de una forma muy intensa), y no a lo largo de siglos de historia. Su inicio data de la década de 1970, aunque ya el término **análisis de datos**, acuñado por J. Tukey en 1962 en su artículo *The Future of Data Analysis* ([Tukey, 1962](#)) se puede considerar como un precedente del término **ciencia de datos**. En dicho artículo, Tukey definió, por primera vez, el análisis de datos como: “procedimientos para analizar datos, técnicas para interpretar los resultados de dichos procedimientos, formas de planificar la recopilación de datos para hacer su análisis más fácil, más preciso o acertado, y toda la maquinaria y los resultados de las estadísticas matemáticas que se aplican al análisis de datos” ([Tukey, 1962](#)). A partir de este momento, toda una serie de acontecimientos fueron consolidando el término **ciencia de datos** como una nueva disciplina. Una breve descripción de los acontecimientos se muestran en la Fig. (1.1).

La ciencia de datos implica la limpieza, la agregación y la manipulación de datos recabados de la web, de teléfonos inteligentes, de clientes, de pacientes, de sensores o de encuestas, entre otras fuentes, para llevar a cabo un análisis de datos avanzado de los mismos, así como su modelización, para ayudar a detectar patrones, tendencias, comportamientos y, por tanto, facilitar la

1.2. ¿Qué es la ciencia de datos?

27

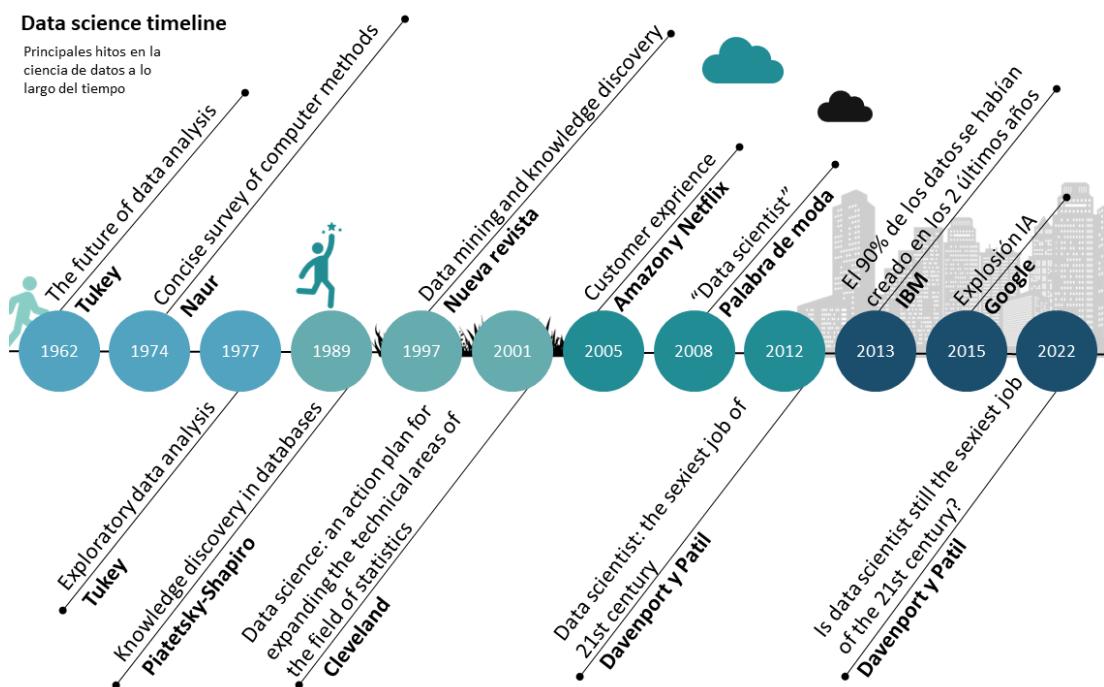


Figura 1.1: Línea del tiempo de la ciencia de datos. Elaboración propia

toma de decisiones. El crecimiento acelerado del volumen de fuentes de datos y, posteriormente, de datos, ha hecho que la ciencia de datos sea uno de los campos de más rápido desarrollo en todas las industrias. Como resultado, no sorprende que surgiera la nueva profesión del **científico de datos**, para ayudar a comprender y a analizar los volúmenes masivos de datos que se acumulaban en ese momento, trabajo que fue calificado como el “trabajo más sexy del siglo XXI” por [Davenport and Patil \(2012\)](#).

La ciencia de datos es, por tanto, una disciplina relativamente nueva que combina la estadística, las matemáticas, la informática y la programación, para obtener valor de los datos.

Se utiliza en una amplia variedad de campos, como la astronomía, la medicina, la economía, el marketing, las finanzas, la biología, la industria, etc. Esta naturaleza transdisciplinaria de la ciencia de datos añade cierta complejidad a su caracterización pues, como se ha avanzado, siendo una única disciplina, subsume en su ejercicio otras disciplinas como las ciencias matemáticas y la estadística y la ciencia de la computación, que a su vez son aplicadas a un amplio rango de dominios de manera integral. La ciencia de datos se sirve de los métodos formales de las matemática y de las aplicaciones prácticas e ingenieriles de las ciencias de la computación para la generación de conocimiento y para la resolución de problemas prácticos en múltiples campos. Esta ubicuidad la sitúa, transversalmente, entre los saberes de primer orden. En otras palabras, la ciencia de datos va adoptando los paradigmas, modelos, teorías o constructos propios del campo sustantivo en el que se ejerce, de forma que para resolver alguna problemática sobre personas, puede recurrir al corpus relativo de la psicología o de la sociología, y para profundizar sobre alguna condición de salud, puede hacer lo propio con la medicina o la biología, por mencionar algunos ejemplos.

1.3. Lo científico de la ciencia de datos

En la Sec. 1.1 se manifestó que un aspecto fundamental de la ciencia es que utiliza el método científico con la finalidad de hallar estructuras generales (principios y leyes) con capacidad predictiva y comprobable (en el sentido amplio del término). Es por ello que el marco general de la **metodología científica** ha sido bien fundamentado a lo largo de las últimas décadas gracias a las contribuciones de diferentes teóricos de la ciencia ([Chalmers et al. \(2000\)](#), [Bunge \(2004\)](#), [Díez and Moulines \(2008\)](#)). Por otra parte, las ciencias se clasifican, según el objeto de estudio ([Bunge, 2018](#)), en:² empíricas y formales. Dado que la ciencia de datos subsume diferentes disciplinas y se aplica a diferentes campos, puede tener características tanto de las ciencias empíricas como de las formales.

Si se analiza el conjunto de saberes científicos se aprecia que tienen en común una serie de características ([Bunge, 2018](#)). Por tanto, la pregunta fundamental en este punto es: ¿comparte la ciencia de datos estas características? De ser satisfechas, conferirían a la ciencia de datos el estatuto de ciencia que comparten otros saberes científicos:

- (1) La actividad científica es **metódica**. Es decir, utiliza un método, se caracteriza por proceder de manera ordenada y planificada. Esta estructuración le otorga solidez y consistencia.

²Las ciencias empíricas emplean una aproximación hipotético-inductivo-deductiva y amplían el conocimiento. Las ciencias formales se fundamentan en la deducción y explicitan el saber, pero no lo amplían. Ejemplo de ciencia empírica es la biología, mientras que la lógica lo es de la ciencia formal.

1.3. Lo científico de la ciencia de datos

29

En ciencia de datos también se actúa de manera metódica, a través de diferentes metodologías, como *Knowledge Discovery in Databases* (KDD), *Sample, Explore, Modify, Model, Assess* (SEMMA) y *CRoss-Industry Standard Process for Data Mining* (CRISP-DM), tal y como se expone en el Cap. 2.

- (II) El conocimiento científico **se fundamenta en hechos**. En general, los científicos disponen de diferentes instrumentos para observar y registrar la realidad sobre la que conjeturan. Esta labor también la realizan los científicos de datos, quienes cuentan con un elevado número de instrumentos y metodologías para la recolección de datos. Tal es el caso de los cuestionarios, escalas psicométricas o datos transaccionales producidos por diferentes tecnologías.
- (III) El saber científico es implica que las afirmaciones científicas puedan ser **contrastadas** a través de los hechos. En ciencia de datos, esto también sucede, ya que, estadísticamente, los resultados a los que se llega no están ligados a la subjetividad del analista, sino a la objetividad de los datos y a las técnicas estadísticas de contrastación.
- (IV) La ciencia es una actividad que **trasciende los hechos**. Es decir, la ciencia parte de evidencias empíricas que tienden a ser superadas, puesto que la explotación de las mismas suele generar nuevas evidencias que, a su vez, pueden contribuir a crear nuevos marcos teóricos explicativos o a ampliar los existentes. La ciencia de datos puede ejercerse en el mismo sentido. Por ejemplo, la construcción de un recomendador, como Netflix, parte de ciertos datos, pero su uso genera nuevos *inputs* comportamentales que pueden ser empleados para optimizar su sustrato algorítmico.
- (V) La investigación científica se caracteriza también por ser una **actividad analítica**. Es decir, tiende a descomponer los problemas en sus partes constitutivas. Cabe observar que la consecuencia de ello es que no se pueda hablar de una ciencia general, sino de especializaciones. Naturalmente, la especialización también existe en esta disciplina; por eso, cuando la ciencia de datos se aplica intensivamente en recursos humanos, por ejemplo, es posible hablar de *Human Resource Analytics*. Lo mismo ocurre en Economía, con el *Business Analytics*, y así en un sinfín de disciplinas.
- (VI) La ciencia es **comunicable** y, para ello, se sirve de sistemas representacionales lógico-formales. Este atributo también se aprecia en la ciencia de datos, puesto que los resultados tienden a ser compartidos a través de diferentes estrategias, entre ellas, la visualización de datos.

La ciencia, sin embargo, no sólo puede describirse mediante sus características constitutivas, sino también **funcionalmente** (Hempel, 2005). De hecho, las características anteriormente citadas son las que posibilitan las funciones **descriptiva**, **explicativa** y **predictiva**.

- (I) La primera, la **descriptiva**, permite recabar información sobre el suceso que se analiza para tratar de conocerlo en mayor profundidad y detalle. En ciencia de datos, usualmente, una de las primeras tareas consiste en describir el conjunto de datos para conocer en detalle sus características, es decir, el número de variables, el número de observaciones, los valores nulos, etc. Esta tarea se conoce como “comprensión de los datos” en la metodología CRIPS-DM (véase Sec. 2.3).

- (II) La segunda, la **explicativa**, determina cómo se relacionan los fenómenos que se observan. En general, cuando un científico de datos emplea un modelo de regresión lineal, lo que hace es establecer una relación explicativa entre la variable dependiente y las independientes. Esta parte se conoce como “modelado” en la metodología CRIPS-DM (véase Sec. 2.3).
- (III) La tercera, la **predictiva**, permite anticipar ciertos eventos en el tiempo o en el espacio. Tal es el caso de los científicos de datos que ejercen su labor en el ámbito comercial y emplean, por ejemplo, el análisis de series temporales para pronosticar las ventas futuras y poder realizar una planificación del aprovisionamiento de existencias con mayor eficiencia. Esta parte está incluida en la fase de “validación” en la metodología CRIPS-DM (véase Sec. 2.3).

A la luz de lo expuesto hasta aquí, se puede sostener, sin lugar a dudas, que la ciencia de datos emplea el método científico y comparte las principales funciones de la ciencia. Ahora bien, la ciencia de datos no puede entenderse plenamente sin presuponer las disciplinas en las que se aplica. Por tanto, uno de los interrogantes que deberán resolver los futuros profesionales es si la ciencia de datos es un saber de primer orden, que lida directamente con la realidad, como la física o la química, o si, por el contrario, es un saber de segundo orden, es decir, una suerte de disciplina que se sirve de otros saberes para desplegarlos y actualizarlos.

Resumen

- Para determinar si la ciencia de datos es, realmente, una ciencia en primer lugar se debe consensuar la definición de ciencia, que va íntimamente ligada a la definición de método científico.
- Las ciencias tienen en común una serie de características, que deben ser satisfechas por la ciencia de datos para adquirir el estatus de ciencia.
- Dado que la ciencia de datos emplea el método científico y comparte las principales funciones de la ciencia, se concluye que la ciencia de datos es una ciencia.

Capítulo 2

Metodología en ciencia de datos

Gema Fernández-Avilés^a y Ramón A. Carrasco^b

^aUniversidad de Castilla-La Mancha

^bUniversidad Complutense de Madrid

2.1. Preliminares

En el Cap. 1 se puso de manifiesto que el método científico es el elemento que define la ciencia. Bunge (2018), al hablar del método científico, lo define como: “un procedimiento para tratar un conjunto de problemas. Cada clase de problemas requiere un conjunto de métodos o técnicas especiales. Los problemas del conocimiento, a diferencia de los del lenguaje o los de la acción, requieren la invención o la aplicación de procedimientos especiales adecuados para los varios estadios del tratamiento de los problemas...”. De acuerdo con su concepción del **método**, Bunge (2018) destaca ocho operaciones en la aplicación de este:

- (I) Enunciar preguntas bien formuladas y verosímilmente fecundas.
- (II) Arbitrar conjeturas, fundadas y contrastables con la experiencia, para contestar las preguntas.
- (III) Derivar consecuencias lógicas de las conjeturas.
- (IV) Arbitrar técnicas para someter las conjeturas a contraste.
- (V) Someter a contraste esas técnicas para comprobar su relevancia y la validez que merecen.
- (VI) Llevar a cabo la contrastación e interpretar sus resultados.
- (VII) Estimar la pretensión de verdad de las conjeturas y la fidelidad de las técnicas.

- (viii) Determinar los dominios en los cuales valen las conjeturas y las técnicas, y formular los nuevos problemas originados por la investigación.

A su vez, sugiere una serie de reglas para la ejecución ordenada de las operaciones anteriores:

- (i) Formular el problema con precisión y, al principio, específicamente.
- (ii) Proponer conjeturas bien definidas y fundadas de algún modo, y no suposiciones que no comprometan cuestiones concretas ni tampoco ocurrencias sin fundamento visible.
- (iii) Someter las hipótesis a contrastación dura, no laxa.
- (iv) No declarar verdadera una hipótesis satisfactoriamente confirmada; considerarla, en el mejor de los casos, como parcialmente verdadera.
- (v) Preguntarse por qué la respuesta es como es y no de otra manera.

Sin embargo, de acuerdo con [Montero \(1997\)](#), estas reglas no son definitivas ni infalibles y necesitan de ulterior perfeccionamiento, que se llevará a cabo a lo largo de la investigación científica. Además, las reglas del método científico no son autosuficientes, necesitan apoyarse en la inteligencia y creatividad humanas.

En resumen, es el tratamiento sistemático de los problemas, de la forma descrita, y no la certeza de los resultados obtenidos o la utilización de las técnicas muy concretas y específicas, el que garantiza el carácter científico de las conclusiones [Cancelo \(1997\)](#). La ciencia de los datos, como no podía ser de otra forma, proporciona una serie de metodologías que guían el trabajo de los científicos de datos. Las principales metodologías se presentan a continuación.

2.2. Principales metodologías en ciencia de datos

En un proyecto de ciencia de datos es muy importante la metodología, pues proporciona al científico de datos una estrategia y un marco con el que trabajar. Desde finales del siglo XX se han ido proponiendo diversas metodologías, centradas en la resolución de problemas concretos mediante el uso de los datos, que hoy podrían englobarse bajo el paraguas común de la ciencia de datos.

Estas metodologías han nacido y se han desarrollado en el ámbito de los problemas de negocio, aunque todas son extrapolables a otros ámbitos de conocimiento (educación, ciencia, salud, etc.). Por tanto, en este capítulo (y, en general, en todo el manual) el término de “negocio” (empleado en las propias metodologías frecuentemente) debe de ser entendido en sentido amplio, abarcando los diversos ámbitos del conocimiento en los que se aplica la ciencia de datos.

Por su amplio uso, destacan tres metodologías:

- (i) Obtención de conocimiento en bases de datos (Knowledge Discovery in Databases-**KDD**), propuesta por [Fayyad et al. \(1996\)](#) e inspirada en un trabajo previo de [Brachman and](#)

2.3. CRISP-DM para ciencia de datos

33

[Anand \(1994\)](#), fue la primera metodología aceptada por la comunidad científica. Se trata del primer intento serio de sistematizar el proceso conocido hoy día como ciencia de datos y en aquellos tiempos como conocimiento basado en bases de datos, pues se centraba en la minería de datos.

- (II) **SEMMA**, acrónimo que coincide con las etapas de las que consta (en inglés, *Sample, Explore, Modify, Model and Assess*) fue desarrollada y mantenida por el Instituto SAS en 2012. Se define como el proceso de selección, exploración y modelización de grandes bases de datos para descubrir patrones de negocio desconocidos.
- (III) **CRISP-DM**, acrónimo en inglés de *Cross Industry Standard Process for Data Mining*, planteada inicialmente en 1996, publicada formalmente en [Chapman et al. \(2000a\)](#) y mantenida durante varios años por la compañía SPSS, posteriormente adquirida por IBM, que se ha encargado de mantenerla y refinarla hasta la actualidad. Esta metodología define una secuencia flexible de seis fases que permiten la construcción e implementación de un modelo de minería de datos para ser utilizado en un entorno real, que contribuya a respaldar la toma de decisiones de negocio. Se considera la metodología más utilizada en la actualidad ([Azevedo and Santos \(2008\)](#) y [Shafique and Qaiser \(2014\)](#), entre otros) y se describe en la siguiente sección.

2.3. CRISP-DM para ciencia de datos

La metodología CRISP-DM consta de seis etapas, que no han variado desde su publicación en 2000 (Fig. 2.1) y una serie de funciones que se han sido refinando en el tiempo ([CRISP-DM, 2021](#)). De manera esquemática, dichas etapas son:

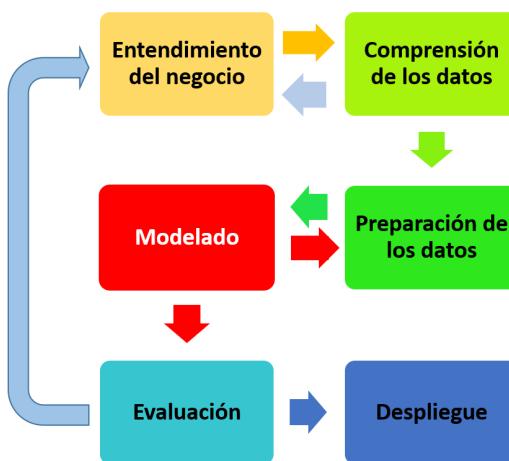


Figura 2.1: Etapas de la metodología CRISP-DM.

1. **Entendimiento del negocio.** Fundamental para el éxito del mismo. Consta de cuatro fases:

- *Determinación de los objetivos de negocio*, consensuados previamente con la organización. Es importante fijar los key performance indicators (KPIs) que permitan medir fidedignamente el grado de consecución de dichos objetivos.
- *Evaluación de la situación actual*. Inventariar las fuentes de datos que estarán disponibles, los recursos materiales y humanos con los que se podrá contar, los factores de riesgo y el plan de contingencia para los mismos.
- *Determinación de los objetivos del proyecto*, que debe alinearse con el correspondiente rendimiento de los modelos (por ejemplo, cuál debe de ser su nivel de precisión).
- *Plan del proyecto*, con los procesos a realizar y recursos asignados.

2. **Comprensión de los datos.** Consta de cuatro fases que giran en torno a los datos:

- *Recopilación*, tanto de datos internos como externos a la organización. Esta fase incluye, si es necesario, la obtención de datos adicionales, y el etiquetado de casos no clasificados con anterioridad.
- *Descripción*, especificando aspectos como la cantidad de datos disponibles, anticipando posibles problemas de rendimiento en el modelado posterior, tipología de las variables (numéricas, categóricas, booleanas, etc.), codificación de las mismas (especialmente para las categóricas), etc.
- *Exploración*, a través del análisis exploratorio de datos (AED). Esta tarea ayuda a formular hipótesis sobre los datos y dirige las posteriores etapas de preparación y modelado.
- *Verificación de la calidad*, detectando problemas como la existencia de valores perdidos, errores en datos (por ejemplo, tipográficos), errores de las mediciones (datos que son correctos pero que están expresados en unidades de medida incorrectas), incoherencias en la codificación (especialmente en las variables categóricas).

3. **Preparación de los datos.** Esta etapa suele ser la que requiere más tiempo y esfuerzo del proyecto (frecuentemente más del 70 %). Consta de cinco fases:

- *Selección*: se toman decisiones sobre los casos o filas que hay que seleccionar y sobre los atributos (variables) o columnas que hay que incluir.
- *Limpieza*: si en la subfase de verificación de la calidad de los datos se han detectado problemas, hay que subsanarlos. Los valores perdidos se pueden excluir o interpolar; los errores en los datos se pueden corregir con algún esquema lógico o manualmente; si hubiera incoherencias en la codificación se podría llevar a cabo una recodificación que sustituyese a la codificación original.
- *Construcción*: a partir de los ya disponibles, de nuevos atributos (variables) o columnas y de nuevas filas o registros.
- *Integración*: necesaria para construir un concepto de negocio unificado (por ejemplo, el concepto de cliente) si se han usado diversas fuentes (tiquet de compra y registros de cliente). La fusión de columnas con algunas claves en común (*join*), adición de filas con las columnas en común (*union*), la agrupación, etc., se utilizan frecuentemente.

2.3. CRISP-DM para ciencia de datos

35

- *Formateo*: orientada a las necesidades de los modelos que se usarán posteriormente. La conversión de variables categóricas a numéricas (usando técnicas de *one hot encoding*) o viceversa, la normalización (usando normalizaciones *min-max* o *z-score*), etc., son tareas comunes en esta etapa.
- (iv) **Modelado**: se trata de que los modelos ingieran dichos datos y aprendan de ellos, de forma automática, cómo resolver el problema de negocio planteado mediante técnicas, especialmente de *machine learning*. Las subfases de las que consta esta fase son:
- *Selección de técnicas de modelado*, si se va a usar *machine learning* supervisado o no supervisado y, específicamente, el tipo de algoritmos a usar en cada una de estas técnicas. Por supuesto, se tienen en cuenta los requisitos fijados en la primera fase, la cantidad y tipo de datos de los que se dispone, los requisitos concretos de cada modelo, etc.
 - *Generación de un diseño de comprobación*, a través de medidas y criterios de bondad del modelo: el área bajo la curva ROC, el criterio de información de Akaike (AIC), el coeficiente de determinación lineal (R^2), la matriz de confusión, etc.
 - *Generación de modelos*, que se entrena oportunamente para seleccionar, posteriormente, el más adecuado.
 - *Validación del modelo*, en función de los modelos generados y del plan de pruebas especificado.
- (v) **Evaluación**. Se debe comprobar que el modelo final generado cumple las expectativas de negocio especificadas en la primera fase. Hay que hacer hincapié en este aspecto ya que suele confundir en la práctica esta fase de evaluación con la subfase de la anterior etapa de validación del modelo. Ahora la evaluación se lleva a cabo desde el punto de vista del negocio. Así, por ejemplo, cabe plantearse si con el modelo elegido se pueden alcanzar las metas de negocio especificadas y medidas con los correspondientes KPIs. Tras esta evaluación de los resultados del modelo se abre un proceso de revisión que permitirá valorar si cumple las expectativas o se tiene que volver a etapas anteriores.
- (vi) **Implementación**. El conocimiento obtenido con el modelado es puesto en valor en esta fase de cara a satisfacer los objetivos de negocio planteados en el proyecto. Este despliegue depende mucho del tipo de proyecto que se esté realizando, aunque generalmente incluye las actividades siguientes:
- *Planificación del despliegue*: del modelado y/o del conocimiento obtenido.
 - *Planificación del control y del mantenimiento*. Así, por ejemplo, hay que verificar que el modelo está cumpliendo con las expectativas para las que se ha desarrollado, comprobar si hay que reentrenarlo o sustituirlo por otro, etc.
 - *Creación del informe final*: para comunicar los resultados del proyecto y los pasos siguientes.
 - *Revisión final del proyecto*: donde se establecen las conclusiones finales y se formalizan las lecciones aprendidas para incorporarlas a futuros proyectos de ciencia de datos.

Para concluir, subrayar que, aunque son varias las metodologías propuestas, CRISP-DM es la más completa, la más desarrollada y, además, puede ser implementada, como todas las propuestas en la literatura, mediante el lenguaje **R**.

Resumen

- El método científico es el elemento clave en la definición de ciencia.
- [Bunge \(2018\)](#) establece una serie de reglas y características para la correcta aplicación de la metodología científica. En un proyecto de ciencia de datos es muy importante la metodología, pues proporciona al científico de datos una estrategia y un marco en el que trabajar. Entre ellas destaca el CRISP-DM como la más aceptada y utilizada por las empresas y científicos.
- El CRISP-DM se basa en la organización flexible de seis pilares: entendimiento del negocio, compresión de los datos, preparación de los datos, modelado, evaluación e implementación.

Capítulo 3

R para ciencia de datos

Emilio L. Cano

Universidad Rey Juan Carlos

3.1. Introducción

El análisis estadístico de datos es una tarea fundamental en la transformación digital de las empresas y organizaciones. Siempre ha estado ahí, pero en la actualidad la disponibilidad de datos, la cantidad de los mismos, y la velocidad con la que se requieren resultados, está haciendo necesario el capacitar a los profesionales para su análisis con nuevas herramientas. Nuevas tendencias (muchas veces malinterpretadas) como Inteligencia Artificial, *Big Data*, Industria 4.0, *Internet of Things* (IoT), o *Data Science*, aumentan el interés por parte de las empresas, los profesionales y los investigadores en estas técnicas.

El tratamiento de datos y su análisis requiere el uso de software avanzado. Aunque algunas tareas como, por ejemplo, mecanizar y almacenar datos, se pueden realizar eficazmente con programas de hoja de cálculo como Excel, se debería utilizar software especializado para el análisis de datos. Existen distintos paquetes estadísticos comerciales, como SPSS, Statgraphics, Stata, SAS, JMP o Minitab. En los últimos años se ha abierto camino como alternativa el [software estadístico y lenguaje de programación R](#) ([R Core Team, 2021](#)). Hay otras alternativas que, en su mayoría, o son parciales, referidas a un ámbito concreto, o son más lenguajes de programación que software estadístico, como Python. **R** es software libre, pero su gratuidad sólo es una de sus ventajas, como se verá a lo largo del libro. Su gran inconveniente es la curva de aprendizaje: no es tan fácil de aprender y usar como un software de ventanas, ya que el uso de **R** se basa en expresiones que hay que ejecutar desde *scripts* (archivos de código).

R es un sistema para **computación estadística**: software de **análisis de datos y lenguaje de programación**. Ha sido ampliamente utilizado en investigación y docencia, y actualmente también en las empresas y organismos públicos. Es la evolución del trabajo de los laboratorios Bell con el lenguaje S ([Venables and Ripley, 2002](#)), llevado al mundo del software libre por Ross

Ihaka y Robert Gentleman en los años 90 ([Ihaka and Gentleman, 1996](#)). La versión R 1.0.0 se publicó el 29 de febrero de 2000.

Uno de los aspectos más espectaculares de **R** es la cantidad de **paquetes** disponibles. Un paquete (*package*) de **R** es un componente con funcionalidad adicional que se puede instalar en el sistema para ser utilizado por **R**. En el momento de compilar este libro, el número de paquetes disponibles en el repositorio oficial es de 19721.

Una vez conocido el mundo de **R**, se plantea la siguiente pregunta: ¿y por qué utilizar **R**? Es imposible dar un único motivo. A continuación se enumeran algunos de ellos:

- Es Free and Open Source Software (FOSS). Gratis y libre. En inglés se suele decir *free as in free beer, and free as in free speech*.
- Tiene una amplia comunidad de usuarios que proporciona recursos.
- Es multiplataforma.
- Se usa cada vez en más empresas e instituciones.
- Es posible obtener soporte comercial, por ejemplo a través de Posit Software PBC¹.
- Se ha alcanzado una masa crítica de usuarios que lo hace confiable.
- Es extensible (desde pequeñas funciones, hasta paquetes).
- Se puede implementar la innovación inmediatamente. En software comercial hay que esperar a nuevas versiones, en el mejor de los casos.
- Posee características de “investigación reproducible”. En el Cap. 44 se tratará qué implica este enfoque. En contextos distintos a la investigación, se puede hablar de informes reproducibles y trazabilidad del análisis.

Por otra parte, el uso de **R** en las empresas está **creciendo exponencialmente** debido, principalmente, a la necesidad de analizar y visualizar datos con herramientas potentes para explotar todo su potencial. Grandes empresas de todos los sectores llevan tiempo utilizándolo, si bien la popularización del software y su conocimiento entre los nuevos titulados está facilitando que empresas de todo tipo y tamaño aprovechen esta herramienta en su estrategia digital. Así, además de la **visualización y presentación efectiva** de los datos, equipos bien formados pueden descubrir relaciones entre variables clave, realizar **predicciones**, tomar mejores decisiones o **mejorar sus procesos** gracias al análisis avanzado de datos más allá de la hoja de cálculo.

3.2. La sesión de R

R es una aplicación de análisis estadístico y representación gráfica de datos, y además un lenguaje de programación. **R** es **interactivo**, en el sentido de que responde a través de un “intérprete” a las **entradas** que recibe a través de la **consola**.

La interfaz de usuario de **R** (R GUI, *Graphical User Interface*) cumple las funciones básicas para interactuar con **R**, pero es muy pobre a la hora de trabajar con ella. En su lugar, es más conveniente utilizar el entorno de desarrollo [RStudio Desktop](#) (o su versión en la nube

¹<https://posit.co>, antes RStudio PBC.

3.3. Instalación de **R**

39

<https://posit.cloud/>), que es como un “envoltorio” del sistema **R** con más funcionalidades y ayudas, pero manteniendo el mismo nivel de interacción: consola y *scripts*². Al igual que **R**, RStudio es una aplicación de software libre, pero, en este caso, desarrollada y mantenida por la compañía privada Posit PBC.

Una cosa muy importante en **R** es que las expresiones son **sensibles a mayúsculas**, y por tanto los objetos **datos** y **Datos** son distintos.

3.3. Instalación de R

Durante todo el libro se utiliza la interfaz RStudio. Pero, como se avanzó anteriormente, RStudio es solo un “envoltorio” de **R**, por lo que previamente hay que tener instalado en el ordenador el sistema “base” de **R**. **R** está disponible para sistemas Windows, MacOS y Linux. Por cuestiones de espacio, no se incluyen detalles en este libro, pero la instalación es sencilla siguiendo las instrucciones en sus correspondientes websites:

1. Instalación de **R**: <http://www.r-project.org>
2. Instalación de RStudio: <https://posit.co>

Para completar la instalación de **R**, se muestra cómo instalar³ los paquetes del **tidyverse**⁴ mediante expresiones en la consola o *script* con la función `install.packages()`:

```
install.packages(pkgs = "tidyverse")
```

Una vez instalado el paquete, se cargará con la instrucción `library("nombre_paquete")` en la sesión de **R** donde se quiera utilizar.

```
library("tidyverse")
```

A veces resulta útil usar directamente la función que se va a utilizar en vez de cargar todo el paquete. Esto se hace con el operador `::`, es decir, `nombre_paquete::funcion()`. La siguiente expresión serviría para usar la función `select()` del paquete `dplyr` sin cargar el paquete entero.

```
dplyr::select()
```

²Lo importante es seguir un estilo consistente en cuanto a nombres de objetos, espacios en blanco y el uso de delimitadores y tabulación en el *script*. Véase por ejemplo la [guía de estilo de Hadley Wickam \(Wickham, 2015\)](#).

³Una vez instalado un paquete no hay que volver a instalarlo.

⁴El **tidyverse** es un conjunto de paquetes que se irán describiendo a medida que se utilicen, especialmente en la Sec. 3.6.

3.4. Trabajar con proyectos de RStudio

La manera más eficiente de trabajar con **R**, es mediante **proyectos** de RStudio. Esto permite abstraerse de los detalles de la sesión de **R** (espacio de trabajo, directorio de trabajo, *environment*), ya que al abrir un proyecto estará todo preparado para seguir el trabajo donde se dejó, o empezar de cero si se acaba de crear. Para crear un proyecto de RStudio, se despliega el menú de proyectos a la derecha en la barra de herramientas y se selecciona “New Project...”. También se puede hacer en el menú “File/New Project...”.

Es aconsejable crear siempre una estructura de carpetas que permita tener todo organizado desde el principio, porque al final los proyectos crecen. La estructura perfecta no existe, y depende del proyecto particular. Las siguientes carpetas pueden ser útiles en un amplio abanico de proyectos, y las tres primeras se pueden usar prácticamente en cualquier proyecto:

- **data**: en esta carpeta se tienen los archivos de datos, tanto aquellos orígenes de datos que se quieran importar, como los que se puedan guardar desde un *script*.
- **R**: para los *scripts*. Es posible que solamente haya un *script* en nuestro proyecto, pero si hubiera más se pueden guardar en esta carpeta.
- **inform**: aquí se pueden guardar los archivos Quarto o R Markdown que se utilicen para generar informes o presentaciones.
- **img**: si en nuestro proyecto se utilizan imágenes de cualquier tipo, es una buena idea tenerlas en una carpeta aparte.
- **test**: si se quieren separar los *scripts* que se utilicen para pruebas y no se quieren mezclar con los “buenos” en la carpeta **R**.
- **aux, tmp, util, notas, doc, ...**: este tipo de carpetas vienen bien cuando hay información que está relacionada o es útil para un proyecto, pero el archivo no es del proyecto de análisis de datos en sí. Por ejemplo, unas especificaciones de un producto o servicio, un artículo científico, fotografías de una fábrica, comunicaciones con clientes, etc.
- **ejercicios, practicas, ...**: si nuestro proyecto forma parte de una asignatura, curso, o similar.

Un aspecto importante cuando se trabaja en proyectos colaborativos es el control de versiones. Este tema se aborda en el Cap. 46.

3.5. Tratamiento de datos con R

En este apartado se van a empezar a utilizar expresiones de **R**. Las expresiones se escribirán en *scripts*, que pueden contener “comentarios” (texto que no se ejecutará) utilizando el símbolo “almohadilla” (#). Muchas de las expresiones que se usan son llamadas a funciones⁵. La ayuda de cualquier función se puede obtener en la consola usando la expresión `?function`, donde `function` es el nombre de la función u objeto del que se quiere obtener ayuda.

⁵Por motivos de espacio, no se incluyen mayores explicaciones de las mismas, pero se anima al lector a explorar la ayuda de cada una de ellas para comprender mejor su funcionamiento.

3.5.1. Estructuras y tipos de datos

Las estructuras y tipos de datos más frecuentes con las que se trabaja en **R** son las que se detallan a continuación.

Tablas de datos. Son colecciones de variables numéricas y/o atributos organizadas en columnas, en las que cada fila se corresponde con algún elemento en el que se han observado las características que representan las variables. La forma más común es el **data.frame**. Cada columna del **data.frame** es, en realidad, otra estructura de datos, en concreto, un **vector**. Un ejemplo de **data.frame** es el conjunto de datos **tempmin_data** del paquete **CDR** que se analiza en el Cap. 40 y del que se muestran a continuación las primeras tres filas con la función **head()**.

```
library("CDR")
head(tempmin_data, 3)
#>      fecha indicativo tmin longitud latitud
#> 1 2021-01-06      4358X -4.7 -5.880556 38.95556
#> 2 2021-01-06      4220X -7.0 -4.616389 39.08861
#> 3 2021-01-06      6106X  4.7 -4.748333 37.02944
```

Un **data.frame** es un objeto de datos en dos dimensiones, en el que las filas son la dimensión 1 y las columnas la dimensión 2. Los datos se pueden “extraer” de un **data.frame** por filas, por columnas o por celdas. Para extraer una de las variables del **data.frame** se suele utilizar el operador **\$** después del nombre del **data.frame**, y a continuación el nombre de la variable.

El operador **<-** asigna al “símbolo” que hay a su izquierda el resultado de la expresión que hay a su derecha, y lo guarda con ese nombre en el espacio de trabajo⁶. Por ejemplo, la siguiente expresión extrae todas las filas de la columna **tmin** o, dicho de otra forma, el vector con todas las temperaturas mínimas registradas y lo guarda en el objeto **temp_min**.

```
temp_min <- tempmin_data$tmin
```

Vectores y matrices. Ya se ha visto que una columna de una tabla de datos es un vector. También se pueden crear vectores con la función **c()** y los elementos del vector separados por comas. Una matriz es un vector organizado en filas y columnas. A modo de ejemplo, la primera de las siguientes expresiones crea un vector llamado **nombres** con dos cadenas de texto, y la segunda crea una matriz numérica llamada **coordenadas** a partir de las columnas 4 y 5 del conjunto de datos **tempmin_data**. Nótese que la extracción de valores de un conjunto de datos o de una matriz se puede realizar también por sus índices de filas y columnas entre corchetes separados por una coma. En este caso se extraen todas las filas (pues no se especifica ninguna en la dimensión 1) de las columnas 4 y 5.

```
nombres <- c("longitud", "latitud")
coordenadas <- as.matrix(tempmin_data[, 4:5])
```

⁶También se puede utilizar el símbolo igual (**=**) para realizar asignaciones. No obstante, en el marco de este libro se recomienda el uso del operador específico **<-**.

Factor. Es un tipo especial de vector para representar variables categóricas (también denominadas atributos o factores). En general, una variable categórica suele tomar un número reducido de valores diferentes (categorías), identificados con etiquetas (*labels*) y que se llaman **niveles** del factor (*levels*). Un ejemplo es el dataset `dp_entr` del paquete CDR que se analiza en el Cap. 24. La columna `ind_pro11` es un indicador que toma los valores S y N, mientras que `des_nivel_edu` toma tres posibles valores.

```
dp_entr[1:5, c(1, 17)]
#>      ind_pro11 des_nivel_edu
#> 1          S      MEDIO
#> 497        N      MEDIO
#> 265        N      BASICO
#> 534        N      MEDIO
#> 415        N      BASICO
levels(dp_entr$des_nivel_edu)
#> [1] "ALTO"   "BASICO" "MEDIO"
```

Listas. Son estructuras de datos que contienen una colección de elementos indexados que, además, pueden tener un nombre. Pueden ser heterogéneas, en el sentido de que cada elemento de la lista puede ser de cualquier tipo.

A modo de ejemplo, se muestran los nombres del objeto `tempmax_data` del paquete CDR, que contiene 6 elementos de distintas clases.

```
names(tempmax_data)
#> [1] "ESP"           "ESP_utm"        "grd_sf"        "grd_sp"
#> [5] "temp_max_utm_sf" "temp_max_utm_sp"
```

Fechas. Son un tipo de datos especial que algunas veces provoca problemas al compartir datos entre programas. El conjunto de datos `tempmin_data` contiene la columna `fecha`, que puede convertirse de manera inmediata a tipo fecha (`Date`) porque viene en un formato estándar (véase la ayuda de `strptime` para especificar otros formatos). El paquete `lubridate` del *tidyverse* contiene funciones para hacer más fácil el trabajo con fechas.

```
tempmin_data$fecha<- as.Date(tempmin_data$fecha)
class(tempmin_data$fecha)
#> [1] "Date"
```

Cadenas de texto. Son estructuras de datos que aparecen en forma de vector de caracteres. La columna `indicativo` del conjunto de datos `tempmin_data` es un ejemplo de este tipo de datos. La ayuda de `?regexp` proporciona la información necesaria sobre cómo extraer texto con expresiones regulares, y la de `?paste` para aprender a unir cadenas de texto. El paquete `stringr` del *tidyverse* contiene funciones para facilitar el trabajo con cadenas de texto.

```
head(tempmin_data$indicativo)
#> [1] "4358X" "4220X" "6106X" "9698U" "4410X" "1331A"
```

3.5.2. Importación de datos

En el apartado anterior se han utilizado tablas de datos que están incluidas en un paquete de **R**. Pero lo habitual es que los datos se tengan que importar de fuentes externas, como ficheros. A continuación, se describen algunas de las formas de importar los tipos de ficheros más habituales⁷.

Excel. Sin duda una forma muy popular de organizar los datos en ficheros es mediante **hojas de cálculo como Microsoft Excel**. Hay varios paquetes con los que se puede trabajar con archivos de Excel. En este libro se utiliza el paquete `readxl` del *tidyverse*. Con la siguiente expresión se puede descargar un archivo Excel de ejemplo⁸.

```
download.file(url = "http://emilio.lcano.com/b/adr/p/datos/RRHH.xlsx",
              destfile = "data/RRHH.xlsx",
              mode = "wb")
```

Una vez el archivo está en el directorio de trabajo de la sesión de **R**, se puede importar su contenido al espacio de trabajo con la siguiente expresión:

```
rrhh <- readxl::read_excel("data/RRHH.xlsx")
```

Texto. Los **archivos de texto** son el formato más utilizado y conveniente para compartir datos. Es también muy común que el equipamiento o el software genere datos en formato de texto. Estos archivos suelen tener extensión `.csv` (*comma separated values*) o `.txt`, aunque pueden tener cualquier otra, o incluso no tener extensión. A modo de ejemplo, con la siguiente expresión se puede descargar un archivo csv.

```
download.file(url = "http://emilio.lcano.com/b/adr/p/datos/ejDatos.csv",
              destfile = "data/ejDatos.csv")
```

Si el archivo tiene extensión `.csv`, como el anterior, vendrá ya con una especificación muy concreta, pudiéndose usar directamente las funciones `read.csv()` o `read.csv2()` para tener la tabla de datos en el espacio de trabajo.

```
merma <- read.csv2("data/ejDatos.csv")
```

⁷Para poder reproducir los ejemplos, se debe tener una carpeta `data` en el directorio de trabajo.

⁸La función `download.file()` permite descargar cualquier archivo disponible en la `url` que se indique. Es obligatorio indicar el archivo de destino con el argumento `destfile`. Si el archivo no es de texto plano, se debe indicar `mode = "wb"`.

La función genérica de **R** para importar datos de texto es `read.table()`, que puede importar cualquier especificación cambiando los argumentos adecuados. Por ejemplo, la siguiente expresión tendría el mismo resultado que se ha obtenido con la función `read.csv2`⁹:

```
merma <- read.table(file = "data/ejDatos.csv",
                     header = TRUE,
                     sep = ";",
                     dec = ",",
                     fileEncoding = "utf-8")
```

Para saber cómo importar datos desde sistemas gestores de bases de datos véase el Cap. 5.

Hay infinidad de otras fuentes de las que se pueden importar datos a **R**. Por ejemplo, el paquete `rvest`, que forma parte del *tidyverse*, se puede utilizar para obtener datos de páginas web y otras fuentes de Internet, lo que se suele llamar *web scraping*. Por ejemplo, supóngase que se quiere importar la tabla con los datos de comunidades y ciudades autónomas españolas del enlace https://www.ine.es/daco/daco42/codmun/cod_ccaa_provincia.htm. Las siguientes expresiones importan esta tabla al conjunto de datos `ccaa_ine`.

```
library("rvest")
url <- "https://www.ine.es/daco/daco42/codmun/cod_ccaa_provincia.htm"
ccaa_ine <- url |>
  read_html() |>
  html_node(xpath = '//*[@id="contieneHtml"]/table') |>
  html_table(fill = TRUE)
```

La ruta o “xpath” se puede obtener usando las herramientas de desarrollo del navegador, y puede que una vez importada la tabla se requiera algún post-procesamiento antes de poder analizar los datos.

3.5.3. Exportación de datos y archivos de datos específicos de R

En algunos proyectos es necesario guardar algunos datos que se han ido creando o transformando, bien para compartir con otras partes interesadas, bien para ser utilizados en el mismo u otros proyectos. Para exportar los datos a Excel, se utiliza la función `write.xlsx()` del paquete `openxlsx` (si no está instalado, se instala de la forma habitual). Si lo que se quiere es exportarlo a texto, se pueden utilizar los equivalentes a las funciones de importación `write.csv()`, `write.csv2()` o `write.table()`.

La siguiente expresión exporta la tabla de datos `tempmin_data` a ficheros Excel y csv (formato en inglés).

⁹Con el argumento `header` se indica si la primera fila tiene encabezados (TRUE) o no (FALSE, opción por defecto). También hay que especificar el separador de columnas `sep` y el símbolo decimal `dec`. `fileEncoding` es la especificación de la codificación de texto; las más habituales son `utf-8` y `'latin1'`.

```
openxlsx::write.xlsx(x = tempmin_data,
                      file = "data/temp_min_Filomena.csv")
write.csv(x = tempmin_data, file = "data/temp_min_Filomena.csv")
```

También se pueden guardar los datos en formato “nativo” de **R**. Los archivos **.RData** almacenan un espacio de trabajo entero, y por tanto pueden guardar varios objetos en el mismo archivo. Cuando posteriormente se importe, los objetos estarán en el espacio de trabajo con su nombre original. Se guardan con la función **save()** y se restauran con la función **load()**, como en el siguiente ejemplo.

```
save(tempmin_data, tempmax_data,
      file = "data/datos_temperaturas.RData")
load("data/datos_temperaturas.RData") #carga de nuevo el objeto
```

Los archivos **.rds** almacenan un único objeto en un archivo. Cuando posteriormente se quieran importar, hay que asignar el resultado al nombre que se quiera. Se guardan con la función **writeRDS()** y se restauran con la función **readRDS()**, como en el siguiente ejemplo.

```
saveRDS(object = tempmin_data,
          file = "data/datos_temperaturas.rds")
nuevo_objeto <- readRDS(file = "data/datos_temperaturas.rds")
```

El paquete **foreign** de **R** base y otros paquetes especializados pueden exportar datos a otros formatos de archivo, que no se tratan en detalle en este capítulo.

3.6. Organización de datos con el *tidyverse*

3.6.1. El *tidyverse* y su flujo de trabajo

El *tidyverse* es, según se define en su propia [página web¹⁰](#), un conjunto de paquetes de **R** “opinables” diseñados para ciencia de datos. Las principales ventajas (opinables) de utilizar el *tidyverse* son tres:

1. Utiliza una gramática, estructuras de datos y filosofía de diseño común.
2. El flujo de trabajo es más fluido y, una vez se comprenden las ideas principales, más intuitivo.
3. Para la mayoría de las operaciones, es computacionalmente más eficiente.

¹⁰“... an opinionated collection of R packages designed for data science”. Incluye actualmente 30 paquetes, véase la lista con **tidyverse::tidyverse_packages(include_self = TRUE)** y la ayuda de cada paquete para saber más. Los que se vayan usando en el libro se irán explicando oportunamente.

Uno de los paquetes más populares del *tidyverse* es **ggplot2**, que proporciona una “gramática de gráficos” (Wickham, 2016) y es una pieza clave del *tidyverse* actual, junto con los paquetes **dplyr** (gramática para la manipulación de datos) y **tidyverse** (herramienta para crear datos *tidy*). El flujo de trabajo propuesto por el *tidyverse* se describe en el libro “R for Data Science” (Wickham and Grolemund, 2016) y se sintetiza en la Fig. 3.1.

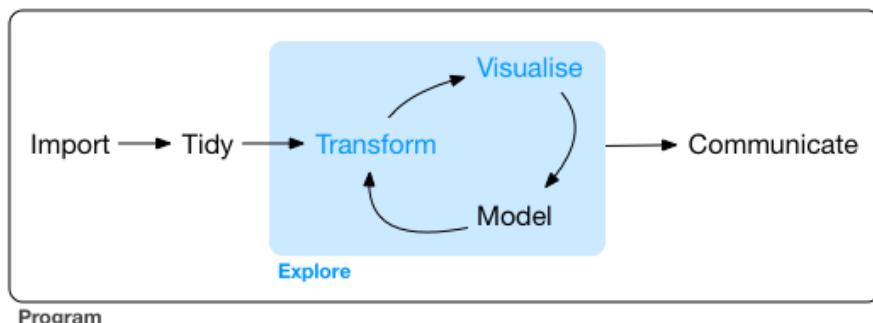


Figura 3.1: Flujo de trabajo en Ciencia de Datos propuesto por el *tidyverse* (fuente: Wickham, H. and Grolemund, G. (2016))

Además del mencionado libro, la web del *tidyverse* (<http://tidyverse.org>) contiene toda la documentación de los paquetes, incluidos artículos para tareas concretas, que merece la pena leer alguna vez. En la web están también las conocidas como *cheatsheets*, algunas de ellas disponibles también en la ayuda de RStudio (menú *Help/Cheatsheets*).

Dentro del flujo de trabajo de la Fig. 3.1, ya se ha tratado la primera etapa (*Import*) en la Sec. 3.5.2. Es importante señalar que, al utilizar las funciones del *tidyverse*, los datos se organizan en objetos de clase **tibble**, que es una extensión del **data.frame** de R base. Las principales diferencias son:

- Permite una representación compacta en la consola al mostrar la tabla de datos.
- La selección con corchetes simples de una única variable siempre devuelve otro **tibble** (a diferencia de un **data.frame**, que devuelve un vector).

Se puede forzar a que una tabla de datos sea de un tipo u otro con las funciones **as.data.frame** (de **tibble** a **data.frame**) y **as_tibble** (de **data.frame** a **tibble**).

Siguiendo con el esquema de la Fig. 3.1, en este apartado se verán algunas tareas de las etapas *Tidy* (organizar) y *Transform* (transformar), que serán ampliadas en los Cap. 8 y 9. La visualización (*Visualise*) se tratará específicamente en el Cap. 11 y transversalmente en muchos otros. La modelización (*Model*) se trata extensamente en los capítulos de las partes IV a IX, y la comunicación (*Communicate*) se verá en los capítulos de la Parte X. Una de las características de la forma en que están programados los paquetes del *tidyverse* es que se puede trabajar¹¹ con *pipes*.

¹¹Existe una guía de estilo del *tidyverse*, que se puede consultar en <https://style.tidyverse.org>. Hay incluso una serie de *Addins* en RStudio para comprobar y aplicar esta guía de estilo a través del paquete **styler**. Los *Addins* son menús adicionales en RStudio para usar la funcionalidad de algunos paquetes de forma interactiva.

3.6. Organización de datos con el tidyverse

47

El *pipe* es, básicamente, un operador compuesto de dos caracteres, `|>`, que se puede obtener con el atajo de teclado **CTRL+MAYUS+M**. El operador se pone en medio de dos expresiones de **R**. Sean `lado_izquierdo` y `lado_derecho` las expresiones que se ponen a izquierda y derecha del *pipe*. Entonces se utiliza de la siguiente manera:

```
lado_izquierdo |> lado_derecho
```

El operador *nativo* de **R**, `|>`, apareció en la versión R-4.1.0. Hay un operador alternativo que proviene del paquete `magrittr`, `%>%`, que había que usar antes de esta versión, y mucha literatura y documentación está escrita usándolo. Hay diferencias, pero a los efectos de este capítulo ambos operadores se pueden utilizar indistintamente.

La expresión `lado_izquierdo` debe producir un valor, que puede ser cualquier objeto de **R**. La expresión `lado_derecho` debe ser una función, que tomará como primer argumento el valor producido en la parte izquierda. Si se desea guardar el resultado final, se debe asignar el resultado a algún nombre de objeto para que se almacene en el espacio de trabajo. La siguiente expresión sería un ejemplo de uso.

```
nombre_objeto <- lado_izquierdo |>
  lado_derecho
```

La ventaja de usar los *pipes* es que se pueden encadenar, de forma que el resultado de cada operación pasa a la siguiente expresión del *pipeline* (secuencia de operaciones con *pipe*), como en el siguiente ejemplo:

```
library("dplyr")
contam_mad |> colnames() |> length()
#> [1] 12
```

3.6.2. Transformación de datos con dplyr

En la gramática del *tidyverse*, dentro del paquete `dplyr` se dispone de una serie de “verbos” (funciones) para una sola tabla, que se pueden agrupar en tres categorías: para trabajar con filas, para trabajar con columnas y para resumir datos.

3.6.2.1. Operaciones con filas

Los verbos definidos para estas operaciones son:

- `filter()`: elige filas en función de los valores de la columna.

```
pm10 <- contam_mad |>
  filter(nom_abv == "PM10")  # se filtra por PM10
```

- **arrange()**: cambia el orden de las filas con algún criterio.

```
zonas<- contam_mad |>
  arrange(desc(zona), daily_mean)
```

- **slice()**: extrae filas por su índice. También hay una serie de funciones “asistentes” (*helpers*) para obtener los índices que se utilizan con frecuencia. Por ejemplo:

- **slice_head()** y **slice_tail()** obtienen las primeras y últimas filas respectivamente (por defecto, una). Se puede especificar **n** (número) o **prop** (proporción) de filas.
- **slice_sample()** obtiene una muestra aleatoria de **n** filas (o proporción **prop**).
- **slice_min()**, **slice_max()** obtienen las filas que contienen los menores o mayores valores respectivamente de la variable indicada en el argumento **order_by**. Si no se especifica **n** o **prop**, se obtienen sólo las filas que contienen el mínimo o el máximo. Nótese que puede haber más de una fila que cumpla la condición.

Véase el resultado de los siguientes ejemplos:

```
pm10 |> slice(10:15) # extrae filas desde la 10 a la 15
pm10 |> slice_tail(n = 3) # extrae las tres últimas filas
pm10 |> slice_max(order_by = daily_mean) # día con mayor valor medio de PM10
set.seed(1) # Para que la muestra aleatoria sea reproducible
pm10 |> slice_sample(n = 4) # muestra 4 registros
```

3.6.2.2. Operaciones con columnas

Los verbos definidos para estas operaciones son:

- **select()**: indica cuando una columna se incluye o no. Se pueden utilizar *helpers* para seleccionar columnas que cumplan cierta condición (por ejemplo, ser numéricas) y también para “quitar” columnas de la selección (con el signo menos (-)).

```
pm10 |> select(longitud, latitud, daily_mean, tipo)
pm10 |> select(where(is.numeric))
pm10 |> select(-c(id:latitud))
```

En cuanto a la **modificación** de datos, existen múltiples posibilidades. Algunas de ellas son:

- **rename()**: cambia el nombre de la columna.

3.6. Organización de datos con el tidyverse

49

- **mutate()**: cambia los valores de las columnas y crea nuevas columnas. La función **transmute()** funciona igual que **mutate()**, pero la tabla de datos resultante sólo contiene las nuevas columnas creadas.
- **relocate()**: cambia el orden de las columnas.

```
pm10 |> rename(zona_calidad_aire = zona)
pm10 |> relocate(fecha, .before = estaciones)
pm10_na <- pm10 |> mutate(isna = is.na(daily_mean))
```

En este punto, es importante señalar que dentro de la función **mutate()** se puede usar cualquier función vectorizada para transformar las variables. Por ejemplo, se podría transformar una columna con las funciones **as.xxx** que se vieron en la Sec. 3.5.1, aplicar formatos a fechas o usar funciones del paquete **lubridate** para trabajar con este tipo de datos. A medida que se avance en el libro irán apareciendo aplicaciones que ahora, quizás, no sean tan evidentes.

3.6.2.3. Operaciones de resumen y agrupación

La primera operación de resumen que puede surgir es “contar” filas. La función **tally()** devuelve el número de filas totales de un **data.frame**. La función **count()** proporciona también este número; si, además, se pasa como argumento alguna variable, lo que devuelve es el número de filas para cada valor diferente de dicha/s variable/s. Estos recuentos se pueden añadir a la tabla de datos con las funciones **add_count()** y **add_tally()**, lo que permite calcular frecuencias absolutas y relativas fácilmente.

```
pm10 |> tally()
#>      n
#> 1 53794
pm10 |> count(zona)
#>      zona      n
#> 1: Interior M30 20690
#> 2: Noreste 12414
#> 3: Noroeste 4138
#> 4: Sureste 8276
#> 5: Suroeste 8276
```

La función **summarise()** (o, equivalentemente, **summarize()**) aplica alguna función de resumen a la/s variable/s que se especifiquen (**mean()**, **max()**, etc.). El paquete **dplyr** tiene algunas funciones de resumen adicionales, como **n()** (número de filas), **n_distinct()** (número de filas con valores distintos) y **first()**, **last()**, **nth()** (primer, último y *n*-ésimo valor, en el orden en el que se encuentran, respectivamente).

En muchas ocasiones, las operaciones de análisis se realizan en grupos definidos por alguna variable de agrupación. La función **group_by()** “prepara” la tabla de datos para realizar operaciones de este tipo. Una vez agrupados los datos, se pueden añadir operaciones de resumen

como las vistas anteriormente. A veces hay que “desagrupar” los datos, para lo que se utiliza la función `ungroup()`.

A continuación, se muestra una expresión un poco más compleja que las anteriores. En el conjunto de datos `contam_mad` del paquete CDR, se filtra por el nombre de contaminante “NOx”. Después se agrupan los datos por zona y se calculan algunos estadísticos resumen para cada zona.

```
contam_mad |>
  filter(nom_abv == "NOx") |> # se filtra por NOx
  group_by(zona) |>
  summarize(
    min = min(daily_mean, na.rm = TRUE),
    q1 = quantile(daily_mean, 0.25, na.rm = TRUE),
    median = median(daily_mean, na.rm = TRUE),
    mean = mean(daily_mean, na.rm = TRUE),
    q3 = quantile(daily_mean, 0.75, na.rm = TRUE),
    max = max(daily_mean, na.rm = TRUE)
  )
#> A tibble: 5 × 7
  zona      min     q1 median   mean     q3    max
  <chr>    <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
#> 1 Interior  M30  0.0833  32.4  54.1  72.9  90.0  759.
#> 2 Noreste    1    23.8   39.6  56.2  68.9  516.
#> 3 Noroeste    0    12.0   20.3  29.7  34.5  352.
#> 4 Sureste    0    29.1   45.4  64.6  77.2  453
#> 5 Suroeste   0.667  33.5  59.6  90.5 114.   666.
```

3.6.3. Combinación de datos

En el apartado anterior se han tratado los “verbos” de una tabla. Es muy común que haya que combinar datos de distintas tablas, para lo cual se utilizan lo que el *tidyverse* considera *two tables verbs*. En esencia, para combinar tablas que contienen información relacionada, hay que saber cuáles son las columnas que se refieren a lo mismo, para hacer las uniones (*joins*) utilizando esas columnas. Hay cuatro tipos de uniones que se pueden realizar, usando las siguientes funciones:

- `inner_join()`: se incluyen las filas de ambas tablas para las que coinciden las variables de unión.
- `left_join()`: se incluyen todas las filas de la primera tabla y sólo las de la segunda donde hay coincidencias.
- `right_join()`: se incluyen todas las filas de la segunda tabla y sólo las de la primera donde hay coincidencias.
- `full_join()`: se incluyen todas las filas de las dos tablas.

Las funciones requieren como argumentos dos tablas de datos y la especificación de las columnas coincidentes. Si no se especifica, hace las uniones por todas las columnas coincidentes en ambas

3.6. Organización de datos con el tidyverse

51

tablas. Para las filas que sólo están en una de las tablas, se añaden valores `NA` donde no haya coincidencias.

A modo de ejemplo, las siguientes expresiones unen dos datasets para combinar datos de municipios con su renta. En el Cap. 8 se verán estas uniones en la práctica.

```
library("sf")
munis_renta <- municipios |>
  left_join(renta_municipio_data) |>
  select(name, cpro, cmun, `2019`)
#> Joining, by = "codigo_ine"
```

Otra forma de unir tablas es, simplemente, añadiendo columnas (que tengan el mismo número de filas) o filas (que tengan el mismo número de columnas). Para ello se usan las funciones `bind_cols()` y `bind_rows()`, respectivamente. Otra forma conveniente de añadir nuevas filas o columnas son las funciones `add_row()` y `add_column()`. Se pueden añadir antes o después de una fila/columna especificada con el argumento `.before`, y pasando los valores como pares “variable = valor” para cada variable en el conjunto de datos.

Como comentario final del paquete `dplyr`, una característica importante es que se pueden usar las funciones vistas sobre tablas de una base de datos, sin necesidad de utilizar sentencias SQL y con la ventaja de que las operaciones se realizan en el motor de la base de datos. En el Cap. 5 se tratarán las cuestiones relacionadas con los gestores de bases de datos y SQL.

3.6.4. Reorganización de datos

A lo largo del capítulo se ha visto la importancia de disponer los datos de forma rectangular, de forma que se tenga una columna para cada variable y una fila para cada observación. Algunas veces es conveniente reorganizar los datos más “a lo ancho” o más “a lo largo” de lo que se encuentran.

Para estas operaciones se utilizan las funciones `pivot_longer()` y `pivot_wider()` del paquete `tidyverse` de la siguiente forma:

- `pivot_longer()`: el argumento `names_to` asigna el nombre de la nueva variable que va a indicar de qué columna vienen los datos; y el argumento `values_to` asigna el nombre de la nueva variable que va a contener el valor de la tabla original.
- `pivot_wider()`: el argumento `names_from` indica el nombre de la variable que contiene los nombres de las nuevas columnas a crear a lo ancho; y el argumento `values_from` indica el nombre de la variable que contiene los valores en la tabla original. Las observaciones deben estar identificadas de forma única por varias variables. Si no es el caso, se puede aplicar una función al estilo de las tablas dinámicas de las hojas de cálculo con el argumento `values_fn`.

Las funciones `pivot_longer()` y `pivot_wider()` admiten otros argumentos `names_xx` y `values_xx` para personalizar la forma de reestructurar los datos. En la mayoría de las ocasiones será suficiente con las comentadas (`xx_from` y `xx_to`). Si fuera necesario, se recomienda consultar la ayuda de las funciones, o la lectura del artículo sobre *pivoting*.

A modo de ejemplo, el conjunto de datos `contam_mad` tiene los datos “mezclados” de varias variables medioambientales en la columna `daily_mean`. La columna `nom_abv` contiene el parámetro al que se refiere la columna de datos. Entonces, interesa “extender” la tabla para tener cada parámetro en una columna, de forma que se pueda hacer un análisis de datos adecuado, como en el siguiente código:

```
library("tidyverse")
extendida <- contam_mad |>
  pivot_wider(names_from = "nom_abv",
              values_from = "daily_mean",
              values_fn = mean)
colnames(extendida)
#> [1] "estaciones"      "id"           "id_name"       "longitud"
#> [5] "latitud"         "nom_mag"       "ud_med"        "fecha"
#> [9] "zona"            "tipo"          "BEN"           "SO2"
#> [13] "NO2"             "EBE"           "CO"            "NO"
#> [17] "PM10"            "PM2.5"         "TOL"           "NOx"
```

Se deja como ejercicio volver a obtener la tabla original usando la función `pivot_longer()` a partir del objeto `extendida`.

El paquete `tidyverse` también contiene funciones para reorganizar las columnas de la tabla uniendo columnas con la función `unite()`, o separando una columna en dos o más con la función `separate()` (véanse los detalles en la ayuda de las funciones).

Para terminar este apartado de reorganización de datos, se da una primera aproximación al tratamiento de valores perdidos, que se tratará en el Cap. 8. En R, un valor perdido se representa por el valor especial `NA` (*not available*). Brevemente, las funciones más utilizadas en este campo son:

- `drop_na()` del paquete `tidyverse`: permite eliminar las filas que tienen valores perdidos en ciertas variables (o en cualquiera, si no se especifica ninguna).
- `replace_na()`: sustituye los valores perdidos en cada variable por el valor especificado.
- `fill()`: permite “rellenar” valores perdidos con los últimos encontrados.

Los datos de contaminación a menudo tienen muchos valores perdidos. La siguiente expresión elimina las filas del conjunto de datos `contam_mad` con valores perdidos y, después, cuenta las filas.

```
contam_mad |>
  drop_na() |> # se omiten los NAs para el análisis
  count()
#>      n
#> 1: 505773
```

Resumen

- R es software libre y gratuito, mantenido por una enorme comunidad.
- La forma de interactuar con R es mediante expresiones, que se escriben en *scripts*, y al ejecutarlas se obtienen los resultados.
- Los objetos de datos que se vayan a usar deben estar en el espacio de trabajo.
- RStudio es un “envoltorio” de R, y por tanto R tiene que estar instalado en el sistema para poder usar RStudio.
- Los paquetes se instalan una sola vez, y deben cargarse con `library()` para usar sus funciones.
- La tabla de datos o `data.frame` es la estructura de datos más adecuada para análisis de datos y cada columna es un `vector`.
- El *tidyverse* es un conjunto de paquetes que facilita las tareas de análisis de datos.
- El operador *pipe*, `|>`, permite “pasar” valores a funciones de forma encadenada.
- Las operaciones básicas con una tabla son filtrado, selección y resumen.
- Para crear nuevas columnas en las tablas de datos se usa la función `mutate`.
- Para combinar tablas con columnas comunes se usan las funciones `xx_join`.

Capítulo 4

Ética en la ciencia de datos

Mónica Villas^a y Bilal Laouah^b

^aOdiseIA ^bAlexandria Business Solutions Based on Data

4.1. ¿Qué es la ética?

La ética es una subdisciplina de la Filosofía que estudia de manera sistemática el comportamiento humano desde las nociones del bien y del mal y en relación con la moral. Ambas disciplinas están muy relacionadas, pero son diferentes. La moral tiene un carácter normativo y prescriptivo: orienta las acciones de acuerdo con algún marco de valores específico (costumbres, creencias, códigos tradicionales, normas no escritas, etc.). Por el contrario, la ética está por encima de cualquier orientación particular, es decir, no se basa en ningún código de mandatos y prohibiciones concreto, sino que pretende establecer los principios a partir de los cuales evaluar las acciones y decisiones. La ética es Filosofía práctica, por eso no transmite juicios, sino que enseña a juzgar.

Las dos preguntas fundamentales que ha tratado de responder la ética a lo largo del tiempo son: ¿qué debemos hacer? y ¿qué es valioso en la vida? La primera pregunta promueve el razonamiento ético, mientras que la segunda sirve para establecer el marco de valores desde el cual juzga y razона el sujeto ético. Desde una perspectiva histórica, Aristóteles es considerado el primer autor occidental en haber sistematizado la ética. Su tratado es comúnmente conocido bajo el título de *Ética a Nicómaco*.

Este capítulo se centra en la ética aplicada, que es la utilización de la ética en la práctica. Algunos ejemplos de ética aplicada son la ética profesional (o deontología), la bioética o la ética medioambiental. La ética aplicada a la ciencia de datos hace referencia a la reflexión que debe acompañar a la toma de decisiones en el contexto de la praxis profesional de los científicos de datos; se puede considerar, por tanto, una concreción de la ética profesional. Y es que los científicos de datos tienen que tomar decisiones a lo largo del ciclo de vida de un proyecto de datos que pueden tener consecuencias sobre las personas. Algunos procesos que pueden ser fuente de

dilemas éticos son: la recopilación de datos, su transformación, la definición de objetivos que se persiguen, el uso de algoritmos y la explicación de los resultados. Estos pasos se pueden ver de manera detallada en el Cap.3. En todos estos pasos, el científico de datos debe usar su pensamiento crítico y tomar decisiones éticas. Así, por ejemplo, si el propósito es automatizar algún proceso, deberá reflexionar y anticipar los posibles impactos negativos que pueden derivarse ya que, si este proceso no se realiza adecuadamente, la toma de decisiones automáticas puede perpetuar algunos de los problemas éticos como son los sesgos. Para ser más específicos: supóngase que se quieren automatizar las contrataciones laborales. Para ello, el científico de datos necesita desarrollar un algoritmo que seleccione a los mejores profesionales para su compañía. Pues bien, algunas cuestiones que debe valorar son las siguientes: primero tiene que entender qué significa “los mejores profesionales” y definir los atributos que los representan. Después, tiene que buscar datos históricos de la compañía, recopilar éstos y estar seguro de que esos datos cumplen con la normativa de privacidad establecida, especialmente si la compañía reside en Europa. Aquí, el científico de datos debe pensar en temas como la procedencia de los datos, ¿cuál es la fuente?, ¿a quién pertenecen los datos?, ¿están los datos anonimizados para que no se pueda identificar a una persona de manera unívoca?, y algunas preguntas similares referidas a la privacidad. Seguidamente, tendrá que asegurarse de que se tiene una muestra de casos cuyos atributos (edad, profesión, experiencia, raza, género, procedencia geográfica, etc.) no tienen sesgos; es decir, que, por ejemplo, el porcentaje de personas de una determinada raza, o edad, o sexo, etc. no es significativamente distinto del que hay en la población de la cual se tomó la muestra. En definitiva, debe asegurarse de que la muestra que tiene es suficientemente representativa de la población con la que va a trabajar y, si no es así, tenerlo en cuenta a la hora de analizar y comunicar los resultados. Además, ha de tener cuidado con los datos personales, como género, edad, raza, etc., dado que, en algunos casos de uso, la toma de decisiones no debería tener en cuenta estos atributos porque podrían inducir a prácticas discriminantes, alejadas de los estándares éticos. Como se puede ver en este sencillo ejemplo, el científico de datos tiene que tomar decisiones, no sólo técnicas, que influyen en el resultado de su trabajo y que pueden afectar a otras personas. Generalmente, los científicos de datos suelen ser profesionales que provienen del mundo técnico, de carreras tecnológicas o relacionadas con las Matemáticas y, a diferencia de otros itinerarios de corte humanista, la presencia de la ética es menos frecuente, razón por la cual conviene fomentar la sensibilización respecto a estas cuestiones.

Mientras que para los profesionales de la salud existen códigos deontológicos bien establecidos y organismos que regulan la práctica de acuerdo a los mejores estándares comportamentales, no existe una guía común para el científico de datos en que se describa cómo debe comportarse. A pesar de todo, la guía de buenas prácticas que publica la asociación ACM (*Association for Computing Machinery*) puede servir de inspiración, si bien, al tratarse de meras recomendaciones, sigue siendo insuficiente para orientar éticamente el comportamiento de éstos.

4.2. Los principios éticos

Un principio no es ni más ni menos que aquello que permite preservar los derechos y libertades de las personas, sin frenar la innovación tecnológica (Olmeda and Ibáñez, 2022). La mayoría de los principios se pueden agrupar en cuatro grandes categorías: **autonomía, justicia, evitar daños y generar beneficios**. Algunos ejemplos de principios que se pueden clasificar en alguna de

4.2. Los principios éticos

57

estas categorías son: **transparencia, explicabilidad, privacidad, accesibilidad o equidad**. Aunque no hay aún un acuerdo a nivel mundial sobre cuáles deberían ser los principios claves de la inteligencia artificial (IA), sí que se están desarrollando proyectos supranacionales como el de la UNESCO¹, que ha sido firmado recientemente por todos sus miembros.

Desde principios de 2010, el crecimiento de la ciencia de datos ha sido exponencial y ha comenzado a usarse en todas las industrias de manera sistemática, entre otras cosas gracias al Big Data. Actualmente, se dispone de más datos que nunca y sólo se analiza un 5 % de ellos. Además, se han producido enormes mejoras en la computación con el surgimiento de nuevos procesadores y también han ocurrido grandes cambios en el área de la algoritmia, teniendo disponibles muchos más algoritmos que nunca, lo que facilita su reutilización. Por ello, la demanda de científicos de datos que conviertan dichos datos en información clave para las empresas ha crecido enormemente en los últimos años.

Asimismo, desde 2016, distintos organismos, asociaciones, empresas y gobiernos han publicado numerosos documentos, donde se resalta la importancia de la necesidad de principios éticos para la ciencia de datos. Google, IBM y Amazon, en el ámbito de las empresas privadas, publicaron sus principios éticos en el 2018. También son muy conocidos los principios de Asilomar de 2016 o la declaración de Toronto de 2017. La mayoría de estos documentos están desarrollados por perfiles multidisciplinares: científicos de datos, abogados o expertos en ética, que resaltan la importancia de tener en cuenta los principios éticos en la toma de decisiones automáticas cuando se utiliza la ciencia de datos.

En definitiva, las cuestiones éticas se están incorporando poco a poco en los proyectos de ciencia de datos en todo el mundo, siendo la regulación europea publicada en abril de 2021 un ejemplo a seguir. Esta regulación, diseñada a lo largo de tres años, parte de un primer documento en 2018² que fue liderado por un grupo de expertos de todos los países miembros: HLEGAI (*High Level Expert Group Artificial Intelligence*). A partir de este primer documento se publicaron otros incluyendo los comentarios y mejoras sugeridas por la sociedad civil, instituciones públicas, empresas e instituciones académicas, hasta, finalmente, publicarse, en abril de 2021, el actual documento de regulación de IA.

En el actual documento regulatorio, se eligió un enfoque basado en riesgos:

- **Riesgo inaceptable**, como el uso de aplicaciones de *social scoring* o de imágenes para procesos de administración de justicia.
- **Riesgo alto**, como el uso de aplicaciones de contratación o médicas, que deberán ser supervisadas por organismos designados antes de su publicación.
- **Riesgo medio**, como el uso de aplicaciones en las que hay que incluir las explicaciones necesarias (dependiendo del tipo de algoritmo de que se trate) para que el usuario pueda entender el proceso de toma de decisiones.
- **Riesgo bajo**, para cualquier otro tipo de aplicación.

¹(<https://en.unesco.org/artificial-intelligence/ethics>)

²<https://op.europa.eu/es/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1>

En el resto del mundo, el progreso en este tipo de regulación está siendo algo más lento, aunque países como Estados Unidos, que hasta ahora no habían puesto el foco en este tipo de regulaciones, están empezando a trabajar en ello desde finales de 2021. Por otro lado, China, conocida mundialmente por su falta de respeto a la privacidad, está empezando a dar algún paso en esta área y comenzando a cambiar su política en este sentido. Como ejemplo, en marzo de 2022, ha lanzado una regulación en la que las empresas tienen que informar mejor a los usuarios sobre sus algoritmos de recomendación. En definitiva, parece que la necesidad de la ética para proyectos de ciencia de datos está avanzando poco a poco en todas las geografías, y Europa es, por el momento, un ejemplo a seguir.

Llegados a este punto, conviene distinguir entre regulaciones legales y principios éticos. Generalmente, las regulaciones legales son coercitivas y su incumplimiento puede tener consecuencias punitivas para quienes no las ratifican e implementan. Estos principios legales, para que sean legítimos, deben fundamentarse e inspirarse en ciertos valores éticos. Ahora bien, es imposible e indeseable regular legalmente todos los aspectos del comportamiento humano, de ahí la necesidad de compartir un marco de valores. La ética permitirá al científico de datos considerar cuál es la mejor decisión cuando exista un vacío legal. El razonamiento ético implica, pues, asumir la responsabilidad de pensar de manera autónoma.

Ahora bien, dado que no hay un acuerdo a nivel mundial sobre cuáles son los principios éticos más importantes para la ciencia de datos, en este capítulo se han seleccionado la **equidad** y la **explicabilidad**, por estar entre los que más deben tener en cuenta los expertos en ciencia de datos. Además, son dos de los principios en los que se centra la regulación europea.

4.3. Equidad: la importancia de los sesgos

El sesgo se puede definir como el resultado de dar un peso desproporcionado a favor o en contra de una persona o cosa en comparación con otra, y normalmente de manera injusta. El término ‘equidad’, se utiliza precisamente para tratar de que las decisiones no estén afectadas por esos sesgos. Si se analiza la literatura al respecto, se pueden encontrar multitud de tipos de sesgos. En la ciencia de datos cuando se habla de sesgo, generalmente se hace referencia a los **sesgos algorítmicos**. Éstos, según la RAE, son “errores sistemáticos en los que se puede incurrir cuando, al hacer muestreros o ensayos, se seleccionan o favorecen unas respuestas frente a otras”.

Este sesgo algorítmico puede darse en cualquiera de los pasos que lleva a cabo el científico de datos (véase Cap. 2). En la Fig. 4.1³, donde se representan los distintos pasos a la hora de diseñar un algoritmo de ciencia de datos, se puede ver cuáles son los momentos críticos en los que, sin percibirlo, se puede caer en este tipo de sesgo. En primer lugar, un sesgo en la **adquisición de los datos**, partiendo de muestras que ya lo tengan. En este punto se encuadran, por ejemplo, los **sesgos históricos** o los **sesgos de representación**. También haber **sesgos de medida**, que son los sesgos algorítmicos derivados de la selección de las características que se eligen para la construcción del modelo. Además, se pueden presentar sesgos en el momento del despliegue, denominados **sesgos de implementación**, que suceden cuando el contexto en el que se despliega el algoritmo es diferente del contexto en que se entrenó.

³<https://www.ibm.com/blogs/research/2018/09/ai-fairness-360/>

4.3. Equidad: la importancia de los sesgos

59

El estudio detallado de estos sesgos algorítmicos está enfocado a evitar que se aumenten o perpetúen sesgos de cualquier tipo, teniendo en cuenta que los algoritmos tienen como objetivo automatizar y generalizar. Como se veía en la sección anterior, mucha de la regulación que se está desarrollando en Europa va enfocada a **mantener el principio de equidad**, es decir, a tratar de evitar los sesgos en la toma de decisiones automáticas realizadas por los algoritmos que se diseñan gracias a la ciencia de datos.

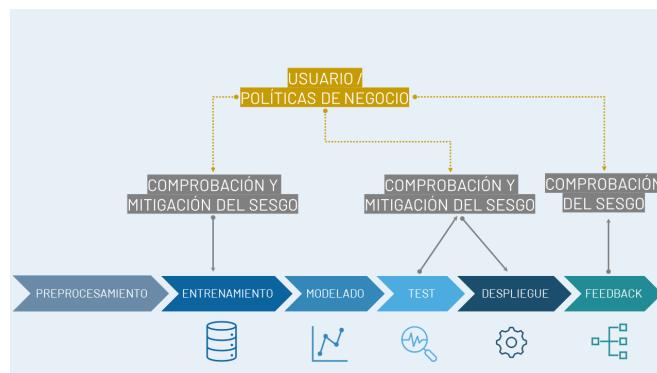


Figura 4.1: Sesgos en el proceso de machine learning. Fuente: Adaptada de IBM

Un ejemplo, que muestra la importancia de los sesgos históricos y de representación, es COMPAS (*Correctional Offender Management Profiling for Alternative Sanction*), una aplicación que da soporte al sistema de justicia americana y que utiliza un algoritmo para evaluar el riesgo potencial de reincidencia de una persona que va a ser juzgada.

COMPAS evalúa, para cada acusado, dos tipos de riesgo: **de reincidir** y **de reincidir con violencia**. El algoritmo que utiliza califica del 1 a 10 la posibilidad de que el acusado vuelva a cometer un delito (sin y con violencia). De 1 a 4, el riesgo se califica de bajo; de 5 a 7, medio; y de 8 a 10, alto. Si la persona puede ser reincidente espera a que ocurra el juicio en la cárcel, y, en caso contrario, no tiene que ir a la cárcel hasta que se celebre el juicio. Diversos estudios y organizaciones analizaron los datos y no parecía que hubiera ningún problema de sesgo inicialmente. Sin embargo, la organización PROPUBLICA, con datos de 7300 personas correspondientes a 2013 y 2014, demostró que la aplicación estaba sesgada. En concreto, demostró que los acusados negros tenían muchas más probabilidades que los acusados blancos de ser clasificados, incorrectamente, como de riesgo de reincidencia elevado, mientras que los acusados blancos tenían más probabilidades que los acusados negros de ser marcados incorrectamente como de riesgo bajo.

El proceso que se siguió fue el siguiente:

1. Partiendo del **proceso de asignación de un riesgo**, se construyó el historial delictivo del acusado.
2. Para determinar la raza, se usó la clasificación establecida, de negros, blancos, hispanos y asiáticos.
3. Se revisó la **definición de reincidencia** y cómo se establecían los riesgos en la aplicación de COMPAS.

4. únicamente se analizaron los riesgos para “reincidencia” y “reincidencia con violencia”.
5. Se analizaron los índices de reincidencia y de reincidencia con violencia en dos años, así como su distribución por raza.
6. Para contrastar la hipótesis de disparidad entre razas en el índice de riesgo, se utilizó una regresión logística que consideraba la raza, la edad, la historia criminal, la reincidencia futura, el grado de los cargos y el género.
7. Para evaluar la exactitud del algoritmo se usó una regresión de Cox.
8. Se utilizó una muestra de unos 7.300 acusados (de los que se tenía datos de 2 años) para analizar la tasa de falsos positivos y falsos negativos.

El modelo logístico concluyó que el factor más predictivo de una puntuación de **riesgo de reincidencia** más alta era la edad. Los acusados menores de 25 años tenían 2,5 veces más probabilidades de obtener una puntuación más alta que los delincuentes de mediana edad, incluso cuando en el modelo se incluía como variable de control el número de los delitos anteriores, la delincuencia futura, la raza y el género. La raza también se consideró muy predictiva de una puntuación más alta. Si bien los acusados negros tenían tasas de reincidencia más altas en general, cuando se ajustaron por esta diferencia y otros factores, tenían un 45 % más de probabilidades de obtener una puntuación más alta que los blancos. En cuanto al sexo, las mujeres tenían un 19,4 % más de probabilidades de obtener una puntuación más alta que los hombres, controlando los mismos factores. Esta conclusión resulta, cuando menos, sorprendente, dados que los niveles de criminalidad de las mujeres eran, en general, más bajos que los de los hombres.

La herramienta predecía bien el **riesgo de reincidencia** en el 60 % de los casos estudiados, pero sólo en el 20 % de ellos cuando se trataba del **riesgo de reincidentir con violencia**. La Tabla 4.1 resume las principales conclusiones obtenidas en el estudio de PROPUBLICA.

Tabla 4.1: Principales conclusiones del estudio de PROPUBLICA

Casuística en el estudio con datos de 2 años	Resultados en porcentaje
A los acusados de raza negra se les asignaba un riesgo más alto de reincidencia	Raza negra: 45 % Raza caucásica: 23 %
A los acusados de raza blanca se les asignaba un riesgo más bajo de reincidencia que a los de raza negra	28 % a los de raza blanca 48 % a los de raza negra
Mayor asignación de riesgo de reincidencia a las personas de raza negra	77 % más de riesgo de reincidir a las personas de raza negra que a las de raza blanca
Se determinó que las variables que tenían mayor importancia para la asignación de riesgo de reincidencia eran la edad, la raza y el género	< 25 años tenía 2.5 veces más de probabilidad de ser asignado un riesgo alto 45 % si eran de raza negra Casi un 20 % si la persona era mujer

En este caso, el problema del sesgo tiene como consecuencia que personas que no reincidentirán permanezcan en la cárcel al asignarseles un índice de reincidencia más alto que el que realmente

4.4. ¿Es necesaria la explicabilidad?

61

les corresponde, y que personas que sí podrían reincidir quedarían en libertad por asignarseles un índice más bajo del que realmente tienen.

Hay multitud de ejemplos publicados respecto al tema de los sesgos. Una de las mejores referencias es [O’Neil \(2016\)](#), que recopila una gran variedad de casos en la que los sesgos pueden llevar a toma de decisiones erróneas y no equitativas.

4.4. ¿Es necesaria la explicabilidad?

La explicabilidad es otro de los principios clave de la propuesta europea de IA confiable y, sin duda, va a ser clave en los próximos años en cuanto la regulación europea de IA entre en vigor.

XAI (Explainable AI) es un término que acuñó DARPA (*Defense Advanced Research Project Agency*) en el año 2017 y que agrupa dentro del término ‘explicabilidad’ no sólo el concepto de interpretabilidad para los algoritmos de *machine learning* sino también los aspectos de la Psicología que están relacionados con proporcionar explicaciones, como se puede ver en la Fig. (4.2). No se trata únicamente de entender la toma de decisión del algoritmo, sino también de dar una explicación adecuada de por qué se toma dicha decisión, en función del tipo de usuario. Si se considera, por ejemplo, de un algoritmo que selecciona imágenes cuando contienen un posible tumor, no serán las mismas explicaciones las que necesitará un científico de datos que un médico. Para el científico de datos será mucho más útil revisar las métricas propias del algoritmo (exactitud, precisión, sensibilidad, etc.) y, además, saber cuáles de los atributos de entrada del algoritmo han tenido más peso en la decisión. En cambio, al médico lo que le interesaría será una explicación menos técnica, más cualitativa, en la que se le explique con detalle, por ejemplo, por qué se seleccionó esa imagen frente a otras, mencionando el tamaño, la forma o características de la imagen, aspectos con los que están familiarizados los profesionales médicos.

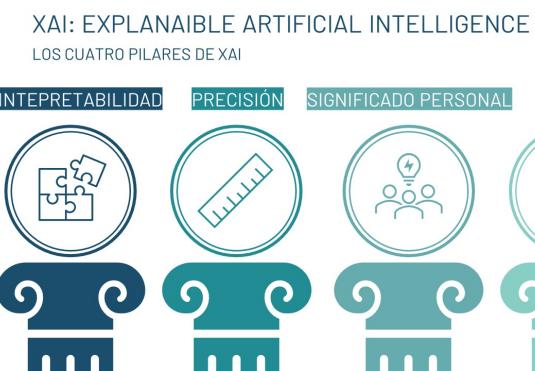


Figura 4.2: Explicabilidad según DARPA

Los algoritmos pueden clasificarse en algoritmos de **caja blanca** o **transparente** (aquellos que son fácilmente interpretables) y **opacos** o **de caja negra** (los que no son interpretables y

que requieren de herramientas adicionales para su interpretación). Normalmente, se tiene que establecer un equilibrio entre la interpretabilidad y la exactitud, dado que son métricas que mantienen una relación inversa (véase Fig. 4.3). A mayor exactitud, menor interpretabilidad, y viceversa. Los algoritmos más interpretables son normalmente los más sencillos, como los algoritmos de clasificación, regresión lineal o los árboles de decisión. Otros, como los modelos de *random forest*, *XGboost* o algoritmos de *deep learning*, son mucho más exactos pero no tan interpretables, lo cual puede llevar a ciertos problemas a la hora de usarlos en la toma de decisiones en las compañías, dado que es más difícil explicar el por qué de la decisión. Cuando las decisiones afectan a áreas clave para las personas (decisiones médicas, de contratación, de concesión de préstamos, etc.,) es cuando es más relevante es proporcionar la explicabilidad adecuada.

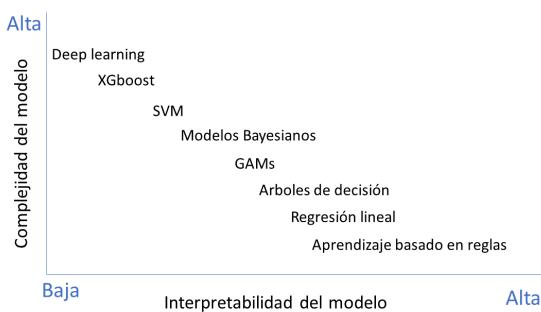


Figura 4.3: Interpretabilidad vs. exactitud

Se está avanzando muy rápido en la interpretabilidad de los algoritmos, y desde 2017 se proporcionan distintas técnicas y herramientas que ayudan a ello, como, por ejemplo, las librerías SHAP (SHapley Additive exPlanation) o LIME (Local Interpretable Model-agnostic Explanations), de código abierto. En la mayoría de las ocasiones, se trata de utilizar algoritmos más sencillos que ayuden a explicar otros más complejos como redes neuronales o XGboost.

Hay muchas taxonomías diferentes para la clasificación de los distintos tipos de algoritmos. Una de las más utilizadas clasifica los algoritmos como sigue:

- **Metodologías globales o locales:** cuando el método utiliza una instancia para la interpretabilidad se denomina local y cuando éste usa todo el modelo se denomina global.
- **Metodologías intrínsecas o post-hoc:** ‘intrínseca’ se refiere a cuando el método es interpretable por si mismo y post-hoc cuando es necesario usar otros algoritmos más sencillos para explicar los más complejos.
- **Metodologías ligadas al modelo o agnósticas del modelo:** las metodologías ligadas al modelo son aquellas que se usan para un tipo de algoritmo concreto, mientras que las metodologías agnósticas permiten trabajar con cualquier tipo de modelo.

4.5. Recursos en R para trabajar en sesgos y explicabilidad

63

Es importante elegir la técnica más adecuada dependiendo del tipo de modelo a interpretar, así como poder combinarlas en aras de conseguir una mejor interpretabilidad. Uno de los mejores libros al respecto que recopila multitud de estas técnicas es Molnar (2020).

4.5. Recursos en R para trabajar en sesgos y explicabilidad

Para un científico de datos es muy relevante conocer las herramientas, tanto *open source* como comerciales, disponibles para ser usadas en labores de análisis de sesgos o explicabilidad. Todas las herramientas en esta área son relativamente recientes. Han ido surgiendo desde 2018 y siguen evolucionando rápidamente.

En el caso de las herramientas para detectar sesgos, los proveedores que empiezan a incluir estos análisis son Microsoft, IBM, Google, Aequitas, Pymetric y Linkedin, siendo el resto *open source*. La mayoría de ellas están abiertas a contribuciones externas y todas ellas utilizan mecanismos para la detección de sesgos, aunque únicamente la de Microsoft e IBM incluyen algoritmos para su mitigación.

En lo relativo a las herramientas sobre explicabilidad, los proveedores más relevantes son Google, IBM, Oracle y H20.ai; el resto, son *open source*. La mayoría de ellas se pueden usar con algoritmos de caja blanca o negra. Respecto a los tipos de explicaciones para distintos usuarios, sólo la herramienta de IBM incluye esta funcionalidad. Se puede resaltar también la facilidad con la que H20.ai permite elegir el nivel de exactitud y explicabilidad en el momento del diseño del algoritmo. Para un mayor detalle se pueden consultar dos tablas comparativas sobre herramientas de explicabilidad y de sesgos que se incluyen en Olmeda and Ibáñez (2022).

Algunas de estas herramientas comerciales incluyen implementaciones en Python o R, de ahí la importancia de revisarlas inicialmente antes de recurrir a otro tipo de recursos.

Recursos en R para equidad

- Tutorial de fairness (2021)⁴: explica las distintas métricas usadas para medir la equidad (paridad demográfica, paridad proporcional, paridad predictiva, etc.) y permite crear la distintas métricas y visualizarlas. El tutorial emplea los datos de COMPAS.
- Librerías de R⁵ incluidas en IBM fairness360 (2020): incluye algoritmos para detectar el sesgo, pero también para mitigarlo.
- Librería EDFFair (2022)⁶: tiene una aproximación distinta, dado que permite al usuario ajustar el nivel de equidad frente al de exactitud, y así mantener el equilibrio requerido. Los detalles pueden verse en Matloff and Zhang (2022).

Recursos en R para explicabilidad:

⁴<https://cran.r-project.org/web/packages/fairness/vignettes/fairness.html><https://kozodoi.me/r/fairness/packages/2020/05/01/fairness-tutorial.html>

⁵<https://developer.ibm.com/blogs/the-aif360-team-adds-compatibility-with-r/>

⁶<https://github.com/matloff/EDFFair>

- Algunas de las herramientas más conocidas en explicabilidad que merecen mención aparte son SHAP⁷ y LIME⁸, disponibles en **R** y en Python y usadas en muchos paquetes comerciales. SHAP (2018) es uno de las librerías más usadas para la explicabilidad. Utiliza los valores de Shapley para poder explicar cualquier tipo de modelo. Los detalles se pueden encontrar en Aas et al. (2021), donde también se proporciona el código⁹. LIME (2017) es otra de las librerías más usadas para la explicabilidad. Para ello, se ajusta un modelo local alrededor de un punto concreto y lo que hace es estudiar los cambios alrededor de este modelo. Se puede encontrar una explicación detallada, junto con el código necesario¹⁰, en Ribeiro et al. (2016).
- Matloff and Zhang (2022) hacen una recopilación de 27 librerías de **R**, incluyendo LIME y SHAP; el código para cada una de ellas puede encontrarse en Github.¹¹
- DALEX(Biecek, 2018) es un paquete de **R** de reciente creación para ayudar a crear explicaciones partiendo de un modelo (el código también puede encontrarse en Github¹²).
- Esta área está evolucionando mucho en los últimos años, y están surgiendo multitud de técnicas nuevas alrededor de la explicabilidad que van a permitir entender mejor el proceso de decisión de los algoritmos más complejos.

Resumen

La ética, subdisciplina de la Filosofía, es el estudio sistemático del comportamiento humano desde las categorías del bien y del mal. Se trata de una rama aplicada que proporciona métodos basados en la racionalidad crítica para evaluar decisiones y acciones en base a ciertos valores compartidos. La aparición de las actuales metodologías que propone la ciencia de datos supone un desafío que no puede resolverse únicamente a partir de criterios técnicos, puesto que muchas de las decisiones que se toman en este campo pueden tener repercusiones sobre las personas.

Para regular ciertas prácticas, se han desarrollado diversas legislaciones en múltiples países y continentes. Estas regulaciones tienen un fundamento ético y ofrecen un marco para valorar qué acciones se ajustan a la legalidad. Ahora bien, ningún cuerpo normativo cubre todas las posibles casuísticas, razón por la cual el razonamiento ético es fundamental para orientar la praxis de los científicos de datos.

Por otro lado, no existe un acuerdo global en relación a los principios que deben regir el comportamiento del científico de datos, pero la numerosa literatura publicada desde 2016 parece estar de acuerdo en que todos ellos pueden agruparse en cuatro categorías: preservar la autonomía humana, generar beneficios, evitar daños y fomentar la justicia. Dado que no hay un acuerdo global a nivel mundial, este capítulo se centra en los principios clave que señala la regulación europea: equidad y explicabilidad.

⁷<https://shap.readthedocs.io/en/latest/>

⁸<https://homes.cs.washington.edu/~marcotcr/blog/lime/>

⁹https://cran.r-project.org/web/packages/shapr/vignettes/understanding_shapr.html

¹⁰<https://cran.r-project.org/web/packages/lime/index.html>

¹¹<https://github.com/MI2DataLab/XAI-tools/blob/master/README.md>

¹²<https://github.com/ModelOriented/DALEX>

Parte II

Bienvenidos a la jungla de datos

Capítulo 5

Gestión de bases de datos relacionales

Ismael Caballero^a, Ricardo Pérez del Castillo^a y Fernando Gualo^{a,b}

^aUniversidad de Castilla-La Mancha y ^bDQTeam SL

5.1. Introducción

El mundo real en el que estamos inmersos es puramente analógico: está lleno de **entidades** que se relacionan entre ellas o consigo mismas a través de unos determinados eventos que representan determinados **hechos**. Tanto **entidades** como **hechos** tienen una colección de características observables (normalmente llamadas **atributos**¹ en el ámbito del diseño de bases de datos), que pueden ser de interés en el contexto de una determinada aplicación.

Para estas aplicaciones que demandan el uso de datos del mundo real, es preciso realizar **observaciones** de esos **atributos** de las **entidades** y de los **hechos** que son relevantes. Al conjunto de entidades y hechos del mundo real que son relevantes para una aplicación se les conoce como **Universo del Discurso** (Piattini et al., 2006). Para poder tener éxito en las aplicaciones, es importante capturar la semántica del Universo del Discurso mediante los modelos correspondientes.

Los datos que son de interés para una determinada aplicación deben ser capturados mediante un **proceso de observación y digitalización** de los valores de los atributos relevantes de las entidades y hechos del mundo real. Durante el proceso de observación y captura se pueden producir errores que pueden derivar en problemas relacionados con la calidad de los datos (Price and Shanks, 2004). Por ejemplo, supóngase que las observaciones requieren una determinada frecuencia mínima de observación en relación con la velocidad en la producción de los hechos;

¹Es importante resaltar que, en el ámbito de la Estadística, **atributo** es una denominación reservada para las variables cualitativas.

si esta frecuencia no es adecuada, la cantidad de observaciones realizada será insuficiente para modelar el hecho, llevando a un estado inconsistente entre lo sucedido y lo observado.

Una vez capturados estos datos, pueden ser usados en los procesos de negocio para una tarea determinada, o bien ser analizados para producir un conocimiento del mundo real que hasta ahora no se tenía (Davenport and Harris, 2017).

5.2. Concepto de base de datos

Para poder habilitar el procesamiento o los análisis de forma automática mediante las potentes técnicas tratadas en el resto de capítulos de este manual, es necesario almacenar previamente los datos en algún lugar como **repositorios** o **bases de datos**, donde se puedan fácilmente añadir, borrar, recuperar o modificar los datos.

De acuerdo con Piattoni et al. (2006), una **base de datos (BD)** es una colección o depósito de datos integrados, almacenados en soporte secundario (no volátil) y con redundancia controlada. En una base de datos, los valores correspondientes a los atributos que han de ser compartidos por diferentes usuarios y aplicaciones deben mantenerse independientes de ellos, y su definición (estructura de la base de datos), única y almacenada junto con los datos, se ha de apoyar en un modelo de datos. Este modelo debe captar las interrelaciones y restricciones existentes en las entidades y hechos del mundo real al que representan. Existen diferentes tipos de modelos que permiten estructurar y representar la semántica de los datos, como por ejemplo, el modelo relacional (Codd, 1970a), que es el fundamento de las bases de datos relacionales en las que se centra este capítulo.

En el ámbito de los Sistemas de Información, se han desarrollado programas que dan soporte a todo el proceso de creación y explotación de las bases de datos. A estos programas se les conoce como **Sistemas Gestores de Bases de Datos (SGBD)**. Como ejemplos de estos SGBD se pueden citar Microsoft Access, Microsoft SQL Server, Oracle Server, MySQL, MariaDB, Infor-mix, MongoDB,... En cualquier caso, como se verá más adelante en este capítulo, los SGBD más utilizados son los conocidos como **relacionales (SGBDR)**, aunque, con el auge del *Big Data* y del *Machine Learning*, esta tendencia está cambiando y empiezan a desplegarse cada vez más SGBD conocidos como *not only structured query language*, **NoSQL** (no solo lenguaje estructurado de consulta) (véase Cap. 6). Para evitar confusiones, es importante diferenciar entre la base de datos propiamente dicha (como una colección de datos almacenada en un fichero de datos) y el software SGBD específico, ya sea relacional o NoSQL: una misma base de datos, con las correspondientes adaptaciones, puede ser gestionada usando diferentes tipos de SGBD. Habitualmente, los tipos de SGBDR más usados son los que tienen capacidades multiproceso/multiusuario, ya que permiten acceder a datos compartidos mediante el uso de interfaces de datos para ejecutar diferentes tipos de análisis, empleando lenguajes de programación más potentes -o versátiles- como **R** software o Python, que los lenguajes típicos de programación.

5.2.1. Gestión de los datos en una base o repositorio de datos

Las organizaciones usan datos para sus procesos de negocio. La Tabla 5.1 muestra la tipología de datos sugerida por Mahanti (2019).

5.2. Concepto de base de datos

69

Tabla 5.1: Tipos de datos

Tipo Datos	Amplitud Datos	Volumen	Frec Actualizació	Nro Co lumnas	Ejemplos	Propietarios	Stma Típico	Registro
Metadatos	Organización	Muy bajo	Casi estática	Varias	Definiciones de atributos	Organización	Repositorio de metadatos	
Datos de referencia core	Organización, casi toda la	Bajo	Muy Baja	Pocas	Códigos de países o de divisas,	Normalmente externos a la organización	Repositorio de datos de referencia corporativo	
Datos maestros comunes	Múltiples sistemas abarcando la organización	Medio	Baja	Muchas	Cliente, empleado, producto, cuentas bancarias	Principalmente usuarios que los crean o los necesitan	Varios escenarios	
Datos transaccionales	Transacciones actuales	Alto	Media	Muchas	Compras, inversiones, ...	Stmas. que generaron las transiciones	Stmas. que generaron las transiciones	

Tipos de Datos	Amplitud de los datos	Volumen	Frecuencia Actualización	Número de Columnas	Ejemplos	Propietarios o Stewards	Sistema Típico de Registro
Metadatos	Organización	Muy baja	Casi estática	Varias	Definiciones de Organización	Organización	Repositorio de Metadatos
Datos de Referencia Core	Organización, casi toda la organización	Baja	Muy Baja	Pocas	SIC Países, Códigos de Divisas, Esquemas ISOO	Normalmente externos a la organización	Repositorio de datos de referencia corporativo
Datos Maestros Comunes	Múltiples sistemas abarcando la organización	Medio	Bajo	Muchas	Cliente, Empleado, Producto, Cuentas Bancarias	Principalmente usuarios que los crean o lo necesitan	Varios escenarios
Datos Transaccionales	Transacciones actuales	Alto	Medio	Muchas	Compras, inversiones, ...	Sistema que generaron las transiciones	Sistema que generaron las transiciones

Para compilar en HTML tabla.html

Característica	Definición	Inherente	Dependiente del sistema
Exactitud	Grado en que los datos tienen atributos que representan correctamente el valor de un atributo	X	
Completeness	Grado en el que existen suficientes valores para todos los atributos necesarios para la representación de una entidad	X	
Consistencia	Grado en que los datos están libres de contradicción y son coherentes con el resto de los datos de su contexto de uso	X	
Credibilidad	Grado en que los datos se consideran ciertos y creíbles por los usuarios	X	
Actualidad	Grado en que los datos tienen atributos con las fechas y tiempos correctos	X	
Accesibilidad	Grado en que los datos pueden ser accedidos por cualquier usuario	X	X
Cumplimiento	Grado en el que los datos están construidos conforme estándares, convenciones o regulaciones	X	X
Confidencialidad	Grado en el que los datos tienen atributos específicos que solo pueden ser accedidos por usuarios autorizados	X	X
Eficiencia	Grado en el que los datos tienen atributos que pueden ser procesados y provistos dentro de los niveles de rendimiento esperados	X	X
Exactitud	Grado en el que los datos tienen atributos que son precisos	X	X
Trazabilidad	Grado en el que los datos tienen atributos que proveen información detallada sobre los cambios realizados en los datos	X	X
Comprensibilidad	Grado en el que los datos se expresan de manera que los usuarios puedan leerlos e interpretarlos fácilmente	X	X
Disponibilidad	Grado en el que los datos están disponibles para ser accedidos por usuarios y/o aplicaciones autorizadas	X	
Portabilidad	Grado en el que los datos pueden ser alejados, reemplazados o movidos desde un sistema a otro	X	
Recuperabilidad	Grado en el que los datos disponen de formas de mantener un nivel especificado de operabilidad incluso cuando se producen fallos	X	

Para poder usar estos datos, las aplicaciones pueden realizar los siguientes cuatro tipo de operaciones (normalmente conocidas como ‘operaciones CRUD’):

1. **Crear datos (Create):** inserta datos en el repositorio de datos.
2. **Leer datos (Read):** recupera datos del repositorio para aprovisionar el proceso de negocio, o bien para realizar alguna operación de análisis específica.
3. **Actualizar datos (Update):** modifica el valor de los atributos correspondientes a los hechos o entidades para actualizarlos a nuevas observaciones.
4. **Borrar datos (Delete):** elimina en bloque o selectivamente los datos almacenados en el repositorio de los datos.

En cualquier caso, estos procedimientos de inserción, actualización, recuperación y borrado deben garantizar siempre la seguridad del conjunto de los datos, de modo que sólo sean accesibles por aquellos usuarios que estén autorizados a trabajar con ellos, y siempre para el propósito establecido para los datos (Piattini et al., 2006).

La forma de implementar estas operaciones depende fuertemente del formato (modelo lógico) en el que estén almacenados los datos. Aunque existen diferentes modelos (estructurados, semi-estructurados, no estructurados), en este capítulo el énfasis se pone en el modelo relacional (Codd, 1970b), ya que es el más ampliamente usado en el ámbito organizacional y el que implementan los SGBDR. Para poder dar soporte a las operaciones CRUD anteriormente citadas en bases de datos relacionales, se desarrolló un lenguaje llamado ‘lenguaje estructurado de consulta’ (*structured query language*, SQL), que se aborda en la siguiente sección.

5.3. SQL: el lenguaje estructurado de consulta

Los principios de SQL están establecidos en el estándar internacional ISO/IEC 9075:1989² como un mecanismo para identificar y regular las expresiones necesarias que permiten manejar bases de datos relacionales. Al ser un estándar, es importante señalar que cada fabricante de SGBDR, como Oracle con [Oracle Database Manager Server](#)³ o con [MySQL](#)⁴, Microsoft con [SQL Server](#)⁵, IBM con [DB2](#)⁶..., implementa en sus productos su propia versión del estándar SQL. Y aunque son prácticamente iguales, hay ligeros matices que les permiten diferenciarse de la competencia y que, por tanto, deben ser conocidos cuando se utilicen los correspondientes productos comerciales. No obstante, existen en el mercado algunas soluciones *open source* como [MariaDB](#)⁷ o [PostgreSQL](#)⁸. En este capítulo, todos los ejemplos que se han desarrollado trabajan contra un servidor MySQL^{8,9}.

SQL tiene diferentes tipos de sentencias o instrucciones que dan soporte a los diferentes aspectos de las interacciones con la base de datos. Cualquier manual de SQL permite tratar en profundidad todos los elementos sintácticos del lenguaje, pero es importante señalar que los detalles específicos de la sintaxis específica dependerán fuertemente del SGBDR empleado. Para los ejemplos propuestos en este libro, puede consultarse el [manual de referencia de SQL de MySQL v8.0](#)¹⁰. Las siguientes secciones proporcionan una visión global de dichos grupos de sentencias.

5.3.1. SQL como lenguaje de definición de datos

Una base de datos relacional tiene una organización en forma de tabla y su concepto fundamental es el de ‘relación’, que no es el que el lector se puede imaginar a primera vista. Una

²<https://www.iso.org/standard/16662.html>

³<https://www.oracle.com/database/>

⁴<https://www.mysql.com/products/enterprise/database/>

⁵<https://www.microsoft.com/es-es/sql-server/>

⁶<https://www.ibm.com/es-es/db2>

⁷<https://mariadb.org/>

⁸<https://www.postgresql.org/>

⁹<https://www.mysql.com/products/>

¹⁰<https://dev.mysql.com/doc/refman/8.0/en/sql-statements.html>

5.3. SQL: el lenguaje estructurado de consulta

71

‘relación’ representa un conjunto de entidades con las mismas propiedades y se compone de filas (o registros; también denominadas tuplas) cuyos valores dependen de los atributos que se representen en las columnas. Por ejemplo, una relación puede ser el conjunto de equipos de fútbol de la primera división española, siendo los atributos, su presupuesto, nombre del entrenador, número de jugadores españoles...

Las bases de datos relacionales se caracterizan por utilizar el lenguaje de consulta estructurado (SQL) y, por ello, son también denominadas bases de datos SQL. En particular, SQL se utiliza para definir todos los elementos necesarios para crear y modificar las tablas de datos. Se tienen tres tipos de instrucciones básicas para gestionar las tablas como estructuras de datos:

- **Create:** permite crear un componente de la base de datos, tal como la base de datos propiamente dicha, una tabla, una vista,... En el siguiente ejemplo, se crea, usando instrucciones SQL, primero una base de datos llamada **Biblioteca**, y luego una tabla **Autor** con seis atributos (**CodAutor**, **Nombre**, **Apellido1**, **Apellido2**, **Pseudonimo** y **Nacionalidad**), donde se podrán almacenar los valores correspondientes a dichos atributos, conformando así la base de datos:

```
create database Biblioteca
create Table Autor (
    CodAutor nvarchar (20) primary key,
    Nombre nvarchar(40) not null,
    Apellido1 nvarchar(50) not null,
    Apellido2 nvarchar(50),
    Pseudonimo nvarchar(50),
    Nacionalidad nvarchar (50)
);
```

- **Alter:** permite modificar la estructura de un componente, añadiendo por ejemplo, atributos a una tabla, restricciones a un atributo, o modificando el tipo de datos de algún atributo existente; también permite eliminar un atributo de una tabla existente. Siguiendo el ejemplo anterior, con la siguiente instrucción se añade un nuevo atributo, **LocalidadNacimiento**, a la tabla **Autor**:

```
alter table Autor add LocalidadNacimiento nvarchar(50) not null
```

- **Drop:** sirve para eliminar un componente específico, como por ejemplo una tabla, una vista,... Pero no sirve para eliminar los valores almacenados en una tabla. En el siguiente ejemplo, se eliminan las tablas **Escribe**, **Autor** y **Libro**:

```
drop table Escribe;
drop table Autor;
drop table Libro;
```

5.3.2. SQL como lenguaje de manipulación de datos

En esta sección se describen las instrucciones más importantes de SQL para el soporte a las operaciones CRUD anteriormente introducidas. Existen, por tanto, cuatro tipos de sentencia para manipular los datos:

- **Create:** implementada mediante la instrucción `insert`, sirve para insertar registros (también llamados **tuplas**) en una base de datos. En los ejemplos siguientes se insertan diversas tuplas en varias tablas, siguiendo el mismo orden en el que se especificaron los atributos cuando se creó la tabla. Así por ejemplo, se crean los códigos “`dbrown`” (como valor para el atributo `CodAutor`) para “*Dan Brown*” y “`cdv`” (como valor para el atributo `CodLibro`) para su libro “*El Código da Vinci*”. El siguiente código SQL muestra las instrucciones necesarias:

```
insert into Autor values ('dbrown', 'Dan', 'Brown', '', '', 'EstadoUnidense');
insert into Libro values ('cdv', 'El Código da Vinci', 'Random House',
→ '2003-04-23');
insert into Escribe values ('dbrown', 'cdv');
```

- **Read:** implementada mediante la instrucción `select`, permite hacer consultas a la base de datos. En los siguientes ejemplos se escribe el código que selecciona (i) el nombre y primer apellido de los autores con nacionalidad española, ordenados por orden alfabético del Apellido1 y (ii) la lista todos los libros que haya escrito el autor “*Pérez Reverte*” (cuyo `CodAutor` es “`perezreverte`”).

```
Select Nombre, Apellido1 from Autor where Nacionalidad like 'Español' order by
→ Apellido1;

Select Libro.Título from Autor, Escribe, Libro where ( Libro.CodLibro =
→ Escribe.CodLibro and Autor.CodAutor = Escribe.CodAutor) and (Escribe.CodAutor
→ ='perezreverte');
```

- **Update:** permite actualizar los valores de las tuplas seleccionadas. En el siguiente ejemplo, se actualiza el valor del atributo `pseudonimo` al valor “*El Manco de Lepanto*” para el autor “*Miguel de Cervantes*”, con `CodAutor` “`mcervantes`”:

```
update Autor set pseudonimo="El Manco de Lepanto" where CodAutor='mcervantes';
```

- **Delete:** su principal objetivo es eliminar, en bloque o de forma selectiva, una o varias tuplas o registros de datos que cumplan una determinada condición. En el siguiente código SQL se borra la(s) tupla(s) que contiene(n) datos del autor cuyo `CodAutores` “`perezreverte`”.

5.4. Acceso y explotación de bases de datos desde **R**

73

```
delete from Autor where CodAutor ='perezreverte';
```

5.3.3. SQL como lenguaje de administración de datos

SQL también puede ser usado también para administrar los usuarios de una base de datos. Esto implica crear usuarios de la base de datos y asignarles diferentes tipos de permisos para realizar los diferentes tipos de operaciones vistos anteriormente sobre los distintos componentes de datos. Por ejemplo, para crear un usuario llamado `Ismael.Caballero` que tenga por contraseña `LibroMDSR` se puede usar la siguiente instrucción:

```
create user 'Ismael.Caballero' identified by 'LibroMDSR';
```

y la siguiente instrucción se usa para asignar al usuario `Ismael.Caballero` los permisos necesarios para el acceso, lectura, selección, inserción, actualización y borrado de los valores de la base de datos `Biblioteca`, así como para poder modificar la estructura de los componentes de la base de datos `Biblioteca` creada anteriormente:

```
GRANT SELECT, INSERT, UPDATE, DELETE, CREATE, DROP, ALTER ON biblioteca.* TO
→ 'Ismael.Caballero'@'%';
FLUSH PRIVILEGES;
```

Recuérdese que los ejemplos mostrados han sido realizados para MySQL 8, aunque la sintaxis no debería ser muy diferente para otros SGBDR.

5.4. Acceso y explotación de bases de datos desde **R**

Como el propósito de este manual es aprender los fundamentos de la Ciencia de Datos usando **R**, en las siguientes secciones se explicará cómo implementar las operaciones CRUD usando sentencias de paquetes específicos de **R**. Dado que en este capítulo se usa MySQL, se utiliza el driver específico de RMySQL ([Ooms et al., 2022](#)). En caso de que se hubiese usado otro sistema gestor de bases de datos, se hubiera tenido que recurrir al paquete específico que contuviera el driver correspondiente. En las siguientes secciones se explica cómo conectarse a una base de datos usando las funciones correspondientes y cómo se implementan las operaciones CRUD con funciones del paquete RMySQL.

5.4.1. Conexión a una base de datos

Antes de poder realizar ninguna operación con las bases de datos gestionadas por MySQL es preciso tener instalado el paquete RMySQL y cargar los paquetes necesarios:

```
install.packages("RMySQL")
library("RMySQL")
library(DBI)
```

Es necesaria también la librería DBI ([R Special Interest Group on Databases \(R-SIG-DB\) et al., 2022](#)) porque proporciona la infraestructura común para todos los drivers de acceso a base de datos. Además del mencionado RMySQL, otros ejemplos de drivers son RPostgres ([Wickham et al., 2023](#)), RMariaDB ([Müller et al., 2022a](#)), odbc ([Hester and Wickham, 2023](#)) o RSQLite([Müller et al., 2022b](#)), por citar algunos.

Para desarrollar las explicaciones, se usa una base de datos llamada `classicmodels`, implementada en MySQL v8.0 y desplegada en un servidor con dirección IP 172.20.48.118 que está escuchando en el puerto 3306. El usuario que se conecta a la base de datos es `Ismael.Caballero`, siendo su contraseña `MdsR.2022`. Otros usuarios tendrían que modificar los parámetros correspondientes para realizar las conexiones a sus propias bases de datos. Se almacenan todos estos datos en variables para hacer más sencillo el mantenimiento de los scripts. Con `dbConnect()` se realiza la conexión. Con `summary()` o con `dbGetInfo()` se pueden mostrar los resultados de la conexión en caso de que ésta se haya realizado con éxito. Una vez terminadas todas las tareas con la base de datos, debería desconectarse mediante la instrucción `dbDisconnect()`.

```
usuario      = 'Ismael.Caballero'
passwd       = 'MdsR.2022#'
nombrebd    = 'classicmodels'

servidor     = '172.20.48.118'
puerto       = 3306

mibbdd = dbConnect(MySQL(), user=usuario, password=passwd, dbname=nombrebd,
                   host=servidor, port= puerto)
summary (mibbdd)
dbGetInfo(mibbdd)
dbDisconnect (mibbdd)
```

5.4.2. Operaciones de lectura / consulta/ selección (*read*) de datos

Entre las operaciones más frecuentes en cualquier tipo de bases de datos están las de lectura o consulta (*read*), implementadas en SQL con las sentencias de tipo `select`. El driver RMySQL ofrece distintas alternativas para realizar consultas de selección en **R**. La elección de la mejor operación dependerá de la complejidad de las consultas que se quieran realizar. Se listan a continuación:

1. `dbReadTable()`, que permite leer una tabla entera de una base de datos MySQL. Es recomendable usar este método si la tabla no es excesivamente grande. Se pueden almacenar los resultados en un `data.frame` para hacer operaciones con ellos después. La ventaja

5.4. Acceso y explotación de bases de datos desde **R**

75

es que, sabiendo manejar `data.frames` en **R**, no se necesita aprender mucho más detalle del lenguaje **SQL**¹¹ como lenguaje de manipulación de datos (LMD); sin embargo, la desventaja es que se acaba perdiendo parte del potencial expresivo de SQL para hacer algunas operaciones más fáciles y eficientes. En el siguiente ejemplo se muestra cómo cargar toda la tabla `customers` en un `data.frame` llamado `tblCustomers`, obteniéndose los resultados con la instrucción `summary()`:

```
tblCustomers <- dbReadTable(mibbdd, "customers")
summary (tblCustomers)
```

2. `dbGetQuery()`, que tiene más flexibilidad que `dbReadTable()` porque permite, mediante una sentencia SQL `select` (véase más información en [el tutorial de SQL en W3C](#)¹² o en la página oficial de “`select`” sobre MySQL¹³), particularizar la consulta a la base de datos. Esto puede implicar la selección de atributos específicos o incluso el uso de filtros sobre los atributos seleccionados. Por ejemplo, si se quisieran recuperar el número y el nombre de los clientes de Madrid, se podría personalizar la consulta añadiendo las condiciones correspondientes en la cláusula `where`¹⁴, como se muestra en siguiente código. Por comodidad, se escribe aparte la consulta SQL, en una variable, para poder manejar más fácilmente la operativa en **R**. Escribir esta consulta puede ser lo que entraña más dificultad. A continuación, se ejecuta la consulta con `dbGetQuery()` y se almacenan los resultados en un `data.frame` para su uso posterior. Nuevamente, se comprueba el resultado con la instrucción `summary()`.

```
SentenciaSQL_Nombres_Clientes ="Select CustomerNumber, CustomerName from customers
→ where city = 'Madrid'"

Consulta_Clientes_Madrid = dbGetQuery (mibbdd, SentenciaSQL_Nombres_Clientes)
summary (Consulta_Clientes_Madrid)
```

Teniendo los resultados en `data.frames`, ya es posible procesarlos en **R** como si fuesen cualquier otro tipo de datos.

Obsérvese que las instrucciones siguientes serían equivalentes:

```
dbReadTable(mibbdd, "customers")
dbGetQuery (mibbdd, "select * from customers")
```

3. `dbSendQuery()` combinado con `dbFetch()`. La principal diferencia entre `dbSendQuery()` y `dbGetQuery()` es que la primera no recupera datos de la base de datos y hay que traerlos explícitamente con la función `dbFetch()`. En el siguiente fragmento de código se muestra la utilización de ambas funciones con un resultado exactamente igual que en el apartado anterior.

¹¹<https://www.w3schools.com/sql/>

¹²https://www.w3schools.com/mysql/mysql_select.asp

¹³<https://dev.mysql.com/doc/refman/5.7/en/select.html>

¹⁴<https://www.mysqltutorial.org/mysql-where/>.

```

SentenciaSQL_Nombres_Clientes ="Select CustomerNumber, CustomerName from customers
→   where city = 'Madrid'"

Consulta <- dbSendQuery(mibbdd, SentenciaSQL_Nombres_Clientes);
dbGetInfo(Consulta)
#> $statement
#> [1] "Select CustomerNumber, CustomerName from customers where city = 'Madrid'"
#>
#> $isSelect
#> [1] 1
#>
#> $rowsAffected
#> [1] -1
#>
#> $rowCount
#> [1] 0
#>
#> $completed
#> [1] 0
#>
#> $fieldDescription
#> $fieldDescription[[1]]
#> NULL

print(paste("Consulta realizada:", dbGetStatement(Consulta)) )

Consulta_Clientes_Madrid_condbSendQuery <- dbFetch(Consulta, n=-1)
print( paste("Número de elementos devueltos en la consulta",dbGetRowCount(Consulta)))

summary (Consulta_Clientes_Madrid_condbSendQuery)

```

En cualquier caso, para considerar la opción más adecuada deben considerarse los siguientes aspectos:

- La información de la consulta generada con `dbSendQuery()` puede mostrarse con la función `dbGetInfo()`.
- Es posible recordar la consulta SQL que se utilizó en `dbSendQuery()` mediante la función `dbGetStatement()`.
- La función `dbFetch()` tiene dos argumentos: la consulta y el número de registros a recuperar; si se quieren recuperar todos los registros que haya podido producir la consulta, debe pasarse el argumento `n=-1`.
- Si se quiere saber el número de elementos que se han traído con la función `dbFetch()` se puede usar la función `dbGetRowCount()`.

La principal ventaja de `dbSendQuery()` combinado con `dbFetch()` es que el filtro se hace en el

5.4. Acceso y explotación de bases de datos desde **R**

77

sistema gestor de bases de datos y sólo llegan a la memoria de **R** los datos que se van a utilizar, que es mejor que descargar toda la tabla a la memoria de **R** y, después, hacer el filtro.

Para extraer información de los resultados de la consulta, se puede usar la función `dbColumnInfo()`:

```
dbColumnInfo(Consulta)
#> name Sclass type length
#> 1 CustomerNumber integer INTEGER 11
#> 2 CustomerName character VAR_STRING 200
```

El driver RMySQL no proporciona funciones para conocer directamente el tipo y tamaño de los atributos de una tabla. Mediante la función `dbSendQuery()` y `dbColumnInfo()` se puede obtener esta información haciendo una consulta que incluya los atributos en los cuales se está interesado. Por ejemplo, para conocer el tipo de datos y tamaño de los atributos de la tabla `Employees` se podría usar el siguiente fragmento de código:

```
SentenciaSQL_Tabla_Employees = "Select * from employees"
Consulta_Employees <- dbSendQuery(mibbdd,SentenciaSQL_Tabla_Employees)

dbColumnInfo(Consulta_Employees)

#> name Sclass type length
#> 1 employeeNumber integer INTEGER 11
#> 2 lastName character VAR_STRING 200
#> 3 firstName character VAR_STRING 200
#> 4 extension character VAR_STRING 40
#> 5 email character VAR_STRING 400
#> 6 officeCode character VAR_STRING 40
#> 7 reportsTo integer INTEGER 11
#> 8 jobTitle character VAR_STRING 200

dbClearResult(Consulta_Employees)
#> [1] TRUE
```

Con `dbColumnInfo()` se muestran los metadatos de implementación (operativos) de los atributos de la tabla `Employees`. Finalmente, con la instrucción `dbClearResult(ConsultaEmployees)` se pueden limpiar los resultados de la consulta para optimizar el sistema.

5.4.3. Operaciones de inserción (*create*) y actualización (*update*) de datos

Antes de almacenar los datos en la base de datos, es necesario crear las estructuras necesarias, que, como se avanzó anteriormente, son las tablas y los atributos. Para ello se utilizan instrucciones especiales SQL como lenguaje de definición de datos (LDD); esto incluye instrucciones

para crear tablas (`create table`)¹⁵, para modificarlas (`alter table`)¹⁶ o para borrarlas (`drop table`)¹⁷.

Para poder hacer operaciones con los datos, es preciso crear usuarios y asignarles los privilegios adecuados sobre las tablas y atributos. Ello también requiere las instrucciones especiales SQL como lenguaje de administración de datos (LAD), que incluye instrucciones para crear usuarios (`create user`)¹⁸, modificar ciertos aspectos de los mismos (`alter user`)¹⁹ y borrarlos (`drop user`)²⁰.

Un usuario de la base de datos que tenga privilegios suficientes sobre las estructuras creadas puede crear (`insert`)²¹ o modificar (`update`)²² registros de datos usando las instrucciones específicas de SQL como lenguaje de manipulación de datos (LMD) -véase cómo otorgar privilegios a un usuario para crear tablas²³.

No obstante, y dado que el software en el que se centra este manual es en **R**, se deja fuera del alcance de este capítulo el uso de los aspectos LDD, LMD y LAD de SQL, y se cubrirán mediante la instrucciones `dbWriteTable()` de **RMySQL** los aspectos de inserción y de actualización de los registros. `dbWriteTable()` se usa por tanto para exportar datos de **R** a una base de datos MySQL, y puede ser usado para las acciones que se exponen a continuación, siempre y cuando el usuario que ejecute las acciones tenga suficientes permisos en el sistema gestor de bases de datos para realizarlas.

5.4.3.1. Crear una nueva tabla con datos

La creación de una nueva tabla de datos se lleva a cabo a partir de un `data.frame` que se puebla con datos iniciales y que tendrá tantas columnas como atributos tenga la tabla. Por ejemplo, a continuación se crea un `data.frame` llamado `dfDatos_Prueba` con dos columnas, una de tipo numérico llamada `CodPrueba` y otra de tipo texto llamada `DatosPrueba`. En este ejemplo, y a modo ilustrativo, los datos son completamente arbitrarios. Después, se construye el `data.frame` `dfDatos_Pruebas` y mediante `dbListTables()` se comprueba que la tabla no existe en la conexión a la base de datos. Finalmente, con `dbWriteTable()`, se crea la nueva tabla. Es importante tener en cuenta las posibles conexiones simultáneas a la base de datos porque se podrían generar problemas. Con `dbWriteTable()` se puede comprobar si la tabla se ha creado correctamente.

```
CodPrueba <- c(1:26)
Nombre_Prueba <- c(letters[1:26])
dfDatos_Prueba <- data.frame(CodPrueba, Nombre_Prueba)
dbListTables(mibbdd)
```

¹⁵<https://dev.mysql.com/doc/refman/8.0/en/creating-tables.html>

¹⁶<https://dev.mysql.com/doc/refman/8.0/en/alter-table.html>

¹⁷<https://dev.mysql.com/doc/refman/8.0/en/drop-table.html>

¹⁸<https://dev.mysql.com/doc/refman/8.0/en/create-user.html>

¹⁹<https://dev.mysql.com/doc/refman/8.0/en/alter-user.html>

²⁰<https://dev.mysql.com/doc/refman/5.6/en/drop-user.html>

²¹<https://dev.mysql.com/doc/refman/8.0/en/insert.html>

²²<https://dev.mysql.com/doc/refman/8.0/en/update.html>

²³<https://dev.mysql.com/doc/refman/5.7/en/grant-tables.html>

5.4. Acceso y explotación de bases de datos desde **R**

79

```
#> [1] "Autor" "DatosPrueba_16" "DatosPrueba_22" "Datos_Prueba_01",
#> [5] "Make" "Pelicula" "customers" "employees",
#> [9] "offices" "orderdetails" "orders""payments",
#> [13] "productlines" "products"

dbWriteTable(mibbdd, "DatosPrueba", dfDatos_Prueba, overwrite = TRUE, row.names =FALSE
← )
#> [1] TRUE
#>
dbListTables(mibbdd)
#> [1] "Autor" "DatosPrueba" "DatosPrueba_16" "DatosPrueba_22",
#> [5] "Datos_Prueba_01" "Make" "Pelicula" "customers" ,
#> [9] "employees" "offices" "orderdetails", "orders" ,
#> [13] "payments" "productlines" "products"
```

Es interesante pensar en la utilidad de este método para duplicar tablas en caso necesario

5.4.3.2. Sobreescribir una tabla existente con datos actualizados

Cuando se trata de actualizar algunos valores de los atributos de la tabla o de añadir nuevos registros a la tabla, la operación es básicamente la misma que antes, pero primeramente habrá que leer la tabla y convertirla en un **data.frame** para actualizar en él los valores o añadir los nuevos valores (en este caso se añade una nueva fila); una vez hecho esto, se vuelve a utilizar el comando **dbWriteTable()** añadiendo los parámetros **overwrite = TRUE** (para sobrescribir toda la tabla) y **row.names = FALSE**. En el siguiente ejemplo se actualizan los valores de una tupla específica.

```
dfDatos_Prueba <- dbReadTable(mibbdd, "DatosPrueba")

dfDatos_Prueba$NombrePrueba[25] <- "en un lugar de la mancha"

dbWriteTable(mibbdd, "DatosPrueba", dfDatos_Prueba, overwrite = TRUE, row.names =FALSE
← )
#> [1] TRUE

dfDatos_Prueba_Modificado <- dbReadTable(mibbdd, "DatosPrueba")
```

5.4.3.3. Añadir nuevos registros a una tabla

Existen dos estrategias para añadir registros a una tabla. La primera es utilizar la técnica de obreescritura descrita anteriormente. Para ello, se procede como antes: se carga la tabla en un **data.frame** (en este caso **dfDatos_Prueba**), se añaden nuevas filas (registros) al **data.frame** (cargadas previamente en el **data.frame** **dfNuevoRegistro**) usando **rbind()**, y a continuación se sobreescrige la tabla usando **dbWriteTable()**. En el siguiente fragmento de código se muestra cómo añadir nuevos registros a una tabla sobreescriéndola completamente.

```
dfDatos_Prueba <- dbReadTable(mibbdd, "DatosPrueba")

dfNuevo_Registro <- as.list(dfDatos_Prueba)

dfNuevo_Registro$CodPrueba <- c(27)
dfNuevo_Registro$NombrePrueba <- c("Un Valor Nuevo")

dfDatos_Prueba <- rbind (dfDatos_Prueba, dfNuevo_Registro)

dbWriteTable(mibbdd, "DatosPrueba", dfDatos_Prueba, overwrite = TRUE, row.names =FALSE)
#> [1] TRUE
```

La opción anterior puede ser interesante si la tabla no tiene muchos registros y el coste computacional no es muy grande. Pero si se tiene muchos registros es preferible usar otra estrategia para añadir un nuevo registro a la tabla. En este caso, se puede hacer creando un `data.frame` compatible con la estructura de la tabla, y ejecutar la instrucción `dbWriteTable()` poniendo el parámetro `append = TRUE`. Esto añadirá el nuevo registro al final de la tabla. El siguiente fragmento de código muestra cómo realizar esta operación.

```
dfDatos_Prueba_Nuevos <- as.list(dfDatos_Prueba)

dfDatos_Prueba_Nuevos$CodPrueba <- 28
dfDatos_Prueba_Nuevos$NombrePrueba <- "Otro valor nuevo"

dfDatos_Prueba_Nuevos <- data.frame (dfDatos_Prueba_Nuevos)

dbWriteTable(mibbdd, "DatosPrueba", dfDatos_Prueba_Nuevos, append = TRUE, row.names
            =FALSE)
#> [1] TRUE
```

5.4.3.4. Inserción con consulta SQL usando la instrucción `dbSendQuery()`

Una última forma de insertar valores en una tabla es mediante la instrucción `dbSendQuery`, utilizando una consulta de inserción `insert`. En el siguiente ejemplo se muestra cómo insertar tuplas o registros mediante `dbSendQuery()`; en este caso, se añaden datos completamente arbitrarios a modo de ejemplo.

```
SentenciaSQL_Insercion ="insert into DatosPrueba value (29, 'Una tercera forma')"

dbSendQuery (mibbdd, SentenciaSQL_Insercion)
#> <MySQLResult:-365007472,0,23>
```

5.4.4. Operaciones de borrado de datos (*delete*)

Finalmente, se describen las operaciones de borrado. Análogamente a como se hacían las operaciones de inserción, se pueden hacer de dos formas:

1. **Borrado de valores usando `dbWriteTable()` con sobreescritura:** esto implica extraer todos los datos de la tabla, borrar el registro o los registros correspondientes y sobreescribir nuevamente la tabla en la base de datos mediante la instrucción `dbWriteTable()` con la opción `overwrite = TRUE`; para ver el resultado se puede usar la función `summary()`. El siguiente fragmento de código muestra cómo hacerlo:

```
dfDatos_Prueba <- dbReadTable(mibbdd, "DatosPrueba")

dfDatos_Prueba <- dfDatos_Prueba[dfDatos_Prueba$CodPrueba < 25, ]

dbWriteTable(mibbdd, "DatosPrueba", dfDatos_Prueba, overwrite = TRUE, row.names = FALSE)
#> [1] TRUE
```

2. **Borrado de registros con consulta SQL en `dbSendQuery()`:** se puede llevar a cabo utilizando una sentencia SQL de borrado `delete`²⁴ con la instrucción `dbSendQuery()` para borrar registros de la base de datos. El siguiente fragmento de código muestra cómo hacerlo:

```
# Se usa una sentencia SQL de borrado. El criterio de borrado es completamente
→ arbitrario a efectos ilustrativos.
SentenciaSQL_Eliminación ="delete from DatosPrueba where CodPrueba > 10"
dbSendQuery (mibbdd, SentenciaSQL_Eliminación)
#> <MySQLResult:1,0,30>
```

Finalmente, si fuera necesario eliminar toda la tabla, se podría usar una sentencia `drop table`²⁵:

```
dbSendQuery(mibbdd, "drop table DatosPrueba")
#> <MySQLResult:0,0,31>
dbDisconnect(mibbdd)
#> [1] TRUE
```

²⁴<https://dev.mysql.com/doc/refman/8.0/en/delete.html>

²⁵<https://dev.mysql.com/doc/refman/8.0/en/drop-table.html>

Resumen

En este capítulo se han presentado los fundamentos de las bases de datos relacionales. Es importante tener presente los siguientes aspectos:

- Los datos en las bases de datos se corresponden a valores de atributos relevantes de entidades del mundo real.
- Los datos de una base de datos son una percepción u observación del mundo real.
- Los datos son la materia prima de los procesos de negocio.
- Los sistemas de información dan soporte a los procesos de negocio.
- Los datos son un elemento fundamental de los sistemas de información.
- SQL es el lenguaje más comúnmente utilizado en operaciones sobre el modelo físico de bases de datos relacionales.
- SQL se puede utilizar como Lenguaje de Definición de Datos (LDD), como Lenguaje de Manipulación de Datos (LMD), y como Lenguaje de Administración de Datos (LAD).
- La sintaxis de SQL depende fuertemente del sistema gestor de bases de datos relacionales que lo implemente.
- **R Software**, a través del driver específico, permite manejar bases de datos implementando las operaciones CRUD.

Capítulo 6

Gestión de bases de datos NoSQL

Ricardo Pérez-Castillo^a e Ismael Caballero^a

^aUniversidad de Castilla-La Mancha

6.1. Introducción al big data

Actualmente se vive en la era de la información, con un teléfono móvil en cada bolsillo, un ordenador portátil en cada mochila y grandes sistemas de tecnología funcionando diariamente mandando datos y datos cada segundo. El mundo tiene más datos que nunca, pero esto no es todo, ya que el volumen aumenta de forma exponencial (López, 2012). Es la era de las bases de datos masivos, en inglés *big data*.

En particular, el volumen de datos disponibles para las empresas aumentó drásticamente desde 2004. En 2004, la cantidad total de datos almacenados en Internet fue de 1 petabyte (equivalente a 100 años de todo el contenido de televisión). En 2011 la cantidad total de información almacenada en todo el mundo ya era de 1 zettabyte (1 millón de petabytes o 36 millones de años de video de alta definición); en 2015 alcanzó los 7.9 zettabytes (o 7.9 millones de petabytes) y en 2020 se disparó a 35 zettabytes (o 35 millones de petabytes). Este gran volumen de datos, y su crecimiento continuo y exponencial, supera las capacidades de las herramientas de datos tradicionales para capturarlos, almacenarlos, administrarlos y analizarlos (Kalyvas and Overly, 2014). Por este motivo, se hace necesario el uso de nuevos métodos, técnicas y herramientas de gestión de datos. Este espacio es el que cubre *big data*.

Big data es un término abstracto que, en cierta medida, se ha puesto de moda en diferentes ámbitos: negocios, marketing, *social media* y diferentes ingenierías como informática, sistemas de información, almacenamiento y recuperación de datos, etc. *Big data* es un término que hace referencia al gran volumen de datos (tanto estructurados como no estructurados) que inundan día a día a cualquier organización. Pero lo más relevante no es la cantidad de datos. Lo que realmente importa es lo que las organizaciones pueden hacer con los datos. Los grandes

volúmenes de datos se pueden analizar, por ejemplo, en busca de ideas conducentes a una mejor toma de decisiones y movimientos comerciales estratégicos ([SAS Institute Inc., 2017](#)).

Cuando se acumulan grandes volúmenes de datos, se plantea la necesidad de ver qué se puede hacer con ellos. Esto implica gestionar los datos con una finalidad organizativa y disponer de tecnología y metodologías específicas. La propia gestión de datos lleva a generar información relevante en el contexto de la organización, es decir, a generar conocimiento para la acción que sea aplicable: por ejemplo, a la toma de decisiones, al diseño de acciones o a la elaboración de planes estratégicos ([García-Alsina, 2017](#)).

Por tanto, cuando se habla de datos masivos, se está hablando también de gestión de la información y de generación de conocimiento para la acción. Este campo científico es el que proporciona las pautas metodológicas para gestionar grandes volúmenes de datos con el fin de crear valor mediante una serie de procesos y procedimientos. Pero exige contar con la tecnología para capturar los datos, procesarlos, analizarlos e interpretarlos de manera eficaz y eficiente ([García-Alsina, 2017](#), [Gómez García and Conesa i Caralt \(2015\)](#)).

Por consiguiente, se puede concluir que *big data* es el “*conjunto de datos masivos heterogéneos que supera la capacidad del software habitual para ser capturados, gestionados y procesados en un tiempo razonable*”. Esta definición tiene en cuenta tres de las V’s del *big data* (véase Sec. 6.2): volumen, variabilidad y velocidad. Así, cuando se habla de *big data* se está haciendo referencia a conjuntos de datos o combinaciones de conjuntos de datos cuyo tamaño (*volumen*), complejidad (*variabilidad*) y *velocidad* de crecimiento dificultan su captura, gestión, procesamiento y análisis mediante tecnologías y herramientas convencionales, tales como las bases de datos relacionales y los paquetes de visualización y técnicas estadísticas convencionales, en el tiempo necesario para que sean útiles.

6.2. Las V’s del big data

El **volumen** se refiere a la cantidad de datos que son generados cada segundo, minuto y día en nuestro entorno. Es la característica más asociada al *big data*, ya que hace referencia a las cantidades masivas de datos que se almacenan con la finalidad de procesar dicha información, transformando los datos en acciones. Las personas están cada vez más conectadas al mundo digital, por lo que se generan más y más datos. Para algunas empresas, el estar en el mundo digital es algo obligatorio, por lo que la cantidad de datos generados es aún mayor. Por ejemplo, a una empresa que vende sus productos únicamente a través de un canal online, le convendría implantar tecnología *big data* para procesar toda aquella información que recoge su página web rastreando todas las acciones que lleva a cabo el cliente: conocer donde cliquea más veces, cuántas veces ha pasado por el carrito de la compra, cuáles son los productos más vistos, las páginas más visitadas, etc.

La **velocidad** se refiere a la rapidez con la que los datos son creados, almacenados y procesados en tiempo real. Así, en procesos como la detección de fraude en una transacción bancaria o la monitorización de un evento en redes sociales, el tratamiento de la información en tiempo real es imprescindible para que resulten útiles y den lugar a acciones efectivas. Otros ejemplos son la gestión de catástrofes naturales o pandemias, o el seguimiento de una campaña analizando

6.3. Tipos de datos en entornos big data

85

comentarios de los actores a quienes ésta va dirigida, para ir reorientándola en función de la retroalimentación que fluye en redes sociales ([García-Alsina, 2017](#)).

La **variedad** se refiere a que los sistemas de procesamiento del *big data* deben ser capaces de procesar datos de diversas formas, tipos y fuentes. Los datos pueden ser estructurados y fáciles de gestionar, como las bases de datos relacionales, o no estructurados, entre los que se incluyen desde documentos de texto, correos electrónicos, datos de sensores, audios, vídeos o imágenes que se tienen en un dispositivo móvil, hasta publicaciones en los perfiles de redes sociales, artículos en blogs, secuencias de click que los usuarios hacen en una misma página, formularios de registro e infinitud de acciones más que se realizan desde un *smartphone*, una *tablet* o un ordenador. Este tipo de datos requiere un herramiental específico, ya que su tratamiento es totalmente diferente al de los datos estructurados. Por ello, las empresas necesitan disponer de las herramientas apropiadas para integrar, observar y procesar este tipo de datos.¹

Con el tiempo, se han ido incorporando, progresivamente, otras V's: las V's de **valor** (de enorme interés en el análisis de datos), **veracidad**, **viabilidad** y **visualización** ([IIC, 2016](#)).

6.3. Tipos de datos en entornos big data

En función de la estructura con la que se organizan los datos (forma en la que se agrupan, almacenan y se relacionan entre sí y manera en la que se puede acceder a ellos, analizarlos o modificarlos), éstos pueden clasificarse en: estructurados, no estructurados o semiestructurados.

- **Datos estructurados:** son aquellos que tienen longitud y formato, como las fechas, los números o las cadenas de caracteres. En esta categoría entran los que se compilan en los censos de población, los diferentes tipos de encuestas, los datos de transacciones bancarias, las compras en tiendas online, etc.
- **Datos no estructurados:** son los que carecen de un formato determinado y no pueden ser almacenados en una tabla. Pueden ser de tipo texto (los que generan los usuarios de foros, redes sociales, documentos de *Word*, etc.), y los de tipo no-texto (cualquier fichero de imagen, audio, vídeo, ...). Este tipo de datos no tiene campos fijos y normalmente se tiene poco control sobre ellos. Su manipulación requiere tecnología de bases de datos *bigdata*, también conocidas como bases de datos NoSQL (*No only SQL*).
- **Datos semiestructurados:** poseen organización interna o marcadores que facilitan el tratamiento de sus elementos. No pertenecen a bases de datos relacionales. Es el caso de documentos XML, HTML o los datos almacenados en bases de datos NoSQL, que tienen una cierta estructura, aunque sin llegar a estar totalmente estructurados. También se pueden incluir en este tipo de datos los multi-estructurados o híbridos (datos de mercados emergentes, *e-commerce*, datos meteorológicos, etc.).

¹Los datos estructurados y no estructurados se abordan en detalle en la siguiente sección.

6.4. ¿Por qué bases de datos NoSQL?

Lo que hace que el *big data* sea tan útil para muchas empresas e instituciones es el hecho de que proporciona respuestas a muchas preguntas que dichas empresas e instituciones ni siquiera sabían que tenían. En otras palabras, proporciona un punto de referencia. Con una cantidad tan grande de información, los datos pueden ser moldeados o probados de cualquier manera que la organización considere adecuada. Al hacerlo, las organizaciones son capaces de identificar los problemas de una forma más comprensible.

La recopilación de grandes cantidades de datos y la búsqueda de tendencias en ellos permite que las organizaciones sean más ágiles y actúen mucho más rápidamente, sin problemas y de manera más eficaz. También les permite eliminar las áreas problemáticas antes de que los problemas acaben con sus beneficios o su reputación.

El análisis de *big data* ayuda a las organizaciones a aprovechar sus datos y utilizarlos para identificar nuevas oportunidades. Eso, a su vez, conduce a movimientos de negocio (o de otro tipo) más inteligentes, operaciones más eficientes, mayores ganancias y clientes más satisfechos. Las organizaciones más exitosas con *big data* consiguen valor de las siguientes formas:

- **Reducción de costes.** Las grandes tecnologías de datos, como Hadoop y el análisis basado en la nube, aportan importantes ventajas en términos de costes cuando se trata de almacenar grandes cantidades de datos, además de identificar maneras más eficientes de hacer negocios.
- **Mejores decisiones y más rápidas.** Con la velocidad de las bases de datos NoSQL y la analítica en memoria, combinada con la capacidad de analizar nuevas fuentes de datos, las organizaciones pueden analizar la información inmediatamente y tomar decisiones basadas en lo que han aprendido.
- **Nuevos productos y servicios.** Con la capacidad de medir las necesidades de los clientes y su satisfacción a través de análisis, viene la posibilidad de dar a los clientes lo que quieren. Con la analítica de *big data*, cada vez son más las organizaciones que están creando nuevos productos para satisfacer las necesidades de los clientes.

Existen ciertas diferencias entre las fuentes de datos tradicionales y las nuevas fuentes de datos que considera *big data*. Un resumen de las mismas puede verse en la Tabla 6.1.

Tabla 6.1: . Diferencias entre tecnologías tradicionales y tecnologías *big data*.

Tradicionales	<i>Big data</i>
Bases de datos relacionales	Bases de datos relacionales + <i>NoSQL</i>
Consultas	Consultas, capturas y procesamientos
Datos homogéneos	Datos heterogéneos
Ámbito de la informática	Todos los ámbitos

En primer lugar, la tecnología tradicional de almacenamiento y gestión de datos (desde final de los años 80) han sido las bases de datos relacionales. Aunque las bases de datos relacionales no

6.5. Bases de datos NoSQL

87

son, ni mucho menos, una tecnología en desuso, los entornos *big data* consideran otras tecnologías como, por ejemplo, las bases de datos NoSQL, que son bases de datos no relacionales optimizadas para modelos de datos sin esquema y de desempeño escalable. También son muy conocidas por su facilidad de desarrollo, baja latencia y resiliencia ([Amazon Web Services, 2018](#)).

A diferencia de las bases de datos basadas en SQL, las bases NoSQL no usan tablas tradicionales con líneas y columnas para almacenar datos, sino que los organizan con técnicas más flexibles, como, por ejemplo, documentos, gráficos, pares de valores y columnas. Por ello, son ideales para aplicaciones en las que se procesan grandes volúmenes de datos y que requieren estructuras flexibles. Como los sistemas NoSQL hacen uso de clústeres de hardware y servidores de nube, las capacidades se distribuyen de manera uniforme y la base de datos funciona con fluidez, aunque el volumen de datos sea grande. A diferencia de las bases de datos relacionales, cuyo rendimiento se reduce notablemente cuando aumenta el volumen de datos, las bases NoSQL suponen una solución potente, flexible y escalable incluso con grandes volúmenes de datos. Otra particularidad de los sistemas NoSQL es el escalamiento horizontal. Las bases de datos SQL relacionales cuentan con un escalamiento vertical y toda su capacidad de rendimiento se basa en un solo servidor. Sin embargo, en general, las soluciones NoSQL distribuyen los datos en varios servidores. Si aumenta el volumen de datos, simplemente se añaden nuevos servidores².

Otra gran diferencia respecto a las tecnologías tradicionales es que los entornos *big data* no sólo se centran en la consulta de datos, sino también en su captura y procesamiento (véase la Tabla 6.1). Además, las fuentes de datos pueden proveer datos heterogéneos con formatos heterogéneos. Estas diferencias hacen que no siempre se suficiente un ámbito de trabajo puramente informático, tendiendo a equipos multidisciplinares cuando se habla de proyectos *big data* (ingeniería, estadística, etc.).

6.5. Bases de datos NoSQL

6.5.1. Fundamentos de las bases de datos NoSQL

El término NoSQL se usa la primera vez en 1998 para referirse a una base de datos relacional sin SQL ([Strozzi, 1998](#)). NoSQL no significa estar en contra de SQL, y de hecho esto suele ser una falacia encontrada en la literatura. Sin embargo, para determinados problemas hay otras soluciones de almacenamiento más apropiadas. En la actualidad el término NoSQL se refiere a bases de datos que no solo tienen SQL (*not only SQL*).

Hay una gran variedad de sistemas de gestión de bases de datos que no usan SQL como principal lenguaje de consultas. Los datos almacenados no requieren estructuras fijas como tablas y no se garantizan completamente los **principios ACID** (*atomicity, consistency, isolation, and durability*) que sí deben cumplir las bases de datos relacionales (SQL):

- **Atomicidad.** Las *transacciones* se ejecutan completamente o no. Si fallan es como si ni siquiera se hubieran intentado ejecutar.

²<https://www.ionos.es/digitalguide/hosting/cuestiones-tecnicas/nosql/>.

- **Consistencia.** Un sistema consistente garantiza que cualquier *transacción* llevará a la base de datos de un estado válido a otro estado válido. Cualquier dato que se escriba en la base de datos tiene que ser válido de acuerdo con todas las reglas de integridad definidas en el modelo.
- **Aislamiento.** Cuando varias *transacciones* se ejecutan en paralelo, cada una de ellas ataca la base de datos de la misma manera a como lo haría si se ejecutara cada operación individual de forma aislada o secuencial.
- **Durabilidad.** El resultado de las transacciones es un cambio en el estado del sistema persistente. Si se apaga la máquina y se arranca de nuevo, el cambio producido por la transacción aún está presente.

Nota

Una **transacción** en una base de datos se refiere a un bloque de operaciones sobre los datos que debe completarse en su conjunto, o de lo contrario, en caso de un error puntual, se restaura el estado de la base de datos a su estado anterior, antes del inicio de la transacción.

A pesar de todo, las bases de datos NoSQL se denominan “no sólo SQL” para subrayar el hecho de que también pueden soportar lenguajes de consulta de tipo SQL.

Se puede decir el término NoSQL aparece con la llegada de la web 2.0, ya que hasta ese momento sólo subían contenido a la red aquellas empresas que tenían un portal. Pero con la llegada de aplicaciones como *Facebook*, *Twitter* o *Youtube*, entre otras, cualquier usuario podía subir contenido, provocando así un crecimiento exponencial de los datos ([acens.com, 2014](#)).

Es en este momento cuando empiezan a aparecer los primeros problemas relacionados con la gestión de toda esa información almacenada en bases de datos relacionales³. En un principio, para solucionar estos problemas de accesibilidad, las empresas optaron por utilizar un mayor número de máquinas, pero pronto se dieron cuenta de que esto no solucionaba el problema, además de ser una solución muy cara. La otra opción era la creación de sistemas pensados para un uso específico que con el paso del tiempo han dado lugar a soluciones robustas, apareciendo así el movimiento NoSQL ([acens.com, 2014](#)).

Por tanto, siguiendo de nuevo a [acens.com \(2014\)](#), hablar de bases de datos NoSQL es hablar de estructuras que permiten almacenar información en aquellas situaciones en las que las bases de datos relacionales generan ciertos problemas debidos, principalmente, al aumento progresivo de la capacidad de almacenamiento (escalabilidad) y rendimiento al darse cita miles de usuarios concurrentes y con millones de consultas diarias.

Además, como se esbozó al principio de la sección, las bases de datos NoSQL son sistemas de almacenamiento de información que no cumplen con el esquema ‘entidad–relación’. Tampoco

³Recuérdese (Cap. 5) que una base de datos relacional se basa en una organización tabular de los datos y que pivota sobre el concepto de ‘relación’ (que no es precisamente el que al lector le viene a la mente de inmediato). Formalmente, una relación representa un conjunto de entidades con las mismas propiedades y se compone de una serie de filas (o registros; también denominados tuplas) y columnas (atributos). Un ejemplo de relación pueden ser los equipos de fútbol de la primera división española, estando en cada fila los distintos equipos y en cada columna los atributos que se consideran (ciudad a la que pertenece, presupuesto, nombre del entrenador, puesto en el último campeonato, número de jugadores españoles...).

6.5. Bases de datos NoSQL

89

utilizan una estructura de datos en forma de tabla donde se van almacenando los datos, sino que para el almacenamiento hacen uso de otros formatos como clave–valor, mapeo de columnas o grafos.

Las bases de datos relacionales modernas normalmente han mostrado poca eficiencia en determinadas aplicaciones que usan los datos de forma intensiva, incluyendo el indexado de un gran número de documentos, la presentación de páginas en sitios que tienen gran tráfico y en sitios de streaming audiovisual. Las implementaciones típicas de los sistemas gestores de bases de datos realacionales (SGBDR) se han afinado, bien para una cantidad pequeña pero frecuente de lecturas y escrituras, bien para un gran conjunto de transacciones que tiene pocos accesos de escritura. Sin embargo, NoSQL puede servir gran cantidad de carga de lecturas y escrituras.

6.5.2. Necesidades no cubiertas por las bases de datos relacionales

Las bases de datos NoSQL son, principalmente, bases de datos distribuidas escalables horizontalmente y que no se basan en esquemas de datos predefinidos, por lo que ofrecen una fácil replicación y un conjunto sencillo de operaciones y consultas para el acceso a un gran volumen de datos.

La necesidad de este tipo de bases de datos surge porque hay fuentes de datos que son difíciles de modelar en bases de datos relacionales. Algunos ejemplos son texto (datos no estructurados), procesado en *streaming* (flujo continuo de datos) y bases de datos científicas (estructuras multidimensionales).

Como se avanzó en la Sec. 6.5.1, las características de las nuevas aplicaciones de Internet, como las redes sociales, juegos online, etc., hacen que las bases de datos NoSQL sean necesarias para conseguir mayor velocidad, escalabilidad, independencia de la localización, disponibilidad y mejor gestión, sea cual sea el tipo de datos:

- **Velocidad.** Para demostrar la importancia de la velocidad en internet, sirva como ejemplo cómo dos grandes compañías monetizan la velocidad de acceso. Por un lado, *Amazon* tiene estudiado que cuando el tiempo de respuesta disminuye 100ms los ingresos aumentan en un 1 %. Por otro lado, *Yahoo* asegura que el tráfico aumenta en un 9 % cuando el rendimiento mejora en 400ms. De ahí la importancia de esta V del *big data* para la evolución hacia bases de datos NoSQL.
- **Escalabilidad.** Al principio, la web se consideró una interfaz más, pero no es sólo eso; se ha convertido en un elemento generador y consumidor de datos (fundamentalmente semiestructurados y no estructurados). En el contexto actual, las compañías necesitan mantener una respuesta rápida, aunque se incremente el número de usuarios simultáneos o el volumen de datos manejado. Además, la arquitectura de las bases de datos NoSQL permite: (i) escalar sin disminuir el rendimiento; (ii) añadir nodos sobre la marcha, es decir, sin interrupciones del servicio; (iii) evitar que se generen cuellos de botella. La Fig. 6.1 (adaptada de Lo (2017)) muestra una representación comparativa de la escalabilidad de las bases de datos NoSQL frente a las relacionales. Como se observa, aunque las bases de datos relacionales tienen un mejor rendimiento para volúmenes de datos reducidos, éste se reduce drásticamente para grandes volúmenes. Mientras, el rendimiento de las bases de datos NoSQL tiende a ser constante, por lo que escalan mejor para datos masivos.

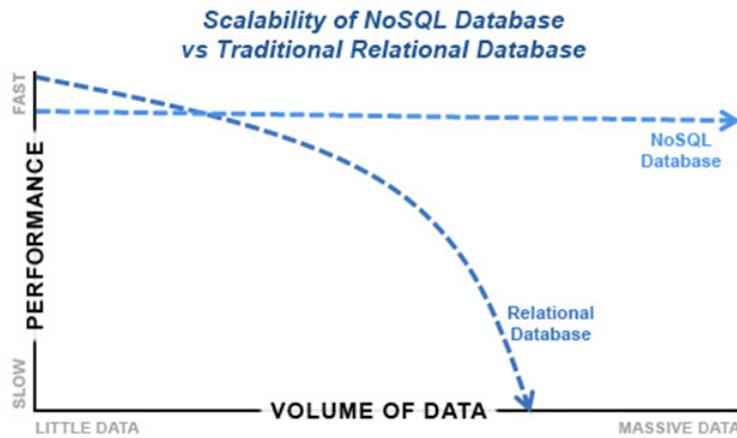


Figura 6.1: Comparativa de escalabilidad entre bases de datos relacionales y NoSQL.

Nota

Escalabilidad es un anglicismo que describe la capacidad de un negocio o sistema para crecer en magnitud.

- **Independencia de la localización.** La globalización del mercado en *World Wide Web* (WWW) obliga a dar servicio rápido y en todas partes del mundo. Las bases de datos no relacionales son distribuidas de acuerdo con diferentes arquitecturas como “*nodo principal y nodo secundario*”, o bien “*peer-to-peer*”.
- **Disponibilidad.** Similar a la independencia de la localización, la disponibilidad en el mercado WWW es uno de los factores más críticos, ya que se espera una disponibilidad de servicio 24x7. Es decir, hay que pasar de una alta disponibilidad a la disponibilidad continua, cuyas características son: (i) diseño que no sigue el modelo *principal-secundario*; (ii) centro multi-datos (*multi-data center*); (iii) disponibilidad *cloud*; (iv) copias de datos y funcionalidad en múltiples localizaciones.
- **Gestión de todos los tipos de datos.** Un factor clave en las bases de datos no relacionales es la necesidad de manejar tanto datos estructurados, como no estructurados y semiestructurados; y todo esto sin perder el enfoque de un almacenamiento eficiente. En ese sentido, a menudo, las bases de datos NoSQL están altamente optimizadas para las operaciones de recuperar y agregar, y normalmente no ofrecen mucho más que la funcionalidad de almacenar los registros (p.ej. almacenamiento clave-valor). La pérdida de flexibilidad en el tiempo de ejecución, comparado con las bases de datos SQL clásicas, se ve compensada por ganancias significativas en escalabilidad y rendimiento cuando se trata con ciertos tipos de almacenamiento de datos.

6.5.3. Tipos de almacenamiento en bases de datos NoSQL

Se pueden distinguir al menos 4 tipos de bases de datos NoSQL ([Hecht and Jablonski, 2011](#)): clave-valor, documental, en grafo y orientadas a columnas.

- **Almacenamiento clave-valor.** Los datos se almacenan de forma similar a los mapas o diccionarios de datos, donde se accede al dato a partir de una clave única. Los valores (datos) son aislados e independientes entre ellos, y no son interpretados por el sistema. Pueden ser variables simples, como enteros o caracteres, u objetos. Por otro lado, este sistema de almacenamiento carece de una estructura de datos clara y establecida, por lo que no requiere un formateo de los datos muy estricto. Son útiles para operaciones simples basadas en las claves. *Apache Cassandra* es la tecnología de almacenamiento clave-valor más reconocida por los usuarios.
- **Almacenamiento documental.** Este tipo de base de datos almacena datos semi-estructurados. Los datos se llaman documentos, y pueden estar formateados en XML (*Extensible Markup Language*), JSON (*JavaScript Object Notation*), BSON (*Binary JSON*) o el que acepte la propia base de datos, pero suele ser un formato de texto. Un ejemplo de cómo se usa es un blog: se almacena el autor, la fecha, el título, el resumen y el contenido del post. Todos los documentos tienen una clave única con la se puede acceder e identificarlos explícitamente. Estos documentos no son opacos al sistema, por lo que se pueden interpretar y lanzar consultas sobre ellos (véase Fig. 6.2). *CouchDB* o *MongoDB* son, quizás, los sistemas de bases de datos más conocidas. Hay que hacer mención especial a *MapReduce*, una tecnología de *Google* inicialmente diseñada para su algoritmo *PageRank*, que permite seleccionar un subconjunto de datos, agruparlos o reducirlos y cargarlos en otra colección, y a *Hadoop*, que es una tecnología de *Apache* diseñada para almacenar y procesar grandes cantidades de datos. Por ejemplo, *MongoDB* es una base de datos orientada a documentos. Los documentos se guardan en BSON, que es una forma de representar de forma binaria objetos JSON. De esta forma, con el comando `insert` y pasando un objeto JSON, *MongoDB* crea automáticamente un documento y lo añade en la base de datos generando un *ObjectId* para el nuevo documento ([Rubenfa, 2014](#)). Este objeto está especialmente pensado para garantizar unicidad en entornos distribuidos como *MongoDB*. El campo está compuesto por 12 bytes. Los cuatro primeros bytes son un *timestamp* con los segundos; los tres siguientes bytes representan el identificador único de la máquina; los dos siguientes el identificador del proceso; y, para finalizar, los últimos tres bytes son un campo incremental. En definitiva, los nueve primeros bytes garantizan un identificador único por segundo, máquina y proceso. Los tres últimos bytes garantizan que cada segundo se pueden insertar $2^{24} = 16.777.216$ documentos con un identificador distinto. Esta composición del *ObjectId* proporciona funcionalidades muy útiles. La primera es que indicar el orden de creación de los documentos. También sirve para obtener la fecha de creación del documento.
- **Almacenamiento en grafo.** Este tipo de almacenamiento maneja datos semi-estructurados y está basado en la teoría de grafos (véase Cap. 39). En las bases de datos NoSQL se establece que la información son los nodos y las relaciones entre la información



Figura 6.2: Ejemplo representativo de base de datos NoSQL documental. Adaptado de @Sanchez2017.

son las aristas (algo similar al modelo relacional). Su mayor uso se contempla en caso de tener que relacionar grandes cantidades de datos que pueden ser muy variables. Por ejemplo, los nodos pueden contener objetos, variables y atributos diferentes en unos y otros. Las operaciones de consulta con `join` se sustituyen por recorridos a través del grafo, y se guarda una lista de adyacencias entre los nodos. A modo de ejemplo, en *Facebook* se considera cada usuario como un nodo, que puede tener aristas de amistad con otros usuarios, o aristas de publicación con nodos de contenidos. Soluciones como *Neo4J* y *GraphDB* son las más conocidas dentro de las bases de datos orientadas a grafos.

- **Almacenamiento orientado a columnas.** Es similar al almacenamiento documental. Su modelo de datos se define como “un mapa de datos multidimensional poco denso, distribuido y persistente” (Hecht and Jablonski, 2011). Se orienta a almacenar datos con tendencia a escalar horizontalmente, por lo que permite guardar diferentes atributos y objetos bajo una misma clave. A diferencia del documental y del clave-valor, en este caso se pueden almacenar varios atributos y objetos, pero no serán interpretables directamente por el sistema. Permite agrupar columnas en familias y guardar la información cronológicamente, mejorando el rendimiento. Esta tecnología se suele usar en casos con 100 o más atributos por clave. Su precursor es *BigTable* de Google, pero han aparecido nuevas soluciones como *HBase* o *HyperTable*.

6.5.4. Limitaciones de las bases de datos NoSQL

Las bases de datos NoSQL no sólo tienen ventajas; también tienen algunas limitaciones, tanto técnicas como de carácter no tecnológico.

Entre las técnicas se encuentran: (i) cómo se modelan los datos correctamente para maximizar las capacidades; (ii) nivel bajo de seguridad; (iii) no soporte de transacciones; (iv) falta de madurez en *Business Intelligence*; y (v) problemas de compatibilidad ya resueltos en los modelos relacionales. Entre las de carácter no tecnológico pueden citarse (i) la falta de expertos; (ii) la

resistencia al cambio; (iii) la disponibilidad del vendedor; y (iv) que el código abierto puede implicar problema de soporte para las empresas.

6.6. Integración de bases de datos NoSQL en R

En esta sección se verá como **R** puede ser utilizado para conectarse a una base de datos NoSQL: en particular a MongoDB. En la Sec. 6.6.1 se presenta una introducción a MongoDB. En la Sec. 6.6.2 se explican los paquetes de **R** utilizados para acceder a MongoDB. La Sec. 6.6.3 indica cómo conectarse a una base de datos MongoDB remota. Las secciones 6.6.4 y 6.6.5 realizan consultas y análisis sobre una colección de viajes realizados por los usuarios de un servicio de bicicletas compartidas con sede en la ciudad de Nueva York.

6.6.1. Introducción a MongoDB

MongoDB (de la palabra inglesa “*humongous*”, que significa *enorme*) es un sistema de base de datos NoSQL orientado a documentos, desarrollado bajo el concepto de código abierto. MongoDB forma parte de la nueva familia de sistemas de bases de datos NoSQL. En lugar de guardar los datos en tablas como se hace en las bases de datos relacionales, MongoDB guarda estructuras de datos en documentos similares a JSON con un esquema dinámico (MongoDB utiliza una especificación llamada BSON), haciendo que la integración de los datos en ciertas aplicaciones sea más fácil y rápida. MongoDB soporta la búsqueda por campos, consultas de rangos y expresiones regulares. Las consultas pueden devolver un campo específico del documento, pero también puede ser una función JavaScript definida por el usuario. Cualquier campo en un documento de MongoDB puede ser indexado, al igual que es posible hacer índices secundarios. El concepto de índices en MongoDB es similar a los encontrados en bases de datos relacionales. Tecnológicamente, MongoDB es una base de datos multiplataforma, orientada a documentos, que brinda alto rendimiento, alta disponibilidad y facilita la escalabilidad.

MongoDB trabaja con el concepto de colección y documento. La Tabla 6.2 muestra la relación de esta terminología respecto a las bases de datos relacionales.

Tabla 6.2: . Diferencias entre conceptos y terminología en bases de datos relacionales y MongoDB.

Bases de datos relacionales	MongoDB
Bases de datos	Bases de datos
Tabla	Colección
Tupla o fila	Documento
Tabla <i>Join</i>	Documentos incrustados (embeded)
<i>Primary Key</i>	Por defecto, <i>key_id</i> (gestionada por MongoDB)

Asimismo, MongoDB proporciona una función, **MapReduce**, que se puede utilizar para el procesamiento por lotes de datos y operaciones de agregación. El framework de agregación permite

realizar operaciones similares a las que se obtienen con el comando SQL GROUP BY. El framework de agregación está construido como un *pipeline* (flujo de trabajo) en el que los datos van pasando a través de diferentes etapas en las cuales estos datos son modificados, agregados, filtrados y formateados hasta obtener el resultado deseado (véase ejemplo esquemático en la Fig. 6.3 (Morgan (2015))). Todo este procesado es capaz de utilizar índices, si existieran, y se produce en memoria.

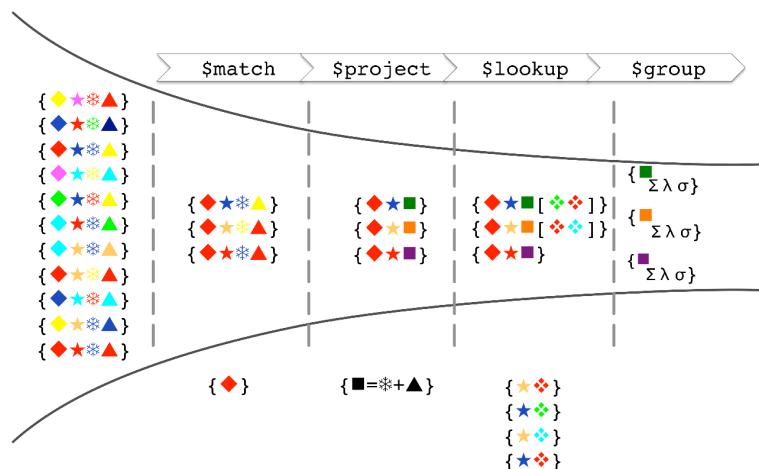


Figura 6.3: Ejemplo esquemático del *pipeline* de agregación en MongoDB. Adaptado de Morgan (2015)

6.6.2. Plataforma tecnológica para el caso práctico

Para la realización del caso práctico se utiliza **Atlas**[^nosql-atlas], un servicio en la nube gratuito para manejar bases de datos MongoDB. Atlas es fácil de configurar y tiene conjuntos de datos de muestra para ejemplos de **R** con MongoDB. Puede cargar conjuntos de datos de muestra usando el botón “...” junto al de colecciones en la página de su cluster (el servidor). No obstante, aunque en Atlas se puede crear un cluster específico, en este ejemplo práctico se parte de uno ya creado.

Adicionalmente, a modo de apoyo, se recomienda utilizar una herramienta **cliente** para conectarse a MongoDB y poder inspeccionar los datos contenidos. Es muy útil para realizar las consultas. Puede considerarse la herramienta para la gestión de la instalación de MongoDB[^nosql-mongod]. Además, si se crea el propio cluster en Atlas, éste tiene una interfaz amigable para inspeccionar los datos.

Como complemento a estas funciones, existe documentación de las colecciones de documentos y la información contenida en esta base de datos[^nosql-ejem] de ejemplo.

Para la resolución de ejercicios puede consultarse el Manual de MongoDB[^nosql-mongo2], que contiene ejemplos y explicaciones de la sintaxis de MongoDB.

6.6.2.1. Paquetes R utilizados

El controlador **R** preferido por la comunidad de MongoDB es **mongolite**; por ello se utiliza en los siguientes ejemplos. Es rápido y tiene una sintaxis similar a la del *shell* MongoDB. **Mongolite** es el controlador **R** para MongoDB más reciente y puede realizar operaciones de indexación, canalizaciones de agregación, cifrado TLS (*transport layer security*) y autenticación SASL (*simple authentication and security layer*), entre otras. Está basado en los paquetes **jsonlite** y **mongo-c-driver**. Se puede instalar desde CRAN o desde RStudio. Existen otros paquetes para conectar MongoDB y **R**, como, por ejemplo **RMongo** y **rmongodb**, aunque no han estado muy activos GitHub últimamente, por lo que carecen de soporte.

Para poder usar el paquete **mongolite** hay que instalarlo previamente, además de importar la librería con el siguiente comando.

```
library('mongolite')
```

6.6.3. Conexión y acceso a MongoDB desde R

La variable **cadenaConexion** representa la cadena de conexión a MongoDB en Atlas. Si se desea, se puede sustituir por otro servidor o cluster en Atlas, o por un servidor local.

```
cadenaConexion <- "mongodb+srv://user01:user01@cluster0.myclc3z.mongodb.net/test"
```

En opciones de seguridad se establece la no validación de certificados SSL (*secure sockets layer*), para evitar que exista un error de conexión a Atlas.

```
opcionesConexion <- ssl_options(weak_cert_validation = T)
```

Nota

En entornos reales de producción está desaconsejado evitar esta comprobación por razones de seguridad.

Después de establecer la conexión a MongoDB se recupera la colección *trips* (colección de viajes de la base de datos **sample_training**) con la función **mongo()** en código **R**. Esta colección contiene datos de viajes realizados por los usuarios de un servicio de bicicletas compartidas con sede en la ciudad de Nueva York.

```
viajes <- mongo(collection = "trips", db = "sample_training", url = cadenaConexion,
                  options = opcionesConexion)
```

Se puede verificar que el código está conectado a la colección mediante la consulta del número total de documentos en esta colección. Para hacerlo, se usa la función **count()**.

```
viajes$count()
#> [1] 10000
```

6.6.4. Obtención de datos en R desde MongoDB

Ahora que hay una conexión establecida con la base de datos, se pueden leer los datos de la misma para ser procesados por **R**. Para recuperar datos de MongoDB y mostrarlos se puede usar la interfaz de usuario de Atlas (en este caso para ver los documentos de *trip_collection*). Se puede obtener cualquier documento de muestra de la colección usando la función `$iterate().$one()`, pudiéndose, así, examinar la estructura de los datos de la colección.

Una vez se conoce la estructura de los documentos, se pueden realizar consultas más avanzadas, como buscar los tres viajes más largos y luego enumerar la duración en orden descendente. La consulta propuesta utiliza operadores de clasificación y límite⁴ para producir este conjunto de resultados.

```
viajes$find(sort = '{"tripduration" : -1}', limit = 3)
#>   tripduration start station id      start station name end station id
#> 1     326222           391      Clark St & Henry St      310
#> 2     279620           3165 Central Park West & W 72 St    3019
#> 3     173357           3155 Lexington Ave & E 63 St    3083
#>   end station name bikeid  usertype birth year
#> 1     State St & Smith St  18591 Subscriber    1979
#> 2     NYCBS Depot - DEL  17547 Customer
#> 3 Bushwick Ave & Powers St  15881 Customer
#>   start station location.type start station location.coordinates
#> 1                   Point      -73.99345, 40.69760
#> 2                   Point      -73.97621, 40.77579
#> 3                   Point      -73.96649, 40.76440
#>   end station location.type end station location.coordinates
#> 1                   Point      -73.98913, 40.68927
#> 2                   Point      -73.98193, 40.71663
#> 3                   Point      -73.94100, 40.71248
#>   start time          stop time
#> 1 2016-01-01 01:58:20 2016-01-04 20:35:23
#> 2 2016-01-02 17:07:26 2016-01-05 22:47:46
#> 3 2016-01-02 15:25:36 2016-01-04 15:34:53
```

6.6.5. Análisis de datos de MongoDB en R

Para analizar datos de MongoDB en **R** con más detalle, se puede usar el marco de agregación de datos de MongoDB . Este marco permite a los operadores crear canalizaciones de agregación que ayuden a obtener los datos con una sola consulta.

⁴<https://jeroen.github.io/mongolite/query-data.html#sort-and-limit>

6.6. Integración de bases de datos NoSQL en R

97

Para saber cuántos suscriptores realizaron viajes de una duración mayor que 240 segundos y regresaron a la misma estación donde comenzaron, la consulta usa la cláusula `$expr`, que compara dos campos en el mismo documento.

```
query <- viajes$find('{"usertype":"Subscriber","tripduration":{$gt:240}, "$expr":  
  {$eq: ["$start station name","$end station name"]}}')
```

Combinando estos operadores con código R, también se puede ver, por ejemplo, qué tipo de usuarios es más común: *suscriptores* o *clientes únicos*. Para ello, se pueden agrupar usuarios por el campo `usertype`, que define el tipo de usuario.

```
tipos_usuario <- viajes$aggregate('[$group:{ "_id": "$usertype", "Count":  
  {$sum:1}}]')
```

Para comparar los resultados, se pueden visualizar (véase Fig. 6.4). Es conveniente convertir los datos obtenidos de `mongolite` en un `data.frame`(marco de datos) y, por ejemplo, usar el paquete `ggplot2`, para trazar estos datos.

```
library("ggplot2")  
  
df <- as.data.frame(tipos_usuario)  
  
ggplot(df, aes(x = reorder(`_id`, Count), y = Count)) +  
  geom_bar(stat = "identity", color = "blue", fill = "green") +  
  geom_text(aes(label = Count)) +  
  coord_flip() +  
  xlab("Tipo de usuario")
```

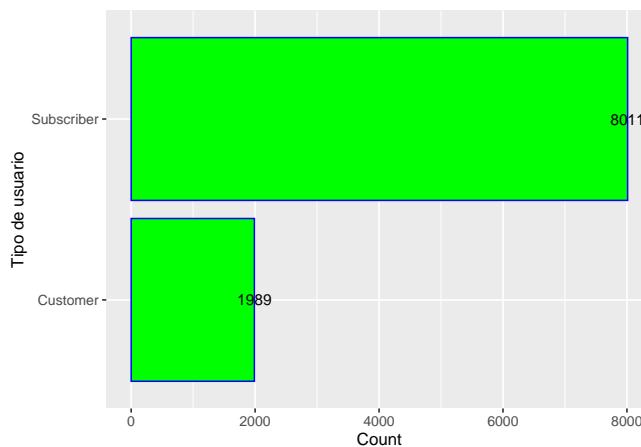


Figura 6.4: Suscripción por tipo de usuario.

Resumen

En este capítulo se presenta el concepto de *big data*, por qué surje y qué aporta respecto a soluciones previas.

En particular, se discute qué son las bases de datos NoSQL y cuáles son sus diferencias con las bases de datos relacionales (más tradicionales). Posteriormente, se explican algunas limitaciones de las bases de datos NoSQL. Finalmente, como ejemplo, se muestra la integración de datos en **R** desde MongoDB, explicando cómo acceder a este tipo de datos y cómo analizarlos en el caso concreto de una base de datos documental como MongoDB.

Capítulo 7

Gobierno, gestión y calidad del dato

Ismael Caballero^a, Ricardo Pérez del Castillo^a, y Fernando Gualo^{a,b}

^aUniversidad de Castilla-La Mancha ^bDQTeam SL

7.1. Introducción

Los datos se han convertido en un elemento vital para el desarrollo económico de las organizaciones, ya que permiten una mayor eficiencia en el uso de los recursos y un aumento de su productividad. Tanto es así, que la Unión Europea establece, a través de la [Estrategia Europea de Datos](#)¹, que en 2030 se establecerá un Espacio Único Europeo de Datos para fomentar un ecosistema con nuevos productos y servicios basados en los datos. Para ello, en esta Estrategia Europea de Datos -que prevé un incremento del 530 % del volumen global de datos- se reclama la necesidad de implantar **mecanismos de gobierno del dato** a través de políticas y directrices consensuadas a alto nivel para alcanzar los objetivos de la [estrategia organizacional](#) y satisfacer tanto aspectos regulatorios genéricos (como las leyes europeas [General Data Protection Regulacy \(GDPR\)](#)² o [Data Governance Act](#)³, o las españolas [Esquema Nacional de Seguridad -ENS](#)⁴, o [Esquema Nacional de Interoperabilidad -ENI](#)⁵) como aspectos sectoriales específicos (como [Solvencia II](#)⁶ para el sector seguros, o Basilea III⁷ para el sector financiero).

¹https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_es

²GDPR: <https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:32016R0679&from=EN>

³Data Governance Act:<https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:52020PC0767&from=EN>

⁴ENS: <https://www.boe.es/buscar/act.php?id=BOE-A-2022-7191>

⁵ENI: <https://www.boe.es/buscar/doc.php?id=BOE-A-2010-1331>

⁶Solvencia II: https://www.eiopa.europa.eu/browse/solvency-2_en

⁷Basilea III:<https://www.bis.org/bcbs/basel3.htm>

Estos mecanismos de gobierno del dato deben abordar aspectos verticales relacionados con la adquisición, tenencia, compartición, uso y explotación de los datos en los procesos de negocio, abordando a la vez aspectos transversales relacionados con su gestión: calidad de los datos, aspectos éticos y privacidad, interoperabilidad, gestión del conocimiento y el control sobre los activos de datos a través de las políticas correspondientes, y despliegue de estructuras organizativas con una conveniente separación de los roles de gobierno del dato de los de gestión del dato (ISO, 2017). Por tanto, puede decirse que el **gobierno del dato** marca la dirección de cómo la organización debe realizar la **gestión del dato** para alcanzar los objetivos establecidos en su(s) estrategia(s) del dato. Esto se consigue mediante la definición e implementación de una serie de políticas del dato.

7.2. Concepto de gobierno del dato

El gobierno de los datos se ha convertido en un habilitador de la economía de los datos (Engels, 2019; Weber et al., 2009), así como también en un pilar básico para la mejora de la transparencia y eficiencia de las administraciones públicas (OECD, 2019; Osimo et al., 2020; Osorio-Sanabria et al., 2020). Aunque existen algunas aproximaciones académicas y profesionales al gobierno del dato, no hay una definición consensuada de este concepto que permita aunar las distintas visiones. No obstante, la definición más aceptada de gobierno del dato es la propuesta por DAMA en DMBoK2 (DAMA, 2017): “colección de prácticas y procesos que ayudan a asegurar la gestión formal de los activos de datos dentro de una organización mediante el ejercicio de autoridad, control y toma de decisiones compartidas, planificadas, monitorizadas y forzadas”. Teniendo en cuenta los matices que introduce, es también interesante la lectura de la propuesta por Soares (2015), que define **gobierno del dato** como “la formulación de políticas para optimizar, conseguir los niveles adecuados de seguridad y protección, y potenciar los datos como activos organizacionales mediante la alineación de los objetivos de diferentes funciones organizacionales; por su naturaleza, el gobierno del dato requiere cooperación interdepartamental para entregar oportuna y fielmente datos con el máximo valor para la toma de decisiones en la organización”.

De alguna manera, se podría entender que gobernar los datos implica **el diseño, implementación y mantenimiento de un sistema de gobierno del dato**. El gobierno del dato tiene tres características destacables (Caballero et al., 2022b):

- **Está dirigido por el valor de los datos:** pues el principal objetivo del gobierno del dato es asegurar que los datos son tratados como activos de datos y que la gestión y uso que se hace de ellos permite alcanzar el máximo valor organizacional que se espera de ellos. Por tanto, todas las acciones están encaminadas a la obtención de este valor organizacional.
- **Está centrado en la arquitectura empresarial:** para poder gobernar los datos adecuadamente, es preciso revisar o incorporar ciertos componentes a la arquitectura empresarial tal que, o bien den el soporte adecuado, o bien forme parte del resultado del gobierno del dato.
- **Es iterativo e incremental:** pues para alcanzar un estado en el que se considere que los activos de datos están perfectamente gobernados es preciso desarrollar un conjunto

de programas de gobierno del dato. Así, a través de la ejecución de diferentes proyectos relacionados entre sí, se conseguirá el desarrollo y la puesta en valor de los artefactos típicos de un sistema de gobierno del dato (véase sección 7.2.2). Esto sólo se puede conseguir en incrementos relevantes (p.ej. la creación de más componentes del sistema de gobierno del dato o la inclusión de nuevos datos a ser gobernados en el alcance de gobierno del dato).

Un aspecto interesante es que, a medida que se avanza en la ejecución de estos programas de gobierno del dato, más sensible se vuelve la organización hacia la importancia de los datos, más aprende a gestionarlos y gobernarlos, y más amplio es el alcance del gobierno del dato; en definitiva, se puede decir que más madura se vuelve la organización en lo que se refiere al gobierno y a la gestión del dato.

7.2.1. Beneficios del gobierno del dato

Cuando se desarrolla un sistema de gobierno del dato, con cada incremento del sistema se espera conseguir uno o una combinación de los siguientes beneficios (ISACA, 2019):

- (1) **Alineamiento estratégico:** optimización del valor organizacional de los datos mediante el alineamiento con la estrategia organizacional.
- (2) **Realización de beneficios:** aseguramiento de que los datos son entregados en condiciones aceptables a los diferentes consumidores del dato.
- (3) **Optimización de riesgos:** paliar o minimizar, dentro de la propensión al riesgo de la organización, los riesgos relacionados con la adquisición, uso y explotación de los datos, asegurando el cumplimiento de la normativa interna y regulatoria.
- (4) **Optimización de recursos:** optimización de las capacidades de los recursos humanos y tecnológicos necesarios, y que son utilizados para dar un soporte más eficiente a las distintas operaciones involucradas en la gestión del dato, minimizando el desperdicio de recursos al gestionar, usar y explotar los datos.

Estos beneficios deben especificarse como parte de la estrategia del dato de la organización. Así, por ejemplo, una organización que considere realizar un alineamiento estratégico y una optimización de riesgos, estará desarrollando una **estrategia defensiva** que debería implementarse a través de un **gobierno técnico**; por otro lado, si una organización quiere, por ejemplo, maximizar la realización de beneficios, podría considerarse que estaría trazando una **estrategia ofensiva** que se podría materializar mediante un **gobierno para el valor**.

7.2.2. Componentes de un sistema de gobierno del dato

Para poder obtener los beneficios descritos anteriormente, las organizaciones deben realizar esfuerzos para implantar los mecanismos de gobierno del dato reclamados en la Estrategia Europea de Datos, particularizándolos a su realidad y en función de su madurez. Estos mecanismos implican el desarrollo de un **sistema de gobierno del dato**, que involucra la creación o mantenimiento de forma interrelacionada y sujeto a las restricciones correspondientes de una serie de componentes. Dependiendo de si se tiene un gobierno técnico o un gobierno para el

valor, la creación y uso de los distintos tipos de componentes será más o menos intensiva. Estos componentes son los siguientes ([Caballero et al., 2022b](#)):

1. **Procesos de gestión del dato, gestión de calidad del dato y gobierno del dato**, que se refieren al diseño y posterior particularización e implantación de las buenas prácticas relacionadas con las tareas típicas de los datos a nivel de las correspondientes disciplinas. Es posible obtener descripciones genéricas de estos procesos en diferentes modelos de referencia de procesos, tales como **Data Maturity Model (DMM)**([Mecca et al., 2014](#)), **The Data Management Capability Assessment Model (DCAM)**([Council, 2020](#)) o el **Modelo Alarcos de Madurez de Datos (MAMD)**([Caballero et al., 2023](#))
2. **Estructuras organizacionales**, que deben recoger las cadenas de responsabilidades y rendición de cuentas, haciendo una adecuada separación entre las responsabilidades propias del gobierno del dato y aquellas propias de la gestión del dato y de su calidad. Los roles que deben asumir estas responsabilidades son típicamente el de *Chief Data Officer* ([Soares, 2015](#); [Treder, 2020](#)), con un punto de vista más ejecutivo/estratégico, y los de *data stewards* ([Plotkin, 2020](#)), desde una perspectiva más táctica/operativa.
3. **Principios, políticas y marcos de referencia**, que deberían incluir todos los principios rectores en los que se basará el uso de los datos (tales como los *Generaly Accepted Information Principles* listados en [Ladley \(2019\)](#)), las directrices o políticas y los controles correspondientes asociados necesarios para modelar y gestionar el valor de los datos, el riesgo a asumir y las restricciones a considerar según se describe en ISO/IEC 38505-2 ([ISO, 2018](#)).
4. **Datos e Información**, que se deben gobernar como las descripciones necesarias a través de los metadatos correspondientes. Para la parte del dato es fundamental poder establecer una adecuada **arquitectura del dato** con los correspondientes modelos que recojan la semántica del entorno de la organización y reflejen el cómo ésta usa los datos para desarrollar su actividad organizacional y/o económica. Para dar soporte al uso correspondiente, deben generarse y mantenerse los metadatos correspondientes, que pueden ser de varios tipos ([DAMA, 2017](#)):
 - **metadatos de negocio**, recogidos típicamente en *glosario de negocio* y que describen la relación del dato con el negocio;
 - **metadatos técnicos**, recogidos habitualmente en los *catálogos de datos* , que describen detalles técnicos de los datos; y
 - **metadatos operacionales**, recogidos típicamente en los *diccionarios de datos* , que recogen aspectos relacionados con el procesamiento y acceso a los datos. Es importante que todos estos metadatos estén reconciliados convenientemente entre ellos, ya que su visión conjunta permitirá una descripción adecuada de los datos bajo el gobierno del dato, y si esta descripción es suficiente será posible usarlos con las suficientes garantías de éxitos.
5. **Cultura, ética y comportamiento**, cuyo objetivo es identificar aquellos aspectos culturales y éticos que deben regular la forma en la que la organización abordará las tareas relacionadas con los datos para que estos tengan el valor organizacional deseado ([Harrison et al., 2019](#)).

6. **Personas, habilidades y competencias**, componente que trata de organizar los roles que deben asumir las diferentes responsabilidades relacionadas con los diferentes procesos; también debe enfocarse en asegurar que esos roles tienen los conocimientos, habilidades y competencias necesarias para abordar las tareas asociadas mediante los programas formativos correspondientes; finalmente, este artefacto incluye asegurar que la organización tiene planes de contingencia ante la eventual rotación funcional de los recursos humanos dedicados a las responsabilidades relacionadas con los datos (Plotkin, 2020).
7. **Servicios, infraestructuras y aplicaciones**: este componente aborda todo lo relacionado con las tecnologías y sistemas de información para dar soporte a las diferentes actividades de los procesos de gestión del dato, así como de la gestión de su calidad y de su gobierno.

7.3. Marcos y metodologías de gobierno del dato

En la literatura, tanto académica como profesional, existen algunas propuestas de creación de sistemas de gobierno del dato. Es interesante resaltar que en el ámbito académico se han desarrollado algunas revisiones sistemáticas de literatura científica para identificar los componentes del gobierno del dato, aunque de forma general, y salvo algunas referencias, se mantienen desconectados de las propuestas profesionales. También es importante mencionar que salvo COBIT 2019 (*Control Objectives for Information and related Technology, Objetivos de control para la información y tecnologías relacionadas*)⁸, la inmensa mayoría de estos marcos no identifican explícitamente el concepto de “**sistema de gobierno del dato**”, sino que se establece bajo un paraguas más genérico de “**gobierno del dato**”. En cualquier caso, la idea es la misma.

En los siguientes párrafos se resumen los aspectos más importantes de los marcos más relevantes, que pueden servir para que el lector encuentre el que mejor se adapta a su circunstancia profesional:

- Abraham et al. (2019) analizan la literatura para identificar los elementos de un marco de trabajo teórico que se clasifican en torno a cuatro áreas del gobierno del dato: (1) alcance organizacional (aspectos intra e inter organizacionales), (2) alcance de los datos (datos tradicionales vs. big data), (3) alcance del dominio (calidad del dato, seguridad del dato, arquitectura del dato, ciclo de vida, metadatos, almacenamiento e infraestructura del dato), y (4) mecanismos de gobierno (estructurales, procedimentales y relacionales).
- Al-Ruithe et al. (2019) identifican, también a través de una revisión sistemática de literatura, las áreas o retos del gobierno del dato donde merece la pena investigar. Éstas son tecnología (seguridad, privacidad, disponibilidad, rendimiento, clasificación de los datos y migración del dato), legalidad, y aspectos organizacionales o del negocio.
- Brous et al. (2016) derivan, de nuevo basándose en una revisión sistemática de la literatura, los principios para desarrollar de forma efectiva estrategias y aproximaciones para el gobierno del dato, que agrupan en torno a cuatro conceptos fundamentales: (1) organización, (2) alineamiento, (3) cumplimiento y (4) entendimiento común de los datos.

⁸<https://www.isaca.org/resources/cobit>

- Carruthers and Jackson (2020) identifican los posibles elementos que deben contemplarse en la transformación digital, la cual debe apoyarse en el gobierno del dato. Estos elementos (personas, datos, procesos, tecnologías) son representados mediante un triángulo, en cuyo centro están los datos. En la continuación de la obra de estos autores presentada en Jackson and Carruthers (2019), se propone un modelo de transformación, convenientemente soportado en el gobierno del dato.
- DCAM (*Data Management Capability Assessment Model*) (Council, 2020) es un modelo de referencia para la evaluación de la capacidad de gestión del dato desarrollado por el EDM Council. El modelo tiene ocho componentes agrupados en cuatro niveles (1) fundamentos (estrategia del dato y casos de negocio; programas de gestión del dato y financiación), (2) ejecución (arquitectura del dato y de negocio; arquitectura del dato y de tecnología; gestión de calidad del dato; gobierno del dato), (3) colaboración (entorno de control del dato) y (4) formalización del diseño e implementación de las actividades analíticas.
- DMBoKv2 (*Data Management Body of Knowledge*) (DAMA, 2017) es un marco de referencia de procesos desarrollado por DAMA que posiciona el gobierno del dato como la función que guía el resto de las acciones relacionadas con la gestión del dato. Identifica una serie de elementos que deben generarse a partir del gobierno del dato: estrategia de gobierno del dato; estrategia del dato; hoja de ruta del gobierno del dato; principios de gobierno del dato, políticas de gobierno del dato, procesos; marco operativo de gobierno del dato; hoja de ruta y guía de implementación; plan de operaciones; glosario de términos; plan de operaciones; cuadro de mando de gobierno del dato; etc.
- Eryurek et al. (2021) identifica los “ingredientes” propios de un sistema de gobierno del dato (herramientas; personas y procesos; cultura del dato), así como las áreas en las que debería enfocarse el gobierno de datos a lo largo del ciclo de vida de los datos (descubrimiento y limpieza del dato; gestión del dato; políticas de privacidad, seguridad y acceso).
- COBIT 2019 (ISACA, 2019) identifica para el sistema de gobierno de tecnologías y de información los siguientes componentes: procesos; estructuras organizacionales; principios, políticas y marcos de referencia; información; cultura, ética y comportamiento; personas, habilidades y competencias; servicios, infraestructuras y aplicaciones.
- ISO 38505-1 (ISO, 2017) e ISO 38505-2 (ISO, 2018) muestran los aspectos claves del gobierno del dato (valor de los datos, riesgo, y restricciones) e introducen seis principios (responsabilidad, estrategia, adquisición, rendimiento, cumplimiento y comportamiento humano). Identifica una serie de procesos (evaluar, dirigir, monitorizar) como áreas propias de actuación del gobierno del dato, distinguiéndolos de las operaciones propias de la gestión del dato y estableciendo las correspondientes relaciones con ellas. Sin embargo, no describen actividades específicas para la creación de sistemas de gobierno del dato.
- Khatri and Brown (2010) aducen que el gobierno del dato implica tomar decisiones sobre activos claves de datos en varios dominios de decisión (principios, gestión de calidad del dato, metadatos, acceso a datos y ciclo de vida de los mismo).
- Janssen et al. (2020) exploran las capacidades de gobierno del dato necesarias para que las organizaciones dirigidas por datos puedan extraer el máximo beneficio de los sistemas algorítmicos basados en Big Data (*Big Data Algorithmic Systems*) y proponen un marco para la creación del gobierno del dato que permita optimizar estos sistemas.
- Ladley (2019) presenta un marco de gobierno del dato basado en cinco pilares: compromiso, estrategia, arquitectura y diseño, implementación y operación, y, por último, gestión del cambio.

7.4. Gestión de calidad del dato

105

- [Lillie and Eybers \(2019\)](#) estudian la literatura existente para identificar los aspectos más interesantes sobre (1) el alcance y los constructos más importantes del gobierno y gestión del dato, y (2) las capacidades ágiles requeridas en el gobierno y la gestión de datos.
- En el llamado **proceso Unificado de Gobierno del dato** de IBM ([Soares, 2010](#)) se identifican cinco **ingredientes claves** que deberían ser cubiertos por cualquier marco de gobierno del dato: (1) fuerte respaldo por parte de la organización con soporte de las TI, (2) centrarse en los elementos de datos críticos, (3) énfasis en los artefactos de datos, (4) alineación en torno a métricas y aplicación de políticas, y (5) celebración de las victorias rápidas conseguidas como hitos en una hoja de ruta a largo plazo.
- La Organización para la Cooperación y el Desarrollo Económicos (*Organisation for Economic Co-operation and Development* (OECD), en su informe sobre gobierno del dato para administraciones públicas ([OECD, 2019](#)), recoge las mejores prácticas llevadas a cabo por diferentes administraciones de los países que la componen en lo que se refiere a transparencia del gobierno del dato e incremento del valor de la información disponible sobre la ciudadanía de cara a una mejor prestación de servicios públicos.
- [Treder \(2020\)](#) identifica algunos componentes específicos que debería tener un sistema de gobierno del dato (cadenas de valor; estrategia de dato; procesos de datos; descripción de los roles y sus responsabilidades; gestión del equipo de la oficina del dato), así como las áreas en las que debe enfocarse el gobierno del dato (casos de negocio; aspectos éticos y cumplimiento; gestión y análisis del dato).

A modo de ejemplo, se van a introducir más detalles sobre un marco de referencia basado en estándares internacionales ISO: el **Modelo Alarcos de Madurez de Datos (MAMD) v4.0** ([Caballero et al., 2023](#)). MAMDv4.0 es un marco de trabajo que se usa para la evaluación y mejora de la capacidad de los procesos de la organización relacionados con la gestión, la gestión de la calidad, y el gobierno del dato. Tiene dos componentes principales:

- Un **modelo de referencia de procesos (MRP)**, que contiene una descripción de los procesos de gestión del dato, de gestión de calidad del dato y de gobierno del dato. Está alineado con los principales estándares en el área (ISO 8000-61 ([ISO, 2016](#)), e ISO/IEC 38505-2 ([ISO, 2018](#))), así como con las buenas prácticas de otros modelos como DAMA, DMM o COBIT 2019 (Véase Fig.[7.1](#)).
- El **modelo de evaluación de procesos (MEP)**, que sigue las directrices de evaluación y los niveles de capacidad y madurez descritas por ISO/IEC 33000 y adaptadas a la evaluación de procesos de datos conforme al modelo de madurez propuesto en ISO 8000-62 ([ISO, 2018](#)) (véase Fig. [7.2](#)).

7.4. Gestión de calidad del dato

Los datos con niveles inadecuados de calidad acaban teniendo un impacto negativo para las organizaciones, bien en términos económicos, bien en términos de reputación ([Redman, 2016](#)).

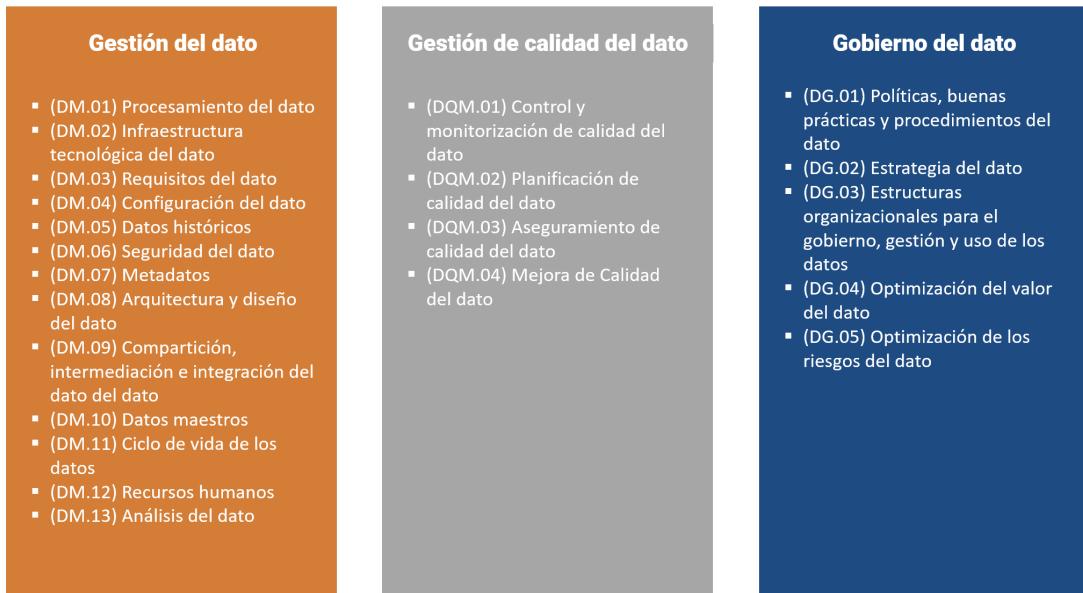


Figura 7.1: Modelo de Referencia de Procesos de MAMD; DM: gestión del dato; DQM: gestión de calidad del dato; DG: gobierno del dato

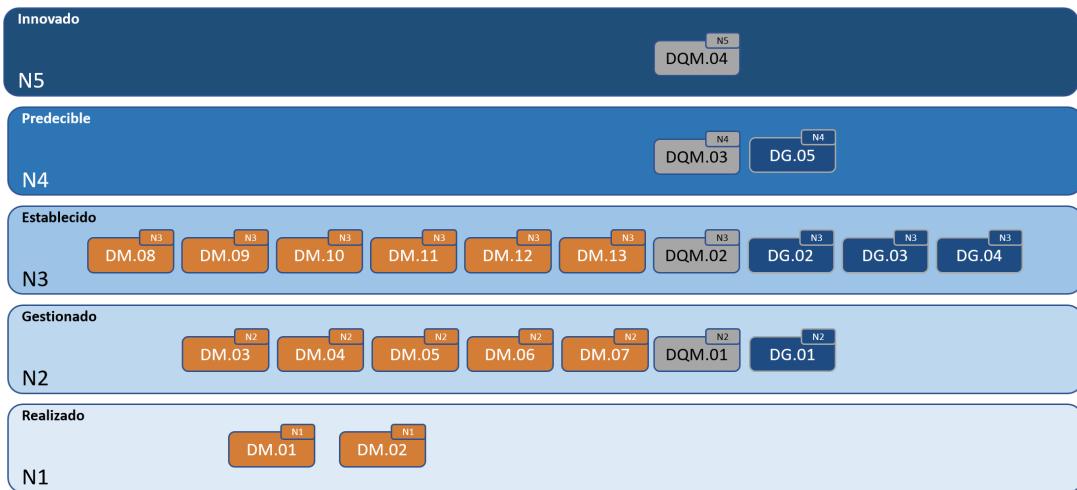


Figura 7.2: Modelo de Madurez Organizacional de MAMD; N: nivel; DM: gestión del dato; DQM: gestión de calidad del dato; DG: gobierno del dato

7.4. Gestión de calidad del dato

107

Por eso es importante que las organizaciones cuiden del nivel de calidad de sus datos y se aseguren que dicho nivel permanece dentro de los permitidos para que la ejecución de los procesos de negocio se haga dentro del margen de riesgo de la organización.

Se dice que un conjunto de datos tiene calidad cuando sirve para el propósito para el que fue recogido (*fitness for use*) (Strong et al., 1997b). Para determinar si un conjunto de datos tiene calidad suficiente para dicho propósito, es preciso identificar y seleccionar un conjunto de criterios (llamados en la literatura **dimensiones** (Wang, 1998), o **características de calidad del dato** (ISO/IEC, 2008a)) que permitan determinar si dicho conjunto cumple los requisitos de calidad que exige el usuario de tales datos. Al conjunto de dimensiones o características de calidad del dato seleccionadas se le denomina **modelo de calidad del dato**. La Tab. 7.1 introduce una descripción de las características de calidad del dato incluidas en el estándar ISO/IEC 25012.

Característica	Definición	Inherente	Dependiente del sistema
Exactitud	Grado en que los datos tienen atributos que representan correctamente el valor de un atributo	X	
Compleitud	Grado en el que existen suficientes valores para todos los atributos necesarios para la representación de una entidad	X	
Consistencia	Grado en que los datos están libres de contradicción y son coherentes con el resto de los datos de su contexto de uso	X	
Credibilidad	Grado en que los datos se consideran ciertos y creíbles por los usuarios	X	
Actualidad	Grado en que los datos tienen atributos con las fechas y tiempos correctos	X	
Accesibilidad	Grado en que los datos pueden ser accedidos por cualquier usuario	X	X
Cumplimiento	Grado en el que los datos están construidos conforme estándares, convenciones o reglas establecidas	X	X
Confidencialidad	Grado en el que los datos tienen atributos específicos que solo pueden ser accedidos por usuarios autorizados	X	X
Eficiencia	Grado en el que los datos tienen atributos que pueden ser procesados y provistos dentro de los niveles de rendimiento esperados	X	X
Exactitud	Grado en el que los datos tienen atributos que son precisos	X	X
Trazabilidad	Grado en el que los datos tienen atributos que proveen información detallada sobre los cambios realizados en los datos	X	X
Comprendibilidad	Grado en el que los datos se expresan de manera que los usuarios puedan leerlos e interpretarlos correctamente	X	X
Disponibilidad	Grado en el que los datos están disponibles para ser accedidos por usuarios y/o aplicaciones autorizadas		X
Portabilidad	Grado en el que los datos pueden ser alojados, reemplazados o movidos desde un sistema a otro		X
Recuperabilidad	Grado en el que los datos disponen de formas de mantener un nivel especificado de operabilidad incluso cuando se producen fallos		X

Como puede observarse, estas características se clasifican en dos grandes bloques: **inherentes** y **dependientes del sistema**. Las **inherentes** se refieren al grado con el que las características de calidad de los datos tienen un potencial intrínseco para satisfacer las necesidades establecidas y necesarias cuando los datos son utilizados bajo condiciones específicas; las **dependientes del sistema**, por otro lado, permiten determinar el grado con el que la calidad del dato es alcanzada y preservada a través de un sistema informático cuando los datos son utilizados bajo condiciones específicas.

Para ilustrar el significado de algunas de estas características (p.ej. **exactitud**, **completitud**, o **consistencia**), a continuación se introducen algunos ejemplos:

- Como ejemplo de nivel inadecuado de **exactitud sintáctica** podría ponerse el hecho de que el atributo **Nombre** de la entidad **Persona** toma un valor o dato “*Marja*” (no existente en los datos de referencia de nombre) en lugar de “*María*” (que sí que está incluido).
- El hecho de que el atributo **Nombre** de la entidad **Persona** tome el valor de “*George*” en vez de “*Jorge*” para almacenar datos de la Persona llamada realmente “*Jorge*” es un ejemplo de nivel inadecuado de **exactitud semántica**. Ambos valores son sintácticamente correctos, pero **George** es otra persona distinta a **Jorge**, y quien capturó y guardó los datos, simplemente se equivocó de persona.

Tabla 7.1: Características de calidad del dato

Característica	Definición	Inherente	Dependiente del sistema
Exactitud	Grado en que los datos representan correctamente el valor de un atributo de una entidad del mundo real	X	
Completilud	Grado en el que existen suficientes valores para todos los atributos necesarios para una aplicación que involucre una representación de una entidad del mundo real	X	
Consistencia	Grado en que los datos están libres de contradicción y son coherentes con el resto de los datos de su contexto de uso	X	
Credibilidad	Grado en que los datos se consideran ciertos y creíbles por los usuarios	X	
Actualidad	Grado en que los datos tienen fechas y tiempos adecuados para el uso previsto	X	
Accesibilidad	Grado en que los datos pueden ser accedidos por cualquier usuario autorizado	X	X
Cumplimiento	Grado en el que los datos están construidos conforme estándares, convenciones o regulaciones	X	X
Confidencialidad	Grado en el que los datos tienen atributos específicos que solo pueden ser accedidos por usuarios autorizados	X	X
Eficiencia	Grado en el que los datos tienen atributos que pueden ser accedidos y procesados dentro de los niveles de rendimiento esperados	X	X
Precisión	Grado en el que los datos son precisos.	X	X
Trazabilidad	Grado en el que los datos tienen atributos que proveen información detallada sobre los cambios realizados en los datos	X	X
Comprendibilidad	Grado en el que los datos se expresan de manera que los usuarios puedan leerlos e interpretarlos correctamente	X	X
Disponibilidad	Grado en el que los datos están disponibles para que los usuarios y/o aplicaciones autorizadas puedan acceder a ellos		X
Portabilidad	Grado en el que los datos pueden ser alojados, reemplazados o movidos desde un sistema a otro		X
Recuperabilidad	Grado en el que los datos disponen de formas de mantener un nivel especificado de operabilidad incluso cuando se producen fallos		X

7.4. Gestión de calidad del dato

109

- Supóngase que, para una determinada aplicación, se necesita recoger valores (o datos) para los siguientes atributos de una entidad Persona: DNI, Nombre, Apellido1, y Apellido2 para ser usados adecuadamente en un contexto de uso. En caso de faltar alguno de ellos (nivel inadecuado de completitud), podría ocurrir que los datos de la persona no se pudieran utilizar; incluso, podrían faltar algunos atributos más, como por ejemplo email, pero si no es relevante para la aplicación, no habría ese problema de completitud.
- Un ejemplo de falta de consistencia puede darse, por ejemplo, cuando el valor (o dato) del atributo FechaNacimiento de la entidad Persona no es posterior a la fecha de hoy.

Diferentes autores han proporcionado diferentes mecanismos para medir y evaluar la calidad de los datos usando las dimensiones o características de calidad seleccionadas. Aunque para dar soporte a este proceso, se han propuesto numerosas metodologías de evaluación (Batini et al., 2016), el principal problema de estas contribuciones es que, normalmente, se han realizado *ad hoc* y no permiten ni generalizar los resultados obtenidos ni compararlos con los obtenidos por otras organizaciones (Loshin, 2011).

Para paliar estos problemas, se han desarrollado estándares que recogen los conocimientos y principios básicos comunes para medir y evaluar la calidad de los datos. Ejemplos de estos estándares pueden ser la mencionada ISO/IEC 25012 (ISO/IEC, 2008a); la ISO 8000-8 (ISO/IEC, 2015) que recogen características de calidad; la ISO/IEC 25024 (ISO/IEC, 2008b) que recoge aspectos específicos de cómo llevar a cabo las mediciones de las características, o la ISO/IEC 25040 (ISO/IEC, 2011), que proporciona una metodología de evaluación de calidad del software que puede ser adaptada a la evaluación rigurosa y sistemática de la calidad de los datos. En este punto es necesario introducir la principal diferencia entre **medir** y **evaluar** la calidad: medir consiste en determinar la cantidad de calidad del dato que tiene un conjunto de datos; mientras que evaluar implica determinar si, de acuerdo al nivel de riesgo que asume la organización, la cantidad de calidad del dato medida es suficiente y adecuada para usar los datos en el contexto de uso establecido para esos datos. La evaluación de calidad del dato requiere primero medir la calidad. Y para medir la calidad, primero deben definirse procedimientos de medición.

En ese sentido, ISO 25024 (ISO/IEC, 2008b) proporciona una serie de propiedades medibles para cada una de las características presentadas en la Tab. 7.1; además, para cada una de estas propiedades medibles, el estándar, proporciona un método de medición genérico, que permitirá, convenientemente particularizado, medir dichas propiedades y luego agruparlas para determinar el valor de la característica de calidad del dato. En la Fig. 7.3 se muestran las propiedades medibles para las características de calidad identificadas como inherentes (véase Tab. 7.1).

Una de las ventajas de usar estas propiedades medibles es que, en caso de niveles inadecuados de calidad de datos, es posible identificar mejor qué está causando que esto ocurra y por tanto, es más fácil actuar directamente sobre dichas causas.

A modo de ejemplo, supóngase quese quiere medir para medir el grado de **exactitud** de un conjunto de datos. Para ello se considera necesario medir las propiedades “*exactitud sintáctica*”, “*exactitud semántica*” y “*rango de exactitud*”; con lo resultados, habrá que hacer algún tipo de agrupación que tenga en cuenta la importancia o peso relativo de cada una de estas propiedades a la hora de evaluar la **exactitud**. Supóngase que una organización, determina que la mejor forma de hacerlo es mediante una media aritmética ponderada de los resultados de la medición

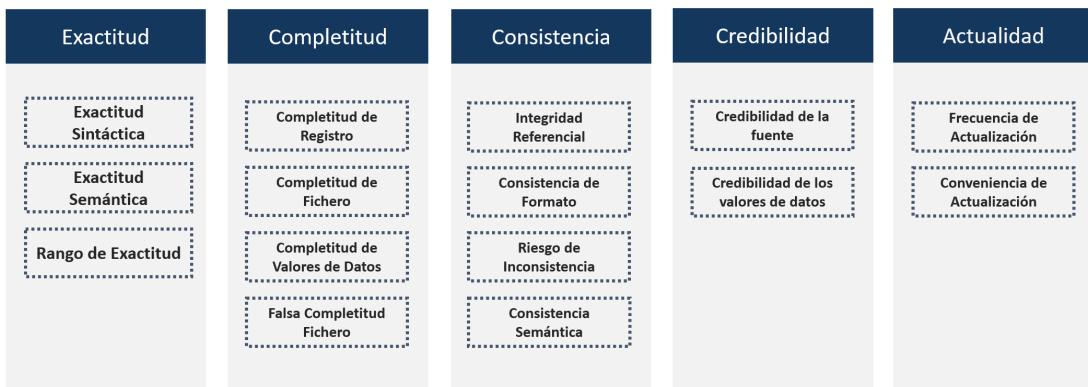


Figura 7.3: Algunas propiedades de las características inherentes de calidad del dato

de las tres propiedades medibles. En base a su nivel de riesgo para un determinado proceso de negocio, supóngase que la organización considera que para una determinada aplicación, puede asignar, para la media aritmética ponderada, los siguientes pesos: 0,4, 0,4 y 0,2 para la exactitud semántica, para la exactitud sintáctica y para el rango de exactitud respectivamente.

A la hora de medir las propiedades medibles correspondientes a las características de calidad del dato, es interesante tener en cuenta que la ISO/IEC 25024 proporciona procedimientos de medición cuya implementación depende fuertemente de la naturaleza de la propiedad y del objeto cuya calidad quiere medirse. La medición de algunas de estas propiedades implica contar el porcentaje de registros que violan las reglas de negocio que regulan la adecuación al uso de los datos en un contexto determinado (Loshin, 2002). Sin embargo, a la hora de la medición, uno de los ejercicios más difíciles es recolectar y validar las reglas de negocio específicas que rigen la validez de los datos (Caballero et al., 2022a). Para el ejemplo propuesto, imagínese que si se pretende medir el nivel de exactitud sintáctica del dato recogido en el atributo DNI se pudiera usar la siguiente regla de negocio “*el DNI tiene que seguir la especificación para DNI, o para el NIE, correspondiente con la expresión regular (d{8}) ([A-Z])*”. Habría que comprobar bien manualmente, bien mediante algún tipo de script, cuántos registros verifican la anterior regla de negocio para su atributo DNI. Como se verá en la siguiente subsección, en algunos entornos académicos y profesionales, se utilizan técnicas de perfilado de datos para realizar mediciones.

Supóngase que para el ejemplo, y tras haber realizado todas las mediciones de las propiedades, y haberlas agrupado realizando la media aritmética ponderada, es posible obtener un resultado de la medición para la *exactitud* de 70.

Una vez realizada la medición, el siguiente paso es la evaluación propiamente dicha. La evaluación consiste en comparar el resultado obtenido (70 en el ejemplo) con el umbral mínimo de aceptación que depende del nivel de riesgo que decida asumir la organización al utilizar estos datos Redman (2016)). Si, por ejemplo, dicho umbral se hubiese establecido en 75 para el uso concreto que se le va a dar a estos datos, se concluiría que no deberían ser utilizado. Esto no significa que los datos no puedan usarse en otro contexto en el que, por ejemplo, el valor umbral se estableciese en 65.

7.4. Gestión de calidad del dato

111

En algunos contextos de uso, como se verá posteriormente en el Cap. 8, antes de usar los datos se realiza un proceso de preparación de los mismos que tiene como objetivo determinar y adecuar los niveles de calidad al uso que se pretende dar mediante un proceso de evaluación y mejora que se centra en la **limpieza de los datos**. Normalmente, en este proceso suele recurrirse a métodos estadísticos, frente a la aproximación basada en la medición de las características de calidad del dato presentada anteriormente. Se pierde entonces de alguna manera la capacidad de establecer una dirección más efectiva y sobre todo alineada a las necesidades reales de la organización de las operaciones de evaluación y limpieza del dato.

Finalmente, es interesante mencionar que, basándose en los estándares ISO/IEC 25012 ([ISO/IEC, 2008a](#)) e ISO/IEC 25024 ([ISO/IEC, 2008b](#)) es posible certificar el nivel de calidad del dato de un repositorio de datos. [Gualo et al. \(2021\)](#) recoge experiencias de medición, evaluación y certificación de calidad del dato.

7.4.1. Medición de calidad de datos vs perfilado del dato

En esta subsección se plantea el **perfilado del dato** como una técnica base para realizar la medición de las propiedades medibles de las características de calidad del dato o para descubrir nuevas reglas de negocio. [Abedjan et al. \(2015\)](#) clasifica los tipos de **perfilado del dato** en las siguientes categorías:

- **Perfilado de columna simple**, que implicaría tareas de identificación de cardinalidades, identificación de patrones y tipos de datos, distribución de valores de datos, clasificación de dominios.
- **Perfilado de columnas múltiples**, que implicaría tareas de correlación y reglas de asociación, identificación de clusters y outliers, elaboración de resúmenes de datos y bocetos.
- **Perfilado de dependencias**, que a su vez implica:
 - **Detección de reglas de unicidad**, tales como la identificación de claves, identificación de condiciones e identificación de sinónimos.
 - **Detección de dependencias de inclusión**, que puede abarcar el descubrimiento de claves ajenas o la identificación de dependencias condicionales de inclusión.
 - **Dependencias funcionales**, como pueden ser las dependencias condicionales.

En **R Software** se puede utilizar el paquete `dlookr` [Ryu \(2022\)](#), que contiene algunas funciones interesantes que pueden ayudar a realizar determinadas tareas de perfilado. Por ejemplo, la función `overview()` da información general sobre un conjunto de datos; resulta muy interesante la función `diagnose()`, que proporciona información realizando un perfilado de los valores únicos y los valores únicos de un conjunto de valores.

En el siguiente fragmento se muestra el tipo de información proporcionada por `diagnose` (`Madrid_POIS$City_Center`): ‘variables’ muestra el nombre de los atributos del conjunto de datos (en este caso Llonde longitud y Lat de latitud); `types` muestra el tipo de dato de cada variable; `missing_count`, `missing_percent`, `unique_count` y `unique_rate` describen respectivamente el conteo de valores nulos, el porcentaje de dichos valores, el número de valores

únicos o no repetidos y su correspondiente porcentaje; `<chr>`, `<int>`, `<dbl>` se hacen referencia al tipo de dato de cada uno de los parámetros anteriores (carácter, *integer*, *double*)

```
library("dlookr")
library("idealista18")
diagnose(Madrid_POIS$City_Center)
```

Si estas funciones de perfilado proporcionan información suficiente y adecuada, es posible usar los resultados para computar las mediciones de las propiedades medibles. Por ejemplo, se puede utilizar el resultado de la columna `missing count` para calcular el grado de completitud de las variables `longitud` y `latitud`, que se pueden establecer en 100% al ser `missing count = 0` para las dos variables. Incluso se pueden utilizar funciones como `plot_na_pareto()` para visualizar un gráfico de Pareto mostrando las variables que no tienen valores nulos. Finalmente, es interesante mencionar que el paquete `dlookr` incluye funciones como `diagnose_paged_report()`, que permiten elaborar informes que contienen información sobre las estructuras de datos del conjunto de datos, avisos, descripción de las variables, valores perdidos, valores únicos de las variables categóricas y numéricas, distribuciones de valores nulos y negativos, posibles outliers, ... El siguiente fragmento de código explica cómo crear un informe de 15 páginas en formato PDF con toda esa información sobre la variable `idealista18::Madrid_POIS$Metro`:

```
#diagnose_paged_report(idealista18::Madrid_POIS$Metro)
```

En ocasiones, y retomando la idea de las reglas de negocio, puede decirse que la información proporcionada por el perfilado del dato, puede usarse para derivar reglas de negocio a partir del estado actual de los datos; y, para recoger información que se puede emplear durante el proceso de medición de determinadas características de calidad del dato. En el Cap. 8 se profundizará en el proceso de estudio de dos características de calidad del dato: completitud y consistencia.

7.4.2. Mejora del dato

Si los datos no tienen el nivel de calidad necesario, es preciso mejorar su calidad para que no arrojen los procesos de negocio. Para ello, a partir de los resultados de las mediciones, los analistas de calidad del dato deben determinar las causas raíces de esos niveles inadecuados de calidad del dato. Strong et al. (1997a) identifican diez posibles obstáculos que pueden hacer que los datos no tengan esos niveles adecuados de calidad:

1. Múltiples fuentes de datos producen diferentes valores para el mismo atributo de la misma entidad.
2. La realización de juicios subjetivos en la producción de los datos, puede llevar a valores diferentes.
3. Errores sistemáticos en la producción de información llevan a la pérdida de información.
4. Grandes volúmenes de información almacenada dificultan su acceso en tiempo razonable.
5. Sistemas heterogéneos distribuidos llevan a definiciones, formatos y valores inconsistentes.
6. La información no numérica es difícil de indexar.

7.4. Gestión de calidad del dato

113

7. El análisis automatizado de los contenidos en colecciones de información pueden no producir resultados adecuados.
8. A medida que las necesidades de los usuarios cambian, la información que es relevante y útil para la realización de una determinada tarea también cambia.
9. Un acceso fácil a la información puede entrar en conflicto con los requisitos de seguridad, confidencialidad y privacidad.
10. La falta de recursos de computación limita el acceso a los datos en circunstancias favorables.

En función de la naturaleza del problema detectado, las acciones correctivas pueden ser de distinta naturaleza:

- **Corrección de causas sistemáticas.** Si se observa que los problemas se suceden de forma sistemática y repetida, entonces las acciones de mejora del dato deben estar orientada a eliminar esas causas sistemáticas (véase [Strong et al. \(1997a\)](#)). Por ejemplo: si los errores de calidad del dato se deben a que un proceso de negocio está mal diseñado, entonces hay que rediseñarlo; si las causas se deben a que hay personas desempeñando ciertos roles para los que no tienen los conocimientos o habilidades adecuadas, entonces hay que darle la formación adecuada; o si se deben a que hay software (por ejemplo, procesos ETL) que falla, entonces hay que realizar el mantenimiento correctivo correspondiente.
- **Corrección de errores debidos a causas aleatorias.** Si no es posible identificar cuáles son las causas raíces, porque son completamente desconocidas o aleatorias, no queda más remedio que actuar sobre los valores de los datos, cambiándolos para asegurarse que se cumplen las reglas de negocio que están establecidas. A este proceso se le suele llamar **depuración o limpieza del datos (*data cleansing*)**. [Ilyas and Chu \(2019\)](#) identifican diversas técnicas de limpieza del dato (que pueden incluir operaciones de **imputación de datos** - véase la sección ?? del Cap. 8, de normalización): esto implica realizar limpieza basadas en reglas de negocio, deduplicación de datos, transformación de datos, o limpieza guiadas por machine learning. En este caso, sería posible utilizar algunas funciones del paquete `dlookr` relacionadas con la transformación de los datos tales como `imputate_na()` o `imputate_outlier()` que genera valores para evitar datos faltantes o valores que garantizan niveles adecuados de exactitud o de consistencia.

Resumen

En este capítulo se presentan los fundamentos del gobierno del dato. Es importante tener en cuenta los siguientes aspectos:

- El gobierno del dato tiene como objetivo asegurar que los datos que se usan y gestionan en las organizaciones están alineadas a las estrategias del dato de la organización, maximizando así su valor organizacional.
- Gobernar los datos implica el diseño, implementación y mantenimiento de un sistema de gobierno del dato. Un sistema de gobierno del dato tiene siete tipos de componentes: procesos de gestión del dato, gestión de calidad del dato y gobierno del dato; estructuras organizacionales; principios, políticas y marcos de referencia; datos y descripción de los datos; cultura, ética y comportamiento; personas, habilidades y competencias; servicios, infraestructuras y aplicaciones.
- Existen modelos de referencias que pueden ser usados como base para la creación de sistemas de gobierno del dato.
- El gobierno del dato persigue cuatro beneficios básicos para la organización: alineamiento estratégico, realización de beneficios, optimización de riesgos, optimización de recursos.
- La gestión de la calidad del dato es el proceso mediante el cual se garantiza que los datos tengan el nivel de calidad adecuado para las tareas para las que fueron recogidos.
- Para evaluar y medir la calidad se necesitan criterios; estos criterios se llaman características o dimensiones de calidad del dato.
- La evaluación y medición de calidad del dato requiere la identificación y clasificación de las reglas de negocio que rigen la validez de los datos. Las técnicas y herramientas de perfilado del dato se pueden utilizar como base para la identificación de reglas de negocio a partir de los datos.
- Cuando los datos no tienen calidad, a partir de las mediciones y evaluaciones realizadas, deben investigarse cuáles son las posibles causas. Si las causas son sistemáticas, entonces hay que enfocar el problema desde un punto de vista organizacional; si las causas son aleatorias, se pueden usar las técnicas de limpieza del dato vistas.

Capítulo 8

Integración y limpieza de datos

Jorge Velasco López^a y José-María Montero^b

^aInstituto Nacional de Estadística de España ^bUniversidad de Castilla-La Mancha

8.1. Introducción

En los proyectos de ciencia de datos, generalmente es necesario realizar un **preprocesamiento** (o **preparación**) de los datos antes de iniciar las fases de modelado. Las labores de preprocesamiento son específicas para cada conjunto de datos, para los objetivos del proyecto y para las técnicas de modelización que se van a utilizar. Sin embargo, hay una serie de tareas comunes, como las de **integración** (combinación de datos de distintas fuentes) a partir de los datos en bruto (o sin procesar) y **limpieza** (identificación y corrección de posibles errores en los datos). Otras tareas que se suelen incluir en el proceso de preparación de datos son: la transformación de la variable objetivo, para cambiar su distribución de probabilidad (normalmente para hacerla gaussiana), la transformación de variables predictoras (o clasificadoras, en su caso) (*feature engineering*), la normalización y la reducción de la dimensionalidad. Estas tareas se abordarán en el Cap. 9.

En **R**, existen varios paquetes para llevar a cabo estos trabajos: **tidyverse**, para la manipulación de ficheros y variables que se han ilustrado en el Cap. 3; **dlookr** (Staniak and Biecek, 2019); **validate**, **errorlocate** y **dcmmodify** (van der Loo and de Jonge, 2019), para realizar validaciones y transformaciones a los datos; **caret** (Kuhn, 2008), para imputar los datos faltantes o perdidos (*missing data*); **sf** (Pebesma et al., 2018), para el manejo de conjuntos de datos espaciales; y **GGally** (Schloerke et al., 2021) y **naniar** (Tierney and Cook, 2018) para labores de visualización.

8.2. Integración de datos

La **integración** es un conjunto de procesos técnicos y de negocio que se utilizan para combinar información proveniente de diferentes fuentes. En términos generales, se puede decir que consiste en acceder a los datos desde todas las fuentes y localizaciones, tanto en entorno local, como en la nube o en una combinación de ambos, de modo que los registros de una fuente de datos enlacen con los registros de otra.

Para ilustrar el proceso de integración, a continuación se integra, por separado,¹ el conjunto de datos **Madrid_Sale** (incluido en el paquete **idealista18**), que contiene el identificador de las viviendas en venta en el municipio de Madrid y 41 variables relativas a dichos inmuebles (como su antigüedad y precio, por ejemplo) con otros dos conjuntos de datos del mismo paquete: **Madrid_POIS**, donde se listan, entre otras, las coordenadas de las estaciones de metro de la ciudad de Madrid; y **Madrid_Polygons**, que contiene los polígonos (en este caso, distritos) del municipio. Ello redundará en un enriquecimiento de los análisis que se lleven a cabo, al disponer en un mismo conjunto de datos un número mayor de variables relativas al problema a solucionar. A modo de ejemplo, la integración de **Madrid_Sale** con **Madrid_POIS** permitirá determinar el número de estaciones de metro a menos de 500 metros de la vivienda y la distancia de cada vivienda a la estación de metro más cercana; la integración de **Madrid_Sale** y **Madrid_Polygons** permitirá la construcción un mapa de precios medios del metro cuadrado de vivienda por distritos. Ambos ejemplos se ilustrarán con detalle en las dos subsecciones siguientes.

La función **glimpse()** permite mostrar la estructura de los tres conjuntos de datos incluidos en el paquete **idealista18**.

```
library("tidyverse")
library("idealista18")
library("sf")
library("GGally")
library("dlookr")

glimpse(Madrid_Sale)
glimpse(Madrid_POIS)
glimpse(Madrid_Polygons)
```

La combinación de conjuntos de datos se realiza, fundamentalmente, con las funciones de unión. En el Cap. 3 se mostraban las cuatro funciones de unión principales del paquete **tidyverse**: **left_join()**, **inner_join()**, **right_join()** y **full_join()**. Sin embargo, también merece la pena mencionar las uniones de filtrado entre dos objetos *x* e *y*, que se llevan a cabo mediante las siguientes funciones:

- **semi_join()**: devuelve todas las filas de *x* con una coincidencia en *y*.
- **anti_join()**: devuelve todas las filas de *x* que no tengan una coincidencia en *y*.
- **nest_join()**: devuelve todas las filas y columnas de *x* con una nueva columna anidada, que contiene todas las coincidencias de *y*.

¹Podría parecer que lo lógico es integrar todos los ficheros en un sólo conjunto de datos. Sin embargo, en muchas ocasiones es conveniente realizar integraciones parciales de los ficheros, para llevar a cabo distintas tareas en cada una de ellas, o, simplemente, por cuestiones de rendimiento.

8.2.1. Integración de los ficheros Madrid_Sale y Madrid_POIS

Como se avanzó anteriormente, dos interesantes resultados que se podrían obtener mediante la integración de estos conjuntos de datos son: (i) la determinación del número de estaciones de metro a menos de 500 metros de la localización de la vivienda de interés, y (ii) la distancia a la estación de metro más cercana. Para la integración entre los dos ficheros, se utiliza la función `st_join()`, función de unión para datos espaciales, del paquete `sf`.²

```
vivs_madrid <- Madrid_Sale |>
  st_join(Madrid_Polygons, left = TRUE)
```

Para proceder a la integración de ambos ficheros, primeramente se crean las variables que indican cuál es el sistema de referencia de coordenadas (SRC) que se va a utilizar y que permite determinar la posición de un punto en relación a otro en base a líneas imaginarias (en el ejemplo que nos ocupa, permite representar la ubicación de las viviendas en la superficie de la Tierra). En este caso, la asignación de coordenadas se realiza a través de las variables `projcrc_src` y `projcrs_dest`, en las que se establecen los parámetros de:

- Nombre de la proyección (`proj`).
- Zona UTM (`zone`) donde se ubica el conjunto de viviendas.
- Nombre del elipsode (`ellips`). La Tierra no es una esfera y tiene accidentes geográficos, por lo cual hay que trabajar con elipsoides y explicitar los parámetros que definen su forma.
- Nombre del datum (`datum`). Define el origen y la orientación de los ejes de coordenadas, es decir, proporciona la información necesaria para dibujar el sistema de coordenadas en el elipsoide. El World Geodetic System (WGS84) es un standard en la industria a nivel mundial; no obstante, existen algunas variantes locales (la más famosa es el North American Datum (NAD83)).
- Tipo de unidades (`units`); en este caso, metros.

Seguidamente, se indica la distancia (en este caso en metros) que se va a usar como radio en la variable `radius_meters`. Finalmente, se lleva a cabo un procesamiento específico para datos espaciales: se crea un objeto espacial, se proyecta a plano para pasar de tres a dos dimensiones (hasta ahora se ha trabajado en la representación de la Tierra en tres dimensiones; sin embargo, estamos acostumbrados a ver mapas, es decir, a ver dos dimensiones), se cambia la geometría y, finalmente, se vuelve al sistema de coordenadas no proyectadas.

```
projcrs_src <- "+proj=longlat +datum=WGS84 +no_defs"
projcrs_dest <- "+proj=utm +zone=30 +ellps=WGS84 +datum=WGS84 +units=m +no_defs"
radius_meters <- 500 # Se marca la distancia que interesa.
pois_metro <- Madrid_POIS$Metro |>
  st_as_sf(coords = c("Lon", "Lat"), crs = projcrs_src) |> # Crear objeto espacial sf
  st_transform(crs = st_crs(projcrs_dest)) |> # Proyectar al plano (st_crs recupera la
  ↵ referencia de la coordenada y st_transform realiza la transformación)
```

²Téngase en cuenta que la vivienda es un bien anclado a una localización geográfica.

```
st_buffer(dist = radius_meters) |> # Cambiar la geometría de punto a polígono
  ↵ (círculo)
st_transform(crs = st_crs(projcrs_src)) # Volver al sistema de coordenadas no
  ↵ proyectadas (Ángulos)
```

A continuación, para cada una de las viviendas, se calcula el número de estaciones de metro a menos de 500 metros (variable `N_METRO_STOPS_500_M`). Para ello, primero se realiza el cálculo del objeto `sf metro_count`, seleccionando la variable `pois_metro` y cruzando con la geometría.

```
metro_count <- vibs_madrid |>
  select(ASSETID) |>
  st_join(pois_metro,
    join = st_intersects,
    left = FALSE
  )
# Se elimina la geometría para evitar ralentizar el cálculo
st_geometry(metro_count) <- NULL
```

Los valores de `N_METRO_STOPS_500_M` se obtienen con el siguiente código:

```
metro_count <- metro_count |>
  group_by(ASSETID) |>
  summarise(N_METRO_STOPS_500_M = n()) |>
  ungroup()
```

Al cruzar `metro_count` con `vibs_madrid`, se observa que hay casi 25.000 registros que no cruzan (25.000 viviendas que no tienen ninguna estación de metro a menos de 500 metros). En consecuencia, se retiran del análisis puesto el objetivo de la integración de estos dos conjuntos de datos es (*i*) la determinación del número de estaciones de metro a menos de 500 metros de la localización de las viviendas incluidas en el fichero `Madrid_Sale`.

```
vibs_madrid <- vibes_madrid |>
  inner_join(metro_count, by = "ASSETID")
```

Posteriormente, se determina la estación más cercana a cada vivienda a la venta con la función `st_nearest_feature()`.

```
pois_metro <- Madrid_POIS$Metro |>
  st_as_sf(coords = c("Lon", "Lat"), crs = projcrs_src) # Crear objeto espacial

mascercano_metro_stops <- pois_metro[st_nearest_feature(vibs_madrid, pois_metro), ] #
  ↵ Cálculo de las paradas cercanas
```

Por último, con la función `st_distance()`, se calcula la distancia de cada vivienda a la estación de metro más cercana, creándose la variable `METRO_STOP_MASCERCANO_DISTANCIA`.

```
vivs_madrid <- vivs_madrid |>
  mutate(METRO_STOP_MASCERCANO_DISTANCIA =
    ↪ as.numeric(st_distance(mascercano_metro_stops, geometry, by_element = T)))
```

8.2.2. Integración de los ficheros Madrid_Sale y Madrid_Polygons

En esta subsección se muestran los detalles para construir un mapa de precio medio del metro cuadrado de la vivienda en la ciudad de Madrid, por distritos, tras la integración de los conjuntos de datos `Madrid_Sale` y `Madrid_Polygons`.

Para proceder a la integración de ambos ficheros, primeramente se realiza la conversión del conjunto de datos `Madrid_Polygons` a objeto espacial y se le asocia la coordenada de referencia (de forma similar a como se hizo en la integración de los ficheros `Madrid_Sale` y `Madrid_POIS`):

```
# Se convierten a objetos sf
Madrid_Polygons_sf <- sf::st_as_sf(Madrid_Polygons, wkt = "WKT") # WKT (Well-known
  ↪ text) es un formato de vectores geométricos
Madrid_Sale_sf <- st_as_sf(Madrid_Sale, coords = c("LONGITUDE", "LATITUDE"))
# se asocia la coordenada de referencia del objeto
st_crs(Madrid_Sale_sf) <- "+proj=longlat +datum=WGS84 +no_defs"
st_crs(Madrid_Polygons_sf) <- "+proj=longlat +datum=WGS84 +no_defs"
```

A continuación, se lleva a cabo la unión entre el objeto espacial de `Madrid_Polygons_sf` y `Madrid_Sale_sf` para calcular su precio por metro cuadrado (`preciopm2`) y el área de la geometría (`tract_area`).

```
Madrid_Sale_Polygons <- Madrid_Polygons_sf |>
  dplyr::mutate(tract_area = st_area(WKT)) |>
  sf::st_join(Madrid_Sale_sf) |>
  dplyr::group_by(LOCATIONNAME) |>
  dplyr::summarize(tract_area = unique(tract_area), preciopm2 = mean(PRICE /
  ↪ CONSTRUCTEDAREA))
```

A partir del resultado de esta integración, se construye la Fig. 8.1, que muestra un el mapa del precio medio del metro cuadrado de las viviendas a la venta en Madrid, a escala de distrito, lo que da una visión clara de las zonas más o menos económicas.

```
plot(Madrid_Sale_Polygons["preciopm2"])
```

8.3. Limpieza de datos

Es más habitual de lo deseable que algunas variables presenten problemas en la calidad de sus datos. En el Cap. 7, se mencionaban una serie de causas y la posibilidad de realizar el

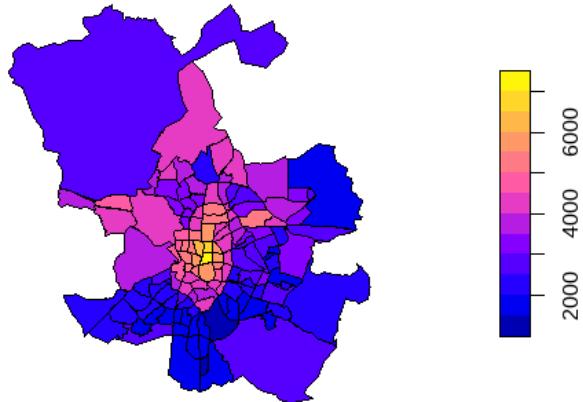


Figura 8.1: Precio por metro cuadrado de viviendas a la venta en Madrid por distrito

perfilado para tener una medición de la calidad de los datos. Si los datos no tienen el nivel de calidad adecuado, deben realizarse tareas de limpieza para transformarlos en datos consistentes, corrigiendo datos incorrectos, corruptos, con formato incorrecto, duplicados o incompletos.

En la Fig. 8.2 se muestra un proceso general de limpieza de datos. Cada rectángulo azul representa datos en un estado determinado, mientras que cada flecha representa las actividades necesarias para pasar de un estado a otro. En el primer estado están los datos tal y como se recogen (**datos en bruto** o **sin procesar**). Pueden carecer de encabezados, contener tipos de datos incorrectos, etiquetas de categoría incorrectas, codificación de caracteres desconocida o inesperada, etc. Una vez realizadas las correcciones necesarias, los datos pueden considerarse **datos técnicamente correctos**. Es decir, en este estado, los datos se pueden leer en un **data.frame** de R, con los nombres, tipos y etiquetas correctos. Sin embargo, esto no significa que los valores estén libres de errores o completos. Los **datos consistentes** son aquellos que están preparados para las fases de modelado.

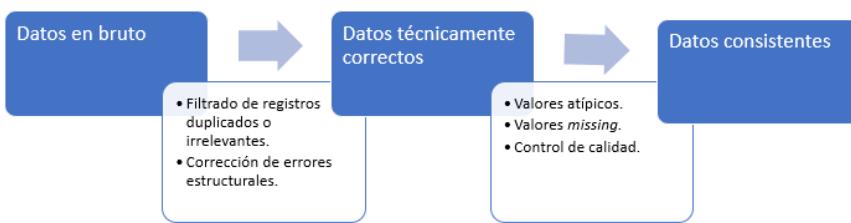


Figura 8.2: Flujo del proceso de limpieza de datos

Si bien las técnicas utilizadas para la limpieza de datos pueden variar según el tipo de datos que se esté procesando, en general, se pueden dividir en cinco grupos:

- Corrección de errores estructurales.

- Filtrado de registros duplicados o irrelevantes.
- Gestión de valores atípicos.
- Gestión de valores faltantes (*missing*).
- Validación y control de la calidad de los datos.

8.3.1. Corrección de errores estructurales

Los **errores estructurales** ocurren cuando se observan formatos erróneos, estructuras incorrectas o errores tipográficos que pueden dar lugar a categorías o clases mal etiquetadas. Por tanto, puede haber errores estructurales a nivel de conjunto de datos (`data.frame`, `sf`, `tibble`,...) y a nivel de variable.

8.3.1.1. A nivel de conjunto de datos

Los **cambios estructurales a nivel de conjunto de datos** consisten en modificar el tipo de objeto o eliminar o agregar variables. Por ejemplo, a continuación se genera el conjunto de datos `Madrid_Sale_int`, con estructura de `data.frame`, a partir del conjunto de datos `vivs_madrid` (de tipo `sf`), fruto del proceso de integración anterior. Además, para este objeto, se elimina la variable `geometry`, que no se va a usar en adelante y ralentiza la computación.

```
Madrid_Sale_int <- as.data.frame(vivs_madrid) |> select(-geometry)
```

También se crea un segundo conjunto de datos reducido, `Madrid_Sale_red`, con una selección de variables que se consideran de interés para ilustrar las tareas de limpieza que se exponen en este capítulo.

```
Madrid_Sale_red <- select(Madrid_Sale_int, ASSETID, PRICE, UNITPRICE, CONSTRUCTEDAREA,
                           ROOMNUMBER, CONSTRUCTIONYEAR, HASNORTHORIENTATION, HASSOUTHORIENTATION,
                           HASEASTORIENTATION, HASWESTORIENTATION, CONSTRUCTIONYEAR, DISTANCE_TO_METRO,
                           METRO_STOP_MASCERCANO_DISTANCIA)
```

Por último, se añade la variable `LOCATIONID1`, que indica el código de localización, al conjunto de datos `Madrid_Polygons`.

```
Madrid_Polygons$LOCATIONID1 <- substr(Madrid_Polygons$LOCATIONID, 1, 10)
```

En la siguiente sección, a partir de estos conjuntos de datos, se lleva a cabo un proceso de diagnosis y exploración.

8.3.1.2. A nivel de variable

Los **errores estructurales a nivel de variable** se centran fundamentalmente en el tipo de dato de las variables.

En primer lugar, se visualizan los datos con la función `diagnose()` de `dlookr`.

Nota

El paquete `dlookr` se usa para tareas de diagnosis y exploración, y es de utilidad para la localización de valores duplicados, faltantes, atípicos, tipología de datos, etc. La función `overview()` permite obtener una visión genérica del conjunto de datos, y la función `diagnose()` proporciona información a nivel de variable, como el tipo de dato (`type`), y sobre valores faltantes y únicos. Otras funciones útiles son `diagnose_numeric()` y `diagnose_category()`, que proporcionan información específica para valores numéricos y categóricos, respectivamente.

```
diagnose(Madrid_Sale_red)
```

Al ejecutar la función, se comprueba que todas las variables, excepto el identificador `ASSETID`, son de tipo numérico (o `integer`), lo que es correcto. En caso de tener que modificar el tipo de dato, por considerarse un error estructural, o porque sea conveniente para las fases de modelado, debería hacerse usando las funciones `as.factor()`, `as.numeric()` y `as.character()`, según el caso.

Corregir **errores estructurales tipográficos de variables categóricas** es especialmente relevante en algunas áreas de la ciencia de datos, como la minería de textos o *text mining* (que se verá con más profundidad en el Cap.38), donde la limpieza de textos consiste en eliminar todo aquello que no aporte información sobre su temática, estructura o contenido. A continuación, se muestra una función creada a partir del paquete `stringr` que permite realizar una limpieza básica de un texto, y que se ejecuta sobre la variable `Madrid_Polygons$LOCATIONNAME`, generando la variable `LOCATIONNAME1`.

```
library("stringr")
limpieza_textos <- function(texto) {
  # El orden de la limpieza no es arbitrario
  # Se convierte todo el texto a minúsculas
  nuevo_texto <- tolower(texto)
  # Eliminación de páginas web (palabras que empiezan por "http." seguidas
  # de cualquier cosa que no sea un espacio)
  nuevo_texto <- str_replace_all(nuevo_texto, "http\\S*", "")
  # Eliminación de signos de puntuación
  nuevo_texto <- str_replace_all(nuevo_texto, "[[:punct:]]", " ")
  # Eliminación de números
  nuevo_texto <- str_replace_all(nuevo_texto, "[[:digit:]]", " ")
  # Eliminación de espacios en blanco múltiples
  nuevo_texto <- str_replace_all(nuevo_texto, "[\\s]+", " ")
```

```

    return(nuevo_texto)
}
Madrid_Polygons$LOCATIONNAME1 <- limpieza_textos(texto = Madrid_Polygons$LOCATIONNAME)
glimpse(Madrid_Polygons)
#> $ LOCATIONNAME <fct> Conde Orgaz-Piovera, Pinar del Rey, Timón, Palacio,
#> ...
#> $ LOCATIONNAME1 <chr> "conde orgaz piovera", "pinar del rey", "timón",

```

8.3.2. Eliminación de observaciones duplicadas o irrelevantes.

Las **observaciones duplicadas** aparecen frecuentemente durante la recogida de datos e integración de las bases de datos, por lo que dichas duplicidades deben ser eliminadas en esta fase de limpieza.

A continuación, se usa la función `overview()` del paquete `dlookr` sobre el conjunto de datos `Madrid_Sale_int`, obtenido en la Sec. 8.3.1.1.

```

head(overview(Madrid_Sale_int), n = 9)
#>      division           metrics   value
#> 1      size      observations 70059
#> 2      size      variables     46
#> 3      size      values    3222714
#> 4      size      memory size 21931336
#> 5 duplicated duplicate observation      0
#> 6 missing   complete observation  26394
#> 7 missing   missing observation  43665
#> 8 missing   missing variables       7
#> 9 missing   missing values    48653

```

Entre otra información, como la existencia de valores faltantes (*missing*) en siete variables, se puede observar que no hay valores duplicados después del proceso de integración. En caso contrario, se podrían usar las funciones `base` de R para (*i*) localizarlos, con `duplicated()`, y (*ii*) extraer los registros únicos, con `unique()`. También se puede usar `distinct()`, del paquete `dplyr`, para eliminar los registros duplicados de un *data.frame*.

Las **observaciones irrelevantes** son aquellas que no encajan en el problema específico que se está analizando. Por ejemplo, si el objeto de estudio son datos de Madrid, se pueden eliminar las observaciones que no correspondan a dicho municipio. A continuación, se puede advertir que todas las observaciones de `Madrid_Polygons$LOCATIONID1` empiezan por el código correspondiente a Madrid (0-EU-ES-28) y, por tanto, no es necesario filtrar registros.

```

head(table(Madrid_Polygons$LOCATIONID1))
#>
#> 0-EU-ES-28
#>      135

```

En caso necesario, se pueden filtrar todos los registros de Madrid en el objeto `Madrid_Polygons1` haciendo:

```
Madrid_Polygons1 <-
  Madrid_Polygons |> filter(substr(Madrid_Polygons$LOCATIONID, 1, 10) != "0-EU-ES-28")
```

8.3.3. Gestión de valores atípicos no deseados

A menudo, hay observaciones distintas que, aparentemente, no encajan en los datos que se están analizando. Si existe una razón coherente para eliminar un valor atípico (un *outlier*), como una entrada de datos incorrecta, hacerlo mejorará el rendimiento que proporcionan los datos con los que se está trabajando. Sin embargo, el hecho de que exista un valor atípico no significa que sea incorrecto. Si un valor atípico resulta ser irrelevante para el análisis, o es un error, debe considerarse su eliminación. El número de posibles valores atípicos en el conjunto de datos `Madrid_Sale_red` se determina con el siguiente código, que avisa de la posibilidad de que existan para cada una de las variables.

```
diagnose_numeric(Madrid_Sale_red)
```

Otra manera de localizar datos atípicos es a través de la **visualización**. Por ejemplo, en la Fig. 8.3 se relaciona el precio de la vivienda por metro cuadrado con su localización, y se observa que la zona más cara es Recoletos y la más barata es San Cristobal. La simple observación aconsejaría un análisis de los casos extremos (muy baratos o caros en cada uno de los distritos).

```
ggplot(Madrid_Sale_int, aes(x = reorder(LOCATIONNAME, PRICE / CONSTRUCTEDAREA, na.rm =
  TRUE), y = PRICE / CONSTRUCTEDAREA)) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1)) +
  labs(x = "Distrito", y = "Precio metro cuadrado")
```

Los box-plots y gráficos de dispersión de variables, para las categorías dadas de otra, así como las correlaciones entre dichas variables, también pueden utilizarse para detectar valores atípicos. Por ejemplo, se puede considerar la relación del precio del metro cuadrado de la vivienda con otras variables, como la superficie construida, la distancia al metro y el número de habitaciones. Para ello, primeramente se crea el conjunto de datos `Madrid_Sale_red2` con la variable derivada `price_bin` (de tipo factor), cuyas categorías o clases (o *bins*) son los cuartiles de la variable `PRICE`.

```
Madrid_Sale_red2 <- mutate(Madrid_Sale_int, price_bin = cut2(PRICE, g=4)) |>
  select(price_bin, CONSTRUCTEDAREA, DISTANCE_TO_METRO, ROOMNUMBER, LOCATIONNAME)
```

A partir del conjunto de datos `Madrid_Sale_red2` se puede crear construir la Fig. 8.4.

8.3. Limpieza de datos

125

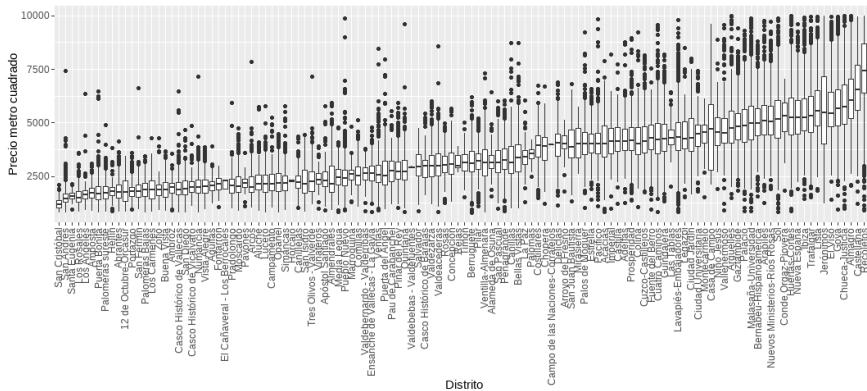


Figura 8.3: Precio medio del metro cuadrado por distritos

```
ggpairs(Madrid_Sale_red2,
  column = 1:4, aes(color = price_bin, alpha = 0.5),
  upper = list(continuous = wrap("cor", size = 2))
)
```

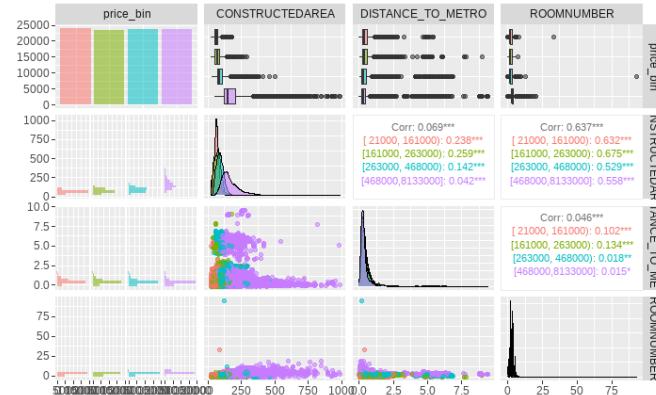


Figura 8.4: Distribuciones y correlaciones cruzadas algunas variables de Madrid-Sale-red

En dicha figura, la diagonal descendente muestra la función de cuantía (para precio medio del metro cuadrado) y las funciones de densidad de CONSTRUCTEDAREA, DISTANCE_TO_METRO y ROOMNUMBER. Los tres últimos paneles de la primera columna muestran los histogramas de las estas tres últimas variables. Los tres últimos paneles de la primera fila, proporcionan los box-plots de estas variables para los cuatro *bins* de la variable price_bin (primer cuartil en rosa, segundo en verde, tercero en azul y cuarto en morado). Los paneles del triángulo lateral derecho muestran sus correlaciones, mientras que los del triángulo inferior izquierdo presentan sus gráficos de dispersión. Dicho lo anterior, por ejemplo, en la primera fila se observa que

las viviendas más económicas suelen tener menos superficie construida (segunda columna), que suelen estar ligeramente más alejadas del metro (tercera) y suelen tener menos habitaciones. Sin embargo, se aprecian algunas cuestiones que llaman la atención. Por ejemplo, que hay una viviendas muy alejadas (a casi 400 kilómetros) de la estación de metro más cercana, lo cual distorsiona algunas de las figuras e impide ver la información que contienen; o que hay viviendas cuyo precio por metro cuadrado pertenece a la primera categoría de la variable `price_bin` (las más económicas) con muchas habitaciones o con mucha superficie construida. A continuación, por ejemplo, se filtran las viviendas con 30 o más habitaciones (aunque la lógica sería válida para muchas menos). Se observa que la superficie construida es de menos de 120 metros lo que, sin mayor conocimiento del conjunto de datos, no parece ser coherente y podrían excluirse (filtrarse) del conjunto de datos, o tratar de recabar la información correcta.

```
Madrid_Sale_red2 |> filter(price_bin == "[ 21000, 168000)", ROOMNUMBER > 30)
#>   price_bin CONSTRUCTEDAREA DISTANCE_TO_METRO ROOMNUMBER LOCATIONNAME
#> 1 [ 21000, 168000)           90       0.3826137      33 Almendrales
```

Finalmente, detectados los valores atípicos, por cualquiera de los procedimientos anteriormente expuestos, el paquete `dlookr`, a través de la función `impute_outlier()`, permite llevar a cabo sofisticadas imputaciones de los mismos, si bien sólo en el caso variables numéricas. Los métodos de imputación que se contemplan son: media, mediana, moda y *capping* (imputar los valores atípicos superiores con el percentil 95, y los inferiores con el percentil 5). Por ejemplo, se podría imputar la variable `CONSTRUCTEDAREA_imp` a partir de `CONSTRUCTEDAREA` con el método media (*mean*): `CONSTRUCTEDAREA_imp <- impute_outlier(Madrid_Sale_red2, CONSTRUCTEDAREA, method = "mean")`.

Otra opción es poner los valores atípicos como valores no disponibles (*not available*, `NA`) y proceder a imputar dichos `NA` tal y como se muestra en el epígrafe siguiente.

```
Madrid_Sale_red2$ROOMNUMBER[Madrid_Sale_red2$ROOMNUMBER >= 30 &
                           Madrid_Sale_red2$price_bin == "[ 21000, 168000)"] <- NA
```

8.3.4. Gestión de datos faltantes (*missing*)

Los datos pueden faltar por multitud de razones, aunque generalmente se suelen agrupar en dos categorías: **valores faltantes informativos** (Kuhn et al., 2013) y **valores faltantes aleatorios** (Little and Rubin, 2019). Los informativos implican una causa estructural, ya sea por deficiencias en la forma en que se recopilaron los datos o por anomalías en el entorno de observación. Los aleatorios son aquellos que tienen lugar independientemente del proceso de recopilación de datos.

Dependiendo de si los valores faltantes son de uno u otro tipo, se procederá de una u otra manera. A los informativos, en general, se les puede asignar un valor concreto (por ejemplo, “Ninguno”), ya que este valor puede afectar a los resultados de las predicciones. Los aleatorios pueden manejarse mediante la eliminación o la imputación. Además, los diferentes algoritmos de aprendizaje automático manejan la falta de información de manera diferente. De hecho, la

8.3. Limpieza de datos

127

mayoría de los algoritmos no incorporan mecanismos para manejarlos (por ejemplo, modelos lineales generalizados y derivados, redes neuronales y *support vector machine*) y, por lo tanto, requieren que se traten previamente. Sólo unos pocos modelos (principalmente basados en árboles) tienen procedimientos incorporados para tratar los valores faltantes.

Como se avanzó anteriormente, en **R**, los valores nulos se representan con el símbolo `NA`. Es importante distinguirlos de los valores indefinidos (p. ej., dividir entre cero), que se representan con el símbolo `NaN` (*Not a Number*). Para visualizar los patrones de datos *faltantes* de la variable `price_bin` del conjunto de datos `Madrid_Sale_red2`, se ejecuta el siguiente código.

```
library("naniar")
gg_miss_fct(x = `Madrid_Sale_red2`, fct = price_bin)
```

En la Fig. 8.5 se puede observar claramente que hay datos faltantes en la variable `LOCATIONNAME`, sobre todo en los dos primeros cuartiles (*bins*). Concretamente, hay 42 valores faltantes. No obstante, aunque el degradado del color morado apenas permite apreciarlo, también hay un valor faltante en `ROOMNUMBER`.

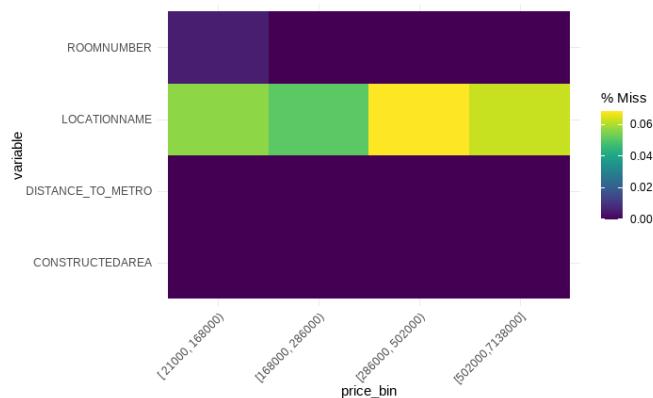


Figura 8.5: Visualización de valores faltantes

La gestión de los valores faltantes debe hacerse considerando la problemática que se quiera resolver. Una primera opción a considerar sería excluirlos, si bien se estaría eliminando información. Para filtrar los registros faltantes, se podría utilizar la función `is.na()`. En el caso de `ROOMNUMBER`:

```
Madrid_Sale_red3 <- Madrid_Sale_red2
Madrid_Sale_red3 |>
  filter(!is.na(ROOMNUMBER))
```

También se puede optar por reemplazarlos, por ejemplo por un 0, de la siguiente manera:

```
Madrid_Sale_red3[is.na(Madrid_Sale_red3)] <- 0
# También puede usarse la función `replace_na()`, que sustituye los valores perdidos en
# cada variable por el valor especificado.
```

No obstante, estas dos opciones no son acciones recomendables en primera instancia, porque eliminar los registros con valores faltantes, o introducir valores que podrían no respetar la semántica de los datos, puede ocasionar un alto impacto negativo en los niveles globales de calidad de datos del conjunto de datos.

Se puede ir más allá de la eliminación de valores faltantes. A través de diversos métodos se pueden imputar valores que, con mayor o menor probabilidad, podrían ser los que realmente correspondieran a estos valores faltantes. Estos métodos se conocen como **métodos de imputación de valores**. Para imputar valores faltantes se pueden usar diversas alternativas, como la función `preProcess()` del paquete `caret` o la función `impute_na()` del paquete `dlookr`. A continuación, se imputan los valores faltantes del conjunto de datos `Madrid_Sale_red2` con dos métodos. En primer lugar, con el algoritmo de KNN (k vecinos más cercanos), que sustituye el valor faltante por la media de los valores de los k vecinos más próximos. Después de realizar el preprocessamiento, se comprueba que las imputaciones han sido realizadas.

```
library("caret")
# Se realiza el preprocessamiento:
pre_knn <- preProcess(Madrid_Sale_red2, method = "knnImpute", k = 2)
# Se obtienen los datos
imputed_knn <- predict(pre_knn, Madrid_Sale_red2)
# Se comprueba que se ha imputado el valor faltante de la variable ROOMNUMBER
diagnose(imputed_knn)
```

A continuación, la imputación se realiza con la mediana (que suele ser preferible a imputar la media, puesto que el promedio puede verse afectado por outliers).

```
# Se realiza el preprocessamiento:
pre_median <- preProcess(Madrid_Sale_red2, method = "medianImpute")
# Se obtienen los datos
imputed_median <- predict(pre_median, Madrid_Sale_red2)
# Se comprueba que se ha imputado el valor faltante de la variable ROOMNUMBER
diagnose(imputed_median)
```

Nota

El paquete `dlookr`, a través de la función `impute_na()`, permite imputar valores faltantes. El predictor admite variables numéricas y categóricas. Los métodos que utiliza son: para numéricas `media`, `moda`, `KNN`, `rpart` y `mice`); y para categóricas: `mode`, `rpart` y `mica`. El paquete `recipes` también es recomendable. Por ejemplo, para la imputación de los valores faltantes con la media de la variable se usaría `step_meanimpute(all_numeric())`.

8.3.5. Validación y control de calidad

Al final del proceso de limpieza de datos, éstos deberían ser consistentes y seguir las reglas apropiadas para su campo de negocio. De no ser así, los modelos que estimen en base a ellos no representarán convenientemente la realidad objeto de estudio y las conclusiones que se obtengan de dichos modelos no serán de utilidad para dicha realidad.

La verificación de si los datos son o no consistentes y si siguen o no las reglas del campo de negocio del cual proceden, se puede llevar a cabo con el paquete `tidyverse`, que permite hacer selecciones, filtrados o tablas de frecuencias, entre otras acciones. A modo de ejemplo, en el caso del precio medio del metro cuadrado de los distritos de la ciudad de Madrid, se puede usar la función `count()` para obtener la distribución de frecuencias de la variable `METRO_STOP_MASCERCANO_DISTANCIA` y comprobar si es consistente con el conocimiento que se tiene de esa variable y del conjunto de datos. Se muestran las distancias a la estación más cercana para las viviendas correspondientes a los seis primeros registros.

```
head(count(as.data.frame(Madrid_Sale_red), METRO_STOP_MASCERCANO_DISTANCIA))
#>   METRO_STOP_MASCERCANO_DISTANCIA n
#> 1                 1.413845 1
#> 2                 1.414032 1
#> 3                 2.586694 1
#> 4                 3.156593 1
#> 5                 4.013776 1
#> 6                 4.128947 1
```

Una opción más sofisticada es el paquete `validate`, donde se pueden introducir las reglas de negocio dentro del propio código o bien desde un fichero externo. A continuación, se realiza un ejemplo con las reglas incrustadas en el propio código. Estas reglas pueden ser avisos o normas que indican error en esos datos. En este ejemplo, se han definido siete reglas: por ejemplo, `PRICE ≥ 0`, o que la suma de las variables `HASNORTHORIENTATION`, `HASSOUTHORIENTATION`, `HASEASTORIENTATION` y `HASWESTORIENTATION` sea la unidad. La salida que se obtiene se presenta a continuación. A modo de ejemplo, la regla `HASNORTHORIENTATION + HASSOUTHORIENTATION + HASEASTORIENTATION + HASWESTORIENTATION = 1` es la número 3, que, como se puede ver, no se cumple en 48.446 ocasiones.

```
library("validate")

Madrid_Sale_int |>
  check_that(
    HASLIFT >= 0,
    PRICE >= 0,
    HASNORTHORIENTATION + HASSOUTHORIENTATION + HASEASTORIENTATION + HASWESTORIENTATION
    == 1,
    is.numeric(PRICE),
    UNITPRICE * CONSTRUCTEDAREA == PRICE,
    if (ROOMNUMBER > 3) PRICE > 100000,
    nrow(.) >= 20000
```

```
) |>
summary()
#> name items passes fails NA error warning
#> 1 V1 70059 70059      0    0 FALSE FALSE
#> 2 V2 70059 70059      0    0 FALSE FALSE
#> 3 V3 70059 21613 48446  0 FALSE FALSE
#> 4 V4      1     1    0  0 FALSE FALSE
#> 5 V5 70059 15280 54779  0 FALSE FALSE
#> 6 V6 70059 70041     18  0 FALSE FALSE
#> 7 V7      1     1    0  0 FALSE FALSE
```

Nota

El proceso de validación puede ser más o menos complejo, según afecte a una única variable en un mismo registro, a más de una variable de un mismo registro o a más de una variable en más de un registro. En el último caso, además, se puede validar en un solo conjunto de datos o en más de uno.

En un esquema tradicional de validación, además de las reglas de validación aportadas por los expertos en el tópico del que se trate, debe incluirse también un listado de reglas de corrección (igualmente aportado por los expertos en la materia) que indiquen cómo hay que corregir un registro cuando no cumple con una determinada regla de validación. Este modo de proceder, además de suponer un doble esfuerzo, puede conducir a inconsistencias o validaciones cíclicas.

El Método de Fellegi y Holt³ (MFH) dà una solución a este problema, evitando dichas inconsistencias, proporcionando un procedimiento que genera un conjunto completo de reglas de validación, incorporando reglas implícitas a las formuladas por los expertos de manera explícita.

En breves palabras, dicho método asegura el cumplimiento de las siguientes tres premisas:

- Minimizar el número de campos a corregir en un registro para hacerlo pasar todas las validaciones.
- Mantener, en la medida de lo posible, la distribución conjunta original del conjunto de datos.
- Derivar las reglas de corrección, directamente y de forma implícita, de las reglas de validación. Por tanto, dichas reglas de corrección no son propuestas el experto o, en su caso, por el validador.

Los detalles sobre el MFH pueden verse en [Boskovitz et al. \(2003\)](#).

El MFH no está exento de limitaciones. La primera es el incremento del coste computacional, que puede llegar a constituir un problema en caso de que el número de reglas implícitas sea muy elevado, lo cual es muy frecuente. De hecho, hay casos en los que hay más reglas implícitas que registros. Para solucionar este problema, denominado “problema de localización del error”,

³El MFH es un estándar internacional en la revisión de la integridad de la información de encuestas y censos.

8.3. Limpieza de datos

131

que consiste, básicamente, en determinar el conjunto mínimo de variables a corregir para cada validación, se han propuesto varias alternativas, que incluyen métodos de investigación de operaciones, árboles binarios y metaheurísticas como algoritmos genéticos y similares.

A efectos prácticos, el MFH se puede aplicar con la función `locate_errors()` del paquete `errorlocate`, determinándose así cuáles son las variables a corregir para solventar los errores en las reglas de negocio establecidas (objeto `rules`). Por ejemplo, en el conjunto de datos `Madrid_Sale_red2` (donde se definía la variable `price_bin`), se establecen ahora unas reglas básicas algo más laxas (específicamente una: más de 10 habitaciones los tres primeros cuartiles), obteniéndose que habría que depurar la variable `ROOMNUMBER` en dos ocasiones para que el conjunto de datos quedase totalmente limpio (o depurado).

```
library("errorlocate")

rules <- validator(if (ROOMNUMBER >= 10) price_bin == "[502000,7138000]")
el <- locate_errors(Madrid_Sale_red2, rules) |>
  summary(el)
el
# el$variable
#   names          errors missing
#
#     price_bin      0      0
#   ROOMNUMBER      2      1
#   CONSTRUCTEDAREA 0      0
#   DISTANCE_TO_METRO 0      0
#   LOCATIONNAME    0     42
```

¿Y qué se debe hacer con los registros que no cumplen las normas de validación? La respuesta es, como norma, “siempre que se disponga de información de negocio, ésta debe preponderar sobre cualquier tipo de imputación”. A partir de este punto se puede proceder a realizar imputaciones determinísticas para solucionar los problemas detectados.

En el ejemplo anterior, se propone imputar el valor `ROOMNUMBER=5` a los casos de los tres primeros cuartiles (todos menos el más caro) que tengan más de 10 habitaciones. Para ello, se utiliza la función `modify_so()` del paquete `dcmmodify`. Para comprobar que la imputación se ha llevado a cabo con éxito, se pueden comparar los conjuntos de datos antes y después de la imputación con la función `compare()`, comprobándose que tal imputación se ha realizado exitosamente en los 2 registros que presentaban problemas con la regla `ROOMNUMBER >= 10`.

```
library("dcmmodify")

out <- Madrid_Sale_red2 |>
  modify_so(if (ROOMNUMBER >= 10 & price_bin != "[502000,7138000]") ROOMNUMBER <- 5)

rules <- validator(if (ROOMNUMBER >= 10) price_bin == "[502000,7138000]")
compare(rules, raw = Madrid_Sale_red2, modified = out)
#> Object of class validatorComparison:
#>
```

```
#> compare(x = rules, raw = Madrid_Sale_red2, modified = out)
#>
# Status           raw modified
# validations      70059   70059
# verifiable       70058   70058
# unverifiable     1        1
# still_unverifiable 1        1
# new_unverifiable 0        0
# satisfied        70056   70058
# still_satisfied  70056   70056
# new_satisfied    0        2
# violated         2        0
# still_violated   2        0
# new_violated     0        0
```

Resumen

- En un proyecto de ciencia de datos deben realizarse procesos de integración y limpieza previos a la fase de modelización, para asegurar niveles adecuados de calidad. Por ello, tras las labores iniciales de depuración, debe comprobarse si los datos son o no consistentes, y si siguen o no las reglas del campo de negocio del cual proceden. En este capítulo se abordan las cuestiones relativas a la integración de conjuntos de datos, su limpieza y depuración, y se proponen procedimientos para la validación de los mismos.
- El conjunto de datos utilizado en este capítulo está disponible en el paquete ‘idealista18'; en concreto, se utilizan los datos de Madrid.
- A partir de estos datos, se muestra un ejemplo de integración de datos espaciales y se diseña un marco de limpieza genérico basado en una serie de pasos básicos.
- Para la realización de estas funciones de integración y limpieza de datos, se proponen distintas funciones de los paquetes **tidyverse** (para la manipulación de ficheros y variables) y **caret** (para la imputación de valores perdidos).
- Para el tratamiento de datos espaciales se utiliza el paquete **sf**. La visualización de los mismos se lleva a cabo con **GGally** y **naniar**.
- Finalmente, se utiliza la función **locate_errors()**, del paquete **errorlocate**, para determinar cuáles son las variables con errores, de acuerdo a las reglas establecidas y **dcmmodify** para realizar imputaciones determinísticas.

Capítulo 9

Selección y transformación de variables

Jorge Velasco López^a y José-María Montero^b

^aInstituto Nacional de Estadística de España ^bUniversidad de Castilla-La Mancha

9.1. Introducción

Como se indicó en el Cap. 8, la preparación de datos, en un contexto de ciencia de datos, consiste en transformarlos de tal forma que se puedan utilizar adecuadamente en las fases posteriores de modelado. Esta preparación o pre-preprocesamiento puede ser un proceso laborioso e incluye tareas como la integración y limpieza de datos, que se detallaron en dicho capítulo.

El presente capítulo aborda las tareas relativas a la **selección de variables** (*feature selection*) y **transformación de variables**. La selección de variables tiene como objetivo elegir el elenco de variables más relevantes para el análisis. La transformación de variables hace referencia, básicamente, al uso de determinados procedimientos para modificar la distribución de la variable objetivo, a la ingeniería de variables (*feature engineering*), a la normalización y a la reducción de la dimensionalidad del problema de interés.

Se usará el conjunto de datos `Madrid_Sale` (disponibles en el paquete de R `Idealista18`), con datos inmobiliarios del año 2018 para el municipio de Madrid, y los paquetes `caret` (Kuhn, 2008), para diversas tareas de preparación de datos y `corrplot` (Wei et al., 2017), para visualizar correlaciones, entre otros.

9.2. Selección de variables

Quizás, el primer gran reto al que se enfrenta el científico de datos cuando maneja grandes conjuntos de datos es la identificación de las variables que proporcionen información valiosa sobre la variable objetivo, bien se trate de un problema de regresión o de clasificación. En caso de que el científico de datos salga exitoso de este primer gran reto, un determinado subconjunto de variables del conjunto de datos de interés proporcionará la misma información sobre la variable objetivo que la totalidad de variables incluidas en el conjunto de datos.

En consecuencia, la selección de variables involucra un conjunto de técnicas cuyo objetivo es seleccionar el subconjunto de variables predictoras más relevante para las fases de modelización. Esto es importante porque:

- Variables predictoras redundantes pueden distraer o engañar a los algoritmos de aprendizaje, lo que posiblemente se traduzca en un menor rendimiento, no solo predictivo (exactitud y precisión), sino también en términos de tiempo de computación.
- Igualmente, la inclusión de variables irrelevantes aumenta el coste computacional y dificulta la interpretabilidad.

Una adecuada selección de variables tiene ventajas importantes: (i) elimina las variables con información redundante; (ii) reduce el grado de complejidad de los modelos; (iii) evita o reduce el sobreajuste; (iv) incrementa de la precisión de las predicciones; y (iv) reduce la carga computacional.

No obstante, es importante señalar que, antes de llevarse a cabo la selección de variables propiamente dicha, debe comprobarse la magnitud de la varianza de las variables candidatas a ser seleccionadas y de sus correlaciones dos a dos, así como si existen combinaciones lineales entre ellas (multicolinealidad). Y ello, porque estas tres comprobaciones sirven para realizar una primera pre-selección de variables, si bien por razones técnicas y no de capacidad de explicación del comportamiento de la variable respuesta.

Los métodos de selección de variables (tras la pre-selección anteriormente mencionada) se suelen clasificar en: (i) los que utilizan la variable objetivo (supervisados); y (ii) los que no (no supervisados). Debido a la complejidad de la cuestión, se pasará revista únicamente a los métodos supervisados más relevantes, que se pueden dividir en:

- **Métodos tipo filtro**, que puntúan de mayor a menor cada variable predictora en base a su capacidad predictiva y seleccionan un subconjunto de ellas en base a dichas puntuaciones (Brownlee, 2020).
- **Métodos tipo envoltura (wrapper)**, que eligen el subconjunto de variables que dan como resultado el modelo con mayores prestaciones en cuanto a calidad de resultados y eficiencia: error de predicción o clasificación, precisión, tiempo de computación...
- **Métodos intrínsecos** (o *embedded*), que seleccionan las variables automáticamente como parte del ajuste del modelo durante el entrenamiento (tal es el caso de algunos modelos de regresión penalizados, como Lasso, árboles de decisión y bosques aleatorios (*random forests*)).

9.2.1. Pre-selección de variables

9.2.1.1. Varianza nula

Uno de los aspectos fundamentales en la selección de variables es comprobar si su varianza es cero o cercana a cero porque, si es así, sus valores son iguales o similares, respectivamente, y, por tanto, esas variables estarán perfectamente o quasi-perfectamente correladas con el término independiente del modelo, con lo cual, en el mejor de los casos, solo añadirán ruido al modelo. Además, este tipo de variables causan problemas a la hora de dividir el conjunto de datos en subconjuntos de entrenamiento, validación y test. Las causas de una nula o muy pequeña variabilidad pueden estar en haber medido la variable en una escala inapropiada para la variable o en haber expandido una variable politómica en varias dicotómicas (una por categoría), entre otras. En el primer caso, un cambio de escala puede evitar el problema de la colinealidad. Otra opción más drástica la eliminación de la variable.

A continuación, se comprueba si las variables del conjunto de datos `Madrid_Sale` tienen **varianza cero**. Para ello se utiliza la función `nearZeroVar()` del paquete `caret`.

Se seleccionan en primer lugar las variables numéricas en el conjunto de datos `Madrid_Sale_num`.

```
library("idealista18")
library("tidyverse")
library("caret")

Madrid_Sale <- as.data.frame(Madrid_Sale)
numeric_cols <- sapply(Madrid_Sale, is.numeric)
Madrid_Sale_num <- Madrid_Sale[, numeric_cols]
```

Se observa que se devuelve el valor `nzv=FALSE` (`nzv: near zero variance`) para casi todas las variables, con la excepción de `PARKINGSPACEPRICE`, `ISDUPLEX`, `ISSTUDIO`, `ISINTOPFLOOR` y `BUILTYPEID_1`, que podrían descartarse como variables predictoras.

```
varianza <- nearZeroVar(Madrid_Sale_num, saveMetrics = T)
# Con el argumento saveMetrics, se guardan los valores que se han utilizado para los
# cálculos.
# Se muestran los primeros resultados
head(varianza, 2)
#>      freqRatio percentUnique    zeroVar    nzv
#> PERIOD  2.019617   0.004218742 FALSE  FALSE
#> PRICE   1.076923   2.911986500 FALSE  FALSE
```

Para filtrar (excluir) las variables que se descartan como predictoras, se procede como sigue:

```
Madrid_Sale_num <- Madrid_Sale_num |>
  select(-c(PARKINGSPACEPRICE, ISDUPLEX, ISSTUDIO, BUILTYPEID_1))
```

9.2.1.2. Correlación entre variables

Como se avanzó anteriormente, otra de las cuestiones a tener en cuenta en el proceso de selección de variables es la magnitud de las **correlaciones entre las variables** candidatas, pues la existencia de correlaciones elevadas tiene consecuencias perversas sobre la fiabilidad de las predicciones (o de la clasificación realizada). En el caso extremo el modelo tendrá problemas de colinealidad o multicolinealidad (véase Sec. 9.2.1.3).

Para detectar las variables con muy elevada correlación entre ellas, se le pasa la función `findCorrelation()` de `caret`, con valor 0,9, a la matriz de correlaciones lineales entre las variables susceptibles de ser seleccionadas.

```
madrid_cor <- cor(Madrid_Sale_num[, 1:20])
alta_corr <- findCorrelation(madrid_cor, cutoff = .9)
```

Con ello, se comprueba que la variable `HASPARKINGSPACE` tiene correlaciones superiores a 0,9 con varias de las variables predictoras, procediéndose a su eliminación.

```
Madrid_Sale_num <- Madrid_Sale_num[, -alta_corr]
```

la Fig. 9.1, generada con el paquete `corrplot`, muestra las correlaciones existentes entre las primeras variables predictoras.

```
library("corrplot")

matriz_corr <- cor(Madrid_Sale_num[, 1:8])
corrplot(matriz_corr, method = "circle")
```

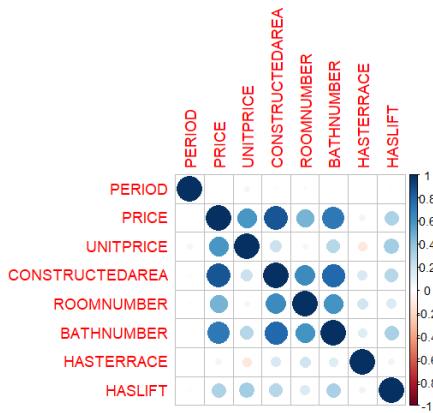


Figura 9.1: Matriz de correlaciones topada en 0,9

Se aprecia que ya no hay variables altamente correlacionadas.

9.2.1.3. Combinaciones lineales

En la práctica, en la mayoría de los casos, por ejemplo en las regresiones lineales, las variables que se utilizan como predictores no son ortogonales sino que tienen cierto grado de dependencia lineal entre ellas. Si dicho grado es moderado, las consecuencias de la no ortogonalidad en la predicción no son graves, pero en los casos de dependencia lineal quasi-perfecta las inferencias resultantes del modelo estimado distan mucho de la realidad. Dichas consecuencias son aún más graves en el caso de que las combinaciones lineales sean perfectas. Por ello, la existencia de colinealidad o combinaciones lineales entre las variables seleccionables también es una circunstancia a evitar. En el caso de que los predictores (o varios de ellos) conformen una o varias combinaciones (o quasi-combinaciones) lineales, no se puede conocer el impacto específico de cada uno de ellos en la variable objetivo, pues dichos impactos se solapan unos con otros. Además, como se ha avanzado, las predicciones no son fiables, entre otras cosas (véase (?)). Y es que se le está pidiendo al conjunto de datos en estudio más información sobre la variable objetivo de la que realmente tiene. Entre otros modelos, la regresión lineal y la regresión logística parten del supuesto de no colinealidad o multicolinealidad entre las variables, por lo que no debería haber variables correlacionadas, ni dos a dos, ni en forma de combinación lineal entre varias de ellas.

Las principales fuentes de multicolinealidad son:

- El método utilizado en la recogida de datos (subespacios).
- Restricciones en el modelo o en la población (existencia de variables correlacionadas).
- Especificación del modelo (polinomios).
- Más variables que observaciones.

En cuanto al detalle de las consecuencias más importantes de la multicolinealidad, hay que señalar las siguientes:

- Los estimadores tendrán grandes varianzas y covarianzas.
- Las estimaciones de los coeficientes del modelos serán demasiado grandes.
- Los signos de los coeficientes estimados suelen ser distintos a los esperados.
- Pequeñas variaciones en los datos, o en la especificación del modelo, provocarán grandes cambios en los coeficientes.

En el ejemplo con los datos del conjunto `Madrid_Sale` se utiliza la función `findLinearCombos()` de `caret` para encontrar, en caso de que las haya, combinaciones lineales de las variables predictoras.

```
Madrid_Sale_num_na <- tidyverse::drop_na(Madrid_Sale_num) # Es necesario eliminar los NA.
combos <- findLinearCombos(Madrid_Sale_num_na)
combos
#$remove
#NULL
```

Como puede comprobarse, no se encuentra ninguna combinación lineal en las variables numéricas de `Madrid_Sale`. En caso de existir, una solución al problema de la multicolinealidad pasa por:

- Eliminar variables predictoras que se encuentren altamente relacionadas con otras que permanecen en el modelo.
- Sustituir las variables predictoras por componentes principales (véase Cap. 32).
- Incluir información externa a los datos originales. Esta alternativa implica utilizar estimadores contraídos (de Stein o ridge) o bayesianos.

A continuación se muestra, caso de existir, cómo se eliminarían las combinaciones lineales:

```
Madrid_Sale_num_na[, -combos$remove]
```

9.2.2. Métodos de selección de variables

Tras la pre-selección de variables llevada a cabo en el epígrafe anterior, procede la selección de variables, propiamente dicha, de entre las que han superado la fase previa, en base, principalmente, a criterios de capacidad predictiva. No obstante, también se utilizan para:

- Simplificar de modelos para hacerlos más interpretables.
- Mejorar la precisión del modelo (si se ha escogido bien el subconjunto de variables).
- Reducir el tiempo de computación; sobre todo, entrenar algoritmos a mayor velocidad.
- Evitar la maldición de la dimensionalidad (o efecto Huges), que se refiere a las consecuencias no deseadas que tienen lugar cuando la dimensionalidad de un problema es muy elevada.
- Reducir la probabilidad de sobreajuste.

9.2.2.1. Métodos tipo filtro

Los **métodos de selección de variables tipo filtro** usan técnicas estadísticas para evaluar la relación entre cada variable predictora (o de entrada, o independiente) y la variable objetivo (o de salida, o dependiente). Generalmente, consideran la influencia de cada variable predictora sobre la variable objetivo por separado. Las puntuaciones obtenidas se utilizan como base para clasificar y elegir las variables predictoras que se utilizarán en el modelo.

La elección de las técnicas estadísticas depende del tipo de variables (objetivo y predictoras). Por ejemplo, si las variables de entrada (predictoras) y salida (objetivo) fueran numéricas, se utilizaría el coeficiente de correlación de Pearson o el de Spearman (dependiendo de si la relación entre la variable predictora y la variable objetivo es lineal o no) o el método de información mutua (véase Vergara and Estévez (2014)). Si ambas fuesen categóricas, podrían usarse medidas de asociación para tablas de contingencia 2×2 o $R \times C$ (véanse Sec. 23.4 y 23.5). Si la de entrada

fuese categórica y la de salida numérica, la técnica adecuada sería el Análisis de la Varianza (ANOVA, véase Sec. 15.4.6.1). Si la categórica fuese la de salida y la numérica la de entrada, entonces habría que acudir a la regresión logística (véase Sec. ??), por ejemplo. Sin embargo, el conjunto de datos no tiene porqué tener sólo un tipo de variable de entrada. Para manejar diferentes tipos de variables de entrada, se pueden seleccionar, por separado, variables de entrada numéricas y variables de entrada categóricas, usando en cada caso las técnicas apropiadas.

Estos métodos suelen eliminar sólo las variables de menor interés a la hora de predecir/clasificar. Permiten ahorrar tiempo y son especialmente robustos para el sobreaprendizaje. Sin embargo, no tienen en cuenta las relaciones entre las variables, lo que puede dar lugar a seleccionar variables redundantes si es que no se ha llevado a cabo una fase de pre-selección.

Nota

Existen diversos paquetes, como `FSelector` (Romanski et al., 2013) y el mismo `caret`, para implementar técnicas de selección de variables. Aquí se utiliza `FSinR`, que contiene una serie de métodos de filtro y envoltura, que se combinan con algoritmos de búsqueda para obtener el subconjunto óptimo de variables, usando funciones para entrenar modelos de clasificación y regresión disponibles en el paquete `caret`. La selección de variables o características se lleva a cabo con la función `FeatureSelection()`, y la del algoritmo de búsqueda que se utilizará en el proceso de selección de funciones se realiza con la función `searchAlgorithm()`. Por su parte, los métodos de filtrado se implementan a través de la función `filterEvaluator()`. No debe olvidarse que, antes de realizar el proceso de selección de variables, el usuario tiene que dividir el conjunto de datos convenientemente para llevar a cabo cada operación sobre el subconjunto correcto (véase Cap. 10). Igualmente, también de manera previa, se tiene que resolver el problema de los datos faltantes.

A continuación se muestra un ejemplo para variables predictoras numéricas. Para ello, se toma una muestra del conjunto de datos `Madrid_Sale_num`, obtenido en la Sec. 9.2.1.1. Una vez en disposición de la muestra, primeramente se transforma la variable objetivo en categórica, siendo las categorías (intervalos) cuatro cortes de la distribución de sus valores; dicha categorización se lleva a cabo mediante **binning**.¹ También se eliminan los registros con datos faltantes.

¹ *Binning* (anglicismo que deriva de la palabra *bin*: cubo, cesta, contenedor) es una técnica de discretización que agrupa datos numéricos en intervalos. Se suele utilizar para simplificar el análisis de datos continuos y aumentar la interpretabilidad del modelo, si bien a costa de reducir las combinaciones de las categorías de las variables predictoras que pueden realizarse, con lo cual el modelo sólo podrá hacer predicciones para unas pocas combinaciones de categorías de las variables predictoras. El *binning* puede ser supervisado o no (agrupamiento automático o manual). En este último caso, hay que tomar muchas precauciones porque, como señala Kuhn et al. (2013), (i) el *binning* en las variables predictoras puede llevar a una pérdida significativa en la capacidad del modelo a la hora de determinar la relación (sobre todo si es compleja) entre los predictores y la variable objetivo; y (ii) en el entorno clasificadorio, puede dar lugar a una alta tasa de falsos positivos. Estas limitaciones pueden superarse en el caso de que el *binning* se lleve a cabo de forma supervisada (tal es el caso de los árboles de regresión y clasificación y de la regresión adaptativa multivariante con splines), si bien debe tenerse en cuenta que, aunque se utilizan todos los predictores para llevar a cabo el proceso de *binning*, la categorización está guiada por un único objetivo (por ejemplo, maximizar la exactitud).

```

library("rsample")

# Se toma una muestra con el paquete rsample
set.seed(7)
Madrid_Sale_num_sample <- sample(1:nrow(Madrid_Sale_num), size = 5000, replace = FALSE)
Madrid_Sale_num_sample <- Madrid_Sale_num[Madrid_Sale_num_sample, ]
# Se realiza binning con cuatro bins
Madrid_Sale_num_sample_bin <- Madrid_Sale_num_sample |>
  mutate(price_bin = cut(PRICE, breaks = c(0, 250000, 500000, 750000, 10000000), labels
    ~ = c("primerQ", "segundoQ", "tercerQ", "c"), include.lowest = TRUE)) |>
  select(price_bin, CONSTRUCTEDAREA, ROOMNUMBER, BATHNUMBER, HASTERRACE, HASLIFT)
# Se eliminan los registros con valores missing
Madrid_Sale_sample_na <- drop_na(Madrid_Sale_num_sample_bin)

```

Una vez discretizada la variable objetivo, se selecciona el conjunto de variables predictoras de la variable objetivo `price_bin`, que es la variable `PRICE` transformada mediante *binning*. Como método tipo filtro se utiliza `minimum description length` (MDLM), que es un método de selección de variables que se basa en una medida de la complejidad del modelo denominada “longitud mínima de la descripción” (de ahí el nombre del modelo), por lo que su objetivo es encontrar el modelo más sencillo que proporcione una explicación aceptable de los datos. Como algoritmo de búsqueda se utiliza `sequential forward selection`.²

```

library("FSinR")

# Método tipo filtro MDLC (Minimum-Description-Length-Criterion)
evaluador <- filterEvaluator("MDLC")
# Se genera el algoritmo de búsqueda
buscador <- searchAlgorithm("sequentialForwardSelection")
# Se implementa el proceso, pasando a la función los dos parámetros anteriores
resultados <- featureSelection(Madrid_Sale_sample_na, "price_bin", buscador, evaluador)
# Se muestran los resultados
resultados$bestFeatures
#>      CONSTRUCTEDAREA ROOMNUMBER BATHNUMBER HASTERRACE HASLIFT
#> [1,]          0          0          0          1          0
resultados$bestValue
#> [1] 355.3439

```

En este caso, con los argumentos propuestos, el modelo seleccionado para explicar el comportamiento de la variable objetivo `price_bin` contiene únicamente el término independiente y una variable predictora: `HASTERRACE`.

²Este método (selección hacia adelante) consiste en ajustar primero un modelo que contenga únicamente un término independiente (o intercepto); es decir, sin variables predictoras. Posteriormente, se ajusta otro con término independiente y una sola variable predictora. Después, se ajusta otro modelo con término independiente y dos variables predictoras. Y así sucesivamente. El criterio de parada suele ser que el valor del criterio de información de Akaike no experimente una reducción significativa al añadir una variable más al modelo. Los otros dos métodos o criterios para moverse en el espacio de búsqueda de subconjuntos de variables predictoras son el `backward` (selección hacia atrás), que funciona justo al revés que el `forward`, por eliminación de variables, y la selección paso a paso, `stepwise` que es una combinación de los dos anteriores.

9.2.2.2. Métodos de selección de variables tipo envoltura (*wrapper*)

Este enfoque realiza una búsqueda a través de diferentes combinaciones o subconjuntos de variables predictoras/clasificadoras para comprobar el efecto que tienen en la precisión del modelo (Saeys et al., 2007).

Hay varias alternativas:

- Evaluar las variables individualmente y seleccionar las n variables principales que obtienen unas buenas prestaciones, aunque se pierde la información de las dependencias entre variables.
- Observar el rendimiento del modelo para todas las combinaciones de variables posibles. En este sentido, se puede utilizar un algoritmo de búsqueda global estocástica, como los algoritmos genéticos que, si bien pueden ser efectivos, también pueden ser computacionalmente muy costosos.

Los métodos **wrapper** son de gran eficacia a la hora de eliminar variables irrelevantes y/o redundantes (cosa que no ocurre en los de tipo filtro porque se centran en el poder predictor de cada variable de forma aislada). Además, tienen en cuenta la circunstancia de que dos o más variables, aparentemente irrelevantes en cuanto a su capacidad predictiva o clasificatoria cuando se consideran una por una, pueden ser relevantes cuando se consideran conjuntamente. Sin embargo, son muy lentos, ya que tienen que aplicar muchísimas veces el algoritmo de búsqueda, cambiando, cada vez, el número de variables, siguiendo, cada vez, algún criterio tanto de búsqueda como de paro. En lo que respecta a los criterios de búsqueda, estos son similares a los de los métodos tipo filtro. Por lo que se refiere a los criterios de paro, los usados en los métodos *wrapper* son menos eficientes que los criterios basados en algún tipo de medida de ganancia de información, distancia o consistencia, entre el predictor y la variable objetivo (o clase) que utilizan los de tipo filtro.

Nota

Las principales diferencias entre los métodos tipo filtro y tipo envoltura son las siguientes:

- Los métodos de filtro cuantifican la relevancia de las variables por su correlación con la variable salida, mientras que los métodos de tipo envoltura cuantifican las prestaciones del modelo para diferentes subconjuntos de variables.
- Los métodos de filtro tienen una carga computacional enormemente inferior a la de los envolventes, ya que no necesitan entrenar ningún modelo.
- Los métodos de filtro utilizan métodos estadísticos para evaluar la selección de variables; los de tipo envoltura utilizan métodos de validación cruzada.
- En la mayoría de ocasiones, la selección de variables realizada por los métodos tipo envoltura suele ser más exitosa que la proporcionada por los métodos de filtro.
- Los métodos de envoltura tienen una probabilidad de sobreajuste mucho mayor que los de filtro.

A continuación, sobre el conjunto de datos `Madrid_Sale_sample_na` y sobre la variable objetivo `price_bin` se establecen tanto los parámetros de los algoritmos de búsqueda como los métodos

de filtrado y se calculan los resultados, usando para ello de `FSinR`. Como método de selección de variables se utiliza un método **wrapper** (con la función `wrapperEvaluator()`de `FSinR`) y como algoritmo de búsqueda **sequential forward selection**.

```
# Se fijan los parámetros
evaluador <- wrapperEvaluator("rpart1SE")
buscador <- searchAlgorithm("sequentialForwardSelection")
# Se evalúan sobre Madrid_Sale_sample_na
results <- featureSelection(Madrid_Sale_sample_na, "price_bin", buscador, evaluador)
resultados$bestFeatures
resultados$bestValue
```

El resultado es el mismo que con el con el método tipo filtro anteriormente utilizado: el modelo seleccionado para explicar el comportamiento de la variable objetivo `price_bin` contiene únicamente el término independiente y una variable predictora: `HASTERRACE`.

Nota

Se puede sofisticar más el modelo ajustando los parámetros del modelo con parámetros de remuestreo, que son los mismos argumentos que se pasan a la función `trainControl()` del paquete `caret`. En segundo lugar, se pueden establecer los parámetros de ajuste, que son los mismos que para la función de `train` de `caret`.

9.2.2.3. Métodos de selección tipo intrínseco (*embedded*)

Finalmente, hay algunos algoritmos de aprendizaje automático que realizan la selección automática de variables como parte del aprendizaje del modelo. Estos son los métodos de selección de tipo intrínseco, que aglutinan las ventajas de los métodos de filtro y envoltura.

Un ejemplo son los relativos a los modelos de regresión penalizados, como Lasso, o *ridge* (que tienen funciones de penalización incluidas para reducir el sobreajuste), árboles de decisión y bosques aleatorios.

En el siguiente ejemplo se modeliza un bosque aleatorio (usando el paquete `randomForest`) y, tras dicha modelización, se identifica el conjunto óptimo de variables con la función `varImp()` de `caret`.

```
library("randomForest")

# Usar random forest para la selección de variables
rf_modelo <- randomForest(price_bin ~ ., data = Madrid_Sale_num_sample_bin)

# Listar las variables más importantes
varImp(rf_modelo)
```

Con este método de selección de variables, el modelo con mayor poder predictivo de la variable salida `price_bin` es el que contiene un término independiente y los predictores `CONSTRUCTEDAREA`, `ROOMNUMBER` y `BATHNUMBER` (ejecútese el código para comprobarlo).

9.3. Transformación de variables

La transformación y creación de variables predictoras a partir de los datos en bruto tiene una componente técnica y otra más creativa; en esta última, son de gran relevancia la intuición y la experiencia en trabajos de modelado, así como el dominio de los datos en cuestión. Para labores de transformación también se utilizará el paquete `caret`.

Nota

`Caret` se ha elegido como herramienta principal para la parte de preprocesamiento por su amplia difusión y porque también se utiliza en la parte de *machine learning* supervisado de este libro. No obstante, se podrían usar otros paquetes, como `recipes`, incluido en `tidymodels`. Este tipo de paquetes, comúnmente llamados metapaquete (*meta-packages*), permiten agrupar varios programas junto a sus dependencias para su instalación de una vez. Por tanto, un metapaquete permite ahorrar tiempo y esfuerzo a la vez que facilita la implementación de múltiples modelos en paralelo para, posteriormente, vincular sus resultados.

La fase de modelización puede condicionar la fase previa de preparación de datos. Por ejemplo, determinadas técnicas imponen requisitos y expectativas sobre el tipo y forma de las variables predictoras (Boehmke and Greenwell, 2019). Así, podría ser necesario que la variable objetivo tenga una distribución de probabilidad específica, o la eliminación de variables predictoras altamente correlacionadas con otras y/o que no estén fuertemente relacionadas con la variable objetivo.

Generalmente, estas transformaciones son más útiles para algoritmos como los de regresión, métodos basados en instancias (también llamados *memory-based learning methods*, como *k*-vecinos más cercanos -KNN- y *Learning Vector Quantization* -LVQ-), máquinas de vectores de soporte -SVM- y redes neuronales -NN-, que para métodos basados en árboles y reglas.³

9.3.1. Transformación de la distribución de la variable objetivo

Aunque no siempre es necesario, la transformación de la distribución de la variable objetivo puede llevar a una mejora predictiva significativa, especialmente en el caso de modelos paramétricos. Por ejemplo, los modelos de regresión lineal ordinarios asumen que el término de error, y, por consiguiente, la variable objetivo, se distribuyen normalmente. Pero puede ocurrir, por

³Una de las varias clasificaciones existentes de los métodos de aprendizaje los divide en basados en instancias (muestras u observaciones del conjunto de entrenamiento) o en modelos. Los algoritmos basados en instancias “memorizan” dichas instancias, y utilizan esta información a la hora de realizar una predicción. El aprendizaje basado en modelos tiene como objetivo la creación de un modelo a partir de los datos de entrenamiento, con el cual se harán las predicciones.

ejemplo, que la variable objetivo tenga valores atípicos y la suposición de normalidad no se cumpla por asimetría.

Para simetrizar la distribución de probabilidad de la variable objetivo (mejorando así la dispersión de valores y, a veces, desenmascarando las relaciones lineales y aditivas entre los predictores y el objetivo) se puede usar una transformación log (entre otras). Para corregir la asimetría positiva de la distribución probabilística de la variable objetivo se suele utilizar una de las dos opciones siguientes:

- **Normalizar con una transformación logarítmica**, que proporciona buenos resultados en la mayoría de los casos. En la Fig. 9.2, se puede comprobar que, en el ejemplo que se viene arrastrando, una transformación logarítmica normaliza, en gran medida, la distribución de la variable PRICE. Nótese que, si la variable objetivo tiene valores negativos o cero, una transformación logarítmica producirá *Nan* y *-Inf*, respectivamente. Si los valores de respuesta no positivos son pequeños (por ejemplo, entre -0,99 y 0), se puede aplicar una pequeña compensación (por ejemplo, la función `log1p()` agrega un 1 al valor antes de aplicar la transformación).

```
respuesta_log <- log(Madrid_Sale$PRICE)
```

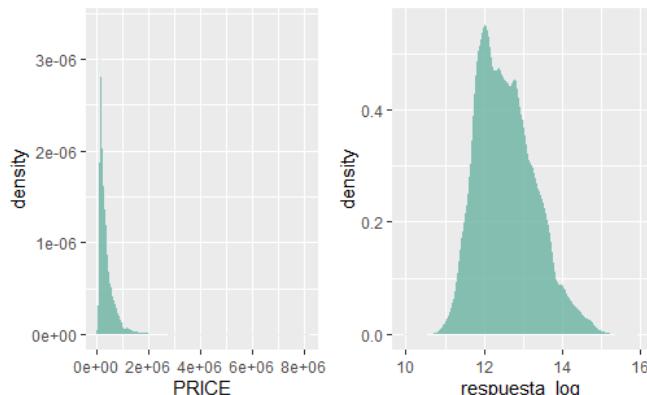


Figura 9.2: Normalización logarítmica

- Como segunda opción, se puede usar una transformación de la familia de transformaciones Box-Cox (o simplemente una **transformación de Box-Cox**), de carácter potencial y con mayor flexibilidad que la transformación logarítmica. Generalmente, se puede encontrar la función adecuada a partir de una familia de transformadas de potencia, que llevarán la distribución de la variable transformada tan cerca como sea posible de la distribución normal [Sakia (1992)]⁴. No obstante, igual que la transformación logarítmica, las transformaciones del tipo Box-Cox también tienen la limitación de ser sólo aplicables a variables

⁴La piedra angular de la transformación de Box-Cox es el exponente de dicha transformación (λ), que varía entre -5 y 5.

cuyos valores sean positivos. Por consiguiente, tanto si se usa una transformación log como una Box-Cox, no se deben centrar los datos primero, ni realizar ninguna operación que pueda hacer que los valores de la variable transformada no sean positivos.

- En caso de valores nulos o negativos, una muy buena opción, la tercera, es la transformación Yeo-Johnson, que es una extensión de la transformación Box-Cox que no está limitada a los valores positivos.

```
respuesta_boxcox <- preProcess(Madrid_Sale_num_sample, method = "BoxCox")
trainBC <- predict(respuesta_boxcox, Madrid_Sale_num_sample)
respuesta_boxcox
#> Created from 5000 samples and 2 variables
#>
#> Pre-processing:
#> - Box-Cox transformation (2)
#> - ignored (0)
#>
#> Lambda estimates for Box-Cox transformation:
#> -0.3, -0.3
```

Hay que tener en cuenta que, cuando se modela con una variable objetivo transformada, las predicciones también estarán en la escala transformada. Es posible que haya que deshacer (o volver a transformar) los valores pronosticados a su escala original, para que los responsables de la toma de decisiones puedan interpretar los resultados más fácilmente.

Nota

El paquete **recipes** (recetas de cocina), incluido en **tidymodels**, permite la transformación de variables de forma secuencial. La transformación de la distribución probabilística de la variable objetivo con **recipes** se lleva a cabo en 4 etapas: (i) **recipe()**, donde se especifica la fórmula (variables predictoras y variable objetivo); (ii) **step()**, donde se definen los pasos a seguir: imputación de valores perdidos, creación de variables ficticias (dummies), normalización, etc.; (iii) **prep** (preparar, o en otros términos, entrenar), donde que se utiliza un conjunto de datos para analizar cada paso en él; y (iv) **bake** (hornear/cocinar), donde, una vez aplicada la receta, se aplica al conjunto de datos. La idea detrás de **recipes** es similar a **caret::preProcess()**. Sin embargo, a diferencia de **caret**, no maneja automáticamente las variables categóricas y requiere crear variables ficticias manualmente.

9.3.2. Cambios de origen y escala en las variables (normalizaciones)

La escala en que se miden las variables individuales no es una cuestión baladí a la hora de la modelización. Los modelos que incorporan funciones lineales en las variables predictoras, son sensibles a la escala de esas variables. Lo mismo puede decirse de los algoritmos que utilizan medidas de distancia, como los de agrupación y clasificación, o los de escalamiento multidimensional, entre otros; o los de reducción de la dimensionalidad. Cuando se estiman modelos,

a menudo es aconsejable modificar la escala de las variables predictoras; el objetivo es evitar que unas variables tengan mayor influencia que otras en el resultado obtenido. Por ejemplo, en el conjunto de datos `Madrid_Sale` la superficie de las viviendas, medida en metros cuadrados, tiene una media y una desviación típica mayores que la antigüedad de la misma, medida en años. En consecuencia, los algoritmos basados en la magnitud de los errores pueden dar más importancia a las variables con mayor desviación típica, pero no porque tengan mayor variabilidad real que las otras, sino porque la medida de dicha variabilidad (la desviación típica) es más grande debido a la distinta escala en la que están medidas dichas variables. La consecuencia: efectos perniciosos indeseados sobre las predicción o la clasificación.

La normalización de variables tiene como objetivo que las comparaciones entre estas variables, en cuanto a su contribución al análisis de interés, sean objetivas; es decir, ponerlas en igualdad de condiciones en lo que respecta a su influencia (más allá de la que realmente tienen) en la variable objetivo.

La estandarización (o normalización *z-score*) es el método de normalización de variables más popular. Consiste en restar la media de la variable a sus los valores y, posteriormente, dividir esta diferencia entre la desviación típica de la variable. De esta manera, las variables (numéricas) transformadas tendrán media nula y varianza unitaria, lo que proporciona una unidad de medida comparable común a todas las variables: la distancia a la media medida en términos de desviaciones típicas.

A modo de ejemplo, a continuación se estandarizan las variables del conjunto de datos `Madrid_Sale` con la función `preProcess()` de `caret` y el `method=c('center', 'scale')`, de tal manera su media sea nula y su desviación típica unitaria.

```
prep_centrado <- preprocess(Madrid_Sale_num, method = c("center", "scale"))
pred_centrado<-predict(prep_centrado, Madrid_Sale_num)[1:3]
head(pred_centrado, n = 3)
#>      PRICE    CONSTRUCTEDAREA      ROOMNUMBER
#1 -1.523435   -0.64763050   -0.5764192
#2 -1.523435   -0.38628625   0.4062338
#3 -1.523435   -0.05541004   0.7717038
```

Otra normalización también popular es la **min-max**, que re-escala los valores de la variable entre 0 y 1, o entre -1 y 1, y cuya expresión general es:

$$X_{norm} = \frac{X - \min(X)}{\max(X) - \min(X)}.$$

Si se desea re-escalar entre dos valores arbitrarios, a y b, la expresión anterior se transforma como sigue:

$$X_{norm} = a + \frac{(X - \min(X))(b - a)}{\max(X) - \min(X)},$$

Otras opciones de normalización pueden verse en la amplia literatura sobre la cuestión.

Finalmente, recordar que, cuando se lleva a cabo un proceso de normalización de variables, hay que hacerlo tanto en el subconjunto de entrenamiento como en el de test, para que ambos se basen en la misma media y varianza.

9.3.3. Ingeniería de variables (*feature engineering*)

La ingeniería de variables consiste en el proceso de conseguir, a partir de la información disponible, las variables idóneas (y el en número apropiado) para que los modelos o clasificadores proporcionen los mejores resultados posibles, dados los datos disponibles y el modelo a ejecutar. En otros términos, es el proceso de transformación de las variables seleccionadas, de forma que se obtenga el mejor rendimiento posible de los modelos de *machine learning*. Por ejemplo, transformar las variables relacionadas con la fecha de tal manera que se diferencie según el tipo de horario (“de oficina” y “de descanso”), o que se considere la cercanía al momento actual (los datos más cercanos contienen más información); los filtros de imagen (desenfocar una imagen) y la conversión de texto en números (utilizando el procesamiento avanzado del lenguaje natural, que asigna palabras a un espacio vectorial) son también ejemplos interesantes.

La mayoría de los modelos requieren que los predictores tengan forma numérica, por lo que, en caso de tener predictores de carácter categórico, hay que transformarlos en numéricos. Para implementar otro tipo de modelos, conviene transformar alguna(s) variable numérica en categórica. En el primer caso, conviene aplicar técnicas de **agrupamiento** (o *binning*), que crean agrupaciones o intervalos a partir de variables continuas; en el segundo, las técnicas de **codificación**, permiten tratar variables categóricas como si fueran continuas. Hay casos, como el de los modelos basados en árboles, que manejan, de manera natural, variables numéricas y categóricas; pero incluso en estos modelos se puede mejorar su rendimiento si se preprocesan las variables categóricas.

La identificación entre las labores de selección y de transformación de variables es bastante frecuente; sin embargo, es errónea, pues, si bien tienen algunos solapamientos, sus objetivos son claramente distintos. La ingeniería de variables tiene como objetivo la construcción de modelos más sofisticados y más interpretables que los que se pueden implementar con los datos tal y como están en el fichero raíz. La selección de variables permite que el modelo sea manejable, mejorando su interpretabilidad sin que por ello se reduzca significativamente el rendimiento del modelo.

El proceso de agrupamiento ya ha sido referido e ilustrado en la Sec. 9.2.2.1. En cuanto al proceso de codificación, se pueden distinguir dos tipos:

- **Codificación de etiquetas:** consiste en asignar a cada etiqueta un entero o valor único según el orden alfabético. Es la codificación más popular y ampliamente utilizada.
- **Codificación one-hot:** consiste en crear una nueva variable ficticia (*dummy*) binaria por cada categoría existente en la variable a codificar. Estas nuevas variables contendrán un 1 en aquellas observaciones que pertenezcan a esa categoría, y un 0 en el resto.⁵

⁵En muchas tareas, como, por ejemplo, en la regresión lineal, es común usar $k - 1$ variables binarias en lugar de k , siendo k el número total de categorías. Esto se debe a que la k -ésima variable binaria es redundante, ya que no es más que una combinación lineal de las otras, y, además, provocará problemas numéricos. Por otra

Para ejemplificar este tipo de codificación, a continuación, en el conjunto de datos `Madrid_Sale_num_sample_bin`, se crean *dummies*, una para cada cada categoría de las variables objeto de codificación. Para ello, se utiliza la función `dummyVars()` de `caret`. El resultado puede verse con la función `predict()`.

```
dummies <- dummyVars(~ ., data = Madrid_Sale_num_sample_bin)
head(predict(dummies, newdata = Madrid_Sale_num_sample_bin))
```

No debe olvidarse, igual que para todas las transformaciones descritas, hacer las mismas transformaciones en el conjunto de test.

9.4. Reducción de dimensionalidad

La **reducción de dimensionalidad** es un enfoque alternativo para filtrar las variables no informativas sin eliminarlas (como se hacía en la Sec. 9.2, que generalmente se usa para variables numéricas). La diferencia es que las técnicas de reducción de la dimensionalidad crean una proyección de los datos que da como resultado variables predictoras completamente nuevas, que son combinaciones lineales independientes formadas a partir de las variables originales, solucionando así, también, los problemas de colinealidad y multicolinealidad (perfecta o quasi-perfecta). Como se explica en el Cap. 32, el espacio de un conjunto de variables puede reducirse proyectándolo a un subespacio de variables de menor dimensión utilizando componentes principales (la técnica de reducción de la dimensionalidad por anonomasia).

Resumen

- Se presentan las principales técnicas y métodos de *feature selection* para llevar a cabo la selección (pre-selección y selección propiamente dicha) de las variables predictoras o clasificadoras más relevantes para obtener predicciones o clasificaciones exitosas.
- Se describen las principales transformaciones que se realizan en la fase de pre-procesamiento de un proyecto de modelado predictivo: las transformaciones de la escala o de la distribución de la variable objetivo, la transformación de variables (*feature engineering*) y la reducción de la dimensionalidad.
- La creación de variables predictoras a partir de los datos en bruto tiene una componente creativa, que requiere de herramientas adecuadas y de experiencia para encontrar las mejores representaciones, apoyándose, en la medida de lo posible en el conocimiento que se tenga de los datos.
- Las labores de selección y transformación de variables se ilustran con el conjunto de datos de `Madrid_Sale`, utilizándose los paquetes `caret` y `rsample`.

parte, la no inclusión de dicha variable no implica pérdida de información alguna, ya que se entiende que, si el resto de las categorías contienen un 0, la categoría correspondiente es la de la categoría eliminada.

Capítulo 10

Herramientas para el análisis en ciencia de datos

José-María Montero^a y Jorge Velasco López^b

^aUniversidad de Castilla-La Mancha ^bInstituto Nacional de Estadística de España

10.1. Introducción

En este capítulo se describen una serie de herramientas necesarias para desarrollar proyectos de ciencia de datos. Son herramientas que se utilizan pre- o post- modelado de los datos y que aumentan significativamente el rendimiento de los modelos. Tal caja de herramientas incluye el particionado del conjunto de datos, el manejo de datos no equilibrados¹, los métodos de remuestreo, el equilibrio entre sesgo y varianza, el ajuste de hiperparámetros y la evaluación de modelos, entre otras.

Para ilustrar el manejo de las herramientas anteriormente mencionadas, se utiliza el conjunto de datos `Madrid_Sale` (disponible en el paquete de **R** `Idealista18`), con datos inmobiliarios del año 2018 para el municipio de Madrid. En cuanto al software **R**, se utiliza `caret` (Kuhn, 2008), para diversas tareas de preparación de datos, y `rsample`, para muestreo.

10.2. Partición del conjunto de datos

El objetivo principal del proceso de ciencia de datos es encontrar el modelo o algoritmo que mejor resuelva la pregunta de investigación o, lo que es lo mismo, que proporcione mejores resultados.

¹El término inglés *unbalanced data* se suele traducir, en español, en lenguaje formal, por “datos no equilibrados”, si bien en la jerga de ciencia de datos, a nivel de divulgación, también se utiliza la expresión “datos desbalanceados”.

Por ejemplo, en el caso de los modelos de predicción (y en general de aquellos en los que el aprendizaje es supervisado), muy populares en la ciencia de datos, el que prediga con mayor exactitud los valores futuros de la variable objetivo a partir de los predictores seleccionados en el conjunto de datos disponible. En otras palabras, un algoritmo que, no sólo ajuste bien los datos pasados sino, lo que es más importante, que proporcione predicciones (futuras) acertadas (y precisas). Para ello, inicialmente, se dividen los datos en dos subconjuntos:

- **de entrenamiento (*train*)**: se utiliza para desarrollar conjuntos de funciones, entrenar algoritmos, ajustar hiperparámetros, comparar modelos y realizar todas las demás actividades necesarias para seleccionar un modelo final.
- **de prueba (*test*)**: se utiliza para validar la precisión del modelo seleccionado en la fase de entrenamiento.

A la hora de dividir el conjunto de datos en los dos subconjuntos anteriores, hay que tomar dos decisiones:

- ¿Qué porcentaje de los datos (casos, observaciones) se incluye en cada subconjunto?
- ¿Cómo se seleccionan los casos u observaciones que van a cada subconjunto?

Por lo que se refiere a la primera decisión, cuanto más grande sea el subconjunto de entrenamiento mejor será el predictor (o clasificador), aunque las mejoras serán cada vez más pequeñas. Por el contrario, cuanto más grande sea el subconjunto de prueba o test, más precisa será la estimación del error de predicción. En otros términos, lo ideal sería tener un conjunto de datos muy grande y que ambos subconjuntos fueran grandes. De esta manera, los errores de predicción serían pequeños y tendrían poca variabilidad. Sin embargo, con frecuencia, este no es el caso en la práctica, y el dilema es elegir un buen predictor (o clasificador) o una buena estimación del error de predicción. En la práctica, lo más frecuente es incluir el 70 % de los datos en el subconjunto de entrenamiento y el 30 % restante en el de test, aunque los repartos 80 %-20 % y 60 %-40 % también son muy populares.

En cuanto a la segunda decisión, la respuesta es: mediante métodos de muestreo, siendo los más utilizados el muestreo aleatorio simple y el muestreo aleatorio estratificado (véase Cap. 13).

10.2.1. Muestreo aleatorio simple

La forma más sencilla de asignar los datos a los subconjuntos de entrenamiento y prueba, es tomar una **muestra aleatoria simple** (m.a.s.) (véase Sec. 13.2) del conjunto de casos u observaciones del tamaño deseado, y asignarlos al subconjunto de entrenamiento, asignándose los restantes al conjunto de test.

Un problema que puede surgir con las m.a.s. es que, cuando el conjunto de datos es pequeño y los valores de uno (o más) de los predictores estén muy desequilibrados (por ejemplo, el predictor es binario y el 95 % de sus valores pertenecen a una clase o categoría y el 5 % restante a la otra), hay una probabilidad nada desdeñable de que en alguno de los dos subconjuntos (sobre todo en el de test) dicho predictor no esté representado. Si esta circunstancia ocurriese

en el conjunto de entrenamiento, algunos algoritmos darían error al aplicarlos al conjunto de test (donde habría datos de un predictor más). Si, por el contrario, ocurriese en el conjunto de test, los problemas surgirían por haber un predictor menos que en el conjunto de entrenamiento. Los problemas se agravarían si la desproporción anterior tuviese lugar en la variable objetivo.

A continuación, se realiza una división² 70%-30% en el conjunto de datos `Madrid_Sale_num`, generado en el Cap. 9. Para que se pueda reproducir, se establece al principio una semilla determinada.

```
library ("caret")

set.seed(123) # para permitir reproducirlo
index <- createDataPartition(Madrid_Sale_num$PRICE, p = 0.7, list = FALSE)
train <- Madrid_Sale_num[index, ]
test <- Madrid_Sale_num[-index, ]
dim(Madrid_Sale_num) # 94815
dim(train) # 66373
dim(test) # 28442
```

Como puede comprobarse, de los 94.815 datos que contiene `Madrid_Sale_num`, 66.373 (el 70%), pasan a formar parte del conjunto de entrenamiento y, el resto, 28.442, constituyen el conjunto de test.

10.2.2. Muestreo estratificado

Si se desea controlar el muestreo para que los subconjuntos de entrenamiento y prueba tengan distribuciones similares en las clases de la variable objetivo³, se puede usar **muestreo estratificado** (véase Sec. 14.3).⁴ Sin embargo, este tipo de muestreo, estratificando por la variable objetivo, no garantiza que ocurra lo mismo con los predictores. Es decir, presenta la misma limitación que el muestreo aleatorio simple en caso de que algún (o algunos) predictores estén muy desequilibrados y se quiera garantizar que en ambos subconjuntos las clases de la variable respuesta estén representadas de forma similar. Una posible solución es la eliminación de dichos predictores (que tendrán varianza próxima a cero), si bien ello implica pérdida de información.

A continuación, se estratifica el conjunto de datos `Madrid_Sale_num_sample_bin` del Cap. 9 por la variable objetivo (`price_bin`, el precio de venta con *binning*), que tiene cuatro categorías.

```
library("rsample")

set.seed(123) # para permitir reproducirlo
```

²Por defecto, salvo que la variable sea categórica, la división se realiza mediante muestreo aleatorio simple; por tanto, a diferencia del caso de muestreo aleatorio estratificado, que se verá a continuación, no se indica el argumento de las funciones `training` y `testing`.

³Es decir, que cada clase esté representada con, aproximadamente, las mismas proporciones en los dos subconjuntos.

⁴El muestreo estratificado también puede ser de utilidad en problemas de predicción cuando el conjunto de datos es pequeño y la distribución probabilística de la variable objetivo se desvía mucho de la normalidad.

```
table(Madrid_Sale_num_sample_bin$price_bin) |> prop.table()
#      0.4776    0.3062    0.1024    0.1116
split_estrat <- initial_split(Madrid_Sale_num_sample_bin, prop = 0.7, strata =
  ~ "price_bin")
train_estrat <- training(split_estrat)
test_estrat <- testing(split_estrat)
```

Como puede comprobarse debajo, al generar muestras aleatorias estratificadas por la variable objetivo, la distribución de éstas en los subconjuntos de entrenamiento y de prueba es aproximadamente igual:

```
table(train_estrat$price_bin) |> prop.table()
# 0.4777015 0.2913093 0.1132075 0.1177816
table(test_estrat$price_bin) |> prop.table()
# 0.4799886 0.3061750 0.1023442 0.1114923
```

10.3. Técnicas para manejar datos no equilibrados

A menudo, los datos utilizados en determinadas áreas tienen menos del 1% de eventos raros, pero precisamente su rareza es lo que los hace “interesantes”: por ejemplo, estafas en operaciones bancarias o usuarios que hacen *clic* en anuncios. En otros términos, una de las clases de la variable objetivo es dominante, pero la clase minoritaria es la que presenta interés. Sin embargo, la mayoría de los algoritmos no funcionan bien con variables cuyas clases están desequilibradas (Kuhn et al., 2013). Hay varias técnicas para manejar este problema:

- ***Downsampling***⁵: equilibra el conjunto de datos reduciendo el tamaño de las clases abundantes para que coincida con el de la clase menos prevalente. Este método es de utilidad cuando el tamaño del conjunto de datos es suficientemente grande para ser aplicado.
- ***Upsampling***⁶: equilibra el conjunto de datos aumentando el tamaño de las clases más raras. En lugar de deshacerse de datos de las clases abundantes, se generan nuevos datos para las clases raras mediante repetición o *bootstrapping*. Este procedimiento es de utilidad cuando no hay suficientes datos en la clase (o clases) rara.
- **Creación de datos sintéticos**: esta técnica consiste en equilibrar el conjunto de entrenamiento generando nuevos registros sintéticos, esto es, inventados, de la clase minoritaria. Existen diversos algoritmos que realizan esta tarea, siendo uno de los más conocidos la técnica de SMOTE (*Synthetic Minority Oversampling Technique*) (Chawla et al., 2002).
- **Otras técnicas**: como que el algoritmo implemente mecanismos para dar mayor peso a los casos de la clase minoritaria, etc.

A modo de ejemplo, a continuación se utiliza *downsampling* en el conjunto de datos *Madrid_Sale_num_sample_bin*, para mejorar la precisión del modelo, mediante el algoritmo *gradient-boosting*:

⁵También denominado *undersampling*.

⁶También denominado *oversampling*.

```
# Se especifica que el modelo se entrene con downsampling
ctrl <- trainControl(
  method = "repeatedcv", repeats = 5,
  classProbs = TRUE,
  sampling = "down"
)
Madrid_Sale_num_sample_bin_downsample <- train(price_bin ~ .,
  data = Madrid_Sale_num_sample_bin,
  method = "gbm",
  preProcess = c("range"),
  verbose = FALSE,
  trControl = ctrl
)
```

No existe una ventaja absoluta de un método sobre otro. La aplicación de estos métodos depende del caso de uso al que se aplique y del conjunto de datos. La función de `caret` para implementar estas técnicas está en `?caret::trainControl()`.

Una alternativa a los métodos anteriores es la implementación de algoritmos que proporcionen un buen rendimiento con variables cuyas clases están desequilibradas, de tal manera que se refuerce el aprendizaje en la clase minoritaria. La idea detrás de esta segunda opción es incluir una penalización o un sesgo que pondere las clases de tal manera que se le dé más importancia a la predicción, clasificación, etc. correcta en la clase minoritaria (para más detalles, véase ([García Abad et al., 2021](#))).

10.4. El enfoque de validación

Anteriormente, se indicaba que los datos deben dividirse en dos subconjuntos, uno de entrenamiento y otro de prueba, y que no debía usarse el subconjunto de prueba para evaluar la exactitud del modelo durante la fase de entrenamiento. Si la exactitud de las predicciones (por ejemplo, porcentaje de casos bien clasificados en un problema de clasificación) en el conjunto de test es (además de elevada) similar a la que se obtiene en el conjunto de entrenamiento, entonces el modelo entrenado generalizará bien para otros conjuntos de datos y puede darse por bueno. En otro caso, el modelo no ha entrenado bien. Por ejemplo, si la exactitud de las predicciones en el conjunto de entrenamiento es del 80 % y en el de test es del 25 % o del 97 %, entonces el modelo no puede darse por válido. En el caso del 97 %, la diferencia de porcentajes está indicando un aprendizaje “excesivo” del conjunto de datos de entrenamiento, de tal manera que el modelo en cuestión puede proporcionar muy buenas predicciones para el conjunto de datos utilizado, pero no para nuevos datos, o conjuntos de datos (esta circunstancia se conoce como sobreajuste u *overfitting*.⁷

En el caso en el que la exactitud de las predicciones sea muy distinta en los subconjuntos de entrenamiento y test, hay varias opciones (no excluyentes) para salvar dicha circunstancia:

⁷Subajuste o *underfitting* indica la imposibilidad de identificar o de obtener resultados correctos debido a un insuficiente tamaño de muestra en el conjunto de entrenamiento, o un entrenamiento muy pobre.

mejorar el modelo, ajustar sus hiperparámetros, incluir más casos en el conjunto de datos, modificar el preprocessado de los datos, ver si hay desequilibrio entre las clases, analizar si las variables predictoras son o no las adecuadas, revisar el proceso de limpieza de datos... y, posteriormente, entrenar de nuevo el modelo y determinar si es o no válido. Otra opción más drástica es, simplemente, cambiar de modelo.

Una mejor opción sería utilizar desde el principio un enfoque de validación, que implica dividir el subconjunto de entrenamiento en dos partes: un subconjunto de entrenamiento propiamente dicho y un conjunto de **validación**. Así, se puede entrenar el modelo en el nuevo subconjunto de entrenamiento y estimar su exactitud en el conjunto de validación. Es importante tener claro que el subconjunto de validación no es un subconjunto que se deje aparte, como el de test, durante la fase de entrenamiento, sino que se utiliza en dicha fase.

En resumen, con el enfoque de validación, para dar por válido un modelo, se procede como sigue:

- Dividir el conjunto de datos en subconjunto de entrenamiento y subconjunto de test.
- Dividir el subconjunto de entrenamiento en un subconjunto de entrenamiento propiamente dicho y un subconjunto de validación.
- Entrenar el modelo con los datos del subconjunto de entrenamiento propiamente dicho.
- Comprobar que la exactitud de las predicciones en dicho subconjunto de entrenamiento y en el de validación es similar (y aceptable para la exigencia que se requiere).
- Realizar predicciones con el conjunto de test y comprobar que se obtiene un porcentaje de buenas predicciones aceptable para los requisitos exigidos.
- Agregar el conjunto de test al de entrenamiento (global) y entrenar de nuevo el modelo (que será el definitivo); de esta manera se aprovecha el 100 % de los datos. Este último entrenamiento debería mejorar el modelo final, aunque la única manera de comprobarlo es mediante su comportamiento en el entorno real.

La limitación del enfoque de validación con un solo subconjunto de reserva (de validación) es que dicha validación puede ser muy variable y poco confiable, a menos que se esté trabajando con conjuntos de datos muy grandes (Molinaro et al., 2005). Y aquí es donde entran en juego los procedimientos de validación que utilizan remuestreo. El procedimiento de validación con remuestreo más utilizado es la validación cruzada (VC) k -grupos (*k-fold cross validation*). También es muy popular el que utiliza remuestreo por *bootstrapping*, que se abordará tras el VC k -grupos.

Para llevar a cabo una VC k -grupos, se divide aleatoriamente el subconjunto de datos de entrenamiento en k grupos (*folds*) de aproximadamente el mismo tamaño. El modelo se ajusta en los $k-1$ primeros grupos y el último se usa como conjunto de validación, para “validar” la bondad del modelo. A continuación, se separa el penúltimo grupo y se ajusta el modelo con los restantes, usándose el penúltimo grupo como subconjunto de validación para validar la bondad del modelo. Después se separa el antepenúltimo grupo, y así sucesivamente, hasta separar el primero. Como resultado, se obtienen k conjuntos de errores, cuyo promedio (véase Sec. @ref(evaluación)) podría servir como estimación de la exactitud y precisión (o error⁸) esperada en un conjunto de datos nuevo.

⁸Boehmke and Greenwell (2019) lo denominan error de generalización puesto que tiene lugar al “generalizar” el modelo entrenado a nuevos conjuntos de datos.

10.4. El enfoque de validación

155

El procedimiento descrito puede repetirse varias veces (VC con repetición), mediante nuevas particiones aleatorias del conjunto de entrenamiento y procediendo igual que en la iteración anterior.

En la práctica, normalmente se usa $k = 5$ o $k = 10$ (las Fig. 10.1 y 10.2 ilustran el caso de CV 5-grupos). No existe una regla formal en cuanto al tamaño de k , pero a medida que k aumenta, la diferencia entre el rendimiento estimado y real precisión estimada y el real que se obtendrá en el conjunto de test, disminuirá. En el lado negativo de la balanza, un k demasiado grande puede aumentar notablemente la carga computacional y, además, no generar mejoras significativas. A este respecto, en Molinaro et al. (2005) se concluye que CV con $k = 10$ funciona de manera similar a CV con $k = n$, la CV más extrema, también conocida como VC “dejando uno fuera” (*leave one out cross validation, LOOCV*).



Figura 10.1: CV 5-grupos (i)

y si la exactitud de las predicciones en el subconjunto de entrenamiento propiamente dicho y en el de validación es similar (y aceptable para la exigencia que se requiere):

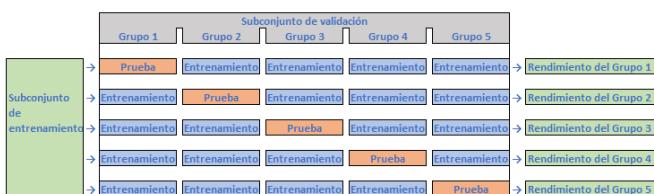


Figura 10.2: CV 5-grupos (ii)

Aunque $k \geq 10$ contribuye a minimizar la variabilidad del error de predicción (es decir, tiende a aumentar la precisión de las predicciones), en general la CV k -grupos suele proporcionar mayores variabilidades que el *bootstrapping* (que se analiza a continuación); no ocurre lo mismo con el sesgo (Boehmke and Greenwell, 2019). Kim (2009) demostró que repetir el CV k -grupos puede ayudar a reducir la estimación del error de generalización.

Una implementación del CV k -grupos, con tres repeticiones, utilizando el conjunto de datos `Madrid_Sale_num_sample_bin`, previa mejora de la precisión del modelo mediante *downsampling*, se llevó a cabo en la Sec. 10.3 con la siguiente orden:

```
control <- trainControl(method = "repeatedcv", number = 10, repeats = 3)
```

Bootstrapping es un procedimiento de muestreo aleatorio con reemplazamiento (Efron and

Tibshirani, 1986). Esto significa que, después de seleccionar un dato para incluirlo en el subconjunto que sea, sigue disponible para una selección posterior. Una muestra *bootstrap* tiene el mismo tamaño que el conjunto de datos original a partir del cual se obtiene. Las observaciones originales seleccionadas (una o varias veces) en la muestra conforman el subconjunto de entrenamiento, mientras que aquellas que no aparecen en ella (se les denomina *out-of-bag*) conforman el subconjunto de test.

La Fig. 10.3, tomada de Boehmke and Greenwell (2019), muestra un esquema de muestreo *bootstrap*, donde cada muestra contiene 12 observaciones, al igual que en el conjunto de datos original. Como puede observarse, el muestreo *bootstrap* lleva aproximadamente a la misma distribución de valores (representados por colores) que el conjunto de datos original.

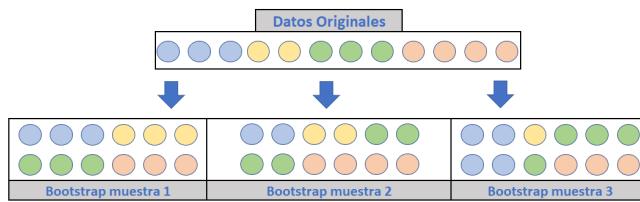


Figura 10.3: Remuestreo bootstrap

El hecho de que *bootstrapping* replique el conjunto de observaciones implica, como se dijo anteriormente, que la variabilidad del error es menor que en CV k -grupos. Sin embargo, dicha replicación puede aumentar el sesgo de la estimación dicho error. Esto puede ser un problema con conjuntos de datos muy pequeños, pero no para la mayoría de los conjuntos de datos, que suelen ser de tamaño medio o grande (por ejemplo, $n \geq 1000$).

Las muestras *bootstrap* pueden crearse fácilmente con `rsample::bootstraps()`, como se ilustra a continuación.

```
bootstraps(Madrid_Sale_num_sample_bin, times = 10)
```

Si se usa la función `trainControl()`, se debe especificar: `method = "boot"`.

10.5. Compensación (*trade off*) entre sesgo y varianza

En el entorno predictivo, el objetivo es que el error de predicción, en términos generales, o por término medio, sea lo más pequeño posible. Sin embargo, no se pueden promediar los errores porque se compensarían los positivos con los negativos. Por ello, se promedian elevados al cuadrado (considerando sólo su magnitud), denominándose dicho promedio “error cuadrático medio”, ECM (en este caso, de predicción): $E(y_j - \hat{y}_j)^2$ en términos probabilísticos, o $\sum_{j=1}^N (y_j - \hat{y}_j)^2 \frac{n_j}{N}$ en términos descriptivos. Pues bien, el ECM se puede descomponer como suma de dos componentes:

10.6. Ajuste de hiperparámetros

157

- Uno debido a la diferencia entre el valor correcto de la variable objetivo o respuesta y el que se espera que proporcione el modelo. Dicha diferencia se denomina sesgo (en inglés, *bias*), y aparece elevado al cuadrado en dicha descomposición.
- Otro debido a que, dado un conjunto de valores de las variables predictoras, la respuesta del modelo no es siempre la misma. Esta variabilidad aparece en la descomposición en forma de varianza, y por ello se denomina varianza del error de predicción o, simplemente, varianza de predicción.

Lógicamente, el incremento/reducción de uno de los componentes implica la reducción/incremento del otro.

Un sesgo muy elevado es un indicador de que el modelo es muy simple y no ha ajustado bien los datos de entrenamiento (*underfitting*), lo cual se traduce en errores de predicción elevados. Una varianza de predicción elevada (es decir, pequeños cambios en los datos de entrada producen salidas muy distintas), es un signo de complejidad en el modelo y sobreajuste (*overfitting*) de los datos de entrenamiento.

Los algoritmos con tendencia a un elevado porcentaje del ECM debido al sesgo, tienen, lógicamente, un porcentaje del ECM debido a la varianza de predicción pequeño; es decir, tienen los problemas que se derivan del infra-ajuste de los datos: malas predicciones (sesgadas) y, encima, muy precisas. Y al contrario, aquellos que tienen un porcentaje del ECM debido al sesgo pequeño, tienen un porcentaje del ECM por varianza de predicción elevado.

Los modelos lineales (regresión lineal, análisis discriminante lineal, regresión logística...) suelen tener errores por sesgo elevados.⁹ Modelos como los árboles de decisión, el *k*-vecinos más cercanos y los *support vector machine* tienen errores por sesgo pequeños (y, por tanto, varianza de predicción grande, siempre en relación al ECM total), son muy adaptables y ofrecen una flexibilidad extrema en cuanto a los patrones a los que pueden ajustarse. Sin embargo, plantean sus propios problemas, especialmente el de sobreajuste de los datos de entrenamiento, cuya consecuencia es que el modelo no se generalizará bien con datos nuevos.¹⁰

Lógicamente el modelo predictivo o clasificador deseado es el que tenga el menor ECM posible.¹¹ Si este no fuera el caso, habría que encontrar la combinación sesgo cuadrático-varianza de predicción que minimizase el ECM. La Fig. 10.4, se ha generado a partir de datos sintéticos de sesgo (al cuadrado) y varianza de predicción. En ella se puede apreciar que el valor mínimo del ECM, que es la suma de los valores del sesgo cuadrático y varianza de predicción determinados por la linea vertical amarilla.

10.6. Ajuste de hiperparámetros

Los denominados “hiperparámetros” de un modelo son los valores de las configuraciones utilizadas durante el proceso de entrenamiento. A diferencia de los parámetros, son valores que no

⁹Sin embargo, rara vez se ven afectados por el error introducido en el remuestreo.

¹⁰Por ello, el remuestreo es clave para reducir este riesgo.

¹¹Sin embargo, hay una parte del ECM que no se puede reducir: aquél que se debe a la aleatoriedad, o a la no inclusión en el modelo de variables relevantes, entre otras.

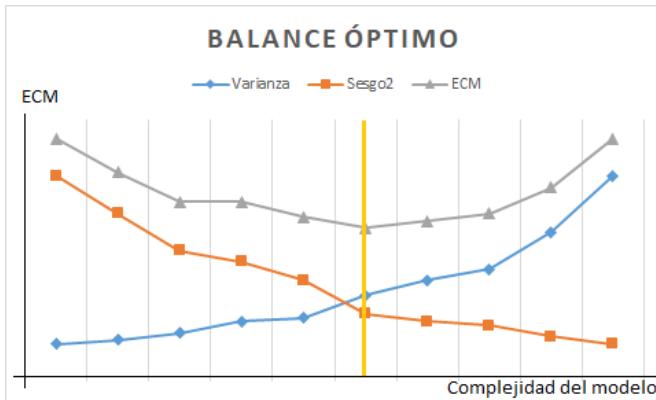


Figura 10.4: Trade-off entre sesgo y varianza

se obtienen a partir de los datos, sino que los propone el científico de datos. Podría decirse que son conjeturas (buenas conjeturas) realizadas sin utilizar las observaciones disponibles.

Los hiperparámetros, a diferencia de los parámetros, se fijan antes del entrenamiento. Siendo más específicos, al entrenar un modelo de aprendizaje automático se fijan los valores de los hiperparámetros para que con estos se estimen los parámetros. Podría decirse que son los ajustes del modelo para que éste pueda resolver de manera óptima el problema de aprendizaje automático.

Algunos ejemplos de hiperparámetros utilizados para entrenar los modelos son la ratio de aprendizaje en el algoritmo del descenso del gradiente, el número de vecinos en el algoritmo de k -vecinos más cercanos, la profundidad máxima en un árbol de decisión, el número de árboles en un bosque aleatorio (*random forest*)... Como puede apreciarse, sirven para controlar la complejidad de los algoritmos de aprendizaje automático y, por tanto, la compensación entre sesgo y varianza.

En conclusión, hiperparámetros y parámetros son conceptos bien diferentes.

A la luz de la definición de hiperparámetro, lo natural sería que el científico de datos los fijase de acuerdo con su experiencia en el pasado en problemas similares, asignándoles los mismos, o parecidos, valores. Sin embargo, existen métodos más sofisticados para resolver el problema de la “optimización de hiperparámetros”, es decir, de la obtención del conjunto óptimo de valores de los mismos que proporciona la configuración que, tras el entrenamiento, dará lugar a los mejores resultados. Entre estos procedimientos cabe destacar los siguientes por ser los que incorporan los algoritmos más populares:

- Optimización bayesiana: utiliza la moda para elegir qué hiperparámetros considerar, en función del rendimiento de las elecciones anteriores.
- Búsqueda en cuadrícula (*grid search*): prueba con todas las combinaciones posibles de la cuadrícula.

10.7. Evaluación de modelos

159

- Búsqueda aleatoria: muestrea y evalúa aleatoriamente conjuntos de una distribución de probabilidad específica.
- Optimización secuencial basada en modelos: son una formalización de la optimización bayesiana.

A modo de ejemplo, los dos hiperparámetros que más influencia tienen en un modelo de *random forest* (véase Cap. ?), son `mtry` (número de variables muestreadas aleatoriamente como candidatas en cada *split*) y `ntree` (número de árboles). Pues bien, a continuación, como viene siendo habitual, se utiliza el conjunto de datos `Madrid_Sale_num_sample_bin` para buscar el valor óptimo de `mtry` mediante la técnica de búsqueda en cuadrícula (se usa la métrica `Accuracy` (exactitud), que hace referencia a la proporción de predicciones correctas, véase Sec. @ref(evaluación)).

```
control <- trainControl(method = "repeatedcv", number = 10, repeats = 3)
seed <- 7
metrica <- "Accuracy"
set.seed(seed)
mtry <- sqrt(ncol(Madrid_Sale_num_sample_bin))
tunegrid <- expand.grid(.mtry = mtry)
```

Nota

La implementación del algoritmo de *random forest* del paquete `randomForest` proporciona la función `tuneRF()`, que busca valores `mtry` óptimos dados los datos disponibles.

Se entrena el modelo y se observa que la mayor exactitud, 68,76 %, se obtiene con un `ntree` de 2.449489.

```
rf_default <- train(price_bin ~ ., data = Madrid_Sale_num_sample_bin, method = "rf",
                     metric = metrica, tuneGrid = tunegrid, trControl = control)
print(rf_default)
# Accuracy
# 0.6876006
```

10.7. Evaluación de modelos

La última fase del proceso de modelización contesta a la pregunta: ¿Cómo de bueno es el modelo entrenado? ¿Cómo de bien generaliza los (buenos) resultados obtenidos en la fase de entrenamiento a nuevos conjuntos de datos (datos *out-of-sample*)? Por ello, en este epígrafe se presentan las métricas más populares de rendimiento de modelos en los entornos de regresión (predicción) y clasificación.

En el entorno de regresión, es prácticamente imposible predecir valores exactos, y hay que conformarse con muy buenas aproximaciones a dichos valores exactos, por lo que las métricas

que se utilizan para medir la bondad del modelo entrenado suelen estar basadas en la diferencia entre los valores reales y los que predice el modelo, es decir, en los errores de predicción. La medida natural sería la media de los errores de predicción, pero, para evitar la compensación de los errores positivos con los negativos, y puesto que el interés se centra en la magnitud de los errores y no en su signo, las métricas de evaluación más populares en el entorno de regresión son:

- **Error cuadrático medio (ECM)**: es la más utilizada en tareas de regresión. Como se avanzó en 10.5, se define como la media de las diferencias cuadráticas de entre los valores objetivo (y_j) y los predichos por el modelo (\hat{y}_i), evitando así la compensación de errores positivos y negativos. Su expresión es, por tanto, $ECM = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$. Al considerar los errores de predicción al cuadrado, exacerba los errores grandes, por lo que hay que utilizar esta métrica con cuidado cuando se tengan valores anómalos en el conjunto de datos y no hayan sido tratados en las fases previas a la modelización y validación. En ese caso, incluso un modelo cuasiperfecto podría tener un ECM elevado. Si todos los errores son inferiores a la unidad hay un riesgo elevado de subestimar lo malo que es el modelo (en caso de que lo fuese).
- **Error absoluto medio (EAM)**: es una alternativa al ECM que se define como la media del valor absoluto de los errores de predicción (para evitar compensaciones): $EAM = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$. Al no elevar al cuadrado, no magnifica los errores grandes, por lo que es menos sensible que el ECM a valores anómalos, si bien tampoco es recomendable en caso de que éstos no hayan sido tratados previamente. Una ventaja que tiene es que su unidad de medida es la misma que la de la variable objetivo.
- **Raíz cuadrada del error cuadrático medio (RECM)**: “deshace”, no en un sentido matemático, sino aproximadamente, la elevación al cuadrado de los errores en el ECM y, por consiguiente, viene dada en las mismas unidades que la salida del modelo, lo que la hace más interpretable. Su expresión es: $RECM = +\sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$.
- **Coeficiente de determinación (R^2)**: se define como $R^2 = 1 - \frac{\sum_{i=1}^n (y_j - \hat{y}_j)^2}{\sum_{i=1}^n (y_j - \bar{y})^2}$ (véanse detalles adicionales en el Cap. 5 de Montero (2007)) y su campo de variación es $[-\infty, 1]$. En la práctica totalidad de las situaciones reales toma valores entre 0 y 1, puesto que sólo toma valores negativos cuando el modelo entrenado sea muy deficiente y prediga peor que cuando se establece como predicción la media de las salidas observadas, sean cuales sean los valores de los predictores. Obviando estas situaciones, y aquéllas en las que la varianza de la variable salida no se pueda descomponer en varianza debida al modelo y varianza debida al error (por ejemplo, en el caso de una regresión potencial)¹², R^2 se puede interpretar como la reducción proporcional en el ECM que tiene lugar al predecir las salidas del modelo mediante los predictores (cualquiera que sea la función que los ligue con la salida) en vez de mediante la media de la variable output (que es la predicción óptima en ausencia de predictores). Mide, por tanto, lo bueno que es disponer de predictores para predecir los valores de la variable output o de salida; o, en otros términos, el porcentaje de varianza de la variable salida que explican los predictores a través del modelo que liga a ambos. Cuanto más cercano esté a la unidad mejor es el modelo a efectos predictivos.

¹²En estos casos es mejor utilizar otra métrica, como el ECM.

10.7. Evaluación de modelos

161

- **Coeficiente de determinación ajustado (R_{adj}^2)**: una limitación importante del R^2 es que su valor puede aumentarse artificialmente mediante la inclusión de más y más variables predictoras, pues la inclusión de las mismas o mantiene o mejora dicha métrica. Esta circunstancia puede dar lugar a confusión, pues el hecho de que un modelo utilice más variables predictoras que otro, no quiere decir que sea mejor. El R_{adj}^2 corrige dicha circunstancia penalizando la complejidad del modelo, entendiéndose que un modelo es más complejo que otro si utiliza un mayor número de variables predictoras que ese otro. Su expresión viene dada por $R_{adj}^2 = 1 - \left(\frac{n-1}{n-p-1} \right) (1 - R^2)$, y su valor nunca supera el del R^2 .
- **Deviance**: es una métrica relacionada con la estimación de modelos (especialmente modelos lineales generalizados) por el método de la máxima verosimilitud. Compara, por cociente, la verosimilitud del modelo estimado con la del modelo saturado (aquel que tiene tantos parámetros como observaciones¹³ y que, por tanto, tiene la máxima verosimilitud alcanzable). Mide el grado en el que un modelo explica la variabilidad en un conjunto de datos cuando se utiliza la estimación de máxima verosimilitud. En términos de log-verosimilitud (l) se define como $D = 2(l_{\text{Modelo saturado}} - l_{\text{Modelo propuesto}})$, y, lógicamente, cuanto menor es la deviance mejor es el modelo.
- **Raíz del error logarítmico cuadrático medio (RELCM)**: similar a RMSE, pero tomando logaritmos en los valores reales y predichos. De especial interés cuando lo que importa es la magnitud relativa (porcentual) de los errores. No se puede utilizar cuando la variable objetivo toma valores negativos. Para salvar la problemática de que la variable objetivo tome el valor cero, generalmente se agrega una constante a los valores reales y predichos de la variable salida antes de aplicar la operación logarítmica. Dependiendo del problema, se puede elegir otro tipo de constante. Su expresión viene dada por: $RMSLE = \sqrt{\frac{1}{n} (\log(y_i + 1) - \log(\hat{y}_i + 1))^2}$.

En este manual se usan estas medidas en repetidas ocasiones. Por ejemplo, en el Cap. 19 se ajusta la regresión *ridge* en el subconjunto de entrenamiento y se evalúa su ECM en el subconjunto de test.

```
ridge_pred <- predict(ridge.mod, s = 1e10, newx = x[test, ])
mean((ridge_pred - y.test)^2)
```

En el entorno clasificadorio, las salidas del modelo pueden ser de clase (tal es el caso de los algoritmos de máquinas de vectores soporte y k -vecinos más cercanos, por ejemplo) o de probabilidad (caso de la regresión logística, los bosques aleatorios, el adaboost...). Dado que pasar de salidas probabilísticas a salidas de clase consiste únicamente en fijar umbrales de probabilidad, y que algunos algoritmos ya proporcionan el paso de salidas de clase a salidas probabilísticas, en lo que sigue no se hará distinción entre ellas.

En dicho entorno clasificadorio, es muy frecuente el uso de la **matriz de confusión**, que compara las clases (niveles categóricos) reales con las predichas en el subconjunto de test (cuyos

¹³Por tanto, pasa por todas ellas.

resultados se conocen). La Fig. 10.5 muestra un ejemplo de clasificación multiclas (en concreto, 3 clases) basado en el famoso conjunto de datos “Flor iris” de Fisher, que considera tres especies (iris setosa, iris virginica e iris versicolor). La predicción de la clase a la que pertenece una flor se hace en función del largo y el ancho del sépalo y del pétalo.

En la diagonal ascendente figura el número de flores, de cada color, cuya clase ha sido correctamente predicha. Los elementos fuera de dicha diagonal indican las flores, de cada clase, que el clasificador utilizado ha clasificado erróneamente. Como puede apreciarse, 47 de las 50 flores iris que se consideran fueron bien clasificadas. Sin embargo, dicho clasificador clasificó una flor versicolor como virgínica, y dos virgínicas como versicolores.

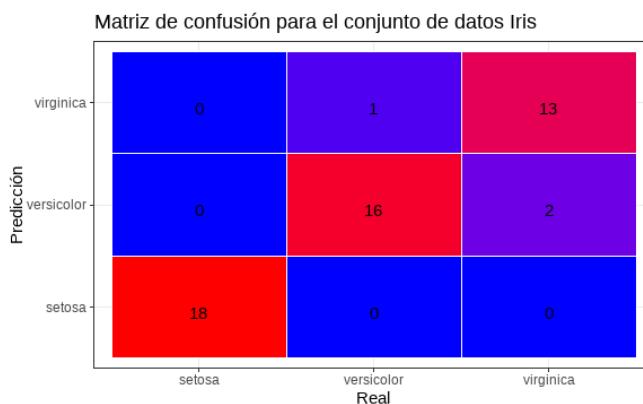


Figura 10.5: Matriz de confusión con tres clases

Aunque el concepto de matriz de confusión es muy sencillo, la terminología que lo rodea no lo es tanto; incluso podría decirse que es confusa. La predicción proporcionada por el modelo puede ser (véase Fig. 10.6):

- Un verdadero positivo (VP): predicción de verdadero y verdadero en realidad.
- Un verdadero negativo (VN): predicción de falso y falso en realidad.
- Un falso positivo (FP): predicción de verdadero y falso en la realidad.
- Un falso negativo (FN): predicción de falso y verdadero en la realidad.

		Predicción	
		Positivo	Negativo
Realidad	Positivo	VP (Verdaderos Positivos)	FN (Falsos Negativos)
	Negativo	FP (Falsos Positivos)	VN (Verdaderos Negativos)

Figura 10.6: Terminología de una matriz de confusión

Como se avanzó anteriormente, esta terminología es confusa y, por ello, se ilustra a continuación con un ejemplo. Supóngase que se está interesado en conocer si un determinado tratamiento

10.7. Evaluación de modelos

163

médico tiene efectos positivos sobre una enfermedad. Echando mano de la teoría de la contrasentación de hipótesis (véase Sec. 13.5), supóngase que se toma como hipótesis nula (H_0): SÍ y como hipótesis alternativa (H_1): NO. Pues bien:

- Si es cierto que el tratamiento en cuestión tiene un efecto positivo en la enfermedad y el modelo no rechaza la hipótesis nula, se tiene un VP.
- Si la hipótesis nula es falsa, es decir, si el tratamiento no tiene un efecto positivo sobre la enfermedad, y el modelo rechaza la hipótesis nula, entonces se tiene un VN.
- Si la hipótesis nula es falsa y el modelo concluye que no se rechaza la hipótesis nula de que el tratamiento cura la enfermedad, entonces se tiene un FP.
- Si es cierto que el tratamiento tiene un efecto positivo en la enfermedad y el modelo rechaza la hipótesis nula, se tiene un FN.

Las siguientes medidas se pueden calcular a partir de una matriz de confusión (es decir, a partir del número de VPs, VNs, FPs y FNs) para un clasificador binario:

- **Exactitud.** Es la proporción de predicciones correctas: $Exactitud = \frac{VP+VN}{Total}$. Responde a la pregunta: ¿Con qué frecuencia funciona correctamente el clasificador? Lógicamente, la tasa de clasificación errónea se obtiene como $\frac{FP+FN}{Total}$. En caso de desequilibrio notable de clase, por ejemplo, la clase A contiene 999 casos y la B tan sólo 1, siendo B la clase rara, la clase positiva, la precisión no sería una métrica fiable para evaluar el rendimiento clasificatorio del modelo. Por ejemplo, el clasificador podría consistir en una regla que dijese que todos los casos pertenecen a la clase A, ¡y acertaría en el 99,9% de los casos! En estos casos, sería mejor utilizar como métricas la precisión y la sensibilidad.
- **Precisión.** Da respuesta a la pregunta: cuando el clasificador predice “SÍ”, ¿con qué frecuencia predice correctamente? Su expresión viene dada por: $Precisión = \frac{VP}{VP+FP}$.
- **Sensibilidad**¹⁴. Restringe el denominador de la precisión a los SÍ reales. Responde a la pregunta: ¿cuándo en realidad es un SÍ, cuál es el porcentaje de aciertos del clasificador? Su expresión es $Sensibilidad = \frac{VP}{VP+FN}$. Por consiguiente, la sensibilidad es una medida de la probabilidad de que un caso real positivo se clasifique como positivo.
- **Especificidad.** Se define como $Especificidad = \frac{VN}{FP+VN}$. Es, por tanto, el porcentaje de verdaderos negativos respecto de todo lo que debería haber sido clasificado como negativo. Su complementario, la 1-especificidad ($\frac{FP}{FP+VN}$) es, básicamente, una medida de la frecuencia (relativa) con la que se producirá una falsa alarma, o la frecuencia con la que un caso real negativo se clasifique como positivo.
- **Puntuación F1.** Es la media armónica de la precisión y la sensibilidad. Su campo de variación es [0, 1], donde 0 indica falta total de precisión y sensibilidad y 1 significa precisión y sensibilidad perfectas. La puntuación F1 penaliza los valores extremos y se suele utilizar en caso de conjuntos de datos desequilibrados.

¹⁴A veces denominada “exhaustividad”.

- **Área bajo la curva de características operativas del receptor (área bajo la curva ROC).** Al graficar la sensibilidad (tasa de verdaderos positivos) frente a la tasa de falsos positivos (también denominada 1-especificidad), se obtiene la curva ROC. La diagonal ascendente representa la aleatoriedad. Cuanto más grande sea el área bajo la curva ROC, mejor será la precisión obtenida. Es una medida recomendable en el caso de clases desequilibradas. Un ejemplo de área bajo la curva ROC puede verse en el Cap. 16, y se reproduce en la parte derecha de la Fig. 10.7.

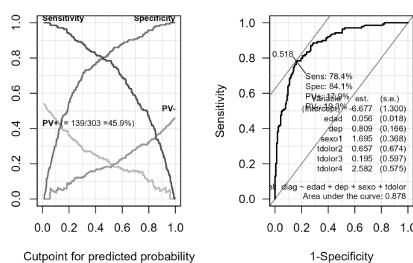


Figura 10.7: Ejemplo de curva ROC a partir de la estimación de un modelo lineal generalizado (parte derecha)

- **Índice de Gini.** Bien conocido en la literatura estadística sobre concentración, se trata de un indicador útil en el caso de clases desequilibradas. Su campo de variación es $[0,1]$, donde 0 representa la igualdad perfecta y 1 la concentración en una única clase. Puede calcularse a partir del área bajo la curva ROC de la siguiente manera: $IG = 2 \text{ Área bajo la curva ROC} - 1$. En caso de $IG=0$, el área bajo la curva ROC es $1/2$ y la curva ROC coincide con la diagonal ascendente. En caso de $IG=1$, el área bajo la curva ROC será 0 y dicha curva vale cero para todos los valores del eje de abscisas, excepto para el último, para el cual vale 1. En caso de que las observaciones no se vayan acumulando de una en una para la configuración de la curva bajo ROC, o para el cálculo directo del índice de Gini, es mejor utilizar una versión del mismo denominada “índice E” (IE, véase Cap. 3 de [Montero \(2007\)](#)), pues el índice de Gini tan sólo proporcionará una aproximación de la concentración existente; buena aproximación, pero aproximación.

- **Índice de Jaccard.** Mide las similitudes entre el conjunto clases reales y predichas por el clasificador. Se define como el cociente entre el número de coincidencias en los conjuntos real y predicho (clases predichas correctamente) y el tamaño de la unión de los dos conjuntos (el doble del número de clases a etiquetar menos el número de clases predichas correctamente): $I_{Jaccard} = \frac{VP+VN}{2Total-(Fp+FN)}$.

Otras medidas de interés son la pérdida logarítmica, el coeficiente de correlación de Matthews, el gráfico de Kolmogorov-Smirnov y el gráfico de ganancia y elevación. Éstas, entre otras, pueden verse en [Dembla \(2020\)](#), [Hernández-Orallo et al. \(2011\)](#) y [Vujović et al. \(2021\)](#), entre otros.

Con tantas métricas, ¿cómo decidir entre ellas? Será el conocimiento de la problemática que se esté estudiando (el negocio) el que normalmente guie la elección de la métrica: si se trata de

un problema de predicción de la producción de un determinado bien cuyo coste de almacenaje es elevado, lo más prudente sería utilizar el ECM para penalizar la sobreproducción; si lo que interesa son valores altos de la precisión y la sensibilidad (tal es el caso en numerosas situaciones del ámbito sanitario), la mejor medida sería la puntuación F1. No obstante, se habrá de tener siempre en cuenta que la exactitud, la precisión, la sensibilidad, la especificidad y el índice de Jaccard son buenas formas de evaluar clasificadores cuando las clases están equilibradas. En caso contrario, el área bajo la curva ROC y el IG (o el IE) son dos buenas alternativas.

Resumen

En este capítulo se describen una serie de herramientas necesarias para desarrollar proyectos de ciencia de datos. Son herramientas que se utilizan se utilizan pre- o post-modelado de los datos y que aumentan significativamente el rendimiento de los modelos. Tal caja de herramientas incluye el particionado del conjunto de datos, el manejo de datos no equilibrados, los métodos de remuestreo, el equilibrio entre sesgo y varianza, el ajuste de hiperparámetros y la evaluación de modelos, entre otras. Se hace especial hincapié en las métricas para evaluar modelos, tanto en el entorno de regresión como en el de clasificación, y, en ambos casos, se ofrece un amplio abanico de ellas. A efectos ilustrativos, se utiliza, fundamentalmente, el paquete `caret` y conjuntos de datos derivados del dataset `Madrid_Sale`.

Capítulo 11

Análisis exploratorio de datos

Emilio L. Cano

Universidad Rey Juan Carlos

11.1. Introducción

El análisis exploratorio de datos (AED), y en particular su visualización, es el primer análisis que se debe hacer sobre cualquier conjunto de datos. El AED se realiza mediante dos herramientas: los resúmenes numéricos y las visualizaciones gráficas. La “historia” que nos esté contando el gráfico de los datos nos guiará hacia las técnicas de aprendizaje estadístico más adecuadas. Incluso, en muchas ocasiones será suficiente el AED para tomar una decisión sobre el problema en estudio.

11.1.1. El cuarteto de Anscombe

Un ejemplo clásico de la importancia del **AED** y, concretamente, de las representaciones gráficas es el “cuarteto de Anscombe” ([Anscombe, 1973](#)), el cual está compuesto por 11 filas de 8 variables numéricas que conforman 4 conjuntos de datos (disponibles en el objeto `anscombe`), con los mismos resúmenes estadísticos pero con propiedades muy distintas, lo que se ve fácilmente cuando se representan en forma gráfica. Si se calcula, por ejemplo, la media y la desviación típica de cada variable, se observa que son prácticamente iguales. Incluso los coeficientes de correlación de cada X con su Y son también prácticamente idénticos.

```
library("dplyr")
anscombe |> summarise(across(.fns = mean))
#> #> x1 x2 x3 x4      y1      y2      y3      y4
#> #> 9  9  9  9    7.5009  7.5009  7.5    7.5009
anscombe |> summarise(across(.fns = sd))
```

```
#>   x1    x2    x3    x4    y1    y2    y3    y4
#> 3.316 3.316 3.316 3.316 2.031 2.031 2.030 2.030
```

Sin embargo, la Fig. 11.1 muestra que, a pesar de tener medias y desviaciones típicas prácticamente iguales, los datos son muy diferentes.

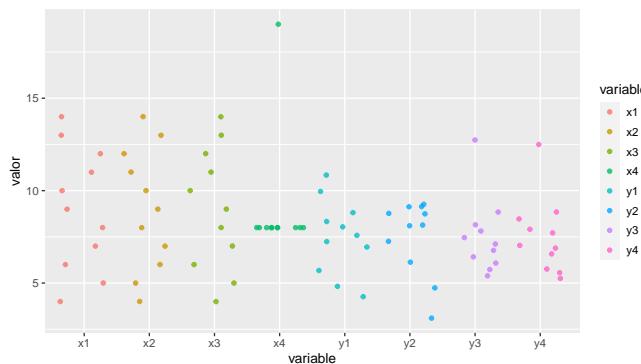


Figura 11.1: Representación de las variables del cuarteto de Anscombe

Si en el análisis por separado ya se ve la necesidad de hacer un gráfico, ésta es más evidente cuando se analizan las variables conjuntamente. La Fig. 11.2 muestra los cuatro gráficos que constituyen “el cuarteto de Anscombe” y que se puede obtener de la propia ayuda del conjunto de datos (`example(anscombe)`). La línea de regresión que se ajusta es prácticamente la misma, y los coeficientes de correlación entre las variables X e Y de los cuatro gráficos, idénticos: 0.8163. Es evidente que la relación entre las variables es muy distinta en cada uno de los casos, y si no se visualizan los datos para elegir el mejor modelo de regresión y después interpretarlo, se pueden tomar decisiones erróneas. El cuarteto de Anscombe es muy ilustrativo, al igual que *The Datasaurus Dozen* (Matejka and Fitzmaurice, 2017) en <https://www.autodeskresearch.com/publications/samestats>.

11.1.2. Conceptos generales

Muy brevemente, se presentan una serie de conceptos esenciales para la mejor comprensión de este manual¹. Los datos que se analizan, provienen de una determinada **población**, y no son más que una **muestra**, es decir, un subconjunto de toda la población. La **Estadística Descriptiva** se ocupa del AED en sentido amplio, que se aplica sobre los datos concretos de la muestra. La **Inferencia Estadística** (véase Cap. 13) hace referencia a los métodos mediante los cuales, a través de los datos muestrales, se toman decisiones, se analizan relaciones o se hacen predicciones sobre la población. Para ello, se hace uso de la **Probabilidad** aplicando el modelo adecuado (véase Cap. 12). Además, es muy importante considerar el método de obtención de la muestra (véase Cap. 14) que, en términos generales, debe ser representativa de la población

¹Para un análisis extenso de los conceptos aquí expuestos puede consultarse, por ejemplo, en Montero Lorenzo (2007).

11.1. Introducción

169

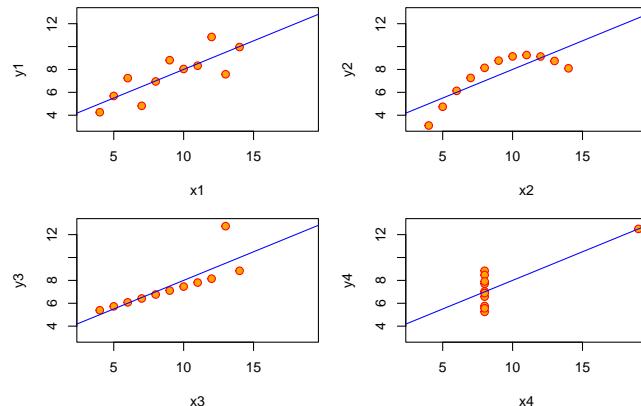


Figura 11.2: Los cuatro gráficos que constituyen el cuarteto de Anscombe junto con un ajuste lineal

para que las conclusiones sean válidas. La Fig. 11.3 representa la esencia de la Estadística y sus métodos.

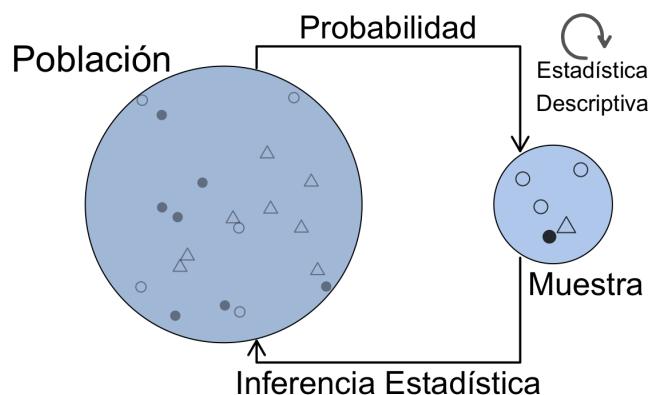


Figura 11.3: La esencia de los métodos estadísticos

Las características a observar en los elementos de una población pueden dar lugar a diferentes tipos de datos o variables. El análisis a realizar dependerá del tipo de variable, que puede ser:

1. **Cuantitativa** (se puede medir o contar). Se denomina **variable cuantitativa** a cualquier característica observable que pueda expresarse en valores numéricos. Se clasifican como **variables discretas** (se puede *contar* el número de valores que toma) y **continuas** (pueden tomar cualquier valor en un intervalo dado).
2. **Cualitativa** (no se puede expresar como un número). Se denomina **variable cualitativa, atributo o factor** a cualquier característica observable que indica una *cualidad* o *atributo*.

Éstas pueden tener varios niveles (polítómicas) o solo dos (dicotómicas). Si en una variable categórica se pueden *ordenar* las categorías, entonces se denomina **variables ordinal**.

11.1.3. Componentes de un gráfico y su representación en R

De los diferentes sistemas que tiene **R** para representar gráficos (los “base”, paquete **graphics**, y los “grid”, paquete **lattice** ([Sarkar, 2008](#))), este capítulo se centra en el paquete **ggplot2** ([Wickham, 2016](#)), que forma parte del *tidyverse*, por su amplio uso y popularidad.

El flujo de trabajo con **ggplot2** se puede resumir en los siguientes pasos:

1. Proporcionar una **tabla de datos** a la función **ggplot**. Es el primer argumento (**data**) y se puede utilizar el operador *pipe*.
2. Proporcionar las **columnas** de la tabla de datos que serán representadas en el gráfico. Este será el segundo argumento (**mapping**) de la función **ggplot**, y se especifica con la función **aes** (*aesthetics*) como una lista de pares *aesthetic = variable*, de forma que el elemento especificado como *aesthetic* será “mapeado” a los valores de la variable. Esta especificación se puede hacer también en las funciones que añaden capas, que se explican a continuación. Los *aesthetics* más comunes (para muchos tipos de gráficos obligatorios) son **x** e **y**, es decir, las columnas que se usarán para el eje horizontal y el eje vertical respectivamente. Además, se pueden especificar columnas para el color, el tamaño, el símbolo de los puntos, el tipo de línea, el texto, y otros específicos del tipo de gráfico. Los *aesthetics* se pueden especificar también de forma “fija” (sin depender de ninguna variable) fuera de la función **aes**.
3. Añadir las **capas** del gráfico con los *geoms*, es decir, los objetos geométricos que representan a cada variable. Esto se indica con el operador **+**, como si se “sumasen” componentes al gráfico mediante funciones **geom_xxx**.
4. Añadir otras capas al gráfico: por ejemplo, una capa de etiquetas del gráfico (función **labs**); de ejes, para modificar los ejes y leyendas creados por defecto (funciones **scale_*_xxx**); de estadísticos, para crear nuevas variables a representar basadas en los datos (funciones **stat_xxx**).
5. Añadir un tema al gráfico: por ejemplo, en blanco y negro, o con especificaciones concretas, como el posicionamiento de la leyenda.
6. Añadir “facetas” (*facets*). De esta forma se divide el gráfico en varios subgráficos basándose en los valores de una o más variables discretas (normalmente categóricas).

En las secciones que siguen se verán ejemplos de todo el proceso. El siguiente es un ejemplo de una expresión que contiene los elementos anteriores. Se anima al lector a que los identifique con dicha lista:

```
ggplot(data, aes(x = variable)) + geom_histogram() +
  labs(title = "Título") + theme_bw() + facet_wrap(~factor)
```

11.2. Análisis exploratorio de una variable

11.2.1. Variables cualitativas

El resumen numérico de variables cualitativas se muestra en la tabla de frecuencias, la cual se puede representar con un gráfico de barras o con un gráfico de sectores². Las frecuencias absolutas son el número de observaciones en cada categoría y las frecuencias relativas son la proporción de observaciones en cada categoría con respecto al total. Por ejemplo, el conjunto de datos `accidentes2020_data` disponible en el paquete CDR describe los datos de accidentes de tráfico con víctimas y/o daños al patrimonio en la ciudad de Madrid registrados por Policía Municipal. Entre sus variables, contiene la variable cualitativa tipología del accidente `tipo_accidente`. Un resumen puede obtenerse tanto con la función `table()` como con el paquete `dplyr`, como se vio en la Sec. 3.6. En variables cualitativas, la categoría más frecuente se denomina **moda** de la variable³.

```
library("CDR")
library("dplyr")
accidentes2020_data |>
  count(tipo_accidente) |>
  mutate(porc = 100 * n / sum(n))
#> #>   tipo_accidente     n      porc
#> #> 1: Alcance 7294 22.4936010
#> #> 2: Atropello a animal 75 0.2312887
#> #> 3: Atropello a persona 2127 6.5593487
#> #> 4: Caída 2118 6.5315940
#> #> 5: Choque contra obstáculo fijo 4667 14.3923274
#> #> 6: Colisión frontal 899 2.7723810
#> #> 7: Colisión fronto-lateral 8081 24.9205909
#> #> 8: Colisión lateral 4386 13.5257656
#> #> 9: Colisión múltiple 2231 6.8800691
#> #> 10: Despeñamiento 2 0.0061677
#> #> 11: Otro 251 0.7740463
#> #> 12: Solo salida de la vía 151 0.4656613
#> #> 13: Vuelco 145 0.4471582
```

Para representar el gráfico de barras con la función `ggplot()`, se añade la capa de geometría con la función `geom_bar()` (véase Fig. 11.4).

²El gráfico de sectores no es recomendable, ya que proporciona la misma información que el gráfico de barras y para el ojo humano es mucho más difícil de distinguir ángulos que alturas.

³Nótese que el orden por defecto que utiliza R es el alfabético. Se puede cambiar este comportamiento reordenando los niveles del factor; por ejemplo, para poner “Otro” en la última posición.

```
library("ggplot2")
library("CDR")
accidentes2020_data |>
  ggplot() +
  geom_bar(aes(y=tipo_accidente), fill = "pink")
```

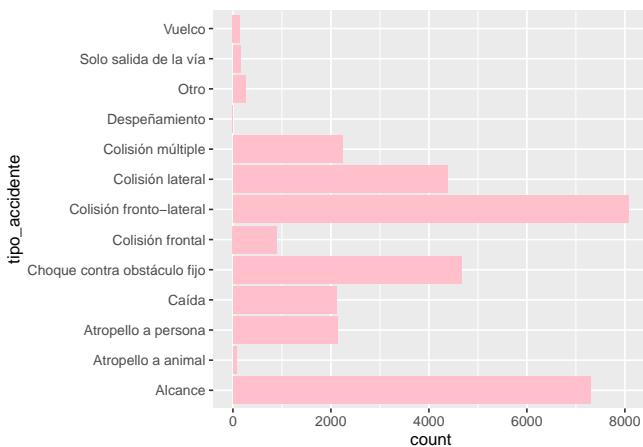


Figura 11.4: Gráfico de barras con ggplot2

El código anterior es la forma más básica de hacer un gráfico con `ggplot2`. Opciones más avanzadas pueden encontrarse en [Wickham and Grolemund \(2016\)](#).

Ya se ha comentado que los gráficos de sectores no se recomiendan a menos que se incluya en ellos información numérica. El paquete `ggstatsplot` realiza gráficos que incluyen análisis estadísticos. Por ejemplo, la función `ggpiestats()` proporciona un gráfico de sectores con algunos tests estadísticos (véase la ayuda de la función) y podría utilizarse para determinar en qué medida un conjunto de 80 ayuntamientos de distinto signo político presta o no un determinado servicio `serv` (véase el conjunto de datos en el paquete del libro `?CDR::ayuntam`). El siguiente código produce el gráfico de la Fig. 11.5.

```
library("ggstatsplot")
ayuntam |>
  ggpiestats(x = serv)
```

Una alternativa a los gráficos de sectores son los *waffle charts* (gráficos de gofre o de tableta de chocolate). La siguiente expresión produce de la Fig. 11.6 usando el paquete `waffle`. Con el argumento `use_glyph` se pueden incluir iconos en vez de cuadrados.

11.2. Análisis exploratorio de una variable

173

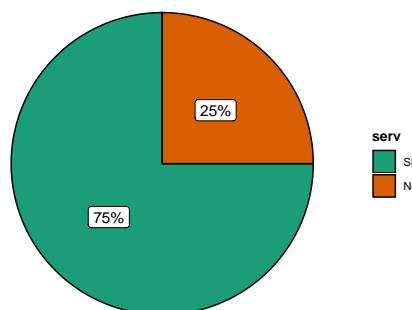
 $\chi^2_{\text{gof}}(1) = 20.00, p = 7.74e-06, \hat{\phi}_{\text{Pearson}} = 0.45, \text{CI}_{95\%} [0.30, 1.00], n_{\text{obs}} = 80$

 $\log_e(\text{BF}_{01}) = -8.04, \alpha_{\text{Gelman-Dickey}} = 1.00$

Figura 11.5: Gráfico de sectores con tests. Prestación o no de un determinado servicio X en ayuntamientos de distinto signo político

```
library("waffle")
freq <- ayuntam |>
  count(serv)
m <- setNames(freq$n, freq$serv)
waffle(m,
  rows = 4, colors = c("red", "green"))
```

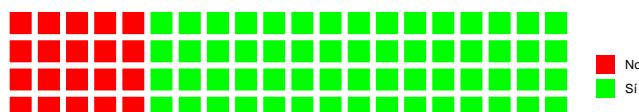


Figura 11.6: Gráfico waffle: Prestación o no de un determinado servicio X en 80 ayuntamientos de distinto signo político

11.2.2. Variables cuantitativas

Los estadísticos descriptivos más importantes que se utilizan en un AED se dividen en tres grandes grupos:

- **Medidas de posición**, que su vez se dividen en (i) centrales: media (`mean()`), mediana (`median()`) y moda y (ii) no centrales: cuantiles `quantile()`, mínimo (`min()`) y máximo (`max()`).
- **Medidas de dispersión**. Las más importantes son: varianza (`var()`), desviación típica (`sd()`), rango intercuartílico (`IQR()`), desviación absoluta mediana (`mad()`) y coeficiente de variación (`sd(x)/mean(x)`).
- **Medidas de forma**: asimetría (*skewness*) y apuntamiento (*kurtosis*).

La función `summary()` de R base es una función de las llamadas “genéricas” y solo aborda las medidas de posición.

```
summary(renta_municipio_data$`2019`)
#>   Min. 1st Qu. Median Mean 3rd Qu. Max. NA's
#> 4053    9914   11595 12247   13690 32183 5697
```

Sin embargo, los estadísticos descriptivos suelen presentarse juntos “describiendo” el conjunto de datos. Existen distintos paquetes, como `summarytools`, que proporcionan un resumen completo de un vector numérico con la función `descr()` así como de un conjunto de datos completo (ver opciones del paquete).

```
library("summarytools")
renta_municipio_data |>
  select(`2019`) |>
  descr()

#>Descriptive Statistics
#>2019
#>N: 55273
#>          2019
#>-----
#>      Mean 12246.84
#> Std.Dev 3562.94
#> Min 4053.00
#> Q1 9914.00
#> Median 11595.00
#> Q3 13690.50
#> Max 32183.00
#> MAD 2742.81
#> IQR 3776.25
#> CV 0.29
```

```
#>      Skewness      1.82
#> SE.Skewness     0.01
#>      Kurtosis      5.77
#>      N.Valid   49576.00
#> Pct.Valid     89.69
```

A continuación se proporciona una breve definición de las anteriores medidas y su fórmula matemática, como referencia general y por su uso en otras partes del libro. En las fórmulas siguientes, x_i representa cada valor de la variable en la muestra y n es el número de observaciones en dicha muestra.

- Media (*Mean*), \bar{x} : es el centro de gravedad de los datos, en torno al cual varían.

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}.$$

- Desviación típica (*Std.Dev*), s : es la raíz cuadrada de la varianza s^2 . Representa la variación promedio alrededor de la media, en las unidades de la variable y en sus unidades al cuadrado respectivamente. La siguiente fórmula corresponde a la varianza muestral, dividiendo por $n - 1$. Para la varianza poblacional, se dividiría por el tamaño de la población N . Para calcular la desviación típica, bastaría con hacer la raíz cuadrada.

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}.$$

- El mínimo (*Min*) y el máximo (*Max*) son los extremos de los datos.
- La mediana (*Median*) es el segundo cuartil Q_2 . Es el punto central de los datos, dejando la mitad de las observaciones por abajo y la otra mitad por arriba. Los cuartiles Q_1 y Q_3 , dividen los datos dejando por debajo de su valor el 25 % y el 75 %, respectivamente.
- La desviación absoluta mediana (*MAD*) es la mediana de las desviaciones a la mediana.
- El rango intercuartílico es la diferencia entre el tercer y el primer cuartil. Es decir, el rango del 50 % de las observaciones centrales:

$$IQR = Q_3 - Q_1.$$

- El coeficiente de variación es el cociente entre la desviación típica y la media (en valor absoluto). Es una medida de variabilidad relativa muy útil para comparar la variabilidad de distintas muestras o poblaciones.

$$CV = \frac{s}{|\bar{x}|}.$$

- Coeficiente de asimetría (*Skewness*). En variables simétricas es nulo. Si es mayor de cero, los datos presentan asimetría positiva (una cola a la derecha, en los valores altos con respecto a la media) y si es menor de cero, los datos presentan asimetría negativa (una cola a la izquierda, en valores bajos con respecto a la media).

$$g_1 = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{s^3}.$$

- Coeficiente de apuntamiento (*Kurtosis*). Si los datos presentan un apuntamiento similar al de la distribución normal, será próximo a cero. Cuanto más grande, más apuntado, y cuanto más pequeño, más aplanado.

$$g_2 = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{s^4} - 3.$$

Los otros tres valores que aparecen en la salida de la función `descr()` son el error típico del coeficiente de asimetría (*SE.Skewness*), el número total de valores válidos (*N.Valid*) y el porcentaje respecto al total de datos (*Pct.Valid*). El complementario de este último, por tanto, es el porcentaje de valores perdidos (*missing*).

La representación gráfica de la tabla de frecuencias de una variable cuantitativa es el **histograma**⁴. Para representarlo, se cuenta el número de observaciones (frecuencia) por intervalo (*bin*). Una posible regla sería el método de Sturges⁵, que se puede hallar con la función `nclass.Sturges()`.

Para obtener la tabla de frecuencias de la renta neta per cápita en 2019 usando el número de intervalos con la regla de Sturges se procede como sigue:

```
renta_municipio_data |>
  mutate(clases_sturges_renta = cut(`2019`,
    breaks = nclass.Sturges(`2019`))
)) |>
  count(clases_sturges_renta)
```

⁴En el caso de las variables discretas con un número de posibles valores pequeño, es mejor proceder igual que si fuera una variable cualitativa, obteniendo una tabla de frecuencias y un gráfico de barras, con la diferencia de que el orden de los posibles valores será el numérico.

⁵Este es el método que utiliza por defecto la función `hist` de **R** base, que además redondea la amplitud del intervalo para facilitar la interpretación. Otra regla muy sencilla es tomar como número de intervalos en torno a la raíz cuadrada del número total de datos.

11.2. Análisis exploratorio de una variable

177

Sin embargo, esta regla no siempre es la más apropiada, como se verá en la Sec. 40.4, pues debe estudiarse bien la naturaleza de la variable a analizar.

El histograma proporciona mucha información sobre la variable: (i) si es aproximadamente simétrica, (ii) si tiene forma de campana (se parece a la distribución Normal), (iii) si hay valores extremos y cómo son de frecuentes, y (iv) si puede haber mezcla de poblaciones (más de una moda).

La función `geom_histogram()` del paquete `ggplot2` añade una capa con un histograma al gráfico. El color de las barras se controla con el *aesthetics fill* y la altura puede representar las frecuencias absolutas (recuentos) o relativas (proporciones). El número de intervalos se indica con el argumento `bins`, o alternativamente la anchura de intervalo con `bin_width`, véase la Fig. 11.7.

```
p <- renta_municipio_data |>
  tidyverse::drop_na() |>
  ggplot(aes(`2019`))

h1 <- p + geom_histogram(color = "yellow", fill = "pink")
h2 <- p + geom_histogram(
  color = "yellow", fill = "pink",
  bins = nclass.Sturges(renta_municipio_data$`2019`)
)
h3 <- p + geom_histogram(color = "yellow", fill = "pink", bins = 20)

library("patchwork")
h1 + h2 + h3
```

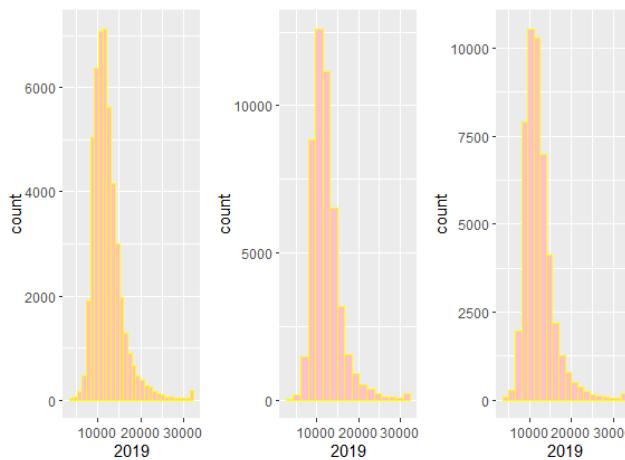


Figura 11.7: Histogramas de la renta neta per cápita en 2019 con distintos bins. Izquierda: bins por defecto ($n=30$); Centro: bins con la regla de Sturges; Derecha: bins = 20

Una representación alternativa al histograma es la línea de densidad, que sustituye las barras por una línea continua, generalmente suavizada. A continuación, se añade la linea de densidad a uno de los histogramas de la Fig. 11.7 y el resultado se puede ver en la Fig. 11.8.

```
p <- renta_municipio_data |>
  tidyverse::drop_na() |>
  ggplot(aes(`2019`))
p + geom_histogram(aes(y = after_stat(density)),
  position = "identity",
  color = "yellow", fill = "pink") +
  geom_density(lwd = 1, colour = 4)
```

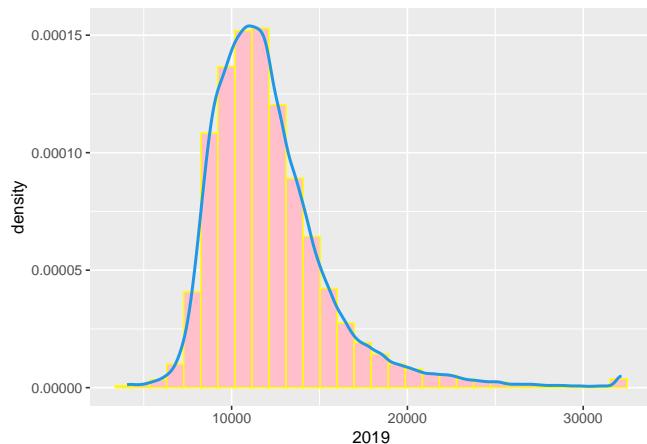


Figura 11.8: Histograma y linea de densidad de la renta neta per capita española en 2019

Otras representaciones gráficas muy útiles de las variables continuas son el gráfico de **caja y bigotes** y el **diagrama de violín**, que se obtienen fácilmente combinando en `ggplot()` las capas `geom_boxplot()` y `geom_violin()`, respectivamente (véase Fig. 11.9).

```
p <- renta_municipio_data |>
  tidyverse::drop_na() |>
  ggplot(aes(x=0, y= `2019`))
boxplot <- p + geom_boxplot(color = "yellow", fill = "pink")
violin <- p + geom_violin(aes(), color = "yellow", fill = "pink")
boxplot + violin
```

Otra visualización básica para una variable numérica es la visualización secuencial de las observaciones, bien a través de puntos (`geom_point()`) o a través de líneas (`geom_line()`). El orden de las observaciones puede indicar cuándo se ha producido un cambio u otros patrones.

11.3. Análisis exploratorio de varias variables

179

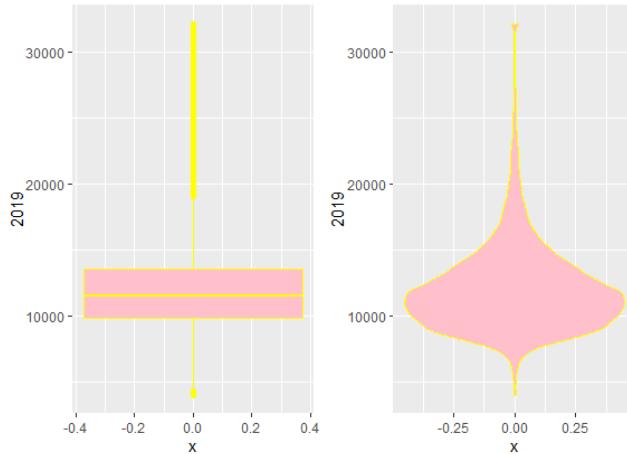


Figura 11.9: Boxplot y violin plot de la renta neta per cápita en 2019

11.3. Análisis exploratorio de varias variables

En la Sec. 11.2 se ha realizado un AED de variables aisladas, pero lo usual es incluir las relaciones entre variables en el AED. Las herramientas estadísticas utilizadas son: (i) las tablas de frecuencias conjuntas, que, en el caso de dos atributos, son tablas de doble entrada, con un atributo en filas y el otro en columnas, para determinar si existe asociación entre dichos atributos, como se verá en el Cap. 23; (ii) los resúmenes numéricos, como la covarianza, el coeficiente de correlación, coeficientes de asociación, etc. y (iii) los gráficos en los que se puede representar más de una variable.

11.3.1. Variables cualitativas

El resumen numérico sigue siendo la tabla de frecuencias, en este caso conjuntas para las distintas combinaciones de los niveles de las variables. Este tipo de tablas se denominan **tablas de contingencia** (véase Cap. 23). Para dos atributos, se puede representar en forma de tabla de doble entrada.

El resultado de la función `table()` se puede utilizar dentro de las funciones `prop.table()` y `addmargins()` para obtener las frecuencias relativas, añadir los totales marginales, o ambas cosas. Para el ejemplo de la prestación de un servicio “X” o no por parte de 80 ayuntamientos, `table()` podría utilizarse para dar respuesta a la siguiente pregunta: ¿La prestación pública del servicio X es independiente del signo político del Ayuntamiento o depende de dicho signo?

```
table(ayuntam$signo_gob , ayuntam$serv)
#>
#>           No  Sí
```

```
#> Avanzados 14 28
#> Ilustrados 6 32
```

No obstante, la representación gráfica más habitual siguen siendo los gráficos de barras, como se muestra en la Fig. 11.10 producida con el siguiente código:

```
p <- ayuntam |>
  ggplot(aes(signo_gob, fill = serv))

frecuencias <- p + geom_bar()
proporciones <- p + geom_bar(position = position_fill())

frecuencias + proporciones
```

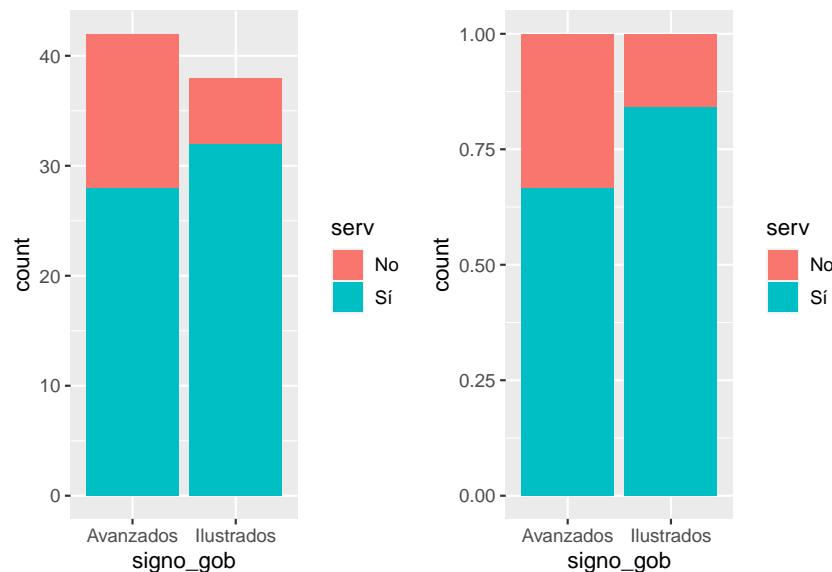


Figura 11.10: Grafico de barras de la prestación pública del servicio X por parte de 80 Ayuntamientos de distinto signo político. Izquierda: frecuencias absolutas. Derecha: frecuencias relativas.

Una visualización interesante de tablas de doble entrada son los gráficos en los que se representan las frecuencias conjuntas por medio de puntos cuya área es proporcional a la frecuencia. La Fig. 11.11 muestra gráficamente la tabla de frecuencias conjunta de los atributos `signo_gob` y `serv` del conjunto de datos `ayuntam`.

11.3. Análisis exploratorio de varias variables

181

```
library('gplots')
balloonplot(table(ayuntam$signo_gob , ayuntam$serv))
```

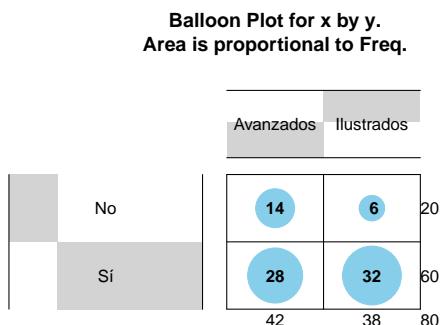


Figura 11.11: Representación gráfica de tabla de frecuencias con la función ‘balloonplot()’

Para representar dos o más factores a la vez en un único gráfico, se dispone de los gráficos de mosaico con la función `mosaicplot()` de R base, o bien el paquete `ggmosaic`, que incluye una función `geom_mosaic()` para usar en gráficos `ggplot2`. El siguiente ejemplo produce el gráfico de la Fig. 11.12:

```
library('ggmosaic')
accidentes2020_data |>
  ggplot() +
    geom_mosaic(aes(x = product(tipo_accidente, sexo),
                    fill=sexo))
```

En cualquier caso, se pueden representar más variables creando “subgráficos” o facetas (*facets*). Basta con añadir una capa al gráfico `ggplot2` con la función `facet_wrap()` y el argumento `facets` una lista de variables (categóricas o discretas) para cuyos valores se quiere hacer un gráfico distinto. Con el siguiente código⁶ se construye un gráfico para cada nivel del factor `tipo_accidente`. Cada uno de estos gráficos es una “faceta” del gráfico, que se muestra en la Fig. 11.13.

```
niveles <- levels(factor(accidentes2020_data$tipo_accidente))
etiquetas <- set_names(str_wrap(niveles, width = 20),
                       niveles)
accidentes2020_data |>
```

⁶Una sintaxis alternativa para especificar las facetas en la función `facet_wrap()` sería la siguiente: `facet_wrap(~tipo_accidente)`.

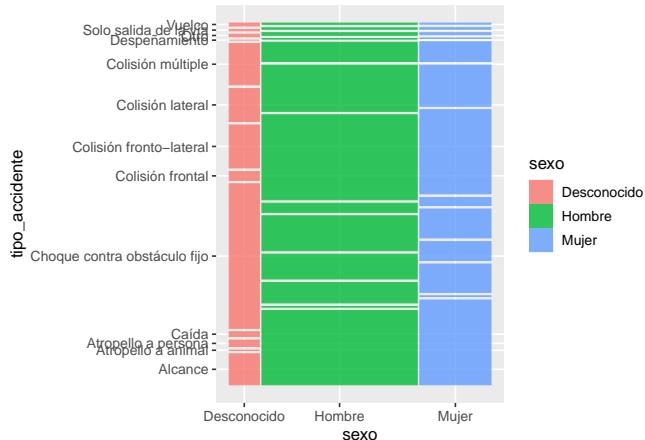


Figura 11.12: Gráfico de mosaico para relacionar el tipo de accidente y el sexo en los datos de accidentes

```
ggplot(aes(sexo, fill = estado_meteorológico)) +
  facet_wrap(vars(tipo_accidente),
             labeller = as_labeller(etiquetas)) +
  geom_bar() +
  labs(fill = "Estado Meteorológico") +
  theme(axis.text.x = element_text(angle = 90))
```

11.3.2. Variables cuantitativas

La descripción conjunta de variables numéricas se puede resumir con el vector de medias (medias de cada variable) y la matriz de varianzas-covarianzas. La covarianza s_{xy} (función `var()`) es una medida del grado de dependencia lineal entre dos variables numéricas. Si la covarianza es cero, no hay relación lineal (pero podría haber otro tipo de relación, recuérdese el cuarteto de Anscombe). Pero la covarianza es una medida que depende de la escala de las variables, por lo que

es más fácil interpretar el coeficiente de correlación lineal r_{xy} (función `cor()`), que está acotado entre -1 y 1. Cuanto más se acerque a 1, en valor absoluto, más fuerte será la dependencia lineal. Las fórmulas para calcular ambos estadísticos son las siguientes:

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}),$$

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y}.$$

11.3. Análisis exploratorio de varias variables

183

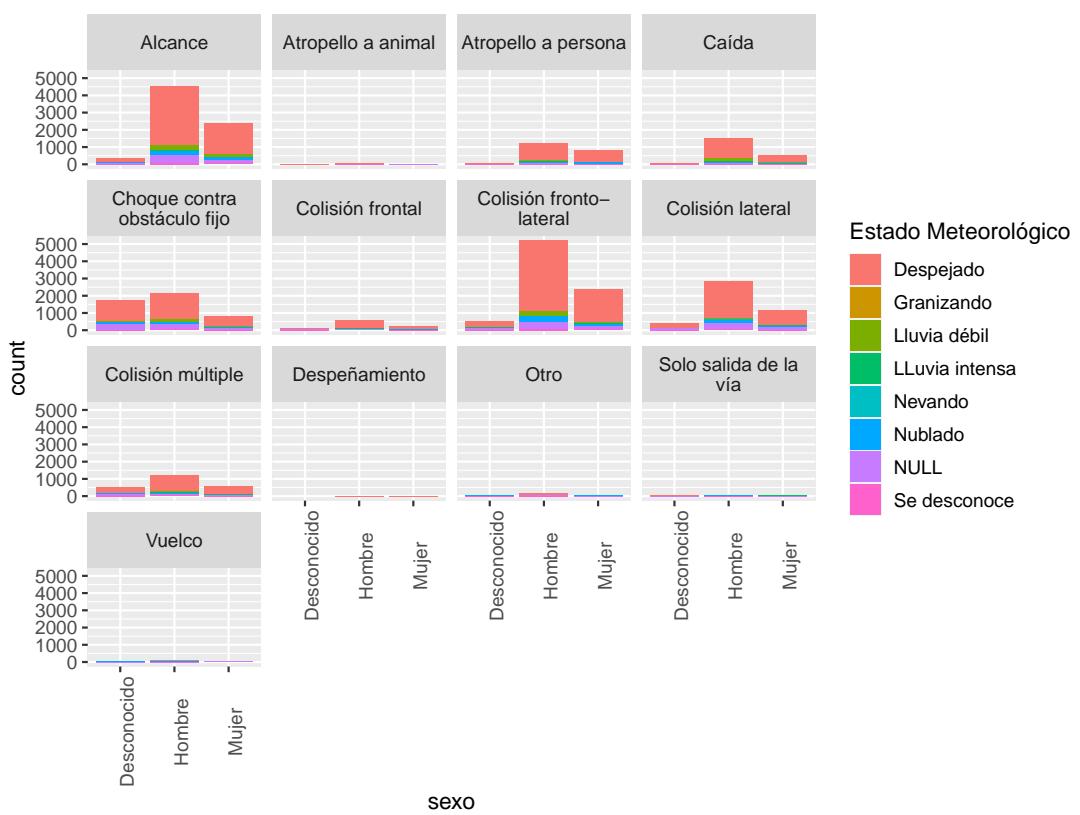


Figura 11.13: Representación de tres atributos mediante gráficos de barras conjuntos y facetas

Además, la matriz de correlación suele ser un punto de partida en las técnicas de reducción de la dimensionalidad (Véanse los Cap. 32, ?? y ??). Si, por ejemplo, se desea calcular la matriz de correlaciones del conjunto de datos TIC2021, que presenta las estadísticas de uso de las TIC en la Unión Europea 2021, se puede utilizar el paquete `corrplot`, que proporciona una forma elegante y versátil de representarla⁷ (véase la Fig. 11.14).

```
library('corrplot')
mcor_tic <- cor(TIC2021)
corrplot.mixed(mcor_tic, order = 'AOE')
```



Figura 11.14: Representación gráfica de la matriz de correlaciones entre las variables del conjunto de datos TIC2021

La matriz de correlaciones se puede representar mediante “mapas de calor” (*heatmap*), es decir, un cuadrado que representa las filas y columnas de la matriz de correlaciones (variables) y donde el color de las celdas es una gradación que depende del valor de las mismas. Un mapa de calor de la matriz de correlaciones guardada anteriormente, `mcor_tic`, puede obtenerse con la expresión `heatmap(mcor_tic)`.

En cuanto a los resúmenes gráficos, el **diagrama de dispersión** es el gráfico más popular. La función `geom_point()` de `ggplot2` añade una capa con los puntos (x, y), que ya nos da una idea de la relación entre las variables, y permite interpretarla conjuntamente con el coeficiente de correlación. Se puede añadir una línea de regresión, incluida una banda de confianza, por diversos métodos (función `geom_smooth()` por defecto, una curva *loess* o *gam* dependiendo del número de filas). Alternativamente a los puntos como objeto geométrico, se pueden representar líneas (`geom_line()`).

Por ejemplo, antes de llevar a cabo un ajuste lineal, o de otro tipo, con los datos `airquality`, tal y como se hará en los Cap. 15 a 19, se podría hacer un AED previo entre las variables `Ozone` y `Temp` con el gráfico de la Fig. 11.15.

⁷Una gran cantidad de ejemplos puede verse ejecutando `example(corrplot)`.

11.3. Análisis exploratorio de varias variables

185

```
airquality |>
  dplyr::select(Ozone,Temp) |>
  ggplot(aes( x= Temp, y=Ozone)) +
  geom_point() +
  geom_smooth()
```

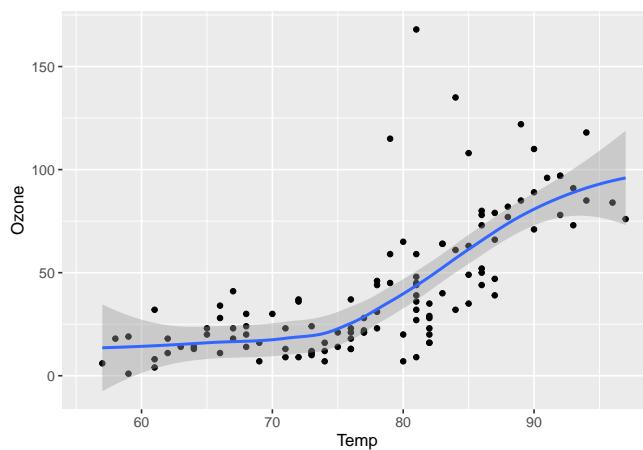


Figura 11.15: Gráfico de dispersión del Ozono frente a la Temperatura

Un caso particular es cuando la variable explicativa es el tiempo. En este caso, se tiene una serie temporal, y la representación con líneas es más adecuada (véase la Fig. 11.16).

```
library("dplyr")
contam_mad |>
  filter(nom_abv == "NOx") |>
  group_by(fecha, nom_mag) |>
  summarise(media_estaciones = mean(daily_mean, na.rm = TRUE)) |>
  ggplot(aes(x = fecha, y = media_estaciones)) +
  geom_line(aes(color = nom_mag)) +
  geom_smooth(linewidth = 0.5, color = "black", se = TRUE) +
  theme(legend.position = "none")
```

11.3.3. Variables cualitativas y cuantitativas

Cuando se trabaja en un proyecto de ciencia de datos, lo normal es tener tanto variables cualitativas como cuantitativas. Para representar conjuntamente ambos tipos de variables existen múltiples posibilidades, algunas de las cuales se enumeran a continuación, con el tipo de gráfico adecuado:

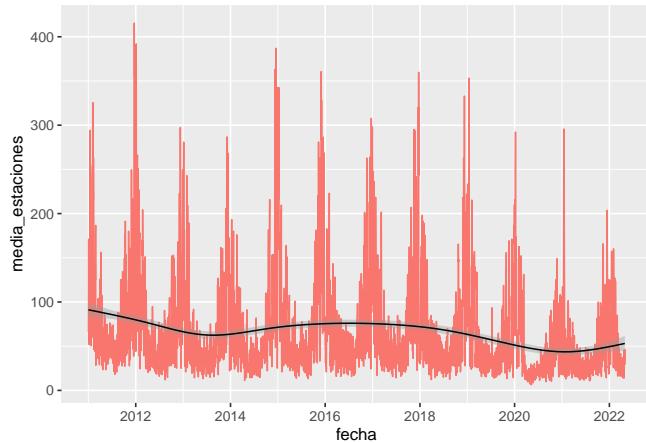


Figura 11.16: Concentración media semanal de NOx en las estaciones de medición de Madrid (enero 2011- marzo 2022)

- Una variable numérica y una variable categórica: gráficos de cajas o de violín para cada nivel de la categórica (Fig. 11.17), o bien gráficos de densidad para cada categoría (Fig. 11.17).

```
contam_mad |>
  na.omit() |>
  filter(nom_abv == "PM10") |>
  filter(between(fecha, left = as.Date("2022-03-10"), right = as.Date("2022-03-20")))
  |>
  ggplot(aes(zona, daily_mean)) +
  geom_violin() +
  geom_jitter(height = 0, width = 0.01) +
  aes(x = zona, y = daily_mean, fill = zona)
```

```
library('ggridges')
contam_mad |>
  filter(nom_abv == "NOx") |>
  ggplot(aes(x = daily_mean, y = tipo, fill = tipo)) +
  geom_density_ridges()
```

- Dos variables numéricas y varias variables categóricas: gráfico de dispersión para las numéricas y mapeado del color, tamaño y símbolo por cada nivel de las categóricas, como en la Fig. 11.19.

11.3. Análisis exploratorio de varias variables

187

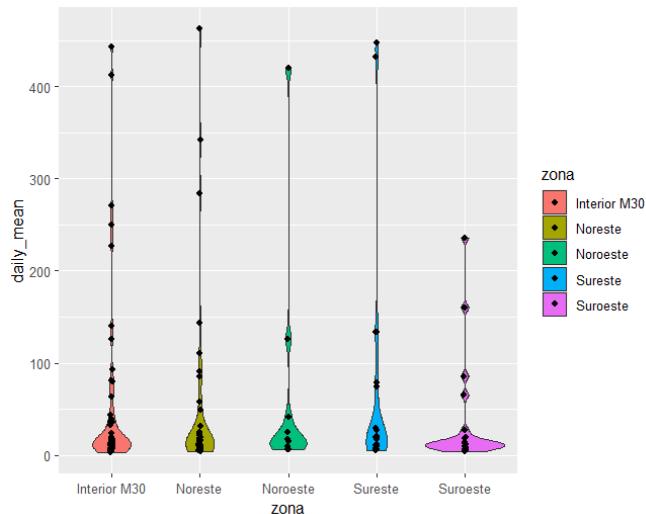


Figura 11.17: Comparación de los niveles de PM10 en las Zonas de la ciudad de Madrid a efectos de Calidad del Aire durante la Calima de marzo de 2022

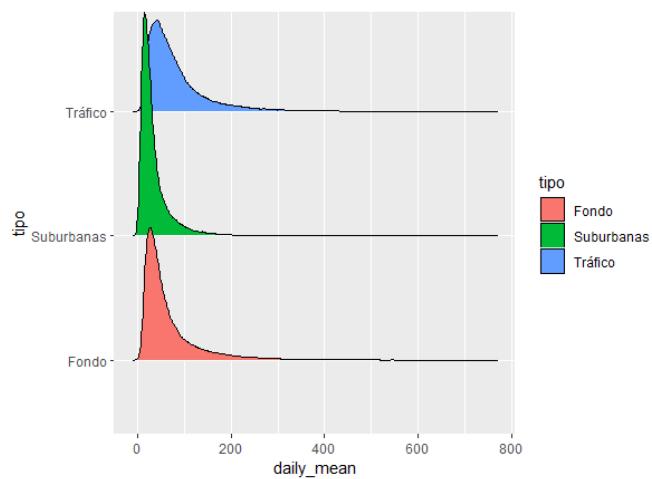


Figura 11.18: Comparación de concentraciones de NOx por tipo de estación de medición

```
# periodo del estado de alarma
pm10_nox_mad <- contam_mad |>
  na.omit() |>
  filter(nom_abv %in% c("PM10", "NOx")) |>
  filter(between(fecha, left = as.Date("2020-03-14"), right = as.Date("2020-06-30")))
  |>
  select(estaciones, zona, tipo, nom_abv, daily_mean, fecha) |>
tidyrr::pivot_wider(names_from = "nom_abv", values_from = "daily_mean", values_fn =
  mean)

pm10_nox_mad |>
ggplot(
  aes(x = PM10, y = NOx, colour = tipo, size = zona )) +
  geom_point()
```

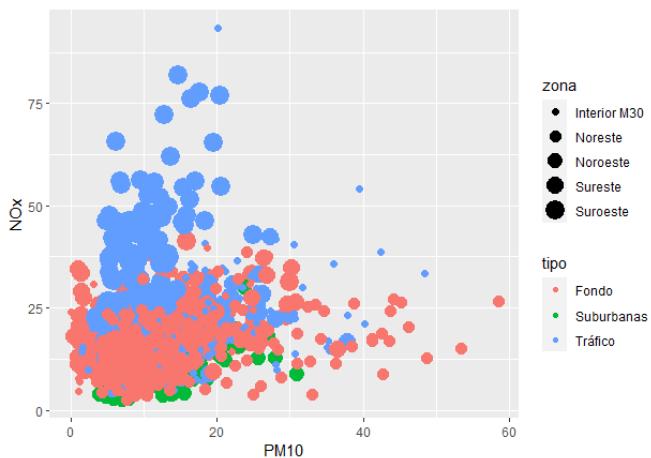


Figura 11.19: Gráfico de dispersión de las variables ‘NOx’, ‘PM10’, ‘zona’ y ‘tipo’ (de emplazamiento) durante el estado de alarma en la ciudad de Madrid (todas las estaciones de medición)

- Más de dos variables numéricas y más de una categórica: gráfico de dispersión y mapeado de las otras variables a otros *aesthetics*. Combinación de geometrías y *aesthetics*. Por ejemplo, añadir puntos con efecto *jitter* a un gráfico de cajas.

En todos los casos anteriores se pueden crear “facetas” para hacer un gráfico por cada combinación de variables categóricas, de forma que se tenga un buen número de variables representadas en un mismo “lienzo”, como en la Fig. 11.20.

```
pm10_nox_mad |>
  tidyrr:: drop_na() |>
```

11.3. Análisis exploratorio de varias variables

189

```
ggplot(aes(y=NOx, x= PM10, colour = tipo, shape = zona)) +
  geom_point() +
  geom_smooth() +
  facet_wrap(vars(estaciones))
```

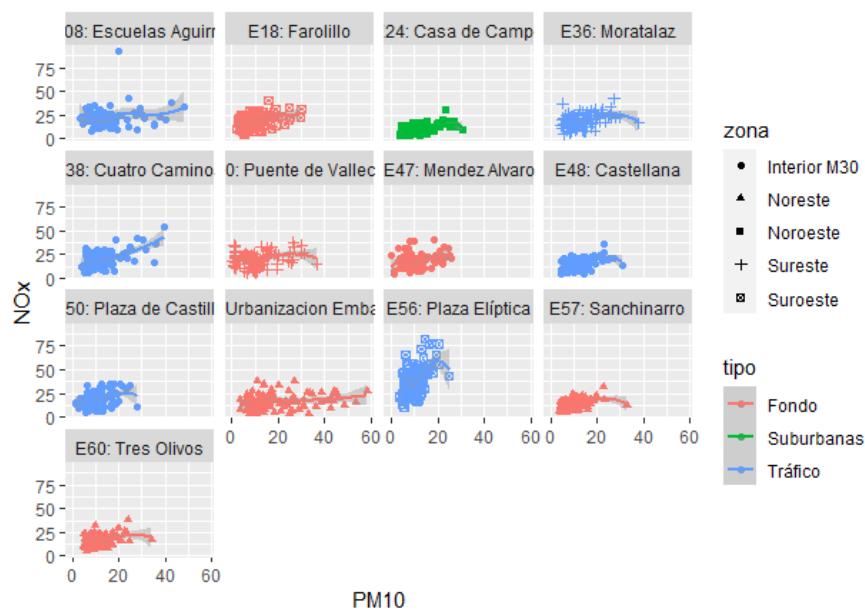


Figura 11.20: Gráfico de dispersión de las variables ‘NOx’, ‘PM10’, ‘zona’ y ‘tipo’ (de emplazamiento) por estación de medición durante el estado de alarma en la ciudad de Madrid

Resumen

Análisis exploratorio de una variable

- El análisis exploratorio es una tarea fundamental antes de abordar cualquier otra técnica estadística.
- Las variables cualitativas se resumen con tablas de frecuencias y gráficos de barras.
- Las variables cuantitativas discretas se pueden resumir también con tablas de frecuencias y gráficos de barras, pero si hay muchos valores distintos también pueden ser apropiados los histogramas.
- Las variables cuantitativas continuas se pueden resumir con tablas de frecuencias por intervalos, medidas de posición y de dispersión, histogramas y gráficos de cajas.
- Los gráficos de caja sirven, además, para identificar valores atípicos.

Análisis exploratorio de varias variables

- Las variables cualitativas se pueden resumir con tablas de frecuencias conjuntas y su representación gráfica y con combinaciones de gráficos de barras.
- La principal medida conjunta de dos variables cuantitativas es el coeficiente de correlación. Para más de dos variables se suelen representar en forma de matriz.
- El gráfico de dispersión es la representación básica para dos variables cuantitativas. Se pueden representar estos gráficos por pares en forma de matriz de gráficos.
- Para añadir más variables, se pueden “mapear” variables a *aesthetics* (tamaño, color, etc.), añadiendo más objetos geométricos, o bien añadiendo “facetas” (subgráficos) para cada variable o para cada posible valor de una variable cualitativa.

Parte III

Fundamentos de estadística

Capítulo 12

Probabilidad

M^a Leticia Meseguer Santamaría^a y Manuel Vargas Vargas^a

^a Universidad de Castilla-La Mancha

12.1. Introducción a la probabilidad

La incertidumbre es inevitable en muchos campos científicos, producto de la imposibilidad de predeterminar el resultado de un fenómeno repetido bajo idénticas condiciones, el desconocimiento de todas o algunas causas que pueden influir en él, o una información limitada sobre los condicionantes que rigen su comportamiento. De hecho, gran parte del avance científico consiste en reducir o controlar el nivel de incertidumbre, mejorando el proceso de obtención e interpretación de datos o estableciendo **modelos** que expliquen los resultados.

Producto de la **incertidumbre**, las decisiones que se toman (o la validez de los resultados que se obtienen) conllevan un **riesgo**, que puede concretarse en enunciados equivocados, modelos con escaso poder predictivo o decisiones con resultados no deseados. Sin embargo, no se prescinde de tomar decisiones en ambientes de incertidumbre, sino que se intenta evaluar y minimizar los riesgos asociados.

Así pues, resulta importante poder “**medir**” la incertidumbre, es decir, cuantificar su magnitud y establecer reglas de medida que permitan su tratamiento –la estimación de riesgo– y ayuden a la toma de decisiones. La **teoría de la probabilidad** se puede entender como un ente que proporciona reglas de comportamiento que ayudan a conseguir los objetivos anteriores, siendo el campo de aplicación tan amplio que puede cubrir cualquier rama de las ciencias sociales, técnicas y naturales.

El concepto de probabilidad apareció en la antigüedad, asociado a los juegos de azar, y se ha ido refinando y formalizando a lo largo de la historia. Sin embargo, la mayoría de las definiciones tradicionales presentan limitaciones que impiden su uso riguroso en cualquier situación. Aún así, siguen estando en el subconsciente colectivo, de forma que se entienden expresiones como

“*es muy probable que llueva mañana*”, “*es improbable que me toque la lotería de Navidad*”, o “*es probable que se obtenga en breve una vacuna contra cierta enfermedad*”, cuando responden a conceptualizaciones diferentes y, en muchos casos, vagas e imprecisas.

Aunque sigue habiendo debates filosóficos y epistemológicos sobre el concepto de probabilidad (véase, por ejemplo, [Hajek and Hitchcock \(2016\)](#)), su uso generalizado en muchos campos científicos está más relacionado con el desarrollo de su carácter de **medida de la incertidumbre** que con un tratamiento matemático que permite su aplicación práctica (véase, por ejemplo, [Ross \(2012\)](#), [Morin \(2016\)](#) o [Balakrishnan et al. \(2019\)](#)). Es este enfoque el que se desarrolla sucintamente en este capítulo.

12.2. Probabilidad: elementos básicos, definición y teoremas

El desarrollo del concepto de probabilidad, entendido como medida de la incertidumbre sobre la ocurrencia de un evento, precisa de algunos requisitos previos que permitan una aplicación operativa.

En primer lugar, es necesario definir en qué situaciones se puede aplicar. Se entenderá por **experimento** cualquier acción u observación de la realidad que pueda repetirse varias veces en idénticas condiciones, dando lugar a resultados identificables y conocidos antes de ser realizado. Cuando, dadas las condiciones, se conoce qué resultado se producirá, se dice que el experimento es **determinista**; en caso contrario, si dadas las condiciones, no se puede saber cuál de los posibles resultados ocurrirá, el experimento se denomina **aleatorio** o **estocástico**. Así pues, sólo se puede hablar de **probabilidad** sobre **experimentos aleatorios**.

Ahora, dado un experimento aleatorio (E) y el conjunto de posibles resultados, denominados genéricamente **sucesos** (Ω), se define **probabilidad** como una medida del grado de creencia en la ocurrencia de cada posible suceso, $S \in \Omega$. Como se ve, la definición es muy amplia, por lo que precisa de algún requisito para evitar que una asignación concreta de grados de creencia produzca inconsistencias. Dicho requisito supone el cumplimiento de una estructura (matemática) concreta, que se adopta de forma axiomática. La más conocida, debida a Andrei Kolmogorov, se puede formalizar de la siguiente forma:

Axiomática de Kolmogorov: se considera un experimento aleatorio E , el conjunto de posibles sucesos Ω , y una función real P que asigna a cada suceso un número real. Se dice que P es una medida de probabilidad si cumple:

- 1. $P(S) \geq 0, \forall S \in \Omega$.
- 2. $P(\Omega) = 1$.
- 3. Dada una sucesión numerable de sucesos $\{S_i\}$ disjuntos dos a dos, es decir $S_i \cap S_j = \emptyset \forall i, j$ (donde \emptyset es el suceso imposible), la probabilidad del suceso unión es la suma de sus probabilidades: $P\left(\bigcup_i S_i\right) = \sum_i P(S_i)$.

12.2. Probabilidad: elementos básicos, definición y teoremas

195

Así, una probabilidad es una medida que cumple esta axiomática, asignando a cada suceso un número real (entre 0 y 1) que expresa el grado de creencia en la ocurrencia de dicho suceso, entendiendo que 0 indica que se cree que no ocurre nunca y 1 que ocurre seguramente (véase, por ejemplo [de Finetti \(2017\)](#) para su fundamentación).

Algunas consecuencias que se derivan de la axiomática de Kolmogorov de forma inmediata son:

- $P(\emptyset) = 0$.
- Denominando \bar{S} al suceso complementario, $P(\bar{S}) = 1 - P(S)$.
- Dados dos sucesos cualesquiera, $P(S_1 \cup S_2) = P(S_1) + P(S_2) - P(S_1 \cap S_2)$.

Sin embargo, esta definición es formal, en el sentido de que indica qué requisitos debe cumplir para evitar inconsistencias, pero no determina qué valor concreto de probabilidad asignar a cada suceso. Históricamente, se han propuesto varias concepciones para resolver este problema:

- **Concepción clásica (o de Laplace):** dado un experimento aleatorio E con n posibles resultados elementales mutuamente excluyentes e igualmente verosímiles, la probabilidad de un suceso S_i es:

$$P(S_i) = \frac{\text{casos favorables a la ocurrencia de } S_i}{\text{casos posibles}} = \frac{n_i}{n} = f_i. \quad (12.1)$$

Por “igualmente verosímiles” se entiende que “no hay razón para afirmar que uno suceda más veces que otro”, conocido como “principio de razón insuficiente”. Es fácilmente comprobable que esta regla cumple la axiomática de Kolmogorov e interpreta la probabilidad como la “frecuencia” de ocurrencia de cada suceso. A pesar de sus limitaciones (utiliza la equiprobabilidad de los sucesos elementales para definir la probabilidad y asume un conjunto finito de ellos), su fácil comprensión y utilidad en casos sencillos hace que esta regla sea muy utilizada (e incluso confundida con una “definición” de probabilidad).

- **Concepción frecuentista:** se consideran n repeticiones de un experimento aleatorio, manteniendo idénticas condiciones. Sea n_i el número de veces que se presenta el suceso S_i ; entonces se le asigna la probabilidad:

$$P(S_i) = \lim_{n \rightarrow \infty} \frac{n_i}{n}. \quad (12.2)$$

Esta concepción extiende la versión clásica, identificando la probabilidad con la frecuencia relativa de cada suceso cuando el experimento se repite un gran número de veces.

El siguiente paso es formalizar cómo influye la ocurrencia de un suceso sobre la probabilidad de que ocurran otros. Así, dado un suceso A con $P(A) > 0$, la probabilidad de que ocurra otro, B , condicionado a que ha ocurrido A , $P(B/A)$, se calcula como:

$$P(B/A) = \frac{P(B \cap A)}{P(A)}. \quad (12.3)$$

es decir, la probabilidad de que ocurran simultáneamente ambos dividida entre $P(A)$ para que $P(\Omega/A) = 1$.

Esta nueva medida se denomina **probabilidad condicionada**¹ y permite obtener resultados fundamentales para el cálculo de probabilidades:

- **Independencia de sucesos:** dos sucesos, A y B, se dicen independientes si $P(A/B) = P(A) \Rightarrow P(A \cap B) = P(A)P(B)$.
- **Teorema de la probabilidad total:** dado un conjunto de sucesos $\{A_i\}$ disjuntos y cuya unión es Ω (denominado **partición** de Ω), la probabilidad de cualquier suceso B compatible con los A_i es:

$$P(B) = \sum_i P(B/A_i)P(A_i). \quad (12.4)$$

Este teorema permite determinar la probabilidad de un suceso B , que puede tener varias causas, o darse bajo diversas alternativas, A_i , mediante la suma de las probabilidades de que aparezca B condicionada a cada una de las causas ponderadas por la probabilidad de cada causa o alternativa.

- **Teorema de Bayes:** dada una partición de Ω , y un suceso B con $P(B) > 0$, la probabilidad de cada elemento de la partición condicionada a que ha ocurrido B es:

$$P(A_i/B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(B/A_i)P(A_i)}{\sum_j P(B/A_j)P(A_j)}. \quad (12.5)$$

Este teorema, aplicación directa de la definición de probabilidad condicionada y del teorema de la probabilidad total, es un resultado tan importante que su uso ha dado nombre a una rama entera de la estadística, la conocida como **estadística bayesiana**.² También es utilizado en la moderna **inteligencia artificial**, en técnicas como Naive Bayes (véase Cap. 27)

12.3. Variable aleatoria y su distribución de probabilidad

Una limitación operativa de la probabilidad, tal como se ha utilizado hasta ahora, es que hace referencia a sucesos y operaciones entre conjuntos, lo que dificulta su tratamiento. Sin embargo, en muchos casos los sucesos están caracterizados por valores numéricos, por lo que podrían ser utilizados en sustitución de los primeros para facilitar los cálculos. A esta idea corresponde la noción de **variable aleatoria** (v.a.), que es una función que asigna un valor numérico a cada

¹ Esta definición cumple la axiomática de Kolmogorov y es la forma de introducir “información” en la determinación de probabilidades. Su versión para distribuciones de probabilidad es muy utilizada en inferencia estadística.

² $P(A_i)$ se denominan **probabilidades a priori**, $P(B/A_i)$ **verosimilitudes** y $P(A_i/B)$ **probabilidades a posteriori**. El teorema establece cómo se modifican las probabilidades cuando se introduce información en forma de verosimilitudes, siendo muy utilizada su versión para distribuciones de probabilidad.

12.4. Modelos de distribución de probabilidad

197

suceso de un experimento aleatorio. Para trabajar con probabilidades sobre números, a cada uno se le asigna la probabilidad de los sucesos que están caracterizados por dicho valor.³

Dada una v.a. X , su **función de distribución** asigna a cada número real x la probabilidad de que la variable tome un valor menor o igual que x ,

$$F_X(x) = P(X \leq x). \quad (12.6)$$

Una variable se dice **discreta** si sólo puede tomar un conjunto finito (o infinito numerable) de valores con probabilidad positiva. A ese conjunto de valores y sus probabilidades $\{x_i; P(X = x_i)\}$ se le denomina **función de cuantía**.

Una variable se denomina **continua**, si su función de distribución es continua y existe su primera derivada y es continua. Como consecuencia, la probabilidad en un valor concreto siempre será cero, $P(X = x_i) = 0$, por lo que sólo habrá probabilidades positivas sobre intervalos. Se denomina **función de densidad** a la derivada de la función de distribución $f(x) = F'(x) = \frac{dF(x)}{dx}$.⁴

Dado que una v.a. X está caracterizada por su distribución de probabilidad (a través de la función de distribución o de la de cuantía-densidad), se han desarrollado **modelos de distribución de probabilidad** que permiten modelizar el comportamiento aleatorio de las v.a. y calcular probabilidades de forma sencilla.

12.4. Modelos de distribución de probabilidad

En esta sección se presentan los modelos más utilizados en la práctica, distinguiendo entre modelos discretos y continuos, según la naturaleza de la v.a. Para una visión completa, se puede consultar, por ejemplo, [Johnson et al. \(2008\)](#).

12.4.1. Modelos discretos

Los modelos de distribución discretos más populares son el binomial, el binomial negativo y el de Poisson. Los dos primeros se asientan sobre el **fenómeno de Bernoulli** con independencia, que, de manera general, consiste en un experimento dicotómico (o que puede considerarse dicotómico), es decir, que se consideran sólo dos posibles resultados (uno identificado con el **éxito** del experimento, cuya probabilidad se denota por p , y el otro con el **fracaso**, con probabilidad $q = (1 - p)$) tal que los resultados producidos por el experimento son independientes de los precedentes.

■ Distribución Binomial $B(n,p)$

³Matemáticamente, una variable aleatoria es una función $X : \Omega \rightarrow \mathbb{R}$ que, para cada valor real, cumple $X^{-1}(x) \in \Omega$, de forma que se pueda asignar $P(x) = P(X^{-1}(x))$.

⁴La función de densidad es el equivalente continuo de la función de cuantía, indicando, de forma intuitiva, dónde se “concentra” la probabilidad de observar valores de X . De hecho, no es raro que se utilice el término general de “densidad” independientemente del tipo de variable que sea.

La distribución binomial (n, p) es una distribución de probabilidad discreta que asigna probabilidades al número de éxitos en una secuencia de n experimentos independientes de Bernoulli con una probabilidad fija de éxito p . Puede tomar los valores $x = 0, 1, \dots, n$ y su función de cuantía es:

$$P(X = x) = \binom{n}{x} p^x q^{n-x} \equiv \frac{n!}{x!(n-x)!} p^x q^{n-x}. \quad (12.7)$$

Su esperanza, valor esperado o media es $E(X) = \mu = np$, y su varianza $Var(X) = \sigma^2 = npq$.

La representación gráfica de las funciones de cuantía y distribución se muestra en la Fig. 12.1

```
par(mfrow=c(1,2))
x<-0:10
dens <- dbinom(0:10, size=10, prob=0.5)
plot(x, y=dens, type="h", xlab="x",ylab="P(x)",main="Función de cuantía", col="red",
      lwd=2)
plot(x, pbinom(x,10,0.5),ylab="F(x)",xlab="x",type="s",main="Función de
      distribución",lwd=2)
```

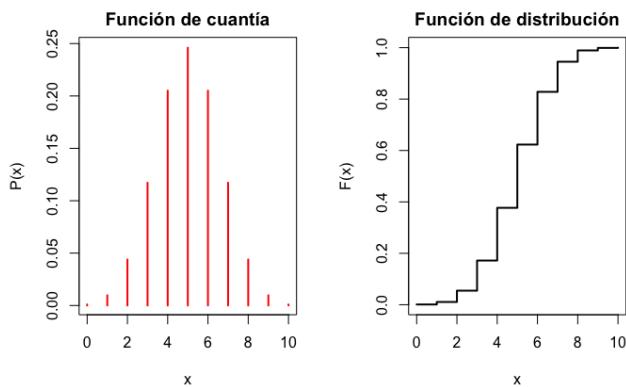


Figura 12.1: Función de cuantía y de distribución para una variable $B(10, 0.5)$

En el **caso particular** de que $n=1$, $B(1,p)$, se denomina **distribución Bernoulli**, $B(p)$.

- **Distribución Binomial negativa o de Pascal $BN(r,p)$**

La distribución binomial negativa surge en el contexto de una serie de experimentos de Bernoulli independientes, con probabilidad constante de éxito p , donde la v.a. X denota el número de experimentos fracasados (x) hasta que se produce un número determinado de éxitos (r). Puede tomar los valores $x = 0, 1, \dots$ y su función de cuantía es:

$$P(X = x) = \binom{x + r - 1}{x} q^x p^r \equiv \frac{(x + r - 1)!}{(r - 1)!x!} q^x p^r, \quad (12.8)$$

12.4. Modelos de distribución de probabilidad

199

con media $E(X) = r \frac{q}{p}$ y varianza $Var(X) = r \frac{q}{p^2}$.

La representación gráfica de las funciones de cuantía y distribución se muestra en la Fig. 12.2

```
par(mfrow=c(1,2))
x<-0:20
plot(x,dnbinom(x,3,0.35),type="h",ylab="P(x)",xlab="x",main="Función de cuantía",
  col="pink", lwd=2)
plot(x, pnbinom(x,3,0.35),ylab="F(x)",xlab="x",type="s",main="Función de
  distribución",lwd=2)
```

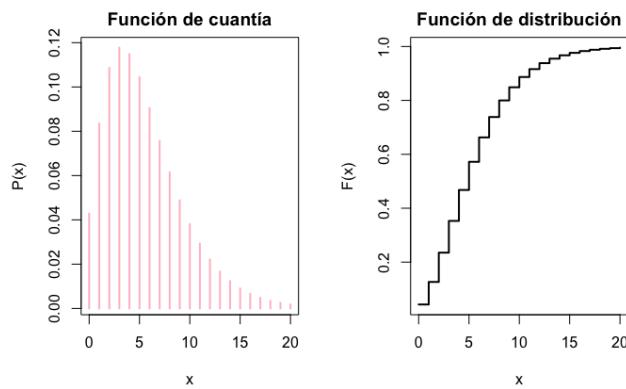


Figura 12.2: Función de cuantía y de distribución para una variable BN(3,0.35)

En el caso particular de que $r = 1$, BN($1,p$), se denomina **distribución geométrica G(p)**.

- **Distribución de Poisson P(λ)**

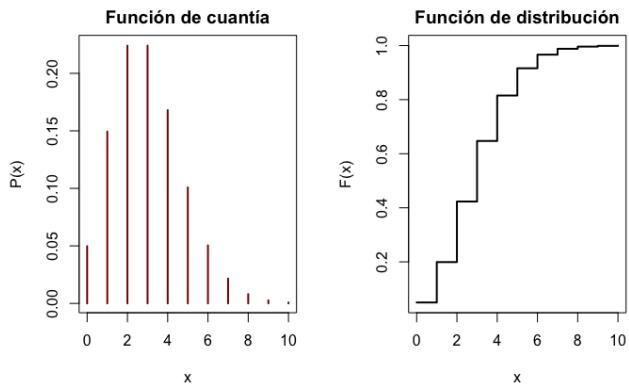
Se denominan **fenómenos de Poisson** a aquéllos en los que la ocurrencia de un suceso se encuentra distribuida a lo largo de un tiempo (o espacio) dado, cumpliendo que el proceso es estable, con una media de ocurrencias λ por unidad de tiempo, ocurrencias que se presentan de forma aleatoria e independiente. La variable que mide el número x de ocurrencias puede tomar los valores $x = 0, 1, 2, \dots$ y se dice que sigue una distribución de Poisson, con función de cuantía:

$$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}. \quad (12.9)$$

Su media es $E(X) = \lambda$ y su varianza $Var(X) = \lambda$.

La representación gráfica de las funciones de cuantía y distribución se muestra en la Fig. 12.3

```
par(mfrow=c(1,2))
x<-0:10
plot(x,dpois(x,3),type="h",ylab="P(x)",xlab="x",main="Función de cuantía",
  col="darkred",lwd=2)
plot(x, ppois(x,3),ylab="F(x)",xlab="x",type="s",main="Función de distribución",lwd=2)
```

Figura 12.3: Función de cuantía y de distribución para una variable $P(2.5)$

12.4.2. Modelos continuos

Los modelos de distribución para variables continuas más habituales son el Normal, el Gamma, el Chi-cuadrado, el t -student y el F -Snedecor.

■ Distribución Normal $\mathbf{N}(\mu, \sigma)$

La distribución Normal, de Gauss o gaussiana, tiene una gran importancia debido a que un gran número de fenómenos aleatorios se pueden modelizar a partir de ella (véase la Sec. 12.5 sobre el teorema central del límite). Además, es la distribución que se toma como supuesto y en la que se basan muchas de las técnicas estadísticas que se ven en este libro.

Una v.a. se dice que sigue una distribución normal de parámetros μ y σ si puede tomar cualquier valor real y su función de densidad es de la forma:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}. \quad (12.10)$$

Su media es $E[X] = \mu$ y su varianza, $Var(X) = \sigma^2$. Por ello, se suele decir que la normal está caracterizada por su media y su desviación típica.

La gráfica de la función de densidad tiene forma de campana (conocida como **campana de Gauss**) y es simétrica respecto de la media, con mayor probabilidad en las colas conforme aumenta la desviación típica, como muestra la Fig. 12.4.

```
par(mfrow=c(1,2))
x<-seq(-5, 5, 0.01)
plot(x,dnorm(x,0,1), ylab="P(x)", xlab="x", main="Función de
→ densidad",type="l",col="blue")
curve(dnorm(x,0,1.5), ylab="P(x)", add=TRUE,type="l",col="red")
curve(dnorm(x,0,2), ylab="P(x)", add=TRUE ,type="l",col="darkgreen")
```

12.4. Modelos de distribución de probabilidad

201

```
plot(x, pnorm(x,0,1),ylab="F(x)",xlab="x",type="s",main="Función de
↓ distribución",col="blue")
curve(pnorm(x,0,1.5), ylab="P(x)", add=TRUE ,type="l",col="red")
curve(pnorm(x,0,2), ylab="P(x)", add=TRUE ,type="l",col="darkgreen")
```

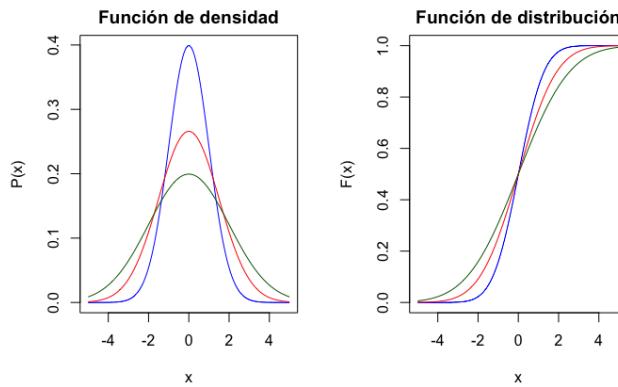


Figura 12.4: Función de densidad y de distribución de variables Normales, con media 0 y desviación típica 1 (azul), 1.5 (rojo), y 2 (verde)

Una característica importante de la distribución normal es que verifica la **propiedad aditiva o reproductiva**, es decir, que las combinaciones lineales de distribuciones normales independientes siguen siendo distribuciones normales. Si se consideran n variables aleatorias independientes con distribuciones $N(\mu_i, \sigma_i)$, cualquier combinación lineal cumple:

$$\beta_0 + \beta_1 X_1 + \dots + \beta_n X_n = \beta_0 + \sum_{i=1}^n \beta_i X_i \sim N\left(\beta_0 + \sum_{i=1}^n \beta_i \mu_i, \sqrt{\sum_{i=1}^n \beta_i^2 \sigma_i^2}\right). \quad (12.11)$$

En particular, si $X \sim N(\mu, \sigma)$, la variable $Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$, y se conoce como **normal estándar** o **normal tipificada**.⁵

- **Distribución Gamma** $\Gamma(\alpha, \beta)$

La **distribución Gamma** es útil en el contexto de los fenómenos de Poisson o cuando se trata de asignar probabilidades al tiempo de espera (o la vida útil) hasta que ocurre un número determinado de sucesos (α), suponiendo que β es el tiempo medio entre ocurrencias de un suceso.

Esta distribución toma valores positivos y su función de densidad viene dada por la expresión:

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, \quad (12.12)$$

⁵También es posible interpretar que toda distribución normal es una transformada de la distribución Z , ya que $X = \mu + \sigma Z$.

donde $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx = (\alpha - 1)!$ si α es un número natural.

Su media es $E(X) = \alpha\beta$ y su varianza, $Var(X) = \alpha\beta^2$. El parámetro α es conocido como **parámetro de forma**; si $\alpha \leq 1$, la función de densidad tiene forma de “L”; si $\alpha > 1$, la distribución es campaniforme, con asimetría positiva, y conforme va aumentando, el centro de la distribución se desplaza hacia la derecha. β se conoce como **parámetro de escala** y determina el alcance de la asimetría positiva. La Fig. 12.5 muestra la representación gráfica de las funciones de densidad y distribución para varias combinaciones de valores de los parámetros ($\beta = 2$; $\alpha = 1$ (azul), $\alpha = 2$ (morado), $\alpha = 5$ (rojo) y $\alpha = 10$ (verde)).

```
par(mfrow=c(1,2))
x<-seq(0, 10, 0.01)
plot(x, dgamma(x,2,2),type="l", ylab="f(x)",main="Función de
→ densidad",col="purple",lwd=2)
curve(dgamma(x,5,2),type="l", add=TRUE,col="red",lwd=2)
curve(dgamma(x,10,2),type="l", add=TRUE,col="green",lwd=2)
curve(dgamma(x,1,2),type="l", add=TRUE,col="blue",lwd=2)
plot(x, pgamma(x,2,2),type="l", ylab="f(x)",main="Función de
→ distribución",col="purple",lwd=2)
curve(pgamma(x,5,2),type="l", add=TRUE,col="red",lwd=2)
curve(pgamma(x,10,2),type="l", add=TRUE,col="green",lwd=2)
curve(pgamma(x,1,2),type="l", add=TRUE,col="blue",lwd=2)
```

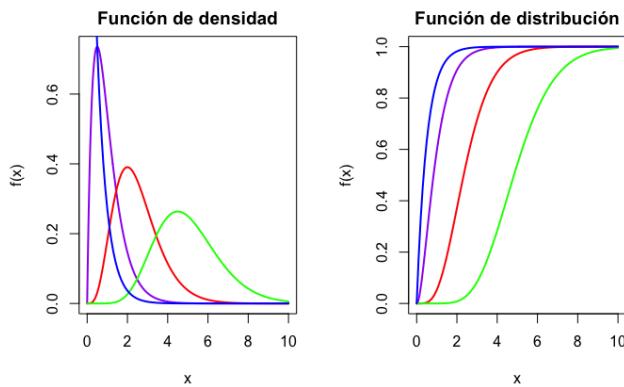


Figura 12.5: Función de densidad y de distribución de variables Gamma

El caso particular de que $\alpha = 1$ se denomina **distribución exponencial** de parámetro β .

- **Distribución χ_n^2 de Pearson**

Sean X_1, X_2, \dots, X_n v.a. independientes, todas distribuidas según una $N(0, 1)$. La suma de sus

12.4. Modelos de distribución de probabilidad

203

cuadrados sigue una distribución $\Gamma(\frac{n}{2}, 2)$, que se denomina **distribución Chi-cuadrado**:⁶

$$\sum_{i=1}^n X_i^2 \sim \chi_n^2 \equiv \Gamma\left(\frac{n}{2}, 2\right). \quad (12.13)$$

Al parámetro n se le llama **grados de libertad**. Su media es $E(X) = n$ y su varianza es $Var(X) = 2n$ y la forma funcional de su densidad y distribución son casos particulares de la Gamma. De hecho, la Fig. 12.5 corresponde a distribuciones χ_n^2 con $n = 2, 4, 10$, y 20 .

- **Distribución $t - Student$**

La distribución $t - Student$ surge, entre otros contextos, en el muestreo de poblaciones normales (véase el Sec. 13.6), asociada al uso de medias. Se dice que una v.a. X sigue una distribución $t - student$ con n **grados de libertad** si es el cociente entre una distribución normal estándar y la raíz de una χ_n^2 dividida entre sus grados de libertad, ambas independientes: $X \sim \frac{N(0,1)}{\sqrt{\chi_n^2/n}}$.

Su función de densidad viene dada por:

$$f(x) = \frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\Gamma(n/2)}(1+x^2/n)^{-(n+1)/2}, \quad (12.14)$$

con media $E(X) = 0$ y varianza $Var(X) = \frac{n}{n-2}$, siendo $n > 2$. Su densidad tiene forma acampanada, simétrica respecto a cero y parecida a la de la Normal, pero con mayor probabilidad en las colas. En la Fig. 12.6 se muestran las funciones de densidad y distribución para tres $t - Student$.⁷

```
par(mfrow=c(1,2))
x<-seq(-3, 3, 0.01)
plot(x, dt(x,df=100),type="l", ylab="f(x)",main="Función de densidad",col="blue",lwd=2)
curve(dt(x,df=10),type="l", add=TRUE,col="red",lwd=2)
curve(dt(x,df=3),type="l", add=TRUE,col="darkgreen",lwd=2)
plot(x, pt(x,df=100),type="l", ylab="f(x)",main="Función de
  distribución",col="blue",lwd=2)
curve(pt(x,df=10),type="l", add=TRUE,col="red",lwd=2)
curve(pt(x,df=3),type="l", add=TRUE,col="darkgreen",lwd=2)
```

- **Distribución F de Snedecor**

Este modelo también está asociado al muestreo sobre poblaciones normales, en este caso, a la comparación de varianzas. Se define una distribución F de Snedecor con n y m grados de

⁶Como se verá en la Sec. 13.6, esta distribución aparece en el muestreo de poblaciones normales, en concreto al trabajar con varianzas, y también resulta fundamental en el Cap. 23. Además de otras aplicaciones, estos dos casos justifican su interés.

⁷Para valores de n mayores que 30, la distribución $t - Student$ y la $N(0,1)$ prácticamente coinciden.

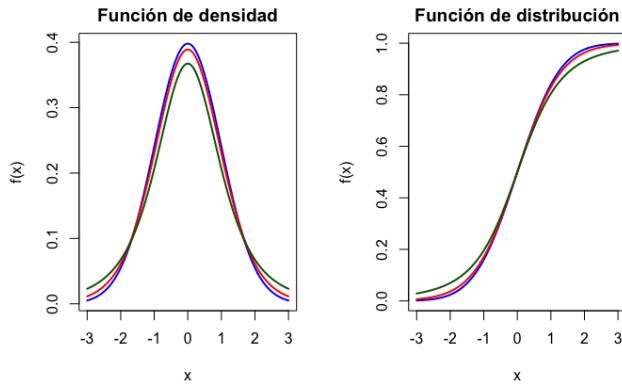


Figura 12.6: Función de densidad y de distribución de variables t-Student, con 3 (verde), 10 (rojo) y 100 (azul) grados de libertad

libertad como el cociente de dos distribuciones χ^2 independientes divididas entre sus grados de libertad, $F_{n,m} = \frac{\chi_n^2/n}{\chi_m^2/m}$.

La función de densidad $F_{n,m}$ viene dada por:

$$f(x) = \frac{\Gamma(\frac{n+m}{2})}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})} \left(\frac{n}{m}\right)^{\frac{n}{2}} \frac{x^{\frac{n-2}{2}}}{(1 + \frac{nx}{m})^{\frac{n+m}{2}}} \quad (12.15)$$

con media $E(X) = \frac{m}{m-2}$, siendo $m > 2$ y varianza $Var(X) = \frac{2m^2(n+m-2)}{n(m-2)^2(m-4)}$, cuando $m > 4$.

La gráfica de la función de densidad es parecida a la de la χ_n^2 . Así, sólo está definida para el semieje positivo y su apariencia variará según los grados de libertad. La Fig. 12.7 muestra las funciones de densidad y distribución para varias distribuciones F-Snedecor.

```
par(mfrow=c(1,2))
x<-seq(0, 4, 0.01)
plot(x, df(x,5,10),type="l", ylab="f(x)",main="Función de densidad",col="blue")
curve(df(x,10,5),type="l", add=TRUE,col="red")
curve(df(x,5,5),type="l", add=TRUE,col="darkgreen")
plot(x, pf(x,5,10),type="l", ylab="f(x)",main="Función de distribución",col="blue")
curve(pf(x,10,5),type="l", add=TRUE,col="red")
curve(pf(x,5,5),type="l", add=TRUE,col="darkgreen")
```

12.5. Teorema central del límite

A veces es difícil encontrar la distribución muestral de algunos estadísticos o estimadores, o, incluso, es imposible determinar la distribución de la variable de interés; entonces es útil aplicar algunos teoremas de convergencia, en especial el **Teorema Central del Límite (TCL)**.

12.6. Distribuciones de probabilidad en **R**

205

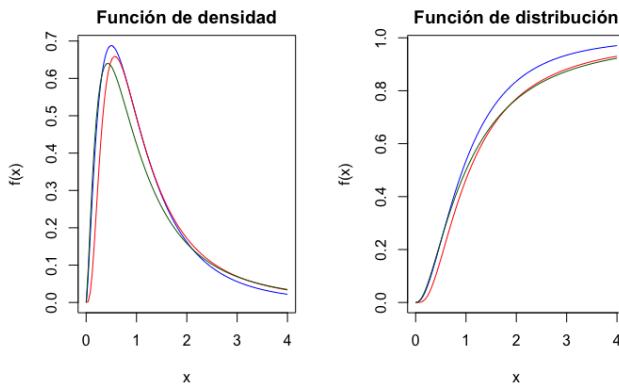


Figura 12.7: Función de densidad y de distribución de variables F-Snedecor, en azul con (5,10) grados de libertad, en rojo con (10,5) y en verde con (5,5)

El TCL permite, bajo ciertas condiciones, usar la distribución normal para aproximar otras distribuciones o para modelizar el comportamiento de variables de las que se desconozca su distribución.

Teorema central del límite: si X_1, \dots, X_n son v.a. independientes e idénticamente distribuidas (iid) con media μ y desviación típica σ , entonces $\sum_{i=1}^n X_i$ tiene asintóticamente una distribución normal de media $n\mu$ y desviación típica $\sqrt{n}\sigma$,

$$\frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma} \xrightarrow{n \rightarrow \infty} N(0, 1). \quad (12.16)$$

El TCL indica que la distribución de la suma de n variables aleatorias independientes tiende a una distribución normal, cuando n es muy grande. Es decir, aunque cada uno de los efectos sea raro o difícil de estudiar, si lo que se quiere estudiar es la suma de los mismos, se sabe que, bajo ciertas condiciones y siempre que sean independientes, ésta se comportará como una distribución normal. Así se explica el hecho constatado de que muchas distribuciones de variables observadas en la naturaleza o en experimentos físicos sean aproximadamente normales; por ejemplo, las medidas del cuerpo humano, altura, peso, longitud de los dedos, etc.

12.6. Distribuciones de probabilidad en R

En **R** están implementadas las distribuciones de probabilidad más importantes, de tal forma que, en el paquete **Rlab**, se aplica a cada nombre del modelo un determinado prefijo para calcular una función específica: *d* para la función de cuantía o densidad, *p* para la función de distribución, *q* para los cuantiles (o percentiles) y *r* para generar muestras pseudo-aleatorias.

En la Tabla 12.1 se exponen los modelos de distribución vistos, indicando el tipo, con su notación y la función utilizada en **R** para su cálculo:

Tabla 12.1: Funciones de distribución en R

Distribución	Tipo de modelo	Notación	Función en R
Binomial	discreto	$B(n,p)$	<code>binom</code>
Binomial negativa	discreto	$BN(r;p)$	<code>nbinom</code>
Geométrica	discreto	$G(p)$	<code>geom</code>
Poisson	discreto	$P(\lambda)$	<code>pois</code>
Normal	continuo	$N(\mu, \sigma)$	<code>norm</code>
Gamma	continuo	$\Gamma(\alpha, \beta)$	<code>gamma</code>
Exponencial	continuo	$Exp(\beta)$	<code>exp</code>
Chi-cuadrado	continuo	χ_n^2	<code>chisq</code>
t-student	continuo	t_n	<code>t</code>
F-Snedecor	continuo	$F_{n,m}$	<code>f</code>

A continuación se realizan dos ejemplos con R, uno para modelos discretos y otro para la normal. La adaptación a cualquier otro modelo de probabilidad consiste, básicamente, en sustituir las funciones de R.

12.6.1. Ejemplo de distribuciones discretas con R

Sea un algoritmo de identificación que trata un número muy elevado de imágenes, teniendo acreditada una tasa de error del 20% en caso de personas y del 5% para el resto de imágenes. Supóngase que las imágenes de personas son el 25% del total de imágenes.

- a) Si se analizan 10 imágenes de personas, calcúlese la probabilidad de que identifique correctamente siete.

Denominando X al número de imágenes de personas correctamente clasificadas, se tiene que $X \sim B(10, 0,8)$. Se pide $P(X = 7)$:

```
dbinom(7,size=10,prob=0.8)
#> [1] 0.2013266
```

- b) Para el resto de imágenes, calcúlese la probabilidad de que identifique correctamente como mucho 50 hasta que se produzca el segundo error.

Denominando Y al número de imágenes correctamente identificadas hasta el segundo error, se tiene que $Y \sim BN(2, 0,05)$. Se pide $P(Y \leq 50)$:

```
pnbinom(50,size=2,prob=0.05)
#> [1] 0.7405031
```

12.6. Distribuciones de probabilidad en R

207

- c) Históricamente, el número medio diario de imágenes incorrectamente clasificadas es de 7. Calcular la probabilidad de que un día seleccionado al azar clasifique incorrectamente entre 6 y 9 imágenes.

Denominando T al número de imágenes incorrectamente identificadas en un día, se tiene que $T \sim P(\lambda = 7)$. Se pide $P(6 \leq T \leq 9) = P(T \leq 9) - P(T \leq 5)$:

```
ppois(9,7,lower.tail = TRUE) - ppois(5,7,lower.tail = TRUE)
#> [1] 0.5297877
```

- d) Calcúlese la probabilidad de que en un lote de 20 imágenes del mismo tipo todas sean correctamente clasificadas.

Como no se especifica el tipo de imágenes, hay que calcular dicha probabilidad condicionada a cada grupo y utilizar el teorema de la probabilidad total: $P(\text{acuerdo}) = P(\text{acuerdo}/\text{personas}) * P(\text{personas}) + P(\text{acuerdo}/\text{otras}) * P(\text{otras})$.

```
acuerdo_personas<-dbinom(0,20,0.2)
acuerdo_otras<-dbinom(0,20,0.05)
acuerdo_total<-acuerdo_personas*0.25+acuerdo_otras*0.75
acuerdo_total
#> [1] 0.2717467
```

- e) Si se han clasificado correctamente las 20 imágenes del lote, calcúlese la probabilidad de que correspondan a imágenes de personas.

En este caso hay que utilizar el teorema de Bayes: $P(\text{personas}/\text{acuerdo}) = \frac{P(\text{acuerdo}/\text{personas}) * P(\text{personas})}{P(\text{acuerdo})}$.

```
acuerdo_personas*0.25/acuerdo_total
#> [1] 0.01060658
```

12.6.2. Ejemplo de una distribución normal con R

Las calificaciones (de 0 a 10) en un curso de estadística siguen una de distribución normal $N(6, 1.25)$, $X \sim N(6, 1.25)$. Calcúlese:

- a) La probabilidad de que una persona obtenga una calificación inferior a 4.

```
pnorm(4,mean=6,sd=1.25)
#> [1] 0.05479929
```

- b) El número esperado de personas que obtendrán sobresaliente (9 o más) en un grupo de 60 personas.

```
p<-pnorm(9,6,1.25,lower.tail=FALSE)
p
#> [1] 0.008197536
# En un grupo de 60 personas, redondeando a un número entero
round(60*p)
#> [1] 0
```

- c) la nota mínima para estar en el 30 % de personas con mejores calificaciones.

```
qnorm(0.7,6,1.25)
#> [1] 6.655501
```

- d) En un curso de informática las calificaciones siguen una de distribución normal $N(5, 1,75)$, independientes de las de estadística. Calcular la probabilidad de que una persona matriculada en ambos cursos saque mayor calificación en estadística.

Llamando $X = C_e - C_i$ a la diferencia entre las calificaciones en estadística (C_e) y en informática (C_i), ambas normales e independientes, la distribución de X será $X \sim N(6 - 5, \sqrt{1,25^2 + 1,75^2})$. Se pide $P(X > 0)$, que se representa en la Fig. 12.8.

```
pnorm(0,mean=1,sd=sqrt(1.25^2+1.75^2),lower.tail = FALSE)
#> [1] 0.6790309
```

```
media<-1
desv<-sqrt(1.25^2+1.75^2)
area_n<-function (media, desv ,lb , ub,...)
{
  x<-seq(media-4*desv, media+4*desv, 0.05)
  if (missing(lb)) {lb<-min(x)}
  if (missing(ub)) {ub<-max(x)}
  plot(x,dnorm(x, media, desv), ylab="P(x)", xlab="x", main="Probabilidad ",type="l",
    lty=1, lwd=2)
  # Nueva rejilla de valores para x2, para el área.
  x2<-seq(0, 10, 0.05)
  y<-dnorm(x2, media, desv)
  polygon(c(0, x2, 10), c(0, y, 0),col="lightblue")
}
area_n(media,desv, 0, 10)
```

12.6. Distribuciones de probabilidad en **R**

209

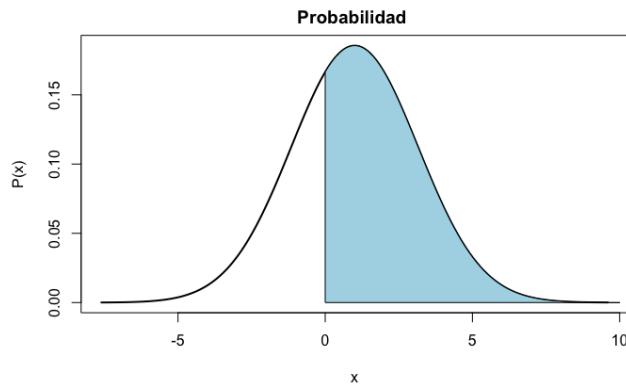


Figura 12.8: $P(X>0)$ representada como el área bajo la función de densidad de X

Resumen

La teoría de la probabilidad proporciona reglas de comportamiento que permiten la ordenación y toma de decisiones en situaciones donde prevalecen condiciones de incertidumbre.

Los modelos de distribución de probabilidad más usuales en la práctica son, para variables discretas, el binomial, el binomial negativo y el Poisson (junto a sus casos particulares). Para el caso de variables continuas, los modelos más frecuentes son el normal, el gamma, el t-Student, el Chi-cuadrado y el F-Snedecor (así como sus casos particulares). Estos modelos facilitan el cálculo de probabilidades en, prácticamente, cualquier proyecto de las ciencias sociales, técnicas y naturales.

Capítulo 13

Inferencia estadística

M^a Leticia Meseguer Santamaría^a y Manuel Vargas Vargas^a

^a Universidad de Castilla-La Mancha

13.1. Introducción

Cuando se estudian fenómenos mediante variables aleatorias, el objetivo estadístico básico es determinar cuáles son las distribuciones probabilísticas que rigen dichas variables o algunas características determinadas por ellas. Es este comportamiento aleatorio el que permite hacer predicciones con unos márgenes de error conocidos, analizar y cuantificar la relación entre variables, evaluar si hipótesis o modelos teóricos son congruentes con los datos disponibles, etc. Así, en la práctica, cuando se estudia una variable X , lo habitual es que se desconozca su distribución probabilística, $F(x)$, pero que se disponga de un conjunto de realizaciones (x_1, \dots, x_n) , también llamado muestra, valores concretos de dicha variable a partir de los cuales “aproximar” la distribución desconocida.

La inferencia estadística proporciona las herramientas y técnicas que permiten, a partir de la información muestral, extrapolar resultados a la distribución poblacional con márgenes de error conocidos. Un primer objetivo (más detallado en el Cap. 14) es analizar qué condiciones debe cumplir la muestra para que su información sea válida y extrapolable a toda la población (es la conocida como **teoría de muestreo**). Un segundo objetivo es establecer los mecanismos que permitan dicha extrapolación manteniendo controlados los errores de muestreo.

Es habitual que se conozca (o se asuma) que la distribución poblacional $F(x)$ pertenezca a alguna familia paramétrica, es decir, que se asuma su forma funcional pero que dependa de algunos parámetros (lo más frecuente es que se asuma la normalidad, pero podría ser cualquiera de los modelos paramétricos existentes). Se habla entonces de **inferencia paramétrica**, ya que se usa la información muestral para determinar los “mejores” valores (bajo algún criterio) de los parámetros que rigen la distribución poblacional, existiendo tres planteamientos básicos: estimación puntual (Sec. 13.3), por intervalo (Sec. 13.4) y contraste de hipótesis (Sec. 13.5).

También hay situaciones en las que la forma funcional de la distribución poblacional es desconocida, o se duda de que la familia paramétrica considerada sea adecuada. En estos casos, bajo el nombre genérico de **inferencia no paramétrica**, se plantean contrastes que buscan determinar cuándo es posible asumir un modelo concreto de distribución, entre los que destacan, por su frecuente uso, los contrastes de normalidad (Sec. 13.8). Otra alternativa que permite aproximar características poblacionales sin asumir ninguna distribución poblacional concreta es el **remuestreo**, fundamentalmente el denominado “*bootstrap*” (se aborda en el Cap. 14).

13.2. Muestreo aleatorio simple

Al estudiar una variable poblacional, X , de la que se desconoce su distribución, llamada **distribución poblacional**, $F(x)$, se utiliza la información suministrada por una **muestra** obtenida por algún método de muestreo probabilístico que garantice que sea representativa de la variable poblacional.

En la mayoría de los casos y técnicas estadísticas se asume que la muestra está obtenida mediante el método básico de muestreo, conocido como **muestreo aleatorio simple**, consistente en seleccionar totalmente al azar y con reemplazo a los individuos de la muestra, por lo que todos tienen la misma probabilidad de formar parte de ella. De esta forma, dada una distribución poblacional $F(x)$, una muestra aleatoria simple (m.a.s.) es una realización de un conjunto de n variables aleatorias independientes e idénticamente distribuidas $X = (X_1, \dots, X_n)$, denominadas **variables muestrales** y cuya **distribución conjunta** es de la forma:

$$F(X_1, \dots, X_n) = F_{X_1}(x_1) \dots F_{X_n}(x_n) = F(x_1) \dots F(x_n). \quad (13.1)$$

Una herramienta básica para la inferencia es la **distribución empírica de la muestra**, definida como:

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{(-\infty, x]}(X_i), \quad (13.2)$$

donde $\mathbb{I}_{(-\infty, x]}(X_i)$ es una función indicadora que toma el valor 1 si $X_i \leq x$ y 0 en caso contrario.¹

La gran ventaja del muestreo aleatorio simple consiste en que, dada una m.a.s. (X_1, \dots, X_n) :

$$\lim_{n \rightarrow \infty} E \left[\left(\hat{F}_n(x) - F(x) \right)^2 \right] = 0, \quad (13.3)$$

expresión conocida como *teorema de Glivenko-Cantelli*. Este resultado es fundamental en inferencia, pues garantiza que el muestreo aleatorio simple produce muestras representativas de la población, ya que, a medida que aumenta el tamaño muestral, la distribución empírica de la

¹Es decir, la distribución empírica de la muestra indica, para cada valor x , la proporción de elementos de la muestra que toman un valor menor o igual que él.

13.2. Muestreo aleatorio simple

213

muestra se aproxima cada vez más a la distribución poblacional (véase Fig. 13.1).² Así, cualquier característica (media, varianza...) de una distribución poblacional puede ser aproximada por su equivalente en la distribución empírica.

```
par(mfrow = c(1, 3))
set.seed(196)
x1 <- rnorm(20)
plot.ecdf(x1, main = "n=20")
curve(pnorm, add = TRUE, col = "red")
x2 <- rnorm(50)
plot.ecdf(x2, main = "n=50")
curve(pnorm, add = TRUE, col = "red")
x3 <- rnorm(200)
plot.ecdf(x3, main = "n=200")
curve(pnorm, add = TRUE, col = "red")
```

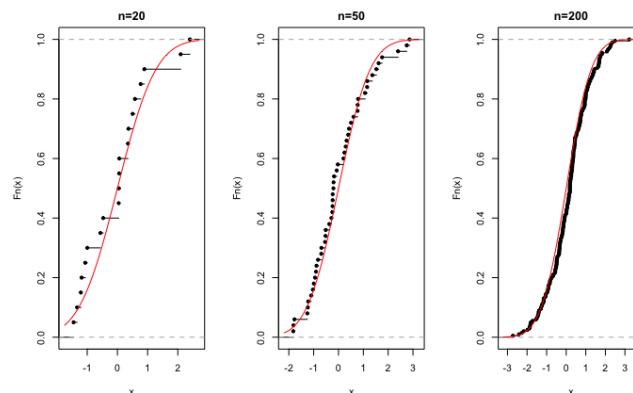


Figura 13.1: Distribución empírica para muestras de diferente tamaño de una distribución Normal

Es muy frecuente que, a efectos de inferencia, no se estudie el comportamiento aleatorio de toda la muestra (su distribución conjunta) sino que interese el comportamiento de una función de la muestra que no dependa de ningún valor desconocido, $T(X) = T(X_1, \dots, X_n)$, llamada genéricamente **estadístico muestral**; dicho comportamiento vendrá determinado por la **distribución en el muestreo** del estadístico $T(X)$. El hecho de utilizar una m.a.s. permite establecer resultados de interés sobre los estadísticos o, en algunos casos, incluso obtener la distribución en el muestreo exacta de los estadísticos más usuales (Sec. 13.6).

Así, dadas una variable poblacional X con varianza finita y una m.a.s., se define la **media muestral** (aleatoria) como:

$$\bar{X} = \frac{X_1 + \dots + X_n}{n}. \quad (13.4)$$

²Un tema que se abordará en el Cap. 14 es la determinación del tamaño muestral necesario para que la aproximación tenga un error menor que uno prefijado.

El hecho de utilizar una m.a.s. garantiza que:

$$E[\bar{X}] = E[X] ; \text{Var}(\bar{X}) = \frac{\text{Var}(X)}{n}. \quad (13.5)$$

Este resultado es muy útil, ya que indica que la variabilidad de la media muestral es más pequeña que la variabilidad de la variable poblacional, siendo inversamente proporcional al tamaño muestral.

Otro estadístico muy utilizado es la **varianza muestral**,³ que se define como:

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}. \quad (13.6)$$

En este caso, su esperanza es:

$$E[S^2] = \frac{n-1}{n} \text{Var}[X], \quad (13.7)$$

que no coincide con la varianza poblacional. Para evitar este hecho, se define la **cuasivarianza muestral** (aleatoria):

$$S_c^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}, \quad (13.8)$$

estadístico para el que sí se cumple que $E[S_c^2] = \text{Var}[X]$, ya que existe una relación de proporcionalidad entre ambos estadísticos $nS^2 = (n-1)S_c^2$.⁴

13.3. Estimación puntual

Sea una población caracterizada por una distribución poblacional, $F(x, \theta)$, de una familia paramétrica de la que se desconoce el valor del parámetro $\theta \in \Theta$, donde Θ es el espacio paramétrico (conjunto de posibles valores de θ). Dada una m.a.s. $X = (X_1, \dots, X_n)$, se considera como *estimador* de θ a un estadístico muestral cuyo resultado sea un posible valor del parámetro:

$$\hat{\theta} = T(X) = T(X_1, \dots, X_n) \in \Theta. \quad (13.9)$$

La siguiente expresión corresponde al **error cuadrático medio** de un estimador:

$$ECM_\theta(\hat{\theta}) = E_\theta \left[(\hat{\theta} - \theta)^2 \right], \quad (13.10)$$

³que, como estadístico, también es aleatoria

⁴Muchos textos, sobre todo anglosajones, no hacen esta distinción, sino que denominan directamente “varianza muestral” a la cuasivarianza. En R, por ejemplo, las funciones `var()` o `sd()` proporcionan la cuasivarianza y cuasidesviación típica muestrales respectivamente, matiz que hay que tener siempre presente.

que proporciona un valor medio del error que se comete al “aproximar” el verdadero valor θ por el resultado del estimador $\hat{\theta}$. Así, el criterio de “mínimos cuadrados” propone utilizar el estimador que minimiza el error cuadrático medio:

$$\hat{\theta}_{MC} = \min_{\hat{\theta}} E_{\theta} \left[(\hat{\theta} - \theta)^2 \right]. \quad (13.11)$$

Desarrollando la expresión del ECM (13.10), éste se puede re-expresar como

$$ECM_{\theta}(\hat{\theta}) = Var_{\theta}(\hat{\theta}) + (E_{\theta}(\hat{\theta}) - \theta)^2 = Var_{\theta}(\hat{\theta}) + b_{\theta}^2(\hat{\theta}), \quad (13.12)$$

donde $b_{\theta}(\hat{\theta}) = (E_{\theta}(\hat{\theta}) - \theta)$ se conoce como **sesgo** del estimador (*bias*, en inglés). Así, el ECM de un estimador depende de su varianza y de su sesgo al cuadrado.

Por tanto, la determinación del “mejor” estimador, bajo el criterio de mínimos cuadrados, se puede llevar a cabo en dos pasos:

- Seleccionar estimadores “inesgados”, es decir, de sesgo cero, o sea, $E(\hat{\theta}) = \theta$ (el valor medio del estimador coincide con el parámetro).
- De entre los estimadores inesgados, seleccionar el de varianza mínima, $Var(\hat{\theta}_{MC}) = \min_{\hat{\theta}} Var(\hat{\theta})$.

Queda fuera del objetivo de este capítulo plantear la obtención del estimador de mínimos cuadrados para cualquier distribución poblacional y parámetros, que el lector interesado puede encontrar en cualquier texto teórico de inferencia estadística ([Casella and Berger \(2007\)](#), [Blais \(2020\)](#), [Almudevar \(2021\)](#)).

Otro planteamiento para encontrar estimadores puntuales se basa en la función de densidad conjunta de la muestra, que depende de ésta y del parámetro que caracteriza a la distribución poblacional:

$$f(x_1, \dots, x_n; \theta) = f(x_1; \theta) \dots f(x_n; \theta) = L(\theta; x_1, \dots, x_n). \quad (13.13)$$

Considerando el parámetro como fijo, la función se interpreta como la densidad de probabilidad de la muestra. Sin embargo, si se considera que la muestra está dada, entonces se puede interpretar como una función del parámetro que mide la **verosimilitud** (*likelihood*, en inglés) de cada valor del parámetro en función de la muestra obtenida. Así, el criterio para determinar el “mejor” estimador puede ser seleccionar aquél que maximiza la función de verosimilitud; se obtiene entonces el conocido como **estimador máximo verosímil**:

$$\hat{\theta}_{MV} = \max_{\theta} L(\theta; x_1, \dots, x_n). \quad (13.14)$$

Para el cálculo del estimador máximo verosímil no se suele utilizar la función de verosimilitud, sino su logaritmo (que alcanza los máximos y mínimos en los mismos puntos), derivando respecto al parámetro e igualando a cero (ecuación de verosimilitud).

Este método suele proporcionar estimadores con buenas propiedades estadísticas y, en muchos casos, suele conducir al mismo resultado que el método de mínimos cuadrados.⁵ En las distribuciones usuales, es relativamente sencillo obtener la ecuación de verosimilitud y resolverla, por lo que se dispone de estimadores máximo verosímiles conocidos. En modelos más elaborados, la resolución de la ecuación de verosimilitud se puede complicar, hasta el extremo de que haya que recurrir a métodos numéricos de aproximación.

Una alternativa computacionalmente más sencilla es la basada en el conocido como **método de los momentos**. El planteamiento básico es expresar el parámetro en función de los momentos poblacionales (esperanza, varianza, etc.) y utilizar como estimador la misma función pero de los momentos muestrales (media muestral, varianza muestral, etc.). En las distribuciones más usuales, los parámetros suelen ser momentos poblacionales o transformadas simples de éstos, por lo que el método de los momentos es muy sencillo. Como contrapartida, es más difícil evaluar las propiedades estadísticas de estos estimadores, salvo que coincidan con los de mínimos cuadrados o de máxima verosimilitud.

En **R**, el paquete **fdistrplus** dispone de la función **fitdist()**, que permite la obtención de los estimadores para las distribuciones usuales por diversos métodos, incluidos el de máxima verosimilitud (**mle**) y el de los momentos (**mme**).

13.4. Estimación por intervalos

Dado que todo estimador es una variable aleatoria, su valor concreto, la “*estimación*” del parámetro $\hat{\theta}$, depende de la muestra. Esta variación muestral ocasiona incertidumbre sobre la estimación. Una forma de incluir esta variabilidad en la estimación puede consistir en sustituir la estimación puntual por un intervalo de valores en el que se tenga un cierto nivel de confianza de que contenga al verdadero valor del parámetro.

El método más extendido para obtener **intervalos de confianza** consiste en utilizar un estimador puntual y su distribución en el muestreo para construir un intervalo que contenga, con cierta probabilidad $(1 - \alpha)$, el verdadero valor θ :

$$IC_{(1-\alpha)} = [LIC, LSC] \text{ tal que } P(LIC \leq \theta \leq LSC) = (1 - \alpha), \quad (13.15)$$

donde los límites inferior (LIC) y superior (LSC) de confianza, denominados **valores críticos**, dependen de la desviación típica del estimador y de constantes asociadas a su distribución y al nivel de confianza $(1 - \alpha)$. En esta ecuación, tanto el LIC como el LSC son variables aleatorias; cuando se utilizan los datos de una muestra, se convierten en valores reales, por lo que no se puede hablar de “*probabilidad de que el parámetro esté dentro del intervalo*”, sino que se habla de “*confianza en que el intervalo contenga el valor del parámetro*”.

⁵En las distribuciones usuales es así, salvo que el estimador de máxima verosimilitud sea sesgado, como es el caso de estimar la varianza en una distibución normal.

En **R**, el paquete **Rlab** permite obtener los valores críticos de las distribuciones usuales a través de los cuantiles, anteponiendo q al nombre de la distribución (véase la Tabla 12.1 “Funciones de distribución en **R**”); por ejemplo, usando las funciones `qbinom()`, `qnorm()`, `qt()`, `qf()`, etc.. Igualmente, el paquete **DescTools** dispone de funciones para calcular intervalos de confianza en poblaciones normales para la media (`MeanCI()`), la diferencia de medias (`MeanDiffCI()`), la mediana (`MedianCI()`), cualquier cuantil (`QuantileCI()`) o la varianza (`VarCI()`). Por último, en el caso de no conocer la distribución en el muestreo del estimador, se puede recurrir al remuestreo por *bootstrap*, que se detallará en el Cap. 14, indicando el método `boot` en las funciones anteriores.

13.5. Contrastes de hipótesis

Hay situaciones donde no interesa tanto estimar el valor de un parámetro sino decidir si la información muestral es congruente con algún valor concreto del parámetro. En estos casos, se puede establecer como **hipótesis** que el parámetro toma un valor concreto y *contrastar* si es verosímil haber obtenido el resultado muestral dado. Este planteamiento se conoce como **contrastes de significación**.

Así, se establece una hipótesis, históricamente conocida como **hipótesis nula**, que determina un valor del parámetro:

$$H_0 \equiv \theta = \theta_0. \quad (13.16)$$

Suponiendo cierta la hipótesis nula, la distribución muestral del estimador permite obtener la probabilidad de observar un valor del estimador más “*distante*” del valor del parámetro fijado en la hipótesis nula que el obtenido en la muestra, probabilidad conocida como **p-valor**: si es muy pequeño, es muy poco probable que se observe el valor obtenido en la muestra cuando la hipótesis es cierta, por lo que la evidencia empírica no es congruente con ella; si no es pequeño, dicho valor es probable que se observe (bajo la hipótesis nula), por lo que no habría evidencia empírica “*en contra*” de ella.

Se habla de **p-valor bilateral** o “a dos colas” cuando la distancia se considera tanto por la derecha como por la izquierda de la distribución del estimador bajo la hipótesis nula. En caso de que se considere sólo por la izquierda o por la derecha, se habla de **p-valor unilateral** (a la izquierda o a la derecha, respectivamente) o “a una cola”. La comparación (distancia) entre el valor del parámetro establecido en la hipótesis nula y el del estimador de dicho parámetro puede llevarse a cabo por diferencia (tal es el caso del contraste de medias) o por cociente (caso de los contrastes de varianzas).

Habitualmente, se considera que un p-valor por debajo de 0.05 ya indica que la evidencia empírica no permite asumir como cierta la hipótesis nula, expresándose como que el valor del parámetro es “*significativamente distinto (menor o mayor)*” que θ_0 . También es posible interpretar el p-valor como “*la probabilidad máxima de cometer el error de rechazar la hipótesis nula cuando es cierta*”, abreviado como “*tamaño del error si se rechaza la hipótesis nula*”.

Estos contrastes de significación, originalmente desarrollados por Ronald Fisher, fueron incluidos en un esquema de toma de decisiones por Jerzy Neyman y Egon Pearson, planteando que,

de no ser cierta la hipótesis nula, se debe plantear una hipótesis alternativa H_1 . La decisión de qué hipótesis resulta más congruente con los datos se basa en la comparación por cociente de las verosimilitudes de la muestra bajo cada una de ellas, decidiendo el rechazo de la hipótesis nula a favor de la alternativa cuando dicho cociente es, en probabilidad, inferior a un valor prefijado, α , conocido como **nivel de significación**. Dependiendo de la estructura de las hipótesis (simples, si sólo determinan un valor del parámetro, o compuestas, si determinan más de uno; a su vez, unilaterales si los valores son todos menores, o mayores, que uno dado, o bilaterales en caso contrario) la regla de decisión del contraste resulta más o menos compleja de obtener.

Cuando se adopta el planteamiento decisional de Neyman-Pearson, el nivel de significación permite evaluar la probabilidad de rechazar la hipótesis nula cuando es cierta (conocida también como **probabilidad de error de tipo I**, α), pero también la probabilidad de aceptar como cierta H_0 cuando es más correcta H_1 (denominada **probabilidad de error de tipo II**, β) o, equivalentemente, su complementario: la probabilidad de rechazar H_0 cuando H_1 es más correcta, probabilidad conocida como **potencia del contraste**, $(1 - \beta)$. Si la hipótesis alternativa es simple, es posible evaluar la potencia, por lo que se tiene una medida probabilística de la magnitud de ambos errores (de tipo I y de tipo II), lo cual permite una valoración completa del resultado de la regla de decisión (contraste de hipótesis). Sin embargo, si la hipótesis H_1 es compuesta, la magnitud de la potencia es una función evaluada en el rango de valores que establezca dicha hipótesis. En este caso, se dispone de una medida probabilística del error de tipo I pero no del error de tipo II, puesto que depende de valores concretos del parámetro que no son especificados en la hipótesis alternativa, H_1 .

Computacionalmente, dada la información muestral, es más fácil calcular el p-valor que plantear el esquema de decisión de Neyman-Pearson, por lo que es la estrategia utilizada en la práctica.

Dado el carácter breve e introductorio de este capítulo, no se profundizará más en este esquema de decisión, que puede consultarse, por ejemplo, en Casella and Berger (2007), Blais (2020) o Almudevar (2021), entre otros muchos.

13.6. Inferencia estadística paramétrica sobre poblaciones normales

Como consecuencia del teorema central del límite (Sec. 12.5), el supuesto de que la distribución poblacional es una normal es el caso más habitual en la práctica, siendo requisito básico en muchísimas técnicas estadísticas. En este caso, las distribuciones muestrales de los estimadores de los parámetros poblacionales, tanto de la media μ como de la varianza σ^2 , son conocidas, lo que facilita la construcción de intervalos de confianza y contrastes de hipótesis.

Así, dada una distribución poblacional normal y una m.a.s. de tamaño n ,

- Para estimar la varianza poblacional, σ^2 , el estimador máximo verosímil es la varianza muestral (13.6), que es sesgado. El estimador insesgado es la cuasivarianza muestral (13.8). Sus distribuciones en el muestreo, en el caso habitual de que la media poblacional sea desconocida, son:

13.6. Inferencia estadística paramétrica sobre poblaciones normales

219

$$\frac{nS^2}{\sigma^2} = \frac{(n-1)S_c^2}{\sigma^2} \sim \chi_{n-1}^2. \quad (13.17)$$

Así, el intervalo de confianza a nivel $(1 - \alpha)$ es:

$$IC_{(1-\alpha)} = \left[\frac{ns^2}{\chi_{n-1,\alpha/2}^2}, \frac{ns^2}{\chi_{n-1,1-\alpha/2}^2} \right], \quad (13.18)$$

o, equivalentemente, usando la proporcionalidad entre varianza y cuasivarianza muestrales:

$$IC_{(1-\alpha)} = \left[\frac{(n-1)s_c^2}{\chi_{n-1,\alpha/2}^2}, \frac{(n-1)s_c^2}{\chi_{n-1,1-\alpha/2}^2} \right], \quad (13.19)$$

donde $\chi_{n-1,\alpha/2}^2$ representa el cuantil en la distribución.⁶

Para el contraste de $H_0 \equiv \sigma^2 = \sigma_0^2$, la “distancia” es $\frac{ns^2}{\sigma_0^2} = \frac{(n-1)s_c^2}{\sigma_0^2}$, lo que permite calcular los p-valores mediante una distribución χ_{n-1}^2 .⁷

Para el caso de querer estimar la desviación típica, basta con calcular la raíz cuadrada del estimador de la varianza, o si se busca un intervalo de confianza, la raíz de los extremos del intervalo para la varianza. Los contrastes de hipótesis son equivalentes, ya que $\sigma^2 = \sigma_0^2 \equiv \sigma = \sigma_0$.

- Para estimar el parámetro μ se utiliza el estimador media muestral $\hat{\mu} = \bar{X}$, en el que coinciden los métodos de mínimos cuadrados, de máxima verosimilitud y de los momentos, siendo insesgado y de varianza mínima.

Si la varianza poblacional es conocida, la distribución en el muestreo de la media muestral es:

$$\bar{X} \sim N \left(\mu, \frac{\sigma}{\sqrt{n}} \right) \equiv \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1). \quad (13.20)$$

El intervalo de confianza a nivel $(1 - \alpha)$ es:

$$IC_{(1-\alpha)} = \left[\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right], \quad (13.21)$$

donde $z_{\alpha/2}$ representa el cuantil en una distribución normal estándar.

⁶Nótese que en los IC se utilizan los valores observados del estimador, s^2 o S_c^2 , según se haya utilizado como estimador la varianza o la cuasi-varianza. Lo mismo ocurre en los demás intervalos

⁷Nótese que la “distancia” no involucra sólo al valor observado del estimador y el valor del parámetro bajo la hipótesis nula, sino también una constante (en este caso n o $n - 1$). Ello se hace porque así el valor de esta “distancia” puede compararse directamente con el facilitado por las tablas de la distribución probabilística correspondiente (en este caso una Chi-cuadrado).

Para el contraste de $H_0 \equiv \mu = \mu_0$, la “distancia” es $\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$, lo que permite calcular los p-valores, directamente, mediante una distribución $N(0, 1)$.

Si la varianza poblacional es desconocida, se sustituye por su estimación, por lo que la distribución en el muestreo de la media muestral es:

$$\frac{\bar{X} - \mu}{S/\sqrt{n-1}} \equiv \frac{\bar{X} - \mu}{S_c/\sqrt{n}} \sim t_{n-1}. \quad (13.22)$$

El intervalo de confianza a nivel $(1 - \alpha)$ es:

$$IC_{(1-\alpha)} = \left[\bar{x} - t_{n-1,\alpha/2} \frac{s}{\sqrt{n-1}}, \bar{x} + t_{n-1,\alpha/2} \frac{s}{\sqrt{n-1}} \right], \quad (13.23)$$

o, equivalentemente:

$$IC_{(1-\alpha)} = \left[\bar{x} - t_{n-1,\alpha/2} \frac{s_c}{\sqrt{n}}, \bar{x} + t_{n-1,\alpha/2} \frac{s_c}{\sqrt{n}} \right], \quad (13.24)$$

donde $t_{n-1,\alpha/2}$ representa el cuantil de la distribución *t-Student*.

Para el contraste de $H_0 \equiv \mu = \mu_0$, la “distancia” es $\frac{\bar{X} - \mu_0}{S/\sqrt{n-1}} = \frac{\bar{X} - \mu_0}{S_c/\sqrt{n}}$, lo que permite calcular los p-valores mediante una distribución t_{n-1} .

A continuación, el interés se centra en la comparación de dos poblaciones normales independientes, X e Y, a partir de muestras (X_1, \dots, X_n) y (Y_1, \dots, Y_m) :

- Para la comparación de las varianzas poblacionales (una es mayor que la otra, o al revés; y que se lleva a cabo mediante el cociente de las correspondientes varianzas o cuasivarianzas muestrales), se tiene que:

$$\frac{\frac{mS_Y^2}{(m-1)\sigma_Y^2}}{\frac{nS_X^2}{(n-1)\sigma_X^2}} \equiv \frac{S_{cY}^2/\sigma_Y^2}{S_{cX}^2/\sigma_X^2} \sim F_{m-1, n-1}, \quad (13.25)$$

lo cual permite calcular intervalos de confianza de forma idéntica a la expuesta a los casos anteriores pero con la distribución *F*. Un caso muy frecuente es querer contrastar si ambas varianzas poblacionales son iguales (el cociente entre ellas es la unidad).

- Para la comparación de las medias poblacionales (que se lleva a cabo mediante la diferencia de las correspondientes medias muestrales), el caso más común es asumir que las varianzas (aunque desconocidas) son iguales, por lo que el estimador es:

$$\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{nS_{cX}^2 + mS_{cY}^2}{n+m-2}} \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2}, \quad (13.26)$$

si se utiliza la varianza muestral como estimador de su homónima poblacional, o:

$$\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{(n-1)S_{cX}^2 + (m-1)S_{cY}^2}{n+m-2}} \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2}, \quad (13.27)$$

si se utiliza la cuasivarianza muestral.

Al utilizarse distribuciones t-Student, los intervalos de confianza y contrastes de hipótesis son similares a los del caso de una única población con las correcciones pertinentes.

13.7. Inferencia sobre poblaciones normales con R

Los datos sobre calidad del aire en la ciudad de Nueva York (`airquality`) incluyen la variable `Wind`, que recoge, en mph, la velocidad del viento entre el día 1 de mayo y el 30 de septiembre DE 1973. Los datos de dicha variable se dividen en dos variables X = Velocidad del viento hasta el 15 de julio e Y = Velocidad del viento desde el 16 de julio. Asumiendo que las distribuciones poblacionales son normales, se propone:

- a) Obtener una estimación de la velocidad media y de la desviación típica de ambas variables, usando el método de máxima verosimilitud.

```
library('Rlab')
library('fitdistrplus')
x <- airquality$Wind[1:76]
y <- airquality$Wind[77:153] # Se partitiona la muestra en los dos períodos
mle_x <- fitdist(x, distr = "norm", method = "mle")
mle_y <- fitdist(y, distr = "norm", method = "mle")
mle_x
#> Fitting of the distribution ' norm ' by maximum likelihood
#> Parameters:
#>     estimate Std. Error
#> mean 10.640789 0.4274723
#> sd    3.726618 0.3022685
mle_y
#> Fitting of the distribution ' norm ' by maximum likelihood
#> Parameters:
#>     estimate Std. Error
#> mean 9.283117 0.3581657
#> sd   3.142891 0.2532613
```

El resultado muestra las estimaciones de la media y la desviación típica de la velocidad del viento (junto al “*error standar*” o desviación típica de los estimadores respectivos) para las dos variables anteriormente creadas.

La orden `plot(mle_x)` permite visualizar la congruencia entre la muestra y la distribución probabilística basada en las estimaciones realizadas; por ejemplo, optando por el primer período, se obtiene la Fig.13.2, que representa el histograma de los valores muestrales junto a la distribución teórica construida con las estimaciones.

```
plot(mle_x)
```

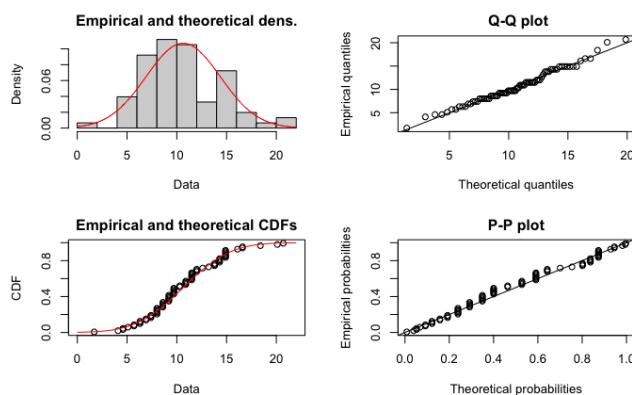


Figura 13.2: Resultados gráficos de la estimación por máxima verosimilitud

- b) Construir un intervalo de confianza para la velocidad media del viento hasta el 15 de julio, con un nivel de confianza del 95 %.

```
library('DescTools')
MeanCI(x, conf.level = 0.95)
#>      mean     lwr.ci     upr.ci
#> 10.640789  9.783563 11.498016
```

- c) Calcular un intervalo de confianza para la desviación típica de la velocidad del viento desde el 16 de julio, con un nivel de confianza del 90 %.

```
sqrt(VarCI(y, conf.level = 0.9))
#>      var    lwr.ci    upr.ci
#> 3.163501 2.795147 3.655468
```

- d) ¿Se puede considerar que las varianzas poblacionales en ambos períodos son iguales, con un nivel de significación del 1 %?

```
var.test(x, y, conf.level = 0.99, alternative = "two.sided")
#>
#> F test to compare two variances
#>
#> data: x and y
#> F = 1.4062, num df = 75, denom df = 76, p-value = 0.1406
#> alternative hypothesis: true ratio of variances is not equal to 1
#> 99 percent confidence interval:
#> 0.7727136 2.5616496
#> sample estimates:
#> ratio of variances
#> 1.406197
```

El estadístico *F-Snedecor* de contraste arroja un valor de 1.4062, con un p-valor de 0.1406. Como este p-valor no es pequeño (es superior al nivel de significación prefijado), no hay suficiente evidencia empírica como para rechazar la hipótesis nula de igualdad de varianzas.

- e) Teniendo en cuenta los resultados del apartado anterior, ¿se puede afirmar que la velocidad media del viento en el primer período es mayor que la del segundo, con un nivel de significación del 1%?

```
t.test(x, y, conf.level = 0.99, alternative = "greater", var.equal = TRUE)
#>
#> Two Sample t-test
#>
#> data: x and y
#> t = 2.4212, df = 151, p-value = 0.008328
#> alternative hypothesis: true difference in means is greater than 0
#> 99 percent confidence interval:
#> 0.03918338 Inf
#> sample estimates:
#> mean of x mean of y
#> 10.640789 9.283117
```

Un p-valor tan bajo (0.008, inferior al nivel de significación prefijado) indica que existe suficiente evidencia empírica como para rechazar la hipótesis nula de igualdad de medias; en otros términos, la evidencia empírica no es suficiente para rechazar, con un nivel de confianza del 99 %, que la velocidad media del viento en el primer período es superior a la del segundo.

13.8. Inferencia estadística no paramétrica: contrastes de normalidad

Hasta ahora, se ha supuesto que la distribución muestral del estimador era “funcionalmente” conocida, aunque dependiente de un parámetro (o varios). Sin embargo, hay situaciones donde

no se conoce cómo se distribuyen los datos, debiendo decidir qué distribución los ha generado. Es lo que se conoce como **inferencia estadística no paramétrica**. En este capítulo no se aborda un planteamiento sistemático de esta rama, sino que se presenta la situación más habitual en la práctica, que es decidir si se puede mantener que una muestra proviene de una distribución normal, supuesto básico en muchas técnicas estadísticas.

Possiblemente el test más potente para contrastar la normalidad sea la prueba de Shapiro-Wilks, que asume como hipótesis nula que los datos están generados por una distribución normal. Un rechazo de esta hipótesis (p-valor muy bajo) debería hacer reflexionar sobre la adecuación de muchas técnicas y la interpretación de los resultados. En **R**, la función `shapiro.test()` proporciona dicho contraste de normalidad.

Una alternativa es el uso del test de Kolmogorov-Smirnov, diseñado para comparar las distribuciones de dos muestras, fijando que una de ellas sea la distribución normal (este test puede ser igualmente utilizado para cualquier otra distribución usual). La función `ks.test()` permite en **R** obtener los resultados de este contraste.

Para ilustrar el uso del test de Shapiro-Wilk en **R** se recurre de nuevo a los datos sobre calidad del aire en la ciudad de Nueva York del ejemplo anterior (Sec. 13.7) y se contrasta si se puede asumir que las variables `Temp` y `Wind` están generadas por distribuciones normales:

```
shapiro.test(airquality$Temp)
#>
#> Shapiro-Wilk normality test
#>
#> data: airquality$Temp
#> W = 0.97617, p-value = 0.009319
shapiro.test(airquality$Wind)
#>
#> Shapiro-Wilk normality test
#>
#> data: airquality$Wind
#> W = 0.98575, p-value = 0.1178
```

Para la variable `Temp` el p-valor (0,0093) es muy bajo en comparación con los niveles de significación habituales (0,01, 0,05), por lo que hay suficiente evidencia empírica como para rechazar que dicha variable tenga una distribución normal. Por el contrario, en el caso de la variable `Wind`, el p-valor (0,1178) no es pequeño, por lo que no hay suficiente evidencia como para rechazar que esté generada por una distribución normal. La Fig. 13.3 muestra la comparación entre los cuantiles empíricos de ambas variables y los teóricos de una distribución normal.

```
par(mfrow = c(1, 2))
qqnorm(airquality$Temp, main = "Normal Q-Q Plot for Temp")
qqnorm(airquality$Wind, main = "Normal Q-Q Plot for Wind")
```

13.8. Inferencia estadística no paramétrica: contrastes de normalidad

225

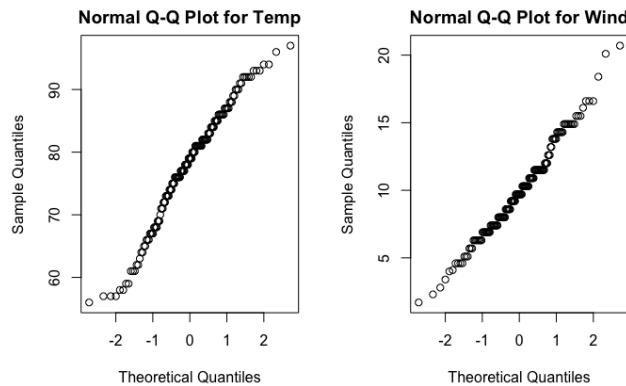


Figura 13.3: Q-Q Plots normales para las variables Temp (izq) y Wind (der)

Resumen

La inferencia estadística permite estimar la distribución poblacional de una variable a partir de la información suministrada por una muestra. Se abordan los métodos de estimación puntual de los principales parámetros poblacionales y la construcción de intervalos de confianza para ellos, así se implementan y resuelven una serie de contrastes de significación sobre diversas hipótesis.

Para el caso de poblaciones normales, se desarrollan las expresiones operativas de los métodos anteriores. Igualmente, en el ámbito de la inferencia no paramétrica, se presenta un contraste de normalidad que permite decidir cuándo el supuesto de normalidad es adecuado o no.

Capítulo 14

Muestreo y remuestreo

M^a Leticia Meseguer Santamaría^a y Manuel Vargas Vargas^a

^a Universidad de Castilla-La Mancha

14.1. Introducción al muestreo

Muchas investigaciones científicas abordan el estudio de características de un conjunto de elementos, que se denomina “*población*”. Sin embargo, no siempre es posible estudiar la totalidad del colectivo, por problemas de accesibilidad, confidencialidad, costes económicos o temporales, etc. En estos casos, se recurre a extraer un subconjunto de la población que se pretende que sea representativo de ésta respecto a las características estudiadas. El objetivo es obtener información relevante que pueda extrapolarse al total de la población, proceso denominado genéricamente “muestreo”.

En otros casos, la conveniencia de muestrear una población está relacionada con la credibilidad de los resultados obtenidos o de las propiedades de los modelos utilizados. Así, muchas técnicas cuantitativas dividen la información (muestra) disponible en un subconjunto de “*entrenamiento*” o “*estimación*” y otro de “*contraste*” o “*validación*”; una elección inadecuada puede alterar los resultados e invalidar las conclusiones. Por último, caso muy común en ciencias sociales, no se puede acceder a la medición directa de los fenómenos (por ser una población muy grande o un fenómeno subjetivo), por lo que se accede a ella mediante encuestas, que precisan de una metodología de muestreo rigurosa y diseñada previamente.

Para fijar terminología, se define **población** como el conjunto de casos de interés a los que se quiere generalizar los resultados de la investigación, denominados genéricamente **individuos**; a veces, sólo es posible acceder a una parte de la población, por lo que se utiliza el término **población objetivo** para el conjunto total y **población muestrable** al conjunto al que se tiene acceso. Por coherencia, cuando ambos colectivos no coincidan, los resultados deben extrapolarse sólo al último de ellos.

Las características de interés de la población se consideran “*variables*”, en el sentido estadístico del término, por lo que son estudiadas mediante su distribución. En algunos casos, ésta será completamente desconocida, siendo de interés su determinación completa; en otros muchos, se buscará determinar sólo algunos aspectos (medidas de posición, dispersión, etc.) o los parámetros que rigen una distribución funcionalmente conocida. En todo caso, el objetivo del estudio es determinar la distribución estadística de estas variables, conocida como “*distribución poblacional*”; a veces, por simplicidad en el lenguaje, a esta distribución o, incluso, a las variables, se les denomina población (por ejemplo, no es infrecuente encontrar enunciados del tipo “*la población sigue una distribución binomial*”, identificando población con variable, o “*la población es una distribución normal*”, identificando población con distribución).

Se define **muestra** como un subconjunto de la población, que será utilizado para caracterizar la distribución poblacional y **unidad muestral** a cada individuo de la muestra. Si la muestra tiene la misma distribución que la población, se dice que es **representativa**, mientras que en caso contrario se denomina **sesgada**.

Siempre será preferible un método de muestreo que proporcione muestras representativas, característica ligada a la forma de seleccionar la muestra y al tamaño de ésta. Así, siempre que sea posible, se recomienda utilizar un **muestreo probabilístico**, basado en la selección aleatoria de la muestra (conociéndose, por tanto, la probabilidad de que cada individuo salga seleccionado), lo que permite extender los resultados a toda la población, cuantificando posibles sesgos y detallando un error máximo dentro de un nivel de confianza seleccionado al inicio del proceso.

No siempre será posible utilizar un muestreo probabilístico, por lo que existe una colección de métodos de muestreo no probabilístico (conveniencia, bola de nieve, cuotas, etc.), que no se detallarán en este capítulo. La característica común a todos ellos es que los resultados obtenidos de la muestra no se deben extrapolar a la población, ya que no está garantizada la representatividad.

En el resto del capítulo se presentarán brevemente los métodos de muestreo probabilístico más usuales. Para mayor detalle, se pueden consultar las referencias [Chaudhuri and Stenger \(2005\)](#), [Arnab \(2017\)](#) o [Wu and Thompson \(2020\)](#).

14.2. Muestreo aleatorio simple

El método básico de muestreo es el conocido como **muestreo aleatorio simple (m.a.s.)**, ya introducido en la Sec. 13.2, consistente en seleccionar totalmente al azar a los individuos de la muestra, por lo que todos tienen la misma probabilidad de formar parte de ella. Si cada individuo sólo puede aparecer una vez en la muestra, se habla de **muestreo aleatorio sin reemplazamiento** o **muestreo aleatorio irrestricto**, mientras que, en caso contrario, se denomina **muestreo con reemplazamiento** o, en general, **muestreo aleatorio simple**.¹

¹Un muestreo aleatorio con reemplazamiento garantiza la independencia entre los elementos de la muestra seleccionada, mientras que un muestreo irrestricto no garantiza tal propiedad. Como consecuencia, el cálculo de las varianzas en ambos casos difiere ligeramente; sin embargo, dicha diferencia converge a cero cuando el tamaño poblacional aumenta. Así, para tamaños grandes, es casi equivalente hablar de muestreo irrestricto (sin reemplazo) que de m.a.s. (con reemplazo). En este epígrafe se ha optado por presentar los resultados en el primer caso y, como límite, los del segundo.

14.2. Muestreo aleatorio simple

229

Este procedimiento es el que se asume en la inmensa mayoría de las técnicas estadísticas convencionales, pero presenta dos inconvenientes. En primer lugar, presupone que existe un registro, o listado, completo de todos los individuos de la población (lo que no siempre es posible), y puede resultar costosa (en medios, tiempo y dinero) su aplicación práctica. En segundo lugar, presupone que la característica estudiada es homogénea en todos los individuos de la población, es decir, la distribución poblacional es idéntica en todos los individuos. Frecuentemente, esta homogeneidad poblacional no se cumple, por lo que sería necesario abordar otros métodos de muestreo que se expondrán más adelante. En todo caso, si se utiliza un m.a.s., la heterogeneidad induce un aumento de la variabilidad muestral, hecho que debe ser tenido en cuenta en la interpretación de resultados.

Además de la forma de selección, el factor que determina la representatividad de una muestra es su tamaño (véase el teorema de Glivenko-Cantelli (13.3)). Al utilizar la información muestral para aproximar los aspectos o parámetros desconocidos en la población se comete el llamado **error muestral**, que representa el margen de error que se está dispuesto a aceptar (por tanto, está íntimamente relacionado con los intervalos de confianza)²; si el tamaño de la muestra está determinado, el margen de error muestral marca el grado de precisión con el que se pueden extrapolar los resultados. Una alternativa es predeterminar un error muestral a cierto nivel de confianza y calcular cuál es el menor tamaño muestral que cumple ese requisito.

El margen de error (o simplemente error muestral), ϵ , depende del aspecto o parámetro poblacional que se quiera conocer (frecuentemente, la media, el total poblacional o la proporción), del estimador utilizado y del nivel de confianza. Si se asume una distribución poblacional normal, la expresión general sería:

$$\epsilon_\alpha = z_{1-\alpha/2} \sigma(\hat{\theta}), \quad (14.1)$$

expresión que es frecuentemente extrapolada a distribuciones poblacionales no normales.

Para el caso de estimar la media poblacional utilizando la media muestral, sustituyendo en la ecuación anterior (14.1), la relación entre error muestral, nivel de confianza y tamaño muestral sería:

$$\epsilon_\alpha = z_{1-\alpha/2} \sqrt{\left(1 - \frac{n}{N}\right) \frac{s^2}{n}}. \quad (14.2)$$

Operando y despejando el tamaño muestral, se obtiene la expresión:

$$n = \frac{z_{1-\alpha/2}^2 N s^2}{N \epsilon_\alpha^2 + z_{1-\alpha/2}^2 s^2}. \quad (14.3)$$

En algunos casos se considera que se está muestreando una población de tamaño infinito, lo que produce una simplificación de la fórmula de obtención del tamaño muestral:

²Por ejemplo, si la proporción de votantes que vota a un determinado partido ha sido estimada en el 60% y el margen de error muestral se fija en el 3%, para un nivel de confianza del 95%, se puede concluir que, con dicho nivel de confianza, el porcentaje de votantes a dicho partido estará entre el 57% y el 63% del total de votantes.

$$\epsilon_\alpha = z_{1-\alpha/2} \sqrt{\frac{s^2}{n}} \implies n = \frac{z_{1-\alpha/2}^2 s^2}{\epsilon_\alpha^2}. \quad (14.4)$$

Si se está interesado en estimar el total poblacional, un procedimiento análogo conduce a las ecuaciones:

$$\epsilon_\alpha = z_{1-\alpha/2} \sqrt{N^2 \left(1 - \frac{n}{N}\right) \frac{s^2}{n}} \implies n = \frac{z_{1-\alpha/2}^2 N^2 s^2}{\epsilon_\alpha^2 + z_{1-\alpha/2}^2 N s^2}. \quad (14.5)$$

Por último, si se desea estimar la proporción poblacional, P , de individuos que cumplen algún criterio, se puede particularizar el caso del estimador de la media poblacional sobre una población binomial, por lo que el resultado obtenido es:

$$n = \frac{z_{1-\alpha/2}^2 N p q}{(N-1) \epsilon_\alpha^2 + z_{1-\alpha/2}^2 p q}, \quad (14.6)$$

siendo $q = 1 - p$.

En la práctica es muy frecuente que se desconozca la varianza poblacional, por lo que se suele recurrir a alguna estimación previa, con lo que se tiene una aproximación al tamaño muestral requerido.

14.2.1. Ejemplo de m.a.s.

Para exemplificar el proceso de obtención de una muestra aleatoria simple en **R**, se usará el paquete **samplingbook** y el conjunto de datos **iris**, correspondiente a las medidas, en centímetros, de largo y ancho de los sépalos y pétalos de 150 flores, equidistribuidas entre las especies *setosa*, *versicolor* y *virginica*.

```
library('samplingbook')
datos_ej<-data.frame(iris)
```

En este caso, por simplicidad, se considerará que la población es el conjunto de las 150 flores disponibles, y que se desea una muestra aleatoria simple con reemplazamiento para determinar la longitud media de los sépalos con un error de 0.3 centímetros al 95% de confianza. La función **sample.size.mean()** permite calcular el tamaño de muestra necesario para cumplir estos requisitos.³

³Dado el tamaño poblacional finito ($N=150$), la función utiliza una corrección de población finita para la obtención del tamaño muestral que garantice el error máximo prefijado.

```
sd <- sd(datos_ej$Sepal.Length) # Se considera como la desviación típica poblacional
N <- nrow(datos_ej) # Tamaño de la población
e <- 0.3 # Margen de error prefijado
sample.size.mean(e, sd, N, level = 0.95)
#>
#> sample.size.mean object: Sample size for mean estimate
#> With finite population correction: N=150, precision e=0.3 and standard deviation
#> S=0.8281
#>
#> Sample size needed: 25
```

Así, basta con una muestra aleatoria simple de tamaño 25 para poder estimar la longitud media de los sépalos con los requisitos dados. Para obtener la muestra concreta, la función `sample` proporciona los valores obtenidos (conjunto de 25 valores aleatorios entre 1 y N=150) y permite seleccionar los casos que conforma la muestra:

```
set.seed(196) # Fija la semilla de aleatorización para poder reproducir los resultados
muestra <- sample(1:N, 25, replace = TRUE) # Si se quisiera un muestreo sin reemplazo,
#> se utilizaría la sentencia replace=FALSE
datos_muestra <- datos_ej[muestra, ] # Se seleccionan los datos que conforman la
#> muestra
head(datos_muestra)
#>   Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
#> 122       5.6      2.8       4.9      2.0  virginica
#> 133       6.4      2.8       5.6      2.2  virginica
#> 104       6.3      2.9       5.6      1.8  virginica
#> 95        5.6      2.7       4.2      1.3 versicolor
#> 95.1      5.6      2.7       4.2      1.3 versicolor
#> 73        6.3      2.5       4.9      1.5 versicolor
```

Para finalizar, usando la función `Smean()` se puede obtener la estimación de la media poblacional, así como su error estándar y un intervalo de confianza. Aunque en la práctica el valor poblacional es desconocido, en este ejemplo sí se puede obtener a partir del conjunto de todos los datos, lo que permite comparar la estimación con el verdadero valor buscado.

```
Smean(datos_muestra$Sepal.Length, N, level = 0.95)
#>
#> Smean object: Sample mean estimate
#> With finite population correction: N=150
#>
#> Mean estimate: 5.976
#> Standard error: 0.1115
#> 95% confidence interval: [5.7575,6.1945]
mean(datos_ej$Sepal.Length) # Valor de la media poblacional
#> [1] 5.843333
```

En este ejemplo, el error cometido sería de $5.976 - 5.843 = 0.133$ cm.

Si interesa estimar el total poblacional, basta con multiplicar la estimación de la media por el tamaño poblacional, N . Por último, si se desea estimar una proporción poblacional, el proceso sería idéntico al descrito, pero usando la función `Sprop()`.

14.3. Muestreo estratificado

Una consecuencia del muestreo aleatorio simple es que “*reproduce*” las características de la población, entre otros aspectos, su variabilidad, que se asumen comunes a todos los individuos. Sin embargo, es frecuente que haya “*grupos*” de individuos que presenten diferencias en los parámetros que rigen la distribución poblacional (por ejemplo, en sus medias o en sus varianzas) de la característica que se quiere estudiar.

En el ejemplo anterior (14.2.1), se ha muestreado para estimar la longitud media de los sépalos de las 150 flores recogidas en los datos. Sin embargo, al calcular la media y la desviación típica agrupando según la especie:

```
library('plyr')
estratos <- ddply(datos_ej, .(Species), summarize, media.sl = mean(Sepal.Length),
                     desv.sl = sd(Sepal.Length))
estratos
#>      Species media.sl   desv.sl
#> 1    setosa     5.006 0.3524897
#> 2 versicolor   5.936 0.5161711
#> 3 virginica   6.588 0.6358796
```

puede observarse que las tres especies no se comportan igual respecto al parámetro de interés (longitud media de los sépalos). La especie *setosa* presenta unos valores menores de longitud media y desviación típica, mientras que la especie *virginica* presenta los valores más elevados.

Si no se considerara la especie, un muestreo aleatorio simple podría sesgar los resultados, por ejemplo, con un predominio de las setosas o de las virginicas. En todo caso, la variabilidad del conjunto de datos es más elevada que en cualquiera de las especies, pues a la variación dentro de cada especie se une la variación entre especies. Este hecho hace aumentar el tamaño muestral necesario para estimar con un margen de error prefijado.

En general, cuando existen grupos de individuos con un comportamiento más homogéneo dentro del grupo y diferenciado entre grupos, no resulta apropiado aplicar un m.a.s. En estos casos, es recomendable el denominado **muestreo estratificado**, donde se realiza previamente una partición de la población en **estratos** y se selecciona una m.a.s. dentro de cada grupo.

La estratificación presenta ventajas, como el aumento de la representatividad de la muestra (se necesita un menor tamaño muestral total que en el m.a.s.), la reducción del error muestral (la variabilidad es menor en cada estrato) y el incremento de probabilidad de representación en la muestra de grupos con características diferenciadas. Por el contrario, no siempre resulta evidente la relación entre la variable de estratificación y las de interés.

Una de las decisiones que se han de tomar en el muestreo estratificado es el reparto de tamaño muestral entre los distintos estratos, procedimiento conocido como **afijación**. Las dos opciones

más utilizadas son la **afijación proporcional**, que reparte el tamaño muestral en función de los tamaños poblacionales de cada estrato, y la **afijación óptima**, que considera también los diferentes valores de la variabilidad dentro de cada estrato.

Para ejemplificar el proceso de muestreo estratificado en el caso de la base de datos utilizada, se procede a considerar cada especie de iris como un estrato, ya que se ha comprobado que presentan distintas distribuciones poblacionales respecto a la variable *longitud del sépalo*. Como en la Sec. 14.2.1, se quiere estimar la longitud media de la variable con un margen de error de 0.3 al 95 % de confianza. Dentro del paquete **samplingbook** se puede utilizar la función **stratasize()** para determinar el tamaño muestral que cumple estos requisitos

```
stratasize(e, Nh = c(50, 50, 50), Sh = estratos[, 3], level = 0.95)
#>
#> stratamean object: Stratified sample size determination
#>
#> type of sample: prop
#>
#> total sample size determined: 11
```

Como se aprecia, para garantizar al 95 % de confianza un margen de error de 0.3 cm es necesario un tamaño muestral de 11, sensiblemente inferior al requerido con un m.a.s. (25). Una vez determinado el tamaño, el criterio de afijación elegido distribuye la muestra entre los estratos.

```
stratasamp(n = 11, Nh = c(50, 50, 50), Sh = estratos[, 3], type = "prop")
#>
#> Stratum 1 2 3
#> Size    4 4 4
stratasamp(n = 11, Nh = c(50, 50, 50), Sh = estratos[, 3], type = "opt")
#>
#> Stratum 1 2 3
#> Size    3 4 5
```

Como los tres estratos tienen el mismo tamaño poblacional, la afijación proporcional distribuye la muestra equitativamente; sin embargo, la afijación óptima, al considerar las diferencias en variabilidad, asigna más muestra al estrato con mayor variabilidad y menos muestra al de menor variabilidad (como los tamaños muestrales son necesariamente números enteros, se puede producir una ligera diferencia entre el tamaño muestral calculado globalmente y la suma de los tamaños de cada estrato).

Con la afijación óptima estimada, se procede a la selección de la submuestra en cada estrato (m.a.s.) y a la obtención de los datos que conforman la muestra.

```
set.seed(195) # Fija la semilla de aleatorización
muestra1 <- sample(1:50, 3, replace = TRUE)
muestra2 <- sample(51:100, 4, replace = TRUE)
muestra3 <- sample(101:150, 5, replace = TRUE) # m.a.s. en cada estrato
muestra_estr <- c(muestra1, muestra2, muestra3)
```

```
datos_muestra_estr <- datos_ej[muestra_estr, ] # Selección de los datos que conforman
#→ la muestra
datos_muestra_estr
#>   Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
#> 26      5.0       3.0      1.6       0.2    setosa
#> 38      4.9       3.6      1.4       0.1    setosa
#> 5       5.0       3.6      1.4       0.2    setosa
#> 61      5.0       2.0      3.5       1.0 versicolor
#> 61.1     5.0       2.0      3.5       1.0 versicolor
#> 65      5.6       2.9      3.6       1.3 versicolor
#> 58      4.9       2.4      3.3       1.0 versicolor
#> 147     6.3       2.5      5.0       1.9  virginica
#> 146     6.7       3.0      5.2       2.3  virginica
#> 134     6.3       2.8      5.1       1.5  virginica
#> 130     7.2       3.0      5.8       1.6  virginica
#> 141     6.7       3.1      5.6       2.4  virginica
```

Finalmente, usando la función `Smean()` se obtiene la estimación de la media poblacional, así como su error estándar y un intervalo de confianza, al igual que se hizo con el m.a.s.

```
Smean(datos_muestra_estr$Sepal.Length, N, level = 0.95)
#>
#> Smean object: Sample mean estimate
#> With finite population correction: N=150
#>
#> Mean estimate: 5.7167
#> Standard error: 0.2393
#> 95 % confidence interval: [5.2476, 6.1857]
```

14.4. Otros tipos de muestreo probabilístico

Existen otros métodos de muestreo probabilístico que buscan simplificar la extracción de una muestra representativa, entre los que destacan el **muestreo por conglomerados** y el **muestreo sistemático**.

Cuando la población es muy grande, es frecuente que se puedan establecer (o construir a partir de alguna variable) subgrupos, o **clusters**, que tengan las mismas características que todo el conjunto respecto a la variable de interés. En esos casos, a efectos de estimación, sería equivalente muestrear toda la población o sólo un cluster, con el consiguiente ahorro de tamaño muestral, tiempo y coste. Es el conocido como **muestreo por conglomerados**.

Por ejemplo, se puede estar interesado en estimar el tiempo medio que el alumnado de E.S.O. dedica a estudiar matemáticas en España. Obtener una muestra para todo el país puede ser costoso en tiempo, recursos materiales y tamaño muestral; sin embargo, se puede asumir que no existen diferencias entre provincias respecto a esta variable, por lo que sería posible muestrear sólo en una provincia (o pocas). En este caso, el muestreo por conglomerados consistiría en una

primera etapa de selección aleatoria de clusters (provincia/s en este ejemplo) y, posteriormente, aplicar un método de muestreo sobre dicha selección (que, a su vez, podría ser un m.a.s. o un muestreo estratificado).

No conviene confundir los conceptos de estrato y cluster, aunque ambos sean subgrupos de la población total. En el primer caso, los individuos de cada estrato son muy homogéneos entre sí y diferenciados del resto de estratos. En el segundo caso, los individuos de cada cluster tienen la misma variabilidad que el conjunto de la población, no habiendo diferencias entre clusters respecto a la variable de interés. Así, la ganancia en el muestreo estratificado proviene de trabajar con menores variabilidades intra-estratos, mientras que en el muestreo por conglomerados proviene de utilizar una subpoblación más pequeña.

En otras situaciones, si se dispone de un marco poblacional (listado completo de los individuos), es posible plantear un mecanismo sencillo de obtención de la muestra. Si se tiene un tamaño poblacional N y se quiere una muestra de tamaño n , se pueden establecer $k = N/n$ bloques, elegir al azar un número entre 1 y k (que permite seleccionar el primer elemento de la muestra) y, a partir de esa posición, dar saltos de magnitud k en el listado para seleccionar el resto de unidades muestrales. Es el método conocido como **muestreo sistemático** o, más técnicamente, **muestreo sistemático uniforme de paso k** .

Como ejemplo, supóngase que se quiere obtener una muestra de 50 individuos en una población de 2000. El paso sería $k = 2000/50 = 40$ unidades y se selecciona aleatoriamente una unidad entre las 40 primeras (supóngase que la número 13); el resto de la muestra se obtendría sumándole el paso a la primera seleccionada (13, 53, 93, 133, 173, y así hasta la 1973).

Por último, los métodos expuestos no son incompatibles, sino que se pueden combinar por etapas, dando lugar a los conocidos como **muestreos polietápicos**. Por ejemplo, la Encuesta de Población Activa, elaborada por el Instituto Nacional de Estadística, adopta un muestreo bietápico, en primer lugar, estratificado entre secciones censales y, en segundo lugar, muestreando entre las viviendas familiares de cada sección.

14.5. Técnicas de remuestreo: Bootstrap

Cuando se infiere una característica poblacional a partir de una muestra, sólo se dispone del valor concreto que el estadístico toma sobre dicha muestra. Salvo en raras ocasiones, no se dispone de su distribución en el muestreo, o sólo se tiene una aproximación asintótica, por lo que no se pueden evaluar sus propiedades estadísticas con tamaños muestrales no elevados. En otros casos, es la complejidad analítica de muchas técnicas actuales de análisis de datos la que dificulta la determinación de la distribución de las estimaciones de los parámetros.

En estos casos, el método **bootstrap** propone sustituir la distribución poblacional (desconocida) por una estimación conocida (como puede ser la distribución empírica o una aproximación paramétrica) que, mediante remuestreo, sirva para generar muestras aleatorias a partir de la muestra original. Se obtiene así una distribución de remuestreo, llamada también **distribución bootstrap**, cuyo comportamiento sobre la estimación aproxima a la de la distribución muestral en torno al parámetro, lo que permite evaluar la precisión de las estimaciones.

El método bootstrap más sencillo, llamado **bootstrap uniforme** o **bootstrap naïve**, parte de la aproximación de la distribución poblacional por la distribución empírica de la muestra. Supóngase que se tiene una muestra $X = (x_1, \dots, x_n)$ que es utilizada para obtener un estimador $T(X) = \hat{\theta}$ para un parámetro poblacional θ . Utilizando su distribución empírica (véase ecuación (13.2)):

- Se genera una primera muestra $X^{*1} = (x_1^{*1}, \dots, x_n^{*1})$, obtenida mediante muestreo aleatorio simple con reemplazamiento de la muestra original, que permite evaluar el estadístico $T^{*1}(X^{*1}) = \theta^{*1}$.
- Siguiendo el mismo procedimiento, se puede generar un número elevado (B) de muestras bootstrap, X^{*1}, \dots, X^{*B} , que permiten obtener el valor del estadístico sobre cada una de ellas $T^{*1}(X^{*1}), \dots, T^{*B}(X^{*B})$
- Con estos valores, se obtiene la distribución bootstrap. Por ejemplo, se puede utilizar esta distribución para calcular la media bootstrap del estadístico $T(X)$, $\bar{T}^* = \frac{1}{B} \sum_{b=1}^B T^*(X^{*b})$, y cuyo error estándar es:

$$\hat{S}_{boot} = \sqrt{\frac{1}{B-1} \sum_{b=1}^B (T(X^{*b}) - \bar{T}^*)^2}, \quad (14.7)$$

expresión que aproxima al error del estadístico $T(X)$ para estimar θ .

Ejemplo: los datos sobre calidad del aire en la ciudad de Nueva York (`airquality`) recogen la variable `Temp` que mide la temperatura, en grados Fahreheit, entre el día 1 de mayo y el 30 de septiembre. Se ha visto en la sección 13.7, que no se puede asumir que dicha variable esté generada por una distribución normal, por lo que no se podrían utilizar los intervalos de confianza mostrados en la Sec. 13.6. El objetivo es calcular una estimación de la temperatura media y dar un intervalo al 95 % de confianza.

En este ejemplo, se utiliza la muestra para obtener un valor del estadístico *media muestral* ($\bar{X} = 77,88$), estimador insesgado de la media poblacional. Sin embargo, al no conocer la distribución en el muestreo (no se asume ningún tipo de distribución poblacional ni se puede hacer uso de aproximaciones asintóticas), no se podría construir un intervalo de confianza.

Aplicando el método bootstrap (en su versión uniforme), se van a obtener 5000 muestras de tamaño 20, mediante remuestreo con reemplazamiento,

```
set.seed(196) #Se fija la semilla para permitir la reproducibilidad
B<-5000 #Se fija el número de remuestras
muestras_boot<-numeric(B) #Se almacenan todos los valores del estadístico
for (k in 1:B) {
  remuestra <- sample(airquality$Temp, 20, replace = TRUE)
  muestras_boot[k] <- mean(remuestra)
}
media_boot<-mean(muestras_boot)
```

14.5. Técnicas de remuestreo: Bootstrap

237

```
media_boot #Media de los 5000 valores medios de las remuestras
#> [1] 77.87309
desv_boot<-sd(muestras_boot)
desv_boot #Desviación típica de los 5000 valores medios de las remuestras
#> [1] 2.090255
```

Para construir los intervalos de confianza, se calculan los valores críticos al 95 % de confianza sobre la distribución empírica de los valores medios remuestreados:

```
val_crit<-quantile(muestras_boot,c(0.025,0.975))
val_crit #Valores críticos
#> 2.5% 97.5%
#> 73.75 81.95
```

Así, con el método bootstrap, no es necesario asumir ninguna distribución en el muestreo. Aún así, la representación gráfica de la distribución empírica de los valores medios remuestreados y los extremos del intervalo de confianza (Fig. 14.1), muestra cómo la distribución de las remuestras sobre el estadístico se asemeja a la distribución del estadístico sobre el parámetro.

```
hist(muestras_boot, freq=FALSE)
lines(density(muestras_boot))
abline(v=val_crit)
```

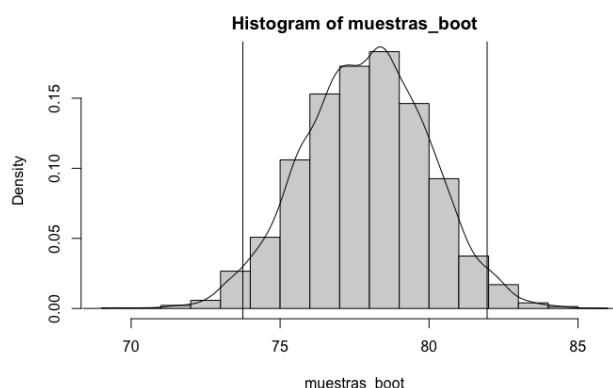


Figura 14.1: Distribución empírica de la media remuestreada

El paquete **boot** de **R** permite también obtener réplicas de un estadístico sobre una muestra. La función básica de este paquete es **boot()**, que permite utilizar distintos métodos de remuestreo. En su estructura más simple, basta con indicar los datos originales, el estadístico que se quiere remuestrear y el número de réplicas.

Así, si se quiere estimar, por ejemplo, la mediana de la población con 1000 remuestras, se puede recurrir a la función **boot()**:

```

library(boot)
estadistico<-function(data,i){
  median(data[i]) #Se especifica aquí el estadístico remuestreado
}
set.seed(196)
mediana_boot<-boot(airquality$Temp,estadistico,R=1000)
mediana_boot
#>
#> ORDINARY NONPARAMETRIC BOOTSTRAP
#>
#>
#> Call:
#> boot(data = airquality$Temp, statistic = estadistico, R = 1000)
#>
#>
#> Bootstrap Statistics :
#>      original   bias   std. error
#> t1*       79   -0.052    1.071655

```

Como resultado, se obtiene el valor del estadístico sobre la muestra original, el sesgo estimado y el error estándar. Para calcular los intervalos de confianza, basta con utilizar la función `boot.ci()` sobre la muestra bootstrap obtenida, indicando el nivel de confianza y el tipo de intervalo (por defecto, `all` proporciona todos los intervalos disponibles; en el ejemplo, se usa el método de los percentiles)

```

boot.ci(mediana_boot,conf=0.95,type="perc")
#> BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
#> Based on 1000 bootstrap replicates
#>
#> CALL :
#> boot.ci(boot.out = mediana_boot, conf = 0.95, type = "perc")
#>
#> Intervals :
#> Level      Percentile
#> 95%        (77, 81)
#> Calculations and Intervals on Original Scale

```

La función `boot()` permite modificaciones del bootstrap uniforme mediante parámetros adicionales. Por ejemplo, el parámetro `strata` se utiliza para generar remuestreos estratificados cuando la muestra original también lo es. Aunque no se han comentado dado el carácter introductorio de este capítulo, existen otros métodos bootstrap, que se pueden obtener especificándolos mediante el parámetro `sim`. Por defecto, el valor es `ordinary`, que corresponde al bootstrap uniforme; otras alternativas pueden ser `parametric` para bootstrap paramétrico, `balanced`, `permutation` o `antithetic` para otros métodos más avanzados.

14.5. Técnicas de remuestreo: Bootstrap

239

Resumen

El muestreo probabilístico busca seleccionar una muestra representativa de una población, que permita inferir la distribución poblacional o alguno de sus parámetros. Las decisiones básicas para un correcto proceso de muestreo son el método utilizado (aleatorio simple, estratificado, polietápico, etc.), que depende de la estructura de la población, y la determinación del tamaño muestral que garantice el margen de error asumible. La técnica **bootstrap** de remuestreo permite aproximar la distribución de estadísticos muestrales sin asumir ninguna hipótesis sobre la distribución poblacional, ventaja muy útil para evaluar la precisión de los estimadores en muchísimas técnicas complejas.

Parte IV

Modelización estadística

Capítulo 15

Modelización lineal

Víctor Casero-Alonso^a y María Durbán^b

^aUniversidad de Castilla-La Mancha

^bUniversidad Carlos III de Madrid

15.1. Modelización

Se acude a los **modelos de regresión** para intentar explicar la relación entre dos o más variables. Para ello se predefine un modelo que pretende explicar el comportamiento de la variable **respuesta o dependiente**, denotada por Y , utilizando la información proporcionada por las **variables explicativas**, también llamadas independientes o predictoras, denotadas por X_1, \dots, X_p . Pero dichas variables pueden ser de distinto tipo. Si la variable respuesta es continua, más concretamente, si se puede asumir que sigue una **distribución de probabilidad Normal**, y al menos una de las variables explicativas es también continua, se puede acudir a la **modelización lineal** que se desarrolla en este capítulo. Sin embargo, si la variable respuesta fuese de otro tipo, por ejemplo, dicotómica, la modelización lineal no sería adecuada. En el Cap. 16, en el que se aborda el **modelo lineal generalizado**, quedará más clara esta distinción.

El primer paso en el proceso de modelización es intentar explicar una variable respuesta, que de aquí en adelante se supone continua y con distribución Normal, a partir de una sola de las variables explicativas, de forma *lineal* (**modelo lineal simple**). Dicho modelo probablemente no será “bueno”, no explicará bien el comportamiento de la variable respuesta si la realidad que se pretende explicar es compleja, pero podría ser *suficiente* para el propósito del estudio¹.

¹La capacidad de explicación la proporciona el coeficiente de **bondad de ajuste** o coeficiente de determinación lineal, R^2 (véase la Sec. 15.2.1).

Nota

Se entiende por **modelo lineal** aquel cuya relación entre las variables viene determinada por una combinación *lineal* de los parámetros, por ejemplo:

- $Y = \beta_0 + \beta_1 X + \epsilon.$
- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon.$
- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \epsilon.$

El último ejemplo refleja un modelo lineal en los parámetros, pero no lineal en las variables, por el término X_1^2 . Ejemplos de **modelos no lineales** en los parámetros son:

- $Y = \beta_0 \cdot e^{\beta_1 X_1} + \epsilon.$
- $Y = \beta_0 + X_1^{\beta_1} + \epsilon.$

Por ejemplo, se sabe que el peso de una persona está relacionado con muchos factores, pero uno de los más determinantes es la altura. Si se recogen datos de pesos y alturas de un conjunto de personas se puede ajustar el modelo y obtener una explicación *suficiente*, aunque parcial, del peso de una persona a partir de su altura. Es claro que la inclusión de otras variables en el modelo puede ayudar a *explicar* mejor la variable respuesta. Se llega así al denominado **modelo de regresión lineal múltiple** que se puede expresar matemáticamente como:

$$Y_j = \beta_0 + \beta_1 X_{1j} + \dots + \beta_p X_{pj} + \epsilon_j, \quad \epsilon_j \sim N(0, \sigma^2), \quad j = 1, \dots, N. \quad (15.1)$$

donde:

- β_0 es el **término independiente o constante** del modelo,
- β_1, \dots, β_p son los **coeficientes de regresión o parámetros** del modelo, que se estimarán a partir de los datos observados $(x_{1j}, \dots, x_{pj}, y_j)$ y reflejan la magnitud del efecto *lineal* (constante) sobre la variable explicada Y de incrementos unitarios en las variables explicativas X_i .
- y ϵ_j es el **término de error** del modelo, la parte de Y que no es capaz de explicar la parte determinista del mismo ($\beta_0 + \beta_1 X_{1j} + \dots + \beta_p X_{pj}$), que se supone sigue una distribución de probabilidad Normal, con media 0 y varianza constante σ^2 ;
- además, se asume que las **observaciones** son **independientes**.

Siguiendo con el ejemplo del peso, añadir alguna variable genética, el sexo u otras, ayudará a mejorar la “bondad” del modelo lineal. Otro ejemplo consiste en pretender explicar el salario en un determinado sector económico en función de los años de experiencia, la formación, la situación familiar, el sexo, etc., de los trabajadores. Nótese que entre las variables explicativas sí puede haber variables de distinto tipo, continuas, categóricas, etc.². Ahora bien, la interpretación de los coeficientes dependerá del tipo de variable al que van asociados, como se verá en los casos prácticos (Sec. 15.4).

²Conviene mencionar la estrecha relación entre regresión lineal múltiple y **ANOVA con varios factores** (véase la Sec. 15.4.6.1).

Un par de referencias para ampliar conocimientos sobre este tema utilizando **R** son [Faraway \(2002\)](#) y [James et al. \(2013\)](#).

15.2. Procedimiento de modelización

15.2.1. Estimación del modelo

Los datos recogidos u observados sirven para **especificar** la relación predefinida de antemano, mediante la **estimación** de los coeficientes β_i que mejor ajustan dicha relación, utilizando el método de **mínimos cuadrados**. Además, los correspondientes contrastes permiten decidir si cada coeficiente es **significativamente distinto de 0**, esto es, si tiene un *efecto* significativo sobre la respuesta,³ en cuyo caso tiene sentido mantener en el modelo la variable a la que va asociado. En la práctica, el coeficiente estimado es *significativo* si su **p-valor** (definido en la Sec. 13.5) asociado es suficientemente pequeño.

Nota

Se acepta, mayoritariamente, como “suficientemente pequeño” un p-valor inferior a 0.05, lo que supone un nivel de confianza en las estimaciones del 95 %. Pero dicho valor es arbitrario y podrían considerarse otros valores de referencia. Por ejemplo, en las salidas de **R** aparecen otros tres niveles de referencia: 0,1, 0,01 y 0,001. En general, cuanto menor sea el p-valor más confianza se tendrá en las conclusiones.

Como se avanzó anteriormente, si algún coeficiente no es significativo, procede eliminar del modelo la variable explicativa asociada. En tal caso, se vuelven a estimar los coeficientes de las variables que se mantienen hasta llegar a un modelo con todos los coeficientes significativos, iterando las veces necesarias⁴. Para facilitar esta labor, se han desarrollado métodos automáticos de selección de variables, basados en la comparación de la varianza residual (haciendo uso del test *F*), mediante el estadístico AIC (criterio de información de Akaike), etc.⁵ Junto con los contrastes, se pueden aportar los intervalos de confianza de los coeficientes, que, si son significativos, no contendrán el valor 0.

A la par del contraste de significación de cada coeficiente, se obtiene el **contraste de significación global del modelo**. La hipótesis nula es que todos los coeficientes β_1, \dots, β_p son 0. Dicho de otro modo, que el conocimiento de las variables X_1, \dots, X_p no aporta información alguna para explicar los valores de Y .

También se ha de obtener la **bondad del ajuste** del modelo, normalmente medida por el **coeficiente de determinación lineal**, R^2 (adimensional, que toma valores entre 0 y 1). Para comparar entre diferentes modelos, se utiliza el **R^2 ajustado/corregido**, que tiene en cuenta la composición/complejidad del modelo (número de variables, etc.). Cuanto mejor ajuste el

³Desde el punto de vista estadístico, la influencia/efecto sobre la respuesta no es fruto del azar.

⁴El proceso debe basarse en la relación entre las variables predefinidas de antemano (por ejemplo, en la formulación teórica del modelo). Por ello, a pesar de la no significatividad estadística de algún coeficiente, en ocasiones, la variable asociada se mantiene en el modelo.

⁵Consultese el Cap. 10 de [Faraway \(2002\)](#).

modelo los datos observados, más próximo a 1 será el valor de R^2 (1 indica una relación lineal perfecta entre la variable respuesta y las predictoras). Por el contrario, un R^2 cercano a 0 indica que el modelo estimado ajusta mal los datos.

Es habitual valorar conjuntamente la significación global del modelo, su bondad de ajuste y la significación de cada uno de los coeficientes, considerándose apropiados aquellos modelos que son globalmente *significativos* y tienen la suficiente “bondad”, aunque tengan coeficientes no significativos.

15.2.2. Validación del modelo

Aunque el modelo sea significativo se debe *validar*, es decir, se deben someter a contraste los supuestos estadísticos que subyacen al modelo. Para ello se utilizan los **residuos** del modelo, la parte de Y que no explica la regresión estimada o, en otros términos, la diferencia entre los valores observados y los estimados. Matemáticamente,

$$e_j = y_j - \hat{y}_j = y_j - (\hat{\beta}_0 + \hat{\beta}_1 x_{1j} + \dots + \hat{\beta}_p x_{pj}), \quad j = 1, \dots, N.$$

Los supuestos a contrastar son:

- los residuos han de tener **varianza constante** (por definición tienen **media cero**).
- los residuos han de seguir la **distribución** de probabilidad **Normal**.
- las observaciones tienen que ser **independientes**.
- la **relación** entre la variable respuesta y las explicativas se supone **lineal**.
- las variables explicativas son linealmente independientes: ninguna puede ser explicada como combinación lineal de las otras. En caso contrario, se tendría el conocido problema de la **multicolinealidad** y debería quitarse del modelo la variable explicada por el resto.

15.2.3. Interpretación de los coeficientes

Una vez validado el modelo, se procede a la interpretación de los coeficientes significativos. Teniendo en cuenta la expresión del modelo de regresión lineal múltiple (15.1), la regla general de interpretación de cada uno de los coeficientes de regresión estimado $\hat{\beta}_i$ es simple y directa: el cambio/impacto *medio* en el valor de la variable respuesta Y ante un cambio unitario de una variable explicativa cuantitativa o ante un cambio de categoría (desde la que se toma como referencia) si la variable es categórica. Y ello *ceteris paribus*, esto es, manteniendo constante el valor de las demás variables explicativas.

Habrá que tener en cuenta la magnitud de cada variable, porque la influencia real en la respuesta podría ser de poca magnitud (quizá por las unidades o escala utilizada), pero estadísticamente significativa .

15.2.4. Predicción

La utilidad del modelo estimado (especificado) queda plasmada en su utilización para **predecir** nuevos valores, \hat{y}_j , a partir del conocimiento/asignación de nuevos valores de las variables explicativas, $\{x_{1j}, \dots, x_{pj}\}$. No obstante, dichas predicciones son valores *esperados (medios)*, pudiéndose construir sus correspondientes intervalos de confianza.

15.3. Procedimiento con **R**: la función `lm()`

R tiene implementada la función `lm()` para ajustar/estimar modelos de regresión lineal múltiple:

```
lm(formula, data = ..., ... )
```

El argumento mínimo necesario es **formula**, donde se predeterminará la relación entre las variables respuesta y explicativas de una forma bastante intuitiva:

- $Y \sim X$, es la fórmula a utilizar para definir un modelo simple donde Y denota la variable respuesta y X la variable explicativa: $Y = \beta_0 + \beta_1 X + \epsilon$.
- $Y \sim X_1 + X_2$, define un modelo lineal múltiple con 2 variables explicativas: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$.
- $Y \sim X_1 + X_2 + X_3 - 1$ elimina el término independiente, β_0 , del modelo lineal múltiple de 3 variables explicativas.
- ...

En el segundo argumento, **data**, se indica el conjunto de datos donde se encuentran las variables de trabajo. No es especificarlo si están en el **Environment**.

Hay que hacer notar que **R** considera, por defecto, las variables explicativas como cuantitativas. Si se tienen variables categóricas codificadas con números hay que indicarle que las trate como categóricas, usando la función `factor()`⁶. De no hacerlo, la función las consideraría numéricas, con el consecuente error de interpretación de los coeficientes asociados.

A partir de los datos disponibles de Y, X_1, \dots, X_p , la función `lm()` estima los coeficientes $\hat{\beta}_i$ asociados a cada variable X_i , mediante el **método de mínimos cuadrados**, y calcula sus errores estándar, con los que obtiene sus estadísticos de contraste (de la t de Student)⁷ y su significación. En el objeto `lm` que se genera también se almacenan los valores ajustados, residuos, etc., que se pueden mostrar a través de funciones genéricas disponibles en **R**. Algunas de ellas son:

- `print()`: muestra un breve resumen.
- `summary()`: proporciona un resumen completo.
- `coef()`: proporciona las estimaciones de los coeficientes del modelo.

⁶Si no está ya definida como `factor()` en el conjunto de datos.

⁷Con los errores estándar también se pueden obtener los intervalos de confianza de los coeficientes.

- `confint()`: construye intervalos de confianza para los coeficientes.
- `fitted.values()`: muestra los valores ajustados del modelo (para cada observación del `data.frame`).
- `residuals()`: calcula los residuos del modelo (también para cada observación del `data.frame`).

15.4. Casos prácticos

En esta Sección se utilizan los datos `airquality`⁸, que consisten en 154 medidas (de 6 variables) de calidad del aire en Nueva York. Las variables consideradas aquí son las cuatro siguientes:

- `Ozone`: Concentración media de ozono en la atmósfera (en ppb, partes por billón).
- `Solar.R`: Radiación solar (en lang, Langley).
- `Wind`: Velocidad media del viento (en mph, millas por hora).
- `Temp`: Temperatura máxima diaria (en grados Fahrenheit).

El objetivo es establecer la relación entre la concentración de ozono en la atmósfera, variable respuesta, y las variables meteorológicas `Solar.R`, `Wind` y `Temp`, variables explicativas. Los valores disponibles de las cuatro variables permiten considerarlas como variables continuas.

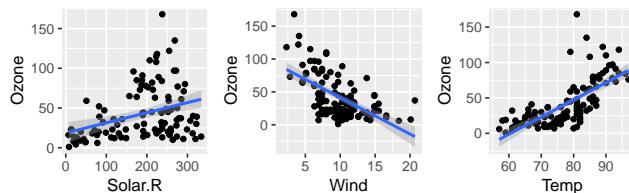


Figura 15.1: Gráficos de dispersión de las variables explicativas frente a la variable respuesta

Antes de proceder con el ajuste múltiple se pueden realizar los ajustes simples, individuales. La Fig. 15.1 representa 3 regresiones lineales simples, de la variable respuesta `Ozone` sobre cada una de las 3 variables explicativas. Cada gráfico muestra un **diagrama de dispersión** de sólo dos variables, la explicativa en el eje X y la respuesta en el eje Y, obteniéndose la popularmente denominada **nube de puntos**. En tales diagramas se puede ver si entre las variables hay

⁸Conjunto de datos incluido con la instalación `base` de R. Más información ejecutando `?airquality`.

relación lineal, o no, y en caso de que la haya, si es positiva/directa (a mayores valores de X , mayores valores de Y) o negativa/inversa. En cada gráfico se ha añadido la correspondiente recta de regresión lineal (con su correspondiente intervalo de confianza), que podría no ser la más apropiada, como parece que ocurre en las regresiones de `Ozone` sobre `Wind` y sobre `Temp`. En ambos casos, la relación parece más bien no lineal, aunque una recta podría ser suficiente (en función del interés del estudio) para explicar relativamente bien el comportamiento del nivel de concentración de ozono. El código para el obtener el primer gráfico es:

```
library("ggplot2")
ggplot(airquality, aes(Solar.R, Ozone)) +
  geom_point() +
  theme(aspect.ratio=1) +
  geom_smooth(method = "lm")
```

En regresión lineal múltiple no es posible visualizar en un sólo gráfico la relación entre la variable respuesta y varias variables explicativas, salvo si son sólo 2, en cuyo caso se tendría un gráfico en 3 dimensiones, generalmente difícil de visualizar.

15.4.1. Estimación de los coeficientes

Se comienza ajustando el siguiente modelo lineal múltiple:⁹:

$$Ozone = \beta_0 + \beta_1 Solar.R + \beta_2 Wind + \beta_3 Temp + \epsilon$$

La definición en **R** del modelo se puede ver como primer argumento de la función `lm()`. El objeto que genera la función `lm()` se guarda bajo el nombre de `airq_lm` y, a continuación, se muestra su resumen con `summary()`:

```
airq_lm <- lm(Ozone ~ Solar.R + Wind + Temp, data=airquality)
summary(airq_lm)
#>
#> Call:
#> lm(formula = Ozone ~ Solar.R + Wind + Temp, data = airquality)
#>
#> Residuals:
#>      Min       1Q   Median       3Q      Max
#> -40.485 -14.219 - 3.551 10.097 95.619
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept) -64.34208 23.05472 -2.791 0.00623 **
#> Solar.R      0.05982 0.02319 2.580 0.01124 *
#> Wind         -3.33359 0.65441 -5.094 1.52e-06 ***
#> Temp          1.65209 0.25353 6.516 2.42e-09 ***
```

⁹Más adelante se introducirá una variable categórica para enriquecer el análisis.

```
#> ---
#> Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 21.18 on 107 degrees of freedom
#> (42 observations deleted due to missingness)
#> Multiple R-squared: 0.6059, Adjusted R-squared: 0.5948
#> F-statistic: 54.83 on 3 and 107 DF, p-value: < 2.2e-16
```

La salida del `summary()` proporciona las estimaciones de los coeficientes del modelo (columna `Estimate`). El término independiente aparece como (`Intercept`) y toma el valor $\beta_0 = -64.3421$, el coeficiente asociado a `Solar.R` es $\beta_1 = 0.0598$, etc. También aparecen sus p-valores asociados (columna `Pr(>|t|)`), pudiéndose comprobar que los 4 coeficientes son significativos al 5 %. Según la leyenda `Signif. codes`, a mayor número de asteriscos mayor significación del coeficiente (menor p-valor). Así, los coeficientes de `Temp` y `Wind` son más significativos que el de `Solar.R`.

También se pueden apreciar (penúltima línea) dos medidas de la bondad del ajuste del modelo considerado: el R cuadrado múltiple y el R cuadrado múltiple ajustado. En el ejemplo, el R^2 (ajustado) es 0.5948, que se podría considerar “suficiente” o no en función del objetivo del estudio, aunque, en este caso, está claro que el modelo no explica suficientemente bien la concentración de ozono.

En la última linea de la salida aparece información sobre el contraste global del modelo: valor del estadístico F , grados de libertad y p-valor asociado. Como se aprecia, el modelo es globalmente significativo (p-valor del orden de 10^{-16}).

15.4.2. Validación

Lo anterior carece de *validez* si no se satisfacen las hipótesis del modelo mencionadas en la Sec. 15.2.2, principalmente las relativas a varianza constante (homocedasticidad) y normalidad. Para ello se realiza un análisis de residuos. La función `autoplot()` del paquete `ggfortify` proporciona los gráficos que se muestran en la Fig. 15.2.

```
library("ggfortify")
autoplot(airq_lm) +
  theme_minimal()
```

Por un lado, el gráfico de residuos frente a valores ajustados (fitted) muestra cierta heterocedasticidad (varianza cambiante con el valor en el eje X) y no linealidad (ya apreciable de forma individual en la Fig. 15.1). Por su parte, el gráfico Normal Q-Q, que enfrenta los residuos estandarizados con los cuantiles de la distribución Normal, indica que los residuos presentan desviaciones de la normalidad en ambas colas.

Para completar el análisis gráfico se puede acudir a contrastes de hipótesis vistos en la Sec. 13.5. El más habitual para contrastar normalidad es el de **Shapiro-Wilk**, que se implementa en **R** con la función `shapiro.test()`¹⁰. Para contrastar la homocedasticidad se puede utilizar

¹⁰Se pueden encontrar otros contrastes de normalidad en el paquete `nortest`.

15.4. Casos prácticos

251

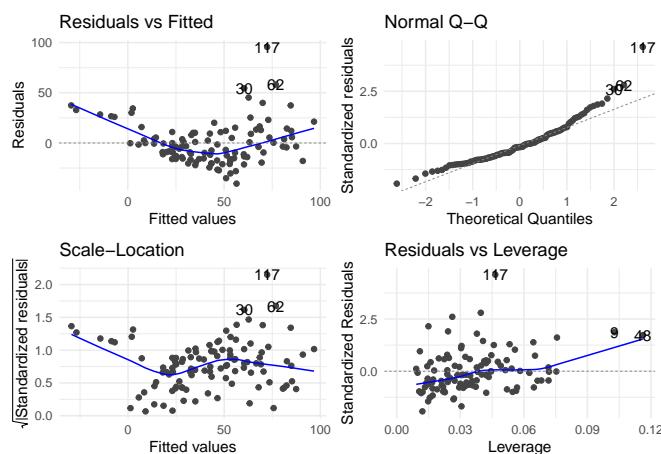


Figura 15.2: Gráficos de residuos

alguno de los tres tests implementados para tal fin en el paquete `lmtest()`: el de Breusch-Pagan `bptest()`, el de Goldfeld-Quandt `gqtest()` o el de Harrison-McCabe `hmctest`.

```
shapiro.test(airq_lm$residuals)
#>
#> Shapiro-Wilk normality test
#>
#> data: airq_lm$residuals
#> W = 0.91709, p-value = 3.618e-06
lmtest::bptest(airq_lm)
#>
#> studentized Breusch-Pagan test
#>
#> data: airq_lm
#> BP = 5.0554, df = 3, p-value = 0.1678
```

El contraste de homocedasticidad lleva a no rechazar tal supuesto ($p\text{-valor} > 0.05$), pero el contraste de Shapiro-Wilk confirma la falta de normalidad ($p\text{-valor} < 0.05$). A este respecto, en la Fig. 15.3 se muestra el histograma de la variable `Ozone`, apreciándose que los datos recogidos presentan asimetría incompatible con la normalidad, asumida por defecto para la variable respuesta. Una posible solución sería el uso de una transformación logarítmica, que produce cierta simetría en la distribución de la variable, acercándola, por tanto, a la normalidad.

Para el análisis de colinealidad se pueden representar gráficos 2 a 2 de las variables explicativas, para comprobar si están o no correlacionadas (Fig. 15.4).

```
library("GGally")
ggpairs(airquality[, 2:4])
```

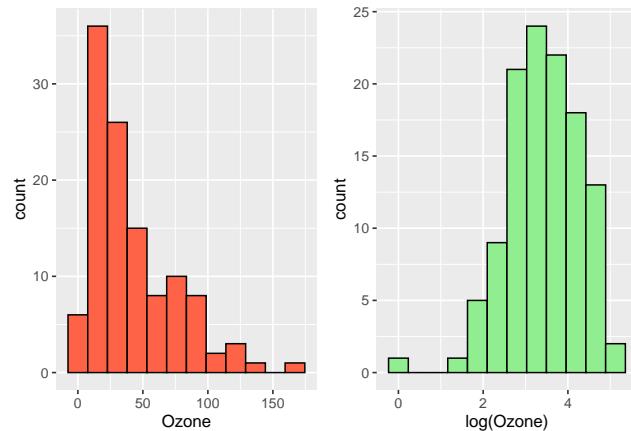


Figura 15.3: Histogramas de las variables ‘Ozone’ y ‘log(Ozone)’

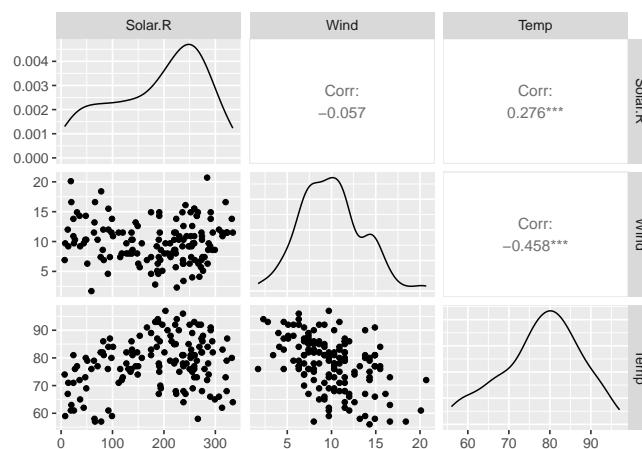


Figura 15.4: Gráfico de dispersión por pares de las variables explicativas

15.4. Casos prácticos

253

Pero este es un análisis parcial, puesto que una de las variables explicativas podría venir explicada por el resto o varias de ellas. Por si este fuera el caso, conviene calcular también los **factores de inflación de la varianza** (VIF), que indican el incremento de la varianza estimada del coeficiente de regresión de una determinada variable explicativa como consecuencia de la colinealidad con las demás (para más detalle, véase el Cap. 3 de James et al. (2013)).

El mínimo valor de VIF es 1, no existiendo límite superior. Una regla general para interpretar los VIF es la siguiente: Si el VIF de una variable explicativa X_i es 1, no hay correlación entre ella y cualquier otra variable explicativa del modelo. Si está entre 1 y 5 la correlación es moderada y no provoca graves problemas. Si es mayor que 5 la correlación es fuerte y, probablemente, las estimaciones de los coeficientes y los p-valores resultantes de la estimación del modelo no sean confiables.

Los VIF se pueden obtener mediante la función `vif()` del paquete `car`:

```
car::vif(airq_lm)
#> Solar.R      Wind      Temp
#> 1.095253 1.329070 1.431367
```

En la Fig. 15.4 se aprecia que los gráficos de dispersión muestran ausencia de correlación entre `Solar.R` y `Wind`; sin embargo, la correlación entre `Wind` y `Temp` no parece despreciable. No obstante, todos los VIF son prácticamente unitarios, por lo que se puede concluir que el modelo no presenta multicolinealidad.

15.4.3. Interpretación de los coeficientes

De acuerdo con lo dicho en la Sec. 15.2.3 se tiene que:

- Un incremento en `Temp` de un grado Fahrenheit, manteniéndose constantes los valores de `Wind` y `Solar.R`¹¹, provoca un aumento (por ser positivo el coeficiente) promedio en el nivel de concentración de ozono en el aire de 1,6521 ppb.
- El coeficiente de `Wind` es negativo, por lo que un aumento en la variable `Wind`, *ceteris paribus*, reduce la concentración de ozono. En concreto, dicha reducción, es de 3,3336 ppb por cada milla por hora que se incremente la variable `Wind`.
- La influencia de `Solar.R` en el nivel de concentración de ozono en la atmósfera es positiva, como la de `Temp`, pero de mucha menor magnitud (por las unidades de una y otra). Concretamente, por cada langley (Ly) que se incremente `Solar.R` el nivel de concentración de ozono se eleva, *ceteris paribus*, en 0,0598 ppb.

Por tanto, el impacto promedio de un incremento unitario en la magnitud de las variables explicativas depende de la variable y de la magnitud de su coeficiente.

Conviene mencionar que las interpretaciones realizadas no deben extrapolarse a valores fuera del rango que toman las variables explicativas, porque en esas regiones podrían darse otros efectos distintos del lineal que presupone el modelo estimado.

¹¹Si estos cambian, tendrán su correspondiente impacto en el nivel de concentración de ozono.

15.4.4. Predicción

Aunque el modelo estimado no es adecuado, por la falta de normalidad, linealidad, etc., detectadas, a continuación se ilustra cómo obtener predicciones con la función `predict()`. Para ello, se asignan los valores de interés a las variables explicativas del modelo, con formato `data.frame`, obteniéndose predicciones del valor medio de la variable respuesta, junto con sus intervalos de confianza o predicción, según se proporcione al argumento `interval` los valores `confidence` o `prediction`, respectivamente. En el siguiente ejemplo se obtienen predicciones de niveles de concentración de ozono para un par de casos elegidos arbitrariamente (el primero corresponde a `Solar.R=50`, `Wind=5` y `Temp=62`):

```
nueva_meteo <- data.frame(Solar.R = c(50, 300),
                            Wind = c(5, 17),
                            Temp = c(62, 90))
predict(airq_lm, newdata = nueva_meteo, interval = "confidence")
#>       fit      lwr      upr
#> 1 24.41075 11.01412 37.80739
#> 2 45.62141 31.46838 59.77444
predict(airq_lm, newdata = nueva_meteo, interval = "prediction")
#>       fit      lwr      upr
#> 1 24.41075 -19.662967 68.48448
#> 2 45.62141   1.311914 89.93090
```

Como se puede observar, en ambos casos la predicción puntual (`fit`) es la misma y se obtiene sustituyendo en el modelo estimado los valores de las variables explicativas para los cuales se desea realizar la predicción. Sin embargo, los intervalos de confianza son distintos. Con `confidence` se obtienen intervalos de confianza para el valor medio de las predicciones correspondientes a los días en los que los valores de las variables predictoras sean unos dados. Con `prediction`, el intervalo de confianza es para la predicción de un valor individual, es decir, para la predicción de un día concreto con esas condiciones meteorológicas. Los intervalos de predicción consideran tanto la incertidumbre de la estimación de un valor (debida a la estimación de los parámetros desconocidos) como la variación aleatoria de los valores individuales muestreados (las observaciones muestrales son variables aleatorias). Esto significa que el intervalo de predicción es siempre más ancho que el intervalo de confianza.

15.4.5. Nuevo ajuste con `log(Ozone)`

Ante los problemas de falta de normalidad de la variable `Ozone`, se ajusta un nuevo modelo con la variable `log(Ozone)` como respuesta (véase Sec. 9.3.1). Se aprovecha para introducir una variable *dicotómica* para explicar su interpretación. Se define `Temp_f` dicotomizando `Temp` (tomando sólo dos valores): 1, si la temperatura está por encima de su mediana; y 0, si está por debajo.

```
mediana <- median(airquality$Temp)
Temp_f <- factor(as.numeric(airquality$Temp>mediana))
```

```

lairq_lm <- lm(log(Ozone) ~ Wind + Solar.R + Temp_f, data=airquality)
summary(lairq_lm)
#>
#> Call:
#> lm(formula = log(Ozone) ~ Wind + Solar.R + Temp_f, data = airquality)
#>
#> Residuals:
#>   Min     1Q Median     3Q    Max
#> -2.55347 -0.29689  0.02409  0.37171  1.18373
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept) 3.3879872  0.2232099 15.178 < 2e-16 ***
#> Wind        -0.0885666  0.0161038 -5.500 2.61e-07 ***
#> Solar.R      0.0030723  0.0005973  5.143 1.23e-06 ***
#> Temp_f1      0.6999123  0.1158384  6.042 2.25e-08 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 0.5572 on 107 degrees of freedom
#>   (42 observations deleted due to missingness)
#> Multiple R-squared:  0.5972, Adjusted R-squared:  0.5859
#> F-statistic: 52.89 on 3 and 107 DF,  p-value: < 2.2e-16

```

Al redefinirse la variable respuesta y una variable explicativa del modelo, las estimaciones de los coeficientes cambian respecto al modelo anterior. Todos los coeficientes siguen siendo significativos al 5 % (incluso al 0.1 %), el modelo global también es significativo y el R^2 (ajustado) es similar al del modelo anterior. En el Cap. 17, en el que se abordan los modelos aditivos generalizados, se verá cómo se puede modelar la relación entre `Ozone` y el resto de variables de una forma más satisfactoria. No obstante, este segundo modelo puede ser útil para ilustrar la relación entre las variables, sin olvidar que se ha de comprobar su validez. Para ello, se haría de nuevo el análisis de residuos (que se deja como tarea al lector, al obtenerse de manera idéntica al anterior). En los gráficos de residuos se observará mayor homocedasticidad, linealidad y normalidad que en el caso anterior.

15.4.6. Coeficientes de variables categóricas

A continuación, se aborda la interpretación de coeficientes asociados a variables categóricas. `Temp_f` toma los valores 0 y 1, según la temperatura sea menor o mayor que la mediana respectivamente. En la salida anterior de `R` sólo aparece `Temp_f1`. El 1 final indica que el coeficiente está asociado a la categoría 1 de `Temp_f`. Para los cálculos con estas variables categóricas, `R` toma una categoría como referencia¹² y proporciona un coeficiente para cada una de las restantes categorías, que representa el cambio *medio* al pasar desde la categoría de referencia a cada una de ellas (técnicamente utiliza variables *dummy*) o diferencia entre la media de la variable

¹²La primera “alfanuméricamente”, si no se especifica expresamente el orden con el argumento `levels` en la función `factor()`.

respuesta en las observaciones correspondientes a una categoría específica y a la categoría que sirve de referencia. La categoría de referencia está considerada en los cálculos del término independiente del modelo. Por lo tanto, el coeficiente de `Temp_f1` indica que, *ceteris paribus*, la concentración media de ozono de los días con temperaturas por encima de la mediana (categoría 1) es 0.6999 ppb mayor que la de los días con temperaturas inferiores a ella (categoría 0).

15.4.6.1. Comparativa: regresión frente a ANOVA

En la Fig. 15.1 se pueden apreciar regresiones simples *puras* (variable continua sobre variable continua). Si se regresa la variable `Ozone` sobre la variable categórica `Temp_f`, no se obtendrá un gráfico similar.¹³ No obstante, el gráfico ayudará a comparar visualmente las medias de la variable respuesta en cada categoría. En realidad, al incluir un *factor* en el modelo se está realizando un contraste *t de Student* para averiguar si existen diferencias entre la media de la variable respuesta para cada categoría con respecto a la categoría de referencia. Técnicamente, tales contrastes dos a dos son equivalentes al contraste ANOVA (análisis de la varianza), aunque este permite comparar si las medias de la variable respuesta en todas las categorías son iguales o no. El **ANOVA** es un caso particular de regresión lineal en los parámetros, concretamente, cuando todas las variables explicativas son categóricas.

15.5. Comentarios finales

En capítulos posteriores se abordarán modelizaciones más complejas, como por ejemplo, los modelos lineales generalizados, GLM (Cap. 16), los modelos aditivos generalizados, GAM (Cap. 17) y los modelos mixtos (Cap. 18). También se verán modelos *sparse* y métodos penalizados de regresión (Cap. 19), como la regresión *ridge*, que permite manejar los problemas que genera la presencia de multicolinealidad.

Queda fuera de este capítulo la “consideración” de variables de confusión. Ejemplos típicos de tal tipo de variables son la edad y el sexo. De ser incluidas en el modelo, la magnitud e interpretación de las estimaciones de los coeficientes, las predicciones, etc. pueden ser erróneas, pues el efecto de las variables de confusión puede mezclarse con el de otras variables explicativa incluidas en el modelo (por ejemplo, la influencia de la edad/sexo en enfermedades). Si se incluyen, se podrá obtener el efecto de cualquier variable X_i en la respuesta Y , *ceteris paribus*, es decir, independientemente de los valores y/o categorías de las variables de confusión (edad, sexo, etc.).

También podrían haberse considerado interacciones entre variables, que se suelen interpretar como sinergias o antagonismos. Pero dada la limitación de espacio y el carácter introductorio de este capítulo no se ha considerado oportuno, pues, además, la interpretación de dichas interacciones suele ser compleja. En el Cap. 3 de James et al. (2013) puede encontrarse un ejemplo.

¹³`ggplot(airquality, aes(Temp_f, Ozone)) + geom_point()`.

Resumen

En este capítulo se introduce el modelo de regresión lineal. En particular:

- Se presenta el modelo de regresión lineal múltiple indicando los pasos del análisis de regresión: estimación, validación, interpretación y predicción. La regresión lineal simple se plantea como un caso particular de la múltiple.
- Se muestra el uso de **R** para el ajuste de este tipo de modelos.
- Se presentan diversos casos prácticos para ilustrar la interpretación de los coeficientes de regresión, tanto asociados a variables continuas como a categóricas, la interpretación de las predicciones y el resto de análisis.
- Se mencionan distintos problemas de modelización que el análisis ayuda a detectar, proponiendo a su vez soluciones para solventarlos.

Capítulo 16

Modelos lineales generalizados

Víctor Casero-Alonso^a y María Durbán^b

^aUniversidad de Castilla-La Mancha

^bUniversidad Carlos III de Madrid

16.1. Introducción

Como se ha mencionado en el Cap. 15 de modelización lineal, el objetivo detrás del uso de modelos es el de intentar explicar el comportamiento de una variable en función del comportamiento de otras que se cree que influyen en él. Por ejemplo, podría interesar **predecir**:

- si un empleado abandonará la empresa, o no, en función de sus años de experiencia, su formación, etc.; o si un paciente sufrirá, o no, una enfermedad en función de su edad, sexo, nivel de colesterol, etc.
- el número de días que un empleado puede estar de baja laboral en función del tipo de enfermedad, su antigüedad, salario, etc.; o el número de días que un paciente puede estar hospitalizado en función de la dolencia por la que acude a urgencias, su edad, sexo, etc.

Estos casos no pueden analizarse correctamente con el modelo de regresión lineal múltiple visto en el Cap. 15, porque la **variable respuesta**, la que interesa predecir en cada ejemplo, no sigue una distribución de probabilidad Normal (supuesto necesario para utilizar la regresión lineal) o ni siquiera es continua. Concretamente:

- abandonar, o no, la empresa o sufrir, o no, una enfermedad se puede modelizar mediante una variable **dicotómica/binaria** asignando los valores 0 y 1 a las dos posibles respuestas, lo que encaja perfectamente con una distribución de probabilidad de Bernoulli (muy distinta de la Normal, ya que es discreta, aunque podría parecerse; véase Cap. 12).

- el número de días de baja (en empleados), de hospitalización (en pacientes)... son variables de tipo **recuento** (sólo cero o valores positivos) modelizables mediante una variable discreta que podría seguir una distribución de Poisson (que es también distinta a la Normal, por ser discreta, aunque también podría parecerse).

En este Capítulo se aborda el **modelo lineal generalizado** (GLM), que generaliza el caso en que la variable respuesta sea Normal a cualquier tipo de distribución de probabilidad perteneciente a la familia exponencial y que permite varianzas no constantes en los errores. En concreto, se centra en la regresión logística y la regresión de Poisson, casos particulares de este modelo, que permiten modelizar correctamente los dos casos planteados anteriormente. Un buen libro de referencia para este Capítulo es [James et al. \(2013\)](#).

16.2. El modelo y sus componentes

El **modelo lineal generalizado**¹ se puede escribir como:

$$\mu = g^{-1}(\eta),$$

en el que se tienen los siguientes componentes:

1. $\mu = E(Y)$, el **componente aleatorio**: la **media** de la variable respuesta Y , que puede seguir cualquier distribución de probabilidad de la familia exponencial. Entre ellas están las más habituales: la Normal (por tanto el modelo de regresión lineal es un caso particular del GLM), la Bernoulli/binomial (utilizada en la regresión logística), la Poisson, la gamma, etc.
2. $\eta = X\beta$, el **componente sistemático**, el **predictor lineal**, la “estructura” que aportan las variables explicativas/predictoras $X = (X_1, \dots, X_p)$, que intentan explicar el comportamiento de la variable respuesta, donde $\beta = (\beta_1, \dots, \beta_p)$ es el vector de coeficientes (parámetros) a estimar del modelo.
3. $g(\cdot)$, la novedad de los GLM, la denominada **función de enlace**, que relaciona los dos componentes anteriores. Esta función puede tomar distintas formas, como se verá en la siguiente Sección.

Igual que en el modelo lineal, las dos partes o etapas fundamentales del análisis de un GLM son:

1. La **especificación** de la relación o estructura predefinida de antemano, mediante la *estimación* de los *coeficientes* que mejor ajustan dicha relación, utilizando el **método de máxima verosimilitud**². Más adelante se verá cómo se interpretan tales coeficientes y cómo se puede comprobar la adecuación del modelo.

¹Generalized linear model. Cuidado: general y generalizado no son sinónimos en este contexto.

²A diferencia de la regresión lineal múltiple, que utiliza el de mínimos cuadrados.

16.3. Procedimiento con R: la función `glm()`

261

2. La utilización del modelo estimado (especificado) para **predecir** nuevas respuestas, según sea el caso: valores, probabilidades de ocurrencia, etc.

Para la correcta aplicación de los GLM es crucial la elección tanto de la variable respuesta como de las explicativas (que podrían ser de distinto tipo: numéricas -continuas o discretas- o categóricas/cualitativas -dicotómicas o polítómicas³), así como de la distribución de probabilidad más apropiada para la respuesta. Como se adelantó en el Cap.15, como muestra sirva percatarse de que una misma variable podría ser explicativa o respuesta, por ejemplo “diabetes” (sí o no), según se tenga interés en explicar la influencia de la diabetes en otra variable o la influencia de otras variables en la diabetes. También una misma variable podría considerarse y utilizarse como variable de distinto tipo; por ejemplo, la edad puede considerarse como variable discreta (años cumplidos) o como categórica ordinal (grupos de edad), aunque es una variable continua (puede tomar cualquier valor en un intervalo).

16.2.1. Función enlace

Cada distribución de probabilidad tiene asociada una **función de enlace canónica**⁴:

- Para la Normal es la identidad: $g(\mu) = \mu$.
- Para la Bernoulli, es la función logit: $g(\mu) = \text{logit}(\mu) = \log(\mu/(1 - \mu))$.
- Para la Poisson, es el logaritmo: $g(\mu) = \log(\mu)$.
- Para la Gamma es la inversa: $g(\mu) = 1/\mu, \dots$

Tanto en el caso de la regresión logística (variable respuesta tipo Bernoulli) como en la regresión de Poisson aparece el logaritmo (neperiano) en la función de enlace, lo que conduce a efectos **multiplicativos** de los factores o covariables X_i sobre la respuesta, como se verá más claramente en la Sec. 16.4.3. Este es un punto que las distingue de la regresión lineal, en la que los efectos son **aditivos**.

16.3. Procedimiento con R: la función `glm()`

En el paquete `stats` (de la distribución `base` de R) se encuentra la función `glm()` que se utiliza para llevar a cabo el ajuste de un GLM:

```
glm(formula, family = ..., data = ..., ...)
```

- **formula:** para definir el *predictor lineal*; por ejemplo, $Y \sim X_1 + X_2 + X_3$.
- **family:** para indicar la distribución de la variable respuesta (`gaussian`, `binomial`, `poisson` ...) que determina la función de enlace (`binomial` → `logit`, etc.; consultese `?family` o `?glm` para más detalles).

³En este contexto, las variables numéricas y categóricas se denominan **factores** y **covariables**, respectivamente.

⁴Podrían considerarse otras para cada distribución, pero esta cuestión excede el nivel de este Capítulo.

Las “herramientas” utilizadas para `lm()` también se pueden usar para `glm()` (aunque algunas interpretaciones varían). Así, se puede usar `summary()` para detectar los predictores importantes, `fitted()` para obtener los valores ajustados, etc.

16.4. Regresión logística

La **regresión logística** es el caso más “famoso” de GLM, de gran relevancia en distintos contextos: Medicina, Economía, etc. Se utiliza cuando la **variable respuesta** es **dicotómica**, del tipo pertenencia, o no, a un determinado grupo (fumadores, enfermos, morosos, ...). Habitualmente se considera que toma el valor $Y = 1$ si la observación pertenece al grupo de interés e $Y = 0$ en caso contrario. Tal tipo de variable se puede modelizar con una **distribución de Bernoulli**, caracterizada por un parámetro p que indica la probabilidad de pertenecer al grupo de interés.

El objetivo principal suele ser predecir el grupo al que pertenece un nuevo individuo/elemento, sobre la base de la información sobre dicho elemento/individuo que proporcionan las variables explicativas. Para ello, se estima el modelo con los datos disponibles, determinándose qué variables **significativamente** influyen en la variable respuesta⁵.

¿Por qué no tiene cabida aquí el uso del modelo de regresión lineal múltiple? Al ajustar un modelo del tipo: $Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \epsilon$, las estimaciones \hat{Y} serán números reales que rara vez coincidirán con 0 ó 1 (que son los valores admisibles de la variable respuesta). Dicho de otro modo, si sólo se tuviese una variable explicativa, al ajustar el modelo de regresión lineal simple, la recta sobrepasaría, o no alcanzaría, los valores posibles de respuesta (0 ó 1), como ocurre en la Fig. 16.1 (izquierda) obtenida a partir de los datos del ejemplo de enfermedad coronaria que se manejará en la Sec. 16.6.1.

El **modelo de regresión logística múltiple** se define como:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \epsilon, \quad (16.1)$$

que permite estimar la ratio entre la probabilidad de pertenecer al grupo de interés, p , y la de no pertenecer a dicho grupo, $1-p$. Utilizando la función de enlace se puede transformar el predictor lineal $\eta = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$ para obtener valores admisibles para una probabilidad. La función logística, definida como

$$p = \frac{e^\eta}{1 + e^\eta} \quad (16.2)$$

y representada en la Fig. 16.1 (derecha), es la más habitual, y por ello da el nombre a la regresión logística. Con ella se obtiene la probabilidad de pertenecer al grupo de interés, $p = P[Y = 1]$, e inmediatamente la de no pertenecer a dicho grupo, $1 - p = P[Y = 0]$.

La ventaja del modelo logístico es que la función logística tiende a cero por la izquierda y a uno por la derecha, por lo que la predicción será siempre un valor válido para una probabilidad. Como contrapartida, a medida que las probabilidades se acercan a cero o a uno la relación entre el predictor y la probabilidad deja de ser lineal, lo que complica la interpretación de los coeficientes.

⁵Desde el punto de vista estadístico, la influencia/efecto no es fruto del azar.

16.4. Regresión logística

263

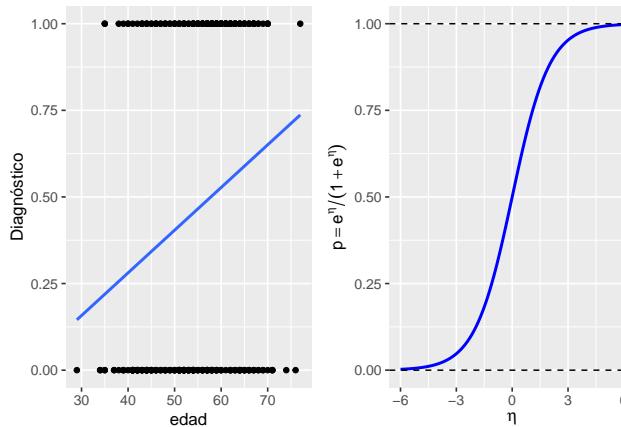


Figura 16.1: Gráfico de dispersión del diagnóstico frente a la edad para el ejemplo de enfermedad coronaria (izquierda) y función logística (derecha)

16.4.1. Procedimiento de ajuste

A partir de los datos disponibles de las variables *predictoras* y *respuesta*:

1. Se estiman los coeficientes β_i del modelo, para especificar la relación entre las variables, contrastando si tales estimaciones, $\hat{\beta}_i$, son o no significativas.
2. Se valora la eliminación de variables no significativas, atendiendo a la significación global del modelo, al modelo teórico subyacente, etc.
3. Se comprueba la adecuación del modelo final obtenido.
4. En caso afirmativo, se interpretan los coeficientes y el modelo queda listo para hacer **predicciones** de probabilidades o para **clasificar** individuos/elementos.

Nota

Como se menciona en el Cap. 15, un coeficiente es *significativo* cuando su *p-valor* asociado es lo *suficientemente pequeño* (como norma general, inferior a 0.05). No obstante, pueden tomarse otros valores de referencia; por ejemplo, en las salidas de **R** aparecen otros tres niveles de referencia, 0.1, 0.01 y 0.001.

16.4.2. Adecuación del modelo

Para comprobar si el modelo estimado es adecuado, se compara con el modelo más simple, el que solo incluye el término independiente. Para ello, se pueden utilizar distintos contrastes basados en la **deviance**,⁶ una medida que juega el papel de la suma de cuadrados de los

⁶En ocasiones traducida como *devianza*, aunque es habitual no traducirla.

residuos. En el modelo de regresión logística, la *deviance* de un modelo es menos dos veces el logaritmo de la **verosimilitud** de dicho modelo. La diferencia entre la *deviance* de un modelo más elaborado y el simple se distribuye como una χ^2 con tantos grados de libertad como restricciones impuestas a los parámetros, lo que permite contrastar cuál de los dos modelos ajusta mejor los datos. P-valores bajos indican que el modelo ajustado es inadecuado, debiendo investigarse otras posibles variables predictoras, si se incumple la hipótesis de linealidad o existe sobredispersión (por ejemplo por exceso de ceros).

Adicionalmente, se puede aplicar el contraste de la **razón de verosimilitudes**. Este test contrasta la significación de cada variable predictora, basándose en la *deviance* que se genera al añadir cada variable secuencialmente al modelo que contiene las anteriores, lo que ayuda a decidir si mantenerla o eliminarla del modelo.

Para contrastar la bondad de ajuste del modelo, el contraste más popular en la literatura es el de *Hosmer-Lemeshow*, aplicable a modelos con al menos una variable cuantitativa. P-valores bajos indican falta de ajuste.

Continuando con las medidas de bondad de ajuste, en el modelo de regresión logística no tiene sentido calcular el coeficiente de determinación lineal, R^2 , pero existen varias alternativas equivalentes para hacerse una idea de la variabilidad de la respuesta explicada por el modelo. Las tres más populares son el Pseudo R^2 de McFadden, el R^2 de Cox y Snell (que por construcción no puede alcanzar el 1) y el R^2 de Nagelkerke (una corrección del de Cox y Snell).

16.4.3. Interpretación de resultados

La interpretación de los coeficientes no es tan directa y sencilla como en el modelo lineal. A partir de las estimaciones del predictor lineal $\hat{\eta}$, modelo (16.1), se puede estimar la probabilidad de que un individuo pertenezca al grupo de interés utilizando la expresión (16.2). Alternativamente se puede estimar el *odds*, con la siguiente expresión, que se deduce de (16.1):

$$\text{odds} = \frac{\hat{p}}{1 - \hat{p}} = e^{\hat{\eta}} = e^{\hat{\beta}_0 + \hat{\beta}_1 X_1 + \dots + \hat{\beta}_k X_k}. \quad (16.3)$$

El *odds* se puede explicar como cuántas veces es más probable pertenecer al grupo de interés ($Y = 1$) que al otro grupo ($Y = 0$). Por ejemplo, si $\hat{p} = 0,75$ el *odds* es $0,75/0,25 = 3$, entonces es tres veces más probable pertenecer al grupo de interés que no pertenecer⁷.

La ecuación (16.3) se puede expresar, equivalentemente, como:

$$\text{odds} = e^{\hat{\beta}_0} \cdot e^{\hat{\beta}_1 X_1} \cdot \dots \cdot e^{\hat{\beta}_k X_k}. \quad (16.4)$$

La interpretación de los $\hat{\beta}_i$ carece de sentido. Lo interpretable son los $e^{\hat{\beta}_i}$, denominados **odd ratios** (OR). Como el modelo (16.4) es multiplicativo (sus términos aparecen multiplicando, no sumando), por lo que valores de $e^{\hat{\beta}_i}$ inferiores a 1 implican disminución en el valor del odds –se reduce la probabilidad de pertenecer al grupo de interés respecto a la de no pertenecer– y

⁷Por esto se traduce *odds* como *ventaja*. Pero cuidado con su uso: si $Y = 1$ es tener un accidente, no parece muy adecuado hablar de ventaja. Por esto mismo, a p tampoco se le denomina aquí probabilidad de éxito.

valores por encima de 1 implican incremento del valor del odds, *ceteris paribus*⁸. En concreto, $100(e^{\hat{\beta}_i} - 1)$ indica la variación porcentual en el *odds* ante un incremento unitario en la variable explicativa X_i , mientras que $100(e^{-\hat{\beta}_i} - 1)$ indica el cambio porcentual que se opera en el *odds* debido a un decremento unitario de dicha variable explicativa, *ceteris paribus*.

El modelo de regresión logística se acompaña del denominado **riesgo relativo** (RR, *relative risk*). Es una razón de probabilidades como el *odds*, pero en vez de comparar la probabilidad de pertenecer a un grupo o a otro de la variable respuesta, compara las probabilidades de pertenecer al grupo de referencia ($Y = 1$) según los valores del predictor categórico X_i . El riesgo relativo es similar al OR de la variable X_i sólo cuando la probabilidad de $Y = 1$ es pequeña; se suele dar como referencia que sea inferior al 10%. Por lo tanto, OR y RR no coinciden siempre.

16.4.4. Predicción. Curva ROC y AUC

Como se ha mencionado anteriormente, el uso habitual de la regresión logística es la predicción, no tanto de probabilidades, sino de la clasificación en un grupo u otro de futuros individuos/elementos en base a la información conocida de ellos a través de las variables explicativas. La regla de clasificación en los grupos depende de la elección del punto de corte de la probabilidad, por ejemplo $\hat{Y} = 1$ si $\hat{p} > 0,7$ e $\hat{Y} = 0$ en otro caso. Para seleccionar dicho punto de corte, se acude al análisis de los casos clasificados correctamente o no por el modelo ajustado: análisis de la sensibilidad y la especificidad del modelo, a su visualización más popular, la **curva ROC** (*receiver operating characteristic*) y la medición del área bajo dicha curva, AUC (*area under the curve ROC*); véase Cap. 9.

La **sensibilidad**, o tasa de verdaderos positivos, se define como el cociente entre estos y la suma de los verdaderos positivos y los falsos negativos. La **especificidad**, o tasa de verdaderos negativos, es el cociente entre estos y la suma de los verdaderos negativos y los falsos positivos. Por consiguiente, la sensibilidad es una medida de la probabilidad de que un caso real positivo ($Y = 1$) sea clasificado correctamente como positivo por el modelo ($\hat{Y} = 1$). Equivalentemente, la especificidad es la probabilidad de clasificar correctamente casos negativos.

Gráficamente, se pueden obtener los valores de la sensibilidad y especificidad para distintos puntos de corte (entre 0 y 1). La Fig. 16.2 (izquierda) muestra este gráfico para uno de los casos prácticos⁹. Pero la forma más popular de analizar la sensibilidad y la especificidad es mediante la curva ROC, que representa, para los distintos puntos de corte considerados, la *sensibilidad* frente a la tasa de falsos positivos, esto es la *1-especificidad* (Fig. 16.2 (derecha)). Su AUC sirve para comparar distintos modelos de regresión logística (u otros modelos con el mismo fin¹⁰). Cuanto mayor poder discriminante tenga el modelo, más próxima a la unidad estará la AUC. Un clasificador aleatorio presentaría la curva ROC coincidente con la diagonal, con una AUC de 0.5.

⁸Manteniendo constante el valor de las demás variables.

⁹La interpretación del gráfico se hace en la sección donde aparece la figura.

¹⁰Véase el Cap. 9

16.5. Regresión de Poisson

Se utiliza cuando la **variable respuesta** es **discreta**, sólo toma valores no negativos y su distribución de probabilidad es modelizable mediante la **distribución de Poisson**. Por ejemplo, las variables de tipo “número de”, como la del ejemplo motivador del principio: número de días de hospitalización. El parámetro que caracteriza la distribución de Poisson es λ , que representa tanto la media como la varianza de la variable aleatoria (a nivel teórico).

Dados los anteriores condicionantes, en esta tesitura tampoco tiene cabida el modelo de regresión lineal, principalmente porque las estimaciones \hat{Y} podrían arrojar valores negativos.

De nuevo el objetivo suele ser la predicción: en el ejemplo, el número de días que un (nuevo) paciente estará hospitalizado ($0, 1, 2, \dots$), en base a la información proporcionada por las variables explicativas. Previo a la predicción, se estima el modelo, a partir de los datos disponibles, para determinar las variables explicativas que influyen significativamente sobre la variable respuesta. Y, nuevamente, los efectos serán *multiplicativos*, dado que la función de enlace es de tipo logarítmico, como se ha visto en la Sec. 16.4.3. Por ello, todo lo visto anteriormente para la regresión logística es válido para la regresión de Poisson.

16.6. Casos prácticos

16.6.1. Ejemplos de regresión logística

Para llevar a cabo estos ejemplos, se utiliza el conjunto de datos `cleveland` incluido en el paquete CDR que acompaña este libro. Se quiere explicar la variable `diag` (diagnóstico; dicotómica: 1, ha sufrido una enfermedad coronaria; 0, no la ha sufrido) a partir de otras variables. La variable respuesta `diag` ya aparece como factor en la base de datos.

Estimación

Primeramente, se considera un primer modelo con dos variables explicativas continuas `edad` y `dep` (depresión en el segmento ST):

```
library("CDR")
reg_log <- glm(diag ~ edad + dep,
                family = "binomial", data = cleveland)
summary(reg_log)
#>
#> Call:
#> glm(formula = diag ~ edad + dep, family = "binomial", data = cleveland)
#>
#> Deviance Residuals:
#>      Min        1Q     Median        3Q       Max
#> -2.3804   -0.8905   -0.6210    1.0021    1.9644
#>
#> Coefficients:
#>             Estimate Std. Error z value Pr(>|z|)
#>
```

16.6. Casos prácticos

267

```
#> (Intercept) -3.08080   0.81939  -3.760  0.00017 ***
#> edad         0.03738   0.01476   2.532  0.01134 *
#> dep          0.86851   0.13791   6.298 3.02e-10 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> (Dispersion parameter for binomial family taken to be 1)
#>
#> Null deviance: 417.98 on 302 degrees of freedom
#> Residual deviance: 350.64 on 300 degrees of freedom
#> AIC: 356.64
#>
#> Number of Fisher Scoring iterations: 4
```

En la salida se muestran las estimaciones de los coeficientes del modelo (columna `Estimate`) y su significatividad (columna `Pr(>|z|)`). En este ejemplo, los dos coeficientes de interés (los correspondientes a `edad` y `dep`) son significativos (a un nivel de significación del 5 %).

Puede observarse que, dicha salida, también incluye la *null deviance* y la *residual deviance*. La primera indica lo “bien” que predice el modelo sin variables explicativas (tan sólo con el término independiente o intercepto); la segunda indica lo “bien” que predice el modelo con variables explicativas. Como se avanzó anteriormente, un contraste Chi-cuadrado permitirá dilucidar si el modelo con variables explicativas predice significativamente mejor que el que solo tiene un término independiente y su predicción, sea cual sea el valor de las variables explicativas, es siempre la media de los valores de la variable respuesta.

Se completa el modelo anterior añadiendo variables categóricas –concretamente, `tdolor` (polítómica, tipo de dolor) y `sexo` (dicotómica)– para poder incluir más adelante la interpretación de los coeficientes asociados a este tipo de variables. Para su correcta interpretación, se deben introducir en el predictor como variables de tipo `factor`¹¹. De lo contrario, el procedimiento las considera numéricas, obteniéndose una salida que llevaría a una interpretación errónea.

```
reg_log2 <- update(reg_log, ~ . + sexo + tdolor)
```

Adecuación del modelo

Para evaluar la bondad de ajuste de los modelos se lleva a cabo el contraste de Hosmer-Lemeshow:

```
library("ResourceSelection")
hoslem.test(reg_log$y, reg_log$fitted.values)
#>
#> Hosmer and Lemeshow goodness of fit (GOF) test
#>
#> data: reg_log$y, reg_log$fitted.values
#> X-squared = 5.631, df = 8, p-value = 0.6885
```

¹¹En este caso ya están definidas como `factor` en el `data.frame`.

```
hoslem.test(reg_log2$y, reg_log2$fitted.values)
#>
#> Hosmer and Lemeshow goodness of fit (GOF) test
#>
#> data: reg_log2$y, reg_log2$fitted.values
#> X-squared = 4.8171, df = 8, p-value = 0.7769
```

En ambos casos, los p-valores son “altos”, indicando un ajuste suficiente.

Para realizar el test de la razón de verosimilitudes con **R** se acude a la función **anova()**:

```
anova(reg_log2, test = "Chisq")
#> Analysis of Deviance Table
#>
#> Model: binomial, link: logit
#>
#> Response: diag
#>
#> Terms added sequentially (first to last)
#>
#>
#>          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
#> NULL           302     417.98
#> edad          1   15.447    301     402.54 8.487e-05 ***
#> dep            1   51.894    300     350.64 5.858e-13 ***
#> sexo           1   23.982    299     326.66 9.726e-07 ***
#> tdolor         3   62.153    296     264.51 2.037e-13 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Las *deviances* correspondientes a añadir secuencialmente cada variable o factor al modelo que contiene los anteriores permite concluir que dicha inclusión secuencial de tales variables predictoras es significativa respecto a los modelos que no las incluyen.

Para completar, se podría contrastar también el efecto de una variable sobre la respuesta comparando la *deviance* del modelo con dicha variable y sin ella:

```
## se elimina edad del segundo modelo
reg_log3 <- update(reg_log2, ~ . - edad)
anova(reg_log3, reg_log2, test = "Chisq")
#> Analysis of Deviance Table
#>
#> Model 1: diag ~ dep + sexo + tdolor
#> Model 2: diag ~ edad + dep + sexo + tdolor
#> Resid. Df Resid. Dev Df Deviance Pr(>Chi)
#> 1      297     274.45
#> 2      296     264.51  1    9.9488  0.00161 **
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

16.6. Casos prácticos

269

El resultado indica que la variable `edad` se considera un predictor significativo en el modelo. Se deja al lector realizar este ejercicio con las restantes variables, que le llevarán a concluir que todas son predictores significativos.

El valor del Pseudo R^2 de McFadden se obtiene como sigue:

```
null <- glm(diag ~ 1, family = "binomial", data = cleveland)
1 - logLik(reg_log) / logLik(null)
#> 'log Lik.' 0.1611088 (df=3)
1 - logLik(reg_log2) / logLik(null)
#> 'log Lik.' 0.3671819 (df=7)
```

Para el primer modelo el Pseudo R^2 de McFadden es 0.16, mientras que para el segundo es 0.37.

Interpretación de los coeficientes

A continuación se muestran los coeficientes estimados del segundo modelo y sus correspondientes OR (sus exponentiales) que son los interpretables:

```
summary(reg_log2)
#>
#> Call:
#> glm(formula = diag ~ edad + dep + sexo + tdolor, family = "binomial",
#>      data = cleveland)
#>
#> Deviance Residuals:
#>      Min        1Q     Median        3Q       Max
#> -2.5346  -0.6604  -0.2436   0.6568   2.3975
#>
#> Coefficients:
#>             Estimate Std. Error z value Pr(>|z|)
#> (Intercept) -6.67669   1.30016 -5.135 2.82e-07 ***
#> edad         0.05621   0.01828  3.075  0.0021 **
#> dep          0.80890   0.16597  4.874 1.10e-06 ***
#> sexo1        1.69477   0.36801  4.605 4.12e-06 ***
#> tdolor2      0.65668   0.67357  0.975  0.3296
#> tdolor3      0.19465   0.59654  0.326  0.7442
#> tdolor4      2.58230   0.57549  4.487 7.22e-06 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> (Dispersion parameter for binomial family taken to be 1)
#>
#> Null deviance: 417.98 on 302 degrees of freedom
#> Residual deviance: 264.51 on 296 degrees of freedom
#> AIC: 278.51
#>
#> Number of Fisher Scoring iterations: 5
exp(coef(reg_log2))[-1]
```

```
#>      edad      dep     sexo1    tdolor2    tdolor3    tdolor4
#> 1.057824 2.245432 5.445391 1.928371 1.214880 13.227480
```

Los coeficientes asociados a las variables continuas son significativos y positivos, por tanto sus OR son significativamente superiores a 1. Indican que, *ceteris paribus*,

- por cada año más, el odds de padecer la enfermedad frente a no padecerla se incrementa¹² un 5.78% (al ser el valor del OR asociado a la edad 1.0578). ¿Y cuál sería el incremento en el *odds* ante un aumento de la edad en 20 años? En tal caso, el odds se multiplicaría por $e^{\hat{\beta}_1 \cdot 20} = 3.078$, es decir, más que se triplica. Se debe ser cuidadoso con la interpretación: no se triplica la probabilidad sino el odds.
- Ante un aumento de una unidad en la variable *dep* más que se duplica el *odds* del paciente, pues su OR es 2.2454.

Nota

En regresión lineal, que un coeficiente sea significativo quiere decir que es significativamente distinto de cero, por el hecho de que el modelo considera *efectos aditivos* de las variables. El modelo de regresión logística es de *efectos multiplicativos* y, por ello, el valor “neutro” es el 1. Por tanto, un coeficiente significativo se interpreta como significativamente distinto de 1.

¿Y cómo se interpretan los coeficientes asociados a las variables categóricas? ¿Qué significan los valores de *sexo1*, *tdolor2*...? Representan el cambio (promedio) en la variable respuesta al pasar de la categoría de referencia a la mostrada.¹³

- el de *sexo1* significa que el odds de padecer la enfermedad en los hombres (*sexo=1*) es más de 5 veces superior que en las mujeres (*sexo=0*).
- el de *tdolor4* indica que los individuos con este tipo de dolor (asintomáticos) presentan un odds 13 veces superior al de los de *tdolor=1* (angina típica).¹⁴

Nótese que no se muestra e^{β_0} , pues su interpretación carece de sentido práctico: sería el OR de las personas de 0 años, *dep=0*, *sexo=0* y *tdolor=1*.

Predicciones

Una vez estimado el modelo y comprobada su bondad, se puede usar para obtener predicciones. Para ello, se deben asignar valores a las variables explicativas: *edad*, *dep*, *sexo* y *tdolor* (se han escogido arbitrariamente).

¹²Por ser el coeficiente mayor que 1.

¹³En las variables definidas como **factor** R toma como referencia la primera categoría al ordenarse los valores de la variable, bien alfabéticamente (a, b, c...) o bien numéricamente, de menor a mayor, salvo que se haya especificado otro orden. De ahí que *sexo=0* sea la categoría de referencia, porque esta variable toma los valores 0 y 1, y sólo aparezca en las estimaciones *sexo1*, reflejando el cambio medio en la variable respuesta al pasar de la categoría de referencia a la categoría 1. Para *tdolor*, al tomar los valores 1, 2, 3 y 4, la categoría de referencia es *tdolor=1*, apareciendo coeficientes para *tdolor=2*, *tdolor=3* y *tdolor=4*.

¹⁴Los coeficientes de *tdolor2* y *tdolor3* no son significativos.

```
individuo <- data.frame(edad = 50, dep = 3,
                         sexo = "1", tdolor = "4" #entre comillas por ser factores
)
```

Con la función `predict()` se puede obtener tanto el valor predicho para el predictor lineal (valor no interpretable), η , como directamente la probabilidad de que dicho individuo sufra la enfermedad coronaria, p :

```
#(eta <- predict(reg_log2, individuo))
(p <- predict(reg_log2, individuo, type = "response"))
#>      1
#> 0.9446843
```

A partir del valor de p (o de η) se puede obtener el *odds* correspondiente:

```
p / (1 - p) # exp(eta)
#>      1
#> 17.07805
```

Es decir, un individuo con 50 años, `dep=3`, `sexo=1` (hombre) y `tdolor=4` (asintomático) tiene una probabilidad de padecer la enfermedad en cuestión 17 veces mayor que de no padecerla.

Riesgo relativo

Se ha visto que el OR para la categoría hombre en el segundo modelo es 5.45. El riesgo relativo de sufrir la enfermedad, en el caso de los varones, para un paciente de 50 años, valor `dep=3` y `tdolor=4` (asintomático) es:

```
hom <- data.frame(edad = 50, dep = 3, sexo = "1", tdolor = "4")
muj <- data.frame(edad = 50, dep = 3, sexo = "0", tdolor = "4")
(ph <- predict(reg_log2, hom, type = "response"))
#>      1
#> 0.9446843
(pm <- predict(reg_log2, muj, type = "response"))
#>      1
#> 0.7582346
ph / pm # riesgo relativo hombre/mujer
#>      1
#> 1.2459
```

El riesgo relativo de sufrir la enfermedad, en el caso de los varones, es $RR=0.9447/0.7582 = 1.2459$, es decir, un varón tiene un 24.6 % más de posibilidades de tener la enfermedad que una mujer con sus mismos valores o categorías en las variables que se usan como predictores. En este caso, el OR y el RR no son valores cercanos; ello es debido a que el diagnóstico 1 es frecuente (concretamente lo presentan el 45.9 % de los pacientes de la base de datos).

Curva ROC

El análisis de sensibilidad y especificidad del modelo, así como la curva ROC y su AUC, para este ejemplo, se pueden obtener con el siguiente código:

```
par(mfrow = c(1, 2))
library("Epi")
ROC(
  form = diag ~ edad + dep + sexo + tdolor, data = cleveland,
  plot = "sp"
)
ROC(
  form = diag ~ edad + dep + sexo + tdolor, data = cleveland,
  plot = "ROC", las = 1
)
```

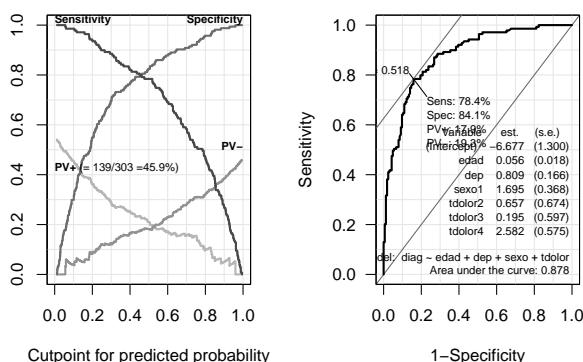


Figura 16.2: Gráfico de sensibilidad y especificidad según puntos de corte de discriminación (izquierda) y curva ROC (derecha) para el segundo modelo de regresión logística

La Fig. 16.2 (izquierda) muestra que tomando como punto de corte una probabilidad cercana a 0.45 se obtienen valores de sensibilidad y especificidad alrededor de 0.8. No obstante, se deben evaluar los costes y riesgos de una mala clasificación (por ejemplo, dar tratamiento cuando no hace falta y no darlo cuando es necesario).

La Fig. 16.2 (derecha) representa la curva ROC de `reg_log2`. La curva está por encima de la diagonal, con lo que es mejor que un clasificador aleatorio. La AUC es de 0.878, con una sensibilidad de 78.4% y una especificidad de 84.1%. Si se realiza el análisis para `reg_log` se comprobará que la AUC es de 0.759, por lo que el segundo modelo es mejor para discriminar. Además, el gráfico también indica el valor del **punto de corte óptimo**, 0.518 (el que proporciona la mayor AUC, 0.878) junto con los correspondientes valores de sensibilidad, especificidad, etc. También muestra las estimaciones y los errores estándar (*s.e.*) de los coeficientes del modelo considerado.

16.6.2. Ejemplo de regresión de Poisson

A partir del mismo conjunto de datos, `cleveland`, ahora se considera como variable a explicar, variable respuesta, `dhosp`, el número de días de hospitalización de un paciente. Como variables explicativas se seleccionan las siguientes: `diag`, `edad`, `sexo` y `tdolor`, que son de distinto tipo, lo que permitirá ilustrar sus distintas interpretaciones.

Visualización ilustrativa

```
p1 <- ggplot(cleveland, aes(dhosp, fill = diag)) +
  geom_bar(position = "dodge")
p2 <- ggplot(cleveland, aes(x = factor(dhosp), y = edad)) +
  geom_boxplot()
library("patchwork")
p1 + p2
```

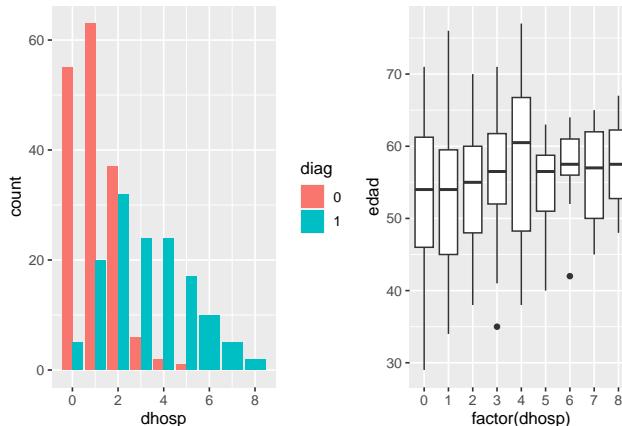


Figura 16.3: Gráfico de barras de ‘dhosp’ según tipo de diagnóstico (izquierda) y gráficos de cajas de ’edad’ según el número de días de hospitalización (derecha)

A la vista de los gráficos de la Fig. 16.3, generados con las sentencias anteriores, `diag` parece ser una buena variable predictora del número de días de hospitalización, mientras que `edad` no tiene una influencia tan clara.

Ajuste e interpretación

```
reg_pois <- glm(dhosp ~ diag + edad + sexo + tdolor,
  data = cleveland, family = "poisson")
```

Al especificar `family = "poisson"`, la función `glm()` selecciona automáticamente la función de enlace apropiada: el logaritmo (efectos multiplicativos).

```
summary(reg_pois)$coef
#>             Estimate Std. Error   z value    Pr(>|z|)
#> (Intercept) -0.2230728726 0.332125111 -0.67165314 5.018045e-01
#> diag1        1.0902902179 0.109731744  9.93595999 2.903654e-23
#> edad         0.00010262777 0.0048771904 0.02103931 9.832143e-01
#> sexo1        0.2160286255 0.103297627  2.09132226 3.649919e-02
#> tdolor2      0.1408464230 0.205232539  0.68627725 4.925383e-01
#> tdolor3      0.1217757202 0.189404790  0.64293897 5.202637e-01
#> tdolor4      0.1215388103 0.178729482  0.68001546 4.964947e-01
exp(coef(reg_pois))[-1]
#> diag1       edad     sexo1   tdolor2   tdolor3   tdolor4
#> 2.975137 1.000103 1.241138 1.151248 1.129501 1.129233
```

De todas las variables introducidas en el modelo, sólo `diag` y `sexo` son significativas (al 5%). Se confirma así lo visto en la Fig. 16.3 (derecha): que `edad` no influye en la respuesta. Al ser las dos variables significativas de tipo dicotómico, su interpretación, como en la regresión logística, se hace en función de la categoría de referencia, y *ceteris paribus* (esto en cualquier caso):

- El coeficiente de `diag` es 1.0903 pero se le debe aplicar la exponencial para tener como unidades de medida días. Así, el número medio de días de estancia en el hospital es 2.98 veces mayor con `diag=1` que con `diag=0` (categoría tomada como referencia).
- Algo similar se puede decir para `sexo`: un hombre (`sexo=1`) se espera que esté en el hospital, en media, 1.24 días por cada día que esté una mujer.
- De la misma manera se interpretarían los coeficientes asociados a `tdolor`, si bien no son significativos: cada uno de ellos expresaría la diferencia con los pacientes del grupo `tdolor=1`.
- Aunque el coeficiente asociado a `edad` tampoco es significativo, y su magnitud es ínfima, se da su interpretación, al ser la única variable continua: por cada año que aumenta la `edad`, el número medio de días en el hospital se ve multiplicado por 1.0001 (¹⁵ínfimo).

Para posteriores comparaciones, se considera otro modelo en el que se elimina `tdolor` (por no ser significativa):

```
reg_pois2 <- glm(dhosp ~ diag + sexo + edad,
  data = cleveland, family = "poisson")
```

Adecuación

En los ajustes `reg_pois` y `reg_pois2` hay información sobre la *Null* y la *Residual deviance*, con las que se puede realizar el contraste de comparación de modelos (el simple frente al elaborado) mencionado en la Sec. 16.4.2.

¹⁵Si el incremento fuese de 20 años (como en el ejemplo de regresión logística), el incremento en los días de hospitalización seguiría siendo despreciable: 0,0021 días.

```
pchisq(reg_pois$deviance, reg_pois$df.residual, lower.tail = F)
#> [1] 0.1325654
pchisq(reg_pois2$deviance, reg_pois2$df.residual, lower.tail = F)
#> [1] 0.155157
```

Al ser los p-valores superiores a 0.05, ambos modelos se pueden considerar que explican “mejor” que el modelo nulo, el que únicamente contiene la constante y el término aleatorio.

Predicción

Con los modelos ajustados, y comprobada su adecuación, se pueden predecir valores, que en este caso son el número medio de días de hospitalización.

```
pacientes <- data.frame(
  diag = c("1", "1", "0", "0"),
  edad = c(50, 50, 50, 50),
  sexo = c("1", "0", "1", "0"),
  tdolor = c("4", "4", "4", "4"))
predict(reg_pois, pacientes, type = "response")
#>      1         2         3         4
#> 3.3532035 2.7017171 1.1270752 0.9080983
```

Se han escogido, arbitrariamente, valores de las variables explicativas para 4 `pacientes`: los pacientes 1 y 2 presentan la enfermedad coronaria (`diag=1`) mientras que los pacientes 3 y 4 no; los pacientes 1 y 3 son hombres (`sexo=1`), mientras que el 2 y el 4 son mujeres; los cuatro tienen 50 años y son asintomáticos (`tdolor=4`). Las predicciones obtenidas indican que el paciente 1 (hombre con enfermedad coronaria) estará hospitalizado más días, en media, que el resto, aunque le sigue de cerca la paciente 2 (mujer con enfermedad coronaria).

También se pueden dibujar las predicciones modelo `reg_pois` de todo el conjunto de datos, que `glm()` ha guardado como `fitted.values`.

```
cleveland$hat <- reg_pois$fitted.values
g1 <- ggplot(cleveland, aes(x = edad, y = hat, colour = diag)) +
  geom_point() +
  labs(x = "Edad", y = "Días hospitalización")
cleveland$hat2 <- reg_pois2$fitted.values
g2 <- ggplot(cleveland, aes(x = edad, y = hat2, colour = diag)) +
  geom_point() +
  labs(x = "Edad", y = "Días hospitalización")
g1 + g2
```

La Fig. 16.4 muestra las predicciones del número de días (promedio) de hospitalización para los 303 individuos considerados en el conjunto de datos y para cada uno de los dos modelos de regresión de Poisson considerados.

Cabe recordar que en el primer modelo se consideran las variables `diag`, `edad`, `sexo` y `tdolor`, mientras que en el segundo se ha eliminado la variable `tdolor`. Lo primero que hay que destacar

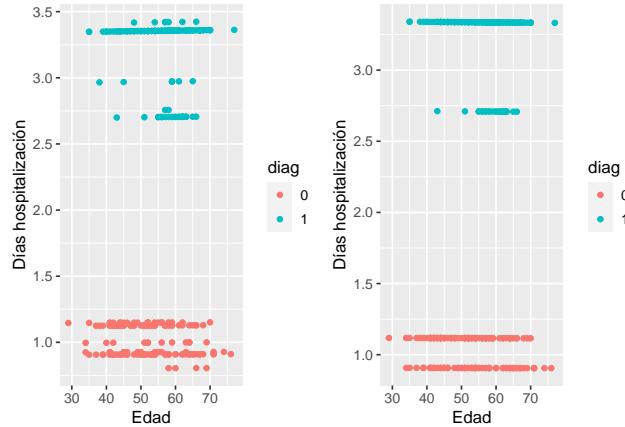


Figura 16.4: Predicciones de todo el conjunto de datos del modelo ‘regpois’ (izquierda) y ‘regpois2’ (derecha)

es que los puntos se “ordenan” en forma de lineas de puntos horizontales ordenadas verticalmente. Ello se debe a que (i) en el eje de abscisas figura la variable `Edad`, que no influye significativamente en los días de hospitalización (lineas horizontales); y (ii) las demás variables son categóricas, por lo que desplazan las líneas verticalmente, sin modificar su pendiente. El desplazamiento vertical diferenciado por colores, mediante la variable `diag`, es claro: la predicción del número de días (promedio) de hospitalización para personas con `diag=1` es un valor en torno a 3, mientras que para aquellos con `diag=0` es un valor en torno a 1. En ambos gráficos no se ha incluido identificación para la variable sexo, podría hacerse añadiendo apropiadamente `facet_wrap(vars(sexo))` lo que permitiría ver que es el sexo el que genera las diferencias verticales (significativas, porque la variable `sexo` es significativa) de las dos líneas horizontales tanto en torno a `hosp=3` como a `hosp=1` (apreciables claramente en el gráfico de la derecha). Por último, como la diferencia entre el gráfico de la izquierda y el de la derecha es la inclusión o no de `tdolor` queda claro que es dicha variable la que genera las distintas líneas horizontales de puntos (gráfico de la izquierda) con poca variabilidad vertical (por no ser significativa).

Resumen

En este capítulo se introduce el modelo de regresión lineal generalizado, indicado cuando las respuestas no son gaussianas (Normales). En particular:

- Se introduce la función de enlace, que juega un papel importante en estos modelos.
- Se describen los casos particulares de regresión logística y de regresión de Poisson.
- Se muestra el uso de **R** para el ajuste de estos modelos.
- Se ilustra la interpretación de los coeficientes, tanto asociados a variables cuantitativas como a categóricas, y los demás resultados obtenidos con **R**, mediante casos prácticos.
- Se incluye el uso de la regresión logística como clasificador y se dan indicaciones al respecto mediante la curva ROC y la AUC.

Capítulo 17

Modelos aditivos generalizados

^aUniversidad de Castilla-La Mancha

^bUniversidad Carlos III de Madrid

17.1. Introducción

Los modelos lineales, o los lineales generalizados (GLM), vistos en los capítulos 15 y 16 tienen la ventaja de ser fáciles de ajustar e interpretar. Además, se dispone de técnicas para contrastar las hipótesis del modelo. Sin embargo, cuando la variable respuesta no está relacionada de forma lineal con las variables explicativas no tiene sentido utilizar modelos lineales (generalizados o no) y hay que acudir a modelos que flexibilicen esta relación, que, en el caso de una única variable explicativa, se puede expresar como sigue:

$$Y = \beta_0 + f(X) + \varepsilon.$$

Puede que la función $f()$ sea conocida de antemano, como ocurre en muchos modelos biológicos, donde existe una dependencia de tipo exponencial, $f(x) = e^{\beta_0 + \beta_1 x}$. En otras ocasiones, dicha función es desconocida y se puede utilizar una aproximación. Por ejemplo, mediante la regresión polinómica, muy utilizada en la práctica:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \dots + \beta_p X^p + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2). \quad (17.1)$$

Sin embargo, la regresión polinómica tiene un gran inconveniente: que no se lleva a cabo de forma local y, por tanto, cada vez que se cambia un coeficiente del modelo, el cambio impacta a los valores ajustados en todo el rango de la variable explicativa. No obstante, es posible utilizar técnicas (como las que se presentan en este capítulo) en las que el valor predicho en un punto dado sólo depende de las observaciones en ese punto y de las observaciones vecinas, es decir, el ajuste se lleva a cabo de forma local.

En el caso de disponer de más de una variable explicativa, la extensión del modelo de regresión múltiple sería el **modelo aditivo** (en el caso de variable respuesta gaussiana), donde no se

asume que la relación entre la variable respuesta y cada una de las variables explicativas tenga que ser lineal:

$$Y = f(X_1) + \dots + f(X_j) + \epsilon, \quad \epsilon \sim N(0, \sigma^2). \quad (17.2)$$

Las funciones $f()$ incluyen también a las funciones lineales vistas en el Capítulo 15. Los **modelos aditivos generalizados**, GAM, extienden el modelo anterior a respuestas no gaussianas, como lo hacen los GLM respecto de los modelos lineales con respuesta gaussiana (véase Cap. 16).

17.2. Splines con penalizaciones

Las funciones de la Eq. (17.2) se estiman mediante técnicas de suavizado o *smoothers*, cuyo objetivo es extraer las tendencias (o señales) existentes en la relación entre la variable respuesta y las variables explicativas, sin presuponer una forma funcional a priori para ellas; sólo se asume que la relación entre Y y X es suave (tiene poco ruido). Las predicciones obtenidas mediante estas técnicas tienen menos variabilidad que la variable respuesta; de ahí que a estas técnicas se les denomine “suavizadores” (la regresión lineal es un suavizador llevado al extremo). Las siguientes son algunas de las técnicas de suavizado más populares:

1. Regresión polinomial local con pesos, *lowess*.
2. Kernels.
3. Splines.

Este capítulo se centra en uso de los **splines**, ya que es la técnica de suavizado más utilizada. Los splines son funciones polinómicas a trozos de la variable explicativa, que se unen en puntos llamados **nodos**. Existen muchos tipos de splines (naturales, cílicos, B-splines, O-splines, etc.). Independientemente del tipo de spline, este capítulo se centra en los splines con penalizaciones (P-splines), que se basan en: (i) hacer una aproximación de la función $f()$ mediante una base de funciones, y (ii) añadir una penalización a la hora de estimar el modelo, de manera que se pueda controlar la variabilidad de la curva que se quiere estimar. Hay muchas maneras de representar una función a través de una base de funciones (un ejemplo sencillo de una base de funciones es el caso de la regresión polinómica, en la que la base de funciones, es decir, la matriz de regresión, es una matriz cuyas columnas son las potencias de la variable explicativa: $[X : X^2 : \dots : X^p]$). Una de las mejores opciones son los B-splines (De Boor, 2001), debido a sus buenas propiedades numéricas. La penalización se añade en la función de verosimilitud y se construye a partir de la derivada de la curva que se quiere penalizar. Generalmente se utilizan penalizaciones de orden 2, lo que implica que se penaliza todo aquello que no es lineal en la función; por tanto, si la penalización es muy grande la curva estimada es simplemente una línea recta. La penalización está controlada por un parámetro llamado **parámetro de suavizado** Véanse Eilers and Marx (2010) y Eilers et al. (2015) para más detalle).

A la hora de ajustar este tipo de modelos hay que tomar dos decisiones importantes:

- El número de nodos del B-spline: generalmente se utiliza esta regla:

$$\text{número de nodos} = \min\{40, \text{valores únicos de } X/4\} \quad (17.3)$$

(por ejemplo, si se tienen 100 observaciones diferentes, se elegirían $100/4=25$ nodos).

17.3. Aspectos metodológicos

281

- El valor del parámetro de suavizado: se puede estimar por distintos métodos: validación cruzada, validación cruzada generalizada, máxima verosimilitud, máxima verosimilitud restringida, etc. Se recomienda re-expresar el modelo como un modelo mixto (véase Cap. 18) y estimarlo mediante el **método de máxima verosimilitud restringida**, REML¹, para así poder estimar el parámetro de suavizado junto con los demás parámetros del modelo.

La Fig. 17.1 muestra el impacto que el parámetro de suavizado tiene en el ajuste final de la curva (los datos corresponden al dataset `fossil` del paquete `Semipar`).

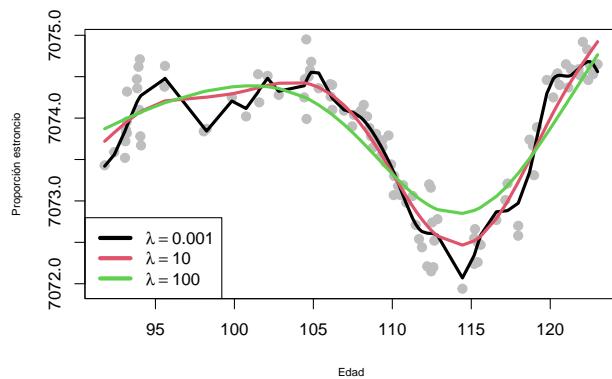


Figura 17.1: Regresión con P-splines para diferentes valores del parámetro de suavizado

17.3. Aspectos metodológicos

Al igual que en el caso de los modelos lineales y los modelos GLM, en los modelos GAM es necesario conocer algunos aspectos metodológicos que son fundamentales para llevar a cabo un ajuste correcto de los modelos y entender los resultados obtenidos en el ajuste. A continuación se muestran los más relevantes.

17.3.1. Estimación de los parámetros del modelo

La estimación de los modelos GAM se lleva a cabo mediante máxima verosimilitud penalizada. Supóngase el caso de una sola variable explicativa y que se quiere ajustar el modelo:

$$Y = f(X) + \epsilon.$$

¹Es igual que el de máxima verosimilitud pero teniendo en cuenta los grados de libertad utilizados para estimar los efectos fijos al estimar los componentes de varianza (el método de máxima verosimilitud no lo hace). Para más detalles consultese [Henderson \(1953\)](#).

Como se comentó anteriormente, los modelos GAM tienen como punto de partida la aproximación de la función a estimar mediante una matriz formada por B-splines; es decir, se busca transformar el modelo lineal o lineal generalizado tradicional de tal forma que $f(X)$ sea el producto de una matriz multiplicada por unos coeficientes (esa matriz está formada por los B-splines). En otros términos, se elige una base (una matriz \mathbf{B}) que permita escribir la función $f(X)$ como una combinación lineal de sus elementos (los elementos de esta base son conocidos ya que se calculan a partir de las variables explicativas):

$$f(X) = \sum_{l=1}^k b_l(X)\theta_l,$$

donde $b_l(X)$ son las funciones B-spline que componen la base. En forma matricial:

$$f(X) = \mathbf{B}\boldsymbol{\theta}.$$

Los parámetros $\boldsymbol{\theta}$ se estiman minimizando la siguiente expresión (en el caso de asumir gaussianidad para los errores, y por tanto para la variable respuesta, los mínimos cuadrados penalizados son equivalentes a la máxima verosimilitud penalizada):

$$(\mathbf{y} - \mathbf{B}\boldsymbol{\theta})'(\mathbf{y} - \mathbf{B}\boldsymbol{\theta}) + \lambda\boldsymbol{\theta}'\mathbf{P}\boldsymbol{\theta},$$

donde \mathbf{P} es la matriz de penalización y λ es el parámetro de suavizado. Dado un valor del parámetro de suavizado, las estimaciones de los parámetros vienen dadas por²:

$$\hat{\boldsymbol{\theta}} = (\mathbf{B}'\mathbf{T}\mathbf{B} + \lambda\mathbf{P})^{-1}\mathbf{B}'\mathbf{y}, \quad (17.4)$$

y las estimaciones de la variable respuesta se obtienen como: $\hat{\mathbf{y}} = \underbrace{\mathbf{B}(\mathbf{B}'\mathbf{B} + \lambda\mathbf{P})^{-1}\mathbf{B}'}_{\mathbf{H}}\mathbf{y}$. La matriz \mathbf{H} juega un papel importante, ya que la suma de su diagonal da una idea de la complejidad de la curva ajustada (la curva más compleja sería la que interpola los datos). Dicha suma se denomina **grados de libertad efectivos** (que no se corresponden con el número de parámetros ajustados).

17.3.2. Inferencia sobre las funciones suaves

Para saber si la relación estimada entre Y y X es o no estadísticamente significativa, se debe proceder al contraste:

$$\begin{aligned} H_0 : f(X) &= 0 && \text{(no efecto)} \\ H_1 : f(X) &\neq 0 && \text{(efecto).} \end{aligned}$$

Dado que la función $f(X)$ depende de los coeficientes que acompañan a las bases de B-splines, el contraste anterior es equivalente al contraste:

$$\begin{aligned} H_0 : \boldsymbol{\theta} &= 0 \\ H_1 : \boldsymbol{\theta} &\neq 0. \end{aligned}$$

²Como se avanzó anteriormente, si el modelo se expresa como un modelo mixto, la estimación REML proporciona la estimación del parámetro de suavizado junto con la de los restantes parámetros del modelo

La distribución del estadístico de contraste dependerá de si la variable respuesta sigue una distribución Normal o no: en caso afirmativo el estadístico de contraste sigue un distribución F . En caso negativo, sigue una distribución χ^2 .

Comparación de modelos

Cuando se trabaja con un modelo aditivo (17.2) en el que hay más de una variable explicativa, puede ser de interés comparar versiones de ese modelo que contengan distintos conjuntos de variables. La comparación dependerá de la relación entre los modelos a comparar:

1. **Modelos anidados.** La comparación se basa, igual que en los GLM, en la diferencia en la *deviance residual*. Si se quieren comparar dos modelos m_1 y m_2 (donde $m_1 \subset m_2$), entonces:

- En el caso de variable respuesta Normal, el estadístico de contraste es:

$$\frac{(DR(m_1) - DR(m_2))/(df_2 - df_1)}{DR(m_2)/(n - df_2)} \approx F_{(df_2 - df_1), (n - df_2)},$$

donde DR es la *deviance residual* (suma de cuadrados residual) y df son los grados de libertad asociados con cada modelo.

- En otro caso, se utiliza como estadístico de contraste el siguiente:

$$DR(m_1) - DR(m_2) \approx \chi^2_{df_2 - df_1}.$$

2. **Modelos no anidados.** En este caso los contrastes anteriores no son válidos y se utilizarán criterios basados en el AIC (criterio de información de Akaike).

17.3.3. Suavizado multidimensional y para datos no Gaussianos

Para el suavizado penalizado en 2 dimensiones (o más) también se necesita una base y una penalización. El modelo sería:

$$Y = \beta_0 + f(X_1, X_2) + \epsilon,$$

donde $f()$ es una función de las dos covariables X_1 y X_2 . Dicha función se aproxima mediante el producto tensorial de las bases de B-splines marginales para cada una de las covariables y la penalización dependerá de dos parámetros de suavizado. Los términos de suavizado multidimensional se pueden combinar con términos unidimensionales y términos lineales. En este caso, la penalización dependería de dos parámetros de suavizado (uno para cada covariable).

La extensión de los modelos de suavizado al caso en el que la variable respuesta no sea Gaussiana, se hace de forma similar al caso lineal, cuando se pasa de un modelo de regresión lineal a un GLM. Al igual que en el caso de los GLMs, $g(\boldsymbol{\mu}) = \boldsymbol{\eta} = f(\mathbf{X}) = \mathbf{B}\boldsymbol{\theta}$, y se añade la penalización a la función de verosimilitud de la distribución correspondiente:

$$\ell_p(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) + \lambda \boldsymbol{\theta}' P \boldsymbol{\theta},$$

donde $\ell(\boldsymbol{\theta})$ es la log-verosimilitud.

17.4. Procedimiento con R: la función `gam()` del paquete `mgcv`

Aunque hay muchas librerías disponibles, la principal es `mgcv`, que implementa una gran variedad de modelos de suavizado a través de la función `gam()` (generalized additive models)³.

```
gam(formula, method="", select="", family=gaussian())
```

- `formula` es el argumento principal de esta función; es la ecuación del modelo: por ejemplo, $y \sim x_1 + x_2 + s(x_3)$.
 - Lo primero que se tiene que elegir es la base a utilizar para representar las funciones suaves, `s(x)` (véase `?s` o `?smooth.terms`), o `te(x1, x2)` en el caso de suavizado bidimensional. Por defecto se usan los llamados *thin plate splines*. El tipo de base usada se puede modificar utilizando el argumento `bs` dentro de `s(x, bs = "ps")`; en este caso `ps` indica el uso de B-splines con penalizaciones. A continuación se describen otras alternativas:

<code>bs</code>	Descripción
<code>tp</code>	Thin plate regression splines
<code>ts</code>	Thin plate regression splines con regularización
<code>cr</code>	Spline cúbicos de regresión
<code>crs</code>	Spline cúbicos de regresión con regularización
<code>cc</code>	Spline cílicos
<code>ps</code>	P-splines

- `m` indica el orden de la penalización; por defecto es 2.
- `k` es el número de nodos para construir la base. El número por defecto suele ser demasiado bajo, por lo que siempre se recomienda que el usuario elija el número utilizando la regla dada en (17.3).
- `by` debe igualarse a una variable numérica o factor de la misma dimensión de cada covariante, para hacer interacciones entre curvas y variables.
- `id` se utiliza para forzar que diferentes términos suaves utilicen la misma base y la misma cantidad de suavizado.
- `method` selecciona método para estimar el parámetro de suavizado. Se puede elegir entre: `REML` (máxima verosimilitud restringida), `ML` (máxima verosimilitud), `GCV.Cp` (validación cruzada generalizada), `GACV.Cp` (validación cruzada aproximada generalizada). En la práctica, como se indicó anteriormente, se prefiere `REML`.
- `family` permite elegir la distribución de la variable respuesta (binomial, Poisson, etc.); por defecto asume Gaussiana.
- `select=TRUE` contrasta si una variable debe entrar o no en el modelo.

³La principal referencia para esta sección es el libro de Wood (2006).

Es importante reseñar que si el método elegido para estimar el parámetro de suavizado es REML, entonces internamente, el modelo se transforma en un modelo mixto y lo estima junto con el resto de los parámetros del modelo (véase 18).

17.5. Casos prácticos

En este apartado se ven una serie de aplicaciones que permiten mostrar los diferentes usos de este tipo de modelos.

17.5.1. Modelo unidimensional con `fossil`

Se empieza ilustrando el uso de la función `gam()` con el conjunto de datos `fossil` del paquete `SemiPar`. El objetivo es estimar la relación entre la edad de los fósiles y la proporción de isotopos de estroncio.

```
library("SemiPar")
data(fossil)
Y <- 10000*fossil$strontium.ratio
X <- fossil$age
plot(X,Y, xlab="Edad", ylab = "Proporción de estroncio")
```

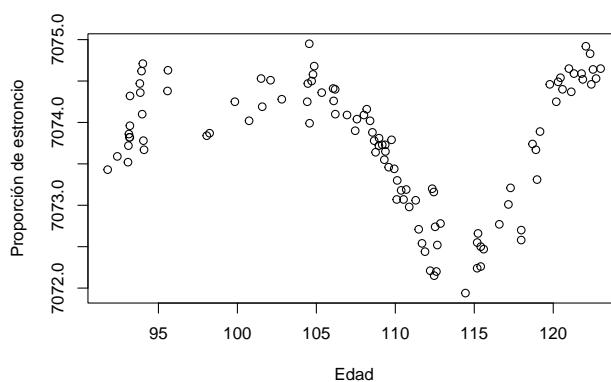


Figura 17.2: Edad de los fósiles con respecto a la proporción de isótopos de estroncio

A la vista de la Fig. 17.2, está claro que se necesita ajustar una curva (y no una línea) para estimar la relación entre ambas variables. Para ello se utiliza la función `gam()`, que devuelve un objeto de tipo "gam" y que se puede usar con las típicas funciones `print()`, `summary()`, `fitted()`, `plot()`, `residuals()`, etc.

```

library("mgcv")
fit_gam <- gam(Y ~ s(X,k=25,bs="ps"), method="REML", select=TRUE)
# se eligen 25 nodos ya que se la variable tiene 106 observaciones
summary(fit_gam)
#>
#> Family: gaussian
#> Link function: identity
#>
#> Formula:
#> Y ~ s(X, k = 25, bs = "ps")
#>
#> Parametric coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept) 7.074e+03 2.435e-02 290504    <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Approximate significance of smooth terms:
#>          edf Ref.df   F p-value
#> s(X) 10.22     24 35.89    <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> R-sq.(adj) =  0.891  Deviance explained = 90.2%
#> -REML = 23.946  Scale est. = 0.062849 n = 106

```

Como se puede ver, la relación entre la variable respuesta (Y , proporción de estroncio) y la variable explicativa (X , edad) se ha especificado mediante un *spline*, `s()`, de tipo penalizado, `ps`, con 25 nodos. Se ha seleccionado REML como método para estimar el parámetro de suavizado (los parámetros del spline se estiman también mediante REML, ya que da lugar a las mismas estimaciones que máxima verosimilitud).

En la primera parte de la salida anterior aparecen los términos que entran linealmente en el modelo (en este caso sólo el término independiente o intercepto); en la parte de abajo se muestran los términos de suavizado. Como se indicó anteriormente, dado que se ha usado `select=TRUE`, se está contrastando si la variable `edad` debe entrar en el modelo o no. En este caso, es claro que ha de entrar ya que el p-valor de `s(x)` es pequeño y los grados de libertad asociados son aproximadamente 10, lo que indica que la relación entre Y y X está lejos de la linealidad.

La función `gam.check()` devuelve los gráficos de residuos usuales (residuos frente a valores ajustados, gráficos de cuantiles para comprobar la normalidad, etc.), pero además proporciona información sobre el proceso de ajuste del modelo.

```
gam.check(fit_gam,cex=1.2)
```

```
#>
```

17.5. Casos prácticos

287

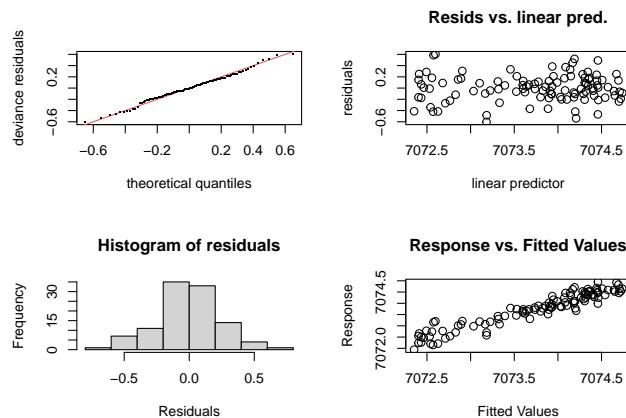


Figura 17.3: Gráficos de residuos obtenidos con ‘gam.check()’

```
#> Method: REML Optimizer: outer newton
#> full convergence after 5 iterations.
#> Gradient range [-4.557319e-06,5.900236e-06]
#> (score 23.94602 & scale 0.06284944).
#> Hessian positive definite, eigenvalue range [4.557347e-06,53.03185].
#> Model rank = 25 / 25
#>
#> Basis dimension (k) checking results. Low p-value (k-index<1) may
#> indicate that k is too low, especially if edf is close to k'.
#>
#>          k'  edf k-index p-value
#> s(X) 24.0 10.2    1.03   0.56
```

El test que aparece en la parte de abajo está contrastando si el número de nodos elegido es suficiente. Si el valor de k está muy próximo al de edf , entonces se debería reajustar el modelo con más nodos.

El comando `plot()` permite dibujar la función suave que relaciona Y con X. La curva estimada que aparece en la Fig. 17.4 está centrada (la función `plot()` siempre lo hace de esta forma), el argumento `shade` hace que se sombre el intervalo de confianza y `seWithMean` hace que la incertidumbre sobre el término independiente se incluya en el cálculo del intervalo de confianza.

```
plot(fit_gam, shade=TRUE, seWithMean=TRUE, pch=19, 1, cex=.55)
```

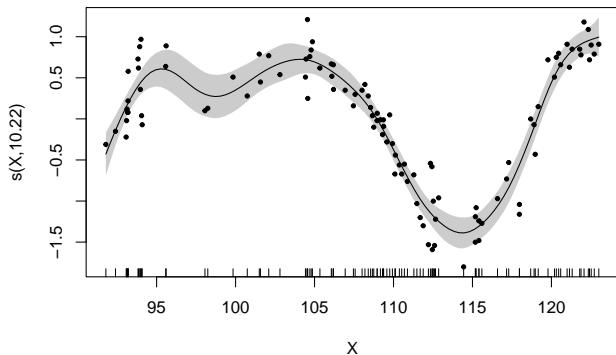


Figura 17.4: Curva ajustada e intervalo de confianza

17.5.2. Modelo aditivo con airquality

En esta sección se analizan de nuevo los datos `airquality` (ver `airquality`⁴), que consisten en 154 medidas de calidad del aire en Nueva York, de mayo a septiembre 1973. El objetivo es establecer la relación entre las variables meteorológicas y el nivel de concentración de ozono en la atmósfera. Ya se ha analizado dicha relación en el Cap. 15, donde los ajustes lineales realizados eran satisfactorios pero se encontraban problemas en los residuos del modelo, lo cual impedía validar la modelización realizada. Allí se sugería que la relación entre la variable respuesta y alguna explicativa fuese no lineal. Además, se consideró la transformación logarítmica de la variable `Ozone`, y con dicha trasformación se obtenía una distribución más similar a la distribución Normal.

En consecuencia, se va a ajustar el modelo incluyendo las variables explicativas sin imponerles linealidad; en particular, se van a incluir las variables `Wind`, `Temp` y `Solar.R`. Las variables `Wind` y `Temp` tienen sólo 31 y 40 valores únicos, respectivamente, aunque el conjunto de datos tiene 154 valores; por eso, para estas dos variables, se decide establecer el número de nodos en 10 y no más; para la variable `Solar.R` el número de nodos se fija en 20.

```
airq_gam=gam(log(Ozone)~s(Wind,bs="ps",k=10) +
  s(Temp,bs="ps",k=10)+s(Solar.R,bs="ps",k=20),
  method="REML",select=TRUE,data=airquality,na.action=na.omit)
summary(airq_gam)
#>
#> Family: gaussian
#> Link function: identity
#>
#> Formula:
```

⁴Conjunto de datos incluido con la instalación base de R.

```
#> log(Ozone) ~ s(Wind, bs = "ps", k = 10) + s(Temp, bs = "ps",
#>   k = 10) + s(Solar.R, bs = "ps", k = 20)
#>
#> Parametric coefficients:
#>   Estimate Std. Error t value Pr(>|t|)
#> (Intercept) 3.41593   0.04586 74.49 <2e-16 ***
#> ---
#> Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Approximate significance of smooth terms:
#>   edf Ref.df F p-value
#> s(Wind)    2.318     9 2.255 3.13e-05 ***
#> s(Temp)     1.852     9 6.128 < 2e-16 ***
#> s(Solar.R) 2.145    19 1.397 2.31e-06 ***
#> ---
#> Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> R-sq.(adj) = 0.689 Deviance explained = 70.7%
#> -REML = 86.106 Scale est. = 0.23342 n = 111
```

Los resultados indican que todas las variables son significativas (p-valores pequeños), estando la variable **Temp** próxima a la linealidad (los grados de libertad efectivos asociados a la variable son 1.8). El R^2 ajustado es 0.69, por lo que el modelo ajusta moderadamente bien los datos.

La Fig. ?? muestra las tres curvas ajustadas junto con sus correspondientes intervalos de confianza. También incluye los denominados *residuos parciales* que corresponden a, por ejemplo, en el caso del gráfico del viento, $\log(Ozone) - \hat{\beta}_0 - \hat{f}(Temp) - \hat{f}(Solar.R)$, es decir, lo que queda sin explicar después de haber ajustado los demás términos del modelo.

```
library("mgcviz")
# getViz es otra opción para dibujar los términos de un modelo gam()
b <- getViz(airq_gam)
pl <- plot(b) + l_points() + l_fitLine(linetype = 2) + l_ciLine(colour = 2)
print(pl, pages=1)
```

17.5.3. Modelo semiparamétrico con onions

Es un caso particular del modelo aditivo, pues en este modelo todas las variables entran de forma lineal excepto una:

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_{p-1} X_{p-1} + f(X_p) + \epsilon.$$

La forma de ajustar el modelo es exactamente igual a la anterior. Pero hay un caso que merece especial interés: cuando en la parte paramétrica se incluye una variable categórica con dos o más niveles. Al igual que en el caso de regresión lineal, se puede plantear si se quieren ajustar dos o más rectas paralelas (modelo aditivo) o no paralelas (modelo con interacción).

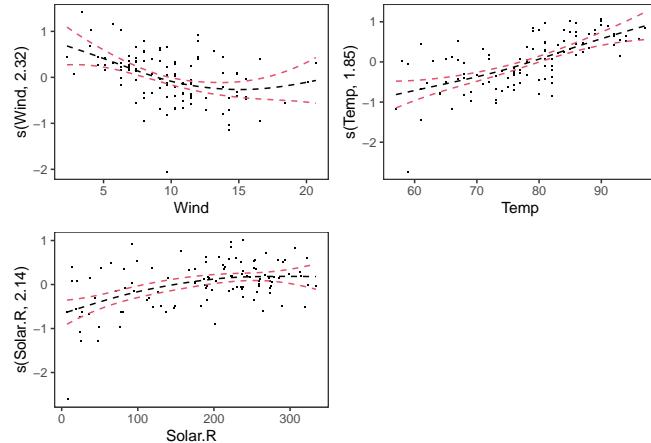


Figura 17.5: Curvas estimadas para 'Wind', 'Temp' y 'Solar'

Para ilustrar este caso se acude al `data.frame onions` (librería `SemiPar`). Contiene 84 observaciones de un experimento sobre la producción de un tipo de cebolla en dos localidades: (Purnong Landing (la localidad de referencia) y Virginia. El objetivo es relacionar el logaritmo de la producción de cebollas con la densidad de plantas por metro cuadrado, `dens`. El modelo lineal básico sería:

$$\log(\text{yield}_j) = \beta_0 + \beta_1 \text{location}_j + \beta_2 \text{dens}_j + \epsilon_j$$

donde

$$\text{location}_j = \begin{cases} 0 & \text{si la observación } j \text{ es de Purnong Landing} \\ 1 & \text{si la observación } j \text{ es de Virginia} \end{cases}$$

Se comienza por ajustar el siguiente modelo:

$$\log(\text{yield}_j) = \beta_0 + \beta_1 \text{location}_j + f(\text{dens}_j) + \epsilon_j$$

```
library("mgcv")
library("SemiPar")
data(onions)
#Se indica a R que la variable locationVirginia es categórica
onions$location <- factor(onions$location)
#Se recodifica la variable
levels(onions$location) <- c("Purnong Landing","Virginia")
fit1 <- gam(log(yield) ~ location + s(dens,k=20,bs="ps"),
            method="REML", select=TRUE, data=onions)
summary(fit1)
#>
#> Family: gaussian
#> Link function: identity
#>
#> Formula:
```

17.5. Casos prácticos

291

```
#> log(yield) ~ location + s(dens, k = 20, bs = "ps")
#>
#> Parametric coefficients:
#>                               Estimate Std. Error t value Pr(>|t|)
#> (Intercept)           4.85011   0.01688 287.39 <2e-16 ***
#> locationVirginia -0.33284   0.02409 -13.82 <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Approximate significance of smooth terms:
#>             edf Ref.df   F p-value
#> s(dens) 4.568     19 72.76 <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> R-sq.(adj) =  0.946  Deviance explained = 94.9%
#> -REML = -54.242  Scale est. = 0.011737 n = 84
```

En este ejemplo se ve que en la parte lineal aparecen dos parámetros, ambos significativos: la ordenada en el origen o intercepto y el coeficiente de la categoría **Virginia** de la variable **location**, que es negativo, indicando que la producción media en Purnong Landing es mayor que en Virginia. El término de suavizado también es significativo.

En este caso, función `plot.gam()` sólo dibuja una curva, pues las curvas para las dos localizaciones son paralelas y la diferencia entre ellas es igual al valor del parámetro correspondiente a **localización**. Para dibujar las curvas para cada localización se utiliza la función `plot_smooth()` de la librería `tidymv`. Los argumentos son, primero el modelo, después la variable explicativa y por último la variable categórica.

```
library("tidymv")
library("ggplot2")
plot_smooths(fit1, dens, location) +
  theme(text = element_text(size = 12))
```

Asumir curvas paralelas para ambas localidades implica que el descenso en la producción de cebollas a medida que aumenta la densidad de plantas es el mismo para ambas localidades, y esto no tiene por qué ser cierto. Para relajar esta hipótesis se puede ajustar un modelo con interacción (de manera similar a lo que se hace en el caso de regresión lineal):

$$\log(\text{yield}_j) = \beta_0 + \beta_1 \text{location}_j + f(\text{dens}_j)L(j) + \epsilon_j$$

donde

$$L(j) = \begin{cases} 0 & \text{si la } j\text{-ésima observación es de Purnong Landing} \\ 1 & \text{si la } j\text{-ésima observación es de Virginia} \end{cases}$$

Para hacerlo en **R**, se introduce el argumento `by=location` dentro de la curva

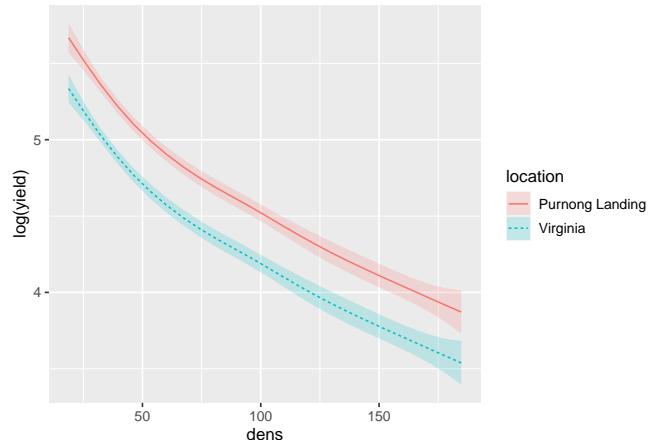


Figura 17.6: curvas ajustadas para ambas localidades

```

fit2 <- gam(log(yield) ~ location + s(dens,k=20,bs="ps",by=location),
             method="REML", data=onions)
summary(fit2)
#>
#> Family: gaussian
#> Link function: identity
#>
#> Formula:
#> log(yield) ~ location + s(dens, k = 20, bs = "ps", by = location)
#>
#> Parametric coefficients:
#>              Estimate Std. Error t value Pr(>|t|)
#> (Intercept) 4.84415   0.01603 302.19  <2e-16 ***
#> locationVirginia -0.33018   0.02270 -14.54  <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Approximate significance of smooth terms:
#>          edf Ref.df     F p-value
#> s(dens):locationPurnong Landing 3.096 3.834 176.9 <2e-16 ***
#> s(dens):locationVirginia       4.742 5.795 153.0 <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> R-sq.(adj) =  0.952  Deviance explained = 95.7%
#> -REML = -58.541  Scale est. = 0.010446 n = 84

```

Ahora aparecen dos términos suaves, uno para cada localidad, de modo que estas curvas no tienen por qué ser paralelas, sino que cada una se ajustará a la forma que tengan los datos. En

17.5. Casos prácticos

293

en este caso, la Fig. 17.7, generada de nuevo con `plot_smooths`, muestra como las curvas se van alejando a medida que aumenta la densidad de plantas.

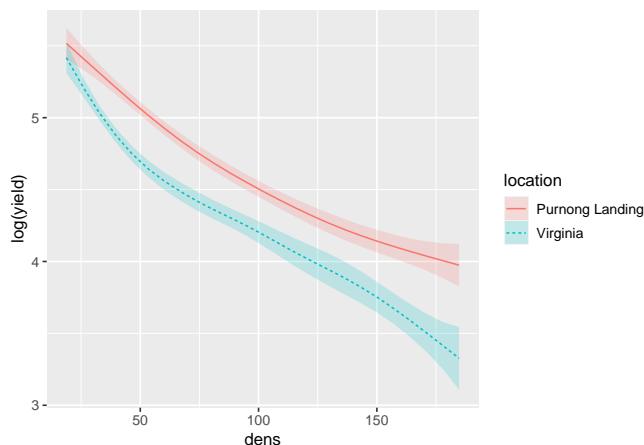


Figura 17.7: Curvas ajustadas para ambas localidades permitiendo que no sean paralelas

Para finalizar se comparan ambos modelos con el criterio AIC.

```
AIC(fit1); AIC(fit2)
#> [1] -125.2307
#> [1] -131.2181
```

Dado que el menor valor se alcanza en el segundo modelo, se escogería el modelo que incluye la interacción entre la variable densidad y la localidad.

17.5.4. Modelo aditivo generalizado y multidimensional con `smacker`

En este epígrafe se analizan los datos `smacker` del paquete `sm`. El objetivo es ver cómo influyen las condiciones del mar (temperatura de agua, etc.) en la ausencia o presencia de huevos de jurel en el mar Cantábrico. Además, se incorporará al modelo la posición geográfica mediante las covariables latitud y longitud; de esta forma se podrá captar el efecto espacial.

```
library("sm")
data(smacker)
library("dplyr")
smacker <- smacker |>
  mutate(Presence = ifelse(Density>0, 1, 0),
         smack.long = -smack.long,
         ldepth = log(smack.depth))
library("maps")
par(pty="s")
```

```

Position <- cbind(smacker$smack.long, smacker$smack.lat)
plot(Position,col=NULL,xlim=c(-10,-1),ylim=c(43,48),cex=1.2,xlab="longitud",
~ ylab="latitud")
map("world",add=TRUE,fill=TRUE,col="grey")
points(Position[smacker$Presence==1],pch=1,cex=.5,col=4)
points(Position[smacker$Presence==0],pch=16,cex=.5,col=2)
legend("topleft",c("Presencia ", "Ausencia"), col=c(4,2),pch=c(1,16),cex=.85)

```

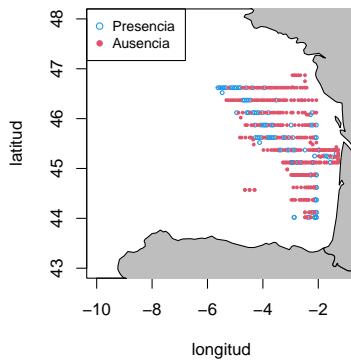


Figura 17.8: Área donde se constató la ausencia/presencia de huevos de jurel

Dado que la variable respuesta es dicotómica, se utiliza un modelo de regresión logística en el que se flexibiliza la relación lineal de las variables explicativas con la respuesta y , además, se usa una superficie para estimar el efecto de la localización como una función en dos dimensiones (latitud y longitud). En este caso, en vez de usar `te()` se usa `s()` también para el caso de 2 dimensiones. La diferencia fundamental con `te()` es que `s()` asume un suavizado isotrópico, es decir, el mismo parámetro de suavizado para la latitud y longitud. No se debe usar `s()` para el suavizado en dos dimensiones si las covariables están medidas en unidades diferentes. En este caso, como tanto la longitud como la latitud están medidas en las mismas unidades, se puede usar el suavizado isotrópico.

```

logit1 <- gam(Presence~s(ldepth)+ s(Temperature)+ s(smack.long, smack.lat,k=60),
family=binomial, select=TRUE, data=smacker)
b <- getViz(logit1)
print(plot(b, allTerms = T), pages = 1)

```

En la Fig. 17.9 se aprecia que la relación entre la probabilidad de presencia de huevos y la temperatura no es lineal, mientras que sí lo es en el caso de la profundidad. El R^2 es tan sólo 0,4, por lo que convendría utilizar más variables explicativas para obtener buenas predicciones.

Las probabilidades predichas se pueden obtener con la función `predict`.

17.5. Casos prácticos

295

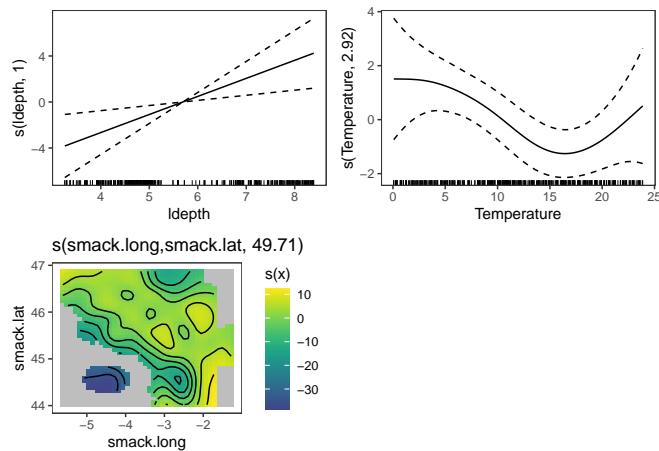


Figura 17.9: Efectos suaves estimados por el modelo para las variables. Efecto de la profundidad y temperatura en la fila superior y efecto espacial en la inferior

```
prob=predict(logit1,type="response")
```

Resumen

En este capítulo se introducen los modelos aditivos generalizados. En particular:

- Se presentan distintos aspectos metodológicos de carácter inferencial en este tipo de modelos.
- Se muestra el uso de R para llevar a cabo su ajuste.
- Se presentan diversos casos prácticos que ilustran la versatilidad de estos modelos para analizar datos complejos.

Capítulo 18

Modelos mixtos

^aUniversidad de Castilla-La Mancha

^bUniversidad Carlos III de Madrid

18.1. Conceptos básicos

Los **modelos mixtos** (MM) para variables de respuesta continuas son modelos estadísticos en los que los residuos siguen una distribución Normal pero puede que no sean independientes o no tengan varianza constante. Son necesarios en muchas situaciones, sobre todo en experimentos donde se realiza algún tipo de muestreo:

1. Estudios con datos agrupados, como por ejemplo, alumnos en una clase, individuos en una ciudad.
2. Estudios longitudinales o de medidas repetidas, donde un elemento o individuo es medido repetidamente a lo largo del tiempo o bajo condiciones distintas.

Este tipo de estudios se pueden encontrar en diferentes áreas como la medicina, biología, ciencias experimentales y sociales.

18.1.1. Tipo y estructura de los datos

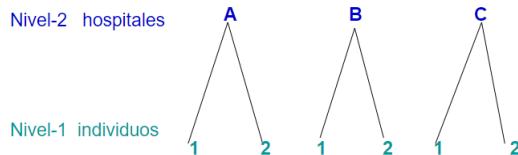
La estructura de los datos con la que se trabaja es el factor determinante para saber si se han de utilizar modelos mixtos, y en su caso, qué tipo de modelo.

18.1.1.1. Datos jerárquicos (o agrupados)

En este tipo de datos, la variable dependiente (de respuesta, de interés) se mide una sola vez en cada unidad de análisis (individuos, objetos, elementos ...), y los individuos¹ están agrupados (o anidados) en unidades mayores. Muchos tipos de datos tienen una estructura jerárquica: alumnos en escuelas, personas en municipios, pacientes en hospitales, plantas en una parcela...

Las jerarquías son una forma de representar la relación de dependencia que hay entre los individuos y los grupos a los que pertenecen. Por ejemplo, supóngase que se quiere hacer un estudio sobre el tiempo de recuperación en pacientes hospitalizados por COVID-19 en diferentes hospitales. Se tiene la siguiente estructura con dos niveles:

- Muchos individuos en el nivel 1 (pacientes).
- Agrupados en unas pocas unidades en el nivel 2 (hospitales).



Las estructuras multinivel pueden aparecer también como consecuencia del diseño del estudio que se está llevando a cabo. Por ejemplo, una encuesta sobre el estado de salud puede dar lugar a un diseño a tres niveles: primero se muestran regiones, luego distritos y después individuos.

En cada nivel de la jerarquía se pueden medir variables. Algunas estarán medidas en su nivel *natural*; por ejemplo, en el nivel “hospital” se podría medir el tamaño y en el nivel “pacientes” situación socio-económica. Además, se pueden mover las variables de un nivel a otro mediante agregación o desagregación:

- **Agregación:** la variable correspondiente al nivel más bajo se mueve a un nivel más alto; por ejemplo, se puede asociar a cada hospital la media del nivel socioeconómico de sus pacientes.
- **Desagregación:** mover las variables a un nivel más bajo; por ejemplo, asignarle a cada paciente una variable que indique el tamaño de su hospital de referencia.

18.1.1.2. Medidas repetidas y datos longitudinales

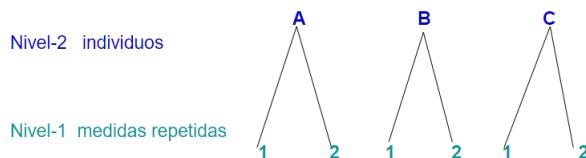
En este tipo de datos, la variable dependiente se mide más de una vez en un mismo individuo ([Singer and Willet, 2003](#)). Por ejemplo, se miden los niveles de glucosa de un enfermo antes y después de haberle inyectado insulina. Este tipo de datos también puede ser considerado como

¹En adelante nos referiremos a las unidades de análisis como individuos.

18.1. Conceptos básicos

299

datos multinivel (o jerárquicos) donde el nivel 2 representa a los individuos y el nivel 1 a las diferentes medidas tomadas. Dado que las medidas se toman a un mismo individuo, es probable que no sean independientes, por lo que utilizar un modelo lineal ordinario no sería apropiado.



Por **datos longitudinales** se entienden datos en los que la variable dependiente se ha medido en distintos instantes de tiempo en cada una de las unidades de análisis. En algunos casos, cuando la variable dependiente se mide a lo largo del tiempo, puede ser difícil identificar si los datos son medidas repetidas o datos longitudinales. Desde el punto de vista del análisis de los datos mediante MM esta distinción no es un elemento crítico. Lo importante es que en ambos tipos de datos la variable dependiente se ha medido repetidas veces en la misma unidad de análisis, y que, por tanto, las observaciones no son independientes.

18.1.2. ¿Efectos fijos o aleatorios?

En un modelo mixto la clave se encuentra en la distinción entre efectos fijos y aleatorios ([Snijers, 2003](#)). Esto es importante porque la inferencia y el análisis de ambos efectos es distinta.

Los **efectos fijos** son variables en las cuales el investigador ha incluido sólo los niveles (o tratamientos) que son de su interés. Por ejemplo, en un experimento se puede estar interesado en comparar dos grupos, uno al que se le aplica un tratamiento y otro de control. En este caso, el estudio compara los grupos y no interesa generalizar los resultados a otros tratamientos que podrían haber sido incluidos. Otro ejemplo sería el caso en el que se hace una encuesta y se eligen 10 ciudades. Si sólo interesan los resultados para esas 10 ciudades y no se quieren generalizar al resto de ciudades que podrían haber sido seleccionadas, la variable ‘ciudad’ es un efecto fijo. Si se eligen las ciudades de forma aleatoria de una población grande de ciudades, la variable ‘ciudad’ es un **efecto aleatorio**.

Una cantidad se considera aleatoria cuando cambia sobre las unidades de una población. Cuando un efecto en un modelo estadístico es considerado aleatorio, se está asumiendo que se quieren extraer conclusiones sobre la población de la cual se han elegido las unidades observadas, y no se tiene interés en esas unidades en particular. En este contexto se habla de **intercambiabilidad**, en el sentido de que se podría cambiar una unidad de la muestra por otra de la población y sería indiferente. Este es el caso de los factores de agrupamiento o diseño, como las parcelas en un experimento agrícola, o los días cuando un experimento se lleva a cabo en días distintos, o el técnico de laboratorio cuando hay varios haciendo el experimento; también lo serían los sujetos en un diseño de medidas repetidas o las localizaciones donde se recogen muestras en un río, si el objetivo es generalizar a todo el río.

Los métodos estándar utilizados para construir tests e intervalos de confianza para los efectos fijos no son válidos para los efectos aleatorios, pues en este último caso los efectos observados son sólo una muestra de todos los posibles efectos.

La clave para distinguir, estadísticamente hablando, entre efectos fijos y aleatorios, es si los niveles de la variable se pueden interpretar como extraídos de una población con una cierta distribución de probabilidad. En el caso de un efecto fijo, normalmente interesa comparar los resultados de la variable dependiente para los distintos niveles de la variable explicativa, es decir, interesa la diferencia entre las medias. En el caso de efectos aleatorios, no interesa específicamente comparar si las medias son distintas, sino cómo el efecto aleatorio explica la variabilidad en la variable dependiente. Por lo tanto, para que un efecto pueda considerarse aleatorio, es necesario que la variable dependiente presente cierta variabilidad no explicada asociada con las unidades del efecto aleatorio.

La Fig. 18.1 puede ayudar a determinar si un efecto es fijo o aleatorio:

1. ¿Cuál es el número de niveles?

Pequeño	Fijo
Grande o infinito	Possiblemente aleatorio

2. ¿Son los niveles repetibles?

Sí	Fijo
No	Aleatorio

3. ¿Hay, conceptualmente, un número infinito de niveles?

No	Possiblemente fijo
Sí	Possiblemente aleatorio

4. ¿Se necesitan realizar inferencias para niveles no incluidos en el muestreo?

No	Possiblemente fijo
Sí	Possiblemente aleatorio

Figura 18.1: Cuestiones para determinar si un efecto es fijo o aleatorio

Por ejemplo, en un estudio sobre satisfacción en el trabajo (variable dependiente) de los empleados (unidades observadas) de un cierto número de empresas (efecto aleatorio), si el nivel de satisfacción de los empleados de unas empresas es mayor que el de otras y el investigador no lo tiene en cuenta, habrá una cierta variabilidad residual asociada con el efecto ‘empresa’. Si esta variabilidad fuera próxima a cero, no sería necesario incluir el efecto aleatorio asociado con la empresa.

¿Por qué hay que utilizar modelos mixtos?

Cuando las observaciones están agrupadas en niveles o siguen una cierta jerarquía, las unidades se ven afectadas por el grupo al que pertenecen. Las jerarquías (o niveles) permiten representar la relación de dependencia entre los individuos y los grupos a los que pertenecen. Los alumnos que están en una misma escuela se parecen más entre sí que si se hubieran seleccionado aleatoriamente de entre toda la población de alumnos. Los modelos mixtos permiten tener en cuenta que las observaciones no son independientes.

18.2. Formulación del modelo con efectos aleatorios o modelos mixtos

El nombre de *modelos mixtos lineales* viene del hecho de que estos modelos son lineales en los parámetros y en las covariables y pueden implicar efectos fijos o aleatorios. Son, por lo tanto, una extensión de los modelos lineales de regresión.

18.2.1. Formulación general

La formulación general de un modelo mixto tiene la siguiente forma:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}, \quad \mathbf{u} \sim N(0, \mathbf{G}), \quad \boldsymbol{\epsilon} \sim N(0, \mathbf{R}), \quad (18.1)$$

donde:

- \mathbf{X} es una matriz $n \times k$ (k es el número de efectos fijos).
- \mathbf{Z} es una matriz $n \times p$ (p es el número de efectos aleatorios).
- $\boldsymbol{\beta}$ es el vector de efectos fijos y \mathbf{u} el de efectos aleatorios.
- \mathbf{G} es la matriz de varianzas-covarianzas de los efectos aleatorios, con dimensión $p \times p$.
- \mathbf{R} es la matriz de varianzas-covarianzas del error.

18.2.1.1. Estimación de $\boldsymbol{\beta}$ y \mathbf{u}

Se hace mediante las llamadas **ecuaciones de Henderson** (Henderson, 1953). Permiten obtener el mejor estimador lineal insesgado de $\boldsymbol{\beta}$ y el mejor predictor lineal insesgado de \mathbf{u} . Se obtienen maximizando la densidad conjunta de \mathbf{y} y \mathbf{u} :

$$f(\mathbf{y}, \mathbf{u}) = f(\mathbf{y}|\mathbf{u})f(\mathbf{u}), \quad \mathbf{y}|\mathbf{u} \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \mathbf{R}) \quad \mathbf{u} \sim N(0, \mathbf{G}). \quad (18.2)$$

Derivando con respecto a $\boldsymbol{\beta}$ y \mathbf{u} e igualando a cero se obtienen las ecuaciones de Henderson, cuya solución es:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \quad (18.3)$$

$$\hat{\mathbf{u}} = \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}), \quad (18.4)$$

donde $\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$. Sin embargo, \mathbf{V} depende de los parámetros de la varianza en el modelo, que forman parte de \mathbf{G} y \mathbf{R} y que es necesario estimar, como se muestra a continuación.

18.2.1.2. Estimación de los componentes de la varianza

Los métodos más comunes para la estimación de los parámetros de las matrices de covarianzas son: máxima verosimilitud (ML) y máxima verosimilitud restringida (REML). No existe una solución cerrada para los estimadores, y se estiman de forma numérica o mediante algoritmos iterativos. REML tiene en cuenta los grados de libertad utilizados para estimar los efectos fijos en el modelo. Si n es pequeño, REML dará mejores estimaciones que ML; si n es grande, no habrá prácticamente ninguna diferencia. El método preferido es REML.

18.2.2. Inferencia y selección del modelo

18.2.2.1. Contrastes de hipótesis para los efectos fijos, β

Utilizando la distribución aproximada:

$$\hat{\beta} \sim N \left(\beta, \underbrace{(\mathbf{X}' \hat{\mathbf{V}}^{-1} \mathbf{X})^{-1}}_{Var(\hat{\beta})} \right)$$

- Si se contrastan parámetros individuales, se utiliza el t-test para un solo efecto.
- Si se contrasta un conjunto de parámetros, se utiliza el F-test para más de un efecto.
- También se pueden comparar modelos usando el test de la razón de verosimilitud, LRT por sus siglas en inglés:

$$LRT = -2 [\ln(l_{H_0}) - \ln(l_{H_1})] \approx \chi^2_{df}. \quad (18.5)$$

Nota

En este caso hay que utilizar ML para estimar los parámetros de la varianza.

18.2.2.2. Contrastes de hipótesis para los parámetros de varianza

Al usar el test LRT (Eq. (18.5)) se ha de tener en cuenta que la distribución asintótica del estadístico del test depende de si el valor del parámetro bajo la hipótesis nula (H_0) está o no en la frontera del espacio paramétrico²

- **Caso 1:** El valor de los parámetros de varianza bajo H_0 no está en la frontera del espacio paramétrico (por ejemplo, al contrastar si los parámetros de varianza de dos efectos aleatorios son iguales o no). En ese caso se utiliza el test normalmente.
- **Caso 2:** El valor de los parámetros de varianza bajo H_0 está en la frontera del espacio paramétrico (por ejemplo, si se quiere contrastar si la varianza del efecto aleatorio es cero o no). La distribución asintótica del estadístico del test es una mixtura entre χ^2_p y χ^2_{p-1} , concretamente $0,5\chi^2_p + 0,5\chi^2_{p-1}$, donde p es el número de parámetros de la varianza que se hacen cero bajo la H_0 .

²Es espacio paramétrico es el conjunto de posibles valores del parámetro. Los valores que están en la frontera son los valores que están en el límite inferior (el mínimo) o el superior (máximo) del conjunto de valores posible. Dado que la varianza es positiva, si se contrasta si el valor es cero, estaría tomando un valor en la frontera.

18.2.3. Diagnosis del modelo

En el caso de modelos mixtos, se ha de contrastar la hipótesis de normalidad tanto para los residuos al nivel más bajo como para los efectos aleatorios.

En este tipo de modelos se utilizan los residuos condicionales, que son la diferencia entre los valores observados y el valor predicho condicional:

$$\hat{\epsilon} = y - X\hat{\beta} - Z\hat{u}.$$

Estos residuos tienden a estar correlados y sus varianzas pueden cambiar de un grupo a otro, aunque en el verdadero modelo los residuos están incorrelados y tienen varianza constante. Para solucionar este problema se pueden escalar los residuos por sus desviaciones estándar (o las estimaciones de éstas), dando lugar a los **residuos estandarizados** (si las desviaciones estándar son conocidas), o a los **residuos studentizados** (si son desconocidas y se utilizan estimaciones de las mismas). Con estos residuos se hace un análisis similar al caso de los modelos de regresión lineal.

18.3. Procedimiento con **R** para ajustar modelos mixtos

Hay varios paquetes de **R** para el ajuste de modelos mixtos. Los más usados son **nlme** y **lme4**. El segundo es una versión del primero que incluye modelos más generales y mejora los gráficos. A continuación se describe la función principal del paquete **lme4**.

18.3.1. La función **lmer()**

Esta función permite el uso de efectos aleatorios anidados y de errores correlados y/o heterocedásticos dentro de los grupos. En general, para definir un modelo mixto se necesita especificar la estructura de la media y de la parte aleatoria del modelo, incluidos los factores de agrupamiento, así como la estructura de correlación (si la hay).

También se puede especificar el método de estimación: “REML” o “ML”.

La parte aleatoria del modelo se incluye entre paréntesis en la ecuación y “|” separa las variables de agrupamiento de las predictoras. Si no hay variables predictoras para la parte aleatoria se pone un 1.

La función **VarCorr()** aplicada a un objeto **lmer** proporciona información sobre la estructura de componentes de varianza.

18.4. Caso práctico

En esta sección se comienza viendo cómo construir diferentes modelos con efectos aleatorios según a qué nivel estén medidas las variables explicativas y se termina dando una guía de construcción de estos modelos en la práctica. Los datos con los que se va a trabajar se encuentran

en el dataframe `Hsb82` del paquete `mlmRev` y provienen de un estudio titulado *High School and Beyond*. Los datos corresponden a 7.185 estudiantes repartidos en 160 escuelas, el número de alumnos por escuela varía entre 14 y 67. La variable de interés, `mAch`, es el nivel estandarizado alcanzado en matemáticas. Una cuestión inicial que se puede plantear es si el nivel socioeconómico (`cse`) del alumno predice las diferencias en el nivel de matemáticas. Para ello se ajusta el modelo:

$$y_j = \beta_0 + \beta_1 x_j + \epsilon_j,$$

que ignora que los alumnos provienen de distintos centros (por eso solo aparece el subíndice j , que es el que representa a las unidades de nivel más bajo, en este caso a los alumnos).

```
library("mlmRev")
Hsb82$school = factor(Hsb82$school, ordered=F)
multi0 <- lm(mAch ~ cses, data = Hsb82)
summary(multi0)
#>
#> Call:
#> lm(formula = mAch ~ cses, data = Hsb82)
#>
#> Residuals:
#>      Min       1Q   Median       3Q      Max
#> -17.8660  -5.1165   0.2966   5.3880  14.8705
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept) 12.74785   0.07933 160.69 <2e-16 ***
#> cses         2.19117   0.12010   18.24 <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 6.725 on 7183 degrees of freedom
#> Multiple R-squared:  0.04429,    Adjusted R-squared:  0.04415
#> F-statistic: 332.8 on 1 and 7183 DF,  p-value: < 2.2e-16
```

La ordenada en el origen es 12.75 y la pendiente 2.19, lo que indica que por cada unidad que aumenta el nivel socio-económico, la puntuación del test aumenta en 2.19 unidades; además se puede ver que el coeficiente es significativo.

Supóngase que ocurre la situación mostrada en la Fig. 18.2:

Los alumnos de la escuela A (rombos negros) sacan, en promedio, mejores notas que las que le asigna el modelo ajustado; con la escuela B (círculos azules) ocurre lo contrario. El gráfico indica que la ordenada en el origen (el intercepto) no debería ser la misma para todos los centros, sino que debería ser distinta para distintos centros. Es decir, el valor predicho debe ajustarse hacia arriba o hacia abajo. Eso se puede conseguir permitiendo que cada escuela tenga su propia ordenada en el origen:

$$y_{ij} = \beta_{0i} + \beta_1 x_{ij} + \epsilon_{ij}$$

Este modelo es similar al anterior añadiendo el subíndice i para identificar el centro al que

18.4. Caso práctico

305

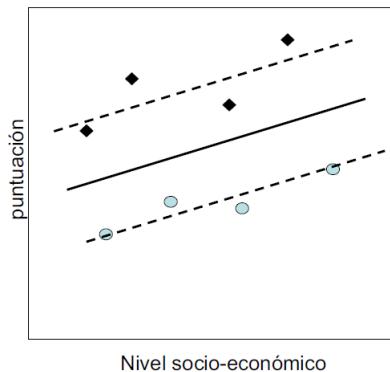


Figura 18.2: Ilustración de posibles escenarios para dos escuelas

p pertenece cada alumno. En realidad, se utiliza una variable categórica con tantas categorías como escuelas.

```
multi1 <- lm(mAch ~ cses + school, data = Hsb82)
```

Se están considerando las escuelas como un efecto fijo y no aleatorio, es decir, implícitamente se está suponiendo que solo interesan estas escuelas en particular.

La situación se pueden complicar más: es posible que el efecto del nivel socio-económico sea distinto para cada centro, es decir, que un aumento de una unidad en ese nivel pueda dar lugar a un aumento distinto en la nota del test en cada centro. En la Fig. 18.3 se ve como la pendiente de la recta para la escuela C es distinta a la dos anteriores.

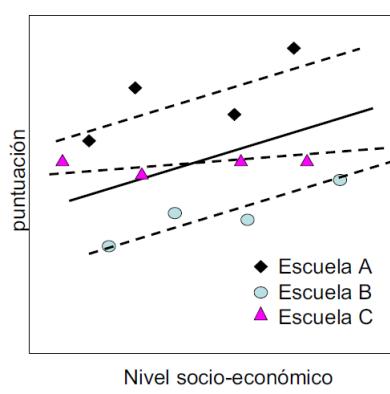


Figura 18.3: Ilustración de posibles escenarios para tres escuelas

El modelo que permite tener en cuenta esta situación es:

$$y_{ij} = \beta_{0i} + \beta_{1i}x_{ij} + \epsilon_{ij}$$

donde aparece ahora el sub-índice i también en la pendiente, lo que indica que cada centro tiene una pendiente diferente.

El código

```
multi2 <- lm(mAch ~ cses*school, data = Hsb82)
```

generaría 159 coeficientes más (uno por cada escuela), que son los que se incluirían con la interacción. Pero no interesan estas escuelas en concreto, sino la población de la que estas escuelas son una muestra.

Con un modelo con efectos aleatorios, sin embargo, se pueden contestar preguntas como: ¿Cuáles son las causas de esta variabilidad? ¿Qué variables pueden explicarla?

18.4.1. Modelo con ordenada en el origen aleatoria

Es el modelo mixto más sencillo. Se considera que los datos tienen una estructura con dos niveles: los alumnos están en el nivel 1 y están agrupados en escuelas, nivel 2. Se empieza suponiendo que no se dispone de ninguna variable explicativa, y que por lo tanto el único interés es la diferencia entre las notas medias del test de matemáticas en los distintos centros.

Los dos niveles del modelo son:

$$\text{Nivel 1: } y_{ij} = \beta_{0i} + \epsilon_{ij}$$

- El subíndice j corresponde a alumnos y el i a escuelas, si se considera a las escuelas como un efecto aleatorio,
- β_{0i} (la media de cada escuela) vendría dada por:

$$\text{Nivel 2: } \beta_{0i} = \beta_0 + u_i,$$

- β_0 es la media de todos los alumnos,
- u_i es la desviación de la media de la escuela i respecto de la media de todas las escuelas.

Poniendo las dos ecuaciones juntas:

$$y_{ij} = \beta_0 + u_i + \epsilon_{ij}, \quad i = 1, \dots, m, \quad j = 1, \dots, n_m \quad (18.6)$$

- La media de y para el grupo i es $\beta_0 + u_i$,
- Los residuos a nivel individual ϵ_{ij} son la diferencia entre el valor de la variable respuesta del individuo j y la media del grupo al que pertenece,
- $u_i \sim N(0, \sigma_u^2)$, $\epsilon_{ij} \sim N(0, \sigma^2)$, y ambos son independientes, es decir, las observaciones que provienen de distintas escuelas son independientes.

En el ejemplo de las escuelas:

18.4. Caso práctico

307

```
library("lme4")
Modelo0 <- lmer(mAch ~ 1+(1 | school), data = Hsb82)
Modelo0
#> Linear mixed model fit by REML ['lmerMod']
#> Formula: mAch ~ 1 + (1 | school)
#> Data: Hsb82
#> REML criterion at convergence: 47116.79
#> Random effects:
#> Groups   Name        Std.Dev.
#> school   (Intercept) 2.935
#> Residual            6.257
#> Number of obs: 7185, groups: school, 160
#> Fixed Effects:
#> (Intercept)
#>          12.64
```

- La media total estimada es 12.64,
- La media estimada para la escuela i es: $12.64 + \hat{u}_i$, donde \hat{u}_i es el efecto aleatorio predicho para dicha escuela.

Para obtener los valores predichos de los efectos aleatorios, y ver si siguen una distribución Normal, se utiliza la función `ranef()`.

La Fig. 18.4**} permite ver los efectos aleatorios junto con sus intervalos de confianza (las escuelas han sido ordenadas atendiendo a su media para apreciar mejor la variabilidad entre las mismas). Para ello se ajusta el modelo con la función `lmer()`:

```
library("lattice")
qqmath(ranef(Modelo0, condVar = TRUE))$school
```

La etiqueta sería Fig. 18.4

Una primera aproximación para contrastar si hay o no diferencias entre los grupos sería calcular el intervalo de confianza para σ_u :

```
confint(Modelo0)
#>           2.5%    97.5%
#> .sig01     2.594729  3.315880
#> .sigma      6.154803  6.361786
#> (Intercept) 12.156289 13.117121
```

pudiéndose apreciar que el intervalo para sig01 no contiene al cero. Sin embargo, la forma más correcta de hacerlo sería utilizando el LRT (véase Eq. (18.5)) con:

$$\begin{aligned} H_0 : \quad \sigma_u^2 = 0 &\Rightarrow y_{ij} = \beta_0 + \epsilon_{ij} \\ H_1 : \quad \sigma_u^2 \neq 0 &\Rightarrow y_{ij} = \beta_0 + u_i + \epsilon_{ij}. \end{aligned}$$

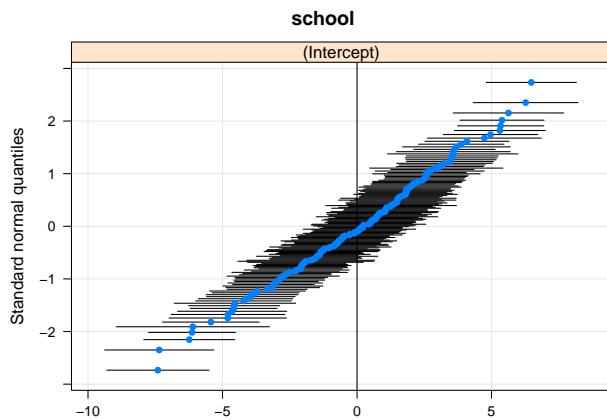


Figura 18.4: Efectos aleatorios junto con sus intervalos de confianza

El resultado del test, se compara con el valor de una mixtura de distribuciones Chi-cuadrado $0,5\chi_0^2+0,5\chi_1^2$, concretamente $0,5\chi_p^2+0,5\chi_{p-1}^2$, donde p es el número de parámetros de la varianza que se hacen cero bajo H_0 .

```
Modelo_NULL <- lm(mAch ~ 1, data = Hsb82)
test = -2*logLik(Modelo_NULL, REML = T) + 2*logLik(Modelo0, REML = T)
mean(pchisq(test, df = c(0, 1), lower.tail = F))
#> [1] 9.320673e-217
```

Conclusión: el efecto aleatorio es necesario en el modelo.

El siguiente paso sería introducir las variables explicativas (en este caso solo hay una), ya estén al nivel 1 o al 2.

18.4.1.1. Variables explicativas en el nivel 1 (individuos)

Como la variable explicativa está medida al nivel 1, se introduce en la ecuación del nivel 1:

- Nivel 1: $y_{ij} = \beta_0 i + \beta_1 x_{ij} + \epsilon_{ij}$
- Nivel 2: $\beta_0 i = \beta_0 + u_i$

Si X es una variable continua, este modelo asume que la pendiente de la recta es la misma para todas las escuelas (por eso β_1 no lleva el subíndice i). Poniendo las dos ecuaciones juntas:

$$y_{ij} = \underbrace{\beta_0 + \beta_1 x_{ij}}_{\text{efectos fijos}} + \underbrace{u_i + \epsilon_{ij}}_{\text{efectos aleatorios}}$$

18.4. Caso práctico

309

En este modelo, la relación global entre Y y X viene representada por la línea recta con ordenada en el origen β_0 y pendiente β_1 . Sin embargo, la ordenada en el origen para una determinada escuela i viene dada por $\beta_0 + u_i$. Será mayor o menor que la ordenada en el origen global β_0 en una cantidad u_i . Aunque la ordenada en el origen varía de grupo a grupo, la pendiente es la misma para todos los grupos. Todas las líneas rectas ajustadas para cada grupo son paralelas.

En el ejemplo de las escuelas, se introduce como variable explicativa `cse`s (nivel socioeconómico centrado en su media):

```
Modelo1 <- lmer(mAch ~ cses+(1 | school), data = Hsb82)
Modelo1
#> Linear mixed model fit by REML ['lmerMod']
#> Formula: mAch ~ cses + (1 | school)
#>   Data: Hsb82
#> REML criterion at convergence: 46724
#> Random effects:
#> Groups   Name        Std.Dev.
#> school   (Intercept) 2.945
#> Residual            6.084
#> Number of obs: 7185, groups: school, 160
#> Fixed Effects:
#> (Intercept)      cses
#>           12.636     2.191
```

Ahora se tienen dos efectos fijos:

$$\hat{\beta}_0 = 12,64 \\ \hat{\beta}_1 = 2,19$$

$\hat{\beta}_0$ es la nota media para alumnos con nivel socioeconómico medio (la variable está centrada). La recta media vendría dada por:

$$E[y|cses] = 12,64 + 2,19 \text{ cses}$$

Para contrastar si la variable `cse`s es significativa se utiliza el LRT (Eq. (18.5)). Primero se tienen que ajustar de nuevo los modelos que se quieren comparar usando máxima verosimilitud (en vez de REML). Si se utiliza la función `lmer()` para ajustar el modelo no es necesario reajustar con ML pues la función `anova` lo hará automáticamente, mientras que si se usa la función `lme()` sí será necesario hacerlo.

```
anova(Modelo0, Modelo1)
#> Data: Hsb82
#> Models:
#> Model0: mAch ~ 1 + (1 | school)
#> Model1: mAch ~ cses + (1 | school)
#>          npar  AIC  BIC logLik deviance Chisq Df Pr(>Chisq)
#> Model0    3 47122 47142 -23558     47116
#> Model1    4 46728 46756 -23360     46720 395.4  1 < 2.2e-16 ***
```

```
#> ---
#> Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Por lo tanto, el nivel socioeconómico afecta a los resultados escolares. Comparado con el modelo sin la variable explicativa (Modelo0), la inclusión del nivel socio-económico (Modelo1) ha reducido la variabilidad a nivel del alumno en un 2.8% ($6.084 - 6.257)/6.257 = -0.028$).

18.4.1.2. Variables explicativas en el nivel 2 (grupos)

Si las variables explicativas se miden al nivel 2, entonces:

$$\begin{aligned} \text{Nivel 1: } y_{ij} &= \beta_{0i} + \epsilon_{ij} \\ \text{Nivel 2: } \beta_{0i} &= \beta_0 + \beta_2 s_i + u_i \end{aligned}$$

$$y_{ij} = \underbrace{\beta_0 + \beta_2 s_i}_{\text{efectos fijos}} + \underbrace{u_i + \epsilon_{ij}}_{\text{efectos aleatorios}}.$$

En nuestro ejemplo, la variable utilizada es `sector` (público o católico):

$$mAch = \beta_0 + \beta_2 \text{sector} + u_i + \epsilon_{ij}$$

Se ajusta el modelo usando la función `lmer()`:

```
Modelo2 <- lmer(mAch ~ sector + (1 | school), data = Hsb82)
Modelo2
#> Linear mixed model fit by REML ['lmerMod']
#> Formula: mAch ~ sector + (1 | school)
#>   Data: Hsb82
#> REML criterion at convergence: 47080.13
#> Random effects:
#> Groups   Name        Std.Dev.
#> school   (Intercept) 2.584
#> Residual           6.257
#> Number of obs: 7185, groups: school, 160
#> Fixed Effects:
#> (Intercept) sectorCatholic
#>           11.393            2.805
```

$$E[y|sector] = 11,39 + 2,8 \text{ sector},$$

o equivalentemente

$$\begin{aligned} E[y|sector = 0] &= 11,39 \\ E[y|sector = 1] &= 11,39 + 2,8 = 14,19. \end{aligned}$$

La nota de un alumno en una escuela católica se espera que sea 2.8 unidades mayor que la de un alumno en una escuela pública (este resultado no sólo es válido para las escuelas de la

muestra sino que se puede generalizar a todas las escuelas, pues se asume que las escuelas son un efecto aleatorio). La varianza del efecto aleatorio de nivel 2, σ_u^2 , ha descendido: $(2,935^2 - 2,584^2)/2,935^2 = 0,22$, es decir, al introducir la variable sector la variabilidad a nivel de escuela se ha reducido en un 22 %.

Para contrastar si la variable sector es significativa se usa de nuevo el test LRT:

```
anova(Modelo0, Modelo2)
#> Data: Hsb82
#> Models:
#>   Modelo0: mAch ~ 1 + (1 | school)
#>   Modelo2: mAch ~ sector + (1 | school)
#>      npar  AIC  BIC logLik deviance Chisq Df Pr(>Chisq)
#> Modelo0     3 47122 47142 -23558     47116
#> Modelo2     4 47087 47115 -23540     47079 36.705  1  1.374e-09 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Por lo tanto, el hecho de que la escuela sea pública o católica influye en el resultado académico de los alumnos.

18.4.2. Modelo con pendiente aleatoria

En este tipo de modelos se supone que la relación entre la variable respuesta y las variables explicativas es distinta para las distintas unidades de nivel 2, es decir, la relación puede cambiar de un centro educativo a otro. Por ejemplo, el efecto del nivel socioeconómico en las notas puede ser distinto en distintos centros, de modo que se puede relajar el modelo anterior, en el que la pendiente era la misma para todos los grupos, permitiendo que la pendiente varíe aleatoriamente entre los grupos.

$$\begin{aligned} \text{Nivel 1: } & y_{ij} = \beta_{0i} + \beta_{1i}x_{ij} + \epsilon_{ij} \\ \text{Nivel 2: } & \beta_{0i} = \beta_0 + u_i \\ & \beta_{1i} = \beta_1 + v_i \end{aligned}$$

Poniendo las dos ecuaciones juntas:

$$y_{ij} = \underbrace{\beta_0 + \beta_1 x_{ij}}_{\text{efectos fijos}} + \underbrace{u_i + v_i x_{ij} + \epsilon_{ij}}_{\text{efectos aleatorios}}, \quad \begin{pmatrix} u_i \\ v_i \end{pmatrix} \sim N(0, \mathbf{G}_i), \quad \mathbf{G}_i = \begin{pmatrix} \sigma_u^2 & \\ \sigma_{uv} & \sigma_v^2 \end{pmatrix},$$

donde σ_{uv} es la covarianza entre las ordenadas en el origen y las pendientes de los grupos (β_{0i} y β_{1i} , respectivamente). Un valor positivo de la covarianza implica que los grupos con un valor del efecto de grupo u_i elevado tienden a tener valores elevados de v_i , o equivalentemente, centros educativos con ordenada en el origen alta, tienen pendiente alta.

El modelo en R sería:

```
Modelo3 <- lmer( mAch ~ cses + (cses | school), data = Hsb82)
Modelo3
#> Linear mixed model fit by REML ['lmerMod']
#> Formula: mAch ~ cses + (cses | school)
#>   Data: Hsb82
#> REML criterion at convergence: 46714.23
#> Random effects:
#> Groups   Name        Std.Dev. Corr
#> school   (Intercept) 2.9464
#>         cses        0.8331  0.02
#> Residual            6.0581
#> Number of obs: 7185, groups: school, 160
#> Fixed Effects:
#> (Intercept)      cses
#>           12.636    2.193
```

El efecto del nivel socioeconómico en la escuela i se estima como $2,19 + \hat{u}_i$, y la varianza de las pendientes para las escuelas es $0,833^2 = 0,694$. Para la *escuela promedio* se predice un aumento de 2,19 en la puntuación cuando el nivel socioeconómico aumenta en una unidad.

Ahora se tienen las siguientes estimaciones de los parámetros de varianza:

$$\hat{\sigma}_u^2 = 8,68 \quad \hat{\sigma}_v^2 = 0,694 \quad \hat{\sigma}_{uv} = \rho\sigma_u\sigma_v = 0,051 \quad \hat{\sigma}^2 = 36,7$$

La varianza de la ordenada en el origen estimada, 8,68, se interpreta como la variabilidad (de la nota) entre las escuelas para un nivel socioeconómico medio (valor nulo de la variable por estar centrada).

El parámetro de covarianza estimado es $\sigma_{uv} = 0,051$, por lo que se puede plantear si es necesario o no.

Para comprobarlo, el contraste de hipótesis sería en este caso:

$$H_0 : \sigma_{uv} = 0 \quad \text{y} \quad H_1 : \sigma_{uv} \neq 0$$

```
Modelo3.1 <- lmer(ses ~ cses + (cses || school), data = Hsb82)
```

Cuando se quiere que haya un efecto aleatorio para la ordenada en el origen y para la pendiente pero que estén incorrelados, en la función solo hay que poner doble barra en vez de simple.

En este caso no es necesario utilizar la mixtura de distribuciones Chi-cuadrado para contrastar $H_0 : \sigma_{uv} = 0$, pues σ_{uv} puede tomar cualquier valor.

```
anova(Modelo3.1, Modelo3)
#> Data: Hsb82
#> Models:
#> Model3.1: ses ~ cses + ((1 / school) + (0 + cses / school))
#> Model3: mAch ~ cses + (cses | school)
```

18.4. Caso práctico

313

```
#>      npar      AIC      BIC logLik deviance Chisq Df Pr(>Chisq)
#> Modelo3.1    5 -226164 -226129 113087   -226174
#> Modelo3     6  46723  46764 -23355    46711      0  1          1
```

Por lo tanto, se puede suponer que la covarianza es 0.

El siguiente paso sería contrastar si es necesario que las rectas tengan pendientes diferentes, es decir, $H_0: \sigma_v^2 = 0$, $H_1: \sigma_v^2 > 0$. En este caso sí se necesita la aproximación:

```
test <- -2 * logLik(Modelo1, REML = T) +
       2 * logLik(Modelo3.1, REML = T)
mean(pchisq(test, df = c(0, 1), lower.tail = F))
#> [1] 0
```

Por lo tanto, la pendiente es diferente en las distintas escuelas.

Además, se puede usar algún criterio de información para comparar los modelos:

```
AIC(logLik(Modelo3))
#> [1] 46726.23
AIC(logLik(Modelo3.1))
#> [1] -226113.6
AIC(logLik(Modelo1))
#> [1] 46732
```

A veces la covariable medida a nivel 2 (a nivel de grupo, en este caso escuelas) puede explicar tanto la variabilidad de la ordenada en el origen como de la pendiente:

$$\begin{aligned} \text{Nivel 1: } y_{ij} &= \beta_{i0} + \beta_{1i}x_{ij} + \epsilon_{ij} \\ \text{Nivel 2: } \beta_{i0} &= \beta_0 + \beta_2 s_i + u_i \\ \beta_{1i} &= \beta_1 + \beta_3 s_i + v_i \\ y_{ij} &= \underbrace{\beta_0 + \beta_1 x_{ij} + \beta_2 s_i + \beta_3 x_{ij} s_i}_{\text{efectos fijos}} + \underbrace{u_i + v_i x_{ij} + \epsilon_{ij}}_{\text{efectos aleatorios}}. \end{aligned}$$

Al introducir la variable medida al nivel 2, la parte fija se modifica (con respecto al Modelo 3), pero no la parte aleatoria:

```
Modelo4 <- lmer(mAch ~ cses*sector + (cses || school),
                 data = Hsb82)
Modelo4
#> Linear mixed model fit by REML ['lmerMod']
#> Formula: mAch ~ cses * sector + ((1 | school) + (0 + cses | school))
#> Data: Hsb82
#> REML criterion at convergence: 46648.85
#> Random effects:
```

```
#> Groups      Name           Std.Dev.
#> school     (Intercept) 2.5971
#> school.1   cses         0.5182
#> Residual    6.0580
#> Number of obs: 7185, groups: school, 160
#> Fixed Effects:
#>             (Intercept)      cses      sectorCatholic
#>             11.393          2.784        2.805
#> cses:sectorCatholic
#>             -1.346
```

Los centros católicos tienen una nota media más alta que los públicos (2.81 puntos más), y una pendiente más suave que la de dichos centros públicos (-1.35). Esto último indica que en un colegio católico la mejora de la nota con respecto al nivel socio-económico es más lenta que un colegio público.

18.4.3. ¿Cómo construir el modelo en la práctica?

1. Se ajusta el modelo con todos los efectos fijos y aleatorios posibles.

```
Modelo5 <- lmer(mAch ~ cses * sector+(cses | school), data = Hsb82)
```

2. Se contrasta qué efectos aleatorios son significativos, sin mover los efectos fijos.

Primero se contrasta si la covarianza entre efectos fijos y aleatorios es cero o no:

```
#Se ajusta el modelo con covarianza = 0
Modelo5.1 <- lmer(ses ~ cses * sector+(cses||school),
                  data = Hsb82)
anova(Modelo5.1, Modelo5)
#> Data: Hsb82
#> Models:
#> Modelo5.1: ses ~ cses * sector + ((1 / school) + (0 + cses / school))
#> Modelo5: mAch ~ cses * sector + (cses / school)
#>           npar      AIC      BIC logLik deviance Chisq Df Pr(>Chisq)
#> Modelo5.1    7 -220069 -220021 110042   -220083
#> Modelo5     8  46650  46705 -23317    46634      0   1       1
```

Como lo es, no se tendría que contrastar nada más. Los efectos aleatorios son los que se han incluido en el Modelo5. Si no hubiera sido significativa, se continuaría contrastando si la pendiente aleatoria es significativa y si la ordenada en el origen lo es.

3. Una vez elegidos los efectos aleatorios que se mantienen en el modelo, se eligen los efectos fijos:

18.4. Caso práctico

315

```

Modelo6 = update(Modelo5, . ~ .-cses:sector)
anova(Modelo6, Modelo5)
#> Data: Hsb82
#> Models:
#> Modelo6: mAch ~ cses + sector + (cses / school)
#> Modelo5: mAch ~ cses * sector + (cses / school)
#>      npar   AIC   BIC logLik deviance Chisq Df Pr(>Chisq)
#> Modelo6    7 46678 46726 -23332     46664
#> Modelo5    8 46650 46705 -23317     46634 29.983  1  4.358e-08 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Dado que la interacción es significativa, se opta por dejar los efectos fijos `cse`s y `sector` en el modelo (en aras de facilitar las interpretaciones), por lo que no sería necesario contrastar su significatividad y este sería el modelo final.

Resumen

En este capítulo se introducen los modelos mixtos o modelos con efectos aleatorios. En particular:

- Se dan las claves para distinguir entre efectos fijos y aleatorios.
- Se presenta la formulación del modelo y indica cómo llevar a cabo la estimación del mismo.
- Se explican las etapas del proceso a seguir para el ajuste de este tipo de modelos.
- Se muestra el uso de **R** para ajustar estos modelos.
- Se ilustra el análisis de modelos multinivel como caso particular de un modelo con efectos aleatorios.

Capítulo 19

Modelos *sparse* y métodos penalizados de regresión

María Durbán

Universidad Carlos III de Madrid

19.1. Introducción

El modelo de regresión lineal múltiple: $y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \varepsilon$, visto en el Cap. 15, a pesar de su simplicidad, tiene importantes ventajas como la **interpretabilidad** y su buen poder **predictivo** en muchas situaciones.

En este capítulo enseña cómo se puede hacer el modelo aún más interpretable y mejor predictor, y para conseguirlo se reemplaza el método de estimación de mínimos cuadrados por un método alternativo. Más concretamente, el objetivo de este capítulo es presentar técnicas para mejorar la:

- **Precisión de la predicción:** en particular, cuando el número de variables es mayor que el número de observaciones: $p > n$ (algo que ocurre con mucha frecuencia hoy en día). En este caso no se pueden utilizar mínimos cuadrados ya que la matriz de diseño no es de rango completo y, por lo tanto, no se puede encontrar una solución única al problema de minimización. Por ello, se necesita reducir el número de variables, que además, evitará que se sobreajusten los datos.
- **Interpretabilidad del modelo:** al eliminar las variables irrelevantes (es decir, haciendo cero los correspondientes coeficientes) se obtendrá un modelo más fácil de interpretar.

En base a lo anterior,a continuación se presentan varios métodos para llevar a cabo de forma automática la reducción de variables en el modelo, actividad también denominada **selección de variables**. Tales métodos son:

- **Selección del mejor subconjunto:** su objetivo es identificar el subconjunto de $k < p$ predictores que contenga sólo los que mejor expliquen el comportamiento de la variable respuesta.
- **Shrinkage:** en este caso no se quieren seleccionar variables explícitamente, sino que se añade una penalización que penaliza el número de coeficientes o su tamaño.
- **Reducción de la dimensión:** el objetivo es proyectar los p -predictores en un subespacio de dimensión más pequeña (mediante el uso de combinaciones lineales de las variables predictoras, las cuales se usarán como “nuevos” predictores). Dichas combinaciones lineales se llaman **componentes principales** y a su análisis se dedica el Cap. 32.

En este Capítulo se ven los dos primeros métodos. Para el tercero, se remite al lector al Cap. 32.

19.2. Selección del mejor subconjunto

Supóngase que se tiene acceso a p variables predictoras, pero se quiere un modelo más simple que involucre sólo a un subconjunto de esos p predictores. La forma lógica de conseguirlo es considerar todos los posibles subconjuntos de los p predictores y elegir el mejor modelo de entre todos los modelos construidos con cada uno de los subconjuntos de variables. Los pasos a seguir serían:

1. Se crea el modelo **nulo**, M_0 , que es aquel que únicamente contiene la ordenada en el origen y ningún predictor. Este modelo simplemente predice la media muestral para cada observación.
2. Para cada valor de $k = 1, 2, \dots, p$, se calculan los $\binom{p}{k}$ modelos que contienen k predictores. Es decir, los p modelos que contienen 1 predictor, los $p \times (p - 1)/2$ modelos que contienen 2 predictores, etc.
3. Para cada valor de k , se elige el mejor entre los $\binom{p}{k} = \frac{p!}{(p-k)!k!}$ posibles modelos y se denota por M_k . Es decir, M_1 sería el mejor modelo entre los p modelos con una única variable, M_2 sería el mejor modelo entre los modelos con dos variables, etc. En este caso, el **mejor** modelo sería aquel cuyo RSS (suma de residuos al cuadrado) sea menor, o equivalentemente, aquel cuyo R^2 sea mayor.
4. Finalmente, entre los modelos: M_1, \dots, M_p se elige el mejor utilizando un criterio como AIC (criterio de información de Akaike), BIC (criterio de información bayesiano) o R^2 ajustado.

Este método se puede usar también en el caso de GLMs, si bien, en este caso, se usa la *deviance* en vez de RSS .

19.2.1. Procedimiento con R: la función `regsubset()`

En esta subsección se aplica el método descrito al conjunto de datos `Hitters` del paquete `ISLR2`. El objetivo es predecir el sueldo, `Salary`, de jugadores de béisbol a partir de varias variables asociadas con su rendimiento el año anterior.

La variable `Salary` no está disponible para algunos de los jugadores. Éstos se pueden identificar con la función `is.na()`. La función `sum()` permite ver cuántos hay. Se utiliza `na.omit()` para eliminarlos.

```
library("ISLR2")
Hitters <- na.omit(Hitters)
```

La función `regsubsets()` del paquete `leaps` lleva a cabo la selección del mejor subconjunto de variables predictoras, identificando el mejor modelo que contiene un número dado de ellas (1,2,3, etc.) atendiendo a *RSS*. La sintaxis usada es similar a la de la función `lm()`.

```
library("leaps")
regfit_full <- regsubsets(Salary ~ ., Hitters)
```

Los resultados se pueden ver usando `summary()`, donde se muestra el mejor modelo para cada número específico de variables. Las variables incluidas en cada modelo se indican con un asterisco. Por ejemplo, el mejor modelo con dos variables incluye `Hits` y `CRBI`. Por defecto, `regsubsets()` solo muestra los resultados de los modelos que contienen hasta ocho variables. La opción `nvmax` se puede usar para incrementar esta cantidad, por ejemplo hasta 19 variables (que es el número de variables predictoras en el conjunto de datos):

```
regfit_full <- regsubsets(Salary ~ .,
  data = Hitters,
  nvmax = 19
)
reg_summary <- summary(regfit_full)
```

La función `summary()` devuelve diferentes medias de bondad de ajuste: R^2 , RSS , R^2 ajustado, C_p y BIC que se utilizan para elegir el *mejor* de entre todos los modelos.

```
names(reg_summary)
#> [1] "which"   "rsq"     "rss"     "adjr2"   "cp"      "bic"     "outmat"  "obj"
reg_summary$adjr2
#> [1] 0.3188503 0.4208024 0.4450753 0.4672734 0.4808971 0.4972001 0.5007849
#> [8] 0.5137083 0.5180572 0.5222606 0.5225706 0.5217245 0.5206736 0.5195431
#> [15] 0.5178661 0.5162219 0.5144464 0.5126097 0.5106270
```

En el ejemplo, el R^2 ajustado mayor corresponde al modelo con 11 variables.

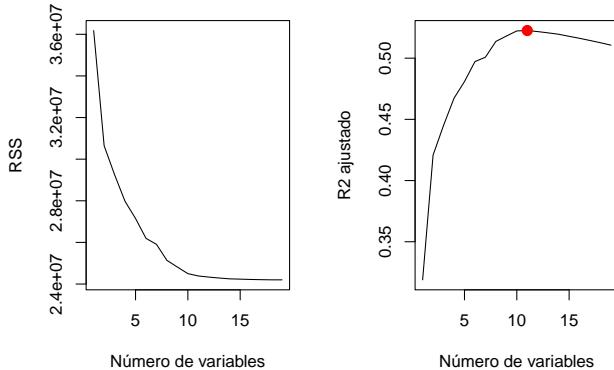


Figura 19.1: Valores de R^2 y R^2 ajustados correspondientes a modelos con distinto número de variables

Los resultados también se pueden mostrar y dibujar simultáneamente; por ejemplo, los valores de RSS y R^2 ajustado de todos los modelos se muestran en la Fig. 19.1.

Otra manera de visualizar los resultados es:

```
plot(regfit_full, scale = "adjr2")
```

La primera fila tiene un cuadrado negro en cada una de las variables explicativas del modelo con mayor R^2 ajustado (en este caso, sería similar para los otros criterios).

Varios modelos tienen un valor de R^2 ajustado próximo a 0,52, pero es el modelo con 11 variables el que alcanza el mayor valor. La función `coef()` permite ver los coeficientes estimados de este modelo.

```
coef(regfit_full, 11)
#> (Intercept)      AtBat       Hits       Walks      CATBat      CRUNS
#> 135.7512195 -2.1277482  6.9236994  5.6202755 -0.1389914  1.4553310
#> CRBI          CWalks     LeagueN   DivisionW    PutOuts     Assists
#>  0.7852528 -0.8228559 43.1116152 -111.1460252   0.2894087  0.2688277
```

19.2.2. Selección *stepwise*

Cuando el número de variables predictoras, p , es grande, el método anterior es computacionalmente muy costoso ya que el número de posibles combinaciones de variables crece de una manera alarmante. En general, la función `regsubset()` puede lidiar con hasta 30-40 variables predictoras. Además, otro problema es el sobreajuste. Si se tienen 40 variables, se estarían ajustando millones de modelos, y puede que el modelo elegido funcione muy bien en los datos

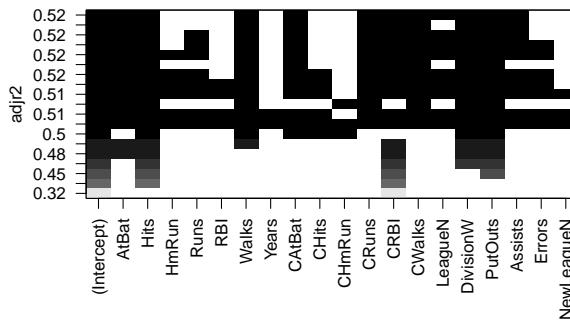


Figura 19.2: Variables seleccionadas en cada uno de los modelos y su correspondiente valor de R^2 ajustado.

utilizados para su construcción, pero no tan bien en un nuevo conjunto de datos. Una alternativa es el método **stepwise**. La idea detrás de este método es similar a la anterior, pero se busca el mejor modelo entre un conjunto mucho más pequeño de modelos.

Hay dos posibilidades de hacer stepwise: **forward** y **backward**. Ambas son bastante parecidas; la principal diferencia es el modelo del que se parte: del modelo sin ninguna variable predictora (**forward**) o del modelo con todas ellas (**backward**).

19.2.2.1. Forward stepwise

En este caso se comienza con el **modelo nulo**, M_0 , y se van añadiendo variables secuencialmente. En particular, en cada paso (*step*) la variable que proporciona la mayor mejora al ajuste es la que se añade al modelo. Los pasos a seguir serían:

1. Se crea el modelo nulo, M_0 .
2. Para cada valor de $k = 0, 1, 2, \dots, p$:
 - a) Se consideran todos los $p - k$ modelos que surgen de aumentar el modelo M_k con un predictor.
 - b) Se elige el **mejor** de esos $p - k$ modelos, que se denota M_{k+1} . El término **mejor** significa tener el RSS más bajo o el R^2 más alto.
3. Se elige el mejor de los modelos M_0, \dots, M_p en función de un criterio que tenga en cuenta la complejidad del modelo, como AIC, BIC o R^2 ajustado.

Este enfoque tiene ventajas computacionales claras, ya que el número de modelos ajustados es mucho menor, pero no garantiza que el modelo elegido sea el mejor modelo posible, especialmente si existe correlación entre las variables predictoras.

19.2.2.2. Backward stepwise

En este caso, se comienza con el modelo que incluye las p las variables predictoras y se van eliminando de forma iterativa hasta llegar al modelo nulo (M_0). Los pasos serían:

1. Se ajusta el modelo M_p , que contiene todas (p) las variable predictoras.
2. Para cada valor de $k = p, p-1, \dots, 1$: *a)* Se consideran todos los k modelos que surgen de reducir en el modelo M_k un predictor, es decir, modelos con $k-1$ variables predictoras.
b) Se elige el **mejor** de esos k modelos, que se denota M_{k-1} . Dicho modelo será el que tenga el RSS más bajo o el R^2 más alto.
3. Entre los modelos M_0, \dots, M_p se elige el mejor en función de un criterio como AIC, BIC o R^2 ajustado.

Tanto en el caso *forward* como en el caso *backward*, se busca el mejor modelo “sólo” entre $1+p(p+1)/2$ modelos, lo que los hace recomendables frente a la selección del mejor subconjunto de variables cuando p es demasiado grande..

El método *backward stepwise* necesita que el número de observaciones n sea mayor que el de variables predictoras p (ya que necesita ajustar el modelo con todas las variables). Por el contrario, el método *forward stepwise* se puede usar incluso cuando $n < p$.

19.2.2.3. Procedimiento con R: la función `regsubset()`

La función `regsubset()` permite utilizar los métodos *forward* y *backward*, usando los argumentos `method = "forward"` o `method = "backward"`:

```
regfit_fwd <- regsubsets(Salary ~ .,
  data = Hitters,
  nvmax = 19, method = "forward"
)
regfit_bwd <- regsubsets(Salary ~ .,
  data = Hitters,
  nvmax = 19, method = "backward"
)
```

Los métodos mejor subconjunto de variables, *forward stepwise* y *backward stepwise* no tienen por qué seleccionar el mismo (mejor) modelo. Ni siquiera tienen por qué seleccionar el mismo modelo para cada número de predictores en la fase previa a la selección final. Así ocurre, por ejemplo, cuando el número de variables predictoras es $k = 2$:

```
coef(regfit_full, 2)
#> (Intercept)      Hits       CRBI
#> -47.9559022   3.3008446   0.6898994
coef(regfit_fwd, 2)
#> (Intercept)      Hits       CRBI
```

19.2. Selección del mejor subconjunto

323

```
#> -47.9559022  3.3008446  0.6898994
coef(regfit_bwd, 2)
#> (Intercept)      Hits       CRuns
#> -50.8174029   3.2257212  0.6614168
```

En la etapa final de selección, el modelo seleccionado no tiene por qué ser el mismo en función de los distintos criterios de selección (aunque normalmente lo es). Lo habitual es decidir un criterio para elegir el mejor modelo (R^2 ajustado, BIC , etc.) y seleccionarlo en función de él. En el ejemplo, seleccionando el criterio del R^2 ajustado, el mejor modelo es el que tiene 11 variables, tanto con el criterio *forward* como con el *backward*¹:

```
which.max(summary(regfit_fwd)$adjr2)
#> [1] 11
which.max(summary(regfit_bwd)$adjr2)
#> [1] 11
```

Con el criterio BIC también se selecciona el modelo con 11 variables predictoras:

```
which.min(summary(regfit_fwd)$bic)
#> [1] 6
which.min(summary(regfit_bwd)$bic)
#> [1] 8
```

Otra posibilidad es utilizar como criterio de selección el error de predicción, y para ello se puede echar mano de algún esquema de validación cruzada. A continuación se ilustra el caso en el que se divide la muestra en dos subconjuntos: *training* y *testing*, pero se puede utilizar cualquier otro método (validación cruzada k-grupos*, etc. Véase Sec. 10.4.

```
set.seed(1)
entreno <- sample(c(TRUE, FALSE), nrow(Hitters), replace = TRUE)
test <- (!entreno)
```

Se utiliza `regsubsets()` en la muestra de entrenamiento para obtener los modelos con distinto número de variables predictoras:

```
regfit_best <- regsubsets(Salary ~ ., data = Hitters[entreno, ], nvmax = 19)
```

Para calcular el error de predicción, dado que la función `regsubset()` no tiene asociada una función `predict()`, se han de calcular “manualmente” los valores predichos para la muestra de test. Para eso se necesita la matriz de diseño del modelo.

¹Recuérdese que con el método del mejor subconjunto de variables predictoras el criterio del R^2 ajustado también seleccionó el modelo con 11 variables.

```
test.mat <- model.matrix(Salary ~ ., data = Hitters[test, ])
```

Ahora, para cada modelo de tamaño k , se extraen los coeficientes de `regfit_best` para el mejor modelo de ese tamaño, se multiplica el vector de coeficientes por la matriz de diseño y se obtienen las predicciones; a continuación se calcula el error cuadrático medio (MSE).

```
val_errors <- rep(NA, 19)
for (i in 1:19) {
  coefi <- coef(regfit_best, id = i)
  pred <- test.mat[, names(coefi)] %*% coefi
  val_errors[i] <- mean((Hitters$Salary[test] - pred)^2)
}
```

Con este criterio, el mejor modelo es el que contiene 7 variables:

```
val_errors
#> [1] 164377.3 144405.5 152175.7 145198.4 137902.1 139175.7 126849.0 136191.4
#> [9] 132889.6 135434.9 136963.3 140694.9 140690.9 141951.2 141508.2 142164.4
#> [17] 141767.4 142339.6 142238.2
```

19.3. Métodos *shrinkage*

Los métodos anteriores se basan en el ajuste de modelos mediante mínimos cuadrados ordinarios. Los métodos *shrinkage*, sin embargo, se basan en una modificación del procedimiento de mínimos cuadrados ordinarios que consiste en añadir una penalización que *encoje* los coeficientes del modelo (normalmente hacia 0). Una de las ventajas de este tipo de métodos es que reduce la varianza de los coeficientes estimados.

Recuérdese que en el ajuste por mínimos cuadrados las estimaciones de $\beta_0, \beta_1, \dots, \beta_p$ son los valores que minimizan:

$$RSS = \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2.$$

19.3.1. Regresión *ridge*

La **regresión ridge**² añade un término de penalización controlado por un parámetro (que habrá que elegir) que penalizará la magnitud de los coeficientes. Cuanto más grande es el coeficiente mayor es la penalización. En consecuencia, en la regresión *ridge* la expresión que se minimiza

²La traducción en español sería regresión contraída o regresión alomada.

para obtener las estimaciones de los parámetros del modelo es:

$$\sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p \beta_j^2 = RSS + \lambda \sum_{j=1}^p \beta_j^2. \quad (19.1)$$

En realidad, lo que se está haciendo es hacer pagar al modelo un precio (en términos de ajuste) por el hecho de que los coeficientes no sean cero, y el precio será tanto mayor cuanto más grande sea la magnitud del coeficiente. A esta penalización se le llama **penalización shrinkage** porque “anima” a los coeficientes a que se *contraigan* hacia 0 (así es como este método favorece la simplicidad de los modelos). La magnitud de dicha contracción está gobernada por lambda, el parametro de afinado o regulación (también conocido en la jerga como “de tuneado”). Si $\lambda = 0$, se está en el caso de mínimos cuadrados ordinarios, y cuanto mayor sea λ , mayor será el precio a pagar para que esos coeficientes sean distintos de 0. Si λ es extremadamente grande, los coeficientes estarán muy próximos a 0, para que el segundo término pequeño (recuérdese que se minimiza RSS más la penalización). Aunque valores más grandes de los coeficientes proporcionasen un mejor ajuste (y por lo tanto un menor RSS), el término de penalización aumentaría se hará grande y no se alcanzaría el mínimo. Por lo tanto λ gobierna el equilibrio entre un buen ajuste del modelo y el tamaño de los coeficientes (y por lo tanto el número de coeficientes distintos de cero).

La elección del valor de λ es un punto crucial de este tipo de regresión. Para su determinación se suelen utilizar procedimientos de validación cruzada.

19.3.1.1. Escalado de variables predictoras

Un punto importante en regresión ridge es si las variables predictoras están escaladas o no.

El método de mínimos cuadrados ordinarios es *invariante a la escala (scale-invariant)*, es decir, que si se multiplica una variable predictora X_j por una constante c , el coeficiente estimado se multiplicada por $1/c$, pero $X_j \hat{\beta}_j$ no cambia. Sin embargo, en el caso de la regresión *ridge* los coeficientes estimados pueden cambiar sustancialmente ante un cambio de escala (es decir, si se multiplica una variable predictora por una constante), ya que todos los coeficientes forman parte del término de penalización. Por lo tanto, antes de utilizar la regresión *ridge* (o cualquier método de regularización) es importante **estandarizar las variables predictoras**, dividiendo cada variable por su desviación estándar, de forma que todas tengan desviación estándar igual a 1:

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{N} \sum_{i=1}^N (x_{ij} - \bar{x}_{ij})^2}}.$$

Con esto se consigue que los coeficientes estén en “igualdad de condiciones”.

En muchas ocasiones la regresión *ridge* da lugar a un menor MSE que el obtenido con mínimos cuadrados ordinarios. Sin embargo, por muy grande que sea λ los coeficientes no serán 0, sino que estarán próximos a cero, por lo que **este método no es realmente un método de selección de variables**.

Sin embargo, la regresión ridge puede ser muy útil cuando hay variables predictoras altamente correlacionadas pero se desea mantener todas en el modelo. En estos casos, la regresión ridge soluciona los problemas de multicolinealidad.

19.3.1.2. Procedimiento con R: la función `glmnet()`

Para llevar a cabo la regresión *ridge* (y para otros métodos de regresión *shrinkage*) se usa el paquete `glmnet`. La función principal en este paquete se llama también `glmnet()`. Esta función tiene una sintaxis un poco diferente a las funciones usuales para el ajuste de distintos modelos en **R**. Es necesario pasarle la matriz \mathbf{X} de variables predictoras (sin la columna correspondiente a la ordenada en el origen) y el vector \mathbf{y} con la variable respuesta. Para ilustrar su uso se utilizan los datos anteriores sobre béisbol.

```
x <- model.matrix(Salary ~ ., Hitters)[, -1]
y <- Hitters$Salary
```

La función `glmnet()` tiene un argumento, `alpha`, que determina el tipo de penalización que se añade en el modelo. En el caso de regresión *ridge*, `alpha=0`.

Por defecto, la función `glmnet()` elige de forma automática el rango de valores de λ . Sin embargo, a modo ilustrativo, se va a elegir la rejilla de valores que van desde $\lambda = 10^{10}$ hasta $\lambda = 10^{-2}$, cubriendo de esta forma una gran gama de escenarios, desde el modelo nulo (solo la ordenada en el origen) hasta el caso de mínimos cuadrados ordinarios. Más adelante se verá que se puede llevar a cabo el ajuste del modelo para un valor determinado de λ que no esté entre los de la rejilla inicial.

```
library("glmnet")
grid <- 10^seq(10, -2, length = 100)
ridge_mod <- glmnet(x, y, alpha = 0, lambda = grid)
```

Por defecto, la función `glmnet()` estandariza las variables predictoras para que estén en la misma escala. Si por alguna razón no se quisiera hacer, se usaría `standardize = FALSE`.

Asociado con cada valor de λ hay un vector de coeficientes estimados mediante regresión ridge almacenados en un matriz accesible utilizando `coef()`. En este caso, el tamaño de la matriz es 20×100 , donde las 20 filas corresponden a cada uno de los predictores más la ordenada en el origen y las 100 columnas a cada valor de λ . Lo esperable es que los coeficientes estimados sean más pequeños cuanto mayor sea el valor de λ . A continuación, se muestra el valor de los coeficientes cuando $\lambda = 11,498$, así como la suma de sus cuadrados, $\sum_{j=1}^p \beta_j^2$:

```
ridge_mod$lambda[50]
#> [1] 11497.57
sum(coef(ridge_mod)[-1, 50]^2)
#> [1] 40.45739
```

Por el contrario, si λ es más pequeño, 705, el valor de su suma de cuadrados es mucho mayor.

19.3. Métodos shrinkage

327

```
ridge_mod$lambda[60]
#> [1] 705.4802
sum(coef(ridge_mod)[-1, 60]^2)
#> [1] 3261.554
```

La Fig. 19.3 muestra el efecto de λ en los coeficientes del modelo:

```
plot(ridge_mod, xvar = "lambda", label = TRUE)
```

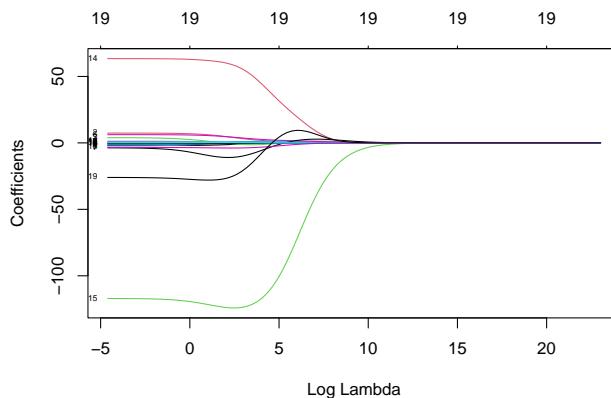


Figura 19.3: Coeficientes estimados para distintos valores del parámetro de penalización (en la escala logarítmica)

El lado izquierdo de la Fig 19.3 corresponde a valores de λ muy pequeños, y por lo tanto no existen restricciones sobre los coeficientes. Conforme aumenta el valor de λ los coeficientes se aproximan rápidamente a cero. Pero no todos se aproximan a cero de la misma manera: hay un conjunto de variables cuyo coeficiente es prácticamente cero para cualquier valor de λ , mientras que para un valor de $\log(\lambda) = 3$ parece que hay sólo 4 coeficientes distintos de 0.

La función `predict()` se puede utilizar con diferentes propósitos. Por ejemplo, se pueden obtener los coeficientes de la regresión `ridge` para un valor específico de λ , por ejemplo $\lambda = 50$:

```
predict(ridge_mod, s = 50, type = "coefficients")[1:20, ]
```

En lo que sigue, se ilustra el ajuste de una regresión ridge y se computa el *MSE* de predicción para distintos valores de λ . Primeramente, se divide el conjunto de datos en un subconjunto de entrenamiento y otro de test:

```
set.seed(1)
entreno <- sample(1:nrow(x), nrow(x) / 2)
test <- (-entreno)
y_test <- y[test]
```

A continuación se ajusta la regresión ridge a los datos del subconjunto de entrenamiento usando un valor específico de λ (por ejemplo $\lambda = 4$). Posteriormente, se evalúa su *MSE* con los datos del subconjunto de test. Para ello se usa la función `predict()`. En este caso, para obtener las predicciones para la muestra de test, se reemplaza `type = "coefficients"` por el argumento `newx`.

```
ridge_mod <- glmnet(x[entreno, ], y[entreno], alpha = 0, lambda = grid)
ridge_pred <- predict(ridge_mod, s = 4, newx = x[test, ])
mean((ridge_pred - y_test)^2)
#> [1] 142226.5
```

El *MSE* es 142,199. Si se usa un valor muy alto de λ , por ejemplo 10^{10} (esto sería equivalente a ajustar un modelo solo con la ordenada en el origen), el resultado es muy distinto:

```
ridge_pred <- predict(ridge_mod, s = 1e10, newx = x[test, ])
mean((ridge_pred - y_test)^2)
#> [1] 224669.8
```

Por lo tanto, en este caso, ajustar un modelo de regresión ridge con $\lambda = 4$ da un *MSE* mucho menor que el obtenido cuando el modelo sólo contiene la ordenada en el origen.

A continuación se compara el resultado para $\lambda = 4$ con el obtenido utilizando mínimos cuadrados ordinarios ($\lambda = 0$).³

```
ridge_pred <- predict(ridge_mod, s = 0, newx = x[test, ], exact = T,
                      x = x[entreno, ], y = y[entreno])
mean((ridge_pred - y_test)^2)
#> [1] 167018.2
```

Se observa que el *MSE* es menor cuando se usa regresión *ridge* (con $\lambda = 4$) que cuando se usan mínimos cuadrados ordinarios.

Hasta ahora se ha elegido el valor $\lambda = 4$ de forma arbitraria. En la siguiente sección se aborda la cuestión de cómo seleccionar el valor de dicho parámetro de una forma automática.

19.3.2. Selección del parámetro de penalización

En la subsección anterior se ha visto que el valor de λ tiene un gran impacto en los resultados obtenidos cuando se utiliza un modelo con penalización.

³Además se ha de añadir ‘exact = T’ en la función ‘predict()’

Una buena manera de elegir λ es usar validación cruzada (*cross-validation*). Por ejemplo, se puede usar validación cruzada con 10 grupos (*k-fold cross-validation*) :

- Se dividen los datos en k grupos, se ajusta el modelo ridge a $k - 1$ de esos grupos (para una rejilla de valores de λ) y se calcula el error de predicción para el otro grupo.
- La acción anterior se repite tomando como muestra de test cada uno de los k grupos y se suman los errores de predicción.
- Al final se dispondrá de una curva con los errores para cada valor de λ y se elegirá el que dé el mínimo error.

En la práctica, el procedimiento anterior se puede hacer con la función `cv.glmnet()`. Por defecto, esta función usa un *10-fold cross-validation*, pero el número de grupos se puede cambiar usando el argumento `nfolds`.

En el ejemplo del béisbol:

```
set.seed(1)
cv_out <- cv.glmnet(x[entreno, ], y[entreno], alpha = 0)
plot(cv_out)
```

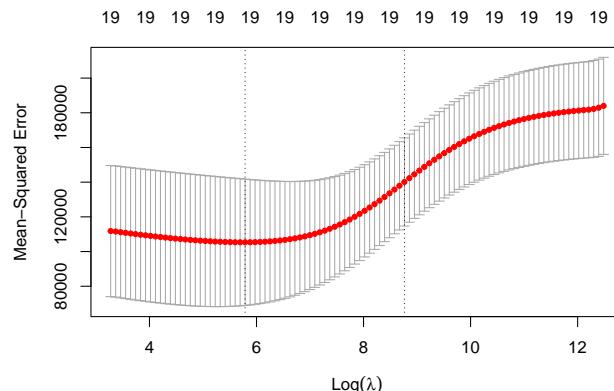


Figura 19.4: Valor del error cuadrático medio y su intervalo de confianza (calculado sobre los 10 grupos) para distintos valores del parámetro de penalización

```
mejorlam <- cv_out$lambda.min
mejorlam
#> [1] 326.0828
```

En la Fig. 19.4, los puntos rojos corresponden a la media del MSE para los 10 grupos y las barras superior e inferior corresponden a esa cantidad más/menos una desviación estándar (el

ancho será tanto menor cuanto mayor sea el número de grupos). La primera línea vertical corresponde al valor de λ que hace mínimo el MSE y la segunda es el valor que está a una distancia de una desviación típica del λ mínimo (usar este último valor podría ser una buena opción para evitar el sobre-ajuste, es decir dejar demasiadas variables en el modelo).

El valor mínimo del MSE se calcula como sigue:

```
ridge_pred <- predict(ridge_mod, s = mejorlam, newx = x[test, ])
mean((ridge_pred - y_test)^2)
#> [1] 139833.6
```

Como se puede apreciar, hay una apreciable mejora en el error de predicción que se había obtenido cuando el parámetro de penalización se había fijado en $\lambda = 4$.

19.3.3. Regresión *lasso*

Uno de los puntos débiles de la regresión *ridge* es que no hace selección de variables (los coeficientes pueden estar próximos a cero pero no ser exactamente cero). En el modelo final se incluyen todos los coeficientes y, por lo tanto, **la regresión *ridge* sólo es útil cuando la mayoría de las variables predictoras tienen un impacto significativo en la respuesta.**

La regresión *lasso* (least absolute shrinkage and selection operator, por sus siglas en inglés) , introducida por Tibshirani (1996), es una alternativa a la regresión *ridge* cuyo objetivo es precisamente corregir la limitación anteriormente mencionada de la regresión *ridge*, y es útil cuando la mayoría de las variables predictoras no son relevantes en el modelo. Los coeficientes *lasso*, $\hat{\beta}^L$, minimizan la siguiente cantidad:

$$\sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p |\beta_j| = RSS + \lambda \sum_{j=1}^p |\beta_j|.$$

Ahora los coeficientes se *contraen* hacia cero utilizando la suma de los coeficientes en valor absoluto en vez de la suma de los cuadrados de dichos coeficientes. A esta norma se le llama l_1 , $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$. El cambio que supone es sutil pero importante. En ambos casos los coeficientes se contraen hacia 0, pero en el caso de la regresión *lasso* cuando λ es suficientemente grande los coeficientes serán 0, de modo que se está haciendo una selección de variables. Por consiguiente, la regresión *lasso* anulará los coeficientes de las variables que no son importantes a la hora de explicar el comportamiento de la variable respuesta mediante un valor de λ es suficientemente grande. En este sentido el modelo de regresión *lasso* es lo que se llama un **modelo sparse** (un modelo con un número *sparse*, o escaso, de parámetros).

¿Por qué *lasso* hace que los coeficiente se contraigan exactamente hacia cero? Para entenderlo se va a ver una formulación equivalente a la de los mínimos cuadrados penalizados en el caso de la regresión *lasso*:

$$\sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 \text{ sujeto a } \sum_{j=1}^p |\beta_j| < s.$$

19.3. Métodos shrinkage

331

Dicha formulación equivalente corresponde a mínimos cuadrados con una restricción, o lo que es lo mismo, con un *presupuesto* en la norma l_1 sobre los coeficientes. Las dos formulaciones son equivalentes en el sentido de que si se tiene un *presupuesto* s , habrá un λ en la primera formulación que corresponda al presupuesto s en la segunda, y viceversa. Supóngase que se hacen mínimos cuadrados y se obtienen las estimaciones de los parámetros (coeficientes) tal que la suma de sus valores absolutos es 10, pero alguien dice que nuestro *presupuesto* es 5 (la suma de los valores absolutos de los coeficientes no puede ser mayor que esa cantidad). Entonces, hay que resolver el problema de mínimos cuadrados pero los coeficientes no pueden tomar cualquier valor, ya que se tiene una restricción sobre los mismos. Cuanto más pequeño sea el *presupuesto*, más próximos a cero serán los coeficientes. Si el *presupuesto* es 0, todos los coeficientes serán también 0. Si el presupuesto es muy alto, hay libertad para que los coeficientes tomen el valor que quieran, y se estaría en el caso de mínimos cuadrados. El *presupuesto* impone que haya un equilibrio entre el ajuste a los datos y el tamaño de los coeficientes.

La Fig. 19.5 (tomada de James et al. (2013)) muestra por qué el modelo de regresión lasso es *sparse*. El gráfico corresponde a un modelo de regresión con dos variables predictoras. El punto donde está el vector de coeficientes, $\hat{\beta}$, es donde se alcanza el valor mínimo de la suma los cuadrados de los residuos del modelo (RSS) y los contornos son combinaciones de valores de β_1 y β_2 que dan lugar al mismo valor de RSS , pero que ya no sería el mínimo. Las regiones de restricción son $|\beta_1| + |\beta_2| < s$ (*lasso*) y $\beta_1^2 + \beta_2^2 < s$ (*ridge*). En el caso de la regresión *ridge*, el *presupuesto* es el radio del círculo y la regresión *ridge* busca el primer lugar en el que el contorno toca a la región de restricción, pero, al ser un círculo, difícilmente uno u otro coeficiente va a ser 0. En el caso de la regresión *lasso*, la región de restricción tiene forma de diamante y, por lo tanto, tiene vértices. Como puede apreciarse, en la Fig. 19.5 el contorno toca a la región de restricción en el caso en que $\beta_1 = 0$.

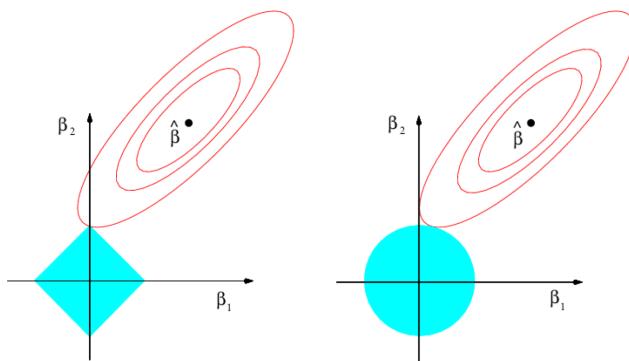


Figura 19.5: Contornos (rojo) de RSS y regiones de restricción (en azul) para la regresión lasso (izquierda) y ridge (derecha)

Se vuelve al ejemplo del béisbol para mostrar la regresión lasso; en este caso el argumento α toma valor 1 (0 en el caso de la regresión *ridge*).

```
lasso_mod <- glmnet(x[entreno, ], y[entreno], alpha = 1, lambda = grid)
plot(lasso_mod)
```

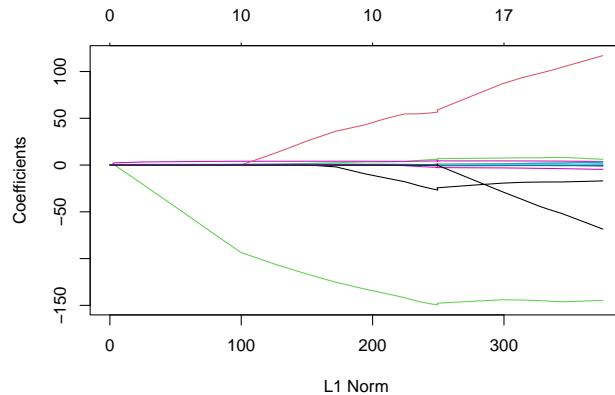


Figura 19.6: Valor de los parámetros estimados para distintos valores de la penalización (que depende del parámetros de penalización)

En la Fig. 19.6 se puede ver que, dependiendo del valor del parámetro de penalización, algunos de los coeficientes se hacen exactamente 0. Para elegir el valor de dicho parámetro y calcular el *MSE* resultante en el conjunto de test se procede como sigue:

```
set.seed(1)
cv_out <- cv.glmnet(x[entreno, ], y[entreno], alpha = 1)
plot(cv_out)
```

```
mejorlab <- cv_out$lambda.min
lasso.pred <- predict(lasso_mod, s = mejorlab, newx = x[test, ])
mean((lasso.pred - y_test)^2)
#> [1] 143673.6
```

Este valor es bastante más bajo que *MSE* en la muestra de test en el caso de mínimos cuadrados ordinarios (224,666,8) y bastante parecido al obtenido con la regresión *ridge* cuando el parámetro de penalización se elige mediante validación cruzada: 139,856,6). Sin embargo, la regresión *lasso* tiene una ventaja importante con respecto a la regresión *ridge* ya que los coeficientes estimados son *sparse*. En los resultados que se muestran a continuación, se puede observar que 10 de los 20 coeficientes estimados son 0. Por lo tanto, el modelo *lasso* con λ elegido mediante validación cruzada contiene sólo nueve variables predictoras.

19.3. Métodos shrinkage

333

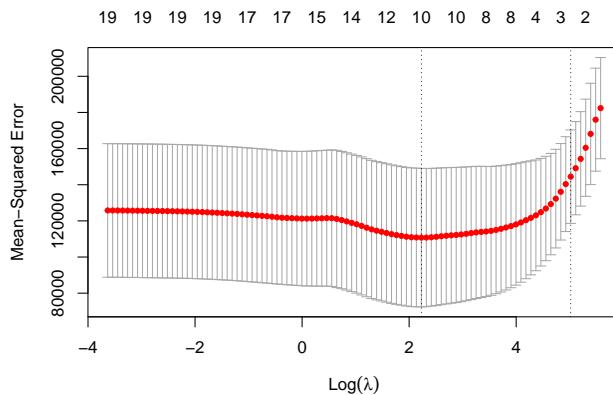


Figura 19.7: Valor del error cuadrático medio y su intervalo de confianza para distintos valores del parámetro de penalización

```
out <- glmnet(x, y, alpha = 1)
lasso_coef <- predict(out, type = "coefficients", s = mejorlab)[1:20, ]
lasso_coef[lasso_coef != 0]
#>   (Intercept)      Hits       Walks      CHmRun      CRuns
#> -3.04787656  2.02551572  2.26853781  0.01647106  0.21177390
#>      CRBI      LeagueN     DivisionW     PutOuts      Errors
#>  0.41944632 20.48456551 -116.59062083  0.23718459 -0.94739923
```

19.3.4. Elastic net

Uno de los problemas de la regresión *lasso* es cuando hay variables predictoras correladas entre sí, pues elegirá una de ellas (y los coeficientes de las demás los hará cero) sin un criterio objetivo. Además, supóngase que se está en una situación en la que el número de variables p es mayor que el número de observaciones n ; en este caso la regresión *lasso* elegiría como mucho n variables; mientras que la regresión *ridge* las utilizaría todas, aumentando la complejidad del modelo (esto en algunos casos puede ser lo deseable o no). *Elastic net* (Zou and Hastie, 2005) es una generalización de los métodos anteriores que combina las penalizaciones de las regresiones *ridge* y *lasso*:

$$\sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda_1 \sum_{j=1}^p \beta_j^2 + \lambda_2 \sum_{j=1}^p |\beta_j|.$$

También aparece en muchas ocasiones de esta otra forma:

$$\sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \left[\frac{1}{2}(1-\alpha) \sum_{j=1}^p \beta_j^2 + \alpha \sum_{j=1}^p |\beta_j| \right],$$

donde $\alpha \in [0, 1]$. El parámetro α es que gobierna la combinación de las dos penalizaciones, mientras que λ es el que controla la cantidad de penalización. Si $\alpha = 0$ se está en el caso de la regresión *ridge*; $\alpha = 1$ lleva a la regresión *lasso*.

La función `glmnet()` también sirve para ajustar *elastic net*, pero el parámetro α hay que elegirlo a priori. Otra opción es utilizar el paquete `caret` para hacer validación cruzada sobre α y λ simultáneamente:

```
set.seed(1)
library("caret")
cv_glmnet <- train(
  x = x[entreno, ],
  y = y[entreno],
  method = "glmnet",
  trControl = trainControl(method = "cv", number = 10),
  tuneLength = 10
)
# modelo con el MSE más pequeño
cv_glmnet$bestTune
#> alpha lambda
#> 9  0.1 99.12337
ggplot(cv_glmnet)
```

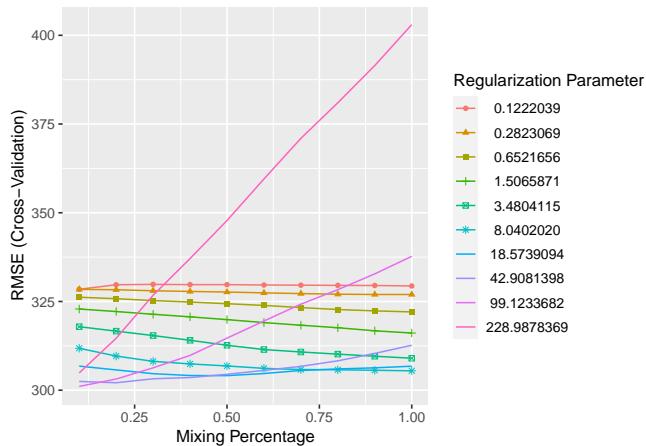


Figura 19.8: Valor de la raíz cuadrada del error cuadrático medio para distintas combinaciones de α y λ

La Fig. 19.8 muestra como la combinación de α y λ da lugar a diferentes MSE (en la figura aparece el $RMSE$, o sea, su raíz cuadrada). Cada línea corresponde a un valor de λ distinto, y en el eje x se representan los valores de α .

A continuación se calcula el valor del MSE en el conjunto de test para el modelo *elastic net* con $\alpha = 0, 1$ y $\lambda = 99, 337$, que son los valores de α y λ que hacen mínimo dicho MSE :

```
elastic_mod <- glmnet(x[entreno, ], y[entreno], alpha = cv_glmnet$bestTune$alpha)
elastic_pred <- predict(elastic_mod, newx = x[test, ], s = cv_glmnet$bestTune$lambda)
mean((elastic_pred - y_test)^2)
#> [1] 141626.1
```

Como puede comprobarse, es peor que el de la regresión *ridge* pero mejor que el de *lasso*.

Por tanto, si no se quiere hacer ningún tipo de selección debe estimar una regresión *ridge* y si se si se quiere reducir al máximo el número de variables predictoras se debe estimar una regresión *lasso* (a costa de que aumente el *MSE* en el conjunto de test). El equilibrio viene de la mano del modelo **elastic-net**, que hace selección de variables pero no aumenta el *MSE* de predicción.

Existen otros métodos de penalización que se derivan de estos, como el *group lasso*, el *sparse group-lasso*, etc. Se puede encontrar información sobre ellos en ([Hastie and Tibshirani, 2015](#)).

Resumen

En este capítulo se introducen una serie de técnicas para mejorar la predicción y la interpretabilidad de los modelos de regresión. En particular:

- Se muestra el uso de la técnica de selección del mejor subconjunto de variables en el modelo, así como los métodos *stepwise*.
- Se presentan 3 métodos tipo *shrinkage*: regresión *ridge*, *lasso* y *elastic net*, bien para la selección de variables, o para solventar problemas de multicolinealidad en el modelo.
- Se muestra cómo seleccionar el parámetro de penalización (o de combinación de penalizaciones en el caso de la regresión *elastic net*) que controla la regresión penalizada.
- Se ilustra el uso de todas las metodologías propuestas en el capítulo mediante el análisis de un caso práctico.

Capítulo 20

Modelización de series temporales

M^a Carmen García Centeno

Universidad San Pablo-CEU, CEU Universities

20.1. Conceptos básicos

El análisis de series temporales es muy útil para analizar el comportamiento de los datos con referencia temporal. Con dicho análisis, se trata de obtener modelos que expliquen su dinámica y puedan utilizarse para predecir valores futuros y tomar decisiones.

Una serie temporal se puede definir como el conjunto de valores observados, en períodos consecutivos de tiempo de la misma amplitud, de una característica de interés en la que su valor en cada instante temporal es una variable aleatoria. Por consiguiente, una serie temporal no es sino la realización de un proceso estocástico; por ejemplo: las precipitaciones diarias, la inflación mensual o el PIB trimestral.

A continuación se exponen algunos de los conceptos clave sobre los que se fundamenta el análisis de series temporales:

1. **Proceso estocástico.** Es una sucesión de variables aleatorias $\{Y_t\}$, donde $t = -\infty, \dots, -2, -1, 0, 1, 2, \dots, \infty$, que dependen de un parámetro. En el caso de las series temporales ese parámetro es el tiempo.¹ Así, en cada momento del tiempo el proceso se particulariza en una variable aleatoria con una determinada distribución de probabilidad. Desde el punto de vista práctico es muy complicado, y a veces imposible, caracterizar un proceso estocástico según su distribución de probabilidad conjunta. Por

¹Pueden depender de más de un parámetro; pero en este capítulo se aborda el caso de un único parámetro: el tiempo

esta razón, se recurre a caracterizarlo de una forma menos completa, pero más sencilla y práctica, basada en los dos primeros momentos de una distribución; en concreto: la media ($E(Y_t) = \mu_t$), la varianza ($Var(Y_t) = E(Y_t - \mu_t)^2 = \sigma_t^2$) y las covarianzas ($\gamma(k) = Cov(Y_t, Y_{t-k}) = E(Y_t - \mu_t)(Y_{t-k} - \mu_{t-k})$).

2. **Estacionariedad.** Un proceso estocástico es estacionario, en sentido simple o amplio, si su media y la varianza se mantienen invariantes a lo largo del tiempo y las covarianzas entre dos variables solo dependen del lapso de tiempo que transcurre entre ellas. Es decir:

- $E(Y_t) = \mu, \quad \forall t,$
- $Var(Y_t) = E(Y_t - \mu)^2 = \sigma^2, \quad \forall t$
- $Cov(Y_t, Y_{t-k}) = E(Y_t - \mu)(Y_{t-k} - \mu) = \gamma(k), \quad \forall t \neq s.$

3. **Función de autocovarianzas.** Está formada por el conjunto de autocovarianzas calculadas para distintos órdenes (k). Tiene como finalidad determinar las relaciones de dependencia lineales que existen entre las variables del proceso con el fin de identificar el modelo que mejor explica su dinámica a lo largo del tiempo.
4. **Función de autocorrelación simple (ACF).** Está formada por el conjunto de coeficientes de correlación lineales calculados para distintos órdenes (k). De forma genérica, para medir la correlación lineal existente entre Y_t e Y_{t-k} , el coeficiente de correlación de orden k , $\rho(k)$, se calcula como el cociente entre la covarianza de orden k y la varianza (ya que, como el proceso es estacionario, las desviaciones típicas de Y_t e Y_{t-k} son iguales).

$$\rho(k) = \frac{Cov(Y_t, Y_{t-k})}{(\sqrt{Var(Y_t)})(\sqrt{Var(Y_{t-k})})} = \frac{\gamma(k)}{\gamma(0)}. \quad (20.1)$$

La representación gráfica de los coeficientes de correlación $\rho(k)$ para $k = 0, 1, 2, \dots$ se conoce como **correlograma**.

5. **Función de autocorrelación parcial (PACF).** Está formada por el conjunto de las correlaciones lineales parciales obtenidas para los distintos valores de k . Así, para dos instantes de una serie temporal t y $(t - k)$, mide la correlación lineal existente entre las variables Y_t e Y_{t-k} asociadas a ellos, ajustada de los valores que toma el proceso temporal en los períodos intermedios ($Y_{t-1}, Y_{t-2}, \dots, Y_{t-(k-1)}$).²
6. **Ergodicidad.** Un proceso estocástico es ergódico cuando, a partir de un determinado desfase temporal entre las variables, la correlación lineal existente entre ellas tiende a desaparecer. Esto implica que las covarianzas y el coeficiente de correlación tienden a cero, es decir,

²En el coeficiente de autocorrelación parcial de orden k se calcula la correlación entre parejas de valores separados por esa distancia temporal pero eliminando el efecto debido a la correlación producida por retardos anteriores a k .

$$\lim_{k \rightarrow \infty} \gamma(k) = 0 \quad y \quad \lim_{k \rightarrow \infty} \rho(k) = 0$$

7. **Ruido blanco.** Es un proceso puramente aleatorio que se puede expresar de la siguiente forma:

$$Y_t = a_t.$$

Se caracteriza por tener esperanza nula, varianza constante y covarianzas nulas. Es decir,

$$E(a_t) = 0 \quad \forall t; \quad E(a_t^2) = \sigma^2 \quad \forall t; \quad E(a_t a_s) = 0 \quad \forall t \neq s.$$

Esto implica que un ruido blanco siempre es estacionario.

20.2. Modelos ARIMA

Para determinar las características del proceso estocástico subyacente a la serie temporal, se van a utilizar los modelos ARIMA (acrónimo del inglés *AutoRegressive Integrated Moving Average*). Estos modelos están formados por tres componentes: **AR** (autorregresivo), **I** (integrado, es decir, número de diferencias necesarias para convertirlo en estacionario cuando no lo es) y **MA** (medias móviles). Fueron propuestos por Box y Jenkins y son un caso particular de procesos estocásticos lineales, estacionarios, ergódicos y discretos. Entre este tipo de procesos lineales, los más frecuentes son:

- Los procesos puramente aleatorios (por ejemplo, un proceso ruido blanco).
- Los procesos puros de medias móviles regulares (MA) o medias estacionales (sMA).
- Los procesos puros autorregresivos (AR) o autorregresivos estacionales (sAR).
- Una combinación de los procesos anteriores que da lugar a los procesos mixtos regulares (ARMA) o mixtos estacionales (sARMA).

Los modelos ARIMA tratan de captar la dinámica y la dependencia existente en los datos de una serie temporal. Por lo tanto, son muy útiles para describir un valor como una combinación o función lineal de valores pasados y errores debidos al azar. Para detalles sobre la cuestión pueden verse, entre muchos otros, [Hamilton \(1994\)](#), [Uriel Jiménez and Peiro Giménez \(2000\)](#), [Cryer and Chan \(2010\)](#), [Pemberton \(2011\)](#), [Mínguez Salido and García Centeno \(2011\)](#), [Brockwell and Davis \(2016\)](#) y [Shumway and Stoffer \(2017\)](#).

En la modelización ARIMA se pueden destacar varias fases. La primera se centra en la **identificación** del modelo ARIMA que haya podido generar los datos de la serie temporal. En esta fase, es necesario decidir los órdenes del proceso (es decir, el desfase temporal entre el mayor y menor periodo de tiempo de las variables incluidas en el modelo) tanto de la parte regular como de la estacional.

La ACF y la PACF desempeñan un papel clave en la identificación de estos órdenes. Antes de calcular la ACF y la PACF, es necesario comprobar que la serie es estacionaria, es decir, que no deambula ni tiene tendencia creciente o decreciente. En el caso de que no lo sea, se realizarán las transformaciones necesarias para convertirla en estacionaria, ya que la modelización ARIMA exige que los datos sean estacionarios. Las dos razones fundamentales por las cuales pueden no ser estacionarios son: la no constancia en el tiempo de la media o la existencia de fluctuaciones de diferente amplitud que hacen que la varianza no se mantenga constante.

Si la serie no es estacionaria en varianza, de entre las transformaciones Box-Cox, se puede utilizar una transformación logarítmica porque reduce el rango dinámico de la variable, corrige la asimetría positiva y acerca a la normalidad la distribución de la variable; además, facilita la interpretación de las estimaciones de los coeficientes del modelo, ya que la diferencia del logaritmo de la serie es, aproximadamente, su tasa de variación porcentual.

Si no es estacionaria en media, puede ser debido a la existencia de tendencia en el tiempo o de estacionalidad. Para eliminar la tendencia se calculan las diferencias regulares (una o dos como máximo según el tipo de tendencia lineal o no lineal, es decir, $Y_t - Y_{t-1} = \Delta Y_t$ o $\Delta^2 Y_t = \Delta(\Delta Y_t)$, respectivamente). Para eliminar la estacionalidad, lo que se calcula es una diferencia estacional (es decir, $Y_t - Y_{t-s} = \Delta_s Y_t$).

Después de proponer el modelo, y tras dividir el conjunto de datos en dos subconjuntos (el de entrenamiento y el de test), se procede, con los datos del subconjunto apropiado, a la **estimación** del modelo, a su **validación** y a la **realización de predicciones**. Primero se obtienen las estimaciones de los parámetros del modelo propuesto, así como sus desviaciones típicas, y los residuos del modelo. Posteriormente, se aborda la fase de **validación** o **diagnóstico** y se realizan los contrastes necesarios para determinar si el modelo propuesto es adecuado o no, es decir si los parámetros estimados son estadísticamente significativos o no; si el modelo estimado es estacionario e invertible; y si sus residuos siguen un proceso ruido blanco o no. En el caso de que no lo sean, será necesario corregir el modelo con la información proporcionada por los residuos hasta obtener un modelo cuyos residuos sean ruido blanco. Finalmente, se procede a la utilización del modelo para **predecir**.

20.3. Análisis de series temporales con R

Algunas de las librerías que normalmente se suelen utilizar en la modelización de series temporales con **R** son:

```
library("tseries")
library("astsa")
library("forecast")
library("lubridate")
library("foreign")
library("quantmod")
library("ggplot2")
```

La librería **tseries** permite manipular datos de series temporales; **astsa** es adecuada para analizar series de tiempo en los dominios de frecuencia y tiempo; **forecast** lleva a cabo (y analiza)

20.3. Análisis de series temporales con **R**

341

predicciones con series temporales; **lubridate** facilita el trabajo con fechas y horas; **foreign** es esencial para crear objetos en **R** importando datos de casi cualquier formato conocido, como por ejemplo SAS, SPSS, Stata, etc.; **quantmod** ayuda, de forma cuantitativa, en el desarrollo de estrategias y modelización mediante la utilización de estadísticas; **readxl** facilita la exportación de datos de Excel a **R**; y “**ggplot2**” es muy útil para la realización de gráficos.

En el caso real que se propone, se utilizan datos mensuales del INE (www.ine.es) correspondientes al índice general de precios al consumo (IPC) en el periodo muestral comprendido entre enero de 2002 y marzo de 2022. Los datos están en el fichero **ipc** del paquete CDR del libro.

```
ipc <- CDR::ipc
```

Para trabajar con series temporales, la fecha es un requisito imprescindible. En caso de los datos no tuviesen fecha, habría que incluirla. Por ejemplo, para series mensuales, el código sería:

```
ipc_ts <- ts(ipc$ipc, start = c(2002, 1), end = c(2022, 3), frequency = 12)
```

En el caso de fichero **ipc** la fecha figura en la primera columna del fichero.

La representación gráfica de la serie original puede ayudar a saber si la serie ha sido generada por un proceso estacionario o no. Para obtener dicha representación, se ejecuta el siguiente código:

```
ipc$Time <- as.Date(ipc$Time)
ggplot(
  data = ipc,
  aes(x = Time, y = ipc)
) +
  theme(
    axis.title.x = element_text(size = 15),
    axis.title.y = element_text(size = 15)
) +
  geom_line(colour = "blue")
```

La descomposición aditiva o multiplicativa en (i) tendencia, (ii) componente estacional, (iii) componente cíclico y (iv) componente irregular de la serie del IPC, así como su representación gráfica, también puede ayudar a determinar si la serie es estacionaria o no. El código para la descomposición aditiva de la serie es el siguiente:

```
componentes_ts <- decompose(ipc_ts)
autoplot(componentes_ts)
```

Tanto en el gráfico de la serie original del IPC (Fig. 20.1) como en el de su descomposición en componentes (Fig. 20.2), se aprecia que la serie tiene tendencia y componente estacional (mensual). Por lo tanto, no es estacionaria en media, es decir, la media es distinta para diferentes

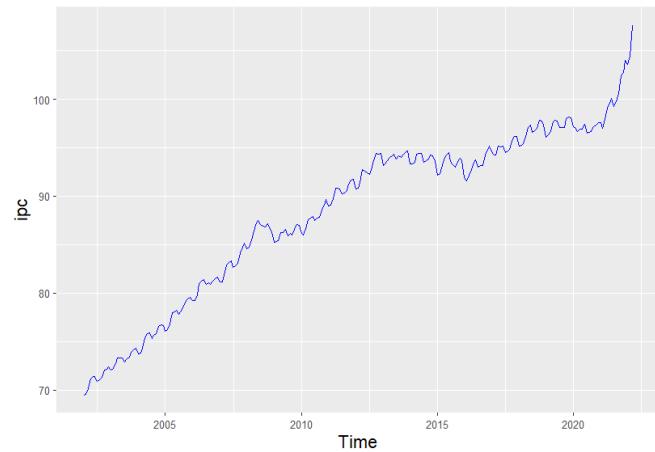


Figura 20.1: Evolución del IPC entre enero 2002 y marzo 2022.

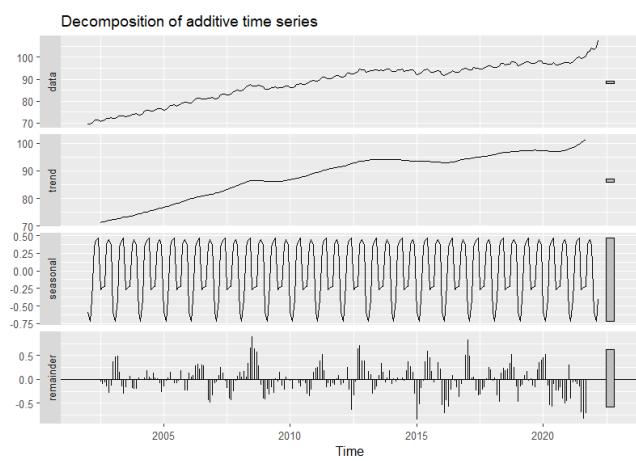


Figura 20.2: Descomposición aditiva del IPC

20.3. Análisis de series temporales con **R**

343

meses y años; tampoco lo es en varianza, ya que la dispersión respecto de la media cambia a lo largo del tiempo.

Si una serie no es estacionaria ni en media ni en varianza es necesario empezar corrigiendo la no estacionariedad en varianza y, posteriormente, la no estacionariedad en media.

Las transformaciones de Box-Cox son las más utilizadas para corregir el problema de no estacionariedad en varianza. De todas estas transformaciones, la más habitual es el logaritmo. Para obtener la representación gráfica del logaritmo de Y_t , se ejecuta el siguiente código:

```
logipc <- log10(ipc_ts)
ts_logipc <- data.frame(value = logipc, Time = time(logipc))
ggplot(
  data = ts_logipc,
  aes(x = Time, y = logipc)
) +
  theme(
    axis.title.x = element_text(size = 15),
    axis.title.y = element_text(size = 15)
  ) +
  geom_line(colour = "red")
```

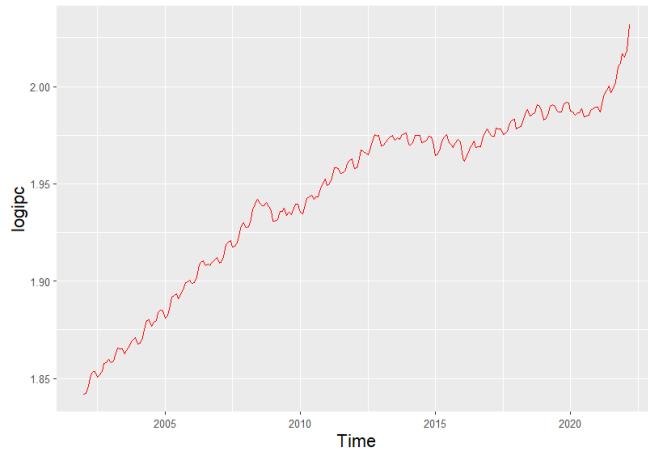


Figura 20.3: Logaritmo del IPC

Corregido el problema de la no estacionariedad en varianza, en la Fig. 20.3 se aprecia que sigue sin ser estacionaria, ya que al tener estacionalidad y tendencia creciente no es estacionaria en media. Para conseguir que la serie sea estacionaria es necesario corregir tanto su tendencia como su estacionalidad.

Es indiferente cuál de los dos problemas se resuelve primero, pues el resultado final de la transformación es el mismo. En este caso, se empieza corrigiendo la tendencia. Para ello, se calcula una diferencia regular del logaritmo del IPC, obteniéndose la serie ($dlogipct = \log(ipct) -$

$\log(ipc_{t-1})$. Al ser datos mensuales, esta transformación representa la tasa de variación relativa mensual del IPC. El código para calcular esta diferencia regular y su representación gráfica (Fig. 20.4) es:

```
dlogipc <- diff(logipc, differences = 1)
ts_dlogipc <- data.frame(value = dlogipc, Time = time(dlogipc))
ggplot(
  data = ts_dlogipc,
  aes(x = Time, y = dlogipc)
) +
  theme(
    axis.title.x = element_text(size = 15),
    axis.title.y = element_text(size = 15)
  ) +
  geom_line(colour = "blue")
```

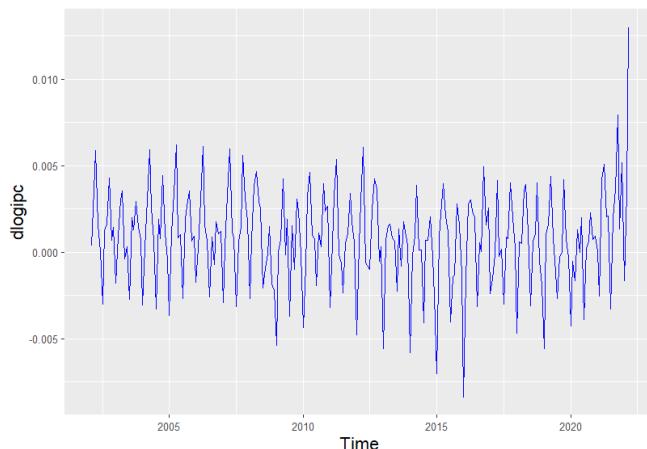


Figura 20.4: Diferencia regular del logaritmo del IPC

Corregida la tendencia, a continuación se corrige su estacionalidad. Para ello, sobre la diferencia regular del logaritmo del IPC, se calculará una diferencia estacional mensual ($d12dlogipc_t = \log(ipc_t) - \log(ipc_{t-12})$). El código para calcular esta diferencia estacional y su representación gráfica (Fig. 20.5) es:

```
d12dlogipc <- diff(dlogipc, 12)
ts_d12dlogipc <- data.frame(value = d12dlogipc, Time = time(d12dlogipc))

ggplot(
  data = ts_d12dlogipc,
  aes(x = Time, y = d12dlogipc)
) +
  theme(
```

20.3. Análisis de series temporales con **R**

345

```

axis.title.x = element_text(size = 15),
axis.title.y = element_text(size = 15)
) +
geom_line(colour = "purple")

```

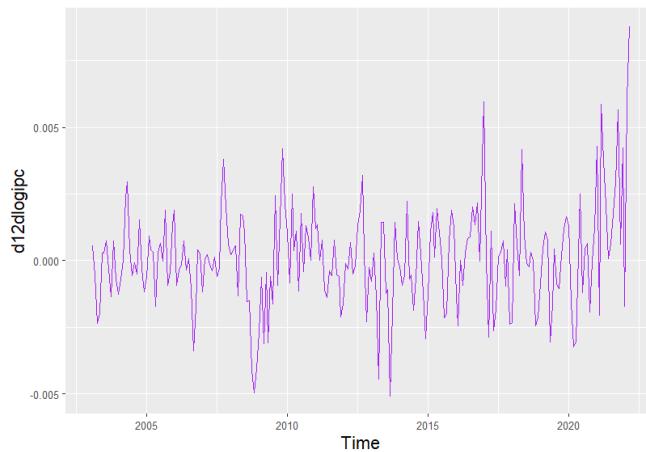


Figura 20.5: Diferencia estacional de la diferencia regular del logaritmo del IPC

Para contrastar si una serie es estacionaria en media o no, también se puede utilizar un test de raíces unitarias[^][Una raíz unitaria es una característica de los procesos que evolucionan a lo largo del tiempo que implica problemas de no estacionariedad en media y también al realizar inferencias. A partir de la ecuación incial de un AR:

$$\Delta Y_t = \alpha + (\phi - 1)Y_{t-1} + a_t,$$

las hipótesis del contraste son:

$$H_0 : (\phi - 1) = 0$$

$$H_1 : (\phi - 1) < 0.$$

En esta ecuación se pueden añadir más retardos de la variable, una tendencia (en series que claramente tengan una tendencia decreciente o creciente) o eliminar la constante (en series con media nula).] En este caso se utiliza el test de raíces unitarias de Dickey-Fuller, que es un test unilateral en el que la hipótesis nula es la existencia de raíces unitarias y, por lo tanto, que la serie no es estacionaria (es decir, es I(1)), mientras que la hipótesis alternativa es que sí es estacionaria (es decir, la serie es I(0)). Si se realiza el test sobre la serie original (`ipc_ts`) y sobre la serie transformada del IPC (`d12dLogIPC`), se comprueba que esta última es la transformación estacionaria. El código utilizado para realizar este contraste es:

```
adf.test(ipc_ts)
#>
#> Augmented Dickey-Fuller Test
#>
#> data: ipc_ts
#> Dickey-Fuller = -1.8396, Lag order = 6, p-value = 0.6433
#> alternative hypothesis: stationary
adf.test(d12dlogipc)
#>
#> Augmented Dickey-Fuller Test
#>
#> data: d12dlogipc
#> Dickey-Fuller = -3.3727, Lag order = 6, p-value = 0.06002
#> alternative hypothesis: stationary
```

A la luz del p-valor para los valores originales del IPC, si se preestablece un nivel de significación del 10 %, no se rechaza la hipótesis nula (y, por tanto, la serie no es estacionaria). Sin embargo, para la serie resultante tras las transformaciones anteriores (`d12dLogipc`) se rechaza la H_0 . Por lo tanto, se puede concluir que dicha transformación la ha convertido en estacionaria.

20.3.1. Identificación o especificación del modelo

A partir de la transformación estacionaria del modelo es necesario calcular la ACF y la PACF muestrales, con el fin de identificar el modelo ARIMA más adecuado, es decir, especificar tanto el modelo como su correspondiente orden. Antes de hacerlo para el caso concreto del IPC, se analizan brevemente los principales modelos teóricos que pueden explicar la dinámica de una serie temporal. Estos modelos, para la parte regular (es decir, de la modelización de la dependencia asociada a observaciones consecutivas de tiempo) pueden ser: autorregresivos puros ($\text{AR}(p)$), medias móviles ($\text{MA}(q)$) o mixtos $\text{ARMA}(p,q)$, donde p y q representan sus correspondientes órdenes, respectivamente. En el caso de la parte estacional (s), es decir, de la modelización de la dependencia asociada a observaciones que distan entre sí s -periodos de tiempo o múltiplos de s , se tiene: $\text{sAR}(P)$, $\text{sMA}(Q)$ o $\text{sARMA}(P,Q)$, donde P y Q representan, respectivamente, los órdenes de los modelos correspondientes a la parte estacional. Destacar que para identificar el modelo de la parte estacional es necesario analizar los coeficientes correspondientes a los retardos estacionales. Así, por ejemplo, en el caso de una serie mensual, habría que fijarse en los coeficientes correspondientes a los retardos 12, 24, 36, etc., para una serie trimestral en los retardos, 4, 8, 12, 16, etc., y así sucesivamente.

Antes de identificar el modelo correspondiente al IPC en el periodo muestral analizado, se muestran algunos ejemplos de modelos ARIMA, su ecuación general y el comportamiento de su ACF y PACF.

- En el caso de $\text{MA}(q)$ o $\text{sMA}(Q)$:
 - La ACF se anula para órdenes superiores a q (o a Q en el caso estacional). Es decir, solo los q primeros coeficientes son significativos (o Q en el caso estacional). El resto de los coeficientes son nulos (o estadísticamente nulos en el caso muestral).

20.3. Análisis de series temporales con **R**

347

- La PACF decrece de forma exponencial o sinusoidal hacia cero.

Ejemplo de las ACF y PACF teóricas de los procesos estacionarios MA(1).

La ecuación que describe la dinámica de un modelo MA(1) o ARMA(0,1) es:

$$Y_t = a_t - \theta a_{t-1},$$

o bien, en forma polinómica:

$$Y_t = (1 - \theta L)a_t,$$

donde L es el operador de retardos y a_t es un ruido blanco, es decir: $E(a_t) = 0 \quad \forall t$; $E(a_t^2) = \sigma^2 \quad \forall t$; $E(a_t a_s) = 0 \quad \forall t \neq s$.

Este modelo siempre es estacionario (ya que se obtiene como una combinación de procesos ruido blanco) y para que sea invertible (es decir, para que se pueda expresar en función del pasado de la variable), es necesario que las raíces del polinomio estén fuera del círculo unidad, o lo que es lo mismo, que $|\theta| < 1$.

Su ACF teórica es:

$$\rho(k) = \begin{cases} \frac{-\theta}{(1+\theta^2)} & \text{si } k = 1 \\ 0 & \forall k > 1, \end{cases}$$

y su PACF (también teórica) viene dada por:

$$\phi_{kk} = \frac{-\theta^k(1-\theta^2)}{1-\theta^{2(k+1)}} \quad \text{para } k \geq 1.$$

- En un AR(p) o sAR(P):

- La ACF decrece rápidamente de forma exponencial o sinusoidal hacia cero.
- La PACF se anula para órdenes superiores a p (o a P , en el caso estacional). Es decir, sólo los p (o P en el caso estacional) primeros coeficientes son significativamente distintos de cero. El resto de los coeficientes son nulos (o estadísticamente nulos en el caso muestral).

Ejemplo de las ACF y PACF teóricas de los procesos estacionarios AR(1).

La ecuación que describe la dinámica de un modelo AR(1) o ARMA(1,0) es:

$$Y_t = \phi Y_{t-1} + a_t,$$

o bien, en forma polinómica: $Y_t(1 - \phi L) = a_t$, donde L es el operador de retardos y a_t es un ruido blanco.

Este modelo siempre es invertible (ya que está expresado en función del pasado de la variable) y para que sea estacionario es necesario que las raíces del polinomio de retardos estén fuera del círculo unidad, o lo que es lo mismo, que $|\phi| < 1$.

Su ACF teórica es:

$$\rho(k) = \phi\rho(k-1) = \phi^k,$$

mientras que su PACF (teórica) viene dada por:

$$\phi_{kk} = \begin{cases} \rho_1 = \phi & \text{si } k = 1 \\ 0 & \forall k > 1. \end{cases}$$

■ En un ARMA(p,q) o sARMA(P,Q):

- La ACF tiene un comportamiento irregular en los q primeros (o en los Q en el caso estacional) coeficientes. A partir del orden q (o Q en el caso estacional), se comporta como la de un AR(p) (o un sAR(P) en el caso estacional).
- La PACF tiene un comportamiento irregular en los p primeros coeficientes (o P en el caso estacional). A partir del orden p (o P en el caso estacional), se comporta como la de un MA(q) (o de un sMA(Q) en el caso estacional).

Ejemplo de las ACF y PACF teóricas de los procesos estacionarios ARMA(1,1).

La ecuación que describe la dinámica de un modelo ARMA(1,1) es:

$Y_t = \phi Y_{t-1} + a_t - \theta a_{t-1}$, o bien, en forma polinómica, $Y_t(1 - \phi L) = (1 - \theta L)a_t$, donde L es el operador de retardos y a_t es un ruido blanco.

Para que el modelo sea estacionario, es necesario que las raíces del polinomio de la parte autorregresiva estén fuera del círculo unidad (es decir, que $|\phi| < 1$).

Para que sea invertible es necesario que las raíces del polinomio de las medias móviles estén fuera del círculo unidad (o que $|\theta| < 1$).

Además, es necesario que no existan raíces comunes, es decir, $\phi \neq \theta$.

Su ACF teórica es:

$$\rho(k) = \begin{cases} \frac{(1-\phi\theta)(\phi-\theta)}{1+\theta^2-2\phi\theta} & \text{si } k = 1 \\ \phi\rho(k-1) & \forall k > 1. \end{cases}$$

Su PACF teórica se obtiene como:

20.3. Análisis de series temporales con **R**

349

$$\phi_{11} = \rho_1$$

$$\phi_{22} = \frac{1 \quad \rho_1}{\begin{vmatrix} \rho_1 & \rho_2 \\ 1 & \rho_1 \end{vmatrix}} = \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2}$$

$$\phi_{33} = \frac{1 \quad \rho_1 \quad \rho_1}{\begin{vmatrix} \rho_1 & 1 & \rho_2 \\ \rho_2 & \rho_1 & \rho_3 \\ 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_1 \\ \rho_2 & \rho_1 & 1 \end{vmatrix}} = \frac{\rho_1^3 - \rho_1 \rho_2 (2 - \rho_2) + \rho_3 (1 - \rho_1^2)}{1 - \rho_2^2 - 2\rho_1^2(1 - \rho_2)}$$

$$\vdots$$
Nota:

La forma de obtener los diferentes coeficientes de la ACF y PACF de los modelos ARIMA estacionales es la misma que para los modelos ARIMA utilizados para la parte regular.

La diferencia fundamental reside en que para la parte regular se calculan los coeficientes de correlación entre las variables que se encuentran en periodos consecutivos de tiempo, mientras que para la parte estacional se hace entre las variables que están separadas entre si s periodos o múltiplos de s (donde, por ejemplo, $s=12$, si la serie es mensual; $s=4$ si la serie fuese trimestral, etc.). Por esta razón, sólo se muestra la forma de calcular la ACF y la PACF correspondiente a la parte regular.

Teniendo en cuenta lo anterior, en el caso concreto del IPC, después de obtener su transformación estacional, la ACF y la PACF muestrales son muy útiles para identificar el modelo ARIMA más adecuado, así como su correspondiente orden en el periodo muestral estudiado.

A continuación se muestra el código **R** para ejecutar los comandos que permiten calcular y representar la ACF y la PACF muestrales de la transformación estacional del IPC (Fig. 20.6 y Fig. 20.7, respectivamente). El número de retardos (lags) se establece en 40 para, así, incluir al menos 3 retardos estacionales (12, 24 y 36):

```
acf(ts(d12dlogipc, frequency = 1), lag.max = 40, main = "")
```

```
knitr:::include_graphics("img/acf1.png")
```

```
pacf(ts(d12dlogipc, frequency = 1), lag.max = 40, main = "")
```

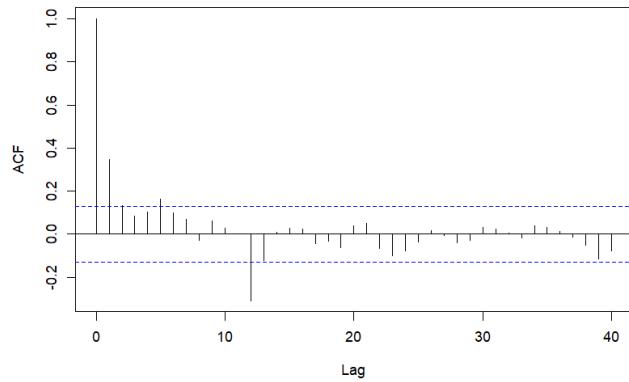


Figura 20.6: ACF. Función de autocorrelación muestral de d12dLogIPC

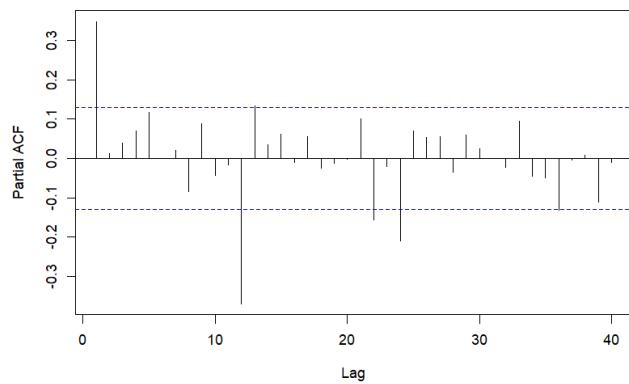


Figura 20.7: PACF. Función de autocorrelación parcial muestral de d12dLogIPC

Observando el comportamiento de estas funciones, para la parte regular se podría proponer un AR(1) y para la parte estacional un MA(1)12. Por lo tanto, ya que se ha calculado una diferencia regular y otra estacional para convertir a la serie original en estacionaria, el modelo ARIMA para el IPC, en el periodo muestral analizado, sería un ARIMA(1,1,0)(0,1,1)12 o sARIMA(1,1,0)(0,1,1). Señalar que **R** permite la identificación automática de los órdenes del modelo.

20.3.2. Estimación del modelo

Los parámetros del modelo se estiman por el método de máxima verosimilitud. Dado que la función de máxima verosimilitud no es lineal, para maximizarla se lleva a cabo un procedimiento iterativo de estimación no lineal. El código en **R** que permite obtener las estimaciones de los parámetros del modelo y sus correspondientes desviaciones típicas es:

```
modelo <- arima(ipc_ts, c(1, 1, 0), c(0, 1, 1))
modelo
#>
#> Call:
#> arima(x = ipc_ts, order = c(1, 1, 0), seasonal = c(0, 1, 1))
#>
#> Coefficients:
#> ar1 sma1
#> 0.4665 -0.8302
#> s.e. 0.0701 0.0860
#>
#> sigma^2 estimated as 0.1082: log likelihood = -77.76, aic = 161.51
```

20.3.3. Validación

En la metodología ARIMA desarrollada por Box y Jenkins es necesario determinar si el modelo propuesto es correcto o no, es decir, si se ajusta o no correctamente a los datos de la muestra. Para ello, en la fase de validación, fundamentalmente, es necesario comprobar que:

1. Los parámetros del modelo estimado cumplen con las condiciones de estacionariedad e invertibilidad.

El modelo estimado para el IPC es invertible y estacionario, ya que (*i*) la parte regular, al modelizarse mediante un AR(1), siempre va a ser invertible; además, dado que la estimación de ϕ es, en valor absoluto, menor que uno, es decir, $|0,4665| < 1$, también es estacionario; (*ii*) la parte estacional sigue un esquema de medias móviles de primer orden (sMA(1)), que siempre es estacionario; como, adicionalmente, el valor absoluto de la estimación de θ es menor que uno ($|-0,8302| < 1$), también es invertible.

Si alguna de las raíces del polinomio de retardos de la parte autorregresiva estuviese próxima a uno, entonces es posible que la serie original este subdiferenciada y, por lo tanto, no sería

estacionaria, lo que implicaría la necesidad de calcular alguna diferencia adicional. Si fuese alguna de las raíces del polinomio de retardos de las medias móviles la que estuviese próxima a uno, entonces el modelo podría estar sobreparametrizado y habría que eliminar alguna diferencia. Además, si existiesen raíces comunes en ambos polinomios de retardos de la parte regular, sería necesario eliminar diferencias tanto en la parte autorregresiva como en la parte de medias móviles y el modelo correcto sería un ARIMA($p-1, d, q-1$). Los comentarios anteriores son válidos para la parte estacional.

2. Los parámetros estimados son estadísticamente significativos.

Para comprobar si los parámetros son estadísticamente significativos o no, (o en otros términos, para comprobar si se ha sobreparametrizado o no el modelo), se realiza un contraste de significatividad individual para cada uno de ellos. Por ejemplo, para el parámetro del AR(1) la hipótesis nula y alternativa serían:

$$\begin{aligned} H_0 &: \phi = 0 \\ H_1 &: \phi \neq 0. \end{aligned}$$

Para realizar el contraste se utiliza el estadístico t , que sigue una distribución t-Student con $(n-k)$ grados de libertad. Este estadístico t se calcula dividiendo el valor estimado del parámetro entre su correspondiente error estándar. En concreto, para los dos parámetros estimados de este modelo (el correspondiente al AR(1) de la parte regular y al sMA de la parte estacional) se tiene:

$$t = \frac{0,4665}{0,0701} = 6,654 \quad y \quad |t| = \left| \frac{-0,8302}{0,0860} \right| = 9,653.$$

Como, para un nivel de significación del 5 % y 248 grados de libertad, el valor crítico de la t-Student es aproximadamente 1,96, se rechaza H_0 , lo que implica que las estimaciones de los dos parámetros del modelo son estadísticamente distintas de cero.

3. Los residuos del modelo son ruido blanco.

Como se avanzó al final de la Sec. 20.1, uno de los supuestos de la modelización ARIMA es que el error del modelo (o perturbación aleatoria) tiene que ser ruido blanco. Como los errores no son observables, las comprobaciones se llevan a cabo sobre los residuos (diferencia entre los valores reales y los predichos por el modelo). Existen varias formas de comprobar si los residuos son ruido blanco o no. Entre ellas, las más utilizadas son: el gráfico de la serie original de residuos, la ACF y la PACF estimadas y el contraste de Portmanteau (planteado inicialmente por Box-Pierce y actualizado posteriormente por Ljung-Box).

En primer lugar, de forma intuitiva, el gráfico de los residuos puede mostrar si la media es constante e igual a cero y si su varianza también es constante. Además, puede reflejar si existen valores atípicos u outliers (se consideran como tales aquellos que superen tres veces su desviación típica). El siguiente código R permite obtener los residuos y su representación gráfica (Fig. 20.8).

20.3. Análisis de series temporales con **R**

353

```

residuos <- residuals(modelo)
ts_residuos <- data.frame(value = residuos, Time = time(residuos))
ggplot(
  data = ts_residuos,
  aes(
    x = Time,
    y = residuos
  )
) +
  geom_line(colour = "red")

```

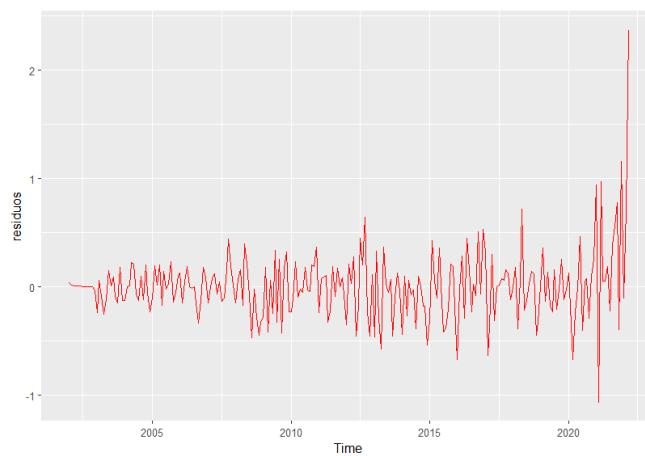


Figura 20.8: Gráfico de los residuos

En la Fig. 20.8 se observa que la media es constante e igual a cero y que a partir de febrero de 2022, como consecuencia de las tensiones inflacionarias, los residuos son mayores que en el resto del periodo muestral.

En segundo lugar, se pueden calcular la ACF y la PACF muestrales de los residuos para comprobar que están incorrelacionados, es decir, que los coeficientes de correlación calculados son estadísticamente nulos. Si éstos fuesen estadísticamente significativos, los residuos no serían ruido blanco. Entonces, para la correcta modelización habría que identificar el proceso e incorporarlo al modelo inicial propuesto para volver a estimarlo. Los correlogramas obtenidos (Fig. 20.9) con el código que se muestra a continuación, indican que no hay coeficientes de correlación estadísticamente significativos (ya que se encuentran dentro de los intervalos de confianza) y, por lo tanto, los residuos son ruido blanco.³

³A veces, pueden observarse 1 o 2 correlaciones significativas en desfases de mayor orden no estacionales. Ello se debe, generalmente, a un error aleatorio y no constituye una señal de incumplimiento del supuesto, por lo que se puede concluir que los residuos son independientes. Tal es el caso en la Fig. @(acfpacfresiduos), concretamente en la PACF.

```
par(mfrow = c(2, 1), mar = c(1, 1, 1, 1) + 0.1)
acf(ts(residuos, frequency = 1), lag.max = 40)
pacf(ts(residuos, frequency = 1), lag.max = 40)
```

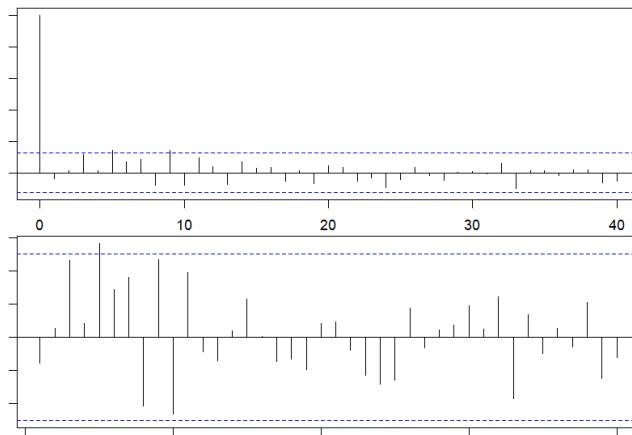


Figura 20.9: ACF y PACF estimadas de los residuos

También se puede utilizar el contraste de Portmanteau (Ljung-Box) para comprobar si los residuos están incorrelacionados y se comportan como un ruido blanco o no. En este contraste global de significación, la hipótesis nula que se plantea, en el caso del IPC, es que los primeros 40 coeficientes de correlación son cero frente a la hipótesis alternativa de que no lo son. Para un nivel de significación del 5 %, la evidencia empírica no es suficiente para rechazar H_0 , ya que el p-valor correspondiente al estadístico del contraste (X-squared o Chi-cuadrado) es 0.5448, mayor que el nivel de significación del 5 % y, por lo tanto, se concluye que los residuos están incorrelacionados. El código R para llevar a cabo el test de Box-Ljung es:

```
Box.test(residuos, type = "Ljung-Box")
#>
#> Box-Ljung test
#>
#> data: residuos
#> X-squared = 0.36682, df = 1, p-value = 0.5447
```

4. Análisis de la bondad del ajuste.

Finalmente, se debe comprobar la capacidad de ajuste del modelo comparando los valores observados y los estimados. Una forma sencilla e intuitiva de hacerlo es a través de su representación gráfica. El código R utilizado para ello es el siguiente:

20.3. Análisis de series temporales con **R**

355

```
plot(ipc_ts)
lines(ipc_ts - modelo$residuals, col = "red")
```

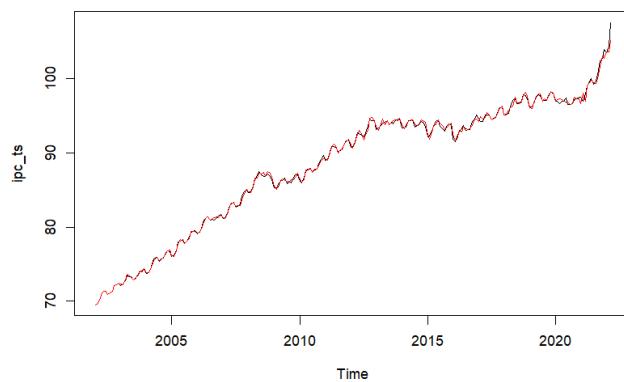


Figura 20.10: Ajuste con el modelo estimado.

La Fig. 20.10 muestra que el modelo estimado se ajusta bastante bien a los valores observados y, por lo tanto, evidencia que el modelo sARIMA $(1,1,0)(0,1,1)$ propuesto capta la dinámica del IPC en el periodo muestral analizado.

Para comprobar la adecuación entre el modelo estimado y los valores observados no suele ser adecuado utilizar el coeficiente de determinación o el coeficiente de determinación corregido, ya que si hay que calcular diferencias para convertir la serie en estacionaria, la variable dependiente cambia (es decir, no es lo mismo Y_t , que $dlog(Y_t)$, que $d12dlog(Y_t)$). Si se quieren comparar diferentes modelos para elegir cuál es el que mejor se ajusta a los datos, hay que utilizar otros métodos de información, como el criterio de información de Akaike (AIC), el criterio bayesiano o de Schwarz (BIC) o el de Hannan-Quinn (HQC). Si, por ejemplo, se eligiese el AIC será mejor el modelo con menor AIC.

20.3.4. Predicción

Después de comprobar que el modelo ARIMA estimado es adecuado, se puede utilizar para obtener valores futuros de la variable objeto de análisis. Las predicciones que se obtienen pueden ser de dos tipos: puntuales o por intervalos. Las predicciones puntuales para un horizonte temporal h se obtienen calculando el valor esperado de la variable dependiente en $T+h$ (\hat{Y}_{T+h}) condicionado al conjunto de información disponible hasta el momento actual (T). Para obtener las predicciones por intervalos es necesario sumar y restar a la predicción puntual la desviación típica del error de predicción multiplicada por el valor crítico tabulado para el nivel de confianza fijado.

Algunas características generales de las predicciones obtenidas con modelos ARIMA son:

- En modelos MA($*q*$) o sMA($*Q*$): si el horizonte temporal de predicción es mayor que el orden del proceso, entonces la predicción es la media del proceso.
- En los modelos AR($*p*$) o sAR($*P*$), a medida que aumenta el horizonte temporal, la predicción tiende a la media del proceso.
- En los modelos ARMA($*p*, *q*$) o sARMA($*P*, *Q*$), para ordenes superiores al medias móviles, la función de predicción se comporta como la de un autorregresivo, lo que implica que tiende a la media del proceso.
- Es importante que la serie sea estacionaria para que las predicciones sean estables, ya que si no es estacionaria el modelo tendrá un comportamiento explosivo.
- Cuanto más alejado esté el horizonte temporal de predicción mayor será la incertidumbre respecto de las predicciones obtenidas.

En el caso concreto del IPC, para obtener 12 predicciones puntuales y sus correspondientes intervalos de predicción al 80 % y 95 % de confianza, así como su representación gráfica (Fig. 20.11), el código R que hay utilizar es:

```
forecast::forecast(modelo, h = 12)
#>   Point Forecast    Lo 80     Hi 80    Lo 95     Hi 95
#> Apr 2022 109.7116 109.2900 110.1332 109.0668 110.3564
#> May 2022 110.5926 109.8443 111.3410 109.4481 111.7371
#> Jun 2022 111.1030 110.0714 112.1346 109.5253 112.6806
#> Jul 2022 110.5518 109.2748 111.8289 108.5987 112.5050
#> Aug 2022 110.7714 109.2787 112.2641 108.4885 113.0543
#> Sep 2022 111.0288 109.3435 112.7140 108.4515 113.6060
#> Oct 2022 111.9631 110.1034 113.8228 109.1189 114.8073
#> Nov 2022 112.1740 110.1540 114.1939 109.0847 115.2632
#> Dec 2022 112.3896 110.2209 114.5583 109.0728 115.7064
#> Jan 2023 111.6049 109.2968 113.9130 108.0750 115.1349
#> Feb 2023 111.6666 109.2270 114.1061 107.9355 115.3976
#> Mar 2023 112.5012 109.9369 115.0656 108.5794 116.4231

predicciones <- forecast::forecast(modelo, h = 12)
autoplot(predicciones)
```

20.3. Análisis de series temporales con **R**

357

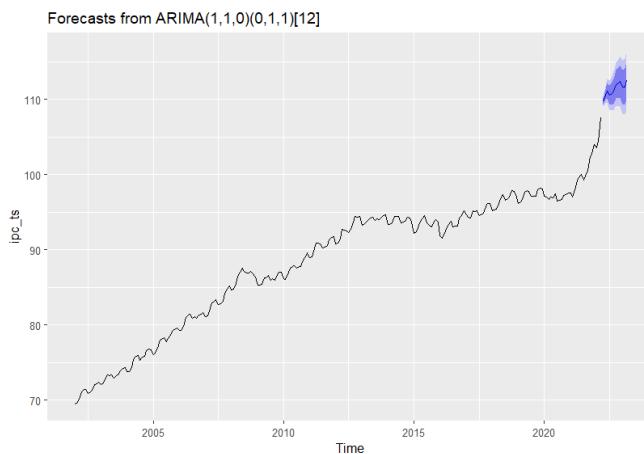


Figura 20.11: Predicciones con el modelo ARIMA(1,1,0)(0,1,1)12 estimado

Resumen

Los modelos univariantes ARIMA analizan las series de tiempo desde un punto de vista moderno y estocástico y son muy útiles para captar su dinámica. Es importante destacar que:

- Hay que tener en cuenta conceptos básicos que son fundamentales para llevar a cabo una correcta modelización ARIMA. Entre ellos, se pueden destacar los siguientes: proceso estocástico, estacionalidad, estacionariedad, invertibilidad y ruido blanco.
- Las principales fases en el análisis de un modelo ARIMA son: especificación, estimación, validación y predicción.
- Para identificar el modelo ARIMA que mejor capta la dinámica de la serie temporal objeto de análisis, es necesario que los datos sean estacionarios en media y en varianza.
- La ACF y la PACF muestrales son muy útiles para identificar el modelo más adecuado para explicar la dinámica de los datos.
- Para comprobar que el modelo estimado es adecuado es necesario realizar un análisis de los parámetros estimados, un análisis de los residuos y medir la bondad del ajuste.
- Validado el modelo, se suele utilizar, fundamentalmente, para predecir y tomar decisiones.

Capítulo 21

Análisis discriminante

M^a Leticia Meseguer Santamaría^a y Manuel Vargas Vargas^a

^aUniversidad de Castilla-La Mancha

21.1. Introducción

El **análisis discriminante (AD)** es una técnica de dependencia orientada a la clasificación de individuos en grupos (o poblaciones) preexistentes y con ciertas características conocidas, utilizando para ello la información proporcionada por un conjunto de variables clasificadoras.¹ La clasificación se realiza mediante **funciones discriminantes**, combinaciones de las clasificadoras originales y que se emplean como criterio para asignar cada individuo a un grupo o población.

De esta forma, en el análisis discriminante se identifican dos **finalidades**: la descriptiva, caracterizando la separación entre grupos al proporcionar la contribución de cada variable clasificadora a dicha separación; y la predictiva, estableciendo el criterio de clasificación de un individuo nuevo a alguno de los grupos conociendo los valores de las variables clasificadoras.

El problema de la *discriminación* puede plantearse de diversas formas y aparece en numerosas áreas de investigación. El **AD** se agrupa dentro de los modelos denominados **supervisados**, puesto que se conoce a priori a qué grupo o población está asignado cada individuo de la muestra, y es utilizado en campos tan diferentes como los sistemas automáticos de concesión de créditos bancarios (*credit scoring*), clasificación de pacientes en función de pruebas diagnósticas, atribución de obras literarias o pictóricas a autores, o en control de calidad, cuando la información es muy costosa o requiera la destrucción de las unidades. En el campo de la ingeniería, la discriminación se conoce como *reconocimiento de patrones (pattern recognition)* y es utilizada para el diseño de máquinas de clasificación automática (reconocimiento de billetes o monedas, sonidos, etc.)

¹La referencia a individuos es en sentido amplio, entendiéndose por individuos no solo personas, sino también objetos, entes, elementos, casos, etc.

Aunque existen varios enfoque diferentes, en este capítulo se adoptará el enfoque clásico de Fisher, que asume la normalidad multivariante de las variables clasificadoras. Como punto de partida, para un **AD** se considera:

- Un conjunto de **N** individuos, de los que se conoce su grupo de pertenencia. Esta información se resume en una variable categórica **Y** cuyas categorías son los distintos grupos.
- Un conjunto de **k** grupos ($k \geq 2$) con, al menos, dos individuos en cada uno de ellos.
- Un conjunto de **p** variables clasificadoras, medidas en intervalo o razón. Estas variables no deben presentar multicolinealidad, es decir, ninguna clasificadora puede ser combinación lineal de otras clasificadoras. Además, dado el enfoque adoptado, se asume que estas variables siguen una distribución normal multivariante.²

El número de variables discriminantes debe ser inferior en más de dos al número de individuos ($p < N - 2$) para poder identificar los parámetros (véase Cap. 15). Además, en la práctica, es útil disponer de algún criterio o método que permita seleccionar qué variables se considerarán clasificadoras. Una alternativa pueden ser los métodos de jerarquización de variables desarrollados en análisis de regresión o la selección de variables (*feature selection*, véase Cap. 9). Como punto inicial, es frecuente que se considere que una variable puede ser clasificadora si presenta diferencias en su distribución entre los grupos, utilizando para ello un **ANOVA**.

Así, el **AD** busca determinar un criterio o *regla discriminante* que clasifique a cada individuo, j , en uno de los k grupos conociendo las observaciones de cada una de las p variables X_i , es decir, el vector $X_j = (X_{1j}, X_{2j}, \dots, X_{pj})'$. Estas reglas discriminantes están basadas en la información muestral y en los supuestos que sobre ésta se hacen; en el planteamiento clásico de Fisher, al asumir la normalidad de las variables, se basan en el comportamiento en los k grupos de los vectores de medias y de las matrices de varianzas-covarianzas (véase Sec. 11.3). Por ello, se suelen distinguir varios casos, que conducen a distintos métodos de obtención de reglas discriminantes, por lo que reciben nombres diferentes:

- El caso más sencillo (e históricamente el más antiguo), además de la normalidad, supone que las matrices de varianzas-covarianzas son iguales en todos los grupos (supuesto de *homocedasticidad*). El método se conoce como **análisis discriminante lineal** (*linear discriminant analysis* o *LDA*). En este caso, detallado en la Sec. 21.2, la diferencia en la distribución de las variables entre los grupos se produce en los vectores de medias, y la función discriminante obtenida es una combinación lineal de las variables clasificadoras que minimiza los errores de clasificación.
- Otra posibilidad es que se asuma la normalidad pero no la igualdad de las matrices de varianzas-covarianzas entre los grupos. En este caso, la función discriminante es una función cuadrática, por lo que el método se conoce como **análisis discriminante cuadrático** (*quadratic discriminant analysis* o *QDA*), detallado en la Sec. 21.3.

²Este supuesto garantiza que el método clásico propuesto por Fisher es óptimo. En la práctica, el AD es robusto frente a incumplimientos de la normalidad p-dimensional, por lo que también se aplica en muchos casos prácticos donde no se puede garantizar este requisito.

Sea cual sea el método elegido, las *reglas discriminantes* que se obtengan para clasificar a un individuo en uno de los grupos deben determinarse minimizando los errores de clasificación, que pueden ser evaluados probabilísticamente al disponer de la distribución de las variables en cada grupo. Así, para cada individuo j y sus valores de las variables clasificadoras $X_j = (x_{1j}, x_{2j}, \dots, x_{pj})'$, se dispone de las verosimilitudes para cada uno de los k grupos, $L_i(X_j; \theta_i)$, $1 \leq i \leq k$.

Conociendo la probabilidad *a priori* de pertenencia de un individuo a cada grupo³ π_i , $1 \leq i \leq k$, aplicando el teorema de Bayes (véase (12.5)), se puede calcular la probabilidad de que el individuo pertenezca a cada grupo, G_i

$$P(G_i/x_j) = \frac{L_i(X_j; \theta_i)\pi_i}{\sum_m L_m(X_j; \theta_m)\pi_m} \quad (21.1)$$

A partir de esta ecuación, la *regla discriminante* consiste en asignar al individuo al grupo más probable. Dado que el denominador de (21.1) es constante para todos los grupos, la regla equivale a asignar al individuo al grupo donde sea *ponderadamente* más verosímil:

$$j \text{ se clasifica en } G_i \text{ si } L_i(X_j; \theta_i)\pi_i = \max_m L_m(X_j; \theta_m)\pi_m \quad (21.2)$$

ecuación que se simplifica en el caso de igual probabilidad *a priori*, resultando la *regla discriminante* en asignar a cada individuo al grupo más verosímil.

En general, se pueden cometer dos tipos de error: no clasificar al individuo en un grupo cuando realmente pertenece a él; o clasificarlo en un grupo al que realmente no pertenece. Si no se conocen dichos costes (o son iguales), no afectan a la *regla discriminante*; sin embargo, si son conocidos y han de ser tenidos en cuenta, la regla se modificaría, ponderando cada verosimilitud por los costes asociados.

En las secciones siguientes se abordarán ambos modelos de **AD** que, aunque no son los únicos, sí representan la gran mayoría de las aplicaciones prácticas.

21.2. Análisis discriminante lineal

Es un modelo de **AD** basado en los supuestos generales expuestos en el epígrafe anterior (N individuos, k grupos y p variables clasificadoras con distribución normal y sin multicolinealidad) y caracterizado por la **igualdad de las matrices de varianza-covarianza** de las variables en todos los grupos. Para la exposición de la metodología, se presentará el caso más sencillo, con sólo dos grupos y probabilidades *a priori* iguales, para generalizarlo posteriormente al caso general de k grupos.

Dos grupos y una variable clasificadora.

³Suele ser frecuente asignar idéntica probabilidad *a priori* a todos los grupos $\pi_i = \frac{1}{k}$ o proporcional al tamaño de cada grupo.

Es el supuesto más simple posible, donde se han de clasificar N individuos en con dos grupos (I y II) a partir de la información de una única variable clasificadora, X . En este caso, las distribuciones de probabilidad de X en los grupos I y II solo difieren en la media, como se muestra en la Fig. 21.1.

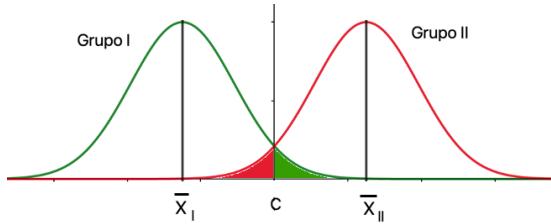


Figura 21.1: Dos grupos y una variable clasificadora.

La *regla discriminante* consistirá en asignar cada individuo al grupo donde su verosimilitud es mayor. Como se aprecia, esta regla divide la recta real en dos partes, cuyo **punto de corte (C)** es:

$$C = \frac{\bar{x}_I + \bar{x}_{II}}{2} \quad (21.3)$$

quedando la asignación de cada individuo:⁴

$$\text{si } x_j < C \in \text{ Grupo I y si } x_j > C \in \text{ Grupo II} \quad (21.4)$$

Las probabilidades de los errores que se pueden cometer en la asignación corresponderían a las áreas resaltadas en rojo (individuo asignado al grupo I cuando realmente pertenece al grupo II) y en verde (individuo asignado al grupo II cuando realmente pertenece al grupo I), constituyendo la zona de error de clasificación.

Dos grupos y dos variables clasificadoras.

Si, bajo los mismos supuestos, se dispone de dos variables clasificadoras, X_1 y X_2 , se proyectan los elipsoides de ambos grupos sobre las dos variables, se obtendría la Fig. 21.2:

Se obtienen, sobre cada variable, zonas de error de clasificación amplias (marcadas en amarillo) que conllevarán errores de clasificación grandes. Sin embargo, si se proyectan ambos elipsoides sobre un nuevo *eje*, obtenido como una combinación lineal de ambas variables clasificadoras ($w_1X_1 + w_2X_2 - D = 0$), es posible reducir la zona de error de clasificación y, como consecuencia, la probabilidad de error de clasificación.

La obtención de la combinación lineal que minimiza la probabilidad de error de clasificación fue resuelto por Fisher buscando una **función discriminante** que maximiza la separación entre

⁴De forma intuitiva, se asigna cada individuo al grupo cuya media está más cercana al valor de la variable. Esta interpretación se generaliza a más variables clasificadoras, asignando cada individuo al grupo cuyo centroide esté más cercano a él. Si la probabilidad *a priori* fuese proporcional al tamaño de los grupos, el punto de corte se calcularía como $C = \frac{n_I\bar{x}_I + n_{II}\bar{x}_{II}}{N}$.

21.2. Análisis discriminante lineal

363

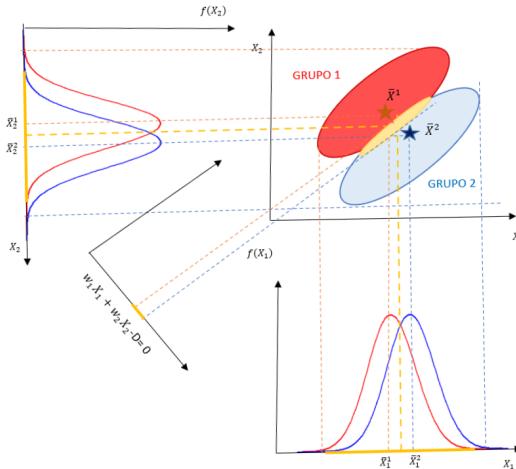


Figura 21.2: Dos grupos y dos variables clasificadoras.

ambos grupos, maximizando la distancia entre sus centroides y minimizando la variabilidad dentro de cada grupo. El procedimiento se detalla para el caso general de p variables.

Dos grupos y p variables clasificadoras.

El objetivo es encontrar una *regla discriminante* que permita *separar* ambos grupos. Se busca la **función discriminante de Fisher**, que se plantea como una combinación lineal de las p variables clasificadoras:

$$D = w_1X_1 + w_2X_2 + \dots + w_pX_p \quad (21.5)$$

que asignaría al individuo j -ésimo una **puntuación discriminante** $D_j = w_1X_{1j} + w_2X_{2j} + \dots + w_pX_{pj}$; expresando matricialmente estas puntuaciones en diferencias respecto a las medias:

$$\begin{pmatrix} D_1 - \bar{D} \\ D_2 - \bar{D} \\ \vdots \\ D_N - \bar{D} \end{pmatrix} = \begin{pmatrix} X_{11} - \bar{X}_1 & X_{21} - \bar{X}_2 & \dots & X_{p1} - \bar{X}_p \\ X_{12} - \bar{X}_1 & X_{22} - \bar{X}_2 & \dots & X_{p2} - \bar{X}_p \\ \vdots & \vdots & \ddots & \vdots \\ X_{1N} - \bar{X}_1 & X_{2N} - \bar{X}_2 & \dots & X_{pN} - \bar{X}_p \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_p \end{pmatrix} \quad (21.6)$$

donde $\bar{D} = w_1\bar{X}_1 + w_2\bar{X}_2 + \dots + w_p\bar{X}_p$ por ser D una combinación lineal de variables normales. En notación abreviada, la ecuación (21.6) se puede expresar como $d = Xw$.

La suma de cuadrados de las desviaciones de la función discriminante respecto a su media quedaría entonces como:

$$\mathbf{d}' \mathbf{d} = \mathbf{w}' \mathbf{X}' \mathbf{X} \mathbf{w} \quad (21.7)$$

donde $\mathbf{X}'\mathbf{X}$ es la matriz simétrica de las desviaciones cuadráticas respecto a sus medias de las variables clasificadoras (o matriz **suma de cuadrados y productos cruzados**, SCPC). Esta matriz se puede descomponer en la suma de dos matrices, la SCPC entregrupos, \mathbf{F} y la SCPC residual o intragrupos, \mathbf{U} , por lo que la ecuación (21.7) se puede reexpresar como:

$$\mathbf{d}' \mathbf{d} = \mathbf{w}' \mathbf{Fw} + \mathbf{w}' \mathbf{Uw} \quad (21.8)$$

Fisher propuso determinar los pesos $\{w_i\}$ buscando que discriminen entre los grupos, maximizando la variabilidad entre grupos respecto a la intragrupos, es decir:

$$\max_{\mathbf{w}} \frac{\mathbf{w}' \mathbf{Fw}}{\mathbf{w}' \mathbf{Uw}} \quad (21.9)$$

Como esta función es invariante frente a cambios de escala, maximizar (21.9) es equivalente a maximizar $\mathbf{w}' \mathbf{Fw}$ con la condición $\mathbf{w}' \mathbf{Uw} = 1$ que, aplicando los multiplicadores de Lagrange, implica:

$$\begin{aligned} L &= \mathbf{w}' \mathbf{Fw} - \lambda(\mathbf{w}' \mathbf{Uw} - 1) \Rightarrow \frac{\partial L}{\partial w} = 2\mathbf{Fw} - 2\lambda\mathbf{Uw} = 0 \Rightarrow \\ &\Rightarrow \mathbf{Fw} = \lambda\mathbf{Uw} \Rightarrow (\mathbf{U}^{-1}\mathbf{F})\mathbf{w} = \lambda\mathbf{w} \end{aligned} \quad (21.10)$$

Así, el vector propio asociado al mayor autovalor de la matriz $\mathbf{U}^{-1}\mathbf{F}$ proporcionará la **función discriminante lineal de Fisher** que mejor separa ambos grupos.

El **punto de corte (C)** se obtiene evaluando la función discriminante en la media de cada grupo y promediando por el tamaño de los grupos:

$$\begin{aligned} \bar{D}_I &= w_1 \bar{X}_{1I} + w_2 \bar{X}_{2I} + \dots + w_p \bar{X}_{pI} \\ \bar{D}_{II} &= w_1 \bar{X}_{1II} + w_2 \bar{X}_{2II} + \dots + w_p \bar{X}_{pII} \end{aligned} \quad (21.11)$$

$$C = \frac{n_I \bar{D}_I + n_{II} \bar{D}_{II}}{N} \quad (21.12)$$

Y el **criterio de asignación para cada individuo, j** , será:

$$\text{si } D_j < C \in \text{ Grupo I y si } D_j > C \in \text{ Grupo II} \quad (21.13)$$

G grupos y p variables:

En caso de existir más de dos grupos, la generalización del caso anterior es relativamente sencilla. Siguiendo la misma idea utilizada para dos grupos, se debería obtener un número de **funciones discriminantes de Fisher** suficiente para separar los k grupos; este número es $T = \min(k - 1, p)$.⁵

⁵Para separar linealmente k grupos hacen falta $k - 1$ hiperplanos, pero su obtención está también limitada por el número p de variables clasificadoras.

21.2. Análisis discriminante lineal

365

Así, cada una de las T funciones discriminantes será una combinación lineal de las p variables clasificadoras:

$$D_t = w_{t1}X_1 + w_{t2}X_2 + \dots + w_{tp}X_p \text{ para } t = 1, \dots, T \quad (21.14)$$

donde se exige que $\text{Corr}(D_i, D_j) = 0, \forall i \neq j$.

La suma de cuadrados de las desviaciones de la matriz \mathbf{D} de funciones discriminantes respecto a sus medias tendría una expresión equivalente a la ecuación (21.7):

$$\mathbf{D}'\mathbf{D} = \mathbf{W}'\mathbf{X}'\mathbf{X}\mathbf{W} \quad (21.15)$$

Para que las funciones discriminen lo máximo posible a los k grupos, las combinaciones lineales han de maximizar la variabilidad entre los grupos respecto a la intragrupos, en un razonamiento análogo la expuesto en la ecuación (21.9):

$$\max \frac{\mathbf{W}'\mathbf{F}\mathbf{W}}{\mathbf{W}'\mathbf{U}\mathbf{W}} \quad (21.16)$$

Al ser una función homogénea, su maximización equivale a maximizar $\mathbf{W}'\mathbf{F}\mathbf{W}$ con la condición $\mathbf{W}'\mathbf{U}\mathbf{W} = 1$ que, aplicando los multiplicadores de Lagrange, implica:

$$\begin{aligned} L &= \mathbf{W}'\mathbf{F}\mathbf{W} - \lambda(\mathbf{W}'\mathbf{U}\mathbf{W} - 1) \Rightarrow \frac{\partial L}{\partial \mathbf{w}} = 2\mathbf{F}\mathbf{W} - 2\lambda\mathbf{U}\mathbf{W} = 0 \Rightarrow \\ &\Rightarrow \mathbf{F}\mathbf{W} = \lambda\mathbf{U}\mathbf{W} \Rightarrow (\mathbf{U}^{-1}\mathbf{F})\mathbf{W} = \lambda\mathbf{W} \end{aligned} \quad (21.17)$$

Así, el vector propio asociado al mayor autovalor de la matriz $\mathbf{U}^{-1}\mathbf{F}$ (generalmente no simétrica) proporcionará la **primera función discriminante lineal de Fisher**, siendo el autovalor la proporción de varianza total explicada por las T funciones discriminantes que recoge la primera función.

Para obtener el resto de funciones discriminantes, basta con ir eligiendo los siguientes vectores propios asociados a los autovalores, ordenados decrecientemente. Como los vectores propios son linealmente independientes, las funciones de discriminación son incorreladas.⁶

De esta forma, la primera función discriminante, D_1 , será la que proporcione mayor discriminación entre los centroides de los grupos; D_2 será la que proporcione mayor discriminación, después de D_1 , y que esté incorrelada con ella; y así sucesivamente, D_t será la que produzca mayor discriminación entre los centroides de los grupos, después de las $t - 1$ anteriores, e incorrelada con todas las anteriores.

⁶Como la capacidad discriminante de la funciones va decreciendo, puede haber casos donde no se consideren relevantes todas, sino el conjunto de las m primeras. En ese caso, la variabilidad explicada sería $\sum_{i=1}^m \lambda_i$, por lo que la proporción de variabilidad atribuible a cada función discriminante D_t sería $\frac{\lambda_t}{\sum_{i=1}^m \lambda_i}$.

21.2.1. Discriminante lineal con R: la función lda()

A continuación, se va a ejemplificar la aplicación de un **discriminante lineal** con **R**. Para ello, se utilizará y cargará la base de datos **iris**, que consta de 150 observaciones y 5 variables, 4 numéricas, que serán las clasificadoras, y una categórica, sobre la que se realiza el análisis, con tres categorías: setosa, versicolor y virginica.

```
library("caret")
library("MASS")
library("klaR")
data("iris")
```

Se clasificarán las flores iris, identificadas con la variable **Species** (especies de iris), utilizando como variables clasificadoras: **Sepal.Length** (Longitud del sépalo), **Sepal.Width** (anchura del sépalo), **Petal.Length** (longitud del pétalo) y **Petal.Width** (anchura del pétalo).

Para evaluar la capacidad predictiva del análisis discriminante, se dividen los datos de la muestra en un 80 % para la estimación (o entrenamiento) y un 20 % para el test.⁷

Las distribuciones univariadas deben ser normales; si no fuera así, se podrían transformar utilizando log y root (distribuciones exponenciales) y Box-Cox (distribuciones sesgadas), como se muestra en la Sec. ???. Igualmente, es conveniente estandarizar las variables para evitar que la diferencia de escalas influya en la importancia relativa de cada variable clasificadora en las funciones discriminantes.

```
# División de los datos: 80% para entrenamiento y 20% para test
set.seed(123)
muestra <- iris$Species |>
  createDataPartition(p = 0.8, list = FALSE)
entrenamiento_d <- iris[muestra, ]
test_d <- iris[-muestra, ]
# Estimación de los parámetros de preprocessamiento (estandarización)
preproc_param <- entrenamiento_d |>
  preProcess(method = c("center", "scale"))
# Transformación de los datos usando los parámetros estimados
entrenamiento_t <- preproc_param |> predict(entrenamiento_d)
test_t <- preproc_param |> predict(test_d)
```

Una inspección previa de los datos puede ayudar a detercar si las variables clasificadoras pueden contribuir a la discriminación entre los grupos. En este ejemplo, la Fig. 21.3 muestra la densidad de cada variable sobre cada grupo para el conjunto de entrenamiento:

⁷ Esta estrategia es muy común en modelos predictivos, y tiene como objetivo evitar el **sobreajuste** a los datos muestrales; así, los datos del conjunto de test son realmente “*nuevos*” para el modelo, porque no han sido utilizados en la estimación.

21.2. Análisis discriminante lineal

367

```

library("ggplot2")
library("ggsignif")

p1 <- ggplot(data = entrenamiento_t, aes(x = Sepal.Length, fill = Species, colour =
  Species)) +
  geom_density(alpha = 0.3) +
  theme_bw()
p2 <- ggplot(data = entrenamiento_t, aes(x = Sepal.Width, fill = Species, colour =
  Species)) +
  geom_density(alpha = 0.3) +
  theme_bw()
p3 <- ggplot(data = entrenamiento_t, aes(x = Petal.Length, fill = Species, colour =
  Species)) +
  geom_density(alpha = 0.3) +
  theme_bw()
p4 <- ggplot(data = entrenamiento_t, aes(x = Petal.Width, fill = Species, colour =
  Species)) +
  geom_density(alpha = 0.3) +
  theme_bw()
ggarrange(p1, p2, p3, p4, ncol = 2, nrow = 2, common.legend = TRUE, legend = "bottom")

```

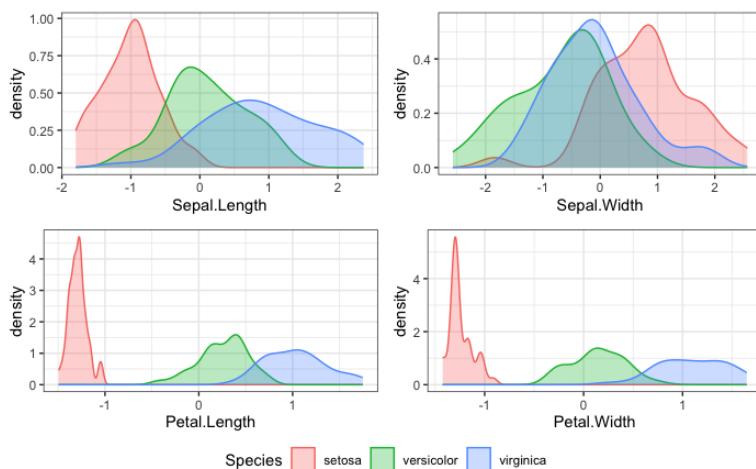


Figura 21.3: Densidad de cada variable clasificadora sobre los grupos.

Igualmente, los gráficos bivariantes pueden ser ayudar a ver si hay *distancias* entre los centroides de los grupos para las variables clasificadoras, como muestra la Fig. 21.4:

```

pairs(x = entrenamiento_t[, -5], col = c("firebrick", "green3",
  "darkblue")[entrenamiento_t$Species], pch = 20)

```

Como se observa en estos gráficos, las variables clasificadoras pueden contribuir a la discriminación entre las tres especies de flores *iris*.

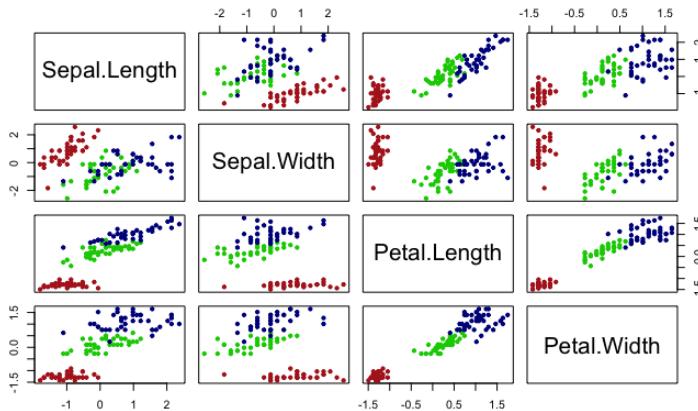


Figura 21.4: Diagramas bivariantes de dispersión de las variables clasificadoras.

Para aplicar la función `lda()` debemos especificar la variable de agrupación (`Species`) y el conjunto de datos (`entrenamiento_t`); de forma opcional, se pueden especificar las probabilidades *a priori* (`prior`, por defecto se usa `proportions`), el método de estimación de las medias y varianzas (`method`, por defecto `moment`) o el argumento `CV` para obtener los grupos pronosticados y las probabilidades a posteriori (por defecto, `CV=FALSE`)

```
options(digits = 4)
modelo_lda <- lda(Species ~ ., data = entrenamiento_t)
modelo_lda
## Call:
## lda(Species ~ ., data = entrenamiento_t)
##
## Prior probabilities of groups:
##   setosa versicolor virginica
## 0.3333 0.3333 0.3333
##
## Group means:
##           Sepal.Length Sepal.Width Petal.Length Petal.Width
## setosa      -1.0113     0.78049    -1.2900    -1.2453
## versicolor     0.1014    -0.68675     0.2566     0.1473
## virginica     0.9099    -0.09374     1.0334     1.0981
##
## Coefficients of linear discriminants:
##                 LD1         LD2
## Sepal.Length  0.6795  0.04464
## Sepal.Width   0.6565 -1.00330
## Petal.Length -3.8365  1.44176
## Petal.Width   -2.2722 -1.96516
##
```

21.2. Análisis discriminante lineal

369

```
#> Proportion of trace:
#>   LD1     LD2
#> 0.9902 0.0098
```

La salida muestra las **probabilidades previas** (*Prior probabilities of groups*) y los **centroides de cada grupo** (*Group means*). A continuación muestra las **funciones discriminantes de Fisher** mediante los respectivos coeficientes w_{it} . En este caso, las dos funciones discriminantes son:

$$D_1 = 0,6795 * SL + 0,6565 * SW - 3,8365 * PL - 2,2722 * PW$$

$$D_2 = 0,0446 * SL - 1,0033 * SW + 1,4418 * PL - 1,9651 * PW$$

con una proporción de discriminación de 0.9902 y 0.0098, respectivamente.

La proyección de los individuos en el plano formado por las dos funciones discriminantes se recoge en la Fig. 21.5:

```
datos_lda <- cbind(entrenamiento_t, predict(modelo_lda)$x)
ggplot(datos_lda, aes(LD1, LD2)) +
  geom_point(aes(color = Species)) +
  ggtitle("Gráfico LDA")
```

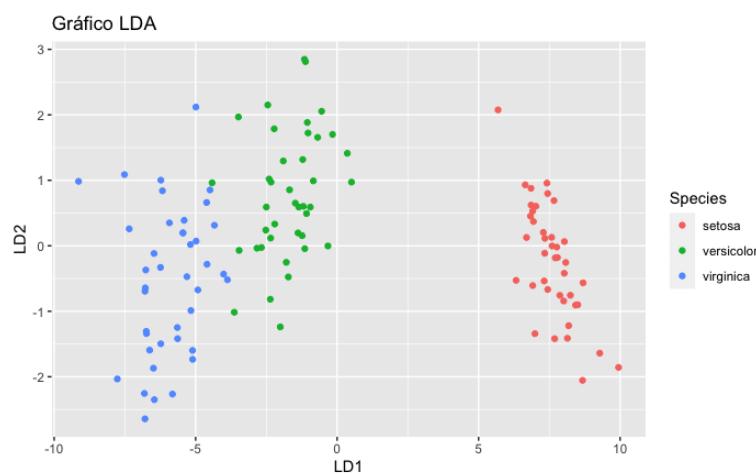


Figura 21.5: Proyección sobre las dos funciones discriminantes de los individuos.

Como se aprecia, la primera función discriminante es la que mayor contribución tiene a la separación entre los grupos, separando muy claramente a la especie *setosa* y, en menor medida, a las especies *virginica* y *versicolor*, grupos entre los que hay un pequeño grado de solapamiento. Por otro lado, la segunda función discriminante, con una proporción de 0.0098 apenas contribuye a la separación entre grupos.

Por último, es posible visualizar cómo quedarían las regiones bivariantes que clasifican a los individuos en cada clase mediante la función `partimat()` del paquete `klaR`, como muestra la Fig. 21.6:

```
partimat(Species ~ ., data = entrenamiento_t, method = "lda", image.colors =
  c("skyblue", "lightgrey", "yellow"), col.mean = "red")
```

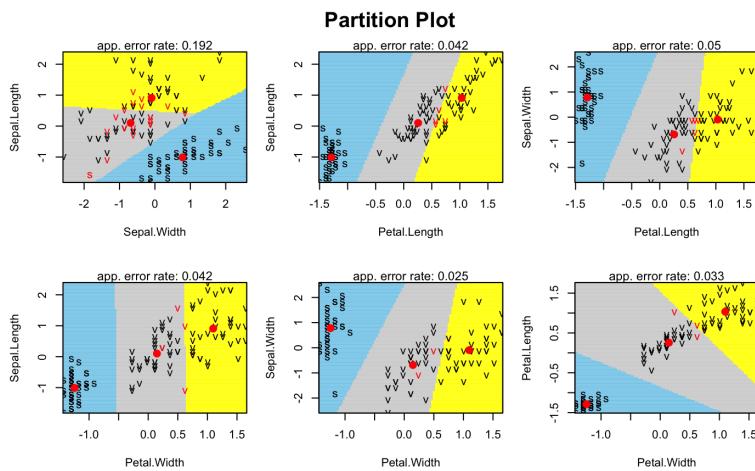


Figura 21.6: Regiones bivariantes de clasificación en cada grupo: setosa (celeste), versicolor (gris) y virginica (amarillo). Centroides en rojo.

Por último, aplicando las funciones discriminantes a los datos reservados para estudiar la capacidad predictiva del modelo:

```
predicciones_lda <- modelo_lda |> predict(test_t)
table(test_t$Species, predicciones_lda$class, dnn = c("Grupo real", "Grupo
  ~ pronosticado"))
#>           Grupo pronosticado
#> Grupo real   setosa versicolor virginica
#>   setosa       10      0      0
#>   versicolor    0     10      0
#>   virginica     0      1     9
mean(predicciones_lda$class == test_t$Species)
#> [1] 0.9667
```

se obtiene la tabla conocida como **matriz de confusión**, donde se compara el grupo real con el pronosticado por el modelo. En este caso, se clasifican correctamente 29 de las 30 “nuevas” flores, indicando un grado de ajuste del 96.9667 %.

21.3. Análisis discriminante cuadrático

En el discriminante lineal visto anteriormente, se asume que las variables clasificadoras tienen idénticas matrices de varianza-covarianza en los distintos grupos, supuesto que garantiza que las funciones discriminantes son combinaciones lineales de las variables.

Es posible eliminar esta restricción, permitiendo diferencias en las matrices de varianzas-covarianzas en los grupos, lo que introduce términos cuadráticos en las funciones discriminantes.

Así, denominando π_t a la probabilidad *a priori* de pertenecer al grupo G_t , μ_t y Σ_t al vector de medias y matriz de varianzas-covarianzas respectivamente de dicho grupo, a partir del vector de observaciones X se puede obtener la **función discriminante cuadrática** como:

$$Q(\mathbf{X}) = \frac{1}{2}\mathbf{X}'(\Sigma_i^{-1} - \Sigma_j^{-1})\mathbf{X} + \mathbf{X}'(\Sigma_i^{-1}\mu_i - \Sigma_j^{-1}\mu_j)\mathbf{X} + \frac{1}{2}\mu_j'\Sigma_j^{-1}\mu_j - \frac{1}{2}\mu_i'\Sigma_i^{-1}\mu_i + \frac{1}{2}\log(|\Sigma_j|) - \frac{1}{2}\log(|\Sigma_i|) \quad (21.18)$$

$\forall i \neq j, i, j = 1, 2, \dots, k$

A partir de aquí, la **regla de clasificación** para un individuo consiste en evaluar la función discriminante (21.18) para cada grupo y asignarlo a aquél que verifique:

$$G = \operatorname{argmax}_t \ln\pi_t + \frac{1}{2}\ln|\Sigma_k| - \frac{1}{2}(X - \mu_t)^t\Sigma_t^{-1}(X - \mu_t) \quad (21.19)$$

En este caso, los límites de la región de clasificación son ecuaciones cuadráticas de x .

21.3.1. Discriminante cuadrático con R: la función qda()

Para ilustrar la realización de un análisis discriminante cuadrático en **R**, se va a ejemplificar la aplicación de la función `qda()` a los datos `iris` utilizados en el modelo lineal. La elección de la misma base de datos responde a un planteamiento didáctico, para poder comparar los resultados de ambos métodos y las diferencias que produce asumir la igualdad de matrices de varianza-covarianza (método lineal) o no asumirlo (método cuadrático).⁸

```
options(digits = 4)
modelo_qda <- qda(Species ~ ., data = entrenamiento_t)
modelo_qda
#> Call:
#> qda(Species ~ ., data = entrenamiento_t)
#>
#> Prior probabilities of groups:
#>      setosa versicolor virginica
#>      0.3333    0.3333    0.3333
#>
```

⁸En una situación real, la estrategia razonable sería decidir previamente sobre la hipótesis de igualdad de las matrices de varianza-covarianza (utilizando, por ejemplo el contraste *M de Box*, aunque es muy sensible al supuesto de normalidad multivariante) y, en función del resultado, optar por uno de las dos alternativas.

```
#> Group means:
#>           Sepal.Length Sepal.Width Petal.Length Petal.Width
#> setosa      -1.0113     0.78049    -1.2900     -1.2453
#> versicolor   0.1014    -0.68675     0.2566     0.1473
#> virginica    0.9099    -0.09374     1.0334     1.0981
```

La representación gráfica de las áreas por las que se clasifican los individuos se representa en la Fig. 21.7:

```
partimat(Species ~ ., data = entrenamiento_t, method = "qda", image.colors =
  c("skyblue", "lightgrey", "yellow"), col.mean = "red")
```

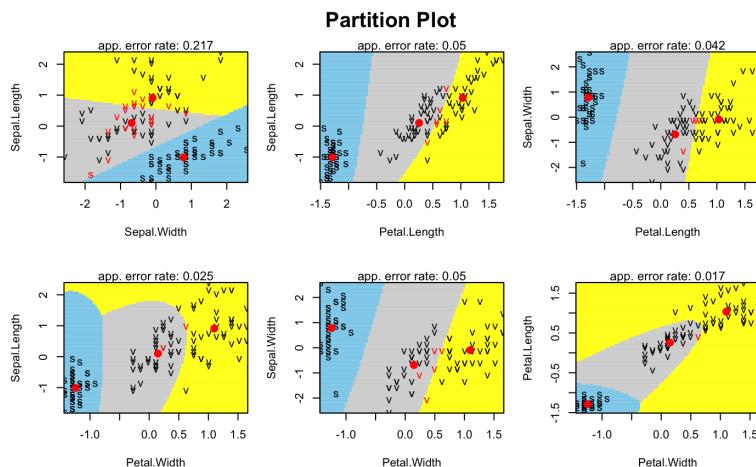


Figura 21.7: Regiones bivariantes de clasificación en cada grupo: setosa (celeste), versicolor (gris) y virginica (amarillo). Centroides en rojo.

Como se aprecia, ahora los contornos de las áreas no son siempre lineales, sino que incluyen fronteras cuadráticas.

```
predicciones_qda <- modelo_qda |> predict(test_t)
table(test_t$Species, predicciones_qda$class, dnn = c("Grupo real", "Grupo
  pronosticado"))
#>           Grupo pronosticado
#> Grupo real  setosa versicolor virginica
#> setosa        10         0         0
#> versicolor     0        10         0
#> virginica      0         1        9
mean(predicciones_qda$class == test_t$Species)
#> [1] 0.9667
```

Resumen

El *análisis discriminante* permite clasificar individuos en distintos grupos preexistentes en relación a una variable cualitativa, a partir de las variables clasificadoras. La información se sintetiza en las funciones discriminantes. Su uso puede tener una finalidad descriptiva, identificar la separación entre grupos y la contribución de cada variable clasificadora; y una finalidad predictiva, para clasificar un individuo nuevo. Los principales tipos son el lineal y el cuadrático, que se desarrollan en **R** con las funciones `lda()` y `qda()`, respectivamente.

Capítulo 22

Análisis conjunto

M^a Leticia Meseguer Santamaría

Universidad de Castilla-La Mancha

22.1. Introducción y conceptos clave

El **análisis conjunto** (o *conjoint analysis*) estudia situaciones de elección múltiple. Funciona dividiendo un producto o servicio en sus componentes (atributos y niveles) y analizando las utilidades parciales de cada uno; después se realizan diferentes combinaciones de éstos para identificar las preferencias del consumidor. Permite conocer las preferencias del público ante el lanzamiento de nuevos productos o servicios, para adaptarlos a ellas y maximizar el éxito. Evalúa la sensibilidad al precio u otras características del producto y predice su comportamiento en el mercado. Mediante este análisis se puede establecer qué atributo y qué categoría (nivel) son los más valorados y cuantificarlos de forma relativa.

Los principales elementos de un análisis conjunto son:

- **Atributos:** características de un producto o servicio sobre las que se basará la elección.
- **Niveles:** valores que puede tomar cada atributo. El número de niveles no tiene por qué ser igual en todos los atributos, pero es conveniente que sea similar para facilitar la elección del entrevistado.
- **Diseño experimental:** proceso estadístico por el que se confeccionan las opciones de las preguntas de la encuesta, y que realiza el investigador.
- **Utilidades parciales:** valoración numérica que representa el grado de preferencia por cada nivel de atributo. Se hace en referencia a las otras opciones. También se conocen como “*partworth utilities*”, “niveles de preferencia”, “niveles relativos de preferencias” o “valoraciones relativas”.
- **Importancia:** valores numéricos que indican la preferencia por cada atributo.

- **Perfil:** combinación concreta de niveles de los atributos de un producto o servicio. También se denomina alternativa.
- **Escenarios:** conjunto de perfiles entre los que el entrevistado tiene que señalar sus preferencias.
- **Pregunta:** conjunto de opciones, normalmente entre 8 y 12.
- **Probabilidad de elección:** probabilidad de tomar una decisión determinada considerando las utilidades parciales.
- **Modelo de comportamiento:** modelo de decisión que se infiere a partir de las probabilidades de elección, después del análisis de las preferencias (utilidades parciales) de los entrevistados.
- **Análisis:** estimación de las utilidades parciales.
- **Valoración:** impacto del nuevo producto o servicio, basado en el modelo de comportamiento que se haya establecido.

Ejemplo: elección de gimnasio. Sea una muestra de gimnasios en la que se identifican 4 atributos (características) con distintos niveles. Se realiza una encuesta sobre qué gimnasio se prefiere. Los participantes eligen una opción entre las distintas combinaciones ofrecidas (**preguntas**). Mediante el análisis de los resultados de la encuesta se extrae el peso de cada atributo y nivel en las respuestas (**utilidades parciales**), que describen las preferencias medias de los encuestados; así mismo, se identifican los atributos y niveles más valorados y su importancia relativa.

Por ejemplo, los resultados pueden indicar que el horario más demandado es el de 09:00 a 23:00 durante 7 días a la semana.

HORARIO
CL_PROGRAMADAS
LIMPIEZA
PRECIO

24 horas (7d)
5 clases/día
3 veces al día
100 €/mes

09:00-23:00 (7d)
3 clases/día
2 veces al día
50 €/mes

07:00-23:00 (5d)
1 clase/día
1 vez al día
25 €/mes

ATRIBUTOS
NIVELES DE CADA ATRIBUTO

Figura 22.1: Elección de gimnasio. Atributos: horario, clases programadas, limpieza y precio

22.2. Tipos de análisis conjunto

En la literatura científica sobre el análisis conjunto se diferencian tres formas de abordar la investigación:

El análisis tradicional de perfil completo (*full profile*)

Este análisis, también llamado *conjoint value analysis (CVA)*, muestra al entrevistado una selección de productos resultantes de la combinación de determinados niveles de una serie de atributos (es decir, una serie de perfiles) y se le pide que los valore según su preferencia en una escala numérica (utilidad). Las preferencias globales de cada individuo se descomponen en

utilidades independientes y compatibles para cada atributo y nivel mediante métodos basados en regresión lineal múltiple (véase Cap. 15), que proporcionan las utilidades para los distintos niveles, *partworth utilities* o, simplemente, *partworths*.

En el diseño original se detallan todos los perfiles hipotéticos (perfiles completos). Sin embargo, su número suele ser alto y se hace uso de técnicas de investigación que indiquen los más significativos. Está limitado a un análisis con muy pocos atributos y niveles.

El análisis conjunto adaptativo o *adaptive conjoint analysis* (ACA)

La recolección de datos se hace en dos fases: en la primera, el encuestado señala la importancia que le da a cada atributo; y en la segunda, asigna utilidades a un número limitado de perfiles. Es decir, se le conduce a través de una investigación sistemática por diferentes secciones, en las que se le presentan uno o pocos atributos, y que se va adaptando a sus respuestas. Se obtienen las valoraciones de los niveles de interés (utilidades parciales). Está limitado por su complejidad y por el uso de software especializado, aunque permite un gran número de atributos y niveles.

El análisis conjunto basado en elecciones o *choice-based conjoint* (CBC)

Pretende un mayor realismo, mostrando a cada entrevistado un grupo de productos distintos y se le pide que elija cuál de ellos prefiere. Además, contempla la posibilidad de no elegir ninguna alternativa. Las utilidades de los atributos-niveles se calculan mediante un modelo de regresión no lineal conocido como modelo de elección discreta. Sus principales inconvenientes son: (i) tiene una mayor complejidad en su diseño y análisis, (ii) requiere muestras muy grandes para tener validez estadística, y (iii) el número máximo de atributos que admite es 10. Su mayor ventaja es que presenta un mejor sistema de elección que las alternativas anteriores.

22.3. Etapas de la realización del análisis conjunto

Para la aplicación correcta de un análisis conjunto es conveniente seguir una serie de etapas que faciliten las elecciones metodológicas que han de hacerse, así como la interpretación de los resultados.

Planteamiento del problema. Se analiza la oportunidad de aplicar esta técnica dados el objetivo y los datos del proyecto. Para ello, se debe especificar tanto el producto o servicio como los atributos que se quieren estudiar.

Elección de la metodología conjoint a aplicar. Ésta dependerá, básicamente, de las características y cantidad de atributos estudiados; también se debe considerar la forma en que se valorarán por parte de los individuos. A modo de guía:

- Se opta por CVA cuando se analizan hasta 6 atributos, de productos o servicios habituales, de forma que el encuestado pueda hacer una elección rápida, sin demasiada reflexión.
- Se elige ACA cuando se analizan más de 10 atributos, procediendo de manera que la elección sea rápida pero reflexiva, ya que se trata de un proceso adaptativo cuyo resultado final depende, en buena medida, de la idoneidad de las primeras valoraciones.

- Se usar CBC si se analizan entre 6 y 10 atributos y con elecciones reflexivas sobre productos, puesto que el encuestado no asigna directamente utilidades, sino que se limita a elegir una (o ninguna) de las opciones presentadas.

La selección de elementos. Se escogen sólo los **atributos** que condicionan la elección, es decir, los que expliquen las preferencias de los individuos y permitan diferenciar bien entre los productos o servicios; deben estar lo más cercanos posible a la independencia entre ellos. Para cada atributo se eligen sus **niveles**; deben ser mutuamente excluyentes y cubrir todo el rango de posibilidades. Por último, y dependiendo de la metodología utilizada, se determinan los **perfles y escenarios**.

Creación de estímulos. Se utilizarán diseños factoriales fraccionados ortogonales, que reducen el número de perfles que se le muestran al entrevistado. Como el número de posibles perfles es la multiplicación del número de niveles de todos los atributos, puede ser imposible en la práctica que el individuo indique sus preferencias entre todos ellos. Por ello, se seleccionan sólo algunos de los perfles, que sean representativos del resto, es decir, que en los perfles incluidos en los estímulos aparezca cada nivel de cada factor combinado con el resto de niveles de forma lo más proporcional posible.¹ La selección se efectúa mediante diseño factorial de experimentos, fraccionado (que jerarquiza los efectos, permitiendo la reducción de perfles) y ortogonal (equilibrado en los niveles).

Forma de presentación. Dependerá de la metodología elegida. La Fig. 22.2 muestra algunos ejemplos.

- CVA: *matriz de comparaciones o trade-off*, en la que el entrevistado valora la combinación de atributos y niveles (sólo es válida para dos atributos). *Perfiles completos para ordenar*, donde se elaboran perfles de cada producto o servicio utilizando sólo un nivel de cada atributo y el encuestado los valora (ordena) según sus preferencias. *Perfil determinado para valorar*, combinación que el encuestado valora según sus preferencias.
- ACA: *comparaciones pareadas*, en las que se comparan dos perfles incompletos.
- CBC: *elección de un perfil*, en el que los encuestados señalan el perfil preferido entre el subconjunto que se les muestra, sin valorarlos ni ordenarlos.

Trabajo de campo y tratamiento de los datos

- CVA: la recogida de datos es en papel u ordenador, y el análisis en ordenador con software especializado.
- ACA y CBC: la recogida y el análisis son con ordenador.

¹Por ejemplo, si se consideran 4 atributos con 3 niveles cada uno, habría un total de 81 perfles diferentes. Se elige un número menor de perfles (suele ser habitual considerar como mínimo el número total de niveles menos el de atributos más uno; en este ejemplo $4 \times 3 - 4 + 1 = 9$ perfles) y se busca que en ellos aparezca, aproximadamente, el mismo número de veces cada nivel de cada atributo.

22.3. Etapas de la realización del análisis conjunto

379

Matriz Trade-off o de comparaciones				Tarjetas perfiles-valoración			
PAN		Forma		Alternativa A		Alternativa B	
		Pan redondo	Barra	Artesano Pan redondo 125 gr	Artesano Pan redondo 250 gr	Artesano Barra 125 gr	Artesano Barra 250 gr
Tipo	Artesano	6	4	Alternativa C	Alternativa D	Alternativa E	Alternativa F
	Industrial	2	5	Industrial Pan redondo 125 gr	Industrial Pan redondo 250 gr	Artesano Molde 125 gr	Artesano Molde 250 gr
				Alternativa G	Alternativa H	Alternativa I	Alternativa J
				Industrial Barra 125 gr	Industrial Barra 250 gr	Industrial Molde 125 gr	Industrial Molde 250 gr
				Alternativa K	Alternativa L		
				Industrial Barra 125 gr	Industrial Molde 250 gr		

Matriz Trade-off o de comparaciones				Comparaciones pareadas			
¿Valore de 1 a 5 su gusto por este pan?				¿Valore de 1 a 5 estas dos opciones de pan?			
<div style="border: 1px solid black; padding: 10px; text-align: center;"> Artesano Pan redondo 125 gr </div>				<div style="display: flex; justify-content: space-around;">   </div>			
<div style="border: 1px solid black; padding: 10px; text-align: center;"> 1 2 3 4 5 Nada Poco Normal Bastante Mucho </div>				<div style="display: flex; justify-content: space-around;"> <div style="border: 1px solid black; padding: 5px; text-align: center;"> Opción A Artesano Pan redondo 125 gr </div> <div style="border: 1px solid black; padding: 5px; text-align: center;"> Opción B Industrial Barra 250 gr </div> </div>			
Prefiero opción A				Indiferencia			
Prefiero opción B				1 2 3 4 5			

Elección de un perfil entre varios					
Alternativa A		Alternativa B		Alternativa C	
Artesano Pan redondo 125 gr	Artesano Barra 250 gr	Industrial Pan redondo 250 gr			
1	2	3			
Alternativa D		Alternativa D			
Industrial Barra 250 gr	No me gusta ninguna opción				
4	5				

Figura 22.2: Formas de presentación

Estimación de las utilidades

Con CVA se utiliza, por lo general, un modelo regresión por mínimos cuadrados ordinarios (OLS), mediante el cual se determinan las utilidades parciales o *partworths*, que indican las preferencias del encuestado mediante un modelo aditivo lineal en relación a los niveles de los atributos considerados como referencia.

Sea la variable X_{jk}^i , que indica si el nivel k del j -ésimo atributo está o no en el i -ésimo perfil. Dicha variable toma sólo los valores $X_{jk}^i = 1$ (si está) y $X_{jk}^i = 0$ si no lo está. Sea Y^i la preferencia que tiene un individuo sobre el i -ésimo perfil. Entonces la función de utilidad a estimar es:

$$Y^i = \beta_0 + \sum_{j,k} U_{jk} X_{jk}^i + \varepsilon^i, \quad (22.1)$$

donde los coeficientes U_{jk} son las utilidades parciales, o *partial partworths*, que indican la utilidad que el individuo asigna a cada nivel de cada atributo, y ε^i denota el término de error.

En el modelo (22.1) habrá multicolinealidad², ya que todos los atributos deben estar presentes en todos los perfiles, es decir $\sum_k X_{jk}^i = 1$. Para evitar las consecuencias indeseadas de la multicolinealidad en la estimación de las utilidades parciales, se elimina uno de los niveles de cada factor (sin pérdida de generalidad, el último, K), estimándose por OLS la función de utilidad (que ahora se denomina “restringida”):

$$Y^i = \delta + \sum_{j,k, k \neq K} \gamma_{jk} X_{jk}^i + \varepsilon^i, \quad (22.2)$$

donde $\delta = \beta_0 + \sum_j U_{jK}$ y $\gamma_{jk} = U_{jk} - U_{jK}$. A partir de la estimación de (22.2), se pueden calcular los valores de la función de utilidad original en (22.1).³

Si hay Z individuos, cada uno de ellos, z , dará una valoración diferente a los perfiles i , es decir los valores de Y_z^i serán diferentes, produciendo una estimación diferente de la función de utilidad (22.1) y, por tanto, diferentes utilidades parciales $\widehat{U}_{jk,z}$. Para obtener las utilidades para el conjunto de individuos, se procede al cálculo de sus valores medios:

$$\widehat{U}_{jk} = \frac{\sum_z \widehat{U}_{jk,z}}{Z} \quad (22.3)$$

Se obtienen, así, las **utilidades** de cada nivel k de cada atributo j , reflejando la importancia que conceden los individuos a cada uno de esos niveles.

Para determinar la importancia de cada atributo j se utiliza la diferencia entre la utilidad más alta y más baja de sus niveles, es decir:

²En otros términos, el modelo no tendrá rango completo y las estimaciones no serán únicas.

³Los valores U_{j1} se pueden calcular imponiendo que $\sum_k U_{jk} = 1$ para cada atributo j . Menos frecuente, aunque también válido, es expresar la utilidades relativas a un nivel de referencia (en este caso, el último) por lo que se consideraría cada $U_{jK} = 0$.

$$Imp_j = \max_k \{U_{jk}\} - \min_k \{U_{jk}\} \quad (22.4)$$

En términos relativos, la importancia de cada atributo j respecto al conjunto de atributos, se puede expresar como:

$$ImpRel_j = \frac{Imp_j}{\sum_{t=1}^J Imp_t}. \quad (22.5)$$

Por último, una vez estimadas las utilidades, la *utilidad total* de un perfil i es la suma de las utilidades de los niveles de cada uno de los atributos que lo definen más la constante de regresión:⁴

$$\widehat{U}_i = \beta_0 + \sum_{j=1}^J \widehat{U}_{jk} \text{ para los niveles presentes en el perfil,} \quad (22.6)$$

donde: \widehat{U}_i es la utilidad total del perfil i ; \widehat{U}_{jk} es la utilidad parcial asociada con el nivel k del atributo j ; y β_0 es la constante de regresión.

En los casos de las metodologías ACA y CBC, la estimación de las utilidades parciales es más compleja, debiendo recurrir a modelos de regresión no lineal. Sin embargo, su interpretación es la misma que en el caso de la metodología CVA.

Interpretación de resultados. Algunos de los resultados del análisis conjunto que frecuentemente se analizan son:

- La importancia de los atributos y niveles. La importancia relativa de cada atributo ($ImpRel_j$) refleja la opinión de los individuos sobre la relevancia (en porcentaje) de dichos atributos a la hora de evaluar los perfiles. Igualmente, para cada atributo j , las utilidades de sus niveles U_{jk} muestran la importancia relativa asignada a dichos niveles en la valoración de cada atributo.
- Las utilidades totales de cada perfil, o la comparación relativa entre un conjunto determinado de perfiles, permiten detectar cuáles son los preferidos por los individuos. Esta información puede orientar en la determinación de la configuración final del producto o servicio analizado.
- La influencia de un determinado atributo sobre la elección de los individuos. Así, es posible estimar el cambio en la utilidad que se produciría por la modificación de sus niveles (cambio generalmente conocido como *elasticidad*); por ejemplo, si un atributo fuese el precio, se podría dar una aproximación a la variación de utilidad producida por un incremento del precio.

⁴Si se desea obtener la utilidad total del perfil i para un individuo concreto, z , bastaría con utilizar la (22.6) para los valores concretos de $\beta_0 z$ y $U_{ij,z}$.

- Al disponer de las utilidades parciales de cada individuo, es posible estudiar si éstas son homogéneas para toda la muestra o si, por el contrario, existen *grupos* de individuos con utilidades parecidas entre ellos y diferenciadas del resto. Este planteamiento podría ayudar a segmentar el mercado y a ofrecer productos o servicios diferenciados a cada nicho de mercado.

22.4. Procedimiento con R: la función Conjoint()

Para exemplificar la puesta en práctica del análisis conjunto, con el método de perfil completo (*CVA*), se utiliza el paquete **conjoint** y su base de datos **tea**, compuesta por 5 listados: niveles (**tlevn**), perfiles (**tprof**), preferencias (**tpref**), preferencias en forma matricial (**tprefm**) y la simulación de perfiles (**tsimp**).

El listado **tprof** recoge los 13 perfiles que se presentan a los individuos, cada uno de ellos con una combinación de niveles de cada factor. El hecho de utilizar un diseño factorial fraccionario ortogonal ha permitido encontrar una muestra de perfiles *representativa* del total (habría $3 \times 3 \times 3 \times 2 = 54$ perfiles distintos) donde cada nivel de cada atributo está representado, aproximadamente, de forma proporcional.

La matriz **tprefm** contiene las preferencias que cada uno de los 100 individuos asigna a cada uno de los perfiles *i* que se le presentan (estas preferencias se recogen ordenadas en el vector **tpref**). Por último, **tsimp** muestra cuatro simulaciones de perfiles distintos a los mostrados a los individuos.

En este ejemplo, se desea conocer el por qué unos téns son preferidos a otros. Para ello, en el contexto de un análisis tradicional con perfiles incompletos, se seleccionan 4 atributos, con sus correspondiente niveles (**tlevn**):

- Precio, distinguiendo entre bajo, medio y alto.
- Variedad, distinguiendo entre negro, verde y rojo.
- Forma de presentación, distinguiendo entre bolsita, granulado y hojas.
- Aroma, distinguiendo entre aromático y no aromático.

De las 54 posibles combinaciones se seleccionan 13 perfiles, valorados de 0 a 10 según las preferencias de cada uno de los 100 encuestados.

Para realizar el análisis conjunto propuesto, primeramente se carga tanto el paquete como los datos:

```
library("conjoint")
data("tea")
```

22.4. Procedimiento con **R**: la función **Conjoint()**

383

La función **Conjoint()** devolverá las utilidades de los niveles, el vector de porcentajes de la importancia de los atributos y las gráficas correspondientes.⁵

```
Conjoint(y=tpref, x=tprof, z=tlevn)
```

En la Tabla 22.1 se exponen otras funciones disponibles para obtener algunos resultados concretos:

Tabla 22.1: Funciones en **R** para cálculos sobre análisis conjunto

Función en R	Utilidad Resultado
caUtilities(y=tprefm , x=tprof, z=tlevn)	Utilidades medias para cada nivel
caPartUtilities(y=tprefm , x=tprof, z=tlevn)	Matriz de utilidades de los niveles para cada individuo
caTotalUtilities(y=tprefm , x=tprof)	Utilidades totales de los perfiles mostrados para cada individuo
colMeans(caTotalUtilities(y=tprefm , x=tprof))	Utilidades totales medias para cada perfil
ShowAllUtilities(y=tprefm , x=tprof, z=tlevn)	Todas las utilidades
caImportance(y=tprefm , x=tprof)	Importancia relativa media de cada atributo
caModel(y=tprefm[20,], x=tprof)	Utilidades para un solo individuo, por ejemplo el 20.

Así, es posible obtener sólo la información relativa a las utilidades, mediante la función **caUtilities()**, o sólo la correspondiente a la importancia de los atributos, usando la función **caImportance()**:

```
caUtilities(y=tprefm , x=tprof, z=tlevn)
#>
#> Call:
#> lm(formula = frml)
#>
#> Residuals:
#>   Min     1Q Median     3Q    Max
#> -5,1888 -2,3761 -0,7512  2,2128  7,5134
#>
#> Coefficients:
#>              Estimate Std. Error t value Pr(>|t|)
#> (Intercept) 3,55336   0,09068 39,184 < 2e-16 ***
#> factor(x$price)1 0,24023   0,13245  1,814   0,070 .
#> factor(x$price)2 -0,14311   0,11485 -1,246   0,213
#> factor(x$variety)1 0,61489   0,11485  5,354 1,02e-07 ***
```

⁵La función **Conjoint()** proporciona una salida matricial donde, además de los resultados numéricos, proporciona una gráfica de utilidad para cada atributo más una para la utilidad media de cada atributo. Dada la cantidad de información, los gráficos no se reproducen en el texto.

```
#> factor(x$variety)2 0,03489 0,11485 0,304 0,761
#> factor(x$kind)1 0,13689 0,11485 1,192 0,234
#> factor(x$kind)2 -0,88977 0,13245 -6,718 2,76e-11 ***
#> factor(x$aroma)1 0,41078 0,08492 4,837 1,48e-06 ***
#> ---
#> Signif. codes: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
#>
#> Residual standard error: 2,967 on 1292 degrees of freedom
#> Multiple R-squared: 0,09003, Adjusted R-squared: 0,0851
#> F-statistic: 18,26 on 7 and 1292 DF, p-value: < 2,2e-16
#> [1] 3.55336207 0.24022989 -0.14311494 -0.09711494 0.61488506 0.03488506
#> [7] -0.64977011 0.13688506 -0.88977011 0.75288506 0.41077586 -0.41077586
caImportance(y=tprof, x=tprof)
#> [1] 24.76 32.22 27.15 15.88
```

Interpretación del resultado:

- La primera parte de la salida corresponde a los coeficientes del modelo de regresión (22.2), donde no se incluye el último nivel de cada atributo (para evitar la multicolinealidad); se señalan con asteriscos los que resultan estadísticamente significativos. También se indica el grado de ajuste del modelo. A continuación, están las utilidades parciales para todos los niveles de todos los atributos:
- El primer valor (3.5534) corresponde al término independiente del modelo, por lo que no está asociado a ningún atributo.
- Para el atributo **precio**, es el nivel bajo (*low*) el que recibe mayor preferencia (0.2402), siendo el nivel medio (*medium*) el menos preferido (-0.1431), algo por debajo de la preferencia del nivel alto (*high*), -0.0971.
- Para el atributo **variedad**, el té negro (*black*) es el nivel con mayor utilidad (0.6149), seguido del té verde (0.0349), quedando el nivel de té rojo con una preferencia mucho menor (-0.6498).
- En el caso del atributo **presentación**, en hojas (*leafy*) es el nivel más preferido (0.7529), seguido de la modalidad en bolsitas (*bags*), con una utilidad de 0.1369; el nivel granulado presenta la preferencia más baja (-0.8898).
- Por último, para el atributo **aroma**, que sea aromático (*yes*) es el nivel más preferido, con una utilidad de 0.4108; al ser un atributo con solo dos niveles, la preferencia del otro nivel (*no*) es -0.4108.

La última parte de la salida recoge la importancia relativa de cada atributo, calculada a partir de la importancia de cada atributo (22.4) pero expresada como proporción de la suma de importancias de todos ellos (22.5).

En este ejemplo, el atributo con más peso es la **variedad de té**, con una importancia relativa del 32.22 %; el siguiente es la **presentación**, con un 27.15 %; después, el **precio**, con un 24.76 %;

22.4. Procedimiento con **R**: la función *Conjoint()*

385

y por último, el **aroma**, con un 15.88 %. Es decir, el conjunto de 100 individuos, al valorar los perfiles que se les presentan, consideran principalmente el atributo variedad, seguido de forma de presentación y precio; por el contrario, que sea aromático o no aporta relativamente poca valoración.

De forma adicional, el análisis conjunto también permite profundizar en el conocimiento de los individuos en función de sus preferencias. Como ejemplo, se puede abordar cuestión de si existen *grupos* de individuos con preferencias similares entre ellos y diferentes de las del resto de grupos, cuestión conocida generalmente como *segmentación*. En este caso, el objetivo del proyecto sería identificar grupos de encuestados con preferencias similares, segmentando el mercado y permitiendo adaptar los atributos de cada producto o servicio a las características concretas de ese nicho de mercado.

Para ello, se puede utilizar la función `caSegmentation()`, que devuelve un análisis de conglomerados dividiendo a los individuos en k *clusters*, usando el método *k-means* (véase Sec. 31). Como ejemplo, se consideran tres clusters.

El vector de agrupación resume las características de los tres grupos formados.

```
segments<-caSegmentation(y=tprof, x=tprof, c=3)
segments$segm
#> K-means clustering with 3 clusters of sizes 40, 40, 20
#>
#> Cluster means:
#>      [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7]   [,8]   [,9]
#> 1 5.480275 2.9381 1.3681 4.540275 1.9731 3.782900 1.382900 0.965750 2.820750
#> 2 4.754975 4.6918 3.6718 6.964975 6.6918 3.500525 4.385525 2.717225 3.062225
#> 3 2.623500 6.6211 4.7511 2.933500 2.5211 4.189050 4.549050 5.066950 2.086950
#>      [,10]  [,11]  [,12]  [,13]
#> 1 0.111225 3.450750 0.442900 0.692900
#> 2 1.840925 6.292225 6.595525 7.105525
#> 3 5.312100 4.266950 4.859050 3.569050
#>
#> Clustering vector:
#> [1] 2 3 2 2 2 1 2 3 2 2 2 2 1 1 1 1 3 1 3 1 1 2 1 2 3 2 3 3 2 2 1 2 3 2 2 2 2
#> [38] 1 1 1 1 3 1 3 1 2 2 1 1 1 2 1 1 1 2 2 1 2 1 3 1 1 2 3 3 2 1 1 1 2 2 1 2 3
#> [75] 2 2 2 1 2 2 3 3 2 2 1 1 1 3 1 3 1 1 2 1 3 2 2
#>
#> Within cluster sum of squares by cluster:
#> [1] 1605.654 2690.267 1131.293
#> (between_SS / total_SS = 41.4%)
#>
#> Available components:
#>
#> [1] "cluster"      "centers"       "totss"        "withinss"     "tot.withinss"
#> [6] "betweenss"    "size"         "iter"         "ifault"
```

La salida contiene, en el apartado *Cluster means*, las utilidades medias de cada nivel (recogidas en las 13 columnas) para cada uno de los tres grupos (en filas), mostrando las diferencias entre

ellas. La composición de cada cluster se muestra en el apartado *Clustering vector*, que recoge el cluster de pertenencia de cada uno de los cien individuos.

Resumen

El **análisis conjunto** estudia situaciones de elección múltiple. Divide un producto o servicio en atributos y niveles y analiza las utilidades parciales de cada uno; después, se realizan diferentes combinaciones de éstos para identificar las preferencias del consumidor, estableciendo qué atributos y niveles son los más valorados, cuantificando dicha valoración de forma relativa. Permite evaluar las preferencias del público ante el lanzamiento de nuevos productos o su sensibilidad de alguna característica, como el precio, el formato, cambios en la imagen del producto, etc.

Capítulo 23

Análisis de tablas de contingencia

José-María Montero

Universidad de Castilla-La Mancha

23.1. Introducción

Las **tablas de contingencia** analizan la relación existente entre variables categóricas, o susceptibles de categorizar, con un número de categorías finito. Dada su naturaleza, no permiten el uso de las tradicionales operaciones aritméticas, con lo cual, en el ámbito de la Estadística Descriptiva su análisis suele basarse en diagramas de barras y porcentajes (véase Cap. 11), y en la esfera de la Inferencia Estadística (Cap. 13) se centra en los contrastes de hipótesis no paramétricos y, básicamente, en el contraste de independencia entre dos o más de estas variables. Una pregunta que suele hacerse toda aquella persona que se acerca por primera vez al análisis de tablas de contingencia es el significado del término “contingencia”. Pues bien, este término fue acuñado por Pearson ([Pearson, 1904](#)) al apuntar: “... Este resultado nos permite partir de la teoría matemática de la independencia probabilística, tal como se desarrolla en los libros de texto elementales, y construir a partir de ella una teoría generalizada de la **asociación** o, como yo la llamo, **contingencia**.”

El análisis de tablas de contingencia (o de asociación) permite dar respuesta, entre otras, a preguntas como: los factores involucrados en una tabla de contingencia, ¿son independientes o están asociados? Si están asociados, ¿qué niveles de dichos factores son los que están asociados?, ¿cuál es la intensidad de dicha asociación?

23.1.1. Notación

Sea una población (o una muestra) de N elementos sobre la que se pretende analizar, simultáneamente, dos (por simplicidad) **atributos** o **factores** (A y B) con R y C **niveles**, **modalidades**

o **categorías**, respectivamente. Sean $\{A_1, A_2, \dots, A_R\}$ y $\{B_1, B_2, \dots, B_C\}$ los niveles anteriormente aludidos. Sea n_{ij} el número de elementos que presentan a la vez las modalidades i y j de los factores A y B , respectivamente. La tabla estadística que describe, conjuntamente, estos N elementos (en otros términos, que muestra las frecuencias conjuntas de los niveles de ambos factores) se denomina **tabla de contingencia**.

		Factor B				
		Nivel B_1	Nivel B_2	...	Nivel B_C	Total
Nivel A_1		n_{11}	n_{12}	...	n_{1C}	$n_{1.}$
Nivel A_2		n_{21}	n_{22}	...	n_{2C}	$n_{2.}$
.	
Factor A	
.	
Nivel A_R		n_{R1}	n_{R2}	...	n_{RC}	$n_{R.}$
Total		$n_{.1}$	$n_{.2}$...	$n_{.C}$	n

A modo de ejemplo, considérese una muestra de 80 ayuntamientos de una CC.AA., anotándose en la base `ayuntam`, incluida en el paquete CRD del libro, el signo político del equipo gubernamental (`signo_gob`) y si prestan o no públicamente el servicio X (`serv`). Los resultados obtenidos fueron los siguientes:

```
library("CDR")
data("ayuntam")

summarytools::ctable(ayuntam$signo_gob, ayuntam$serv, headings = TRUE) |>
  print()
#> Cross-Tabulation, Row Proportions
#> signo_gob * serv
#> Data Frame: ayuntam
#>
#> -----
#>           serv      No       Sí     Total
#>   signo_gob
#>   Avanzados        14 (33.3%)  28 (66.7%)  42 (100.0%)
#>   Ilustrados         6 (15.8%)  32 (84.2%)  38 (100.0%)
#>   Total            20 (25.0%)  60 (75.0%)  80 (100.0%)
#> -----
```

Del estudio de las distribuciones marginales de ambos factores se deduce que el 52,5 % de los ayuntamientos de la CC.AA. están regidos por los Avanzados y el 47,5 % por los Ilustrados. Y, más interesante, que en el 75 % de los ayuntamientos prestan el servicio X .

El análisis de la tabla de contingencia daría respuesta a las siguientes preguntas: ¿La prestación pública del servicio X es independiente del signo político del ayuntamiento o depende de dicho signo? En este último caso: ¿Qué signo político está asociado con la prestación pública y cuál

no?, ¿la asociación entre los factores “Signo político del equipo gubernamental” y “Prestación pública del servicio X ” es muy intensa? Pero dicho análisis se abordará posteriormente.

En función del número de factores involucrados en la tabla y del número de niveles de cada uno de ellos se tiene la siguiente tipología de tablas de contingencia:

- Tablas $R \times C$: 2 factores, el primero con R niveles y el segundo con C niveles.
- Tablas $R \times C \times M$: 3 factores, con R , C y M niveles, respectivamente.
- Y así sucesivamente.

Dentro de las tablas $R \times C$ se distinguen las tablas 2×2 de las demás, por su especial interés en la realidad y por criterios pedagógicos, al ser las más sencillas.

23.1.2. Diseños experimentales o procedimientos de muestreo que dan lugar a una tabla de contingencia

Una cuestión a la que no se le da la suficiente importancia es la forma en la que se toma la información contenida en la tabla (el diseño del experimento o procedimiento de muestreo). Dada una determinada tabla de contingencia, ésta puede haber sido obtenida mediante uno u otro diseño de experimento o procedimiento de muestreo, y esta circunstancia no es baladí, puesto que condiciona su análisis, sobre todo cuando el tamaño muestral es pequeño.

Sin ánimo de exhaustividad, los diseños experimentales o procedimientos de muestreo más habituales que dan lugar a una tabla de contingencia son los siguientes:¹

- **Tipo 1: se fijan los totales marginales de ambos factores**

Ejemplo: se desea investigar si la preferencia de la larva de gorgojo por el tipo de judía es independiente de la cubierta de la semilla o depende de ésta. Para ello se toman 22 judías de tipo A y 18 de tipo B , que se introducen en un recipiente con 33 larvas. Dadas las condiciones de densidad, no entrará más de una larva por judía. Pasado un tiempo prudencial para que las larvas entren en las judías, se cuentan las que han sido atacadas de cada tipo y las que no.²

		Presencia de larva		Total
		NO	SÍ	
Tipo de	A	N_{11}	N_{12}	22
	B	N_{21}	N_{22}	18
	Total	7	33	40

¹Otros procedimientos de muestreo o diseños experimentales no habituales pueden verse en Ruiz-Maya et al. (1995).

²Como en el resto del manual, las variables aleatorias se escriben en mayúsculas y los valores que toman en minúscula. En este caso, los totales marginales han sido fijados y son valores pero no ocurre lo mismo con las frecuencias absolutas en las cuatro celdas de las tablas, que son variables aleatorias.

Como puede apreciarse, los totales marginales de ambos factores han sido fijados en el diseño del experimento.

- **Tipo 2: sólo se fijan los totales marginales de uno de los factores**

Ejemplo: en un municipio se desea investigar si el desempleo es o no independiente del sexo del desempleado. Se seleccionan aleatoriamente 100 varones y 100 mujeres y se les pregunta por su situación laboral (trabajando; en paro). [^tablas_conting-marginales columnas]

[^tablas_conting-marginales columnas]: En este caso los totales marginales por columnas son variables.

		Situación	laboral	
		Trabajando	En paro	Total
Sexo	Varón	N_{11}	N_{12}	100
	Mujer	N_{21}	N_{22}	100
	Total	$N_{.1}$	$N_{.2}$	200

- **Tipo 3: únicamente se fija el tamaño muestral**

Ejemplo: un estudio transversal sobre la prevalencia de osteoporosis y su relación con dietas pobres en calcio incluyó a 400 mujeres entre 50 y 54 años. Cada una de ellas realizó una densíometría de columna y llenó un cuestionario sobre sus antecedentes dietéticos para determinar si su dieta era o no pobre en calcio. [^tablas_conting-marginales filas]

[^tablas_conting-marginales filas]: Ahora tanto los totales marginales por columnas como por filas son variables.

		Dieta pobre	en calcio	
		NO	SÍ	Total
Osteoporosis	SI	N_{11}	N_{12}	$N_{1.}$
	NO	N_{21}	N_{22}	$N_{2.}$
	Total	$N_{.1}$	$N_{.2}$	400

23.2. Contraste de independencia en tablas 2×2

Como se avanzó en la Sec. 23.1, la primera pregunta a la que debe dar respuesta el análisis de tablas de contingencia es si los factores involucrados en la tabla son independientes o, por el contrario, están asociados. La respuesta a esta pregunta exige llevar a cabo un contraste de independencia y, para ilustrarlo, se aborda, inicialmente, el caso de las tablas 2×2 . Dicho contraste se lleva a cabo de tres formas: (i) exacta, (ii) aproximada, y (iii) aproximada con corrección de continuidad.

23.2.1. Planteamiento general del contraste exacto de independencia

■ Hipótesis:

- H_0 : los factores son independientes.
 - H_1 : están asociados.³
- Filosofía del contraste: se trata de un contraste de significación. Por tanto, la tabla observada será “rara” (bajo H_0) si su probabilidad, más la probabilidad de obtener tablas más alejadas de H_0 que ella, es inferior al nivel de significación, α , prefijado para el contraste. En ese caso, se rechaza la hipótesis de independencia entre los factores involucrados en la tabla.

23.2.2. Algoritmo para la realización del contraste exacto de independencia

De acuerdo con la filosofía de los contrastes de significación (Sec. 13.5), el algoritmo para la realización del contraste de independencia en tablas de contingencia es como sigue:

1. Selección de las tablas del espacio muestral que se alejen de la hipótesis de independencia, en la dirección marcada por la hipótesis alternativa, tanto o más que la tabla observada, incluida esta última.
2. Cálculo, bajo la hipótesis de independencia, de la probabilidad de ocurrencia de cada una de las tablas seleccionadas en el punto 1.
3. Suma de dichas probabilidades y comparación con el α prefijado.
4. Toma de la decisión relativa al rechazo o no de la hipótesis de independencia.

Nótese que (i) los pasos 1 y 2 dependen del diseño del experimento o procedimiento de muestreo llevado a cabo; (ii) en ausencia del software adecuado, la realización de un test exacto es un procedimiento laborioso (a veces un reto), con lo cual, si ese fuera el caso, los test aproximados de independencia son bienvenidos.

A continuación, se expone el contraste de independencia, en sus versiones exacta, aproximada y aproximada con corrección de continuidad, cuando el procedimiento de muestreo o diseño experimental es el de tipo 1. En la Sec. 23.2.4 se comentan algunas cuestiones de interés cuando el diseño de muestreo es de tipo 2 o tipo 3.

³También puede establecerse como hipótesis alternativa la asociación en un determinado sentido: “el nivel 1 del factor A está asociado con el 1 del B y el 2 del A con el 2 del B” (asociación positiva); o “el nivel 1 del factor A está asociado con el 2 del B y el 2 del A con el 1 del B” (asociación negativa). En estos dos casos el contraste no sería bilateral sino unilateral.

23.2.3. Contraste de independencia: diseño tipo 1

23.2.3.1. Contraste exacto (test exacto de Fisher)

Considérese el ejemplo del diseño tipo 1 expuesto en 23.1.2⁴. Supóngase que el resultado obtenido fue el siguiente:

```
datos_jud = c(1,6,21,12)
tabla = cbind(expand.grid(list(Tipo_de_judía = c("A","B"),
                                Presencia_larva = c("No","Sí"))),
               count = datos_jud)
tabla_jud <- ftable(xtabs(count~Tipo_de_judía+Presencia_larva, tabla))
```

		Presencia de larva		Total
		NO	SI	
Tipo de	A	1	21	22
	B	6	12	18
	Total	7	33	40

Según el algoritmo expuesto en la Sec. 23.2.2, el contraste es como sigue:

1. **Selección de las tablas que se alejan de H_0 tanto o más que la observada.**⁵ Como se señalaba en Pearson (1904), la teoría de la independencia probabilística indica que, bajo la hipótesis de independencia, el porcentaje de judías de tipo A y de tipo B no atacadas (o atacadas) por una larva de gorgojo tiene que ser el mismo. En otros términos, bajo la hipótesis de independencia, en cada una de las cuatro celdas se tiene que verificar que: $N_{ij} = \frac{N_i N_j}{N}, \forall i, j$, donde $\frac{N_i N_j}{N} = E_{ij}$ se denomina frecuencia esperada bajo la hipótesis de independencia (en este caso, al estar los totales marginales fijos, $E_{ij} = \frac{n_i n_j}{n}$). Denominando $D_{ij} = N_{ij} - E_{ij}, \forall i, j, i = 1, 2, j = 1, 2$, se puede que comprobar que en una tabla 2×2 , $D_{11} = D_{22} = -D_{12} = -D_{21}$, con lo cual, tomando de referencia, por ejemplo, la celda {1,1}, las tablas que se alejan tanto o más que la observada de la hipótesis de independencia son aquellas que verifican, en valor absoluto, que $D_{11} = N_{11} - E_{11} \geq n_{11} - \frac{n_{11} n_{12}}{n}$.

En el ejemplo que se considera, las D_{11} son las siguientes (en negrita las de la tabla observada y aquellas otras que se alejan tanto o más que ella de H_0):

$T_0: -3,85; T_1: -2,85; T_2: -1,85; T_3: -0,85; T_4: 0,15; T_5: 1,15; T_6: 2,15; T_7: 3,15$,

donde el subíndice de T indica el valor de N_{11} en dicha tabla.

Nótese que el criterio anterior no es otro que el criterio general de seleccionar las tablas en las que la diferencia de porcentajes, por ejemplo, por fila, en valor absoluto, sea superior a la de la tabla observada, puesto que $\left| \frac{N_{11}}{n_{11}} - \frac{N_{21}}{n_{21}} \right| = |D_{11}| \frac{n}{n_{11} n_{21}}$.

⁴Se trata de un ejemplo clásico de Sokal and Rolf (2012).

⁵El contraste se ha llevado a cabo de forma bilateral. En caso de una alternativa unilateral (asociación positiva o en el sentido de la diagonal principal; o negativa, en el sentido de la diagonal no principal), el procedimiento sería el mismo y las tablas seleccionadas serían, en el primer caso, todas menos la T_7 , y en el segundo, T_0 y T_1 .

2. Cálculo, bajo la hipótesis de independencia, de la probabilidad de ocurrencia de cada una de las tablas seleccionadas en 1. La probabilidad de ocurrencia de una tabla de contingencia con los totales marginales fijos se puede obtener como el cociente entre el número de disposiciones de las frecuencias observadas favorables a dicha tabla y el número de disposiciones posibles. El número de disposiciones favorables coincide con el coeficiente multinomial (maneras de que de n frecuencias observadas, n_{11} caigan en la celda {1,1}, n_{12} lo hagan en la celda {1,2}, n_{21} lo hagan en la celda {2,1} y n_{22} lo hagan en la celda {2,2}): $\frac{n!}{n_{11}!n_{12}!n_{21}!n_{22}!}$.

El número de disposiciones posibles, supuesta H_0 , es: $\binom{n}{n_{11}} \binom{n}{n_{12}} = \frac{n!}{(n_{11}!n_{12}!) (n_{21}!n_{22}!)}.$

Por tanto, el cociente entre ambas es: $P = \frac{n_{11}!n_{12}!n_{21}!n_{22}!}{n!n_{11}!n_{12}!n_{21}!n_{22}!}.$

En consecuencia, las probabilidades de las tablas seleccionadas en el punto 1 son: $T_0 : 0,0017; T_1 : 0,0219; T_7 : 0,0091$.

3. Suma de dichas probabilidades: 0,0327.

4. Comparación con α y decisión sobre el rechazo o no de la hipótesis de independencia: La decisión depende del valor de α . Si fuera, por ejemplo, 0,05, se rechazaría la independencia entre el tipo de judía y si es o no atacada por la larva de gorgojo.

El código R necesario para tomar llevar a cabo el test exacto de Fisher anterior es:

```
#Ho: Los factores son independientes.
#H1: Los factores están asociados
fisher <- fisher.test(tabla_jud, alternative = "two.sided")
fisher$p.value
#[1] 0.0327607
```

```
#Ho: Los factores son independientes.
#H1: Existe asociación negativa.
fisher_less <- fisher.test(tabla_jud, alternative = "less")
fisher_less$p.value
#[1] 0.02361309
```

```
#Ho: Los factores son independientes.
#H1: Existe asociación positiva.
fisher_greater <- fisher.test(tabla_jud, alternative = "greater")
fisher_greater$p.value
#[1] 0.998293
```

Como puede apreciarse, se rechaza la hipótesis de independencia frente a la de asociación (test bilateral). Esto no significa que las larvas ataquen a un tipo de judía y no al otro. Atacan a ambos tipos, ¡y bastante! Este es el primer hecho que se constata. Sin embargo, atacan más a las judías de tipo A (un 95 % son atacadas) que a las de tipo B (dos terceras partes son atacadas). Esa diferencia porcentual de judías atacadas se considera significativa bajo el supuesto de independencia y, en ese sentido, se dice que existe asociación entre el tipo de judía y la presencia o no de larva atacante. La asociación sería A-SI y B-NO. Sin embargo, ¡cuidado!,

las larvas atacan siempre. La asociación anterior debe entenderse como “el porcentaje de ataque es muy grande en ambos casos, pero en A (mucho) más que en B . Este es el segundo hecho importante que se constata: las larvas muestran una preferencia significativa por las judías tipo A . Aunque ya se ha visto la dirección de la asociación (en el sentido de la diagonal ascendente), en la Sec. 23.4, dedicada a las medidas de asociación en tablas 2x2, se cuantificará su intensidad.

23.2.3.2. Contraste aproximado

En este caso (tipo 1), bajo H_0 : independencia, la frecuencia conjunta de una celda, N_{ij} , cualquiera que sea, se distribuye según una ley hipergeométrica con $E(N_{ij}) = \frac{N_{ij}}{n_{i\cdot} n_{\cdot j}}$ y $V(N_{ij}) = \frac{n_{i\cdot} n_{\cdot j} (n - n_{i\cdot}) (n - n_{\cdot j})}{n^2(n-1)}$. Por consiguiente,

$$P \left(\left(N_{11} - \frac{n_{1\cdot} n_{\cdot 1}}{n} \right)^2 \geq \left(n_{11} - \frac{n_{1\cdot} n_{\cdot 1}}{n} \right)^2 \right) = P \left(\frac{(N_{11} - \frac{n_{1\cdot} n_{\cdot 1}}{n})^2}{\frac{n_{1\cdot} n_{\cdot 1} n_{2\cdot} n_{\cdot 2}}{n^2(n-1)}} \geq \frac{(n_{11} - \frac{n_{1\cdot} n_{\cdot 1}}{n})^2}{\frac{n_{1\cdot} n_{\cdot 1} n_{2\cdot} n_{\cdot 2}}{n^2(n-1)}} \right).$$

Y si ninguna $\hat{E}_{ij} = \frac{n_{ij}}{n_{i\cdot} n_{\cdot j}}$ es inferior a 5, la probabilidad anterior puede aproximarse (teorema central del límite) por:

$$P \left(\chi_1^2 \geq \frac{\left(n_{11} - \frac{n_{1\cdot} n_{\cdot 1}}{n} \right)^2}{\frac{n_{1\cdot} n_{\cdot 1} n_{2\cdot} n_{\cdot 2}}{n^2(n-1)}} \right) = P \left(\chi_1^2 \geq \frac{(n-1)(n_{11}n_{22} - n_{21}n_{12})^2}{n_{1\cdot} n_{\cdot 1} n_{2\cdot} n_{\cdot 2}} \right),$$

donde el estadístico $\frac{(n-1)(n_{11}n_{22} - n_{21}n_{12})^2}{n_{1\cdot} n_{\cdot 1} n_{2\cdot} n_{\cdot 2}}$ se denomina chi-cuadrado ajustado (χ_{adj}^2) y es tal que $\chi_{adj}^2 = \frac{n-1}{n} \frac{n(n_{11}n_{22} - n_{21}n_{12})^2}{n_{1\cdot} n_{\cdot 1} n_{2\cdot} n_{\cdot 2}}$, donde $\frac{n(n_{11}n_{22} - n_{21}n_{12})^2}{n_{1\cdot} n_{\cdot 1} n_{2\cdot} n_{\cdot 2}}$ es el estadístico chi-cuadrado (χ^2) que proporcionan todos los softwares de contraste de independencia en tablas de contingencia.

En el ejemplo propuesto:

```
chisq.test(tabla_jud)$expected
#>      [,1] [,2]
#> [1,] 3.85 18.15
#> [2,] 3.15 14.85
chisq.test(tabla_jud, correct=FALSE)
#>
#> Pearson's Chi-squared test
#>
#> data: tabla_jud
#> X-squared = 5.6828, df = 1, p-value = 0.01713
```

Como $\chi^2 = 5,6828$, entonces $\chi_{adj}^2 = 5,54073$ y $P(\chi_{adj}^2 \geq 5,54073) = 0,0186$.⁶

⁶Este “cálculo extra” es un pequeño peaje que hay que pagar en aras de la exactitud. Y es importante, porque pudiera hacer cambiar la decisión resultante del contraste.

Nótese que la probabilidad exacta de obtener una tabla tan alejada o más de la hipótesis de independencia que la observada (incluida esta) es 0,0327, mientras que la probabilidad aproximada es 0,0186. La aproximación no es muy buena, y ello se debe a la existencia de frecuencias esperadas menores que 5.

23.2.3.3. Contraste aproximado con corrección de continuidad

Como se vio en la subsección anterior, al aproximar la probabilidad de obtención de tablas tanto o más alejadas de H_0 que la observada (que se calcula con una distribución hipergeométrica, que es discreta) mediante una distribución χ^2_1 (que es continua), se comete un “error de continuización”. Dicho error se intenta corregir incluyendo en el contraste una **corrección de continuidad**. Hay varias correcciones que han tenido cierto éxito en la literatura. La más popular es la corrección de Yates, si bien sólo se recomienda cuando las E_{ij} sean múltiplos de 0,5.

En el contraste aproximado con corrección de Yates, se rechaza H_0 si:

$$P\left(\chi^2_1 \geq \frac{(n-1)(|n_{11}n_{22} - n_{21}n_{12}| - 0,5n)^2}{n_{1\cdot}n_{\cdot1}n_{2\cdot}n_{\cdot2}}\right) \leq \alpha,$$

donde el estadístico $\frac{(n-1)(|n_{11}n_{22} - n_{21}n_{12}| - 0,5n)^2}{n_{1\cdot}n_{\cdot1}n_{2\cdot}n_{\cdot2}}$ se denomina estadístico chi-cuadrado ajustado corregido de continuidad de Yates ($\chi^2_{adj,CCY}$).

En el ejemplo propuesto, el test chi-cuadrado con corrección de continuidad de Yates se obtiene directamente con la función `chisq.test()` que, por defecto, incluye el argumento `correct = TRUE`.

```
chisq.test(tabla_jud)
#>
#> Pearson's Chi-squared test with Yates' continuity correction
#>
#> data: tabla_jud
#> X-squared = 3.8637, df = 1, p-value = 0.04934
```

Como el estadístico Chi-cuadrado corregido de continuidad χ^2_{CCY} vale 3,8337, entonces $\chi^2_{adj,CCY} = \frac{39}{40} \times 3,8337 = 3,7636$; y como $P(\chi^2_1 \geq 3,7636) = 0,0524$, H_0 se rechazaría cuando $\alpha > 0,0524$. Nótese que si, por ejemplo, $\alpha = 0,05$, la decisión sobre el rechazo o no de H_0 es distinta con χ^2_{CCY} y $\chi^2_{adj,CCY}$; de ahí la importancia de utilizar el estadístico ajustado.

Por tanto, la corrección de Yates ha transformado la infraestimación de la probabilidad exacta en una sobreestimación de más o menos el mismo tamaño. Ello se debe a que en la tabla observada, hay freecuencias esperadas (E_{ij}) que distan mucho de ser multiplos de 0,5. La corrección de Yates es la que incluye la librería utilizada (`stats`). Otras correcciones pueden verse en Ruiz-Maya et al. (1995) y Montero (2002).

23.2.4. Contraste de independencia: diseños tipo 2 y tipo 3

En el caso tipo 2, para la realización del test exacto, las tablas que se alejan tanto o más que la observada de la hipótesis de independencia son las que verifican:

$$\left| \frac{N_{11}}{n_1} - \frac{N_{21}}{n_2} \right| \geq \left| \frac{n_{11}}{n_1} - \frac{n_{21}}{n_2} \right|,$$

y la probabilidad de ocurrencia de una tabla de contingencia viene dada por:

$$P(N_{11} = n_{11}; N_{12} = n_{12}; N_{21} = n_{21}; N_{22} = n_{22} | N = n) = \binom{n_1}{N_{11}} \binom{n_2}{N_{21}} \left(\frac{N_{\cdot 1}}{n} \right)^{N_{\cdot 1}} \left(\frac{N_{\cdot 2}}{n} \right)^{N_{\cdot 2}}. \quad (23.1)$$

El estadístico de contraste en el test aproximado viene dado por $\chi^2 = \frac{n(n_{11}n_{22} - n_{21}n_{12})^2}{n_1 n_{\cdot 1} n_2 n_{\cdot 2}}$, y por $\chi^2_{CC} = \frac{n((|n_{11}n_{22} - n_{21}n_{12}| - \frac{f}{2})^2)}{n_1 n_{\cdot 1} n_2 n_{\cdot 2}}$ en el caso de estar corregido de continuidad, siendo f el mayor factor común de los tamaños muestrales fijados.

En el caso tipo 3, las tablas que se alejan tanto o más que la observada de la hipótesis de independencia son las que verifican la condición expuesta en el tipo 2 (y tipo 1), siendo su probabilidad de ocurrencia:

$$P(N_{11} = n_{11}; N_{12} = n_{12}; N_{21} = n_{21}; N_{22} = n_{22} | N = n) = \frac{n!}{n_1! n_2! n_{\cdot 1}! n_{\cdot 2}!} \left(\frac{n_{\cdot 1}}{n} \frac{n_{\cdot 2}}{n} \right)^{n_{11}} \left(\frac{n_1}{n} \frac{n_2}{n} \right)^{n_{12}} \left(\frac{n_2}{n} \frac{n_{\cdot 1}}{n} \right)^{n_{21}} \left(\frac{n_1}{n} \frac{n_{\cdot 2}}{n} \right)^{n_{22}}. \quad (23.2)$$

El test aproximado, en este caso, es un test razón de verosimilitudes donde el estadístico de contraste, $G = -2 \ln \frac{n_{11}^{n_{11}} n_{22}^{n_{22}} n_{\cdot 1}^{n_{\cdot 1}} n_{\cdot 2}^{n_{\cdot 2}}}{n_{11}^{n_{11}} n_{21}^{n_{21}} n_{12}^{n_{12}} n_{22}^{n_{22}} n^n}$, también se distribuye como una χ^2_1 en caso de independencia. Apenas hay literatura sobre correcciones de continuidad en este modelo y la poca que hay sugiere la aplicación de la corrección de Yates.

Nota

En el diseño de muestreo tipo 3 se recomienda no usar ninguna corrección de continuidad, salvo que el tamaño muestral sea muy pequeño y sea imprescindible la realización del test.

El código R para llevar a cabo estos dos contrastes aproximados puede verse en la Sec. 23.3.1.

23.3. Contraste de independencia en tablas $R \times C$

El análisis de tablas $R \times C$ puede considerarse, en principio, una generalización del caso de tablas 2×2 . Ahora bien, en el caso $R \times C$ los test exactos, recomendados en el caso de que

$E_{ij} \leq 5$ en más del 20 % de las celdas, [Reynolds (1984)]⁷ son un auténtico reto y aún no están disponibles en el software convencional.

Si no se cumple el requisito anterior, una solución es agrupar categorías, con sentido común y coherencia.⁸ Si la agrupación de categorías no pudiese hacerse, por carecer de sentido o cualquier otro motivo, lo más honesto sería no realizar el contraste hasta disponer de una base de datos mejor.

A la luz de lo anteriormente expuesto, en el caso de tablas $R \times C$ la atención se centra en los tests aproximados.

23.3.1. Contrastes aproximados

Cuando el procedimiento de muestreo o el diseño experimental es de tipo 1 o 2 el contraste aproximado de independencia es el denominado contraste Chi-cuadrado. La filosofía de dicho contraste es la siguiente: parece lógico que el contraste se base en las diferencias (cuadráticas, para que no se compensen las negativas con las positivas) entre las frecuencias observadas y las esperadas bajo la hipótesis de independencia. Si los factores son independientes, dichas diferencias serán pequeñas y atribuibles a fluctuaciones aleatorias. Si están asociados, serán grandes y atribuibles a la asociación existente entre sus niveles. Pearson propuso el siguiente estadístico de contraste:

$$\chi^2 = \sum_{i=1}^R \sum_{j=1}^C \frac{(N_{ij} - E_{ij})^2}{E_{ij}},$$

que, si no se incumple el requisito expuesto en las primeras líneas de la Sec. 23.3, y bajo la hipótesis de independencia, se distribuye como una $\chi^2_{(R-1)(C-1)}$. En caso de que $P\left(\chi^2_{(R-1)(C-1)} \geq \sum_{i=1}^R \sum_{j=1}^C \frac{(n_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}}\right)$ sea inferior al nivel de significación prefijado, se rechaza la hipótesis de independencia.⁹ Téngase en cuenta que en el caso $R \times C$ la hipótesis alternativa es “al menos un nivel de un factor está asociado con un nivel del otro factor”.

La razón de que las diferencias $(N_{ij} - \hat{E}_{ij})^2$ se dividan por \hat{E}_{ij} en el estadístico chi-cuadrado de Pearson es la siguiente: la misma diferencia $N_{ij} - \hat{E}_{ij}$ puede significar cosas bien diferentes. Una diferencia de 5 no es nada si $\hat{E}_{ij} = 1000$; pero es muchísimo si $\hat{E}_{ij} = 2$. Por eso la diferencia (cuadrática) se pone en relación con la frecuencia esperada.

En el caso de que el procedimiento de muestreo sea del tipo 3, aunque puede aplicarse el contraste chi-cuadrado, es recomendable proceder con el contraste de independencia de razón

⁷La razón es que, sea cual sea el diseño de muestreo, si la probabilidad de que un elemento de la tabla caiga en una determinada celda $\{i,j\}$ es muy pequeña (se suele utilizar el límite del 5 %), entonces la distribución de probabilidad de N_{ij} es muy asimétrica, y para que se simetrique lo suficiente y se pueda aproximar por una Normal (que después, al cuadrado, participará en una chi-cuadrado), el tamaño muestral, en el tipo 3, y los totales marginales fijados, en los tipos 1 y 2, tienen que ser grandes (se suele fijar el valor de 100), de ahí el límite de $E_{ij} \leq 5$; por eso es recomendable que el tamaño muestral sea grande y los totales marginales no estén desequilibrados. En todo caso, lo ideal es que requisito $E_{ij} \leq 5$ se cumpla en todas las celdas de la tabla.

⁸En tablas 2×2 no se pueden agrupar filas, y la única solución son los tests exactos.

⁹ $\hat{E}_{ij} = \frac{n_i \cdot n_j}{n}$ es el valor de E_{ij} calculado con los datos de la tabla observada. Igualmente, n_{ij} son las frecuencias de la tabla observada.

de verosimilitudes, que compara por cociente las frecuencias esperadas bajo la hipótesis de independencia y las observadas. Se basa en la razón de la verosimilitud de la hipótesis de independencia a la luz de la muestra obtenida y del máximo de la función de verosimilitud, Λ . Bajo el supuesto de independencia la razón será cercana a la unidad, atribuyéndose la diferencia a fluctuaciones aleatorias; el logaritmo neperiano de dicha razón (negativo) estará cercano a cero. En caso contrario, el cociente de verosimilitudes (negativo) disminuye, tanto más cuanto más diferencia hay entre la verosimilitud de la hipótesis de independencia y el máximo de la función de verosimilitud. En Wilks (1935) se demostró que, cuando la hipótesis de independencia es cierta, $G = -2\ln\Lambda$, con $\Lambda = \prod_{i=1}^R \prod_{j=1}^C \left(\frac{E_{ij}}{N_{ij}}\right)^{N_{ij}}$ se distribuye como una $\chi^2_{(R-1)(C-1)}$. Ambos estadísticos de contraste, χ^2 y G , son asintóticamente equivalentes.

A modo de ejemplo, se quiere contrastar si en la Comunidad de Madrid la opinión sobre la presidenta Dña. Isabel Díaz Ayuso depende de la zona geográfica o si por el contrario, es independiente de ella. Para ello se encuestan, por algún procedimiento aleatorio 2795 personas con derecho a voto en la comunidad y se eliminan las respuestas “NS/NC/me es indiferente”. Los resultados obtenidos fueron los siguientes (dataset `ayuso` del paquete CDR):

```
data("ayuso")
tabla_ayuso<-table(ayuso)
tabla_ayuso
#>           opinion
#> zona          n1_nefasta n2_mala n3_buena n4_excelente
#> n1_mad_muni      25     500     50     1000
#> n2_metropol       10     280     50     460
#> n3_extraradio      5     130     25     260
chisq.test(tabla_ayuso)$expected
#>           opinion
#> zona          n1_nefasta n2_mala n3_buena n4_excelente
#> n1_mad_muni   22.540250 512.7907 70.43828   969.2308
#> n2_metropol    11.449016 260.4651 35.77818   492.3077
#> n3_extraradio   6.010733 136.7442 18.78354   258.4615
chisq.test(tabla_ayuso, correct=FALSE)
#>
#> Pearson's Chi-squared test
#>
#> data: tabla_ayuso
#> X-squared = 19.486, df = 6, p-value = 0.003418
```

```
library("DescTools")
GTest(tabla_ayuso,correct = "none")
#>
#> Log likelihood ratio (G-test) test of independence without correction
#>
#> data: tabla_ayuso
#> G = 19.357, X-squared df = 6, p-value = 0.003602
```

Como puede verse, sea cual sea el estadístico de contraste, la hipótesis de independencia se rechaza para cualquiera de los valores de α utilizados en la práctica (1%, 2,5%, 5%, 10%).

23.3.2. Contraste aproximado con corrección de continuidad

Afortunadamente, en la mayoría de las ocasiones el tamaño muestral es grande y los totales marginales no están muy desequilibrados, con lo que los estadísticos chi-cuadrado y chi-cuadrado corregido de continuidad son prácticamente iguales, sobre todo si el número de niveles de ambos factores es elevado. En caso de utilizar una corrección de continuidad, hay unanimidad en utilizar la de Yates, sea cual sea el procedimiento de muestreo y el test (chi-cuadrado o G), si bien dicha unanimidad tiene mucho que ver con que es la única que está programada en el software convencional sobre tablas de contingencia.

23.4. Medidas de asociación en tablas 2×2

Si no se rechaza la hipótesis de independencia, el análisis de la tabla se puede dar por finalizado. En caso contrario, el nuevo objetivo es determinar la dirección de la asociación detectada (o las fuentes de asociación en el caso $R \times C$) y su intensidad, y para ello se utilizan las denominadas medidas de asociación. Igual que en el contraste de independencia, se distinguirán los casos 2×2 y $R \times C$, en esta ocasión no tanto por motivos pedagógicos sino porque las situaciones son bien diferentes.

En el caso 2×2 , los tipos de asociación en función de su dirección (positiva y negativa) ya se definieron en la Sec. 23.2.1. Por lo que se refiere a los límites de su intensidad, se dice que la asociación es perfecta cuando al menos uno de los niveles de uno de los factores queda determinado por un nivel del otro factor. La asociación perfecta puede ser estricta o implícita de tipo 2:¹⁰

- Estricta: dado el nivel de un factor, el nivel del otro queda inmediatamente determinado.
- Implícita de tipo 2: dado un nivel de un factor, el nivel del otro queda inmediatamente determinado; dado el otro nivel, no queda determinado el nivel del otro factor.

23.4.1. La Q de Yule

En caso de independencia, las frecuencias observadas coinciden con las esperadas. A medida que las primeras se separan de las segundas, se produce un alejamiento de dicha hipótesis y los niveles de los factores aumentan la intensidad de su asociación. Por consiguiente, las diferencias D_{ij} entre las frecuencias observadas y las esperadas bajo el supuesto de independencia pueden ser la base de una magnífica medida de asociación. A mayores diferencias, mayor asociación. Mas sencillo todavía: una única diferencia, por ejemplo la D_{11} , podría servir como medida de asociación porque, como bajo la hipótesis de independencia, $D_{ij} = 0$ y $D_{11} = D_{22} = -D_{12} = -D_{21}$, entonces se tiene que:

¹⁰La asociación perfecta implícita de tipo 1, que no puede darse en tablas 2×2 , consiste en que, dado un nivel de un factor, el nivel del otro queda inmediatamente determinado; pero dado el nivel de este último no se puede determinar el nivel del primero.

- En caso de independencia: $D_{11} = D_{22} = 0$ y $D_{12} = D_{21} = 0$, o simplemente, $D_{11} = 0$.
- En caso de asociación positiva: $D_{11} = D_{22} \geq 0$ y $D_{12} = D_{21} \leq 0$, o simplemente, $D_{11} \geq 0$.
- En caso de asociación negativa: $D_{11} = D_{22} \leq 0$ y $D_{12} = D_{21} \geq 0$, o simplemente, $D_{11} \leq 0$.

Por tanto, D_{11} determina muy fácilmente la dirección de la asociación. Sin embargo, en cuanto a la intensidad de la misma, el campo de variación de D_{11} , $[-\frac{N_{12}N_{21}}{n}; \frac{N_{12}N_{21}}{n}]$, depende de los valores de las frecuencias observadas (esto es un problema a la hora de la interpretación) y la máxima intensidad asociativa se da cuando la diagonal descendente o la diagonal ascendente sólo contienen ceros, es decir en caso de asociación perfecta estricta (negativa o positiva).

Para solucionar el problema anterior, se define la Q de Yule como:

$$Q = \frac{nD_{11}}{N_{11}N_{22} - N_{12}N_{21}} = \frac{N_{11}N_{22} - N_{12}N_{21}}{N_{11}N_{22} + N_{12}N_{21}}.$$

El campo de variación de Q es $[-1; 1]$ y:

- En caso de independencia, $Q = 0$.
- En caso de asociación positiva, $Q < 0$.
- En caso de asociación negativa, $Q > 0$.

Por tanto, cuando se sustituyen los datos de la tabla observada en Q , obteniéndose su valor muestral u observado: $\hat{Q} = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}}$, se actúa como sigue:

- Cuando $\hat{Q} = 0$ se dice que hay independencia.
- Cuando $\hat{Q} < 0$ se dice que hay asociación negativa.
- Cuando $\hat{Q} > 0$ se dice que hay asociación positiva.

Lógicamente, a mayor valor absoluto de \hat{Q} mayor intensidad de la asociación.

En el ejemplo utilizado para ilustrar el diseño experimental de tipo 1, el valor observado de Q es $\hat{Q} = 0,83$:

```
YuleQ(tabla_jud)
#> [1] -0.826087
```

A la luz del valor del valor de \hat{Q} se concluye la existencia de una fuerte asociación negativa.

23.4.2. Otras medidas de asociación para tablas 2×2

23.4.2.1. Cuadrado medio de la contingencia de Pearson

La primera medida de asociación que se nos viene a todos a la cabeza es el propio estadístico de contraste χ^2 . Sin embargo, no puede utilizarse como medida de asociación porque es siempre positivo y, sobre todo, porque su valor máximo, $n(k - 1)$, depende del tamaño muestral N y de k , el número más pequeño de filas o columnas. En el caso 2×2 depende únicamente de n porque $k - 1 = 1$. Para eliminar el efecto tamaño muestral, se define el cuadrado medio de la contingencia de Pearson como:

$$\phi^2 = \frac{\chi^2}{n} = \frac{(N_{11}N_{22} - N_{12}N_{21})^2}{N_{1\cdot}N_{2\cdot}N_{\cdot1}N_{\cdot2}},$$

y se estima como

$$\hat{\phi}^2 = \frac{(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1\cdot}n_{2\cdot}n_{\cdot1}n_{\cdot2}}.$$

a partir de la tabla observada.

Su campo de variación es $[0; 1]$, tomando el valor 0 en caso de independencia y 1 cuando hay asociación perfecta y estricta. Cuanto mayor sea el valor del coeficiente, mayor es intensidad de la asociación.

No proporciona la dirección de la asociación, si bien se puede saber por el signo de $n_{11}n_{22} - n_{12}n_{21}$. Otra consideración importante es que, si se codifican los niveles de los factores como $(0;1)$, ϕ^2 coincide con el coeficiente de determinación lineal entre los factores. Por tanto, la asociación que mide es “lineal” (de ahí que su valor suela ser más bajo que el de Q). Su raíz cuadrada es conocida como “la V de Cramer”. En el ejemplo utilizado se tiene que:

```
(Phi(tabla_jud))^2
#> [1] 0.1420701
```

23.4.2.2. Odds ratio o cociente de posibilidades¹¹

Se define como $\alpha = \frac{P_{11}/P_{12}}{P_{21}/P_{22}}$, donde P_{ij} es la probabilidad de que un elemento de la tabla pertenezca al i -ésimo nivel de A y al j -ésimo de B , y se estima como $\hat{\alpha} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$.

Su campo de variación es $[0; \infty)$, asimétrico, y por consiguiente difícil de interpretar. En todo caso:

- Si $\alpha < 1$, la probabilidad (aquí denominada posibilidad) de pertenecer al nivel 1 del factor B es menor en el nivel 1 del factor A que en el 2.
- Si $\alpha = 1$, la probabilidad de pertenecer al nivel 1 del factor B es la misma en ambos niveles del factor A.

¹¹Para no confundirlo con el riesgo relativo.

- Si $\alpha > 1$, la probabilidad pertenecer de al nivel 1 del factor B es mayor en el nivel 1 del factor A que en el 2.

Una posible solución a la dificultad de interpretación es definir $\ln\alpha$, que es una medida simétrica en $(-\infty; +\infty)$. Sin embargo, su interpretación, dada la gran amplitud del campo de variación, continúa siendo muy difusa.

Una ventaja que tiene respecto a α es que no cambia si las filas se convierten en columnas y las columnas en filas.¹² Por ello, la razón de posibilidades se puede utilizar no sólo en estudios retrospectivos, sino también en aquéllos prospectivos y transversales. Finalmente, nótese que (i) la razón de posibilidades y el riesgo relativo (P_1/P_2) se relacionan como sigue: $\alpha = \frac{P_1(1-P_2)}{P_2(1-P_1)}$; y que (ii) ambos son similares cuando la probabilidad de éxito P_i está cerca de cero en ambos grupos.

El código siguiente proporciona el valor muestral de α ($\hat{\alpha}$) en el ejemplo que nos ocupa, así como su intervalo de confianza del 95 %:

```
library("epiR")
epi.2by2(tabla_jud, method = 'cohort.count')$massoc.summary[2,]
#>      var      est      lower      upper
#> 2 Odds ratio 0.0952381 0.01021365 0.8880565
```

23.5. Medidas de asociación en tablas $R \times C$

En caso de rechazo de la hipótesis de independencia en una tabla $R \times C$, se concluye que al menos un nivel de uno de los factores está asociado con uno del otro factor. En ese caso, se utilizarán las medidas de asociación para determinar la intensidad de la misma. Las asociaciones de determinados niveles del factor A con determinados niveles del factor B que llevan al rechazo de la independencia de ambos se denominan **fuentes de asociación**, y se determinan mediante los residuos estandarizados ajustados.

23.5.1. Medidas derivadas del estadístico Chi-cuadrado

Como se avanzó en la Sec. 23.4.2.1, el estadístico χ^2 no puede utilizarse como medida de asociación porque su valor máximo, $n(k - 1)$, siendo n el tamaño muestral y k el número más pequeño de filas o columnas, depende tanto de N como del número de niveles de los factores.

El cuadrado medio de la contingencia, ϕ^2 , elimina el efecto “tamaño muestral”, pero no el efecto “número de niveles de los factores”. Igual le ocurre al coeficiente de contingencia; y a la T de Tschuprow, salvo en las tablas cuadradas. La única medida derivada del estadístico χ^2 que corrige ambos efectos es la V de Cramer:

$$V = \sqrt{\frac{\chi^2}{kn}},$$

¹²Es decir, no es necesario identificar la variable respuesta para utilizar esta medida.

23.5. Medidas de asociación en tablas $R \times C$

403

con $k = \min(R - 1; C - 1)$. Su campo de variación es $[0, 1]$ y alcanza su máximo en caso de asociación perfecta. En tablas cuadradas $V = T$.

En el ejemplo utilizado en la Sec. 23.3.1:

```
CramerV(tabla_ayuso)
#> [1] 0.0590406
```

Aunque se rechaza la hipótesis de independencia, la asociación existente entre la opinión sobre la presidente y la zona geográfica es muy pequeña. Y ello porque, sea cual sea la zona geográfica, aunque hay ligerísimas variaciones, la opinión es muy favorable: para alrededor del 60% es excelente, para la tercera parte muy buena y tan solo para el 5% es mala (alrededor del 4%) o muy mala (apenas el 1%).

23.5.2. Medidas basadas en la reducción proporcional del error: λ de Goodman y Kruskal

Al contrario que las medidas basadas en el estadístico Chi-cuadrado, exigen determinar cuál es el factor explicativo y cuál el factor a explicar. Sea A el factor explicativo y B el factor a explicar: supóngase que se selecciona aleatoriamente uno de los elementos de la tabla. Este elemento pertenecerá a un nivel “ i ” de A y a un nivel “ j ” de B . Supóngase que se quiere predecir el nivel de B al que pertenece, (i) sin utilizar el hecho de saber a qué nivel de A pertenece, y (ii) utilizando dicho hecho. Lógicamente, tanto en el caso (i) como en el (ii) se comete un error ($P(i)$ y $P(ii)$, respectivamente). La probabilidad de error será la misma si los factores son independientes. Sin embargo, si están asociados, el conocimiento del nivel de A al que pertenece el elemento seleccionado ayudará en la predicción del nivel de B al que pertenece (tanto más cuanto más asociados estén los factores) y la probabilidad de error disminuirá respecto al caso (i). La reducción proporcional que se opera en el error es:

$$\lambda = \frac{P(i) - P(ii)}{P(i)},$$

donde $P(i)$ y $P(ii)$ se estiman como sigue: $\hat{P}(i) = n - \max_j n_{.j}$ y $\hat{P}(ii) = n - \sum_{j=1}^C \max_j n_{.j}$.

En el caso en que A sea el factor a explicar, $\hat{P}(i) = n - \max_i n_{i.}$ y $\hat{P}(ii) = n - \sum_{i=1}^R \max_i n_{i..}$.

En caso de no tener claro cuál es el factor a explicar, se utiliza la media agregativa de las dos medidas anteriores:

$$\hat{\lambda} = \frac{\sum_{j=1}^C \max_i n_{i.} - \max_i n_{i.} + \sum_{i=1}^R \max_j n_{.j} - \max_j n_{.j}}{2n - \max_i n_{i.} \max_j n_{.j}}.$$

Su campo de variación es $[0, 1]$. En caso de independencia, $\lambda = 0$. Ahora bien, que $\hat{\lambda} = 0$ no implica necesariamente que A y B tengan que ser independientes, puesto que λ también toma el valor 0 cuando en uno de los niveles del factor a explicar las frecuencias son superiores a

las de los demás niveles, y ello para todos los niveles del factor explicativo, aunque los factores no sean independientes. En caso de asociación, $0 < \lambda \leq 1$, alcanzándose la unidad en caso de asociación perfecta.

Una limitación de λ (además de la anterior y de que exige determinar el factor explicativo y el factor a explicar) es su sensibilidad a totales marginales desequilibrados; en este caso, toma valores anormalmente bajos. Tal es el caso del ejemplo que nos ocupa, donde $\hat{\lambda} = 0$, con factor a explicar la opinión, y, sin embargo, los factores no son independientes. Y es que, sea cual sea la zona geográfica, las frecuencias de la categoría de opinión “excelente” son siempre las más elevadas.

```
Lambda(tabla_ayuso, direction = "row")
#> [1] 0
```

23.5.3. Determinación de las fuentes de asociación

En el caso de tablas $R \times C$, el rechazo la hipótesis de independencia no indica que cada nivel de uno de los factores esté asociado con uno de los niveles del otro factor, como en las tablas 2×2 . Lo que indica es que al menos uno de los niveles de uno de los factores está asociado con un nivel del otro. Por tanto, puede ser, y así es normalmente, que dicho rechazo se deba a que algunos niveles de uno de los factores (incluso sólo uno) están asociados con alguno de los del otro factor. Ya no hay dirección de la asociación. Hay fuentes de asociación.

Para identificar las fuentes de asociación lo lógico es fijarse en cada celda en las diferencias entre la frecuencia observada y la esperada bajo el supuesto de independencia (tales diferencias juegan el papel de un término de error). Pero su interpretación depende del tamaño de la frecuencia esperada, y por ello se estandarizan, es decir, se ponen en relación a la raíz cuadrada de las correspondientes frecuencias esperadas.

Como para decidir si tales diferencias estandarizadas son significativamente grandes (asociación) o no (independencia), se necesita conocer su distribución de probabilidad bajo la hipótesis de independencia, y para ello se dividen por su desviación típica porque de esta manera tienen aproximadamente una distribución $N(0;1)$. Cuando estas diferencias estandarizadas divididas por sus desviaciones típicas se calculan (se estiman) a partir de los resultados observados y dispuestos en la tabla, se denominan residuos estandarizados ajustados (o de Haberman), y son los que se utilizan para identificar las fuentes de asociación.

Por tanto, la estimación de las diferencias (los residuos) son:

$$\hat{R}_{ij} = n_{ij} - \hat{E}_{ij},$$

y los de las diferencias estandarizadas (los residuos estandarizados) vienen dados por:

$$\hat{R}_{ij}(est) = \frac{n_{ij} - \hat{E}_{ij}}{\sqrt{\hat{E}_{ij}}},$$

23.6. Contrastes de independencia en tablas multidimensionales

405

mientras que la siguiente expresión corresponde a las estimaciones las dierencias estandarizadas ajustadas, que no son otras que los denominados residuos estandarizados ajustados:

$$\hat{R}_{ij}(est; adj) = \frac{\hat{R}_{ij}(est)}{\sqrt{\left(1 - \frac{n_{i\cdot}}{N}\right)\left(1 - \frac{n_{\cdot j}}{N}\right)}}.$$

Habrá una fuente de asociación en cada celda $\{i;j\}$ que verifique: $|\hat{R}_{ij}(est; adj)| \geq k$, con $k = 2, 33; 1, 96; 1, 64$ para $\alpha = 0, 01; 0, 05; 0, 10$, respectivamente.

En el ejemplo que nos ocupa, los residuos estandarizados ajustados son:

```
library(questionr)
chisq.residuals(tabla_ayuso, digits = 2, std = TRUE)
#>           opinion
#> zona      n1_nefasto n2_mala n3_buena n4_excelente
#> n1_mad_muni      0.79   -1.04   -3.77      2.41
#> n2_metropol     -0.51    1.74    2.88     -2.78
#> n3_extraradio    -0.45   -0.76    1.59      0.17
```

Asumiendo $\alpha = 0, 05$, las fuentes de asociación son “Madrid municipio-excelente” a costa de “buena”, y “Madrid metropolitano-buena” a costa de “excelente”.

23.6. Contrastes de independencia en tablas multidimensionales

En tablas con más de dos factores (el objetivo aquí es el caso $R \times C \times M$, por simplicidad), no sólo se puede contrastar la hipótesis de independencia global sino que, en caso de ser rechazada, también se pueden contrastar las hipótesis de (i) independencia parcial: dos factores están asociados y el tercero es independiente de ellos, e (ii) independencia condicional: dos de los factores son independientes para cada nivel del tercero pero están asociados con él.

En los tres casos, el estadístico de contraste (contraste aproximado) es

$$\chi^2 = \sum_{i=1}^R \sum_{j=1}^C \sum_{m=1}^M \frac{(N_{ijm} - E_{ijm})^2}{E_{ijm}},$$

con los siguientes grados de libertad (g.l.) y E_{ijm} bajo la correspondiente H_0 :

Independencia global:

$$dl = (R \times C \times M) - (R - 1) - (C - 1) - (M - 1) - 1 \text{ y } E_{ijm} = \frac{N_{i..} N_{.j.} N_{...m}}{n^2}$$

Independencia parcial:

A y B asociados entre sí pero independientes de C :

- $g.l. = (R \times C \times M) - (R \times C - 1) - (M - 1) - 1$ y $E_{ijm} = \frac{N_{i..} N_{.j..}}{n}$

A y C asociados entre sí pero independientes de B :

- $g.l. = (R \times C \times M) - (R \times M - 1) - (C - 1) - 1$ y $E_{ijm} = \frac{N_{i..} N_{.j..}}{n}$

B y C asociados entre sí pero independientes de A :

- $g.l. = (R \times C \times M) - (C \times M - 1) - (R - 1) - 1$ y $E_{ijm} = \frac{N_{.jm} N_{i..}}{n}$

Independencia condicional

A y B son independientes entre sí, pero están asociados con C :

- $g.l. = (R \times C \times M) - (R \times M - 1) - (C \times M - 1) - 1$ y $E_{ijm} = \frac{N_{i..} N_{.jm}}{n}$

A y C son independientes entre sí, pero están asociados con B :

- $g.l. = (R \times C \times M) - (R \times C - 1) - (M \times C - 1) - 1$ y $E_{ijm} = \frac{N_{ij..} N_{.jm}}{n}$

B y C son independientes entre sí, pero están asociados con A :

- $g.l. = (R \times C \times M) - (C \times R - 1) - (M \times R - 1) - 1$ y $E_{ij..} = \frac{N_{i..} N_{.jm}}{n}$

Para calcular el valor muestral del estadístico de contraste para las diferentes hipótesis, basta con sustituir las N_{ij} por las frecuencias observadas (las n_{ij}) y los totales marginales ($N_{i..}$, $N_{.j..}$, etc.) por los totales marginales observados y que figuran en la tabla que surge de los datos en estudio ($n_{i..}$, $n_{.j..}$, $n_{i..}$, etc.).

También son interesantes las relaciones de segundo orden o superior (por ejemplo, si la asociación entre dos de los factores difiere en dirección y/o intensidad para distintos niveles del tercero), pero se estudian mediante **modelos logarítmico lineales**.

Resumen

Las tablas de contingencia analizan la relación entre variables categóricas. Su análisis responde preguntas como: los factores involucrados en la tabla, ¿son independientes o están asociados? Si están asociados, ¿qué niveles de dichos factores son los que están asociados?, ¿cuál es la intensidad de dicha asociación? Se aborda ampliamente el caso de tablas bifactoriales y se proponen test exactos y aproximados para el contraste de la hipótesis de independencia (para tres procedimientos de muestreo diferentes) y una selección de medidas de asociación. Finalmente, se hace una breve incursión en el ámbito de las tablas multidimensionales.

Parte V

Machine learning supervisado

Capítulo 24

Árboles de clasificación y regresión

Ramón A. Carrasco^a e Itzcóatl Bueno^{b,a}

^aUniversidad Complutense de Madrid ^bInstituto Nacional de Estadística

24.1. Introducción

Los árboles de decisión son modelos que se utilizan principalmente para la resolución de problemas de clasificación, en los que hay que predecir las distintas categorías de la variable objetivo o dependiente, aunque también son aplicables a la predicción de valores numéricos de dicha variable objetivo, esto es, como modelos de regresión. De ahí que sean conocidos como árboles de clasificación y regresión (CART, Classification and Regression Trees). Algunos ejemplos de árboles de decisión son:

- **Clasificación:** en la medida que la variable objetivo debe ser categórica se podrían usar por ejemplo para tomar la decisión de qué empleados deberían de promocionar (variable con dos categorías: sí promocionar o no promocionar) en base a sus méritos, capacidades, edad, etc. Otro ejemplo podría ser su uso para decidir si se juega o no un partido de tenis en base a la climatología prevista. Este ejemplo se muestra gráficamente en la Fig. 24.1. En este último caso, el algoritmo que se utilice indicará la decisión a tomar en base a los registros climatológicos de los partidos que ya se hayan jugado. Así, si un determinado día se quiere jugar al tenis, se deberán tomar como variables de entrada las previsiones de Tipo de día (soleado, nublado o lluvioso), la fuerza del Viento y la Humedad. En caso de ser un día nublado, el algoritmo sugerirá que se juegue. En caso de ser soleado, comprobará el nivel de Humedad y, si no es muy elevada, recomendará que se juegue el partido. Lo mismo pasará si la previsión es de lluvia pero la fuerza del Viento prevista no es lo suficientemente intensa como para impedir el normal desarrollo del partido.

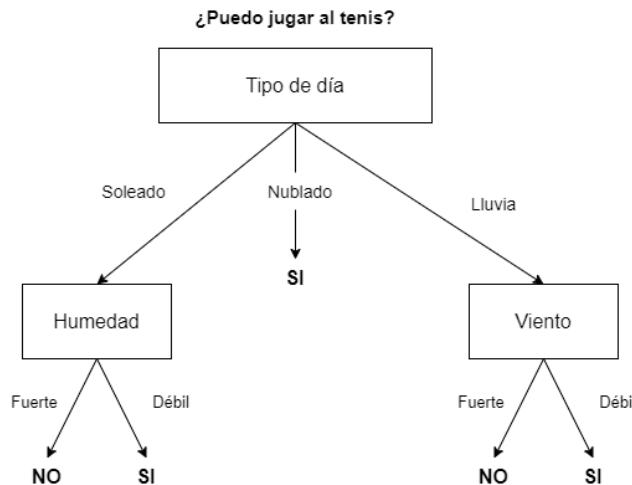


Figura 24.1: Ejemplo de árbol de decisión.

- **Regresión:** siguiendo con el ejemplo del partido de tenis, también se puede aplicar un árbol de decisión para determinar cuántas horas jugar de acuerdo a las condiciones climatológicas. En la Fig. 24.1 se sustituirían la predicciones dicotómicas SI/NO por valores numéricos, como se muestra en la Fig. 24.2. Por ejemplo, el algoritmo puede sugerir jugar 5 horas si el día está soleado pero la Humedad es del 30 % de vapor de agua por m^3 , y 3,5 horas si está soleado pero la Humedad es del 80 %. También puede decidir que si el día está nublado se jueguen 4 horas. O en caso de lluvia, podría decidir que el partido dure 0,75 horas si la fuerza del Viento es de 62km/h y 1,15 horas si es de 27km/h.

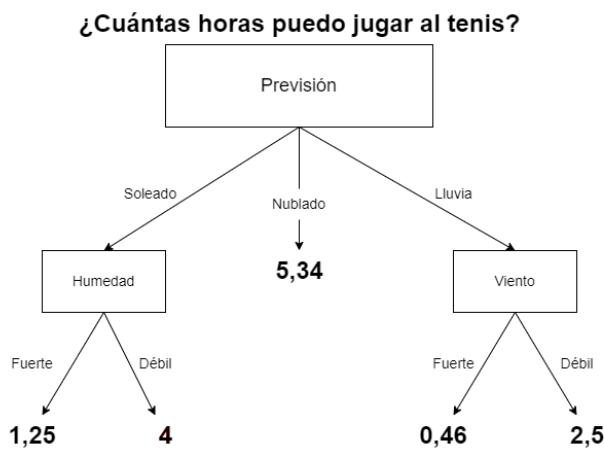


Figura 24.2: Ejemplo de árbol de regresión

Como se ha comentado, CART es un término genérico para describir este tipo de algoritmos de árbol y también un nombre específico para el algoritmo original de (Breiman et al., 1984) de construcción de árboles de clasificación y regresión. Sin embargo, existen otros como el ID3 (Induction Decision Trees), o el C4.5 que está basado en el ID3. En la Tabla 24.1 se muestra una pequeña comparativa de estos tres algoritmos:

Tabla 24.1: Características de los principales algoritmos de árboles de decisión.

Algoritmo	Criterio de división	Tipo de variables input	Estrategia de poda
ID3	Ganancia de información	Solo categóricas	No poda
CART	Índice de Gini	Categóricas y numéricas	Poda basada en el coste de complejidad
C4.5	Ratio de ganancia	Categóricas y numéricas	Poda basada en el error

Los árboles de decisión tienen múltiples ventajas. Entre ellas destacan:

- Son fáciles de entender e interpretar. Su visualización clara permite interpretar la salida del modelo y entender su proceso como un conjunto de condicionantes.
- El mismo algoritmo incorporado en R (CART) es válido tanto para problemas de clasificación como de regresión y, por tanto, la variable objetivo puede ser continua o categórica. Respecto al resto de variables de entrada, las independientes, comentar que puede ser tanto categóricas como numéricas. Al contrario que ocurre con otros algoritmos, este último tipo de variables no requieren la estandarización, puesto que se basa en reglas y no en el cálculo de distancias entre observaciones.
- Tratan mejor que otros algoritmos el problema de la no linealidad.
- Respecto a los datos, hacen un tratamiento automático de valores ausentes (en la mayoría de los árboles de clasificación) y no se ven afectados con las observaciones atípicas.

Sin embargo, también tienen ciertas desventajas:

- Son inestables ya que la inclusión de una nueva observación en la fase de entrenamiento obliga a reconstruirlo, pudiendo modificar la estructura del árbol final.
- No son recomendables en caso de grandes conjuntos de datos, puesto que el modelo entrenado puede estar sobreajustado. Este sobreajuste es el principal problema de los árboles de decisión, ya que modelos demasiados complejos pueden ajustar muy bien los datos observados, pero también pueden cometer muchos errores en la fase de predicción. Cuando esta circunstancia se da, el modelo ha aprendido los datos de entrenamiento pero no la generalidad del problema que es lo que normalmente se pretende. El sobreajuste da lugar también a una varianza elevada.

24.2. Procedimiento con R: la función rpart()

En el paquete **rpart** de **R** se encuentra la función **rpart()** que se utiliza para entrenar un árbol de decisión:

```
rpart(formula, data, ...)
```

- **formula**: Refleja la relación entre la variable dependiente Y y los predictores tal que $Y \sim X_1 + \dots + X_p$.
- **data**: Conjunto de datos con el que entrenar el árbol de acuerdo a la fórmula indicada.

24.3. Árboles de clasificación

Formalmente, un árbol de decisión es un grafo acíclico (un grafo sin ciclos, siendo un ciclo un circuito completo) que se inicia en un **nodo raíz**, el cual se divide en **ramas**, también conocidas como **aristas**. De las ramas salen las **hojas**, también denominadas **nodos**. Estos nodos pueden ser nodos finales o **puntos de decisión** (si de ellos no salen nuevas ramas con nuevos nodos) o no (de ellos salen nuevas ramas con nuevas hojas o nodos) y así hasta que todos los nodos sean puntos de decisión. En el ejemplo de la Fig. 24.1 el nodo raíz es la caja *Tipo de día*. Las ramas o aristas, son sus tres niveles o categorías: *Soleado*, *Nublado* o *Lluvia*. Cada una de estas ramas conecta con una nueva hoja o nodo: *Humedad* o *Viento* en los casos de soleado o lluvia, respectivamente. Sin embargo, en ese ejemplo, *Nublado* representa un nodo terminal, puesto que, llegado a ese punto, la salida que proporcionaría el árbol es “*Jugar al tenis*”. Este proceso se repite utilizando el conjunto de datos disponible en cada hoja, generándose una clasificación final cuando una hoja no tenga ramas nuevas, en cuyo caso recibe la denominación de nodo final. El objetivo es que el árbol sea lo más general y pequeño posible. Esto se consigue seleccionando, en cada paso, la variable que optimice la división de los datos en conjuntos homogéneos, de tal forma que se prediga mejor la clase objetivo.

24.3.1. ¿Cómo se va formando el árbol de clasificación?

Como ya se ha mencionado, en la construcción de un árbol de decisión se va dividiendo en nuevas ramas de forma recursiva, es decir, cada división está condicionada por las anteriores. El objetivo en cada hoja, es encontrar la variable más adecuada para dividir los datos de ese nodo en dos nuevas hojas, de tal forma que el error global entre la clase observada y la predicha por el árbol se minimice. Para la construcción de árboles de clasificación, el algoritmo CART utiliza la medida de impureza de Gini para generar las particiones, mientras que los algoritmos ID3 y C4.5 están basados en las de entropía.

24.3.1.1. Impureza de Gini

La **Impureza de Gini**, utilizada por el algoritmo CART, es una medida de la frecuencia con la que una observación elegida aleatoriamente del conjunto se asignaría a la clase errónea si se

24.3. Árboles de clasificación

413

etiqueta al azar en una de las clases que se consideran. Formalmente, sea X un conjunto de datos con κ clases, y sea p_i la probabilidad de que una observación pertenezca a la clase i . La Impureza de Gini para X se define como:

$$Gini(X) = 1 - \sum_{i=1}^{\kappa} p_i^2 \quad (24.1)$$

Como se ha comentado, a la hora de construir el árbol se selecciona el atributo con la menor impureza de Gini para dividir el conjunto de datos en el nodo en dos. Si un conjunto de datos X se divide en un atributo φ en dos subconjuntos X_1 y X_2 con tamaños n_1 y n_2 , respectivamente, la impureza ponderada de Gini se define como:

$$Gini_{\varphi}(X) = \frac{n_1}{n} Gini(X_1) + \frac{n_2}{n} Gini(X_2) \quad (24.2)$$

En el ejemplo de la Fig. 24.1 considérese la siguiente situación:

Tabla 24.2: Datos para decidir si se juega el partido

Día	Tipo de día	Humedad	Viento	Decisión
1	Soleado	Fuerte	Débil	NO
2	Soleado	Fuerte	Fuerte	NO
3	Lluvia	Fuerte	Débil	SI
4	Nublado	Fuerte	Débil	SI
5	Lluvia	Débil	Débil	SI
6	Lluvia	Débil	Fuerte	NO
7	Soleado	Fuerte	Débil	NO
8	Nublado	Débil	Fuerte	SI
9	Soleado	Débil	Débil	SI
10	Lluvia	Débil	Débil	SI
11	Soleado	Débil	Fuerte	SI
12	Nublado	Fuerte	Fuerte	SI
13	Nublado	Débil	Débil	SI
14	Lluvia	Fuerte	Fuerte	SI
15	Soleado	Fuerte	Fuerte	NO

Para el *Tipo de día*, los datos se agruparían como muestra la Tabla 24.3, permitiendo el cálculo de la impureza de Gini para cada una de sus categorías.

Tabla 24.3: Días que se juega o no de acuerdo al *Tipo de día*

Tipo de día	SI	NO	# observaciones
Soleado	2	4	6
Nublado	4	0	4

Tipo de día	SI	NO	# observaciones
Lluvia	4	1	5

$$Gini(Soleado) = 1 - \left(\frac{2}{6}\right)^2 - \left(\frac{4}{6}\right)^2 = 0,45$$

$$Gini(Nublado) = 1 - \left(\frac{4}{4}\right)^2 = 0$$

$$Gini(Lluvia) = 1 - \left(\frac{4}{5}\right)^2 - \left(\frac{1}{5}\right)^2 = 0,32$$

Ahora, se calcula la suma ponderada de la impureza de Gini para la variable *Tipo de día*:

$$Gini(\text{Tipo de día}) = 0,45 \cdot \left(\frac{6}{15}\right) + 0 \cdot \left(\frac{4}{15}\right) + 0,32 \cdot \left(\frac{5}{15}\right) = 0,29$$

Del mismo modo, se puede calcular la impureza de Gini para el resto de variables. La Tabla 24.4 y la Tabla 24.5 presentan los resultados para *Humedad* y *Viento*, respectivamente.

Tabla 24.4: Impureza de Gini para las categorías de Humedad

Humedad	SI	NO	# observaciones	p_{SI}	p_{NO}	Impureza de Gini
Fuerte	4	4	8	0,50	0,50	0,50
Débil	6	1	7	0,86	0,14	0,76

$$Gini(Humedad) = 0,5 \cdot \left(\frac{8}{15}\right) + 0,76 \cdot \left(\frac{7}{15}\right) = 0,62$$

Tabla 24.5: Impureza de Gini para las categorías de Viento

Viento	SI	NO	# observaciones	p_{SI}	p_{NO}	Impureza de Gini
Fuerte	4	3	7	0,57	0,43	0,49
Débil	6	2	8	0,75	0,25	0,38

$$Gini(Viento) = 0,49 \cdot \left(\frac{7}{15}\right) + 0,38 \cdot \left(\frac{8}{15}\right) = 0,43$$

En la Tabla 24.6 se puede ver que la impureza de Gini para las tres variables incluidas en el ejemplo. La variable con la menor impureza de Gini, el *Tipo de día*, es la elegida para ser el nodo raíz del árbol de clasificación.

24.3. Árboles de clasificación

415

Tabla 24.6: Impureza de Gini para las variables de entrada

Variable	Impureza de Gini
Tipo de día	0,29
Humedad	0,62
Viento	0,43

Al entrenar un árbol de decisión, se repite este proceso, y a la hora de dividir cada nodo, se elige el atributo que proporcione el menor $Gini_\varphi(X)$.

Para obtener la ganancia de información para una variable, las impurezas ponderadas de los nodos hijos se restan de la impureza del nodo padre. La ganancia de Gini para la variable X, $\Delta Gini()$, se calcula así:

$$\Delta Gini(\varphi) = Gini(X) - Gini_\varphi(X) \quad (24.3)$$

Siguiendo el ejemplo del árbol de clasificación, para saber si se puede jugar al tenis o no, se tendría que obtener la impureza de Gini para el nodo *Humedad* o el nodo *Viento*. Repitiendo el proceso anteriormente mostrado, dado que el *Tipo de día* sea soleado, se obtienen los resultados de la Tabla 24.7.

Tabla 24.7: Impureza de Gini para las variables en días soleados

Variable	Impureza de Gini
Humedad	0,00
Viento	0,44

Entonces, la ganancia de Gini para cada variable será:

$$\begin{aligned} \Delta Gini(Humedad) &= 0,45 - 0 = 0,45 \\ \Delta Gini(Viento) &= 0,45 - 0,45 = 0 \end{aligned}$$

Puede observarse que la ganancia de información al dividir por *Humedad* es mayor que al hacerlo por *Viento*, por lo que el árbol se dividirá respecto a la *Humedad*, como se observó en la Fig. 24.1.

24.3.1.2. Entropía

La entropía es un concepto matemático que mide la incertidumbre de una fuente de información, es decir, la varianza en los datos entre diferentes clases. Para cada nodo y su partición, la entropía se calcula como:

$$E = -p_1 \log_2(p_1) - p_2 \log_2(p_2) \quad (24.4)$$

donde p_1 y p_2 representan la probabilidad de pertenecer a cada una de las clases en ese nodo. En teoría de la información, la base logarítmica varía dependiendo de la aplicación, y con ella varía la unidad de medida. En este caso, la ganancia de información se obtiene como:

$$IG = E_{\pi} - E_{\pi+1}, \quad (24.5)$$

donde E_{π} representa la entropía en el nodo padre, mientras que $E_{\pi+1}$ es la entropía en el nodo que resulta de dividir el nodo padre. Entonces, siguiendo el ejemplo basado en los datos de la Tabla 24.3 se tendría que la entropía en origen es:

$$E = -\frac{10}{15} \log_2\left(\frac{10}{15}\right) - \frac{5}{15} \log_2\left(\frac{5}{15}\right) = 0,9183$$

Si se obtiene la entropía para cada variable, se determinará el nodo raíz para aquel que aporte una mayor ganancia de información. En el caso de la variable *Tipo de día* se calcula:

$$\begin{aligned} E_{Soleado} &= -\frac{2}{6} \log_2\left(\frac{2}{6}\right) - \frac{4}{6} \log_2\left(\frac{4}{6}\right) = 0,9183 \\ E_{Nublado} &= -\frac{4}{4} \log_2\left(\frac{4}{4}\right) - \frac{0}{4} \log_2\left(\frac{0}{4}\right) = 0 \\ E_{Lluvia} &= -\frac{4}{5} \log_2\left(\frac{4}{5}\right) - \frac{1}{5} \log_2\left(\frac{1}{5}\right) = 0,7219 \end{aligned}$$

Y por tanto:

$$E_{\text{Tipo de día}} = \frac{6}{15} \cdot 0,9183 + \frac{4}{15} \cdot 0 + \frac{5}{15} \cdot 0,7219 = 0,608$$

Repetiendo el mismo procedimiento con las variables *Viento* y *Humedad* se puede comprobar que $E(Viento) = 0,893$ y $E(Humedad) = 0,809$. A partir de esto se puede obtener la ganancia de información como:

$$IG_{\text{Tipo de día}} = E - E_{\text{Tipo de día}} = 0,918 - 0,608 = 0,31$$

$$IG_{\text{Viento}} = E - E_{\text{Viento}} = 0,918 - 0,893 = 0,025$$

$$IG_{\text{Humedad}} = E - E_{\text{Humedad}} = 0,918 - 0,809 = 0,109$$

Se puede comprobar que la disminución de la aleatoriedad, o la ganancia de información, es mayor para la variable *Tipo de día* y por tanto se elige para ser el nodo raíz. Repitiendo este proceso se va construyendo el árbol hasta alcanzar los nodos terminales.

24.3.2. Sobreajuste

Ya se ha comentado en la Sec. @ref(intro_decisiontree) que una de las principales desventajas de los árboles de decisión es su propensión a sobreajustar el modelo al conjunto de datos de entrenamiento y, por tanto, hay que prestar especial atención a la complejidad del modelo. Basándose en las observaciones utilizadas en la fase de entrenamiento, un árbol de decisión puede extraer los patrones presentes en el conjunto de observaciones de entrenamiento y ser muy preciso en el ajuste de dichas observaciones. Sin embargo, puede ocurrir que el árbol resultante no sea capaz de clasificar correctamente ni el conjunto de validación ni nuevas observaciones. Esta circunstancia puede ocurrir porque haya patrones no observados en los datos de entrenamiento que el modelo no es capaz de detectar, o porque la división de los datos entre entrenamiento y validación no se realizó correctamente siendo los datos de entrenamiento no representativos del conjunto de datos completo. Intentando que el árbol entrenado tenga la capacidad de aprender patrones muy complejos, se puede producir este sobreajuste materializado con árboles muy profundos. La forma de evitar el sobreajuste es controlar el crecimiento del árbol para evitar que se vuelva excesivamente complejo.

24.3.3. ¿Cuánto debe crecer un árbol de clasificación?

En cada paso de construcción del árbol se determina la variable óptima para realizar la división de las observaciones de un nodo padre en sus nodos hijos. La pregunta es: ¿cuándo se detiene?, ¿cuál es el criterio de parada? Por ejemplo, se puede utilizar como criterio de parada que el árbol alcance un tamaño o profundidad determinado, para que no sea excesivamente complejo y así no tengan lugar las consecuencias derivadas del sobreajuste.

En consecuencia, se debe llegar a un equilibrio entre la profundidad y complejidad del árbol para optimizar la predicción de futuras observaciones. Este equilibrio se puede lograr siguiendo alguno de los siguientes enfoques: la parada temprana o la poda.

24.3.3.1. La parada temprana

La parada temprana restringe el crecimiento del árbol, tanto de clasificación como de regresión, de forma explícita. Existen distintas maneras de establecer esta restricción al árbol, pero dos de las técnicas más populares son las de restringir la profundidad a un cierto nivel o la de establecer un número mínimo de observaciones permitidas en un nodo terminal. En el primer caso, el árbol deja de dividirse al llegar a cierta profundidad. Así, cuanto menos profundo sea el árbol, menos variación habrá en las predicciones que proporcione. Sin embargo, existe el riesgo de introducir mucho sesgo al modelo al no ser capaz de captar interacciones y patrones complejos en los datos. El segundo enfoque lo que provoca es que no se dividan nodos intermedios con pocas observaciones. En el caso extremo, si se permite que un nodo terminal sólo contuviese una observación esta actuaría como predicción. De este modo, los resultados probablemente no serían generalizables y tendrían mucha variabilidad. En el otro extremo, si se exigen un gran número de observaciones en el nodo terminal se reduce el número de divisiones y, por lo tanto, se reduce la varianza.

24.3.3.2. La poda

El otro enfoque es el de la poda que consiste en construir un árbol muy profundo y complejo y después podarlo para encontrar el subárbol óptimo. Este subárbol se obtiene utilizando un hiperparámetro de complejidad (ζ) que penaliza la función objetivo de la partición por el número de nodos terminales del árbol (τ), es decir, se busca minimizar:

$$R_\zeta(\tau) = R(\tau) + \zeta|\tau| \quad (24.6)$$

Donde $R(\tau)$ es el error total de entrenamiento de los nodos, $|\tau|$ es el número total de nodos y ζ es el hiperparámetro de complejidad. A medida que ζ aumenta, más ramas del árbol son podadas. Mientras que a valores más bajos, los modelos producidos son más complejos y en consecuencia más grandes. En conclusión, a medida que un árbol crece, el error de entrenamiento debe tener una reducción mayor que la penalización por la complejidad.

24.3.4. Ejemplo: Árbol de clasificación para determinar la intención de compra

A continuación se describe el caso que se va a resolver mediante modelos de clasificación tanto en este como en los siguientes capítulos. Existen diversas aserciones para definir Comercio Electrónico (CE). Entre ellas, la Organización para la Cooperación y el Desarrollo Económico (OCDE) lo define como el proceso de compra, venta o intercambio de bienes, servicios e información a través de redes de comunicación, comúnmente Internet. La clasificación más básica del CE se hace en base al tipo de entes que se relacionan: empresas (businesses, B), consumidores (consumers, C) y entes públicos (governments, G). De esta forma, una empresa de CE convencional suele ser B2B si vende a otras empresas, B2G si su relación comercial es con administraciones o B2C si vende a consumidores finales.

En este caso, se puede considerar que la empresa “Beauty eSheep” lleva a cabo un CE de tipo B2C. Su producto estrella es una crema hidratante unisex, denominada internamente como “Crema Luxury”, con mucho éxito entre su clientela. A partir de este producto inicial, la empresa ha ido ofreciendo un catálogo de productos tanto de belleza como de bienestar y salud.

Hace tiempo la empresa instauró una estrategia relacional, centrada en el cliente, de tal manera que ha ido recabando diversos datos sobre los mismos, incluidas las distintas compras que han realizado.

Basándose en los datos recopilados para cada cliente, la empresa quiere realizar una campaña para impulsar la venta de tensiómetros digitales. La empresa tiene acceso a un stock muy flexible en fechas de envío de estos productos y el precio de los tensiómetros es muy bueno, por lo que se espera una buena rentabilidad en su venta.

Por tanto, en este proyecto hay que identificar el público objetivo susceptible de comprar dicho producto para ofrecérselo a través de la plataforma de CE de la compañía, SMS y/o webmail durante el periodo que dura la campaña.

La tabla con los datos integrados a nivel de cliente, incluyendo el consumo de los distintos productos de la empresa, es **dp_ENTR**, incluida en el paquete **CDR**, y que se resume en la

24.3. Árboles de clasificación

419

Tabla 24.8. Este ejemplo se va a replicar en el resto de capítulos de machine learning supervisado para clasificación.

Tabla 24.8: Descripción de las variables del conjunto de datos **dp_entr**.

COLUMNA	TIPO	DESCRIPCIÓN
CLS_PRO_pro13	Factor	Clase objetivo, es un indicador de si el cliente es consumidor de ese producto “Tensiómetro Digital” (‘S’) o no (‘N’)
ind_pro11	Factor	Indicador de si el cliente es consumidor del producto “Fragancia Luxury” (‘S’) o no (‘N’)
ind_pro12	Factor	Indicador de si el cliente es consumidor del producto “Depiladora Eléctrica” (‘S’) o no (‘N’)
ind_pro14	Factor	Indicador de si el cliente es consumidor del producto “Crema Luxury” (‘S’) o no (‘N’)
ind_pro15	Factor	Indicador de si el cliente es consumidor del producto “Smartwatch Fitness” (‘S’) o no (‘N’)
ind_pro16	Factor	Indicador de si el cliente es consumidor del producto “Kit Pesas Inteligentes” (‘S’) o no (‘N’)
ind_pro17	Factor	Indicador de si el cliente es consumidor del producto “Estimulador Muscular” (‘S’) o no (‘N’)
importe_pro11	Doble	Importe neto global gastado por el cliente en ese producto en euros
importe_pro12	Doble	Importe neto global gastado por el cliente en ese producto en euros
importe_pro14	Doble	Importe neto global gastado por el cliente en ese producto en euros
importe_pro15	Doble	Importe neto global gastado por el cliente en ese producto en euros
importe_pro16	Doble	Importe neto global gastado por el cliente en ese producto en euros
importe_pro17	Doble	Importe neto global gastado por el cliente en ese producto en euros
edad	Entero	Edad del cliente
tamano_fam	Entero	Número de miembros de la unidad familiar a la que pertenece el cliente incluyéndolo a él mismo
anos_exp	Entero	Años de trabajo del cliente

COLUMNA	TIPO	DESCRIPCIÓN
ingresos_ano	Doble	Ingresos anuales del cliente en euros
des_nivel_edu	Factor	Descripción del nivel de educación del cliente

Se construye un árbol de clasificación utilizando el conjunto de entrenamiento, como se ha comentado, sin transformar (en su escala original) mediante el algoritmo CART implementado en el paquete `rpart` con Árboles de Regresión y Partición Recursiva (Recursive Partitioning and Regression Trees, RPART) que se puede usar tanto para regresión como para clasificación.

```
library("CDR")
library("reshape")
library("caret")
library("rpart")
library("rpart.plot")
library("ggplot2")

data("dp_entr")
head(dp_entr)

  ind_pro11 ind_pro12 ind_pro14 ind_pro15 ind_pro16 ind_pro17 importe_pro11
1      S      N      S      S      S      N      157
497    N      N      S      N      S      N       0
265    N      N      S      S      S      S       0
534    N      S      S      N      N      N       0
415    N      S      S      N      S      N       0
298    S      N      S      N      N      N      115
  importe_pro12 importe_pro14 importe_pro15 importe_pro16 importe_pro17 edad
1          0        40      200      180       0     49
497        0       240       0      180       0     38
265        0       425      200      180     300     61
534      120       60       0       0       0     47
415      120      133       0      180       0     34
298        0       220       0       0       0     43
  tamano_fam anos_exp ingresos_ano des_nivel_edu CLS_PRO_pro13
1         4      24    30000      MEDIO       S
497       2      12    53000      MEDIO       N
265       4      37   172000    BASICO       S
534       3      21    38000      MEDIO       N
415       1      10    38000    BASICO       N
298       2      18    60000      ALTO       N

trControl <- trainControl(
  method = "cv",
  number = 10,
  classProbs = TRUE,
  summaryFunction = twoClassSummary
)
```

24.3. Árboles de clasificación

421

En primer lugar, se carga la librería necesaria para entrenar el modelo, así como los datos de compras de los clientes. En este caso se usa el método de remuestreo de validación cruzada con 10 folds, visto en el Cap. 9. A continuación, se determina la semilla aleatoria para que los resultados sean replicables, y se entrena el modelo.

```
# se fija una semilla aleatoria
set.seed(101)

# se entrena el modelo
model <- train(CLSPRO_pro13 ~ ., # . equivale a incluir todas las variables
                data=dp_entr,
                method="rpart",
                metric="ROC",
                trControl=trControl)
```

```
model
CART

558 samples
17 predictor
 2 classes: 'S', 'N'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 502, 502, 502, 503, 503, 502, ...
Resampling results across tuning parameters:

  cp        ROC      Sens      Spec
0.05017921 0.8172123 0.9214286 0.7026455
0.10394265 0.7559406 0.8386243 0.6914021
0.51971326 0.6347222 0.8564815 0.4129630

ROC was used to select the optimal model using the largest value.
The final value used for the model was cp = 0.05017921.
```

```
ggplot(melt(model$resample[,-4]), aes(x = variable, y = value, fill=variable)) +
  geom_boxplot(show.legend=FALSE) +
  xlab(NULL) + ylab(NULL)
```

Los resultados de validación cruzada quedan recogidos en los boxplot, por lo que se puede ver los valores entre los que oscilan las principales medidas en los 10 folds del proceso de validación. Estas medidas (ROC, sensibilidad y especificidad) se definieron en el Cap. 9, y en el caso de árboles de clasificación se utilizan para medir la precisión del modelo. A continuación se

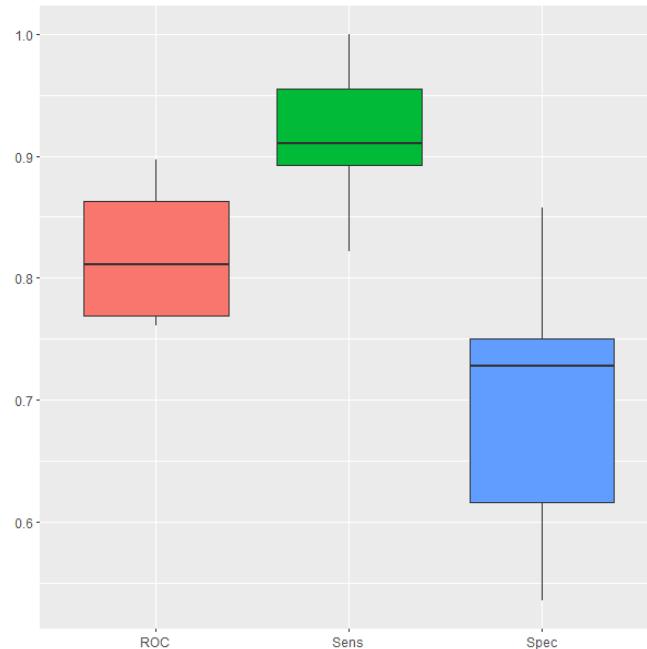


Figura 24.3: Resultados del modelo durante la validación cruzada.

muestra el árbol generado. Se puede observar que este árbol es muy sencillo, y por tanto es fácil obtener su interpretación. En primer lugar decide si un cliente que compra el *smartchwatch fitness* comprará el nuevo producto. En caso de no comprar el *smartchwatch fitness* (No a `ind_pro15S=1`), pero sí compra la *depiladora eléctrica* (Yes a `ind_pro12S=1`) sí comprará el *tensiómetro digital*. Si no compra ninguno de esos dos productos no comprará el nuevo producto.

```
# Gráfico del árbol obtenido
rpart.plot(model$finalModel)
```

Este modelo se puede mejorar ajustando automáticamente el hiperparámetro incluido en `rpart` para el entrenamiento de árboles de decisión. Los hiperparámetros son los valores utilizadas durante el proceso de entrenamiento en la configuración del modelo. Por consiguiente, primero es necesario conocer el hiperparámetro a optimizar en el algoritmo implementado en R que estemos usando. Esto se consigue mediante la siguiente instrucción incluida en el paquete `caret`:

```
modelLookup("rpart")
model parameter          label forReg forClass probModel
1 rpart      cp Complexity Parameter   TRUE    TRUE    TRUE
```

El hiperparámetro a optimizar es la complejidad del árbol, `cp`, que es un hiperparámetro que se aplica en la fase de parada durante la construcción del árbol. Según se ha comentado, esta

24.3. Árboles de clasificación

423

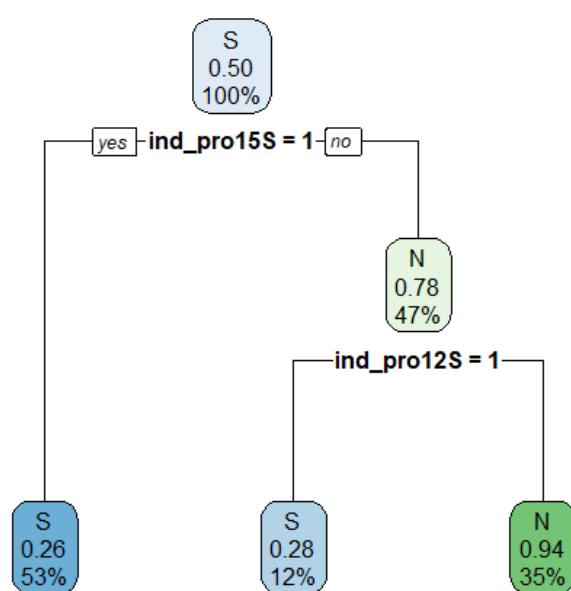


Figura 24.4: Árbol de clasificación sin ajuste automático de hiperparámetros.

fase tiene como función principal evitar desarrollar divisiones que no valgan la pena. Se puede entender `cp` como la mejora mínima necesaria en cada nodo del modelo. Es necesario definir los valores de `cp` que se quieren evaluar con el objetivo de obtener su valor óptimo.

```
# Se especifica un rango de valores típicos para el hiperparámetro
tuneGrid <- expand.grid(cp = seq(0.01,0.05,0.01))
```

```
# se entrena el modelo
set.seed(101)

model <- train(CLSPRO_pro13 ~ .,
                data=dp_entr,
                method="rpart",
                metric="ROC",
                trControl=trControl,
                tuneGrid=tuneGrid)
```

```
model
CART

558 samples
 17 predictor
  2 classes: 'S', 'N'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 502, 502, 502, 503, 503, 502, ...
Resampling results across tuning parameters:
```

cp	ROC	Sens	Spec
0.01	0.8962254	0.8678571	0.8167989
0.02	0.8663454	0.9000000	0.7667989
0.03	0.8458097	0.9392857	0.7310847
0.04	0.8449381	0.9214286	0.7383598
0.05	0.8172123	0.9214286	0.7026455

```
ROC was used to select the optimal model using the largest value.
The final value used for the model was cp = 0.01.
```

De forma automática se construyen diversos árboles para cada uno de los valores explicitados del parámetro `cp`. Para cada uno de esos árboles se obtienen las correspondientes métricas de precisión: el área bajo la curva (denotada como ROC, por las siglas en inglés de *Receiver Operating Characteristic*), sensibilidad (Sens) y especificidad (Spec), todas ellas definidas en el Cap. 9. El valor ROC es el utilizado para la elección del valor óptimo de `cp`, por lo que se determina que finalmente el óptimo es $cp = 0,01$ al maximizar el valor ROC alcanzando un 89,6 %. Por tanto, ajustando el hiperparámetro se ha aumentado la precisión del modelo en casi un 8 % respecto al 81,7 % que tenía el modelo sin ajustar automáticamente el valor de `cp`.

24.3. Árboles de clasificación

425

En la Fig. 24.5 se puede ver el rendimiento de cada una de las métricas del árbol entrenado utilizando validación cruzada. Dicha figura se obtiene con la siguiente instrucción:

```
ggplot(melt(model$resample[,-4]), aes(x = variable, y = value, fill=variable)) +
  geom_violin(show.legend=FALSE) +
  xlab(NULL) +
  ylab(NULL)
```

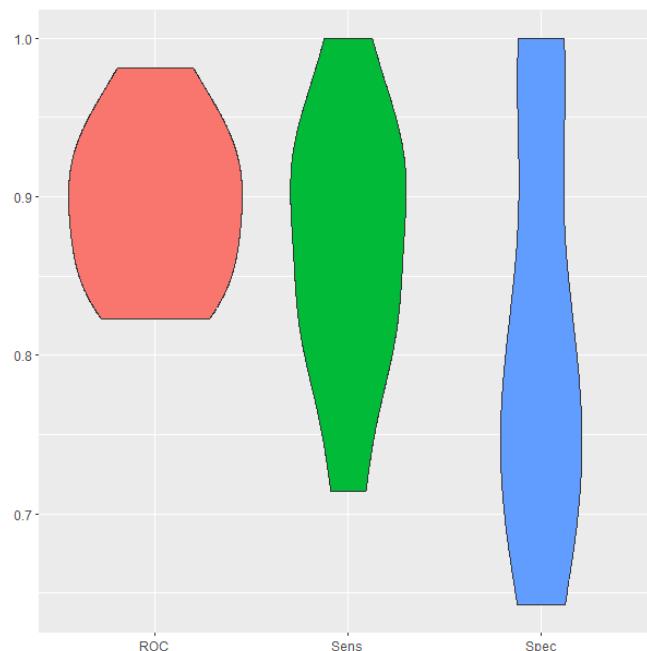


Figura 24.5: Resultados del modelo con ajuste automático durante la validación cruzada

En la Fig. 24.6 se muestra el árbol generado. Dicha visualización se ha obtenido con el siguiente código:

```
# Gráfico del árbol obtenido
rpart.plot(model$finalModel)
```

Con el objetivo de aumentar la generalidad del árbol y facilitar su interpretación, se procede a reducir su tamaño podándolo. Para ello se establece que un nodo terminal tenga como mínimo 50 observaciones, dando lugar al árbol que se muestra en la Fig. 24.7.

```
set.seed(101)
prunedtree <- rpart(CLSPRO_pro13 ~ ., data=dp_entr,
                      cp= 0.01, control = rpart.control(minbucket = 50))
```

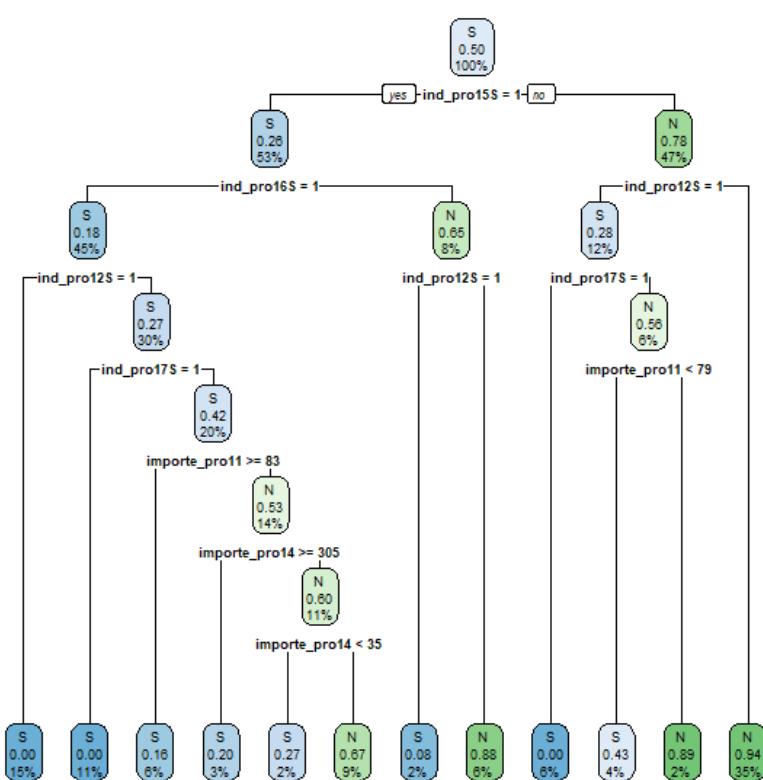


Figura 24.6: Árbol de clasificación con ajuste automático.

24.4. Árboles de regresión

427

```
rpart.plot(prunedtree)
```

El árbol ha reducido el número de nodos terminales, en tres de ellos el árbol predice que un cliente comprará el nuevo producto si:

1. Compra el *smartwatch fitness* (*ind_pro15* = S - Yes) y la *depiladora eléctrica* (*ind_pro12* = S - Yes).
2. Compra el *smartwatch fitness* (*ind_pro15* = S - Yes) y el *estimulador muscular* (*ind_pro17* = S - Yes), pero no la *depiladora eléctrica* (*ind_pro12* = S - No).
3. No compra el *smartwatch fitness* (*ind_pro15* = S - No), pero si la *depiladora eléctrica* (*ind_pro12* = S - Yes).

Sin embargo, dos nodos terminales predicen que el cliente no compra el nuevo producto si:

1. Compra el *smartwatch fitness* (*ind_pro15* = S - Yes), pero no la *depiladora eléctrica* (*ind_pro12* = S - No) ni el *estimulador muscular* (*ind_pro17* = S - No).
2. No compra el *smartwatch fitness* (*ind_pro15* = S - No) ni la *depiladora eléctrica* (*ind_pro12* = S - No).

24.4. Árboles de regresión

Como se ha comentado, los árboles de decisión también pueden ser usados para resolver problemas de regresión. En este caso, la idea es que la predicción dada en cada hoja sea un valor numérico en lugar de un valor de una categoría. En la Tabla 24.9 se muestran los datos para un problema de regresión equivalente al presentado en secciones anteriores para clasificación. Como ya se ha mencionado, la variable objetivo (*Horas jugadas*) ahora es continua en lugar de categórica, como ocurría en el ejemplo anterior con la variable *Decisión*.

Tabla 24.9: Datos de Horas jugadas dada la climatología del día

Día	Tipo de día	Humedad	Viento	Horas jugadas
1	Soleado	Fuerte	Débil	2,3
2	Soleado	Fuerte	Fuerte	1,5
3	Lluvia	Fuerte	Débil	1,3
4	Nublado	Fuerte	Débil	2,4
5	Lluvia	Débil	Débil	1,9
6	Lluvia	Débil	Fuerte	2,4
7	Soleado	Fuerte	Débil	2,3
8	Nublado	Débil	Fuerte	2,2

Día	Tipo de día	Humedad	Viento	Horas jugadas
9	Soleado	Débil	Débil	1,3
10	Lluvia	Débil	Débil	1,8
11	Soleado	Débil	Fuerte	1,2
12	Nublado	Fuerte	Fuerte	2,9
13	Nublado	Débil	Débil	2,2
14	Lluvia	Fuerte	Fuerte	1,5
15	Soleado	Fuerte	Fuerte	1,5

Se pueden calcular medidas descriptivas de la variable respuesta, *Horas jugadas*, como la media, varianza, desviación típica y coeficiente de variación siendo estas:

$$\bar{x}_{\text{Horas jugadas}} = \frac{1}{n} \sum x = 1,91 \quad (24.7)$$

$$\sigma_{\text{Horas jugadas}}^2 = \frac{\sum (x - \bar{x})^2}{n} = 0,25 \quad (24.8)$$

$$\sigma_{\text{Horas jugadas}} = \sqrt{\sigma^2} = 0,50 \quad (24.9)$$

$$CV_{\text{Horas jugadas}} = \frac{\sigma}{\bar{x}} = 0,26 \quad (24.10)$$

24.4.1. ¿Cómo se va formando el árbol de regresión?

Mientras que en los árboles de clasificación se utilizaba la entropía o la impureza de Gini para medir la homogeneidad de un nodo, en los árboles de regresión se utiliza como métrica la desviación típica (σ). Por tanto, cuando se selecciona una variable para hacer la división, se calcula la desviación típica para cada una de las ramas, y se obtiene una media ponderada en función del número de elementos de cada una de ellas. Esta media ponderada se calcula del siguiente modo:

$$\sigma_X = \sum_{r \in X} P(r) \cdot \sigma_r \quad (24.11)$$

Donde X es la variable de la cual se quiere obtener la desviación típica y r son las ramas que crecen desde este nodo. Para llevar a cabo la división, en primer lugar se debe calcular para cada nodo su desviación típica. A continuación, se seleccionan posibles variables para hacer la división y se obtiene su desviación típica. Para cada una de estas variables se calcula el decremento de la desviación, y se selecciona aquel que introduzca la mayor reducción.

Por ejemplo, para los datos mostrados en la Tabla 24.9, la desviación típica es 0,50 Horas jugadas, como se calculó en la ecuación (24.9), y el árbol se construiría como se muestra a

24.4. Árboles de regresión

429

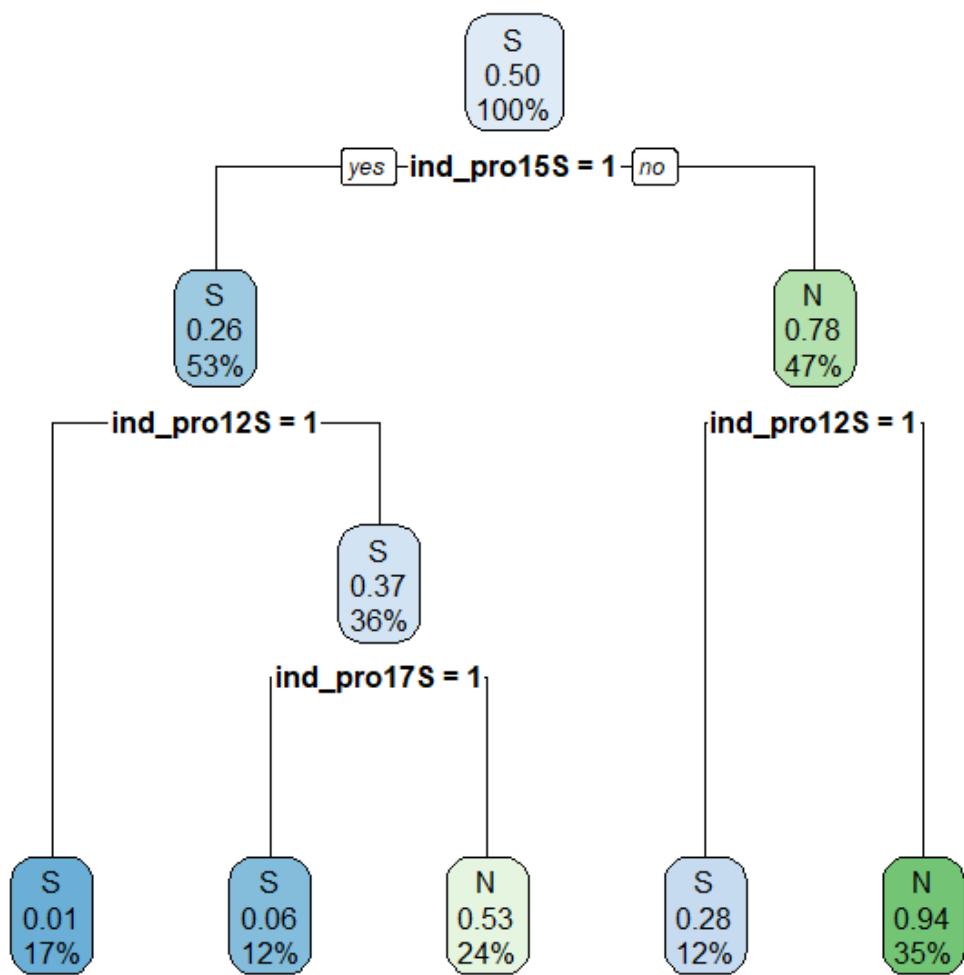


Figura 24.7: Árbol de clasificación con ajuste automático y podado.

continuación. En primer lugar, se selecciona *Tipo de día*, *Humedad* y *Viento* como candidatos a nodo raíz y se obtiene su desviación típica tal que:

Tabla 24.10: Desviación típica en las ramas de la variable *Tipo de día*

Tipo de día	# observaciones	$\sigma_{\text{Horas jugadas}}$
Soleado	6	0,45
Nublado	4	0,29
Lluvia	5	0,38

Tabla 24.11: Desviación típica en las ramas de la variable *Humedad*

Humedad	# observaciones	$\sigma_{\text{Horas jugadas}}$
Fuerte	8	0,55
Débil	7	0,43

Tabla 24.12: Desviación típica en las ramas de la variable *Viento*

Viento	# observaciones	$\sigma_{\text{Horas jugadas}}$
Fuerte	7	0,57
Débil	8	0,42

A partir de las desviaciones típicas en las ramas de cada variable, se puede obtener la desviación típica de cada variable de acuerdo a la ecuación (24.11). Además, se calcula la reducción de desviación típica como la diferencia entre la desviación de la variable respuesta y la desviación si se divide el conjunto de datos en base a alguna de las variables. Tanto la desviación típica de cada variable como el decremento en la desviación que producen se muestran en la Tabla 24.13.

Tabla 24.13: Desviación típica y decremento de desviación de cada variable

Variable	$\sigma_{\text{Horas jugadas}}$	Decremento
Tipo de día	0,38	0,12
Humedad	0,49	0,00
Viento	0,49	0,00

Dado que la variable *Tipo de día* es la que produce una mayor reducción en la desviación típica, resulta elegida como nodo raíz. El árbol seguiría creciendo repitiendo este proceso. Por ejemplo, se muestra cuál sería la siguiente división desde la rama soleado que crece desde nodo *Tipo de día*.

24.4. Árboles de regresión

431

Tabla 24.14: Desviación típica en las ramas de la variable *Humedad* en días soleados

Humedad	# observaciones	$\sigma_{\text{Horas jugadas}}$
Fuerte	4	0,4
Débil	2	0,05

Tabla 24.15: Desviación típica en las ramas de la variable *Viento*

Viento	# observaciones	$\sigma_{\text{Horas jugadas}}$
Fuerte	3	0,14
Débil	3	0,47

En la Tabla 24.16 se muestra la desviación típica para cada variable así como la reducción de desviación que produce. Por tanto, la siguiente división se realizaría con la variable *Humedad*.

Tabla 24.16: Desviación típica y decremento de desviación de cada variable en la rama soleado

Variable	$\sigma_{\text{Horas jugadas}}$	Decremento
Humedad	0,28	0,17
Viento	0,31	0,14

24.4.2. ¿Cuánto debe crecer el árbol de regresión?

Como ocurría en los árboles de clasificación, es necesario establecer reglas que pongan fin al proceso de crecimiento del árbol. Además de los criterios de parada que se utilizan en árboles de clasificación (número de elementos mínimos en un nodo y nivel máximo del árbol), en árboles de regresión se detiene su crecimiento estableciendo un *threshold* (umbral de decisión) sobre el coeficiente de variación del nodo. En el ejemplo expuesto sobre Horas jugadas, se puede ver qué nodos podrían seguir creciendo si se establece que el árbol continúe creciendo en nodos con un coeficiente de variación de un 15 % o más, y que tenga al menos 5 observaciones en el nodo.

Tabla 24.17: Medidas para decidir si el árbol sigue creciendo

Nodo padre	Rama	CV en nodo hijo	# observaciones
Tipo de día	Nublado	11,80 %	4
Tipo de día	Lluvia	21,14 %	5
Humedad	Fuerte	21,04 %	4
Humedad	Débil	4,04 %	2

En este ejemplo, el árbol seguiría creciendo por la rama *Lluvia* donde habría que seleccionar la siguiente variable de división. En el resto de ramas, no se supera el número mínimo establecido de observaciones en el nodo, y en ocasiones tampoco se alcanza el coeficiente de variación mínimo. Por otra parte, en los árboles de regresión la poda se lleva a cabo del mismo modo que para árboles de clasificación. En la ecuación (24.6) se mediría el error de entrenamiento a través de la suma de los cuadrados de los errores (en inglés *Sum of Squared Estimate of Errors*, SSE), es decir:

$$SSE_\zeta(\tau) = SSE(\tau) + \zeta|\tau| \quad (24.12)$$

24.4.3. Árbol de regresión para estimar el número de días de hospitalización

En este ejemplo se utilizan los datos `cleveland`, contenidos en el paquete CDR, y que han sido utilizados en el Cap. 16 para estimar la variable *dhosp*. El conjunto de datos contiene información sobre pacientes que llegan a un hospital con dolor de pecho y de los cuales se han recogido distintas características. Se pretende predecir el número de días de hospitalización que necesitará un paciente en base al resto de características observadas: si el paciente está diagnosticado de accidente coronario, su edad, su sexo, el tipo de dolor que padece y la depresión en el segmento ST inducida por ejercicio en relación al reposo.

```
# se cargan los datos
data("cleveland")

# se entrena el modelo
set.seed(101)
model <- rpart(dhosp ~ diag + edad + sexo + tdolor + dep,
                data=cleveland, method="anova")
```

```
model$cptable

      CP nsplit rel error     xerror      xstd
1 0.37275022    0 1.0000000 1.0128283 0.09213359
2 0.01674747    1 0.6272498 0.6427926 0.06048143
3 0.01132433    4 0.5770074 0.6788431 0.06681871
4 0.01007684    6 0.5543587 0.6825792 0.06505426
5 0.01000000    7 0.5442819 0.6843192 0.06514439
```

Se observa que para valores muy altos del hiperparámetro de complejidad, el SSE es muy elevado. Esto es, produce modelos muy sencillos pero con nula potencia predictiva. En el otro extremo, para $\zeta = 0,01$ el SSE se minimiza hasta llegar a $SSE = 0,54$, por lo que el árbol se poda de acuerdo a la ecuación (24.12) con dicho valor de ζ . El resultado del modelo se muestra en el árbol de la Fig. 24.8. La interpretación de este árbol sería:

24.4. Árboles de regresión

433

1. Si el paciente no tiene diagnóstico de accidente coronario, solo necesitará un día de hospitalización.
2. En el caso de tener este diagnóstico, y una depresión mayor o igual a dos en el segmento ST inducida por ejercicio en relación al reposo, necesitará 2,8 días de hospitalización.
3. En un último ejemplo, si la depresión en el segmento ST inducida por ejercicio en relación al reposo está entre 0,35 y 2 entonces el paciente necesitará 3,8 días de hospitalización. Si por el contrario, la depresión en el segmento ST inducida por ejercicio en relación al reposo es menor a 0,35, el número de días de hospitalización depende del sexo del paciente: los hombres necesitarán 3,2 días y las mujeres tan solo 1,9 días.

```
# se pinta el árbol obtenido
rpart.plot(model)
```

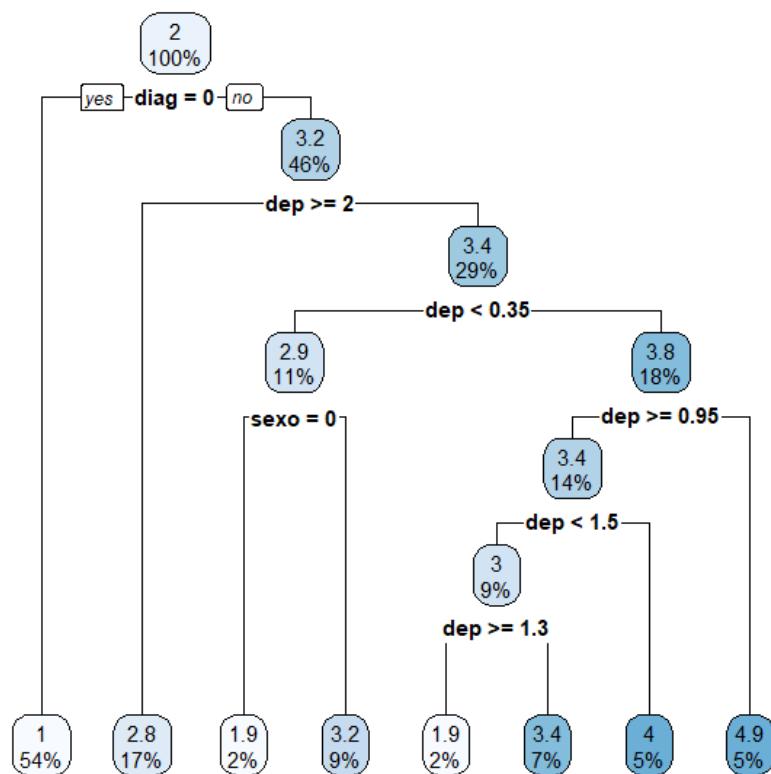


Figura 24.8: Árbol de regresión para predecir el número de días de hospitalización.

24.4.4. Árbol de regresión para la predicción del precio unitario de la vivienda en Madrid

En este ejemplo se va a entrenar un árbol de regresión para predecir el precio unitario de la vivienda en Madrid. Para ello, se van a utilizar los datos de viviendas a la venta en Madrid publicadas en Idealista durante el año 2018. Estos datos están incluidos en el paquete `idealista18`. Para facilitar la interpretación del modelo, sólo se van a utilizar 8 de las variables incluidas en el conjunto de datos: superficie construida, número de dormitorios, número de baños, si tiene terraza, si tiene ascensor, si el precio incluye el parking, distancia al centro de Madrid y distancia a una parada de metro.

```
library("idealista18")
data("Madrid_Sale")

Madrid_Sale <- Madrid_Sale |>
  dplyr::select(UNITPRICE, CONSTRUCTEDAREA, ROOMNUMBER, BATHNUMBER,
    HASTERRACE, HASLIFT, ISPARKINGSPACEINCLUDEDINPRICE,
    DISTANCE_TO_CITY_CENTER, DISTANCE_TO_METRO)

head(Madrid_Sale)

  UNITPRICE CONSTRUCTEDAREA ROOMNUMBER BATHNUMBER HASTERRACE HASLIFT
1 2680.851        47         1         1         0         1
2 4351.852        54         1         1         0         0
3 4973.333        75         2         1         0         0
4 5916.667        48         1         1         0         1
5 4560.000        50         0         1         0         0
6 3921.260       127         3         2         0         1
  ISPARKINGSPACEINCLUDEDINPRICE DISTANCE_TO_CITY_CENTER DISTANCE_TO_METRO
1                         0             8.0584293          0.8720746
2                         0             0.8763693          0.1163821
3                         0             0.9074793          0.1391088
4                         0             0.8454622          0.1442990
5                         0             1.2502313          0.3370982
6                         0             0.5417727          0.1614363

# Se entrena el modelo
library("rpart")
set.seed(101)
model <- rpart(UNITPRICE ~ ., Madrid_Sale, method = "anova")
```

Como en el ejemplo anterior, para $\zeta = 0,01$ el SSE se minimiza hasta llegar a $SSE = 0,56$, por lo que el árbol se poda de acuerdo a la ecuación (24.12) con dicho valor de ζ . El resultado del modelo se muestra en el árbol de la Fig. 24.9. La interpretación de este árbol sería:

1. Si una vivienda con ascensor se encuentra a menos de 3,2km del centro de Madrid y a menos de 0,46km de una estación de Metro, el precio unitario predicho para esa vivienda será de $5.248\text{€}/m^2$.

24.4. Árboles de regresión

435

2. Si una vivienda se encuentra a más de 3,2km del centro de Madrid y no tiene ascensor, el precio unitario predicho será de $2.160\text{€}/m^2$.

3. Si una vivienda se encuentra a menos de 3,2km del centro de Madrid y a más de 0,46km de una estación de Metro, el precio unitario predicho para esa vivienda será de $3.873\text{€}/m^2$.

```
# se pinta el árbol obtenido
rpart.plot(model)
```

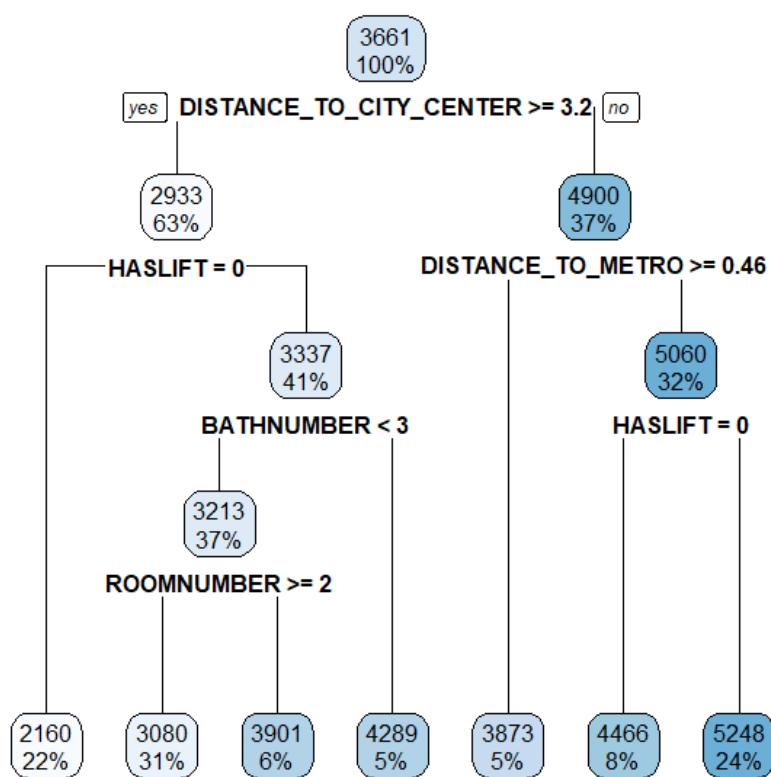


Figura 24.9: Árbol de regresión para predecir el precio unitario de las viviendas en Madrid.

Resumen

En este capítulo se introduce al lector en los árboles de decisión para clasificación y regresión, en particular:

- Se presenta la lógica para la construcción de árboles de decisión, ya sean de regresión o clasificación.
- Se contemplan diferentes medidas con las que el árbol decide avanzar hacia un nuevo punto de decisión.
- Se presentan los conceptos de sobreajuste y complejidad del árbol, así como la forma de controlarlos.
- Se muestra el uso de **R** para la clasificación de clases binarias y para la predicción de variables respuesta numéricas en casos aplicados.

Capítulo 25

Máquinas de vector soporte

Ramón A. Carrasco^a e Itzcóatl Bueno^{b,a}

^aUniversidad Complutense de Madrid ^bInstituto Nacional de Estadística

25.1. Introducción

Aunque las máquinas de vector soporte se desarrollaron en los años 90 dentro de la comunidad informática ((Boser et al., 1992), (Cortes and Vapnik, 1995)) como un método de clasificación binaria, su aplicación se ha extendido a problemas de clasificación múltiple y regresión. Como técnica de clasificación, las máquinas de vector soporte (SVM por sus siglas inglés *Support Vector Machines*) son similares a la regresión logística pero la SVM enfatiza en un margen de error aceptable en torno a la frontera de decisión.

En la Fig. 25.1 se muestra que la regresión logística divide las observaciones en dos clases de tal forma que se minimice la distancia entre los puntos y la **frontera de decisión** (A). Por otro lado, la frontera de decisión (B) de la SVM separa los datos en dos clases, pero maximizando la distancia entre esta y los puntos de ambas clases. El **margen** es la distancia entre la frontera de decisión y los puntos más cercanos. El margen es una parte clave del SVM, puesto que evita clasificaciones erróneas de casos futuros como podría pasar en el caso de la regresión logística y como se ilustra en la Fig. 25.2.

En resumen, los nuevos datos pueden ser clasificados dentro del margen. Cuanto mayor sea este margen, mayor será la capacidad para clasificar correctamente estos puntos. Por tanto, para obtener una clasificación errónea en la SVM es necesario que una observación se clasifique aún más allá del margen que en cualquier otro discriminante lineal. En problemas reales, es difícil que los discriminantes lineales, vistos en el Cap. 21, logren una línea que divida perfectamente las categorías a clasificar. Sin embargo, en la SVM, se incluye en la función objetivo (que mide la calidad del ajuste de los datos de entrenamiento) una penalización a los puntos que queden del lado equivocado del límite de decisión. En caso de que los datos puedan ser divididos linealmente, no se cometerá ninguna penalización y se maximizará el margen. Mientras que si

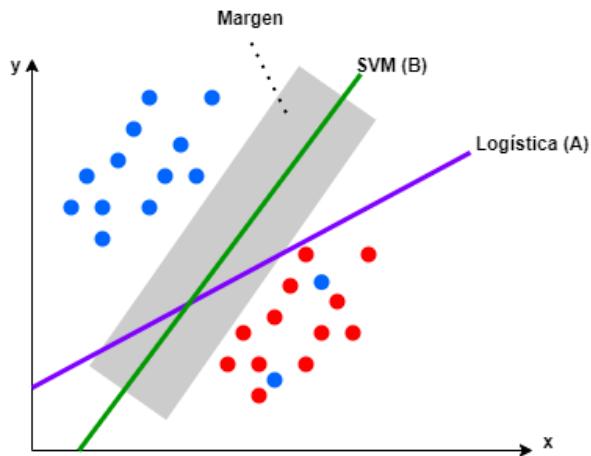


Figura 25.1: SVM vs Regresión logística.

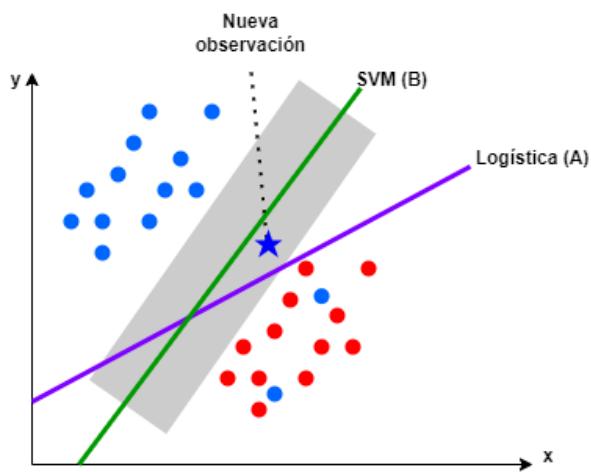


Figura 25.2: Nueva observación clasificada en SVM vs Regresión logística.

los datos no son linealmente separables, el mejor ajuste vendrá dado por el equilibrio entre una penalización del error total bajo y un margen de decisión grande. La penalización a una observación mal clasificada es proporcional a la distancia desde la frontera de decisión.

Sin embargo, la SVM también tiene desventajas reseñables. En primer lugar, la SVM no es adecuada en conjuntos de datos grandes porque la complejidad de entrenamiento es elevada. Además, la SVM no funciona bien cuando los datos tienen mucho ruido, es decir, cuando las clases se superponen. Finalmente, si el conjunto de datos de entrenamiento tiene más variables que observaciones, el rendimiento del modelo disminuirá.

25.2. Algoritmo SVM para clasificación binaria

El algoritmo por el que se obtiene un modelo SVM ([Vapnik, 1997](#)) se basa en la ecuación del hiperplano compuesta por dos hiperparámetros: un vector de números reales ω de la misma dimensión que el vector de variables de entrada x , y un número real b tal que:

$$\omega x - b = 0 \quad (25.1)$$

Donde ωx es $\omega^{(1)}x^{(1)} + \omega^{(2)}x^{(2)} + \dots + \omega^{(p)}x^{(p)}$ siendo p el número de variables incluidas en x . De este modo, la predicción para una instancia de x viene dada por:

$$y = sign(\omega x - b) \quad (25.2)$$

Siendo *sign* el operador que devuelve +1 para cualquier valor positivo y -1 para los valores negativos. Por tanto, el objetivo es ajustar los valores óptimos de ω y b para el algoritmo. Estos hiperparámetros se obtienen resolviendo un problema de optimización sujeto a las siguientes restricciones:

$$\omega x_i - b \geq 1 \text{ si } y_i = +1 \text{ y} \quad (25.3)$$

$$\omega x_i - b \leq 1 \text{ si } y_i = -1 \quad (25.4)$$

Además, el objetivo del problema de optimización es maximizar el margen en torno a la frontera de decisión. Para conseguir esto es necesario minimizar la norma euclídea, y, por tanto, el problema a resolver es:

$$\min ||\omega|| \text{ sujeto a}$$

$$y_i(\omega x_i - b) \geq 1 \text{ para } i = 1, \dots, N$$

25.3. ¿Y si tengo más de dos clases?

Hasta ahora se ha presentado la SVM como un algoritmo solo aplicable a la clasificación de dos clases pero ¿y si se tienen más de dos clases? En general, hay dos enfoques para resolver esto: **uno contra todos** (OVA, por *One Vs All*) y **uno contra uno** (OVO, por *One Vs One*). En el enfoque OVA, se ajusta una SVM para cada clase, es decir una clase contra las demás y se clasifica a la clase para la cual el margen es mayor. En cambio, en el enfoque OVO se ajustan todas las SVM por pares y se clasifica a la clase que gane las competiciones por pares.

25.4. Truco del *kernel*: tratando con la no linealidad

Las SVM funcionan muy bien si la separación entre clases es lineal. Sin embargo, si la separación es más compleja se intenta transformar el espacio en otro de mayor dimensionalidad donde las clases sí sean separables linealmente. Para ello, el modelo SVM se extiende incluyendo la función de pérdida (ℓ) “hinge” (([Gentile and Warmuth, 1998](#)),([Lee and Lin, 2013](#))) definida como:

$$\ell(y_i) = \max(0, 1 - y_i(\omega x_i - b)) \quad (25.5)$$

En machine learning, esta función de pérdida se utiliza para entrenar clasificadores, más concretamente para la clasificación por el margen máximo (métodos de clasificación binaria que se utiliza cuando hay una frontera lineal que separa perfectamente los datos de entrenamiento de una categoría de los de la otra), sobre todo para las SVM. La función de pérdida es cero cuando se cumplen las restricciones, es decir, si ωx_i es clasificado en el lado correcto de la frontera de decisión. Por otro lado, si un dato es mal clasificado, el valor obtenido con la función de pérdida es proporcional a la distancia hasta la frontera de decisión. Por tanto, el objetivo es minimizar la función de coste:

$$C\|\omega\|^2 + \frac{1}{N} \sum_{i=1}^N \max(0, 1 - y_i(\omega x_i - b)) \quad (25.6)$$

Donde C es un hiperparámetro que controla la compensación entre incrementar el tamaño de la frontera de decisión y asegurar que cada x_i sea clasificado en el lado correcto de la frontera de decisión.

Un modelo SVM que optimiza la función de pérdida se denomina SVM *soft-margin* mientras que el modelo original es conocido como SVM *hard-margin*. La ecuación (25.6) muestra que para valores grandes de C el segundo término es despreciable, por lo que el algoritmo ignorará por completo la clasificación errónea y tratará de obtener el mayor margen posible. Si se reduce el valor de C , se penaliza más cada error de clasificación, por lo que se cometerán menos errores sacrificando amplitud del margen.

A veces no es posible separar los datos por un hiperplano en su espacio original. Sin embargo, el **truco del kernel** utiliza una función que implícitamente transforma el espacio original a un espacio de mayor dimensión durante la optimización de la función de coste, como se muestra

25.4. Truco del kernel: tratando con la no linealidad

441

en la Fig. 25.3. Así, es posible transformar un espacio de datos bidimensional no separable linealmente en un espacio de datos tridimensional linealmente separable usando un mapeo específico definido por $\phi : x \rightarrow \phi(x)$ donde $\phi(x)$ es un vector de mayor dimensión que x . Sin embargo, no se conoce la función de mapeo que funcionará en los datos. Si se prueban todas las transformaciones posibles, podría ser ineficiente y no llegar a la resolución del problema de clasificación planteado.

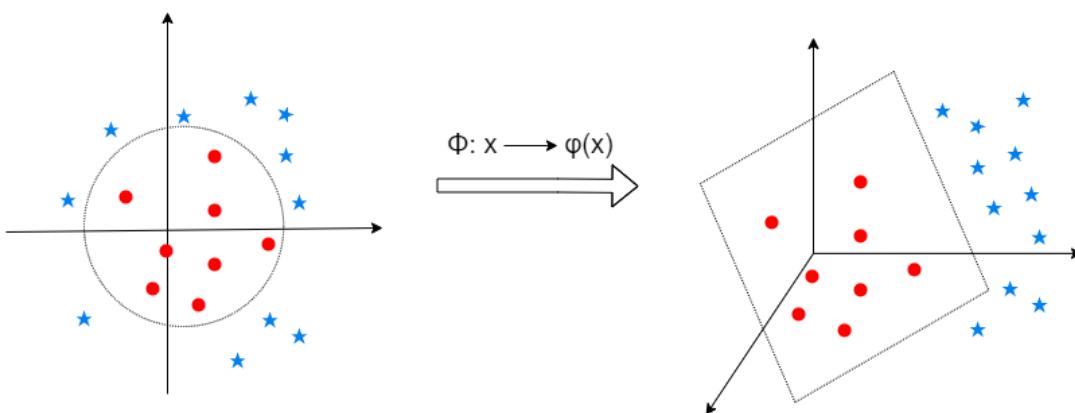


Figura 25.3: Izquierda: Las dos clases en el espacio original (2-D). Derecha: Las dos clases en el espacio de sobredimensionado (3-D).

Se puede trabajar eficientemente en espacios de mayor dimensión sin necesidad de hacer las transformaciones explícitamente. Utilizando el truco del *kernel* se puede evitar este proceso costoso de transformación de tal manera que se evita calcular el producto escalar reemplazándolo por una operación más simple con las variables originales que proporciona el mismo resultado. A continuación se explican algunos de estos operadores especiales, llamados *kernels*, que permiten llevar a cabo dicha transformación.

25.4.1. Algunos *kernels* populares

Los *kernels* ((Schölkopf et al., 1997), (Scholkopf et al., 1997)) más populares en el entrenamiento de SVM están incluidos dentro de la función `svm()` del paquete `e1071` donde se puede especificar en el hiperparámetro `kernel`. Estos kernel son:

- lineal: $K(u, v) = \langle u, v \rangle$
- polinomial de grado δ : $K(u, v) = \gamma(k_1 + \langle u, v \rangle)^\delta$
- base radial: $K(u, v) = e^{\gamma \|u-v\|^2}$
- sigmoidal: $K(u, v) = \tanh(\gamma \langle u, v \rangle + k_1)$

Donde $\langle u, v \rangle = \sum_{i=1}^n u_i v_i$ es el producto escalar. Cada uno de estos kernels tiene sus propios hiperparámetros, como δ o γ , que es necesario tunear para optimizar el rendimiento de la SVM.

En machine learning, el término **tunear**, hace referencia al hecho de ajustar automáticamente tratando de optimizar los hiperparámetros del algoritmo. A la hora de ajustar en **R** un modelo SVM se puede conocer los hiperparámetros a ajustar utilizando la función `modelLookup` de `caret`. Por ejemplo, para una SVM con kernel de base lineal se usaría así:

```
modelLookup("svmLinear")
  model parameter label forReg forClass probModel
1 svmLinear          C Cost    TRUE    TRUE    TRUE
```

El hiperparámetro que se puede ajustar en un modelo SVM con kernel de base lineal en **R** es el coste (*C*), el cual representa a la constante *C* en la ecuación (25.6).

25.5. Procedimiento con R: la función `svm()`

En el paquete `e1071` de R se encuentra la función `svm()` que se utiliza para entrenar un modelo máquinas vector soporte:

```
svm(x, y, scale = TRUE, type = NULL, kernel = ..., ...)
```

- *x*: conjunto de datos de entrenamiento que contiene los predictores
- *y*: vector respuesta con las clases o valores de la variable respuesta.
- `scale`: booleano que indica si es necesario escalar las variables.
- `type`: indica si se pretende resolver un problema de clasificación o de regresión.
- `kernel`: *kernel* utilizado durante el entrenamiento y la predicción.

25.6. Aplicación de un modelo SVM Radial con ajuste automático en R

Los datos utilizados para entrenar el modelo SVM en este capítulo se cargan desde la librería `CDR`. Además, para su entrenamiento se requieren las librerías `caret` y `e1071`.

```
library("CDR")
library("caret")
library("e1071")
library("reshape")
library("ggplot2")

data(dp_entr_NUM)
```

Se entrena un modelo SVM con kernel radial utilizando el conjunto de entrenamiento con todas las variables numéricas. Previamente, se aplica una normalización z-score, presentadas en el

25.6. Aplicación de un modelo SVM Radial con ajuste automático en R

443

Cap. 9, al conjunto de entrenamiento. De este modo, las variables que inicialmente tenían distintas escalas de medida, ahora todas se miden en la misma escala. Además, se ajustan automáticamente los hiperparámetros de dicho algoritmo durante el proceso de entrenamiento.

```
trControl <- trainControl(
  method = "cv",
  number = 10,
  classProbs = TRUE,
  preProcOptions = list("center"),
  summaryFunction = twoClassSummary
)

# Se especifica un rango de valores para los hiperparámetros
tuneGrid <- expand.grid(sigma = seq(from=0.1, to=0.2, by=0.05),
                         C = 10**(-2:4))
```

Se define como procedimiento de muestreo una validación cruzada, como la presentada en el Cap. 9, de 10 folds. Además, se le indica al modelo que debe calcular las probabilidades de clase en cada remuestreo en caso de estar entrenando un modelo de clasificación. Con el argumento `summaryFunction = twoClassSummary` se le indica al modelo que para resumir los resultados se calculen la sensibilidad, especificidad y el área bajo la curva ROC. Como se ha comentado, conviene estandarizar los datos, esto se le indica a la función a través del argumento `preProcOptions` con la opción `center`. A su vez, se define una red de hiperparametros a optimizar. A través de la función `train()` se ajusta automáticamente el modelo con los hiperparametros óptimos.

```
# Se fija la semilla aleatoria
set.seed(101)

# Se entrena el modelo
model <- train(CLSPRO_pro13 ~ .,
                data=dp_entr_NUM,
                method="svmRadial",
                metric="ROC",
                trControl=trControl,
                tuneGrid=tuneGrid)

model

Support Vector Machines with Radial Basis Function Kernel

558 samples
 19 predictor
  2 classes: 'S', 'N'

No pre-processing
Resampling: Cross-Validated (10 fold)
```

```
Summary of sample sizes: 502, 502, 502, 503, 503, 502, ...
Resampling results across tuning parameters:
```

sigma	C	ROC	Sens	Spec
0.10	1e-02	0.9553241	0.8785714	0.7071429
0.10	1e-01	0.9566327	0.8924603	0.8247354
0.10	1e+00	0.9434902	0.8604497	0.8496032
0.10	1e+01	0.9227230	0.8460317	0.8423280
0.10	1e+02	0.8804894	0.8567460	0.8279101
0.10	1e+03	0.8645692	0.8674603	0.8206349
0.10	1e+04	0.8548469	0.8423280	0.8242063
0.15	1e-02	0.9527636	0.8535714	0.6642857
0.15	1e-01	0.9513653	0.9105820	0.8105820
0.15	1e+00	0.9310091	0.8783069	0.8494709
0.15	1e+01	0.8941421	0.8531746	0.8387566
0.15	1e+02	0.8602088	0.8781746	0.8242063
0.15	1e+03	0.8369331	0.8458995	0.8134921
0.15	1e+04	0.8369284	0.8637566	0.8064815
0.20	1e-02	0.9443925	0.8535714	0.6321429
0.20	1e-01	0.9440098	0.9250000	0.7384921
0.20	1e+00	0.9199310	0.8818783	0.8387566
0.20	1e+01	0.8752031	0.8674603	0.8207672
0.20	1e+02	0.8477324	0.8674603	0.8063492
0.20	1e+03	0.8308296	0.8638889	0.8134921
0.20	1e+04	0.8308296	0.8638889	0.8099206

```
ROC was used to select the optimal model using the largest value.
The final values used for the model were sigma = 0.1 and C = 0.1.
```

Los argumentos que requiere la función son la `formula`, es decir, indicar la variable respuesta y qué predictores intervienen en el modelo. Los datos que se van a utilizar, así como el algoritmo a entrenar, en este caso la SVM con kernel de base radial. Además, se indica una métrica para el rendimiento del modelo, en caso de no indicarlo `R` asigna la más acorde de acuerdo a la variable respuesta. Finalmente, se incluyen las opciones de entrenamiento y la red de hiperparámetros a probar para determinar la combinación óptima. Los hiperparámetros del modelo entrenado son $\sigma = 0,1$ y $C = 0,1$. Este resultado queda definido en la salida del modelo, pero también es representable como en la Fig. 25.4. En este gráfico el eje y mide el rendimiento del modelo para ciertos valores de los hiperparámetros. Cada línea representa un valor para el hiperparámetro `sigma`, y se mide su rendimiento variando distintos niveles del parámetro coste (`C`), que queda representado en el eje x. Así, se observa que la línea roja (`sigma=0,1`) alcanza el mayor nivel de precisión en el valor $C=0,1$.

```
ggplot(model)
```

Los boxplot de la Fig. 25.5 muestran un resumen del rendimiento del modelo en las distintas repeticiones del proceso de validación cruzada. De esta manera, se observa como la sensibilidad es superior al 75 % y la especificidad supera valores del 70 %, esto indica que el modelo entrenado

25.6. Aplicación de un modelo SVM Radial con ajuste automático en R

445

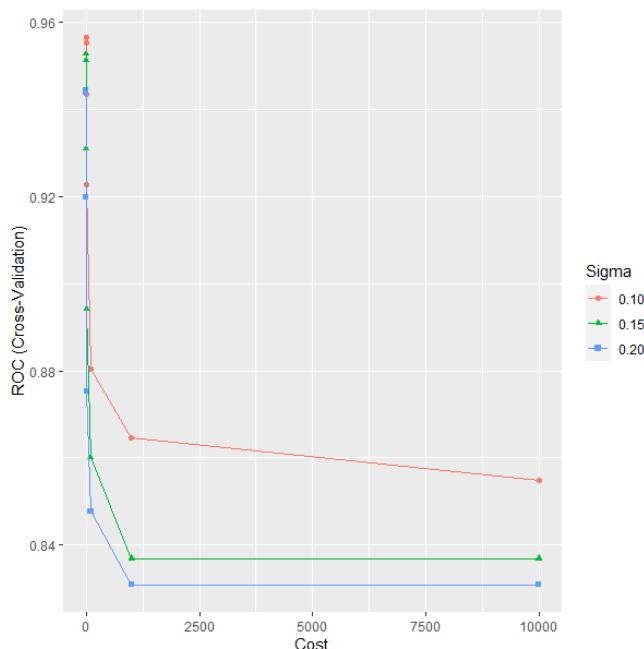


Figura 25.4: Optimización de los parámetros C y sigma de una SVM.

es capaz de predecir correctamente tanto a los clientes que van a comprar el nuevo producto como los que no lo van a hacer.

```
ggplot(melt(model$resample[,-4]), aes(x = variable, y = value, fill=variable)) +
  geom_boxplot(show.legend=FALSE) +
  xlab(NULL) + ylab(NULL)
```

25.6.1. Importancia de las variables

En machine learning, muchos de los algoritmos de caja negra (definidos en el Cap. 4), no proporcionan información sobre la importancia que tiene cada variable en el modelo. Este es el caso de las máquinas de vector soporte. Tanto para la SVM como para otros algoritmos, es posible cuantificar la importancia de cada variable utilizando paquetes de **R** como DALEX, iml o vip.

Este último paquete incluye una función con el mismo nombre `vip()`. Para medir la importancia, se indica qué métrica se utilizó en el proceso de entrenamiento del modelo, en el caso de la SVM se indicará que fue el área bajo la curva (`metric=.auc`). En el argumento `pred_wrapper` se indica una función de medida que contenga tanto los valores observados como los valores predichos. Dado que la SVM entrenada utiliza AUC para medir el rendimiento del modelo ajustado, la función de medida indicada en `pred_wrapper` deberá devolver la probabilidad

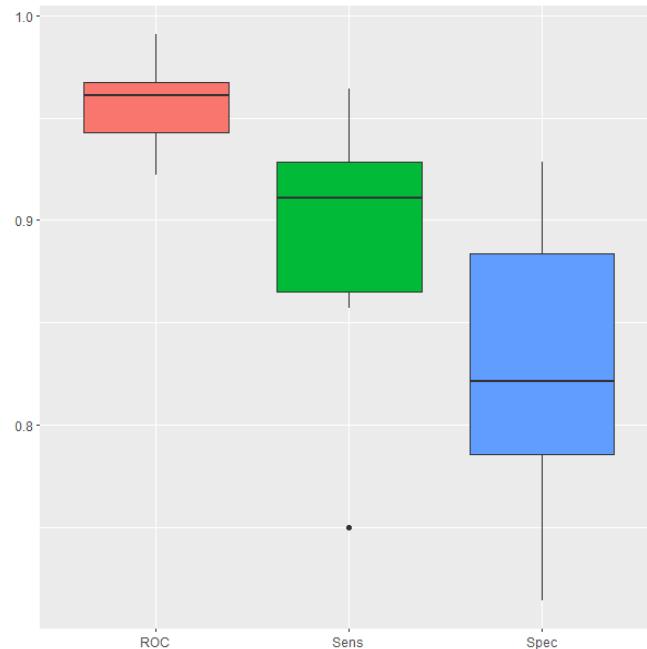


Figura 25.5: Resultados del modelo obtenidos durante la validación cruzada.

de que el modelo asigne una observación a la clase de referencia. En este ejemplo, la clase de referencia es “SI”, puesto que interesa saber si un cliente comprará el nuevo producto. Entonces, la función de predicción se define como:

```
prob_si <- function(object, newdata) {
  predict(object, newdata = newdata, type = "prob")[, "S"]
}
```

Ejecutando la función `vip()` con los argumentos mencionados se genera la Fig. 25.6. Este gráfico indica la importancia de cada variable en el modelo de más a menos importante. En este caso, la variable más importante es el importe gastado en el *smartchwatch fitness*, seguida muy de cerca por la variable que indica si el cliente compra o no el *smartchwatch fitness*. En el otro extremo, se observa que las variables que indican si el cliente tiene un nivel de educación básico o no, no son muy relevantes en la SVM entrenada.

```
library("vip")

set.seed(101)
vip(model, train = dp_entr_NUM, target = "CLS_PRO_pro13", metric = "auc",
  reference_class = "S", pred_wrapper = prob_si, method="permute",
  aesthetics = list(color = "steelblue2", fill = "steelblue2"))
```

25.6. Aplicación de un modelo SVM Radial con ajuste automático en R

447

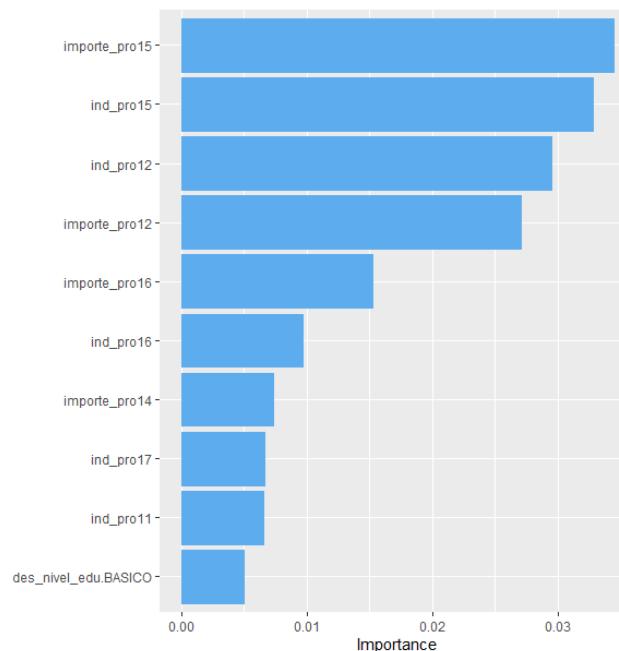


Figura 25.6: Importancia de las variables incluidas en la SVM.

A partir de la Fig. 25.6 se puede concluir que para predecir si un cliente comprará o no el *tensiómetro digital* las variables que más importancia tienen: son el importe que gastó en el *smartwatch fitness*, si compró o no el *smartwatch fitness*, el importe que gastó en la *depiladora eléctrica* y si compró o no la *depiladora eléctrica*.

De forma similar se podrían probar el resto de *kernels* disponibles para el algoritmo SVM.

Resumen

En este capítulo se introduce al lector en el algoritmo de máquinas vector soporte, en particular:

- Se presenta el concepto de margen de decisión, y las ventajas de la SVM respecto a otras técnicas de clasificación.
- Se explica el truco del kernel cuando los datos no son separables por un hiperplano en su espacio original
- Se da un repaso a los kernels más utilizados.
- Se presenta la aplicación de una SVM con kernel radial en **R** para la clasificación de datos con respuesta binaria, en el que se ajustan automáticamente los hiperparámetros.
- Se obtiene la importancia de las variables del modelo final.

Capítulo 26

Clasificador k-vecinos más próximos

Ramón A. Carrasco^a e Itzcóatl Bueno^{b,a}

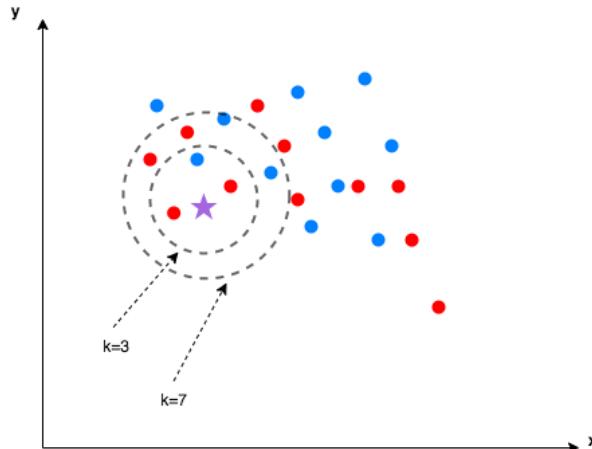
^aUniversidad Complutense de Madrid ^bInstituto Nacional de Estadística

26.1. Introducción

El k-vecinos más próximos (KNN, por sus siglas en inglés *k-Nearest Neighbors*) es un algoritmo de aprendizaje no paramétrico. Un algoritmo no paramétrico no presupone la forma concreta del modelo a entrenar, siendo más flexible. Sin embargo, esto se consigue a costa de necesitar más datos de entrenamiento y siendo más lentos que los algoritmos paramétricos. Al contrario que otros algoritmos de aprendizaje que permiten deshacerse de los datos de entrenamiento una vez se entrena el modelo, el modelo KNN guarda las observaciones de entrenamiento en memoria. Esto es, al incorporar una nueva observación x , el algoritmo KNN encuentra las k observaciones del conjunto de datos de entrenamiento más similares a la nueva y proporciona la clase mayoritaria (en el caso de clasificación) o el valor medio (en el caso de regresión).

El número de casos (k) a utilizar para clasificar las nuevas observaciones es un parámetro crucial para este algoritmo (James et al., 2013). Si, por ejemplo, $k = 3$, el modelo KNN utilizará las tres observaciones más similares (vecinos) al nuevo caso para clasificarlo. Es recomendable probar distintos valores de k para conseguir el mejor ajuste del modelo, y por tanto, es conveniente evitar valores extremos de k . Si se establece un valor muy bajo de k aumentará el sesgo y llevará a clasificaciones erróneas. Mientras que valores muy elevados de k harán que el algoritmo sea computacionalmente costoso y además tampoco será un buen clasificador. Además, también se recomienda establecer valores impares de k para evitar puntos muertos estadísticos (empate entre categorías) y un resultado no válido.

La escala de las variables puede impactar en el resultado del modelo KNN. Por ello, el conjunto de datos debe escalarse para que aquellas variables con unidades de medida grandes no tengan

Figura 26.1: Ejemplo de k -vecinos más próximos.

más importancia en el cálculo que otras con magnitudes menores. Así se reduce la importancia de las variables debido a sus unidades de medida y se estandariza la varianza.

Pese a que el modelo KNN es fácil de entender y generalmente preciso, almacenar el conjunto de datos de entrenamiento, así como calcular la distancia entre cada nueva observación a clasificar y las observaciones del conjunto de datos, supone la necesidad de recursos computacionales altos. Esto implica que cuanto mayor es la cantidad de observaciones en el conjunto de datos, mayor es el tiempo para la ejecución de una sola predicción, y, por tanto, esto puede dar lugar a tiempos de procesamiento lentos. Por este motivo, no se recomienda el uso del algoritmo KNN cuando se dispone de conjuntos de datos muy grandes. Otra desventaja a tener en cuenta es la dificultad de aplicar KNN a conjuntos de datos con un gran número de variables, puesto que calcular las distancias entre observaciones con múltiples dimensiones también incrementará la necesidad de recursos computacionales y podría dificultar que se clasifique de forma precisa.

26.2. Decisiones a tener en cuenta

La elección de la función de distancia, así como el número de vecinos k son decisiones que debe tomar el investigador antes de ejecutar el algoritmo. Siendo este último el hiperparámetro del modelo que la función `modelLookup()` de `caret` nos devuelve considerando que es primordial ajustarlo:

```
modelLookup("knn")
  model parameter      label forReg forClass probModel
  1   knn             k #Neighbors    TRUE     TRUE      TRUE
```

26.2.1. Función de distancia a utilizar

El modelo KNN determina la cercanía entre dos observaciones a través de una función de distancia. Generalmente, se utilizan la distancia euclídea, mostrada en la ecuación (26.1), y la distancia de Manhattan, mostrada en la ecuación (26.2). Otras funciones de distancia, como las presentadas en el Cap. ??, también pueden ser utilizadas para el entrenamiento de este algoritmo.

$$d(x_i, x_k) = \sqrt{\sum_{j=1}^p (x_i^{(j)} - x_k^{(j)})^2} \quad (26.1)$$

$$d(x_i, x_k) = \sum_{j=1}^p |x_i^{(j)} - x_k^{(j)}| \quad (26.2)$$

En el caso de querer incluir tanto variables cuantitativas como variables cualitativas en el cálculo de la distancia, el coeficiente de disimilitud de Gower es la función de distancia más popular para esta situación. El coeficiente de disimilitud de Gower se define como:

$$d(i, j) = \frac{\sum_{k=1}^p \omega_k \delta_{ij}^{(k)} d^{(k)}_{ij}}{\omega_k \delta_{ij}^{(k)}} \quad (26.3)$$

El coeficiente de Gower es una media ponderada de las distancias $d_{ij}^{(k)}$ con ponderaciones $\omega_k \delta_{ij}^{(k)}$.

26.2.2. Número de vecinos (k) seleccionados

Como se ha reiterado, la elección de cuántos vecinos (k) intervienen en el ajuste del algoritmo es determinante para su rendimiento. Si se escogen demasiado pocos vecinos, se producirá sobreajuste en el modelo. En el extremo en el que sólo se utilizará un vecino ($k = 1$), la predicción se basará en la observación con la menor distancia al elemento a clasificar. Por otro lado, un número alto de vecinos (k) hace que el modelo no ajuste bien al tener en cuenta un vecindario más grande. En este sentido, en el caso extremo de elegir todas las observaciones como vecinos más próximos ($k = n$), se obtendrá el valor medio (en el caso de la regresión) o la clase mayoritaria (en el caso de la clasificación) como valor predicho para todas las observaciones del conjunto de entrenamiento.

No existe una regla general para la elección óptima de k , puesto que en gran medida dependerá del conjunto de datos utilizado. Cuando el conjunto de datos tiene pocas variables que no aporten información, valores pequeños de k tienden a funcionar mejor. Cuantas más variables sin importancia se incluyen en el conjunto de datos, mayor deberá ser el valor de k para suavizar su efecto.

26.3. Procedimiento con R: la función knn()

En el paquete `class` de R se encuentra la función `knn()` que se utiliza para entrenar el modelo k-vecinos más próximos:

```
knn(train, test, cl, k = 1, ...)
```

- `train`: conjunto de datos con las observaciones de entrenamiento.
- `test`: conjunto de datos con las observaciones de validación. Un vector se interpreta como una única observación a validar.
- `cl`: clases de las observaciones de entrenamiento.
- `k`: número de vecinos a considerar

26.4. Aplicación del modelo KNN en R

En este ejemplo se entrena un modelo KNN para clasificar qué clientes comprarán el *tensiómetro digital* teniendo en cuenta sus características y el resto de sus compras. Este conjunto de datos está incluido en el paquete CDR con el nombre `dp_entr_NUM`. En este conjunto de datos todas las variables son cuantitativas (excepto la clase objetivo) pero dichas variables tienen distintas escalas de medida (euros, años, unidades, etc.) por lo que es necesario indicar en la función `trainControl()` que se haga un preprocessamiento para estandarizar las variables. Además, se define como método de remuestreo la validación cruzada, como la presentada en el Cap. 9, con 10 folds.

```
library("CDR")
library("class")
library("caret")
library("reshape")
library("ggplot2")

data(dp_entr_NUM)

head(dp_entr_NUM)

  ind_pro11 ind_pro12 ind_pro14 ind_pro15 ind_pro16 ind_pro17 des_nivel_edu.ALTO
1       1       0       1       1       1       0       0
2       0       0       1       0       1       0       0
3       0       0       1       1       1       1       0
4       0       1       1       0       0       0       0
5       0       1       1       0       1       0       0
6       1       0       1       0       0       0       1
  des_nivel_edu.BASICO des_nivel_edu.MEDIO importe_pro11 importe_pro12 importe_pro14
1             0             1         157            0            40
2             0             1            0            0            240
3             1             0            0            0            425
```

26.4. Aplicación del modelo KNN en R

453

```

4          0          1          0        120        60
5          1          0          0        120      133
6          0          0        115        0      220
    importe_pro15 importe_pro16 importe_pro17 edad tamano_fam anos_exp ingresos_ano
    ← CLS_PRO_pro13
1          200        180        0      49       4     24   30000
    ← S
2          0        180        0      38       2     12   53000
    ← N
3          200        180      300      61       4     37 172000
    ← S
4          0          0        0      47       3     21   38000
    ← N
5          0        180        0      34       1     10   38000
    ← N
6          0          0        0      43       2     18   60000
    ← N

# Definimos un método de remuestreo
cv <- trainControl(
  method = "repeatedcv",
  number = 10,
  repeats = 5,
  classProbs = TRUE,
  preProcOptions = list("center"),
  summaryFunction = twoClassSummary
)

```

Antes de obtener el modelo definitivo, es necesario la selección del número óptimo de vecinos k entrenando distintos modelos variando dicho hiperparámetro. Para facilitar este trabajo arduo, en el paquete `caret` de R se puede definir una red de posibles valores sobre los que evaluar el modelo KNN y que de forma automática se determine el valor que mejor rendimiento proporcione. A continuación se definen los posibles valores de k que se quieren evaluar.

```
# Definimos la red de posibles valores del hiperparámetro
hyper_grid <- expand.grid(k = c(1:10,15,20,30,50,75,100))
```

Una vez se que se ha definido tanto el método de remuestreo como la red de posibles valores del hiperparámetro se puede entrenar el modelo:

```

set.seed(101)
# Se entrena el modelo ajustando el hiperparámetro óptimo
model <- train(
  CLS_PRO_pro13 ~ .,
  data = dp_entr_NUM,

```

```

method = "knn",
trControl = cv,
tuneGrid = hyper_grid,
metric = "ROC"
)

model

k-Nearest Neighbors

558 samples
 17 predictor
  2 classes: 'S', 'N'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 5 times)
Summary of sample sizes: 502, 502, 502, 503, 503, 502, ...
Resampling results across tuning parameters:

      k    ROC      Sens      Spec
1  0.6584524  0.6466402  0.6702646
2  0.6769109  0.6179101  0.6131217
3  0.6828893  0.6216402  0.6496561
4  0.6851087  0.6404233  0.6394709
5  0.6951129  0.6540212  0.6666138
6  0.6914664  0.6216402  0.6543386
7  0.6982592  0.6252381  0.6953439
8  0.6974556  0.6281481  0.6960053
9  0.6992229  0.6159524  0.7117725
10 0.6994133  0.6037037  0.7052910
15 0.6875879  0.5749206  0.7232011
20 0.6731477  0.5722751  0.7010582
30 0.6752986  0.5529630  0.7024603
50 0.6890259  0.5163757  0.7605556
75 0.6852886  0.5092593  0.7670106
100 0.6719378  0.5049471  0.7820106

ROC was used to select the optimal model using the largest value.
The final value used for the model was k = 10.

```

En la Fig. 26.2 se observa que el número óptimo de vecinos es $k = 10$, donde se alcanza el rendimiento óptimo del modelo.

```

ggplot(model) + geom_vline(xintercept =
  unlist(model$bestTune), col="red", linetype="dashed") + theme_light()

```

El boxplot de los resultados obtenidos durante el proceso de validación cruzada muestra que el AUC del modelo oscila entre un 60 % y un 85 % aproximadamente.

26.4. Aplicación del modelo KNN en R

455

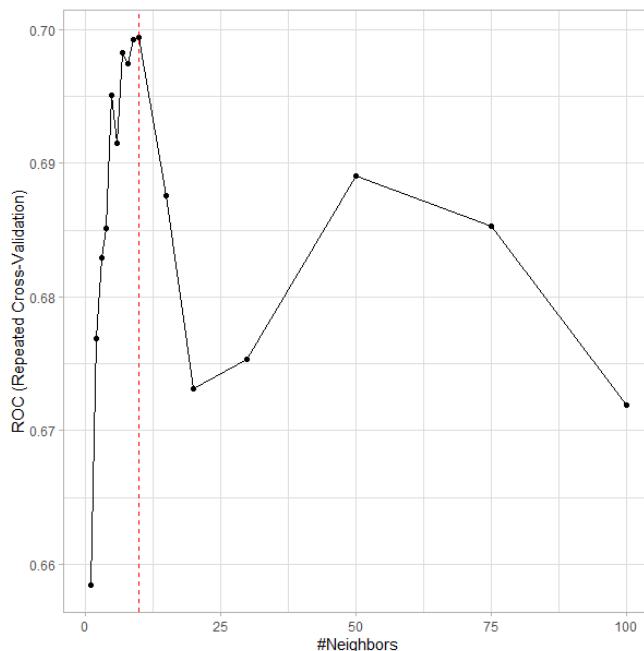


Figura 26.2: Número óptimo de vecinos (k).

```
ggplot(melt(model$resample[,-4]), aes(x = variable, y = value, fill=variable)) +
  geom_boxplot(show.legend=FALSE) +
  xlab(NULL) + ylab(NULL)
```

El modelo se resiente en su rendimiento al tener dificultades en predecir correctamente la clase positiva, más concretamente se puede observar en la Fig. 26.3 que la sensibilidad oscila entre el 40 % y el 75 %; resultados ligeramente peores que los que obtiene al predecir observaciones de la clase negativa, los cuales oscilan entre el 50 % y el 85 %.

Resumen

En este capítulo se introduce al lector en el algoritmo de aprendizaje supervisado conocido como k-vecinos más próximos, destacando:

- Las decisiones a tener en cuenta antes de entrenar el modelo de k-vecinos más próximos.
- Se exponen algunas de las distancias más utilizadas para el entrenamiento de este modelo.
- Se especifican las ventajas y desventajas del número de vecinos a tener en cuenta.

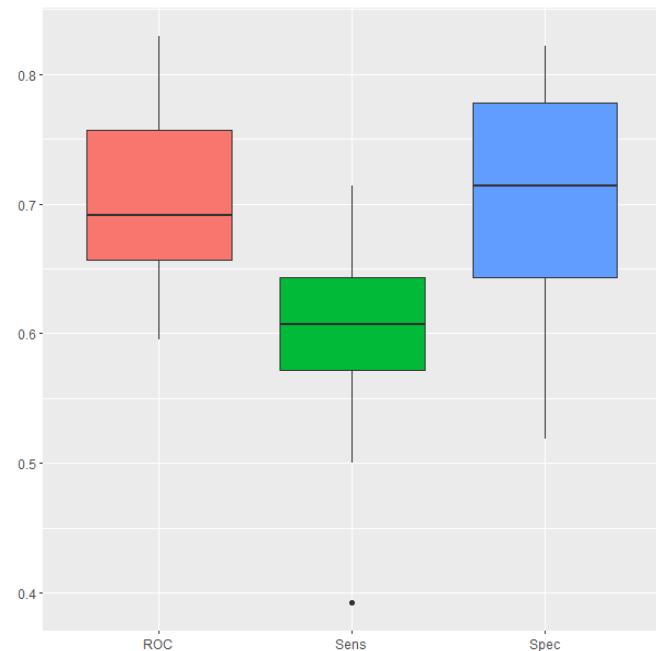


Figura 26.3: Resultados obtenidos durante el proceso de validación cruzada.

Capítulo 27

Naive Bayes

Ramón A. Carrasco^a e Itzcóatl Bueno^{b,a}

^aUniversidad Complutense de Madrid ^bInstituto Nacional de Estadística

27.1. Introducción

Naive Bayes es un algoritmo de aprendizaje supervisado que se utiliza principalmente para la clasificación. Como otros algoritmos de aprendizaje supervisado, este algoritmo se entrena con variables de entrada y la categoría asociada a cada observación y que el modelo debe predecir. Sin embargo, se denomina ‘naive’ dado que asume que las variables de entrada que se incluyen en el modelo son independientes entre sí. Por lo tanto, si se cambia una de las variables de entrada, las demás no se verán afectadas.

Aunque el algoritmo *Naive Bayes* es sencillo, destaca por su facilidad de implementación y su potencia predictiva. Su ventaja principal es que utiliza un enfoque probabilístico, lo que implica que todos los cálculos se realizan en tiempo real y, por tanto, los resultados se obtienen inmediatamente, como se detalla más adelante. Además, cuando el conjunto de datos tiene un gran número de observaciones, el algoritmo *Naive Bayes* es ventajoso respecto a algoritmos como la SVM (Cap. 25) o el Random Forest (Cap. 28) debido a su mejor tiempo de computación.

Al utilizar un enfoque probabilístico, el algoritmo *Naive Bayes* está construido sobre conceptos de probabilidad, presentados en el Cap. 12, y en especial, este algoritmo hace uso del **Teorema de Bayes**. A continuación se repasan los conceptos fundamentales en los que está basado el algoritmo.

27.2. Teorema de Bayes

Sean dos eventos A y B definidos en un espacio muestral, se puede definir la probabilidad condicional de que ocurra el evento A dado que previamente se haya observado B como:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad (27.1)$$

Siempre que $P(B) \neq 0$ y donde $P(A \cap B)$ es la probabilidad de que ocurran ambos eventos a la vez. Los eventos son intercambiables de tal forma que $P(A \cap B) = P(B|A)P(A)$ y si se reemplaza en la primera ecuación tenemos:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \quad (27.2)$$

Esta fórmula es la definición del teorema de Bayes. El algoritmo de clasificación *Naive Bayes* (NB) está basado en este teorema. Para ampliar los conceptos estadísticos aquí presentados pueden consultarse en más detalle en el Cap. 12.

27.3. El algoritmo *naive* Bayes

Si se adapta el teorema de Bayes a un problema de clasificación, se tendría:

$$P(C = c|\ell) = \frac{P(\ell|C = c) \cdot P(C = c)}{P(\ell)} \quad (27.3)$$

En este caso, $P(C = c|\ell)$ representa el objetivo de estimación en un problema de clasificación, es decir, la probabilidad de que un individuo pertenezca a la clase c después de haber observado la evidencia ℓ (incluida en las variables del modelo). Esta es la denominada **probabilidad a posteriori**. El resto de elementos de la fórmula, se definen como:

- $P(C = c)$ es la **probabilidad a priori** de pertenecer a la clase c , es decir, la probabilidad que un individuo tiene de ser asignado a esa clase sin observar sus características previamente.
- $P(\ell|C = c)$ es la verosimilitud de observar una instancia particular de las variables incluidas en el modelo cuando el individuo pertenece a la clase c .
- $P(\ell)$ es la verosimilitud de observar una instancia particular de las variables incluidas en el modelo, independientemente de a qué clase pertenezca el individuo.

Sin embargo, una gran dificultad para aplicar esta ecuación es la necesidad de conocer que $P(\ell|c)$ es igual a $P(\ell_1 \cap \ell_2 \cap \dots \cap \ell_k|c)$. La existencia de un ejemplo concreto en el conjunto de datos de entrenamiento que coincida a la perfección con ℓ es complicado, y en el caso de existir, no se tendrían suficientes ejemplos para poder estimar una probabilidad de forma fiable. La forma de solucionar este problema es incluir una suposición de independencia particularmente fuerte, que como ya se mencionó en la Sec. 27.1, es lo que aporta la denominación de ‘naive’ al algoritmo.

La *independencia condicional* implica que conocer un evento no aporta información sobre otro evento. Esto es equivalente a:

$$P(AB|C) = P(A|C) \cdot P(B|C) \quad (27.4)$$

De este modo, el problema de clasificación en el que era difícil estimar $P(\ell_1 \cap \ell_2 \cap \dots \cap \ell_k|c)$, ahora se tendría:

$$P(\ell|c) = P(\ell_1|c) \cdot P(\ell_2|c) \cdots P(\ell_k|c) \quad (27.5)$$

Y cada uno de los elementos $P(\ell_i|c)$ puede obtenerse directamente de los datos. Combinando este resultado con la regla de Bayes aplicada a un problema de decisión, se obtiene la ecuación dada por el algoritmo *Naive Bayes*:

$$P(c|\ell) = P(\ell_1|c) \cdot P(\ell_2|c) \cdots P(\ell_k|c)P(c) \quad (27.6)$$

El algoritmo *Naive Bayes* clasifica una nueva observación estimando la probabilidad de que pertenezca a cada clase y asignándole a aquella que tenga la mayor probabilidad.

En definitiva, el clasificador *Naive Bayes* es muy eficiente en términos de espacio de almacenamiento necesario, así como tiempos de procesamiento. Además, a pesar de ser muy simple, tiene en cuenta las características observadas. Otra de las ventajas de este clasificador es que es un modelo de aprendizaje incremental. Esto quiere decir que es una técnica de inducción que se actualiza con cada nueva observación de entrenamiento, es decir, no es necesario volver a procesar todo el conjunto de entrenamiento cuando se dispone de nuevas observaciones.

El ejemplo presentado en el Cap. 24 en el que se buscaba predecir si se podría jugar o no al tenis bajo unas condiciones meteorológicas determinadas, puede desarrollarse utilizando el modelo *Naive Bayes*. En este caso, el procedimiento puede resumirse en tres pasos:

- Resumir los datos en una tabla de frecuencias.
- Generar una tabla de verosimilitud obteniendo las probabilidades de las variables.
- Aplicar el teorema de Bayes para calcular la probabilidad a posteriori.

De este modo, las 15 observaciones registradas con el tipo de día (soleado, nublado, lluvioso) y si ese día se jugó, deben resumirse en una tabla de frecuencias como la Tabla 27.1. En este primer paso no se tiene en cuenta la información sobre humedad o viento.

Tabla 27.1: Tabla de frecuencias - Tipo de día vs Jugar partido

	SI	NO	Total
Soleado	2	4	6
Nublado	4	0	4
Lluvia	4	1	5

	SI	NO	Total
Total	10	5	15

En un segundo paso, se obtienen las probabilidades de cada categoría a partir de la Tabla 27.1 resultando en la Tabla 27.2.

Tabla 27.2: Tabla de verosimilitud - Tipo de día vs Jugar partido

	SI	NO	$P(\text{Tipo de día}_i)$
Soleado	2	4	$\frac{6}{15} = 0,40$
Nublado	4	0	$\frac{4}{15} = 0,27$
Lluvia	4	1	$\frac{5}{15} = 0,33$
$P(\text{Jugar})$	$\frac{10}{15} = 0,67$	$\frac{5}{15} = 0,33$	

A partir de la Tabla 27.2 se obtiene la probabilidad de cada tipo de día dado que con esa climatología se jugó o no, es decir, $P(\text{Tipo de día} | \text{Jugar})$. Obteniendo las probabilidades mostradas en la Tabla 27.3

Tabla 27.3: Probabilidad de Tipo de día sabiendo si se jugó el partido

	$c = \text{SI}$	$c = \text{NO}$
$P(\text{Soleado} C=c)$	$\frac{2}{15}$	$\frac{4}{15}$
$P(\text{Nublado} C=c)$	$\frac{4}{15}$	$\frac{0}{15}$
$P(\text{Lluvia} C=c)$	$\frac{4}{15}$	$\frac{1}{15}$

Este proceso se repite de forma independiente para las variables *viento* y *humedad* obteniendo la Tabla 27.4 y la Tabla 27.5 respectivamente.

Tabla 27.4: Probabilidad fuerza del Viento sabiendo si se jugó el partido

	$c = \text{SI}$	$c = \text{NO}$
$P(\text{Débil} C=c)$	$\frac{6}{15}$	$\frac{2}{15}$
$P(\text{Fuerte} C=c)$	$\frac{4}{15}$	$\frac{3}{15}$

27.4. Procedimiento con R: la función `naive_bayes()`

461

Tabla 27.5: Probabilidad nivel de Humedad sabiendo si se jugó el partido

	c = SI	c = NO
P(Débil C=c)	$\frac{6}{15}$	$\frac{1}{15}$
P(Fuerte C=c)	$\frac{4}{15}$	$\frac{4}{15}$

Finalmente, aplicando el teorema de Bayes se podría predecir si se juega o no el partido ante la previsión de un nuevo día. Por ejemplo, ¿cuál es la probabilidad de no jugar al tenis si el día se espera soleado, con fuertes rachas de viento y escasa humedad? Esto es, $\ell=[\text{Soleado}, \text{Fuerte}, \text{Débil}]$ y, de acuerdo al teorema de Bayes, esta pregunta se respondería a través de:

$$P(c|\ell) = \frac{P(\ell|c) \cdot P(c)}{P(\ell)}$$

A partir de las probabilidades previamente obtenidas y de la asunción de independencia entre las variables, se puede calcular la probabilidad de jugar como:

$$P(Si|\ell) = P(\text{Soleado}|Si) \cdot P(\text{Fuerte}|Si) \cdot P(\text{Débil}|Si) \cdot P(Si) = \frac{2}{15} \frac{4}{15} \frac{2}{15} \frac{10}{15} = 0,0032 \quad (27.7)$$

$$P(No|\ell) = P(\text{Soleado}|No) \cdot P(\text{Fuerte}|No) \cdot P(\text{Débil}|No) \cdot P(No) = \frac{4}{15} \frac{3}{15} \frac{1}{15} \frac{5}{15} = 0,0012 \quad (27.8)$$

La probabilidad de jugar es superior a la probabilidad de no jugar y, por tanto, dado un día con esas condiciones climáticas se clasificará como un día en el que se puede jugar.

27.4. Procedimiento con R: la función `naive_bayes()`

En el paquete `naivebayes` de R se encuentra la función `naive_bayes()` que se utiliza para entrenar un modelo *Naive Bayes*:

```
naive_bayes(formula, data, prior = ..., ...)
```

- **formula**: refleja la relación lineal entre la variable dependiente y los predictores $Y \sim X_1 + \dots + X_p$.
- **data**: conjunto de datos con el que se entrena el modelo.
- **prior**: vector con las probabilidades a priori de las clases.

27.5. Clasificación de clientes utilizando el modelo *Naive Bayes*

Como en los capítulos precedentes, en este ejemplo se pretende entrenar un modelo *Naive Bayes* utilizando el conjunto de datos de compras realizadas por clientes incluido en el paquete CDR. Este conjunto de datos cuenta con unas variables predictoras que indican qué productos han comprado los clientes, el importe que han gastado y otras características como su edad y su nivel educativo. Se utiliza el conjunto de datos sin transformar (`dp_entr`), es decir, en su escala original y con las variables categóricas sin codificar. La variable objetivo indica si un cliente comprará o no el nuevo producto (*tensiómetro digital*).

```
library("caret")
library("naivebayes")
library("reshape")
library("ggplot2")
library("CDR")

data("dp_entr")

# se fija la semilla aleatoria
set.seed(101)

# se entrena el modelo
model <- train(CLSPRO_pro13 ~ .,
                 data=dp_entr,
                 method="nb",
                 metric="Accuracy",
                 trControl=trainControl(classProbs = TRUE,
                                         method = "cv",
                                         number = 10))

# se muestra la salida del modelo
model
```

```
Naive Bayes

558 samples
17 predictor
2 classes: 'S', 'N'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 502, 502, 502, 503, 503, 502, ...
Resampling results across tuning parameters:

  usekernel Accuracy Kappa
  FALSE      0.8512662 0.7026716
  TRUE       0.8512338 0.7025165
```

27.5. Clasificación de clientes utilizando el modelo Naive Bayes

463

```
Tuning parameter 'fL' was held constant at a value of 0
Tuning parameter
  'adjust' was held constant at a value of 1
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were fL = 0, usekernel = FALSE and adjust = 1.
```

Los resultados del proceso de entrenamiento muestran que, en este caso, es indiferente indicar el argumento `usekernel` como FALSE o TRUE, los resultados de precisión son equivalentes. El resumen del modelo muestra que la precisión media obtenida durante la validación cruzada alcanza el 85,1 %, lo cual indica que el modelo ajusta bastante bien la intención de compra de nuevos clientes.

```
confusionMatrix(model)
Cross-Validated (10 fold) Confusion Matrix

(entries are percentual average cell counts across resamples)

      Reference
Prediction   S   N
      S 41.8  6.6
      N  8.2 43.4

Accuracy (average) : 0.8513
```

En la matriz de confusión del modelo se observa para cada celda el promedio porcentual entre remuestreos. Así, se observa que en media el modelo predice mejor cuando un cliente no va a comprar el nuevo producto que cuando sí lo hace, aunque no con mucha diferencia (menos de un 2 %). En ambos casos, las clasificaciones erróneas no suponen ni el 10 %.

```
ggplot(melt(model$resample[,-4]), aes(x = variable, y = value, fill=variable)) +
  geom_boxplot(show.legend=FALSE) +
  xlab(NULL) + ylab(NULL)
```

Se puede observar como la precisión oscila entre el 75 % y el 95 %, aunque en uno de los resultados se obtuvo un 96 % de precisión, el cual se marca como un resultado atípico.

Resumen

En este capítulo se introduce al lector en el algoritmo de *Naive Bayes*, en concreto:

- Se presentan los fundamentos del algoritmo bayesiano, particularmente el Teorema de Bayes.
- Se explica el funcionamiento del algoritmo *Naive Bayes* y su relación con dicho Teorema de Bayes.
- Se demuestra su aplicabilidad a un caso real de clasificación a través de R.

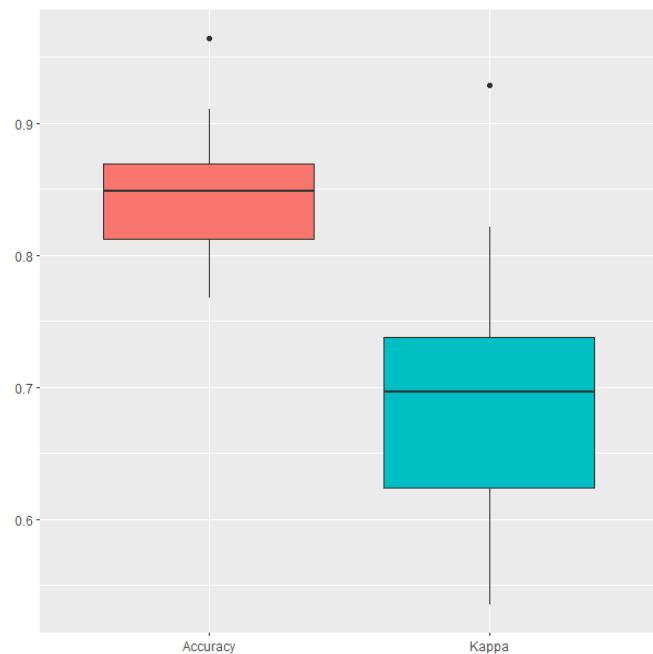


Figura 27.1: Resultados del modelo Naive Bayes obtenidos durante el proceso de validación cruzada.

Capítulo 28

Métodos ensamblados: bagging y random forest

Ramón A. Carrasco^a e Itzcóatl Bueno^{b,a}

^aUniversidad Complutense de Madrid ^bInstituto Nacional de Estadística

28.1. Introducción a los métodos ensamblados

Puede ocurrir que ninguno de los algoritmos hasta ahora presentados (Caps. 24, 25, 26 y 27) proporcionen resultados convincentes para el problema que se quiere modelar. El *aprendizaje ensamblado* (Zhou, 2012) es un paradigma que, como muestra la Fig. 28.1, en lugar de entrenar un modelo muy preciso, se centra en entrenar un gran número de modelos con menor precisión, y después combinar sus predicciones para obtener un metamodelo de una precisión más alta.

A los modelos de menor precisión se les suele nombrar como algoritmos “débiles”, es decir, algoritmos con menor capacidad de aprender patrones complejos en los datos. Por tanto, generalmente, son rápidos tanto en tiempo de entrenamiento como de procesamiento. Existen dos paradigmas de aprendizaje ensamblado: el *bagging* y el *boosting* (Cap. 29).

28.2. Bagging

En lugar de buscar la división más eficiente en cada capa, como ocurre en el árbol de decisión, una alternativa sería construir un metamodelo combinando los resultados de múltiples árboles de decisión. Esta técnica se conoce como *bagging* y consiste en construir varios árboles utilizando una selección aleatoria de los datos que se utilizan para cada árbol y, finalmente, combinar la predicción de cada uno de ellos a través de la media (en el caso de regresión) o mediante un sistema de votación (en el caso de un problema de clasificación).

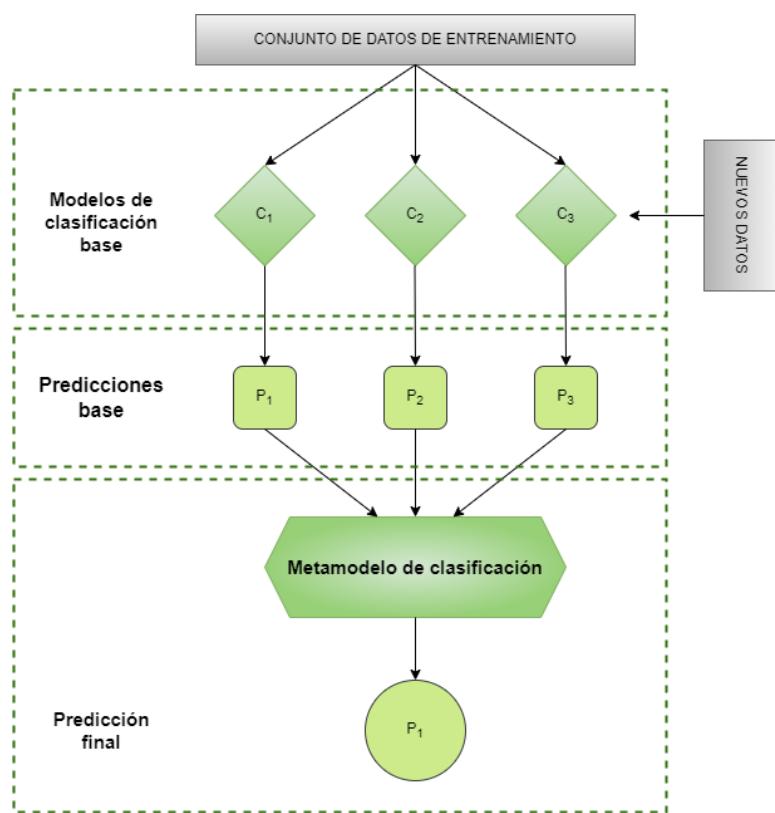


Figura 28.1: Esquema de un metamodelo.

La principal característica del *bagging* es el llamado muestreo bootstrap. La idea básica del bootstrap es que la inferencia sobre una población se haga a partir de una muestra, tomando el papel de población y se remuestree, permitiendo comparar valor poblacional y el valor muestral. En el *bagging*, el objetivo de este remuestreo es que cada árbol esté entrenado con una muestra única, y por tanto, generen respuestas únicas, esto es modelos débiles distintos. Para ello, debe existir aleatoriedad y variación en cada árbol que conforme el modelo final, puesto que no tendría sentido construir varios árboles idénticos. Como se ha comentado, este problema queda resuelto por el muestreo bootstrap, el cual extrae una muestra aleatoria de los datos en cada ronda. En el caso del *bagging*, se extraen distintas muestras de datos para el entrenamiento de cada árbol. Aunque esto no elimina la problemática del sobreajuste, los patrones presentes en el conjunto de datos aparecerán en la mayoría de los árboles entrenados y, por tanto, en la predicción final. Es por ello que el *bagging* es una técnica de gran eficacia para el tratamiento de los valores atípicos y para la reducción de la varianza que generalmente afecta a un modelo compuesto por un único árbol de decisión.

28.2.1. Procedimiento con R: la función `bagging()`

En el paquete `ipred` de R se encuentra la función `bagging()` que se utiliza para entrenar un modelo *bagging*:

```
bagging(formula, data, ...)
```

- **formula:** Refleja la relación lineal entre la variable dependiente y los predictores $Y \sim X_1 + \dots + X_p$.
- **data:** Conjunto de datos con el que se entrena el modelo.
- **nbagg:** Número de replicaciones bootstrap.
- **coob:** Indica si se debe calcular una estimación del ratio de error de predicción.

28.2.2. Implementando *bagging* en R

Es posible la implementación de un modelo de predicción de agregación bootstrap en R. Para ello, se pueden utilizar múltiples funciones como la ya mencionada en la Sec. 28.2.1 `bagging()`. En este ejemplo se utilizan los datos sobre compras de clientes `dp_entr` del paquete CDR, cuyo objetivo es clasificar a los clientes entre quienes comprarían un nuevo producto y quienes no.

```
library("CDR")
library("ipred")
library("caret")
library("reshape")
library("ggplot2")

data("dp_entr")
```

```
# se fija la semilla aleatoria
set.seed(101)

# Se entrena el modelo
bag_model <- bagging(
  formula = CLS_PRO_pro13 ~ .,
  data = dp_entr,
  nbagg = 100,
  coob = TRUE,
  control = rpart.control(minsplit = 2, cp = 0)
)

bag_model

Bagging classification trees with 100 bootstrap replications

Call: bagging.data.frame(formula = CLS_PRO_pro13 ~ ., data = dp_entr,
  nbagg = 100, coob = TRUE, control = rpart.control(minsplit = 2,
  cp = 0))

Out-of-bag estimate of misclassification error:  0.1416
```

El error de clasificación de este modelo es del 14,16 %, o lo que es equivalente, el modelo tiene una precisión del 85,84 %. Desafortunadamente, `bagging()` no selecciona el número óptimo de replicaciones reduciendo el error de clasificación. Para seleccionar el número de replicaciones que minimice el error, se puede graficar la curva de error por número de replicaciones como en la Fig. 28.2. Se itera el modelo variando los valores del hiperparámetro `nbagg` (en este ejemplo entre 10 y 150, incrementándose de cinco en cinco). Se observa que el error mínimo (13,79 %) se obtiene al establecer el hiperparámetro igual a 60.

```
missclass <- c() # vector vacío para recopilar el error en cada iteración
for (n in seq(10,150,5)) { # valores a probar para nbagg
  # se establece la semilla aleatoria
  set.seed(101)
  # se entrena el modelo
  bag_model <- bagging(
    formula = CLS_PRO_pro13 ~ .,
    data = dp_entr,
    nbagg = n,
    coob = TRUE,
    control = rpart.control(minsplit = 2, cp = 0)
  )
  # se agrega el error de esta iteración
  missclass <- c(missclass, bag_model$err) # se agrega el error de esta iteración
}
```

28.2. Bagging

469

```
plot(seq(10,150,5),missclass,type = "l",xlab = "Número de árboles",
  ylab="Missclassification error")
```

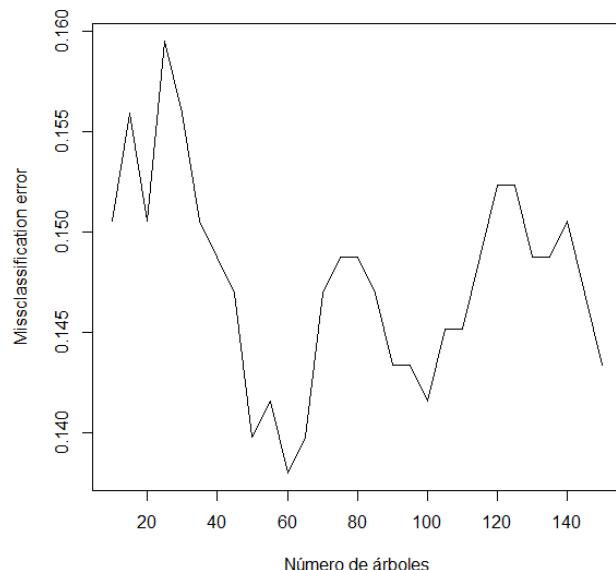


Figura 28.2: Número de replicaciones vs Error de clasificación.

La función `train()` del paquete `caret` es otro método para entrenar un algoritmo de *bagging* en R. Para ello, el argumento `method` debe tomar el valor "`treebag`". Sin embargo, este algoritmo no incluye hiperparámetros a optimizar. Dado que se ha obtenido recursivamente el número óptimo de replicaciones, se puede entrenar el modelo con el valor obtenido y comprobar que el error dado coincide. Se observa que si se entrena un modelo *bagging* con 60 replicaciones, la precisión del modelo es del 86,93 %. Esto es aproximadamente el resultado obtenido anteriormente en el que para 60 replicaciones el modelo tenía un error de clasificación del 13,79 %.

```
set.seed(101)
model_bag <- train(
  CLS_PRO_pro13 ~ .,
  data = dp_entr,
  method = "treebag",
  trControl = trainControl(method = "cv", number = 10),
  nbagg = 60,
  control = rpart.control(minsplit = 2, cp = 0)
)
```

```

model_bag

Bagged CART

558 samples
17 predictor
2 classes: 'S', 'N'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 502, 502, 502, 503, 503, 502, ...
Resampling results:

Accuracy   Kappa
0.8692532  0.7385449

```

28.2.3. Interpretación de variables en el *bagging*

Una de las principales desventajas de los algoritmos ensamblados (incluido el *bagging*) es que mientras que los modelos base son interpretables, el metamodelo resultante no lo es. Pese a esto, aún es posible hacer inferencia de cómo cada una de las variables influye en el modelo entrenado. La manera de medir la importancia de las variables incluidas en un árbol es registrar para cada variable la reducción de la función de pérdida que se le atribuye en cada partición. Dado que una variable puede utilizarse varias veces para dividir el árbol, la importancia total de esa variable será la suma de la reducción de la función de pérdida que se le atribuya por todas las particiones en las que intervenga. Este proceso es similar para el *bagging*. En este caso, para cada árbol se calcula la reducción de la función de pérdida en todas las divisiones. Tras esto, se agrega esta medida en todos los árboles que forman el metamodelo. El paquete `ipred`, en el que se encuentra la función `bagging()`, no captura la información requerida para calcular la importancia de las variables. Sin embargo, el paquete `caret` sí lo hace y se puede construir un gráfico de importancia utilizando la función `vip()` del paquete `vip`.

```

library("vip")
vip(model_bag, num_features = 15,
    aesthetics = list(color = "skyblue", fill = "skyblue"))

```

La Fig. 28.3 muestra que las variables más importantes en el modelo *bagging* entrenado para predecir si un cliente comprará o no el *tensiómetro digital* son: si ha comprado la *depiladora eléctrica*, cuánto importe ha gastado en ese producto, si ha comprado el *estimulador muscular* y si ha comprado el *smartwatch fitness*.

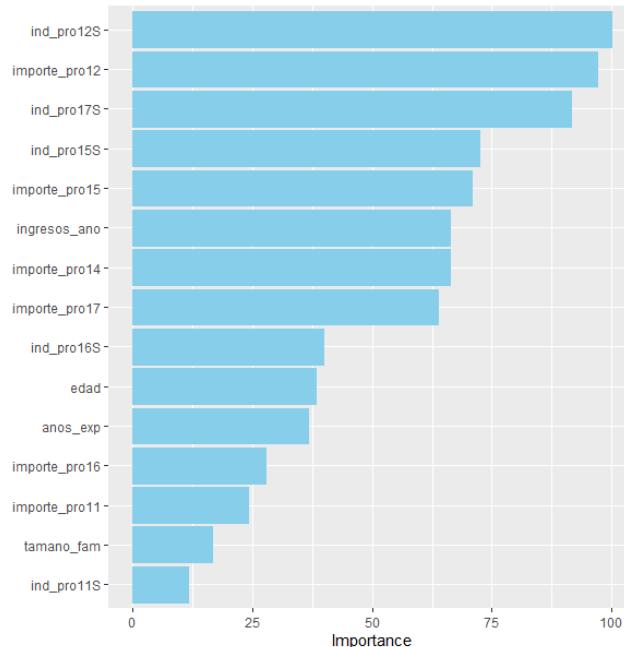


Figura 28.3: Importancia de las variables incluidas en el modelo bagging.

28.3. Random Forest

El *bagging* es el paradigma tras el algoritmo de *random forest*. Este algoritmo fue desarrollado por primera vez por (Ho, 1995). Sin embargo, fueron (Cutler and Zhao, 1999) y (Breiman, 2001) quienes desarrollaron una versión extendida del modelo y registraron **Random Forest** como marca comercial. Este algoritmo básico de *bagging* funciona del siguiente modo: a partir del conjunto de datos de entrenamiento se generan K muestras aleatorias \mathbb{S}_k , se entrena un modelo de árbol de decisión (f_k) utilizando la muestra \mathbb{S}_k como conjunto de entrenamiento. Tras el entrenamiento, se dispone de K árboles de decisión, como se observa en la Fig. 28.4. La predicción de una nueva observación x se obtiene como la media de las K predicciones:

$$y \leftarrow \hat{f}(x) = \frac{1}{K} \sum_{k=1}^K f_k(x) \quad (28.1)$$

En el caso de regresión, o por la votación por mayoría en el caso de clasificación.

Tanto el *bagging* como el *random forest* desarrollan múltiples árboles y utilizan el muestreo bootstrap para la aleatorización de los datos. Sin embargo, el *random forest* establece una limitación artificial a la selección de variables al no considerar todas en cada árbol.

El *bagging* considera las mismas variables para construir cada árbol con el objetivo de minimizar su entropía, y, por tanto, todos los árboles suelen tener un aspecto similar. Esto lleva a que las

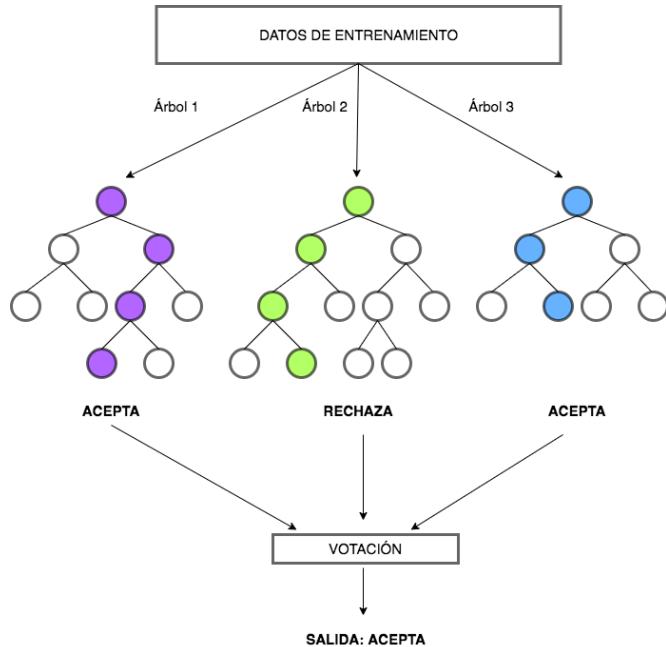


Figura 28.4: Ejemplo de Random Forest.

predicciones dadas por los árboles estén altamente correlacionadas. El modelo *random forest* evita este problema estableciendo la obligación, en cada división, de utilizar un subconjunto de las variables. Esto proporciona a algunas variables mayor probabilidad de ser seleccionadas, y al generar árboles únicos y no correlacionados se consigue una estructura de decisión final más fiable.

En general, es mejor que el *random forest* esté formado por una gran cantidad de árboles (por lo menos 100) para suavizar el impacto de valores atípicos. Sin embargo, la tasa de efectividad disminuye a medida que se incorporan más árboles. Llegado a cierto punto, los nuevos árboles no aportan una mejora significativa al modelo, pero si incrementan los tiempos de procesamiento.

El modelo *random forest* es rápido de entrenar y es una buena técnica para obtener un modelo de referencia. Aunque estos modelos funcionan bien en la interpretación de patrones complejos y son versátiles, otras técnicas, como por ejemplo el *gradient boosting* (Cap. 29), proporcionan una mayor precisión en las predicciones en muchos casos.

Estos modelos se han vuelto populares porque tienden a proporcionar un muy buen rendimiento con los parámetros predeterminados en las distintas implementaciones. En efecto, a pesar de tener muchos hiperparámetros que pueden ser ajustados, los valores por defecto de dichos hiperparámetros tienden a ofrecer buenos resultados en la predicción. Los hiperparámetros más importantes que hay que ajustar al entrenar un modelo *random forest* son: el número de árboles (K), el número de variables incluidos en el subconjunto aleatorio en cada división (`mtry`), la complejidad de cada árbol, el esquema de muestreo y la regla de división a utilizar durante la construcción del árbol.

28.3.1. Número de árboles (K)

El primer hiperparámetro a ajustar es el número de árboles que componen el modelo de *random forest*. Su valor debe ser lo suficientemente grande como para que la tasa de error se estabilice. La regla general es que el valor mínimo de árboles sea igual a 10 veces el número de variables incluidas en el modelo. Sin embargo, cuando se tienen en cuenta otros hiperparámetros para optimizar, es posible que el número de árboles se vea afectado. El tiempo de procesamiento aumenta linealmente con la cantidad de árboles incluidos, pero cuantos más se incluyan, se obtendrán estimaciones de error más estables.

28.3.2. Número de variables a considerar (`mtry`)

`mtry` se refiere al hiperparámetro encargado de controlar la aleatorización de variables utilizadas para las particiones de los árboles. Este hiperparámetro ayuda a equilibrar la baja correlación del árbol con los demás, y una razonable fuerza predictiva. Existe un valor predeterminado para este hiperparámetro el cual se puede utilizar en caso de no querer o no poder ajustarlo. En el caso de la regresión, se determina que $mtry = \frac{p}{3}$ siendo p el número de variables incluidas en el modelo. Y en problemas de clasificación, el valor predeterminado es $mtry = \sqrt{p}$. Cuando hay pocas variables relevantes, es decir, los datos son muy ruidosos, tiende a funcionar mejor que el valor de `mtry` sea alto, pues hace que sea más probable seleccionar esas variables. En cambio, cuando muchas variables son importantes, funciona mejor un valor bajo de `mtry`.

28.3.3. Complejidad de los árboles

Un modelo *random forest* se construye con árboles de decisión a los que se les puede controlar su profundidad y su complejidad como se vio en el Cap. 24. Esto se puede hacer ajustando los hiperparámetros de profundidad máxima permitida, tamaño del nodo o la cantidad máxima de nodos terminales.

El tamaño del nodo es el hiperparámetro más común para controlar la complejidad del árbol y la mayoría de las implementaciones usan los valores predeterminados de 1 para árboles de clasificación y 5 para los árboles de regresión, dado que estos valores tienden a producir buenos resultados. Si se quiere controlar el tiempo de procesamiento, se pueden conseguir reducciones significativas del tiempo aumentando el tamaño del nodo impactando de manera marginal en la estimación del error.

28.3.4. Esquema de muestreo

Por defecto, el *random forest* tiene como esquema de muestreo el bootstrapping, explicado anteriormente, en el cual todas las observaciones se muestran con reemplazo. Todas las repeticiones de bootstrap tienen el mismo tamaño que el conjunto de datos de entrenamiento. Sin embargo, el esquema de muestreo se puede ajustar tanto en el tamaño de la muestra como en el diseño muestral (con o sin reposición). El hiperparámetro de tamaño de muestra determina cuántas observaciones se extraen para el entrenamiento de cada árbol. Cuanto menor sea el

tamaño muestral, menor será la correlación entre los árboles, lo cual puede llevar a mejores resultados de precisión en la predicción. La forma de determinar el tamaño muestral óptimo puede hallarse evaluando algunos valores que oscilen entre el 25 % y el 100 %, y en el caso de que haya variables no balanceadas respecto a los valores de las categóricas se puede intentar muestrear sin reposición.

28.3.5. Regla de división

Por defecto, la regla de división que utilizan los árboles de decisión que conforman un *random forest* es la que se presentó en el Cap. 24. Esto es, en el caso de regresión seleccionar la división que minimiza la desviación típica (σ); y en el caso de clasificación la división que minimiza la impureza de Gini o la entropía.

28.3.6. Procedimiento con R: la función `randomForest()`

En el paquete `randomForest` de R se encuentra la función `randomForest()` que se utiliza para entrenar un modelo de este tipo:

```
randomForest(formula, data=..., ...)
randomForest(x, y, xtest, ytest, ntree=500, mtry, ...)
```

- **formula:** Refleja la relación entre la variable dependiente Y y los predictores tal que $Y \sim X_1 + \dots + X_p$.
- **data:** Conjunto de datos con el que entrenar el árbol de acuerdo a la fórmula indicada.
- **x:** Conjunto de datos de entrenamiento que contiene los predictores
- **y:** Vector respuesta con las clases o valores de la variable respuesta.
- **xtest:** Conjunto de datos que contiene los predictores del conjunto de datos de validación.
- **ytest:** Variable respuesta del conjunto de datos de validación.
- **ntree:** Número de árboles a construir en el modelo.
- **mtry:** Número de variables muestreadas aleatoriamente como candidatas en cada partición.

28.3.7. Aplicación del modelo *random forest* en R

En esta sección se aplica el modelo *random forest* al ejemplo de datos de retail incluido en el paquete CDR. Se carga el paquete y con ello, los datos `dp_entr`. Se busca predecir si un cliente va a comprar o no el nuevo producto de acuerdo a los productos que ha consumido, el importe que gasta en ellos y otras características como, por ejemplo, su nivel educativo.

```
library("CDR")
library("randomForest")
library("caret")
library("reshape")
```

```
library("ggplot2")
data(dp_entr)
```

Este algoritmo al estar basado en árboles de clasificación tiene los mismos requisitos para el entrenamiento que tenían dichos árboles, así se construye el modelo usando el conjunto de datos de entrenamiento.

```
# se fija la semilla aleatoria
set.seed(101)

# se entrena el modelo
model <- train(CLSPRO_pro13~, data=dp_entr_NUM,
                 method="rf", metric="Accuracy", ntree=500,
                 trControl=trainControl(method="cv",
                                         number=10,
                                         classProbs = TRUE))
```

```
model
Random Forest

558 samples
19 predictor
 2 classes: 'S', 'N'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 502, 502, 502, 503, 503, 502, ...
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
    2    0.8602922  0.7206238
   10   0.8620455  0.7241029
   19   0.8620130  0.7240248

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 10.
```

Los resultados de la validación cruzada se pueden ver en el siguiente boxplot. Se observa como la precisión oscila entre el 80 % y el 95 %. Además, se puede ver en el resultado del modelo que el hiperparámetro *mtry* se ha ajustado a 10 variables.

Finalmente, aunque el *random forest* generado está compuesto por 500 árboles, se puede acceder a cualquiera de ellos para estudiarlos en profundidad. Para ello, es necesario instalar el paquete `reptrree` desde el repositorio <https://github.com/araastat/reptrree>.

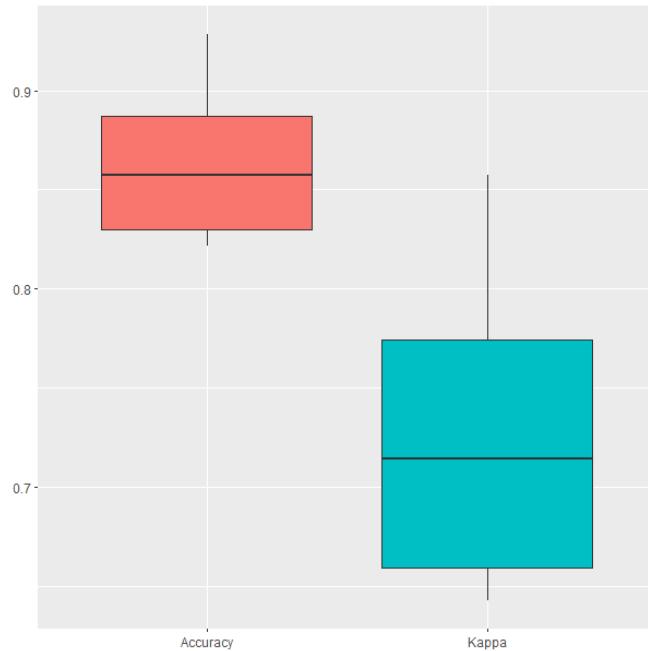


Figura 28.5: Resultados del modelo random forest durante el proceso de validación cruzada.

```
library("devtools")
if(!('reprtree' %in% installed.packages())){
  devtools::install_github('araastat/reprtree')
}
```

Se pueden observar las decisiones que se toman en el árbol de forma tabulada, indicando qué variable se utiliza para la partición, cuál es el valor que decide la división, indicando si es un nodo terminal (-1) o no (1) y la predicción del nodo, el cual es NA si no es un nodo terminal.

```
set.seed(101)
rf <- randomForest(CLSPRO_pro13~, data = dp_entr_NUM, ntree=500,
                     mtry=unlist(model$bestTune))

# se observa el árbol número 205
tree205 <- getTree(rf, 205, labelVar=TRUE)

head(tree205[,-c(1,2)])
  split var split point status prediction
1 importe_pro15      100     1      <NA>
2 importe_pro12       60      1      <NA>
3 importe_pro16       90      1      <NA>
4 ingresos_ano    156500     1      <NA>
```

28.3. Random Forest

477

```

5 importe_pro17      150     1      <NA>
6     anos_exp       33      1      <NA>

tail(tree205[,-c(1,2)])
  split var split point status prediction
120      <NA>      0.0    -1        N
121      <NA>      0.0    -1        S
122 des_nivel_edu.BASICO  0.5     1      <NA>
123      <NA>      0.0    -1        S
124      <NA>      0.0    -1        S
125      <NA>      0.0    -1        N

```

Este árbol se muestra en la Fig. 28.6.

```

library("rpart")
plot.getTree(rf, k=205)

```

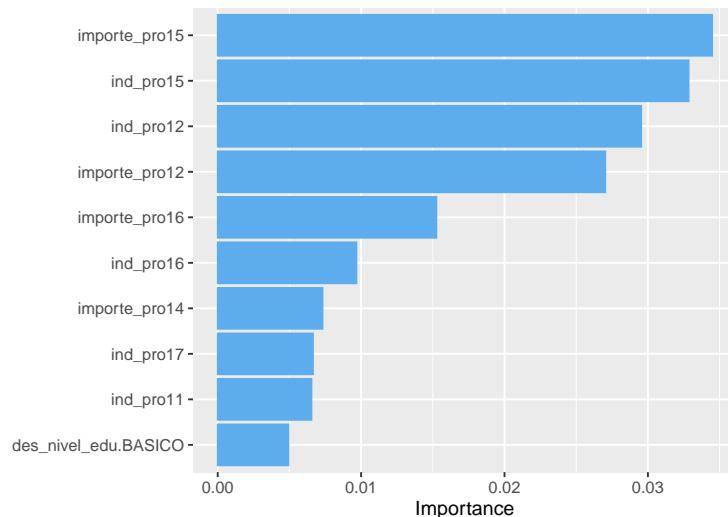


Figura 28.6: Árbol número 205 del random forest entrenado.

Sin embargo, el método por el que se representa gráficamente no es muy claro y puede llevar a confusión o dificultar la interpretación del árbol. Si se desea estudiar hasta cierto nivel del árbol, se puede incluir el argumento `depth` como en el ejemplo abajo mostrado, y que representa el mismo árbol con una profundidad de 5 ramas en la Fig. 28.7.

```

plot.getTree(rf, k=205, depth = 5)

```

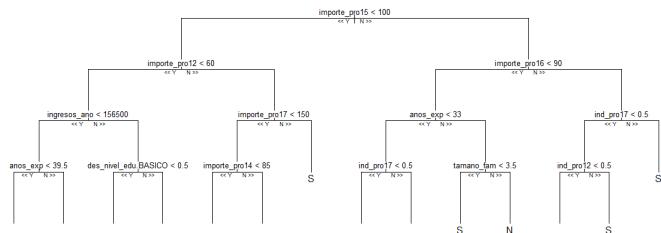


Figura 28.7: Árbol número 205 del random forest entrenado hasta la capa 5.

28.3.7.1. Aplicación del modelo *random forest* con ajuste automático

En este segundo ejemplo, se pretende mejorar la precisión del modelo anterior. Para ello, se ajusta de forma automática los hiperparámetros de dicho algoritmo. De los mencionados anteriormente, solo se va a ajustar automáticamente `mtry`, que es el único incluido el método `rf`.

```

modelLookup("rf")
  model parameter           label forReg forClass probModel
1   rf      mtry #Randomly Selected Predictors  TRUE     TRUE     TRUE

```

Para ajustar el número de árboles y el resto de hiperparámetros, se puede iterar el modelo y probar distintos valores. En una red de opciones se incluyen los valores a probar para el hiperparámetro `mtry`.

```

# Se especifica un rango de valores posibles de mtry
tuneGrid <- expand.grid(mtry = 1:18)

```

A continuación, se entrena el modelo para que se ajuste al valor de `mtry` que maximice el rendimiento predictivo del modelo.

```

# se fija la semilla aleatoria
set.seed(101)

# se entrena el modelo
model <- train(CLSPRO_pro13 ~ ., data=dp_entr_NUM,
                 method = "rf", metric = "Accuracy",
                 tuneGrid = tuneGrid,
                 trControl = trainControl(classProbs = TRUE))

```

```

model

Random Forest

558 samples
19 predictor
2 classes: 'S', 'N'

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 558, 558, 558, 558, 558, 558, ...
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
  1     0.8641354  0.7283098
  2     0.8650087  0.7298376
  3     0.8629614  0.7256812
  4     0.8635609  0.7268514
  5     0.8639559  0.7276250
  6     0.8612659  0.7222420
  7     0.8604934  0.7206476
  8     0.8610116  0.7216937
  9     0.8590645  0.7177882
  10    0.8589073  0.7174718
  11    0.8607248  0.7211179
  12    0.8583609  0.7163903
  13    0.8587296  0.7170933
  14    0.8587384  0.7171642
  15    0.8583195  0.7163106
  16    0.8585407  0.7167355
  17    0.8573597  0.7144030
  18    0.8581404  0.7159558

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 2.

```

Mientras que en el ejemplo anterior el algoritmo sólo probó tres valores de `mtry`, esta vez se realiza una prueba exhaustiva de valores. En el primer ejemplo, el valor del hiperparámetro era `mtry=10`, pero ahora se ha reajustado a `mtry=2`. Esto es equivalente a decir que 2 variables seleccionadas en cada partición son suficientes, y que no son necesarias 10 como en el ejemplo anterior. Finalmente, se puede observar en la Fig. 28.8 los resultados obtenidos durante la validación cruzada. Se observa cómo no sólo la precisión es mayor que en el ejemplo anterior, sino que además los resultados tienen menos dispersión.

```

ggplot(melt(model$resample[,-4]), aes(x = variable, y = value, fill=variable)) +
  geom_boxplot(show.legend=FALSE) +
  xlab(NULL) + ylab(NULL)

```

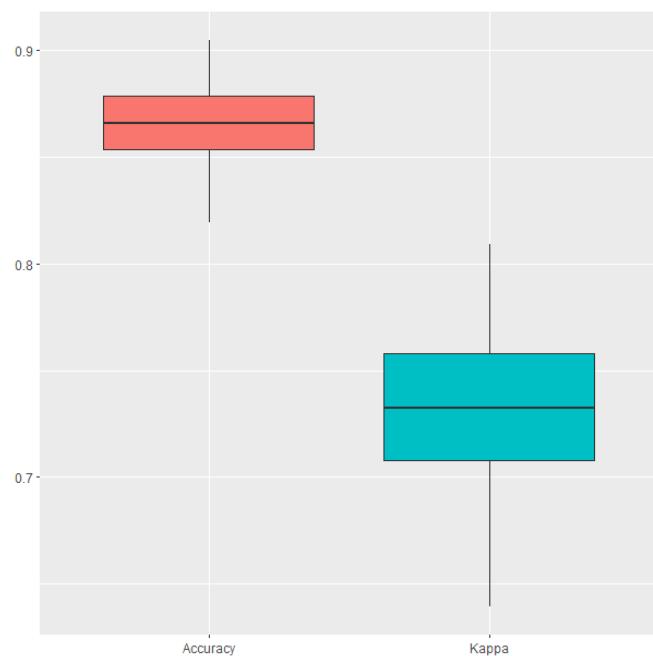


Figura 28.8: Resultados obtenidos por el random forest con ajuste automático durante el proceso de validación cruzada.

Resumen

En este capítulo se introduce al lector en el *bagging* y el algoritmo de aprendizaje supervisado conocido como *random forest*, en concreto:

- Se presenta el concepto de aprendizaje ensamblado, y se profundiza en uno de sus paradigmas: el *bagging*.
- Se implementa el *bagging* en R a través de un caso de clasificación binaria.
- Se expone cómo medir la importancia de las variables incluidas en un modelo *bagging* para facilitar su interpretación.
- Se explica el modelo *random forest*, fundamentado en los árboles decisión y en el *bagging*. Así como los hiperparámetros más importantes para ajustar el modelo de mayor precisión.
- Se presenta un ejemplo de clasificación binaria utilizando el modelo *random forest* en R.

Capítulo 29

Boosting y el algoritmo XGBoost

Ramón A. Carrasco^a e Itzcóatl Bueno^{b,a}

^aUniversidad Complutense de Madrid ^bInstituto Nacional de Estadística

29.1. Métodos ensamblados: bagging vs boosting

En el Cap. 28 se presentó la idea de métodos ensamblados. Una serie de modelos útiles cuando ningún modelo de aprendizaje supervisado es capaz de explicar bien la variable dependiente de interés. En el aprendizaje ensamblado se entrena un gran número de modelos con menor precisión, y se combinan sus predicciones para obtener un metamodelo de una precisión más alta. Los dos modelos de aprendizaje ensamblado más populares son el *bagging* (Cap. 28) y el *boosting*.

La principal diferencia entre ellos radica en cómo se combinan los modelos individuales para obtener una predicción final. Por ejemplo, si se quisiera organizar una fiesta y se tuviese que tomar una decisión sobre el tema para la decoración, se podría hacer tanto con *bagging* como con *boosting*. Si se utilizase el *bagging*, se pediría opinión a distintos grupos de amigos. Después, se combinarán todas sus ideas para tomar una decisión final sobre la decoración de la fiesta. Por otro lado, si se utiliza *boosting*, en lugar de preguntarle a diferentes grupos de amigos al mismo tiempo, se pediría a un amigo en particular su opinión. Si su respuesta no es convincente, entonces se busca la ayuda de otro amigo, quien se enfocará en mejorar la respuesta anterior. Este proceso continúa secuencialmente, solicitando la ayuda de diferentes amigos y construyendo sobre las respuestas anteriores para obtener una decisión final más precisa y refinada sobre la decoración de la fiesta.

29.2. ¿Qué es el boosting?

El *boosting* (Schapire and Freund, 2012) es el otro de los paradigmas de aprendizaje ensamblado, presentado en el Cap. 28. Como el *bagging*, el *boosting* agrega múltiples modelos con menor precisión (débiles) combinando sus predicciones para obtener un metamodelo con una precisión más alta. Los árboles de decisión son los modelos base o débiles que se usan más frecuentemente. En este caso, para llegar al metamodelo a partir de los modelos base, es necesario introducir ponderaciones a los árboles basándose en las clasificaciones erróneas del árbol entrenado previamente a dicho árbol.

El *boosting* reduce el problema del sobreajuste utilizando menos árboles que un modelo *random forest*. Mientras que agregar más árboles al *random forest* ayuda a compensar el sobreajuste, también puede llevar a un aumento del mismo y, por ello, hay que ser cauteloso a la hora de agregar nuevos árboles. Sin embargo, el *boosting* aprende iterativamente de los errores en árboles anteriores, pudiendo llevar a sobreajustar el modelo. Aunque este enfoque produce predicciones más precisas, muchas veces mejores a la mayoría de algoritmos, puede llevar a ajustar las observaciones atípicas. Es por esto que el *random forest* es una técnica más recomendada cuando se trabaja con conjuntos de datos muy complejos con un gran número de observaciones atípicas.

Otra de las grandes desventajas del *boosting* es que su tiempo de procesamiento es muy elevado, puesto que su entrenamiento sigue una lógica secuencial. En el proceso de entrenamiento, un árbol debe esperar a que el inmediatamente anterior sea entrenado, para iniciar su entrenamiento, y esto limita la escalabilidad del modelo. Mientras tanto, un *random forest* entrena los árboles en paralelo, lo que hace que su tiempo de procesamiento sea más rápido.

Tanto los algoritmos de *boosting* como a los de *bagging* presentan el inconveniente de la dificultad de interpretación que tienen respecto a los árboles de decisión. En este aspecto estos algoritmos de ensamblado se pueden considerar como de caja negra.

29.3. Gradient Boosting (GB)

Uno de los algoritmos de *boosting* más conocidos es el **gradient boosting**. Mientras que el *random forest* seleccionaba combinaciones aleatorias de variables en cada proceso de construcción de un árbol, el *gradient boosting* selecciona variables que mejoren la precisión con cada nuevo árbol. Por lo tanto, la construcción del modelo es secuencial, puesto que cada nuevo árbol se construye utilizando información derivada del árbol anterior y, en consecuencia, la construcción de estos árboles no son independientes. En cada iteración se registran los errores cometidos en los datos de entrenamiento y se tienen en cuenta para la siguiente ronda de entrenamiento. Además, se incorporan ponderaciones, como se observa en la Fig. 29.1 a los datos basándose en los resultados de la iteración anterior. Las ponderaciones más altas se aplicarán a las observaciones que fueron erróneamente clasificadas, y no se dará tanta atención a las bien clasificadas. Este proceso se repite hasta que se llega a un nivel bajo de error. El resultado final se obtiene a través de la media ponderada de las predicciones de los árboles de decisión.

Matemáticamente, un algoritmo *gradient boosting* para clasificación sigue los pasos que a continuación se detallan. Sea un problema de clasificación binaria y, asumiendo que se tienen K

29.3. Gradient Boosting (GB)

485

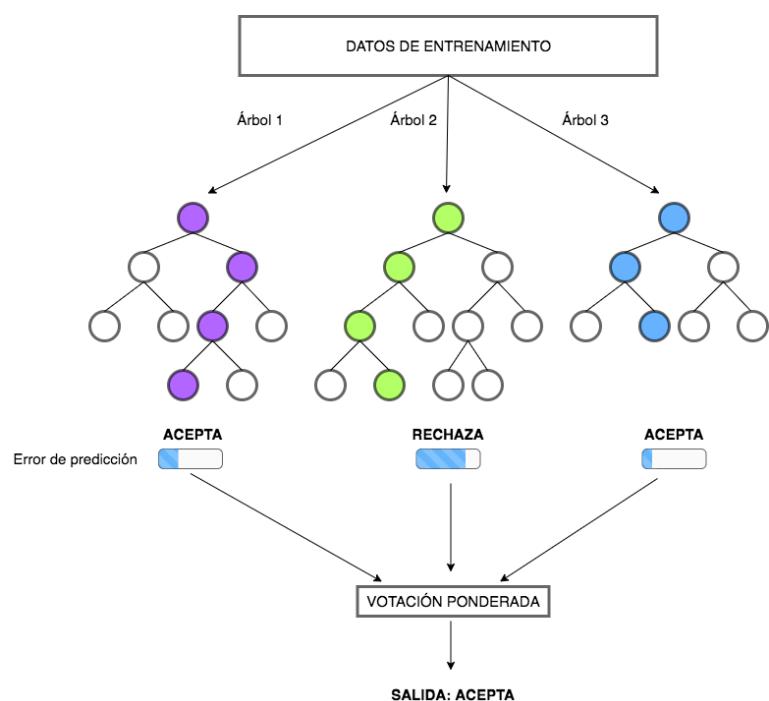


Figura 29.1: Ejemplo de Boosting.

árboles de decisión de clasificación, la predicción del modelo ensamblado se obtiene utilizando la función sigmoidal, como en la regresión logística (Cap. 16), tal que:

$$P(y = 1|x, f) = \frac{1}{1 + e^{-f(x)}} \quad (29.1)$$

Donde $f(x) = \sum_{\kappa=1}^K f_\kappa(x)$ y f_m es un árbol de decisión. De nuevo, como en la regresión logística, se aplica el principio de máxima verosimilitud tratando de hallar una f que maximice $\mathcal{L}_f = \sum_{i=1}^N \ln(P(y_i = 1|x_i, f))$

El algoritmo, en origen, es un modelo constante de la forma $f = f_0 = \frac{p}{1-p}$ donde $p = \frac{1}{N} \sum_{i=1}^N y_i$. Tras cada iteración se añade un nuevo árbol f_κ al modelo. Para encontrar el mejor árbol f_κ , la primera derivada parcial g_i del modelo actual se obtiene para $i = 1, \dots, N$:

$$g_i = \frac{\delta \mathcal{L}_f}{\delta f} \quad (29.2)$$

Donde f es el modelo de clasificación ensamblado construido en la iteración previa. Se necesita obtener las derivadas de $\ln(P(y_i = 1|x_i, f))$ con respecto a f para todo i para poder calcular g_i . Nótese que:

$$\ln(P(y_i = 1|x_i, f)) = \ln\left(\frac{1}{1 + e^{-f(x_i)}}\right) \quad (29.3)$$

Y, por tanto, la derivada respecto a f es igual a:

$$\frac{\delta \ln(P(y_i = 1|x_i, f))}{\delta f} = \frac{1}{e^{f(x_i)} + 1} \quad (29.4)$$

Después, se reemplaza en el conjunto de entrenamiento la categoría original y_i por su correspondiente derivada parcial g_i y se construye un nuevo modelo f_κ utilizando el conjunto de entrenamiento transformado. Tras esto, se obtiene la actualización óptima (ρ_κ) como:

$$\rho_\kappa = \arg \max_{\rho} \mathcal{L}_{f+\rho f_\kappa} \quad (29.5)$$

Al terminar la iteración κ , se actualiza el modelo ensamblado f añadiendo el nuevo árbol f_κ :

$$f \leftarrow f + \alpha \rho_\kappa f_\kappa \quad (29.6)$$

Se itera hasta que $\kappa = K$, entonces el proceso se detiene y se obtiene el modelo ensamblado final f .

29.3.1. Hiperparámetros del modelo *gradient boosting*

Un modelo de *gradient boosting* tiene dos tipos de hiperparámetros:

- Hiperparámetros de *boosting*.
- Hiperparámetros del árbol.

29.3.1.1. Hiperparámetros de *boosting*

Los hiperparámetros de *boosting* son principalmente dos: el número de árboles y la tasa de aprendizaje.

El primero indica el número de árboles a construir y que, como se ha comentado, es importante optimizar para evitar el sobreajuste del modelo. A diferencia de los modelos *random forest* o *bagging*, en el *boosting* los árboles crecen en secuencia para que cada árbol corrija los errores del anterior. El número de árboles necesarios para que el modelo sea buen predictor puede verse incrementado en función de los valores que tomen los otros hiperparámetros.

La tasa de aprendizaje es el hiperparámetro con el que se determina la contribución de cada árbol en el resultado final y controla la rapidez con la que el algoritmo avanza por el descenso del gradiente, es decir, la velocidad a la que aprende. Este hiperparámetro toma valores entre 0 y 1, aunque los valores habituales oscilan entre 0,001 y 0,3. El modelo es más robusto a las características específicas de cada árbol, permitiendo una buena generalización, cuando la tasa de aprendizaje toma valores bajos. Estos valores también facilitan la parada temprana antes del sobreajuste del modelo. Sin embargo, utilizar estos valores vuelve al modelo más exigente computacionalmente y dificulta alcanzar el modelo óptimo con un número fijo de árboles. En resumen, cuanto menor sea este valor, más preciso puede ser el modelo, pero también requerirá más árboles en la secuencia.

29.3.1.2. Hiperparámetros de árbol

Los principales hiperparámetros de árbol son: la profundidad del árbol y el número mínimo de observaciones en nodos terminales, como se vio en el Cap. 24.

El primer hiperparámetro controla la profundidad de los árboles individuales. Los valores habituales de profundidad oscilan entre 3 y 8. Los árboles de menor profundidad son eficientes computacionalmente, pero menos precisos. Sin embargo, los árboles de mayor profundidad permiten que el algoritmo capture interacciones únicas, aunque aumentan el riesgo de sobreajuste.

El segundo hiperparámetro, además de controlar el número mínimo de observaciones en los nodos terminales, controla la complejidad de cada árbol. Los valores típicos de este hiperparámetro suelen estar entre 5 y 15. Los valores más altos ayudan a evitar que un modelo aprenda relaciones que pueden ser muy específicas de la muestra particular seleccionada para entrenar el árbol, evitando así el sobreajuste. Sin embargo, los valores más pequeños pueden ayudar con clases desbalanceadas en problemas de clasificación.

29.3.2. Estrategia de ajuste de hiperparámetros

A diferencia del *random forest*, los modelos *gradient boosting* pueden variar mucho en su precisión de acuerdo a su configuración de hiperparámetros. Por ello, el ajuste puede requerir seguir una estrategia. Un buen enfoque para esto es:

- Elegir una tasa de aprendizaje relativamente alta. El valor predeterminado es 0,1 y generalmente funciona. Sin embargo, para la mayoría de problemas funcionan valores entre 0,05 y 0,2.
- Determinar el número óptimo de árboles para la tasa de aprendizaje elegida.
- Ajustar los hiperparámetros del árbol y la tasa de aprendizaje y evaluar la velocidad frente al rendimiento.
- Ajustar los hiperparámetros específicos del árbol para determinar la tasa de aprendizaje.
- Una vez que se ajustan los parámetros específicos del árbol, se reduce la tasa de aprendizaje para evaluar cualquier mejora en la precisión.
- Utilizar la configuración final de hiperparámetros y aumentar los procedimientos de validación cruzada para obtener estimaciones más robustas. Si se utiliza validación cruzada en los pasos anteriores, entonces este paso no es necesario.

29.3.3. Procedimiento con R: la función gbm()

En el paquete **gbm** de **R** se encuentra la función con el mismo nombre **gbm()** que se utiliza para entrenar un modelo *gradient boosting*:

```
gbm(formula, data=..., ...)
```

- **formula**: Refleja la relación entre la variable dependiente Y y los predictores tal que $Y \sim X_1 + \dots + X_p$.
- **data**: Conjunto de datos con el que entrenar el árbol de acuerdo a la fórmula indicada.

29.3.4. Aplicación del modelo *gradient boosting* en R

A través de los datos de compras **dp_entr** incluidos en el paquete **CDR** se va a aplicar el modelo *gradient boosting* para clasificar qué clientes van a comprar un nuevo producto (*tensiómetro digital*) y quienes no. Se entrena el modelo utilizando el conjunto de datos de entrenamiento sin transformar (en su escala original). Así, en lugar de tener las variables categóricas transformadas mediante one-hot-encoding se usan en su escala original, como ocurre con el caso de la variable que mide el nivel educativo.

```
library("CDR")
library("caret")
library("gbm")
library("reshape")
```

29.3. Gradient Boosting (GB)

489

```
library("ggplot2")
data(dp_entr)

# se determina la semilla aleatoria
set.seed(101)

# se entrena el modelo
model <- train(CLSPRO_pro13 ~.,
                 data=dp_entr,
                 method="gbm",
                 metric="Accuracy",
                 trControl = trainControl(classProbs = TRUE,
                                           method = "cv", number = 10)
               )
```

```
model
Stochastic Gradient Boosting

558 samples
 17 predictor
 2 classes: 'S', 'N'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 502, 502, 502, 503, 503, 502, ...
Resampling results across tuning parameters:
```

	interaction.depth	n.trees	Accuracy	Kappa
1	50	0.8564610	0.7130031	
1	100	0.8690909	0.7383556	
1	150	0.8762338	0.7526413	
2	50	0.8690909	0.7382344	
2	100	0.8762338	0.7526413	
2	150	0.8799026	0.7599227	
3	50	0.8763636	0.7528004	
3	100	0.8781494	0.7563575	
3	150	0.8835390	0.7671499	

```
Tuning parameter 'shrinkage' was held constant at a value of 0.1
Tuning
parameter 'n.minobsinnode' was held constant at a value of 10
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were n.trees = 150, interaction.depth =
3, shrinkage = 0.1 and n.minobsinnode = 10.
```

El modelo resultante del proceso de entrenamiento es un *gradient boosting* que ha ajustado los hiperparámetros a 150 árboles y una profundidad igual a 3. Además, los valores tanto del

número mínimo de observaciones en nodos como de la tasa de aprendizaje, toman los valores por defecto de 10 y 0,1, respectivamente. Los resultados en el proceso de validación cruzada se muestran en la Fig. 29.2, en el que se observa como la precisión oscila entre el 84% y el 93% en las iteraciones.

```
ggplot(melt(model$resample[,-4]), aes(x = variable, y = value, fill=variable)) +
  geom_boxplot(show.legend=FALSE) +
  xlab(NULL) + ylab(NULL)
```

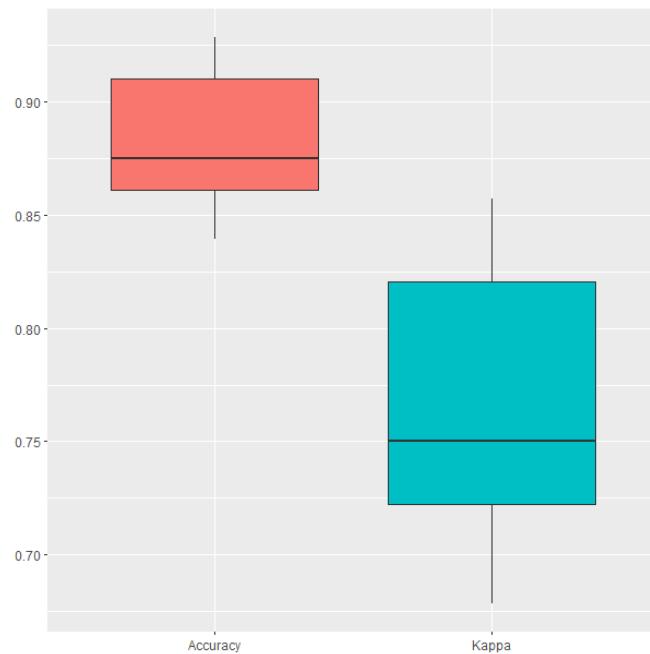


Figura 29.2: Resultados del modelo GB obtenidos durante el proceso de validación cruzada.

29.3.5. *Gradient Boosting* con ajuste automático

Se repite el procedimiento para el ejemplo anterior. Sin embargo, en este ejemplo se ajustan de forma automática los hiperparámetros más relevantes de dicho algoritmo para mejorar los resultados respecto al modelo anterior. Se observa como los hiperparámetros a ajustar para el método `gbm` son: el número de árboles, la profundidad, la tasa de aprendizaje y el número de observaciones en un nodo.

```
modelLookup("gbm")
model      parameter          label forReg forClass probModel
1  gbm        n.trees    # Boosting Iterations   TRUE    TRUE    TRUE
```

29.3. Gradient Boosting (GB)

491

2	gbm	interaction.depth	Max Tree Depth	TRUE	TRUE	TRUE
3	gbm	shrinkage	Shrinkage	TRUE	TRUE	TRUE
4	gbm	n.minobsinnode	Min. Terminal Node Size	TRUE	TRUE	TRUE

Siguiendo la estrategia descrita se definen rangos de posibles valores para los principales hiperparámetros a optimizar.

```
# Se especifica un rango de valores posibles para los hiperparámetros
tuneGrid <- expand.grid(interaction.depth = c(4,6,8),
                         n.trees = c(10*ncol(dp_entr),300,500),
                         shrinkage = c(0.05,0.1,0.2),
                         n.minobsinnode = c(5,10,15))
```

Esta red de posibles valores para los hiperparámetros del modelo se incorporan a la función de entrenamiento. Cuanto más exhaustivo sea el ajuste de estos valores, mayor será el tiempo de ajuste del modelo. La red presentada está formada por 81 combinaciones de los posibles cuatro hiperparámetros.

```
# se fija la semilla aleatoria
set.seed(101)

# se entrena el modelo
model <- train(CLSPRO_pro13~.,
                 data=dp_entr,
                 method="gbm",
                 metric="Accuracy",
                 trControl=trainControl(classProbs = TRUE,
                                         method="cv", number=10),
                 tuneGrid=tuneGrid)
```

El modelo que mejores resultados proporciona es aquel que ajusta los hiperparámetros a los siguientes valores: 180 árboles, una profundidad igual a 6, una tasa de aprendizaje de 0.05 y un tamaño mínimo de los nodos de 10 observaciones.

```
model$bestTune
  n.trees interaction.depth shrinkage n.minobsinnode
13      180             6       0.05        10
```

En la Fig. 29.3 se muestran los resultados obtenidos durante el proceso de validación cruzada. Se puede ver que los resultados son similares a los del modelo anterior, aunque hay diferencias importantes. En primer lugar, se alcanza un valor máximo de precisión mayor al anterior, pues en este caso la precisión oscila entre el 84 % y el 95 %. En segundo lugar, vemos que el valor mediano de la precisión ha subido del 87.5 % del modelo anterior hasta el 90 % de este modelo. Por último, que el rendimiento haya variado tan poco desde el modelo por defecto a un modelo

en el que se ha intentado ajustar los hiperparámetros, confirma lo ya expuesto sobre el buen rendimiento de un modelo de *gradient boosting* con los parámetros por defecto.

```
ggplot(melt(model$resample[,-4]), aes(x = variable, y = value, fill=variable)) +
  geom_boxplot(show.legend=FALSE) +
  xlab(NULL) + ylab(NULL)
```

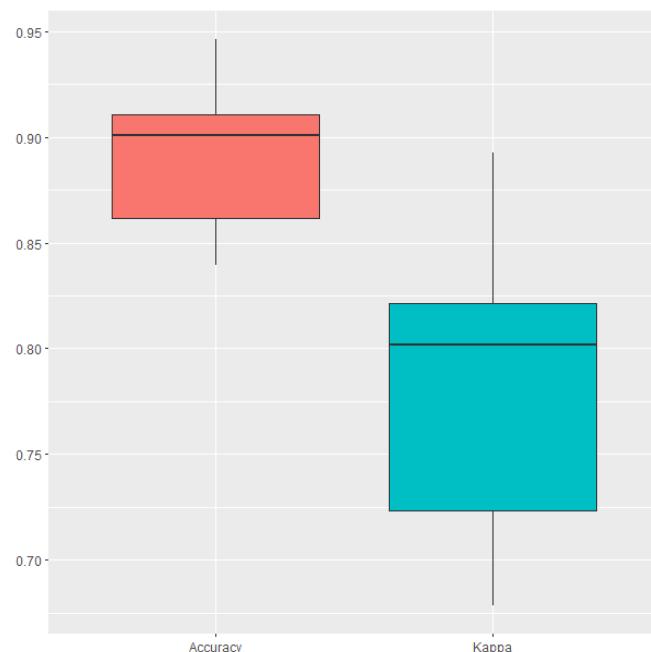


Figura 29.3: Resultados del modelo GB con ajuste automático obtenidos durante el proceso de validación cruzada.

29.4. eXtreme Gradient Boosting (XGB)

El *eXtreme Gradient Boosting* es una implementación eficiente y escalable del modelo *gradient boosting*. Este modelo, abreviado como XGBoost, es un paquete de código abierto en C++, Java, Python ([Wade, 2020](#)), R, Julia, Perl y Scala. En R el modelo se incluye dentro del paquete `xgboost` ([Chen et al., 2015](#)). El paquete incluye un procedimiento para la solución eficiente de modelos lineales y un algoritmo de aprendizaje de árboles.

El paquete es compatible con funciones objetivo de regresión, clasificación y ranking. Además, tiene varias características importantes:

1. Velocidad: `xgboost` puede realizar automáticamente cálculos paralelos. Por lo general, es 10 veces más rápido que el modelo *gradient boosting*.

2. Tipo de entrada: `xgboost` toma varios tipos de datos de entrada en R:

- Matriz densa (`matrix`)
- Matriz dispersa (`Matrix::dgCMatrix`)
- Archivo de datos locales
- Un tipo de datos propio del paquete: `xgb.DMatrix`

3. Dispersion: `xgboost` acepta datos de entrada dispersos para los modelos incluidos.

4. Personalización: `xgboost` admite tanto funciones de objetivo y funciones de evaluación personalizadas.

5. Rendimiento: `xgboost` alcanza generalmente una mayor precisión.

29.4.1. Hiperparámetros del modelo XGBoost

El modelo XGBoost proporciona los hiperparámetros que ya incluía el modelo *gradient boosting* referentes tanto al *boosting* como a los árboles. Sin embargo, `xgboost` también proporciona hiperparámetros adicionales que pueden ayudar a reducir las posibilidades de sobreajuste, lo que lleva a una menor variabilidad de predicción y, por lo tanto, a una mayor precisión. Estos hiperparámetros son: la regularización y el dropout.

Los parámetros de regularización se incluyen para ayudar a evitar el sobreajuste y reducir la complejidad del modelo. Existen tres hiperparámetros que tienen esta funcionalidad: gamma (γ), alpha (α) y lambda (λ). Gamma es un hiperparámetro de pseudo-regularización conocido como multiplicador Lagrangiano y controla la complejidad de un árbol dado. Este hiperparámetro establece que para hacer una partición adicional en un nodo es necesaria una reducción de pérdida mínima especificada por `gamma`. Al especificarlo, el modelo XGBoost hace crecer los árboles hasta una profundidad máxima establecida, pero en un paso de poda eliminará las divisiones que no cumplen con la regularización γ . Este hiperparámetro toma valores entre 0 e infinito (∞), siguiendo la regla de que a mayor valor, mayor será la regularización. Los otros hiperparámetros de regularización, α y λ , son más clásicos. Mientras que α proporciona una regularización L_1 , λ proporciona una regularización L_2 . Estos parámetros de regularización establecen un límite a cómo de extremos pueden llegar a ser los pesos de los nodos en un árbol. Sus valores se encuentran, al igual que los de γ , entre 0 y ∞ .

El dropout es un enfoque alternativo para reducir el sobreajuste. Cuando se entrena un modelo de *gradient boosting*, los primeros árboles tienden a dominar el rendimiento del modelo, mientras que los que se agregan después suelen mejorar la predicción solo para un pequeño grupo de variables. Esto puede llevar a que se incremente el riesgo de sobreajuste. Con el dropout, se descartan árboles aleatoriamente en el proceso de entrenamiento.

En su implementación en R, el modelo XGBoost incluye principalmente los siguientes parámetros para ser optimizados: número de iteraciones, profundidad máxima de los árboles, tasa de aprendizaje y la regularización γ .

```
head(modelLookup("xgbTree"),4)
  model parameter           label forReg forClass probModel
1 xgbTree    nrounds  # Boosting Iterations  TRUE   TRUE   TRUE
2 xgbTree  max_depth      Max Tree Depth  TRUE   TRUE   TRUE
3 xgbTree      eta        Shrinkage  TRUE   TRUE   TRUE
4 xgbTree     gamma Minimum Loss Reduction  TRUE   TRUE   TRUE
```

29.4.2. Procedimiento con R: la función xgboost()

En el paquete `xgboost` de **R** se encuentra la función `xgboost()` que se utiliza para entrenar un modelo *extreme gradient boosting*:

```
xgboost(data = ..., label = ..., ...)
```

- `data`: Conjunto de datos con el que entrenar el modelo.
- `label`: Vector con la variable respuesta.

29.4.3. Aplicación del modelo XGBoost en R

Se entrena este modelo utilizando el conjunto de entrenamiento sin transformar (en su escala original). Se continúa así el ejemplo expuesto durante la aplicación del modelo *gradient boosting* sin y con ajuste automático de sus hiperparámetros. Se repite el procedimiento de entrenar el modelo para los hiperparámetros por defecto que proporciona R.

```
# se determina la semilla aleatoria
set.seed(101)

# se entrena el modelo
model <- train(CLSPRO_pro13~.,
                 data=dp_entr,
                 method="xgbTree",
                 metric="Accuracy",
                 trControl=trainControl(classProbs = TRUE,
                                         method = "cv",
                                         number=10))
```

Por defecto, el entrenamiento establece valores constantes para la regularización γ (0) y para el tamaño mínimo del nodo (1). En cambio, ajusta los hiperparámetros del modelo dentro de los valores por defecto de la función. Así, el modelo XGBoost resultante tiene 50 iteraciones, una profundidad máxima igual a 2 y una tasa de aprendizaje de 0,3. Los resultados de la validación cruzada muestran que la precisión obtenida oscila entre el 85 % y el 95 %, resultado similar al del *gradient boosting* con hiperparámetros ajustados. Sin embargo, el valor mediano de la precisión es del 88 %, ligeramente inferior a la observada en el modelo *gradient boosting* con ajuste automático.

29.4. eXtreme Gradient Boosting (XGB)

495

```
ggplot(melt(model$resample[,-4]), aes(x = variable, y = value, fill=variable)) +
  geom_boxplot(show.legend=FALSE) +
  xlab(NULL) + ylab(NULL)
```

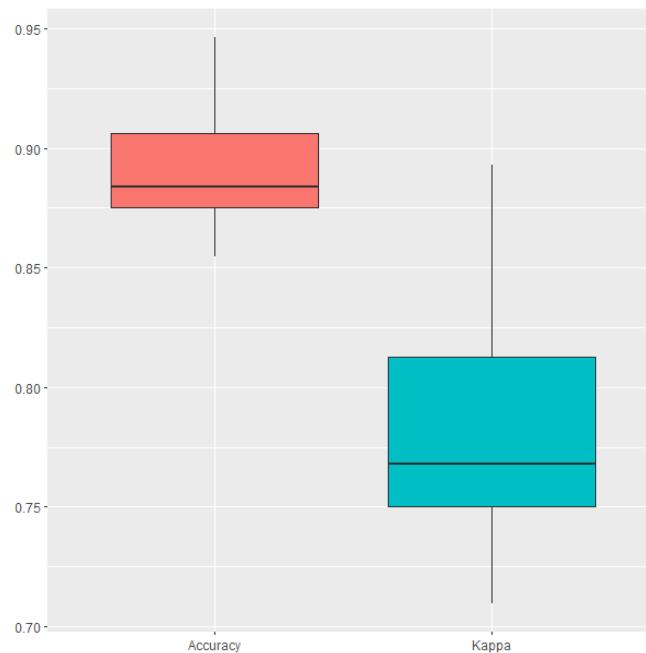


Figura 29.4: Resultados del modelo durante la validación cruzada.

29.4.4. XGBoost con ajuste automático

Se continúa el ejemplo aplicado a los datos sobre compra de un nuevo producto por parte de los clientes utilizando un modelo XGBoost en R. Sin embargo, se quieren mejorar los resultados obtenidos, y por ello ajustan automáticamente los hiperparámetros más relevantes de dicho algoritmo generando una red de posibles valores para dichos hiperparámetros. Por motivos computacionales, ésta no se hace excesivamente exhaustiva para evitar largos tiempos de entrenamiento. Si se dispone de tiempo suficiente para el entrenamiento, es aconsejable tratar de estudiar más valores para los hiperparámetros a optimizar.

```
# Se especifica un rango de valores típicos para los hiperparámetros
tuneGrid <- expand.grid (nrounds=c(50,100,500),
                         max_depth = c(3,4,8),
                         eta =c(0.05,0.1,0.2,0.3),
                         gamma=c(0,0.5,5),
                         colsample_bytree=c(0.8),
```

```

    min_child_weight=c(5),
    subsample=c(0.5))

# se determina la semilla aleatoria
set.seed(101)

# se entrena el modelo
model <- train(CLSPRO_pro13~.,
                 data=dp_entr,
                 method="xgbTree",
                 metric="Accuracy",
                 trControl=trainControl(classProbs = TRUE,
                                         method = "cv",
                                         number = 10),
                 tuneGrid=tuneGrid)

model$bestTune[,1:4]
  nrounds max_depth eta gamma
71      100          4 0.2     5

```

El modelo resultante establece que se utilicen 100 iteraciones, que los árboles tengan una profundidad máxima de 4, una tasa de aprendizaje del 0,2 y que la regularización γ tome el valor 5.

Los resultados obtenidos durante la validación cruzada muestran que la precisión es muy similar a la del modelo por defecto, al encontrarse entre el 85 % y el 95 %. Sin embargo, se observa en el valor mediano de la precisión una ligera mejoría, al aumentar hasta el 90 %.

Resumen

En este capítulo se introduce al lector en el algoritmo de aprendizaje supervisado conocido como *gradient boosting*, en concreto:

- Se presenta el otro paradigma principal de aprendizaje ensamblado: el *boosting*.
- Se explica el modelo basado en este paradigma, el *gradient boosting*, así como sus diferencias con el *random forest* (basado en *bagging*).
- Se exponen los hiperparámetros más relevantes a la hora de optimizar un modelo de *gradient boosting*.
- Se presenta el *eXtreme gradient boosting*, una implementación eficiente y escalable del modelo *gradient boosting*. Así como los hiperparámetros de regularización y otros parámetros importantes en esta implementación.
- Se aplican ambos algoritmos en R en un caso práctico para la clasificación binaria de datos.

29.4. *eXtreme Gradient Boosting (XGB)*

497

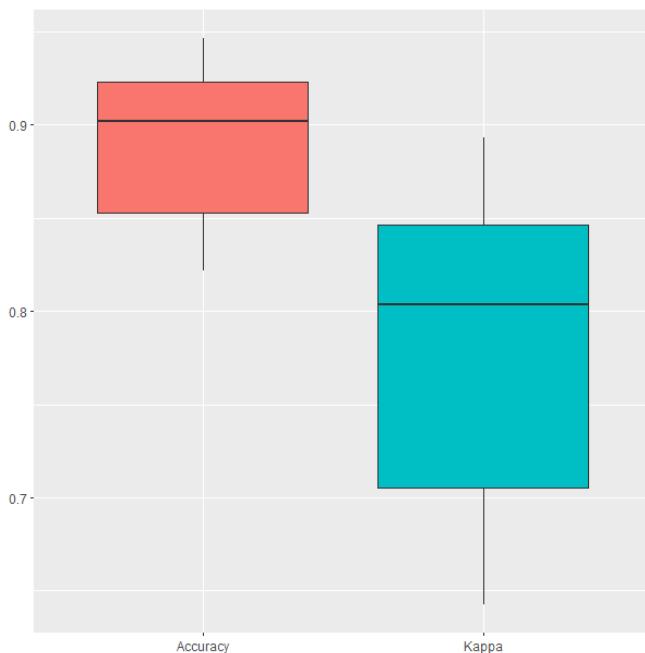


Figura 29.5: Resultados del modelo durante la validación cruzada.

Parte VI

Machine learning no supervisado

Capítulo 30

Análisis cluster: clusterización jerárquica

José-María Montero^a y Gema Fernández-Avilés^a

^aUniversidad de Castilla-La Mancha

30.1. Introducción

El origen de la actividad agrupatoria, hoy en día conocida como análisis cluster o de conglomerados (AC), taxonomía numérica o reconocimiento de patrones, entre otras denominaciones, se remonta a tiempos de Aristóteles y su discípulo Teofrasto. Por tanto, tiene unas profundas raíces y hoy en día se aplica en todos los campos del saber. Se ha evitado la palabra “clasificación” porque existe una pequeña diferencia entre agrupación y clasificación. En la actividad clasificatoria se conoce el número de grupos y qué observaciones del conjunto de datos pertenecen a cada uno, siendo el objetivo clasificar nuevas observaciones en los grupos ya existentes. En la actividad agrupatoria, el número de grupos puede ser conocido (normalmente no lo es), pero no las observaciones que pertenecen a cada uno de ellos, siendo el objetivo la asignación de dichas observaciones a diferentes grupos. Este y el siguiente capítulo se centran en este último problema, al cual se hará referencia por su denominación más popular: AC.

El AC está orientado a la síntesis de la información contenida en un conjunto de datos, normalmente una muestra relativa a objetos, individuos o, en general, elementos, definidos por una serie de características, con vistas a establecer una agrupación de los mismos en función de su mayor o menor homogeneidad. En otros términos, el AC trata de agrupar dichos elementos en grupos mutuamente excluyentes, de tal forma que los elementos de cada grupo sean lo más parecidos posible entre sí y lo más diferentes posible de los pertenecientes a otros grupos (Fig. 30.1).

Para llevar a cabo un AC, se deben tomar una serie de decisiones:

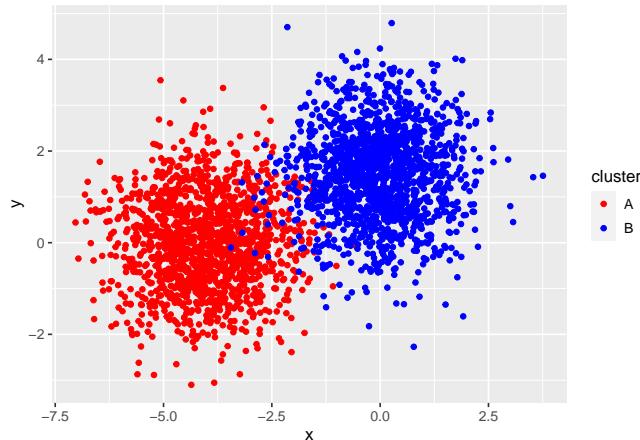


Figura 30.1: Datos simulados que presentan clusters

- Selección de las variables en función de las cuales se van a agrupar los elementos.
- Elección del tipo de distancia o medida de similitud que se va a utilizar para medir la disimilitud entre los elementos objeto de clasificación.
- Elección de la técnica para formar los grupos o conglomerados.
- Determinación del número óptimo de clusters (si no se determina a priori).

En este capítulo se abordarán la primera y, sobre todo, la segunda cuestión, dejando las otras dos para el capítulo siguiente.

Como ilustración práctica, se utilizará la base de datos TIC2021 del paquete CDR, relativa a las estadísticas de uso de las TIC en la Unión Europea en 2021.

30.2. Selección de las variables

La selección de las p variables o características, $\{X_1, X_2, \dots, X_p\}$, en función de las cuales se va a proceder a la agrupación de los n elementos disponibles es crucial, ya que determina la agrupación final, independientemente de los procedimientos técnicos utilizados. Una vez determinadas éstas, la información disponible, para los elementos objeto de agrupación, será:

Tabla 30.1: Información muestral

	X_1	X_2	X_3	...	X_p
Elemento 1	x_{11}	x_{12}	x_{13}	...	x_{1p}
Elemento 2	x_{21}	x_{22}	x_{23}	...	x_{2p}
Elemento 3	x_{31}	x_{32}	x_{33}	...	x_{3p}
...

30.3. Elección de la distancia entre elementos

503

	X_1	X_2	X_3	\cdots	X_p
Elemento n	x_{n1}	x_{n2}	x_{n3}	\cdots	x_{np}

En definitiva, la información de partida es una matriz $\mathbf{X}_{n \times p}$ donde cada elemento viene representado por un punto en el espacio p -dimensional de variables, es decir, una matriz que proporciona los valores de las variables para cada elemento.¹

Una cuestión a tener en cuenta es el número de variables a considerar en el AC. La exclusión de variables relevantes generará una agrupación deficiente. La inclusión de variables irrelevantes complicará el proceso de agrupamiento sin procurar ganancias sustantivas. Dado que el miedo del investigador vendrá por el lado de la exclusión de variables relevantes, tenderá a incluir un número excesivo de variables (muchas de ellas correlacionadas). Por ello, se recomienda realizar previamente un ACP (véase Cap. 32), lo que reduce la dimensionalidad del problema, y llevar a cabo el AC a partir de las componentes principales retenidas (incorreladas, evitando así redundancias). La eliminación de información redundante es una cuestión importante en el proceso de clusterización, porque dicha información estaría sobreponderada en el resultado obtenido. Una solución menos drástica a este problema es la utilización de la distancia de Mahalanobis, que, como se verá posteriormente, corrige estas redundancias.

Otra cuestión importante en este momento es decidir si las variables (o componentes principales en su caso) seleccionadas se utilizarán estandarizadas o no. No existe consenso sobre la cuestión, si bien se suele recomendar su estandarización para evitar consecuencias no deseadas derivadas de la distinta escala y/o unidades de medida. No obstante, autores tan relevantes como Edelbrock (1979) y Brian (1993), están en contra y proponen las siguientes alternativas: (i) recategorizar todas las variables en variables binarias, y aplicar a éstas una distancia apropiada para ese tipo de medidas; (ii) realizar distintos AC con grupos de variables homogéneas (en cuanto a su métrica) y sintetizar después los diferentes resultados; y (iii) utilizar la distancia de Gower, que es aplicable con cualquier tipo de métrica.

30.3. Elección de la distancia entre elementos

Una vez se dispone de la matriz de información $\mathbf{X}_{n \times p}$, la segunda etapa en el AC consiste en la creación de una nueva matriz $\mathbf{D}_{n \times n}$ cuyos elementos $\{d_{ij}\}$ sean las distancias o disimilaridades entre los elementos objeto de agrupamiento.

En caso de variables cuantitativas, la distancia entre dos elementos en un espacio de p dimensiones, $d(\mathbf{x}_i; \mathbf{x}_j)$, se define como una función que a cada dos puntos de \mathbb{R}^p le asocia un número real y que verifica:²

- $d(\mathbf{x}_i; \mathbf{x}_j) \geq 0$,

¹También podría observarse \mathbf{X} por columnas (la j -ésima columna muestra los valores de la j -ésima variable para cada elemento de la muestra). Aparentemente, no hay razón alguna para no poder agrupar las variables que describen cada elemento (cluster de variables en vez de elementos).

²Si se cumple el cuarto requisito la función distancia suele llamarse distancia métrica.

- $d(\mathbf{x}_i; \mathbf{x}_j) = 0$ si y sólo si $\mathbf{x}_i = \mathbf{x}_j$,
- $d(\mathbf{x}_i; \mathbf{x}_j) = d(\mathbf{x}_j; \mathbf{x}_i)$,
- $d(\mathbf{x}_i; \mathbf{x}_j) + d(\mathbf{x}_j; \mathbf{x}_k) \geq d(\mathbf{x}_i; \mathbf{x}_k), \quad \forall \mathbf{x}_k \in \mathbb{R}^p$,

Con variables cualitativas, la similitud entre dos elementos, $s(\mathbf{x}_i; \mathbf{x}_j)$, es una función que a cada dos puntos de \mathbb{R}^p le asocia un número real, y que verifica:

- $s(\mathbf{x}_i; \mathbf{x}_j) \leq s_0$, donde s_0 es un número real finito arbitrario (normalmente 1).
- $s(\mathbf{x}_i; \mathbf{x}_j) = s_0$ si y sólo si $\mathbf{x}_i = \mathbf{x}_j$,
- $s(\mathbf{x}_i; \mathbf{x}_j) = s(\mathbf{x}_j; \mathbf{x}_i)$,
- $|s(\mathbf{x}_i; \mathbf{x}_j) + s(\mathbf{x}_j; \mathbf{x}_k)|s(\mathbf{x}_i; \mathbf{x}_k) \geq d(\mathbf{x}_i; \mathbf{x}_j)s(\mathbf{x}_j; \mathbf{x}_k) \quad \forall \mathbf{x}_k \in \mathbb{R}^p$.

Son numerosas las formas de medir las distancias o similaridades entre dos elementos que satisfacen las condiciones expuestas. Las más populares son las siguientes:

Variables cuantitativas

- **Distancia euclídea.** Se define como:

$$d_e(\mathbf{x}_i; \mathbf{x}_j) = \sqrt{\sum_{k=1}^p (x_{ik} - x_{jk})^2}. \quad (30.1)$$

Ignora las unidades de medida de las variables y, en consecuencia, aunque es invariante a los cambios de origen, no lo es a los cambios de escala. También ignora las relaciones entre ellas. Resulta de utilidad con variables cuantitativas incorreladas y medidas en las mismas unidades. El cuadrado de la distancia euclídea también suele utilizarse como distancia. Para el conjunto de datos TIC2021, la distancia euclídea se obtiene ejecutando el siguiente código:

```
library("CDR")
data("TIC2021")
library("factoextra")

tic <- scale(TIC2021) # estandariza las variables
d_euclidea <- get_dist(x = tic, method = "euclidean")
as.matrix(d_euclidea)[1:5, 1:5]
#> BE      BG      CZ      DK      DE
#> BE 0.000000 6.421631 2.417212 1.870962 2.304686
#> BG 6.421631 0.000000 4.616177 7.988106 4.871235
#> CZ 2.417212 4.616177 0.000000 3.765714 1.366011
#> DK 1.870962 7.988106 3.765714 0.000000 3.607589
#> DE 2.304686 4.871235 1.366011 3.607589 0.000000
```

30.3. Elección de la distancia entre elementos

505

La Fig. 30.2 muestra un *heatmap*³ de distancias euclídeas entre los países de la UE27 a partir de las estadísticas de uso de las TIC en 2021.

```
fviz_dist(dist.obj = d_euclidea, lab_size = 10)
```

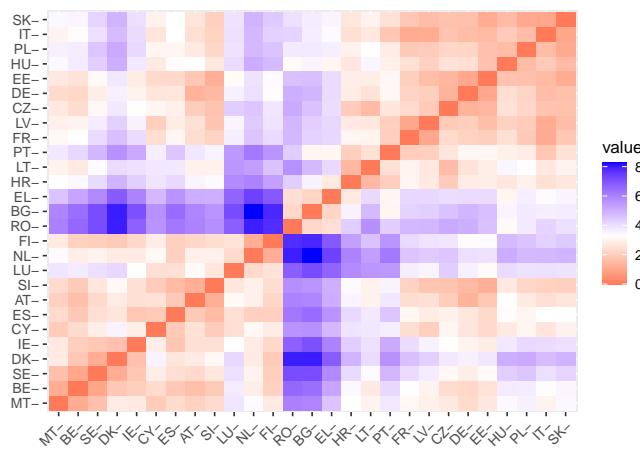


Figura 30.2: *Heatmap* de distancias euclídeas: datos ‘TIC2021’ del paquete ‘CDR’

- **Distancia Manhattan o city block.** Se define como:

$$d_{MAN}(\mathbf{x}_i; \mathbf{x}_j) = \sum_{k=1}^p |x_{ik} - x_{jk}| . \quad (30.2)$$

Viene afectada por los cambios de escala en alguna de las variables y es menos sensible que la distancia euclídea a los valores extremos. Por ello, es recomendable cuando las variables son cuantitativas, con las mismas unidades de medida, sin relaciones entre ellas y con valores extremos.

- **Distancia de Minkowski.** Se define como:

$$d_{MIN}(\mathbf{x}_i; \mathbf{x}_j) = \left(\sum_{k=1}^p |x_{ik} - x_{jk}|^\lambda \right)^{\frac{1}{\lambda}} . \quad (30.3)$$

³Un *heatmap* es una representación visual de fácil lectura e interpretación basada en un código de colores, típica del análisis de páginas web. Normalmente proporciona un patrón visual en forma de "F", que es el del seguimiento ocular de un sitio o plataforma tecnológica.

Las distancias euclídea y Manhattan son casos particulares de la distancia de Minkowski. En la distancia euclídea $\lambda = 2$ y en la Manhattan $\lambda = 1$.

- **Norma del supremo o distancia de Chebychev.** Su expresión es:

$$d_{CHE}(\mathbf{x}_i; \mathbf{x}_j) = \max_{1 \leq k \leq p} \sum_{k=1}^p |x_{ik} - x_{jk}|. \quad (30.4)$$

Únicamente influye en ella la variable con los valores más extremos y, en este sentido, es muy sensible a los cambios de escala en una de las variables.

- **Distancia de Mahalanobis.** Se define como:

$$d_{MAH} = (\mathbf{x}_i; \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)' \mathbf{S}^{-1} (\mathbf{x}_i - \mathbf{x}_j) \quad (30.5)$$

Coincide con la distancia euclídea calculada sobre las componentes principales. Es invariante a cambios de origen y de escala (por tanto, la matriz de covarianzas entre las variables agrupadoras, \mathbf{S} , se puede sustituir por su homónima de correlaciones, \mathbf{R}). Además, tiene en cuenta, explícitamente, las correlaciones lineales que puedan existir entre las variables, corrigiendo así el efecto redundancia. Es, por tanto, una distancia apropiada cuando se trabaja con variables cuantitativas con relaciones aproximadamente lineales. Su principal desventaja es que \mathbf{S} involucra, conjuntamente, a todos los elementos, y no únicamente, y de forma separada, a los elementos de cada cluster.

- **Coeficiente de correlación de Pearson.** Se define como:

$$d_P(\mathbf{x}_i; \mathbf{x}_j) = \frac{\sum_{k=1}^p (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum_{k=1}^p (x_{ik} - \bar{x}_i)^2 \sum_{k=1}^p (x_{jk} - \bar{x}_j)^2}}. \quad (30.6)$$

No es una distancia sino un indicador de similitud. Por tanto, valores altos indican elementos similares y valores bajos elementos distintos.

Su campo de variación es $[-1, 1]$, por lo que se toma su valor absoluto. Cuando las variables están centradas, se denomina coeficiente de congruencia o distancia coseno, puesto que coincide con el coseno formado por los vectores representativos de cada pareja de elementos. Tiene un inconveniente importante: un valor unitario no significa que los dos elementos sean iguales, puesto que también pueden obtenerse valores unitarios cuando los valores de las p variables en uno de los elementos sean combinación lineal de los valores de las p variables del otro.

Se utiliza, en ocasiones, preferentemente con datos cuantitativos y con el algoritmo de distancia mínima. Los coeficientes de correlación por rangos de Kendall y Spearman se utilizan, también, en casos de variables ordinales.

A efectos prácticos, cambiando el argumento `method` de la función `get_dist` (`euclidean`, `maximum`, `manhattan`, `minkowski`, `pearson`, `spearman`, `kendall`) se obtienen distintas matrices de distancias entre los elementos.

Variables cualitativas (dicotómicas)

En este caso, se pueden establecer distintas medidas de similaridad en base a la siguiente tabla de contingencia 2×2 :

		Elem. j		Total
		Presencia	Ausencia	
Elem. i	Presencia	n_{11}	n_{12}	$n_{1\cdot}$
	Ausencia	n_{21}	n_{22}	$n_{2\cdot}$
	Total	$n_{\cdot 1}$	$n_{\cdot 2}$	p

A partir de la tabla anterior, la similaridad entre dos elementos se puede medir a partir de las coincidencias, ya sea de presencias y ausencias como de solo presencias.

Entre las medidas de similaridad que involucran tanto presencias como ausencias comunes están:

- El **coeficiente de coincidencias simple**: $c_{cs} = \frac{(n_{11}+n_{22})}{2}$
- El **coeficiente de Rogers-Tanimoto**: $c_{RT} = \frac{(n_{11}+n_{22})}{2(n_{11}+n_{22})+n_{12}+n_{21}}$

Estos dos coeficientes tienen una relación monotónica (si la distancia entre dos elementos es igual o superior a la distancia entre otros dos con una de las medidas, también lo es con la otra). Esto es importante, dado que algunos procedimientos de agrupación no se ven afectados por la medida utilizada siempre y cuando el ordenamiento establecido por ellas sea el mismo.

Entre aquellas que identifican similaridad con presencias destacan:

- El **coeficiente de Jaccard**: $c_J = \frac{n_{11}}{n_{11}+n_{12}+n_{21}}$
- El **coeficiente de Czekanowski**: $c_C = \frac{2n_{11}}{2n_{11}+n_{12}+n_{21}}$
- El **coeficiente de Sokal y Sneath**: $c_{SS} = \frac{n_{11}}{n_{11}+2(n_{12}+n_{21})}$
- El **coeficiente de Russell y Rao**: $c_{RR} = \frac{n_{11}}{p}$

Los tres primeros coeficientes disfrutan de la relación de monotonía en el sentido anteriormente apuntado, siendo los dos primeros las más utilizados en la práctica.

También se usan como indicadores de similitud las medidas de asociación para tablas 2×2 , sobre todo Q y ϕ (Sec. ref(mejoradas)).

Variables cualitativas (polítómicas)

Cuando todas las variables sean cualitativas y alguna sea polítómica, se generan para estas últimas tantas variables dicotómicas como categorías tienen, denotando con 1 la presencia y con 0 la ausencia.

Variables cuantitativas y cualitativas

Si las variables no son del mismo tipo, se utiliza la medida de similaridad de Gower:

$$S_{ij}(\mathbf{x}_i; \mathbf{x}_j) = \frac{\sum_{k=1}^p s_{ik}}{\sum_{k=1}^p w_{ik}} \quad (30.7)$$

donde w_{ik} vale siempre la unidad, salvo para variables binarias si los dos elementos presentan el valor cero. En cuanto al valor de S_{ij} , se distinguen tres casos:

- Variables cualitativas de más de dos niveles: 1 si ambos elementos son iguales en la k -ésima variable; 0 si son diferentes.
- Variables dicotómicas: 1 si la variable considerada está presente en ambos elementos; 0 en los demás casos.
- Variables cuantitativas: $1 - \frac{|x_{ik} - x_{jk}|}{R_k}$, donde R es el rango de la variable k .

No es recomendable cuando las variables cuantitativas sean muy asimétricas. En este caso, hay dos procedimientos aproximados: (i) calcular medidas separadas para las variables cuantitativas y cualitativas y combinarlas estableciendo algún tipo de ponderación; (ii) pasar las variables cuantitativas a cualitativas y utilizar las medidas propuestas para este tipo de variables.

30.4. Técnicas de agrupación jerárquicas

30.4.1. Introducción

Una vez se han seleccionado las variables en función de las cuales se van a agrupar los elementos disponibles en clusters o conglomerados, así como se ha decidido qué distancia utilizar para tal propósito, el siguiente paso del AC es la selección de un criterio o técnica de agrupamiento para formar los conglomerados. Dichas técnicas se pueden clasificar en (i) jerárquicas y (ii) no jerárquicas.

TÉCNICAS DE CLUSTERIZACIÓN:

1. Jerárquicas:

■ Aglomerativas:

- Vecino más cercano o encadenamiento simple
- Vecino más lejano o encadenamiento completo
- Método de la distancia media
- Método de la distancia entre centroides
- Método de la mediana
- Método de Ward
- Encadenamiento intra-grupos
- Método flexible de Lance y Williams

■ Divisivas:

- Vecino más cercano o encadenamiento simple
- Vecino más lejano o encadenamiento completo
- Método de la distancia media
- Método de la distancia entre centroides
- Método de la mediana
- Método de Ward
- Encadenamiento intra-grupos
- Análisis de la asociación
- Detector automático de interacciones

2. No jerárquicas:

■ Técnicas de reasignación:

- Basadas en centroides: Método de Forgy, k -medias
- Basadas en medoides: k -medoides, PAM, CLARA, CLARANS
- Basadas en medianas: k -medianas

■ Técnicas basadas en la densidad de elementos (mode-seeking):

- Aproximación tipológica: Análisis modal, métodos TaxMap, de Fortin, de Gitman y Levine, de Catel y Coulter
- Aproximación probabilística: método de Wolf
- DBSCAN

■ Otras técnicas no jerárquicas

- Métodos directos: *block-*; *bi*; *co-*; *two – mode clustering*
- Métodos de reducción de la dimensionalidad: modelos Q - y R -factorial
- Clustering difuso
- Métodos basados en mixturas de modelos

Los procedimientos jerárquicos no partitionan el conjunto de elementos de una sola vez, sino que realizan particiones sucesivas a distintos niveles de agrupamiento; es decir, establecen una jerarquía de clusters, de ahí su nombre. Forman los conglomerados, bien agrupando los elementos

en grupos cada vez más grandes, fusionando grupos en cada paso, (jerárquicos aglomerativos), o bien desagregándolos en conglomerados cada vez más pequeños (jerárquicos divisivos).

Las técnicas no jerárquicas se caracterizan porque (i) el número de clusters se suele determinar a priori; (ii) utilizan directamente los datos originales, si necesidad de calcular una matriz de distancias o similaridades; y (iii) los clusters resultantes no están anidados unos en otros, sino que están separados. La caja informativa proporciona un detalle mayor de la tipología de técnicas de agrupación que aborda el presente capítulo⁴. En lo que sigue, el objetivo son las técnicas jerárquicas, abordando las no jerárquicas en el Cap. 31.

30.4.2. Técnicas jerárquicas aglomerativas

Las técnicas jerárquicas aglomerativas, de amplia utilización, parten de tantos conglomerados como elementos y llegan a un único conglomerado final.

Se parte de un conglomerado constituido por los dos elementos más próximos, de tal manera que en la segunda etapa el conglomerado formado actuará a modo de elemento (como si se tuvieran $n - 1$ elementos). En la segunda etapa, de nuevo se agrupan de nuevo los dos elementos más cercanos, que pueden ser dos elementos simples o uno simple y otro compuesto (el conglomerado anterior); en el primer caso, se tendrían dos conglomerados (cada uno de ellos formado por dos elementos) y en el segundo, un conglomerado con tres elementos y otro con uno. Sea cual sea el caso, al final de la segunda etapa se tienen $n - 2$ elementos, dos de los cuales son conglomerados. En las etapas siguientes se procede de idéntica manera: agrupación de los dos elementos (sean elementos simples o conglomerados formados en las etapas anteriores) más cercanos, y así sucesivamente hasta formar un único conglomerado integrado por todos los elementos. Es importante resaltar que un elemento, una vez forma parte de un conglomerado, ya no sale de él.

La pregunta que surge en este momento es: en el proceso de agrupamiento descrito, ¿cómo se mide la distancia de un elemento a un conglomerado, o entre dos conglomerados?⁵ Los métodos más populares son los siguientes:

- **Método del encadenamiento simple o vecino más cercano.**

Utiliza el criterio de “*la distancia mas cercana*”. Por tanto, (i) la distancia entre un elemento y un conglomerado es la menor de las distancias entre dicho elemento y cada uno de los elementos del conglomerado; (ii) la distancia entre dos conglomerados viene dada por la distancia entre sus dos elementos más cercanos. Una vez computada la matriz de distancias se seleccionan los conglomerados más cercanos.

- **Método del encadenamiento completo o vecino más lejano.**

Funciona igual que el anterior, pero ahora el criterio es “*la distancia más lejana*”.

⁴Elaboración propia en base a Kassambara (2017).

⁵El criterio de inclusión de los elementos en dichos conglomerados citado en 30.1.

30.4. Técnicas de agrupación jerárquicas

511

Nótese que, mientras que con el método del vecino más cercano la distancia entre los elementos más próximos de un cluster es siempre menor que la distancia entre elementos de distintos clusters, con el criterio del vecino más lejano la distancia entre los dos elementos más alejados de un cluster es siempre menor que la distancia entre cualquiera de sus elementos y los elementos más alejados de los demás clusters. Nótese también que, mientras que el método del vecino más cercano tiende a separar a los individuos en menor medida que la indicada por sus disimilitudes iniciales (es espacio-contrativo), el criterio del vecino más lejano es espacio-dilatante, es decir, tiende a separar a los individuos en mayor medida que la indicada por sus disimilitudes iniciales (Gallardo-San Salvador and Vera-Vera, 2004).

- **Método de la distancia media.**

Surge como una solución a la constrección o dilatación del espacio que provocan los dos métodos anteriores (por eso se dice que es espacio-conservativo y es muy utilizado), utilizando “*la distancia promedio*”, es decir, la distancia entre un elemento y un conglomerado es la media aritmética de las distancias de dicho elemento a cada uno de los elementos del conglomerado. En caso de dos conglomerados, la distancia entre ellos viene dada por el promedio aritmético de las distancias, dos a dos, tomándose un elemento de cada conglomerado. Igual que los dos métodos precedentes, es invariante a transformaciones monótonas de la distancia utilizada.

En la Fig. 30.3 se puede ver la constrección, dilatación y conservación del espacio que producen los métodos del vecino más cercano, más lejano y de la distancia media, respectivamente. En este caso se utiliza como representación gráfica el dendrograma (diagrama de árbol). En figuras posteriores se utilizarán otras alternativas al dendrograma, con el objetivo de mostrar las más populares.

```
hc_simple <- hcut(tic, k = 3, hc_method = "single")
hc_completo <- hcut(tic, k = 3, hc_method = "complete")
hc_promedio <- hcut(tic, k = 3, hc_method = "average")

library("patchwork")
d1 <- fviz_dend(hc_simple, cex = 0.5, k = 3, main = "Vecino más cercano")
d2 <- fviz_dend(hc_completo, cex = 0.5, k = 3, main = "Vecino más lejano")
d3 <- fviz_dend(hc_promedio, cex = 0.5, k = 3, main = "Distancia promedio")

d1 + d2 + d3
```

- **Método de la distancia entre centroides.**

Según este método, la distancia entre dos grupos o conglomerados es la distancia entre sus centroides, entendiendo por centroide del grupo g : $c_g = (\bar{x}_{1g}, \bar{x}_{2g}, \dots, \bar{x}_{pg})$, donde \bar{x}_{jg} es la media de la j -ésima variable en dicho grupo.

Igual que el método de la media, este método es también espacio-conservativo. Sin embargo, tiene la limitación de que cuando se agrupan dos conglomerados de diferente tamaño, el conglomerado resultante queda más cerca del conglomerado mayor y más alejado del menor, de

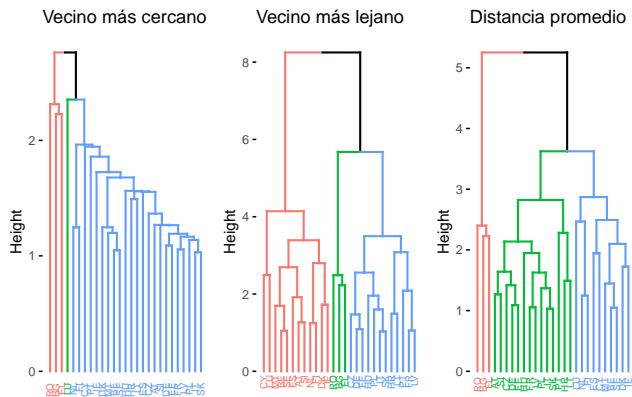


Figura 30.3: Clusterización jerárquica con distancias euclídeas (dendrograma): métodos del vecino más cercano, vecino más lejano y distancia media

forma proporcional a la diferencia de tamaños, lo que lleva a que a lo largo del proceso de clusterización se vayan perdiendo las propiedades de los conglomerados pequeños ([Gallardo San-Salvador, 2022](#)).

- **Método de la mediana.**

Viene a superar la limitación del método del centroide. Para ello, la estrategia natural es suponer que los grupos son de igual tamaño. Dicha estrategia se plasma en suponer que la distancia entre un elemento (o un conglomerado, k) y el conglomerado formado por la agrupación de los conglomerados i y j viene dada por la mediana del triángulo formado por sus centroides (de ahí su nombre). Se trata de un método espacio conservativo, pero, igual que el método del centroide, no es invariante a transformaciones monótonas de la distancia utilizada.

La Fig. 30.4, un tanglegrama o diagrama de laberinto, muestra las agrupaciones producidas por los métodos del centroide y la mediana. En ella se puede observar como el método de la mediana corrige la limitación del método del centroide. `index{método! de la mediana}` `index{método! del centroide}`

```
library("dendextend")
library("cluster")
hc_cent_dend <- as.dendrogram(hclust(d_euclidea, method = "centroid"))
hc_med_dend <- as.dendrogram(hclust(d_euclidea, method = "median"))
tanglegram(hc_cent_dend, hc_med_dend)
```

- **Método de Ward.**

30.4. Técnicas de agrupación jerárquicas

513

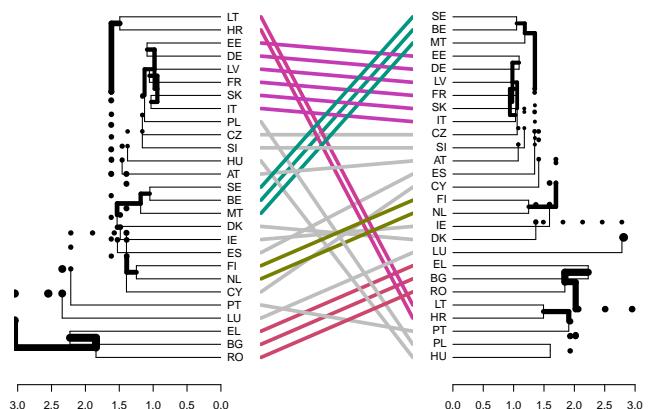


Figura 30.4: Clusterización jerárquica con distancias euclídeas (tanglegrama): método del centroide vs. método de la mediana

El método de Ward agrupa, en cada etapa, los dos clusters que producen el menor incremento de la varianza total intra-cluster: $W = \sum_g \sum_{i \in g} (x_{ig} - \bar{x}_g)'(x_{ig} - \bar{x}_g)$, donde \bar{x}_g es el centroide del grupo g . Así, los grupos formados no distorsionan los datos originales.⁶

Es muy utilizado en la práctica, dado que tiene casi todas las ventajas del método de la media y suele ser más discriminatorio en la determinación de los niveles de agrupación. También suele crear conglomerados muy compactos de tamaño similar. Dado que el menor incremento de W es proporcional a la distancia euclídea al cuadrado entre los centroides de los grupos fusionados, W no es decreciente, solventándose los problemas de los otros métodos basados en centroides.

La Fig. 30.5, muestra el filograma, diagrama filético en forma de árbol filogenético, generado por la librería `igraph` con el método de agrupación de Ward.

```
library("igraph")
set.seed(5665)
hc_ward <- hcut(tic, k = 3, hc_method = "ward.D2")
fviz_dend(
  x = hc_ward,
  k = 3,
  type = "phylogenetic"
)
```

- **Método del encadenamiento intra-grupos.**

Según el método de la distancia promedio (o vinculación entre grupos) la distancia entre dos conglomerados se obtenía calculando las distancias de cada elemento de uno de los grupos con

⁶Específicamente, la propuesta de Ward es que la pérdida de información que se produce al integrar los distintos individuos en clusters sea la mínima posible.

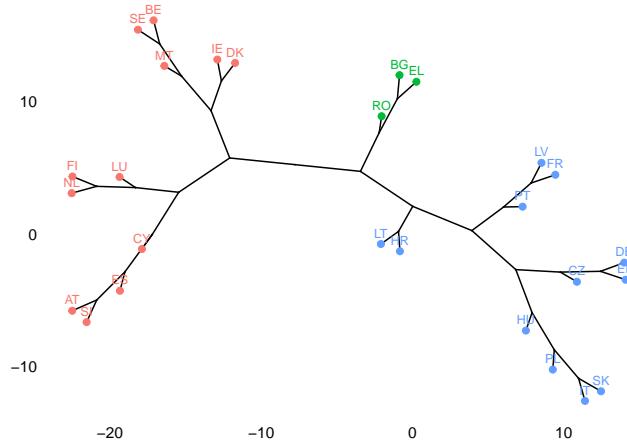


Figura 30.5: Clusterización jerárquica con distancias euclídeas al cuadrado (filograma): método de Ward

todos los del otro y computando, posteriormente, la media aritmética de dichas distancias. Con el método de la vinculación intra-grupos se computa la distancia media entre la totalidad de los elementos de los conglomerados susceptibles de agrupación, con independencia de si pertenecen al mismo conglomerado inicial o a distinto conglomerado. Por ejemplo: si un conglomerado está formado por los elementos a y b , y otro por los elementos c y d , la distancia inter-grupos entre los dos conglomerados es:

$$d_{inter-grupos} = \frac{d_{(a;c)} + d_{(a;d)} + d_{(b;c)} + d_{(b;d)}}{4}$$

mientras que la distancia intra-grupos vendrá dada por la media de las distancias entre los elementos a, b, c y d :

$$d_{intra-grupos} = \frac{d_{(a;b)} + d_{(a;c)} + d_{(a;d)} + d_{(b;c)} + d_{(b;d)} + d_{(-c;d)}}{6}$$

■ Método flexible de Lance y Williams.

Calcula la distancia entre dos conglomerados (el primero formado por la unión de otros dos en la etapa previa) a partir de la siguiente expresión:

$$d_{(g_1 \cup g_2);g_3} = \alpha_1 d_{(g_1;g_3)} + \alpha_2 d_{(g_2;g_3)} + \beta d_{(g_1;g_2)} + \gamma |d_{(g_1;g_2)} - d_{(g_2;g_3)}|,$$

donde $\alpha_1 + \alpha_2 + \beta = 1$; $\alpha_1 = \alpha_2$; $\beta < 1$; $\gamma = 0$, si bien Lance y Williams sugieren adicionalmente un pequeño valor negativo de β . Por ejemplo $\beta = -0,25$.

Los métodos anteriormente expuestos son casos particulares de éste. Denominando n_1 , n_2 y n_3 a los tamaños de los grupos g_1 , g_2 y g_3 , respectivamente, se tiene:

30.4. Técnicas de agrupación jerárquicas

515

Tabla 30.3: Valores de α_1 , α_2 , β y γ para distintos procedimientos de agrupación

Método	α_1	α_2	β	γ
Vecino más cercano	0,5	0,5	0	-0,5
Vecino más lejano	0,5	0,5	0	0,5
Distancia media	$\frac{n_1}{n_1+n_2}$	$\frac{n_2}{n_1+n_2}$	0	0
Distancia entre centroides	$\frac{n_1}{n_1+n_2}$	$\frac{n_2}{n_1+n_2}$	$\frac{-n_1 n_2}{(n_1+n_2)^2}$	0
Método de la mediana	0,5	0,5	-0,25	0
Ward	$\frac{n_1+n_3}{n_1+n_2+n_3}$	$\frac{n_2+n_3}{n_1+n_2+n_3}$	$\frac{-n_3}{n_1+n_2+n_3}$	0
Flexible	$0,5(1-\beta)$	$0,5(1-\beta)$	β	0

30.4.3. Técnicas jerárquicas divisivas

En este caso, la secuencia de acontecimientos es justo la inversa. Se parte de un único conglomerado formado por todos los elementos y se llega a n conglomerados formados, cada uno de ellos, por un único elemento (a veces el proceso termina cuando se llega a un número de grupos preestablecido). Ahora bien, dado que ahora se trata de subdividir conglomerados, es decir, de identificar los elementos más distantes, o menos similares, para separarlos del resto del conglomerado, la estrategia a seguir estará basada en maximizar las distancias (o minimizar las similitudes). En el proceso disociativo surge una cuestión importante: cuándo debe dejar de dividirse un cluster determinado y pasar a dividir otro, cuestión que se resuelve por el procedimiento propuesto por MacNaughton-Smith et al. (1964). Las técnicas divisivas (también llamadas partitivas o disociativas), pueden ser monotéticas o politéticas. En el primer caso, las divisiones se basan en una sola característica o atributo. En el segundo, se tienen en cuenta todas.

Las técnicas divisivas son menos populares que las aglomerativas. Sin embargo, la probabilidad de que lleven a decisiones equivocadas (debido a la variabilidad estadística de los datos) en las etapas iniciales del proceso, lo cual distorsionaría el resultado final del mismo, es menor que en las aglomerativas. En este sentido, los métodos partitivos, al partir del total de elementos, se consideran más seguros que los aglomerativos. Los métodos disociativos más populares son los siguientes:

- **Método de la distancia promedio**

Dentro de las técnicas politéticas, entre las que se cuentan todas las vistas en la clusterización jerárquica aglomerativa, quizás la más popular es la que utiliza para la partición el método de la distancia promedio. Para ilustrarla, supóngase que se tienen 5 elementos y que su matriz de distancias es la siguiente:

$$\mathbf{X} = \begin{pmatrix} \cdot & \cdot & \cdot & \cdot & \cdot \\ 8 & \cdot & \cdot & \cdot & \cdot \\ 7 & 4 & \cdot & \cdot & \cdot \\ 6 & 1 & 4 & \cdot & \cdot \\ 3 & 4 & 5 & 4 & \cdot \end{pmatrix}$$

En la primera etapa hay que dividir el grupo de cinco elementos en dos conglomerados. Hay $2^{2n-1} - 1$ posibilidades, pero según el método de la distancia promedio, se calcula la distancia de cada elemento a los demás y se promedia, desgajándose el elemento con distancia promedio máxima. En este caso, se desgajaría el primer elemento, y en la segunda etapa se partiría de dos grupos: $\{e_1\}$ y $\{e_2, e_3, e_4, e_5\}$.

A partir de la segunda etapa, se procede como sigue (véase Tabla 30.4):

- (i) Se calculan las (4) distancias promedio de cada elemento del conglomerado principal al elemento desgajado;
- (ii) Se calculan las (4) distancias promedio de cada elemento del conglomerado principal al resto de elementos del mismo;
- (iii) Se computan las diferencias (i) – (ii) para cada uno de los 4 elementos del conglomerado principal;
- (iv) De entre aquellos elementos del grupo principal en los que (i) – (ii) < 0 se selecciona aquel para el cual es máxima. Tras esta segunda etapa los conglomerados son $\{e_1, e_5\}$ y $\{e_2, e_3, e_4\}$.

Tabla 30.4: Distancias entre conglomerados: segunda etapa

Elemento	Distancia promedio al grupo desgajado $\{e_1\}$	Distancia promedio al grupo principal	Diferencia
$\{e_2\}$	8	3	5
$\{e_3\}$	7	3	4
$\{e_4\}$	6	3	3
$\{e_5\}$	3	4,33	-1,33

En las siguientes etapas se procede de igual manera, hasta que todas las diferencias sean positivas (en el caso que se considera, esto ocurre en la tercera etapa; véase Tabla 30.5).

Tabla 30.5: Distancias entre conglomerados: tercera etapa

Elemento	Distancia promedio al grupo desgajado $\{e_1, e_5\}$	Distancia promedio al grupo principal	Diferencia
$\{e_2\}$	6	2,5	3,5
$\{e_3\}$	6	4	2

30.4. Técnicas de agrupación jerárquicas

517

Elemento	Distancia promedio al grupo desgajado $\{e_1, e_5\}$	Distancia promedio al grupo principal	Diferencia
$\{e_4\}$	5	2,5	2,5

Cuando esto ocurre, es decir, cuando todos los elementos del conglomerado principal están más cerca de los demás que lo componen que de los del conglomerado disociado, se vuelve a iniciar el algoritmo, pero esta vez para cada uno de los dos conglomerados generados ([MacNaughton-Smith et al., 1964](#)). En caso que nos ocupa, en $\{e_1, e_5\}$ la única partición posible es $\{e_1\}, \{e_5\}$. En $\{e_2, e_3, e_4\}$ se desgaja el elemento con mayor distancia promedio a los demás del grupo. Como $\frac{d_{(2,3)}+d_{(2,4)}}{2} = 2,5$, $\frac{d_{(3,2)}+d_{(3,4)}}{2} = 4$ y $\frac{d_{(4,2)}+d_{(4,3)}}{2} = 2,5$, se desgaja $\{e_3\}$.

A continuación se aplica el algoritmo anteriormente expuesto a cada elemento del grupo principal $\{e_2, e_4\}$ y $\{e_3\}$ (Tabla 30.6) y, como todas las distancias son positivas, se divide $\{e_2, e_4\}$ en $\{e_2\}$ y $\{e_4\}$.

Tabla 30.6: Distancia entre conglomerados: etapa final

Elemento	Distancia promedio al grupo desgajado $\{e_3\}$	Distancia promedio al grupo principal	Diferencia
$\{e_2\}$	4	1	3
$\{e_4\}$	4	1	3

El algoritmo DIvisive ANAlysis (DIANA) permite llevar a cabo la partición anterior utilizando el diámetro de los clusters para decidir el orden de partición clusters cuando se tienen varios con más de un elemento (véase capítulo 6 de [Kaufman and Rousseeuw \(1990\)](#)). Proporciona (i) el coeficiente divisivo (véase `?diana.object`), que mide la cantidad de estructura de agrupamiento encontrada; y (ii) la pancarta, una novedosa presentación gráfica (véase `?plot.diana`).

Para el ejemplo de los datos TIC, DIANA proporciona el coeficiente divisivo (valores cercanos a 1 sugieren una estructura de agrupación fuerte), y el dendrograma, en este caso circular, representado en la Fig. 30.6.

```
hc_diana <- diana(tic, metric = "euclidean")
hc_diana$dc
#> [1] 0.8043393
fviz_dend(
  x = hc_diana,
  k = 3,
  type = "circular",
  ggtheme = theme_minimal()
)
```

- Análisis de la asociación

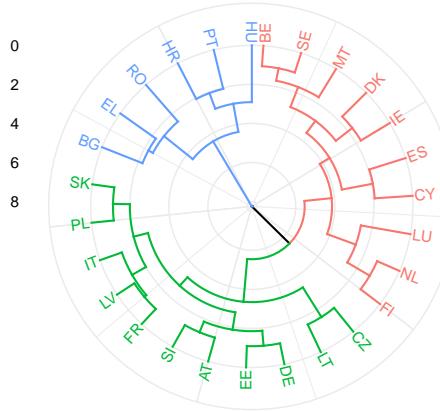


Figura 30.6: Clusterización jerárquica divisiva con DIANA

En caso de que los elementos vengan caracterizados por variables cualitativas o factores dicotómicos, F_1, F_2, \dots, F_n (si alguno fuese politómico, cada una de sus categorías se consideraría como un factor dicotómico), el método del análisis de la asociación (o suma de estadísticos chi-cuadrado) es una técnica monotética muy utilizada que procede como sigue:

- (i) Considérese F_1 y divídase el conjunto de elementos en dos grupos o categorías: uno con los elementos en los que F_1 esté presente y otro con aquellos en los que esté ausente. Hágase lo mismo con los demás factores.
- (ii) Constrúyanse las $n \times (n - 1)$ tablas de contingencia 2×2 que cruzan cada factor con cada uno de los demás (véase Sec. 23.1.1).

		Presencia	Factor j		Total
			SI	NO	
Presencia	SI	n_{11}	n_{21}	$n_{1\cdot}$	$n_{1\cdot}$
	NO	n_{21}	n_{22}	$n_{2\cdot}$	
Total		$n_{\cdot 1}$	$n_{\cdot 2}$	n	

donde $i \neq j$.

- (iii) Calcúlese el estadístico chi-cuadrado ($\chi^2_{ij} = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2}}$) para una de dichas tablas (véase eígrafe 23.2.4) y compútese $\sum_{i \neq j} \chi^2_{ij}$.
- (iii) Desgájese del conglomerado inicial en dos: uno con los elementos que contienen el factor con la máxima $\sum_{i \neq j} \chi^2_{ij}$ y otro con el resto de los elementos (donde dicho factor está ausente).
- (iv) Procédase así iterativamente.

30.5. *Calidad de la agrupación y número de clusters*

519

■ **Método del detector automático de interacciones (AID)**

No es propiamente un método de AC, sino de la esfera de los modelos lineales de rango no completo. Sin embargo, se menciona, siquiera mínimamente, porque se utiliza en algunas ocasiones para combinar categorías de los factores utilizados con la finalidad de generar grupos que difieran lo más posible entre sí respecto de los valores de una variable dependiente medida en una escala métrica (con una escala proporcional o de intervalo) o ficticia (dicotómica con valores 0 y 1). Específicamente, el AID realiza divisiones secuenciales dicotómicas de la variable a explicar mediante un ANOVA, dividiendo inicialmente el conjunto de elementos objeto de agrupación en dos grupos según la variable que mejor explica las diferencias en el comportamiento a estudiar (en cada etapa se busca la partición que maximiza la varianza inter-grupos y minimiza la varianza intra-grupos); cada uno de los dos grupos formados se vuelve a subdividir de acuerdo con la variable que mejor explica las diferencias entre ellos; este proceso continúa hasta que el tamaño de los grupos dicotómicos alcanza un mínimo pre-establecido o hasta que las diferencias entre los valores medios de los grupos sean no significativas.

En este algoritmo, el proceso de subdivisión del conjunto de elementos en grupos dicotómicos continúa hasta que se verifica algún criterio de parada.

Las limitaciones más importantes del AID son las siguientes:

- Tiende a seleccionar como más explicativas las variables con mayor número de categorías. Por eso no conviene utilizarlo cuando las variables explicativas difieren mucho en el número de categorías.
- Las particiones resultantes dependen de la variable elegida en primer lugar, condicionando las sucesivas particiones.
- Su naturaleza exclusivamente dicotómica también es una limitación importante. Si se llevasen a cabo particiones con tres o más ramas producirían una mayor reducción de la varianza residual y, además, permitirían una mejor selección de otras variables.

El AID basado en tablas de contingencia y el estadístico chi-cuadrado (CHAID) corrige la mayoría de estas limitaciones. Aunque inicialmente fue diseñado para variables categóricas, posteriormente se incluyó la posibilidad de trabajar con variables categóricas nominales, categóricas ordinales y continuas, permitiendo generar tanto árboles de decisión, para resolver problemas de clasificación, como árboles de regresión. Además, los nodos se pueden dividir en más de dos ramas.

30.5. Calidad de la agrupación y número de clusters

30.5.1. El coeficiente de correlación lineal cofenético

Dado que las técnicas jerárquicas imponen una estructura sobre los datos y pueden producir distorsiones significativas en las relaciones entre los datos originales, una vez realizada la jerarquización de los elementos objeto de clusterización, surge la siguiente pregunta: ¿en qué medida la estructura final obtenida representa las similitudes o diferencias entre dichos objetos? En

otros términos, ¿en qué medida el dendrograma representa la matriz de distancias o similitudes original?

El coeficiente de correlación lineal cofenético da respuesta a dichas preguntas. Se define como el coeficiente de correlación lineal entre los $n(n - 1)$ elementos del triángulo superior de la matriz de distancias o similitudes y sus homónimos en la matriz cofenética, \mathbf{C} , cuyos elementos $\{c_{ij}\}$ son las distancias o similitudes entre los elementos (i, j) tras la aplicación de la técnica de jerarquización. Obviamente, se utilizará la técnica jerárquica que origine el mayor coeficiente.

En el ejemplo TIC, el mayor coeficiente cofenético corresponde al método del promedio o del centroide, si bien el de las otras técnicas de agregación es bastante parecido.

```
# comparación con la distancia euclídea: d_euclidea
cof_simp <- cophenetic(hc_simple)
cof_comp <- cophenetic(hc_completo)
cof_prom <- cophenetic(hc_promedio)
cof_ward <- cophenetic(hc_ward)
cof_dia <- cophenetic(hc_diana)
coef_cofeneticos <- cbind(d_euclidea, cof_simp, cof_comp, cof_prom, cof_ward, cof_dia)

round(cor(coef_cofeneticos)[1, ], 2)
#> d_euclidea   cof_simp   cof_comp   cof_prom   cof_ward   cof_dia
#>      1.00     0.71     0.61     0.77     0.60     0.65
```

30.5.2. Número óptimo de clusters

Acabado el procedimiento de clusterización de los n elementos disponibles, sea por un procedimiento jerárquico aglomerativo o divisivo, hay que tomar una decisión sobre el número de óptimo de clusters, k . Esta decisión es ardua y requiere un delicado equilibrio. Valores grandes de k pueden mejorar la homogeneidad de los clusters; sin embargo, se corre el riesgo de sobreajuste. Lo contrario ocurre con un k pequeño.

Para tomar esta decisión, además del sentido común y el conocimiento que se tenga del fenómeno en estudio, se puede echar mano de distintos procedimientos heurísticos:

- El primero se basa en el **dendrograma** y, en concreto, en la representación de las distintas etapas del algoritmo y las distancias a la que se producen las agrupaciones o particiones de los clusters. Para cada distancia, el dendrograma produce un número determinado de clusters que aumenta (o disminuye) con la misma. Por tanto, el número de clusters dependerá de la distancia a la que se corte el dendrograma (eje de ordenadas del dendrograma, *height*). Dicha distancia debería elegirse de tal forma que los conglomerados estuviesen bien determinados y fuesen interpretables. En las primeras etapas del proceso las distancias no varían mucho, pero en las etapas intermedias y, sobre todo, finales, las distancias aumentan mucho entre dos etapas consecutivas. Por ello, se suele cortar el dendrograma a la distancia a la cual las distancias entre dos etapas consecutivas del proceso empiecen a ser muy grandes, indicador de que los grupos empiezan a ser muy distintos.

30.5. Calidad de la agrupación y número de clusters

521

- Otra posibilidad es utilizar el **gráfico de sedimentación** (Sec. 32.4), que relaciona la variabilidad entre clusters (eje de ordenadas) con el el número de clusters (eje de abscisas). Normalmente, decrece bruscamente al principio, y posteriormente más despacio, hasta llegar a la parte de sedimentación (el codo del gráfico), donde el decrecimiento es muy lento. Pues bien, el número óptimo de conglomerados es el correspondiente al codo o comienzo del área de sedimentación del gráfico.

El algoritmo del gráfico de sedimentación es como sigue:

1. Clusterícese variando el número de grupos, k , por ejemplo, de 1 a 10.
2. Para cada valor de k , compítese la suma de cuadrados intra-grupo (WSS).
3. Trácese la gráfica de WSS vs. k .
4. Determínese el número óptimo de grupos.

Con conjuntos de datos de tamaño pequeño a moderado, este proceso se puede realizar convenientemente con `factoextra::fviz_nbclust()`.

- Otra opción es el *ancho de silueta promedio*. El coeficiente o ancho de silueta compara, por cociente, la distancia media a elementos en el mismo grupo con la distancia media a elementos en otros grupos.

Este método calcula el ancho de silueta promedio (`avg.sil.wid.`) de los elementos objeto de agrupación para diferentes valores de k . Como un valor alto del ancho promedio indica una buena agrupación, el número óptimo de conglomerados es el que lo maximiza. El campo de variación del ancho de silueta es $[-1, 1]$, donde 1 significa que los elementos están muy cerca de su propio cluster y lejos de otros clusters, mientras que -1 indica que están cerca de los clusters vecinos.

- El **criterio del gap (brecha)**, similar al método del codo, tiene como finalidad encontrar la mayor diferencia o distancia entre los diferentes grupos de elementos que se van formando en el proceso de clusterización y que se representan normalmente en un dendrograma. Se computan las distancias de cada uno de los enlaces que forman el dendrograma y se observa cuál es la mayor de ellas. El máximo del gráfico de estas diferencias vs. el número de clusters indica el número óptimo de clusters.

```
p1 <- fviz_nbclust(tic,
  FUN = hcut, method = "wss",
  k.max = 10
) +
  ggtitle("Elbow")
p2 <- fviz_nbclust(tic,
  FUN = hcut, method = "silhouette",
  k.max = 10
) +
  ggtitle("Silhouette")
p3 <- fviz_nbclust(tic,
  FUN = hcut, method = "gap_stat",
  k.max = 10
) +
```

```
ggttitle("Gap")
p1 + p2 + p3
```

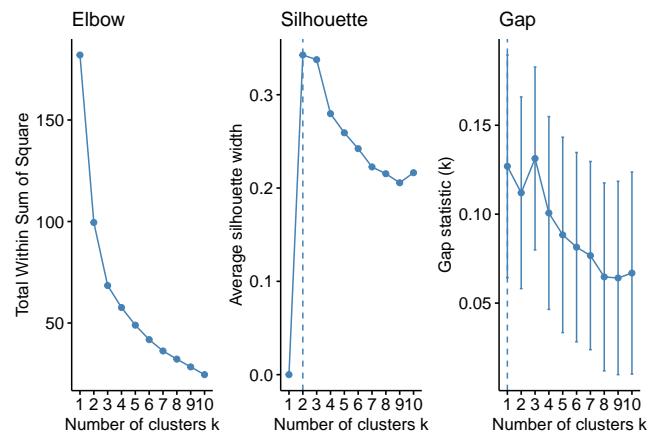


Figura 30.7: Métodos heurísticos para la determinación del número óptimo de clusters

- Finalmente, el **índice de Dunn** es el cociente entre la mínima distancia inter-grupos y la máxima distancia intra-grupos. A mayor índice, mayor calidad de clusterización.

```
library("clValid")
cut2_hc_prom <- cutree(hc_promedio, k = 2)
cut3_hc_prom <- cutree(hc_promedio, k = 3)
cut4_hc_prom <- cutree(hc_promedio, k = 4)
cut5_hc_prom <- cutree(hc_promedio, k = 5)

dunn(d_euclidea, cut2_hc_prom)
#> [1] 0.4465593
dunn(d_euclidea, cut3_hc_prom)
#> [1] 0.3751942
dunn(d_euclidea, cut4_hc_prom)
#> [1] 0.4074884
dunn(d_euclidea, cut5_hc_prom)
#> [1] 0.4366356
```

En el ejemplo TIC, el gráfico de sedimentación y criterio del *gap* indican un número óptimo de clusters de 3. El ancho de silueta alcanza su máximo con dos clusters, si bien la altura del gráfico para tres clusters es prácticamente la misma. Por ello, se opta por 3 clusters a pesar de que el índice de Dunn también se decanta por dos. El primero lo forman Rumanía, Bulgaria y Grecia, la franja sudeste de la UE27, que se caracteriza por tener los peores guarismos en

30.5. *Calidad de la agrupación y número de clusters*

523

dotación y uso de las TIC, tanto a nivel de hogar como de empresa. El segundo lo integran el resto de la franja este más las tres primeras economías de la Unión y Portugal. Tienen unos elevados porcentajes en todas las variables, pero no los mayores, que corresponden a los demás países de la UE27, el tercer conglomerado.

Además de los procedimientos anteriores, hay otros, no tan populares, (*i*) basados en el contraste de hipótesis, suponiendo que los datos siguen alguna distribución multivariante (casi siempre la normal) o (*ii*) procedentes de la abstracción de procedimientos inherentes al análisis multivariante paramétrico; los detalles pueden verse en [Gallardo San-Salvador \(2022\)](#). El paquete **NbClust** de **R** contiene la función **NbClust()**, que calcula 30 índices para valorar el número óptimo de clusters.

Resumen

El análisis cluster está orientado a la agrupación de un conjunto de elementos en grupos, en función de una serie de características, tal que los elementos de cada grupo sean lo más parecidos posible entre sí y lo más diferentes posible de los de otros grupos. Este proceso implica (*i*) la selección de las variables en función de las cuales se van a agrupar; (*ii*) la elección de la distancia o medida de similitud entre ellos; (*iii*) la elección de la técnica para formar los grupos; y (*iv*) la determinación del número óptimo de clusters, cuando sea menester. Estas son las cuestiones que se estudian en este capítulo, si bien, por cuestiones de espacio, en (*iii*) solo se abordan las técnicas de clusterización jerárquicas, estudiándose las no jerárquicas en el siguiente capítulo.

Capítulo 31

Análisis cluster: clusterización no jerárquica

José-María Montero^a y Gema Fernández-Avilés^a

^aUniversidad de Castilla-La Mancha

Como se avanzó en 30.4.1, aunque las técnicas de agrupación jerárquicas son muy utilizadas, existen otras, también muy populares, que se aglutan bajo la denominación de no jerárquicas y que se pueden clasificar, sin ánimo de exhaustividad, en (i) **de optimización o reasignación**; (ii) **basadas en la densidad de elementos**; y (iii) **otras**, como los *métodos directos* (por ejemplo, el *block-*; *bi-*; *co-*; *two-mode cluster*), los *de reducción de la dimensionalidad* (como el *Q-* y el *R-factorial*), los *métodos de clusterización difusa*, o los *basados en mixturas de modelos*.

Las técnicas no jerárquicas proceden con el criterio de la inercia, maximizando la varianza inter-grupos y minimizando la intra-grupos. Se caracterizan porque:

- El número de clusters se suele determinar a priori.
- Utilizan directamente los datos originales, si necesidad de computo de una matriz de distancias o similaridades.
- Los elementos pueden cambiar de cluster.
- Los clusters resultantes no están anidados unos en otros.

31.1. Métodos de reasignación

Los métodos de reasignación permiten que un elemento asignado a un grupo en una determinada etapa del proceso de clusterización sea reasignado a otro grupo, en una etapa posterior, si dicha reasignación implica la optimización del criterio de selección. El proceso finaliza cuando no hay ningún elemento cuya reasignación permita optimizar el resultado conseguido. Estas técnicas suelen asumir un número determinado de clusters a priori y se diferencian entre sí en la manera

de obtener la partición inicial y en la medida a optimizar en el proceso. Respecto a esta última cuestión, los procedimientos más populares son: (i) la minimización de la traza de la matriz de covarianzas intra-grupos; (ii) la minimización de su determinante; (iii) la maximización de la traza del producto de las matrices de covarianzas inter-grupos e intra-grupos; (iv) medidas de información o de estabilidad.

31.1.1. Técnicas basadas en centroides: métodos de Forgy y k-medias

Los algoritmos de reasignación más populares son el de Forgy y, sobre todo, el *k*-medias. La literatura sobre este tipo de técnicas no es clara y, frecuentemente, se confunden el método de Forgy y el *k*-medias, así como el *k*-medias con algunas de sus otras denominaciones (dándose a entender que son técnicas distintas). Sin embargo, la historia es la siguiente: originalmente, [Forgy \(1965\)](#) propuso un algoritmo consistente en la iteración sucesiva, hasta obtener convergencia, de las dos operaciones siguientes: (i) representación de los grupos por sus centroides; y (ii) asignación de los elementos al grupo con el centroide más cercano. Posteriormente, [Diday \(1971\)](#), [Diday \(1973\)](#), [Anderberg \(1973\)](#), [Bock \(1974\)](#) y [Späth \(1975\)](#) desarrollaron una variante del método de Forgy, que solo se diferencia de él en que los centroides se recalculan después de asignar cada elemento (con la técnica de Forgy primero se llevan a cabo todas las asignaciones y posteriormente se recalculan los centroides). Diday la llamó método de las nubes dinámicas o clusters dinámicos, Anderberg se refirió a ella como el criterio de inclusión en el grupo del centroide más cercano, Bock la denominó particionamiento iterativo basado en la mínima distancia, y Späth la llamó HMEANS, una versión por lotes del procedimiento de los autores anteriores. Sin embargo, fue [MacQueen \(1967\)](#) quien previamente acuñó la denominación de “*k*-medias” que se usa hasta la fecha.

K-medias¹ requiere la especificación previa del número de grupos, *k*, en los que se va a dividir el conjunto de elementos. El algoritmo (i) selecciona *k* elementos por algún procedimiento; (ii) asigna los restantes elementos al elemento más cercano de los previamente seleccionados; (iii) sustituye los elementos seleccionados en (i) por los centroides de los grupos que se han formado; (iv) asigna el conjunto de elementos al centroide más cercano del punto (iii); (v) repite iterativamente los dos últimos pasos hasta que la asignación de elementos a los centroides no cambia. Los grupos entonces formados maximizan la distancia inter-grupos y minimizan la distancia intra-grupos. Recuérdese que con el método de Forgy la etapa (iii) no comienza hasta que no se hayan asignado todos los elementos a un cluster en la etapa (ii), mientras que en “*k*-medias” los centroides se recomputan cada vez que un elemento es asignado a un grupo.

La partición que se obtiene es un óptimo local (pequeños cambios en la reasignación de elementos no lo mejoran), pero no se puede asegurar que sea el global, pues se trata de un método heurístico. Sí se puede asegurar que la partición es de calidad.

K-medias es eficiente y sencillo de implementar, pero tiene algunas desventajas: (i) necesita conocer a priori el número de grupos; (ii) la agrupación resultante puede depender de la asignación inicial (normalmente aleatoria) de los centroides, pudiendo converger a mínimos locales, por lo que se recomienda repetir la clusterización 25-50 veces y seleccionar la que tenga menor varianza intra-grupos; (iii) no es robusto a valores extremos; y (iv) no trabaja con datos nominales.

¹Recuérdese que el centro de un conglomerado viene dado por el centroide, vector de medias.

31.1. Métodos de reasignación

527

En el ejemplo TIC, se ha usado el algoritmo AS 136 de [Hartigan and Wong \(1979\)](#), una versión eficiente del de [Hartigan \(1975\)](#) que no busca óptimos locales (varianza intra-grupos mínima en cada grupo), sino soluciones tales que ninguna reasignación de elementos reduzca la varianza (global) intra-grupos (véase Fig. 31.1).

```
set.seed(123)
kmeans_tic <- eclust(tic, "kmeans", k = 3)
```

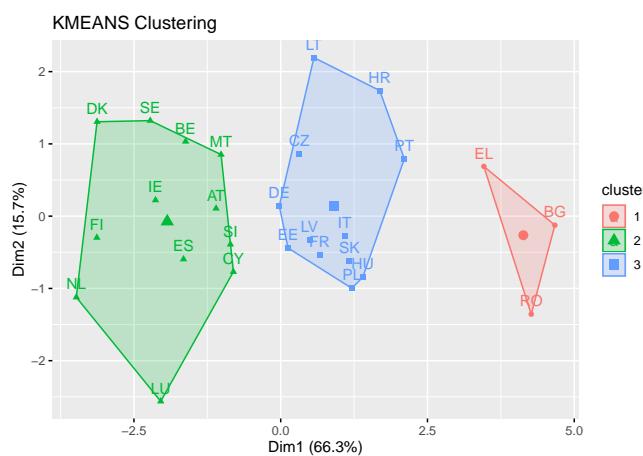


Figura 31.1: Clusterización no jerárquica con k -medias

Algunas versiones del k -medias como el k -medias difuso, el k -medias recortadas, el k -medias armónicas, el k -medias sparse y el k -medias sparse robusto pueden verse en [Carrasco-Oberto \(2020\)](#).

31.1.2. Técnicas basadas en medoides

31.1.2.1. K-medoides (PAM)

Es un método de clusterización similar al k -medias que también requiere la especificación a priori del número de grupos. La diferencia es que, en k -medoides, cada grupo está representado por uno de sus elementos, denominados medoides (o centrotipos)². En el k -medias están representados por sus centroides, que no tienen por qué coincidir con ninguno de los elementos a agrupar. Se trata pues, de formar grupos *particionando el conjunto de elementos alrededor de los medoides* (PAM).

El algoritmo k -medoides es más robusto al ruido y a valores grandes de los datos (de hecho es invariante a los outliers) que el k -medias, ya que minimiza la suma de diferencias por parejas (utiliza la distancia Manhattan) en lugar de la suma de los cuadrados de las distancias

²El medoide de un conglomerado es el elemento del conglomerado con la menor disimilitud promedio entre él y todos los demás miembros del grupo. Es el elemento más céntrico del grupo.

euclídeas³. Además, sus agrupaciones no dependen del orden en que han sido introducidos los elementos, cosa que puede ocurrir con otras técnicas no-jerárquicas, y, como se avanzó anteriormente, propone como centro del cluster un elemento del mismo.

PAM funciona muy bien con conjuntos de datos pequeños (por ejemplo 100 elementos en 5 grupos) y permite un análisis detallado de la partición realizada, puesto que proporciona las características del agrupamiento y un gráfico de silueta, así como un índice de validez propio para determinar el número óptimo de clusters. El algoritmo PAM puede verse al completo en [Kaufman and Rousseeuw \(1990\)](#); para un muy buen resumen véase [Amat Rodrigo \(2017\)](#).

```
set.seed(123)
pam_tic <- eclust(tic, "pam", k = 3)
```

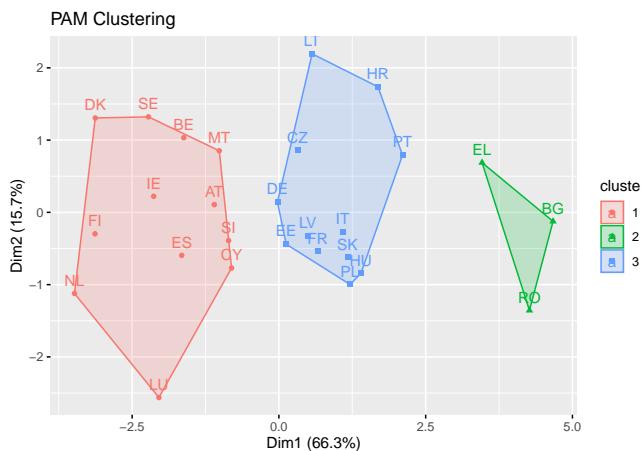


Figura 31.2: Clusterización no jerárquica con PAM

31.1.2.2. CLARA

La ineeficiencia de PAM para bases de datos grandes, junto con su complejidad computacional, llevó al desarrollo de CLARA (clustering Large Applications)⁴

La diferencia entre PAM y CLARA es que el segundo se basa en muestrazos. Solo una pequeña porción de los datos totales es seleccionada como representativa de los datos y los medoides son escogidos (en la muestra) usando PAM. CLARA, pues, combina la idea de k -medoides con el remuestreo para que pueda aplicarse a grandes volúmenes de datos. De acuerdo con [Amat Rodrigo](#)

³Las técnicas basadas en la minimización de promedios de distancias o de residuos en valor absoluto son más robustas que las basadas en sumas de cuadrados.

⁴No obstante, hay una interesante modificación de PAM cambiando el orden de anidamiento de los bucles. La idea es encontrar el mejor intercambio de elementos para cada medoide y ejecutar tantos como sea posible en cada iteración, lo que reduce el número de iteraciones necesarias para la convergencia sin pérdida de calidad ([Schubert and Rousseeuw \(2021\)](#)). También se puede aplicar a los algoritmos CLARA y CLARANS que se verán a continuación.

31.1. Métodos de reasignación

529

(2017) (una descripción completa puede verse en [Kaufman and Rousseeuw \(1990\)](#)), CLARA selecciona una muestra aleatoria y le aplica el algoritmo de PAM para encontrar los clusters óptimos dada esa muestra. Alrededor de esos medoides se agrupan los elementos de todo el conjunto de datos. La calidad de los medoides resultantes se cuantifica con la suma total de distancias intra-grupos. CLARA repite este proceso un número predeterminado de veces con el objetivo de reducir el sesgo de muestreo. Por último, se seleccionan como clusters finales los obtenidos con los medoides que minimizaron la suma total de distancias intra-grupo.

```
set.seed(123)
clara_tic <- eclust(tic, "clara", k = 3)
```



Figura 31.3: Clusterización no jerárquica con CLARA

31.1.2.3. CLARANS

CLARANS (Clustering Large Applications based upon Randomized Search) es una mezcla de PAM y CLARA. Como CLARA puede dar lugar a una mala clusterización si uno de los medoides de la muestra está lejos de los mejores medoides, CLARANS trata de superar esta limitación. El algoritmo puede verse en [Ng and Han \(2002\)](#).

31.1.3. Técnicas basadas en medianas: k-medianas

Igual que el k -medoides, es una variante del k -medias que utiliza como centros las medianas, para que no le afecten ni el ruido ni los valores atípicos. La diferencia con el k -medoides es que la mediana de un grupo no tiene por qué ser una de las observaciones. K -medianas utiliza la distancia Manhattan.

Volviendo al ejemplo TIC, como se ha podido comprobar, la clusterización de los países de la UE27 en función del uso de las TIC es prácticamente la misma con técnicas jerárquicas

aglomerativas como el vecino más lejano, el método de Ward, el del centroide, o algoritmos divisivos como DIANA, que con las técnicas no jerárquicas con pre-selección de 3 grupos (k -medias, PAM y CLARA).

31.2. Métodos basados en la densidad de elementos

Utilizan indicadores de frecuencia, construyendo grupos mediante la detección de aquellas zonas del espacio de las variables (que caracterizan a los elementos) densamente pobladas (clusters naturales) y de aquellas otras con un escasa densidad de elementos. Los elementos que no forman parte de un conglomerado se consideran ruido. Emulan, pues, el funcionamiento del cerebro humano.

La identificación de los grupos (y los parámetros que los caracterizan, cuando se manejan modelos probabilísticos) se lleva a cabo haciéndolos crecer hasta que la densidad del grupo más próximo sobrepase un cierto umbral. Por tanto, imponen reglas para evitar el problema de obtener un solo grupo cuando existen puntos intermedios. Se suele suponer que la densidad de elementos en los grupos es Gaussiana si las variables son cuantitativas, y Multinomial si son cualitativas.

Se suelen clasificar en:

- (i) Las que tienen un **enfoque tipológico**: los grupos se construyen buscando las zonas con mayor concentración de elementos. Pertenece a este tipo el *análisis modal de Wishart*, que supone clusters esféricos y dada la complejidad de su algoritmo no tuvo mucho éxito, el *método TaxMap*, que introduce un valor de corte en caso de que los grupos no estén claramente aislados (ello lleva a que los resultados tengan un cierto grado de subjetividad), y el *método de Fortin*, también con muy escasa difusión en la literatura.
- (ii) Las que tienen un **enfoque probabilístico**: las variables que caracterizan los elementos siguen una distribución de probabilidad cuyos parámetros cambian de un grupo a otro. Se trata, pues, de agrupar los elementos que pertenecen a la misma distribución. Un ejemplo es el *método de las combinaciones de Wolf*.

No obstante, estos algoritmos, y otros como, por ejemplo, los de Gitman y Levine, y Catel y Coulter, aunque muy citados en la literatura en español, tuvieron poco éxito.

Mayor éxito han tenido otros algoritmos como *expectation-maximization* (EM), *model based clustering* (MCLUST), *density-based spatial clustering of applications with noise* (DBSCAN), *ordering point to identify clustering structure clustering* (OPTICS), que es una generalización de DBSCAN, *wavelet-based cluster* (WAVECLUSTER) y *density-based clustering* (DENCLUE), entre otros.

DBSCAN⁵ es, quizás, el más popular. Incluso ha recibido premios por sus numerosísimas aplicaciones a lo largo del tiempo. Soluciona los problemas de los métodos de reasignación, que son

⁵La densidad se refiere al número de elementos en una misma zona. Sin embargo, es un concepto subjetivo, porque (i) ¿cuántos puntos son necesarios para considerar a una zona como densa? y (ii) ¿cómo de distantes pueden estar dichos elementos entre sí? Por ello, ambos (número de puntos y distancia) son los dos hiperparámetros del modelo.

buenos para clusters con forma esférica o convexa que no tengan demasiados *outliers* o ruido, pero que fallan cuando los clusters tienen formas arbitrarias. De acuerdo con [Amat Rodrigo \(2017\)](#), DBSCAN evita este problema siguiendo la idea de que, (i) para que una observación forme parte de un cluster, tiene que haber un mínimo de observaciones vecinas dentro de un radio de proximidad y (ii) que los clusters están separados por regiones vacías o con pocas observaciones. Consecuentemente, DBSCAN necesita dos parámetros: el radio (ϵ) que define la región vecina a una observación (ϵ -neighborhood); y el número mínimo de puntos (minPts) u observaciones en ella.⁶

Los elementos objeto de agrupación se pueden clasificar, en función de su ϵ -neighborhood y minPts, como: (i) *elementos centrales*, si el número de elementos en su ϵ -neighborhood es igual o mayor que minPts; (ii) *elementos frontera*, si no son elementos centrales pero pertenecen al ϵ -neighborhood de otro elemento que sí es central; y (iii) *elementos atípicos o de ruido*, si no verifican ni (i) ni (ii).

A partir de la clasificación anterior, y para ϵ -neighborhood y minPts dados, se origina otra: (i) un elemento Q es *denso-alcanzable directamente* desde el elemento P si Q está en el ϵ -neighborhood de P y P es un elemento central; Q es *denso-alcanzable* desde P si existe una cadena de objetos $\{Q_1 = P, Q_2, Q_3, \dots, Q_n\}$ tal que Q_{i+1} es denso-alcanzable directamente desde Q_i , $\forall 1 \leq i \leq n$; (iii) Q está *denso-conectado* con P si hay un elemento R desde el cual P y Q son denso-alcanzables.

Los pasos del algoritmo DBSCAN son los siguientes ([Amat Rodrigo \(2017\)](#)):

- Para cada elemento u observación x_i calcúlese su distancia con el resto de observaciones. Márquese como central si lo es y como visitado si no lo es.
- Para cada observación marcada como elemento central, si aún no ha sido asignada a ningún grupo, créese un grupo nuevo y asígnesele a él. Búsquese, recursivamente, todas las observaciones denso-conectadas con ella y asígnense al mismo grupo.
- Itérese el mismo proceso para todas las observaciones no visitadas.
- Aquellas observaciones que tras haber sido visitadas no pertenecen a ningún cluster se marcan como *outliers*.

Como resultado del algoritmo DBSCAN se generan clusters que verifican: (i) todos los elementos que forman parte de un mismo cluster están denso conectados entre ellos; y (ii) si un elemento P es denso-alcanzable desde cualquier otro de un cluster, entonces P también pertenece al cluster.

El éxito de DBSCAN se debe a sus importantes ventajas. De nuevo siguiendo a [Amat Rodrigo \(2017\)](#), no requiere la especificación previa del número de clusters; no requiere esfericidad (ni ninguna forma determinada) en los grupos; y puede identificar valores atípicos, por lo que la clusterización resultante no vendrá influenciada por ellos. También tiene algunas desventajas, como que no es un método totalmente determinista puesto que (i) los puntos frontera que son denso-alcanzables desde más de un cluster pueden asignarse a uno u otro dependiendo del orden

⁶Véase [Amat Rodrigo \(2017\)](#) para una discusión sobre el valor de ambos hiperparámetros.

en el que se procesen los datos; y (ii) no genera buenos resultados cuando la densidad de los grupos es muy distinta, ya que no es posible encontrar un ϵ -neighborhood y un minPts válidos para todos a la vez.

Dado el escaso número de datos de la base de datos TIC2021 del paquete CDR no se puede utilizar DBSCAN. Sin embargo, para ilustrar su utilización, la Fig. 31.4 muestra la agrupación en 5 clusters de la base de datos `multishapes` de la librería `factoextra` mediante DBSCAN (función `dbSCAN`) y k -medias. Se trata de una base de datos que contiene observaciones pertenecientes a 5 grupos distintos y con cierto ruido (*outliers*); en consecuencia, los grupos no deberían ser esféricos y DBSCAN sería un algoritmo de agrupación adecuado.

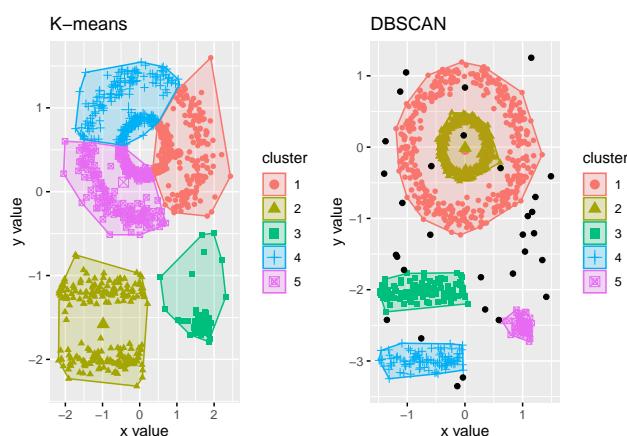


Figura 31.4: Comparación entre los algoritmos k -means y DBSCAN para el conjunto de datos simulado ‘multishapes’

31.3. Otros métodos

Por último, en el cajón de sastre de otras técnicas de clusterización no jerárquicas, merece la pena siquiera mencionar los métodos directos, los de reducción de la dimensionalidad, los de clustering difuso y la clusterización basada en modelos.

Los **métodos directos** agrupan simultáneamente los elementos y las variables. El más conocido es el cluster por bloques (*block-*; *bi-*; *co-*; o *two mode clustering*). El paquete `bicluster` proporciona varios algoritmos para encontrar clusters en dos dimensiones. Además, es muy recomendable para el pre-procesamiento de los datos y para la visualización y validación de los resultados.

Las **técnicas de reducción de la dimensionalidad** buscan factores en el espacio de los elementos (modelo *Q*-factorial) o de las variables (modelo *R*-factorial) haciendo corresponder un cluster a cada factor. Centrándonos en el modelo *Q*-factorial, el método parte de la matriz de correlaciones entre los elementos y rota ortogonalmente los factores encontrados. Dado que

los elementos pueden pertenecer a varios clusters y, por tanto, los clusters pueden solaparse, su interpretación se hace muy difícil.

Las técnicas de clustering difuso precisamente permiten la pertenencia de un elemento a varios clusters, estableciendo un grado de pertenencia a cada uno de ellos. El algoritmo de clustering difuso más popular es fuzzy c -medias, muy similar al k -medias pero que calcula los centroides como una media ponderada (la ponderación es la probabilidad de pertenencia) y, lógicamente, proporciona la probabilidad de pertenencia a cada grupo.

La **clusterización basada en modelos** tiene un enfoque estadístico y consiste en la utilización de una mixtura finita de modelos estocásticos para la construcción de los grupos. - Un vector aleatorio \mathbf{X} procede de una mixtura finita de distribuciones paramétricas si, $\forall \mathbf{x} \subset \mathbf{X}$ su función de densidad conjunta se puede escribir como $f(\mathbf{x}|\psi) = \sum_{g=1}^G \pi_g f_g(\mathbf{x}|\theta_g)$, donde π_g son las proporciones asignadas a cada grupo en la mixtura, tal que $\sum_{g=1}^G \pi_g = 1$; $f_g(\mathbf{x}|\theta_g)$ es la función de densidad correspondiente al g -ésimo grupo y $\psi = (\pi_1, \pi_2, \dots, \pi_G, \theta_1, \theta_2, \dots, \theta_G)$. Las funciones de densidad $f_g(\mathbf{x}|\theta_g)$ suelen ser idénticas para todos los grupos.

En términos menos formales, el clustering basado en modelos considera que los datos observados (multivariantes) han sido generados a partir de una combinación finita de modelos componentes (distribuciones de probabilidad, normalmente paramétricas). A modo de ejemplo, en un modelo resultante de una mixtura de normales multivariantes (el caso habitual), cada componente (cluster) es una normal multivariante y el componente responsable de la generación de una observación específica determina el grupo al que pertenece dicha observación. Para la estimación de la media y matriz de covarianzas, se suele recurrir al algoritmo *expectation-maximization*, una extensión del k -medias⁷. El paquete `mclust` utiliza la estimación máximo verosímil para estimar dichos modelos con distintos número de clusters, utilizando el *Bayesian information criterion* (BIC) para la selección del mejor.

Sus limitaciones fundamentales son (*i*) considerar que las características de los elementos son independientes y (*ii*) que no es recomendable para grandes bases de datos o distribuciones de probabilidad que impliquen un elevado coste computacional.

Una revisión de la evolución de la clusterización basada en modelos desde sus orígenes en 1965 puede verse en [McNicholas \(2016\)](#). Para una idea intuitiva, véase [Amat Rodrigo \(2017\)](#).

31.4. Nota final

La elección de que técnica de clusterización, jerárquica o no, es una decisión del investigador, y dependerá de cómo quiera realizar la agrupación, la métrica de las variables y la distancia o medida de similaridad elegida. No obstante, ambos tipos de técnicas tienen sus ventajas y desventajas, y deberán ser tenidas en cuenta a la hora de decidir. Las jerárquicas adolecen de cierta inestabilidad, lo que plantea dudas sobre la fiabilidad de sus resultados. Además, a veces es difícil decidir cuántos grupos deben seleccionarse. Suelen recomendarse en caso de conjuntos de datos pequeños. En caso de grandes conjuntos de datos, la literatura suele recomendar las no jerárquicas; además, tienen una gran fiabilidad, ya que al permitir la reasignación de los elementos, una incorrecta asignación puede ser corregida posteriormente.

⁷Esta es la razón para incluir este tipo de clusterización entre las técnicas no jerárquicas.

Resumen

En este capítulo se pasa revista a las principales técnicas y algoritmos de agrupación no jerárquicas. Primeramente, se abordan los principales métodos de reasignación, y en particular los basados en centroides (método de Forgy y k -medias), medoides (k -medoides, PAM, CLARA, CLARANS) y medianas (k -medianas). Posteriormente, se exponen las técnicas basadas en la densidad de puntos desde las perspectivas tipológica (análisis modal, métodos TaxMap, de Fortin, de Gitman y Levine, y de Catel y Coulter) y probabilística (método de Wolf), así como se estudia el DBSCAN. Finalmente, se muestran otras técnicas de agrupación no jerárquicas como los métodos directos (*block-; bi; co-; two mode-\$ clustering*), los de reducción de la dimensionalidad (modelos Q - y R -factorial), el clustering difuso y los métodos basados en mixturas de modelos.

Capítulo 32

Análisis de componentes principales

José-María Montero^a y José Luis Alfaro Navarro^a

^aUniversidad de Castilla-La Mancha

32.1. Introducción

En el estudio de cualquier problema de interés, lo ideal es tomar información del mayor número de variables posible, lo cual, actualmente, no es un impedimento. Sin embargo, trabajar con muchas variables es incómodo (por ejemplo, si fueran 30 y se estuviese interesado en su correlación dos a dos, habría que calcular 435 coeficientes). Además, tener muchas variables no implica necesariamente tener mucha información. Si están correlacionadas entre ellas (que suele ser el caso en la realidad), parte de la información que proporcionan es redundante. Por consiguiente, el reto es reducir la dimensionalidad del problema sin reducir la cantidad de información proporcionada por las variables originales, midiéndose dicha cantidad de información a través de su variabilidad, en consonancia con el concepto de entropía. En concreto, se adopta como medida de la variabilidad de las variables originales la suma de sus varianzas.

Pues bien, el análisis de componentes principales (ACP, perteneciente al ámbito del aprendizaje no supervisado) es una técnica de reducción de la dimensionalidad, un problema importante en ciencia de datos, tanto en el aprendizaje supervisado como no supervisado. ACP opera sustituyendo las variables originales por un número reducido de combinaciones lineales de ellas, incorreladas, denominadas **componentes principales** (c.p.), que capturan un elevado porcentaje de la variabilidad de las variables originales (Hothorn and Everitt, 2014; Boehmke, 2020). ACP es el primer intento de reducción de la dimensionalidad y el único utilizado a tal fin hasta el advenimiento del escalamiento multidimensional (aunque no es su función principal) y otras técnicas más complicadas pertenecientes al ámbito del aprendizaje múltiple (*manifold learning*).

La reducción de la dimensionalidad no solo es útil en el estudio de fenómenos complejos con un elevado número de dimensiones, sino también para facilitar la implementación de otros métodos de aprendizaje no supervisado, como el análisis cluster¹ (reduciendo el número de dimensiones a utilizar para configurar los clusters), o supervisado, como, por ejemplo, la regresión (reduciendo el número de regresores y haciéndolos incorrelados, evitando así información redundante y la multicolinealidad); o la técnica de *partial least squares* (PLS, similar a la regresión con c.p. pero que, en vez de ignorar la variable respuesta en la determinación las combinaciones lineales, busca aquellas que, además de explicar la varianza de las variables originales, predicen la variable respuesta lo mejor posible).² También es muy útil para representar gráficamente relaciones multivariantes.

En **R** hay varias opciones para la realización de un ACP: `princomp()`, `prcomp()` y `PCA()`, de la librería **FactoMineR** (Lê et al., 2008), entre otras. Se ha optado por la última porque (i) incorpora notables mejoras gráficas; (ii) permite el ACP con *missing values*, imputando dichos valores (paquete `missMDA`); (iii) proporciona una descripción e interpretación automática de los resultados, seleccionando los mejores gráficos, mediante el paquete **FactoInvestigate**; (iv) permite la implementación de técnicas híbridas (por ejemplo, clusterización con c.p.); y (vi) posibilita la predicción de las coordenadas de individuos y variables adicionales utilizando únicamente inputs del ACP previo.

Como ilustración práctica del ACP, se abordará la reducción de la dimensionalidad de un problema del ámbito de la sociedad de la información en la UE-27, en 2021. Se dispone, para 2021 y a nivel de país, de información sobre 7 variables: 4 relacionadas con el uso de las TIC por parte de las empresas y 3 relativas al uso de dichas tecnologías por parte de las personas y a la equipación TIC de los hogares. Dicha información, así como la descripción de las variables, puede consultarse en la base de datos **TIC2021** del paquete **CDR**.

32.2. Obtención de las componentes principales

32.2.1. Descripción formal del proceso

Sea $\mathbf{X}' = (X_1, \dots, X_p)$ un vector p -dimensional de variables aleatorias con vector de medias $\boldsymbol{\mu}$ y matriz de covarianzas conocida $\boldsymbol{\Sigma}$. Puesto que los cambios de origen no afectan a la covarianza, las variables originales se consideran centradas, de tal manera que $\boldsymbol{\mu} = \mathbf{0}$ y $\boldsymbol{\Sigma} = E(\mathbf{X}'\mathbf{X})$. Se trata de encontrar un conjunto de p combinaciones lineales incorreladas de dichas variables, $Y_j = a_{1j}X_1 + a_{2j}X_2 + \dots + a_{pj}X_p = \mathbf{a}'_j\mathbf{X}$, $j = 1, 2, \dots, p$, denominadas c.p., que recojan la variabilidad existente en los datos. La idea es ordenar las c.p. tal que $V(Y_1) > V(Y_2) > \dots > V(Y_p)$, y seleccionar m de ellas (las m primeras), $m < p$, que capturen un elevado porcentaje de la variabilidad de los datos.

¹Aunque cluster es una palabra inglesa, no se escribirá en cursiva por ser un término popular y muy utilizado en la jerga de la ciencia de datos en español.

²Como señala Amat Rodrigo (2017), PLS puede considerarse como una versión supervisada de la regresión con c.p.

Nota

Geométricamente, las c.p. representan un nuevo sistema de coordenadas obtenido mediante la rotación de los ejes originales. Los nuevos ejes representan las direcciones de máxima variabilidad y proporcionan una descripción más simple de la estructura de covarianza.

La varianza de cada componente y la covarianza entre ellas vienen dadas por:

$$\begin{aligned} \text{Var}(Y_j) &= \mathbf{a}'_j \boldsymbol{\Sigma} \mathbf{a}_j, \quad \forall j = 1, 2, \dots, p, \\ \text{Cov}(Y_j, Y_k) &= \mathbf{a}'_j \boldsymbol{\Sigma} \mathbf{a}_k, \quad \forall j, k \{j \neq k\} = 1, 2, \dots, p. \end{aligned} \quad (32.1)$$

Obtención de la primera componente principal

La primera c.p., Y_1 , se obtiene seleccionando el vector \mathbf{a}_1 que maximice su varianza. Sin embargo, dado que la varianza de cada c.p. puede incrementarse arbitrariamente multiplicando \mathbf{a}_1 por una constante, se impone la condición $\mathbf{a}'_1 \mathbf{a}_1 = 1$; es decir, se normalizan los vectores, de tal forma que tengan longitud unitaria. Por tanto, se trata de encontrar el vector \mathbf{a}_1 que maximiza $\text{Var}(Y_1) = \mathbf{a}'_1 \boldsymbol{\Sigma} \mathbf{a}_1$ sujeto a que $\mathbf{a}'_1 \mathbf{a}_1 = 1$. En otros términos, se selecciona el vector \mathbf{a}_1 que maximiza el lagrangiano:

$$\mathcal{L}(\mathbf{a}_1) = \mathbf{a}'_1 \boldsymbol{\Sigma} \mathbf{a}_1 - \lambda(\mathbf{a}'_1 \mathbf{a}_1 - 1). \quad (32.2)$$

Para ello, se deriva respecto a \mathbf{a}_1 y λ , y se igualan a cero dichas derivadas:

$$\begin{aligned} \frac{\partial \mathcal{L}(\mathbf{a}_1)}{\partial \mathbf{a}_1} &= 2\boldsymbol{\Sigma} \mathbf{a}_1 - 2\lambda \mathbf{a}_1 = (\boldsymbol{\Sigma} - \lambda \mathbf{I}) \mathbf{a}_1 = \mathbf{0}, \\ \frac{\partial \mathcal{L}(\mathbf{a}_1)}{\partial \lambda} &= \mathbf{a}'_1 \mathbf{a}_1 - 1 = 0. \end{aligned} \quad (32.3)$$

La primera ecuación tendrá solución distinta del vector nulo cuando $(\boldsymbol{\Sigma} - \lambda \mathbf{I})$ sea singular. Es decir, cuando $|\boldsymbol{\Sigma} - \lambda \mathbf{I}| = 0$, o en otros términos, cuando λ sea un autovalor de $\boldsymbol{\Sigma}$. Dado que,

- $\boldsymbol{\Sigma}$ es semidefinida positiva y, en general, tendrá p autovalores no negativos,
- y que en el proceso de optimización, premultiplicando $(\boldsymbol{\Sigma} - \lambda \mathbf{I}) \mathbf{a}_1 = \mathbf{0}$ por \mathbf{a}'_1 y teniendo en cuenta que $\mathbf{a}'_1 \mathbf{a}_1 = 1$, resulta que $\lambda = \mathbf{a}'_1 \boldsymbol{\Sigma} \mathbf{a}_1 = V(Y_1)$,³

se seleccionará el mayor de los autovalores de $\boldsymbol{\Sigma}$, obteniéndose el autovector \mathbf{a}_1 de tal forma que cumpla la condición $\mathbf{a}'_1 \mathbf{a}_1 = 1$.

Obtención de la segunda componente principal

$Y_2 = \mathbf{a}'_2 \mathbf{X}$ se obtiene igual que Y_1 , pero añadiendo la condición de incorrelación con Y_1 : $\text{Cov}(Y_1, Y_2) = \mathbf{a}'_2 \boldsymbol{\Sigma} \mathbf{a}_1 = 0$, o equivalentemente, $\mathbf{a}'_2 \mathbf{a}_1 = 0$ (\mathbf{a}_1 y \mathbf{a}_2 ortogonales).⁴

Por tanto, el lagrangiano a maximizar es:

³La condición $\mathbf{a}'_i \mathbf{a}_i = 1$ lleva a que los autovalores de $\boldsymbol{\Sigma}$ coincidan con las varianzas de las c.p.

⁴Si todos los autovalores de $\boldsymbol{\Sigma}$ son distintos, los autovectores son ortogonales. En caso contrario, los autovectores asociados a autovalores comunes se eligen de forma que sean ortogonales.

$$\mathcal{L}(\mathbf{a}_2) = \mathbf{a}'_2 \Sigma \mathbf{a}_2 - \lambda(\mathbf{a}'_2 \mathbf{a}_2 - 1) - \gamma(\mathbf{a}'_2 \mathbf{a}_1 - 0). \quad (32.4)$$

Derivando respecto a \mathbf{a}_2 e igualando a cero:

$$\frac{\partial \mathcal{L}(\mathbf{a}_2)}{\partial \mathbf{a}_2} = 2\Sigma \mathbf{a}_2 - 2\lambda \mathbf{a}_2 - \gamma \mathbf{a}_1 = 2(\Sigma - \lambda \mathbf{I})\mathbf{a}_2 - \gamma \mathbf{a}_1 = \mathbf{0}. \quad (32.5)$$

Premultiplicando por \mathbf{a}'_1 y considerando la condición de ortogonalidad, se tiene que $\gamma = 2Cov(Y_1, Y_2) = 0$, con lo que $\frac{\partial \mathcal{L}(\mathbf{a}_2)}{\partial \mathbf{a}_2} = 2\Sigma \mathbf{a}_2 - 2\lambda \mathbf{a}_2 = 0$, que implica que $(\Sigma - \lambda \mathbf{I})\mathbf{a}_2 = 0$.

Siguiendo el mismo razonamiento que en la obtención de la primera componente, se elige el segundo mayor autovalor de Σ^5 , λ_2 , siendo \mathbf{a}_2 el autovector asociado a él.

Obtención del resto de las componentes principales

Repetiendo este procedimiento, se obtienen las p c.p., siendo los coeficientes de la j -ésima los componentes del autovector asociado al j -ésimo mayor autovalor de Σ .

El vector de c.p. se puede expresar como $\mathbf{Y} = \mathbf{A}'\mathbf{X}$, donde $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_p]$ es la matriz de autovectores (ortogonales) obtenidos.

32.2.2. Cuestiones importantes en el análisis de componentes principales

32.2.2.1. Varianza de las variables originales y las componentes principales

La matriz de covarianzas de las c.p., $\mathbf{V}(\mathbf{Y}) = \mathbf{A}'\Sigma\mathbf{A}$, coincide con Λ , que es una matriz diagonal, puesto que las c.p. están incorreladas y sus varianzas (los valores de la diagonal principal) son los autovalores de Σ . En consecuencia:

$$\sum_{i=1}^p Var(Y_i) = tr(\Lambda) = tr(\mathbf{A}'\Sigma\mathbf{A}) = tr(\Sigma\mathbf{A}\mathbf{A}') = tr(\Sigma) = \sum_{i=1}^p Var(X_i), \quad (32.6)$$

pudiéndose comprobar que la suma de las varianzas de las variables originales⁶ coincide con la suma de las varianzas de las c.p.

Por tanto, la j -ésima c.p. captura un porcentaje de la variabilidad de las variables originales cifrado en $\frac{\lambda_j}{\sum_{j=1}^p \lambda_j} 100$, siendo $\frac{\sum_{j=1}^m \lambda_j}{\sum_{j=1}^p \lambda_j} 100$ la proporción capturada por las m primeras componentes.

⁵El mayor no puede ser, ya que coincidirían \mathbf{a}_1 y \mathbf{a}_2 , en cuyo caso Y_1 e Y_2 no estarían incorrelacionadas.

⁶Recuérdese que se adopta como medida de la variabilidad de las variables originales la suma de sus varianzas.

32.2.2. Componentes principales a partir de variables estandarizadas

A menudo, no solo se centran las variables originales sino que también se estandarizan, para que tengan varianza unitaria.⁷ La razón es que, si las variables originales presentan grandes diferencias en sus escalas de medida o en los rangos de las unidades de medida (edad en años, altura en metros, longitud en kilómetros...), sus combinaciones lineales tendrán poco significado, porque las variables que las conforman no son “igualmente importantes” y en la primera componente tendrá un gran peso la variable original con mayor magnitud (Chatfield and Collins, 1980). Si no fuera el caso, es mejor partir de Σ ; además, la teoría muestral de las c.p. es mucho más compleja cuando las variables están estandarizadas que cuando no lo están (Morrison, 1976).

El mecanismo de obtención de las c.p. no cambia en absoluto, pero su punto de arranque ya no es Σ sino \mathbf{P} , la matriz de correlaciones de dichas variables. Los autovectores de \mathbf{P} son, en general, distintos a los de Σ . Además, la suma de los autovalores, como coincide con la suma de las varianzas de las variables originales, es p , luego el porcentaje de la variación total capturada por la componente j -ésima es $\frac{\lambda_j}{p} 100$, siendo $\frac{\sum_{j=1}^m \lambda_j}{p} 100$ la proporción capturada por las m primeras componentes .

32.2.2.3. Correlación entre las variables originales y las componentes principales

Considérese la variable original X_i y la c.p. $Y_j = a_{1j}X_1 + a_{2j}X_2 + \dots + a_{pj}X_p = \mathbf{a}'_j \mathbf{X}$. Dado que $\mathbf{X}' = [X_1, \dots, X_p]$, entonces $X_i = \mathbf{e}'_i \mathbf{X}$, donde \mathbf{e}_i es un vector de ceros excepto un 1 en la i -ésima posición.

Entonces, como $(\Sigma - \lambda_j \mathbf{I})\mathbf{a}_j = 0$, se tiene que $\Sigma \mathbf{a}_j = \lambda_j \mathbf{a}_j$ y que:

$$Cov(X_i, Y_j) = Cov(\mathbf{e}'_i \mathbf{X}, \mathbf{a}'_j \mathbf{X}) = \mathbf{e}'_i \Sigma \mathbf{a}_j = \mathbf{e}'_i \lambda_j \mathbf{a}_j = \lambda_j a_{ij}, \quad (32.7)$$

$$r_{X_i, Y_j} = \frac{Cov(X_i, Y_j)}{\sqrt{Var(X_i)} \sqrt{Var(Y_j)}} = \frac{\lambda_j a_{ij}}{\sqrt{\sigma_{ii}} \sqrt{\lambda_j}} = \frac{\sqrt{\lambda_j} a_{ij}}{\sigma_{ii}}, \quad (32.8)$$

donde σ_{ii} es el elemento i -ésimo de la diagonal principal de Σ .

Si se parte de variables estandarizadas, entonces se tiene que $r_{Z_i, Y_j} = \sqrt{\lambda_j} a_{ij}$, donde ahora λ_j es el j -ésimo autovalor de \mathbf{P} y a_{ij} es el elemento i -ésimo de su autovector asociado. Sin embargo, el coeficiente de correlación lineal no varía por el hecho de haber estandarizado las variables originales.

Como se verá posteriormente, estos dos coeficientes, r_{X_i, Y_j} y r_{Z_i, Y_j} , serán de gran utilidad en la interpretación de las c.p. Además, como r_{X_i, Y_j} coincide con el coseno del ángulo que forma X_i con Y_j (que es la proyección o coordenada de X_i en el eje de Y_j), resulta de gran ayuda para representar las variables originales en el espacio de las componentes y, por consiguiente, para la interpretación de estas últimas. A mayor coseno, mayor correlación lineal entre X_i e Y_j . Matricialmente, y denominando \mathbf{A}^* a la matriz de coeficientes de correlación lineal entre las

⁷Las variables estandarizadas se denotan por Z_1, Z_2, \dots, Z_p .

variables originales estandarizadas y las c.p., se tiene que $\mathbf{A}^* = \mathbf{A}\Lambda^{\frac{1}{2}}$. \mathbf{A}^* (imprescindible en la interpretación de las c.p.) no cambia por el hecho de estandarizar también las c.p.

Los cuadrados de los elementos de \mathbf{A}^* expresan la proporción de varianza de la variable X_i explicada por la componente Y_j . Por tanto, la suma de los cuadrados de las filas de \mathbf{A}^* será la unidad. Se denomina *contribución* (individual) de X_i a la componente Y_j a la cantidad

$$\frac{r_{X_i, Y_j}^2}{\sum_{i=1}^p r_{X_i, Y_j}^2} = \frac{\cos(X_i, Y_j)}{\sum_{i=1}^p \cos(X_i, Y_j)}.$$

Otras dos expresiones interesantes que involucran \mathbf{A}^* son $\mathbf{A}^*\mathbf{A}^{*\prime} = \mathbf{R}$ y $\mathbf{A}^{*\prime}\mathbf{A}^* = \Lambda$.

32.3. Estimación de las componentes principales

Hasta el momento, se han derivado las c.p. suponiendo conocida la matriz de covarianzas poblacional Σ (o la de correlaciones \mathbf{P}). Sin embargo, este no suele ser el caso en la práctica, por lo que se sustituyen por sus homónimas muestrales $\mathbf{S} = \frac{1}{N}\mathbf{X}'\mathbf{X}$ (o \mathbf{R}). Nada cambia en el proceso de obtención de las c.p., salvo que el punto de partida es \mathbf{S} (o \mathbf{R}) y que los valores de los autovalores y autovectores asociados son estimaciones.

En el ejemplo que nos ocupa, \mathbf{R} puede verse en la Fig. 32.1. Puede apreciarse que la correlación entre las variables es notable en la mayoría de los casos, lo que invita a analizar el problema con menos variables e incorreladas, es decir, mediante ACP.

```
library("CDR")
data("TIC2021")
TIC <- TIC2021

library("corrplot")
corrplot.mixed(cor(TIC))
```

```
library("FactoMineR")
acp <- PCA(TIC, ncp = 7, graph = FALSE)
```

32.4. Número de componentes a retener

Dado que la finalidad de la técnica de componentes principales es la reducción de la dimensionalidad, una decisión clave es el número m de componentes a retener. Los criterios más populares para tomar esta decisión son:

- a) **Seleccionar un número de componentes que capturen, entre todas, un porcentaje de la variabilidad total determinado**

Dicho porcentaje suele estar alrededor del 80 %, si bien, si el número de c.p. es elevado, su interpretación es muy difícil.

32.4. Número de componentes a retener

541

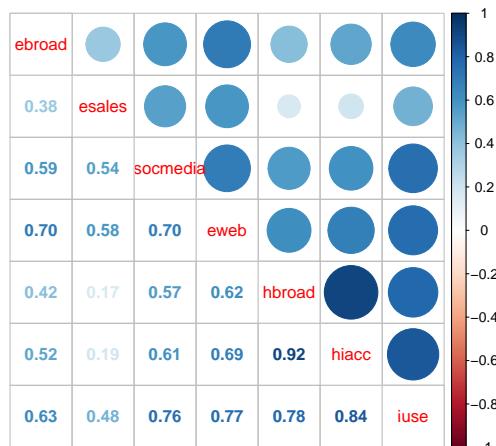


Figura 32.1: Matriz de correlaciones

b) Criterio de la media aritmética o criterio de Kaiser

Dado que la variabilidad total coincide con la suma de los autovalores, se seleccionan aquéllas c.p. cuya varianza exceda la varianza media. Es decir, se selecciona la componente j -ésima si $\lambda_j > \bar{\lambda}$ (si se parte de Σ) o si $\lambda_j > 1$ (si se parte de \mathbf{R}). En caso de valores anómalos (*outliers*) es recomendable utilizar la media geométrica en vez de la aritmética.

c) Criterio de Catell

Se basa en la representación gráfica de los autovalores vs. su número de orden, que se denomina **gráfico de sedimentación** porque se asemeja a la ladera de una montaña con su correspondiente zona de sedimentación. Se seleccionan las c.p. asociadas a los autovalores previos a la zona de sedimentación. En general, el criterio de Catell tiende a incluir demasiadas c.p., al contrario que el de la media, que tiende a incluir demasiado pocas (sobre todo si $p < 20$) (Mardia et al., 1979b).

d) Otros criterios

Otros criterios menos populares son la validación cruzada, el test de esfericidad o igualdad de autovalores de Anderson (requiere normalidad multivariante) y el criterio del bastón roto (véase Cuadras (2007) para los dos últimos). Para grandes conjuntos de datos, Jobson (1992) propone un criterio basado en la partición de la muestra en submuestras mutuamente excluyentes, similar a la validación cruzada.

La Fig. 32.2 muestra el gráfico de sedimentación en el ejemplo que nos ocupa. Puede apreciarse que con tan solo las dos primeras c.p. se captura el 82,07 % de la variabilidad total de las siete variables originales.

```
round(acp$eig[1:7, 1:2], 3)
#>      eigenvalue percentage of variance
#> comp 1      4.644      66.341
#> comp 2      1.101      15.731
#> comp 3      0.547       7.814
#> comp 4      0.328       4.679
#> comp 5      0.191       2.731
#> comp 6      0.124       1.768
#> comp 7      0.066       0.937
library("factoextra")
fviz_eig(acp, addlables = TRUE)
```

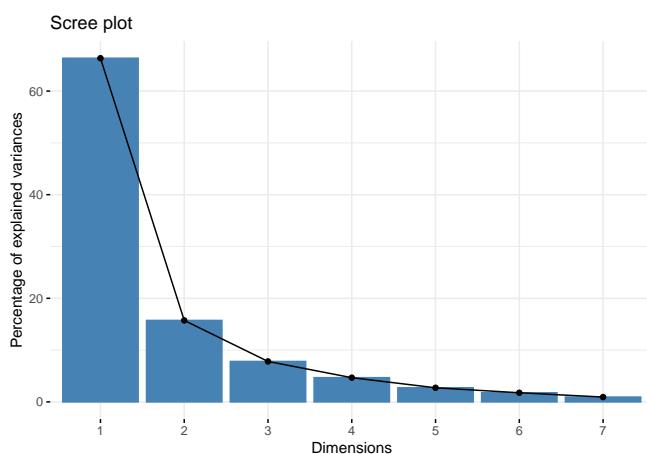


Figura 32.2: Gráfico de sedimentación

32.5. Interpretación de las componentes principales

Una primera vía consiste en analizar el signo y la magnitud de los coeficientes (cargas o *loadings*) de cada variable original en cada componente.

```
loadings <- sweep(acp$var$coord, 2, sqrt(acp$eig[1:7, 1]))
round(loadings, 3)
#>           Dim.1 Dim.2 Dim.3 Dim.4 Dim.5 Dim.6 Dim.7
#> ebroad     -1.410 -0.854 -1.358 -0.561 -0.303 -0.271 -0.259
#> esales     -1.604 -0.318 -0.411 -0.404 -0.310 -0.262 -0.225
#> esocmedia   -1.317 -0.858 -0.645 -1.062 -0.536 -0.311 -0.244
#> eweb        -1.264 -0.865 -0.825 -0.356 -0.773 -0.422 -0.284
#> hbroad     -1.343 -1.550 -0.570 -0.495 -0.413 -0.159 -0.386
#> hiacc      -1.290 -1.496 -0.675 -0.495 -0.413 -0.348 -0.053
#> iuse        -1.217 -1.135 -0.652 -0.587 -0.255 -0.608 -0.331
```

32.5. Interpretación de las componentes principales

543

- Una segunda vía es el análisis de los $r_{X_i, Y_j}, \forall i, j$.

```
round(acp$var$cor, 3)
#>           Dim.1  Dim.2  Dim.3  Dim.4  Dim.5  Dim.6  Dim.7
#> ebroad     0.745  0.195 -0.618  0.012  0.134  0.081 -0.003
#> esales      0.551  0.731  0.328  0.169  0.128  0.090  0.031
#> esocmedia   0.838  0.191  0.095 -0.490 -0.099  0.040  0.012
#> eweb        0.891  0.185 -0.085  0.217 -0.336 -0.070 -0.028
#> hbroad      0.812 -0.501  0.170  0.077  0.024  0.193 -0.129
#> hiacc       0.865 -0.446  0.065  0.077  0.024  0.003  0.203
#> iuse         0.938 -0.086  0.087 -0.014  0.183 -0.256 -0.075
fviz_pca_var(acp,
  col.var = "contrib",
  gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
  repel = TRUE
)
```

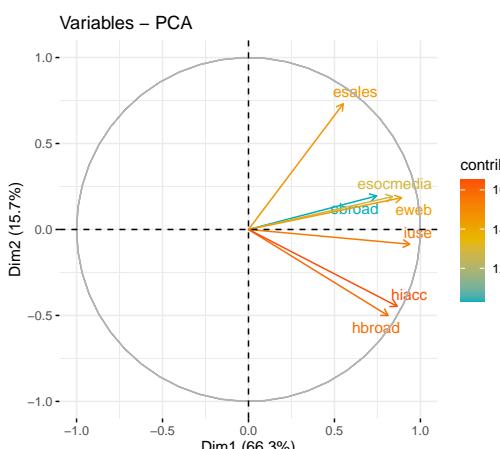


Figura 32.3: Gráfico de cosenos o coeficientes de correlación variables-componentes

Se puede utilizar cualquiera de las dos vías, pues los resultados no serán contradictorios. Sin embargo, algunos autores recomiendan no utilizar sólo la segunda, pues los r_{X_i, Y_j} sólo tienen en cuenta la variable original considerada y no el resto; es decir, se estarían interpretando las componentes desde una perspectiva univariante.

Nota

También es de interés la siguiente consideración: cuando las variables originales están correlacionadas positivamente, la primera c.p. tiene todas sus coordenadas del mismo signo y puede interpretarse como un promedio ponderado de todas ellas.

En la matriz de cargas (o *loadings*) se aprecia que la primera c.p. es una media ponderada (con ponderaciones similares) de las variables originales, mientras que *esales*, *hbroad* y *hiacc*

cargan fuertemente en la segunda (*esales* positivamente y las otras dos de forma negativa). La interpretación desde la perspectiva univariante de los coeficientes de correlación lineal es prácticamente la misma. Por ello, cabe interpretar la primera c.p. como un indicador general del uso de las TIC, mientras que la segunda, positivamente relacionada con la dotación TIC de las empresas pero con una fuerte relación negativa con la de los individuos y hogares, pudiera estar relacionada con las ayudas públicas a la implantación de TICs en el tejido empresarial.

```
acp1 <- fviz_contrib(acp, choice = "var", axes = 1, top = 10)
acp2 <- fviz_contrib(acp, choice = "var", axes = 2, top = 10)
library(patchwork)
acp1 + acp2
```

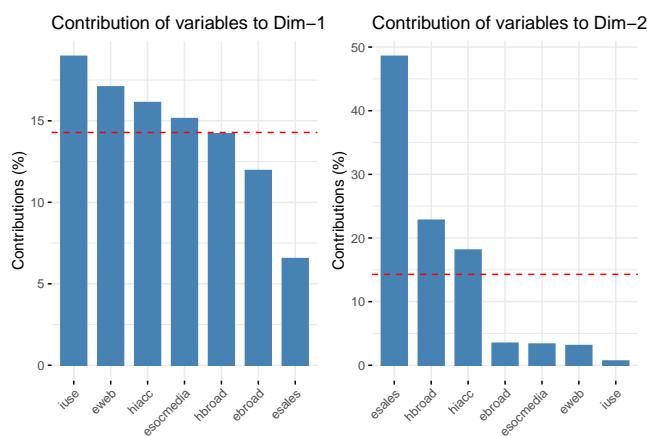


Figura 32.4: Contribución de las variables originales a las componentes retenidas

32.6. Reproducción de los datos tipificados y de la matriz de

En la práctica, el punto de partida del ACP es la matriz **R**, y en tal caso se suelen estandarizar también las c.p. Pues bien, se tiene que:

$$\mathbf{Y}^* = \mathbf{X}\mathbf{A}\Lambda^{-\frac{1}{2}} = \mathbf{X}\mathbf{A}\Lambda^{\frac{1}{2}}\Lambda^{-1} = \mathbf{X}\mathbf{A}^*\Lambda^{-1} = \mathbf{X}\mathbf{F},$$

donde **F** es la matriz de puntuaciones de las c.p. La expresión anterior proporciona las coordenadas de los N elementos en el espacio de las c.p. y, por tanto, sirve de ayuda en la interpretación de éstas. La estandarización de las c.p. asegura que los m ejes (componentes) tengan una métrica homogénea que facilitará la visualización e interpretación. Dichas coordenadas también constituyen el input de técnicas híbridas como, por ejemplo, regresión con c.p., cluster o PLS.

En este caso:

32.7. Limitaciones del análisis de componentes principales

545

```
puntuaciones <- acp$ind$coord
round(puntuaciones[1:5, ], 3)
#> Dim.1 Dim.2 Dim.3 Dim.4 Dim.5 Dim.6 Dim.7
#> BE 1.651 1.053 0.310 -0.238 -0.196 0.296 -0.196
#> BG -4.759 -0.127 -0.128 -0.223 0.023 -0.013 -0.106
#> CZ -0.324 0.875 -0.564 0.870 -0.082 -0.032 -0.345
#> DK 3.188 1.331 0.497 0.262 0.343 -0.300 0.319
#> DE 0.024 0.144 -0.183 0.560 -0.871 -0.485 0.115
```

De la expresión anterior se deduce que $\mathbf{X} = \mathbf{Y}\Lambda^{-\frac{1}{2}}\Lambda^{\frac{1}{2}}\mathbf{A}' = \mathbf{Y}^*_{N \times m}\mathbf{A}_{m \times p}^{*'}\mathbf{Y}_{N \times m}\mathbf{A}_{m \times p}$, expresión que permite reproducir la matriz \mathbf{X} a partir de las m primeras c.p. estandarizadas. En consecuencia, la reproducción de \mathbf{R} a partir de las m primeras c.p. estandarizadas se lleva a cabo como sigue:

$$\begin{aligned} \mathbf{R}_{p \times p} &= \frac{1}{N}\mathbf{X}'_{p \times N}\mathbf{X}_{N \times p} = \mathbf{A}_{p \times m}^*\frac{1}{N}\mathbf{Y}_{m \times N}^*\mathbf{Y}_{N \times m}\mathbf{A}_{m \times p}^{*'} \\ &= \mathbf{A}_{p \times m}^*\mathbf{I}_{m \times m}\mathbf{A}_{m \times p}^{*'} = \mathbf{A}_{p \times m}^*\mathbf{A}_{m \times p}^{*'}. \end{aligned} \quad (32.9)$$

Debajo se muestran las tres primeras filas de la reproducción de \mathbf{R} a partir de las dos primeras c.p. De la comparación de sus valores con los de la verdadera \mathbf{R} (Fig. 32.1) se concluye que se trata de una buena reproducción.⁸

```
matrix <- acp$var$coord[, 1:2] %*% t(acp$var$coord)[1:2, ]
round(matrix[1:3, ], 3)
#> ebroad esales esocmedia eweb hbroad hiacc iuse
#> ebroad 0.593 0.553 0.662 0.700 0.507 0.557 0.682
#> esales 0.553 0.839 0.602 0.626 0.081 0.151 0.455
#> esocmedia 0.662 0.602 0.740 0.782 0.585 0.640 0.770
```

32.7. Limitaciones del análisis de componentes principales

Una primera limitación es que su implementación sólo es posible si todas las variables se trabajan bajo un nivel de análisis numérico. Otra limitación importante es el supuesto subyacente de que los datos observados son combinación lineal de una cierta base. Es decir, sólo se consideran las combinaciones lineales de las variables originales. Otros métodos de reducción de la dimensionalidad como, por ejemplo, el *t-distributed stochastic neighbor embedding* (t-SNE), o la versión Kernel de la técnica, que también funcionan con no linealidad, superan esta limitación.

Además, el hecho que todas las c.p. sean combinaciones lineales de todas las variables originales dificulta su interpretación. Para superar esta limitación, han surgido algunas alternativas, como el *sparse PCA*, que obtiene las c.p. como un problema minimización del error de reconstrucción forzando a que los autovectores tengan una gran parte de sus componentes nula.

⁸Con tres c.p. la reproducción es casi perfecta. Se han retenido sólo dos para poder mostrar representaciones bidimensionales y para evitar la interpretación de una tercera componente.

El t-SNE no es la única alternativa no lineal procedente de la comunidad de *machine learning*. Otras, denominadas actualmente “aprendizaje múltiple” (*manifold learning*) , incluyen el *Sammon's mapping*, el *curvilinear component analysis* (CCA) y sus variantes: los *Laplacian eigenmaps* y el *maximum variance unfolding* (MVU); véase [Wismüller et al. \(2010\)](#).

Finalmente, señalar que el ACP es una técnica matemática que no requiere que las variables originales sigan una distribución normal multivariante, aunque, si así fuera, se podría dar una interpretación más profunda de las c.p.

Resumen

El ACP es una técnica de reducción de la dimensionalidad que captura un gran porcentaje de la variabilidad de un conjunto de variables correladas a partir de un número mucho menor de componentes latentes (las componentes principales) incorreladas. La piedra angular de la construcción de estas componentes son los autovalores de la matriz de covarianzas (o de correlaciones) de las variables originales. En el ACP son cuestiones importantes, entre otras, (i) la determinación del número de componentes a retener, (ii) su interpretación y (iii) la cuantificación del valor de las componentes para cada observación (puntuaciones), que constituyen el input de técnicas híbridas como, por ejemplo, regresión con componentes principales, cluster o *partial least squares*.

Capítulo 33

Análisis factorial

José-María Montero^a y José Luis Alfaro Navarro^a

^aUniversidad de Castilla-La Mancha

33.1. Introducción

Según el trabajo pionero de Harman (1976), el objeto del **Análisis Factorial** (AF) es la representación de una variable X_j en términos de varios factores subyacentes no observables¹. En el marco lineal, y considerando p variables², hay varias alternativas dependiendo del objetivo que se pretenda:

- La captura de la mayor cantidad posible de la varianza de dichas variables (o “explicación” de su varianza).
- La mejor reproducción (o “explicación”) de sus correlaciones observadas.

A modo introductorio, supónganse dos variables polítómicas que surgen de las respuestas de N futbolistas profesionales a dos preguntas: (1) ¿Está usted a gusto en el club? y (2) ¿Se quedaría en el club la siguiente temporada? Las posibles respuestas son: 1, 2, 3, 4, 5 (1 “en total desacuerdo” y 5 “totalmente de acuerdo”).

Cada variable tiene su varianza (nula si todos los futbolistas opinan igual y máxima si la mitad marcase el 1 y la otra mitad el 5). Esta varianza puede ser *común* o *compartida* por las dos variables, o no. Lo normal es que cuanto más a gusto estén los futbolistas en su club mayor sea su deseo de permanecer en él la siguiente temporada, por lo que gran parte de la variabilidad de cada una de las variables sería compartida (ya que la relación -lineal- entre ellas es positiva). El resto de la variabilidad sería *específica* de cada variable (puede que un futbolista

¹Se asumirá la representación lineal, por sencillez, pero puede ser cualquier otra.

²Por las mismas razones que en el análisis de componentes principales (ACP), se trabaja con las variables estandarizadas; véase Cap. 32.

esté muy bien en el club, pero quiera ir a otro más prestigioso; o que esté mal, pero a su familia le encante la ciudad) o *residual* (normalmente debida a factores de medida). El porcentaje de *varianza compartida* se mide a través del coeficiente de determinación lineal, r^2 . El resto, hasta la varianza unidad, o el 100 %, es *varianza única* de cada variable, que incluye tanto la específica como la residual.

De acuerdo con [De la Fuente \(2011\)](#), en el AF caben dos enfoques:³

- El análisis de toda la varianza (común y no común).
- El análisis, únicamente, de la varianza común.

Ambos caben bajo el paraguas genérico del AF; ambos se basan en las relaciones entre las variables para identificar grupos de ellas asociadas entre sí. Sin embargo, del primero se ocupa el ACP (Cap. 32) y, si se parte de la matriz de correlaciones (cuyas entradas fuera de la diagonal principal, al cuadrado, indican la proporción de varianza compartida por las variables que se cruzan en dicha entrada), ésta lleva unos en la diagonal principal. Al segundo se le aplica la denominación de AF y en la matriz de correlaciones se sustituyen los unos de la diagonal principal por la varianza que cada variable comparte con las demás (su **comunalidad**). Por eso se dice que el objetivo del AF es la explicación de la varianza compartida o común de las variables en estudio mediante una serie de **factores comunes** latentes.⁴

El AF puede ser exploratorio o confirmatorio. En el primero no se establecen consideraciones a priori sobre el número de factores comunes a extraer, sino que éste se determina a lo largo del análisis. Por el contrario, en el segundo se trata de contrastar hipótesis relativas al número de factores comunes, así como sobre qué variables serán agrupadas o tendrán más peso en cada factor. Una práctica habitual es validar mediante el *análisis factorial confirmatorio* los modelos teóricos basados en los resultados del *análisis factorial exploratorio*. Sin embargo, [Pérez-Gil et al. \(2000\)](#) alertan de los peligros de esta práctica. Este capítulo considera la versión exploratoria del AF.

A efectos prácticos, se utilizará la base de datos TIC2021 del paquete CDR, ya trabajada en Cap. 32 para el ACP, relativa al uso (por empresas e individuos) y equipación (de los hogares) de las TIC en los países de la UE-27, así como la librería psych ([Revelle, 2022](#)) de R.

```
library("psych")
library("CDR")
data("TIC2021")
```

33.2. Elementos teóricos del análisis factorial

³No son los únicos.

⁴Ambos enfoques dan resultados similares cuando hay más de 30 variables y las communalidades (véase Sec. 33.2.1) son superiores a 0,70, y se interpretan de manera casi idéntica.

33.2.1. Modelo básico y terminología

Considérense p variables $\{X_1, X_2, \dots, X_p\}$ y N elementos, objetos o individuos, siendo las matrices de datos, \mathbf{X} , y datos estandarizados, \mathbf{Z} , las siguientes:

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1N} \\ x_{21} & x_{22} & \cdots & x_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pN} \end{pmatrix}, \quad \mathbf{Z} = \begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1N} \\ z_{21} & z_{22} & \cdots & z_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ z_{p1} & z_{p2} & \cdots & z_{pN} \end{pmatrix},$$

donde el primer subíndice indica la variable y el segundo el elemento.

Mientras que el enfoque de componentes principales está representado por:

$$Z_j = a_{j1}F_1 + a_{j2}F_2 + \cdots + a_{jp}F_p, \quad j = 1, 2, \dots, p, \quad (33.1)$$

en el enfoque AF clásico el modelo teórico es:

$$Z_j = a_{j1}F_1 + a_{j2}F_2 + \cdots + a_{jk}F_k + b_jSP_j + c_jE_j, \quad j = 1, 2, \dots, p, \quad (33.2)$$

donde Z_j , $j = 1, 2, \dots, p$, se modeliza, linealmente, en términos de (i) $k \ll p$ **factores comunes**, F_m , $m = 1, 2, \dots, k$, que dan cuenta de la correlaciones entre las variables Z_j , $j = 1, 2, \dots, p$, y (ii) un **factor específico**, SP_j , $j = 1, 2, \dots, p$, y un término de error, E_j , $j = 1, 2, \dots, p$, que dan cuenta de la **varianza no compartida** (específica y residual, respectivamente). Los coeficientes a_{jm} se denominan **cargas factoriales** y, aunque su notación es igual que en el modelo de componentes principales, no tienen por qué coincidir; el problema básico del AF es precisamente la estimación de dichas cargas. En lo que sigue, se aunará el factor específico y el término de error de Z_j en un **factor único**, U_j , con lo que:

$$Z_j = a_{j1}F_1 + a_{j2}F_2 + \cdots + a_{jk}F_k + d_jU_j, \quad j = 1, 2, \dots, p, \quad (33.3)$$

Los supuestos del modelo (33.3) son los siguientes:

- Como en la práctica los factores comunes y únicos son desconocidos, sin pérdida de generalidad pueden suponerse con media cero y varianza unitaria;
- Los factores únicos se suponen independientes entre sí y de los factores comunes;
- Y dado que los factores involucrados en el modelo se consideran variables aleatorias, si se asume normalidad, e independencia de los factores comunes, $\{F_1, F_2, \dots, F_k\}$ sigue una distribución normal multivariante y Z_j , $j = 1, 2, \dots, p$, una distribución normal.

En términos de valores observados, el modelo AF (33.3) viene dado por:⁵

$$z_{ji} = a_{j1}f_{1i} + a_{j2}f_{2i} + \cdots + a_{jk}f_{ki} + d_ju_{ji}, \quad i = 1, 2, \dots, N; \quad j = 1, 2, \dots, p, \quad (33.4)$$

⁵Aunque en el modelo figuran explícitamente los valores de los factores, en la práctica hay que estimarlos. En otras versiones del AF estos valores se estiman conjuntamente con los parámetros.

El modelo AF es muy parecido al de regresión lineal: una variable se describe como una combinación lineal de otro conjunto de variables más un residuo. Sin embargo, en el análisis de regresión las variables son observables, mientras que en el AF son construcciones hipotéticas que sólo pueden estimarse a partir de los datos observados. Los propios factores se estiman en una etapa posterior del análisis.

En términos matriciales, y considerando:

$$\mathbf{z} = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_p \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} F_1 \\ F_2 \\ \vdots \\ F_k \end{pmatrix}, \quad \mathbf{u} = \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_p \end{pmatrix},$$

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pk} \end{pmatrix}, \quad \mathbf{D} = \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_p \end{pmatrix},$$

el modelo (33.3) puede expresarse como $\mathbf{z} = \mathbf{Af} + \mathbf{Du}$.

Centrándonos en el modelo (33.3), la varianza de Z_j viene dada por:

$$V(Z_j) = 1 = \sum_{m=1}^k a_{jm}^2 + 2 \sum_{m < q} a_{jm} a_{jq} r_{(F_{mi}, F_{qi})} + d_j^2, \quad (33.5)$$

y si los factores comunes están incorrelados, $V(Z_j) = 1 = \sum_{m=1}^k a_{jm}^2 + d_j^2$.

De la expresión (33.5) surgen las siguientes definiciones:

- a_{jm}^2 es la contribución de F_m a la varianza de Z_j .
- $V_m = \sum_{j=1}^p a_{jm}^2$ es la contribución de F_m a suma de las varianzas de todas las variables Z_j , $j = 1, 2, \dots, p$.
- $V = \sum_{m=1}^k V_m$ es la contribución de todos los factores comunes a la varianza de todas las variables Z_j , $j = 1, 2, \dots, p$.
- $\frac{V}{p}$ es un indicador de la *completitud* del análisis factorial.
- $h_j^2 = a_{j1}^2 + a_{j2}^2 + \dots + a_{jk}^2$ es la communalidad de Z_j , $j = 1, 2, \dots, p$, es decir la contribución de los factores comunes a la variabilidad de Z_j .
- d_j^2 es la *unicidad* (o varianza única) de Z_j , $j = 1, 2, \dots, p$, o contribución de U_j a la varianza de Z_j . Es un indicador de la medida en la que los factores comunes fracasan a la hora de representar la varianza (unitaria) de Z_j .
- Cuando se descompone el factor único en sus dos componentes (modelo (33.2)), b_j^2 se denomina *especificidad* (o varianza específica) de Z_j y es la varianza de Z_j debida a la particular selección de las variables en el estudio, mientras que c_j^2 es la que se debe al error (normalmente de medida), que mide la “falta de fiabilidad”.

33.2.2. Patrón y estructura factoriales

Se denomina **patrón factorial** a la siguiente expansión del modelo (33.3),

$$\begin{aligned} Z_1 &= a_{11}F_1 + a_{12}F_2 + \cdots + a_{1k}F_k + d_1U_1 \\ Z_2 &= a_{21}F_1 + a_{22}F_2 + \cdots + a_{2k}F_k + d_2U_2 \\ &\vdots \quad \vdots \quad \ddots \quad \vdots \quad \vdots \\ Z_p &= a_{p1}F_1 + a_{p2}F_2 + \cdots + a_{pk}F_k + d_pU_p \end{aligned} \tag{33.6}$$

o simplemente a la tabla, o matriz, con los coeficientes a_{jm} y d_j (o únicamente, caso habitual, a la *matriz de cargas A*). k determina la “complejidad del modelo”.

Se denomina **estructura factorial** al siguiente conglomerado de $k+1$ conjuntos de p ecuaciones lineales, en $\{a_{jm}\}$, los k primeros, y en $\{d_j\}$, el último, $j = 1, 2, \dots, p$; $m = 1, 2, \dots, k$:⁶

$$\begin{aligned} r_{Z_j F_1} &= a_{j1}r_{F_1 F_1} + a_{j2}r_{F_1 F_2} + \cdots + a_{jm}r_{F_1 F_m} + \cdots + a_{jk}r_{F_1 F_k} \\ &\vdots \quad \vdots \quad \vdots \quad \ddots \quad \vdots \quad \ddots \quad \vdots \\ r_{Z_j F_m} &= a_{j1}r_{F_m F_1} + a_{j2}r_{F_m F_2} + \cdots + a_{jm}r_{F_m F_m} + \cdots + a_{jk}r_{F_m F_k} \\ &\vdots \quad \vdots \quad \vdots \quad \ddots \quad \vdots \quad \ddots \quad \vdots \\ r_{Z_j F_k} &= a_{j1}r_{F_k F_1} + a_{j2}r_{F_k F_2} + \cdots + a_{jm}r_{F_k F_m} + \cdots + a_{jk}r_{F_k F_k} \\ &\vdots \quad \vdots \quad \vdots \quad \ddots \quad \vdots \quad \ddots \quad \vdots \\ r_{Z_j U_j} &= d_j \\ &\vdots \end{aligned} \tag{33.7}$$

En la práctica, viene dada por una tabla, o matriz, Γ , con los coeficientes $\{r_{jm}\}$. Cuando todos los factores están incorrelados el patrón y la estructura coinciden.

El conjunto patrón factorial más estructura factorial se denomina **solución factorial completa**. El patrón factorial muestra la relación lineal de las variables en términos de los factores, como si de una regresión lineal se tratase, y puede usarse para reproducir la correlación entre las variables (y, por tanto, para determinar la bondad de la solución). La estructura factorial es útil para la identificación de los factores y la posterior estimación de las **puntuaciones factoriales**.

En términos matriciales, denominando

$$\mathbf{F} = \begin{pmatrix} f_{11} & f_{12} & \cdots & f_{1N} \\ f_{21} & f_{22} & \cdots & f_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ f_{k1} & f_{k2} & \cdots & f_{kN} \end{pmatrix}, \quad \boldsymbol{\Gamma} = \begin{pmatrix} r_{Z_1 F_1} & r_{Z_1 F_2} & \cdots & r_{Z_1 F_k} \\ r_{Z_2 F_1} & r_{Z_2 F_2} & \cdots & r_{Z_2 F_k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{Z_p F_1} & r_{X_p^* Z_p F_2} & \cdots & r_{Z_p F_k} \end{pmatrix},$$

⁶En caso de variables dicotómicas se utiliza el coeficiente ϕ (véase 23.4) como medida de correlación momento-producto.

el patrón factorial viene dado por $\mathbf{Z} = \mathbf{AF} + \mathbf{DU}$. Multiplicando por \mathbf{F}' y realizando simples operaciones se tiene que $\mathbf{\Gamma} = \mathbf{A}\Phi$, donde Φ es la matriz de correlaciones entre los factores comunes. Si los factores comunes están incorrelados, $\mathbf{\Gamma} = \mathbf{A}$.

Por último, resaltar que el AF es indeterminado, es decir, dado un conjunto de correlaciones, los coeficientes del patrón factorial no son únicos (dado \mathbf{R} , se pueden encontrar infinitos sistemas de factores incorrelados u ortogonales)⁷ consistentes con ella. Por ello, normalmente, tras obtener una solución que ajuste bien los datos originales, se lleva a cabo una rotación de la misma (que ajusta igual de bien dichos datos) que facilite la *interpretación de los factores*.⁸

33.3. El análisis factorial en la práctica

33.3.1. Pre-análisis factorial

33.3.1.1. ¿Procede la realización de un análisis factorial?

Antes de comenzar con el AF, conviene determinar si procede o no; es decir, si las variables se encuentran fuertemente intercorrelacionadas o no. En caso negativo, el AF no tendría sentido. Para ello, se pueden utilizar procedimientos sencillos como observar si el determinante de \mathbf{R} es bajo (correlaciones altas) o elevado (correlaciones bajas); o calcular la *matriz de correlaciones anti-imagen*, cuyos elementos son los coeficientes de correlación parcial cambiados de signo. En la diagonal muestra la *medida de adecuación muestral* para esa variable, MSA_j . Para que se den las condiciones de realización del AF, la mayoría de los elementos no diagonales deben ser pequeños y los diagonales deben estar próximos a la unidad.

Otras alternativas más sofisticadas incluyen las dos siguientes:

Contraste de esfericidad de Bartlett

Exige normalidad multivariante. Contrasta la incorrelación de las variables, es decir, $H_0 : \mathbf{R} = \mathbf{I}$ frente a $H_1 : \mathbf{R} \neq \mathbf{I}$ (o $H_0 : |\mathbf{R}| = 1$ frente a $H_1 : |\mathbf{R}| \neq 1$). El estadístico de contraste es $d_{\mathbf{R}} = -(N - 1 - \frac{1}{6}(2p + 5)) \ln|\mathbf{R}|$ y, bajo H_0 , sigue una $\chi^2_{\frac{p(p-1)}{2}}$, siendo nulo en caso de incorrelación.

```
n <- dim(TIC2021)[1]
cortest.bartlett(cor(TIC2021),n)$chisq
#> [1] 149.7113
cortest.bartlett(cor(TIC2021),n)$p.value
#> [1] 1.992514e-21
```

⁷La historia es más larga: el AF es indeterminado porque dada una matriz $\mathbf{C}_{k \times k}$ no singular, si se define otro vector de factores comunes $\mathbf{f}^* = \mathbf{C}^{-1}\mathbf{f}$ y otra matriz $\mathbf{A}^* = \mathbf{AC}$, entonces $\mathbf{Z} = \mathbf{Af} + \mathbf{Du} = \mathbf{A}^*\mathbf{C}^{-1}\mathbf{C}\mathbf{f}^* + \mathbf{Du} = \mathbf{A}^*\mathbf{f}^* + \mathbf{Du}$ y ambos son equivalentes. Una solución es exigir la incorrelación de los factores comunes ($\Phi = \mathbf{I}$), con lo que la indeterminación se reduciría sólo a cuando \mathbf{C} sea ortogonal. En este caso, la solución será única salvo rotaciones ortogonales.

⁸Una solución determina el espacio k -dimensional que contiene los k factores comunes, pero no su posición exacta.

Medida de adecuación muestral de Kaiser, Meyer y Olkin (KMO)

Se basa en la idea de que, entre cada par de variables, el coeficiente de correlación parcial (que mide la correlación existente entre cada par de ellas eliminando el efecto que el resto de variables tiene sobre las dos consideradas), debe ser cercano a cero, puesto que es una estimación de la correlación entre sus factores específicos, que se suponen incorrelados. Por tanto, si el número de coeficientes de correlación parcial no nulos es elevado, la solución factorial no es compatible con los datos.

En otros términos, cuando las variables incluidas en el análisis comparten gran cantidad de información debido a la presencia de factores comunes, la correlación parcial entre cualquier par de variables debe ser reducida. Por el contrario, cuando dos variables comparten gran cantidad de información entre ellas, pero no la comparten con las restantes variables (ni, consecuentemente, con los factores comunes), la correlación parcial entre ellas será elevada, lo cual es un mal síntoma de cara a la idoneidad del AF.

El índice KMO se define como $KMO = \frac{\sum_{j \neq i} r_{ji}^2}{\sum_{j \neq i} r_{ji}^2 + \sum_{j \neq i} r_{ji}^{*2}}$, donde r_{ji}^* es el coeficiente de

correlación parcial entre las variables Z_j y Z_i . Se considera que valores por encima 0,9 implican elevadísimas correlaciones en **R**; entre 0,5 y 0,9 permiten el AF; y por debajo de 0,5 resultan inaceptables para el AF.

Las MSA_j mencionadas anteriormente, son la versión del índice KMO limitado a cada variable:

$$MSA_j = \frac{\sum_{i \neq j} r_{ji}^2}{\sum_{i \neq j} r_{ji}^2 + \sum_{i \neq j} r_{ji}^{*2}}.$$

La interpretación es similar a la de KMO, pero mide la adecuación de cada variable para realizar un AF, lo que permite no considerar aquellas variables con menor MSA de cara a mejorar la KMO. No obstante, para eliminar una variable del estudio es aconsejable tener en cuenta también las comunidades de cada variable, los residuos del modelo e interpretar los factores obtenidos.

```
round(KMO(TIC2021)$MSA,3)
#> [1] 0.83
round(KMO(TIC2021)$MSAi,3)
#>   ebroad   esales esocmedia      eweb    hbroad     hiacc      iuse
#>   0.850    0.671    0.934    0.856    0.808    0.764    0.875
```

Como puede apreciarse, en nuestro ejemplo TIC, tanto el test de Barlett como el índice *KMO* y las MSA_j indican que el AF se puede llevar a cabo con garantías.

33.3.1.2. El problema de la communalidad y/o del número de factores comunes

El objetivo del AF es encontrar una *matriz de correlaciones reproducida* a partir de los resultados obtenidos, \mathbf{R}^{rep} , con menor rango que la original, \mathbf{R} , tal que su diferencia, la *matriz de correlaciones de los residuos*, \mathbf{R}^{res} , se atribuya únicamente a errores muestrales. \mathbf{R} es una matriz gramiana: simétrica de números reales y de diagonal principal dominante, con lo cual es semidefinida positiva y sus autovalores son nulos o positivos. Por tanto, el número de factores comunes será igual al de autovalores positivos ($k \leq p$). Si el punto de partida en el análisis es \mathbf{R} , rara vez se obtienen menos factores comunes que variables originales, con lo cual el AF realmente es un ACP. Ahora bien, como el número de factores comunes coincide con el rango de \mathbf{R}^{rep} , y éste se ve afectado por los valores de la diagonal principal, al sustituir los unos por las estimaciones de las communalidades (en este caso se está realizando un AF), \mathbf{R}^{rep} no será, en general, gramiana y $k < p$. En conclusión: como la solución factorial ($k < p$) pasa por el conocimiento del rango de \mathbf{R} o de las communalidades, o se hipotetiza sobre dicho rango o se hipotetizan o estiman las communalidades. Normalmente se sigue uno de estos dos caminos:

- (1) Se parte de un k prefijado, se lleva a cabo el AF y se contrasta la hipótesis H_0 : número de factores comunes = k .
- (2) Se estiman las communalidades y se obtienen los factores comunes .

En cuanto a prefijar un número k de factores, se pueden seguir los criterios expuestos en el Cap. 32 para determinar el número de componentes principales a retener (criterio de Kaiser, gráfico de sedimentación, porcentaje mínimo de varianza explicada, ...).

En cuanto a la estimación de las communalidades, de las múltiples posibilidades existentes, las siguientes son interesantes por su sencillez y buenos resultados:

- Una de las más sencillas, si el número de variables es grande, es aproximar la communalidad de una variable por su correlación más alta con las demás variables: $\hat{h}_j^2 = \max(r_{j1}, r_{j2}, \dots, r_{j(j-1)}, r_{j(j+1)}, \dots, r_{jp})$.
- Otra posibilidad es $\hat{h}_j^2 = \frac{r_{js}r_{jt}}{r_{ts}}$, donde Z_s y Z_t son, por este orden, las dos variables más correlacionadas con Z_j . Este procedimiento modera el efecto que tendría en el anterior una correlación excepcionalmente elevada.
- En la misma línea, otra posibilidad es el promedio de los coeficientes de correlación entre la variable en cuestión y las restantes: $\hat{h}_j^2 = \frac{\sum_{j \neq s} r_{js}}{p-1}$.
- Otra alternativa es realizar un ACP y tomar como communalidad de cada variable la varianza explicada por los factores retenidos con el criterio de autovalor mayor que la unidad.
- También se puede utilizar el coeficiente de determinación lineal múltiple de cada variable con las demás como estimación de la cota inferior de sus communalidades: $\hat{h}_j^2 \geq r_{Z_j;(Z_{j1}, \dots, Z_{j-1}, Z_{j+1}, \dots, Z_p)}^2 = 1 - \frac{1}{r^{jj}}$, donde r^{jj} es el j -ésimo elemento de la diagonal de \mathbf{R}^{-1} .

Un valor alto de la communalidad, próximo a $V(X_j)$, significa que dicha variable está bien representada en el espacio de factores.

33.3.2. Análisis factorial

33.3.2.1. Métodos de extracción de los factores

Método de componentes principales

Su objetivo es el análisis de toda la varianza, común y no común, (modelo (33.1)). Por consiguiente, las entradas de la diagonal de \mathbf{R} ⁹ son unitarias y no se requiere la estimación a priori de las communalidades ; tampoco se requiere la estimación a priori del numero de factores comunes, que se determinan a posteriori. Para la exposición del método, así como para su ejemplificación con la base de datos TIC2021 del paquete CDR, se remite al lector al Cap. 32. Aunque en el Cap. 32 se utilizó la función PCA de la librería FactoMineR, también se puede utilizar la función `principal` de la librería psych.

Este método tiene la ventaja de que siempre proporciona una solución. Sin embargo, al no estar basado en el modelo (33.3), puede dar estimaciones de las cargas factoriales muy sesgadas, sobre todo cuando hay variables con communalidades pequeñas.

Método de los factores principales

Es la aplicación del método de componentes principales a la *matriz de correlaciones reducida*, \mathbf{R}^* , es decir, con communalidades en la diagonal en vez de unos. Exige, por tanto, la estimación previa de las communalidades y su objetivo es el análisis de la varianza compartida por todas las variables (modelo (33.3)). Se trata de un procedimiento iterativo que consta de las siguientes etapas:

1.- Cálculo de la matriz de correlaciones muestrales.

2.- Estimación inicial de las communalidades utilizando el coeficiente de determinación lineal múltiple de cada variable con las demás.¹⁰

3.- Cálculo de la matriz de correlaciones reducida:

$$\mathbf{R}^* = \begin{pmatrix} \hat{h}_{1(0)}^2 & r_{12} & \cdots & r_{1p} \\ r_{21} & \hat{h}_{2(0)}^2 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & \hat{h}_{p(0)}^2 \end{pmatrix}.$$

4.- Cálculo de los autovalores y autovectores asociados a \mathbf{R}^* (matriz no necesariamente semidefinida positiva) y, a partir de ellos, obtención de las estimaciones de la matriz de cargas factoriales $\mathbf{A}_{(0)}$. En este paso hay que determinar el número de factores utilizando los criterios del ACP.

⁹Cuando se utiliza este método, se debe utilizar \mathbf{R} , porque, de otra forma, los factores comunes no tendrían media cero y varianza unitaria.

¹⁰Una estimación inicial de las communalidades equivale a una estimación inicial de las varianzas únicas: $\hat{d}_{j(0)}^2 = \hat{\sigma}_j^2 - \hat{h}_{j(0)}^2$; y si las variables originales están tipificadas: $\hat{d}_{j(0)}^2 = 1 - \hat{h}_{j(0)}^2$.

5.- A partir de la estimación de $\mathbf{A}_{(0)}$, obtención de una nueva estimación de las comunalidades: $\hat{h}_{j(1)}^2 = \hat{a}_{j1(1)}^2 + \hat{a}_{j2(1)}^2 + \cdots + \hat{a}_{jk(1)}^2$ y, por tanto, de una nueva estimación de la varianza única (o unicidad) $\hat{d}_{j(1)}^2 = 1 - \hat{h}_{j(1)}^2$.

6.- Comparación de $\hat{h}_{j(1)}^2$ con $\hat{h}_{j(0)}^2$, $j = 1, 2, \dots, p$. Si hay diferencia significativa se vuelve al paso 3, y si la discrepancia no supera una cantidad prefijada se aceptan como válidas las últimas estimaciones disponibles.

En el software **R**, el método de los factores principales se implementa con la función **fa** de la librería **psych**, que parte de **R*** (véase Harman (1976) para el procedimiento iterativo y Revelle (2022) para los detalles sobre la manera cómo **fa** parte de **R*** y lleva a cabo la extracción de los factores).

```
af_facprin <- fa(cor(TIC2021), nfactors=2, rotate="none", fm='pa')
```

```
print(af_facprin, digits=3)
#> Factor Analysis using method = pa
#> Call: fa(r = cor(TIC2021), nfactors = 2, rotate = "none", fm = "pa")
#> Standardized loadings (pattern matrix) based upon correlation matrix
#>          PA1      PA2    h2   u2   com
#> ebroad    0.678   0.189 0.495 0.5050 1.16
#> esales    0.503   0.547 0.553 0.4474 1.99
#> esocmedia 0.796   0.212 0.678 0.3218 1.14
#> eweb      0.872   0.239 0.818 0.1822 1.15
#> hbroad    0.816   -0.452 0.869 0.1306 1.56
#> hiacc     0.888   -0.439 0.982 0.0181 1.46
#> iuse      0.935   -0.023 0.875 0.1248 1.00
#>
#>          PA1      PA2
#> SS loadings  4.435 0.835
#> Proportion Var 0.634 0.119
#> Cumulative Var 0.634 0.753
```

En la salida anterior, *SS loadings* son los autovalores de **R***, que coinciden con la suma de los cuadrados de las cargas de las variables en cada factor (suma de las cargas al cuadrado por columnas). *h2* son las comunalidades (suma de las cargas al cuadrado por filas; sólo se muestran las cargas para los dos primeros factores puesto que entre ambos ya acumulan una varianza explicada de más del 75%). *u2* son las varianzas de los factores únicos; finalmente, *com_j* (en la salida *com*) es el número de factores comunes necesarios para explicar la variable Z_j : $com_j = \frac{\left(\sum_{j=1}^p a_{jm}^2\right)^2}{\sum_{j=1}^p a_{jm}^4}$. Cuanto mayor es *com_j* mejor es la calidad de la variable para participar en la extracción factorial. El promedio de los *com_j* se denomina *índice de complejidad de Hofmann*. Una solución de estructura simple perfecta tiene una complejidad de uno (cada variable carga solo en un factor); una solución con elementos distribuidos uniformemente tiene una complejidad mayor que 1. Interesa que la estructura no sea simple y perfecta porque entonces no tendría sentido la reducción dimensional. Por tanto, el índice de Hofmann deberá ser superior a la unidad.

33.3. El análisis factorial en la práctica

557

Las communalidades y unicidades son:

```
round(af_facprin2$communality,3)
#> ebroad esales esocmedia eweb hbroad hiacc iuse
#> 0.495 0.553 0.678 0.818 0.869 0.982 0.875
round(af_facprin2$uniquenesses,3)
#> ebroad esales esocmedia eweb hbroad hiacc iuse
#> 0.505 0.447 0.322 0.182 0.131 0.018 0.125
```

****Ojo, *af_facprin2* se crea más abajo, este objeto es *af_facprin* o que ha pasado aquí****

Nótese que con el método de los factores principales, al aplicar ACP sobre \mathbf{R}^* , los factores obtenidos están incorrelados y la estructura coincide con el patrón.

Los resultados son, en signo, aunque no tanto en valor, similares a los obtenidos por el método de componentes principales. Además, como era de esperar, no permiten una interpretación clara de los factores comunes. Para facilitar dicha interpretación, dichos factores deberán ser rotados (véase Sec. 33.3.2.2).

Método de máxima verosimilitud

Exige normalidad multivariante y la determinación a priori del número de factores comunes, pero no la estimación de las communalidades. Obedece al modelo (33.3) y consiste en obtener las estimaciones máximo verosímiles de \mathbf{A} y \mathbf{D} . Dado que cualquier transformación ortogonal de \mathbf{A} puede ser una solución, se impone la condición de que $\mathbf{A}'(\mathbf{DD}')^{-1}\mathbf{A}$ sea diagonal. La log-verosimilitud viene dada por $l = -\frac{N}{2}(\log|2\pi\Sigma| + \text{tr}\Sigma^{-1}\mathbf{S})$, donde $\Sigma = \mathbf{AA}' + \mathbf{DD}'$ y \mathbf{S} son las matrices de covarianzas poblacional y muestral, respectivamente, de las variables X_j , $j = 1, 2, \dots, p$. \mathbf{DD}' es la matriz de covarianzas (diagonal) de los factores únicos, donde los elementos de la diagonal representan la parte de la varianza única de cada variable, y en la literatura sobre AF es conocida como Ψ (por ello, $d_j^2 = \psi_{jj}$).

La decisión sobre el número de factores comunes, k , que en este método debe hacerse al principio, es muy importante, pues dos soluciones, una con k factores y otra con $k + 1$, pueden tener matrices de cargas factoriales muy diferentes, al contrario que en el método de componentes principales, donde los primeros k componentes son siempre iguales. Pues bien, una ventaja del método de máxima verosimilitud es que lleva asociado un test estadístico secuencial para determinar el número de factores (véase Sec. 33.3.3). Otra ventaja es que las estimaciones máximo verosímiles son invariantes ante cambios de escala; en consecuencia, las matrices de covarianzas teórica y muestral de la log-verosimilitud pueden ser sustituidas por sus homónimas de correlación sin variación alguna en los resultados. Una desventaja es que puede haber un problema de grados de libertad; o, en otros términos, el número de factores k debe ser compatible con un número de grados de libertad positivo.

El método máximo verosímil se puede implementar en **R** con la librería **pysch** y la función **fa** (con **fm=m1**). Otra posibilidad es utilizar la función **factanal**. En ambos casos hay comprobar el cumplimiento de la hipótesis de normalidad. En el ejemplo TIC no procede su implementación al no cumplirse tal hipótesis.

Otros métodos

Razones de espacio impiden comentar otros procedimientos de extracción de los factores. No obstante, hay que señalar que la función `fa` de la librería `psych` también permite implementar los métodos (*i*) minres (mínimo residuo), que estima las cargas factoriales minimizando (sin ponderaciones) los cuadrados de los residuos no diagonales de \mathbf{R}^{res} ; parte de una estimación de k y, como el método máximo verosímil, no precisa estimar las comunidades, que se obtienen como subproducto tras la estimación de las cargas; y (*ii*) alpha, que maximiza el alpha de Cronbach para los factores. Aunque `fa` también proporciona el método del centroide o la descomposición triangular (que exigen la estimación de las comunidades, o el análisis imagen (que requiere el número de factores), en la actualidad están en desuso).

Otros métodos de extracción de los factores son los métodos de mínimos cuadrados no ponderados y mínimos cuadrados generalizados, que minimizan la suma de las diferencias cuadráticas entre las matrices de correlación observada y reproducida, en el último caso ponderando los coeficientes de correlación inversamente a la unicidad de las variables (alta unicidad supone baja communalidad). Ambos son proporcionados por `fa` y `FAiR`, que también es una librería muy recomendable.

33.3.2.2. Rotaciones en el modelo de análisis factorial

La interpretación de los factores se lleva a cabo a través de la estructura factorial , que, si los factores comunes están incorrelados, coincide con el patrón factorial. Sin embargo, aunque el modelo obtenido sea representativo de la realidad, en ocasiones la interpretación de los factores es harto difícil, porque presentan correlaciones similares con un gran número de variables. Como la solución AF no es única (si \mathbf{A} es una solución factorial, también lo es cualquier transformación ortogonal), con la rotación se trata de que cada variable tenga una correlación próxima a 1 con un factor y a 0 con el resto, facilitando la interpretación de los factores.

Geométricamente, la j -ésima fila de la matriz de cargas contiene las coordenadas de un punto (elemento, observación) en el espacio de las cargas correspondientes a X_j . Al realizar la rotación, se obtienen las coordenadas respecto a unos nuevos ejes, siendo el objetivo situarlos lo más cerca posible del mayor número de puntos. Esto asociaría cada grupo de variables con un sólo factor, haciendo la interpretación más objetiva y sencilla.

Sea \mathbf{T} una matriz ortogonal ($\mathbf{T}'\mathbf{T} = \mathbf{T}\mathbf{T}' = \mathbf{I}$), denominada *matriz de transformación*. Entonces, el modelo (33.3) puede escribirse como $\mathbf{Z} = \mathbf{ATT}'\mathbf{F} + \mathbf{U} = \mathbf{BT}'\mathbf{F} + \mathbf{U}$. Se trata de llegar a una *estructura simple*, que se caracteriza porque en \mathbf{B} :

- Cada fila tiene al menos un cero.
- Cada columna tiene, al menos, tantos ceros como factores comunes (k).
- Cada par de columnas debe ser tal que, para varias variables, una tenga cargas despreciables y la otra no.
- Si $k \geq 4$, cada par de columnas debe tener un número elevado de variables cuyas cargas sean nulas en ambas variables.
- Para cada par de columnas, el número de variables con cargas no nulas en ambas columnas debe ser muy pequeño.

Como se avanzó, se trata de que las variables se aglomeren lo más posible en torno a los factores comunes, y de la manera más discriminatoria posible. Así mejora la interpretación de éstos y, por lo general, aumenta su significado teórico.

Las rotaciones pueden ser ortogonales u oblicuas, dependiendo de si los nuevos factores siguen estando incorrelacionados (ejes perpendiculares) o no (ejes oblicuos).

33.3.2.2.1. Rotaciones ortogonales

Preservan la perpendicularidad de los ejes y no varían las comunidades, pues $\mathbf{B}\mathbf{B}' = \mathbf{A}\mathbf{T}'\mathbf{A}' = \mathbf{A}\mathbf{A}'$. Tampoco modifican los cuadrados de las comunidades ni, por tanto, la suma de sus cuadrados (para todas las variables): $\sum_{j=1}^p \sum_{m=1}^k b_{jm}^4 + 2 \sum_{m < r=1}^k \sum_{m=1}^k b_{jm}^2 b_{jr}^2$. Y como esta expresión se mantiene invariante, minimizar el segundo término implica maximizar el primero.

Las rotaciones ortogonales más usadas son:

Rotación VARIMAX

Se define *simplicidad* del factor m -ésimo como la varianza de los cuadrados de las cargas factoriales (rotadas) b_{ji} , $j = 1, 2, \dots, p$: $SMPL_m = \frac{\sum_{j=1}^p b_{jm}^4}{p} - \left(\frac{\sum_{j=1}^p b_{jm}^2}{p} \right)^2$. Cuanto mayor es la simplicidad de los factores, más sencilla es su interpretación. Por ello, el objetivo es que \mathbf{T} sea tal que se maximice la varianza del cuadrado de las cargas en cada columna del patrón factorial, es decir, en cada factor.

Dicho lo anterior, la rotación VARIMAX consiste en la obtención de una \mathbf{T} que maximice la suma de las simplicidades de todos los factores, $V = \sum_{m=1}^k SMPL_m$.¹¹

Sin embargo, las variables con mayor communalidad, y por tanto con mayores cargas factoriales, tendrán mayor influencia en la solución final, porque la communalidad no se ve afectada por la rotación ortogonal. Para evitar esto, Kaiser propuso la rotación VARIMAX normalizada¹², donde las cargas se dividen entre la raíz cuadrada de la communalidad de la variable correspondiente. Los valores obtenidos son los elementos de \mathbf{B} .

El procedimiento de cálculo de las cargas de los factores rotados es iterativo, rotándose los factores por parejas hasta que se consigue maximizar la suma de simplicidades normalizadas.

La rotación VARIMAX es muy popular por la robustez de sus resultados, si bien se recomienda para un número no muy elevado de factores comunes .

Rotación QUARTIMAX

Su objetivo es maximizar la varianza de los cuadrados de todas las cargas factoriales, es decir, maximizar $Q = \frac{\sum_{j=1}^p \sum_{m=1}^k b_{jm}^4}{pk} - \left(\frac{\sum_{j=1}^p \sum_{m=1}^k b_{jm}^2}{pk} \right)^2$.

¹¹Este criterio es compatible con el hecho de que, en cada fila, uno de los elementos esté próximo a cero y los demás a uno, porque la suma de los cuadrados de los elementos de una fila, es la communalidad fija de la variable correspondiente.

¹²La normalización Kaiser se aplica también en los demás tipos de rotación.

Nótese que, como la rotación ortogonal no modifica las comunidades, $h_j^2 = \sum_{m=1}^k b_{jm}^2$, el segundo término de la expresión anterior no se verá modificado, por lo que el criterio anterior equivale a maximizar $\frac{\sum_{j=1}^p \sum_{m=1}^k b_{jm}^4}{pk}$.

QUARTIMAX es recomendable cuando el número de factores es elevado. Tiende a generar un factor general, el primero, sobre el que la mayor parte de las variables tienen cargas elevadas, lo cual contradice los objetivos que persigue la rotación.

Rotación ORTOMAX

Es una clase general de los métodos de rotación ortogonal que se construye como una composición ponderada de las dos rotaciones anteriores: $\alpha Q + \beta V$, donde V se multiplica por p por conveniencia, ya que una constante multiplicativa no afecta a la solución. Por tanto, su objetivo es maximizar la expresión: $ORT = \sum_{m=1}^k \left(\sum_{j=1}^p b_{ji}^4 - \left(\frac{\theta}{p} \sum_{j=1}^p b_{ji}^2 \right)^2 \right)$, $0 < \theta = \frac{\alpha}{\alpha+\beta} < 1$.

Si $\theta = 1$, se tiene el criterio VARIMAX; si $\theta = 0$, se tiene el criterio QUARTIMAX; si $\theta = 0,5$, se tiene un criterio igualmente ponderado denominado BIQUARTIMAX; y si $\theta = \frac{k}{2}$, se tiene el criterio EQUAMAX, recomendado por parte de la literatura.

Nótese que QUARTIMAX pone el énfasis en la simplificación de la descripción por filas (variables) de la matriz factorial, mientras que VARIMAX lo pone en la simplificación por columnas (factores), para satisfacer los requisitos de *estructura simple*; así, aunque se pueda conseguir la simplicidad de cada variable y que, a la vez, las cargas respecto del mismo factor sean grandes, tal factor queda excluido por la restricción impuesta por la simplificación sobre cada factor (Harman, 1976).

33.3.2.2. Rotaciones oblicuas Superan la incorrelación u ortogonalidad de los factores y se suelen aplicar cuando: (i) se sospecha que, en la población, los factores tienen una fuerte correlación y/o (ii) cierta correlación entre los factores permite una gran ganancia en la interpretación de los mismos. Podrían aplicarse siempre, como norma general, puesto que, en realidad, la ortogonalidad es un caso particular de la oblicuidad.

Los procedimientos que proporcionan soluciones con estructura simple oblicua emanan de los mismos criterios objetivos que los que proporcionan soluciones con estructura simple ortogonal. De hecho, si se relajan las condiciones de ortogonalidad, algunos procedimientos de rotación ortogonal pueden adaptarse al caso oblicuo (tal es el caso, por ejemplo, del método OBLIMAX, a partir del criterio QUARTIMAX). Por otra parte, los métodos de rotación oblicua no solo son directos, sino que también pueden introducir los principios de estructura simple que se requieren para la solución factorial primaria de forma indirecta (métodos indirectos). Las rotaciones oblicuas exigen nuevos conceptos y nueva nomenclatura:

- *Factores de referencia*, G_m , $m = 1, 2, \dots, k$: para cada factor rotado se puede encontrar un nuevo factor incorrelado con los rotados. A esos nuevos factores les llama factores de referencia. En caso de rotación ortogonal, los factores de referencia coinciden con los primeros.
- *Estructura factorial de referencia*: hasta ahora, se denominaba estructura factorial a la matriz de correlaciones entre las variables Z_j , $j = 1, 2, \dots, p$ y los factores rotados, que

33.3. El análisis factorial en la práctica

561

en el caso ortogonal coincide con la matriz de cargas factoriales rotadas. Pues bien, se denomina estructura factorial de referencia a la matriz de correlaciones entre las variables Z_j y los factores de referencia. Si la rotación es ortogonal, coincide con la estructura factorial.

- *Matriz de transformación*: en el caso oblicuo se representa por Λ .
- *Estructura factorial oblicua*: \mathbf{V} , tal que $\mathbf{V} = \mathbf{A}\Lambda$; sus elementos son v_{jm} .
- *Cargas*: en el caso oblicuo el término “carga” se utiliza para denotar la correlación de la variable con el eje de referencia: $v_{jm} = r_{Z_j; \Lambda_m}$.

Mientras las rotaciones ortogonales intentan encontrar la estructura factorial más simple, las oblicuas hacen lo mismo pero con la estructura de referencia.

El método (directo) OBLIMAX maximiza la expresión $K = \frac{\sum_{j=1}^p \sum_{m=1}^k v_{jm}^4}{(\sum_{j=1}^p \sum_{m=1}^k v_{jm}^2)^2}$. Nótese que se trata del criterio QUARTIMAX ortogonal, pero incorporando el denominador, puesto que en la rotación oblicua ya no es constante.

El QUARTIMIN directo, también derivado del QUARTIMAX ortogonal, minimiza el criterio $N = \sum_{j=1}^p \sum_{m \leq q=1}^k v_{jm}^2 v_{jq}^2$, y recibe este nombre por minimizar términos de cuarto grado.

La generalización del criterio “minimizar $H = \sum_{j=1}^p \sum_{m < q=1}^k b_{jm}^2 b_{jq}^2$ ” para factores oblicuos se denomina OBLIMIN, y da lugar a métodos indirectos. Entre ellos, destaca el COVARIMIN, que se obtiene relajando la condición de ortogonalidad en el VARIMAX, minimizando las covarianzas de los cuadrados de los elementos de \mathbf{V} : $C^* = \sum_{m \leq q=1}^k \left(p \sum_{j=1}^p v_{jm}^2 v_{jq}^2 - \sum_{j=1}^p v_{jm}^2 \sum_{j=1}^p v_{jq}^2 \right)$. La versión COVARIMIN normalizada minimiza $C = \sum_{m \leq q=1}^k \left(p \sum_{j=1}^p \frac{v_{jm}^2}{h_j^2} \frac{v_{jq}^2}{h_j^2} - \sum_{j=1}^p \frac{v_{jm}^2}{h_j^2} \sum_{j=1}^p \frac{v_{jq}^2}{h_j^2} \right)$.

Se ha comprobado empíricamente que QUARTIMIN tiende a ser demasiado oblicuo y COVARIMIN demasiado ortogonal. Una solución intermedia es la rotación BIQUARTIMIN, que consiste en minimizar $B^* = H + \frac{C^*}{p}$, donde $\frac{C^*}{p}$ es la expresión completa del COVARIMIN. Una generalización de la rotación BIQUARTIMIN es $B^* = \alpha H + \beta \frac{C^*}{p}$. Sencillas operaciones aritméticas llevan a $B^* = \sum_{m < q=1}^k \left(p \sum_{j=1}^p v_{jm}^2 v_{jq}^2 - \gamma \sum_{j=1}^p v_{jm}^2 \sum_{j=1}^p v_{jq}^2 \right)$, con $\gamma = \frac{\beta}{\alpha + \beta}$. La rotación QUARTIMIN se obtiene con $\gamma = 0$, la BIQUARTIMIN con $\gamma = 0,5$ y la COVARIMIN con $\gamma = 1$. También se pueden obtener versiones normalizadas sin más que normalizar las cargas (dividirlas por h_{jm}^2).

El criterio BINORMALMIN (normalizado) es una alternativa al BIQUARTIMIN para corregir el sesgo de oblicuidad del criterio COVARIMIN. Minimiza $D = \sum_{m < q=1}^k \left(\frac{\sum_{j=1}^p \frac{v_{jm}^2}{h_j^2} \frac{v_{jq}^2}{h_j^2}}{\sum_{j=1}^p \frac{v_{jm}^2}{h_j^2} \sum_{j=1}^p \frac{v_{jq}^2}{h_j^2}} \right)$.

BINORMALMIN suele ser mejor con datos muy simples o muy complejos; BIQUARTIMIN es más recomendable con datos moderadamente complejos.

El método de rotación OBLIMIN directo, en vez de proceder como B^* , que depende de los valores de la estructura, minimiza directamente una función de la matriz del patrón factorial

primario: $F(\mathbf{A}) = \sum_{m < q=1}^k \left(\sum_{j=1}^p a_{jm}^2 a_{jq}^2 - \frac{\delta}{p} \sum_{j=1}^p a_{jm}^2 \sum_{j=1}^p a_{jq}^2 \right)$. Cuando $\delta = 0$, se tiene el QUARTIMIN directo.

Hay otros tipos de transformaciones oblicuas, pero únicamente se mencionarán (i) la ORTO-BLICUA, que llega a la solución oblicua mediante una serie de transformaciones ortogonales intermedias; y (ii) el la PROMAX, muy popular, que actúa alterando los resultados de una rotación ortogonal (concretamente elevando las cargas de la rotación ortogonal a una potencia entre 2 y 4) hasta crear una solución con cargas factoriales lo más próximas a la estructura ideal. Cuanto mayor es esta potencia más oblicua es la solución obtenida.

33.3.2.2.3. ¿Rotaciones ortogonales u oblicuas?

La selección del método de rotación, ortogonal u oblicua, depende del objetivo perseguido. Si se pretende reducir el número de variables originales a un conjunto mucho menor de variables incorrelacionadas para su uso posterior en otra técnica, por ejemplo regresión, la rotación debe ser ortogonal. Si el objetivo es obtener unos factores teóricos significativos, puede resultar apropiada la aplicación de una rotación oblicua.

En R es muy sencillo implementar una rotación ortogonal u oblicua. Basta, por ejemplo, con utilizar la librería GPArotation (Bernaards and Jennrich, 2005) e indicarlo en el argumento `rotate` de la función `fa`. A modo de ejemplo, extrayendo los factores por el método de los factores principales y utilizando una rotación VARIMAX normalizada, sería:

****AQUÍ SE CREA `af_facprin2`, ARRIBA YA SE LLAMABA****

```
library("GPArotation")
af_facprin2 <- fa(cor(TIC2021), nfactors=2, rotate="varimax", fm="pa", digits=3)

af_facprin2 # el objeto contiene información adicional no relevante en estos momentos
#> Factor Analysis using method = pa
#> Standardized loadings (pattern matrix) based upon correlation matrix
#>          PA1 PA2   h2   u2 com
#> ebroad    0.38 0.59 0.50 0.505 1.7
#> esales    0.02 0.74 0.55 0.447 1.0
#> esocmedia 0.46 0.68 0.68 0.322 1.7
#> eweb      0.50 0.75 0.82 0.182 1.7
#> hbroad    0.91 0.20 0.87 0.131 1.1
#> hiacc     0.96 0.26 0.98 0.018 1.1
#> iuse      0.72 0.60 0.88 0.125 1.9
#>
#>          PA1  PA2
#> SS loadings  2.87 2.40
#> Proportion Var 0.41 0.34
#> Cumulative Var 0.41 0.75
```

Nótese que la salida por defecto es la normalizada. También se puede utilizar la librería `stats` indicando T o F en el argumento `normalize`, dependiendo de que se quiera o no, respectivamente, una rotación VARIMAX (u otra) normalizada .

```
library(stats)
varimax(loadings(af_facprin),normalize=T)
```

En el ejemplo del uso las TIC en los países de la UE-27, la rotación VARIMAX ha conseguido facilitar la interpretación de los factores comunes , ya que, tras la rotación, las variables relacionadas con el uso de las TIC a escala individual y de hogar cargan en el primer factor, mientras que las relacionadas con el uso de las TIC a nivel empresarial cargan en el segundo. Por tanto, ambos factores pueden considerarse indicadores de la dotación y uso de las TIC en los ámbitos familiar y empresarial, respectivamente. El lector puede probar (y comparar) con otras rotaciones sin más que incluirlas en el argumento `rotate`.

33.3.3. Post-análisis factorial

Realizado el AF, los siguientes procedimientos permiten comprobar la bondad del modelo obtenido:

Análisis de las correlaciones residuales

Se entiende por bondad de la solución factorial la medida del grado en que los factores del modelo explican las correlaciones entre las variables. Por ello, parece natural que tal medida se base en la comparación entre las correlaciones observadas y las que se derivan del modelo factorial (reproducidas) o, en términos matriciales, en la magnitud de las entradas de la matriz de correlaciones residuales $\mathbf{R}^{res} = \mathbf{R} - \mathbf{R}^{rep}$, donde $\mathbf{R} = \frac{1}{N}\mathbf{Z}\mathbf{Z}'$ y $\mathbf{R}^{rep} = \mathbf{A}\Phi\mathbf{A}' = \boldsymbol{\Gamma}\mathbf{A}'$ (relación fundamental entre el patrón y la estructura factorial; en caso de incorrelación entre los factores, $\Phi = \mathbf{I}$ y $\mathbf{R}^{rep} = \mathbf{A}\mathbf{A}'$). La matriz \mathbf{R}^{rep} se obtiene sin más que sustituir \mathbf{Z} por \mathbf{AF} en la expresión de \mathbf{R} .

Ahora bien, ¿cuál es el criterio apropiado para concluir si una solución factorial es aceptable o no? Para que sea aceptable, los elementos (los residuos) de \mathbf{R}^{res} deben ser cercanos a cero, y como todos los factores comunes han sido considerados, se supone que no existen más vínculos entre las variables y que la distribución de dichos residuos debe ser como la de correlación cero en una muestra del mismo tamaño. Por tanto, como $\sigma_{r=0} = \frac{1}{\sqrt{N-1}}$, entonces $S_{r_{res}} \leq \frac{1}{\sqrt{N-1}}$:¹³

- Si $S_{r_{res}} \gg \frac{1}{\sqrt{N-1}}$, es razonable pensar que existen relaciones adicionales significativas entre las variables y hay que modificar la solución factorial.
- Si $S_{r_{res}} \ll \frac{1}{\sqrt{N-1}}$, es razonable pensar que la solución factorial incluye relaciones que no están justificadas.
- Si $S_{r_{res}} \leq \text{pero no } \ll \frac{1}{\sqrt{N-1}}$, la solución es aceptable.

Otra posibilidad, también muy sencilla, propuesta por [Revelle \(2022\)](#), es utilizar $fit = 1 - \frac{\sum(\mathbf{R}-\mathbf{FF}')^2}{\sum(\mathbf{R})^2}$, que indica la reducción proporcional en la matriz de correlación debida al modelo factorial. Nótese que esta medida es sensible al tamaño de las correlaciones originales. Es decir,

¹³Este criterio tiene como ventaja la simplicidad. Sin embargo, sería conveniente que tuviese en cuenta, al menos, el número de variables.

si los residuos son pequeños, pero las correlaciones son pequeñas, el ajuste es malo. Las medidas clásicas como el RMSE (raíz cuadrada del error cuadrático medio), o similares, también son susceptibles de uso.

En el ejemplo TIC seguido en este capítulo el ajuste realizado es muy bueno:

```
round(af_facprin2$residual, 3)
#>      ebroad esales esocmedia  eweb hbroad  hiacc  iuse
#> ebroad    0.505 -0.068     0.008  0.068 -0.045  0.002  0.003
#> esales   -0.068  0.447     0.026  0.015  0.004 -0.012  0.021
#> esocmedia  0.008  0.026     0.322 -0.047  0.014 -0.005  0.017
#> eweb      0.068  0.015    -0.047  0.182  0.012  0.015 -0.042
#> hbroad   -0.045  0.004     0.014  0.012  0.131 -0.005  0.010
#> hiacc     0.002 -0.012    -0.005  0.015 -0.005  0.018  0.002
#> iuse      0.003  0.021     0.017 -0.042  0.010  0.002  0.125
af_facprin2$rms
#> [1] 0.02907475
af_facprin2$fit
#> [1] 0.9715865
```

NOTA IMPORTANTE

Como se avanzó en la introducción, el AF está enfocado al ajuste de las correlaciones entre las variables observadas mediante el patrón factorial correspondiente al modelo (33.2) (con los factores comunes y el factor único). Pues bien, si en el proceso reproductivo se utiliza el modelo sólo con los factores comunes, la matriz de correlaciones que se reproduce es \mathbf{R} , lo que implica el modelo ACP (modelo (33.1)). Si se incluye también el factor específico, la matriz de correlaciones que se reproduce es \mathbf{R}^* (modelo AF). Si en dicha reproducción se utilizasen los factores comunes y el término de error, se reproduciría \mathbf{R} con una diagonal principal cuyas entradas serían la unidad menos las estimaciones de las communalidades.

Test de bondad de ajuste

Se trata de un contraste de razón de verosimilitudes que se puede llevar a cabo cuando se extraigan los factores por el método de máxima verosimilitud. La hipótesis nula es la suficiencia de k factores comunes para explicación de las correlaciones entre las variables originales y de la varianza que comparten.

El estadístico del contraste es $-2\ln\lambda = np(\hat{a} - \ln\hat{g} - 1]$, donde \hat{a} y \hat{g} son las medias aritmética y geométrica, respectivamente, de los autovalores de la matriz $\hat{\Sigma}_{H_0}^{-1}\mathbf{S}$. Bajo H_0 , se distribuye asintóticamente como una χ_{df}^2 , con $df = \left(p + \frac{p(p+1)}{2}\right) - \left(p + pk + p - \frac{k(k-1)}{2}\right) = \frac{1}{2}(p-k)^2 - \frac{1}{2}(p+k)$.¹⁴

Este test se aplica de manera secuencial: se formula como hipótesis nula $k = 0$. Si no se rechaza, no hay factores comunes subyacentes. Si se rechaza, se sigue con $k = 1$. Si no se rechaza $k = 1$,

¹⁴ df indica la medida en que el modelo factorial ofrece una interpretación más simple que Σ .

se concluye que el modelo con un factor es una adecuada representación de la realidad; si se rechaza, se formula la hipótesis nula de que $k = 2$, y el proceso continúa hasta que no se rechace la hipótesis nula, siempre que el valor de k sea compatible con un número de grados de libertad positivo.

33.3.4. Puntuaciones factoriales

Las puntuaciones factoriales son las estimaciones de los valores de los factores aleatorios no observados, es decir, de los elementos de $\mathbf{F}_{m \times m}$. Así, \hat{f}_{im} será la estimación del valor del m -ésimo factor para la i -ésima observación (elemento, individuo, objeto...). Cuando se extraen los factores por componentes principales las puntuaciones son exactas.

Estas estimaciones pueden ser usadas como inputs para posteriores análisis (regresión, cluster, etc.) en los que se trabaje con los mismos elementos o individuos, sustituyendo las variables originales por los nuevos factores obtenidos. La cuestión es: ¿cómo calcular estas puntuaciones?, porque tanto los factores como los errores no son observables sino aleatorios.

Los métodos más populares para obtener la estimación de las puntuaciones factoriales son:

- El de regresión por mínimos cuadrados ordinarios (MCO), donde $\hat{\mathbf{F}} = (\mathbf{A}'\mathbf{A})^{-1} \mathbf{A}'\mathbf{Z}$.
- El de Bartlett, basado en el método de estimación por mínimos cuadrados generalizados (MCG), con $\hat{\mathbf{F}} = (\mathbf{A}'\Psi^{-1}\mathbf{A})^{-1} \mathbf{A}'\Psi^{-1}\mathbf{Z}$. El mismo estimador se puede obtener por máxima verosimilitud asumiendo normalidad multivariante.
- El de Thompson (con un enfoque bayesiano), donde $\hat{\mathbf{F}} = (\mathbf{I} + \mathbf{A}'\Psi^{-1}\mathbf{A})^{-1} \mathbf{A}'\Psi^{-1}\mathbf{Z}$.
- El de Anderson-Rubin (que obtiene estimaciones MCG imponiendo la condición $\mathbf{F}'\mathbf{F} = \mathbf{I}$ ($\hat{\mathbf{F}} = (\mathbf{A}'\Psi^{-1}\mathbf{R}\Psi^{-1}\mathbf{A})^{-1} \mathbf{A}'\Psi^{-1}\mathbf{Z}$)).

Las ventajas y desventajas de cada uno de ellos pueden verse en [Mardia et al. \(1979a\)](#) y [De la Fuente \(2011\)](#).

En el ejemplo de las TIC, las puntuaciones de los dos factores extraídos con el método de los factores principales y rotados con VARIMAX (la rotación no afecta a las puntuaciones), calculadas por el método de regresión, para los países de la UE-27 (se muestran los de Bélgica, Bulgaria y la República Checa), se obtienen en **R** como sigue:

```
af_facprin3 <- fa(cor(TIC2021), nfactors=2, rotate="VARIMAX", fm="pa",
                     scores="regression")
factor.scores(TIC2021, af_facprin3)$scores[1:3,]
#>          PA1        PA2
#> BE  0.6256359  1.01289866
#> BG -2.1820404 -0.03439974
#> CZ -0.2189723  1.08635525
```

33.4. Relaciones y diferencias entre el AF y el ACP

ACF y AF son aparentemente muy similares, pero en realidad son muy diferentes. Tanto ACP como AF son técnicas de reducción de la dimensionalidad que aparecen juntas en los paquetes estadísticos y persiguen objetivos muy similares, lo cual, en determinadas ocasiones, lleva al lector a pensar que son intercambiables entre sí, cuando ello no es cierto. Por ello, este capítulo finaliza con un breve comentario sobre las diferencias más relevantes entre ambos enfoques.

La primera es que ACP es una mera transformación de los datos en la que no se hace ningún supuesto sobre la matriz de covarianzas o de correlaciones. Sin embargo, AF asume que los datos proceden de un modelo bien definido, el modelo (33.3), en el que los factores subyacentes satisfacen unos supuestos bien definidos.

En segundo lugar, en ACP el énfasis se pone en el paso desde las variables observadas a las componentes principales, mientras que en AF se pone en el paso desde los factores latentes a las variables observadas. Es cierto que en ACP se pueden retener k componentes y a partir de ellas aproximar (reproducir) las variables observadas; sin embargo, esta manera de proceder parece menos natural que la aproximación de las variables observadas en términos de los factores comunes y, además, al no tener en cuenta la unicidad de las variables, sobreestima las cargas factoriales y la dimensionalidad del conjunto de variables originales.

Una tercera diferencia es que, mientras que ACP obtiene componentes en función de las variables originales (los valores de las variables pueden ser estimados a posteriori en función de dichas componentes o factores), en AF las variables son, ellas mismas, combinaciones lineales de factores desconocidos. Es decir, mientras que en ACP la solución viene de la mano de la descomposición en valores singulares, en AF requiere procedimientos de estimación, normalmente iterativos.

La cuarta es que ACF es un procedimiento cerrado mientras que AF es abierto, en el sentido de que explica la varianza común y no toda la varianza.

Finalmente, como pudo verse en 33.3.2.1, cuando las varianzas de los factores únicos son prácticamente nulas, el método de los factores principales es equivalente a ACP, y cuando son pequeñas ambos dan resultados similares. Sin embargo, cuando son grandes, en ACP las componentes principales (tanto las retenidas como las que no se retienen) las absorben, mientras que el AF las considera y les da su lugar.

RESUMEN

El Análisis Factorial es una técnica de reducción de la dimensionalidad que trata de dar una explicación de la varianza compartida, o común, de las variables objeto de estudio (no de toda la varianza, como hace el análisis de componentes principales) mediante un número mucho menor de factores comunes latentes. Por consiguiente, solo tiene sentido implementarlo si dichas variables se encuentran fuertemente correlacionadas. Tras introducir al lector en los principales elementos teóricos del Análisis Factorial (el modelo básico y la solución factorial completa), se abordan las distintas etapas del procedimiento en su vertiente práctica: (i) el pre-análisis factorial, que responde a la pregunta de si procede o no llevarlo a cabo; (ii) el análisis factorial propiamente dicho, prestando especial atención a los métodos de extracción de los factores y a las rotaciones de los mismos para facilitar su interpretación; y (iii) el post-análisis factorial, que incluye una serie de procedimientos para determinar si la solución factorial obtenida es o no aceptable. Posteriormente, se aborda la cuestión de cómo estimar los valores de los factores obtenidos para cada elemento o individuo involucrado en el análisis, pues estas estimaciones pueden usarse como inputs en análisis posteriores (regresión, cluster, etc.) sustituyendo las variables originales por los factores obtenidos. El capítulo finaliza con algunos comentarios sobre las diferencias entre el análisis factorial y el de componentes principales, aparentemente muy similares, pero en realidad muy diferentes.

Capítulo 34

Escalamiento multidimensional

José Luis Alfaro Navarro^a y Manuel Vargas Vargas^a

^a Universidad de Castilla-La Mancha

34.1. Introducción

El escalado multidimensional (EMD) fue propuesto por primera vez a la Universidad de Princeton por Warren S. Torgerson a principios de la década de 1950 siendo un investigador importante en este campo Joseph Bernard Kruskal. El EMD engloba una variedad de técnicas multivariadas cuya finalidad es obtener la estructura (factores o dimensiones) de los individuos (o variables) subyacente a una matriz de datos empíricos, lo que se consigue al representar dicha estructura en una forma geométrica bi o tridimensional.

Por tanto, la idea del EMD es **representar los datos en baja dimensión** (usualmente 2 dimensiones) utilizando la información proporcionada por las distancias entre los datos. Esta técnica surge ya que cada vez con más frecuencia los datos particulares de los que se dispone y el objetivo del análisis hacen difícil su tratamiento con las medidas clásicas, por lo que se han ido diseñando nuevas medidas de distancia entre datos. Estas medidas se pueden utilizar para diferentes tareas: agrupamiento de casos, clasificación, detección de patrones o dimensiones subyacentes, recuperación de información, etc. Por lo tanto, EMD aborda algunas problemáticas que pueden ser analizadas con otras técnicas como, por ejemplo, análisis de componentes principales o factorial cuando el objetivo es representar muchas variables en pocas dimensiones mediante la identificación de la estructura interna de los datos, dimensiones o factores en base a la matriz de correlaciones como medida de proximidad entre las variables o el análisis cluster cuando el objetivo es analizar la proximidad entre los objetos, personas, productos, etc. estudiados.

El EMD analiza matrices de proximidad (similitud, disimilitud o distancia), por ello, es una alternativa más flexible que otros métodos multivariantes con los que comparte objetivos, ya que sólo requiere de una matriz con las proximidades entre los datos, que pueden representar

valoraciones personales, grado de acuerdo entre juicios, parecido entre objetos, frecuencias de aparición de rasgos, diferencias entre tratamientos, etc. La idea central es que las distancias que median entre los puntos se corresponden con las proximidades entre los objetos por medio de una función de ajuste resultante de un proceso iterativo de optimización, pudiéndose describir las relaciones entre los objetos sobre la base de las proximidades observadas ([López-González and Hidalgo-Sánchez, 2010](#))

En **R**, existen distintas funciones para desarrollar el EMD, desde las clásicas funciones `cmdscale()` del paquete **base** e `isoMDS()` del paquete **MASS** hasta el enfoque más actual, usado en este documento, recogido en el paquete **smacof** ([de Leeuw and Mair, 2009; Mair et al., 2022](#)) que proporciona al usuario una gran flexibilidad para especificar EMD. Utiliza siempre matrices de disimilaridad y, desde la primera versión, se han implementado varios enfoques adicionales de EMD y despliegue, así como varias extensiones y funciones de utilidad.

A modo de ejemplo se va a usar la información relacionada con 7 variables de la sociedad de la información disponibles para 27 países europeos en la base de datos **TIC2021**, cuatro relacionadas con el uso de las TIC por parte de las empresas y tres de aspectos relacionados con el uso por parte de las personas y la equipación de los hogares. Dicha información, así como la descripción de las variables, puede consultarse en la base de datos **TIC2021** del paquete **CDR**.

34.2. Medición de distancias y similitudes

Tanto para el EMD como para muchas otras técnicas multivariantes, el concepto de distancia, entendida como medida de diferenciación entre objetos, constituye la base fundamental de la obtención y presentación de sus resultados. También son frecuentes los conceptos de “disimilaridad”, muy parecido al de distancia, o de “similaridad”, dual en su sentido al de distancia. Se nombre como se nombre, la característica que hay que tener siempre presente es si la medida indica “alejamiento” entre los objetos (distancia o disimilaridad) o “cercanía” (similaridad o proximidad).

Básicamente, se considera una medida de distancia a una función que asigna a cada par de objetos (o_i y o_j), que pueden contener mediciones de variables x e y, un número real, $d(o_i, o_j) = \delta_{ij}$, que debe cumplir las siguientes condiciones (para un análisis más detallado véase Sec. [30.3](#)):

- a) No negatividad $\delta_{ij} \geq 0$
- b) Simetría, $\delta_{ij} = \delta_{ji}$
- c) Identificación del objeto, $\delta_{ii} = 0$

Si además es semidefinida positiva y cumple la desigualdad triangular se dice que δ es una distancia métrica.

Aunque no se va a profundizar en ello, existen diferencias matemáticas en los requisitos que debe cumplir una medida para ser considerada una distancia o una distancia métrica, así como las condiciones para ser considerada una similaridad. Básicamente, se considera una medida de similaridad a una aplicación que asigna a cada par de objetos (o_i y o_j) un número real, s_{ij} ,

34.3. *Modelo de escalamiento multidimensional*

571

que cumple las mismas condiciones que la distancia salvo la condición c para la que tiene que cumplir que $s_{ij} \leq s_{ii}$.

Esta condición es más difícil de cumplir por lo que se emplean mucho más las medidas de distancia al ser más sencillo formular la propiedad c pues simplifica mucho el poder atribuir un valor de referencia cero para definir la distancia de un individuo a sí mismo. La similitud carece de este valor de referencia, siendo posible que la similitud de un individuo a sí mismo sea diferente de unos a otros. A pesar de esta dificultad, las medidas de similitud surgen de modo natural en muchos problemas relacionados con valoraciones subjetivas de similitud.

Para un conjunto finito de objetos, la **matriz de similaridad** es:

$$\mathbf{S} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1n} \\ s_{21} & s_{22} & \cdots & s_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n1} & s_{n2} & \cdots & s_{nn} \end{pmatrix} \quad (34.1)$$

El paso de una medida de similaridad (s_{ji}) a una distancia (δ_{ij}) se puede hacer de diversas formas. Las más usuales son:

- a) $\delta_{ij} = 1 - s_{ij}$
- b) $\delta_{ij} = \sqrt{1 - s_{ij}}$
- c) Si los valores de la diagonal de \mathbf{S} no son la unidad, $\delta_{ij} = \sqrt{s_{ii} + s_{jj} - 2s_{ij}}$

En general, cuando las características que se miden sobre los objetos son variables cuantitativas p -dimensionales, las distancias más usadas son la euclídea, la city-block, la de Minkowski o la de Mahalanobis (véase Sec. 30.3).

Cuando las variables son binarias (0 y 1), los coeficientes de similaridad más utilizados son el coeficiente de Jaccard o el de Sokal-Sneath; por último, en el caso más general en el que existan variables cuantitativas, binarias y/o cualitativas, se suele utilizar la distancia de Gower (véase Sec. 30.3).

Como se aprecia, las características de los datos que se quieren analizar influyen determinantemente en qué tipo de medida de proximidad utilizar. A su vez, la elección de una medida concreta puede modificar la configuración de los datos y, consecuentemente, los resultados de los análisis que se hagan a partir de ellos.

34.3. *Modelo de escalamiento multidimensional*

El EMD parte de una matriz de proximidades entre n objetos:

$$\Delta_{n \times n} = \begin{pmatrix} \delta_{11} & \delta_{12} & \cdots & \delta_{1n} \\ \delta_{21} & \delta_{22} & \cdots & \delta_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{n1} & \delta_{n2} & \cdots & \delta_{nn} \end{pmatrix} \quad (34.2)$$

y busca una representación de los n objetos en un espacio de menor dimensión, m , donde x_{ij} es la coordenada del objeto i en la dimensión j :

$$\mathbf{X}_{n \times m} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix} \quad (34.3)$$

de forma que se puede calcular la distancia euclídea entre cada par de objetos:

$$d_{ij} = \sqrt{\sum_{t=1}^m (x_{it} - x_{jt})^2} \quad (34.4)$$

y, construir una matriz de **distancias “reproducidas”**

$$\mathbf{D}_{n \times n} = \begin{pmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{pmatrix} \quad (34.5)$$

que aproxime, en la medida de lo posible, a la matriz de proximidades, Δ .

El concepto básico del EMD es que las distancias entre los objetos en la configuración X , d_{ij} deben corresponder a las proximidades originales, δ_{ij} mediante una transformación, $d_{ij} = f(\delta_{ij})$, donde f es de algún tipo determinado.

En la práctica, no se suele encontrar un ajuste perfecto, por lo que existe un cierto grado de error. Por ello, se define el **Stress de Kruskal** como una medida de la bondad de ajuste del modelo:

$$Stress = \sqrt{\frac{\sum (f(\delta_{ij}) - d_{ij})^2}{\sum d_{ij}^2}} \quad (34.6)$$

En caso de un ajuste perfecto, el Stress sería 0, aumentando conforme más grandes sean los errores (diferencias entre las distancias “reproducidas” y las originales). Así, la solución proporcionada por el EMD será “mejor” cuanto más pequeña sea la medida del Stress. También es frecuente utilizar una variante, llamada **S-Stress**, definida como:

$$S - Stress = \sqrt{\frac{\sum (f(\delta_{ij})^2 - d_{ij}^2)^2}{\sum (d_{ij}^2)^2}} \quad (34.7)$$

Otra medida que se suele utilizar como grado de ajuste es el coeficiente de **correlación al cuadrado** (RSQ), que indica la proporción de variabilidad de los datos explicada por el modelo:

$$RSQ = \frac{[\sum (d_{ij} - d_{..})(f(\delta_{ij}) - f(\delta_{..}))]^2}{[\sum (d_{ij} - d_{..})^2][\sum (f(\delta_{ij}) - f(\delta_{..}))^2]} \quad (34.8)$$

Este valor oscila entre 0 y 1, y se interpreta de forma contraria a las medidas de Stress: mientras mayor sea el RSQ, mejor ajuste del modelo.

34.4. Tipos de escalamiento multidimensional

La elección de la función f que relaciona las proximidades originales y las distancias “reproducidas” produce dos tipos básicos de EMD, el EMD métrico (o clásico) y el EMD no métrico . En el primero, se considera que los datos están medidos en escala de intervalo o de razón y existe una relación funcional entre las distancias originales y las reproducidas; mientras que el segundo se suele aplicar cuando los datos están en escala ordinal o no se asume una relación funcional entre las distancias originales y las reproducidas, sino que sólo se conserva su ordenación.

34.4.1. Escalado multidimensional métrico

En el modelo de escalamiento métrico se asume que la relación entre las proximidades y las distancias es de tipo lineal, $d_{ij} = a + b\delta_{ij}$. De esta forma, se conserva la métrica de distancia original, entre puntos, lo mejor posible, siendo adecuado para el caso de variables cuantitativas. También se conoce como EMD clásico o como análisis de coordenadas principales.

En el ejemplo introductorio, se va a aplicar un EMD métrico con un doble objetivo: por un lado se usa la matriz de correlaciones entre la variables con el objetivo de analizar la similitud entre las mismas y, por otro lado, se determina la distancia entre observaciones con el objetivo de analizar la similitud existente entre observaciones, en este caso los países europeos.

En el primer caso, los “objetos” son las siete variables de la base de datos TIC2021 y se ha utilizado como medida de proximidad (similitud) el coeficiente de correlación entre las variables, por lo que se plantea un EMD métrico.

```
library('smacof')
library('CDR')
correlacion<- cor(TIC2021)
round(correlacion[1:3,],3)
#>           ebroad esales esocmedia eweb hbroad hiacc iuse
#> ebroad     1.000  0.377   0.587 0.704  0.422 0.521 0.632
```

```
#> esales    0.377  1.000    0.542 0.585  0.167 0.195 0.479
#> esocmedia 0.587  0.542    1.000 0.698  0.567 0.609 0.756
```

Los pasos a seguir son:

- Calcular la matriz de disimilaridades sobre la que actúa `smacof()`. En este ejemplo, se usa la matriz de correlaciones que se debe convertir en matriz de disimilaridades, mediante la función `sim2diss()`.

```
datos<-sim2diss(correlacion,method="corr",to.dist=TRUE)
```

La conversión de similitudes (correlaciones) en disimilaridades se ha hecho por el método “corr”, que utiliza la expresión general $\delta_{ij} = 1 - s_{ij}$. Existen otros métodos en la función `sim2diss()` para cuando la matriz de proximidades no sea de correlaciones. El argumento `to.dist=TRUE` permite convertir el resultado en un objeto de la clase `dist`.

- Una vez que disponemos de la matriz de disimilaridades, aplicamos el EMD métrico mediante la función `mds()`, versión equivalente a la función `smacofSym()`.

```
res<-mds(datos,ndim=2,type="ratio")
res
#>
#> Call:
#> mds(delta = datos, ndim = 2, type = "ratio")
#>
#> Model: Symmetric SMACOF
#> Number of objects: 7
#> Stress-1 value: 0.161
#> Number of iterations: 42
```

La Fig. 34.1 que representa los objetos según sus distancias “reproducidas” muestra la configuración final de los siete objetos:

```
plot(res)
```

La información numérica detallada se podría obtener con la información de la salida dada por: `res$conf` que mostraría las coordenadas de los objetos en las dos dimensiones; `res$confdist` que muestra la matriz de distancias reproducidas; `res$stress` para obtener la medida de stress de Kruskal y `res$spp` con la contribución porcentual de cada objeto al stress.

En este caso los resultados muestran que la medida de stress es razonablemente baja con un valor de 0.16, indicando una buena “reproducción” de las proximidades originales. Además, la “contribución” relativa al stress de cada uno de los objetos (delitos) es bastante homogénea, siendo la variable porcentaje de empresas con banda ancha (EBROAD) la que más contribuyen al stress y el nivel de acceso a internet de los hogares (HIACC), la que menos. Para ver gráficamente el grado de ajuste, se usa el gráfico de Shepard (Fig. 34.2) que compara las proximidades originales y las distancias obtenidas:

34.4. Tipos de escalamiento multidimensional

575

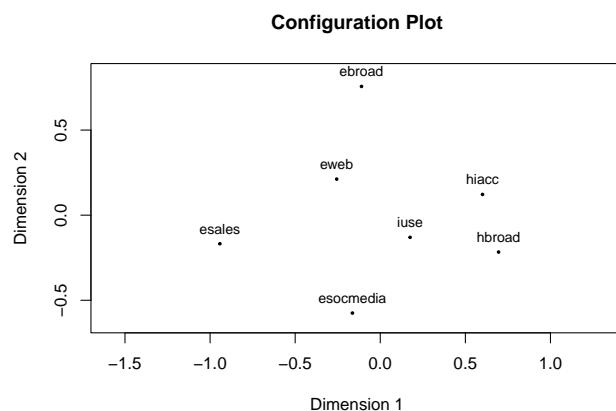


Figura 34.1: Gráfico de objetos en el plano de las distancias reproducidas.

```
plot(res, plot.type="Shepard")
```

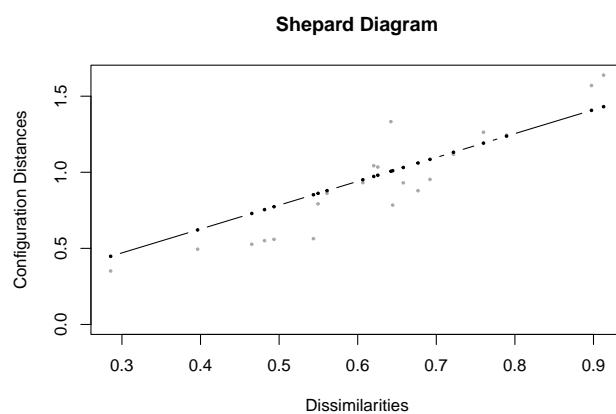


Figura 34.2: Gráfico de Shepard con método 'ratio'.

El diagrama de Shepard incluye las proximidades originales entre pares de objetos (en gris claro) y las obtenidas por el EMD (en negro). También representa el método elegido; en este ejemplo, al usar el argumento `method="ratio"` estamos imponiendo una relación proporcional entre ambos tipos de similitudes, por lo que aparece una recta (que pasa por el origen). Dadas las diferencias que se ven en el gráfico, quizás la elección del método `ratio` (opción por defecto), no sea la más adecuada. Probando con el método `interval`, que impone una relación lineal entre ambos tipos de similitudes (recta que no tiene que pasar necesariamente por el origen), se obtendría la Fig. 34.3:

```
res2<-mds(datos,type="interval")
plot(res2,plot.type="Shepard")
```

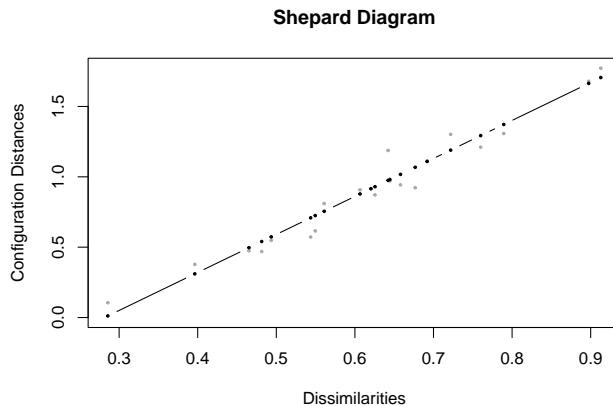


Figura 34.3: Gráfico de Shepard con método ‘interval’.

La medida de stress se ha reducido a la mitad, 0.087, indicando mejor ajuste, y el gráfico de Shepard muestra más concordancia entre las proximidades originales de los pares de objetos y las distancias reproducidas. Otra opción sería usar el método `mspline`, pero en este caso las diferencias son menores. Los tres métodos son métricos, puesto que se fija una forma funcional para relacionar las proximidades originales y las disimilitudes del modelo (un ratio, una función lineal o una función spline).

- Por último, para “interpretar” el sentido de las dimensiones en las que se representan los objetos, se recurre a ver cuáles están en los extremos. En la parte izquierda de la dimensión 1 están las variables relacionadas con el uso en las empresas mientras que en la parte derecha están las relacionadas con los hogares y las personas; se podría decir, entonces, que es una dimensión relacionada con el ámbito de uso de las TIC. En la dimensión 2, con menos distancias, y una interpretación menos clara, aparecen en la parte superior las variables relacionadas con el tipo de conexión y la existencia de web en las empresas y en la parte inferior las relacionadas con las redes sociales, las ventas o la frecuencia de uso de internet por parte de los individuos; se podría decir, **que es una dimensión asociada al uso dado a las TIC**.

Los pasos para desarrollar el mismo análisis agrupando países, observaciones en lugar de variables, serían similares pero usando la matriz de distancias en lugar de la matriz de correlaciones, por lo tanto:

```
library('factoextra')
d_euclidea <- get_dist(x = TIC2021, method = "euclidea")
res3<-mds(d_euclidea,ndim=2,type="ratio")
res3
#>
```

34.4. Tipos de escalamiento multidimensional

577

```
#> Call:
#> mds(delta = d_euclidea, ndim = 2, type = "ratio")
#>
#> Model: Symmetric SMACOF
#> Number of objects: 27
#> Stress-1 value: 0.112
#> Number of iterations: 77
```

```
plot(res3)
```

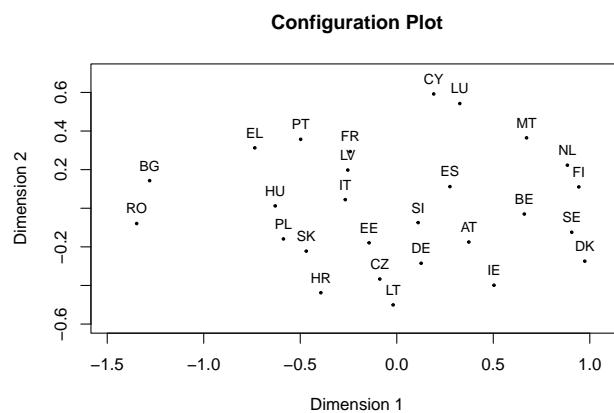


Figura 34.4: Gráfico de países sobre el plano de las distancias reproducidas.

En la Fig. 34.4, la interpretación de la dimensión 1 muestra la existencia de una diferencia clara entre los países del norte y los de última adhesión a la Unión Europea mientras que en la dimensión 2, con unas diferencias menores, aparecen en la parte superior los países de Chipre y Luxemburgo y en la parte inferior Lituania, Croacia, Irlanda y la República Checa.

34.4.2. Escalado multidimensional no métrico

En el modelo de escalamiento no métrico (también conocido como EMD ordinal) no se asume ninguna fórmula métrica que relacione las proximidades originales con las distancias reproducidas, sino que sólo describe un patrón creciente (o decreciente) entre ellas. No es significativo el valor de distancia, sino su relación con las distancias entre otros pares de objetos, por lo que construye distancias ajustadas que están en el mismo orden de rango que la proximidad original. Por ejemplo, si la distancia de los objetos separados A y B ocupa el tercer lugar en los datos de proximidades originales ordenadas, entonces también debería ocupar el tercer lugar en los datos de distancias reproducidas ordenadas. Así, el EMD no métrico busca conservar, no tanto los valores de las proximidades originales, sino la ordenación de los objetos en función de

dichas proximidades; es, por tanto, un modelo que se ajusta mejor a datos cualitativos que el métrico, aunque también se utiliza cuando se busca mayor flexibilidad en el ajuste.

Se va a aplicar un MDS no métrico a la matriz de correlaciones de las tasas anuales de siete delitos en 50 estados de EE.UU. (asesinatos, violaciones, robos, asaltos, allanamiento de morada, hurtos y sustracciones de vehículos). Los datos están disponibles en la librería `smacof` bajo el nombre de `crimes`:

```
data("crimes")
```

En este ejemplo, los “*objetos*” son los siete tipos de delito, y se ha utilizado como medida de proximidad (similitud) el coeficiente de correlación entre las tasas de delito (variables cuantitativas).

- El primer paso consiste en calcular la matriz de disimilaridades sobre la que actúa `smacof`, utilizando la función `sim2diss()`.

```
options(digits=3)
data<-sim2diss(crimes,method="corr",to.dist=FALSE)
data
#>           Murder Rape Robbery Assault Burglary Larceny Auto.Theft
#> Murder      0.000 0.693  0.812  0.436   0.849   0.970   0.943
#> Rape        0.693 0.000  0.671  0.548   0.566   0.632   0.748
#> Robbery     0.812 0.671  0.000  0.663   0.616   0.748   0.616
#> Assault     0.436 0.548  0.663  0.000   0.693   0.825   0.819
#> Burglary    0.849 0.566  0.616  0.693   0.000   0.447   0.548
#> Larceny     0.970 0.632  0.748  0.825   0.447   0.000   0.671
#> Auto.Theft  0.943 0.748  0.616  0.819   0.548   0.671   0.000
```

La conversión de similitudes (correlaciones) en disimilaridades se ha hecho por el método `corr`, que utiliza la expresión general $\delta_{ij} = \sqrt{1 - s_{ij}}$. Existen otros métodos en la función `sim2diss()` para cuando la matriz de proximidades no sea de correlaciones. El argumento `to.dist=TRUE` permite convertir el resultado en un objeto de la clase `dist` si fuese necesario.

Frente a los métodos métricos, que fijan una forma funcional para relacionar las proximidades originales y las disimilaridades del modelo (un ratio, una función lineal o una función spline), una alternativa más flexible es asumir una relación no métrica, donde sólo se conserve la ordenación de las proximidades originales. En este caso, se busca que las distancias reproducidas ordenen a los pares de objetos de forma idéntica a la original.

Para ello, se utiliza el método `ordinal` dentro de la función `mds()`:

```
res4<-mds(data,ndim=2,type="ordinal")
res4
#>
#> Call:
#> mds(delta = data, ndim = 2, type = "ordinal")
```

34.4. Tipos de escalamiento multidimensional

579

```
#>
#> Model: Symmetric SMACOF
#> Number of objects: 7
#> Stress-1 value: 0.002
#> Number of iterations: 15
```

Como se aprecia, la medida de stress es muy baja (0.002), indicando un muy buen ajuste, que resulta significativo como indica la Fig. 34.5 basada en la función `permtest()`:

```
ptestnm<-permtest(res4,nrep=50)
```

```
plot(ptestnm)
```

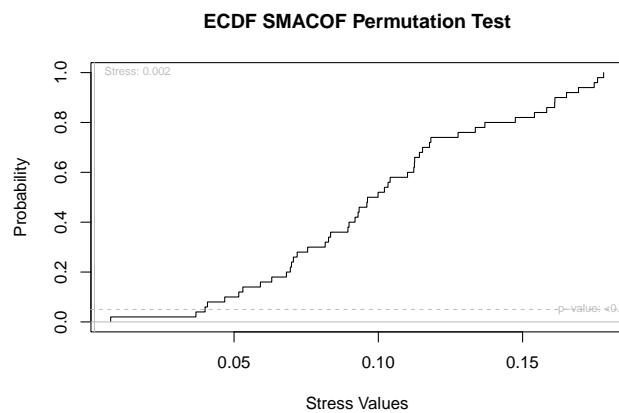


Figura 34.5: Test de permutaciones para evaluación de la significatividad de la medida de stress.

El gráfico de Shepard, representado en la Fig. 34.6, incluye las proximidades originales entre pares de objetos (en gris claro) y las obtenidas por el MDS (en negro), mostrando una alta concordancia entre las ordenaciones original y reproducida:

```
plot(res4,plot.type="Shepard")
```

La contribución porcentual de cada delito a la medida de stress y las coordenadas bidimensionales reproducidas serían:

```
res4$spp #Contribución porcentual de cada objeto al stress
#>      Murder      Rape     Robbery    Assault   Burglary    Larceny Auto.Theft
#>     8.064     6.443    39.270     0.104    10.618     9.023   26.478
res4$conf #Coordenadas de los objetos en dos dimensiones
```

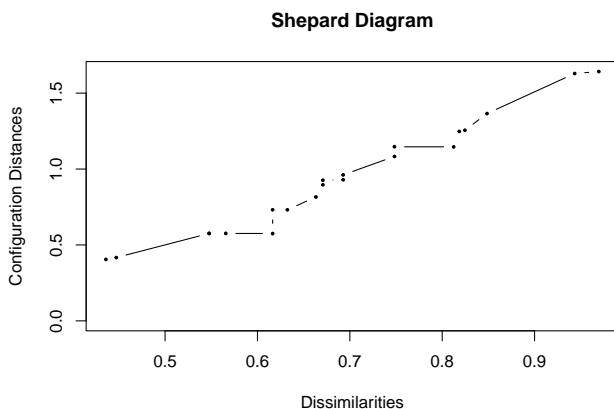


Figura 34.6: Gráfico de Shepard: concordancia entre las ordenaciones original y reproducida.

```
#>           D1      D2
#> Murder    -0.9673  0.01136
#> Rape      -0.1225 -0.37592
#> Robbery    0.0496  0.53424
#> Assault   -0.5630 -0.00483
#> Burglary   0.3925 -0.11326
#> Larceny    0.6015 -0.47400
#> Auto.Theft 0.6093  0.42240
```

El delito de robo (39.27 %) y el de sustracción de vehículos (26.48 %) son responsables del 65.75 % del stress, seguidos muy de lejos por los otros cinco tipos de delito.

La Fig. 34.7 muestra la representación bidimensional de la configuración final de los siete delitos según sus distancias “reproducidas”:

```
plot(res4)
```

Por último, para “interpretar” el sentido de las dimensiones en las que se representan los objetos, se recurre a ver cuáles están en los extremos. En la parte izquierda de la dimensión 1 están los delitos de asesinato y asalto, mientras que en la parte derecha están los de hurto, sustracción de vehículos y allanamiento; se podría decir, entonces, que es una dimensión relacionada con el grado de “personalización” del delito: los que afectan a personas directamente frente a los que no. La dimensión 2, con menos distancias (véase las escalas) y de más difícil interpretación, contrapone en la parte superior los delitos que más “beneficios” económicos producen (robos, sustracción de vehículos) frente a los que menos (violación o hurto); se podría decir, entonces, que es una dimensión asociada a la repercusión económica del delito.

Por último, destacar que puede ser interesante analizar la estabilidad de las soluciones del EMD (bien mediante jackknife, bootstrap, o elipses de pseudo-confianza). También puede interesar

34.4. Tipos de escalamiento multidimensional

581

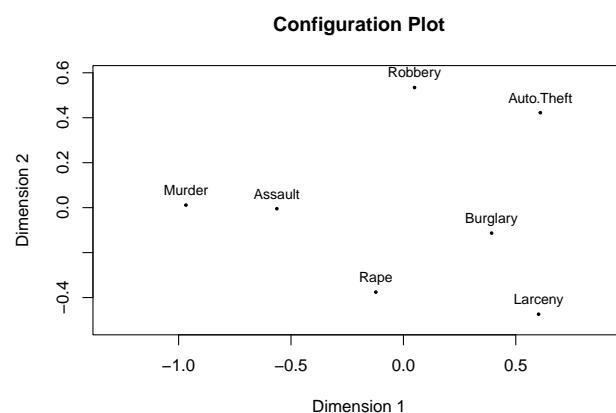


Figura 34.7: Representación bidimensional de las disimilitudes entre los siete tipos de delito.

plantear un modelo de EMD que permita valorar las diferencias individuales (abordable con la función `smacofIndDiff()`). Otra alternativa puede ser abordar un desplegamiento multidimensional, que representa conjuntamente objetos e individuos.

Resumen

La aplicación del análisis EMD implica tres pasos consecutivos:

- La determinación de las proximidades originales entre los objetos. Esta fase depende de las características de los objetos y del tipo de relación que se quiera/pueda establecer entre ellos. Actualmente, **R** dispone de paquetes que permiten estimar las matrices de proximidad a partir de los datos en bruto.
- La conversión de las proximidades en similaridades (si fuese necesario) y el ajuste entre las originales y las reproducidas por el EMD. Se debe elegir el tipo de EMD a utilizar, que depende de la función de ajuste: se puede optar por funciones de tipo `ratio`, `interval` o `mspline` (EMD métricos) o `ordinal` (EMD no métrico). La elección estará relacionada con el tipo de datos usados y el grado de ajuste (stress) de los modelos.
- La interpretación de los resultados a partir de la configuración obtenida, tanto del significado de las dimensiones como de la estructura de los objetos (cuáles se parecen, si existen grupos, etc.)

Capítulo 35

Análisis de correspondencias

Román Minguez Salido^a y Manuel Vargas Vargas^a

^a Universidad de Castilla-La Mancha

35.1. Introducción

El **análisis de correspondencias** es un método gráfico descriptivo de reducción de la dimensión incluido entre los algoritmos de aprendizaje no supervisado. La idea principal es equivalente al método de componentes principales, pero aplicado a variables cualitativas. El objetivo es representar los valores (**niveles** en **R**) de variables cualitativas (**factores** en **R**) en ejes cuantitativos cuyas coordenadas representen la cercanía o lejanía entre los niveles de los factores. Es decir, es un método de reducción de la dimensionalidad para factores representables en pocas dimensiones.

Por sencillez, el punto de partida será una **tabla de contingencia RxC**, T , (véase Cap. 23) que recoge la frecuencia de cada par de niveles A_1, A_2, \dots, A_R del factor A y B_1, B_2, \dots, B_C del factor B :

Tabla 35.1: Ejemplo de tabla de contingencia RxC

	B_1	B_2	...	B_C	Total
A_1	n_{11}	n_{12}	...	n_{1C}	$n_{1\cdot}$
A_2	n_{21}	n_{22}	...	n_{2C}	$n_{2\cdot}$
...
A_R	n_{R1}	n_{R2}	...	n_{RC}	$n_{R\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$...	$n_{\cdot C}$	N

Cada fila representa el **perfil condicional** del nivel A_i , siendo la última el **perfil marginal**

del factor A . Igualmente, cada columna representa el **perfil condicional** del nivel B_j , siendo la última el **perfil marginal** del factor B .

Como se vió en el Cap. 23, si los factores fueran independientes, el **valor esperado** en cada casilla sería $E_{ij} = \frac{n_i \cdot n_{\cdot j}}{N}$, por lo que la diferencia tipificada, $r_{ij} = \frac{n_{ij} - E_{ij}}{\sqrt{E_{ij}}}$ es una medida de asociación entre las modalidades A_i y B_j . La matriz formada por estos “residuos estandarizados” (véase sección 23.5.4), $R = \{r_{ij}\}$ resume la asociación entre los atributos, y es el objetivo básico del análisis de correspondencias; básicamente, se realiza una proyección de las filas y columnas de la tabla de frecuencias relativas (transformadas) para obtener las coordenadas en ejes cuantitativos, representables en la forma habitual como diagramas de puntos.

Para un estudio en profundidad de esta técnica pueden consultarse [Greenacre \(2008\)](#) (en español) o [Beh and Lombardo \(2014\)](#). En el resto del capítulo se hará una breve exposición de la metodología y se exemplificará con el análisis de dos tablas de contingencia.

35.2. Metodología del análisis de correspondencias

Dada una tabla de contingencia T , a partir de las frecuencias observadas n_{ij} , se definen las **distancias** entre los perfiles:

- para los perfiles fila, $d_{ii^*} = \sum_{k=1}^C \frac{1}{n_{\cdot k}} \left(\frac{n_{ik}}{n_{i^* \cdot}} - \frac{n_{i^* k}}{n_{i^* \cdot}} \right)^2$
- para los perfiles columna, $d_{jj^*} = \sum_{k=1}^R \frac{1}{n_{k \cdot}} \left(\frac{n_{kj}}{n_{\cdot j^*}} - \frac{n_{k j^*}}{n_{\cdot j^*}} \right)^2$

Estas distancias aumentan cuanto más se “*diferencien*” unos perfiles de otros. El análisis de correspondencias busca construir **dimensiones** (habitualmente, dos) y obtener las coordenadas de los niveles de ambos factores en dichas dimensiones

$$\mathbf{A} = \begin{pmatrix} \mathbf{a}'_1 \\ \vdots \\ \mathbf{a}'_R \end{pmatrix}, \text{ con } \mathbf{a}_i = (a_{i1} \ a_{i2})', \text{ y } \mathbf{B} = \begin{pmatrix} \mathbf{b}'_1 \\ \vdots \\ \mathbf{b}'_C \end{pmatrix}, \text{ con } \mathbf{b}_j = (b_{j1} \ b_{j2})' \quad (35.1)$$

siendo \mathbf{a}_i las coordenadas del nivel fila A_i y \mathbf{b}_j las del nivel columna B_j en el plano, de forma que “*reproducen*” las distancias entre perfiles fila y columna y los residuos estandarizados (asociaciones):

$$\begin{aligned} d(\mathbf{a}_i, \mathbf{a}_{i'}) &= \sqrt{(a_{i1} - a_{i'1})^2 + (a_{i2} - a_{i'2})^2} \approx d_{ii'} \\ d(\mathbf{b}_j, \mathbf{b}_{j'}) &= \sqrt{(b_{j1} - b_{j'1})^2 + (b_{j2} - b_{j'2})^2} \approx d_{jj'} \\ \mathbf{a}'_i * \mathbf{b}_j &\approx r_{ij} \end{aligned} \quad (35.2)$$

Con las coordenadas contenidas en las matrices \mathbf{A} y \mathbf{B} , es posible “*visualizar*” la posición relativa de cada factor en las nuevas dimensiones. Esta estructura permite ver, tanto las “*distancias*” que hay entre los niveles de cada factor (mediante la distancia de representación en el plano),

como las “*asociaciones*” entre niveles de ambos factores (ya que mientras más asociación haya, más cerca se representarán en el plano).

Para resolver el problema de estimación de las matrices \mathbf{A} y \mathbf{B} , se busca una descomposición de la matriz de $\mathbf{R} = \{r_{ij}\}$ en valores singulares. Según la importancia que se da al ajuste de uno de los perfiles o a la matriz de residuos, se tienen diferentes métodos de selección, llamados **normalizaciones**, que pueden consultarse en [Greenacre \(2008\)](#).

35.2.1. Proyecciones fila, columna y simétrica

El punto de partida es la matriz de frecuencias relativas \mathbf{P} cuyas entradas son n_{ij}/N , también llamada **matriz de correspondencias**. Definiendo el vector de unos $\mathbf{1}$, con la dimensión adecuada, las masas, o frecuencias marginales, de filas y columnas, $r_i = \sum_{j=1}^C p_{ij}$ y $c_j = \sum_{i=1}^R p_{ij}$, respectivamente, se pueden expresar matricialmente como $\mathbf{r} = \mathbf{P}\mathbf{1}$ y $\mathbf{c} = \mathbf{P}'\mathbf{1}$ o, en forma de matrices diagonales,

$$\mathbf{D}_R = \text{diag}(r) \equiv \text{diag}(f_1, \dots, f_R) \quad \text{y} \quad \mathbf{D}_C = \text{diag}(c) \equiv \text{diag}(f_{,1}, \dots, f_{,C})$$

Se calcula la matriz de **residuos estandarizados** (véase Sec. @contaprox) como

$$\mathbf{S} = \mathbf{D}_R^{-\frac{1}{2}} (\mathbf{P} - \mathbf{rc}') \mathbf{D}_C^{-\frac{1}{2}} \quad (35.3)$$

La matriz \mathbf{S} se descompone en valores singulares, calculando matrices las \mathbf{U} , \mathbf{D} y \mathbf{V} tales que:

$$\begin{aligned} \mathbf{S} &= \mathbf{UDV}' \\ \mathbf{UU}' = \mathbf{V}'\mathbf{V} &= \mathbf{I}, \quad \mathbf{U}_{(R \times K)}, \quad \mathbf{V}_{(C \times K)}, \quad K = \min(R - 1, C - 1) \\ \mathbf{D} &= \text{diag}(\mu_1, \dots, \mu_K) \end{aligned} \quad (35.4)$$

donde los μ_i son los llamados **valores singulares**, estando ordenados de forma decreciente $\mu_1 \geq \mu_2 \geq \dots \geq \mu_K$.

A partir de esta descomposición se pueden obtener:

- las **coordenadas estándar de las filas**, $\Phi = \mathbf{D}_R^{-\frac{1}{2}} \mathbf{U}$, y sus **coordenadas principales**, $\mathbf{F} = \Phi \mathbf{D}$.
- las **coordenadas estándar de las columnas**, $\Gamma = \mathbf{D}_C^{-\frac{1}{2}} \mathbf{V}$, y sus **coordenadas principales**, $\mathbf{G} = \Gamma \mathbf{D}$.
- las **inercias principales**, $\lambda_i = \mu_i^2$.

Las coordenadas principales son las utilizadas para definir las **proyecciones fila** y **proyecciones columna**, que representan, en menor dimensión, los perfiles correspondientes, formando los llamados **mapas asimétricos**.

Por último, las matrices $\mathbf{A} = \mathbf{D}_R^{-\frac{1}{2}} \mathbf{UD}$ y $\mathbf{B} = \mathbf{D}_C^{-\frac{1}{2}} \mathbf{VD}$ representan las coordenadas de ambos perfiles en un espacio común, llamado **mapa simétrico**.

35.3. Procedimiento con R: la función ca()

Para realizar un análisis de correspondencias simple con **R** se puede utilizar el paquete **ca**, que contiene la función **ca()**. Esta función acepta como argumento de entrada o bien directamente una tabla de contingencia, o bien los datos originales como objeto matriz o data-frame. Incluso, el argumento puede ser una fórmula del tipo $\sim F1 + F2$ donde $F1$ y $F2$ son factores. Entre los argumentos adicionales se pueden incluir el número de dimensiones en el output así como filas o columnas suplementarias.

35.3.1. Caso práctico 1: tareas del hogar.

Como primer ejemplo, se van a utilizar los datos **housetasks**, contenidos en el paquete **factoextra**, que representan una tabla de contingencia con la frecuencia de ejecución de 13 tareas del hogar por los miembros de la pareja.

```
library('ca')
library('factoextra')
data('housetasks')
```

En primer lugar, la aplicación del test χ^2 de independencia (véase 23) permite contrastar si los factores son independientes o, por el contrario, están asociados:

```
chisq.test(housetasks)
#>
#> Pearson's Chi-squared test
#>
#> data: housetasks
#> X-squared = 1944.5, df = 36, p-value < 2.2e-16
```

Con un valor de $\chi^2 = 1944,5$ y un p-valor de $2.2e-16$, hay suficiente evidencia como para rechazar la hipótesis nula de independencia, indicando asociación entre ambos factores, por lo que tiene sentido analizar más en profundidad la estructura de dicha asociación.

La función **ca()** proporciona los valores singulares y, tanto para filas como para columnas, las masas (valores “*Mass*”); las distancias chi-cuadrado, que representan las distancias en esa métrica de cada fila respecto a la fila centroide (dada por la masa de las columnas, promedio de los vectores fila); las inercias explicadas, que representan la distancia cuadrática χ^2 respecto al perfil promedio (sin calcular raíces), ponderada por la masa (de la fila o columna) correspondiente; así como las coordenadas en el espacio proyectado:

```
options(digits = 2)
ca_house <- ca(housetasks, nd = 2)
ca_house
#>
#> Principal inertias (eigenvalues):
```

35.3. Procedimiento con **R**: la función **ca()**

587

```
#>      1       2       3
#> Value    0.542889 0.445003 0.127048
#> Percentage 48.69% 39.91% 11.4%
#>
#>
#> Rows:
#>   Laundry Main_meal Dinner Breakfast Tidying Dishes Shopping Official
#> Mass     0.10    0.088  0.062     0.080  0.070  0.065  0.069  0.055
#> ChiDist  1.15    1.017  0.786     0.716  0.594  0.550  0.466  0.984
#> Inertia  0.13    0.091  0.038     0.041  0.025  0.020  0.015  0.053
#> Dim. 1   -1.35   -1.188 -0.940    -0.690 -0.534 -0.256 -0.160  0.308
#> Dim. 2   -0.74   -0.735 -0.462    -0.679  0.651  0.663  0.605 -0.380
#>           Driving Finances Insurance Repairs Holidays
#> Mass      0.08    0.065   0.080   0.095   0.092
#> ChiDist  1.13    0.675   0.853   1.819   1.463
#> Inertia  0.10    0.030   0.058   0.313   0.196
#> Dim. 1   1.01    0.367   0.878   2.075   0.343
#> Dim. 2   -0.98   0.926   0.710  -1.296   2.151
#>
#>
#> Columns:
#>   Wife Alternating Husband Jointly
#> Mass    0.34    0.146   0.22    0.29
#> ChiDist 0.94    0.899   1.32    1.04
#> Inertia 0.30    0.118   0.38    0.31
#> Dim. 1  -1.14   -0.084   1.58    0.20
#> Dim. 2  -0.55   -0.437   -0.90   1.54
```

Las dos primeras dimensiones explican el 48.69 % y 39.91 % de la inercia respectivamente, por lo que la representación en un plano engloba al 88.6 % de la inercia global.

Las distancias chi-cuadrado indican lo cerca o lejos que está cada fila respecto al centroide de las mismas. En este ejemplo, la fila más distante del centroide de filas es “Repairs” (1.819), mientras que la columna más distante respecto del centroide de columnas es “Husband” (1.321).

Como las inercias miden la variabilidad de los perfiles, en este ejemplo, respecto a las filas, el nivel que mayor contribuye es “Repairs” (0.312874) mientras que por columnas es “Husband” (0.381373). Esto no es sorprendente ya que ambos niveles eran los más alejados del centro.

Con las coordenadas de las dimensiones se puede realizar un gráfico de las mismas utilizando la función **plot()**, pudiéndose optar por la proyección sólo de las filas (usando los argumentos **map = “rowprincipal”**, **what = c(“all”, “none”)**) o de las columnas (**map = “colprincipal”**, **what = c(“none”, “all”)**), tal como se muestra en la Fig. 35.1:

```
par(mfrow = c(1, 2))
plot(ca_house, map = "rowprincipal", what = c("all", "none"), xlab = "Perfiles fila")
plot(ca_house, map = "colprincipal", what = c("none", "all"), xlab = "Perfiles
→ columna")
```

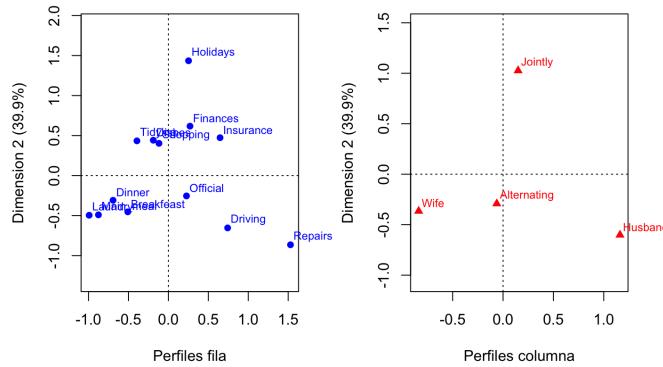


Figura 35.1: Proyecciones de los perfiles fila y columna

Respecto a las filas, se aprecian varios grupos: el compuesto por “*Breakfast*”, “*Dinner*”, “*Main_meal*” y “*Laundry*”; otro por “*Shopping*”, “*Dishes*” y “*Tidying*”; uno tercero por “*Insurance*” y “*Finance*”; y el compuesto por “*Driving*” y “*Official*”. Los niveles “*Holiday*” y “*Repairs*” están alejados del resto.

Las coordenadas simétricas permiten la representación de ambos factores a la vez (*map*= “*symmetric*”, *what*=*c*(“*all*”, “*all*”)), como se muestra en la Fig. ??.

```
plot(ca_house, map = "symmetric", what = c("all", "all"), xlab = "Proyección común de
↪ ambos factores")
```

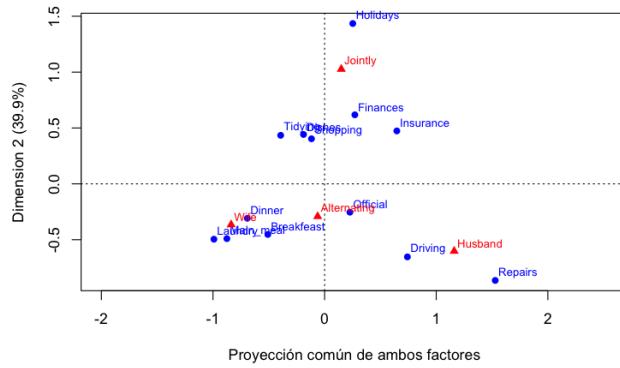


Figura 35.2: Proyección simétrica de ambos factores

El gráfico conjunto permite observar qué niveles de filas y columnas pueden estar más cercanos (aproximación a la asociación entre ellos). El grupo de “*Driving*” y “*Repairs*” está cercano a “*Husband*”; el grupo de “*Dinner*”, “*Breakfast*”, “*Laundry*” y “*Main_meal*” está cercano

a “*Wife*”; mientras que el nivel “*Jointly*” parece estar asociado a “*Holidays*”, “*Finance*”, e “*Insurance*”.

35.3.2. Caso práctico 2: accidentes 2020.

Como segundo ejemplo, se van a utilizar los datos `accidentes2020_data`, contenidos en el paquete CDR, en concreto, la información sobre “*tipo_accidente*” y “*estado_meteorológico*”. Para evitar pares de niveles con frecuencia nula, se eliminan los niveles “*Atropello a animal*”, “*Despeñamiento*” y “*Otro*” del factor “*tipo_accidente*” y los niveles “*Granizando*”, “*Nevando*”, “*NULL*” y “*Se desconoce*” del factor “*estado_meteorológico*”.

```
library('CDR')
library('dplyr')
data('accidentes2020_data')
datos <- data.frame(
  V1 = as.factor(accidentes2020_data$tipo_accidente),
  V2 = as.factor(accidentes2020_data$estado_meteorológico)
)
levelsV1 <- c("Alcance", "Choque contra obstáculo fijo", "Colisión frontal", "Colisión
  ↪ fronto-lateral", "Colisión lateral", "Colisión múltiple")
levelsV2 <- c("Despejado", "Lluvia débil", "Lluvia intensa", "Nublado")
datos_depu <- droplevels(filter(datos, (V1 %in% levelsV1) & (V2 %in% levelsV2)))
```

		<i>V2</i>			
		<i>Despejado</i>	<i>Lluvia débil</i>	<i>Lluvia intensa</i>	<i>Nublado</i>
#>	<i>V1</i>				
#>	<i>Alcance</i>	5525	403	84	449
#>	<i>Choque contra obstáculo fijo</i>	3258	308	43	224
#>	<i>Colisión frontal</i>	711	42	5	48
#>	<i>Colisión fronto-lateral</i>	6359	398	51	494
#>	<i>Colisión lateral</i>	3241	169	30	277
#>	<i>Colisión múltiple</i>	1619	173	29	111

Se comprueba que existe asociación y se obtienen los resultados del análisis de correspondencias:

```
tabla <- table(datos_depu)
chisq.test(tabla)
#>
#> Pearson's Chi-squared test
#>
#> data: tabla
#> X-squared = 104, df = 15, p-value = 2e-15
ca_tabla <- ca(tabla, k = 2)
ca_tabla
#>
#> Principal inertias (eigenvalues):
```

```

#>      1       2       3
#> Value  0.003804 0.000493 2.7e-05
#> Percentage 87.97% 11.4% 0.62%
#>
#>
#> Rows:
#>      Alcance Choque contra obstáculo fijo Colisión frontal
#> Mass    0.26864          0.1594        0.03351
#> ChiDist 0.03200          0.0817        0.06591
#> Inertia 0.00028          0.0011        0.00015
#> Dim. 1 -0.16016          0.2818        0.72969
#> Dim. 2  1.36554         -0.9207       -1.84795
#>      Colisión fronto-lateral Colisión lateral Colisión múltiple
#> Mass            0.3036        0.15455      0.0803
#> ChiDist         0.0446        0.07682      0.1284
#> Inertia         0.0006        0.00091      0.0013
#> Dim. 1          0.6586        1.22996     -2.0813
#> Dim. 2          -0.8221       0.52856      0.1210
#>
#>
#> Columns:
#>      Despejado Lluvia débil LLuvia intensa Nublado
#> Mass    0.86121        0.0621        0.01006 0.06665
#> ChiDist 0.01344        0.2145        0.28954 0.08391
#> Inertia 0.00016        0.0029        0.00084 0.00047
#> Dim. 1  0.20552        -3.4619       -3.67116 1.12291
#> Dim. 2  -0.19007       -0.8381       8.05783 2.02007

```

Las dos primeras dimensiones explican el 87.97% y 11.4% respectivamente, por lo que la representación en un plano engloba al 99.37% de la inercia.

La representación gráfica de la proyección simétrica se muestra en la Fig. 35.3:

```

plot(ca_tabla, map = "symmetric", what = c("all", "all"), xlab = "Proyección común de
→ ambos factores")

```

Se observa que “*Lluvia intensa*” no está especialmente asociada a ningún tipo de accidente; “*Lluvia débil*” con “*Colisión múltiple*” y “*Choque contra obstáculo fijo*”; “*Despejado*” con “*Colisión fronto-lateral*”, “*Colisión frontal*” y “*Alcance*”; y “*Nublado*” con “*Colisión lateral*”.

35.3. Procedimiento con **R**: la función **ca()**

591

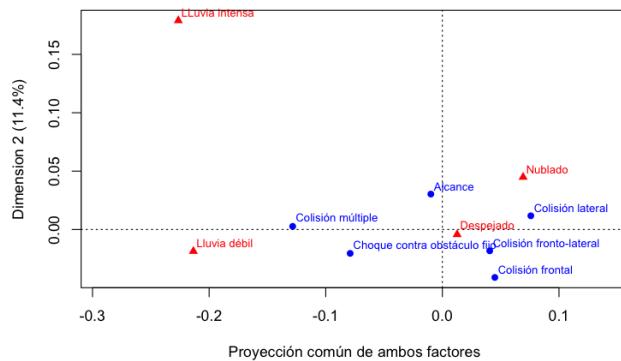


Figura 35.3: Proyección simétrica de ambos factores

Resumen

Dada una tabla de contingencia, el análisis de correspondencias reproduce: (i) las distancias entre niveles de cada factor en un espacio de menor dimensión, permitiendo la comparación gráfica entre ellos; (ii) la representación de los niveles de ambos factores en un espacio común.

En el primer caso, permite una visualización de la composición interna de cada factor, identificando los niveles que más se distancian del centroide. En el segundo, permite la representación de la asociación entre niveles de cada uno de los factores.

Parte VII

Deep learning

Capítulo 36

Redes neuronales artificiales

Noelia Vállez Enano^a y José Luis Espinosa Aranda^a

^aUniversidad de Castilla-La Mancha

36.1. ¿Qué es el *deep learning*?

La inteligencia artificial es una disciplina científica que se ocupa de crear programas informáticos que ejecutan operaciones comparables a las que realiza la mente humana, como el aprendizaje o el razonamiento lógico (de la Real Academia Española, 2023). Entre otros ejemplos se pueden encontrar en la actualidad tanto robots que son capaces de realizar tareas de manera similar a un humano en una fábrica, las denominadas como casas inteligentes o los vehículos autónomos.

Dentro de las técnicas utilizadas para la inteligencia artificial, se encuentran las técnicas clásicas de *machine learning*, ya explicadas en capítulos anteriores de este libro, las cuales tienen la habilidad de aprender sin haber sido explícitamente programadas para una tarea en particular, pudiendo ser utilizadas para varios fines y aplicaciones.

A su vez, dentro de estos algoritmos, se pueden enmarcar como un subconjunto de las mismas las técnicas de *deep learning*, las cuales intentan simular tanto la arquitectura como el comportamiento del sistema nervioso humano, en particular, de las redes de neuronas que componen el encéfalo y que se encargan de realizar tareas específicas (Fig. 36.1). Para ello, estas técnicas se basan en el concepto de redes neuronales, que intentan emular la forma de aprendizaje de los humanos (Goodfellow et al., 2016).

36.1.1. Diferencias entre las técnicas de *machine learning* tradicional y el *deep learning*

Como se vio en el Cap. 2, la metodologías de ciencia de datos tienen una etapa llamada preparación de datos que incluye la tarea de elección de variables, la cual ha sido tratada ampliamente

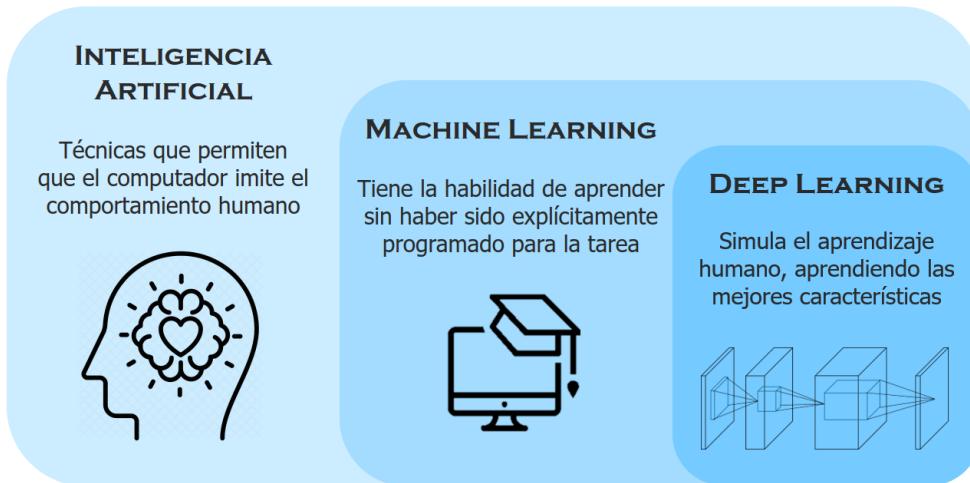


Figura 36.1: Inteligencia Artificial vs Machine learning vs Deep Learning

en el Cap. 9, para realizar una selección de las mejores características que representen el problema a resolver, y que puedan ser comprendidas por el algoritmo de *machine learning* seleccionado de tal forma que sea capaz de solucionar el problema planteado.

Por ejemplo, en el caso de querer detectar una cara dentro de una imagen, sería necesario definir qué tipo de características servirían para detectar la misma, como podrían ser, a bajo nivel, determinados tipos de bordes de la imagen (Fig. 36.2). Estas características proporcionarían la base para detectar a nivel medio elementos de la cara como ojos, narices, orejas, etc. y, definitivamente, a alto nivel, reconocer donde hay una cara dentro de la imagen.



Figura 36.2: Detección de bordes de una imagen mediante el método de Scharr

Esta elección de características requiere en muchas ocasiones de la intervención humana, por lo que puede llevar mucho tiempo y diversos experimentos de prueba y error hasta poder encontrar una combinación de características que permita resolver el problema planteado.

Las técnicas de *deep learning*, a diferencia de las técnicas de *machine learning* tradicional, son capaces de aprender cuales son las mejores características que permitirán representar el

36.1. ¿Qué es el deep learning?

597

problema que se quiere resolver sin necesidad de la interacción humana a la misma vez que buscan la solución al mismo.

Continuando con el ejemplo anterior de la detección de caras, mientras que en las técnicas de *machine learning* sería necesario introducirle al algoritmo qué características base componen una cara para que sea capaz de reconocerlas, al utilizar *deep learning* únicamente sería necesario mostrarle suficientes imágenes de caras para conseguir que el algoritmo sea capaz de aprender a identificar una cara por sí mismo, identificando de forma automática las características más importantes de una cara.

Esta capacidad de aprender las mejores características necesarias por sí mismo hace que a nivel teórico las técnicas de *deep learning* puedan llegar a ser más potentes que el *machine learning* clásico, pero debido a la mayor complejidad del problema y, por consiguiente, al proceso de entrenamiento, también lleva a que que sean necesarios muchos más datos y una mayor potencia de cómputo para entrenarlas.

Este hecho explica que, aunque las bases de las técnicas de *deep learning* como el algoritmo del descenso del gradiente (Kiefer and Wolfowitz, 1952), el perceptrón (Rosenblatt, 1958), los algoritmos de retropropagación y el perceptrón multicapa (Rumelhart et al., 1986) y la primera red neuronal convolucional (LeCun et al., 1995), datan de varios años atrás, no sea hasta hace relativamente poco tiempo, cuando se ha podido empezar a utilizar estas técnicas. Esto se debe a diversos factores:

1. **La evolución en el hardware de procesamiento.** En particular, debido a la mejora de la capacidad de paralelismo masivo durante el cómputo que proporcionaron las nuevas tarjetas gráficas (GPU) al incorporar una gran cantidad de microprocesadores específicos, han podido ser utilizadas para las técnicas de *deep learning*. Originalmente su principal uso era representar modelos complejos 3D en los monitores, pero su utilización para técnicas de *deep learning* ha llevado recientemente al desarrollo de tarjetas específicas para este fin. Además, es posible disponer bajo demanda de estos recursos de computación como servicios a través de Internet. Esto es lo que se conoce como *cloud computing*.
2. **El Big data.** La gran cantidad de datos que se generan y almacenan en la actualidad, así como la mayor facilidad a la hora de trabajar con esos conjuntos de datos (gracias a las nuevas herramientas disponibles), han permitido cubrir la necesidad del gran volumen de datos iniciales necesarios.
3. **La evolución del software.** Recientemente ha habido un amplio interés tanto en buscar nuevos modelos para resolver todo tipo de problemas, como para mejorar las técnicas utilizadas para entrenar dichas redes neuronales. Esto ha llevado a la creación y mejora de diversos frameworks, librerías y aplicaciones relacionadas con el entrenamiento y despliegue de redes neuronales. Entre ellos, serían destacables Keras, Tensorflow, Pytorch, Caffe2, Matlab y OpenVINO.

36.2. Aplicaciones del *deep learning*

Las posibles aplicaciones de las técnicas de *deep learning* son muy diversas y, gracias a la continua investigación desarrollada en el área en la actualidad, no hacen más que aumentar. A continuación se comentan algunas de ellas:

- 1. Clasificación de imágenes.** Aunque la clasificación de imágenes dentro del área de la visión por computador o artificial es una realidad hace muchos años, es con las técnicas de *deep learning* con las que se han logrado los mayores avances, en particular, utilizando las redes neuronales convolucionales. Estas redes permiten determinar a qué clase, perteneciente al conjunto de categorías utilizado para entrenar, se corresponde una determinada imagen.
- 2. Detección de objetos.** Permite localizar los objetos contenidos en una imagen mediante un rectángulo, clasificándolo a su vez por su tipología. Por ejemplo, utilizando una cámara de seguridad instalada en una calle con este tipo de modelos sería posible localizar y diferenciar entre peatones y vehículos (Fig. 36.3).



Figura 36.3: Detección de peatones y vehículos utilizando una cámara térmica y técnicas de deep learning

- 3. Segmentación semántica/de instancias.** De forma similar a la detección de objetos, la segmentación semántica permite localizar objetos contenidos en una imagen, además de su tipología, pero en este caso se marcan utilizando una máscara a nivel de píxel. La segmentación de instancias además es capaz de diferenciar entre diferentes instancias de una misma clase aun cuando se encuentren situadas de forma contigua.
- 4. Reconocimiento del habla.** Permite a un computador procesar y comprender el habla humana. En la actualidad existen varios asistentes inteligentes basados en esta tecnolo-

gía que además son capaces de interpretar órdenes o instrucciones sencillas y actuar en consecuencia.

5. **Traducción automática.** Consiste en utilizar las técnicas de *deep learning* para traducir un texto automáticamente de una lengua a otra sin la necesidad de intervención humana. En la actualidad, no se limita únicamente a la traducción literal, palabra por palabra, del texto, si no que también tiene en cuenta el significado que tendría en el idioma original para adaptarlo al idioma destino (Fig. 36.4).

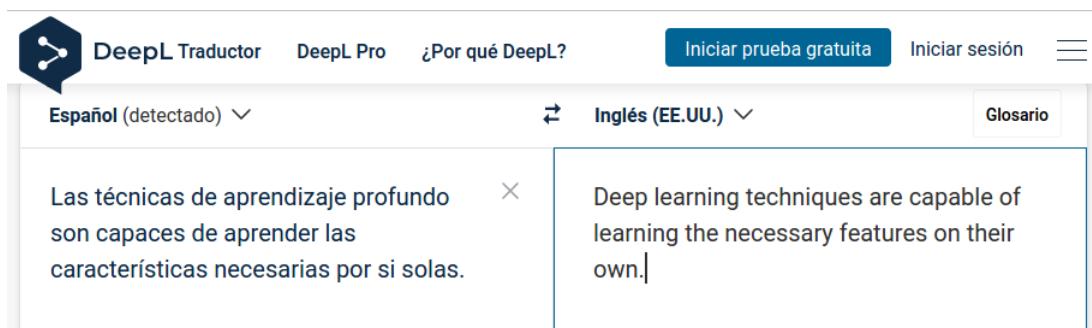


Figura 36.4: Traductor automático basado en Deep Learning

6. **Generación automática de imágenes/texto.** Permite obtener desde una imagen un texto descriptivo que indique el contenido de la imagen, o al contrario, a partir de un texto descriptivo generar una imagen basada en dicha descripción. Un ejemplo de este último caso sería Dall-E (Borji, 2022) (Fig. 36.5).

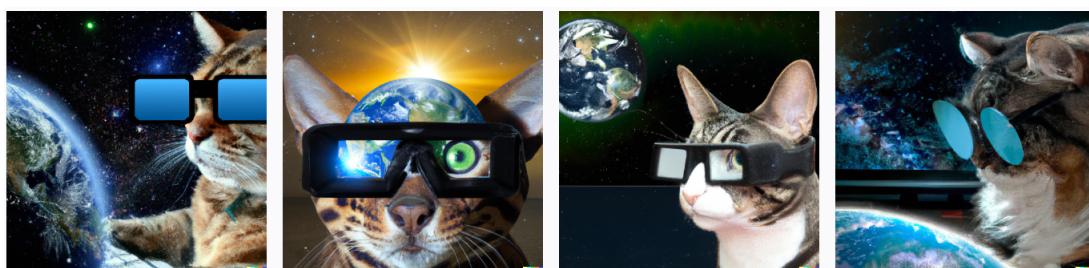


Figura 36.5: Algunas salidas posibles del generador de imágentes a partir de texto Dall-E, para el texto a “cat with glasses studying computer vision in the space with the Earth in the background”

7. **Automóvil autónomo.** Las técnicas de *deep learning* están siendo claves para el desarrollo del vehículo autónomo, capaz de circular sin la necesidad de la interacción de un conductor humano. Para lograr definitivamente un vehículo con estas características,

es necesario que sea capaz de ver, tomar decisiones y conducir al mismo tiempo. Esto se consigue en la actualidad integrando la información de gran cantidad de sensores que obtienen datos en tiempo real sobre el entorno, como serían cámaras, LIDAR, radares o ultrasónicos entre otros, y que son procesados por varias redes neuronales con el fin de que sea capaz de tomar una decisión en cuestión de milisegundos (Fig. 36.3).

8. **Chatbots con inteligencia artificial.** Son aplicaciones software que, utilizando la inteligencia artificial conversacional, son capaces de conversar mediante un chat escrito como si fueran un ser humano. Caben destacar los asistentes virtuales existentes en diversas páginas web y el reciente *ChatGPT* (OpenAI, 2022), el cual es capaz de mantener conversaciones con el usuario, resolver problemas sencillos, generar textos y resúmenes sobre cualquier tema o generar código en diversos lenguajes de programación a partir de una petición realizada mediante lenguaje natural.

36.3. Redes neuronales

Las redes neuronales artificiales (en inglés Artificial Neural Network (ANN)) tienen su origen en la definición de neurona artificial de (McCulloch and Pitts, 1943) y en el diseño del perceptrón por parte de Frank Rosenblatt (Rosenblatt, 1958). Cada ANN está formada por un conjunto de elementos conocidos como “neuronas” cuya organización está inspirada en la que siguen las redes neuronales de los seres vivos. Entre dos neuronas adyacentes existe una serie de conexiones a través de las cuales se envía la información como si de pulsos eléctricos se tratase. De forma aislada, cada neurona procesa la información recibida para producir un resultado que será utilizado por las siguientes neuronas con las que está conectada.

Cada ANN tiene como objetivo resolver una tarea concreta. Por ejemplo, una ANN podría estar diseñada para reconocer un dígito o una letra a partir de una imagen. Para conseguir resolver dicha tarea, la red sigue un proceso de aprendizaje automático. Este proceso se conoce como “entrenamiento” y requiere que se disponga de un conjunto de datos representativos de la tarea a resolver.

36.4. Perceptrón o neurona

El elemento básico de toda ANN es la neurona artificial, inspirada en las neuronas biológicas. Cada neurona tiene una serie de entradas y produce una única salida. Las entradas pueden ser variables extraídas de la tarea que se debe resolver o salidas de otras neuronas de la red.

Para calcular la salida, cada neurona realiza una suma ponderada de sus entradas utilizando una serie de pesos, \mathbf{w} donde $w_i \in \mathbb{R}$, y añade un término constante, $w_0 \in \mathbb{R}$. Por tanto, cada neurona actúa como un clasificador lineal que puede separar dos conjuntos diferentes dependiendo de si la salida es positiva o negativa (Figura 36.6).

Para cada vector de entrada, \mathbf{x} , la neurona aplicará los pesos, \mathbf{w} , como el producto escalar de ambos vectores:

36.4. Perceptrón o neurona

601

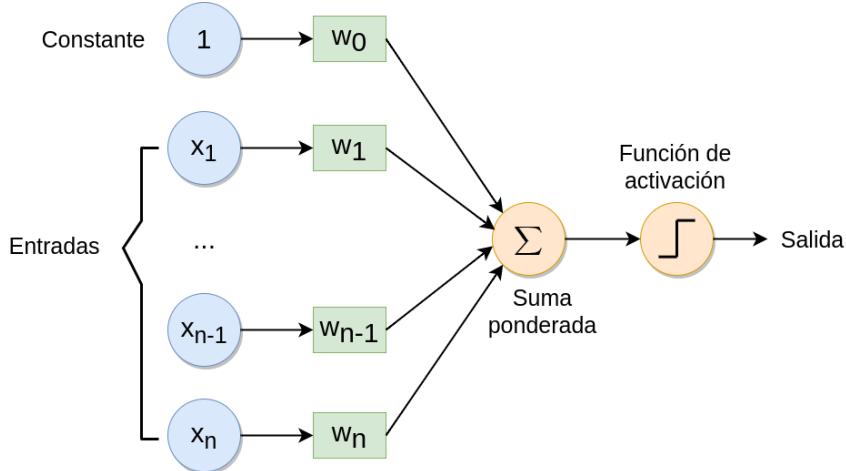


Figura 36.6: Estructura del perceptrón o neurona

$$\mathbf{w}'\mathbf{x} = w_0 \cdot 1 + w_1 \cdot x_1 + w_2 \cdot x_2 + \dots + w_n \cdot x_n. \quad (36.1)$$

Una vez obtenida la suma ponderada, típicamente se puede separar las entradas en dos conjuntos, obteniéndose como salida final un valor binario, siguiendo la fórmula:

$$f(\mathbf{w}'\mathbf{x}) = \begin{cases} 1 & \text{si } \mathbf{w}'\mathbf{x} > 0 \\ 0 & \text{en otro caso} \end{cases}. \quad (36.2)$$

36.4.1. Aprendizaje

Durante el proceso de aprendizaje, el perceptrón busca el ajuste automático de los valores de los pesos. Éstos deben seleccionarse de forma que minimicen el error de clasificación cometido sobre un conjunto de entrenamiento. El conjunto de entrenamiento estará compuesto por un conjunto de muestras del que se conoce su clase:

$$D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_m, y_m)\}, \quad (36.3)$$

donde cada muestra, $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{in})$, pertenece a una de las dos clases, $y_i = \{0, 1\}$.

El primer paso del aprendizaje o entrenamiento consiste en la inicialización de cada peso w_j a 0 o a algún otro valor aleatorio.

Tras ello, se calcula la clase estimada, \hat{y} , en un momento determinado, t , para cada muestra \mathbf{x}_i del conjunto de datos:

$$\hat{y}_i(t) = f(\mathbf{w}(t)^T \mathbf{x}_i) = f(w_0(t) + w_1(t) \cdot x_{i1} + \dots + w_n(t) \cdot x_{in}). \quad (36.4)$$

Tras obtener la salida para todas las muestras de entrenamiento, cada uno de los pesos, w_j , de la neurona se actualiza siguiendo la fórmula:

$$w_j(t+1) = w_j(t) + \lambda \cdot |y_i - \hat{y}_i(t)| \cdot x_{ij}. \quad (36.5)$$

donde $|y_i - \hat{y}_i(t)|$ será 0 cuando la clase predicha coincide con la clase real de la muestra y λ es la tasa de aprendizaje. La tasa de aprendizaje debe seleccionarse de antemano y controla la variación de los pesos entre iteraciones. En algunos casos el valor de λ es 0 o varía durante el proceso de entrenamiento.

Los dos pasos anteriores se repiten hasta que el error de clasificación es menor que un cierto umbral o el número de iteraciones alcanza un cierto valor fijado. Normalmente se suele utilizar el número de iteraciones como criterio de paro puesto que no siempre es posible alcanzar una tasa de error más baja que la deseada.

36.4.2. Convergencia

El teorema de la convergencia del perceptrón dice que, en los problemas en los que haya dos clases linealmente separables, es siempre posible encontrar unos pesos que realicen la separación en un número finito de iteraciones (Novikoff, 1962). Sin embargo, en la mayoría de los casos, no se está ante problemas linealmente separables, esto es, no es posible obtener un conjunto de variables que separen perfectamente las muestras de ambas clases. Por ello, es necesario el uso de ciertas estrategias que solucionen el problema de convergencia en estos casos. Algunas de las estrategias más utilizadas son:

- Algoritmo Pocket: Guarda la mejor solución obtenida hasta el final del entrenamiento.
- Algoritmo Maxover: Halla el margen de separación máximo permitiendo clasificaciones incorrectas.
- Algoritmo de Voto: Se utilizan múltiples perceptrones combinando sus salidas.

36.5. Perceptrón multiclas

Una extensión lógica del uso del perceptrón es su empleo en la resolución de tareas de clasificación donde existan más de dos clases (Haykin, 1999). En ese caso se tendrá un conjunto de entrenamiento, D , de m muestras:

$$D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_m, y_m)\}, \quad (36.6)$$

donde cada muestra $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{in})$ pertenezca a una de las c clases posibles:

$$y_i = \{0, 1, \dots, c-1\}. \quad (36.7)$$

A diferencia del problema binario, en su versión multiclas lo que se definen son varios modelos, F , uno para cada una de las c clases:

$$F = \{f_0, f_1, \dots, f_{c-1}\} f_j : \mathbb{R}^n \rightarrow \mathbb{R}. \quad (36.8)$$

En este caso la salida no se selecciona en función de si el valor obtenido es positivo o negativo, sino que se asigna la clase del modelo que obtenga el valor más alto tras aplicar los pesos a la muestra. Esta estrategia recibe el nombre de “uno contra todos”:

$$\hat{y}_i = \operatorname{argmax}_j(f_j(\mathbf{x}_i)) j \in \{0, 1, \dots, c - 1\}. \quad (36.9)$$

En muchas ocasiones lo que se obtiene no es un único valor con la clase asignada como salida, sino que se obtiene un vector con las salidas binarias de cada uno de los modelos empleados. En ese caso, el vector contendrá un 1 en la posición de la clase asignada y un 0 en el resto de clases. Por ejemplo, el vector [0, 1, 0, 0, 0] representaría que una muestra ha sido asignada a la segunda clase en un problema de clasificación donde existen 5 clases posibles:

$$[(f_1(\mathbf{x}_i)), (f_2(\mathbf{x}_i)), \dots, (f_c(\mathbf{x}_i))]. \quad (36.10)$$

36.6. Funciones de activación

Además de los pesos, toda neurona tiene asociada una función de activación. Esta función se encarga de transformar la suma ponderada de las entradas en el resultado final. En las secciones anteriores se ha utilizado una función de activación con umbral 0, pero existen muchas otras. Algunas de las más utilizadas se enumeran a continuación.

Para algunas de ellas, se ha implementado una función, `plot_activation_function()`, que permite dibujarlas en **R**, y que se puede ver a continuación:

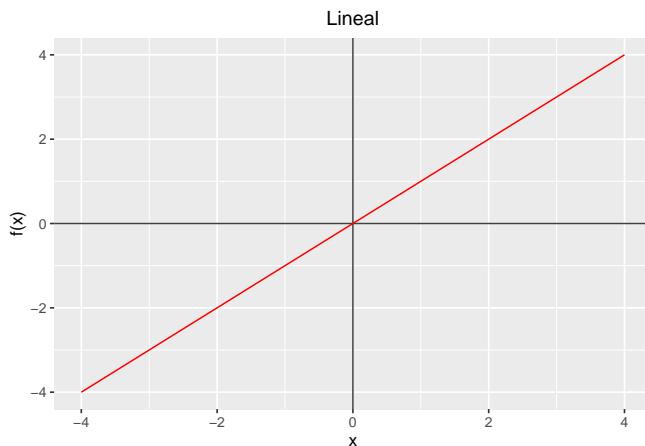
```
require(ggplot2)
plot_activation_function <- function(f, title, range){
  ggplot(data.frame(x=range), mapping=aes(x=x)) +
    geom_hline(yintercept=0, color='black', alpha=3/4) +
    geom_vline(xintercept=0, color='black', alpha=3/4) +
    stat_function(fun=f, colour = "red") +
    ggtitle(title) +
    scale_x_continuous(name='x') +
    scale_y_continuous(name='f(x)') +
    theme(plot.title = element_text(hjust = 0.5))
}
```

- **Función lineal.** Se trata de una función identidad donde la salida tiene el mismo valor que la entrada. Normalmente se aplica en problemas de regresión lineal. Por ejemplo, si se quiere predecir el número de días que lloverá en un mes determinado.

$$f(x) = x \quad (36.11)$$

Y se representa gráficamente de la siguiente forma:

```
f <- function(x){ x }
plot_activation_function(f, 'Lineal', c(-4,4))
```



- **Función umbral.** Esta función recibe también el nombre de función escalón. Si el valor de entrada es menor que el umbral la salida será 0. En caso contrario, la salida será 1. Si el umbral es 0, la función se reduce a mirar el signo del valor analizado.

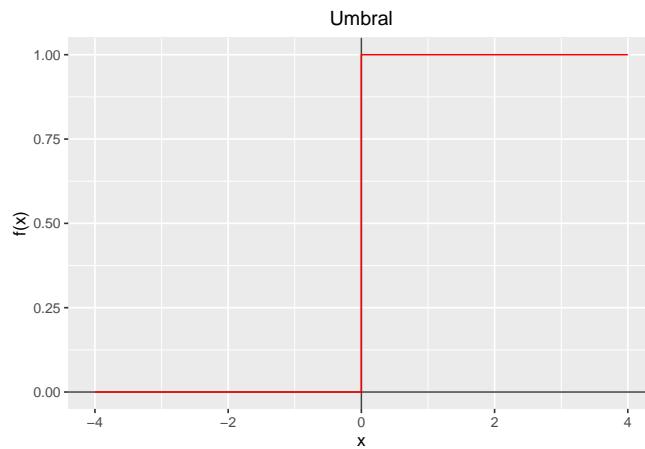
$$f(x) = \begin{cases} 0 & \text{si } x < u \\ 1 & \text{en otro caso} \end{cases} \quad (36.12)$$

Se representa gráficamente mediante el siguiente código, el cual se corresponde con una modificación de la función `plot_activation_function`, ya que la versión original no mostraría de forma correcta la gráfica al requerir representar dos valores en la posición 0, el valor 0 y el valor 1 del escalón:

```
df <- data.frame(x=c(-4, -3, -2, -1, 0, 1, 2, 3, 4), f=c(0,0,0,0,1,1,1,1,1))
ggplot(data=df, aes(x=x, y=f, group=1)) +
  theme(plot.title = element_text(hjust = 0.5)) +
  ggtitle("Umbral") +
  scale_y_continuous(name='f(x)') +
  geom_hline(yintercept=0, color='black', alpha=3/4) +
  geom_vline(xintercept=0, color='black', alpha=3/4) +
  geom_step(color='red')
```

36.6. Funciones de activación

605

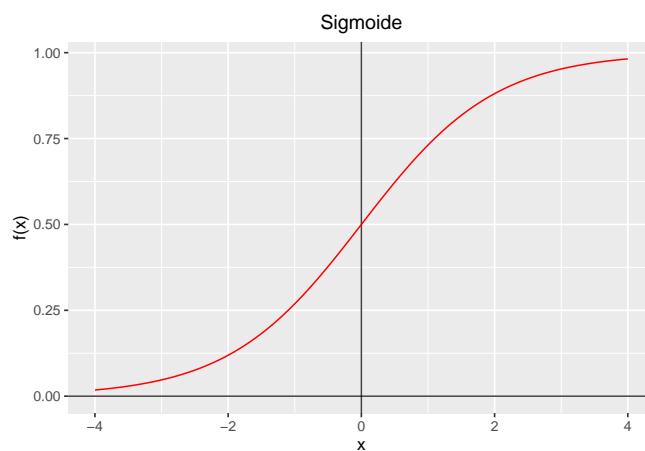


- **Función sigmoide.** También conocida como función logística, se trata de una de las funciones más utilizadas para asignar una clase. Si el punto de evaluación de la función es un valor negativo muy bajo, la función dará como resultado un valor muy cercano a 0, si se evalúa en 0, el resultado es 0,5 y si se evalúa en un valor positivo alto el resultado será aproximadamente 1.

$$f(x) = \frac{1}{1 + e^{-x}} \quad (36.13)$$

Representándose gráficamente de la siguiente forma:

```
f <- function(x){1 / (1 + exp(-x))}
plot_activation_function(f, 'Sigmoid', c(-4,4))
```

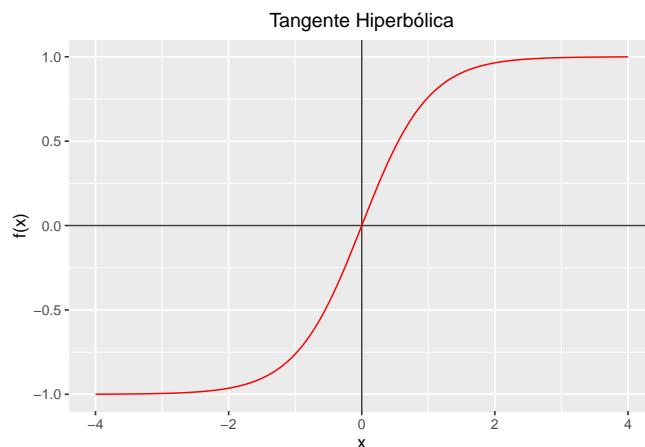


- **Función tangente hiperbólica.** El rango de valores de salida es [-1, 1], donde los valores altos tienden de manera asintótica a 1 y los valores muy bajos tienden de manera asintótica a -1 de forma similar a la sigmoid.

$$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (36.14)$$

Siendo su representación gráfica de la siguiente forma:

```
tanh_func <- function(x){tanh(x)}
plot_activation_function(tanh_func, 'Tangente Hiperbólica', c(-4,4))
```

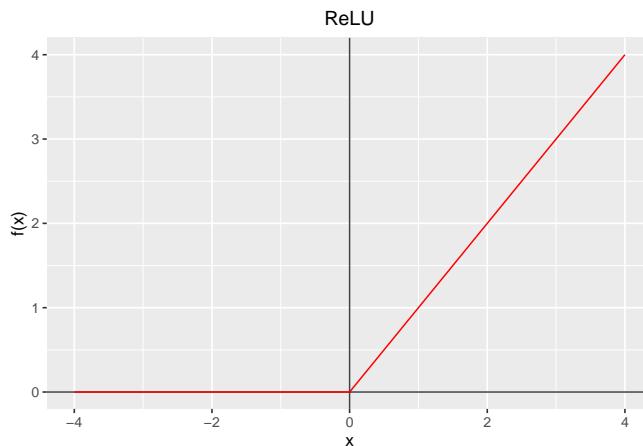


- **Función ReLU.** Se trata de la unidad lineal rectificada (del inglés Rectified Linear Unit). Es posiblemente la función de activación más utilizada actualmente en deep learning ([Nair and Hinton, 2010](#)).

$$f(x) = \begin{cases} 0 & \text{si } x \leq 0 \\ x & \text{en otro caso} \end{cases} \quad (36.15)$$

Y se representaría gráficamente de la siguiente manera:

```
rec_lu_func <- function(x){ ifelse(x < 0 , 0, x )}
plot_activation_function(rec_lu_func, 'ReLU', c(-4,4))
```



36.7. Perceptrón multicapa

Aunque el perceptrón puede aprender muchos tipos de lógica, no es posible que aprenda la operación XOR (OR exclusivo) que se diferencia del OR en que asigna un 1 a la salida cuando las dos entradas son distintas (Minsky and Papert, 1969). El perceptrón multicapa o, en inglés, Multilayer Perceptron (MLP) surge para dar una solución a este problema que es un paradigma de los problemas linealmente no separables, que realmente son la mayoría en el mundo real.

Un MLP está compuesto por varias capas con neuronas. La primera capa será la de entrada, que recibirá las variables que representan los elementos del problema a resolver. Por otro lado, la última capa representará las clases de salida (en las que hay que clasificar las entradas), esto es, la salida del MLP. Entre ambas capas existirán una o más capas “ocultas”. Las neuronas de una capa intermedia tienen como entrada la salida de la capa anterior y su salida es la entrada de las neuronas de la siguiente capa (Figura 36.7). Este tipo de capas también son llamadas *densas* o *totalmente conectadas*.

36.7.1. Aprendizaje

El MLP entra en la categoría de los algoritmos de propagación hacia adelante o *feedforward* ya que las entradas de las neuronas de una capa se combinan mediante la suma ponderada, pasan por una función de activación y el resultado es propagado a las neuronas de la capa siguiente. Este proceso se lleva a cabo desde la capa de entrada hasta la capa de salida.

Dado un conjunto de muestras de entrenamiento $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\}$ donde cada $\mathbf{x}_i \in \mathbb{R}^d$ e $y_i \in \{0, 1\}$ (siendo d el número de características), la salida de la primera capa, \mathbf{z}_1 , para una entrada \mathbf{x} vendrá dada por la expresión:

$$\mathbf{z}_1 = \mathbf{W}'_{(1)} \mathbf{x} + \mathbf{b}_1, \quad (36.16)$$

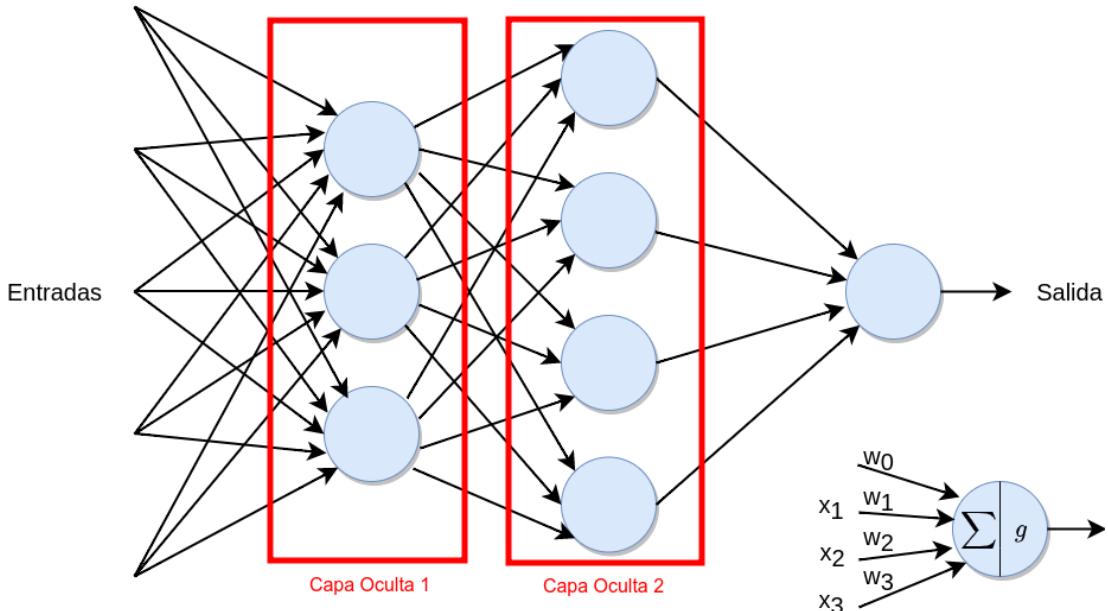


Figura 36.7: Estructura del perceptrón multicapa (MLP)

donde $\mathbf{b}_1 \in \mathbb{R}^h$ es un vector con las constantes de la primera capa, siendo h el número de variables de cada capa, y $\mathbf{W}_{(1)} \in \mathbb{R}^{h \times d}$ son los pesos de la capa. Tras aplicar la función de activación, $g(\cdot)$, al vector intermedio, $\mathbf{z} \in \mathbb{R}^h$, se obtiene:

$$\mathbf{h}_1 = g(\mathbf{z}_1). \quad (36.17)$$

La salida de una capa intermedia, $\mathbf{h}_i \in \mathbb{R}^h$, también está formada por variables intermedias que sirven de entrada a la siguiente capa. La función a calcular en la siguiente capa será por tanto:

$$\mathbf{h}_2 = g(\mathbf{W}'_{(2)} \mathbf{h}_1 + \mathbf{b}_2). \quad (36.18)$$

Siguiendo el mismo razonamiento, la salida de la última capa, \hat{y} , y por tanto de la red, vendrá dada por:

$$\hat{y} = g(\mathbf{W}'_{(n)} \mathbf{h}_{n-1} + \mathbf{b}_n). \quad (36.19)$$

Por ejemplo, si se tiene una red de tres capas la salida podrá calcularse como:

$$\hat{y} = g(\mathbf{W}'_{(3)} g(\mathbf{W}'_{(2)} g(\mathbf{W}'_{(1)} \mathbf{x} + \mathbf{b}_1) + \mathbf{b}_2) + \mathbf{b}_3). \quad (36.20)$$

Para entrenar y ajustar los pesos de este tipo de redes es necesario realizar el ajuste de la combinación de todos los pesos de la red. De forma similar a la búsqueda de los pesos de una

sola neurona, será necesario encontrar la combinación de valores que clasifiquen bien todas las muestras del conjunto de entrenamiento o, en su defecto, que fallen en el menor número de muestras posible o minimicen alguna otra función de coste. En este punto es donde entra en juego la propagación hacia atrás o *backpropagation*.

La propagación hacia atrás es el mecanismo por el que el MLP ajusta de forma iterativa los pesos de la red con el objetivo de minimizar una función de coste que mide lo bueno o malo que es el resultado obtenido en un momento determinado (Rumelhart et al., 1986). Su único requisito de aplicación es que todas las operaciones de la red (incluidas las funciones de activación) sean diferenciables ya que se utiliza el algoritmo del descenso del gradiente para optimizar la función de coste.

El MLP utiliza diferentes funciones de coste o pérdida según el tipo de problema a resolver. Para los problemas de clasificación, la función de coste más utilizada es la Entropía Cruzada Media (en inglés Average Cross-Entropy). Para un problema binario esta función de coste se calcula como;

$$C(\hat{y}, y, \mathbf{W}) = -\frac{1}{n} \sum_{i=0}^n (y_i \ln \hat{y}_i + (1 - y_i) \ln (1 - \hat{y}_i)) + \frac{\alpha}{2n} \|\mathbf{W}\|_2^2, \quad (36.21)$$

donde $\alpha \|\mathbf{W}\|_2^2$ con $\alpha > 0$ es un término de regularización, L2, también conocido como penalización ya que penaliza los modelos complejos. α es un hiperparámetro cuyo valor se establece manualmente.

Para los problemas de regresión, la función de coste se basa en el Error Cuadrático Medio (*Mean Squared Error*):

$$C(\hat{y}, y, \mathbf{W}) = \frac{1}{2n} \sum_{i=0}^n \|\hat{y}_i - y_i\|_2^2 + \frac{\alpha}{2n} \|\mathbf{W}\|_2^2. \quad (36.22)$$

Cada iteración en el proceso de aprendizaje estará compuesta entonces por dos etapas, una de propagación hacia adelante y otra de propagación hacia atrás. En la primera etapa se introducen los valores de entrada a la red y se propagan las operaciones y los resultados hasta obtener la salida final de la red. En la segunda, el gradiente de la función de coste es propagado hacia atrás para actualizar los valores de los pesos de todas las capas y acercarse más a los valores que minimizan la función de coste.

En el algoritmo del descenso del gradiente, $\nabla C_{\mathbf{W}}$ se calcula y deduce de \mathbf{W} . Formalmente esto puede expresarse como:

$$\mathbf{W}^{t+1} = \mathbf{W}' - \lambda \nabla C_{\mathbf{W}}^t, \quad (36.23)$$

donde t es el estado de la red en una iteración determinada y λ es la tasa de aprendizaje cuyo valor debe ser superior a 0.

Al igual que en el caso del perceptrón único, el entrenamiento terminará cuando se alcance un número máximo de iteraciones o la mejora en la función de coste entre dos iteraciones consecutivas no supere cierto umbral.

Durante el proceso de aprendizaje, es necesario guardar en memoria los resultados de cada una de las muestras del conjunto de entrenamiento. Si el número de muestras o el tamaño de la red son grandes, es posible que no se disponga del suficiente espacio. Para resolver este problema, en una iteración no se utiliza todo el conjunto de entrenamiento, sino que se utiliza un subconjunto del mismo llamado *batch*. El conjunto de entrenamiento se divide en cada iteración, por tanto, en un número de *batches* disjuntos con un número de muestras por *batch*. Atendiendo a esta división, es posible definir una serie de hiperparámetros:

- Tamaño del *batch*. Número de muestras utilizadas en cada iteración para actualizar los pesos.
- Número de épocas. Número de pasadas completas sobre el conjunto de entrenamiento hasta terminar el proceso de aprendizaje.
- Número de iteraciones por época. Será el resultado de dividir el número total de muestras por el tamaño del *batch*.

Por ejemplo, si se tiene un conjunto de 55000 muestras y el tamaño del *batch* es de 100, cada época tendrá 550 iteraciones.

36.8. Instalación de librerías de *deep learning* en R: Tensorflow/Keras

El framework que se va a utilizar en este libro para trabajar con técnicas de *deep learning* será Tensorflow/Keras, debido a que es uno de los más completos en la actualidad, permitiendo realizar una configuración completa del proceso de entrenamiento y trabajar con diversos tipos de redes neuronales.

Para poder utilizar Tensorflow/Keras en **R**, es necesario realizar la instalación de la librería fuera de **R**. Por ello, si ya se dispone de una instalación del mismo sería posible utilizarla. No obstante, se recomienda seguir los pasos indicados a continuación para tener una instalación nativa de Tensorflow/Keras asociada directamente a R.

- **Paso 1** - Librería de Tensorflow en **R**

El primer paso será instalar el paquete de **tensorflow** en **R** [].

```
install.packages("tensorflow")
```

A continuación, será necesario tener una instalación de Conda en el sistema. Los usuarios tanto de Windows como de Linux/Mac podrán realizar directamente la instalación de una versión de Conda denominada Mini-Conda en el instalador del siguiente paso, la cual sería la opción recomendada para no tener que realizar una instalación externa de manera adicional.

NOTA

Otra manera disponible para los usuarios de Windows, pero no recomendada por los autores de este libro salvo que ya se disponga de Anaconda instalado, sería la de utilizar el programa y la librería directamente dentro de Anaconda, instalando una versión de R directamente en el sistema a través del siguiente link:

<https://docs.anaconda.com/anaconda/install/windows/>

- **Paso 2 - Instalación de tensorflow y keras**

Para continuar la instalación se activará la librería de Tensorflow y se ejecutará la función *install_tensorflow*

```
library(tensorflow)
install_tensorflow()
```

Al ejecutar esta función, los usuarios deberán marcar “Y” para aceptar la instalación de Mini-Conda, descartando aceptar la utilización de cualquier otro sistema Conda que pueda estar instalado previamente.

También se puede ejecutar la función *install_keras* del paquete **keras** para instalar Tensorflow.

```
install.packages("keras")
library(keras)
install_keras()
```

- **Paso 3 - Confirmar la instalación**

Para confirmar la instalación, se puede comprobar con los siguientes comandos (la salida puede variar según el equipo, pero la línea final tiene que ser similar a la indicada):

```
library(tensorflow)
tf$constant("Hello Tensorflow")  
  
tf.Tensor(b'Hello Tensorflow', shape=(), dtype=string)
```

36.9. Ejemplo de red para clasificación en R

En esta sección se entrena una red neuronal artificial para reconocer o clasificar los dígitos manuscritos del conjunto de datos MNIST (https://en.wikipedia.org/wiki/MNIST_database). Cada una de las imágenes de este conjunto de datos tiene un tamaño de 28×28 píxeles en escala de grises. En vez de extraer una serie de variables a partir de cada imagen, en este caso se utilizan cada uno de los $28 \times 28 = 784$ píxeles como variable de entrada (Figura 36.8).

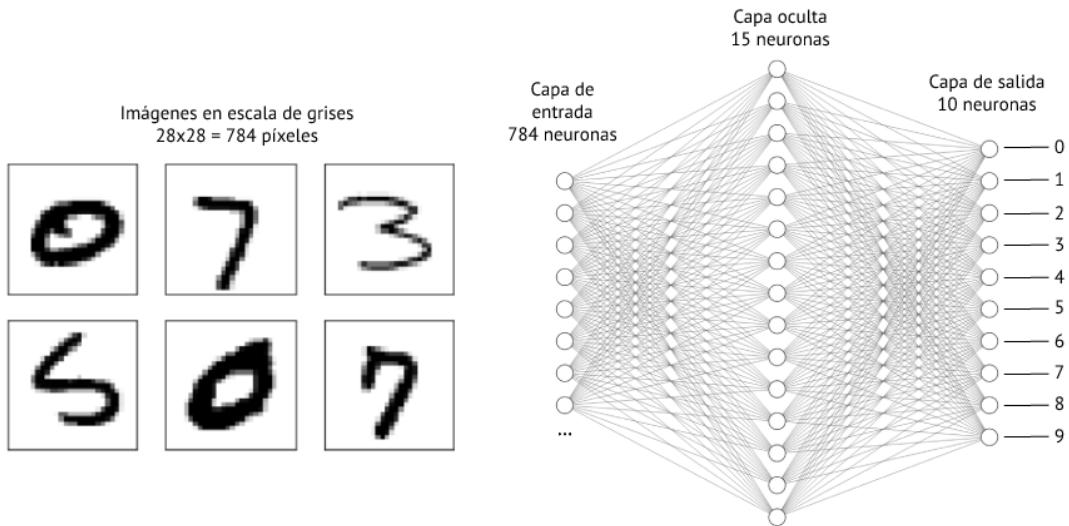


Figura 36.8: MLP para reconocimiento de dígitos manuscritos

36.9.1. Carga y visualización de los datos

El primer paso será cargar la librería **keras** que permite crear redes neuronales y conjunto de imágenes que se encuentra disponible públicamente:

```
library(keras)
mnist <- dataset_mnist()
```

A continuación, se puede ver el contenido de las variables generadas, donde cabe destacar que el conjunto de datos MNIST ya viene separado en dos subconjuntos, uno para entrenamiento y otro para test, compuestos por 60000 y 10000 imágenes respectivamente. En ambos casos, estos datos se almacenan en la variable de nombre *x*.

```
names(mnist)
#> [1] "train" "test"
dim(mnist$train$x)
#> [1] 60000    28    28
dim(mnist$train$y)
#> [1] 60000
dim(mnist$test$x)
#> [1] 10000    28    28
dim(mnist$test$y)
#> [1] 10000
```

Además, las imágenes de cada subconjunto vienen acompañadas de la clase a la que pertenecen

36.9. Ejemplo de red para clasificación en **R**

613

(dígito contenido en la imagen). En ambos casos, esta etiqueta se almacena en la variable con nombre *y*. A continuación se muestra un pequeño ejemplo que permitirá visualizar alguna de las imágenes contenidas en el conjunto de datos de entrenamiento junto con la etiqueta representando el dígito contenido:

```
par(mfcol=c(4, 4))
par(mar=c(0, 0, 3, 0), xaxs='i', yaxs='i')
for (j in 1:16) {
  im <- mnist$train$x[j, , ]
  im <- t(apply(im, 2, rev))
  image(x=1:28, y=1:28, z=im, col=gray((0:255)/255),
        xaxt='n', main=paste(mnist$train$y[j]))
}
```

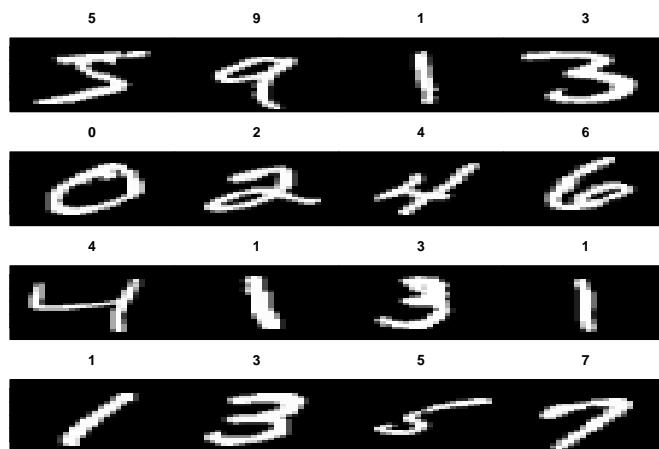


Figura 36.9: Algunas imágenes del conjunto de entrenamiento

36.9.2. Preprocesamiento

Una vez cargados los datos y comprobado su contenido, es posible realizar algún tipo de preprocesado. Dependiendo del tipo de problema se podrán realizar unas operaciones u otras. Por ejemplo, cuando se trabaja con imágenes es muy típico estandarizar los valores de color de las imágenes para mitigar las diferencias producidas por las diferentes condiciones de iluminación.

En este caso, solo se va a transformar los valores originales de la imagen (en rango de 0 a 255) a valores entre 0 y 1 dividiendo cada valor por el máximo, 255:

```
mnist$train$x <- mnist$train$x/255
mnist$test$x <- mnist$test$x/255
```

36.9.3. Generación de la red neuronal

El siguiente paso consiste en la generación de la red neuronal. Para ello, se define primero la estructura utilizando la interfaz *sequential* proporcionada por Tensorflow/Keras a través de la función *keras_model_sequential*:

```
model <- keras_model_sequential() |>
  layer_flatten(input_shape = c(28, 28)) |>
  layer_dense(units = 15, activation = "relu") |>
  layer_dense(10, activation = "softmax")
```

Como se puede observar, la red definida está compuesta por una capa de tipo *flatten* que se encarga de transformar los 28x28 valores a un vector de 784 elementos, para que a continuación una capa oculta *dense* de 15 neuronas con activación *relu* se encargue de realizar las primeras operaciones con esos datos. Al final, una última capa *dense* se encarga de obtener la probabilidad de que la imagen represente cada una de las posibles clases mediante una activación *softmax*¹:

```
summary(model, line_length=64)
```

```
#> Model: "sequential"
#>
#>   Layer (type)        Output Shape       Param #
#>   -----
#>   flatten (Flatten)    (None, 784)          0
#>   dense_1 (Dense)      (None, 15)           11775
#>   dense (Dense)        (None, 10)           160
#>   -----
#>   Total params: 11,935
#>   Trainable params: 11,935
#>   Non-trainable params: 0
#>   -----
```

Finalmente, es necesario compilar el modelo, indicando algunos de los parámetros de configuración necesarios para el proceso de entrenamiento, como la función de coste o pérdida, el optimizador a utilizar y las métricas a obtener:

```
model |>
  compile(
    loss = "sparse_categorical_crossentropy", # función utilizada para problemas de
    ↪   clasificación con varias clases
    optimizer = "sgd", # stochastic gradient descent
    metrics = "accuracy" # Precisión
  )
```

¹La activación *softmax* convierte un vector de números reales en una distribución de probabilidad de tal manera que la probabilidad de pertenecer a cada una de las categorías de salida siempre sume el 100 %.

36.9.4. Entrenamiento

Una vez generada la estructura de la red neuronal y definida la anterior configuración, es posible entrenarla mediante la función *fit()*. Para ello, se le debe indicar el conjunto de imágenes de entrenamiento, *x*, que debe utilizar y sus clases correspondientes, *y*. Además de otros parámetros, se podrá configurar el número de épocas, *epochs*, a entrenar (pasadas sobre el conjunto completo de entrenamiento), el tamaño del *batch* que se utilizará en cada iteración con *batch_size* (número de imágenes por iteración), qué porcentaje de elementos del conjunto de datos se utilizarán para validar el modelo con *validation_split* (imágenes utilizadas durante el entrenamiento pero solo para obtener una estimación real del error cometido) o la tasa de aprendizaje, *learning_rate*.

```
training_evolution <- model |>
  fit(
    x = mnist$train$x, y = mnist$train$y,
    epochs = 10, batch_size = 128,
    validation_split = 0.2,
    learning_rate = 0.1,
    verbose = 2
  )

#> Epoch 1/10
#> 375/375 - 2s - loss: 1.6313 - accuracy: 0.5266 - val_loss: 1.0455 - val_accuracy:
#>   0.7510 - 2s/epoch - 6ms/step
#> Epoch 2/10
#> 375/375 - 1s - loss: 0.8433 - accuracy: 0.7881 - val_loss: 0.6409 - val_accuracy:
#>   0.8434 - 1s/epoch - 3ms/step
#> Epoch 3/10
#> 375/375 - 1s - loss: 0.6022 - accuracy: 0.8427 - val_loss: 0.5031 - val_accuracy:
#>   0.8712 - 1s/epoch - 3ms/step
#> Epoch 4/10
#> 375/375 - 1s - loss: 0.5047 - accuracy: 0.8656 - val_loss: 0.4381 - val_accuracy:
#>   0.8830 - 1s/epoch - 3ms/step
#> Epoch 5/10
#> 375/375 - 1s - loss: 0.4526 - accuracy: 0.8767 - val_loss: 0.4019 - val_accuracy:
#>   0.8909 - 1s/epoch - 3ms/step
#> Epoch 6/10
#> 375/375 - 1s - loss: 0.4201 - accuracy: 0.8854 - val_loss: 0.3764 - val_accuracy:
#>   0.8959 - 1s/epoch - 3ms/step
#> Epoch 7/10
#> 375/375 - 1s - loss: 0.3976 - accuracy: 0.8896 - val_loss: 0.3593 - val_accuracy:
#>   0.8996 - 1s/epoch - 3ms/step
#> Epoch 8/10
#> 375/375 - 1s - loss: 0.3809 - accuracy: 0.8939 - val_loss: 0.3463 - val_accuracy:
#>   0.9022 - 1s/epoch - 3ms/step
#> Epoch 9/10
#> 375/375 - 1s - loss: 0.3678 - accuracy: 0.8975 - val_loss: 0.3359 - val_accuracy:
#>   0.9050 - 1s/epoch - 3ms/step
#> Epoch 10/10
```

```
#> 375/375 - 1s - loss: 0.3571 - accuracy: 0.8997 - val_loss: 0.3289 - val_accuracy:
→ 0.9064 - 1s/epoch - 3ms/step
```

Tras el entrenamiento es posible ver su evolución mediante las gráficas de coste/pérdida y precisión:

```
plot(training_evolution)
```

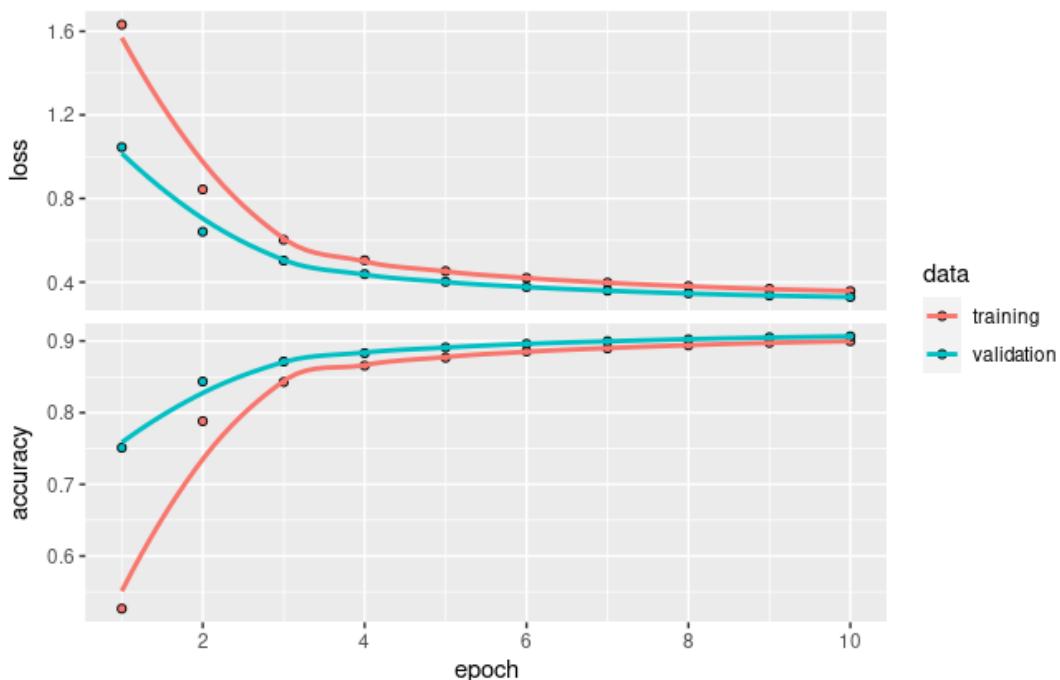


Figura 36.10: Evolución durante el entrenamiento de la función de precisión y de coste/pérdida de los conjuntos de entrenamiento y validación

Como se puede observar, la red entrenada tiene alrededor de un 90 % de precisión (porcentaje de aciertos al clasificar las imágenes) para las imágenes en los conjuntos de entrenamiento y validación. En el caso de la función de pérdida o coste, que mide el error cometido al realizar la clasificación, podemos ver como se reduce conforme la precisión del modelo aumenta.

36.9.5. Test

Una vez entrenado el modelo, es posible aplicarlo sobre el conjunto de test. Para ello, se puede realizar la predicción sobre cualquiera de las imágenes mediante la función *predict*, obteniendo la probabilidad de que pertenezca a una determinada clase:

36.9. Ejemplo de red para clasificación en **R**

617

```
predictions <- predict(model, mnist$test$x)
head(round(predictions, digits=3), 5)

#>      [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9]  [,10]
#> [1,] 0.000 0.000 0.000 0.003 0.000 0.000 0.000 0.995 0.000 0.002
#> [2,] 0.009 0.000 0.836 0.024 0.000 0.009 0.119 0.000 0.003 0.000
#> [3,] 0.000 0.962 0.013 0.006 0.001 0.001 0.003 0.002 0.010 0.002
#> [4,] 0.999 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
#> [5,] 0.001 0.000 0.007 0.000 0.836 0.004 0.011 0.012 0.017 0.111
```

También se puede utilizar la función *evaluate* para calcular tanto el coste o pérdida como la precisión de la red neuronal sobre el conjunto de test. Como se puede observar, se obtienen valores muy similares a los obtenidos durante el entrenamiento:

```
model |>
  evaluate(mnist$test$x, mnist$test$y, verbose = 0)

#> loss accuracy
#> 0.3310305 0.9045000
```

Con la función *predict* se puede también generar la matriz de confusión de la red para evaluar aciertos y fallos para cada clase:

```
prediction_matrix <- model |> predict(mnist$test$x) |> k_argmax()
confusion_matrix <- table(as.array(prediction_matrix), mnist$test$y)
confusion_matrix

#>
#>      0   1   2   3   4   5   6   7   8   9
#> 0 953  0 11  4  2 16 16  3  8  7
#> 1  0 1108 10  2  6  1  3 21 10  5
#> 2  4  3 901 27  5 11 14 27 13  6
#> 3  2  2 16 903  0 46  1  4 29 10
#> 4  1  0 16  0 899 16 12  9 11 43
#> 5  6  1  1 29  1 726  8  1 24 13
#> 6  9  4 19  3 10 21 902  0 10  0
#> 7  2  2 12 17  2 10  0 916 11 18
#> 8  3 15 35 20 10 38  2  3 839  9
#> 9  0  0 11  5 47  7  0 44 19 898
```

En la diagonal principal podemos observar el número de aciertos que obtiene el modelo entrenado para el conjunto de test, mientras que el resto de valores indican en cuantas ocasiones una clase es clasificada de manera incorrecta como otra diferente. A partir de esta matriz de confusión se puede calcular el valor de **accuracy** calculado mediante la función **evaluate** previa.

36.9.6. Guardado y reutilización del modelo

Finalmente, es posible almacenar el modelo entrenado mediante la función `save_model_tf`, que genera una carpeta con la red que se puede cargar y reutilizar mediante la función `load_model_tf`.

```
save_model_tf(object = model, filepath = "model")
reloaded_model <- load_model_tf("model")
round(predict(reloaded_model, mnist$test$x[1,1:28,1:28]), digits=4)

#>      [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9]  [,10]
#> [1,] 2e-04   0 1e-04 0.0028   0 1e-04   0 0.9948   0 0.002
```

36.10. Ejemplo de red para regresión en R

En esta sección se entrena una red neuronal artificial para predecir el precio de la vivienda según sus características en Madrid. Para ello se usará el dataset de `Madrid_Sale` disponibles en el paquete de *R* **Idealista18**, con datos inmobiliarios del año 2018 y que fue utilizado en el Cap. 9. Para ello, se tomarán las siguientes 7 variables que se usarán para realizar la estimación:

- *CONSTRUCTEDAREA*: metros cuadrados construidos.
- *ROOMNUMBER*: número de habitaciones.
- *BATHNUMBER*: número de baños.
- *HASLIFT*: si tiene ascensor.
- *DISTANCE_TO_CITY_CENTER*: distancia al centro de la ciudad.
- *DISTANCE_TO_METRO*: distancia a la parada de metro más cercana.
- *DISTANCE_TO_CASTELLANA*: distancia a la Castellana.

36.10.1. Carga y visualización de los datos

Considerando que ya se ha cargado previamente la librería `keras`, se carga el conjunto de datos indicando las variables a considerar:

```
library(idealista18)
data("Madrid_Sale")

variables <- c("CONSTRUCTEDAREA", "ROOMNUMBER", "BATHNUMBER",
              "HASLIFT", "DISTANCE_TO_CITY_CENTER", "DISTANCE_TO_METRO",
              "DISTANCE_TO_CASTELLANA")
x_madrid <- Madrid_Sale[variables]
x_madrid_mat <- unname(data.matrix(x_madrid))
y_madrid <- Madrid_Sale$PRICE
y_madrid_mat <- matrix(y_madrid, nrow = length(y_madrid), byrow = TRUE)
```

36.10. Ejemplo de red para regresión en **R**

619

El conjunto de datos contiene un total de 94815 elementos, que se dividirán en un 90 % para entrenamiento y un 10 % para test:

```
ind <- sample(c(TRUE, FALSE), length(y_madrid), replace=TRUE, prob=c(0.9, 0.1))
madrid_dat_train_x <- x_madrid_mat[ind, ]
madrid_dat_test_x <- x_madrid_mat[!ind, ]
madrid_dat_train_y <- y_madrid_mat[ind, ]
madrid_dat_test_y <- y_madrid_mat[!ind, ]
```

36.10.2. Preprocesamiento

Una vez cargados los datos y comprobado su contenido, es recomendable la normalización de las variables contenidas en el conjunto de datos debido a su heterogeneidad. Aunque sería posible para la red neuronal el adaptarse a esta situación, ciertamente puede complicar el proceso de entrenamiento haciéndola más imprecisa. Para ello, se utilizará la función `scale()` sobre las variables predictoras y se dividirá la variable del precio entre 100000 para reducir su escala:

```
madrid_dat_train_x <- scale(madrid_dat_train_x)
madrid_dat_test_x <- scale(madrid_dat_test_x)
madrid_dat_train_y <- madrid_dat_train_y/100000
madrid_dat_test_y <- madrid_dat_test_y/100000
```

36.10.3. Generación de la red neuronal

El siguiente paso consiste en la generación de la red neuronal. Para ello, al igual que en la sección 36.9.3, se define primero la estructura utilizando la interfaz *sequential* proporcionada por Tensorflow/Keras a través de la función `keras_model_sequential()`:

```
model <- keras_model_sequential() |>
  layer_dense(units=128, activation="relu", input_shape=7) |>
  layer_dense(units=64, activation="relu") |>
  layer_dense(units=16, activation="relu") |>
  layer_dense(units=1)
```

Como se puede observar, la red está compuesta por varias capas ocultas tipo *dense*, en las que las tres primeras tienen una activación *relu*. Al final, una última capa *dense* se encarga de obtener el valor de la estimación y, al contrario que en el ejemplo previo, no incluye ningún tipo de función de activación debido a que el valor de la misma ya es comprensible tanto para el modelo como para su interpretación. Esto sería equivalente a utilizar la función de activación lineal.

```
summary(model, line_length=64)
```

```
#> Model: "sequential_1"
#>
#>   Layer (type)        Output Shape       Param #
#>   =====
#>   dense_5 (Dense)     (None, 128)        1024
#>   dense_4 (Dense)     (None, 64)         8256
#>   dense_3 (Dense)     (None, 16)         1040
#>   dense_2 (Dense)     (None, 1)          17
#>   =====
#> Total params: 10,337
#> Trainable params: 10,337
#> Non-trainable params: 0
#> -----
```

Finalmente, se compila el modelo indicando los parámetros de configuración necesarios para el proceso de entrenamiento. En este caso la función de coste o pérdida se corresponderá con el Error Medio Cuadrático y la métrica con el error medio absoluto:

```
model |>
  compile(
    loss = "mse", # mean squared error
    optimizer = "sgd", # stochastic gradient descent
    metrics = "mae" # mean average error
  )
```

36.10.4. Entrenamiento

Una vez generada la estructura de la red neuronal y definida la anterior configuración, se entrena la misma utilizando la función `fit()`, configurando el resto de parámetros de forma similar a como se vio en la sección 36.9.4:

```
training_evolution <- model |>
  fit(
    x = madrid_dat_train_x, y = madrid_dat_train_y,
    epochs = 50, batch_size = 512,
    validation_split = 0.2,
    learning_rate = 0.1,
    verbose = 2
  )
```

Tras el entrenamiento es posible ver su evolución mediante las gráficas de coste/pérdida y error:

```
plot(training_evolution)
```

36.10. Ejemplo de red para regresión en **R**

621

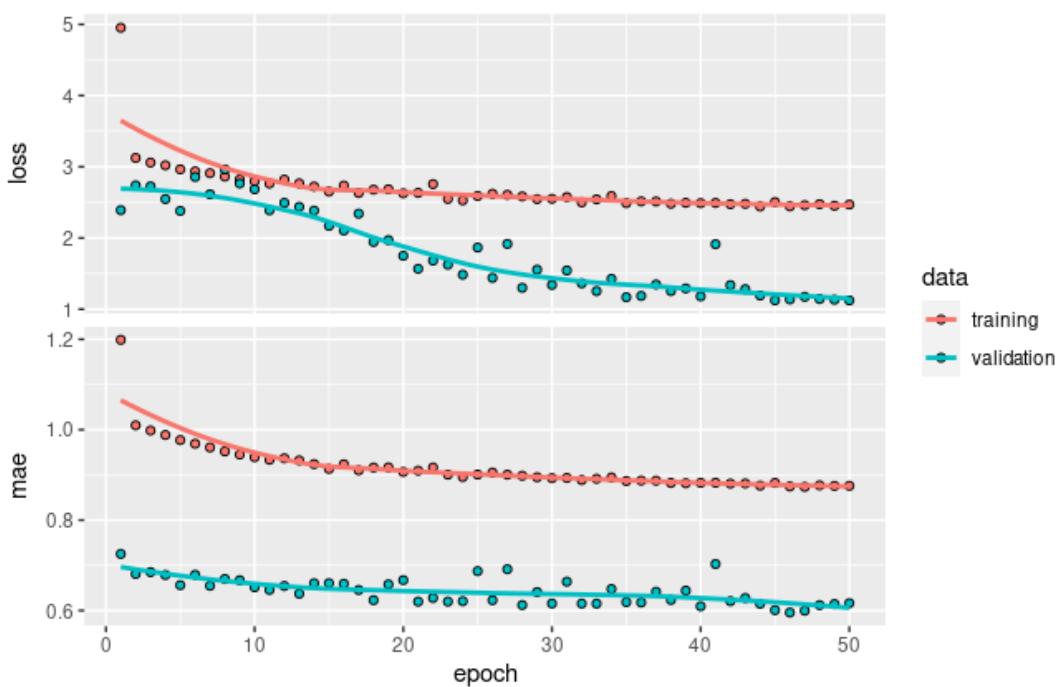


Figura 36.11: Evolución durante el entrenamiento de la precisión y la pérdida de los conjuntos de entrenamiento y validación

Como se puede observar, en este caso el modelo tiene aún posibilidad de mejora, ya que la pérdida sigue siendo alta y no se ha estancado, por lo que incrementando el número de épocas y el tiempo de entrenamiento se podría obtener un mejor resultado.

36.10.5. Test

Una vez entrenado el modelo, es posible aplicarlo sobre el conjunto de test mediante la función `predict()`, obteniendo la estimación para cada una de las viviendas:

```
predictions <- predict(model, madrid_dat_test_x)
head(predictions, 5)
#> [1]
#> [1,] 6.669374
#> [2,] 5.895504
#> [3,] 3.887646
#> [4,] 6.390513
#> [5,] 5.721725
```

Y mediante la función `evaluate()` se calcula tanto el coste o pérdida como el error de la red neuronal sobre el conjunto de test, el cual tendremos que multiplicar por 100000 para obtener el resultado en la escala original del conjunto de datos:

```
model |>
  evaluate(madrid_dat_test_x, madrid_dat_test_y, verbose = 0)
#> loss mae
#> 2.4195166 0.9227165
```

Resumen

- En este capítulo se ha explicado en detalle el concepto de redes neuronales artificiales, incluyendo los elementos que la componen, desde el perceptrón o neurona básica hasta el perceptrón multicapa, pasando el perceptron multiclase, junto al proceso de aprendizaje de los mismos.
- Además, se han definido las funciones de activación clásicas utilizadas en las redes neuronales artificiales, las cuales se encargan de transformar la suma ponderada de las entradas en el resultado final de la capa.
- Finalmente, se han explicado los pasos necesarios para poder entrenar una red neuronal artificial utilizando la librería Tensorflow/Keras en **R**, resolviendo el problema de clasificación de dígitos manuscritos representado en el conjunto de datos MNIST y un problema de regresión para estimar el precio de viviendas según sus características representado en el conjunto de datos de Idealista18.

Capítulo 37

Redes neuronales convolucionales

Noelia Vállez Enano^a y José Luis Espinosa Aranda^a

^aUniversidad de Castilla-La Mancha

37.1. Introducción

Las redes neuronales convolucionales (en inglés *Convolutional Neural Network*, CNN) son una extensión de las redes neuronales artificiales en las que se incluyen capas convolucionales para aprender a extraer, de forma automática, las características de los datos de entrenamiento al inicio de la arquitectura (Fig. 37.1).

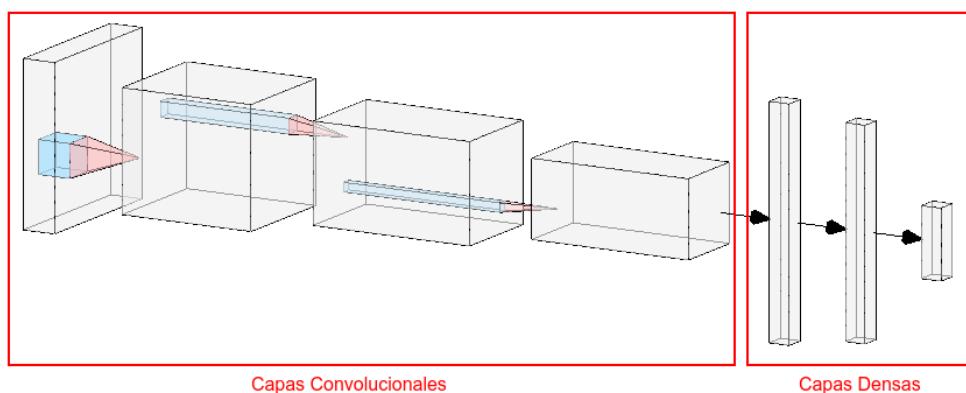


Figura 37.1: Estructura general de una CNN

Las primeras capas convolucionales de la red aprenden a extraer características generales de los datos de entrada mientras que las últimas capas convolucionales extraen características

mucho más específicas. Cuanto más larga es la red (o más *profunda*) mayor cantidad de detalles podrá aprender a distinguir. Esto es lo que ha propiciado la aparición del término “aprendizaje profundo” ([Goodfellow et al., 2016](#)).

Tras las capas convolucionales suelen encontrarse las capas *densas* o *totalmente conectadas* de la misma tipología de las vistas en el Cap. 36. Esta parte de la red será la encargada de realizar la clasificación de las muestras según los valores de las características extraídas en la parte convolucional. Por tanto, se dice que este tipo de redes tiene dos partes: una parte de extracción de características (realizada por la red convolucional) y una parte de clasificación o regresión (como las vistas en el Cap. 36).

37.2. Convolución

Aunque las redes neuronales artificiales (ANN) pueden utilizarse con los valores de color de una imagen como variables para reconocer qué hay en ella (ver Cap. 36), no es posible extraer información espacial de esta forma. Para lidiar con este problema, las CNN incorporan capas convolucionales para extraer características de las muestras de entrada, incluyendo información de la estructura espacial ([LeCun et al., 1995](#)).

Las convoluciones realizan una tarea similar al sistema visual humano, de hecho, se inspiran en cómo el ser humano percibe y procesa las características de los objetos. Aunque se diseñaron principalmente para ayudar a resolver tareas de visión por computador donde la entrada de la red es una imagen, es posible utilizarlas también con entradas vectoriales o series temporales.

Una convolución aplica un filtro sobre la entrada siguiendo un proceso de ventana deslizante. El filtro (o *kernel*) no es otra cosa que una matriz con unos pesos que se centrará en cada uno de los valores de la entrada para realizar una media ponderada de los valores de la entrada por los valores del filtro ([Garcia et al., 2015](#)). El tamaño de los filtros suele ser, por tanto, impar.

La Figura 37.2 muestra el resultado de aplicar la operación de convolución con un filtro de tamaño 3x3 sobre una matriz de entrada de tamaño 5x5. La salida será una matriz, M , de tamaño 3x3, donde cada elemento será la suma ponderada de multiplicar los elementos del filtro centrado en esa posición de la entrada por los valores de la entrada.

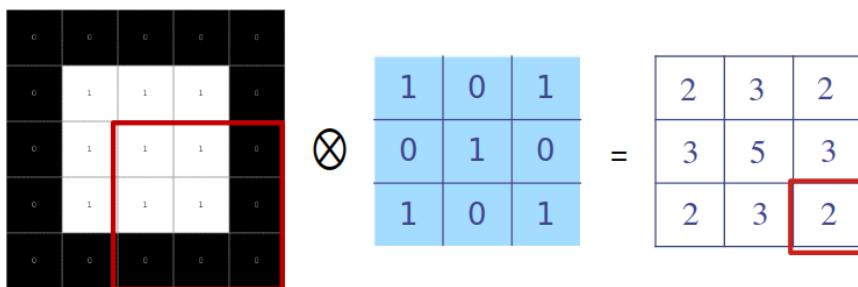


Figura 37.2: Ejemplo de convolución. De izquierda a derecha: entrada (negro = 0, blanco = 1), filtro y salida.

37.2. Convolución

625

En general, una convolución en dos dimensiones se define como:

$$\mathbf{M}[x, y] = \sum_{s=-a}^a \sum_{t=-b}^b \mathbf{F}[s, t] \mathbf{I}[x - s, y - t], \quad (37.1)$$

donde \mathbf{F} es el filtro a aplicar, \mathbf{I} es la matriz de entrada, \mathbf{M} es la matriz de resultado que recibe también el nombre de “mapa de características” y a y b son los desplazamientos desde el centro del filtro a cualquier otro valor.

Por tanto, cada valor del ejemplo de la Figura 37.2 se obtiene como:

$$\begin{aligned} M_{1,1} &= 1 \cdot 0 + 0 \cdot 0 + 1 \cdot 0 + 0 \cdot 0 + 1 \cdot 1 + 0 \cdot 1 + 1 \cdot 0 + 0 \cdot 1 + 1 \cdot 1 = 2 \\ M_{1,2} &= 1 \cdot 0 + 0 \cdot 0 + 1 \cdot 0 + 0 \cdot 1 + 1 \cdot 1 + 0 \cdot 1 + 1 \cdot 1 + 0 \cdot 1 + 1 \cdot 1 = 3 \\ &\dots \\ M_{2,2} &= 1 \cdot 1 + 0 \cdot 1 + 1 \cdot 1 = 5 \\ &\dots \\ M_{3,3} &= 1 \cdot 1 + 0 \cdot 1 + 1 \cdot 0 + 0 \cdot 1 + 1 \cdot 1 + 0 \cdot 0 + 1 \cdot 0 + 0 \cdot 0 + 1 \cdot 0 = 2 \end{aligned} \quad (37.2)$$

La elección de los valores del filtro obtendrá matrices de salida que realcen o suavicen ciertas partes de la entrada. Por ejemplo, si la entrada es una imagen, es posible definir filtros que realcen los bordes, que los suavicen o incluso que detecten dichos bordes y cómo de marcados están. La Figura 37.3 muestra el resultado de aplicar distintos filtros a una imagen de entrada.

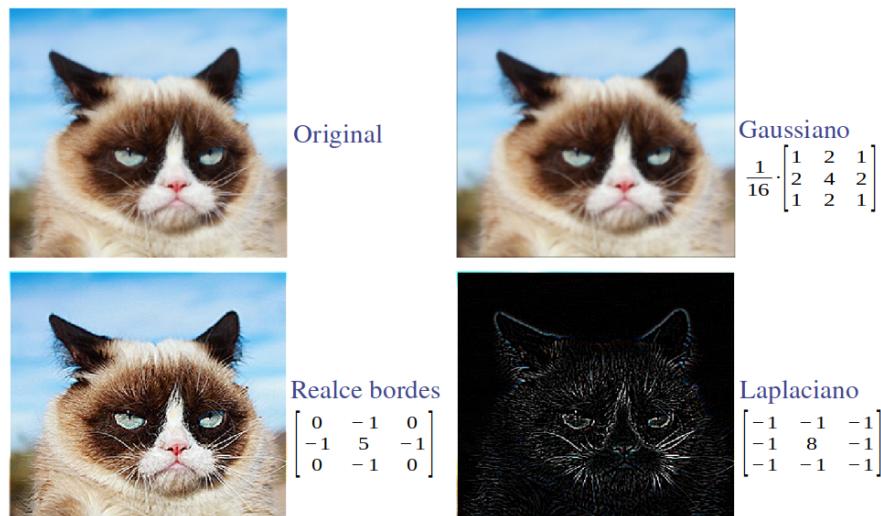


Figura 37.3: Resultado de aplicar diferentes filtros de convolución sobre una imagen dada.

Los valores (o pesos) de los filtros, que se ajustaban tradicionalmente, en un inicio, de forma manual según el problema a resolver. En los frameworks actuales estos filtros se ajustan durante

el proceso de entrenamiento de la CNN junto al resto de pesos de la red. Esto permite encontrar valores que maximicen la precisión final de la red.

37.3. Neuronas convolucionales

Las capas convolucionales de la CNN no estarán compuestas por perceptrones, sino por neuronas convolucionales que realizan las operaciones comentadas. Estas neuronas cuentan con matrices de pesos y no con vectores de pesos como lo hace el perceptrón. En este caso, tanto la entrada como la salida de la neurona son matrices. Para una neurona j , la salida \mathbf{Y}_j se calcula como la combinación lineal de las salidas de las neuronas de la capa anterior \mathbf{Y}_i operando cada una de ellas con el filtro \mathbf{F}_{ij} correspondiente a esa conexión de forma que:

$$\mathbf{Y}_j = g(\mathbf{B}_j + \sum \mathbf{F}_{ij} \otimes +\mathbf{Y}_i), \quad (37.3)$$

donde \mathbf{B}_j y g representan el *bias* y la función de activación respectivamente. La mayoría de las CNN utilizan la ReLU como activación o alguna variante de ésta. Esta activación funciona muy bien con el método del descenso del gradiente utilizado para encontrar los pesos.

Cada neurona dará lugar a un *mapa de activaciones*. La salida de una capa convolucional será entonces un conjunto de estos mapas (Fig. 37.4)

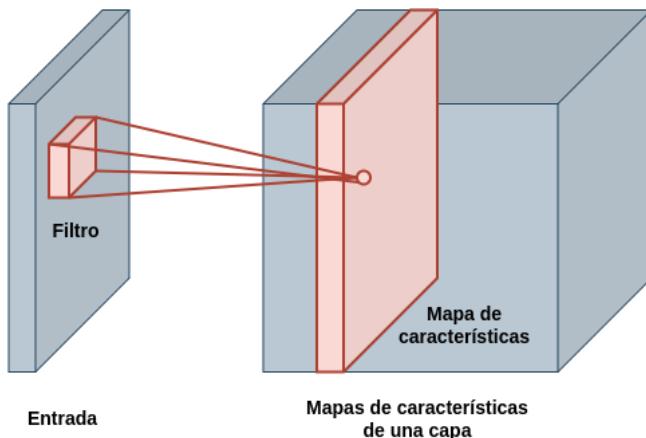


Figura 37.4: Conjunto de mapas de activaciones de una determinada capa. Cada filtro de la capa da lugar a un mapa diferente.

En el caso de que la entrada no sea una matriz 2D sino que sea una matriz 3D como, por ejemplo, una imagen, los filtros contarán con una tercera dimensión. La Figura 37.5 muestra algunos de los rellenos más empleados.

37.4. Relleno del borde

627

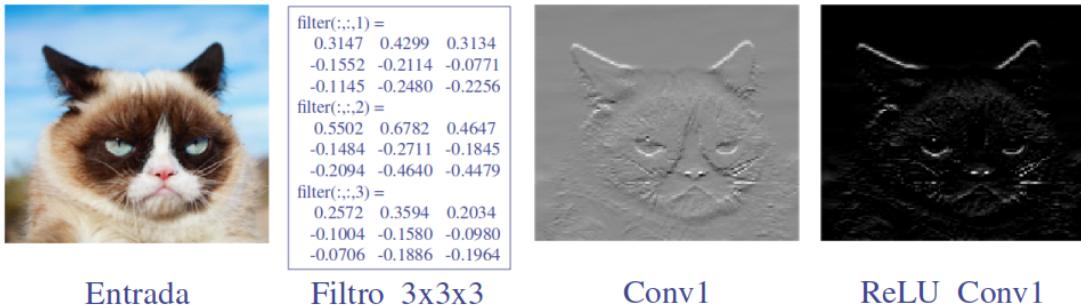


Figura 37.5: Resultado de aplicar un filtro 3D a una imagen antes y después de pasar por el filtro de activación

37.4. Relleno del borde

Si se aplica el filtro convolucional a una entrada, la salida será algo más pequeña al no poder centrar el filtro en los bordes de la matriz. Para poder hacerlo, se suele incrementar la entrada de la capa con un relleno (en inglés *padding*). El relleno se puede realizar con ceros, con algún valor, con el valor más cercano del borde, etc. La Figura 37.6 muestra algunos de los rellenos más empleados.

Valor Constante	Valores del borde	Espejo																																																																																				
<table border="1"> <tr><td>3</td><td>6</td><td>8</td><td></td><td></td><td></td></tr> <tr><td>1</td><td>9</td><td>4</td><td></td><td></td><td></td></tr> <tr><td>5</td><td>2</td><td>7</td><td></td><td></td><td></td></tr> <tr><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> </table>	3	6	8				1	9	4				5	2	7				0	0	0	0	0	0	<table border="1"> <tr><td>3</td><td>3</td><td>6</td><td>8</td><td>8</td><td></td></tr> <tr><td>3</td><td>3</td><td>6</td><td>8</td><td>8</td><td></td></tr> <tr><td>1</td><td>1</td><td>9</td><td>4</td><td>4</td><td></td></tr> <tr><td>5</td><td>5</td><td>2</td><td>7</td><td>7</td><td></td></tr> <tr><td>5</td><td>5</td><td>2</td><td>7</td><td>7</td><td></td></tr> </table>	3	3	6	8	8		3	3	6	8	8		1	1	9	4	4		5	5	2	7	7		5	5	2	7	7		<table border="1"> <tr><td>9</td><td>1</td><td>9</td><td>4</td><td>9</td><td></td></tr> <tr><td>6</td><td>3</td><td>6</td><td>8</td><td>6</td><td></td></tr> <tr><td>9</td><td>1</td><td>9</td><td>4</td><td>9</td><td></td></tr> <tr><td>2</td><td>5</td><td>2</td><td>7</td><td>2</td><td></td></tr> <tr><td>9</td><td>1</td><td>9</td><td>4</td><td>9</td><td></td></tr> </table>	9	1	9	4	9		6	3	6	8	6		9	1	9	4	9		2	5	2	7	2		9	1	9	4	9	
3	6	8																																																																																				
1	9	4																																																																																				
5	2	7																																																																																				
0	0	0	0	0	0																																																																																	
3	3	6	8	8																																																																																		
3	3	6	8	8																																																																																		
1	1	9	4	4																																																																																		
5	5	2	7	7																																																																																		
5	5	2	7	7																																																																																		
9	1	9	4	9																																																																																		
6	3	6	8	6																																																																																		
9	1	9	4	9																																																																																		
2	5	2	7	2																																																																																		
9	1	9	4	9																																																																																		

Figura 37.6: Distintos tipos de relleno del borde

37.4.1. Desplazamiento

El desplazamiento (en inglés *stride*) básico con el que se aplica un filtro convolucional es de 1. Sin embargo, la aplicación de muchos filtros repartidos en capas a lo largo de la red hace que sea especialmente difícil mantener todos los datos generados en un momento determinado del entrenamiento. Para reducir este volumen de datos, se suelen aplicar las convoluciones con un desplazamiento mayor que 1. Esto reduce el tamaño del mapa de activaciones obtenido por una determinada capa (Fig. 37.7).

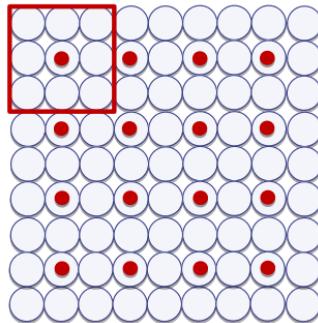


Figura 37.7: Desplazamiento 2x2 del filtro. El punto es el centro de la zona en la que se aplica el filtro en cada momento.

37.5. Capas de agrupación

La ejecución en secuencia de varias capas convolucionales es muy efectiva a la hora de decidir si ciertas características están o no presentes en la entrada. Sin embargo, una de sus ventajas y a la vez limitaciones es que mantiene la localización espacial de las características. Aunque es necesario cierta información espacial como, por ejemplo, el que hubiera unos bigotes cerca de una boca sería característico de una imagen que contuviese un gato, pequeños movimientos del contenido de la imagen producirían mapas de características diferentes.

Una forma de mitigar este problema es usar capas de agrupación (en inglés *pooling*). Estas capas agrupan un número de valores adyacentes de los mapas de características obteniendo un nuevo conjunto de mapas más pequeños. Es posible emplear distintos tipos de operaciones con las que realizar la agrupación. Los más empleados suelen ser el *max pooling* y el *average pooling* (Goodfellow et al., 2016) que seleccionan el máximo de los valores u obtienen su media respectivamente (Fig. 37.8). El tamaño más típico es 2x2.

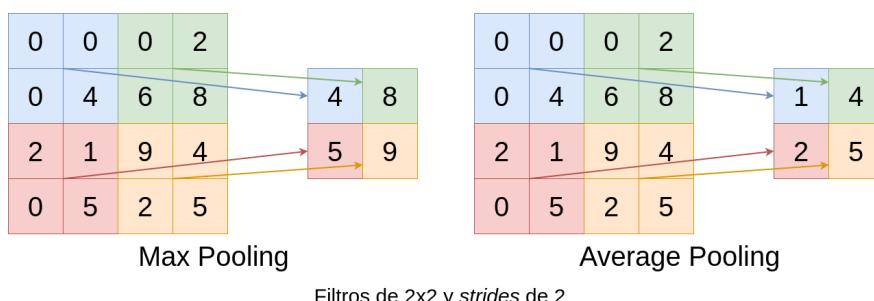


Figura 37.8: Resultado de emplear dos métodos de agrupación diferentes para reducir la dimensión de los datos

37.6. Desvanecimiento del gradiente

La primera red convolucional fue propuesta en 1982 (Fukushima and Miyake, 1982). Esta arquitectura recibió el nombre de Neocognitron y ya constaba de capas convolucionales y capas de *pooling*. Siguiendo la misma idea, en 1998 se diseñó otra CNN para resolver el problema de reconocimiento de dígitos manuscritos, MNIST (LeCun et al., 1998). A esta arquitectura de CNN se la conoce con el nombre de LeNet y es una de las arquitecturas más pequeñas que podemos definir para resolver un problema de clasificación (Fig. 37.9). El extracto de características, consta de dos capas convolucionales alternadas con 2 capas de *pooling* que obtienen un total de 400 variables. La parte final con el clasificador está compuesta por 3 capas densas de 120, 84 y 10 neuronas.

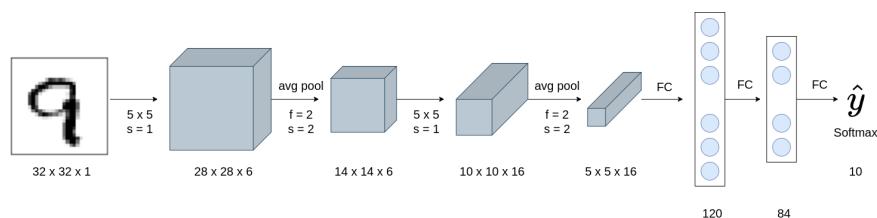


Figura 37.9: Arquitectura LeNet

A pesar de los buenos resultados obtenidos por la arquitectura, el uso de estos métodos para resolver problemas reales estaba aún lejos debido a la carga computacional requerida para su entrenamiento. No fue hasta el año 2012, cuando los ganadores del concurso ImageNet Challenge presentaron una nueva arquitectura llamada AlexNet, que las CNN volvieron a estar en el punto de mira de los investigadores (Deng et al., 2012). A partir de ese momento, y teniendo en cuenta los grandes avances computacionales de las tarjetas gráficas (GPU) que permitían ejecutar operaciones matriciales de forma eficiente, se empezaron a desarrollar cada vez más arquitecturas diferentes.

Durante los primeros años, las arquitecturas desarrolladas tenían cada vez más capas y más filtros en cada una de ellas para extraer la mayor cantidad de información posible de la entrada. Sin embargo, las arquitecturas más profundas se encontraron con un problema: el desvanecimiento del gradiente.

Ciertas funciones de activación como, por ejemplo, la sigmoide, comprimen el espacio de entrada entre 0 y 1. Esto hace que grandes cambios en la entrada produzcan cambios muy pequeños en la salida, haciendo que la derivada sea pequeña. Como los gradientes de la red se calculan durante la propagación hacia atrás capa a capa siguiendo la regla de la cadena, si los valores son muy cercanos a 0, la multiplicación de muchos de estos valores hará que el gradiente de la red caiga rápidamente. Un gradiente muy pequeño hará que los pesos de las capas iniciales apenas se modifiquen con cada iteración y no lleguen a obtener valores adecuados durante el entrenamiento.

Algunas soluciones a este problema son:

- El uso de activaciones tipo ReLU que no obtienen valores muy pequeños en su derivada.

- Capas de normalización. Si se normalizan los datos de entrada ya no habrá grandes cambios entre ellos y los valores estarán lejos de los extremos de la sigmoide.
- Uso de bloques con conexiones residuales que suman el valor de la entrada del bloque a su salida.

37.7. Sobreajuste

Cuanto mayor es el número de parámetros de la red, mayor probabilidad hay de que “memorice” los datos de entrenamiento. Esto se debe a la cantidad de características que la red es capaz de extraer y medir. Si la red es muy profunda, aprenderá cosas muy concretas del conjunto de entrenamiento, lo que dará lugar a modelos que no generalizan bien con nuevos datos (Tetko et al., 1995).

Además de esto, la no linealidad que añaden las funciones de activación puede hacer que se encuentren fronteras de decisión que modelen datos que no son linealmente separables, pero también facilita que se produzca el sobreajuste (Fig. 37.10).

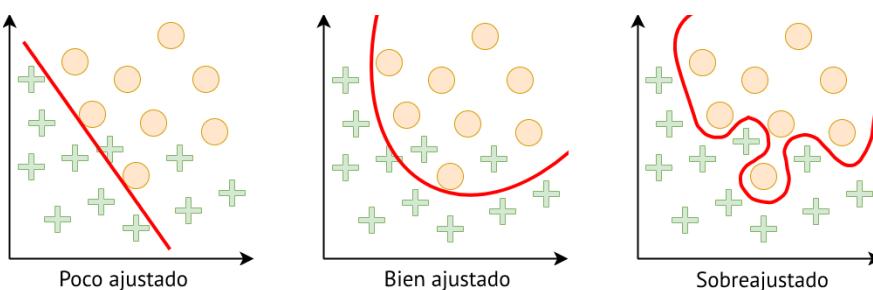


Figura 37.10: Tipos de ajuste del modelo a los datos

Para evitar que se produzca el sobreajuste se suelen emplear técnicas de regularización. Se trata de técnicas que impiden que los modelos sean demasiado complejos mejorando su capacidad de generalización. Algunas de estas técnicas son:

- *Dropout*. Durante el entrenamiento, algunas activaciones se ponen a 0 de forma aleatoria (entre el 10 % y el 50 %). Esto hace que una capa de la red no dependa siempre de los mismos nodos anteriores.
- *Early Stopping*. Se trata de parar el entrenamiento antes de que se produzca el sobreajuste y seleccionar ese modelo como final. Para ello se utilizan dos conjuntos: uno de entrenamiento y otro de validación. Cuando las curvas de pérdida de ambos conjuntos comienzan a diverger, se para el entrenamiento y se selecciona el modelo resultante del momento anterior al comienzo de la divergencia (Fig. 37.11).
- *Regularización L1*. Penaliza los pesos grandes por lo que fuerzan a los pesos a tener valores cercanos a 0 (sin ser 0). Añade un término de penalización a la función de coste sumando

37.8. Generación de datos de entrenamiento artificiales

631

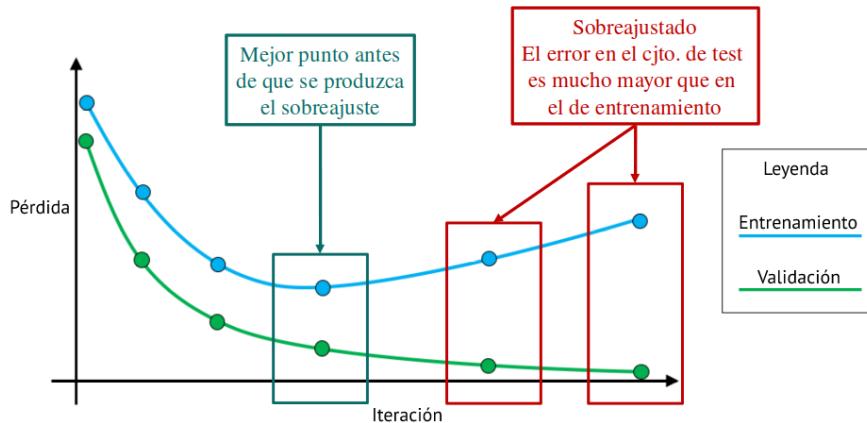


Figura 37.11: Selección del modelo antes del sobreajuste

todos los pesos de la matriz y multiplicado por un valor α que es otro hiperparámetro que debe ser seleccionado manualmente:

$$\alpha \|\mathbf{W}\|_1 = \alpha \sum_i \sum_j |w_{ij}|. \quad (37.4)$$

- **Regularización L2 o weight decay.** Parecida a la regularización L2 pero con una expresión algo diferente:

$$\frac{\alpha}{2} \|\mathbf{W}\|_2^2 = \frac{\alpha}{2} \sum_i \sum_j w_{ij}^2. \quad (37.5)$$

37.8. Generación de datos de entrenamiento artificiales

Como se ha comentado anteriormente, las técnicas de *deep learning* suelen requerir de gran cantidad de datos para su correcto funcionamiento. En muchas situaciones, se dispone de un conjunto limitado para poder entrenar los modelos de forma correcta, por lo que para tratar de suplir la falta de datos se recurre a la generación de datos artificiales, técnica conocida con el nombre de, con la expresión en inglés, *data augmentation* ([Shorten and Khoshgoftaar, 2019](#)).

Esta técnica realiza pequeñas variaciones en los datos del conjunto de entrenamiento del que se dispone para obtener nuevos, manteniendo el significado semántico de los mismos. Esto también permite mejorar la generalización de los modelos. Por ejemplo, si se tienen imágenes donde una de ellas contiene un elemento de la clase *gato*, las modificaciones que se realicen deben permitir poder reconocer esa misma clase a partir de las imágenes modificadas.

Algunos ejemplos de técnicas de *data augmentation* en imagen pueden ser: la realización de rotaciones, modificación del contraste o cambios en la iluminación, reescalados, adición/eliminación de ruido o cambio en las proyecciones de las mismas.

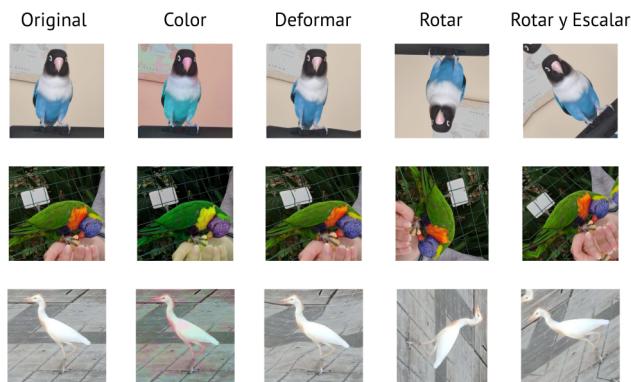


Figura 37.12: Ejemplos de técnicas de generación de datos artificiales

Para agregar diferentes tipologías de esta técnica de *data augmentation* en **R**, se pueden incluir capas de preprocessado en el modelo secuencial, que serán ejecutadas de manera aleatoria únicamente durante el entrenamiento. En el siguiente ejemplo se realizarán rotaciones aleatorias, volteados horizontales y acercamientos a la imagen:

```
data_augmentation <-
  keras_model_sequential() |>
  layer_random_rotation(0.1) |>
  layer_random_flip("horizontal") |>
  layer_random_zoom(0.1)
```

A continuación se muestran los diferentes tipos de preprocessado disponibles para imagen y redes neuronales convolucionales:

```
layer_random_crop()
layer_random_flip()
layer_random_flip()
layer_random_translation()
layer_random_rotation()
layer_random_zoom()
layer_random_height()
layer_random_width()
layer_random_contrast()
```

NOTA

Otros tipos de *data augmentation* disponibles en **keras** y **R** para otro tipo de datos pueden consultarse en
https://tensorflow.rstudio.com/guides/keras/preprocessing_layers

37.9. Ejemplo en R para el conjunto de datos CIFAR10

En esta sección se verá cómo entrenar una red neuronal convolucional para ser capaces de clasificar las clases contenidas en el conjunto de datos **CIFAR10**. La descarga debe hacerse a través del siguiente enlace https://drive.google.com/file/d/1-pFGg-bkooss1fNp5UNYR0-hLUMDP_XO/view?usp=sharing y el conjunto tiene que guardarse en una carpeta **data** dentro del proyecto de trabajo para asegurar la reproducibilidad del capítulo. Cada una de las imágenes contenidas en el mismo contiene un único elemento que puede ser clasificado como avión, coche, pájaro, gato, ciervo, perro, rana, caballo, barco o camión.

Nota

Existe otra versión del conjunto de datos denominada como **CIFAR100**, en la cual se definen un total de 100 posibles categorías en las que las imágenes contenidas pueden ser clasificadas. El ejemplo a continuación puede ser replicado con este mismo conjunto de datos.

https://www.rdocumentation.org/packages/keras/versions/2.9.0/topics/dataset_cifar100

Cada una de las imágenes contenidas en el conjunto tiene un tamaño de 32x32 píxeles en color, representándose mediante los 3 canales RGB, siendo diferente al ejemplo del capítulo **36**, en el cual se trabaja con imágenes en escala de grises y, por tanto, un único canal.

A continuación, se verán los pasos seguidos, siendo de forma general muy similares a los ya descritos en el Cap. **36**, pero adaptando la red al tipo de dato utilizado.

37.9.1. Carga y visualización de los datos

El primer paso será cargar la librería **keras**, para así poder crear las redes neuronales necesarias y también para cargar el conjunto de imágenes CIFAR10 que se encuentra disponible públicamente:

```
library(keras)
load("data/cifar10.RData")
```

A continuación, se puede ver el contenido de las variables generadas, donde se puede observar como el conjunto de datos CIFAR10 ya viene separado en dos subconjuntos que pueden ser utilizados para entrenamiento y para test. Además se puede ver que el conjunto de entrenamiento

está compuesto por 50000 imágenes, mientras que el conjunto de test por 10000. En ambos casos, estas imágenes se almacenan en la variable x .

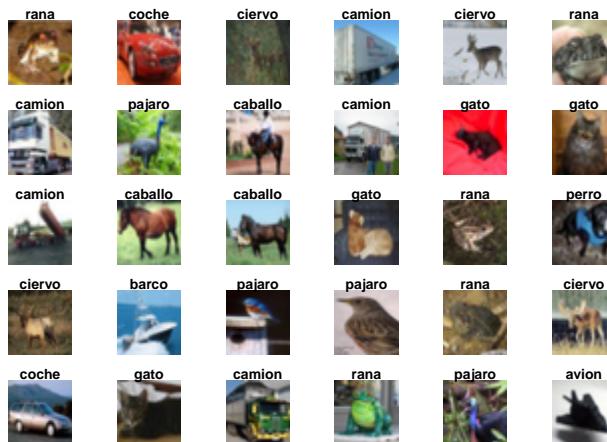
```
names(cifar)
#> [1] "train" "test"
dim(cifar$train$x)
#> [1] 50000    32    32     3
dim(cifar$train$y)
#> [1] 50000      1
dim(cifar$test$x)
#> [1] 10000    32    32     3
dim(cifar$test$y)
#> [1] 10000      1
```

Además, las imágenes de cada subconjunto tienen definida la clase a la que pertenecen, en este caso, cualquiera de las 10 clases indicadas anteriormente. En ambos casos, esta etiqueta se almacena en la variable y . A continuación, se muestra un pequeño ejemplo que permitirá mostrar alguna de las imágenes contenidas en el conjunto de datos de entrenamiento junto con su etiqueta:

```
class_names <- c('avion', 'coche', 'pajaro', 'gato', 'ciervo',
                 'perro', 'rana', 'caballo', 'barco', 'camion')

index <- 1:30

par(mfcol = c(5,6), mar = rep(1, 4), oma = rep(0.2, 4))
cifar$train$x[index,,,] |>
  purrr::array_tree(1) |>
  purrr::set_names(class_names[cifar$train$y[index] + 1]) |>
  purrr::map(as.raster, max = 255) |>
  purrr::iwalk(~{plot(.x); title(.y)})
```



37.9.2. Preprocesamiento

Una vez cargados los datos y comprobado su contenido, al igual que se explicó en el Cap. 36, es posible realizar algún tipo de preprocesado. Al estar trabajando con imágenes, es muy típico estandarizar los valores de color de las imágenes para mitigar las diferencias producidas por las diferentes condiciones de iluminación.

En este caso, al igual que en el Cap. 36, se va a transformar los valores originales de la imagen (en rango de 0 a 255) a valores entre 0 y 1 dividiendo cada valor por el máximo, 255:

```
cifar$train$x <- cifar$train$x/255
cifar$test$x <- cifar$test$x/255
```

37.9.3. Generación de la red neuronal

En esta ocasión se creará la red neuronal convolucional en dos pasos, para además mostrar cómo se pueden utilizar las funciones proporcionadas por la librería **keras** para definir una CNN en varias partes, combinándolas poco a poco.

En el primero, se definirá, utilizando la interfaz *sequential* proporcionada por Tensorflow/Keras a través de la función *keras_model_sequential*, la base convolucional de la red combinando varias capas *Conv2d* con *MaxPooling2D*. Esta es la parte de la red que se encargará de aprender las características necesarias que permitirán representar el contenido de la imagen. Otra de las diferencias principales que se puede observar en esta red es que, al aceptar imágenes de 3 canales, RGB, en vez de imágenes en escala de grises, el tamaño de la entrada de la primera capa tiene que reflejar esto *input_shape = c(32,32,3)*.

```
model <- keras_model_sequential() |>
  layer_conv_2d(filters = 32, kernel_size = c(3,3), activation = "relu",
                 input_shape = c(32,32,3)) |>
  layer_max_pooling_2d(pool_size = c(2,2)) |>
  layer_conv_2d(filters = 64, kernel_size = c(3,3), activation = "relu") |>
  layer_max_pooling_2d(pool_size = c(2,2)) |>
  layer_conv_2d(filters = 64, kernel_size = c(3,3), activation = "relu")
```

Como se puede observar, en esta parte de la red se reduce la dimensión de la información de manera paulatina en cada capa, obteniendo las características representativas del objeto contenido en cada imagen hasta llegar a un tamaño de $4 \times 4 \times 64$.

```
summary(model, line_length=74)
```

```
#> Model: "sequential_2"
#> -----
#> Layer (type)          Output Shape         Param #
#> ======
```

```
#> conv2d_2 (Conv2D)           (None, 30, 30, 32)      896
#> max_pooling2d_1 (MaxPooling2D) (None, 15, 15, 32)      0
#> conv2d_1 (Conv2D)           (None, 13, 13, 64)     18496
#> max_pooling2d (MaxPooling2D) (None, 6, 6, 64)        0
#> conv2d (Conv2D)             (None, 4, 4, 64)      36928
#> =====
#> Total params: 56,320
#> Trainable params: 56,320
#> Non-trainable params: 0
#> -----
```

Ahora, será necesario añadir capas que permitan transformar los resultados de la parte convolucional de la red implementada a un valor de probabilidad de que la imagen represente cada una de las posibles clases de la imagen.

Para ello, primero se inserta una capa de tipo *flatten* que se encarga de transformar la salida de la última capa convolucional $4 \times 4 \times 64$ a un vector de 1024 elementos. A continuación, una capa oculta *dense* de 64 neuronas con activación *relu* se encarga de realizar las primeras operaciones con esos datos y de reducir la dimensionalidad. Finalmente, una última capa *dense* con activación *softmax* se encarga de obtener la probabilidad de que la imagen represente cada una de las 10 posibles clases:

```
model |>
  layer_flatten() |>
  layer_dense(units = 64, activation = "relu") |>
  layer_dense(units = 10, activation = "softmax")
```

A continuación, se puede observar como quedaría la estructura final del modelo implementado:

```
summary(model, line_length=74)
```

```
#> Model: "sequential_2"
#> -----
#> Layer (type)          Output Shape       Param #
#> =====
#> conv2d_2 (Conv2D)      (None, 30, 30, 32)      896
#> max_pooling2d_1 (MaxPooling2D) (None, 15, 15, 32)      0
#> conv2d_1 (Conv2D)      (None, 13, 13, 64)     18496
#> max_pooling2d (MaxPooling2D) (None, 6, 6, 64)        0
#> conv2d (Conv2D)        (None, 4, 4, 64)      36928
#> flatten_1 (Flatten)   (None, 1024)            0
#> dense_7 (Dense)       (None, 64)              65600
#> dense_6 (Dense)       (None, 10)              650
#> -----
#> Total params: 122,570
#> Trainable params: 122,570
```

37.9. Ejemplo en **R** para el conjunto de datos CIFAR10

637

```
#> Non-trainable params: 0
#> -----
```

NOTA

Un detalle a tener en cuenta con respecto al ejemplo del Cap. 36 es el parámetro *Total params*. Este valor indica el número de parámetros que contiene nuestra red neuronal y, en cierta manera, el tamaño de la misma. Se puede observar que en este caso tiene un mayor tamaño al contar con un total de 122570 parámetros con respecto a los 11935 del ejemplo previo.

Finalmente, es necesario compilar el modelo, indicando algunos de los parámetros de configuración necesarios para el proceso de entrenamiento, como serían el optimizador a utilizar, la función de coste y las métricas a calcular para poder evaluar la red entrenada:

```
model |> compile(
  optimizer = "sgd", # stochastic gradient descent
  loss = "sparse_categorical_crossentropy", # función utilizada para problemas de
  # clasificación con varias clases
  metrics = "accuracy" # Precisión
)
```

37.9.4. Entrenamiento

Una vez generada la estructura de la red neuronal convolucional, es posible entrenarla para resolver el problema de clasificación mediante la función *fit*. Para ello, se le debe indicar el conjunto de imágenes de entrenamiento, *x*, que debe utilizar y sus etiquetas correspondientes, *y*. Además de otros parámetros, se podrá configurar el número de *epochs* a entrenar (pasadas sobre el conjunto completo de entrenamiento), el tamaño del batch que se utilizará en cada iteración con *batch_size* (número de imágenes por iteración), qué porcentaje de elementos del conjunto de datos se utilizarán para validar el modelo con *validation_split* (imágenes utilizadas durante el entrenamiento pero solo para obtener una estimación real del error cometido) o la tasa de aprendizaje, *learning_rate*, entre otros.

```
training_evolution <- model |>
  fit(
    x = cifar$train$x, y = cifar$train$y,
    epochs = 10, batch_size = 32,
    validation_split = 0.2,
    learning_rate = 0.1,
    verbose = 2
  )
```

NOTA

Como se puede observar, el *batch_size* configurado es menor que el del Cap. 36 (32 vs 128). Esto es debido a que, el número máximo de imágenes que un mismo equipo utilizado para entrenar podrá procesar en una iteración vendrá determinado por el tamaño de la red neuronal, es decir, por la variable *Total params* indicada en la Nota anterior. Cuanto mayor sea el tamaño de la red, menor será el número máximo de imágenes que podrá tener el batch.

```
#> Epoch 1/10
#> 1250/1250 - 12s - loss: 2.1097 - accuracy: 0.2316 - val_loss: 1.9339 - val_accuracy:
→ 0.2958 - 12s/epoch - 9ms/step
#> Epoch 2/10
#> 1250/1250 - 8s - loss: 1.7478 - accuracy: 0.3667 - val_loss: 1.6987 - val_accuracy:
→ 0.3965 - 8s/epoch - 6ms/step
#> Epoch 3/10
#> 1250/1250 - 8s - loss: 1.5464 - accuracy: 0.4399 - val_loss: 1.4731 - val_accuracy:
→ 0.4707 - 8s/epoch - 7ms/step
#> Epoch 4/10
#> 1250/1250 - 9s - loss: 1.4304 - accuracy: 0.4866 - val_loss: 1.3653 - val_accuracy:
→ 0.5149 - 9s/epoch - 7ms/step
#> Epoch 5/10
#> 1250/1250 - 8s - loss: 1.3477 - accuracy: 0.5199 - val_loss: 1.3407 - val_accuracy:
→ 0.5257 - 8s/epoch - 6ms/step
#> Epoch 6/10
#> 1250/1250 - 7s - loss: 1.2784 - accuracy: 0.5437 - val_loss: 1.2563 - val_accuracy:
→ 0.5564 - 7s/epoch - 6ms/step
#> Epoch 7/10
#> 1250/1250 - 7s - loss: 1.2118 - accuracy: 0.5705 - val_loss: 1.2331 - val_accuracy:
→ 0.5720 - 7s/epoch - 6ms/step
#> Epoch 8/10
#> 1250/1250 - 8s - loss: 1.1539 - accuracy: 0.5954 - val_loss: 1.1807 - val_accuracy:
→ 0.5882 - 8s/epoch - 6ms/step
#> Epoch 9/10
#> 1250/1250 - 7s - loss: 1.1015 - accuracy: 0.6135 - val_loss: 1.1516 - val_accuracy:
→ 0.5935 - 7s/epoch - 6ms/step
#> Epoch 10/10
#> 1250/1250 - 7s - loss: 1.0526 - accuracy: 0.6286 - val_loss: 1.1014 - val_accuracy:
→ 0.6128 - 7s/epoch - 6ms/step
```

Tras el entrenamiento, se puede observar la evolución del mismo mediante las gráficas de coste/perdida y precisión.

```
plot(training_evolution)
```

Como se puede observar, la red entrenada es capaz de alcanzar un 60 % de precisión tanto en los conjuntos de entrenamiento como los de validación

37.9. Ejemplo en **R** para el conjunto de datos CIFAR10

639

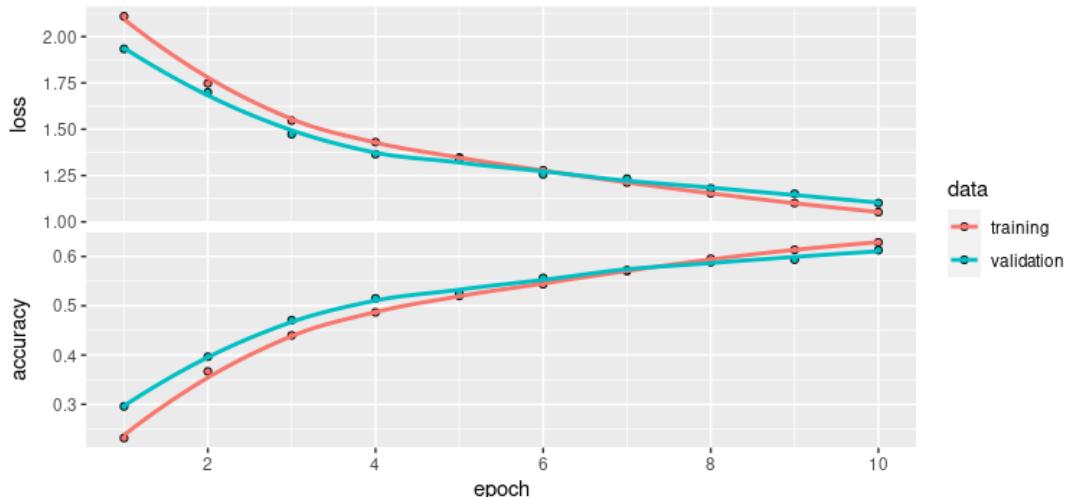


Figura 37.13: Evolución durante el entrenamiento de la precisión y la pérdida de los conjuntos de entrenamiento y validación

37.9.5. Test

Una vez entrenado el modelo, es posible aplicarlo sobre el conjunto de test proporcionado. Para ello, se puede realizar la predicción sobre cualquiera de las imágenes mediante la función `predict()`, obteniendo la probabilidad de que pertenezca a una determinada clase:

```
predictions <- predict(model, cifar$test$x)
head(round(predictions, digits=2), 5)
```

```
#>      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
#> [1,] 0.03 0.00 0.18 0.47 0.01 0.14 0.16    0 0.00 0.00
#> [2,] 0.04 0.07 0.00 0.00 0.00 0.00 0.00    0 0.89 0.00
#> [3,] 0.08 0.28 0.00 0.00 0.00 0.00 0.00    0 0.55 0.07
#> [4,] 0.82 0.01 0.04 0.00 0.00 0.00 0.00    0 0.11 0.00
#> [5,] 0.00 0.00 0.05 0.11 0.11 0.03 0.69    0 0.00 0.00
```

También se puede utilizar la función `evaluate()` para calcular tanto el coste o pérdida como la precisión de la red neuronal sobre el conjunto de test. Como se puede observar, se obtienen valores muy similares a los obtenidos durante el entrenamiento:

```
evaluate(model, cifar$test$x, cifar$test$y, verbose = 0)
```

```
#>      loss accuracy
#> 1.094381 0.611100
```

Con la función *predict* se puede también generar la matriz de confusión de la red para evaluar qué pares de clases está cometiendo un mayor número de errores:

```
prediction_matrix <- model |> predict(cifar$test$x) |> k_argmax()
confusion_matrix <- table(as.array(prediction_matrix), cifar$test$y)
confusion_matrix
```

```
#>
#>      0   1   2   3   4   5   6   7   8   9
#> 0 650  25  57  12  39  12  3  21  82  35
#> 1  43 790  15  18  13  11  16  12  43 179
#> 2 119  20 676 180 325 170 112  94  33  26
#> 3  23  15  57 427  76 184  54  48  18  18
#> 4   1   0  13  20 236  12   5  21   3   2
#> 5   6   7  52 145  41 488  24  82   5  10
#> 6  14   5  71 110 119  42 755  13   4  14
#> 7   7   6  30  46 126  60  16 676   9  18
#> 8 114  58  15  19  20  13   6   5 777  62
#> 9  23  74  14  23   5   8   9  28  26 636
```

37.9.6. Otros ejemplos de interés

- Transfer learning and fine tuning. Explicación de estas técnicas para clasificar imágenes que contienen perros y gatos. https://tensorflow.rstudio.com/guides/keras/transfer_learning
- <https://tensorflow.rstudio.com/guides/>
- <https://tensorflow.rstudio.com/examples/>

Resumen

- Este capítulo presenta las características de las redes neuronales convolucionales y sus diferencias con el perceptrón multicapa.
- Además, se exponen los principales problemas a la hora de diseñar este tipo de redes profundas y sus posibles soluciones.
- Finalmente, se han explicado los pasos necesarios para poder entrenar una red neuronal convolucional utilizando la librería Tensorflow/Keras en **R**, resolviendo el problema de clasificación de las 10 clases del conjunto de datos CIFAR10.

Parte VIII

Ciencia de datos de texto y redes

Capítulo 38

Minería de textos

Víctor Casero-Alonso^a, Ángela Celis^a y María Lozano Zahonero^b

^aUniversidad de Castilla-La Mancha y ^b Università degli Studi di Roma Tor Vergata.

38.1. Introducción

En la actualidad, entre el 80 % y el 90 % de los datos que se generan diariamente son datos no estructurados (vistos en el Cap. ??). Un ejemplo típico de datos no estructurados son los textos, desde los comentarios o mensajes de las redes sociales, reseñas, blogs y microblogs, chats o whatsapp hasta las noticias periodísticas, los discursos políticos o las obras literarias. En consecuencia, aprender a procesar y analizar datos exige aprender a procesar y analizar textos.

Los textos precisan, sin embargo, un tratamiento especial. A diferencia de la mayoría de los datos que se tratan en este libro, que son datos estructurados, los datos textuales requieren que se les otorgue un orden y estructura para su manejo y análisis con el software R. Además, al utilizar un lenguaje natural –es decir, un idioma como, por ejemplo, el español, el chino o el inglés–, los textos no pueden ser procesados directamente por un ordenador. Es preciso “traducirlos” antes a un lenguaje formal que los ordenadores puedan entender.

La **minería de textos** (en inglés, *text mining*), también conocida como **análisis de textos** (en inglés, *text analysis*), puede definirse como el proceso para detectar, extraer, clasificar, analizar y visualizar la información no explícita que contienen los textos, transformando los datos textuales en datos estructurados y el lenguaje natural en lenguaje formal a fin de determinar, después, de manera automática, patrones recurrentes y desviaciones de los mismos. La minería de textos utiliza muchas técnicas y métodos diferentes, la mayor parte de los cuales proceden del **procesamiento del lenguaje natural** (PLN), un ámbito de la inteligencia artificial que se ocupa de la comunicación entre los seres humanos y las máquinas mediante el tratamiento computacional del lenguaje humano.

Este capítulo constituye una primera aproximación a la minería de textos con **R**. Su objetivo es proporcionar un marco teórico y aplicado básico de este ámbito. Para ello, en la Sec. 38.2, se presentan los conceptos y fases fundamentales de la minería de textos. La Sec. 38.3 está dedicada al análisis de sentimientos, que constituye uno de los campos de la minería de textos de mayor desarrollo en la actualidad. La Sec. 38.4 indica paquetes de **R** que permiten realizar análisis textuales de distintos tipos. Cierra el capítulo un ejemplo, en el que se aplica y se amplía lo estudiado anteriormente. Dos referencias útiles sobre el tema son [Fraudejas Rueda \(2022\)](#) y [Jockers \(2014\)](#).

38.2. Conceptos y tareas fundamentales

Lo primero que se necesita para hacer un análisis de textos son los textos. Esta afirmación podría parecer banal, pero no lo es. El volumen de textos en circulación es ingente, pero, en la mayor parte de los casos, es necesario realizar una serie de operaciones complejas para poder extraer y recopilar los datos textuales que se quiere analizar. Es también difícil muchas veces acceder después a estos datos, ya que los textos pueden presentar formatos muy heterogéneos, no siempre interpretables o fáciles de convertir en un formato interpretable. Baste pensar, por ejemplo, en una nota escrita a mano. Dado que este capítulo es una primera aproximación a la minería de textos, se parte del supuesto de que el texto o los textos están disponibles ya en un fichero, denominado **corpus**, legible por **R**. En este contexto, *corpus* es la colección de textos con el mismo origen, por ejemplo, el *corpus* de las obras de un autor, que para poder manejarse requiere metadatos con detalles adicionales.

38.2.1. Preparación de los datos

Una vez constituido el *corpus*, la primera fase es la **preparación de los datos**. Los textos suelen contener un cierto grado de “suciedad”, es decir, elementos que alteran o impiden el análisis. De una buena “limpieza” inicial, dependerá en gran parte la validez de los resultados que se obtengan. Entre las operaciones de “limpieza” generales figuran una serie de transformaciones cuya finalidad es evitar el recuento incorrecto de palabras, como el cambio de mayúsculas por minúsculas y la eliminación de los signos de puntuación, los números y los espacios en blanco en exceso.

La siguiente operación de preparación, que tiene un importante peso en el análisis, es la eliminación de las **palabras vacías** (en inglés, **stopwords**). En la lengua no todas las palabras tienen el mismo tipo de significado. Las palabras con significado léxico, como *mesa* o *corpus*, son palabras a las que corresponde un concepto que se puede definir o explicar. Otras palabras, sin embargo, son palabras funcionales, cuyo contenido es puramente gramatical. Son palabras como el artículo *el*, la preposición *de* o la conjunción *o*: se puede explicar cómo se usan, pero no definirlas asociándolas a un concepto porque carecen de contenido léxico-semántico.

Las palabras vacías son, con gran diferencia respecto de las palabras léxicas, las más frecuentes de la lengua, pero, dado su escaso o nulo significado léxico, en los análisis de tipo semántico, como el análisis de sentimientos o el modelado de temas, carecen de valor informativo, por lo que es conveniente eliminarlas. No es aconsejable eliminarlas, sin embargo, en otros tipos de

análisis, como los análisis estilométricos, donde tienen un importante valor informativo como se verá en la Sec. 38.2.4. Las palabras vacías pertenecen a clases cerradas, es decir, a clases de palabras con un número de elementos limitado, finito. Es posible confeccionar, por tanto, listas de palabras vacías para facilitar su eliminación. En el ejemplo de aplicación que se verá en la Sec. 38.5, se aprenderá a usar estas listas y se podrá apreciar con detalle la diferente información que proporciona una tabla de frecuencias con y sin palabras vacías.

38.2.2. Segmentación del texto: tokenización

La segunda fase de la minería de textos consiste en la **segmentación del texto**, denominada también **tokenización**. El texto se divide en ***tokens***, secuencias de texto con valor informativo. De esta manera, se pasa del lenguaje natural a un lenguaje formal comprensible por el software, dándole formato de ‘vector’ o ‘tabla’. Así se pueden aplicar algunas de las herramientas que se utilizan con datos numéricos para manejar el texto y obtener resúmenes y visualizaciones que muestren la información no explícita contenida en él en forma de patrones recurrentes.

Generalmente, los *tokens* son **palabras**, es decir, secuencias de caracteres entre dos espacios en blanco y/o signos de puntuación, pero pueden ser también **oraciones**, **líneas**, **párrafos** o **n-gramas**. Como se verá en el ejemplo de aplicación, un primer análisis del significado consiste en eliminar las palabras vacías y obtener las frecuencias¹ de las palabras con valor informativo para responder a la pregunta “¿Qué se dice?” (Silge and Robinson, 2017).

Nota

También puede ser útil obtener la **tasa de riqueza léxica** (TTR, del inglés *type-token ratio*). Esta mide la relación entre el número de palabras diferentes que contiene un texto (*types*) dividido entre las palabras totales de ese texto (*tokens*)^a.

$$TTR = \frac{\text{Types}}{\text{Tokens}}$$

^aVéase <https://www.fundeu.es/consideraciones-teoricas/>

38.2.2.1. N-gramas

El análisis puede proseguir estudiando la frecuencia de los **n-gramas**, secuencias de *n* palabras consecutivas en el mismo orden. Se tienen así bigramas o 2-gramas (secuencias de dos palabras), trigramas o 3-gramas (secuencias de tres palabras), etc. El estudio de los *n-gramas* responde al principio de Firth: “*You shall know a word by the company it keeps*” (Firth, 1957, 11). Este principio es el fundamento del llamado **análisis de colocaciones**: para conocer el significado de una palabra es preciso conocer las palabras con las que aparece, el contexto relevante. En un sentido amplio, el análisis de colocaciones consiste en examinar los contextos izquierdo y/o derecho de una palabra. La segmentación en *n-gramas* permite tener en cuenta este contexto relevante que indicará, por ejemplo, que *banco* es, con toda probabilidad, un asiento en las

¹Frecuencias relativas si se comparan distintos textos.

secuencias *banco de madera* o *banco en la terraza*, pero no lo es en secuencias como *banco de peces*, *banco de arena*, *banco de inversiones*, *banco de datos* o *banco de pruebas*. La división en *n-gramas* permitirá también considerar en el análisis, al menos hasta cierto punto, el peso de la ambigüedad, la negación o el distinto significado que pueden tener las palabras según el ámbito temático. Por ejemplo, la forma *larga* no tiene el mismo significado en los bigramas *falda larga*, *mano larga* y *cara larga*, ni tiene tampoco el mismo valor informativo en *es larga / no es larga* o en *de larga experiencia* (valor positivo) y en *se me hizo larga* (valor negativo). En el ejemplo de aplicación (Sec. 38.5), se verá la segmentación en *n-gramas* en la práctica, y cómo la visualización de redes contribuye a complementar el análisis.

38.2.3. *Stemming* y lematización

La tokenización se puede refinar mediante el ***stemming***, o reducción de las palabras “flexionadas” a su raíz, y la **lematización**, o extracción del lema de cada palabra. Un ejemplo de *stemming* sería reducir las palabras *texto*, *textos*, *textual* y *textuales*, que R cuenta como cuatro palabras diferentes, a la raíz “text”. El *stemming* puede proporcionar un recuento más preciso en algunos casos, pero en otros, al eliminar los sufijos de las palabras, puede crear confusión. Además, como en el ejemplo anterior, las raíces pueden no coincidir con palabras existentes, lo que hace que sean difíciles de interpretar y resulten extrañas si se visualizan en nubes de palabras. Con la lematización se reducen las formas flexionadas de una misma palabra al lema, que es la forma que encabeza la entrada de la palabra en el diccionario. Por ejemplo, si se quiere buscar el significado de la palabra *niñas* no se encontrará como tal sino bajo el lema *niño* y si se quiere buscar *iremos* se tendrá que buscar el lema *ir*. En el caso anterior, la lematización reduciría las formas *texto*, *textos*, *textual* y *textuales* a dos lemas: *texto* y *textual*. La lematización evita la dispersión de significado en varias formas, pero a veces es compleja y puede conducir a la pérdida de información pertinente.

38.2.4. Campos de aplicación de la minería de textos

La minería de textos tiene varios campos de aplicación. Entre ellos destacan tres:

1. El **análisis de sentimientos** se tratará con detalle en la Sec. 38.3 y en el ejemplo de aplicación (Sec. 38.5.4).
2. El **modelado de temas o tópicos** (en inglés, *topic modelling*), como su propio nombre indica, tiene por objeto identificar los temas principales sobre los que versa el texto haciendo uso de técnicas de clasificación no supervisada del campo del aprendizaje automático, como por ejemplo LDA (*Latent Dirichlet Allocation*). Se ilustrará en el Cap. 59.
3. La **estilometría** o **análisis estilométrico** es una aplicación de la minería de textos cuya finalidad consiste en determinar las relaciones existentes entre el estilo de los textos y los metadatos incluidos en ellos. Se utiliza principalmente en la atribución de autoría. El concepto base es el de **huella lingüística**, constituida por el conjunto de rasgos lingüísticos que caracterizan el estilo de un autor como un estilo individual y único y permiten identificarlo. Un punto clave es que, contrariamente a lo que podría pensarse, los rasgos

que conforman en mayor medida la huella lingüística son los que tienen un mayor índice de frecuencia. La mayor parte de los enfoques utilizan el vector de las “palabras más frecuentes” (MFW, por sus siglas en inglés), que son, como se ha visto antes, las palabras vacías y no las palabras con significado léxico, para determinar el estilo de un autor. Esto es debido fundamentalmente a que las palabras vacías se usan de manera involuntaria e inconsciente, configurando de esta manera, sin ningún tipo de filtros racionales, una clave estilística idiosincrásica (Lozano Zahonero, 2020). De lo anterior se deduce fácilmente que en este tipo de análisis no deben eliminarse las palabras vacías.

En la actualidad, el análisis estilométrico se usa en ámbitos muy dispares: desde la criminología o los servicios de inteligencia para identificar a los autores de mensajes o notas en casos de asesinatos, terrorismo, secuestro o acoso, por ejemplo, hasta el derecho civil o la literatura en cuestiones de derechos de autor o detección de plagio, entre muchas otras cuestiones.

38.3. Análisis de sentimientos

El **análisis de sentimientos** (en inglés, *sentiment analysis*) es una aplicación de la minería de textos que tiene como finalidad la detección, extracción, clasificación, análisis y visualización de la dimensión subjetiva asociada a los temas o tópicos presentes en los textos. La dimensión subjetiva comprende no solo los sentimientos, sino también las **emociones**, sensaciones y estados afectivos y anímicos, así como las opiniones, creencias, percepciones, puntos de vista, actitudes, juicios y valoraciones. De ahí que reciba también el nombre de **minería de opinión** (en inglés, *opinion mining*) (Lozano Zahonero, 2020).

El análisis de sentimientos asigna a esta dimensión subjetiva una polaridad, que puede ser positiva o negativa (Pang and Lee, 2008). Algunas técnicas añaden además una polaridad neutra. En algunos casos, el análisis de sentimientos se refina hasta llegar a las emociones básicas: este subcampo del análisis de sentimientos se conoce como **detección de emociones**.

La primera aplicación del análisis de sentimientos fue la investigación de mercados. A partir del año 2000, se registra un crecimiento exponencial de textos como reseñas, chats, foros, blogs, microblogs o comentarios y mensajes de las redes sociales, en los que predomina la expresión de emociones y opiniones personales. Mediante el análisis de sentimientos se extrae de ellos información que permite conocer los gustos del consumidor y diseñar productos a su medida. Esta idea se extenderá después a otros ámbitos, en especial a aquellos en los que predomina la comunicación persuasiva como las campañas publicitarias o políticas. Recientemente, ha empezado a utilizarse también con fines predictivos y preventivos en muchas esferas: desde cuáles son los políticos, las empresas, las películas, canciones u obras literarias que obtendrán un mayor rendimiento, mejores resultados o más votos o ventas hasta cómo detectar y prevenir, por ejemplo, conductas suicidas mediante el análisis de mensajes en las redes sociales.

En el análisis de sentimientos y la detección de emociones existen dos enfoques principales: el enfoque basado en el aprendizaje automático (*machine learning*), en el que se usan algoritmos de aprendizaje supervisado, y el enfoque semántico basado en diccionarios o **lexicones**. Este último enfoque es el que se verá en detalle en el ejemplo de aplicación.

En **R** están implementados varios lexicones para el análisis de sentimientos. Dos de los más utilizados son **bing**, de Bing Liu y colaboradores ([Liu, 2015](#)), y **NRC**, de Saif Mohammad y Peter Turney, ambos incluidos tanto en el paquete **tidytext** como en **syuzhet** ([Jockers, 2017](#)). Estos lexicones tienen en común que están basados en unigramas, es decir, en palabras sueltas, y que tienen como idioma original el inglés, si bien hay versiones traducidas automáticamente a distintas lenguas. La diferencia principal entre los dos lexicones es que **bing** clasifica las palabras de forma binaria en polaridad positiva/negativa, mientras que **NRC** además de la polaridad positiva/negativa permite detectar también ocho emociones básicas (*ira, miedo, anticipación, confianza, sorpresa, tristeza, alegría, asco*). En el ejemplo de aplicación se compararán ambos diccionarios. Como se verá, los resultados del análisis dependerán en buena medida del lexicón elegido, así como del idioma del texto y de si el lexicón se elaboró originalmente en ese idioma o es una versión traducida automáticamente de otra lengua.

38.4. Minería de textos en R

En **R** existen diversos paquetes y funciones que facilitan la minería de textos, entre los que destacan:

- **tidytext**: con la filosofía del **tidyverse**, puede combinarse con los conocidos paquetes **dplyr**, **broom**, **ggplot2**, etc. Se puede destacar la función **unnest_tokens()**, que automatiza el proceso de *tokenización* y el almacenamiento en formato *tidy* en un único paso.
- **tm**: destaca por tener soporte *back-end* de base de datos integrada, gestión avanzada de metadatos y soporte nativo para leer en varios formatos de archivo.
- **tokenizers**: incluye *tokenizadores* de palabras, oraciones, párrafos, *n*-gramas, *tweets*, expresiones regulares, así como funciones para contar caracteres, palabras y oraciones, y para dividir textos más largos en documentos separados, cada uno con el mismo número de palabras.
- **wordcloud**: permite visualizar **nubes de palabras**. Las palabras más frecuentes aparecen en mayor tamaño permitiendo de un vistazo obtener las palabras clave del texto.
- **quanteda**: maneja **matrices de documentos-términos** y destaca en tareas cuantitativas como recuento de palabras o sílabas.
- **syuzhet**: incluye distintas funciones que facilitan el análisis de textos, en particular el *análisis de sentimientos* de textos literarios.
- **gutenbergr**: almacena las obras del proyecto Gutenberg²; muy útil si se quieren analizar textos literarios.

²Proyecto desarrollado por Michael Hart en 1971 para crear una biblioteca de libros electrónicos gratuitos, y accesibles en internet, a partir de libros en soporte físico, generalmente de dominio público. Cuenta con más de 50 000 libros.

38.5. Ejemplo de aplicación

38.5.1. Declaración institucional del Estado de Alarma 2020

La “Declaración institucional del presidente del Gobierno anunciando el Estado de Alarma en la crisis del coronavirus” (en adelante, “la Declaración”), dada en La Moncloa el 13 de marzo de 2020 es el objeto de análisis. Esta se puede encontrar en el paquete CDR que acompaña este libro. Se le van a aplicar las operaciones y técnicas mencionadas en la Sec. 38.2.

```
library("CDR")
data("declaracion")
```

38.5.2. Segmentación en palabras y oraciones

Las primeras tareas del análisis son la preparación, limpieza y segmentación o tokenización de los textos, como se vió en las Sec. 38.2.1 y 38.2.2. A continuación, se verá una segmentación en palabras individuales. La función `tokenize_words()` del paquete `tokenizers` prepara el texto convirtiéndolo a minúsculas, elimina todos los signos de puntuación y finalmente segmenta el texto en palabras.

```
library("tokenizers")
palabras <- tokenize_words(declaracion)
tokenizers::count_words(declaracion)
#> [1] 922
```

Con la última sentencia se obtiene la longitud de la Declaración, el número de palabras utilizadas: 922.

La frecuencia de cada palabra se puede obtener y presentar con el código de abajo. La primera sentencia crea la tabla de frecuencias, la tercera la transforma en el tipo `tibble`, creando la columna recuento, y ordena la tabla de forma descendente, de mayor a menor frecuencia.

```
library("tidyverse")
tabla <- table(palabras[[1]])
( tabla <- tibble(palabra = names(tabla),
                  recuento = as.numeric(tabla)) |>
    arrange(desc(recuento)) )
#> # A tibble: 390 x 2
#>   palabra      recuento
#>   <chr>        <dbl>
#> 1 de            43
#> 2 y             41
#> 3 la            35
#> 4 a             31
#> 5 los           26
```

```
#> 6 en      22
#> 7 que     20
#> 8 el      17
#> 9 al      14
#> 10 para   14
#> # ... with 380 more rows
```

En la primera fila de la salida se indican las dimensiones de la `tibble`, por lo que se puede ver que en esta Declaración hay 390 “palabras” distintas (considera los números como palabras).

El resultado son las palabras más utilizadas en el texto, que, como puede apreciarse, son palabras vacías. Esto no debería sorprender porque, como ya se ha visto, estas palabras son las más frecuentes. En la siguiente Sección, se verá cómo eliminarlas para obtener datos con valor informativo.

Para otras formas de segmentar el texto (oraciones, párrafos, *tweets*, etc.): véase `?tokenize_words`. Por ejemplo, para segmentar en oraciones:

```
oraciones <- tokenize_sentences(declaracion)
count_sentences(declaracion)
#> [1] 44
```

Las tres primeras oraciones y la última se obtienen con el siguiente código.

```
oraciones[[1]][1:3] # primeras 3 oraciones
#> [1] "Buenas tardes."
←
#> [2] "Estimados compatriotas."
←
#> [3] "En el día de hoy, acabo de comunicar al Jefe del Estado la celebración, mañana,
← de un Consejo de Ministros extraordinario, para decretar el Estado de Alarma en
← todo nuestro país, en toda España, durante los próximos 15 días."
oraciones[[1]][count_sentences(declaracion)] # última oración
#> [1] "Buenas tardes."
```

También podría medirse la longitud de cada oración, en número de palabras, normalmente para comparaciones con otros textos. Para ello hay que separar cada oración en palabras y obtener la longitud de cada oración, con la función `sapply()`, que puede verse en la Fig. 38.1.

```
palabras_oracion <- tokenize_words(oraciones[[1]])
longitud_o <- sapply(palabras_oracion, length)
head(longitud_o)
#> [1] 2 2 39 33 33 32
```

38.5. Ejemplo de aplicación

651

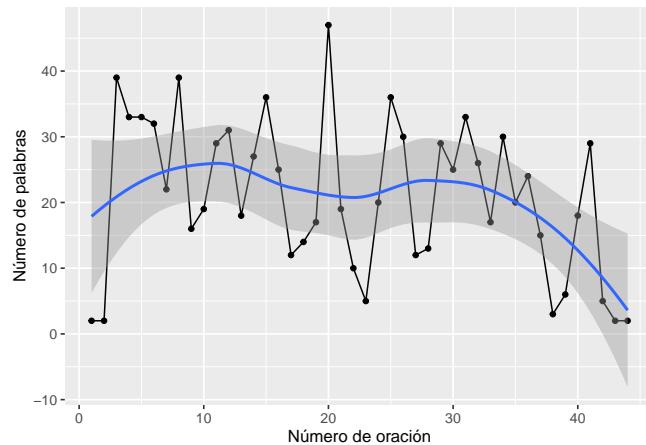


Figura 38.1: Número de palabras en cada oración de la Declaración

38.5.3. Análisis exploratorio

38.5.3.1. Eliminación de palabras vacías

Se llevará a cabo con el paquete `stopwords`, que contiene listas de *palabras vacías* en diferentes idiomas. Para el ejemplo, se define una tabla con la misma estructura que la tabla de la Declaración con las 308 palabras vacías españolas que tiene el paquete:

```
library("stopwords")
tabla_stopwords <- tibble(palabra = stopwords("es"))
```

La siguiente sentencia ‘limpia’ la tabla de la Declaración quitando las palabras vacías españolas. Además, se hace uso de la función `kable()` para una visualización más sofisticada de la tabla (con la longitud que se desee):

```
tabla <- tabla |> anti_join(tabla_stopwords)
knitr::kable(tabla[1:10],  
            caption = "Palabras más frecuentes (sin palabras vacías)")
```

El resultado, Tabla 38.1, se puede considerar el primer análisis léxico con valor informativo: la palabra más frecuente es *virus*, seguida de *recursos* y *social*. Se podría ver que en total hay 319 palabras no vacías distintas.

El método de eliminar palabras con el paquete `stopwords` no es perfecto. Por ejemplo, *va* y *cada* (posiciones 9 y 10 de la tabla) no son muy informativas. En estos casos, como se ha visto antes, se pueden utilizar listas de palabras vacías de otros paquetes como, por ejemplo `tidytext` o `tokenizers` o el listado en español propuesto por [Fradejas Rueda \(2022\)](#), o pueden confeccionarse listas *ad hoc*.

Tabla 38.1: Palabras más frecuentes (sin palabras vacías)

palabra	recuento
virus	9
recursos	7
social	5
alarma	4
conjunto	4
emergencia	4
españa	4
semanas	4
va	4
cada	3

38.5.3.2. Nubes de palabras

Una manera habitual de mostrar la información de forma visual es con las denominadas **nubes de palabras**, acudiendo a la función `wordcloud()` del paquete con el mismo nombre. Al tener dicha función un componente aleatorio, se fija con `set.seed()` (para la reproducibilidad del gráfico por parte del lector).

```
set.seed(12)
library("wordcloud")
wordcloud(tabla$palabra, tabla$recuento,
          max.words = 50, colors = rainbow(3))
```



Figura 38.2: Nube de palabras más frecuentes de la Declaración

El resultado se muestra en la Fig. 38.2. Como se puede observar, el tamaño de letra de la

palabra, y en este caso también el color, están relacionados con su frecuencia.

38.5.4. Análisis de sentimientos y detección de emociones

38.5.4.1. Lexicón bing

El diccionario `bing`, como se ha visto en la Sec. 38.3, es uno de los repertorios léxicos para el *análisis de sentimientos* que se pueden encontrar en **R**. Es un diccionario de polaridad (positiva/negativa) cuyo idioma original es el inglés. Se puede obtener con la función `get_sentiments()` del paquete `tidytext`. Contiene 2005 palabras positivas y 4781 palabras negativas, por lo que hay un marcado sesgo hacia la polaridad negativa.

Para ilustrar el uso de `bing`, se ha traducido al inglés (automáticamente) la Declaración. A continuación se carga el texto y se genera el objeto `tabla`, replicando el procedimiento descrito arriba de preparación, limpieza, segmentación en palabras, eliminación de palabras vacías (obviamente, en idioma inglés).

```
data("EN_declaracion")
tabla <- table(tokenize_words(EN_declaracion)[[1]])
tabla <- tibble(word = names(tabla),
                 recuento = as.numeric(tabla))
tabla <- tabla |> anti_join(tibble(word=stopwords("en"))) |>
  arrange(desc(recuento))
```

Los sentimientos positivos de la Declaración se pueden obtener con:

```
library("tidytext")
pos <- get_sentiments("bing") |>
  dplyr::filter(sentiment=="positive")
pos_EN <- tabla |> semi_join(pos)
knitr::kable(pos_EN)
```

Análogamente, se pueden obtener los sentimientos negativos. Las siete palabras más frecuentes de cada tipo que aparecen en la Declaración se presentan conjuntamente en la Tabla 38.2.

38.5.4.2. Lexicón NRC

Para poder observar las similitudes y diferencias en el análisis según el diccionario elegido, se aplica también NRC a la Declaración (véase la Tabla 38.2).

Tabla 38.2: Palabras más frecuentes de la Declaración utilizando bing y NRC

	positivas bing	fr	negativas bing	fr	positivas NRC	fr	negativas NRC	fr
extraordinary	6		virus	9	resources	7	virus	9
protect	4		alarm	4	extraordinary	6	alarm	4
work	4		emergency	4	protect	4	emergency	4
like	3		vulnerable	3	maximum	3	government	3
decisive	2		difficult	2	public	3	discipline	2
good	2		hard	2	council	2	avoid	1
adequate	1		unfortunately	2	good	2	combat	1

Con el léxico NRC pueden detectarse emociones. La misma palabra puede tener asociada distintas emociones/sentimientos. En la Fig. 38.3 se puede observar la dispar frecuencia de palabras de cada tipo:

```
emo <- get_sentiments("nrc")
emo |> ggplot(aes(sentiment)) +
  geom_bar(aes(fill=sentiment), show.legend = FALSE)
```

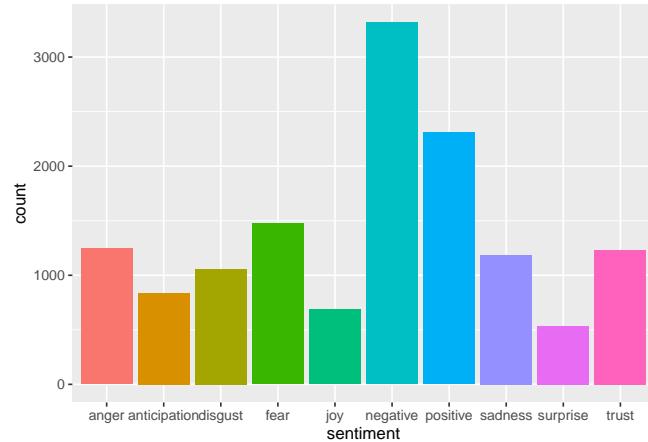


Figura 38.3: Gráfico de barras con la frecuencia de las emociones del lexicón NRC

El análisis de sentimientos y la detección de emociones de la Declaración mediante NRC se puede realizar con el siguiente código, mediante el cual se obtiene la tabla de frecuencias por emociones y sentimientos:

38.5. Ejemplo de aplicación

655

```
emo_tab <- tabla |> inner_join(emo)
head(emo_tab, n=7)
#> # A tibble: 7 x 3
#>   word      recuento sentiment
#>   <chr>      <dbl> <chr>
#> 1 virus        9 negative
#> 2 resources    7 joy
#> 3 resources    7 positive
#> 4 resources    7 trust
#> 5 extraordinary 6 positive
#> 6 alarm         4 fear
#> 7 alarm         4 negative
```

Como se ha mencionado antes, algunas palabras tienen asociados distintos sentimientos, por ejemplo, *resources*. La información de la tabla se puede visualizar bien con un gráfico de barras (Fig. 38.4) bien con nubes de palabras (Fig. 38.5).

```
emo_tab |>
  dplyr::count(sentiment) |>
  ggplot(aes(x=sentiment, y=n)) +
  geom_bar(stat = "identity", aes(fill=sentiment), show.legend = FALSE) +
  geom_text(aes(label = n), vjust=-0.25)
```

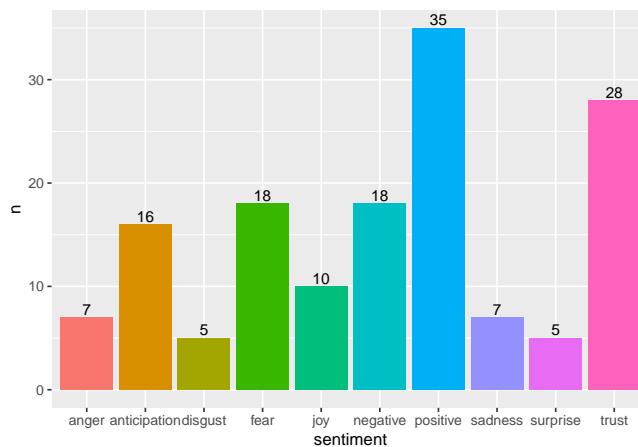


Figura 38.4: Frecuencia de emociones de la Declaración utilizando NRC

Entre las distintas opciones para dibujar nubes de palabras para el análisis de sentimientos es interesante la que se obtiene con el paquete **syuzhet** dado que permite visualizar las palabras agrupadas por emociones. Su obtención requiere distintos pasos en los que primero las palabras se agrupan por emoción y después se organizan en una **matriz de documentos** con la función **TermDocumentMatrix()** del paquete **tm**. Finalmente la función **comparison.cloud()** permite

visualizar el gráfico (tiene distintos argumentos opcionales que admiten distintas posibilidades). En el ejemplo que figura a continuación solo se han escogido tres emociones³:

```
library("syuzhet")
palabras_EN2 <- get_tokens(EN_declaracion)
emo_tab2 <- get_nrc_sentiment(palabras_EN2, lang = "english" )
emo_vec <- c(
  paste(palabras_EN2[emo_tab2$anger > 0], collapse = " "),
  paste(palabras_EN2[emo_tab2$anticipation > 0], collapse = " "),
  paste(palabras_EN2[emo_tab2$disgust > 0], collapse = " "))
library("tm")
corpus <- Corpus(VectorSource(emo_vec))
TDM <- as.matrix(TermDocumentMatrix(corpus))
colnames(TDM) <- c('anger', 'anticipation', 'disgust')
set.seed(1)
comparison.cloud(TDM, random.order = FALSE,
                 colors = c("firebrick", "forestgreen", "orange3"),
                 title.size = 1.5, scale = c(3.5, 1), rot.per = 0)
```



Figura 38.5: Nube de palabras de tres emociones NRC seleccionadas

38.5.5. *N-gramas*

El siguiente código muestra la obtención de *n-gramas* con `tokenizers`.

³Se deja al lector el análisis de la Declaración con más emociones, en castellano, etc.

38.5. Ejemplo de aplicación

657

```
bigramas <- tokenize_ngrams(declaracion, n = 2,
                             stopwords = tabla_stopwords$palabra)
head(bigramas[[1]], n = 3)
#> [1] "buenas tardes"           "tardes estimados"      "estimados compatriotas"
trigramas <- tokenize_ngrams(declaracion, n = 3,
                             stopwords = tabla_stopwords$palabra)
head(trigramas[[1]], n = 3)
#> [1] "buenas tardes estimados" "tardes estimados compatriotas"
#> [3] "estimados compatriotas día"
```

Se ha procedido a eliminar de los bigramas y trigramas aquellas combinaciones con al menos una palabra vacía (*stopword*).

Se procede ahora a obtener los bigramas con `tidytext`. Para el resto de *n-gramas* el procedimiento es análogo, haciendo las modificaciones oportunas. En el último paso se ordenan por frecuencia (de mayor a menor):

```
declara2 <- tibble(texto = declaracion)
bigramas <- declara2 |>
  unnest_tokens(bigram, texto, token = "ngrams", n = 2) |>
  dplyr::count(bigram, sort = TRUE)
bigramas[1:5, ]
#> # A tibble: 5 x 2
#>   bigram          n
#>   <chr>        <int>
#> 1 todos los      6
#> 2 de la         5
#> 3 de los         5
#> 4 del estado     5
#> 5 estado de       5
```

Una forma de eliminar las palabras vacías es:

```
bigramas_limpios <- bigramas |>
  tidyrr::separate(bigram, c("word1", "word2"), sep = " ") |>
  dplyr::filter(!word1 %in% tabla_stopwords$palabra) |>
  dplyr::filter(!word2 %in% tabla_stopwords$palabra) |>
  tidyrr::unite(bigram, word1, word2, sep = " ")
bigramas_limpios[1:5, ]
#> # A tibble: 5 x 2
#>   bigram          n
#>   <chr>        <int>
#> 1 autoridades sanitarias    2
#> 2 buenas tardes            2
#> 3 disciplina social       2
#> 4 haga falta               2
#> 5 ministros extraordinario 2
```

38.5.5.1. Significado y contexto

Como se ha visto en la Sec. 38.2.2, con los **n-gramas** se puede hacer un análisis de colocaciones para extraer los distintos significados y valores informativos a partir del contexto. En este caso, se puede ver cómo la palabra *atender* cambia de sentido cuando va precedida de *no* o *sin*. A continuación, se filtran los bigramas cuya primera palabra es *no*:

```
bigramas_no <- bigramas |>
  tidyr::separate(bigram, c("word1", "word2"), sep = " ") |>
  dplyr::filter(word1 == "no") |>
  dplyr::count(word1, word2, sort = TRUE)
bigramas_no
#> # A tibble: 3 x 3
#>   word1 word2     n
#>   <chr>  <chr> <int>
#> 1 no     atiende    1
#> 2 no     cabe      1
#> 3 no     es        1
```

Estos resultados se pueden utilizar para el análisis de sentimientos y la detección de emociones.

38.5.6. Análisis de redes

En esta Sección se proporcionan las instrucciones para realizar un **análisis básico de redes** (ver Cap. 39), utilizando los paquetes **igraph** y **ggraph**. Dada la corta extensión de la Declaración no es posible obtener conclusiones. En la Fig. 38.6 se pueden ver los gráficos de redes de bigramas, tanto sin palabras vacías como con ellas.

```
library("igraph")
library("ggraph")
set.seed(1)
graf_bigramas_1 <- bigramas_limpios |>
  tidyr::separate(bigram, c("first", "second"), sep = " ") |>
  dplyr::filter(n > 1) |>
  graph_from_data_frame()
g1 <- ggraph(graf_bigramas_1, layout = "fr") +
  geom_edge_link(arrows = arrow(length = unit(4, 'mm'))) +
  geom_node_point(size=0) +
  geom_node_text(aes(label = name))
graf_bigramas <- bigramas |>
  tidyr::separate(bigram, c("first", "second"), sep = " ") |>
  dplyr::filter(n > 2) |>
  graph_from_data_frame()
g2 <- ggraph(graf_bigramas, layout = 'fr') +
  geom_edge_link0() +
  geom_node_point(size=0) +
  geom_node_label(aes(label = name))
```

38.5. Ejemplo de aplicación

659

```
library("patchwork")
g1+g2
```

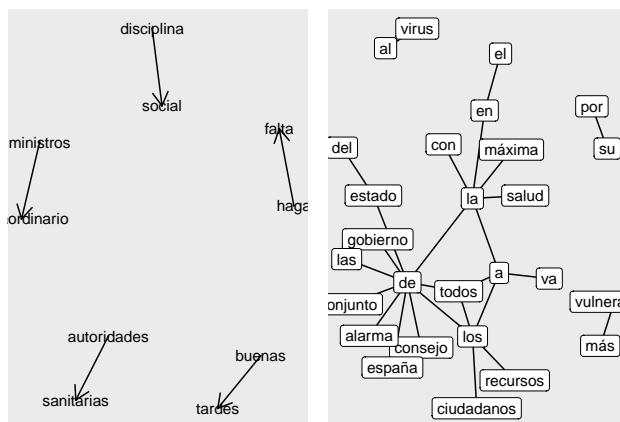


Figura 38.6: Redes de bigramas sin palabras vacías y con ellas

Resumen

En este capítulo se introduce al lector en la minería de textos, en particular:

- Se presentan los conceptos y tareas fundamentales de este ámbito, así como sus principales campos de aplicación. Se pone de relieve la importancia de la preparación de los datos y su segmentación (a distintos niveles) para obtener buenos resultados, acordes con el objetivo de la investigación.
- Se muestra el uso de **R** para el análisis de textos y de sentimientos.
- Se presenta un ejemplo de aplicación para ilustrar las técnicas de minería de textos.
- Se mencionan otros análisis plausibles de minería de textos, como la estilometría o el modelado de temas (véase el Cap. 59).

Capítulo 39

Análisis de grafos y redes sociales

José J. Galán

Universidad Complutense de Madrid

39.1. Introducción

El origen de la teoría de grafos se debe al problema de los siete puentes de Königsberg (Paul Euler, 1736), que es considerado el primer artículo sobre teoría de grafos. El problema se centra en la ciudad Königsberg en Prusia, ahora Kaliningrado (Rusia), donde existen varios puentes y el problema plantea trazar una ruta que cruce todos los puentes una única vez (ver Fig. 39.1). Euler mediante el uso de grafos demostró que no era posible.

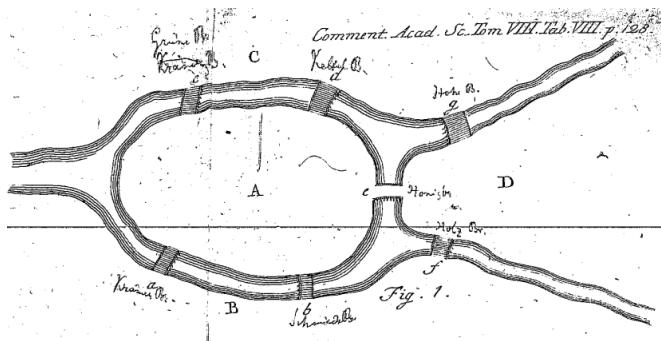


Figura 39.1: Siete puentes de Königsberg, Euler (1736).

Pero, ¿qué relación tiene un concepto creado en 1736, el de grafo, con algo tan reciente como las redes sociales?. Informalmente se puede hablar de las redes sociales (RRSS) como las relaciones existentes entre personas, un hilo invisible que une a las personas en relación con algo que tienen

en común. En algunos casos es muy evidente porque se crean grupos específicos de personas que comparten una afición y en otros casos es menos evidente porque, por ejemplo, pueden compartir un amigo en común sin saberlo. Estos hilos “invisibles” se unen y forman una red que se puede representar como un grafo, el mismo concepto de grafo que describió Euler, y que permite establecer diferentes caminos para unir a las personas que forman la red.

39.2. Teoría de grafos

Informalmente se puede decir que un **grafo** es un conjunto de **nodos** (vértices) que pueden estar unidos por **aristas** (enlaces).

Si se piensa en cada nodo como una persona y en cada arista como la relación que los une, entonces se podría representar mediante grafos una **red social** (ver Fig. ??).

Antes de ver el primer ejemplo se muestran las librerías necesarias en este capítulo.

```
library("igraph")
library("CDR")
```

En este primer ejemplo en **R** se puede observar cómo se obtienen los datos de una red social. El conjunto **datos_facebook** está incluido en el paquete **CDR**. A continuación, se representan las relaciones de los miembros que la componen mediante un grafo.

```
grafo_facebook <- graph.data.frame(datos_facebook, directed = F)
plot(grafo_facebook, vertex.label = NA, vertex.size = 8)
```

Más formalmente una **red social** puede modelizarse con una estructura de red invisible (relación familiar, amistad, trabajo ...) que une mediante relaciones a distintos actores a través de sus intereses o valores comunes, estableciendo una relación personal entre individuos o grupos de individuos conectados.

Existen distintos grafos dependiendo de las características de la red social representada, algunos ejemplos de estos grafos son:

- El grafo de amistad (ver Fig. @ref{fig:grafo-comunidades2}), donde cada nodo representa una persona y la arista conecta dos personas si dentro de la red social son amigos.

```
amistades <- data.frame(
  persona = c("A", "B", "C", "D", "E"),
  amigo = c("B", "C", "A", "E", "A"))
grafo_amistades <- graph_from_data_frame(amistades)
```

- El grafo de comunidades (ver Fig. @ref{fig:grafo-comunidades2}), donde también cada nodo representa una persona y la arista les conecta si dentro de la red social pertenecen a la misma comunidad, entendiendo por comunidad un grupo de individuos que comparten intereses o características en común.

39.3. Elementos de un grafo

663

```
comunidades <- data.frame(
  persona = c("A", "B", "C", "D", "E"), comunidad = c("1", "2", "1", "2", "1") )
grafo_comunidades <- graph_from_data_frame(comunidades)
```

```
par(mfrow=c(1,2))
plot(grafo_amistades, vertex.label = V(grafo_amistades)$name, main="Grafo amistades")
plot(grafo_comunidades, vertex.label = V(grafo_comunidades)$name, main="Grafo
→ comunidades")
```

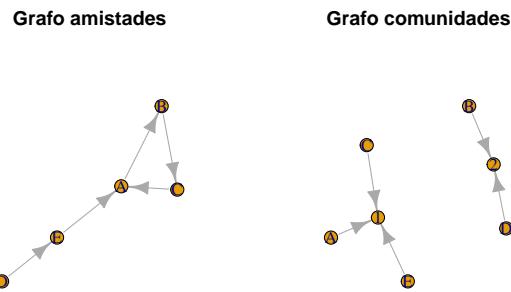


Figura 39.2: Grafo de amistades y comunidades.

39.3. Elementos de un grafo

El análisis de RRSS mediante la teoría de grafos requiere conocer previamente una serie de conceptos básicos ([Perez Sola, 2021](#)) que se enumeran a continuación.

- Los **vértices** representan nodos que se unen mediante aristas. En una red social cada vértice representa una de las personas de dicha red, unidas en ocasiones por intereses comunes a otras.
- Las **aristas** son las relaciones que unen los nodos. Son **dirigidas** (Fig.39.3) si tienen un sentido definido y **no dirigidas** (Fig. 39.3) en caso contrario.

El siguiente código computa un grafo dirigido:

```
grafodirigido <- graph.data.frame(datos_grafos, directed = T)
```

Para un grafo no dirigido, basta con especificar `directed = F` en la función `graph.data.frame()`:

```
grafonodirigido <- graph.data.frame(datos_grafos, directed = F)
```

En otras RRSS, como **LinkedIn**, las aristas podrían representar la relación que une las personas. Las personas forman parte de un grupo con intereses comunes, formando un grafo no dirigido. Pero también se pueden seguir a alguien sin necesariamente ser seguido; en ese caso las relaciones se pueden representar como un grafo dirigido.

- Un **grafo** es un conjunto de vértices y aristas que se puede representar mediante $G = (V, E)$, donde V es el conjunto de nodos o vértices del grafo y E es un conjunto de pares de vértices llamado arista, arco o edge.

Se presenta un grafo “sencillo” en la Fig. 39.3 (sólo se indican las aristas y **R** es capaz de interpretar los vértices), sobre el cual se explicará la matriz de adyacencia, grado y camino:

```
grafo <- graph(edges = c(1, 2, 1, 3, 1, 4, 2, 4, 3, 5, 4, 5))
```

```
par(mfrow=c(1,3))
plot(grafodirigido, vertex.label = V(grafodirigido)$name, main="Grafo dirigido")
plot(grafonodirigido, vertex.label = V(grafonodirigido)$name, main="Grafo no dirigido")
plot(grafo, main="Grafo \"sencillo\"")
```

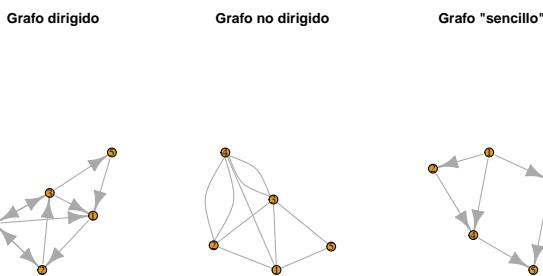


Figura 39.3: Grafo dirigido, no dirigido y sencillo.

39.3. Elementos de un grafo

665

- La información recogida en un grafo también se puede expresar mediante números, organizados en una matriz denominada **matriz de adyacencia**, $A_{n \times n}$, lo que facilita los cálculos computacionales en grandes redes. Cada entrada de la matriz, a_{ij} , indica el número de aristas que comparten los vértices (o nodos) i-ésimo y j-ésimo (si no existe relación entre ellos, entonces $a_{ij} = 0$); cada fila de la matriz indica el número de aristas que comparte el vértice i-ésimo con cada uno de los otros vértices. La suma de todas las entradas a_{ij} de una fila es el grado del vértice correspondiente. En los grafos no dirigidos $A_{n \times n}$ es simétrica, ya que si el vértice o nodo 1 conecta con el 2, entonces el 2 también conecta con el 1. En los grafos dirigidos, donde cada arista tiene una orientación, esto no tiene por qué ocurrir: el vértice o nodo 1 puede conectar con el 1 pero no al revés. En este tipo de grafos la matriz de adyacencia no es simétrica.
- El **grado** o valencia de un nodo x es el numero de aristas que concurren en dicho nodo, y se representa mediante $grado(x)$, $g(x)$ o $gr(x)$, lo cual en R se calcula con la función `degree`, siendo un vértice de grado 0 un vértice aislado. En un grafo G hay un grado máximo $\Delta(G)$ y un grado mínimo $\delta(G)$; el grado del grafo, $g(G)$, es la suma de los grados de todos sus vértices. En una red social representa el número de relaciones que existen; en una red social como Facebook podría significar conocer cuántos amigos tiene cada persona.

A continuación se muestra la matriz de adyacencia del grafo visto previamente:

```
matriz_adyacencia <- get.adjacency(grafo, sparse = FALSE)
matriz_adyacencia
#>      [,1] [,2] [,3] [,4] [,5]
#> [1,]     0    1    1    1    0
#> [2,]     0    0    0    1    0
#> [3,]     0    0    0    0    1
#> [4,]     0    0    0    0    1
#> [5,]     0    0    0    0    0
```

Ahora muestra el grado del mismo grafo:

```
degree(grafo)
#> [1] 3 2 2 3 2
```

- Un **camino** es un conjunto de aristas no recursivas. Entre dos vértices puede existir más de un camino, además puede haber varios y se puede incluir el mismo vértice en el camino más de una vez. Evidentemente, siempre habrá un **camino más corto**: que será aquel que menos aristas ha recorrido. Si entre todos los pares de vértices existe un camino, entonces el grafo se denomina *conexo*.

El siguiente código se utiliza para mostrar el camino más corto entre los vértices 2 y 5 de grafo utilizado de ejemplo.

```
# Camino más corto entre el nodo 2 y el 5
caminos <- get.shortest.paths(grafo, from = "2", to = "5")
V(grafo)[caminos$vpath[[1]]]
#> + 3/5 vertices, from 300ab8b:
#> [1] 2 4 5
```

39.4. Procedimiento con R: el paquete igraph

Existen diversos paquetes en **R** para representar grafos, pero el más utilizado y popularizado, por sencillez y eficacia, es **igraph** (Csardi and Nepusz, 2006). Se trata de un paquete que permite crear y manipular grafos para analizar redes en **R** de forma muy sencilla (Fig. @ref{fig:grafobasico}).

A continuación se muestra un sencillo ejemplo sobre cómo crear un grafo dirigido con la librería **igraph**

```
nodes <- data.frame("nodos" = c("A", "B", "C", "D", "E"))
edges <- data.frame(
  "from" = c("A", "C", "B", "A", "A", "A"),
  "to" = c("B", "D", "C", "C", "D", "E"))
red <- graph_from_data_frame(edges, directed = TRUE, vertices = nodes)
plot(red, vertex.size = 50)
```

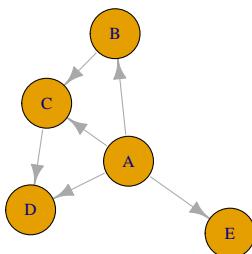


Figura 39.4: Ejemplo de grafo con ‘igraph’

Para crear un grafo, Fig. @ref{fig:grafobasico}, a partir de un *data frame* se ha usado la función **graph_from_data_frame()** con los siguientes argumentos:

graph_from_data_frame(edges, directed = TRUE, vertices = nodes) donde:

39.4. Procedimiento con **R**: el paquete *igraph*

667

- **edges**: es un data frame donde las dos primeras columnas representan una lista de aristas.
- **directed**: es un valor lógico que indica si es un grafo dirigido o no dirigido.
- **vertices** es un data frame con los valores de los vértices o NULL.

El siguiente código muestra las relaciones entre los actores de dos películas. **nodes** contiene el nombre de cada actor y su descripción, es imprescindible que los nombres que más adelante se introducen en edges existan en nodes. Al mismo tiempo no es obligatorio declarar los nodos ya que pueden ser extraídos de las relaciones. **edges** contiene la relaciones, **from** y **to**, además de la película donde coinciden. Siendo este último dato descriptivo y no necesario.

```
nodes <- data.frame("actores" = c(
  "Jim Carrey", "Arnold Swarzenegger", "George Clooney",
  "Cameron Diaz"),
  "descripcion" = c("actor", "actor", "actor", "actriz"))
edges <- data.frame(
  "from" = c("Jim Carrey", "Jim Carrey", "George Clooney", "Jim Carrey"),
  "to" = c(
    "Arnold Swarzenegger", "George Clooney", "Arnold Swarzenegger",
    "Cameron Diaz"),
  "pelicula" = c(
    "Batman y Robin", "Batman y Robin",
    "Batman y Robin", "La mascara"))
red <- graph_from_data_frame(edges, directed = F, vertices = nodes)
plot(red, vertex.size = 50)
```

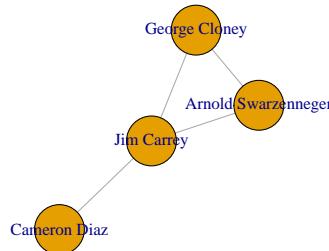


Figura 39.5: Grafo representativo de la relación de actores respecto a películas

En la Fig. @ref{fig:grafoactores} se puede observar como el actor Jim Carrey tuvo relación con todos los actores de la red propuesta, mientras que la actriz Cameron Diaz solo participó con uno de ellos (el propio Jim Carrey).

39.5. Análisis de influencia en un grafo aplicado a RRSS

Existen paquetes para obtener información de distintas RRSS; por ejemplo, en **R** se puede utilizar el paquete **Rfacebook** para conectarse a **Facebook** y obtener información de los contactos existentes. Para ello será necesario activar la API, Interfaz de Programación de Aplicaciones, desde <https://developers.facebook.com>. La información necesaria se puede encontrar en su página web <https://developers.facebook.com/docs>. Para ilustrar un ejemplo didáctico, sin que el lector necesite conocimientos de desarrollo para descargar los datos, se ha generado un fichero Excel que simula la relación entre amigos de una red social como podría ser Facebook generando un grafo dirigido, tipo de grafo habitual en este tipo de redes.

Se incorporan los datos y se muestra su cabecera.

En primer lugar, se utiliza el siguiente código para recoger los datos de un fichero CSV, con dos columnas, separadas por un espacio. Cada columna mediante un número identificador representa una persona, la unión de estas dos personas es el resultado de una relación, estas relaciones pueden visualizarse con el siguiente código.

```
datos_facebook <- graph.data.frame(datos_facebook, directed = F)
#datos_facebook # descomentar para ver las relaciones
```

En esta ocasión se utiliza la función `graph.data.frame()` del paquete **igraph** para crear un objeto de tipo grafo, dirigido en este caso, a partir de un data frame en **R**. Seguidamente, mediante `plot()` se muestra el grafo al mismo tiempo que se establecen sus propiedades. Ver Fig. 39.6. Nótese que la estructura de los datos de entrada para construir el grafo es diferente a la función `graph_from_data_frame()` vista en el apartado anterior, pero ambas cumple el objetivo de construir un grafo y por ello se presentan ambas opciones.

```
grafo_facebook <- graph.data.frame(datos_facebook, directed = T)
plot.igraph(grafo_facebook,
           layout = layout.fruchterman.reingold,
           vertex.label = NA, vertex.label.cex = 1, vertex.size = 3, edge.curved = TRUE
)
```

Siguiendo con el ejemplo, se ven las relaciones de la red social, se puede observar que el número de aristas que concurren en cada vértice indica el número de personas con las que se relaciona el individuo, representado por dicho vértice. Indica, por tanto, el número de relaciones que mantiene dicho individuo.

```
table(degree(grafo_facebook))
#>
#>   9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
#>   1  2  4  4  7  4  8 14 19 15 14 19 22 15 11 13  9  5  3  3  4  1  4
```

Ahora sobre el mismo ejemplo se personalizan los datos. Las RRSS son enormes y, por tanto, es útil centrarse en una subred para estudios concretos. A modo de ejemplo, para focalizar este caso

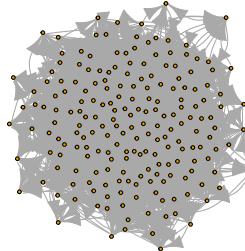


Figura 39.6: Aplicación de ‘igraph’ a RRSS.

de estudio el ejemplo se centrará en aquéllos cuyo grado sea igual o superior a 26, asignándoles su nombre.

```
bad_network <- V(grafo_facebook)[degree(grafo_facebook) <= 26]
grafo_facebook <- delete.vertices(grafo_facebook, bad_network)
V(grafo_facebook)$name <- c(
  "Gema", "Patricia", "Ramon", "José", "Maria", "Ángeles",
  "Gabriel", "Javier", "Victor", "Leonor", "Ana", "Isabel", "Cristóbal", "Rosa",
  ↪ "Aurora"
)
plot(grafo_facebook, vertex.size = 20)
```

En la Fig. 39.7 se pueden ver las relaciones entre las personas incluidas en el grafo (a quién siguen y por quiénes son seguidas). Por ejemplo, a Gema no la sigue nadie. Leonor, otro caso extremo igual que Gema, es seguida por Aurora, Ángeles y Gabriel, pero ella no sigue a nadie.

39.5.1. Centralidad

Un grafo no tiene una centralidad real porque no tiene coordenadas (Easley, 2010), pero existen distintas medidas de centralidad que, en una red/grafo social, permitirán identificar el networking social de cada individuo, es decir, su influencia.

En teoría de grafos y análisis de RRSS, el concepto de **centralidad** refiere a la importancia o prominencia de los vértices o nodos en un determinado grafo o red social. En el caso de una red social de amigos, como, por ejemplo, Facebook, la centralidad de un nodo (persona) representa el número de amigos que tiene.

Son innumerables las medidas de **centralidad** (generalmente normalizadas o estandarizadas) que pueden encontrarse en la literatura sobre la cuestión para determinar y comparar, de forma

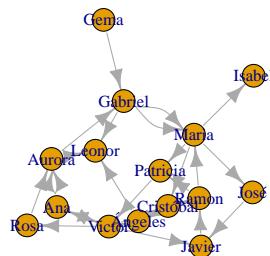


Figura 39.7: Aplicación de ‘Igraph’ a RRSS.

cuantitativa, la importancia relativa de un nodo en el conjunto de la red. La **centralidad** no es un atributo intrínseco de los nodos, sino un atributo estructural: un valor que depende de las relaciones de dicho nodo con los demás de la red. Generalmente, el nodo central suele tener la mayor centralidad, mientras que la mínima suele corresponder a los nodos periféricos.

En los grafos dirigidos, cuantas más aristas reciba un nodo (persona) más personas están intentando interactuar con ella y más prestigio tendrá en la red. Pero si la interacción hacia esta persona no es directa y se realiza a través de un camino más largo pasando por más nodos quiere decir que su influencia es elevada, ya que más personas han recibido esa influencia.

Existen diversas técnicas de obtener la centralidad ([Wasserman, 1995](#)), entre las que destacamos la centralidad por intermediación y la de vector propio por ser dos conceptos diferentes en el análisis de redes. Ahora se vera más en detalle como la primera mide la influencia de un nodo en la transmisión de información, mientras que la segunda es una medida mas amplia de la centralidad global de un nodo en un grafo.

- La técnica de **centralidad de intermediación (betweenness)** se basa en el número de caminos mínimos (camino más corto entre dos vértices en un grafo ponderado) en los que un nodo está involucrado. Por lo tanto, en una red social una persona tendrá mayor influencia cuanto mayor betweenness tenga, porque comunicará mucha información a través de los nodos de la red. Si puede llegar a un grupo grande, aunque sea a través de un nodo a quien nadie sigue, como Gema, puede alcanzar un gran nivel de viralización porque su información llegara hasta muchas personas.

El siguiente código muestra la centralidad mediante intermedio correspondiente a los nodos del ejemplo actual

39.5. Análisis de influencia en un grafo aplicado a RRSS

671

```
betweenness_centrality <- betweenness(grafo_facebook)
betweenness_centrality
#>      Gema Patricia Ramon José María Ángeles Gabriel Javier
#> 0.000000 33.500000 30.200000 10.500000 53.300000 7.000000 37.800000 27.200000
#>   Victor Leonor Ana Isabel Cristóbal Rosa Aurora
#> 42.700000 0.000000 9.333333 0.000000 6.000000 4.666667 23.800000
```

Maria es el nodo con mayor porcentaje, ello quiere decir que es quien tiene un mayor número de enlaces o conexiones con otros nodos, siendo el nodo de mayor importancia en términos de conectividad.

- La **Centralidad de vector propio (eigenvector)**. Se basa en la centralidad de los nodos con los que se relaciona. En concreto la centralidad de valor propio es proporcional a la suma de las centralidades de sus nodos vecinos. Se representa mediante $c_i = \lambda \sum_j a_{ij} c_j$, donde λ es la constante de proporcionalidad y a_{ij} es el valor de la fila i y la columna j de la matriz de adyacencia \mathbf{A} de la red social.

El siguiente código muestra la centralidad mediante vector propio correspondiente a los nodos del ejemplo actual

```
eigencentrality <- eigen_centrality(grafo_facebook)$vector
eigencentrality
#>      Gema Patricia Ramon José María Ángeles Gabriel Javier
#> 0.1618269 0.6090993 0.7768737 0.3148866 0.9257523 0.6787138 0.7180398 0.4714266
#>   Victor Leonor Ana Isabel Cristóbal Rosa Aurora
#> 1.0000000 0.4725741 0.7663131 0.2086397 0.7620632 0.3831565 0.7000984
```

Observamos que según esta centralidad, basada en el número de enlaces y conexiones con otros nodos importantes, el mayor porcentaje es obtenido por el nodo Victor. Este nodo es por tanto central en términos de conexión con otros nodos importantes de la red siendo identificado como el nodo líder o de mayor influencia de la red.

39.5.2. Detección de comunidades

En el análisis de una red social es importante detectar las distintas comunidades que la componen, entendiendo por comunidad un grupo de personas afines, gracias a las comunidades podremos estudiar los distintos grupos que lo forman en función de sus relaciones y afinidad (Missaouri, 2015).

Existen diversos algoritmos en **R** para detectar comunidades (Borgatti, 2022), pero ninguno ha demostrado que pueda actuar a la perfección con todos los grafos debido a la gran tipología que existe. Así, según el ejemplo anterior podemos detectar las siguientes comunidades:

- La **detección de comunidades mediante betweenness** tiene implementado en el paquete **igraph** un algoritmo para detectar comunidades por lo que a continuación se expone.

```
communities <- cluster_edge_betweenness(grafo_facebook)
# head(communities, 10) # descomentar para ver las comunidades
```

- La **detección de comunidades mediante walktrap** es también muy utilizada en RRSS, se basa en el concepto de que las caminatas aleatorias cortas permanecen a la misma comunidad y no esta basado en una medida de centralidad.

A continuación, se presenta el código del algoritmo walktrap, mediante el cual se detectan dos comunidades:

```
communities <- walktrap.community(grafo_facebook)
```

39.5.3. Representación con grafos

Ahora que se ha observado la detección de comunidades se realiza su representación final mediante grafos.

- El siguiente código muestra la red mediante un **grafo de tipo Eigenvector** donde se han detectado mediante este algoritmo tres comunidades. Según este método, no es tan importante que tengas muchos amigos, lo importante es que tus amigos sean muy influyentes (ver Fig. 39.8)

```
cl_g_network <- leading.eigenvector.community(grafo_facebook)
plot(cl_g_network, grafo_facebook,
     edge.arrow.size = 0.25, main =
     "Leading Eigenvector Community", vertex.size = 50
)
```

- A continuación el código que muestra la red mediante el **grafo de tipo betweenness** donde en la Fig.39.9 se observan siete comunidades.

```
betweenness_grafo <- edge.betweenness.community(grafo_facebook)
```

- Ahora se puede observar el código para obtener el **grafo de tipo Walktrap**. Se pueden observar dos comunidades detectadas con **Walktrap**, ver Fig.39.9

```
Walktrap_grafo <- walktrap.community(grafo_facebook, steps = 5, modularity = TRUE)
#Debe indicarse el largo de la caminata aleatoria, se recomienda usar 5 caminatas.
```

```
par(mfrow=c(1,2))
plot(betweenness_grafo, grafo_facebook, edge.arrow.size = 0.25, vertex.size = 50,
     main="Grafo Betweenness")
plot(Walktrap_grafo, grafo_facebook, edge.arrow.size = 0.25,
     vertex.label = (grafo_facebook)$name, vertex.size = 50, main="Grafo Walktrap")
```

39.5. Análisis de influencia en un grafo aplicado a RRSS

673

Leading Eigenvector Community

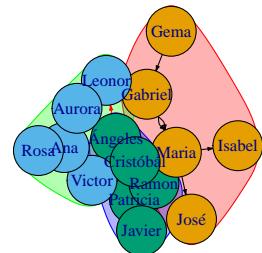
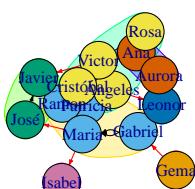


Figura 39.8: Aplicación de ‘Igraph’ a RRSS

Grafo Betweenness



Grafo Walktrap

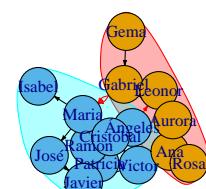


Figura 39.9: Grafo Betweenness y Walktrap.

39.6. Entorno social en el universo cinematográfico Marvel

Se va a analizar el Universo Cinematográfico de Marvel como una red social donde cada héroe tiene un grado de relación con otro. Se realiza este estudio utilizando los datos `marvel.edges` del paquete CDR que contiene dos columnas formateadas para representar la red social del Universo Marvel, donde la primera columna es el nombre de un personaje del Universo cinematográfico Marvel y la segunda el nombre de otro con el que coincide en alguna película, representado cada relación una arista entre dos nodos. Con estos datos se forma el grafo correspondiente a su red social, donde las coincidencias en la misma película entre héroes representan relaciones y cada héroe un nodo. En el siguiente código cargamos el fichero y formamos el `grafo_marvel`.

```
grafo_marvel <- graph.data.frame(marvel_edges, directed = F)
plot(grafo_marvel)
```



Figura 39.10: Grafo original sobre el Universo cinematografico Marvel

Son muchas las relaciones mostradas en la 39.10, por lo que a continuación se muestran aquellos héroes que tienen dos o más relaciones. Esto crea un grafo más visible, incluido un subgrafo de dos nodos (véase 39.11).

```
nodos_poca_realacion <- which(degree(grafo_marvel) < 2)
grafo <- delete.vertices(grafo_marvel, nodos_poca_realacion)
```

A continuación, se presenta el número de relaciones que tiene cada nodo, cada héroe, mediante la centralidad de grado en R usando la función `degree()` del paquete `igraph`. Se puede apreciar como Iron Man y Capitán América son quienes mayor número de relaciones tiene y por lo tanto quienes más gozan de popularidad e influencia.

```
grado_nodos <- degree(grafos)
#sort(grado_nodos) # descomentar para ver la centralidad de grado de los héroes.
```

Ahora, se analizan las comunidades que tiene, identificando cada una con un color distinto. En esta ocasión se utiliza el algoritmo Louvain, se considera el algoritmo más popular por su fácil interpretación, flexibilidad, alta calidad de las comunidades y eficiencia en tiempo de ejecución. El resultado podemos verlos en la [39.11](#).

```
comunidades <- cluster_louvain(grafo)
```

```

par(mfrow=c(1,2))
plot(grafo, vertex.label = V(grafo)$name, main="Relaciones de héroes")
plot(grafo, vertex.color = comunidades$membership, vertex.label = V(grafo)$name,
      main="Comunidades de héroes")

```

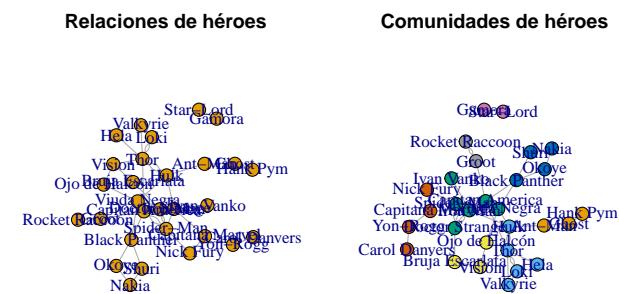


Figura 39.11: Grafo de relaciones y comunidades de héroes.

Por otra parte, se observan las comunalidades, es decir, grupos de personas que tienen algo en común, en este caso héroes que comparten escenas en películas. De esta manera recorremos el grafo anterior para mostrar por separado cada una de las comunidades encontradas, rápidamente se observa que la Comunidad 3 es la comunidad con más miembros por lo tanto es la comunidad más popular y la que posee un mayor interés compartido.

```
comunidades <- cluster_louvain(grafico)
num_comunidades <- length(unique(comunidades$membership))

for (i in 1:num_comunidades) {
  nodos_comunidad <- which(comunidades$membership == i)
  subgrafo <- induced_subgraph(grafico, nodos_comunidad)
}
```

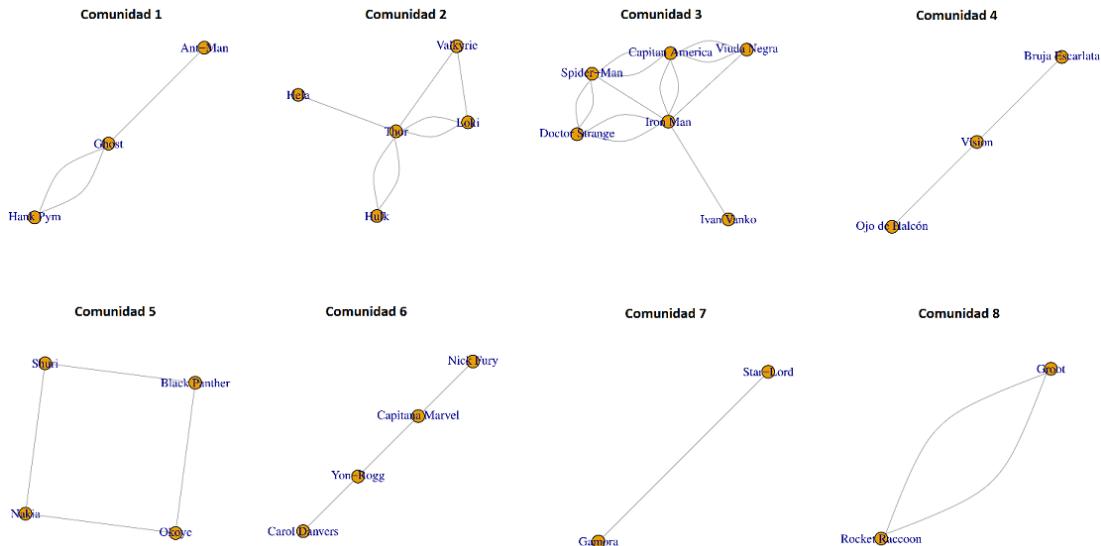


Figura 39.12: Distintas comunidades

A continuación, el análisis se centra en la comunidad con más relaciones (Fig. @ref{fig:grafo-relaciones-color}). Para ello, se identifica la comunidad más grande y se genera un subgrafo, mediante la función `induced_subgraph()` del paquete `igraph`, para su visualización.

```
comunidades <- cluster_louvain(grafo)
tamanos_comunidades <- table(comunidades$membership)
indice_comunidad_max <- which.max(tamanos_comunidades)
nodos_comunidad_max <- which(comunidades$membership == indice_comunidad_max)

subgrafo <- induced_subgraph(grafo, nodos_comunidad_max)
```

La Fig. @ref{fig:grafo-relaciones-color} muestra el tamaño de los héroes según el número de relaciones y con un color distinto por cada héroe, lo que hace que el gráfico final sea más llamativo y fácil de interpretar.

```
# Calcular el grado de cada nodo
grados <- degree(subgrafo)

# Ajustar el tamaño de los nodos proporcionalmente a su grado
tamanos <- 80 * grados / max(grados)

# Generar un vector de colores aleatorios
colores <- sample(colors(), vcount(subgrafo), replace = TRUE)
```

39.6. Entorno social en el universo cinematográfico Marvel

677

```
par(mfrow=c(1,2))
plot(subgrafo, main="(a)")
plot(subgrafo, vertex.color = colores, vertex.size = tamaños, main="(b)")
```

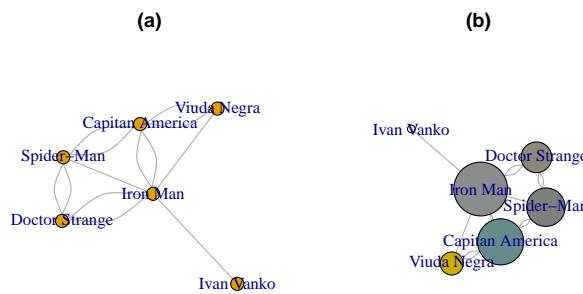


Figura 39.13: Grafo estandar con más relaciones (a) y grafo adaptado con color y tamaño (b)

Las comunidades pueden ser representadas por otros algoritmos. A continuación se representan los algoritmos ya comentados, Betweenness, el cual es útil cuando los nodos que conectan distintas comunidades son los más importantes y se quiere asegurar que se incorporen en una comunidad y el algoritmo Walktrap, el cual detecta eficazmente comunidades de tamaños similares.

```
# edge_betweenness
comunidades <- cluster_edge_betweenness(grafo)
# walktrap
comunidades <- cluster_walktrap(grafo)
```

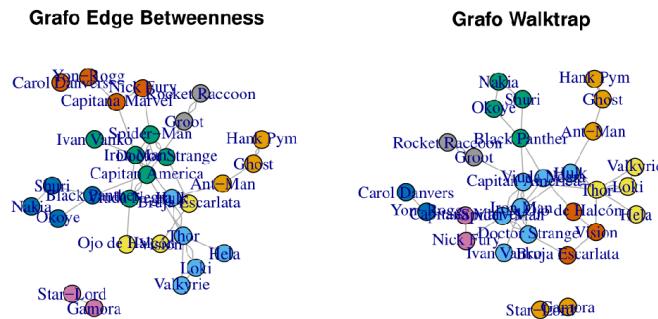


Figura 39.14: Comunidades según el algoritmo utilizado

Reumen

Este capítulo ha introducido la teoría de grafos y su relación con las RRSS, destacando:

- Los conceptos elementales de la teoría de grafos: vértice, arista, gráfico dirigido y no dirigido, grado y camino entre otros.
- El procedimiento con **R** para el análisis de grafos a través del paquete **igraph**, mostrando la estructura necesaria para componer un grafo.
- El análisis de influencia en un grafo aplicado a RRSS, introduciendo los conceptos de centralidad y comunidad.
- El análisis del entorno social de los personajes del universo cinematográfico Marvel, aplicando los conocimientos teóricos presentados a lo largo del capítulo.

Parte IX

Ciencia de datos espaciales

Capítulo 40

Trabajando con datos espaciales

Gema Fernández-Avilés¹

Universidad de Castilla-La Mancha

40.1. Introducción

Los **datos espaciales**, también conocidos como **datos geográficos** o **datos georeferenciados**, son aquellos datos relacionados o que contienen información de una localización o área geográfica de la superficie de la Tierra. El primer análisis de datos geoespaciales fue hecho por el médico John Snow en 1854. Éste produjo un famoso mapa que muestra las muertes causadas por un brote de cólera (que mató a 127 personas en 3 días) en Soho, Londres así como la ubicación de las bombas de agua en el área (Fig. 40.1). Snow descubrió que había un agrupamiento significativo de muertes alrededor de una determinada bomba, y al quitar la manija de la bomba se detuvo el brote. Los datos con los que trabajó Snow y aquellos que contienen coordenadas son considerados datos espaciales.

El análisis espacial de Snow es considerado el antecedente más antiguo conocido de la ciencia de datos (Baumer et al. (2021)): (i) la información clave se obtuvo mediante la combinación de tres fuentes de datos (las muertes por cólera, las ubicaciones de las bombas de agua y el mapa de calles de Londres); (ii) se puede crear un modelo espacial directamente a partir de los datos y (iii) el problema solo se resolvió cuando la evidencia basada en datos se combinó con un modelo plausible que explicaba el fenómeno físico. Es decir, Snow era médico y su conocimiento sobre la transmisión de enfermedades fue suficiente para convencer a sus colegas de que el cólera no se transmitía por el aire.

¹Quisiera agradecer a Diego Hernández la ayuda prestada en la elaboración de este capítulo.

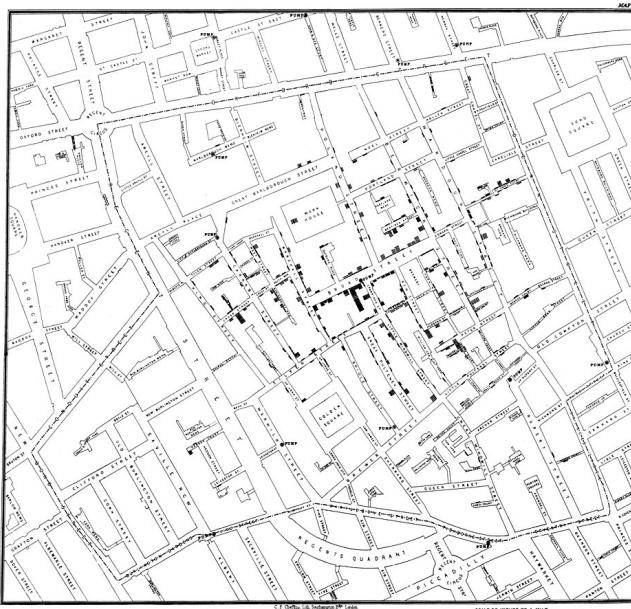


Figura 40.1: Mapa de cólera en Londres según Snow. Fuente: Wikipedia

40.1.1. Estadística para datos espaciales

El área que se encarga de estudiar y analizar los datos espaciales es la **estadística espacial** o la estadística para datos espaciales ([Cressie \(1993\)](#), [Montero et al. \(2015\)](#)).

Debido a que los datos espaciales surgen en una gran variedad de campos y aplicaciones, también hay una gran variedad de tipos de datos espaciales, estructuras y escenarios ([Schabenberger and Gotway, 2005](#), p. 6). La Fig. 40.2 representa la clasificación de datos espaciales proporcionada por [Cressie \(1993\)](#) basada en la naturaleza del dominio espacial en estudio. Cressie distingue tres tipos de datos espaciales:

- (I) datos geoestadísticos,
- (II) datos de patrones de puntos y
- (III) datos lattice o reticulares.

El estudio de los datos geoestadísticos se aborda en el Cap.[41](#), el análisis de los datos *lattice* se lleva a cabo en el Cap. [42](#) dedicado a la Econometría espacial y los datos de patrones de puntos se analizan en el Cap. [43](#).

40.2. Conceptos clave

683

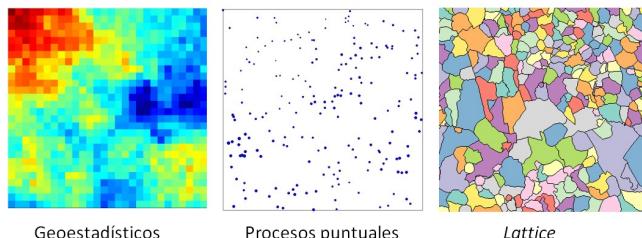


Figura 40.2: Clasificación de datos espaciales propuesta por Cressie (1993)

40.2. Conceptos clave

Visto el contexto original de los datos espaciales y antes de entrar en detalle en su análisis, se debe tener en cuenta una serie de conceptos clave. La Fig. 40.3, representa la localización de los accidentes de tráfico registrados en la ciudad de Madrid durante el año 2020. Sin embargo, tal representación no aporta información útil para su análisis. Por ejemplo, sería interesante añadir un mapa de carreteras junto con la localización de los accidentes.

```
library("CDR")
library("tidyverse")
ggplot(data = accidentes2020_data,
       aes(x = coordenada_x_utm, y = coordenada_y_utm)) +
  geom_point(col="blue", size = 0.1, alpha = 0.3) +
  coord_fixed()
```

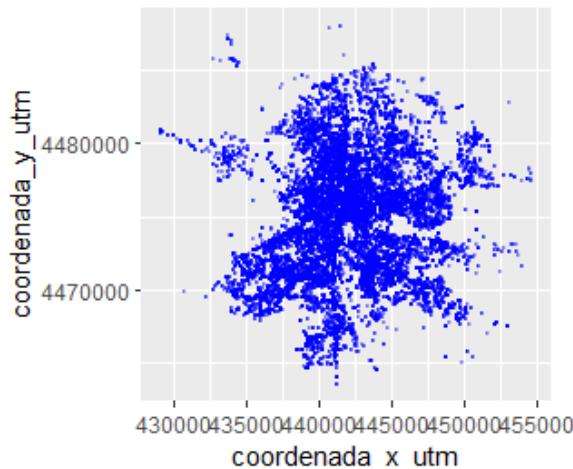


Figura 40.3: Accidentes de Tráfico en Madrid (2020)

Ademas de las coordenadas, en la representación de geodatos es importante el marco o contexto

espacial, así como el conocimiento del (i) **Sistema de referencia de coordenadas** o Coordinate reference system (**CRS**) en el que están georeferenciadas o proyectadas las coordenadas y (ii) el tipo de datos con el que se está trabajando: vectores o ráster.

```
library("sf")
accidentes2020_sf <- st_as_sf(accidentes2020_data,
  coords = c("coordenada_x_utm", "coordenada_y_utm"),
  crs = 25830 # proyección ETRS89/ UTM zone 30N. Área de uso: Europa
)

library("mapSpain")
madrid <- esp_get_munic(munic = "^Madrid$") |>
  st_transform(25830)

# descara imagen de un de mapa estático de las carreteras de Madrid
tile <- esp_getTiles(madrid, "IDErioja", zoommin = 2)

ggplot() +
  tidyterra::geom_spatraster_rgb(data = tile) +
  geom_sf(data = accidentes2020_sf,
    col = "blue", size = 0.1, alpha = 0.3) +
  coord_sf(expand = FALSE)
```

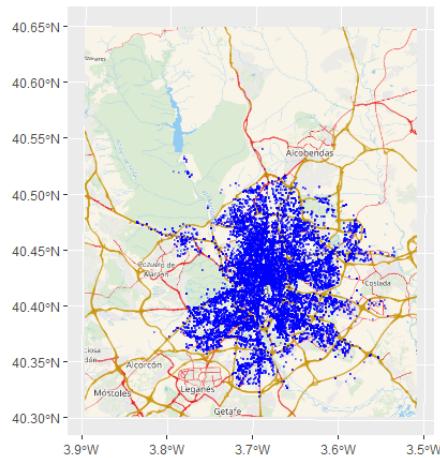


Figura 40.4: Accidentes de tráfico en Madrid proyectados y con mapa de carreteras (2020)

La Fig. 40.4 permite observar ciertos patrones en la ocurrencia de accidentes. Por ejemplo, apenas se producen accidentes en la Casa de Campo o en el Monte del Pardo, y parece observarse cierta concentración en la ciudad y en las autopistas de salida de la ciudad.

40.2.1. Sistema de referencia de coordenadas

Los CRS permiten identificar con exactitud la posición de los datos sobre el globo terráqueo. Cuando se trabaja con datos espaciales provenientes de distintas fuentes de información es necesario comprobar que dichos datos se encuentran definidos en el mismo CRS. Ésto se consigue transformando (o proyectando) los datos a un CRS común. Una buena referencia para profundizar este tema es el Cap. 2 de [Pebesma and Bivand \(2022\)](#).

En la Fig. 40.5 se muestran los puertos en un mapa mundial. Todos los vienen representados por el punto rojo. ¿A qué se debe? A que los datos están en distintos CRS.

```
library("giscoR")

paises <- gisco_get_countries()
puertos <- gisco_get_ports()
paises_robin <- st_transform(paises, st_crs("ESRI:54030")) #Proyección Robinson

plot(st_geometry(paises_robin), main = " ")
plot(st_geometry(puertos), add = TRUE, col="2", pch=20, lwd=2.5)
```



Figura 40.5: Localización de los puertos en el mapamundi (distinto CRS en los puertos y el mapa)

Los dos tipos de CRS que existen se describen a continuación:

- (i) **Geográficos**: aquellos en los que los parámetros empleados para localizar una posición

espacial son la latitud (Norte-Sur [-90°,90°]) y la longitud (Este-Oeste [-180°,180°]). Están basados en la geometría esférica. En este caso las distancias entre dos puntos son **distancias angulares**.

- (ii) **Proyectados:** permiten reducir la superficie de la esfera terrestre (3D) a un sistema cartesiano (2D). Para ello, es necesario transformar las coordenadas longitud y latitud en coordenadas cartesianas X e Y . La unidad de distancia, habitualmente, es el **metro**.

Tras proyectar los puertos al mismo CRS que el mapamundi utilizando la proyección de Robinson (la proyección cartográfica más convencional para mapamundis), la Fig. 40.6 muestra adecuadamente el mapa de la Fig. 40.5.

```
st_crs(puertos) == st_crs(paises_robin) # Comprueba CRS
#> [1] FALSE
puertos_robin <- st_transform(puertos, st_crs(paises_robin))
plot(st_geometry(paises_robin), main = " ")
plot(st_geometry(puertos_robin), add = TRUE, col=4, pch=20)
```

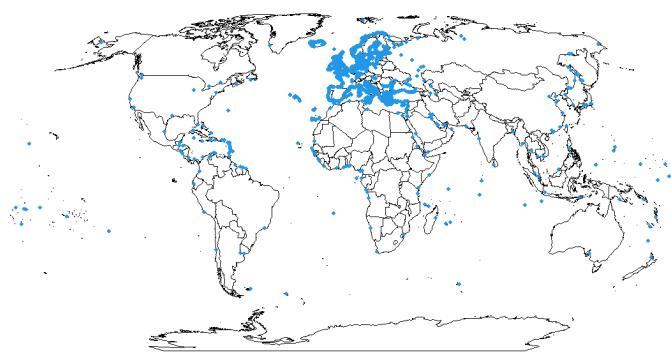


Figura 40.6: Localización de los puertos en el mapa mundi (mismo CRS puertos y mapa)

¿Qué proyección uso? El CRS adecuado para cada análisis depende de la localización y el rango espacial de los datos. El paquete `crssuggest` (Walker, 2022) facilita la labor, sugiriendo la escala de estudio o el tipo de análisis más adecuado para cada zona.

40.2.2. Formatos de datos espaciales

En el ámbito del análisis espacial, el formato de datos espaciales se puede clasificar en función del modelo de datos. Se pueden distinguir dos tipos de modelos de datos (Lovelace et al., 2019): vectoriales y ráster².

40.2.2.1. Datos de vectores

Este modelo está basado en puntos georeferenciados. Los **puntos** pueden representar localizaciones específicas, como la localización de los Hospitales y Centros de Salud de la ciudad de Toledo (Fig. 40.7).

```
ggplot() +
  geom_sf(data = hosp_toledo,
          aes(fill = "Hospitales y Centros Sanitarios",
              color = "blue") +
  labs(title = NULL, fill = NULL) +
  theme_minimal() +
  theme(legend.position = "right")
```

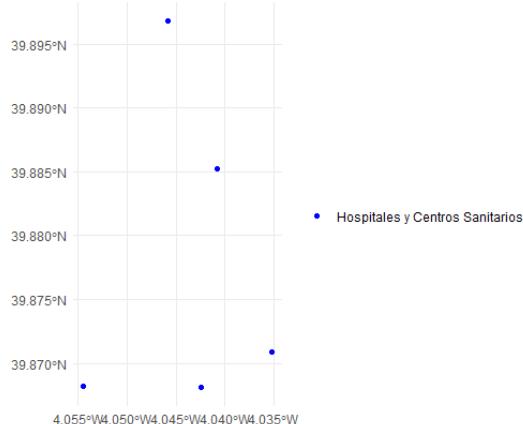


Figura 40.7: Hospitales y Centros de Salud en Toledo

Los puntos también pueden estar conectados entre sí, de manera que formen geometrías más complejas, como **líneas** y **polígonos**.

En la Fig. 40.8, el río Tajo está representado como una línea (`tajo`, sucesión de puntos unidos entre sí) y la ciudad de Toledo como un polígono (`toledo`, línea de puntos cerrada formando un continuo).

²Un análisis detallado puede verse en Hernangómez and Fernández-Avilés (2022)

```
ggplot(toledo) +
  geom_sf(fill = "cornsilk2") +
  geom_sf(data = tajo, col = "lightblue2", lwd = 2, alpha = 0.7) +
  geom_sf(data = hosp_toledo, col = "blue") +
  coord_sf(xlim = c(-4.2, -3.8), ylim = c(39.8, 39.95)) +
  theme_minimal()
```

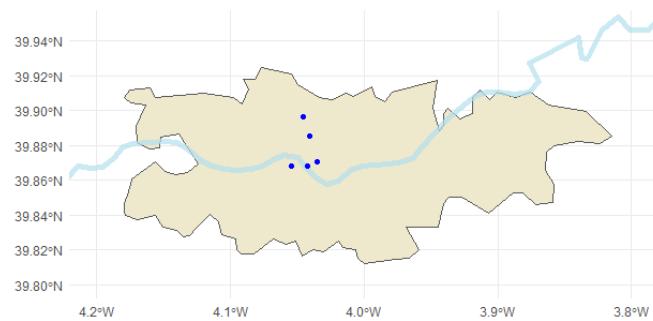


Figura 40.8: Datos vector: Puntos, líneas y polígonos

Las extensiones más habituales de los archivos que contienen datos de vectores se muestran a continuación:

Tabla 40.1: Ficheros con datos vector

Tipo	Extensión
Shapefile	.shp, .shx, .dbf
GeoPackage vector	.gPKG
GeoJson	.geojson
GPX	.gpx
Geography Markup Language	.gml
Keyhole Markup Language	.kml
Otros	.csv, .txt, .xls

ESRI Shapefile surgió como uno de los primeros formatos de intercambio de datos geográficos y en la actualidad es, quizás, el formato más empleado. Sin embargo, tiene una serie de limitaciones: es un formato multiarchivo y el CRS es opcional.

40.2.2.2. Datos ráster

Los datos raster son datos proporcionados en una rejilla de píxeles (regulares o no) denominada **matriz**. El caso más popular de un ráster es una fotografía, donde la imagen se representa como una serie de celdas, determinadas por la resolución de la imagen, es decir, el tamaño del píxel (por ejemplo, 5 x 5 unidades, si es regular, 5 x 10 unidades, si es irregular) y el valor del pixel (RGB, por ejemplo) que determina el color que presenta cada uno de estos píxeles. En el ámbito de los datos espaciales, un archivo ráster está formado por una malla de píxeles georreferenciada, tal y como muestra la Fig. 40.9. Aquí se visualiza el conjunto de datos `elev` dentro del paquete CDR que representa los datos de la altitud de la provincia de Toledo en metros.

```
library("terra")
elev <- rast(system.file("external/Toledo_DEM.asc", package = "CDR"))
plot(elev, main = " ")
polys <- as.polygons(elev, dissolve=FALSE)
plot(polys, add = TRUE, border = "grey90")
plot(st_geometry(Tol_prov), add = TRUE)
```

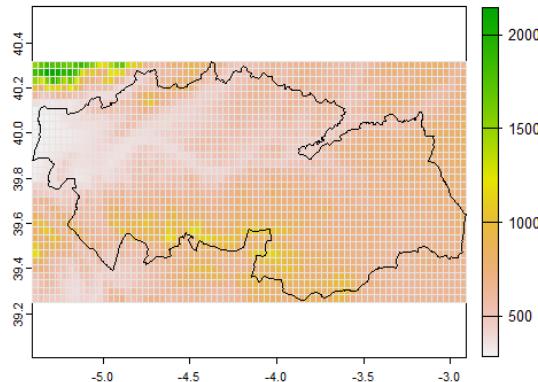


Figura 40.9: Datos ráster. Altitud de la provincia de Toledo

En la Fig. 40.9, el objeto ráster `elev` tiene únicamente una capa. Eso implica que cada píxel tiene asociado un único valor, en este caso, la altitud media del terreno observado. Las extensiones más habituales de los archivos que contienen datos ráster se muestran a continuación:

Tabla 40.2: Ficheros con datos raster

Tipo	Extensión
ASCII Grid	.asc

Tipo	Extensión
GeoTIFF	.tif, .tiff
Enhanced Compression Wavelet	.ecw

40.3. Mi primer mapa

Definidos los elementos clave de los datos espaciales se llevará a cabo la representación en un mapa de la distribución municipal de la renta neta per cápita (`renta_municipio_data`) por municipio (`municipios`) en el periodo 2019 en España³. Los datos están incluidos en el paquete CDR.

Ambos conjuntos deben tener, al menos, un campo en común, `codigo_ine` en este caso, para su unión.

```
library("CDR")
library("sf")
munis_renta <- municipios |>
  left_join(renta_municipio_data) |>    # une datasets
  select(name, cpro, cmun, `2019`)        # selecciona variables
#> Joining, by = "codigo_ine"

ggplot(munis_renta) +
  geom_sf(aes(fill = `2019`), color = NA) +
  scale_fill_continuous(
    labels = scales::label_number(
      big.mark = ".", decimal.mark = ",", suffix = " €")) +
  theme_minimal()
```

La Fig. 40.10 presenta un mapa temático o de coropletas, es decir, una visualización sencilla de cómo varía la distribución de una variable (en este caso la renta neta media por persona) en un área geográfica (España). Adicionalmente, una serie de elementos gráficos característicos de los objetos espaciales puede verse en la información contenida en el objeto `munis_renta`: los datos son de tipo vector, el tipo de geometría es MULTIPOLYGON, el CRS es ETRS89 y una leyenda explica el significado de la variable.

```
head(munis_renta)[1:3, ]
#> Simple feature collection with 3 features and 4 fields
#> Geometry type: MULTIPOLYGON
#> Dimension: XY
#> Bounding box: xmin: -3.140179 ymin: 36.73817 xmax: -2.741701 ymax: 37.24562
#> Geodetic CRS: ETRS89
#> name cpro cmun 2019 geom
```

³Un análisis detallado puede verse en ([Hernangómez and Fernández-Avilés, 2022](#)).

40.4. ¿Cómo (no) mentir con la visualización?

691

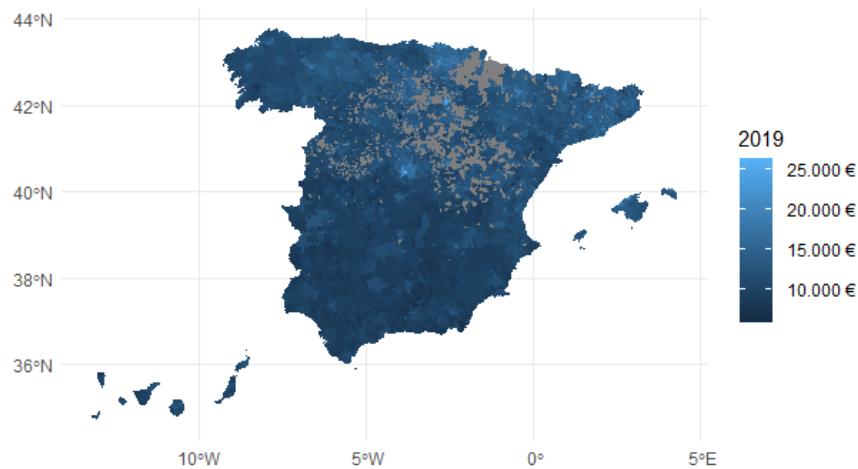


Figura 40.10: Distribución de la renta neta media por persona (€) en 2019

```
#> 1 Abla 04 001 10192 MULTIPOLYGON (((-2.775594 3...
```

```
#> Abrucena 04 002 10021 MULTIPOLYGON (((-2.787566 3...
```

```
#> Adra 04 003 8192 MULTIPOLYGON (((-3.051988 3...
```

40.4. ¿Cómo (no) mentir con la visualización?

Si se realiza un mapa de coropletas como el de la Fig. 40.10, puede que la información aparezca distorsionada. Algunas consideraciones básicas en visualización son:

- La escala de color.
- La distribución de los datos.
- La definición de intervalos.

¿Cómo es la distribución de la variable renta? La variable no sigue una distribución Normal (la renta sigue una distribución Gamma⁴), y si el objetivo es mostrar patrones espaciales de la variable, para una mejor representación será necesario dividir los datos en clases con el paquete `classInt` (Bivand, 2020). De entre las distintas posibilidades que ofrece la función

⁴Las características de la distribución Gamma pueden verse en el Cap. 12.

`classIntervals()`, se utiliza el método de Fisher-Jenks, que consiste en un mapa de cortes naturales que utiliza un algoritmo no lineal para agrupar observaciones de modo que se maximice la homogeneidad dentro del grupo. Este algoritmo está desarrollado específicamente para la clasificación de datos espaciales y su visualización en mapas. Además, se eliminan los municipios sin datos (sombreados en color gris) y se elige una escala de color adecuada. El mapa de la Fig. 40.11 proporciona ahora una visualización adecuada de la variable renta.

```
munis_renta_clean <- munis_renta |>
  filter(!is.na(`2019`))

# crea Fisher-Jenks clases
library(classInt)
fisher <- classIntervals(munis_renta_clean$`2019`,
  style = "fisher", n = 10
)

ggplot(munis_renta_clean) +
  geom_sf(aes(fill = cut(`2019`, fisher$brks)), color = NA) +
  scale_fill_viridis_d(option= "A",
    labels= scales::label_number(suffix= "€")) +
  guides(fill = guide_colorsteps()) +
  labs(fill= "Fisher-Jenks") +
  theme_minimal()
```

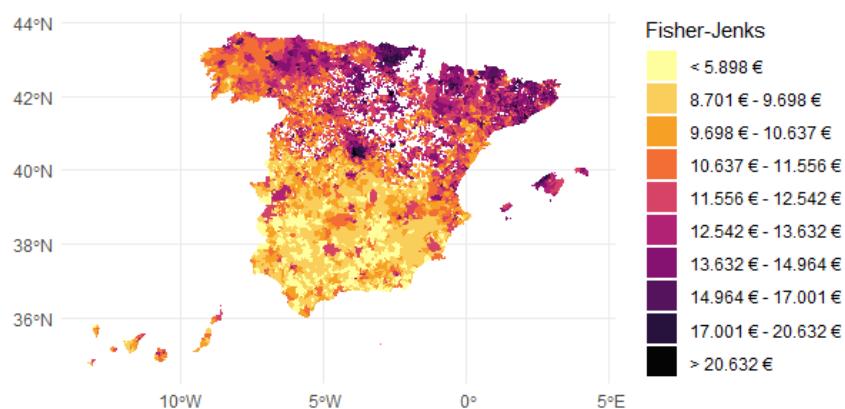


Figura 40.11: Renta neta per cápita (€) por tramos según Fisher-Jenks

40.5. Mapas espacio-temporales

La dimensión temporal es cada vez más importante en el ámbito espacial, por ello, es importante representar en el tiempo los procesos espaciales. La Fig. 40.13 representa la temperatura mínima registrada en España del 6 al 10 de Enero de 2021, CDR::tempmin_data, durante la [Borrasca Filomena](#).

```
tmin_sf <- st_as_sf(tempmin_data,
  coords = c("longitud", "latitud"),
  crs = 4326 # coordenadas geográficas longitud/latitud WGS84
)

esp <- esp_get_ccaa() |> # sf objeto, contorno de España
  filter(ine.ccaa.name != "Canarias") # excluye Canarias
```

La primera pregunta se debe formular es: ¿tengo el CRS de las estaciones de monitoreo en la misma proyección que el contorno de España?

```
st_crs(tmin_sf) == st_crs(esp)
#> [1] FALSE
esp2 <- st_transform(esp, st_crs(tmin_sf))
st_crs(tmin_sf) == st_crs(esp2)
#> [1] TRUE
```

Comprobado el CRS, es habitual representar las coordenadas con las que se trabaja. La Fig. 40.12 muestra la localización de las estaciones de monitoreo en España que registran la temperatura.

```
ggplot(esp2) +
  geom_sf() +
  geom_sf(data = tmin_sf) +
  theme_light()
```

Por último, se representa el mapa espacio-temporal con la función `ggplot()` indicando en el argumento `facet_wrap()` la dimensión temporal.

Nota

Los paquetes **tmap** ([Tennekes, 2018](#)) y **mapsdf** ([Giraud, 2022](#)) son referentes para mapas temáticos y pueden utilizarse como alternativa.

```
# definición de intervalos
cortes <- c(-Inf, seq(-20, 20, 2.5), Inf)
colores <- hcl.colors(15, "PuOr", rev = TRUE)
```

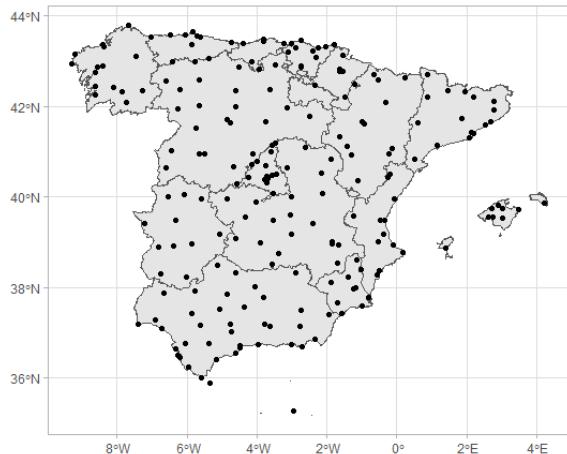


Figura 40.12: Estaciones de AEMET en la Península Ibérica

```
tmin_sf_sptem <- tmin_sf |>
  arrange(fecha, desc(tmin))

ggplot() +
  geom_sf(data = esp2, fill = "grey95") +
  geom_sf(data = tmin_sf, aes(color = tmin), size=3, alpha= .7) +
  facet_wrap(vars(fecha), ncol = 3) +
  labs(color = "Temp. mín") +
  scale_color_gradientn(
    colours = colores,
    breaks = cortes,
    labels = ~str_c(. , " °"),
    guide = "legend")
```

40.6. Mapas interactivos

El desarrollo de la informática ha propiciado también el desarrollo de la geocomputación, que está relacionada con los desarrollos webs, y permite, entre otras cosas, la representación de mapas interactivos.

A modo de ejemplo, el mapa de la Fig. 40.14 representa el mapa Fig. 40.1 de forma interactiva con la librería `leaflet`. Estos mapas dinámicos, ampliables y desplazables, son más informativos que los mapas estáticos y, además, son una alternativa que pueden proporcionar una experiencia diferente y una mayor interacción al usuario.

40.6. Mapas interactivos

695

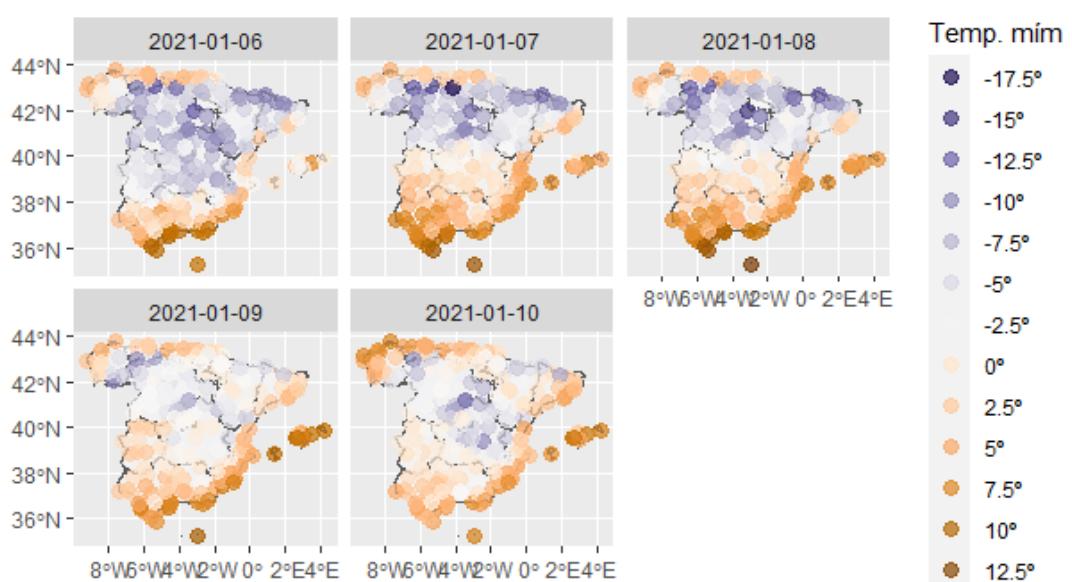


Figura 40.13: Temperatura mínima en España (6-10 enero 2021)

```
library("leaflet")
library("isdas")
data("snow_deaths")
data("snow_pumps")

## crea mapa interactivo
snow_map <- leaflet() |>
  setView(lng = -0.136, lat = 51.513, zoom = 16) |>
  addTiles() |>
  addMarkers( data = snow_deaths, ~long, ~lat,
    clusterOptions = markerClusterOptions(),
    group = "Deaths" ) |>
  addMarkers(data = snow_pumps, ~long, ~lat,
    group = "Pumps" )
snow_map
```

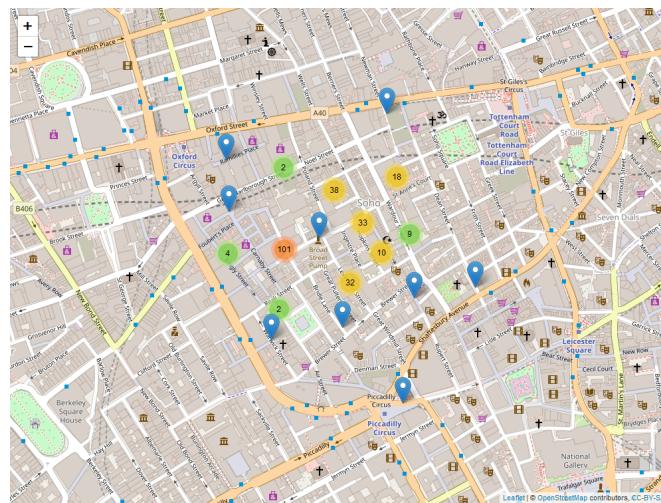


Figura 40.14: Mapa interactivo de las muertes por cólera en Londres según Snow en 1854

Reumen

Los datos espaciales son aquellos que contienen información de una zona geográfica de la tierra. Vienen definidos por coordenadas y por un sistema de referencia de coordenadas que debe tenerse en cuenta para su representación.

Existen dos tipos de formatos de datos: vector y ráster.

Los datos espaciales pueden clasificarse en: geoestadísticos, reticulares y puntuales.

Capítulo 41

Geoestadística

Gema Fernández-Avilés^a y José-María Montero^a

^aUniversidad de Castilla-La Mancha

41.1. Introducción

El término “geoestadística” apareció por primera vez en Matheron (1962), y en él “geo” enfatiza la referencia a las Ciencias de la Tierra, extendiendo el ámbito de la estadística tradicional, cuyo objetivo es el uso de métodos probabilísticos-inferenciales¹, con la incorporación del componente geográfico.

La geoestadística estudia los fenómenos regionalizados, que son aquellos que:

- Se extienden en el espacio, siendo el dominio espacial, D , continuo (se puede observar en cualquiera de sus puntos) y fijo (las ubicaciones observadas no son estocásticas; se seleccionan, por el procedimiento que sea, a juicio del investigador)².
- Presentan una organización o estructura debida a la dependencia espacial existente.

El objetivo fundamental de la geoestadística es sacar provecho de la dependencia espacial existente para llevar a cabo predicciones (interpolaciones) óptimas en ubicaciones o áreas de interés (en este sentido se habla de predicciones puntuales o por bloques, respectivamente), o la realización de mappings sobre todo el dominio o parte de él. Al ser D continuo, no se puede

¹Véanse Caps. 12 y 12.

²Los datos geoestadísticos son tan solo una parte de los datos espaciales: otra parte de ellos, son los datos “lattice”, poligonales o regionales, donde D es discreto (códigos postales, provincias, regiones, países...) y las ubicaciones observadas no son estocásticas. De su estudio se suele encargar la econometría espacial (véase Cap. 42). También hay otro tipo de datos espaciales que surgen en dominios que pueden ser continuos o discretos, pero donde la selección de las ubicaciones observadas no depende del investigador (en este sentido D es aleatorio). Se trata de los denominados procesos de puntos (véase Cap.43).

hacer una representación exhaustiva del fenómeno, pero sí se puede reconstruir a partir de las observaciones disponibles.

Las consecuencias de utilizar la estadística clásica, que no considera la dependencia espacial, cuando la hay, son muy graves y pueden verse en [Montero et al. \(2015\)](#).

El ámbito de aplicación de la geoestadística es enorme: minería, industria petrolífera, geología, meteorología, control de la calidad del aire, ecología, epidemiología, salud pública, criminología, economía, etc. Así, por ejemplo, en el ámbito del control de la calidad del aire en las grandes urbes, la concentración de ozono en aire se mide en una serie de estaciones de seguimiento, y a partir de dichas mediciones se reproduce el comportamiento del proceso sobre toda la urbe.

En conclusión, las dos partes del análisis geoestadístico son: el análisis estructural de la dependencia espacial y la predicción (que se suele acompañar del calificativo “*krigeada*”). Pero antes de estudiarlas, detengámonos en algunos preliminares.

41.2. Preliminares

Dado que los procedimientos geoestadísticos no pueden ser aplicados directamente sobre los fenómenos regionalizados como tales, porque son realidades físicas, se necesita una descripción matemática de los mismos a la que puedan ser aplicados: la *variable regionalizada* (*v.r.*) o *regionalización*, definida en un espacio geográfico, y que se supone que mide y representa correctamente dicho fenómeno.

Formalmente, cuando \mathbf{s} recorre D , el conjunto $z(\mathbf{s}), \mathbf{s} \in D$, se denomina *v.r.*, siendo $z(\mathbf{s}_i), i = 1, 2, 3, \dots$ una colección de valores regionalizados.

Desde la perspectiva probabilística, cada uno de los valores que toma *v.r.* puede interpretarse como el resultado de un mecanismo aleatorio, la variable aleatoria, *v.a.* ([véase 15.3.](#)). Si se toman valores regionalizados en todos los puntos del dominio, D , es decir, si se considerase *v.r.*, ésta podría ser vista como un conjunto infinitamente grande de *v.a.*, una en cada punto de D , que se conoce como *función aleatoria* (*f.a.*), *proceso estocástico* o *campo aleatorio* espacial, $Z(\mathbf{s}), \mathbf{s} \in D$, donde Z representa el fenómeno de interés. Pues bien, *v.r.* se interpreta como una realización de una *f.a.* espacial, y esta es una decisión metodológica clave en geoestadística.

Es importante tener en cuenta que, (*i*) frecuentemente, *v.r.* es muy irregular a escala local, lo que impide su representación mediante una función determinista; y (*ii*) muestra cierta organización o estructura espacial. La interpretación de *v.r.* como una realización de una *f.a.* espacial permite considerar estos dos aspectos:

- En cada localización \mathbf{s} , $Z(\mathbf{s})$ es una *v.a.* (de ahí el aspecto errático).
- Para un conjunto de puntos dado, $\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_k$, las *v.a.* $Z(\mathbf{s}_1), Z(\mathbf{s}_2), \dots, Z(\mathbf{s}_k)$ están ligadas por una red de correlaciones espaciales que son las responsables de la similitud en los valores que toman (de ahí el aspecto estructurado).

Las *f.a.* $Z(\mathbf{s})$ pueden ser estacionarias (en sentido estricto o de segundo orden), intrínsecamente estacionarias o no estacionarias, y el hecho de que tengan uno u otro tipo de estacionariedad determina el análisis geoestadístico.

41.3. Análisis estructural de la dependencia espacial

699

Una *f.a.* espacial es *estrictamente estacionaria* si las familias de *v.a.* $Z(\mathbf{s}_1), Z(\mathbf{s}_2), \dots, Z(\mathbf{s}_k)$, tienen la misma distribución de probabilidad conjunta que $Z(\mathbf{s}_1 + \mathbf{h}), Z(\mathbf{s}_2 + \mathbf{h}), \dots, Z(\mathbf{s}_k + \mathbf{h})$, $\forall k$, $\forall \mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_k$ y $\forall \mathbf{h} \in \mathbb{R}^d$ (donde \mathbf{h} es un vector de traslación), siempre que $\mathbf{s}_1 + \mathbf{h}, \mathbf{s}_2 + \mathbf{h}, \dots, \mathbf{s}_k + \mathbf{h} \in D$. Es decir, la distribución de probabilidad conjunta de $Z(\mathbf{s}_1 + \mathbf{h}), Z(\mathbf{s}_2 + \mathbf{h}), \dots, Z(\mathbf{s}_k + \mathbf{h})$ no se ve afectada por una traslación \mathbf{h} , y por tanto, ni ella, ni las funciones de densidad de dimensión inferior a k , dependen de las localizaciones consideradas.

La estacionariedad estricta es una condición muy restrictiva. Por ello, en la práctica lo que se suele asumir es la *estacionariedad de segundo orden*, que limita la estacionariedad a los dos primeros momentos de la *f.a.*³

Si una *f.a.* es estrictamente estacionaria, también es estacionaria de segundo orden. Sin embargo, la relación inversa no tiene por qué ser cierta.

La estacionariedad de segundo orden implica la existencia de la varianza de la *f.a.*, y deja fuera los fenómenos con infinita capacidad de variación. En este caso, si las diferencias $Z(\mathbf{s} + \mathbf{h}) - Z(\mathbf{s})$ son estacionarias de segundo orden, se dice que la *f.a.* es *intrínsecamente estacionaria*.

Aquellas *f.a.* cuya esperanza y/o varianza dependan de la localización (no son invariantes a las traslaciones) se denominan *no estacionarias*.

Salvo indicación de lo contrario, se asumirá la estacionariedad de segundo orden.

Finalmente, unos breves comentarios sobre la importancia de la estacionariedad. Es imposible inferir la ley de probabilidad que gobierna la *f.a.* espacial a partir de una sola realización de la misma (una sola regionalización), pues sería como tener una muestra de tamaño 1. Pero en la práctica ese será el caso. Bueno, ni siquiera eso. Solo se dispondrá de una parte de la regionalización: la correspondiente a las localizaciones observadas. La solución a tan importante limitación es adoptar la hipótesis de estacionariedad u homogeneidad espacial. Es decir, sustituir la repetición de realizaciones de la *f.a.* espacial por repeticiones en el espacio; dicho de otra forma, suponer que los valores observados en distintas localizaciones de D tienen las mismas características estadísticas y pueden ser considerados, en términos estadísticos, como realizaciones de la misma *f.a.*⁴ Por tanto, la hipótesis de estacionariedad significa que la ley espacial que gobierna *f.a.*, o parte de ella, es invariante a traslaciones; no depende de las localizaciones específicas observadas sino solo de \mathbf{h} .

La hipótesis de estacionariedad permitirá actuar como si todas las *v.a.* que conforman la *f.a.* tuviesen la misma distribución de probabilidad (o los mismos momentos), haciendo posible el proceso inferencial. Por eso se le da tanta importancia a que la *f.a.* sea estacionaria, del tipo que sea.

41.3. Análisis estructural de la dependencia espacial

³En geoestadística lineal el interés se centra en los dos primeros momentos de la *f.a.*, por lo cual tan sólo es necesaria la estacionariedad de segundo orden. Es decir, la esperanza y la varianza existen, son constantes y no dependen de la localización \mathbf{s} . La covarianza existe para cada par de *v.a.* $Z(\mathbf{s})$ y $Z(\mathbf{s} + \mathbf{h})$ y sólo depende de \mathbf{h} .

⁴Estas realizaciones no son independientes, y se suele asumir también la hipótesis de ergodicidad (véase 21.1).

41.3.1. Semivariograma

La estadística espacial se basa en la suposición de que las unidades georeferenciadas cercanas están relacionadas (son dependientes) de alguna manera (Getis, 1999), y tanto más cuanto más cercanas estén (Tobler, 1970a).

Los procesos con dependencia espacial se reconocen, visualmente, porque muestran un patrón en el espacio; en los que no la tienen, el patrón es el de la aleatoriedad. La Fig. 41.1 muestra una simulación de una *f.a.* con dependencia espacial (panel izquierdo) frente a unos datos totalmente aleatorios (panel derecho).

```
library(geoR)
library(fields)
par(mfrow = c(1, 2))
set.seed(728)

sim_dep <- grf(401, grid = "reg", cov.pars = c(1, 0.8), messages = FALSE)
points.geodata(sim_dep,
  main = "Dependencia espacial",
  col = tim.colors(), cex.max = 2
)

sim_indep <- grf(401, grid = "reg", cov.pars = c(0.01, 0), messages = FALSE)
points.geodata(sim_indep,
  main = "Aleatoriedad",
  col = tim.colors(), cex.max = 2
)
```

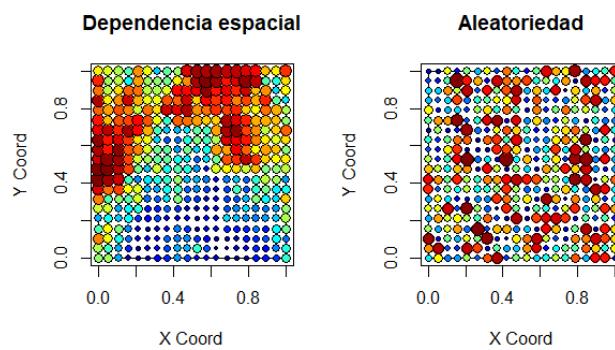


Figura 41.1: Dependencia espacial frente a aleatoriedad

Pasando del terreno de las simulaciones a la realidad, la Fig. 41.2 muestra la temperatura

41.3. Análisis estructural de la dependencia espacial

701

máxima en España el 6 de agosto de 2022⁵, en plena ola de calor (este es el ejemplo real que se utilizará a lo largo del capítulo). En ella puede observarse claramente una estructura de dependencia espacial, con máximas cercanas a 40 grados en la meseta central y Extremadura, de 30 grados o menos en la cordillera cantábrica y las costas atlántica, cantábrica y andaluza, y entre 30 y 35 grados en el resto del país (básicamente Murcia, Comunidad Valenciana y Cataluña).

```
library(CDR)
#summary(CDR::tempmax_data)

# renombra objetos por simplicidad en el análisis
ESP <- tempmax_data$ESP
ESP_utm <- tempmax_data$ESP_utm
grd_sf <- tempmax_data$grd_sf
grd_sp <- tempmax_data$grd_sp
temp_max_utm_sf <- tempmax_data$temp_max_utm_sf
temp_max_utm_sp <- tempmax_data$temp_max_utm_sp

library(ggplot2)
br_paper <- c(-Inf, seq(17.5, 45, 2.5), Inf)
pal_paper <- hcl.colors(15, "YlOrRd", rev = TRUE)

ggplot(ESP_utm) +
  geom_sf() +
  geom_sf(data = temp_max_utm_sf, aes(col = tmax), size = 4) + # temp_max_utm
  theme_light() +
  scale_color_gradientn(colours = pal_paper)
```

Ahora bien, para poder llevar a cabo predicciones geoestadísticas es necesario representar, previamente, los patrones de dependencia espacial observados mediante funciones que indiquen cuál es la estructura de dicha dependencia espacial. Dichas funciones son los semivariogramas. Dado que la identificación de la estructura de la dependencia espacial existente en el fenómeno de interés es la clave del éxito del proceso predictivo, al semivariograma se le considera la piedra angular de la predicción geoestadística (Montero et al., 2015).

Un semivariograma se define como la semivarianza de los incrementos de la *f.a.*:

$$\gamma(\mathbf{s}_i - \mathbf{s}_j) = \frac{1}{2}V[Z(\mathbf{s}_i) - Z(\mathbf{s}_j)], \forall \mathbf{s}_i, \mathbf{s}_j \in D. \quad (41.1)$$

que, en el caso habitual de *f.a.* estacionarias de segundo orden o intrínsecamente estacionarias (sin deriva), se transforma en:

$$\gamma(\mathbf{h}) = \frac{1}{2}V(Z(\mathbf{s} + \mathbf{h}) - Z(\mathbf{s})) = \frac{1}{2}E\left((Z(\mathbf{s} + \mathbf{h}) - Z(\mathbf{s}))^2\right), \quad (41.2)$$

Nótese que:

⁵Los datos ya procesados para el análisis se encuentran en `CDR::tempmax_data`. Una descripción puede verse con `summary(CDR::tempmax_data)`. Estos datos han sido descargados con la librería `climaemet` y `mapSpain`. Un desarrollo completo de manipulación de datos espaciales puede verse en Pizarro et al. (2021).

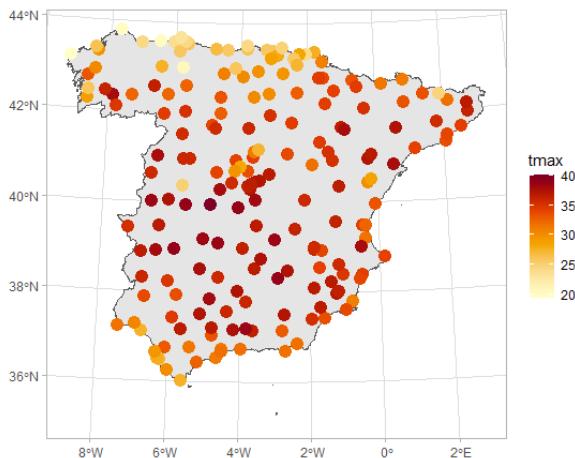


Figura 41.2: Temperatura máxima en España peninsular, 6 de agosto de 2022

- Si hay dependencia espacial (positiva⁶, es lo normal), la diferencia entre los valores de la *f.a.* en los puntos separados por una pequeña distancia será poca y más o menos la misma, es decir, dichas diferencias serán poco variables, y el valor del semivariograma, a pequeñas distancias, será pequeño.
- Si aumenta la distancia, la dependencia espacial se reduce y la diferencia entre los valores de la *f.a.* en los puntos separados por distancias intermedias y grandes no será tan parecida como en el caso anterior, sino mayor; y variará más: Es decir, el valor del semivariograma aumenta con la distancia.
- Si la distancia aumenta lo suficiente como para que ya no haya dependencia espacial, las diferencias entre los valores de la *f.a.* separados por tal distancia alcanzarán la variabilidad de la *f.a.* en estudio, y si ésta es estacionaria de segundo orden, el semivariograma se estabilizará en torno a ella.

En el caso estacionario, las funciones de covarianza, $C(\mathbf{h})$, también pueden ser utilizadas para representar la estructura de la dependencia espacial, si bien se prefiere el semivariograma porque no requiere el conocimiento de la media de la *f.a.* en estudio. Además, el semivariograma cubre un espectro más amplio de fenómenos regionalizados que la función de covarianza, ya que ésta no puede definirse en el caso de estacionariedad intrínseca. Los detalles pueden verse en Montero et al. (2015) y Montero and Larraz (2008).

Cuando el semivariograma depende tanto de la dirección como de la longitud del vector \mathbf{h} que une las localizaciones \mathbf{s} y $\mathbf{s} + \mathbf{h}$, se denomina *anisotrópico*. Cuando solo depende de la distancia ($|\mathbf{h}| = \|\mathbf{h}\|$, porque en el espacio euclídeo el módulo y la norma coinciden) se denomina *isotrópico*.

⁶Valores cercanos similares. Si es negativa, los valores vecinos son diferentes.

41.3. Análisis estructural de la dependencia espacial

703

Un semivariograma no puede ser cualquier función. Tiene que ser nulo en el origen, no negativo, verificar que $\gamma(\mathbf{h}) = \gamma(-\mathbf{h})$, debe ser una función condicionalmente definida negativa y tener un ritmo de crecimiento inferior a $|\mathbf{h}|^2$, es decir, $\lim_{|\mathbf{h}| \rightarrow \infty} \frac{\gamma(\mathbf{h})}{|\mathbf{h}|^2} = 0$ cuando el proceso es no estacionario (sin deriva), siendo finito en caso de procesos estacionarios de segundo orden.

El análisis del comportamiento de un semivariograma a pequeñas, medias y grandes distancias es de sumo interés, como se verá a continuación.

En general, a distancias medias y grandes, los semivariogramas asociados a *f.a.* estacionarias de segundo orden crecen, monótonamente, desde el origen con la distancia, hasta alcanzar un valor límite, la *varianza a priori* de la *f.a.* (o covarianza para $\mathbf{h} = 0$, $C(\mathbf{0})$), bien de forma exacta o asintóticamente. Dicho valor límite se denomina *meseta*, m , y la distancia a la cual se alcanza se conoce como *alcance* o *rango*, a . Por tanto, el rango es la distancia a partir de la cual ya no hay dependencia espacial. Cuando m se alcanza asintóticamente, el alcance no queda perfectamente definido y se toma como alcance, a efectos prácticos, a' , la distancia a la cual el semivariograma toma el valor 0,95m.

En el caso no estacionario (por ejemplo, si hay deriva) o intrínsecamente estacionario el semivariograma no tiene meseta.

El comportamiento a pequeñas distancias, sobre todo cerca del origen, que es donde más dependencia espacial hay, está muy relacionado con el grado de continuidad y regularidad de *f.a.* Cuanto más continua y regular sea, más suaves y estructuradas serán las realizaciones que produzca, y más regular será el comportamiento del semivariograma cerca del origen (Fig. 41.3, panel izquierdo; representación bidimensional).

Los semivariogramas con un comportamiento lineal cerca del origen son típicos de *v.r.* continuas, al menos por partes, pero no diferenciables. Su representación gráfica tridimensional está llena de picos. La amplitud de las fluctuaciones aumenta con la distancia entre localizaciones y es proporcional a la pendiente de la tangente en el origen (Fig. 41.3, panel derecho; representación bidimensional).

```
library(fields)
library(geoR)
par(mfrow = c(1, 2))
set.seed(123)

fa_gauss <- grf(1225, grid = "reg", cov.pars = c(1, .25), cov.model = "gaussian")
image(fa_gauss, col = tim.colors())

fa_sph <- grf(1225, grid = "reg", cov.pars = c(1, .25), cov.model = "spherical")
image(fa_sph, col = tim.colors())
```

Las *v.r.* regulares (aquellas cuya gráfica tridimensional no tiene picos) se identifican con un comportamiento semivariográfico parabólico en el origen. Si dicho comportamiento persiste a largas distancias, puede que exista una fuerte deriva.

Las discontinuidades en el origen (que teóricamente no pueden darse) son frecuentes en la práctica. Su amplitud se denomina “efecto pepita” (nugget effect) y son típicas de variables

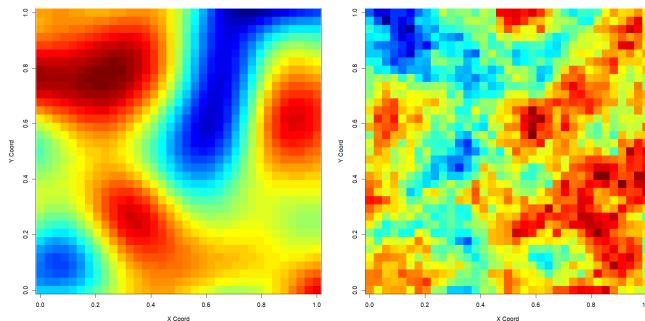


Figura 41.3: Representación bidimensional de dos *f.a.*: con semivariograma parabólico en el origen (izquierda); con semivariograma lineal en el origen (derecha)

regionalizadas muy irregulares y, quizás, discontinuas. Las causas más frecuentes del “efecto pepita” son la existencia de una estructura con alcance inferior a la distancia más corta entre localizaciones y los errores de posicionamiento o de medida (véase Chilès and Delfiner (1999) para más detalle).

El caso límite del efecto pepita es el “efecto pepita puro”. En ese caso, el semivariograma es constante cualquiera que sea la distancia, indicando ausencia de dependencia espacial.

La Fig. 41.4 muestra gráficamente los principales elementos de un semivariograma.

```
library(geoR)
semivar <- function(x, ...) {
  1 - cov.spatial(x, ...)
}
curve(semivar(x, cov.pars = c(0.8, 0.4), cov.model = "sph"), 0.0, 1,
  xlab = "Distancia",
  ylab = expression(bold(gamma("|h|"))), lwd = 4, lty = 1, col = "4", xlim = c(0.03,
  ↪ 1), ylim = c(0, 1)
)
abline(v = 0.4, col = 2, lty = 2, lwd = 2) # alcance
abline(h = 1, col = 3, lty = 2, lwd = 2) # meseta
legend(-0.05, 0.15, "Efecto pepita")
legend(0, 0.95, "Meseta")
legend(0.25, 0.5, "Alcance")
legend(0.5, 0.75, "Ausencia de dependencia")
```

```
knitr::include_graphics("img/semivar-parts.png")
```

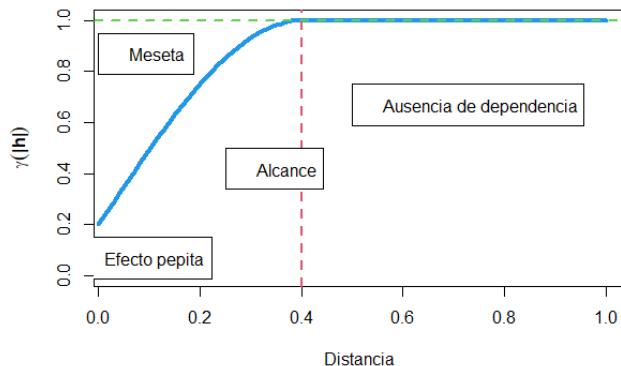


Figura 41.4: Elementos del semivariograma (meseta unitaria)

41.3.2. Modelos de semivariogramas válidos

Las funciones que verifican las condiciones que debe cumplir un semivariograma (véase 41.3.1) se conocen como semivariogramas válidos. El incumplimiento de alguna de ellas tiene perversas consecuencias en el proceso predictivo (por ejemplo, varianzas de los errores de predicción negativas). Su tipología, siguiendo el enfoque sugerido en [Journel and Huijbregts \(1978\)](#) y [Montero et al. \(2015\)](#), es la siguiente:⁷

41.3.2.1. Semivariogramas con meseta

Están asociados a *f.a.* estacionarias de segundo orden. Los más utilizados son:

- **Esférico.** Válido en \mathbb{R}^1 , \mathbb{R}^2 y \mathbb{R}^3 , y viene dado por:

$$\gamma(|\mathbf{h}|) = \begin{cases} m \left(1,5 \frac{|\mathbf{h}|}{a} - 0,5 \left(\frac{|\mathbf{h}|}{a} \right)^3 \right) & \text{si } |\mathbf{h}| \leq a \\ m & \text{si } |\mathbf{h}| > a. \end{cases} \quad (41.3)$$

Tiene un comportamiento lineal cerca del origen, indicando continuidad y cierto grado de irregularidad en la *f.a.* A grandes distancias alcanza la meseta cuando $|\mathbf{h}| = a$. Estas dos características son propias de muchas regionalizaciones observadas en la realidad; de ahí su popularidad.

⁷Aquí se presentan los semivariogramas isotrópicos más utilizados en la práctica. Nos centramos en ellos porque los semivariogramas anisotrópicos pueden ser representados por semivariogramas isotrópicos sin más que llevar a cabo una transformación lineal de las coordenadas, y porque las anisotropías (anisotropías) pueden representarse separadamente mediante semivariogramas isotrópicos. Para un análisis detallado, consultese, por ejemplo, [Montero et al. \(2015\)](#).

- **Exponencial.** Válido en $\mathbb{R}^d, d \geq 1$ y viene dado por:

$$\gamma(|\mathbf{h}|) = m \left(1 - \exp \left(-\frac{|\mathbf{h}|}{a} \right) \right). \quad (41.4)$$

Igual que el esférico, cerca del origen exhibe un comportamiento lineal, siendo menor la pendiente. A diferencia de él, solo alcanza la meseta asintóticamente. A efectos prácticos, se toma como alcance la distancia para la cual el semivariograma alcanza el valor del 95 % de la meseta, $a' \approx 3a$.

- **Gausiano.** Válido en $\mathbb{R}^d, d \geq 1$. Está definido por:

$$\gamma(|\mathbf{h}|) = m \left(1 - \exp \left(-\frac{|\mathbf{h}|^2}{a^2} \right) \right). \quad (41.5)$$

A diferencia del esférico y el exponencial, tiene un comportamiento parabólico cerca del origen. Por consiguiente, está asociado con *f.a.* estacionarias de segundo orden infinitamente diferenciables y, en consecuencia, muy regulares. Igual que el modelo exponencial, alcanza la meseta sólo asintóticamente, con $a' \approx a\sqrt{3}$.

- **Efecto pepita puro.** Refleja la ausencia de dependencia espacial:

$$\gamma(|\mathbf{h}|) = \begin{cases} m & \text{si } |\mathbf{h}| = 0 \\ 0 & \text{si } |\mathbf{h}| > 0 \end{cases}, \quad m > 0. \quad (41.6)$$

- **K-Bessel.** Válido in $\mathbb{R}^d, d \geq 1$. Su expresión es:

$$\gamma(|\mathbf{h}|) = m \left(1 - \frac{1}{2^{\alpha-1}\Gamma(\alpha)} \left(\frac{|\mathbf{h}|}{a} \right)^\alpha K_\alpha \left(\frac{|\mathbf{h}|}{a} \right) \right), \quad \alpha > 0, \quad (41.7)$$

donde K_α es la función de segunda especie de orden α . Este modelo puede representar cualquier tipo de comportamiento cerca del origen. Por ejemplo, para $\alpha = 0,5$ se obtiene el modelo exponencial.

41.3.2.2. Semivariogramas con meseta y efecto hoyo

- **J-Bessel.**

Un semivariograma no tiene por qué ser necesariamente una función monótona no decreciente, sino que puede tener “ondas” (efecto hoyo). Tal es el caso del modelo J-Bessel, que puede ser utilizado en presencia de dependencia espacial negativa o, específicamente, en caso de alternancia entre dependencia positiva y negativa. Válido en $\mathbb{R}^d, d \leq 2(\alpha + 1)$, su expresión viene dada por:

$$\gamma(|\mathbf{h}|) = m \left(1 - \left(\frac{2a}{|\mathbf{h}|} \right)^\alpha \Gamma(\alpha + 1) J_\alpha \left(\frac{|\mathbf{h}|}{a} \right) \right), \quad (41.8)$$

41.3. Análisis estructural de la dependencia espacial

707

donde α es un parámetro de forma, a es un parámetro de escala, Γ es la función de Euler que interpola el factorial y J_α es la función J-Bessel de primera especie de orden α .

La Fig. 41.5 muestra una representación gráfica de los anteriores semivariogramas.

```
library(gstat)
show.vgms(models = c("Sph", "Exp", "Gau", "Nug", "Bes", "Wav"))
```

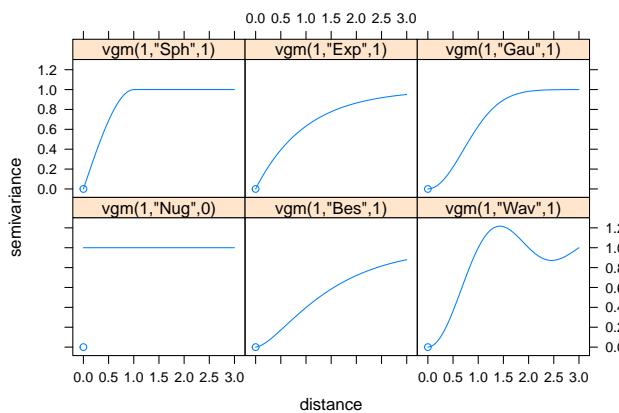


Figura 41.5: Representación de semivariogramas con meseta válidos (meseta y alcance efectivo unitarios; a excepción del efecto pepita puro)

41.3.2.3. Semivariogramas sin meseta

Estos modelos van más allá de la hipótesis estacionaria de segundo orden y corresponden a *f.a.* con una capacidad ilimitada de dispersión espacial, es decir, a *f.a.* intrínsecamente estacionarias, pero no estacionarias de segundo orden.

- **Potencial.** Válido en $\mathbb{R}^d, d \geq 1$ y definido por:

$$\gamma(|\mathbf{h}|) = (|\mathbf{h}|)^\alpha, \quad \text{con } 0 < \alpha < 2, \quad (41.9)$$

- **Logarítmico.** Válido en $\mathbb{R}^d, d \geq 1$ y con expresión:

$$\gamma(|\mathbf{h}|) = b \log |\mathbf{h}| \quad \text{si } |\mathbf{h}| \geq 0, \quad (41.10)$$

donde b es una constante.

Una representación gráfica de ambos semivariogramas puede verse en la Fig. 41.6.

```
library(gstat)
show.vgms(models = c("Pow", "Log"), sill = 1, range = c(2, 1), nugget = 0)
```

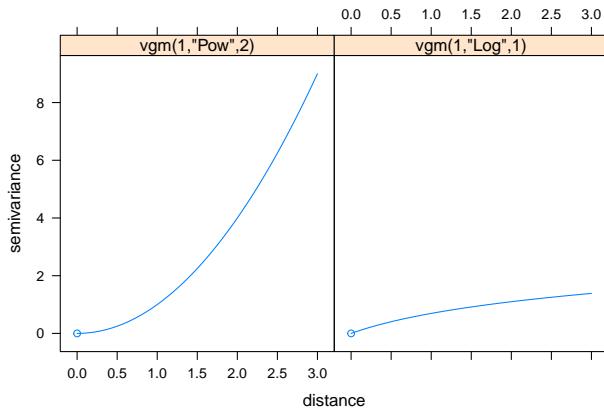


Figura 41.6: Representación de semivariogramas sin meseta válidos

41.3.3. Semivariograma empírico

Dado que la única información de la que se dispone es una realización observada de la *f.a* objeto de estudio, en la práctica la estructura de la dependencia espacial se estima mediante el semivariograma empírico.

En el marco de la estacionariedad intrínseca (que incluye la estacionariedad estricta y de segundo orden), y en \mathbb{R}^d , $d \geq 1$, se estiman (insegadamente) los valores semivariográficos para un número determinado de distancias, por el método de los momentos:

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2\#N(\mathbf{h})} \sum_{N(\mathbf{h})} (Z(\mathbf{s}_i + \mathbf{h}) - Z(\mathbf{s}_i))^2, \quad (41.11)$$

donde $\#N(\mathbf{h})$ es el número de parejas de localizaciones separadas por el vector \mathbf{h} .

La función que mejor ajusta las estimaciones de los valores semivariográficos anteriormente referidos se denomina *semivariograma empírico*, y también se suele denotar por $\hat{\gamma}(\mathbf{h})$.

Los valores semivariográficos se suelen computar para distancias inferiores a la mitad del diámetro de D , porque, para distancias superiores, el número de parejas de localizaciones suele ser pequeño para proporcionar estimaciones fiables. En la práctica, como en muchas de las direcciones no hay un número de parejas suficiente para calcular el semivariograma con cierta fiabilidad, lo habitual es construir un *semivariograma empírico omnidireccional*, es decir, que depende solo de la distancia (longitud del vector \mathbf{h}) y no de la dirección. Para ello se crean regiones de tolerancia, que no se solapen, basadas en intervalos de distancia (normalmente de la

41.3. Análisis estructural de la dependencia espacial

709

misma longitud) y un angulo de tolerancia. En concreto, la tolerancia se especifica en el módulo de \mathbf{h} ($\pm\Delta|\mathbf{h}|$) y su dirección ($\pm\Delta\theta$). Para más detalles y ejemplos, véase [Montero et al. \(2015\)](#).

La Fig. 41.7 muestra la ubicación de los puntos semivariográficos, indicando el número de parejas a cada distancia, en el caso de las temperaturas máximas en España (06/08/2022).

```
vgm_tmax <- variogram(tmax ~ 1, temp_max_utm_sf,
  cutoff = 250000 # 250 km
)
plot(vgm_tmax, plot.numbers = TRUE, pch = "+", lwd = 2, cex = 2)
```

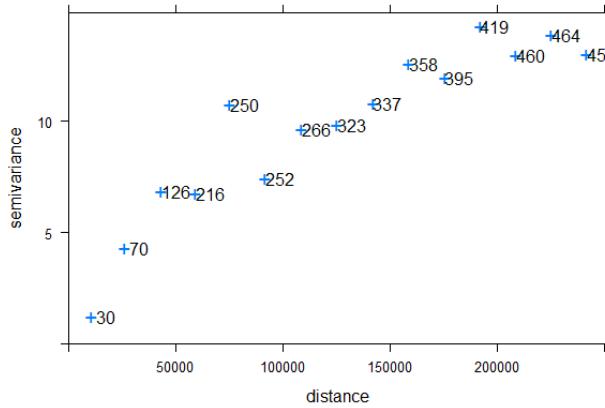


Figura 41.7: Valores semivariográficos. Temperaturas máximas (06/08/2022)

En el ejemplo ilustrado, las distancia mínima entre dos estaciones meteorológicas es 1.125m y la máxima 1.027.597m. Sin embargo, dada la geometría del mapa de España (aunque de Huelva a Gerona hay 987 km en linea recta, más de dos terceras partes de las ciudades españolas no están separadas más de 500 km.), se consideró 250.000m (1/4 de la distancia máxima) como distancia máxima a la hora de calcular los valores semivariográficos, ya que a partir de dicha distancia el número de parejas no es lo suficientemente grande como para obtener resultados fiables. Por convenio, `gstat` divide la distancia en 15 intervalos (`geoR` se divide en 13 porque los autores lo hicieron un viernes 13).

41.3.4. Ajuste semivariográfico

Cualquier función que dependa de una distancia y una dirección no es necesariamente un semivariograma, pues para ello tienen que cumplir los requisitos especificados en 41.3.1. Esta es la razón por la que el semivariograma empírico no puede utilizarse directamente para realizar predicciones geoestadísticas. Por ello, a los valores semivariográficos estimados se les ajusta

una función que represente un semivariograma válido. Sin embargo, esta tarea, clave para el éxito del posterior proceso predictivo, no es sencilla ni existe consenso en torno a ella. El ajuste puede ser *manual*, utilizando métodos visuales y gráficos, o *automático*, que usa procedimientos estadísticos. Una combinación de ambos es muy recomendable.

El ajuste manual pudiera parecer “no muy científico” pero, dado que lo más importante a la hora del ajuste no es tanto la bondad del ajuste para todos los puntos semivariográficos sino lo bien que un semivariograma válido representa las principales características del fenómeno, especialmente el tipo de estacionariedad (comportamiento a largas distancias) y, sobre todo, el tipo de continuidad (comportamiento cerca del origen), resulta ser un procedimiento muy práctico si se guía por las anteriores consideraciones. En este sentido, cualquier conocimiento sobre el fenómeno en estudio es bienvenido.

El ajuste automatizado mediante procedimientos estadísticos incluye los *métodos de mínimos cuadrados* (tanto ordinarios como generalizados y ponderados), que son los más populares en la práctica, y los *métodos basados en máxima verosimilitud*, que incluyen, entre otros, el tradicional método máximo verosímil, la máxima verosimilitud restringida y el método de la verosimilitud compuesta.

La Fig. 41.8 muestra el semivariograma empírico correspondiente a los puntos semivariográficos de la Fig. 41.7. De todos los modelos con meseta, el semivariograma ajustado ha sido un exponencial con alcance 76.404,64 metros y meseta 13,74.

```
fit_vgm_tmax <- fit.variogram(vgm_tmax,
                                model = vgm(model = c("Sph", "Exp", "Gau", "Nug", "Bes",
                                "Wav")), fit.sills = TRUE, fit.ranges = TRUE,
                                fit.kappa = TRUE, fit.method = 7)
fit_vgm_tmax
#>   model    psill     range
#> 1  Exp 13.74102 76404.64
attr(fit_vgm_tmax, "SSER")
#> [1] 6.200657e-07
plot(vgm_tmax, fit_vgm_tmax, lwd = 2, col = "2", pch = "*", cex = 3)
```

El método de ajuste ha sido el que figura por defecto en la función `vgm`: mínimos cuadrados ponderados con ponderaciones $\frac{N_{|h|}}{|h|^2}$, que funciona bien en la práctica y selecciona el semivariograma que mejor ajuste cuando el número de parejas es elevado y la distancia pequeña, que es la parte del semivariograma que hay que ajustar bien porque a pequeñas distancias es cuando más dependencia espacial hay. Respecto a los parámetros iniciales, aunque el investigador puede especificar los que considere convenientes, se recomienda utilizar los que tiene la función por defecto: (i) alcance igual a 1/3 de la distancia máxima en la muestra; (ii) como efecto pepita se toma la media de los tres primeros valores semivariográficos; y (iii) como meseta parcial (meseta menos efecto pepita), la media de los cinco últimos valores semivariográficos.

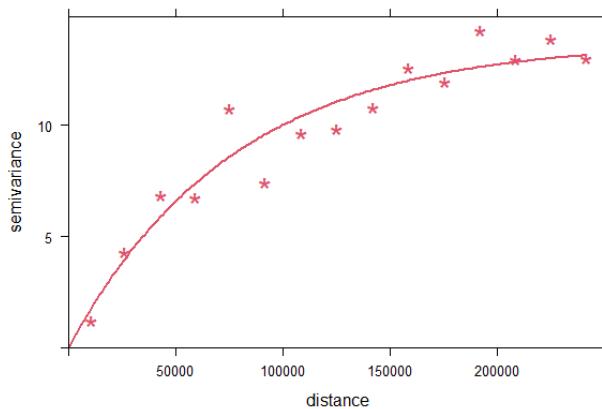


Figura 41.8: Semivariograma empírico. Temperaturas máximas (06/08/2022)

41.4. Kriging

Seleccionado el semivariograma válido que mejor se ajusta a los puntos semivariográficos, se aborda el proceso predictivo. El método predictivo que usa la geoestadística es conocido como *kriging* en honor al ingeniero de minas D.G. Krige.

El kriging tiene como objetivo predecir el valor de una *f.a.*, $Z(\mathbf{s})$, en uno o más puntos (o bloques) no observados, a partir de la regionalización observada (pueden ser puntos o bloques) en un dominio D , y proporciona el mejor predictor lineal insesgado de la *v.r.* de interés en tales puntos o bloques no observados⁸. La limitación a la clase de predictores lineales obedece a que, bajo estacionariedad de segundo orden, solo se requiere el conocimiento de los momentos de segundo orden de la *f.a.* Con más información estructural, pueden definirse predictores no lineales.

Las principales ventajas del kriging sobre los métodos de interpolación espacial deterministas (método de la distancia inversa, splines, regresión polinomial, etc.), es que (*i*) considera la estructura de la dependencia espacial (dando lugar a mejores predicciones), (*ii*) proporciona, junto con la predicción, la varianza del error de predicción y (*iii*) es un interpolador exacto.

Dependiendo del tipo de estacionariedad que se considere en la *f.a.* el kriging puede ser: universal (no estacionariedad en media) u ordinario (estacionariedad de segundo orden o intrínseca). Nos centraremos en el kriging ordinario (*KO*). La generalización al caso universal (hay derivas: la media depende de las localizaciones en vez de ser constante) puede verse en [Montero et al. \(2015\)](#).

En términos formales, *KO* se plantea como sigue: Sea $Z = \{Z(\mathbf{s}), \mathbf{s} \in D\}$ una *f.a.* con estacionariedad de segundo orden o intrínseca y con media desconocida (cuando se conoce, *KO*

⁸En lo que sigue, la exposición se centrará en la predicción puntual a partir de datos puntuales (es decir, en el soporte puntual). La generalización a bloques puede verse en [Montero et al. \(2015\)](#).

se denomina kriging simple). Sea el predictor lineal krigeado $Z^*(\mathbf{s}_0) = \sum_{i=1}^n \lambda_i Z(\mathbf{s}_i)$, donde las ponderaciones $\lambda_i, i = 1, 2, \dots, n$, se obtienen imponiendo al error de predicción las condiciones de esperanza nula y mínima varianza.

El sistema de ecuaciones que proporciona dichas ponderaciones óptimas es:

$$\begin{cases} \sum_{j=1}^n \lambda_j \gamma(\mathbf{s}_i - \mathbf{s}_j) + \alpha = \gamma(\mathbf{s}_i - \mathbf{s}_0), & i = 1, \dots, n \\ \sum_{i=1}^n \lambda_i = 1, \end{cases} \quad (41.12)$$

siendo la varianza del error de predicción: $\sigma_{OK}^2(\mathbf{s}_0) = \sum_{i=1}^n \lambda_i \gamma(\mathbf{s}_i - \mathbf{s}_0) + \alpha$, donde α es el multiplicador de Lagrange involucrado en el proceso de optimización.

Retomando el ejemplo de las temperaturas máximas en la España peninsular el 6 de agosto de 2022, a continuación se muestra el código necesario para la creación de un *mapping* de predicción de dichas temperaturas.

```
kriged_tmax <- kriging(tmax ~ 1,
  temp_max_utm_sp,
  grd_sp,
  model = fit_vgm_tmax
)
#> [using ordinary kriging]

kriged_df <- as.data.frame(kriged_tmax, xy = T, na.rm = T)

ggplot() +
  geom_tile(data = kriged_df,
    aes(x = coords.x1, y = coords.x2, fill = var1.pred)
  ) +
  geom_sf(data = ESP_utm, col = "black", fill = NA) +
  scale_fill_gradientn(colours = pal_paper,
    breaks = br_paper,
    labels = function(x) {
      paste0(x, "°")
    },
    guide = guide_legend(reverse = TRUE, title = "Temp. max.")
  ) +
  theme_light() +
  theme(panel.background = element_blank(),
    panel.border = element_blank(),
    axis.title = element_blank(),
  )

```

El *mapping* de la Fig. 41.9 tiene poco valor si no se acompaña de otro que muestre la desviación típica de los errores de predicción.

41.4. Kriging

713

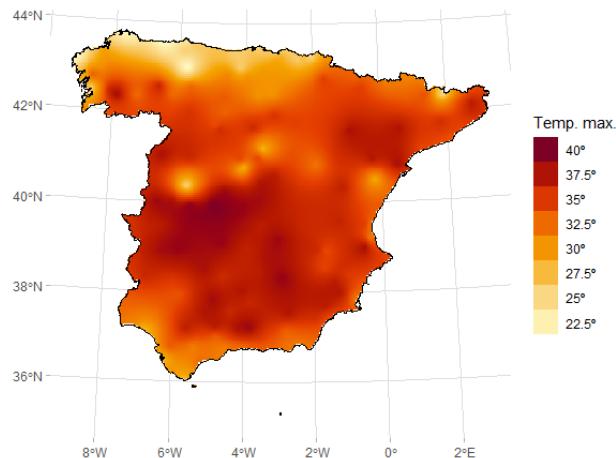


Figura 41.9: Mapping de temperaturas máximas (06/08/2022).

```
ggplot(kriged_df) +
  geom_contour_filled(aes(coords.x1, coords.x2, z = sqrt(var1.var)),
    breaks = c(0, 2, 2.5, 3, 3.5, 4, max(sqrt(kriged_df$var1.var))))
  ) +
  geom_sf(data = ESP_utm, col = "black", fill = NA) +
  geom_sf(data = temp_max_utm_sf, col = "blue", shape = 4) +
  scale_fill_manual( # paleta colores
    values = c("springgreen", hcl.colors(8, "PuRd", rev = TRUE)),
    guide = guide_legend(title = "Desv. típica\n error predicción")
  ) +
  theme_light() +
  theme(panel.background = element_blank(),
    panel.border = element_blank(),
    axis.title = element_blank(),
  )
```

Como se aprecia en la Fig. 41.10, cuanto mayor es el número de localizaciones observadas alrededor del punto de predicción, menor es la desviación típica del error de predicción.

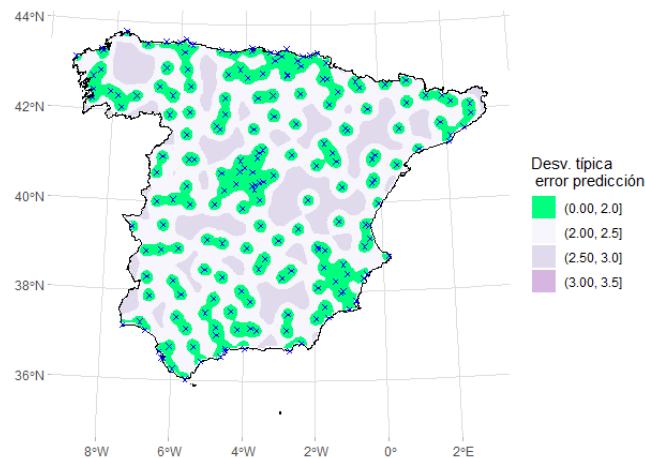


Figura 41.10: Desviaciones típicas del error de predicción

RESUMEN. La geoestadística estudia de fenómenos regionalizados, que son aquellos que se extienden en el espacio y presentan una organización o estructura debida a la dependencia espacial existente. Su objetivo es sacar provecho de dicha dependencia espacial para llevar a cabo predicciones (interpolaciones) óptimas en ubicaciones o áreas de interés, o la realización de *mappings* sobre todo el dominio o parte de él. Las dos partes del análisis geoestadístico son: (*i*) el análisis estructural de la dependencia espacial y (*ii*) la predicción “krigeada”. La estructura de dependencia espacial se representa mediante un semivariograma. La elección del semivariograma entre el elenco de funciones semivariográficas válidas es la clave del éxito de la predicción geoestadística, y por ello al semivariograma se le considera la piedra angular de la geoestadística. La técnica que utiliza la geoestadística para predecir se denomina *kriging*, y presenta un buen número de ventajas sobre los tradicionales métodos de interpolación espacial deterministas al considerar la estructura espacial de las observaciones.

Capítulo 42

Modelos econométricos espaciales

Andrés Vallone^a y Coro Chasco^{b,c}

^a Escuela de Ciencias Empresariales-Instituto de Políticas Públicas Universidad Católica del Norte ^b Departamento de Economía Aplicada, Universidad Autónoma de Madrid ^c Universidad de Neubria

42.1. La dependencia espacial

En muchas ocasiones, los fenómenos de estudio no son independientes del espacio geográfico en el cual se producen. Esto se refleja en la **primera ley de la Geografía** enunciada por [Tobler \(1970b\)](#) “Todas las cosas están relacionadas entre sí, pero las cosas más próximas en el espacio tienen una relación mayor que las distantes” ([Tobler, 1970b](#), p 236). Esta situación, produce una violación del supuesto básico de independencia de las variables aleatorias requerido por el método de estimación de mínimos cuadrados ordinarios (MCO).

En este contexto, los MCO ya no son óptimos y, por tanto, los estadísticos de contraste t y F pueden llevar a conclusiones erróneas([Anselin, 1988](#)). Por ello es necesario encontrar la manera de incorporar el espacio geográfico en los procesos de modelación. En este capítulo se abordará esta cuestión, mostrando primero los métodos de exploración de datos espaciales, para luego presentar las formas de modelización del espacio y los métodos de estimación.

Los modelos de econometría espacial se centran en manejar las situaciones de **dependencia espacial**. Existe dependencia espacial cuando lo que sucede en una locación i está influenciado por lo que sucede en una locación j y viceversa ([Anselin, 2013](#)). La dependencia espacial se traduce en que los valores de la variable en las locaciones i y j con $i \neq j$ están correlacionados entre sí, hecho que se conoce como **autocorrelación espacial** ([Anselin, 2013](#)). La autocorrelación espacial puede ser positiva, cuando las locaciones con valores similares tienden a estar juntas (altos con altos, bajos con bajos) o negativa, cuando las unidades espaciales tienden a estar rodeadas de vecinos con valores diferentes. Los patrones espaciales formados por la existencia

de autocorrelación se muestran en la Figura 42.1. La ausencia de algún tipo de autocorrelación es lo que se entiende como aleatoriedad espacial (Anselin, 2013).

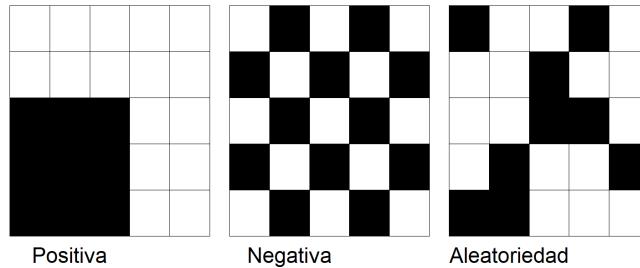


Figura 42.1: Patrones de autocorrelación espacial

42.1.1. Modelización del espacio

El espacio puede jugar un rol importante en la determinación de los procesos a modelizar. Por ello, resulta relevante encontrar una forma de incorporar el espacio en los procesos de estimación. Para modelizar la interacción de una variable consigo misma es natural pensar en el concepto de autocorrelación. No obstante, a diferencia de la autocorrelación temporal, que es unidireccional (sólo el pasado puede afectar el presente), en el caso del espacio la influencia es multidireccional en el entorno o vecindario de la localidad de análisis y, por tanto, es crucial definir el **vecindario**.

La matriz de vecindad o contigüidad $\mathbf{W}_{n \times n}$ muestra la relación entre las n localidades analizadas, y por tanto la interacción existente entre ellas. Es una matriz simétrica y binaria, de forma que $w_{ij} = 1$ si las localidades i y j son vecinas y cero si no lo son. Por tanto, $w_{ii} = 0$ puesto que una localidad no puede ser vecina de sí misma. Existen distintos criterios de definición de vecindad dependiendo del proceso que se desee modelizar y las características de los datos. Si se cuenta con un mapa de polígonos, entonces podemos utilizar alguno de los criterios que se presentan en la Figura 42.2 para configurar la matriz \mathbf{W} .

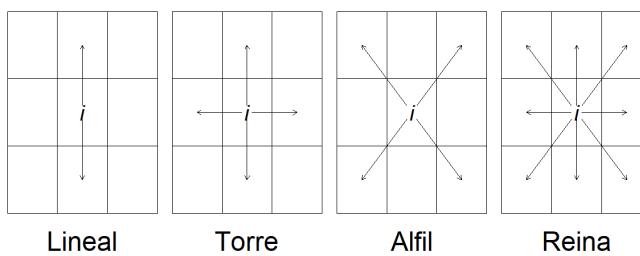


Figura 42.2: Criterios de vecindad

Las matrices \mathbf{W} generadas bajo el **criterio lineal** consideran como vecinas a la localidad i todas aquellas localidades que comparten un borde situadas en la misma dirección cardinal, norte sur

o este oeste, de esta localidad. El resto de los criterios de contigüidad siguen los movimientos de las piezas del ajedrez para definir la vecindad de la localidad i . La construcción de una matriz de vecindad bajo el **criterio de la torre** implica considerar como vecinos de la localidad i aquellas localidades situadas al norte, sur, este u oeste y que comparten un borde en común con dicha localidad. El uso del **criterio de alfil** considera como vecindad de la localidad i aquellas localidades situadas al noreste, noroeste, sureste o suroeste de la localidad i y que tengan, al menos, un punto en común. El **criterio de la reina** considera como vecindario de la unidad espacial i a las localidades en todas las direcciones cardinales y que tengan al menos un punto en común con ella (Martori et al., 2008).

Dependiendo del fenómeno que se analice, la matriz de contigüidad puede ser construida considerando un vecindario más amplio; por ejemplo, considerando como vecinos de la localidad i a los vecinos de los vecinos de dicha localidad, en este caso se dice que la matriz de vecindad es de orden 2 (los vecinos y los vecinos de los vecinos). Las matrices \mathbf{W} , se utilizan para capturar el efecto del vecindario a partir de medias ponderadas basadas en la cercanía de las unidades espaciales. Es por ello que las matrices de vecindad se estandarizan por filas. Los elementos de la matriz estandarizada se obtienen de la siguiente manera:

$$w_{ij}^e = \frac{w_{ij}}{\sum_{j=1}^n w_{ij}} \quad (42.1)$$

En palabras simples, se divide cada elemento de una fila de la matriz \mathbf{W} por la suma de dicha fila. Esto asegura que cada elemento de la matriz \mathbf{W} estandarizada se encuentre entre 0 y 1, y que la suma de cada una de sus filas sea siempre 1. Las matrices de vecindad estandarizadas llevan el nombre de matrices de pesos espaciales. A partir de ahora, cuando se haga referencia a la matriz \mathbf{W} se estará haciendo referencia a una matriz de pesos espaciales.

Para el cálculo de las matrices de vecindad se utilizará el paquete `spdep` (Bivand, 2022). La función `poly2nb()` construye la relación de vecindad a partir de los polígonos de un objeto espacial según el criterio y el orden que se indique; la función `nb2listw()` transforma la relación de vecindad en una lista de pesos espaciales. Para el ejemplificar la construcción de la matriz \mathbf{W} , se utilizará el conjunto de datos del estudio de Guerry (1833) utilizados en (Anselin, 2017). Esta base de datos contiene información respecto a estadísticas morales, criminales y sociales de las distintas provincias de Francia en 1830¹.

```
library("spdep")
library("CDR")

reina<-poly2nb(guerry, queen=TRUE)
w_reina<-nb2listw(reina, style="W", zero.policy=TRUE)
w_reina
#> Characteristics of weights list object:
#> Neighbour list object:
#> Number of regions: 85
#> Number of nonzero links: 420
```

¹Una descripción completa de la base de datos está disponible en <https://geodacenter.github.io/data-and-lab/Guerry/>

```
#> Percentage nonzero weights: 5.813149
#> Average number of links: 4.941176
#>
#> Weights style: W
#> Weights constants summary:
#>   n  nn  S0      S1      S2
#> W 85 7225 85 37.2761 347.6683
```

Para construir una matriz de contigüidad de la torre de orden 1 se utilizan las mismas funciones, cambiando el parámetro `queen` de la función `poly2nb()`

```
torre<-poly2nb(guerry, queen=FALSE)
w_torre<-nb2listw(torre, style="W", zero.policy=TRUE)
```

Cuando se trabaja con datos a nivel puntual, o cuando existen situaciones geográficas de no contigüidad como, por ejemplo, una isla, la construcción de la matriz de contigüidad no es tan evidente. En estos casos, resulta más oportuno definir la matriz de vecindad a partir de **criterios de distancia**. Las matrices de \mathbf{W} basadas en distancias pueden tener configuraciones continuas de la matriz respecto a la distancia d entre las localidades i y j de tal manera que $w_{ij} = 1/d_{ij}$ o $w_{ij} = 1/d_{ij}^2$ y $w_{ii} = 0$. Otras configuraciones implican considerar un numero k de vecinos más cercanos a cada localidad de tal manera que:

$$\begin{cases} w_{ij}(k) = 0 & i = j, \forall k \\ w_{ij}(k) = 1 & d_{ij} \leq d_i(k) \\ w_{ij}(k) = 0 & d_{ij} > d_i(k) \end{cases} \quad (42.2)$$

donde $d_i(k)$ es la k -esima menor distancia entre las localidades i y j . Utilizando funciones de [Bivand \(2022\)](#) y [Pebesma \(2022\)](#) y los datos de [Vallone and Chasco \(2020\)](#) se calculará la matriz de 5 vecinos más próximos de las áreas urbanas chilenas con más de 2000 habitantes.

```
library("sf")
#Se extraen las coordenadas de las ciudades
coord <- st_coordinates(cities)
# Calcula la vecindad de 5 vecinos más cercanos
w5knn <- knearneigh(coord, k=5, longlat= T) |> knn2nb()
```

El uso de matrices de k vecinos puede forzar la vecindad entre localidades, considerando vecinas a localidades que estén muy distantes entre ellas. Para evitar este problema se puede usar una configuración de vecindad basada en una distancia límite d_{max} , de tal manera que:

$$\begin{cases} w_{ij} = 1 & d_{ij} \leq d_{max} \\ w_{ij} = 0 & d_{ij} > d_{max} \end{cases} \quad (42.3)$$

42.2. Medidas de autocorrelación espacial

719

El problema de este criterio de vecindad es la posibilidad de generar unidades espaciales aisladas cuando d_{max} se fija en un valor muy bajo. Este problema se evita fijando la distancia máxima (d_{max}) de tal manera que se asegure que todas las unidades espaciales tengan al menos un vecino.

```
#Calcula la k=1 matriz W
knn1 <- knearneigh(coord) |> knn2nb()
# Obtiene la distancia critica
distancia_critica <- max(unlist(nbdists(knn1,coord)))
#genera la matriz de vecindad de distancia w_ij < d_max
k1 <- dnearneigh(coord, 0, distancia_critica)
w_dist <- nb2listw(k1)
```

Debe considerarse que la configuración basada en k vecinos y en la distancia censurada dan lugar a matrices binarias, mientras que las matrices \mathbf{W} basadas en distancias, no. Para realizar el cálculo de la matriz \mathbf{W} basado en la distancia inversa ($1/d_{ij}$) se utilizará en mismo conjunto de datos que en las matrices \mathbf{W} basadas en distancias. Dos elementos deben considerarse: utilizar una función decreciente respecto distancia (en este caso una hipérbola) para satisfacer la ley de Tobler (Tobler, 1970b) y segundo, dado que la incidencia que puede tener una localidad j que se encuentre muy lejana a la localidad i tiende a cero, (Tobler, 1970b), la matriz suele censurarse a una distancia máxima d_{max} a partir de la cual la incidencia entre unidades espaciales es nula, el criterio para la fijación de d_{max} es el mismo que el utilizado para evitar la existencia de unidades espaciales aisladas.

```
knn1 <- knearneigh(coord) |> knn2nb()
distancia_critica <- max(unlist(nbdists(knn1,coord)))
k1 <- dnearneigh(coord, 0, distancia_critica)
#Calcula la distancia entre los vecinos
dist_list<- nbdists(k1, st_coordinates(cities))
#Calcula la distancia inversa
i_dist_list <- lapply(dist_list, function(x) 1/x)
#Crea la matriz W
w_dist_i <- nb2listw(k1, glist=i_dist_list, style="W")
```

Se ha indicado anteriormente que la matriz \mathbf{W} se utiliza para capturar los efectos del espacio a partir de medias ponderadas, es decir mediante la matriz \mathbf{W} se puede de construir el *retardo* espacial. El retardo espacial $\mathbf{W}\mathbf{y}$ de la variable \mathbf{y} se obtiene al multiplicar dicha variable por la matriz \mathbf{W} ; por tanto, cada elemento del retardo espacial puede ser interpretado como la media ponderada de las observaciones de la variable \mathbf{y} en el vecindario de cada localidad i .

42.2. Medidas de autocorrelación espacial

Una buena herramienta para entender y comprender las medidas de autocorrelación espacial es el **diagrama de Moran** (Anselin, 1996). El diagrama de Moran relaciona una variable con lo que sucede en su entorno mediante su retardo espacial en una gráfico de puntos. La Fig. 42.3

presenta un diagrama de Moran para la base *clergy* del conjunto de datos Guerry, esta variable muestra el ratio de sacerdotes católicos sobre la población de cada provincia francesa. La linea horizontal discontinua en la Figura 42.3 muestra el promedio del retardo espacial, mientras que la linea vertical discontinua indica el promedio de la variable *clergy*. El dividir el diagrama a partir de dichos promedios permite generar cuatro zonas: el área “HH” que contiene a las localidades cuyo valor de la variable *clergy* es superior al promedio y su vecindario también. Las localidades que se sitúen en el área “LL” presentan valores de la variable *clergy* inferiores al promedio y su entorno también. El área “LH” contiene a las localidades cuyo valor de la variable *clergy* es inferior al promedio, pero su vecindario supera el valor promedio, lo contrario sucede en el área “HL”.

A partir del diagrama es posible observar la situación de una variable respecto a su entorno. Si las localidades se sitúan mayoritariamente en las zonas “HH” y “LL”, las localidades con altos valores (superiores al promedio) de la variable de interés están rodeadas por localidades con altos valores de dicha variable, y las localidades con valores bajos (inferiores al promedio) están rodeadas de localidades con valores bajos, lo cual es una señal de existencia de autocorrelación espacial positiva. Si la concentración tiene lugar en las áreas “HL” y “LH” la autocorrelación espacial negativa.

```
library("spdep")
library("CDR")
w_reina_francia<-poly2nb(guerry, queen=TRUE) |>
  nb2listw()
## Diagrama de Moran
moran.plot(guerry$clergy,w_reina_francia, xlab="Clergy",
            ylab="Retardo espacial de Clergy")
```

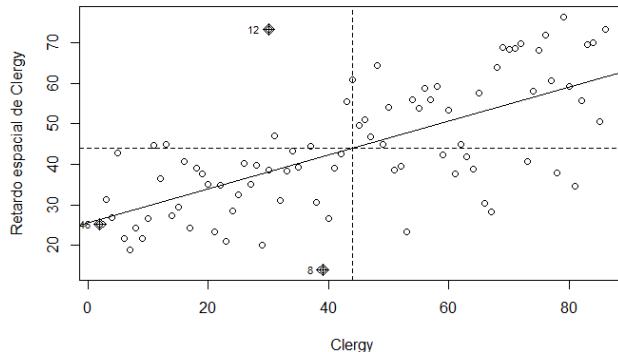


Figura 42.3: Diagrama de Moran de la variable Clergy

La Fig. 42.3 da indicios de la existencia de autocorrelación positiva, es decir que las localidades francesas tienden a estar rodeadas de localidades con numero similares de clérigos. El diagra-

ma de Moran es una herramienta gráfica, para comprobar estadísticamente la existencia de autocorrelación espacial se utilizará el indicador I de Moran.

42.2.1. El indicador I de Moran

Se define $\mathbf{z} = \mathbf{y} - \bar{\mathbf{y}}$, el indicador de Moran se calcula como:

$$I = \frac{n}{\sum_{i=1}^n \sum_{j=1}^n w_{ij}} \frac{\sum_{i=1}^n \sum_{j=1}^n z_i w_{ij} z_j}{\sum_{i=1}^n z_i^2} \quad (42.4)$$

Su campo de variación es $[-1, 1]$ y el signo coincide con el tipo de autocorrelación: valores positivos son indicativos de autocorrelación positiva y valores negativos son indicadores de la existencia de autocorrelación negativa. Nótese que I no es sino el cociente entre la covarianza de la variable \mathbf{y} y su retardo espacial $\mathbf{W}\mathbf{y}$, y la varianza de la variable \mathbf{y} . Por tanto, el coincide con el coeficiente de una regresión lineal de $\mathbf{W}\mathbf{y}$ sobre \mathbf{y} . La linea con pendiente positiva presente en 42.3 es precisamente el resultado de la regresión lineal de $\mathbf{W}\mathbf{y}$ e \mathbf{y} y por tanto su pendiente es el indicador I de Moran. En este sentido, cuanto mayor sea la pendiente de esta recta mayor será el grado de autocorrelación espacial existente. La significatividad estadística del indicador I se establece bajo la hipótesis nula de **aleatoriedad espacial**. La aleatoriedad espacial implica la inexistencia de autocorrelación en la variable analiza; es decir, considera que la variable que se analiza está distribuida en forma aleatoria entre las localidades. En este contexto, p -valores bajos permiten rechazar la hipótesis de aleatoriedad espacial, indicando la existencia de autocorrelación espacial en la variable estudiada (Anselin, 2013).

```
moran.test((guerry$clergy),w_reina_francia,randomisation=TRUE
            ,alternative="two.sided")
#>
#> Moran I test under randomisation
#>
#> data: (guerry$clergy)
#> weights: w_reina_francia
#>
#> Moran I statistic standard deviate = 6.1632, p-value = 7.13e-10
#> alternative hypothesis: two.sided
#> sample estimates:
#> Moran I statistic      Expectation      Variance
#> 0.421118648     -0.011904762     0.004936422
```

El p -valor permite rechazar la hipótesis nula de aleatoriedad espacial a favor de la existencia de autocorrelación positiva.

42.3. Modelos econométricos espaciales de corte transversal

Los modelos espaciales deben ser identificados, antes de proceder a su estimación y contraste. Para ello, es importante disponer de una estrategia de identificación propia, que permita al investigador estimar los parámetros poblacionales a partir de la observación de una muestra de datos.

Tradicionalmente, la econometría espacial ha resuelto este problema asumiendo que la especificación de los modelos es algo que se conoce a priori, bien a partir de la teoría económica existente o bien aplicando ciertas estrategias consistentes en la comparación de varios modelos competitivos. Dentro de esta última opción, se pueden destacar dos estrategias de modelización ampliamente utilizadas: la que va de lo particular (modelo básico sin efectos de autocorrelación espacial) a lo general (modelo con variables explicativas espacialmente retardadas), y la que parte de un modelo general para terminar en un modelo de autocorrelación espacial más sencillo. A partir estos los dos enfoques previos, es posible plantear la estrategia híbrida de [Elhorst \(2010\)](#).

Según se presenta en la Fig 42.4, la estrategia híbrida comienza con la estimación de un **modelo básico sin efectos espaciales**:

$$\mathbf{y} = \alpha \boldsymbol{\iota}_n + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad (42.5)$$

siendo \mathbf{y} el vector de la variable dependiente, de orden $(n \times 1)$; \mathbf{X} la matriz de variables explicativas, de orden $(n \times k)$; $\boldsymbol{\iota}_n$ un vector formado por unos, de orden $(n \times 1)$; α , $\boldsymbol{\beta}$ son el conjunto de $(p+1)$ parámetros a estimar; y $\boldsymbol{\epsilon}$ es el vector de perturbaciones aleatorias, de orden $(n \times 1)$, que se distribuye como $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma_\epsilon \mathbf{I}_n)$, siendo \mathbf{I}_n la matriz identidad de orden n .

Este modelo se estima por MCO y luego se llevan a cabo dos contrastes LM del Multiplicador de Lagrange sobre los errores de la regresión para contrastar si son ruido blanco desde el punto de vista espacial. Se trata de dos tests que contrastan una única hipótesis alternativa: el LMLAG, para la hipótesis de variable dependiente espacialmente retardada, y el LMERR, para la hipótesis de dependencia residual. La hipótesis básica se rechaza en cuanto que alguno de los estadísticos de contraste, que se distribuyen como una Chi cuadrado con 1 grado de libertad (χ^2_1) bajo la hipótesis nula, resulte estadísticamente significativo.

En primer lugar, si alguno de los tests LM resulta significativo, se recomienda seleccionar el **modelo Durbin espacial** (SDM), que es un modelo general ([Anselin, 1988](#)):

$$\mathbf{y} = \rho \mathbf{W}\mathbf{y} + \alpha \boldsymbol{\iota}_n + \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon} \quad (42.6)$$

siendo ρ un parámetro y $\boldsymbol{\theta}$ un vector de p parámetros autorregresivos espaciales.

Este modelo general incluye dos tipos de interacción espacial: los efectos endógenos ($\mathbf{W}\mathbf{y}$) y exógenos ($\mathbf{W}\mathbf{X}$). La variable endógena espacialmente retardada ($\mathbf{W}\mathbf{y}$) referida al mismo momento del tiempo que la variable dependiente (\mathbf{y}) produce en los estimadores MCO una situación

42.3. Modelos econométricos espaciales de corte transversal

723

de simultaneidad y, por tanto, sesgo, ineficiencia e inconsistencia. Por eso, se recomienda su estimación por el método de máxima verosimilitud, o “maximum likelihood” en inglés (ML), que supone normalidad en la distribución de los errores (ver [Anselin \(1988\)](#), Cap. 6).

La estimación ML de este modelo permite utilizar la ratio de verosimilitud o “likelihood ratio” (LR), cuya distribución sigue una Chi al cuadrado con k grados de libertad, como estadístico para contrastar las hipótesis nulas $H_0(\boldsymbol{\theta} = 0)$ y $H_0(\rho = 0)$, siendo las hipótesis alternativas las opciones contrarias. En este punto, se pueden dar tres casos:

- 1) Si no se rechaza la primera hipótesis, pero sí la segunda, siempre y cuando los valores de los tests LMLAG > LMERR, el SDM debería simplificarse a un modelo del retardo espacial o modelo autorregresivo espacial de orden 1 (SAR):

$$\mathbf{y} = \rho \mathbf{W}\mathbf{y} + \alpha \boldsymbol{\iota}_n + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad (42.7)$$

Este modelo se estima por ML si los errores de la estimación por MCO se distribuyen como una normal.

- 2) Si no se rechaza la segunda hipótesis, pero sí la primera, y los valores de los tests LMERR > LMLAG, debería seleccionarse el **modelo del error espacial** (SEM):

$$\begin{cases} \mathbf{y} = \alpha \boldsymbol{\iota}_n + \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \\ u = \lambda \mathbf{W}\boldsymbol{\epsilon} + \boldsymbol{\epsilon} \end{cases} \quad (42.8)$$

siendo λ un parámetro autorregresivo espacial a estimar. La estimación MCO produciría estimadores insesgados, consistentes, pero ineficientes. Por eso, se considera aceptable estimar el modelo SEM por MCO realizando una inferencia robusta de la matriz de varianzas y covarianzas de los estimadores por el método KP-HET propuesto por [Kelejian and Prucha \(2010\)](#), que tiene en cuenta la existencia conjunta de heteroscedasticidad y autocorrelación espacial en los errores de la regresión.

- 3) Si se rechazan ambas hipótesis nulas o no hubiera acuerdo entre los resultados del test LR y los tests LM, entonces el SDM sería el modelo que mejor describiría los datos.

En segundo lugar, si tras la estimación MCO del modelo básico ninguno de los tests LM fuera estadísticamente significativo, entonces dicho modelo tendría que ser reestimado como un **modelo espacial regresivo transversal** (SLX):

$$\mathbf{y} = \alpha \boldsymbol{\iota}_n + \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon} \quad (42.9)$$

Este modelo puede estimarse por MCO ya que, si las variables explicativas son exógenas, también lo serán sus correspondientes retardos espaciales. Este modelo puede considerar todas las

variables exógenas espacialmente retardadas o un subconjunto de ellas, para contrastar la hipótesis nula $H_0(\boldsymbol{\theta} = 0)$. Si esta hipótesis fuese rechazada debería elegirse el modelo básico como el que mejor describe los datos, es decir, no existiría evidencia alguna de la necesidad de efectos de autocorrelación espacial para explicar la variable dependiente. Pero si, por el contrario, la hipótesis $H_0(\boldsymbol{\theta} = 0)$ fuese rechazada, habría que estimar el modelo SDM con las variables \mathbf{WX} estadísticamente significativas, para contrastar, de nuevo, la hipótesis nula $H_0(\rho = 0)$. Si se rechaza esta hipótesis, el modelo seleccionado sería el SDM, pero en caso contrario, sería el modelo SLX el que mejor describiría los datos.

Todos estos modelos pueden también estimarse con metodología bayesiana utilizando el enfoque Markov Chains Monte Carlo (MCMC), tal y como se explica en [LeSage and Pace \(2009\)](#) Cap. 5.

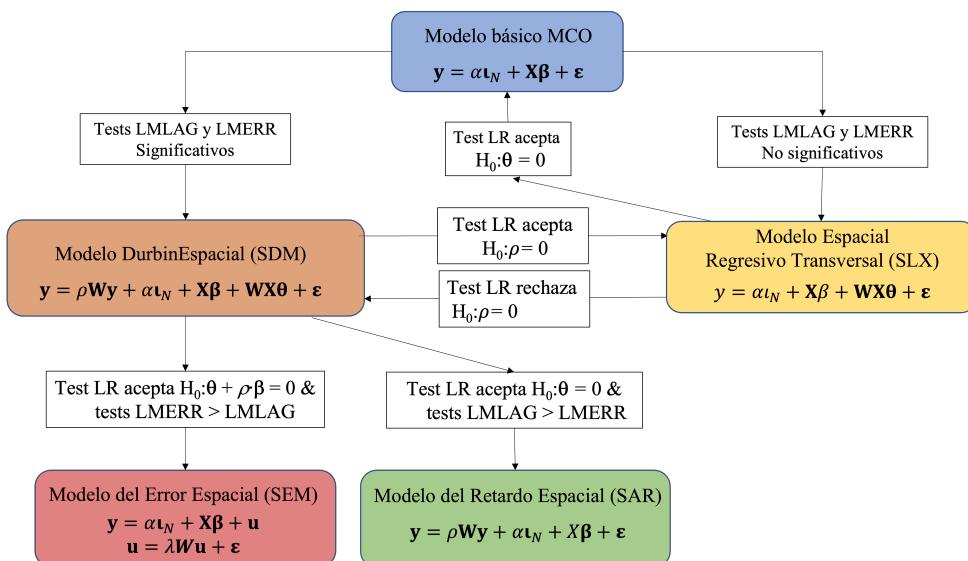


Figura 42.4: Estrategia de especificación híbrida Elhorst (2010)

El siguiente conjunto de secuencias de código muestran cómo estimar los modelos que intervienen en la estrategia de modelización de Elhorst. Para ello, se utiliza un conjunto de datos de los 120 municipios grandes de España (capitales de provincia y ciudades con población superior a 50.000 habitantes) que forman parte de las áreas urbanas de España ([Mella and Chasco \(2006\)](#)). Con estos datos, se formula un modelo de crecimiento económico urbano en España, en el que la tasa media de variación del PIB per cápita, en logaritmos, durante el período 1985-2003 (LPGH), se explica en función del PIB per capita en logaritmos de 1985 (LGH85), la tasa de variación del número de entidades bancarias en el período 1985-2003 (BANK), el porcentaje de personas con educación secundaria y universitaria sobre la población de 16 y más años en el año 2001 (UNI01) y la tasa del número de patentes por habitante en el año 2000 (PAT00).

```
library("spatialreg")
library("tseries")
```

42.3. Modelos econométricos espaciales de corte transversal

725

```

# Matriz de pesos espaciales
coord <- sp::coordinates(gdpmap)
k6 <- knearneigh(coord, k=6)
dmins <- knn2nb(k6) |> nb2listw(style="W")
# Estimación modelo básico por MCO
gdp_ols <- lm(LPGH~LGH85+BANK+UNI01+PAT00, data=gdpmap)
summary(gdp_ols)
jarque.bera.test(gdp_ols$res) # Test normalidad de residuos
lm.LMtests(gdp_ols, dmins, test="all") # Grupo de tests LM

# Estimación modelo SDM por ML
gdp_sdm <- lagsarlm(LPGH~LGH85+BANK+UNI01+PAT00, data=gdpmap, listw=dmins,
                     type="mixed")
summary(gdp_sdm)

# Estimación modeelo SAR por ML
gdp_sar <- lagsarlm(LPGH~LGH85+BANK+UNI01+PAT00, data=gdpmap, listw=dmins)
summary(gdp_sar)
LR.Sarlm(gdp_sdm, gdp_sar) # Test LR: SDM vs. SAR

# Estimación del modelo SEM por ML
gdp_err <- errorsarlm(LPGH~LGH85+BANK+UNI01+PAT00, data=gdpmap, listw=dmins,
                       tol.solve=1e-16)
summary(gdp_err)
LR.Sarlm(gdp_sdm, gdp_err) # Test LR: SDM vs. SEM

# Estimación del modelo SLX por MCO

# Cálculo retardos espaciales
gdpmap$WLGH85 <- lag(dmins, gdpmap$LGH85)
gdpmap$WBANK <- lag(dmins, gdpmap$BANK)
gdpmap$WUNI01 <- lag(dmins, gdpmap$UNI01)
gdpmap$WPAT00 <- lag(dmins, gdpmap$PAT00)

gdp_slx <- lm(LPGH~LGH85+BANK+UNI01+PAT00+WLGH85+WBANK+WUNI01+WPAT00, data=gdpmap)
summary(gdp_slx) # Modelo SLX completo

gdp_slx2 <- lm(LPGH~LGH85+BANK+UNI01+PAT00+WLGH85+WPAT00, data=gdpmap)
summary(gdp_slx2) # Modelo SLX parsimonioso
LR.Sarlm(gdp_sdm, gdp_slx2) # Test LR: SDM vs. SLX

```

42.3.1. Estimación SAR

A continuación se muestra la salida de la estimación del modelo SAR, pudiéndose observar que todos los parámetros estimados, incluido ρ , son estadísticamente significativos.

```

library("CDR")
library("spatialreg")
library("tseries")
library("spdep")

# Matriz de pesos espaciales
coord <- sp::coordinates(gdpmap)
k6 <- knearneigh(coord, k=6)
dmins <- knn2nb(k6) |> nb2listw(style="W")
# Estimación modelo SAR por ML
gdpsar <- lagsarlm(LPGH~LGH85+BANK+UNI01+PAT00, data=gdpmap,
                     listw=dmins)
summary(gdpsar)

#> Call:lagsarlm(formula = LPGH ~ LGH85 + BANK + UNI01 + PAT00, data = gdpmap,
#> listw = dmins)
#>
#> Residuals:
#> Min 1Q Median 3Q Max
#> -0.0184657 -0.0034523 0.0012278 0.0032544 0.0194746
#>
#> Type: lag
#> Coefficients: (asymptotic standard errors)
#> Estimate Std. Error z value Pr(>|z|)
#> (Intercept) 2.5745e-01 3.0703e-02 8.3851 < 2.2e-16
#> LGH85 -3.2962e-02 4.4472e-03 -7.4119 1.246e-13
#> BANK 6.6493e-05 1.3086e-05 5.0811 3.753e-07
#> UNI01 4.6884e-04 8.4080e-05 5.5762 2.459e-08
#> PAT00 4.7256e-02 1.8113e-02 2.6090 0.009082
#>
#> Rho: 0.36545, LR test value: 16.353, p-value: 5.2578e-05
#> Asymptotic standard error: 0.078929
#> z-value: 4.6302, p-value: 3.6538e-06
#> Wald statistic: 21.438, p-value: 3.6538e-06
#>
#> Log likelihood: 447.309 for lag model
#> ML residual variance (sigma squared): 3.7602e-05, (sigma: 0.006132)
#> Number of observations: 122
#> Number of parameters estimated: 7
#> AIC: -880.62, (AIC for lm: -866.27)
#> LM test for residual autocorrelation
#> test value: 6.4996, p-value: 0.01079

```

42.3.2. Comparando SAR con SDM

A continuación se muestra el resultado del test comparando el modelo SDM con el SAR. A la luz de los valores de la LR se rechaza que el valor de los parámetros restringidos sea cero y, por tanto, el modelo SDM es más adecuado para explicar esta variable que el modelo SAR.

42.3. Modelos econométricos espaciales de corte transversal

727

```

# Matriz de pesos espaciales
coord <- coordinates(gdpmap)
k6 <- knearneigh(coord, k=6)
dmins <- knn2nb(k6) |> nb2listw(style="W")

# Estimación modelo SDM por ML
gdpsdm <- lagsarlm(LPGH~LGH85+BANK+UNI01+PAT00, data=gdpmap,
                      listw=dmins, type="mixed")

# Estimación modelo SAR por ML
gdpsar <- lagsarlm(LPGH~LGH85+BANK+UNI01+PAT00, data=gdpmap,
                      listw=dmins)

LR.Sarlm(gdpsdm, gdpsar) # Test LR: SDM vs. SAR

## Likelihood ratio for spatial linear models
## data:
## Likelihood ratio = 11.559, df = 4, p-value = 0.02095
## sample estimates:
## Log likelihood of gdpsdm Log likelihood of gdpsar
## 453.0885 447.3090

```

42.3.3. Interpretación de los estimadores de los modelos de autocorrelación espacial

Sólo en los modelos de autocorrelación espacial en los que el efecto endógeno ($\mathbf{W}\mathbf{y}$) no está presente en la parte derecha del modelo, los coeficientes estimados ($\hat{\beta}$) pueden interpretarse de forma directa, como en el modelo básico sin efectos espaciales. Es decir, el efecto marginal de un cambio del valor de una variable explicativa continua en la variable explicada coincide con la estimación del coeficiente correspondiente a dicha variable, para todas y cada una de las localizaciones.

$$\frac{\partial y_i}{\partial x_{i,k}} = \beta_k \quad (42.10)$$

En los modelos SAR y SDM, la correcta interpretación de los estimadores implica antes pasar de su forma estructural a su forma reducida. Así, por ejemplo, en el modelo SAR de la expresión (42.7) la forma reducida sería (bajo ciertas condiciones de invertibilidad):

$$\mathbf{y} = (\mathbf{I} - \rho\mathbf{W})^{-1}(\alpha\boldsymbol{\iota}_n + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) \quad (42.11)$$

El término $(\mathbf{I} - \rho\mathbf{W})^{-1}$ se denomina multiplicador espacial y, utilizando la expansión potencial, puede expresarse también del modo siguiente:

$$(\mathbf{I} - \rho \mathbf{W})^{-1} = \mathbf{I} + \rho \mathbf{W} + \rho^2 \mathbf{W}^2 + \dots \quad (42.12)$$

Si se utiliza esta nueva expresión en la ecuación (42.11) se observa más claramente que el valor de \mathbf{y} en una determinada localización i es función no sólo del valor de las variables explicativas en esa localización, sino también del valor de las explicativas en las localizaciones vecinas (a través de término $\rho \mathbf{W} \mathbf{X} \boldsymbol{\beta}$), del valor de las explicativas en las localizaciones vecinas a las vecinas (a través del término $\rho^2 \mathbf{W}^2 \mathbf{X} \boldsymbol{\beta}$), etc., hasta llegar a los límites del sistema espacial en estudio.

$$E(y|\mathbf{X}) = \mathbf{X} \boldsymbol{\beta} + \rho \mathbf{W} \mathbf{X} \boldsymbol{\beta} + \rho^2 \mathbf{W}^2 \mathbf{X} \boldsymbol{\beta} + \rho^3 \mathbf{W}^3 \mathbf{X} \boldsymbol{\beta} + \dots \quad (42.13)$$

[LeSage and Pace \(2009\)](#) presentan el cambio (también llamado efecto o impacto) experimentado por \mathbf{y} en una localización i , sea cual sea i , debido a un cambio en el valor \mathbf{x}_k sobre otra localización, j , sea cual sea j . Dicho conjunto de impactos o efectos se presenta en una matriz completa, $\mathbf{S}_k(\mathbf{W})_{ij}$, de orden $n \times n$. Así, cada variable explicativa \mathbf{x}_k del modelo tendrá una matriz completa propia de impactos sobre la variable dependiente.

$$\mathbf{S}_k(\mathbf{W})_{ij} = \frac{\partial y_i}{\partial x'_k} = (\mathbf{I}_n - \rho \mathbf{W})^{-1} \mathbf{b}_k \quad (42.14)$$

En los modelos SAR y SDM, también podemos distinguir efectos directos e indirectos. El efecto directo sería el efecto causado por cambios en el valor de \mathbf{x}_k , en una localización i sobre el valor de \mathbf{y} en esa misma localización. Estos efectos son los valores de la diagonal principal de la matriz, $\mathbf{S}_k(\mathbf{W})_{ii}$. El efecto indirecto viene dado por el resto de los valores de la matriz $\mathbf{S}_k(\mathbf{W})_{ij}$, que serían los “bucles de retroalimentación” en los que el valor de \mathbf{x}_k , en una localización j afecta al valor de \mathbf{y} en la localización i , y viceversa, pudiéndose dar recorridos más largos en los que el efecto en una localización podría llegar a la última localización observada n y luego volver de nuevo al punto de partida.

Por ejemplo, $\mathbf{S}_k(\mathbf{W})_{11}$ es el efecto directo de un cambio unitario en el valor de la variable \mathbf{x}_k en la localización 1 (\mathbf{x}_{k1}) sobre el valor de la variable \mathbf{y} en esa misma localización (y_1), mientras que el valor de $\mathbf{S}_k(\mathbf{W})_{12}$ sería el efecto indirecto de un cambio unitario en el valor de la variable \mathbf{x}_k , en la primera localización, sobre el valor de la variable \mathbf{y} en la segunda (y_2). En las filas, la matriz $\mathbf{S}_k(\mathbf{W})_{ij}$ tiene los efectos de un cambio unitario en \mathbf{x}_k en la variable \mathbf{y} desde cada localización i “hacia” todas y cada una de las localizaciones j , mientras que las columnas representan el efecto de un cambio unitario en \mathbf{x}_k en la variable \mathbf{y} , provocado “desde” todas y cada una de las localizaciones i sobre la localización j .

Dado que no es posible contrastar si todos los impactos directos e indirectos contenidos en la matriz $\mathbf{S}_k(\mathbf{W})_{ij}$ son significativamente distintos de cero, o construir intervalos de confianza para ellos, LeSage y Pace proponen llevar a cabo el proceso inferencial sobre el valor medio de los efectos directos y totales, extrayendo los efectos indirectos por diferencia:

$$\bar{M}(k)_{directo} = \text{tr}(\mathbf{S}_k(\mathbf{W})) / n \quad (42.15)$$

$$\bar{M}(k)_{total} = \boldsymbol{\iota}'_n \mathbf{S}_k(\mathbf{W}) \boldsymbol{\iota}_n / n \quad (42.16)$$

$$\bar{M}(k)_{indirecto} = \bar{M}(k)_{total} - \bar{M}(k)_{directo} \quad (42.17)$$

42.3. Modelos econométricos espaciales de corte transversal

729

donde \bar{M} indica que se trata de un efecto promedio.

El siguiente conjunto de secuencias presentan el cálculo de las matrices de efectos directos, indirectos y totales, y la inferencia para los modelos SAR y SDM correspondientes al ejemplo del modelo estimado para los municipios urbanos de España.

```

library("coda")
# Cálculo de los efectos para el modelo SAR (LeSage y Pace)
Wsp <- as(as_dgRMatrix_listw(dmins), "CsparseMatrix")
trMat <- trW(Wsp, type="mult")
set.seed(1234) # Simulaciones para el proceso inferencial
gdpsar_impacts <- impacts(gdpsar, tr=trMat, R=1000)
summary(gdpsar_impacts, zstats=TRUE, short=TRUE)
HPDinterval(gdpsar_impacts, choice="direct")
HPDinterval(gdpsar_impacts, choice="indirect")
HPDinterval(gdpsar_impacts, choice="total")
plot(gdpsar_impacts, choice="direct")
plot(gdpsar_impacts, choice="indirect")
plot(gdpsar_impacts, choice="total")
plot(gdpsar_impacts, trace=TRUE, density=FALSE, choice="total")

# Cálculo de la matriz de impactos para la variable LGH85
clear.pr <- rep(NA,dim(gdpmap)[1])
names(clear.pr) <- gdpmap$MUNICIPIO
svec <- rep(0,dim(gdpmap)[1])
eye <- matrix(0,nrow=dim(gdpmap)[1],ncol=dim(gdpmap)[1])
diag(eye) <- 1
for(i in 1:length(clear.pr)){
  cvec <- svec
  cvec[i] <- 1
  res <- solve(eye - gdpsar[["rho"]]*Wsp) %*% cvec*gdpsar[["coefficients"]][["LGH85"]]
  clear.pr[i] <- res[i]
}
mult <- solve(eye - gdpsar[["rho"]]*Wsp)
deriv_LGH85 <- solve(eye - gdpsar[["rho"]]*Wsp)*gdpsar[["coefficients"]][["LGH85"]]

# Cálculo de los efectos para el modelo SDM (LeSage y Pace)
set.seed(1234) # Simulaciones para el proceso inferencial
gdpsdm_impacts <- impacts(gdpsdm, tr=trMat, R=1000)
summary(gdpsdm_impacts, zstats=TRUE, short=TRUE)
HPDinterval(gdpsdm_impacts, choice="direct")
HPDinterval(gdpsdm_impacts, choice="indirect")
HPDinterval(gdpsdm_impacts, choice="total")
plot(gdpsdm_impacts, choice="direct")
plot(gdpsdm_impacts, choice="indirect")
plot(gdpsdm_impacts, choice="total")
plot(gdpsdm_impacts, trace=TRUE, density=FALSE, choice="total")

```

42.3.4. Impacto del SDM

A continuación se presenta la salida del cálculo de los efectos para el modelo SDM (LeSage and Pace, 2009) pudiéndose observar que todos son estadísticamente significativos y, salvo en el caso de la variable LGH85, todos ellos son positivos. El signo negativo del coeficiente de LGH85 demuestra la existencia de convergencia en renta en el grupo de grandes ciudades españolas. El impacto total de un crecimiento del 10 % del PIB per cápita en una ciudad en el período inicial (1985) supuso una caída de la tasa media de variación del PIB per cápita en el período 1985-2003 del -0,63 % en dicha ciudad. Este impacto es la suma del efecto directo causado por el crecimiento del PIB per cápita en la propia ciudad (-0,43), que es el efecto directo, y el efecto indirecto proveniente del crecimiento del PIB per cápita en el resto de ciudades (-0,20). Por su parte, el efecto total del crecimiento de 10 patentes por habitante supuso un crecimiento del PIB per cápita en el período del 4,38 %, del cual un 0,5 % procedía del crecimiento de las patentes per cápita en la propia ciudad efecto directo y el 3,88 % restante fue causado indirectamente por el crecimiento de las patentes en el resto de ciudades.

```
# Cálculo de los efectos para el modelo SAR (LeSage y Pace)
Wsp <- as(as_dgRMatrix_listw(dmins), "CsparseMatrix")
trMat <- trW(Wsp, type="mult")
set.seed(1234) # Simulaciones para el proceso inferencial
gdp_sdm_impacts <- impacts(gdpsdm, tr=trMat, R=1000)
summary(gdp_sdm_impacts, zstats=TRUE, short=TRUE)
#> Impact measures (mixed, trace):
#> Direct Indirect Total
#> LGH85 -3.967704e-02 -1.656718e-02 -5.624422e-02
#> BANK 7.484574e-05 -3.878932e-05 3.605642e-05
#> UNI01 5.459202e-04 3.900067e-04 9.359269e-04
#> PAT00 5.029859e-02 1.621204e-01 2.124189e-01
#> =====
#> Simulation results ( variance matrix):
#> =====
#> Simulated standard errors
#> Direct Indirect Total
#> LGH85 4.835913e-03 1.205378e-02 1.208788e-02
#> BANK 1.289395e-05 4.475846e-05 4.830235e-05
#> UNI01 8.434442e-05 3.397464e-04 3.621312e-04
#> PAT00 2.037202e-02 1.317136e-01 1.441993e-01
#>
#> Simulated z-values:
#> Direct Indirect Total
#> LGH85 -8.246556 -1.3907437 -4.6859617
#> BANK 5.826543 -0.9252065 0.6980269
#> UNI01 6.504595 1.2126777 2.6527099
#> PAT00 2.454730 1.2664303 1.5035704
#>
#> Simulated p-values:
#> Direct Indirect Total
#> LGH85 2.2204e-16 0.16430 2.7865e-06
#> BANK 5.6587e-09 0.35486 0.4851604
```

42.3. Modelos econométricos espaciales de corte transversal

731

```
#> UNI01 7.7903e-11 0.22525 0.0079848
#> PAT00 0.014099 0.20536 0.1326920
```

```
plot(gdp_sdm_impacts, choice="direct")
plot(gdp_sdm_impacts, choice="indirect" )
plot(gdp_sdm_impacts, choice="total")
```

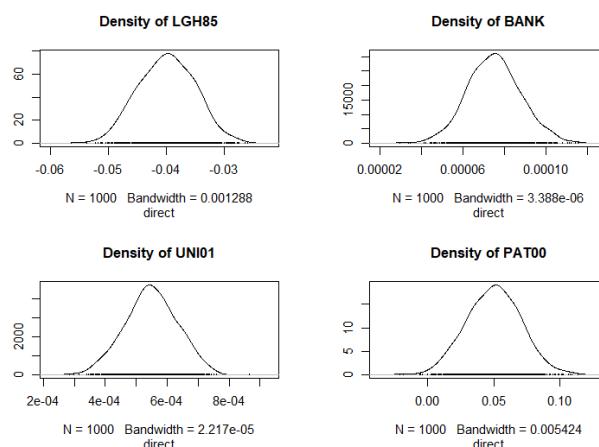


Figura 42.5: Impactos directos (SDM)

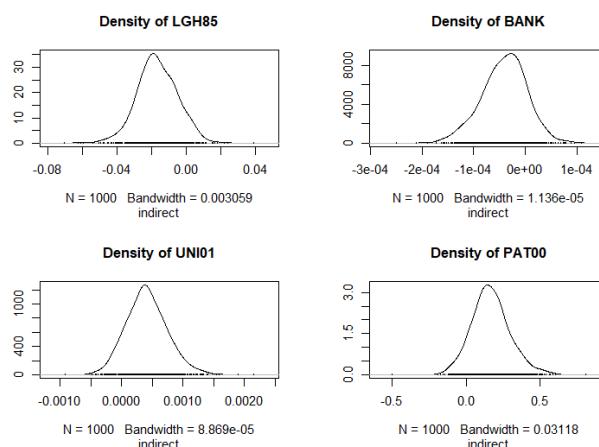


Figura 42.6: Impactos indirectos (SDM)

Como puede observarse en la Fig. 42.5, la Fig. 42.6 y la Fig. 42.7, para cada variable explicativa se estiman tres estimadores, de forma que el efecto total causado por el cambio unitario en el

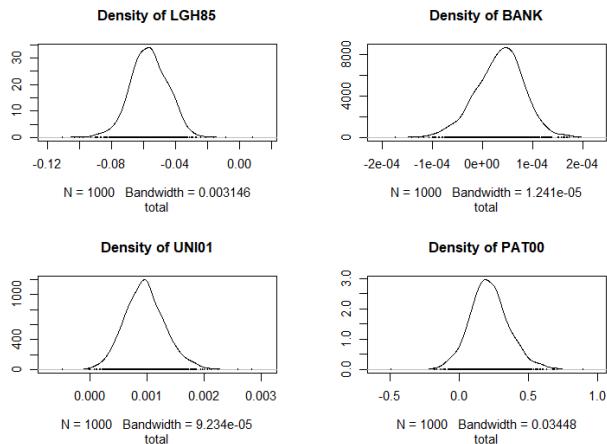


Figura 42.7: Impactos totales (SDM)

valor de dicha variable sobre el valor de la variable explicada, en una ciudad determinada, es la suma de dos efectos, uno directo, ocasionado por el cambio acaecido en la propia ciudad, y otro indirecto, proveniente del cambio acaecido en el resto de ciudades de España, existiendo tantos efectos como ciudades.

Resumen

En este capítulo se introduce a la componente espacial en la estimación econométrica y, en particular, el efecto de dependencia espacial inherente en alguna de las variables involucradas en el proceso de modelización. Primero, se observa la heterogeneidad espacial de los datos a partir de los mapas temáticos y se presenta el indicador de autocorrelación espacial de Moran. Posteriormente, se construye la matriz de pesos espaciales bajo distintas especificaciones. Por último, se muestra la taxonomía de los modelos econométricos espaciales, presentando la estrategia de especificación híbrida y la interpretación de los coeficientes estimados.

Capítulo 43

Procesos de puntos

Jorge Mateu^a y Mehdi Moradi^b

^aUniversidad Jaume I ^bUmeå Universitet

43.1. Introducción

La **estadística espacial** es una rama de la estadística que se ha desarrollado rápidamente durante los últimos treinta años, tanto en el plano teórico como en el práctico. A ello ha contribuido, de manera significativa, la creciente disponibilidad de potencia computacional y variedad en software, que han estimulado la capacidad de resolver problemas cada vez más complejos. Lo cierto es que estos problemas tienen como elemento común la estructura espacial. En general, se observa un desarrollo científico que ha ocurrido en el campo de la estadística espacial: problemas bien definidos con un carácter común saltaron a la agenda del investigador, y la disponibilidad de datos motivaron nuevos desarrollos teóricos.

La estadística espacial reconoce y explota las ubicaciones espaciales de los datos al diseñar, recopilar, administrar, analizar y mostrar dichos datos. Los datos espaciales suelen ser dependientes, y se necesitan clases de modelos espaciales que permitan la predicción de procesos y la estimación de parámetros. **Patrones espaciales** ocurren en una variedad sorprendentemente amplia de disciplinas científicas: los ecólogos estudian las interacciones entre plantas y animales, los silvicultores y agricultores deben investigar la capacidad de las plantas y tener en cuenta las variaciones del suelo en sus experimentos. Así pues, cualquier disciplina que trabaje con datos recopilados de diferentes ubicaciones espaciales, necesita desarrollar modelos que indican cuándo hay dependencia entre mediciones en diferentes lugares. Referencias modernas sobre estadística espacial incluyen los libros de [Diggle \(2013\)](#); [Cressie and Wikle \(2015\)](#); [Montero et al. \(2015\)](#); [Wikle et al. \(2019\)](#); [Diggle and Giorgi \(2019\)](#) entre otros.

Este capítulo se centra en **patrones espaciales de puntos**. Datos en forma de un conjunto de puntos, distribuidos irregularmente dentro de una región del espacio, surgen en muchos contextos diferentes; por ejemplo localizaciones de incendios forestales (Fig. 43.1), delitos (Fig.

[43.2](#)), árboles en un bosque, nidos en una colonia de cría de pájaros, ubicación de núcleos en una sección microscópica de tejido, depósitos de oro mapeados en un estudio geológico, estrellas en un cúmulo estelar, accidentes de tráfico, terremotos, llamadas de teléfonos móviles, avistamientos de animales o casos de una enfermedad rara.

Se llama **patrón espacial de puntos**, cualquier conjunto de datos de este tipo. La disposición espacial de los puntos es el principal foco de investigación. Son muchos los campos de la ciencia donde este tipo de estructuras son de interés; por ejemplo, en ecología, epidemiología, geociencia, astronomía, econometría e investigación criminal. El análisis estadístico de la disposición espacial de los puntos puede revelar características importantes, como una tendencia a que los yacimientos de oro se encuentren cerca de una gran falla geológica, o a que los casos de una enfermedad sean más frecuentes cerca de una fuente de contaminación.

El análisis de los datos de patrones de puntos ha proporcionado evidencia fundamental para importantes investigaciones, desde la transmisión del cólera hasta el comportamiento de los asesinos en serie y la estructura a gran escala del universo. Los puntos en un patrón de puntos pueden tener todo tipo de atributos. Un estudio forestal podría registrar cada ubicación, especie y diámetro del árbol; un catálogo de estrellas puede dar sus posiciones en el cielo, masas, formas y colores; las ubicaciones de los casos de enfermedades pueden estar vinculadas a registros clínicos detallados. Esta información auxiliar adjunta a cada punto en el patrón de puntos se llama **marca** y en ese caso se habla de un patrón de puntos marcado. La colección de localizaciones de un patrón puntual puede venir definida en una **región plana** (Sec. [43.2](#)) o bien en una **red lineal** (Sec. [43.3](#)), haciendo que las distancias dejen de ser euclidianas para pasar a ser del camino más corto. Esto introduce ciertos cambios metodológicos en cuanto a las construcciones de ciertas características, que en el caso de intensidades de primer orden trataremos en este capítulo.

43.2. Patrones puntuales espaciales en \mathbb{R}^2

La teoría de procesos puntuales espaciales constituye la base para el análisis de eventos observados geográficamente a través de sus coordenadas (longitud, latitud) en un espacio bidimensional. Esta rama de los procesos puntuales pertenece al campo de la estadística espacial en conjunción con la de procesos estocásticos. De hecho, un proceso puntual espacial es un proceso estocástico cuyas realizaciones consisten en un conjunto numerable de puntos en el plano (patrón puntual). Heurísticamente, se trata de un conjunto de datos que se encuentran en una región concreta (o área de estudio).

Sea $\mathbf{x} = \{x_1, x_2, \dots, x_n\}, 0 \leq n < \infty$, una realización (patrón puntual) observada de un proceso puntual simple (i.e. sin múltiples eventos por localización) y finito X en \mathbb{R}^2 en la región $W \subset \mathbb{R}^2$ y con la métrica (distancia) asociada $d(u, v)$. En general, las realizaciones consisten en un conjunto numerable de puntos (llamados en muchas ocasiones eventos). Consultar las Figs. [43.1](#) y [43.2](#) para ver algunos ejemplos de patrones puntuales. Para cualquier conjunto arbitrario $A \subset \mathbb{R}^2$, el cardinal de X viene dado por la función de conteo

$$N(X \cap A) = \sum_{x \in X} \mathbf{1}\{x \in A\} < \infty.$$

Además, y gracias a la fórmula de Campbell ([Baddeley et al., 2015](#)), para cualquier función medible $f : \mathbb{R}^2 \rightarrow [0, \infty)$ se cumple que

$$\mathbb{E} \left[\sum_{x \in X} f(x) \right] = \int_{\mathbb{R}^2} f(u) \lambda(u) du,$$

donde $\lambda(\cdot)$ determina la **función de intensidad** de X , y gobierna su distribución espacial. De hecho, $\lambda(u)$ proporciona el valor esperado de eventos por unidad de área en un entorno de $u \in \mathbb{R}^2$. Teniendo en cuenta que $f(x) = \mathbf{1}\{x \in A\}$, se puede observar fácilmente la relación entre la función de intensidad $\lambda(\cdot)$ y la de conteo N , establecida como

$$\mathbb{E}[N(X \cap A)] = \int_A \lambda(u) du.$$

Si la función de intensidad $\lambda(\cdot)$ es constante, i.e. $\lambda(\cdot) = \lambda$, se dice que el proceso X es homogéneo, mientras que, en caso contrario, se dice que es inhomogéneo; en este último caso, la distribución espacial varía a lo largo de la región soporte. Para el lector con un mayor interés en conceptos y desarrollos, se aconseja consultar [Møller and Waagepetersen \(2003\)](#); [Illian et al. \(2008\)](#); [Diggle \(2013\)](#) y [Baddeley et al. \(2015\)](#).

En la práctica se suele observar sólo una única realización, y por ello es importante disponer de una estimación de $\lambda(\cdot)$ que pueda imitar la distribución espacial del proceso subyacente, el cual ha generado el patrón observado. Por ello, se consideran diferentes tipos de **estimadores no paramétricos de la intensidad**.

43.2.1. Estimación de la intensidad basada en funciones núcleo

Dos estimadores no paramétricos, basados en **funciones núcleo**, de la función de intensidad ampliamente utilizados en patrones puntuales en \mathbb{R}^2 , vienen dados por

$$\hat{\lambda}_\sigma^U(u) = \frac{1}{c_{\sigma,W}(u)} \sum_{i=1}^n \kappa_\sigma(u - x_i), \quad u \in W, \tag{43.1}$$

y

$$\hat{\lambda}_\sigma^{JD}(u) = \sum_{i=1}^n \frac{\kappa_\sigma(u - x_i)}{c_{\sigma,W}(x_i)}, \quad u \in W, \tag{43.2}$$

donde κ_σ es una función de densidad de probabilidad en \mathbb{R}^2 con parámetro de suavizado (ancho de banda) σ , y

$$c_{\sigma,W}(u) = \int_W \kappa_\sigma(u - v) dv, \quad u \in W, \tag{43.3}$$

es el área del núcleo centrado en $u \in W$, y equivale a un corrector de borde que compensa por la falta de información fuera de W . Hay que recordar que, en la práctica, sólo se observa una realización de X en la región acotada W . Más allá de la elección de σ , el estimador (43.1) es insesgado si la función de intensidad es constante ([Diggle, 1985](#)), mientras que el estimador

(43.2) conserva la masa total (Jones, 1993). Los estimadores (43.1) y (43.2) suelen ser llamados ‘uniformly-edge-corrected’ y ‘Jones-Diggle’ (Rakshit et al., 2019b). En este capítulo, se considera en todo momento la función núcleo Gaussiana (Silverman, 1986).

En términos prácticos, la adecuación de los estimadores basados en núcleos depende del parámetro de suavizado, de forma que un suavizamiento pequeño lleva a un sesgo (por debajo) y varianza alta, mientras que un parámetro de suavizado alto resulta en un sesgo alto y poca varianza. Para un cierto patrón puntual \mathbf{x} , los estimadores (43.1) y (43.2) pueden ser calculados utilizando la función `density.ppp()` de `spatstat.core` especificando `diggle=FALSE` y `diggle=TRUE`, respectivamente.

43.2.1.1. Selección del parámetro de suavizado

Scott (1992) propuso elegir este parámetro a través de una regla un tanto naive (llamada *rule of thumb*), de la forma

$$(s_x n^{-1/6}, s_y n^{-1/6}),$$

para cada coordenada cartesiana x, y , donde s_x, s_y son las desviaciones típicas de las coordenadas x, y de los eventos. Este procedimiento es útil para análisis exploratorios. La función ‘bw.scott’ de ‘spatstat.explore’ proporciona este estimador. Nótese que, en el caso de Scott, el parámetro de suavizado es, por construcción, un vector de dos componentes para suavizar ambas coordenadas cartesianas.

Cronie and Van Lieshout (2018) propusieron encontrar el parámetro óptimo minimizando

$$CvL(\sigma) = \left(|W| - \sum_{i=1}^n 1/\widehat{\lambda}_\sigma^*(x_i) \right)^2,$$

donde $\widehat{\lambda}_\sigma^*(x_i)$ es un estimador de la intensidad sin corregir (bien sea (43.1) o (43.2) pero sin el término de corrección) evaluado en x_i y con parámetro de suavizado σ . La idea de este estimador proviene de la fórmula de Campbell, ya que

$$\mathbb{E} \left[\sum_{x \in X} 1/\lambda(x) \right] = \int_W (1/\lambda(x)) \lambda(x) du = |W|.$$

Para un patrón puntual \mathbf{x} , la función ‘bw.CvL’ de ‘spatstat.explore’ calcula el parámetro de suavizado mediante el método de Cronie y van Lieshout (se denominará por Cronie–van Lieshout).

43.2.2. Ejemplos prácticos

En esta sección se hace uso de los estimadores de la intensidad anteriormente mostrados y de los diferentes métodos de selección del parámetro de suavizado para analizar la distribución espacial de dos conjuntos de datos: incendios forestales en Nepal (Fig. 43.1), y eventos de crímenes en Medellín, Colombia (Fig. 43.2). En este capítulo, haremos uso de las librerías ‘spatstat, versión 2.3-0,’ (Baddeley and Turner, 2005; Baddeley et al., 2015) para el análisis de patrones puntuales y ‘raster, versión 3.5-15,’ (Hijmans, 2022) para ciertas representaciones gráficas. Nótese

que la librería ‘spatstat’ ha sido recientemente dividida en una familia de sub-librerías ‘spatstat.utils’, ‘spatstat.data’, ‘spatstat.sparse’, ‘spatstat.geom’, ‘spatstat.random’, ‘spatstat.core’, ‘spatstat.linnet’, ‘spatstat.explore’, ‘spatstat.model’, de forma que ‘spatstat’ actúa como una librería paraguas de todas ellas. Los lectores deben estar atentos a posibles futuros cambios en ‘spatstat’ para satisfacer ciertas restricciones de CRAN en relación con los tamaños de sus librerías.

43.2.2.1. Ejemplo 1: Incendios forestales en Nepal

Por cortesía de Ganesh Prasad Sigdel, se dispone de localizaciones georeferenciadas de incendios forestales en Nepal durante 2016, datos cedidos por la institución ICIMOD-Nepal. En 2016, Nepal sufrió 5757 incendios, de los cuales 475 ocurrieron en el distrito de Surkhet, en la provincia de Karnali, en el medio-oeste de Nepal. Se comienza llamando a algunas librerías de **R**, útiles para nuestros propósitos.

```
#library("raster")
#library("spatstat")
#library("CDR")
```

Utilizando los métodos descritos en la Sec. 43.2.1.1, se estima el correspondiente parámetro de suavizado. La regla de Scott proporciona los valores (50253.47 m, 21158.42 m) y el método de validación cruzada de Cronie–van Lieshout estima el valor como 36513.16 m. Mediante el argumento ‘ns’ de ‘bw.CvL’, se puede controlar mejor la búsqueda del parámetro óptimo a través de un grid más fino.

```
data(nepal)
scott_nepal <- bw.scott(nepal) # Scott's rule
CvL_nepal <- bw.CvL(nepal) # Cronie and van Lieshout's criterio
```

Conocido el parámetro de suavizado, se estima la intensidad mediante los estimadores (43.1) y (43.2). La función ‘density.ppp’ proporciona una estimación basada en funciones núcleo para patrones en \mathbb{R}^2 , teniendo en cuenta que, por defecto, esta función hace uso del estimador con corrección uniforme para los bordes (‘uniformly-edge-corrected estimator’) (43.1) con un núcleo Gaussiano. Se fija ‘leaveoneout=FALSE’ para no calcular el estimador *leave-one-out*, mientras que se establece ‘positive=TRUE’ para forzar valores positivos en la densidad. Esto último obedece a que, debido a errores numéricos en el cálculo de la Transformada Rápida de Fourier, se pueden obtener valores negativos en ciertas áreas (ver la ayuda de ‘density.ppp’).

```
d_scott_nepal <- density.ppp(nepal, sigma = scott_nepal, leaveoneout = FALSE, positive
                                = TRUE)
d_cvl_nepal <- density.ppp(nepal, sigma = CvL_nepal, leaveoneout = FALSE, positive =
                                TRUE)
```

Se estima ahora la intensidad mediante el estimador de Jones-Diggle (43.1) escribiendo ‘diggle=TRUE’ en ‘density.ppp’.

```
d_scott_dig_nepal <- density.ppp(nepal, sigma = scott_nepal, leaveoneout = FALSE,
→ positive = TRUE, diggle = TRUE)
d_cvl_dig_nepal <- density.ppp(nepal, sigma = CvL_nepal, leaveoneout = FALSE, positive
→ = TRUE, diggle = TRUE)
```

Tras obtener diferentes estimadores de la intensidad bajo diferentes métodos de selección del parámetro de suavizado, a continuación se muestran estas estimaciones y se comentan sus discrepancias. Para una mejor representación gráfica, se convierten las imágenes de intensidad dadas en la clase ‘im’ a objetos de clase ‘raster’ para luego juntarlas en un ‘RasterStack’. La Fig. 43.1 muestra estas estimaciones, observándose una mayor intensidad en el sur y sur-oeste de Nepal, indicando una clara distribución no uniforme de dicha intensidad, lo que, a su vez, indica un alto grado de inhomogeneidad.

```
sp_int_nepal <- stack(raster(d_scott_nepal), raster(d_cvl_nepal),
→ raster(d_scott_dig_nepal), raster(d_cvl_dig_nepal))
sp_int_nepal <- sp_int_nepal * 10^7
names(sp_int_nepal) <- c("scott_gaus_U", "CvL_gaus_U", "scott_gaus_JD", "CvL_gaus_JD")

at <- c(seq(0, 1.4, 0.2))
pts_nepal <- as.data.frame(nepal)
coordinates(pts_nepal) <- ~ x + y
library("latticeExtra")
spplot(sp_int_nepal, at = at, scales = list(draw = FALSE), col.regions =
→ rev(topo.colors(20)), colorkey = list(labels = list(cex = 3)), par.strip.text =
→ list(cex = 3)) + layer(sp.points(pts_nepal, pch = 20, col = 1))
```

43.2.2.2. Ejemplo 2: Crímenes en Medellín

Medellín es la segunda ciudad con más población en Colombia (DANE, 2019), con un territorio urbano de 105 km², que ha sufrido de múltiples acciones criminales durante muchos años, como es bien conocido. En 2018, la Secretaría de Seguridad de Medellín reportó que el 40% de los ciudadanos se sentía inseguro, proporcionando, a modo de ejemplo, 20607 quejas de robos (Restrepo, 2019). Adicionalmente, el departamento de policía reconocía la necesidad de contratar al menos 2000 policías más para luchar contra los homicidios, robos y micro-tráfico (Monsalve, 2019).

En esta sección, sólo se analiza la distribución espacial de los eventos georeferenciados de crímenes ocurridos en Medellín durante 2005 (Sanabria et al., 2022). En 2005, ocurrieron 910 crímenes, de los cuales el porcentaje de víctimas varones fue del 66%, 28% fueron cometidos durante los fines de semana, el porcentaje de robos fue del 42%, y el de víctimas con edades entre 20 y 40 fue del 60%.

Nótese notar que el conjunto de localizaciones de estos crímenes no necesariamente ocurrió en las calles de la ciudad, y por tanto se considera que el patrón puntual tiene como dominio de definición todo \mathbb{R}^2 .

43.2. Patrones puntuales espaciales en \mathbb{R}^2

739

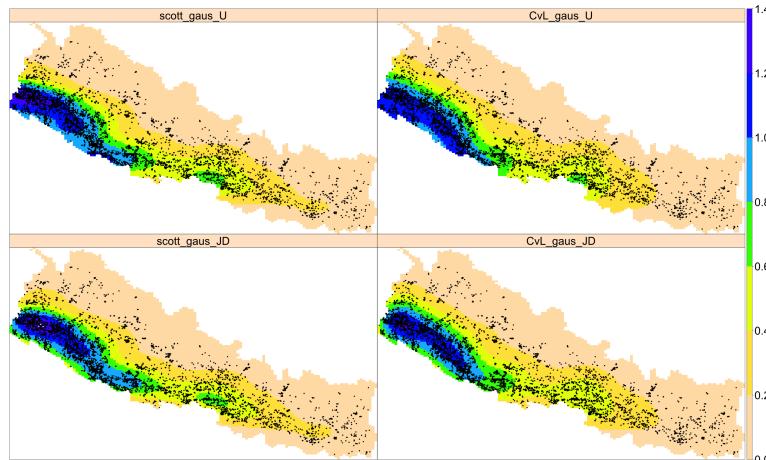


Figura 43.1: Estimación basada en funciones núcleo para los incendios forestales (puntos negros) en Nepal en 2016. Las etiquetas de los nombres comienzan con el método de suavizado, seguido del núcleo utilizado y de la corrección de borde. Los valores de la intensidad indican número de incendios por diez mil km cuadrados. Se usa JD y U para indicar los estimadores de 'Jones-Diggle' y 'uniformly-edge-corrected'.

```
data(medellin)
scott_med <- bw.scott(medellin) # Scott's rule
CvL_med <- bw.CvL(medellin) # Cronie and van Lieshout's criterio
```

La regla de Scott estima el parámetro de suavizado en (691.31m, 954.20m) mientras que el criterio de validación cruzada (CvL) nos lleva a 692.31m. Se hace uso de la función ‘density.ppp’ para obtener los correspondientes estimadores de la intensidad (43.1) y (43.2) bajo los mismos escenarios que en la Sección 43.2.2.1.

```
d_scott_med <- density.ppp(medellin, sigma = scott_med, leaveoneout = FALSE, positive =
  TRUE)
d_cvL_med <- density.ppp(medellin, sigma = CvL_med, leaveoneout = FALSE, positive =
  TRUE)

d_scott_dig_med <- density.ppp(medellin, sigma = scott_med, leaveoneout = FALSE,
  positive = TRUE, diggle = TRUE)
d_cvL_dig_med <- density.ppp(medellin, sigma = CvL_med, leaveoneout = FALSE, positive =
  TRUE, diggle = TRUE)
```

La Fig. 43.2 muestra la intensidad estimada bajo diferentes parámetros de suavizado. Se observa, en general, una distribución no homogénea de los crímenes. Independientemente del método utilizado, se observan dos grandes hotspots en la zona central de Medellín, aunque con diferentes magnitudes. El efecto de la corrección de borde es sólo marginal.

```

sp_int_med <- stack(raster(d_scott_med), raster(d_cvl_med), raster(d_scott_dig_med),
                     raster(d_cvl_dig_med))
sp_int_med <- sp_int_med * 10^5
names(sp_int_med) <- names(sp_int_nepal)
at <- seq(0, 3, by = 0.2)
pts <- as.data.frame(medellin)
coordinates(pts) <- ~ x + y

sp::spplot(sp_int_med, at = at, scales = list(draw = FALSE), col.regions =
  rev(topo.colors(20)), colorkey = list(labels = list(cex = 3)), par.strip.text =
  list(cex = 3)) + layer(sp.points(pts, pch = 20))

```

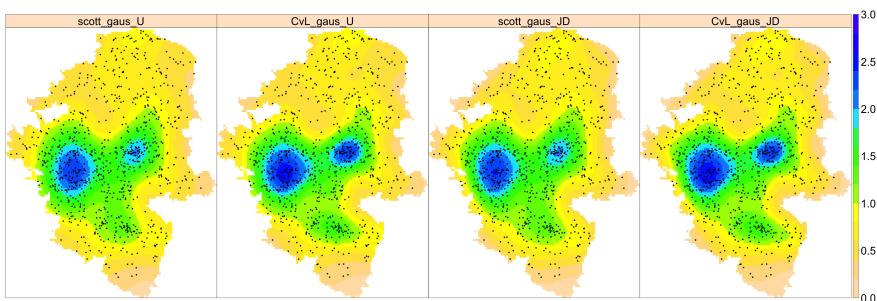


Figura 43.2: Estimación de la intensidad basada en funciones núcleo para los datos de Medellín (puntos negros), durante 2005. Las etiquetas de los nombres comienzan con el método de suavizado, seguido del núcleo utilizado y de la corrección de borde. Los valores de la intensidad indican número de crímenes por cien km cuadrados. Se usa JD y U para indicar los estimadores de ‘Jones-Diggle’ y ‘uniformly-edge-corrected’.

43.2.3. Estimación de la intensidad basada en funciones núcleo en dominios irregulares

Los estimadores (43.1) y (43.2) pueden mostrar deficiencias importantes como no cumplir la condición de que la integral sea el número de puntos, sesgo cerca de las fronteras o presentar suavizamientos artificiales que lleven a resultados inverosímiles en ciertas ocasiones (Baddeley et al., 2022). Estos problemas son más aparentes en caso de dominios irregulares. Como remedio, Baddeley et al. (2022) propusieron estimar la intensidad via una **función núcleo-calor** (*heat kernel*), la cual puede ser definida como una densidad de probabilidad de transición de un movimiento Browniano en W que respeta las fronteras. De hecho, su propuesta, llamada **estimador de difusión**, toma la forma

$$\hat{\lambda}_t(u) = \sum_{i=1}^n \kappa_t(u|x_i), \quad (43.4)$$

donde $t = \sigma^2$ (σ es el parámetro de suavizado en (43.1) y (43.2)) y $\kappa_t(\cdot|x_i)$ es el núcleo-calor. Este estimador es insesgado (bajo homogeneidad) y preserva la masa (es decir, integra el número de

puntos). Baddeley et al. (2022) proponen algunos nuevos métodos de selección del parámetro de suavizado, adaptados a su estimador de difusión, incluyendo el de Cronie–van Lieshout. El estimador de difusión se puede calcular con la función ‘densityHeat.hpp’, y el criterio de Cronie–van Lieshout viene en la función ‘bw.CvLHeat’. Todas estas funciones pertenecen al ‘spatstat.explore’.

A continuación, se utiliza el estimador de difusión para analizar su comportamiento comparándolo, con el estimador (43.1) sobre unos datos de incendios activos en EEUU y América Central (sin considerar las islas) desde el 24 de Febrero al 3 de Marzo 2022¹. Hay que hacer notar que las localizaciones, en este caso, no necesariamente confirman la existencia de un incendio, sino más bien píxeles susceptibles de existencia de incendio, los cuales han sido clasificados por medio de algoritmos preparados para ello. Este formato está relacionado con el contexto de datos en *Near Real-Time (NRT)*.

Los parámetros de suavizado para los estimadores (43.1) y (43.4) siguen el criterio de Cronie–van Lieshout. Nótese que al considerar un área mucho más grande que en los ejemplos precedentes, se considera ‘ns=50’, es decir, se usa un vector de tamaño 50 para buscar el parámetro de suavizado óptimo (por defecto es 16), y ‘dimyx=512’ para obtener imágenes de intensidad con una mejor resolución (por defecto, las imágenes son de tamaño 128×128 píxeles). Los parámetros de suavizado elegidos para calcular (43.1) y (43.4) son 556.3km y 104.9km.

```
data(activefires)
CvL_northcentre <- bw.CvL(activefires, ns = 50)
d_CvL_northcentre <- density.ppp(activefires, sigma = CvL_northcentre, leaveoneout =
  FALSE, dimyx = 512)

heat_CvL_northcentre <- bw.CvLHeat(activefires, ns = 50)
dheat_CvL_northcentre <- densityHeat.hpp(activefires, sigma = heat_CvL_northcentre,
  leaveoneout = FALSE, dimyx = 512)
```

Ambas estimaciones se juntan en un objeto `RasterBrick` que se representa en la Fig. 43.3. Obsérvese que el dominio no es regular pues los estados de Florida, California del Sur y América Central hacen que la región objeto de estudio sea ciertamente irregular. En este caso, sería poco realista si el estimador utilizado proporciona intensidad en zonas como el Golfo de California/Mexico. El mapa de intensidad que se muestra a la izquierda de la Fig. 43.3 muestra que el estimador con corrección uniforme (‘uniformly-edge-corrected’) distribuye la masa total por toda la región, provocando una sobre-suavización. Sin embargo, el mapa de la derecha, construido con el estimador de difusión, muestra una situación más realista, distribuyendo la masa de la intensidad acorde a los sucesos ocurridos.

```
d_northcentre_stack <- stack(raster(d_CvL_northcentre), raster(dheat_CvL_northcentre))
names(d_northcentre_stack) <- c("CvL_gaus_U", "Diffusion")
pts_northcentre <- as.data.frame(activefires)
coordinates(pts_northcentre) <- ~ x + y
d_northcentre_stack <- d_northcentre_stack * 10^6
```

¹https://firms.modaps.eosdis.nasa.gov/active_fire/

```
spplot(d_northcentre_stack, scales = list(draw = FALSE), col.regions =
  rev(terrain.colors(20)), colorkey = list(labels = list(cex = 5)), par.strip.text =
  list(cex = 5)) + layer(sp.points(pts_northcentre, pch = 20, col = 1))
```

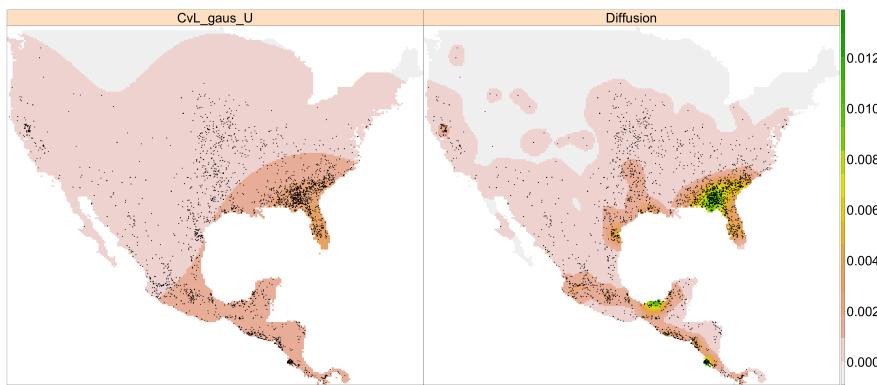


Figura 43.3: Estimación basada en función núcleo para incendios (puntos negros) en EEUU y Centro América (sin las islas) desde el 24 de Febrero hasta el 3 de Marzo de 2022. Izquierdo: estimador con corrección uniforme con núcleo Gaussiano. Derecha: estimador de difusión. El parámetro de suavizado fue obtenido con el criterio de Cronie–van Lieshout. Los valores de la intensidad son fuegos por mil km cuadrados.

43.2.4. Estimadores basados en teselaciones de Voronoi

Como se ha visto, el comportamiento de los estimadores basados en funciones núcleo depende del parámetro de suavizado, e incluso en situaciones en las que hay cambios abruptos en la distribución espacial de los puntos, un único valor constante de este parámetro no puede representar el suavizamiento necesario en toda la región. Para dar una solución a este problema, se propuso un parámetro con variación espacial (adaptable a la estructura espacial) aunque a costa de una mayor complejidad (Davies and Baddeley, 2018; Baddeley et al., 2022). Sin embargo, y como alternativa a esta propuesta, se pueden utilizar estimadores basados en teselaciones de Voronoi, que son no paramétricos (Barr and Schoenberg, 2010).

Para cada $x \in \mathbf{x}$, su celda de Voronoi/Dirichlet \mathcal{V}_x , consistente en todos los $u \in W$ que están más cercanos a x que a cualquier otro elemento $y \in \mathbf{x} \setminus \{x\}$, viene dada por

$$\mathcal{V}_x = \{u \in W : d(x, u) \leq d(y, u) \text{ para todo } y \in \mathbf{x} \setminus \{x\}\}. \quad (43.5)$$

El **estimador basado en teselaciones de Voronoi**, evaluado en cualquier punto arbitrario $u \in W$, es de la forma

$$\hat{\lambda}^V(u) = \sum_{x \in \mathbf{x}} \frac{\mathbf{1}_{\{u \in \mathcal{V}_x\}}}{|\mathcal{V}_x|}. \quad (43.6)$$

El estimador $\widehat{\lambda}^V(u)$ conserva la masa (al igual que $\widehat{\lambda}_\sigma^{\text{JD}}(u)$), y es insesgado si la intensidad real es constante (igual que $\widehat{\lambda}_\sigma^{\text{U}}(u)$), propiedades compartidas por el estimador de difusión. Sin embargo, Moradi et al. (2019) demostraron que $\widehat{\lambda}^V(u)$ tiene una varianza alta, lo que implica una infrasuavización en áreas densas de puntos y una sobre-suavización en áreas con pocos puntos. Por tanto, estos autores proponen corregir el problema de $\widehat{\lambda}^V(u)$ mediante un sub-muestreo de $m \geq 1$ copias re-escaladas \mathbf{x} a través de adelgazamientos independientes (*independent p-thinning*). Su propuesta viene dada por

$$\widehat{\lambda}_{p,m}^V(u) = \frac{1}{m} \sum_{i=1}^m \frac{\widehat{\lambda}_i^V(u)}{p}, \quad u \in W, \quad (43.7)$$

donde $\widehat{\lambda}_i^V(u)$ es el estimador de Voronoi del i -ésimo patrón adelgazado. La idea es que este nuevo estimador balancee mejor la varianza en función de la cantidad de puntos presentes en la subregión, y este efecto se consigue con muestras de menor tamaño procedentes del patrón original. El estimador $\widehat{\lambda}_{p,m}^V(u)$ se conoce como **estimador de remuestreo-suavizado** (*resample-smoothed*), y adicionalmente a las propiedades estadísticas de $\widehat{\lambda}^V(u)$, tiene una varianza bastante más pequeña. En este caso, también se debe seleccionar a priori (m, p) ; sin embargo, Moradi et al. (2019) proponen tanto una *rule-of-thumb* ($m = 400$ y $p \leq 0.2$) como una validación cruzada. Ambos estimadores (43.6) y (43.7) son accesibles por medio de la función ‘densityVoronoi.ppp’ de ‘spatstat.explore’ y en la que los argumentos ‘f’ y ‘nrep’ controlan la probabilidad p y el número de adelgazamientos m . Fijando ‘f=1’ se puede obtener el estimador basado en Voronoi (43.6).

A modo de ejemplo, se estima la intensidad de los incendios en Nepal (Sec. 43.2.2.1) mediante el método de Voronoi resample-smoothed (43.7) considerando diferentes probabilidades de retención para el adelgazamiento correspondiente.

```
d_vor_1_nepal <- densityVoronoi.ppp(nepal, f = 1, nrep = 1)
d_vor_2_nepal <- densityVoronoi.ppp(nepal, f = 0.8, nrep = 400)
d_vor_3_nepal <- densityVoronoi.ppp(nepal, f = 0.6, nrep = 400)
d_vor_4_nepal <- densityVoronoi.ppp(nepal, f = 0.5, nrep = 400)
d_vor_5_nepal <- densityVoronoi.ppp(nepal, f = 0.4, nrep = 400)
d_vor_6_nepal <- densityVoronoi.ppp(nepal, f = 0.2, nrep = 400)
d_vor_7_nepal <- densityVoronoi.ppp(nepal, f = 0.1, nrep = 400)
d_vor_8_nepal <- densityVoronoi.ppp(nepal, f = 0.05, nrep = 400)
```

Las estimaciones obtenidas, al igual que las que proceden de `density.ppp`, son de clase `im` y se unen en un objeto `RasterBrick` para su representación gráfica.

```
sp_int_nepal_v <- stack(raster(d_vor_1_nepal), raster(d_vor_2_nepal),
                         raster(d_vor_3_nepal), raster(d_vor_4_nepal), raster(d_vor_5_nepal),
                         raster(d_vor_6_nepal), raster(d_vor_7_nepal), raster(d_vor_8_nepal))
names(sp_int_nepal_v) <- NULL
names <- as.character(sort(c(seq(.2, 1, .2), 0.1, 0.05, 0.5), decreasing = TRUE))
names <- paste("p =", names)
```

```

sp_int_nepal_v <- sp_int_nepal_v * 10^7
at <- c(0, 0.3, 0.7, seq(2, 5, 1), 30)

spplot(sp_int_nepal_v, at = at, colorkey = list(labels = list(cex = 3)), col.regions =
  ~ topo.colors(20), scales = list(draw = FALSE), par.strip.text = list(cex = 3),
  ~ names.attr = names)

```

La Fig. 43.4 muestra las intensidades procedentes de los estimadores Voronoi *resample-smoothed* para los incendios de Nepal, y para diferentes probabilidades de retención. Se puede observar un menor suavizamiento y mayor varianza para altas probabilidades. Asimismo, se puede ver que para probabilidades de retención menores que 0.2, el estimador proporciona mejores suavizamientos locales que los basados en suavizamientos fijos.

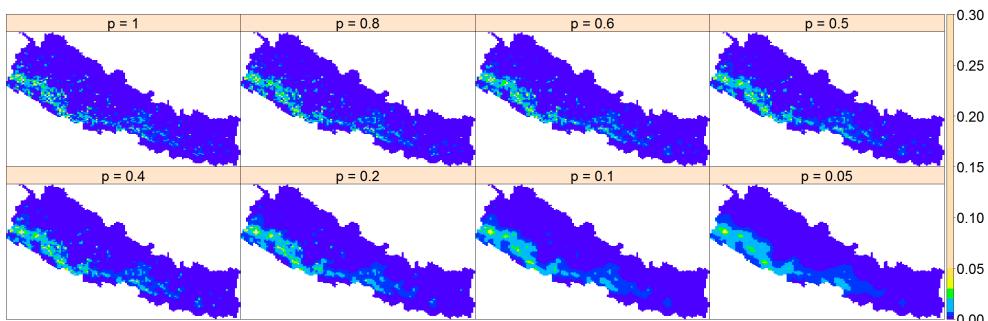


Figura 43.4: Estimaciones de la intensidad de Voronoi resample-smoothed para los incendios en Nepal en 2016 para diferentes probabilidades de retención. La intensidad proporciona el número de incendios por diez mil km cuadrados.

43.2.5. Características de segundo orden: la función K de Ripley

La función de intensidad presentada en las secciones anteriores describe el número esperado de puntos por unidad de espacio, y no tiene en cuenta la estructura de dependencia entre dichos puntos. Esta estructura, sin embargo, viene caracterizada a través de lo que se llaman características de segundo orden. Las funciones de segundo orden determinan la estructura de dependencia espacial (o en su caso espacio-temporal, si interviene el tiempo) inherente al patrón puntual. La literatura ha propuesto varias funciones de segundo orden, de entre las cuales la función K de Ripley es posiblemente la más utilizada. Esta función se define de forma pragmática como el número medio de eventos en un radio r alrededor de cualquier otro evento. Dicho de otra forma, la función $K(r)$ representa el número medio de eventos dentro de un círculo de radio r alrededor de un evento típico del patrón (sin contar dicho evento central). De esta forma, $K(r)$ describe características del proceso de puntos a muchas escalas (tantas como diferentes r se consideren). Esta función puede venir corregida por la intensidad de primer orden en el caso de procesos inhomogéneos. Ambas versiones de la función K vienen implementadas en ‘spatstat’

a través de las funciones ‘Kest’ y ‘Kinhom’ para los casos homogéneo e inhomogéneo. Una propiedad interesante de esta función es que tiene una forma cerrada bajo el caso de aleatoriedad espacial completa, es decir, bajo la situación en la que el patrón de puntos es totalmente aleatorio, sin dependencia espacial alguna (llamado, en este caso, **proceso de Poisson**). Como bajo esta suposición, $K(r) = \pi r^2$ se puede contrastar si un cierto patrón es o no aleatorio, construyendo bandas de confianza sobre la función K evaluada bajo simulaciones de aleatoriedad y evaluando la función K empírica procedente de los datos. La función ‘envelope’ construye tales intervalos de confianza.

También se han utilizado otras funciones para describir y contrastar patrones espaciales; estas funciones están basadas en la distribución de distancias entre puntos que existiría en un patrón de Poisson, como por ejemplo, la función de distribución de distancias al vecino más próximo, la función de distribución de distancias a un punto fijo aleatorio, o la función J , una combinación de las anteriores. Todas estas funciones, incluida la función K , son en cierta forma funciones de distribución ya que, a cada escala o distancia r , todos los pares de puntos separados por una distancia menor que r se usan para estimar el valor de la correspondiente función. En ocasiones, puede ser necesario disponer de una función que caracterice de forma no acumulativa el patrón, es decir, que tenga en cuenta tan sólo los pares de puntos que se encuentran separados por una distancia exactamente igual o similar a la distancia r . La función de correlación de par $g(r)$ (*pair correlation function*) es la herramienta apropiada en este caso (Baddeley et al., 2015).

A continuación se muestra el código y resultados de llevar a la práctica la estimación de la función K (inhomogénea) y los correspondientes intervalos de confianza bajo aleatoriedad para los casos de incendios en Nepal y delitos en Medellín.

```
d_nepal <- density.ppp(nepal, bw.scott, leaveoneout = TRUE)
en_nepal <- envelope(nepal, fun = Kinhom, correction = "border", nsim = 99, simulate =
  expression(rpoispp(d_nepal)), sigma = bw.scott, normpower = 2)

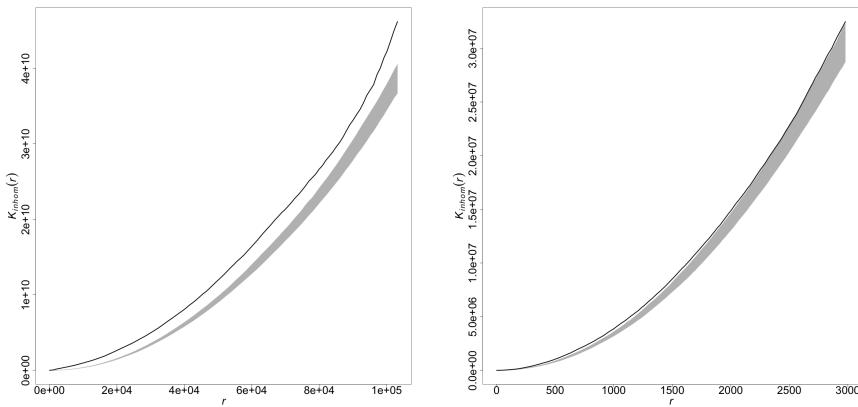
d_med <- density.ppp(medellin, bw.scott, leaveoneout = TRUE)
en_med <- envelope(medellin, fun = Kinhom, correction = "border", nsim = 99, simulate =
  expression(rpoispp(d_med)), sigma = bw.scott, normpower = 2)

en_nepal$mmean <- NULL
plot(en_nepal, main = "", lwd = 3, cex.axis = 2.5, cex.lab = 2.5, legend = FALSE)

en_med$mmean <- NULL
plot(en_med, main = "", lwd = 3, cex.axis = 2.5, cex.lab = 2.5, legend = FALSE)
```

43.3. Patrones puntuales espaciales sobre redes lineales

En los últimos diez años, los patrones de puntos en redes lineales han recibido mucha atención científica. Una red lineal es un conjunto de segmentos (o aristas) unidos por nodos con un formato lineal, tipo una combinación convexa entre dos nodos. La explicación inicial detrás de la consideración de redes lineales, como espacios de estado de algunos procesos puntuales,

Figura 43.5: Funciones K de Ripley para incendios en Nepal y delitos en Medellín

podría estar en el hecho de que los objetos definidos en tales estructuras no pueden usar todo el espacio, y sus movimientos dependen fuertemente de su libertad sobre tales estructuras (Okabe and Sugihara, 2012). En consecuencia, entre otras cosas, la distribución espacial de los puntos, así como la correlación entre ellos, debe estudiarse con respecto a la red subyacente. Sin embargo, no ha sido tan fácil lidiar con este cambio de soporte cuando se pretende adaptar metodologías estadísticas para el análisis de patrones de puntos en redes lineales. Los principales desafíos no fueron solo matemáticos/estadísticos, sino también computacionales (Moradi, 2018; Baddeley et al., 2021).

Una red lineal es una unión de segmentos de línea $l_i = [u_i, v_i] = \{tu_i + (1-t)v_i : 0 \leq t \leq 1\} \subset \mathbb{R}^2$, y una elección común de métrica sobre dicha estructura ha sido inicialmente la distancia de ruta más corta (*shortest-path distance*) $d_L(u, v)$, aunque más tarde Rakshit et al. (2017) propusieron otros tipos de distancias, incluida la distancia euclídea. La idea es que moverse por la red de segmentos implica respetar la geometría de dicha red y por tanto las líneas rectas (que sería el caso de usar distancias euclídeas) no son adecuadas. La distancia de ruta más corta si que permite adaptarse a esta geometría. Sea Y un proceso puntual en una red lineal L , la fórmula de Campbell (43.2) se adapta como

$$\mathbb{E} \left[\sum_{y \in Y} f(y) \right] = \int_L f(z) \lambda(z) d_1 z,$$

donde d_1 denota integración con respecto a la longitud de arco. En este caso, $\lambda(z)$ proporciona el número esperado de puntos por unidad de longitud de L en una vecindad de $z \in L$. Se han desarrollado distintos estimadores de la intensidad para patrones en redes considerando métricas adecuadas y resolviendo ciertos obstáculos matemáticos. El lector puede leer más al respecto en Moradi (2018) y Baddeley et al. (2021); en particular, se recomienda leer sobre el método de estimación no paramétrica basado en convoluciones bi-dimensionales de (Rakshit et al., 2019b). Dada una realización $y = \{y_1, y_2, \dots, y_n\}$ de un proceso puntual Y sobre una

43.3. Patrones puntuales espaciales sobre redes lineales

747

red lineal L , estos autores propusieron

$$\hat{\lambda}_\sigma^U(z) = \frac{1}{c_{\sigma,L}(z)} \sum_{i=1}^n \kappa_\sigma(z - y_i), \quad z \in L, \quad (43.8)$$

con una corrección uniforme, y

$$\hat{\lambda}_\sigma^{JD}(z) = \sum_{i=1}^n \frac{\kappa_\sigma(z - y_i)}{c_{\sigma,L}(y_i)}, \quad z \in L, \quad (43.9)$$

con la corrección de Jones-Diggle, donde κ_σ es una función núcleo bivariante con suavizado σ , y

$$c_{\sigma,L}(z) = \int_L \kappa_\sigma(z - v) d_1 v,$$

es una corrección de borde.

Los dos estimadores anteriores tienen propiedades estadísticas similares a las de sus análogos para patrones de puntos espaciales en \mathbb{R}^2 (es decir, los estimadores (43.1) y (43.2)), y se pueden calcular rápidamente incluso en redes grandes y para grandes anchos de banda (parámetros de suavización). El cálculo rápido se logra mediante la transformada rápida de Fourier (FFT) (Silverman, 1982). Además, Rakshit et al. (2019b) propusieron utilizar las versiones adaptadas de la regla de Scott, a la cual se puede acceder a través de la funciones ‘bw.scott.iso’ de ‘spatstat.linnet’, para obtener un ancho de banda óptimo. Nótese que el cálculo rápido de los estimadores anteriores simplifica aún más el cálculo de los estimadores de intensidad basados en el núcleo adaptativo y el riesgo relativo sobre las estructuras de red (Rakshit et al., 2019b).

También se recuerda que Moradi et al. (2019) propusieron su enfoque de sub-muestreo basado en Voronoi para procesos de puntos generales; para patrones de puntos en redes lineales puede calcularse mediante la función ‘densityVoronoi.lpp’ de ‘spatstat.linnet’.

Como ejemplo práctico para esta sección, se estudia la distribución espacial de delitos callejeros en Valencia. Valencia es la tercera ciudad más grande de España, siendo la capital de la Comunidad Valenciana. El territorio urbano de Valencia encierra un área de 134,65 km², con más de 800000 habitantes en el municipio. El conjunto de datos consta de las ubicaciones de 90247 delitos callejeros como agresión (55610 casos), robo (25342 casos), robo contra la mujer con violencia (454 casos) y otros tipos de delitos (8841 casos). Estos delitos se cometieron entre 2010 y 2020. Sin embargo, en lo que sigue, el análisis únicamente se centra en los datos correspondientes al año 2020, que incluye 6868 casos, de los cuales 4077 son agresiones, 2060 son robos y 66 se relacionan con delitos contra la mujer con violencia. Este conjunto de datos es propiedad de la Generalitat Valenciana (GV), se obtuvieron a través del teléfono de emergencias 112, y se puso a disposición de los autores gracias a un convenio entre GV y la Universidad Jaume I.

A continuación, estimar el parámetro de suavizado utilizando la regla general de Scott, que da 584,1m. La función ‘densityQuick.lpp’ de ‘spatstat.linnet’ se usa para calcular cualquiera de los estimadores (43.8) y (43.9) en los que su valor predeterminado calcula el estimador de borde uniforme corregido (43.8).

```
data(valencia)
scott_valencia <- bw.scott.iso(valencia) # Scott rule
d_scott_valencia <- densityQuick.lpp(valencia, sigma = scott_valencia, leaveoneout =
  FALSE, positive = TRUE, dimyx = 512)
d_scott_valencia <- d_scott_valencia * 1000
```

Las imágenes obtenidas son de tipo 'linim', y se convierten en objetos de clase 'im' antes de pasarlas a objetos 'raster'.

```
par(mfrow = c(1, 2))
plot(valencia$domain>window, lwd = 4)
plot(valencia, pch = 20, main = "", lwd = 4, cex = 1, add = T, cols = "red", col =
  "blue")
plot(raster(as.im(d_scott_valencia)), main = "", axis.args = list(cex.axis = 4),
  legend.width = 2, zlim = c(0, 6))
plot(valencia$domain>window, add = TRUE, lwd = 4)
par(mfrow = c(1, 1))
```

La Fig. 43.6 muestra la intensidad estimada junto con los eventos de delitos. Dicha intensidad identifica las zonas central y norte de la ciudad de Valencia como áreas de alto riesgo junto con otras zonas de bajo riesgo como son el este y la costa de la ciudad.

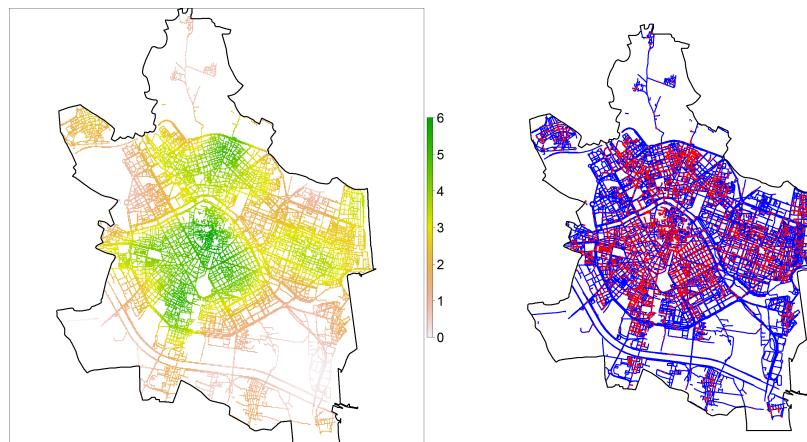


Figura 43.6: Intensidad estimada por función núcleo, usando un estimador borde uniforme corregido (izquierda), para los datos de delitos (puntos rojos) en Valencia durante 2020 (derecha). Los valores de intensidad muestran número de crímenes por km lineal.

```
d_vlc <- densityQuick.lpp(valencia, sigma = scott_valencia, leaveoneout = TRUE,
  positive = TRUE, at = "points", dimyx = 512)
d_vlc_im <- densityQuick.lpp(valencia, sigma = scott_valencia, leaveoneout = TRUE,
  positive = TRUE, dimyx = 512)
```

43.3. Patrones puntuales espaciales sobre redes lineales

749

Finalmente, se muestra la función K de Ripley y el intervalo de confianza bajo un proceso de Poisson en una red lineal (ver Fig. 43.7 (Ang et al., 2012; Rakshit et al., 2019a)). Se observa que la función K empírica cae dentro de las bandas indicando que el tipo de delitos considerados, en 2020, es compatible con un proceso aleatorio. Obsérvese que al no considerar el tiempo, podemos detectar clusters espaciales que no existen en realidad pues estos desaparecerían con la evolución temporal.

```
sim_vlc <- rpoislpp(lambda = d_vlc_im, L = net_vlc, nsim = 199)
library(spatstat.Knet)
K_vlc <- Knetinhom(valencia, lambda = as.numeric(d_vlc))
r <- K_vlc$r

K_sim <- lapply(X = 1:199, function(i) {
  sigma <- bw.scott.iso(sim_vlc[[i]])
  lambda <- densityQuick.lpp(sim_vlc[[i]], sigma = sigma, leaveoneout = TRUE, positive
  ← = TRUE, at = "points", dimyx = 512)
  Ksim <- Knetinhom(sim_vlc[[i]], lambda = as.numeric(lambda), r = r)
  return(Ksim)
})
```

```
K_nsim_df <- as.data.frame(do.call(cbind, d_nsim))
K_nsim_df_est <- K_nsim_df[, seq(3, 399, by = 2)]

maxn <- function(n) function(x) order(x, decreasing = TRUE)[n]
minn <- function(n) function(x) order(x, decreasing = FALSE)[n]

Kmin <- apply(K_nsim_df_est, 1, function(x) x[minn(5)(x)])
Kmax <- apply(K_nsim_df_est, 1, function(x) x[maxn(5)(x)])
```

```
plot(r, Kmin, type = "n", col = "grey", ylim = c(0, 270), xlab = "r", ylab =
← expression(italic(K[inhom])))
points(r, Kmax, type = "n", col = "grey")
polygon(c(r, rev(r)), c(Kmax, rev(Kmin)), col = "grey", border = "grey")
points(r, K_vlc$est, type = "l")
```

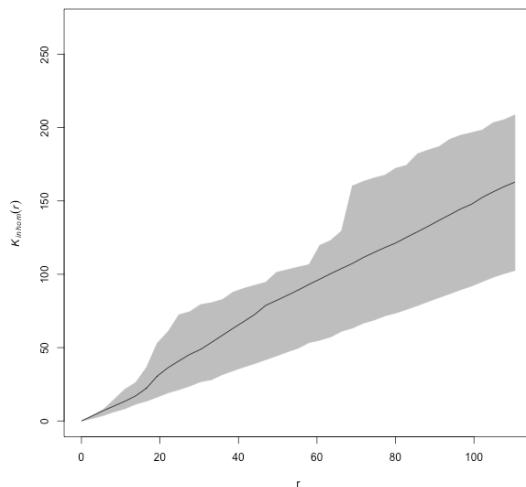


Figura 43.7: Función K para los delitos en Valencia, junto con la envoltura bajo un proceso de Poisson.

Resumen

La teoría de procesos puntuales espaciales constituye la base para el análisis de eventos observados geográficamente a través de sus coordenadas (longitud, latitud) en un espacio bi-dimensional. Esta es una de las ramas del campo de la estadística espacial en conjunción con la de procesos estocásticos. De hecho, un proceso puntual espacial es un proceso estocástico cuyas realizaciones consisten en un conjunto numerable de puntos (llamados en muchas ocasiones eventos). Heurísticamente, se trata de un conjunto de datos que se encuentran en una región concreta (área de estudio). Los puntos pueden representar cualquier población espacialmente explícita, como localizaciones de animales, nidos de aves, epicentros de terremotos, galaxias, crímenes, etc. El modelo estadístico más conocido para el análisis de patrones puntuales espaciales es el proceso puntual espacial de Poisson (asociado a la condición de aleatoriedad espacial completa). A partir del modelo de Poisson se construyen modelos más complejos. La modelización pasa por determinar las intensidades de primer y segundo orden que caracterizarán las propiedades básicas del comportamiento de los puntos. En este capítulo se proponen varios estimadores de la función de intensidad de primer orden junto con sus elementos asociados relacionados con funciones núcleo, parámetro de suavizado y correcciones de borde. Se consideran también algunos aspectos de medidas de segundo orden. El capítulo finaliza con aspectos sobre el cambio de soporte del plano euclídeo a redes lineales.

Parte X

Comunica y colabora

Capítulo 44

Informes reproducibles con R Markdown y Quarto

Emilio L. Cano

Universidad Rey Juan Carlos

44.1. ¿Por qué informes reproducibles?

El resultado final de un proyecto de análisis de datos terminará comunicándose a distintos niveles, tanto *aguas arriba* como *aguas abajo*. Esta comunicación es “la última milla” del flujo de análisis que se esquematizaba en la Fig. ???. Se llama genéricamente “informe” a cualquiera de estos resultados que se pueden producir en distintos formatos de destino. Estos informes estarán compuestos de múltiples elementos como texto, gráficos, resultados numéricos, tablas, etc. Además, es posible que haya que generarlos en distintos formatos (por ejemplo html o pdf, entre otros), para diferentes destinos, como la Web, documentos imprimibles o presentaciones. Finalmente, con una alta probabilidad varias personas intervendrán en el proceso, y la **trazabilidad** (reproducibilidad) del análisis mejorará el proyecto de forma global.

La forma de abordar el problema es típicamente con un enfoque *corta-pegá*, en el que primero se realiza todo el análisis de datos con el software estadístico y después se utilizan los resultados del análisis como base de un informe escrito, posiblemente con algunas iteraciones del proceso si el proyecto tiene cierta envergadura. El software estadístico comercial suele incluir formas de generar resultados listos para integrar en un informe, pero habitualmente bajo este paradigma de incluir el resultado a posteriori ([Leisch, 2002](#)).

Esta forma de trabajar genera un alto coste de mantenimiento (debido a la regeneración manual del informe) a la vez que provoca inconsistencias (por ejemplo entre unos grupos y otros, entre diferentes analistas, etc.), errores, contenidos desactualizados o no reproducibles. Este enfoque es propenso a fallos de organización y gestión de un proyecto, lo se traduce en vulnerabilidades

especialmente en la ejecución de software, simulaciones, etc. Además, cada vez que hay que hacer un cambio, hay que hacerlo en muchos sitios, con la consiguiente pérdida de tiempo y posibles errores.

El enfoque de la **investigación reproducible**¹ supera muchos de los obstáculos a la hora de preparar informes de análisis de datos. El objetivo es unir instrucciones para análisis de datos con datos experimentales de forma que los resultados se puedan volver a obtener automáticamente, entender mejor y verificar.

Un concepto muy relacionado que se utiliza en **R** es la **programación literaria**², mediante la cual se combina un lenguaje de programación como **R** con documentación de todo tipo (por ejemplo comentarios en el código fuente o inclusión de ficheros *readme*).

Con el enfoque de la investigación reproducible, lo que se hace es darle la vuelta al enfoque *corta-pega*, de forma que se escribe el informe a la vez que se realiza el análisis, incrustando el código dentro del propio informe. Obviamente, es necesario un sistema que consolide el informe con los resultados del código, y esto es lo que nos permite **R** y **RStudio** mediante archivos **R Markdown** y su evolución reciente a **Quarto**. El desarrollo de **Quarto** está patrocinado por la empresa Posit, PBC, donde anteriormente crearon **R Markdown** que compartía los mismos objetivos, pero estaba dirigido principalmente a usuarios del lenguaje **R**. El mismo equipo central trabaja tanto en **R Markdown** como en **Quarto**³.

Las ventajas de utilizar un enfoque reproducible se pueden resumir en:

- Si el mismo analista tiene que volver al análisis en el futuro, los resultados se pueden volver a obtener automáticamente de nuevo fácil y comprensiblemente.
- En el caso de que en el proyecto participen más analistas, toda la explicación está a mano.
- Cualquier cambio en un punto del análisis (por ejemplo, añadir una variable a un modelo) se puede realizar de una sola vez y los cambios en los resultados y gráficos se actualizarán automáticamente.
- Los resultados se pueden verificar por terceros en caso necesario. Un caso paradigmático fue el escándalo de los ensayos de cáncer en Duke en 2011⁴. No obstante es un tema que cada vez se demanda más en otros campos fuera de la investigación clínica (por ejemplo en publicaciones de cualquier tipo).

El flujo de trabajo sería el siguiente: los contenidos se encuentran en ficheros de texto plano, con código y texto explicativo. Estos ficheros fuente, se compilán y producen los materiales en los formatos necesarios. Los cambios se hacen una vez, y todos los materiales son actualizados adecuadamente.

¹El objetivo de la investigación reproducible es vincular instrucciones específicas a los análisis y datos experimentales, de modo que los informes puedan recrearse entendidos mejor y verificados. (Kuhn, 2019).

²La programación literaria es una metodología que combina un lenguaje de programación con un lenguaje de documentación (Knuth, 1984).

³<https://quarto.org/about.html>

⁴<http://www.nytimes.com/2011/07/08/health/research/08genes.html>

A continuación se abordará el enfoque reproducible. Para el otro enfoque, simplemente basta copiar los resultados de la consola y los gráficos de la pestaña *Plots* del panel inferior derecho en cualquier editor de documentos.

44.1.1. Markdown, R Markdown, Quarto y RStudio

Markdown es un lenguaje de marcas ligero que fue creado con la intención de ser más legible y fácil de escribir que el código HTML, aunque actualmente se utiliza para otros formatos de salida. Al ser ficheros de texto plano los ficheros se pueden leer bajo cualquier circunstancia, con una sintaxis muy sencilla que permite leerlo directamente por las personas, o ser convertido por un ordenador en otro formato más elaborado, como por ejemplo HTML (página web), pdf o Microsoft Word. En **RStudio**, se pueden crear ficheros **R Markdown**⁵ utilizando esta sintaxis para las explicaciones del análisis, e incluir dentro “trozos” (*chunks*) de código, de forma que, al generar el informe, el resultado de ese código queda incluido en el documento de salida. Así, si una vez terminado el informe se ha olvidado, por ejemplo, incluir un gráfico, sólo hay que añadir las líneas de código que lo crean y volver a generar el informe.

R Markdown ha evolucionado a un nuevo formato denominado **Quarto**⁶, que extiende aún más la funcionalidad de Markdown y está pensado para ser usado con otros lenguajes de programación. En esencia, y a los efectos de este capítulo, hay pocas diferencias.

Para poder utilizar las capacidades de **R Markdown** y **Quarto**, es necesario tener instalado el paquete **knitr** (Xie, 2017), que utiliza también otros paquetes como **rmarkdown**. Aunque **knitr** no forma parte del **tidyverse**, sí es un enfoque moderno de **R** que vino a hacer más fácil la generación de documentos que se hacía en **R** base con la función **Sweave** (Leisch, 2002). Para usar **Quarto** se necesita también el paquete **quarto** (Allaire, 2022) y tener instalado el software **quarto** en el ordenador⁷.

Para crear un nuevo documento **Quarto**, se selecciona *Quarto Document...* en el ícono de nuevo archivo de la barra de herramientas o en el menú *File*. Entonces se abre el cuadro de diálogo *New Quarto Document*. Hay varios tipos de archivos que se pueden crear, que producirán formatos diferentes: *Document* (documento), *Presentation* (presentación de diapositivas) e *Interactive* (aplicación web interactiva con Shiny u Observable JS). De momento se verán los documentos. Se puede crear un archivo **Quarto** vacío si no se quiere crear la estructura. Para que se cree con una estructura mínima, se necesita un título del documento y un autor, que después se podrán cambiar. También se selecciona un formato de salida por defecto, que puede ser HTML (para ver en el navegador), PDF, o Word. Esto también se podrá cambiar después, por lo que la forma más eficiente de trabajar es empezar con HTML, cuya previsualización es más rápida, y cuando esté el resultado final generar el archivo en el formato deseado.

Se puede seleccionar la *Engine* entre **knitr** (**R**) y **Jupyter** (**Python**), y también elegir si se quiere utilizar el editor visual (por defecto). Con el editor visual se pueden utilizar menús para editar el texto y dar el formato Markdown sin esfuerzo. Al hacer clic en el botón *Create* de la ventana *New Quarto Document*, se abre en el editor de **RStudio** un documento quarto (extensión .qmd)

⁵<https://rmarkdown.rstudio.com>

⁶<https://quarto.org>

⁷La instalación es trivial para cualquier sistema desde la web de quarto, <https://quarto.org>.

con una estructura básica a modo de plantilla. Los elementos principales de un archivo **Quarto** aparecen en esta plantilla:

- **Encabezado YAML:** constituyen la configuración del documento, y controlan sobre todo las opciones del formato de salida, es decir, cómo se verá el resultado final. Este encabezado se encuentra entre dos líneas con tres guiones (---), donde se expresan las opciones como **opcion: valor**, y estos valores además se pueden anidar. Dispone de ayuda contextual, de forma que pulsando la combinación de teclas CTRL+ESPACIO aparecen las opciones que se pueden configurar y los posibles valores. Esta parte del documento es constituye el **formato** del documento.
- **Texto formateado:** con una sintaxis muy sencilla, se puede dar formato al texto, como negritas, listas, etc. En el editor visual se puede hacer con los menús y botones de la barra de herramientas del editor.
- **Fragmentos de código (*chunks*):** al generar el documento, se ejecutará el código dentro de estos fragmentos, y en el documento resultante se mostrará el resultado. En cada fragmento de código aparecen dos botones que sirven para ejecutar todos los *chunks* anteriores y para ejecutar el *chunk* actual. Junto con el texto formateado constituyen el **contenido** del documento.

La barra de herramientas del editor ofrece algunas opciones:

- El botón *Render* convierte (“renderiza” en lenguaje informático) el documento **Quarto** produciendo el archivo de salida configurado. Se puede desplegar un menú para cambiar el formato de salida y otras opciones. Al crear el documento solo aparece el formato de salida elegido, pero se puede cambiar el encabezado para poder convertir el documento a distintos formatos. Por ejemplo si se cambia el encabezado que se ha creado por defecto por el siguiente, se puede generar el archivo de salida en html o word seleccionando en la lista desplegable junto al botón *Render*:

```
---
title: "Título del informe"
format:
  html: default
  docx: default
  editor: visual
---
```

- El botón de opciones permite cambiar la forma en que se mostrarán las salidas al ejecutar el código desde el editor.
- El botón *Insert a new code chunk* nos permite insertar un nuevo fragmento de código.

- Las flechas de navegación permiten moverse entre los *chunks* del documento. También se puede usar el selector de esquema en la parte inferior para ir a un fragmento de código o apartado concreto del documento.
- Desde el menú *Run* se puede ejecutar el código de los distintos *chunks*.
- El menú *Publish* nos permitiría publicar el documento en algún servicio como [RPubs](#)
- El botón *Outline* muestra un esquema para navegar por el documento, donde aparecerán los encabezados formateados con Markdown.
- Se puede cambiar entre el editor visual y el del código fuente con los botones *Source* y *Visual* en la parte superior del editor.

Para generar el documento, se guarda el archivo en cualquier carpeta de nuestro proyecto y se utiliza el ícono de conversión (“renderizado”). La Fig. 44.1 muestra en el panel izquierdo el archivo fuente en el editor visual, con alguna opción adicional añadida a la plantilla por defecto, y en el panel derecho el informe renderizado. Si en vez de pulsarlo directamente se selecciona el triángulo de la derecha, se puede seleccionar el formato de salida (html, pdf o Word) si se han incluido esos formatos en el encabezado YAML como se ha indicado. El formato PDF requiere tener instalada una distribución del sistema de edición libre LaTeX⁸. El archivo de destino, con extensión .html, .pdf o .docx según el caso, quedará guardado en la carpeta donde se encuentre el archivo quarto. Dependiendo de las opciones configuradas, el archivo se abrirá automáticamente en una ventana nueva de **RStudio** (por defecto), o en la pestaña *Viewer* del panel inferior derecho, o en el visor de pdf integrado en **RStudio** (pdf). Para poder abrir archivos .docx será necesario tener instalado **Microsoft Word** o algún otro programa que pueda abrirlo, como **LibreOffice**.

Se puede compilar el informe tantas veces como se quiera con el ícono *Render*. Para trabajo en curso, se recomienda ir previsualizando en formato HTML, y una vez sea definitivo generar el formato de destino final. Para la conversión de formatos, **RStudio** integra la aplicación **pandoc**.

Hay una guía rápida de Markdown (*Markdown Quick Reference*) disponible en el menú de ayuda de **RStudio**, así como enlaces a dos *Cheatsheets*: *R Markdown Cheatsheet* y *R Markdown Reference Guide*. Esta última es la más completa y donde se encuentran todas las opciones disponibles (que sirven para los documentos **Quarto** aunque en la propia web de **Quarto** hay una documentación más completa). En los siguientes apartados se revisan las opciones más habituales que cubren un amplio abanico de proyectos.

44.2. Documentos Quarto

En esta sección se detalla cómo añadir contenido y configuración a un documento Quarto con algunas de las opciones más interesantes para la Ciencia de Datos reproducible.

⁸Se puede instalar una distribución ligera de LaTeX con el paquete `tinytex` ejecutando `tinytex::install_tinytex()` habiendo instalado previamente dicho paquete.

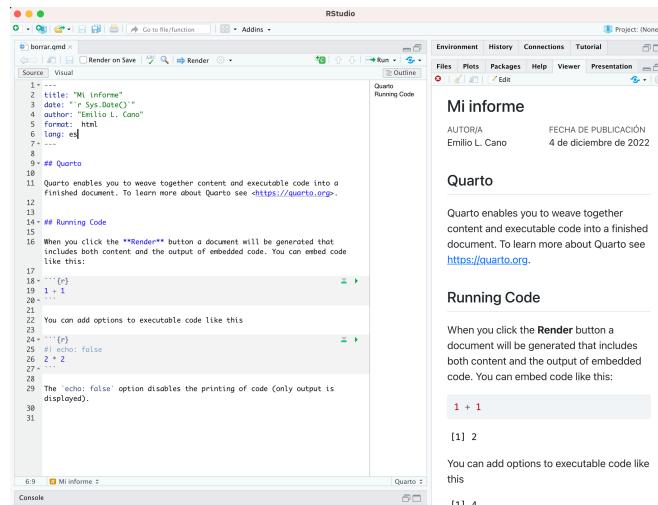


Figura 44.1: Informe Quarto y el documento de salida que produce su conversión (renderizado") con algunas opciones adicionales

44.2.1. Encabezado YAML y configuración

Las opciones de configuración del documento se establecen en este encabezamiento. Al crear el documento con la plantilla, se crea el siguiente encabezado:

```
---
title: "Título del informe"
format: html
editor: visual
---
```

Ya se ha visto cómo se pueden añadir más formatos, poniendo uno en cada línea e “indentando” con el tabulador las distintas opciones. Cada formato a su vez puede incluir opciones, que de nuevo se indican con nuevas líneas que se “indentan” debajo del formato.

La “indentación” se refiere al número de espacios en blanco (o tabulaciones) al principio de cada línea. Las opciones que estén “dentro” de otra, deben tener la misma “indentación” (el mismo número de espacios en blanco al principio de la línea). Véase un ejemplo más completo debajo. En el encabezado YAML la incorrecta “indentación” puede provocar errores al generar el informe.

Además del título (opción `title`), se pueden incluir autor (opción `author`) y fecha (opción `date`). Estos elementos son cadenas de texto que aparecerán al principio del documento de salida. Se debe cuidar que estén entre comillas para evitar posibles errores. El elemento `format` indica el formato de salida.

La cantidad de opciones que se pueden incluir en el encabezado YAML es enorme y no tiene cabida en este capítulo. Algunas de las más utilizadas son `lang` para indicar el idioma del documento (“es” para español), `bibliography` para indicar un fichero bibtex de bibliografía, o `toc` para incluir una tabla de contenidos. Algunas son específicas del formato. Por ejemplo, una muy útil es `reference-doc` para documentos de Word, con la que se puede indicar una plantilla personalizada para usar colores corporativos u otras opciones de diseño del informe⁹. Para documentos html se puede incluir una hoja de estilos con la opción `css`. La lista completa para cada uno de los formatos soportados por **Quarto** se puede consultar en la guía de referencia en <https://quarto.org/docs/reference/>.

El siguiente encabezado YAML fijaría el ancho y el alto de las figuras (en pulgadas) para el formato de salida html, además de las otras opciones comentadas:

```
---
title: "Título del informe"
format:
  html:
    fig-width: 8
    fig-height: 6
    css: estilos.css
  docx:
    toc: true
    reference-doc: plantilla.docx
    pdf: default
lang: "es"
bibliography: bibliografia.bib
---
```

Una explicación detallada del uso de hojas de estilo CSS queda fuera del alcance de este libro. Un ejemplo sencillo para formatear bloques con identificador (nombres precedidos por `#`) y bloques con clase (nombres precedidos por `.`) sería el siguiente:

```
#parrafoazul {
  color: blue;
}

.enfatizar {
  font-size: 1.2em;
}
```

44.2.2. Formateado de texto

Incluir títulos, énfasis en el texto y listas es muy sencillo, y a menudo no se necesita mucho más para realizar un informe. En el informe que se crea con la plantilla ya se ven algunas opciones:

⁹La plantilla se debe crear a partir de un archivo generado con **Quarto**, modificando los estilos del documento y añadiendo elementos como encabezados y pies de página.

- Los encabezados se crean poniendo al principio de la línea tantos símbolos almohadilla (##) como nivel de título se desee (dos almohadillas, apartado, tres almohadillas, subapartado, etc.) Una almohadilla sería para el título del informe, si no se especificara en el encabezado.
- Para poner texto en **negrita**, se incluye entre dos asteriscos a cada lado: ****negrita****.
- Para poner texto en formato **monoespaciado**, tipo código, se pone entre tildes graves (*backticks*, `): `monoespaciado`.
- Los enlaces se crean con: [texto del enlace] (<http://ejemplo.com>).

Existen otras opciones sencillas, que se pueden ver en la *Markdown Quick Reference* del menú *Help*:

- *Cursiva* rodeando el texto con un solo asterisco a cada lado (o guión bajo): cursiva.
- Listas poniendo al principio de la línea un asterisco, guión o signo más:
 - * Primer elemento de la lista
 - * Segundo elemento de la lista
 - + Primer elemento dentro del segundo
 - + Segundo elemento dentro del segundo
 - * ...
- Listas ordenadas poniendo un número y un punto al principio de la línea:
 1. Primer elemento
 2. segundo elemento
 3. ...
- Imágenes de cualquier tipo como: .
- Superíndices^{sup} y subíndices_{sub} con el texto entre símbolos ^ y ~ respectivamente.
- Ecuaciones en formato *LaTeX*, por ejemplo $\sum x_i$ sería \$\\sum x_i\$.
- Saltos de línea (añadiendo más de dos espacios al final de la línea) y saltos de página (tres o más asteriscos, *, o guiones medios, -, en una línea).
- Tablas, usando guiones medios y barras verticales para separar filas y columnas:

Primer encabezado	Segundo encabezado
-----	-----
Contenido de la celda	Contenido de la celda
Contenido de la celda	Contenido de la celda

Con estas opciones se cubren las necesidades de la práctica totalidad de proyectos de análisis. No obstante, dependiendo del formato de salida se pueden añadir otras opciones de formato.

44.2.3. Inclusión de código en el documento

Se pueden crear archivos **Quarto** sin incluir nada de código, simplemente para crear documentos editables fácilmente. Sin embargo, la verdadera potencia de **Quarto** es la posibilidad de incluir código de **R** (y también de otros lenguajes) en los documentos. Como ya se avazó, el código se incluye, principalmente, en forma de *chunks* o bloques de código.

Un *chunk* consta de unos marcadores de inicio y final del *chunk*, entre los cuales se insertan expresiones de **R** que se ejecutarán al generar el documento de salida. El marcador de inicio son tres símbolos de tilde grave seguidos de unas llaves con la letra **r** dentro. El marcador de cierre del *chunk* son de nuevo tres tildes graves, sin más. Y dentro del *chunk* se pueden poner expresiones de **R** de la misma forma en que se trabaja con los *scripts*. Al convertir (“renderizar”) el documento, el código se ejecutará con las opciones que se indiquen como se explica más adelante, y el archivo de salida incluirá el resultado de la ejecución del *chunk*. El siguiente sería un ejemplo de código para incluir gráficos en el informe.

```
```  
library(CDR)
library(corrplot)
mcor_tic <- cor(TIC2021)
corrplot.mixed(mcor_tic, order = 'AOE')
```
```

En todo caso, no hay que escribir los marcadores de inicio y final, ya que se dispone del atajo de teclado **CTRL+ALT+I** que lo hace automáticamente, o, alternativamente, el ícono *Insert a new code chunk* de la barra de herramientas del editor. Una vez se tiene el cursor dentro de un *chunk*, se puede ejecutar una expresión como del mismo modo que se hace en un *script* (**CTRL+ENTER**), o el *chunk* completo (**MAYUS+CTRL+ENTER**).

A veces es necesario incluir algún resultado de **R** en medio del texto y no como un bloque. En esos casos se puede insertar un bloque en línea poniendo, entre dos tildes graves, la letra **r** como primer carácter, y después una expresión de **R** que se pueda “imprimir” como cadena de texto:

```
`r expresion_de_R`
```

Una opción muy interesante de los informes de **Quarto** es la parametrización. Esta opción es muy útil para informes automatizados que pueden cambiar dependiendo de algún valor, por ejemplo del fichero de datos, la fecha, o cualquier otro valor. Estos parámetros se crean como elementos del encabezado YAML de la forma:

```
params:  
  parametro: valor
```

que después se pueden usar en los *chunks* de código como `params$parametro`. La verdadera potencia de esta característica es cuando se convierte el documento desde un *script* en el que

los parámetros son resultados de algún tipo de operación en los datos (por ejemplo, un informe de análisis de inventario solo de una tienda donde se han producido roturas de stock el día x). En vez de utilizar el botón *Render*, en estos casos se usa la función `quarto_render()`, una de cuyas opciones es `execute_params`, donde se pasarían los valores de los parámetros en forma de lista cuyos elementos tienen el nombre de los parámetros.

44.2.4. Opciones de los bloques de código (*chunks*)

Al renderizar un informe que contiene *chunks* sin configurar ninguna opción, el informe mostrará por defecto el código de entrada y las salidas (textos y gráficos), así como todos los mensajes que se pueden producir.

Opcionalmente, justo después del marcador de inicio del *chunk* se pueden añadir opciones del mismo mediante líneas que comienzan por el llamado *hashpipe*, que es una almohadilla seguida de la barra vertical, `#|` y a continuación la opción y su valor, de la misma forma que se hacía en el encabezado YAML para las opciones del documento, es decir, `opcion: valor`.

El *chunk* que se muestra a continuación tiene como identificador “ejemplo”, y como opciones `echo: false` y `fig.align: 'center'`, lo que indica que el código no se mostrará en el informe final y que el gráfico producido se alinearán en el centro del texto.

```
---
#| label: "Ejemplo"
#| echo: false
#| fig-align: 'center'
plot(cars)
---
```

Las opciones de *chunk* se pueden incluir de forma global en el documento estableciéndolas en el encabezado YAML del mismo, por ejemplo para mantener las opciones anteriores en todos los *chunks* por defecto:

```
---
title: "Mi documento"
format: html
knitr:
  opts_chunk:
    echo: false
    fig-align: 'center'
---
```

Es importante señalar que las opciones establecidas en los *chunks* tienen prioridad a las opciones establecidas en el documento.

Hay varias opciones de *chunk* que tienen que ver con la presentación en la salida. Por defecto, si se produce un error, el proceso se detiene y no se genera el archivo de destino. Este comportamiento, y otras muchas opciones, se pueden configurar como opciones del *chunk*. Las más

habituales son: `error: true` para mostrar los errores y no detener la generación del documento; `warning: false` y `message: false` para no mostrar `warnings` ni mensajes respectivamente; `include: false` para ejecutar el código pero no mostrar ningún tipo de salida; `eval: false` para no ejecutar el código; `results: "hide"` para indicar que no se muestren los resultados (otras opciones son `asis`, `hold` o `markup`); `comment: simbolo` para cambiar el símbolo que se usará como comentario del output (a veces es conveniente simplemente no poner comentario, es decir, `.`). Estas opciones del `chunk` se pueden incluir a nivel global en el encabezado YAML como se ha indicado anteriormente. La lista completa de opciones se encuentra en la *R Markdown reference guide* que está disponible en el menú *Help/Cheatsheets*, o a la [documentación de Quarto sobre opciones de ejecución](#)¹⁰.

Una característica muy cómoda es usar la ayuda contextual al escribir las opciones del `chunk`. Al comenzar a escribir, o pulsando la combinación de teclas **CTRL+ESPACIO**, se muestran las opciones disponibles, y al seleccionar una opción, si tiene varios posibles valores aparecen también para seleccionar.

44.2.5. Referencias cruzadas y formateo de tablas

La salida tabular por defecto de la consola normalmente no es adecuada para un informe. En su lugar, lo que se desea es tener una tabla formateada adecuadamente para el formato de salida. Se pueden incluir en los informes de **Quarto** tablas formateadas de calidad. Para ello, se debe utilizar alguna función que formatee la tabla de acuerdo al formato de salida (HTML, PDF, Word) y, a veces, configurar la opción `results` del `chunk` como '`asis`'. Muchas de estas funciones preparan automáticamente el formato según el fichero de salida que se está generando. Por ejemplo, el siguiente `chunk` generaría una tabla en cualquiera de los formatos de salida usando la función `kable` del paquete `knitr`:

```
```{r}
knitr::kable(TIC2021)
```
```

Hay otros paquetes con multitud de opciones de formato y presentación para las tablas como `xtable`, `flextable`, `kableExtra`, `DT`, o `gt`. Se anima al lector a consultar la documentación de estos paquetes para aprender a crear tablas de calidad que comuniquen adecuadamente los resultados de los análisis.

Tanto las tablas como las figuras se deben referenciar adecuadamente en los informes. Para ello se utilizan las **referencias cruzadas** de los informes **Quarto**. Para poder referenciar un gráfico creado en un `chunk`, es necesario: i) que el `chunk` tenga una etiqueta (`label: 'etiqueta'`); ii) que el `chunk` tenga una opción `fig-cap` para el título de la figura. Entonces el gráfico se puede referenciar en cualquier lugar del documento **Quarto** simplemente escribiendo `@etiqueta`. Por ejemplo, el siguiente `chunk` crearía el gráfico de la figura 44.2, y en el texto se referenciaría como “Figura `@fig-tic`”.

¹⁰<https://quarto.org/docs/computations/execution-options.html>

```
```{r}
#| label: "fig-tic"
#| fig-cap: "Ventas vs. % empresas con banda ancha"
library('ggplot2')
library('CDR')
TIC2021 |>
 ggplot(aes(ebroad, esales)) +
 geom_point() +
 geom_smooth(method = "lm")
```

```

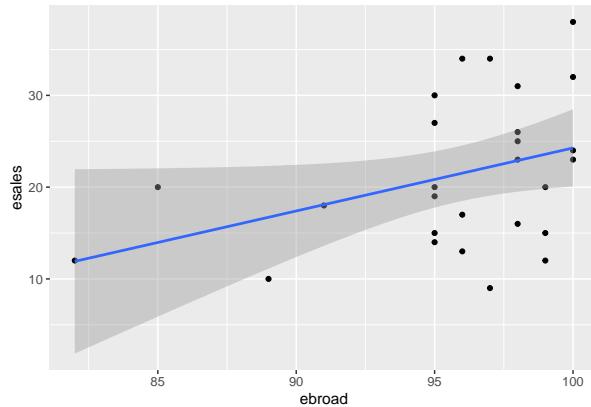


Figura 44.2: Ventas vs. % Empresas con banda ancha

En cuanto a las tablas, igualmente el *chunk* que las crea debe tener una etiqueta. El título de la tabla en este caso lo proporcionará la propia función que la crea. A modo de ejemplo, el siguiente *chunk* crearía la Tabla 44.1 ya formateada con la función `flextable()` del paquete homónimo ([Gohel and Skintzos, 2022](#)), y en el texto se referenciaría como “Tabla @tab-tic”.

```
```{r}
#| label: "tab-tic"
#| fig-cap: "Contaminación media NOx según tipo de estación"
library(dplyr)
library(flextable)
contam_mad |>
 filter(nom_abv == "NOx") |>
 group_by(tipo) |>
 summarise(Media = mean(daily_mean, na.rm = TRUE), n = n(),
 Desv.Tip = sd(daily_mean, na.rm = TRUE),
 Perdidos = sum(is.na(daily_mean))) |>
 flextable() |>
 set_caption("Contaminación media NOx según tipo de estación.") |>
```

```

```
autofit()
`--
```

Tabla 44.1: Contaminación media NOx según tipo de estación.

| tipo | Media | n | Desv.Tip | Perdidos |
|------------|----------|--------|----------|----------|
| Fondo | 64.11108 | 49,656 | 61.48603 | 70 |
| Suburbanas | 32.87574 | 12,414 | 32.13426 | 39 |
| Tráfico | 81.27392 | 33,104 | 68.13974 | 83 |

44.3. Otros formatos

El formato **Quarto** es solo uno de los que se pueden utilizar para aplicar la reproducibilidad que motiva este capítulo. Como se ha comentado, es la nueva generación de **R Markdown**, archivos que pueden seguir creándose con **RStudio**. En **R** base se pueden crear archivos **Sweave**, con sintaxis **LaTeX** para la narrativa y bloques de código **R**. También es posible crear archivos **R** **HTML**, con la narrativa en **HTML**. Los identificadores de los bloques son ligeramente distintos, así como las opciones disponibles, aunque la filosofía es la misma. **R Markdown** y ahora **Quarto** han ido desplazando estos otros formatos al ser más versátil (puede generar cualquiera de los otros formatos, y además otros como **.docx**).

En cuanto a los formatos de salida, hay una cantidad de opciones muy interesante que queda fuera de los objetivos de este libro. A continuación se relacionan algunos de ellos, si bien se puede consultar el libro de [Xie et al. \(2019\)](#) y la documentación de **Quarto** para ver detalles.

- **Notebooks:** es un tipo especial de salida **HTML**, más indicado cuando se quieren ir probando cosas guardando el resultado parcial de lo ejecutado en el **html**.
- **Presentaciones:** es posible crear presentaciones **PowerPoint** y usar una plantilla, como se vio con los documentos de **Word**. Se utiliza una sintaxis muy sencilla y se puede incluir código y resultados igual que en un informe. Otros formatos de presentaciones son **Reveal JS** y **beamer** (**LaTeX**) y con **Quarto**, además, **ioslides**, **slidify** y **xaringan** (**HTML**).
- **Tableros (dashboards):** pueden ser estáticos, usando el paquete **flexdashboard**, útiles para comunicar resultados en un par de pantallazos.
- **Shiny:** aplicaciones web interactivas que responden a inputs del usuario (*reactive*). Estas aplicaciones se tratan en detalle en el Cap. 45.
- **Websites:** websites sencillos con páginas enlazadas en el mismo directorio.
- **Blog:** directamente como proyecto **Quarto**, o con el paquete **blogdown**, se pueden generar sitios web completos al estilo de un blog con páginas.

- **Libros:** directamente como proyecto **Quarto**, o con el paquete **bookdown**, se pueden crear libros en varios formatos. El material de este libro está creado con **bookdown**.
- **Tutoriales:** aparecieron en **RStudio** 1.3; se pueden crear documentos interactivos con preguntas y navegación usando sintaxis **R Markdown**.
- **Tufte Handouts:** un tipo especial de documento con anotaciones al margen que comunica muy bien.

Resumen

- En la comunicación de resultados, es esencial seguir un enfoque reproducible.
- **R Markdown** y su evolución **Quarto** es el formato más versátil para crear informes reproducibles que permitan una trazabilidad de los análisis.
- **RStudio** permite trabajar eficientemente con **R Markdown** y **Quarto** a través de ayuda y opciones.
- El encabezado YAML del informe contiene la configuración global, que puede incluir parámetros para automatización.
- La narrativa del informe se escribe en Markdown, con una sintaxis extremadamente sencilla.
- El código se puede incluir en forma de bloques (*chunks*) o en línea.
- En las opciones del *chunk* se puede personalizar la forma en que se ejecutará y mostrará en el informe.
- Los informes **R Markdown** y **Quarto** se pueden generar en formatos HTML, PDF y Word, entre otros.
- En los informes **Quarto** se pueden hacer referencias cruzadas a tablas, figuras y otros elementos del documento.
- Hay otros formatos más elaborados que merece la pena explorar.

Capítulo 45

Creación de aplicaciones web interactivas con Shiny

Aurora González Vidal

45.1. Introducción

Shiny es un paquete de **R** que permite crear aplicaciones web interactivas que cuentan con todos los elementos de **R**. Shiny se ha convertido en un referente ya que, para aquellos que tienen conocimiento de **R**, es muy sencillo crear una aplicación en cuestión de horas (Winston et al., 2020). Para crear una aplicación mínima, no se necesitan conocimientos de HTML (*HyperText Markup Language*), CSS (*Cascading Style Sheets*) o JavaScript y sus dependencias. Además, no es necesario pensar en elementos técnicos para hacerla accesible en la web como, por ejemplo, el puerto, ya que Shiny se encarga de esos detalles si no se cambian las opciones por defecto. Ésas son algunas de las razones principales por las cuales Shiny se ha vuelto tan popular a lo largo de los años, ya que se pueden crear pruebas de concepto de un producto, mostrar algoritmos o presentar resultados de investigación con claridad a través de interfaces de usuario accesibles, reproducibles y amigables (Fay et al., 2021).

El primer paso para usar Shiny consiste en instalar el paquete que está disponible en CRAN:

```
install.packages("shiny")
```

Para asegurarse de que la versión instalada es igual o superior a la 1...5.0. hay que usar `packageVersion("shiny")`. A continuación, se puede cargar el paquete y ver algunos ejemplos que se incluyen directamente en el mismo utilizando distintas opciones para el argumento `example`.

```
library("shiny")
runExample(example = "01_hello")
# otras: 02_text, 03_reactivity, 04_mpg, 05_sliders, 06_tabsets, 07_widgets, 08_html,
→ 09_upload, 10_download, 11_timer.
```

45.2. Componentes mínimos de una aplicación Shiny y disposición básica

Las aplicaciones **Shiny** tienen dos componentes:

1. Una interfaz de usuario **ui**, que es un script y
2. Un **server** que es un script de servidor o secuencia de comandos de servidor.

Estos componentes pueden encontrarse en el mismo script o estar separadas en dos scripts con nombres fijos: **ui.R** y **server.R**. En este caso, se ha elegido la segunda opción para ilustrar los ejemplos con mayor claridad. Una aplicación **Shiny** es un directorio que contiene estos scripts y otros ficheros adicionales (conjuntos de datos, fichero donde se definen funciones no dinámicas, etc.).

El código mínimo para crear una aplicación con un título, panel lateral y panel principal es el que sigue:

■ **ui.R**

```
shinyUI(fluidPage(
  titlePanel("TÍTULO", # panel de encabezado TÍTULO
  sidebarPanel(), # panel lateral
  mainPanel() # panel principal
))
```

■ **server.R**

```
shinyServer(function(input, output) {})
```

La aplicación se puede lanzar de dos maneras diferentes. La primera mediante el comando **runApp()**, que tiene como argumento la ruta del directorio que almacena los ficheros que componen la aplicación.

```
library(shiny)
runApp("ruta al directorio")
```

La segunda es lanzar la aplicación directamente desde Rstudio mediante el botón RunApp que aparece en cualquiera de los dos scripts `ui.R`, `server.R` reemplazando al Run habitual.

En este capítulo, además de ver los distintos componentes de **Shiny**, se construye una aplicación para la visualización de algunos gráficos presentados en el Cap. 52. Los datos relacionados se encuentran en el paquete CDR del libro.

Además del `ui.R` y el `server.R`, puede ser muy útil tener un fichero donde recoger las funciones, paquetes y datos necesarios para el funcionamiento de la aplicación. Este fichero se puede cargar mediante la función `source` desde el `ui.R` o desde el `server.R`. Una recomendación es denominarlo `source.R` para mejor organización y legibilidad.

45.3. Diseño de una aplicación *Shiny*

Shiny incluye una serie de opciones para el diseño o la disposición de los distintos componentes de una aplicación. En esta sección se ven dos muy componentes sencillos:

- `sidebarLayout()`: para colocar un `sidebarPanel()`, es decir, un panel lateral de entradas junto a un `mainPanel()` de contenido de salida.
- `tabsetPanel()` y `navlistPanel()` para la segmentación de diseños.

Hasta ahora se ha utilizado el primero, sin introducirlo específicamente, para mostrar distintos ejemplos, por ser el más sencillo.

45.3.1. Diseño de las páginas: `fluidPage()`

Un diseño de página fluido `fluidPage()` consiste en filas que a su vez incluyen columnas. Las filas tienen como propósito asegurar que sus elementos aparezcan en la misma línea (si el navegador tiene el ancho adecuado). El objetivo de las columnas es definir cuánto espacio horizontal, dentro de una cuadrícula de 12 unidades de ancho, deben ocupar sus elementos. Las páginas fluidas escalan sus componentes en tiempo real para llenar todo el ancho disponible del navegador.

Una `fluidPage()` tiene 2 argumentos: `headerPanel()` con el título de la aplicación, y `sidebarLayout()`, que es un punto de partida útil para la mayoría de las aplicaciones. Éste a su vez tiene 2 argumentos más: `sidebarPanel()`, que es una barra lateral para las entradas y `mainPanel()`, un gran área principal para la salida.

```
shinyUI(fluidPage(
  headerPanel("Evolución del paro"), # panel de encabezado
  sidebarLayout(
    sidebarPanel( # panel lateral
      radioButtons(
        "vble", "Variable", # botones circulares: nombre y etiqueta
        c(
```

770

Capítulo 45. Creación de aplicaciones web interactivas con Shiny

```

    "sexo" = "sexo",
    "tramo_edad" = "tramo_edad",
    "tiempo_búsqueda_empleo_agregado" = "tiempo_búsqueda_empleo_agregado",
    "sector" = "sector",
    "tiempo_búsqueda_empleo" = "tiempo_búsqueda_empleo"
),
"sexo"
)
),
mainPanel(
  plotOutput("gra1")
)
)
))

```

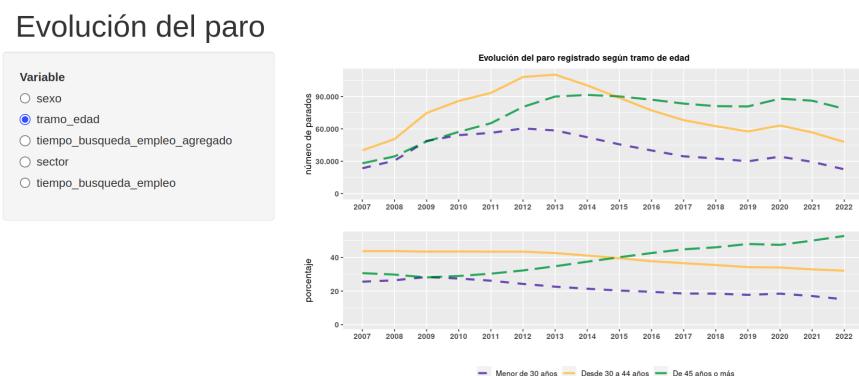


Figura 45.1: Aplicación shiny con sidebarPanel posicionado por defecto a la izquierda

La barra lateral puede posicionarse a la izquierda (por defecto) o a la derecha del área principal. Por ejemplo, para posicionar la barra lateral a la derecha se debe utilizar `position = 'right'` como se aprecia en la Fig. 45.2, donde el resto de argumentos son igual que para generar la Fig.45.1 .

```
shinyUI(fluidPage(  
  headerPanel("Evolución del paro"),  
  sidebarLayout(position = "right", ...)  
)
```

Las funciones `radioButtons()` y `plotOutput()` se introducirán en detalle en las respectivas secciones de este capítulo.

45.3. Diseño de una aplicación *Shiny*

771

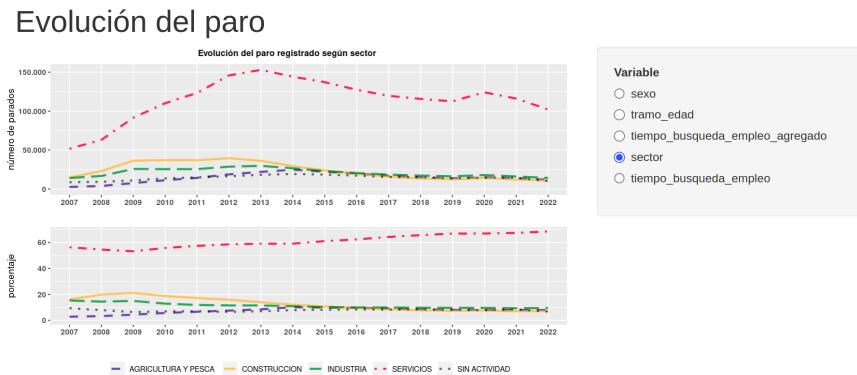


Figura 45.2: Aplicación shiny con sidebarPanel a la derecha

45.3.2. Segmentación de diseños: `tabsetPanel()` y `navlistPanel()`

Para subdividir el panel principal en varias secciones discretas, es decir, crear pestañas, se puede usar `tabsetPanel()` y `tabPanel()` como sigue:

```
mainPanel(
  tabsetPanel(
    tabPanel(
      "Elección con botones circulares",
      radioButtons(
        "vble", "Variable", # botones circulares: nombre y etiqueta
        c(
          "sexo" = "sexo",
          "tramo_edad" = "tramo_edad",
          "tiempo_búsqueda_empleo_agregado" = "tiempo_búsqueda_empleo_agregado",
          "sector" = "sector",
          "tiempo_búsqueda_empleo" = "tiempo_búsqueda_empleo"
        ), "sexo"
      ),
      plotOutput("gra1")
    ),
    tabPanel("Tramo edad fijo", plotOutput("gra2")),
    tabPanel("Tiempo búsqueda empleo fijo", plotOutput("gra3"))
  )
)
```

En el ejemplo de la Fig. 45.3, se aprecia que hay 3 ventanas y se muestra la tercera que es la evolución del paro considerando el tiempo de búsqueda de empleo y que no es una gráfica reactiva sino estática.

`navlistPanel()` es una alternativa a `tabsetPanel()` cuando existan muchas separaciones. Un `navlist` presenta los distintos componentes como una lista de la barra lateral en lugar de utilizar pestañas y no se hace en el `mainPanel`.

Evolución del paro

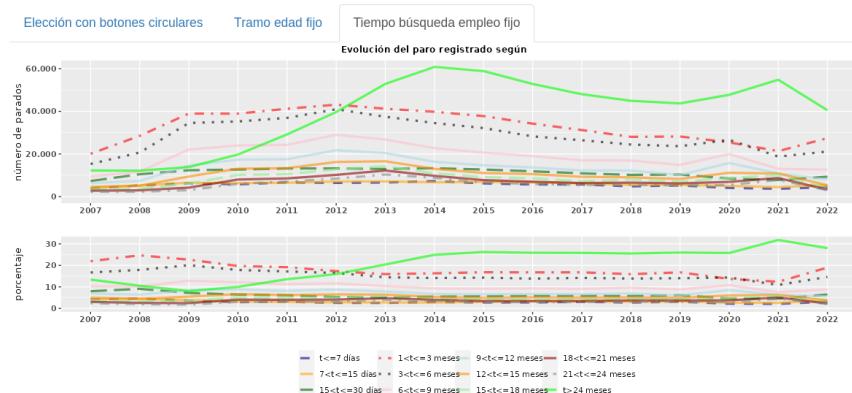


Figura 45.3: Aplicación shiny con varias ventanas

```
ui <- fluidPage(
  titlePanel("Application Title"),
  navlistPanel(
    "Header A",
    tabPanel("Component 1"),
    tabPanel("Component 2"),
    "Header B",
    tabPanel("Component 3")
  )
)
```

45.4. Elementos para la introducción de datos

Para que el usuario de la aplicación Shiny introduzca datos manualmente, hay diversos elementos que se enumeran a continuación:

- **Control deslizante:** un control deslizante permite que el usuario seleccione entre un intervalo de valores moviendo un control de posición por una pista. En Shiny se crea con la función `sliderInput()` que tiene, entre otros, los siguientes argumentos:
 - `inputId`: la entrada que se utiliza para acceder al valor
 - `label`: etiqueta o nombre que aparece en la interfaz
 - `min`: el mínimo del control deslizante
 - `max`: el máximo del control deslizante
 - `value`: el valor inicial
 - `step`: el intervalo entre cada valor seleccionable. NULL significa que va de uno en uno

45.4. Elementos para la introducción de datos

773

- **animate**: booleano que indica si los valores se cambian automáticamente para animar la aplicación

```
sliderInput(inputId, label, min, max, value, step = NULL, animate = FALSE)
```

Sus características incluyen:

- La posibilidad de introducir un único valor y rangos
- Formatos personalizados (por ejemplo, para entradas relativas al dinero)
- Pueden ser animados y recorrer los valores de forma automática (argumento **animate**)

Algunos ejemplos son:

```
sliderInput("enteros", "Enteros:", min = 0, max = 1000, value = 500)
sliderInput("decimales", "Decimales:", min = 0, max = 1, value = 0.5, step = 0.1)
sliderInput("rango", "Rango:", min = 1, max = 1000, value = c(200, 500))
sliderInput("animacion", "Animacion:", 10, 200, 10, step = 10, animate =
  animationOptions(loop = T))
```

- **Botón circular**: un botón circular es un tipo de selector que proporciona una lista de opciones entre las cuales solo se puede seleccionar una. En **Shiny** se crean con la función **radioButtons** que tiene, entre otros, los siguientes argumentos autoexplicativos:

```
radioButtons(inputId, label, choices, selected = NULL)
```

Un ejemplo donde la variable “sexo” está elegida por defecto se puede ver en el primer trozo de código de la subsección [45.3.1 sidebarLayout\(\)](#).

- **Selección múltiple**: un cuadro de selección múltiple es un tipo de selector que proporciona una lista de opciones entre las cuales se pueden seleccionar varias. En **Shiny** se crean con la función **selectInput** que tiene, entre otros, los siguientes argumentos autoexplicativos:

```
selectInput(inputId, label, choices, multiple = FALSE)
```

```
selectInput("año", "Año:",
  c(
    "año1" = "2007",
    "año2" = "2013",
    "año3" = "2019",
    "año4" = "2022"
  ),
  multiple = TRUE
)
```

■ checkboxGroupInput

Muy similar al anterior, este componente crea un grupo de casillas que se pueden utilizar para alternar varias opciones de forma independiente.

```
checkboxGroupInput(inputId, label, choices, multiple = FALSE)
```

```
checkboxGroupInput(  
  "variable", "Variables to show:",  
  c(  
    "año1" = "2007",  
    "año2" = "2013",  
    "año3" = "2019",  
    "año4" = "2022"  
  )  
)
```

■ Entrada numérica

```
numericInput("obs", "Número de observaciones:", 10)
```

■ Entrada de texto

```
helpText("aclaraciones")
```

Otras opciones de entrada que se invita al lector a analizar se relacionan con las fechas: `dateInput()`, `dateRangeInput()` y con un área de texto: `textAreaInput()`.

■ Lectura de ficheros de datos

También es posible introducir información a través de la lectura de ficheros de datos, con la función `fileInput()`. Se pueden combinar valores por defecto de la función utilizada para la lectura de datos con algunos de los elementos anteriores para definir las características del dataset (separador, decimal, cabecera). En el siguiente ejemplo se utiliza la función `read.csv()`, y se da a elegir si tiene cabecera o no con el `checkboxInput()`, así como el tipo de decimal con el `radioButtons`. Otra particularidad del ejemplo es que se asume que los datos están separados por punto y coma, tal y como se aprecia en el argumento `sep` del `read.csv()`.

```
shinyUI(fluidPage(  
  headerPanel("Lectura de datos"),  
  sidebarPanel(  
    h4("Cargar fichero CSV"),  
    fileInput("file1", "",
```

Elementos para la introducción de datos

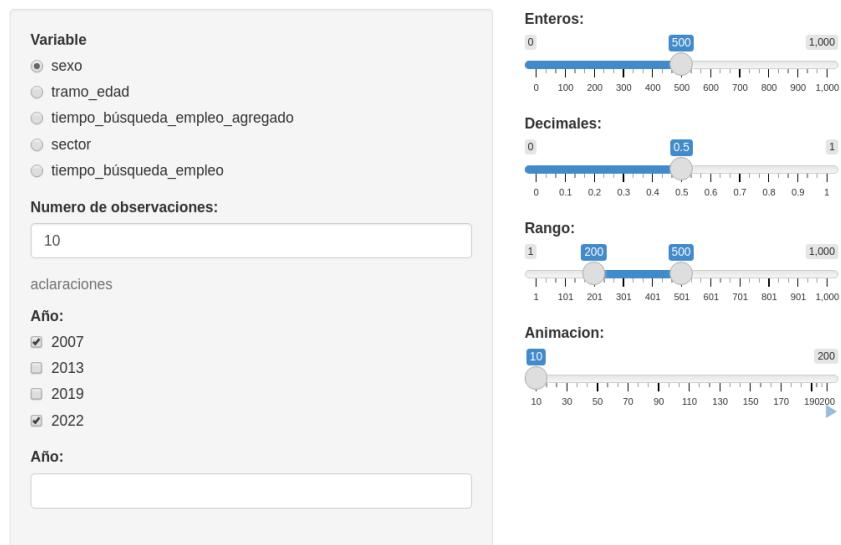


Figura 45.4: Distintos elementos para la introducción de datos

```

accept = c("text/csv", "text/comma-separated-values", "text/plain", ".csv")
),
checkboxInput("header", "Header (el csv tiene nombres de variables)", TRUE),
radioButtons(
  "dec", "Separador de decimales:",
  c(
    "Punto" = ",",
    "Coma" = ","
  )
),
mainPanel(
  tabsetPanel(
    tabPanel(
      "CSV",
      h4("Vista del fichero CSV"),
      tableOutput("contents")
    )
  )
))
)

shinyServer(function(input, output) {
  output$contents <- renderTable({

```

```

inFile <- input$file1

if (is.null(inFile)) {
  return(NULL)
}

read.csv(inFile$datapath, header = input$header, dec = input$dec, sep = ";")
})
}

```

Lectura de datos

The screenshot shows a Shiny application window. On the left, there is a sidebar titled "Cargar fichero CSV" with a "Browse..." button, a file input field containing "parados_subconjunto.csv", and an "Upload complete" button. Below these are checkboxes for "Header" (checked) and "Separador de decimales" (radio button selected for "Punto"). On the right, a "CSV" tab is selected, showing a table titled "Vista del fichero CSV" with the following data:

| año | sector | sexo | edad | tiempo_búsqueda_empleo | parados |
|------|---------------|--------|------|------------------------|---------|
| 2022 | SIN ACTIVIDAD | HOMBRE | 16 | <= 7 DIAS | 7.00 |
| 2022 | SIN ACTIVIDAD | MUJER | 16 | <= 7 DIAS | 8.25 |
| 2022 | SIN ACTIVIDAD | HOMBRE | 16 | > 7 <= 15 DIAS | 10.25 |
| 2022 | SIN ACTIVIDAD | MUJER | 16 | > 7 <= 15 DIAS | 6.25 |
| 2022 | SIN ACTIVIDAD | HOMBRE | 16 | > 15 <= 30 DIAS | 16.25 |
| 2022 | SIN ACTIVIDAD | MUJER | 16 | > 15 <= 30 DIAS | 8.50 |
| 2022 | SIN ACTIVIDAD | HOMBRE | 16 | > 1 <= 3 MESES | 38.50 |
| 2022 | SIN ACTIVIDAD | MUJER | 16 | > 1 <= 3 MESES | 28.00 |
| 2022 | SIN ACTIVIDAD | HOMBRE | 16 | > 3 <= 6 MESES | 24.00 |
| 2022 | SIN ACTIVIDAD | MUJER | 16 | > 3 <= 6 MESES | 16.75 |
| 2022 | SIN ACTIVIDAD | HOMBRE | 16 | > 6 <= 9 MESES | 7.75 |

Figura 45.5: Lectura de datos

45.5. Elementos para visualización (salida)

Tras la introducción de ciertos parámetros en el `ui.R`, éstos se pueden utilizar en el script `server.R` mediante la expresión `input`. El código de **R** que construye el objeto basado en esos datos se desarrolla en el servidor, y para generar dicho objeto se utilizan las funciones `renderX`, donde X es el tipo de objeto a devolver. Por último, este objeto se referencia nuevamente en el `ui.R` en el lugar que se desea mostrar (panel) a través de la expresión `X$output`.

El hecho de colocar una función en `ui` le dice a **Shiny** dónde mostrar su objeto. A continuación, hay que decirle a **Shiny** cómo construir el objeto. Esto se hace proporcionando el código **R** que construye el objeto en la función del servidor.

En concreto, algunas posibilidades se pueden ver en la Tabla ??.

- **Gráficos:** para generar la aplicación de la Fig. ??, se utiliza `renderPlot` en el `server.R` como sigue:

| Server | Ui | Crea |
|-------------|--------------------|----------------|
| renderImage | imageOutput | Imagen |
| renderPlot | plotOutput | Gráfico |
| renderTable | tableOutput | Tabla |
| renderText | textOutput | Texto |
| | htmlOutput | HTML |
| | verbatimTextOutput | Texto verbatim |

```
source("source.R")
shinyServer(function(input, output) {
  output$gra1 <- renderPlot({
    graf_evol(input$vble)
  })
})
```

Se ha llamado `gra1` a la variable que es el gráfico y que se crea con `renderPlot()`. En el interior se utiliza una función denominada `graf_evol()` que es compleja y se crea en `source.R` que se carga al principio. Lo que interesa de esta función es que tiene como único argumento el nombre de la variable de interés cuya evolución se desea mostrar. Ésta puede ser una de las diversas opciones que se dan a través del `radioButton` denominado `vble` que ha sido creado anteriormente y que, como vemos, se utiliza `input$vble` para invocar a la selección realizada en la interfaz de usuario.

- **Tablas:** para mostrar la tabla de la Fig. 45.5 se utiliza `renderTable()` en el server y `tableOutput()` en el ui.

```
shinyServer(function(input, output) {
  output$contents <- renderTable({
    inFile <- input$file1
    if (is.null(inFile)) {
      return(NULL)
    }
    read.csv(inFile$datapath, header = input$header, dec = input$dec, sep = ";")
  })
})
```

45.6. Reactividad

La programación reactiva es un paradigma de programación que se encarga de los flujos de datos y la propagación de los cambios. Ésto significa que cuando un flujo de datos es emitido

por un componente, el cambio se propagará a otros componentes.

El modelo de reactividad que utiliza Shiny es el siguiente: hay una fuente reactiva, un conductor reactivo y un punto final de la reactividad (Wickham, 2021). La fuente reactiva suele ser lo que el usuario introduce y el punto de parada lo que se muestra por pantalla. A lo que el usuario introduce se accede con el objeto input y a lo que se muestra por pantalla con el objeto output. Un ejemplo que ya se ha usado es el siguiente:

```
output$gra1 <- renderPlot({
  graf_evol(input$vble)
})
```

El objeto `output$gra1` es un punto final de la reactividad, y usa la fuente reactiva `input$vble`. Cuando `input$vbles` cambia, a `output$gra1` se le notifica que necesita ejecutarse de nuevo.

45.6.1. Conductores reactivos y control de la reactividad

También es posible crear componentes reactivos que conecten los inputs y los outputs. En el siguiente ejemplo se ha creado un objeto reactivo `datos` que genera datos que siguen una distribución que el usuario selecciona a través del radioButton `dist` y cuya muestra tiene tantos elementos como el usuario haya especificado en el numericInput `obs`.

Shiny, además, permite controlar la reactividad a través de los actionButtons. Se pueden modificar las entradas sin obtener una respuesta hasta que se apriete dicho botón.

Se ha creado un panel nuevo dentro del mainPanel y éste contiene, además del plot previo, un `actionButton` con la etiqueta `Presiona`.

```
shinyUI(fluidPage(
  headerPanel("Controlar reactividad"),
  sidebarPanel(
    radioButtons("dist", "Tipo de distribucion:",
      c(
        "Normal" = "norm",
        "Uniforme" = "unif",
        "Log-normal" = "lnorm",
        "Exponencial" = "exp"
      ),
      selected = "Exponencial"
    ),
    numericInput("obs", "Numero de observaciones:", 10),
  ),
  mainPanel(
    tabPanel(
      "Histograma distribucion RadioButton",
      "Plot", plotOutput("plot"), actionButton("botonReac", "Presiona")
    )
  )
))
```

```
)  
))
```

A continuación, se hace referencia a ese botón para cada una de las expresiones reactivas que se aislarán con la función `isolate()`:

```
shinyServer(function(input, output) {  
  datos <- reactive({  
    if (input$botonReac == 0) {  
      return(dist(rexp(input$obs)))  
    }  
    isolate({  
      dist <- switch(input$dist,  
        norm = rnorm,  
        unif = runif,  
        lnorm = rlnorm,  
        exp = rexp,  
        rnorm  
      )  
  
      dist(input$obs)  
    })  
  })  
  
  output$plot <- renderPlot({  
    if (input$botonReac == 0) {  
      return(NULL)  
    }  
    isolate({  
      hist(datos(),  
        main = paste("r", input$dist, "(", input$obs, ")", sep = "")  
    })  
  })  
})
```

45.7. Publicación de la aplicación en la web

Después del desarrollo de una aplicación Shiny, suele ser interesante publicarla para su explotación científica o empresarial. Rstudio ofrece diversas soluciones que se analizarán, con distintos niveles de complejidad y libertad, para poder publicar la aplicación web: (i) shinyapps.io, (ii) Shiny Server y (iii) RStudio Connect.

A continuación se introduce, muy brevemente, cada uno de ellos, y se proporcionan los enlaces para que el lector pueda indagar en profundidad.

- Shinyapps.io

Rstudio ofrece un servicio de hospedaje denominado Shinyapps.io que permite subir la aplicación directamente desde la sesión de **R** a un servidor que se mantiene por Rstudio. Hay un control casi completo sobre la aplicación, incluyendo la administración del servidor. Tiene distintos planes, desde gratuito hasta profesional, siendo el primero más restringido en cuanto a servicios (número de aplicaciones, horas activo...) y el último más completo (autenticación, personalización, etc).

Lo único que se necesita es:

- Un entorno de desarrollo de **R**, como RStudio IDE.
- La última versión del paquete **rsconnect**.

En la web shinyapps.io en el apartado “Dashboard” se realiza el registro. Shinyapps.io genera de forma automática un token que el paquete **rsconnect** utiliza para acceder a la cuenta.

```
rsconnect::setAccountInfo(name = "<ACCOUNT>", token = "<TOKEN>", secret = "<SECRET>")
```

Para desplegar la aplicación se utiliza **deployApp()** como sigue:

```
library(rsconnect)
deployApp()
```

Existen opciones gratuitas, que carecen de ciertas ventajas, como la posibilidad de restringir el acceso a la aplicación shiny, es decir, la aplicación no será privada con el plan gratuito aunque sí existen las opciones de autentificación con otros planes. Para más información sobre este método, consultese la página <https://shiny.rstudio.com/articles/shinyapps.html>.

- Shiny Server

Shiny server construye un servidor web diseñado para hospedar aplicaciones **Shiny**. Es gratuito, de código abierto y está disponible en GitHub.

Para usar el **Shiny** Server, es necesario tener un servidor Linux que tenga soporte explícito para Ubuntu 12.04 or superior (64 bit) y CentOS/RHEL 5 (64 bit). Aunque no se esté utilizando una distribución con soporte explícito, también se puede utilizar, si bien construyéndolo desde el paquete fuente.

En el mismo **Shiny** Server se pueden hospedar múltiples aplicaciones **Shiny**. Para ver instrucciones detalladas para su instalación y configuración, se recomienda la guía **Shiny** Server <https://docs.rstudio.com/shiny-server>.

La seguridad y privacidad quedarán supeditadas a los conocimientos del usuario, ya que dependerán de su propio servidor.

- RStudio Connect

Cuando **Shiny** se utiliza en entornos con fines lucrativos, existen herramientas de servidor que se pueden comprar y que vienen equipadas con los programas habituales de un servidor de pago:

- Soporte SSL
- Herramientas de administrador
- Soporte prioritario
- Privilegios de usuario
- Opción con Docker

Para tener dichas herramientas de servidor, la plataforma de publicación RStudio Connect puede ser una solución. Esta herramienta permite compartir aplicaciones **Shiny**, informes RMarkdown, cuadros de mando, gráficos, Jupyter Notebooks y más. Con RStudio Connect se puede programar la ejecución de informes y políticas de seguridad flexibles.

Además, RStudio Connect permite seleccionar privilegios de usuario en aplicaciones Shiny. La aplicación Shiny puede reconocer a un usuario basándose en la información de inicio de sesión y ofrecerle contenido personalizado, de manera que se puede controlar quién ve qué contenido y cuándo lo ve.

45.8. Extensiones de Shiny

Shiny es una herramienta totalmente expansible. Lo que se ha mostrado en este capítulo hasta ahora es un aperitivo en relación a todas las posibilidades que existen en el mundo de **Shiny**. Hay repositorios que recopilan información sobre paquetes que proveen de mejoras a las aplicaciones **Shiny** en su estilo y funcionalidad ([Xiao, 2018](#), [Gilmore et al. \(2017\)](#)). En esta sección se mencionan algunos de ellos pero, sobre todo, se recomienda al lector visitar dichos repositorios para una mayor profundidad en este tema.

- **shinydashboard**, **shinydashboardPlus** y **flexdashboards**

En temas de estilo, merece la pena destacar estos tres paquetes. Los dos primeros presentan una serie de plantillas predefinidas para la creación de las aplicaciones **Shiny**, de manera que los colores combinan y los elementos visuales tienen cierta armonía.

Por su parte, **flexdashboards** tiene como base un documento *R Markdown* y los distintos niveles del mismo definen los paneles de la aplicación a crear.

- **shinyWidgets**

Este paquete ofrece *widglets*¹ personalizados y diversos componentes para mejorar las aplicaciones. Se pueden reemplazar los **checkboxes** por **switch buttons**, añadir colores a los

¹Para el caso de estudio que se desarrolla se ha usado la librería **rtweet** para acceder a la **API Standard (V1.1)** (accesible a todo el mundo). Para obtener las credenciales que permiten trabajar en **modo usuario** se facilita el script de **python make_token_Twitter.ipynb** en github https://github.com/cogostro/token_API_V1.1**. Este script se puede ejecutar en el entorno Google Colab <https://colab.research.google.com/>.

`radioButtons` y al grupo de casillas de verificación (`checkboxGroupInput`), etc. Cada widget tiene un método de actualización para cambiar el valor de una entrada del server.

- `shinycssloaders`

Cuando una salida de `Shiny` (un gráfico, una tabla, etc.) se está calculando, permanece visible pero en gris. Si hay procesos algo más complejos, pueden tardar en mostrarse. Utilizando `shinycssloaders`, se puede añadir una rueda de carga (*spinner*) a las salidas en lugar de hacerlas grises. Envuelviendo una salida `Shiny` en `withSpinner()`, el *spinner* aparecerá automáticamente mientras la salida se recalcula.... Hay ocho tipos de animación incorporadas y personalizables en color y tamaño, pero también se pueden cargar otras animaciones.

- Visualizaciones interactivas

Paquetes como `heatmaply` o `leaflet` se pueden combinar perfectamente con `Shiny` para crear mapas de calor y mapas geográficos interactivos, y utilizarlos en las aplicaciones.

Resumen

`Shiny` es un paquete de `R` que permite crear aplicaciones web interactivas requiriendo únicamente conocimientos de `R`. En la primera parte de este capítulo se muestran los elementos básicos de una aplicación `Shiny`: user interface (`ui.R`) y servidor (`server.R`), así como los posibles diseños en relación a los componentes que una aplicación puede tener: barra lateral, paneles discretos, paneles de navegación, etc. A continuación, se repasan los elementos de introducción de datos en una aplicación `Shiny`, incluyendo la carga de conjuntos de datos y también los elementos de salida, como gráficas y tablas. También se aborda el modelo reactividad, es decir, cómo al cambiar algo en los parámetros de entrada de forma dinámica cambia la salida y cómo controlar éste proceso y se muestran distintas opciones para la publicación de las aplicaciones `Shiny`. Por último, se mencionan algunas de las posibles extensiones al paquete.

Capítulo 46

Git y GitHub R

Michal Kinel

46.1. ¿Qué es Git y GitHub?

Git es un sistema de control de versiones distribuido, diseñado para registrar y rastrear los cambios realizados en un archivo o conjunto de archivos a lo largo del tiempo ([Chacon, 2009](#)). Al utilizar Git, se pueden ver y restaurar versiones anteriores de un archivo, así como fusionar cambios realizados por diferentes personas en una única versión actualizada.

Por otro lado, GitHub es una plataforma de alojamiento de código online que utiliza Git como sistema de control de versiones subyacentes. Esta plataforma permite a los desarrolladores compartir y colaborar en proyectos de software, alojando el código fuente en la nube ([Astigarraga and Cruz-Alonso, 2022](#)). Además de alojar repositorios de Git, GitHub ofrece herramientas adicionales como seguimiento de problemas, solicitudes de extracción, y wiki de proyectos, lo que la hace una herramienta popular para el desarrollo de software colaborativo y de código abierto.

El uso de Git y GitHub se ha extendido a otros campos más allá del desarrollo de software, como la ciencia de datos, la documentación técnica y la colaboración en general. Su popularidad se debe en gran parte a la facilidad de uso, la flexibilidad y la capacidad de trabajar en proyectos de software colaborativos de manera eficiente y segura tanto.

46.2. ¿Por qué usar Git y GitHub?

Git y GitHub son herramientas para el desarrollo de software moderno, y su uso se ha extendido a otros campos como la ciencia de datos, la documentación técnica y la colaboración entre los desarrolladores. A continuación, se presentan tres ventajas clave de uso de Git y GitHub:

1. **Control de versiones y colaboración eficiente** Git es un sistema de control de versiones distribuido que permite registrar y rastrear los cambios realizados en un archivo o conjunto de archivos a lo largo del tiempo. Esto es especialmente útil cuando se trabaja en proyectos de software colaborativos, donde múltiples personas pueden estar editando el mismo archivo al mismo tiempo. Con Git, se pueden ver y restaurar versiones anteriores de un archivo, y también fusionar cambios realizados por diferentes personas en una única versión actualizada. Además, GitHub ofrece herramientas adicionales como seguimiento de problemas, solicitudes de extracción, y wiki de proyectos.
2. **Mejora la eficiencia y la seguridad en el desarrollo de software** El uso de Git y GitHub permite a los desarrolladores trabajar de manera más eficiente y segura en proyectos de software. Al utilizar un sistema de control de versiones como Git, los desarrolladores pueden colaborar de manera más efectiva y reducir el riesgo de conflictos o errores en el código. Además, GitHub ofrece características como la integración continua y la entrega continua (CI/CD), que automatizan y agilizan el proceso de desarrollo de software.
3. **Fomenta la transparencia y la comunidad** GitHub es una plataforma de alojamiento de código abierto, lo que significa que los proyectos alojados en ella son de acceso público y pueden ser revisados y mejorados por otros desarrolladores. Esto fomenta la transparencia de código abierto como privado dentro de una empresa. Además, GitHub cuenta con una gran comunidad de desarrolladores que pueden ofrecer soporte y retroalimentación a otros miembros de la comunidad.

46.3. Instalación y/o actualización de R y RStudio

R es un lenguaje de programación utilizado en la estadística y la ciencia de datos para realizar análisis, modelado y visualización de datos. **RStudio**, por otro lado, es un entorno de desarrollo integrado (IDE) que proporciona una interfaz gráfica de usuario para trabajar con R. Instalar o actualizar **R** y RStudio es un proceso relativamente sencillo, y se pueden seguir los siguientes pasos:

1. Descargar e instalar **R** Lo primero que se debe hacer es descargar **R** desde la [página oficial de R](#). Dependiendo del sistema operativo, se debe elegir la versión correcta de **R** para descargar. Una vez que se haya descargado el archivo, se debe seguir el asistente de instalación para instalar **R** en el equipo.
2. Descargar e instalar RStudio: una vez instalado **R**, se puede proceder a instalar RStudio desde su [página oficial](#). Al igual que con **R**, se debe elegir la versión adecuada para el sistema operativo y seguir el asistente de instalación para instalar RStudio en el equipo.
3. Actualización de **R** y RStudio Para actualizar **R**, se debe abrir **R** y ejecutar el siguiente comando en la consola:

```
install.packages("installr")
library(installr)
updateR()
```

Esto instalará el paquete `installr` y actualizará automáticamente **R** a la última versión disponible. Para actualizar RStudio, se debe abrir RStudio y verificar si hay una actualización disponible en el menú “Help” -> “Check for Updates”. Si hay una actualización disponible, se debe seguir el asistente de actualización para instalar la última versión de RStudio.

En resumen, la instalación o actualización de **R** y RStudio es un proceso sencillo que se puede realizar siguiendo los pasos mencionados anteriormente. Mantener estas herramientas actualizadas es importante para asegurarse de tener acceso a las últimas características y correcciones de errores.

46.4. Configuración de Git y GitHub

Configurar Git y GitHub es un paso importante antes de comenzar a trabajar en proyectos de software colaborativos. Se pueden seguir los siguientes pasos para configurar Git y GitHub:

1. Instalar Git . En primer lugar, es necesario instalar Git en el equipo. Git puede descargarse desde la página oficial de [Git](#). Una vez que se haya descargado el archivo, se debe seguir el asistente de instalación para instalar Git en el equipo. Además, en la página oficial se encuentra un manual completo sobre el uso de Git.
2. Configurar Git. Una vez que se ha instalado Git, se debe configurar el nombre de usuario y la dirección de correo electrónico para que los cambios que se realicen en los repositorios estén correctamente etiquetados. Para hacer esto, se debe abrir la línea de comandos, Git Bash o la terminal de RStudio, y ejecutar los siguientes comandos:

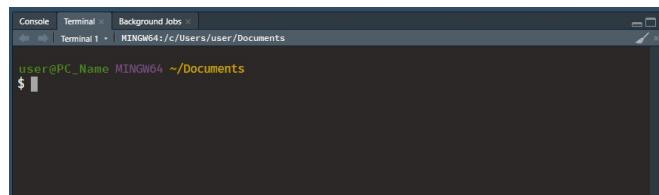


Figura 46.1: Terminal de RStudio

```
$ git config --global user.name "Su Nombre"
$ git config --global user.email "su.correo@ejemplo.com"
```

Esto configurará el nombre de usuario y la dirección de correo electrónico de forma global en Git.

3. Crear una cuenta en GitHub. Para utilizar GitHub, es necesario crear una cuenta en la página oficial de GitHub (<https://github.com/join>). Una vez que se haya creado la cuenta, se debe iniciar sesión en GitHub.
4. Configurar la clave SSH (protocolo Secure Shell), una credencial de acceso para el protocolo de red que permite el acceso remoto a través de una conexión cifrada. Para autenticar las conexiones con GitHub de manera segura (véase capítulo 10 de (Jenny Bryan, 2021)), se recomienda configurar una clave SSH en el equipo y agregarla a la cuenta de GitHub. Para ello, se debe abrir la línea de comandos, Git Bash o la terminal de RStudio, y ejecutar los siguientes comandos:

```
$ ssh-keygen -t rsa -b 4096 -C "su.correo@ejemplo.com"
```

Esto generará una clave SSH. A continuación, se debe agregar la clave SSH al agente de SSH:

```
$ eval "$(ssh-agent -s)"
$ ssh-add ~/.ssh/id_rsa
```

Finalmente, se debe copiar la clave SSH al portapapeles:

```
$ clip < ~/.ssh/id_rsa.pub
```

y agregarla a la cuenta de GitHub siguiendo las instrucciones en la página de configuración de la cuenta de GitHub:

- En la esquina superior derecha de la página del inicio, haga clic en la foto del perfil y, luego, en “Settings” (Configuración).
- En la sección “Access” de la barra lateral, haga clic en “SSH and GPG keys”.
- Haga clic en “New SSH key” para agregar la clave SSH.

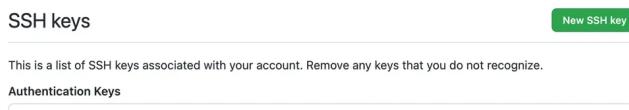


Figura 46.2: Llaves SSH en GitHub

- En el campo “Title” (Título), agregue una etiqueta descriptiva para la clave nueva. Por ejemplo, si está utilizando un portátil personal, puede llamar a esta clave “Portátil personal”.
- Seleccione el tipo de clave, ya sea de autentificación o de firma. Para obtener más información sobre la firma de una confirmación, consulte [aquí](#).

46.4. Configuración de Git y GitHub

787



Figura 46.3: Añadir llave SSH en GitHub

- Pegue su clave pública en el campo “Key”.
- Haga clic en “Add SSH key” para agregar la clave SSH.
- Si se le solicita, confirme su contraseña en GitHub. Para obtener más información, véase [Modo sudo](#).

Además de utilizar una clave SSH para autenticar las conexiones con GitHub, también se puede utilizar la autenticación basada en token de acceso personal, PAT , de GitHub. Esta forma de autenticación es recomendada por GitHub como una forma segura de autenticar conexiones, especialmente cuando se trabaja con aplicaciones y herramientas que requieren acceso a repositorios de GitHub. Para más información sobre el token de acceso personal, PAT, consulte el [capítulo 9](#) de ([Jenny Bryan, 2021](#)).

A continuación, se describen los pasos para utilizar la autenticación basada en token de acceso personal de GitHub:

1. Generar el token PAT: existen dos librerías `usethis` y `gitcreds` que facilitan la generación del PAT y almacenarlo, para ello, se introduce en la consola de RStudio:

```
library(usethis)
usethis::create_github_token()
```

2. Seguir las instrucciones en GitHub: a continuación se abrirá el sitio web de GitHub, se ingresa mediante el usuario y contraseña, con el cuadro para generación del PAT, *New personal access token (classic)*. En *Note* se introduce una nota identificativa al igual que en el procedimiento anterior y se selecciona el tiempo de validez del PAT en la pestaña *Expiration*, dejando las demás opciones por defecto. Se hace clic en *Generate token* para crear el token. En la nueva ventana se copia el token para posteriormente introducir en la consola:

```
library(gitcreds)
gitcreds::gitcreds_set()
```

En *password* se pega el token copiado anteriormente.

3. Para verificar que el nuevo PAT está configurado se introduce en la consola:

```
gitcreds::gitcreds_get(use_cache = FALSE)
```

Si la autentificación fue correcta se generará una salida similar a la siguiente:

```
<gitcreds>
  protocol: https
  host     : github.com
  username: mi_usuario
  password: <-- hidden -->
```

Una vez conectado RStudio y GitHub mediante SSH o PAT se puede proceder con la creación del repositorio en Git.

46.5. Conectar Git y GitHub con Rstudio

46.5.1. Rstudio primero

Este apartado se centra en el enfoque de creación de un nuevo proyecto en un ordenador local para posteriormente subirlo a GitHub, en remoto.

Una vez instalados y configurado Git en nuestro sistema y con la cuenta de GitHub hay que seguir los siguientes pasos para conectar Git y GitHub con RStudio:

1. Configurar Git en RStudio: Una vez que Git está instalado en el sistema, se debe configurar Git en RStudio. Para ello, se debe ir a la pestaña “Tools” en la barra de menú principal, seleccionar “Global Options” y luego seleccionar “Git/SVN”. Desde allí, se debe configurar la ubicación del ejecutable de Git en el sistema.
2. Verificar la versión de Git, introduciendo en la Terminal:

```
$ git --version
```

Si la salida es la versión de Git entonces la instalación fue ejecutada correctamente.

46.5. Conectar Git y GitHub con Rstudio

789

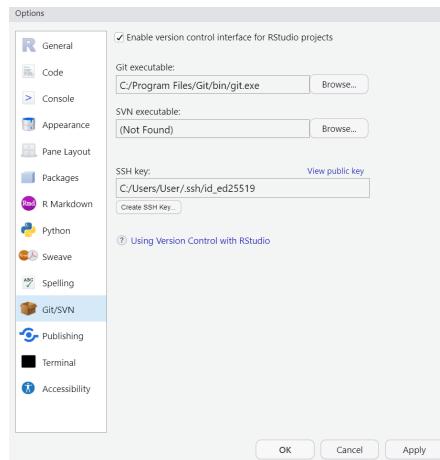


Figura 46.4: Tools de Rstudio

3. Crea un proyecto nuevo desde “File” -> “New project”

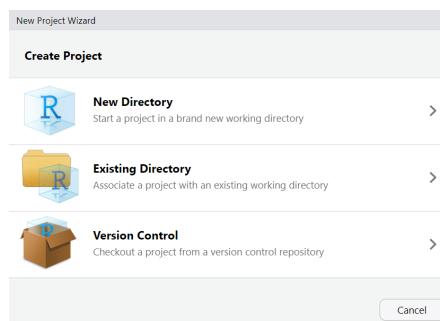


Figura 46.5: Nuevo proyecto de Rstudio

4. En el siguiente cuadro se procede dando clic en “New directory” y en la siguiente ventana se rellenan los datos como el nombre del proyecto y se marca la opción “Create a git repository” para crear un nuevo proyecto con repositorio de Git.

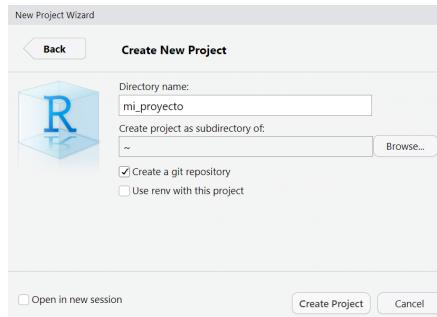


Figura 46.6: Nuevo proyecto en un directorio nuevo

- En el ícono de Git en la parte superior se accede a la ventana de revisión de cambios, se añaden los ficheros pinchando en los ticks, se añade el mensaje de confirmación y se hace clic en “commit”.

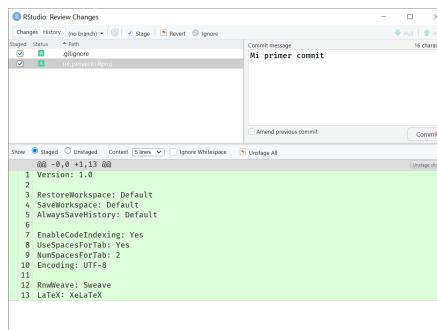


Figura 46.7: Revisión de cambios

- Alternativamente se puede utilizar la pestaña de Git, marcando los ficheros modificados o creados y confirmando mediante clic en “commit” tras el cual se abrirá el cuadro de diálogo anterior.

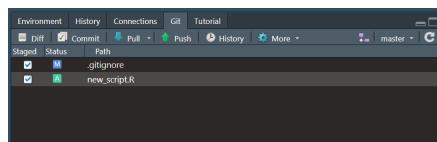


Figura 46.8: Revisión de cambios

- Para subir los cambios realizados en el proyecto recién configurado en RStudio, habiendo configurado Git y GitHub en los pasos anteriores, se ejecuta en la consola los siguientes pasos

```
library("usethis")
usethis::use_github()
```

La función `usethis::use_github()` en sus valores por defecto creará un repositorio público con el nombre de proyecto en la cuenta asociada. Para ver más opciones acuda a la ayuda de la función, ejecutando en la consola `?usethis::use_github`.

46.5.2. GitHub Primero

Este apartado se centra en el enfoque de creación de un nuevo proyecto en un ordenador local a partir de un repositorio disponible en GitHub, en remoto. Antes de todo hay que verificar si se tiene instalado Git, basta introduciendo en la terminal de RStudio al igual que en el apartado anterior.

```
$ git --version
```

Cuando la salida de la terminal arroje la versión de Git entonces la instalación fue correcta. En el caso de que la salida no arroje la versión vuelva la Sec. 46.4 o consulte el manual de la página oficial de Git en: <https://git-scm.com>.

A continuación, se describe paso a paso sobre cómo conectar GitHub y RStudio a partir de un proyecto ya existente en GitHub y con Git configurado previamente:

1. Abra RStudio y seleccione la opción “New Project” en la pestaña “File” del menú principal. Posteriormente haga clic en la opción “Version Control”.

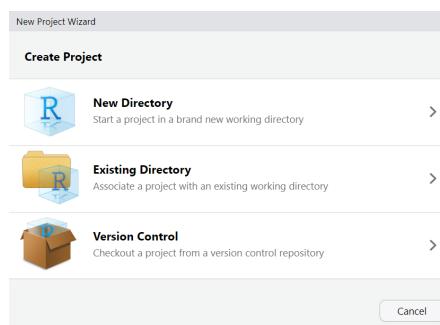


Figura 46.9: Nuevo proyecto de Rstudio

2. En la ventana emergente que aparece, elija “Git”.

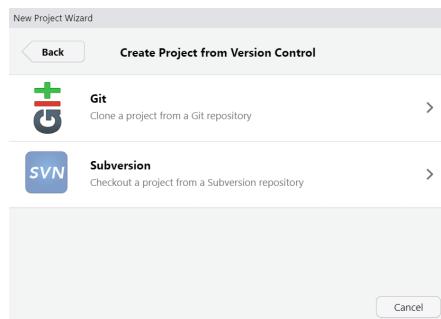


Figura 46.10: Crear proyecto desde control de versiones

3. En la siguiente ventana, pegue la URL del repositorio que deseé clonar y presione “Create Project”. RStudio le preguntará en qué carpeta desea guardar el proyecto, una vez que hayas elegido una ubicación, el proyecto se clonará en tu computadora.

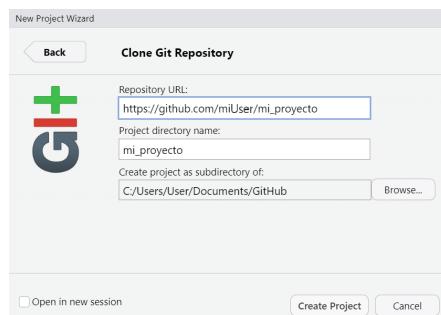


Figura 46.11: Nuevo proyecto desde un repositorio de Git

Los cambios realizados en el repositorio se realizan de la misma forma que en el apartado anterior.

46.6. Flujo de trabajo general de Git y GitHub en RStudio

A continuación se describe un flujo básico de trabajo, comenzando desde RStudio:

1. **Iniciar un repositorio local:** Lo primero que hay que hacer es inicializar un repositorio local en RStudio. Para ello, abra RStudio y seleccione la opción “New Project” en la

46.6. Flujo de trabajo general de Git y GitHub en RStudio

793

pestaña “File” del menú principal. Luego, seleccione la opción “New Directory” y elija una ubicación para su proyecto. A continuación, seleccione “Version Control” y luego “Git”. RStudio le preguntará si desea inicializar un repositorio en este directorio; haga clic en “Yes”. Tal y como se ha descrito en el punto 1 de la Sec. 46.5.1.

2. **Añadir archivos al repositorio:** Ahora, debe añadir los archivos de su proyecto al repositorio. Para ello, haga clic en la pestaña “Git” en la parte superior derecha de RStudio, y luego seleccione los archivos que desea agregar al repositorio. Haga clic en el botón “Stage”, y los archivos seleccionados pasarán a la sección “Staged” en la parte inferior de la pestaña “Git”. Si desea agregar todos los archivos del proyecto al repositorio, haga clic en el botón “Stage All”.
3. **Hacer un “commit” de los cambios:** Una vez que los archivos están en la sección “Staged”, debe hacer un “commit” para registrar los cambios. Para hacerlo, escriba un mensaje breve que describa los cambios que ha realizado en la sección “Commit message”. Luego, haga clic en el botón “Commit”. Los cambios se registrarán en el repositorio local.
4. **Crear una rama (opcional):** Si desea trabajar en una nueva función o corregir un error sin afectar la rama principal (master o main), debe crear una nueva rama. Para ello, haga clic en el botón “New Branch” en la pestaña “Git”. Escriba un nombre para la nueva rama y haga clic en “Create”. Ahora ya es posible hacer cambios en los archivos en la nueva rama sin afectar la rama principal.
5. **Subir los cambios al repositorio remoto:** Una vez que ha hecho un “commit” o confirmación de sus cambios, es hora de subirlos al repositorio remoto en GitHub. Para ello, haga clic en el botón “Push” en la pestaña “Git”. Los cambios se subirán al repositorio remoto en GitHub, que fue configurado en la Sec. 46.4.
6. **Solicitar un pull request (opcional):** Si trabaja en un proyecto colaborativo con otros usuarios, debe solicitar un “pull request” antes de fusionar los cambios en la rama principal. Para hacerlo, haga clic en la pestaña “Pull Requests” en la interfaz de GitHub. Luego, haga clic en el botón “New Pull Request” y siga las instrucciones para crear la solicitud.
7. **Fusionar los cambios en la rama principal (opcional):** Si trabaja en una nueva rama y desea fusionar los cambios en la rama principal, debe crear una solicitud de “pull request”. Si la solicitud es aceptada por el propietario del repositorio, los cambios se fusionarán en la rama principal.

Repository local y remoto:

- **Repository local** en Git es una copia completa de un proyecto que se encuentra en el equipo del usuario. Con un repositorio local, los usuarios pueden trabajar en un proyecto sin conexión a Internet y luego enviar los cambios al repositorio remoto cuando estén conectados.
- **Repository remoto** en GitHub es una versión en línea del proyecto que está almacenada en los servidores de GitHub. Los usuarios pueden clonar un repositorio remoto a su equipo para tener una copia local del proyecto y trabajar en ella. Los cambios realizados en la copia local pueden ser enviados al repositorio remoto para compartirlos con otros usuarios.

En resumen, el flujo de trabajo general de Git y GitHub en RStudio implica inicializar un repositorio local, añadir archivos al repositorio, hacer un “commit” de los cambios, crear una nueva rama si es necesario, subir los cambios al repositorio remoto en GitHub.

Todas las operaciones se pueden realizar desde la terminal de RStudio. Aquí hay algunos de los comandos más comunes que se utilizan en Git:

- **git init**: Este comando se utiliza para crear un nuevo repositorio de Git. Se ejecuta en el directorio raíz del proyecto y establece la estructura necesaria para que Git rastree los cambios en el código fuente.
- **git clone**: Este comando se utiliza para clonar un repositorio existente de Git. Es útil cuando se desea trabajar en un proyecto que ya está en GitHub o en otro servicio de alojamiento de repositorios de Git.
- **git add**: Este comando se utiliza para agregar archivos nuevos o modificados al área de preparación “Stage” de Git. La preparación es el primer paso para confirmar los cambios en Git.
- **git commit**: Este comando se utiliza para confirmar los cambios realizados en el repositorio de Git. Los cambios confirmados se guardan en la base de datos de Git y se etiquetan con un mensaje que describe los cambios.
- **git push**: Este comando se utiliza para enviar los cambios confirmados a un repositorio remoto, como GitHub. Esto actualiza el repositorio remoto con los cambios realizados en el repositorio local.
- **git pull**: Este comando se utiliza para actualizar el repositorio local con los cambios realizados en el repositorio remoto. Es útil cuando se está trabajando en un proyecto colaborativo y otros colaboradores han realizado cambios en el repositorio remoto.
- **git branch**: Este comando se utiliza para crear, listar y eliminar ramas en el repositorio de Git. Las ramas son una forma de trabajar en diferentes versiones del proyecto sin afectar la rama principal.

46.6. Flujo de trabajo general de Git y GitHub en RStudio

795

- **git merge:** Este comando se utiliza para fusionar ramas diferentes del repositorio de Git. Esto se utiliza comúnmente cuando se trabaja en diferentes ramas y se desea integrar los cambios realizados en una rama en la rama principal.
- **git status:** Este comando se utiliza para verificar el estado del repositorio de Git. Proporciona información sobre los archivos que se han modificado y los archivos que se han agregado al área de preparación.
- **git log:** Este comando se utiliza para ver un registro detallado de los cambios confirmados en el repositorio de Git. Muestra información como el autor del cambio, la fecha y la descripción del cambio.

Para conocer más a fondo la mecánica de Git es muy recomendable el manual ([Chacon, 2009](#)) o la hoja resumen proporcionada por GitHub, disponible en https://training.github.com/downloads/es_ES/github-git-cheat-sheet.pdf.

Resumen

- Git es un sistema de control de versiones distribuido utilizado para rastrear cambios en archivos a lo largo del tiempo, mientras que GitHub es una plataforma de alojamiento de código que utiliza Git como su sistema de control de versiones subyacente.
- La instalación y configuración de Git y GitHub es sencilla y permite una colaboración eficiente y un control de versiones en el desarrollo de software.
- Conectar GitHub y RStudio implica configurar las credenciales de Git, hacer cambios en los archivos y enviar los cambios al repositorio de GitHub.
- El flujo de trabajo general en Git y GitHub implica inicializar un repositorio local, agregar archivos, comprometer cambios, crear una nueva rama si es necesario, enviar cambios al repositorio remoto, solicitar una solicitud de extracción si se trabaja en colaboración y fusionar cambios en la rama principal.

Capítulo 47

Geoprocесamiento en nube

Dominic Royé

Fundación de la Investigación del Clima

47.1. Introducción

Cuando se plantea un problema basado en datos desde diversos proveedores, habitualmente implica la descarga de grandes volúmenes. La actual proliferación de servicios de Open Data, despliegues de sensores y diversas fuentes incluyendo los satélites dificulta su procesamiento en equipos personales. El gran crecimiento en grandes volúmenes de datos espacio-temporales de tipo vectorial o ráster lleva a la necesidad en trabajar con servicios en nube para ahorrar tiempo computacional y espacio de almacenamiento. En la actualidad existen diferentes servicios de geoprocесamiento en nube que ayudan a hacer análisis online sin necesidad de descargar los datos ni preocuparse por el rendimiento computacional. Uno de estos servicios es *Google Earth Engine* (GEE), donde se combina un catálogo de varios petabytes de imágenes satelitales y conjuntos de datos geoespaciales multidimensionales (vectorial y ráster) de alta resolución con capacidades de análisis a escala planetaria. Este servicio gratuito para uso no comercial incluye incluso la posibilidad en crear aplicaciones.

GEE consiste en una API con bibliotecas de cliente para JavaScript y Python, que traducen los análisis geoespaciales y hacen posible acceder a los datos. No es necesario descargar grandes volúmenes de datos ni configurar la computación. Para el lenguaje de R se puede hacer uso del paquete `rgee` que hace puente entre **R** y la API GEE.

Se hará uso del dataset con el nombre “NOAA CDR OISST v02r01”, una interpolación óptima diaria de temperatura de la superficie del mar (OISST, por sus siglas en inglés) con una resolución de 1/4 grados (27 km). Los datos los proporciona la National Oceanic and Atmospheric Administration (NOAA) con campos completos de temperatura del océano construidos mediante la combinación de observaciones ajustadas por sesgo de diferentes plataformas

(satélites, barcos, boyas) en una cuadrícula global regular, con lagunas estimadas por interpolación (https://developers.google.com/earth-engine/datasets/catalog/NOAA_CDR_OISST_V2_1) (Reynolds et al. (2008)).

El objetivo de este capítulo es mostrar el potential del uso de APIs directamente en **R**. El resultado se empleará en el Cap. 56.

47.2. Sintaxis de Google Earth Engine

Con ayuda de GEE se preprocesan los datos de tal manera que el resultado son las anomalías estivales en forma de ráster para cada año entre 1981 y 2022. El primer paso consiste en crear el usuario en earthengine.google.com. Además, es necesario instalar *CLI* de *gcloud* (<https://cloud.google.com/sdk/docs/install?hl=es-419>).

Antes se deben conocer algunos conceptos fundamentales sobre la sintaxis en GEE. En general, el lenguaje nativo es Javascript el que se caracteriza por la forma combinando funciones y variables usando el punto, el que se sustuye por el \$ en R. Todas las funciones GEE en **R** empiezan por el prefijo ee_* (`ee_print()`, `ee_image_to_drive()`). Los términos más relevantes son los siguientes:

- *ImageCollection*: serie temporal de imágenes.
- *Geometry*: dato vectorial.
- *Functions*: `map()` aplica funciones sobre *ImageCollections*, `ee.Date()` define una fecha, `filterDate()` permite filtrar por fecha una *ImageCollection*, `select()` selecciona una banda, etc.

Muchas funciones son similares a las de **tidyverse**.

Se puede obtener más ayuda en <https://r-spatial.github.io/rgee/reference/rgee-package.html> y en la propia página de GEE.

47.3. Primeros pasos

Después de darse de alta en GEE y haber instalado *CLI gcloud* en el sistema operativo, se crea un entorno virtual de Python con todas las dependencias de GEE usando la función `ee_install()`.

```
library(rgee)

ee_install() # crear entorno virtual de Python; ¡sólo una vez!
ee_check() # comprobar si todo está correcto
```

Antes de pasar a programar con la sintaxis propia de GEE, se debe autenticar e inicializar GEE empleando la función `ee_Initialize()`.

```
ee_Initialize(drive = TRUE) # autenticar e inicializar GEE
ee_user_info() # inf sobre usuario
```

Hay que tener en cuenta que, únicamente cuando se envían tareas, GEE ejecuta el cálculo en los servidores enviando todos los objetos creados. En la mayoría de los pasos se crean objetos *EarthEngine*, que se usan una vez que se construyó un mapa interactivo, la exportación o la impresión en consola de un objeto.

Por ejemplo, se puede seleccionar la banda NDVI del producto MODIS MOD13A2 e imprimir los metadatos del primer día disponible con `ee_print()`. Existe un límite 5000 elementos que se podrían ver usando esta función.

```
# imageCollection NDVI
img <- ee$ImageCollection('MODIS/006/MOD13A2')$select('NDVI')

# metadatos
ee_print(img$first())
```

47.4. Cálculo de anomalías

47.4.1. Definiciones previas

Los datos NOAA CDR OISST contienen la temperatura superficial de los océanos a nivel global, por eso, se fija un rectángulo que cubre la extensión del Mar Mediterráneo como objeto de estudio.

```
# extensión del Mar Mediterráneo
geom <- ee$Geometry$Polygon(coords = list(
  c(-6.046418548121442, 46.733937391710846),
  c(-6.046418548121442, 29.680544334046786),
  c(42.469206451878556, 29.680544334046786),
  c(42.469206451878556, 46.733937391710846)
),
proj = "EPSG:4326",
geodesic = FALSE)

geom #vemos que es un objeto EarthEngine de tipo geometría
str(geom) # construcción javascript
```

En el siguiente paso se define el período de interés, desde el año 1982 hasta el 2022.

```
startDate <- ee>Date('1982-01-01') # fecha inicio
endDate <- ee>Date('2023-01-16') # fecha final
```

Se puede acceder a todas las colecciones (*ImageCollection*) indicando su identificación. Además, se filtran y se recortan los datos con respecto al periodo y a la extensión fijada. Finalmente se selecciona la banda o variable de interés “sst” (*surface sea temperature*).

```
collection_era5 <- ee$ImageCollection("NOAA/CDR/OISST/V2_1")$filterDate(startDate, endDate)$filterBounds(geom)$select('sst')
```

Finalmente, se procede a calcular el número de años en el período fijado.

```
number0fyears <- endDate$difference(startDate, 'years')$round()
```

47.4.2. Promedio estival

Después de las anteriores definiciones se crea una nueva colección con el promedio estival de cada año durante el periodo objeto de estudio. Para ello se crea una lista de los años sobre la que se mapea otra función. Esta función se debe pasar con `ee_utils_pyfunc()`, que traduce una función R a una de Python.

En la función personalizada se filtran los meses de verano, se calcula el promedio y se multiplica por 0,01, un factor de escala. Cuando se crean nuevas colecciones es importante fijar la nueva fecha con `set()`.

```
yearly <- ee$ImageCollection(
  ee>List$sequence(0, number0fyears$subtract(1L))$map(ee_utils_pyfunc(function(dayOffset) {
    yr = startDate$advance(dayOffset, 'years')$get('year')
    start = ee>Date$fromYMD(yr, 12L, 1L)
    end = ee>Date$fromYMD(yr$add(1L), 2L, 28L)
    return(collection_era5$filterDate(start, end)$mean()$multiply(0.01)$set('system:time_start', start$millis()))
  })))
)
```

En el siguiente paso se calcula la temperatura media estival entre 1982 y 2010, como período de referencia para las anomalías.

47.4. Cálculo de anomalías

801

```
msst <- collection_era5$filterDate('1982-01-01', '2010-12-31')$  
  filter(ee$Filter$calendarRange(12L, 2L, 'month'))$  
  mean()$  
  multiply(0.01)
```

Se aplica otra función personalizada sobre las medias estivales de todos los años, en la que se resta la temperatura del periodo de referencia, obteniéndose así las diferencias entre la temperatura media de cada año en el periodo estival y la temperatura media global del periodo estival en el periodo 1982-2010.

```
anom <- yearly$map(ee_utils_pyfunc(function (im) {  
  return(im$subtract(msst)$set('system:time_start',  
                                im$get('system:time_start')))  
}))
```

Es puede crear un mapa interactivo de un año concreto aplicando la función `Map.addLayer()` (Fig. 47.1). En este paso es la primera vez que GEE calcula lo que se ha creado anteriormente.

```
# metadatos  
ee_print(anom$first()) # año 1982  
  
# mapa interactiva del año 1982  
Map$setCenter(9, 40, 5) # centrar mapa en el mediterráneo con nivel de zoom 5  
  
# crear mapa con leyenda  
Map$addLayer(  
  eeObject = anom$first(),  
  visParams = list(  
    palette = rev(RColorBrewer::brewer.pal(11, "RdBu")),  
    min = -3,  
    max = 3  
  ),  
  name = "MED_SST"  
) +  
Map$addLegend(list(min = -3, max = 3,  
                  palette = rev(RColorBrewer::brewer.pal(11, "RdBu"))),  
             name = "SST Anomaly",  
             position = "bottomright",  
             bins = 4)
```

Hasta este momento, no se ha enviado una tarea para que GEE la realice. Para ello hay que exportar las anomalías de cada año en formato `geotiff`. La función `ee_imagecollection_to_local()` facilita la exportación de todas las capas de una `ImageCollection`. En cambio, la función `ee_image_to_drive()` exporta datos individuales de una única imagen a `Google Drive`. El argumento `scale` indica con qué resolución se exporta. Aunque la resolución de los datos originales es de 27 km, se fija una resolución de 2 km, lo que implica una interpolación en la exportación por razones de estética en la visualización.

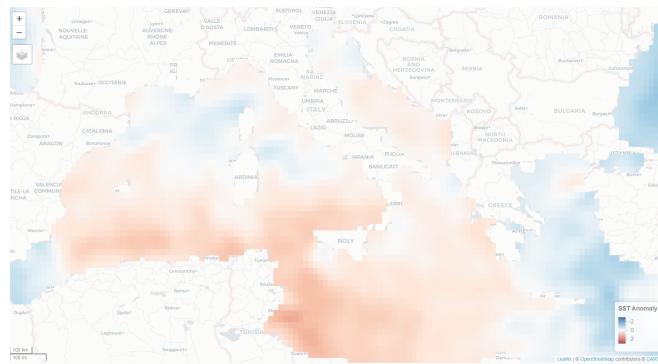


Figura 47.1: Mapa interactivo del mar Mediterráneo (1982)

```
tmp <- tempdir() # carpeta temporal o cualquier otra ruta

ic_drive_files_2 <- ee_imagecollection_to_local(
  ic = anom,
  region = geom,
  fileFormat = "GEO_TIFF",
  scale = 2000,
  lazy = FALSE,
  dsn = file.path(tmp, "rast_"), # prefijo del archivo
  add_metadata = TRUE
)
```

Este ejemplo ha mostrado una pequeña parte de la capacidad del geoprocесamiento en manejar grandes volúmenes de datos sin que implique su descarga ni el cálculo localmente. Pero también es posible manejar datos vectoriales u otros datos de alta resolución. Incluso se puede llegar a crear aplicaciones basadas en GEE.

Resumen

El crecimiento de volúmenes de datos en la actualidad lleva a la necesidad de emplear servicios de geoprocесamiento en nube que ayuden a hacer análisis online sin necesidad de descargar los datos ni preocuparse por el rendimiento computacional. Uno de estos servicios es *Google Earth Engine*, donde se combina un catálogo de imágenes satelitales y conjuntos de datos geoespaciales multidimensionales de alta resolución con capacidades de análisis a escala planetaria. En este ejemplo se ha mostrado cómo procesar la temperatura superficial del mar calculando las anomalías estivales de la cuenca mediterránea desde 1982 a 2022.

Parte XI

Casos de estudio en ciencia de datos

Capítulo 48

Análisis de una red criminal

F. Liberatore^a, L. Quijano-Sánchez^b, M. Camacho-Collados^c

^aCardiff University, ^bUniversidad Autónoma de Madrid, ^cMinisterio del Interior

48.1. Introducción

En este capítulo se plantea la idea de realizar un **análisis de una red de crimen organizado**. Para ello, se estudia la red derivada de un *dataset* real, relativo a la operación *Oversize*. El estudio se llevará a cabo usando la librería `igraph`.

48.2. El conjunto de datos *Oversize*

Los datos que se van a analizar se han obtenido de la operación *Oversize* (Berlusconi et al., 2016) (Tribunale di Milano, Ufficio del giudice per le indagini preliminari, 2006) (Tribunale di Lecco, 2a Sezione Penale, 2009), un proceso italiano contra un grupo mafioso. La investigación duró del 2000 al 2006, y se enfocó en más de 50 sospechosos involucrados en tráfico internacional de drogas, homicidios y robos. El juicio empezó en el 2007 y duró hasta el 2009, cuando se dictó la sentencia y los principales sospechosos fueron condenados con penas de 5 a 22 años de cárcel. La mayoría de los sospechosos eran afiliados de la '*Ndrangheta*, una mafia de Calabria (una región del sur de Italia) con ramificaciones en otras regiones y en el extranjero.

En particular, se va a estudiar la red obtenida de las escuchas telefónicas. Los datos hacen referencia a todas las conversaciones telefónicas transcritas por la policía y consideradas relevantes. En esta red, los nodos representan sospechosos (los datos son anónimos y los nombres asignados en la red se han generado de forma aleatoria). Las aristas conectan los sospechosos que han tenido al menos una conversación telefónica relevante al caso durante la investigación.

48.3. Creación de la red mafiosa

El dataset `Oversize_nodes` contiene el listado de nodos con sus propiedades, en este caso el nombre (ficticio) del sospechoso. `Oversize_edges` contiene las aristas del grafo, representadas como parejas de nodos, a su vez identificados por su ID. A partir de estos datasets la librería `igraph` permite crear un grafo, tal y como se ilustra a continuación.

```
library('igraph')
library('CDR')
data(oversize_edges, oversize_nodes)
net <- graph_from_data_frame(d=oversize_edges,
                             vertices=oversize_nodes,
                             directed=F)

net
#> IGRAPH 3e2ce9f UN-- 182 247 --
#> + attr: name (v/c)
#> + edges from 3e2ce9f (vertex names):
#> [1] Casto Ben          --Gustavo Mango
#> [2] Casto Ben          --Metrofane Abbatiello
#> [3] Uranio Natoli     --Fidenziano Marcellino
#> [4] Lancilotto Di Biasi--Romolo Gemignani
#> [5] Lancilotto Di Biasi--Fidenziano Marcellino
#> [6] Senesio Rabito     --Pacifico Caliri
#> [7] Senesio Rabito     --Michelangelo Piccininni
#> [8] Romolo Gemignani   --Alighiero Mazzarella
#> + ... omitted several edges
```

La vista previa del grafo indica lo siguiente:

- El grafo es no dirigido (`UN`) y está compuesto por 182 nodos y 247 aristas.
- El único atributo es el nombre de los nodos (`attr: name (v/c)`).
- Finalmente, se proporciona una previsualización de un subconjunto de aristas, indicando para cada una los dos nodos conectados (ej. `Casto Ben --Gustavo Mango`).

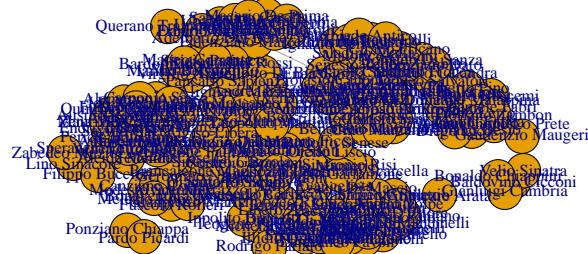
48.4. Visualización de la red mafiosa

Para hacerse una idea de que aspecto tiene el grafo, se procede a su visualización, usando el comando `plot()` de **R**.

```
plot(net, asp=0)
```

48.4. Visualización de la red mafiosa

807

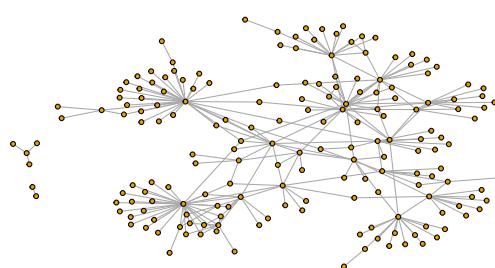


Como se puede apreciar, el resultado no es muy claro. Todos los nodos tienen el mismo tamaño y se solapan entre ellos. Además, se muestran los nombres de todos los actores dentro de la red, lo cual dificulta ulteriormente su interpretación.

Se puede mejorar esta presentación usando unos parámetros de `plot()`, específicos de `igraph`. En particular:

- `vertex.size`: determina el tamaño de los nodos.
- `vertex.label`: define el texto asociado a cada nodo. Por defecto se asume que es su nombre. En el ejemplo de abajo, se excluyen los nombres de la visualización.

```
plot(net, vertex.size=2, vertex.label=c(''), asp=0)
```



En la Fig. ?? se puede ver cómo el grafo permite una mejor valoración de la distribución de los actores dentro de la red. Por ejemplo, hay dos grupos pequeños (de cuatro y dos actores) completamente desconectados de la red principal.

48.5. Importancia de los actores (delincuentes)

Las medidas de centralidad permiten asignar un valor a cada actor que establece su importancia relativa a los demás. Existen diversas medidas, cada una con sus características y finalidad. En este ejemplo se van a usar las siguientes:

- **Grado:** número de aristas que llegan al nodo o salen de él. Cuanto más alto sea este valor, más vecinos tendrá el nodo.
- **Intermediación:** cuantifica el número de veces que un nodo se encuentra en el camino más corto entre otros actores. Cuanto más alto este valor, más información pasará por el nodo.

```
dgr <- degree(net) # Centralidad de grado
btwn <- betweenness(net) # Centralidad de intermediación
```

A continuación, se muestran los actores con los valores más altos en cada medida de centralidad.

```
head(sort(dgr, decreasing = T))
#>      Gustavo Mango      Pacifico Caliri Metrofane Abbatiello
#>            32                  31                      18
#>      Olindo Iacona      Arturo Gizzi      Guido Minervini
#>            17                  16                      16
```

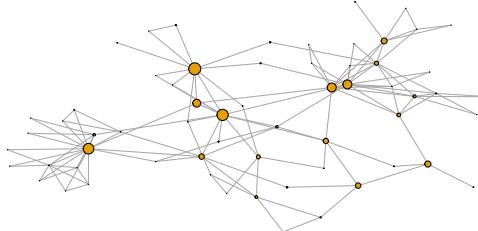
```
head(sort(btwn, decreasing = T))
#>      Gustavo Mango      Bino Lana      Pacifico Caliri
#>        4602.167      4292.902      4056.435
#>      Olindo Iacona Metrofane Abbatiello      Donato Di Santi
#>        3397.907      3387.931      2978.427
```

Las medidas de centralidad se pueden usar para mejorar la visualización del grafo. Primero, se filtran todos los nodos que tengan grado menor que dos, ya que representan actores muy marginales en la red. Luego, se representa el tamaño de cada nodo en función de su valor de intermediación, escalando con un tamaño máximo de cinco.

```
vertex_filter <- dgr > 1 # detección actores marginales
scaled_btwn = 0.1 + 4.9*btwn/max(btwn) # Escalado del tamaño del nodo en función de la
#> → intermediación
net2 = induced.subgraph(net, which(vertex_filter)) # creación subgrafo
plot(net2,
      vertex.size=scaled_btwn[vertex_filter],
      vertex.label=c(''),
      rescale=T,
      asp = 0) # visualización subgrafo
```

48.6. Identificación de comunidades de la mafia

809



Como se puede apreciar en la Fig. ??, gracias a las medidas de centralidad se puede tener una mejor idea de cómo se configura la red respecto a sus actores más importantes.

48.6. Identificación de comunidades de la mafia

A continuación, se procede a identificar las comunidades existentes en el grafo de la operación *Overdrive*. `igraph` proporciona una gran variedad de algoritmos para la detección de comunidades en redes sociales. En el siguiente ejemplo, se usa el algoritmo Louvain (Blondel et al., 2008) que es el más popular.

```
louvain_partition <- cluster_louvain(net) # Ejecucion del algoritmo Louvain
net$community <- louvain_partition$membership # Asignacion de las comunidades al grafo
unique(net$community) # Visualizacion de las comunidades encontradas
#> [1] 1 2 3 4 5 6 7 8 9
```

El algoritmo identifica distintas comunidades, cada una con su número asignado.

48.7. Visualización de comunidades de la mafia

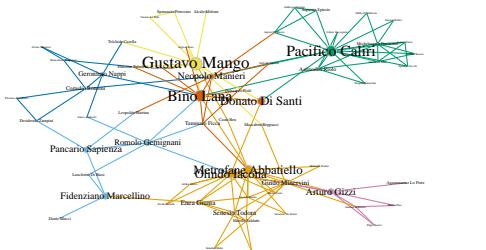
Se procede a visualizar las comunidades detectadas en el subgrafo, representando cada una de ellas en un color distinto. Además, para mejorar la calidad de la información representada, se resalta la importancia de cada actor representando los nodos asociados y sus nombres en tamaños proporcionales a su centralidad en toda la red.

```
V(net2)$size <- scaled_btwn[vertex_filter] # Tamaño del nodo en funcion de su
#> centralidad
V(net2)$frame.color <- "grey"
```

```
V(net2)$color <- net$community[vertex_filter] # Color del nodo en función de su
→ comunidad
V(net2)$label <- V(net2)$name
V(net2)$label.cex <- (1+scaled_btwn[vertex_filter])/6 # Escalado del nombre en función
→ de su centralidad
V(net2)$label.color <- 'black'

# Definición del color de las aristas en función de la comunidad de origen
edge.start <- ends(net2, es = E(net2), names = F)[,1]
E(net2)$color <- V(net2)$color[edge.start]

plot(net2, asp=0) # Los resultados puede ser distintos con cada ejecución
```



Se puede mejorar aún más el aspecto del grafo. Para ello, se va a experimentar con una disposición diferente de los nodos. En este ejemplo, se usa el algoritmo Fruchterman-Reingold ([Fruchterman and Reingold, 1991](#)). Además, se aplica un efecto de curvatura a las aristas asignando un valor positivo al parámetro `edge.curved`. El resultado se puede ver en la Fig. ??.

```
l1 <- layout_with_fr(net2) # algoritmo Fruchterman-Reingold
plot(net2,
      layout=l1,
      asp = 0,
      edge.curved=0.5) # Los resultados pueden ser distintos con cada ejecución
```

48.7. Visualización de comunidades de la mafia

811



Finalmente, se puede exportar el grafo como PDF usando la función `pdf()` de **R**.

```
pdf('grafo_final.pdf')
plot(net2, layout=11, asp = 0, edge.curved=0.5) # Los resultados puede ser distintos
#> con cada ejecucion
dev.off()
#> pdf
#> 2
```

Como se ha podido observar tras las acciones anteriores, en la red se aprecian siete distintas comunicades. Tres destacan por su importancia, lideradas por Gustavo Mango, Bino Lana y Pacifico Caliri. Bino Lana, en particular, tiene especial relevancia ya que actúa como un puente entre Gustavo Mango y Pacifico Caliri.

Capítulo 49

Optimización de inversiones publicitarias

Carlos Real Ugena

Deloitte

49.1. Metodologías para optimizar las inversiones publicitarias

Uno de los principales retos a los que se enfrentan los departamentos de marketing de cualquier compañía es cuantificar el impacto de la publicidad en el negocio. Esta medición es clave en la optimización de las inversiones que destinan a cada medio publicitario, existiendo múltiples métodos para medir el ROI (*Return On Investment*), es decir, el retorno que se obtiene por cada euro invertido en publicidad. Antes de revisar un ejemplo práctico, es necesario entender bien las características que presentan cada uno de ellos. Los métodos de cuantificación del impacto de la publicidad en el negocio, según la investigación llevada a cabo por la consultora Gartner¹, se pueden clasificar en cuatro grandes grupos:

1. **Marketing Mix Modeling (MMM)**: modelos de series temporales que sirven para estimar la contribución del marketing u otras variables explicativas a las ventas desde un punto de vista estratégico.
2. **Multitouch Attribution (MTA)**: modelos de atribución que valoran cada punto de contacto (*touchpoint*) del recorrido del cliente (*customer journey*), asignando a cada uno un peso en la conversión (venta, descarga de folleto, etc). Son modelos tácticos que normalmente se centran en el canal online.

¹ Artículo “Gartner Identifies Four Methods for Measuring Marketing’s Impact.” disponible en <https://www.gartner.com/en/newsroom/press-releases/2020-03-04-gartner-identifies-four-methods-for-measuring-marketing-s-impact>

3. **Holdout Testing o Experiments (EXP)**: modelos causales que evalúan el impacto de una campaña publicitaria a partir de una muestra de control y otra de test.
4. **Unified Measurement Approaches (UMA)**: combinación de los modelos anteriores (MMM+MTA+EXP) con el objetivo de tener una visión unificada de los resultados.

En la comparativa (Fig. 49.1) que se muestra a continuación se resume el objetivo de cada uno de los modelos, así como las preguntas que permiten responder:

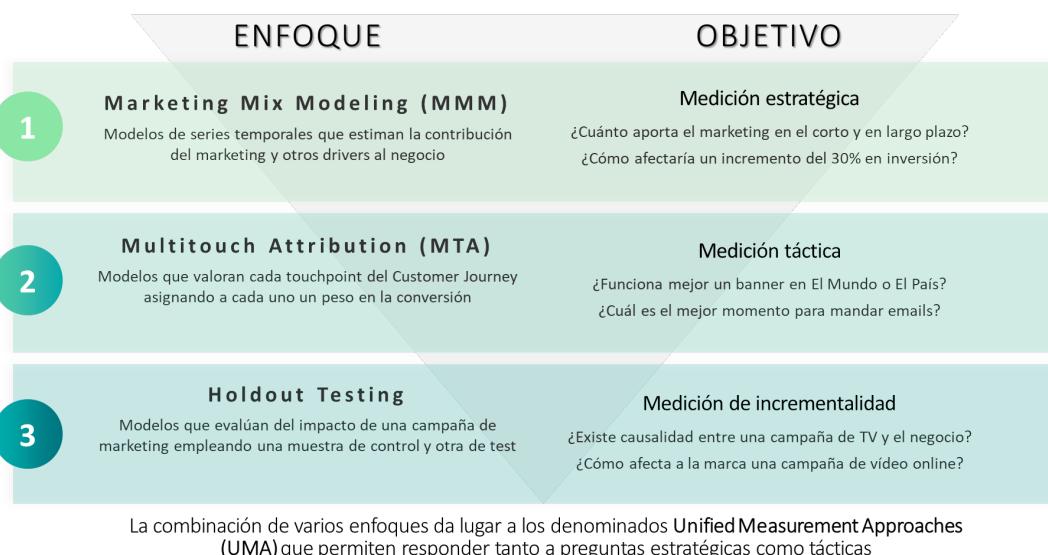


Figura 49.1: Comparativa de metodologías de medición

En los últimos años se han producido una serie de cambios en el entorno de la industria del marketing digital enfocados a garantizar el más estricto control de la privacidad de los usuarios. Desde el lanzamiento de la nueva ley de protección de datos “GDPR” en 2018, hasta la reciente confirmación por parte de Google de la prohibición de uso de cookies de tercera parte en “Chrome”, la posibilidad de acceder a datos de identificación personal para la medición y optimización del impacto de las campañas de publicidad digital está cada vez más limitada.

Esta situación está provocando que el primer tipo de enfoque, el MMM, esté siendo el gran beneficiado, dado que es una técnica que no depende del acceso a datos a nivel de individuo. El ejemplo más claro sobre el protagonismo que está alcanzando el MMM es que grandes compañías como Meta, Google y Uber están desarrollando soluciones *open-source* basadas en técnicas de *Machine Learning* que integran grandes avances y mejoras sobre las metodologías tradicionales. Entre las metodologías que han desarrollado, hay claras diferencias en las bases teóricas sobre las que se rigen, así como diferencias en tiempos de computación, lenguaje en el que se han desarrollado o capacidades funcionales. Se detalla a continuación las principales características de cada una de ellas:

- **Robyn**²: desarrollado por Meta, **Robyn** ([Facebook Marketing Science \(2021\)](https://facebookexperimental.github.io/Robyn/)) está pensado para *datasets* con gran cantidad de variables independientes dado que trabaja con regresiones *ridge* (cubiertas en el Cap. 19), las cuales están pensadas para lidiar con problemas de multicolinealidad, muy presentes en este tipo de análisis. Cabe destacar que, entre las pruebas que realiza, ofrece una serie de *outputs* visuales avanzados que permiten al usuario seleccionar el que mejor se adapta al contexto de negocio y necesidad. Requiere tiempos de cómputo alto, pudiendo llegar hasta las tres horas, y permite hacer optimizaciones de presupuesto.
- **Lightweight**³: desarrollado por Google, sus fundamentos teóricos se basan en modelos bayesianos. La particularidad de esta solución es la posibilidad de incluir datos geográficos para su posterior segregación. Lightweight también considera distintos tipos de *adstock* (recuerdo publicitario), así como la tendencia y estacionalidad de la serie a explicar. El uso de esta herramienta es más sencillo, aunque los resultados son expuestos en un notebook y no posee *outputs* más elaborados como el resto de soluciones. Por último, también permite realizar optimizaciones de presupuesto.
- **Orbit**⁴: desarrollado por Uber, se basa en modelos bayesianos. Permite al usuario medir los retornos a lo largo del tiempo, lo cual hace que sea un modelo adecuado para compañías con grandes picos de ventas. Incluye análisis de estacionalidad mediante la descomposición de la serie en series de Fourier. Es el modelo más complejo de desarrollar, sin embargo, el tiempo de ejecución es bajo. Cabe destacar que es la librería más estable y, por lo tanto, se podría utilizar para realizar reportes con mayor frecuencia. No incluye la posibilidad de optimizar presupuestos.

49.2. Robyn como alternativa *open-source* en R

Robyn está ganando en popularidad debido a las mejoras continuas que están aplicando desde Meta, así como gracias a su adaptabilidad a la realidad del anunciante. Además, existe un paquete en **R** validado y subido al CRAN que lo sitúa como una gran alternativa de código abierto en **R** para iniciarse en el campo de la medición y optimización del retorno de las inversiones publicitarias. La fórmula empleada es la siguiente:

$$y_t = \beta_0 + S\text{Curve}(x_j) + \beta_{hol}hol_t + \beta_{sea}sea_t + \beta_{trend}trend_t + \dots + \beta_{ETC}ETC_t + \epsilon_t \quad (49.1)$$

donde:

²Dirección Web Robyn: <https://facebookexperimental.github.io/Robyn/>.

³Dirección Web Lightweight: https://github.com/google/lightweight_mmm.

⁴Dirección Web Orbit: <https://www.uber.com/en-ES/blog/orbit/>.

y_t = ventas en el instante t

t = instante de las variables (por ejemplo, semanas)

j = subíndice del medio (por ejemplo, TV o Display)

β_0 = intercepto

$x_{decay_{t,j}} = x_{t,j} + \theta_j x_{decay_{t,j-1}}$ (adstock, es decir, el recuerdo publicitario)

x_j = inversión publicitaria en cada medio j

θ_j = tasa fija de decrecimiento en cada medio j

$$SCurve(x, j) = \beta_j \times \frac{x_{decay_{t,j}}^\alpha}{x_{decay_{t,j}}^\alpha + \gamma^\alpha} \text{ transformación no lineal de curva en S} \quad (49.2)$$

α, γ = hiperparámetros que definen la S-Curve

γ : implementada en la SCurve, donde $\gamma_{tran} = cuantil(x_{decay_j}, \gamma)$

β_j = coeficientes de cada medio j

hol = festivos

sea = estacionalidad

$trend$ = tendencia

ETC = resto de variables independientes (precio, promociones, etc)

ϵ_t = término de error en el instante t

Las variables de medios no suelen presentar un efecto lineal sobre la variable de negocio que se modeliza, sino que la publicidad tiende a presentar rendimientos decrecientes no lineales. Para modelizar este efecto se utiliza la función biparamétrica de *S Curve* o curva de rendimientos decrecientes. Esta curva permite optimizar los repartos presupuestarios en todos los canales de medios, ya que su forma en “S” indica tanto el umbral a partir del cual los resultados del gasto presupuestario mejoran significativamente el objetivo (por ejemplo, ventas), como en qué punto está saturando y, por lo tanto, perdiendo eficacia.

A continuación, se aplica Robyn sobre un ejemplo con información simulada del sector hotelero, donde el objetivo es predecir el número de reservas de un hotel en función de una serie de predictores (entre ellos, las inversiones publicitarias). Toda la información sobre cómo aplicar esta metodología se puede consultar en Github⁵. La versión utilizada en este ejemplo práctico es la 3.6.3, disponible en el CRAN o descargable desde Github⁶. Este ejercicio traza el camino más corto que se puede seguir hasta llegar a los principales *outputs* y a la interpretación de los mismos. Sin embargo, para conocer en detalle la metodología, se recomienda seguir profundizando a través de la realización de pruebas adicionales más complejas.

Para comenzar, se carga el paquete Robyn, comprobando que se está trabajando con la versión 3.6.4 y forzando el uso de multicore al utilizar Rstudio:

⁵Metodología de Robyn disponible en <https://github.com/facebookexperimental/Robyn>

⁶Versión 3.6.3 de Robyn disponible en <https://github.com/facebookexperimental/Robyn/releases/tag/v3.6.3>

49.2. Robyn como alternativa open-source en R

817

```
library(Robyn)
packageVersion("Robyn")
Sys.setenv(R_FUTURE_FORK_ENABLE = "true")
options(future.fork.enable = TRUE)
```

Se carga una librería de Python llamada `nevergrad` (Facebook Research AI (2019)), necesaria en el proceso de estimación de los hiperparámetros, invocándola desde R. Hay varias opciones para llevar a cabo este proceso, una de ellas es instalar primero el paquete `reticulate` (Ushey et al. (2022)) y, a continuación, `nevergrad` vía pip:

```
install.packages("reticulate")
library(reticulate)
virtualenv_create("r-reticulate")
use_virtualenv("r-reticulate", required = TRUE)
py_install("nevergrad", pip = TRUE)
py_config()
```

En caso de encontrar alguna dificultad al cargar `nevergrad`, existen distintas alternativas para su instalación que se pueden consultar en Github⁷.

Posteriormente, se carga la tabla que recoge la información simulada del sector hotelero con la que se medirá el impacto de la publicidad:

```
library("CDR")
data('hotel_tablonsemanal')
head(hotel_tablonsemanal)
```

Se detalla a continuación la información que contiene cada variable:

1. *semana*: lunes de la semana de referencia.
2. *reservas*: número de reservas que ha conseguido la cadena hotelera en cada una de las semanas bajo análisis.
3. *turismo*: número de pernoctaciones.
4. *covid_mov*: movilidad desde el comienzo de la pandemia.
5. *notoriedad*: conocimiento espontáneo de la marca a lo largo del tiempo.
6. *temperatura*: temperatura media.
7. *tv_grps20* y *tv_inv*: métrica de impactos (GRPs) e inversión realizada en TV.
8. *resto_off_inv*: resto de inversiones offline realizadas.
9. *paidsearch_imp* y *paidsearch_inv*: métrica de impactos (impresiones) e inversión realizada en Paid Search.
10. *display_imp* y *display_inv*: métrica de impactos (impresiones) e inversión realizada en Display.

⁷ Alternativas a la instalación de `nevergrad` disponibles en <https://github.com/facebookexperimental/Robyn/issues/189>.

11. *onlinevideo_imp* y *onlinevideo_inv*: métrica de impactos (impresiones) e inversión realizada en Online Video
12. *competidores*: inversión realizada por la competencia.
13. *eventos*: recoge eventos que tienen impacto en la serie de reservas. En este caso se considera el Black Friday.

El objetivo de este ejercicio es estimar el impacto de cada una de las variables detalladas sobre las reservas, poniendo especial foco en las variables de medios —TV, Resto Medios Offline, Paid Search, Display y Online Video— para medir sus efectos y, posteriormente, optimizar sus inversiones.

Con este fin, se cargan los festivos de una tabla auxiliar:

```
data('festivos')
head(festivos)
```

Se fija dónde se guardarán los resultados:

```
robyn_object <- "data/MyRobyn.RDS"
ruta_outputs <- "data"
```

Y se define la configuración inicial del modelo:

```
hotel_tablonsemanal$eventos <- hotel_tablonsemanal$eventos |> replace_na('na')
InputCollect <- robyn_inputs(
  dt_input = hotel_tablonsemanal
  ,dt_holidays = festivos
  ,date_var = "semana" # tiene que tener este formato "2020-01-01"
  ,dep_var = "reservas" # sólo una variable dependiente
  ,dep_var_type = "conversion" # "revenue" o "conversion". En nuestro caso son
  → reservas.
  ,prophet_vars = c("trend", "season", "holiday") #
  → "trend=tendencia", "season=estacionalidad", "weekday=día de la semana" &
  → "holiday=festivos"
  ,prophet_signs = c("default", "default", "default")
  ,prophet_country = "ES"# selección un país. España en nuestro caso
  ,context_vars = c("eventos", "turismo", "covid_mov", "notoriedad", "temperatura",
  → "competidores") # variables de contexto que no sean medios
  ,context_signs = c("default", "positive", "negative", "positive", "default",
  → "negative")# signos fijados a priori (por ejemplo, el turismo debe tener un signo
  → positivo puesto que estamos analizando reservas de hotel)
  ,paid_media_spends = c("display_inv", "onlinevideo_inv", "paidsearch_inv"
  → , "resto_off_inv", "tv_inv") # variables de inversión
  ,paid_media_signs = c("positive", "positive", "positive", "positive", "positive")
  ,paid_media_vars = c("display_imp", "onlinevideo_imp" , "paidsearch_imp"
  → , "resto_off_inv" , "tv_grps20") # variables de impacto si están disponibles. Si no
  → están disponibles utilice el coste como en el caso de Resto_off_inversion
  ,factor_vars = c("eventos") # especifique variables que son factores. En nuestro
  → caso, sólo la variable de Eventos
```

49.2. Robyn como alternativa open-source en R

819

```

,window_start = "2018-10-01">#fecha de inicio del modelo. En nuestro caso, octubre
← 2010 porque previamente no se tienen datos de reservas
,window_end = "2021-09-27" #fecha fin del modelo
,adstock = "geometric" # tipo de adstock. Seleccione el adstock geométrico para
← reducir tiempos de cómputo
)
print(InputCollect)

```

Después, se obtienen los nombres de los hiperparámetros a ajustar:

```
hyper_names(adstock = InputCollect$adstock, all_media = InputCollect$all_media)
```

A continuación, se definen los rangos en los que se moverán los hiperparámetros que definen la *S-curve* de cada medio. En este ejemplo se utilizan los mismos límites para α y γ , que controlan la forma de la curva y el punto de inflexión, respectivamente. Por otro lado, θ_j define el *adstock*, que se puede acotar teniendo en cuenta los intervalos recomendados por Meta: TV (entre 0,3 y 0,8), Resto Medios Offline (entre 0,1 y 0,4) y Digital (entre 0 y 0,3). En este caso, se deja θ libre entre 0 y 0,99 para todos los medios:

```

hyperparameters <- list(
  display_inv_alphas = c(0.0001, 3),display_inv_gammas = c(0.3, 1),display_inv_thetas =
  ← c(0, 0.99)

  ,onlinevideo_inv_alphas = c(0.0001, 3),onlinevideo_inv_gammas = c(0.3,
  ← 1),onlinevideo_inv_thetas = c(0, 0.99)

  ,paidsearch_inv_alphas = c(0.0001, 3),paidsearch_inv_gammas = c(0.3,
  ← 1),paidsearch_inv_thetas = c(0, 0.99)

  ,resto_off_inv_alphas = c(0.0001, 3),resto_off_inv_gammas = c(0.3,
  ← 1),resto_off_inv_thetas = c(0, 0.99)

  ,tv_inv_alphas = c(0.0001, 3),tv_inv_gammas = c(0.3, 1),tv_inv_thetas = c(0, 0.99)
)

```

Se añaden los hiperparámetros al *input* para ajustar los modelos:

```

InputCollect <- robyn_inputs(InputCollect = InputCollect, hyperparameters =
  ← hyperparameters)
print(InputCollect)

```

Y se ejecuta el modelo definiendo el número de iteraciones y *trials*. En este ejercicio, el número de iteraciones será 2000 y el de *trials* 10, obteniendo un total de 20000 posibles soluciones del modelo:

```
OutputModels <- robyn_run(
  InputCollect = InputCollect
  , iterations = 2000
  , trials = 10
  , outputs = FALSE # se desactivan los outputs que se guardarán después
)
print(OutputModels)
```

Los mejores resultados de la modelización según el frente de Pareto (conjunto de óptimos de Pareto que minimizan el error cuadrático medio normalizado (*Normalized Root Mean Square Error*, NRMSE) y la descomposición de la suma cuadrática de la distancia⁸ (*Decomposition Root Sum of Squared Distance*, DECOMP.RSSD)) se guardan en la carpeta seleccionada:

```
OutputCollect <- robyn_outputs(
  InputCollect, OutputModels
  , pareto_fronts = 3
  , csv_out = "all"
  , clusters = TRUE
  , plot_pareto = TRUE
  , plot_folder = ruta_outputs #ruta para exportar los resultados
)
print(OutputCollect)
```

Para finalizar, se revisan los distintos modelos obtenidos y el resumen gráfico proporcionado por Robyn. Se recomienda profundizar en la interpretación de uno de los modelos obtenidos con buen ajuste. A través del *one-pager* (Fig. 49.2), se puede consultar toda la información relativa al impacto de la publicidad, incluyendo métricas relativas a la bondad del ajuste como el coeficiente de determinación o R2 y el NRMSE en la parte superior:

Los principales *outputs* a evaluar de los modelos se visualizan en los gráficos del *one-pager*:

1. **Descomposición en cascada de la respuesta por predictor** (*Response Decomposition Waterfall by Predictor*): representa el peso de cada una de las variables en el modelo. En el ejemplo, el peso de la publicidad es del 54,5% (suma de los pesos de las variables de inversiones).
2. **Respuesta real vs. estimada** (*Actual vs. Predicted Response*): muestra el ajuste del modelo. En este caso, la línea del ajuste (azul) y la real (naranja) se aproximan bastante, indicando que el modelo es capaz de explicar la mayor parte de la variabilidad de la serie de reservas.
3. **Cuota de gasto frente a cuota de efecto con CPA total** (*Share of Spend vs Share of Effect with total CPA*): muestra la relación entre la cuota de inversión y la cuota de contribución generada por las inversiones publicitarias. En este ejercicio, se puede observar

⁸Indicador clave de Robyn que representa la diferencia entre la cuota de gasto y la cuota de efecto para las variables de medios.

49.2. Robyn como alternativa open-source en R

821

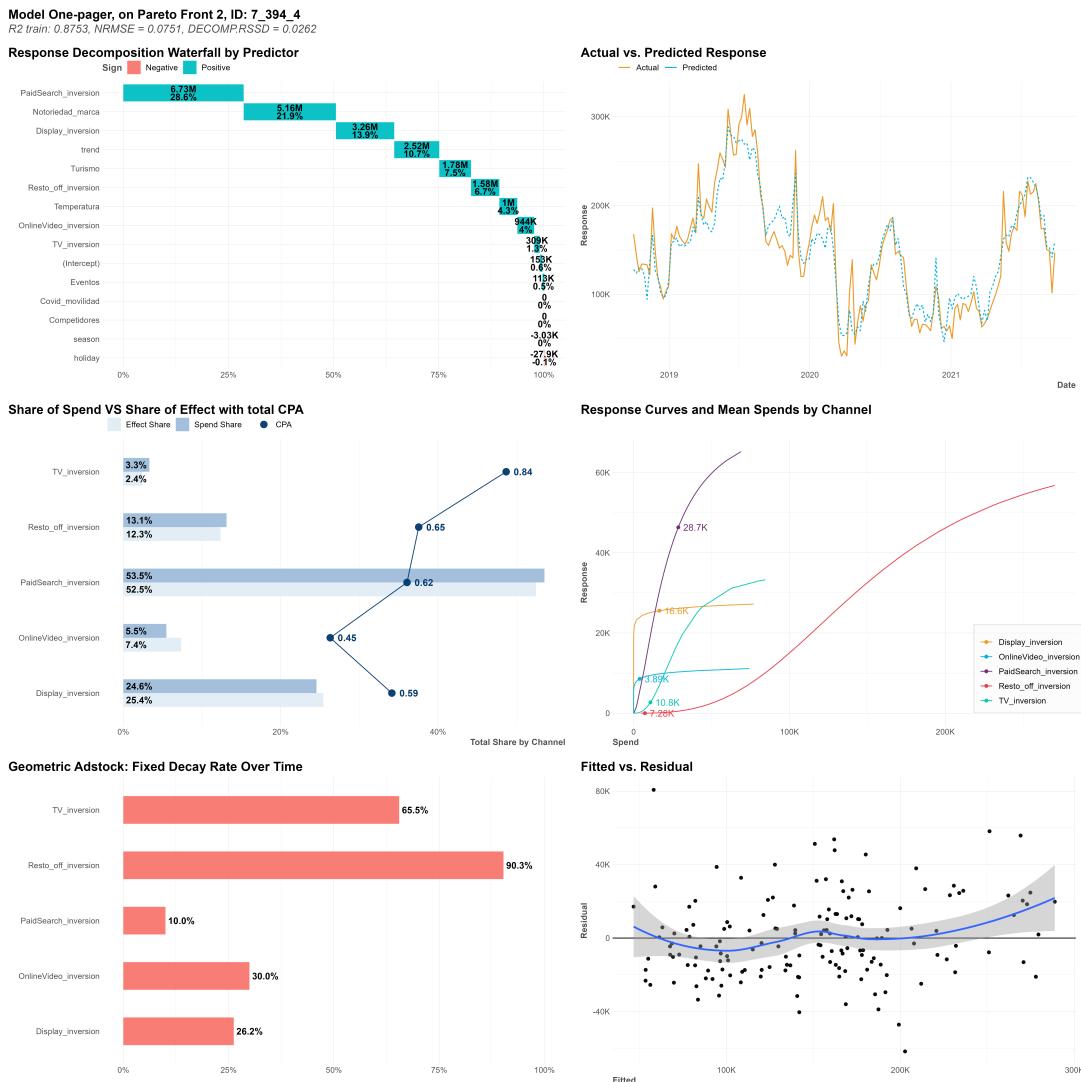


Figura 49.2: One-pager de resultados de Robyn

que tanto Online Video como Display obtienen un aporte mayor de lo esperado por su cuota de inversión, lo que hace que el Coste Por Adquisición o CPA (valores en azul) sean menores para estos medios. La TV se encuentra en el polo opuesto, con el máximo CPA, es decir, es el medio menos eficiente a la hora de generar reservas.

4. **Curvas de respuesta y gastos medios por canal** (*Response Curves and Mean Spends by Channel*): muestra la relación no lineal existente entre las inversiones y las reservas. Es el input principal a la hora de optimizar las inversiones.
5. **Adstock geométrico: Tasa constante de decrecimiento en el tiempo** (*Geometric Adstock: Fixed Decay Rate Over Time*): muestra el efecto recuerdo (*adstock*) de cada medio publicitario. Se puede observar que los medios offline (TV y Resto Medios Offline) son aquellos que presentan un efecto recuerdo más prolongado. No se debe olvidar que la configuración del modelo permite fijar el rango de valores que puede tomar el *adstock* en cada uno de los medios.
6. **Ajustados frente a residuos** (*Fitted vs. Residual*): muestra la nube de puntos para los datos ajustados (eje x) y los residuos (eje y). Este gráfico debe mostrar que los puntos están aleatoriamente ubicados alrededor del eje horizontal.

Y, ¿cómo se pueden optimizar las inversiones en medios empleando estos resultados? En primer lugar, se pueden utilizar las curvas obtenidas (parámetros guardados en el fichero *all_hyperparameters.csv*) para simular distintos escenarios y seleccionar el que genere un mayor número de reservas. La segunda opción es utilizar la documentación de código ⁹ incluida en el Step 5 para generar el reparto óptimo de inversión basado en un presupuesto pre-establecido.

En este capítulo se dan los primeros pasos en la medición y optimización del ROI de las inversiones publicitarias. Se anima al lector a que siga profundizando en este campo tan apasionante y complejo para basar las decisiones futuras de inversión en análisis desarrollados en **R**.

⁹Documentación de Robyn disponible en <https://github.com/facebookexperimental/Robyn/blob/main/demo/demo.R>

Capítulo 50

¿Cómo twitea Elon Musk?

Mariluz Congosto

Universidad Carlos III de Madrid

50.1. Introducción

El objetivo de este caso de uso es arrojar luz, de manera objetiva, sobre un **fenómeno social** de interés general: el uso de Twitter por parte de **Elon Musk**. Este multimillonario adquirió la red social el 28 de octubre de 2022 y, desde entonces, ha tomado decisiones drásticas, como reducir la plantilla y lanzar nuevos servicios de pago de manera apresurada. Su constante cambio de rumbo queda reflejado en su actividad frenética en Twitter, donde es un usuario muy activo.

Este caso de estudio aborda, una vez descargados los *tweets* mediante las **APIs de Twitter**¹, cómo adaptarlos mediante minería de textos (38) para su **análisis y visualización**. Se representa el contenido de estos mensajes mediante nubes de palabras, *scatters plots* y *timelines*. Este conjunto de gráficas ofrecerán distintas vistas de los datos que, sin duda, ayudarán a comprender los cambios de comunicación desde que adquirió Twitter.

50.2. Análisis visual de los *tweets* de Elon Musk

Las librerías que se utilizan son las siguientes:

¹Para el caso de estudio que se desarrolla se ha usado la librería `rtweet` para acceder a la **API Standard (V1.1)** (accesible a todo el mundo). Para obtener las credenciales que permiten trabajar en **modo usuario** se facilita el script de `python make_token_Twitter.ipynb` en github https://github.com/congosto/token_API_V1.1**. Este script se puede ejecutar en el entorno Google Colab <https://colab.research.google.com/>.

```
library("tidyverse")      # manipulación de datos
library("lubridate")       # formato de fechas
library("scales")          # manejo de escalas numéricas
library("tidytext")         # manipulación de textos
library("ggwordcloud")     # creación de una nube de palabras
library("RColorBrewer")     # paleta de colores
```

El rango de fechas elegido para delimitar temporalmente los tweets de Elon Musk va desde el 16-06-2022 hasta el 22-12-2022. Este rango es adecuado para visualizar la actividad e impacto del perfil de Musk antes y después de la adquisición de Twitter.

Se cargan los datos de la librería CDR del libro.

```
tweets_user <- CDR::elon_musk |>
  # Cambiar a formato fecha
  mutate(created_at = as.POSIXct(created_at, format = "%Y-%m-%dT%H:%M:%S", tz = "UTC"
  → ))
```

Una vez obtenidos los datos, se les puede dar forma. Los datos incluyen fechas, textos, tipos de tweets y métricas que pueden ser representados mediante gráficos. A continuación, se definen unos parámetros generales a todas las gráficas: la fecha de la compra de Twitter, el color de los distintos tipos de mensajes (*original*, *quoted*, *reply*, *retweeted*).

```
# Fecha en la que Elon Musk compró Twitter
compra_twitter <- as.POSIXct("2022-10-28")

# Se ordena la leyenda del tipo de tweet
order_tipo_tweet <- c("original", "quoted", "reply", "retweeted")
tweets_user$tipo_tweet <- factor(tweets_user$tipo_tweet, levels = order_tipo_tweet)

# Se define el color de los elementos de la leyenda del tipo de tweet
my_color <- c("retweeted" = "purple", "reply" = "blue", "quoted" = "green", "original"
  → = "red")
```

50.2.1. ¿Cuáles son los temas más recurrentes?

Para representar los términos más frecuentes en los tweets de Elon Musk se utiliza una nube de palabras. Esta representación gráfica se crea mediante la librería `ggwordcloud`, que funciona en el entorno `ggplot`.

El texto de los mensajes se encuentra en la variable `full_text`, la cual se limpia eliminando las URLs y los *handles* de los usuarios. Además, se añade una columna para distinguir los textos anteriores y posteriores a la compra de Twitter.

Posteriormente, se descomponen los textos en palabras independientes, se eliminan las *stop words* y se calcula la frecuencia de aparición de cada palabra.

50.2. Análisis visual de los tweets de Elon Musk

825

```

data(stop_words) # Descarga las stop words de la librería tidytext

corpus_text <- tweets_user |>
  mutate(text_plain = gsub("http\\S+\\s*", "", full_text)) |> # Quita las URL
  mutate(text_plain = gsub("RT @\\w+", "", text_plain)) |> # Quita los RTs
  mutate(text_plain = gsub("&", "&", text_plain)) |> # Rectifica el &
  mutate(text_plain = gsub("@\\w+", "", text_plain)) |> # Quita las menciones
  # Crea una columna para distinguir el periodo antes/después de la compra
  mutate(periodo = ifelse(created_at < compra_twitter,
    "Antes de la compra", "Después de la compra")) |>
  select(text_plain, periodo) |>
  unnest_tokens(word, text_plain) |> # Convierte las frases en un conjunto de palabras
  anti_join(stop_words) |> # Elimina las stop words
  group_by(word, periodo) |> # Agrupa por palabras
  summarise( freq = n(), .groups = "drop" ) |> # Calcula la frecuencia de cada palabra
  ungroup() |>
  arrange(desc(freq)) # Ordena de mayor a menor frecuencia de aparición

# print (corpus_text) # Descomentar para ver el resultado final

```

Para generar la nube de palabras de Elon Musk, se utiliza la función `geom_text_wordcloud_area()`. Esta función toma como entrada la lista de palabras y su frecuencia, y genera una comparación entre antes y después de la compra de Twitter (Fig. 50.1).

La proporción del tamaño de las palabras en la nube está en función de su frecuencia y se utiliza la librería `RColorBrewer` para definir la paleta de colores.

El resultado muestra que, antes de la compra de Twitter, Musk centraba su atención en sus empresas y en la guerra de Rusia-Ucrania. Sin embargo, tras la adquisición, su temática se relaciona con su nueva propiedad.

```

paleta <- brewer.pal(8, "Dark2")

ggplot() +
  geom_text_wordcloud_area(
    data = corpus_text |>
      top_n(300),
    aes(label = word, group = periodo, size = freq, color = freq),
    angle = 0.35
  ) +
  scale_size_area(max_size = 24) + # tamaño de las letras según frecuencia
  scale_color_gradientn(colors = paleta) +
  facet_wrap(~periodo) + # desdobra gráfica
  theme_minimal() +
  theme (strip.text = element_text(color = "grey50", size = 18))

```

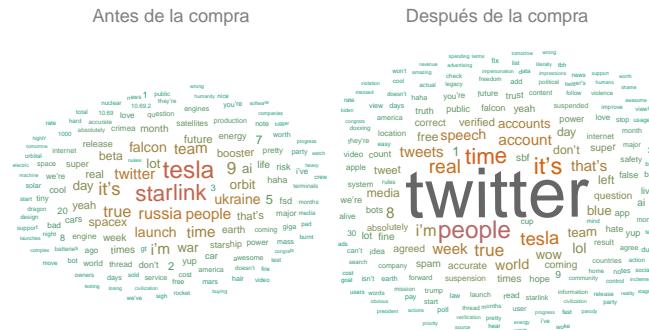


Figura 50.1: Palabras más frecuentes de Elon Musk en Twitter, antes y después de su compra

50.2.2. ¿Quiénes son los usuarios con los que más conversa?

Es posible visualizar con quiénes ha conversado Elon Musk con mayor frecuencia. Para ello, se pueden utilizar las respuestas que ha dado a otros usuarios en Twitter. Estas respuestas se obtienen de la variable `full_text`.

Para identificar con quiénes ha interactuado más Musk, se extraen los *handles* de los comentarios y se añade una columna para distinguir las menciones antes y después de la adquisición de Twitter. A continuación, se calcula la frecuencia de aparición de cada *handle*.

```

data(stop_words)
corpus_menciones <- tweets_user |>
  # Extrae los handles de los comentarios con una expresión regular "@\\w+"
  mutate(mentions = ifelse(tipo_tweet == "reply", str_extract(full_text, "@\\w+"), NA))
  |>
  # Crea una columna para distinguir el periodo antes/después de la compra
  mutate(periodo = ifelse(created_at < compra_twitter,
    "Antes de la compra", "Después de la compra")) |>
  filter(!is.na(mentions)) |>    # elimina las filas vacías
  select(mentions, periodo) |>    # selecciona menciones y periodo
  group_by(mentions, periodo) |>
  summarise(freq = n(), .groups = "drop") |> # calcula freq. de palabra
  ungroup() |>
  arrange(desc(freq)) # ordena de mayor a menor freq. de aparición

# print (corpus_menciones) # Descomentar para ver el resultado final

```

Una vez que los datos han sido procesados, se utiliza la función `geom_text_wordcloud_area()` para generar la nube de palabras correspondiente a las menciones en los tweets de Elon Musk.

Para ello, se toma la lista de menciones y su frecuencia, y se utiliza la misma operación que se realizó con la nube de palabras anterior.

El resultado (Fig. 50.2) muestra que algunos interlocutores se mantienen, otros pierden protagonismo y aparecen otros nuevos. Se mantienen @BillyM2k (comediante) y @WholeMars-Blog (relacionado con temas de Marte). Pierden protagonismo @teslaownersSVm, @EvaFoxU, @PPathole y @Teslarati (relacionados con Tesla). Ganan protagonismo @stillgray (*influencer*), @micsolana (capital riesgo) y @Jason (emprendedor).

```
paleta <- brewer.pal(8, "Dark2")
ggplot() +
  geom_text_wordcloud_area( # dibuja la nube de palabras
    data = corpus_menciones |> top_n(50),
    aes(label = mentions, size = freq, color = freq), angle = 0.35
  ) +
  scale_size_area(max_size = 12) +
  scale_color_gradientn(colors = paleta) +
  facet_wrap(~periodo) +
  theme_minimal() +
  theme(strip.text = element_text(color = "grey50", size = 18))
```



Figura 50.2: Usuarios con los que dialoga Elon Musk antes y después de la compra de Twitter

50.2.3. ¿Cuál es su rutina de publicación?

Para analizar la distribución horaria de los tweets de Elon Musk, se examina la frecuencia de publicación de *tweets* cada hora de cada día. Dado que la residencia declarada de Musk es Austin (Texas), se ajustará la hora de los tweets al huso horario de esta ciudad, ya que la hora proporcionada por Twitter está en GMT.

Debido a que los datos abarcan un período largo, desde junio hasta diciembre, se acotarán a 15 días antes y después de la compra de Twitter. Es importante tener en cuenta que la fecha de creación de los tweets (`created_at`) se presenta en formato fecha-hora, y cada día consta de 86.400 segundos (60 segundos * 60 minutos * 24 horas).

```

tweets_user_hour <- tweets_user |>
  # Cambiamos al huso horario de Texas
  mutate(created_at = lubridate::with_tz(created_at, "US/Central")) |>
  # Filtra los tweets anteriores a la compra de Twitter
  filter(created_at >= (compra_twitter - (60 * 60 * 24 * 15))) |>
  filter(created_at <= (compra_twitter + (60 * 60 * 24 * 15))) |>
  # Creamos una nueva columna para la fecha
  mutate(time_in_days = as.POSIXct(floor_date(created_at, "day"))) |>
  # Creamos una nueva columna para la hora
  mutate(hour_tweet = hour(created_at)) |>
  # Agrupamos el número de tweets por tipo y hora
  group_by(time_in_days, hour_tweet, tipo_tweet) |>
  # Calculamos el número de tweets por día, hora y tipo
  summarise( num_tweets = n(), .groups = "drop" ) |>
  ungroup()

# print (tweets_user_hour) # Descomentar para ver el resultado final

```

A continuación, se recalcan los días de la semana que son festivos en color rojo para apreciar si hay distinta rutina.

```

festivos <- tweets_user |>
  # Cambiamos al huso horario de Texas
  mutate(created_at = lubridate::with_tz(created_at, "US/Central")) |>
  # Filtramos los tweets anteriores a la compra de Twitter
  filter(created_at >= (compra_twitter - (60 * 60 * 24 * 15))) |>
  filter(created_at <= (compra_twitter + (60 * 60 * 24 * 15))) |>
  # Creamos una columna con el tiempo en días
  mutate(time_in_days = floor_date(created_at, "1 day")) |>
  # Agrupamos por día
  group_by(time_in_days) |>
  # calculamos el número de tweets por día
  summarise( num_tweets = n(), .groups = "drop" ) |>
  ungroup() |>
  # Creamos una columna con el día de la semana
  mutate(week_day = wday(time_in_days)) |>
  # Creamos una columna para colorear los días según sean festivos o no
  mutate(festivo = ifelse(wday(time_in_days) == 7 |
    (wday(time_in_days) == 1), "red", "black"))

# print (festivos) # Descomentar para ver el resultado final

```

Finalmente, se representa un gráfico de dispersión (*scatter plot*) con las coordenadas de las horas del día (eje X) y los días seleccionados (eje Y), utilizando la función `geom_point()`. El tamaño del punto es proporcional al número de *tweets* en esa hora y día, y el color el tipo de tweet (*original, reply, quoted* y *retweeted*). Se marca una línea horizontal con la función `geom_hline()` en la fecha de compra de Twitter y se crea un eje X doble para que sea más fácil ver las horas debido a la altura de la gráfica.

50.2. Análisis visual de los tweets de Elon Musk

829

La Fig. 50.3 muestra que no hay una rutina clara en la publicación de tweets de Elon Musk. Esto podría deberse a que viaja mucho. La mayoría de sus mensajes son comentarios y han aumentado considerablemente desde la compra de Twitter. El máximo número de *tweets* por hora fue 10.

```
ggplot() +
  geom_point(
    data = tweets_user_hour,
    aes(
      x = hour_tweet,
      y = time_in_days,
      size = num_tweets,
      color = tipo_tweet
    ),
    alpha = 0.5
  ) +
  # separa las fechas antes y después de la compra
  geom_hline(aes(yintercept = compra_twitter), linetype = 2) +
  # define una etiqueta de tiempo por día
  scale_y_datetime(
    date_labels = "%d-%b-%y(%a)", # formato fecha (día semana abreviado)
    date_breaks = "1 day", # una marca de tiempo cada día
    expand = c(0, 0, 0.02, 0.02)
  ) + # ajustes de márgenes
  # Definimos una etiqueta para cada hora
  scale_x_continuous(
    breaks = seq(0, 23, 1), # crea un vector de 0 a 23
    sec.axis = dup_axis() # duplica el eje X
  ) +
  labs( x = "", y = "", color = "", size = "N. tweets") +
  # ajusta las leyendas en dos filas para que no se trunquen
  guides(color = guide_legend(nrow = 2, override.aes = list(size = 4))) +
  theme_minimal() +
  # indica la posición de la leyenda y el color de las fechas
  theme(
    panel.grid.major.x = element_line(),
    legend.position = "top",
    axis.text.y = element_text(colour = festivos$festivo)
  )
```

50.2.4. ¿Cuál es su *timeline* de publicación?

Ahora se analiza cómo se distribuyen los *tweets* en el tiempo por tipo de tweet. Se resaltará la fecha de compra de Twitter con una anotación para facilitar la comparación de la frecuencia anterior y posterior a esta fecha.

Se crea una columna con la fecha redondeada a días, se agrupan los *tweets* por fecha y el tipo de tweet y se calcula su número para cada día.

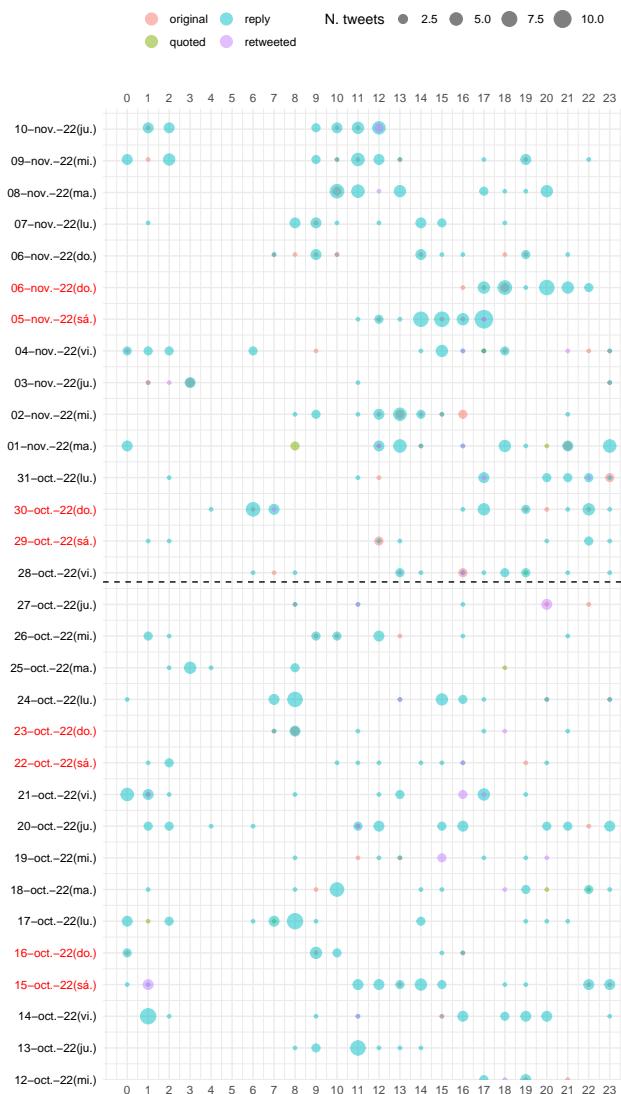


Figura 50.3: Rutina de publicación de Elon Musk. (huso horario de Texas)

50.2. Análisis visual de los tweets de Elon Musk

831

```

tweets_user_day <- tweets_user |>
# Creamos una columna con el tiempo en días
mutate(time_in_days = floor_date(created_at, "1 day")) |>
# Agrupamos el número de tweets por día y tipo
group_by(time_in_days, tipo_tweet) |>
# calculamos el número de tweets por día y tipo
summarise( num_tweets = n(), .groups = "drop" ) |>
ungroup()

# print (tweets_user_day) # descomentar para ver el resultado

```

En la Fig. 50.4 se puede observar un incremento en el número de publicaciones después de la compra de Twitter. De hecho, se publicaron casi el doble de tweets en comparación con el periodo anterior a la adquisición de la plataforma. Asimismo, se puede ver, al igual que en la Fig. 50.3, que la mayoría de los *tweets* de Elon Musk fueron comentarios.

```

ggplot(data = tweets_user_day) +
  geom_col(aes(x = time_in_days, y = num_tweets, fill = tipo_tweet), alpha = 0.7 ) +
  geom_vline(aes(xintercept = compra_twitter), linetype = 2) + # compra de Twitter
  geom_label( # señala el evento
    aes(
      x = compra_twitter - (60 * 60 * 24 * 25),
      y = max(num_tweets),
      label = "Elon Musk\ncompra Twitter"
    ),
    color = "gray45"
  ) +
  geom_curve( # flecha con curva para señalar el evento
    aes(
      x = compra_twitter - (60 * 60 * 24 * 10),
      y = max(num_tweets),
      xend = compra_twitter,
      yend = max(num_tweets) * 0.80
    ),
    arrow = arrow(length = unit(0.08, "inch")), linewidth = 0.5,
    color = "gray20", curvature = -0.3
  ) +
  scale_x_datetime( # ajusta la escala de tiempo y su formato
    date_labels = "%d\n%b",
    date_breaks = "2 week"
  ) +
  scale_y_continuous( # ajusta el formato del eje Y
    name = "Num. Tweets por día",
    labels = label_number(scale_cut = cut_short_scale())
  ) +
  scale_color_manual(values = my_color) + # aplica colores definidos
  labs( x = "", y = "Num. Tweets por día", fill = "" ) +
  theme_minimal() +
  theme(legend.position = "top")

```

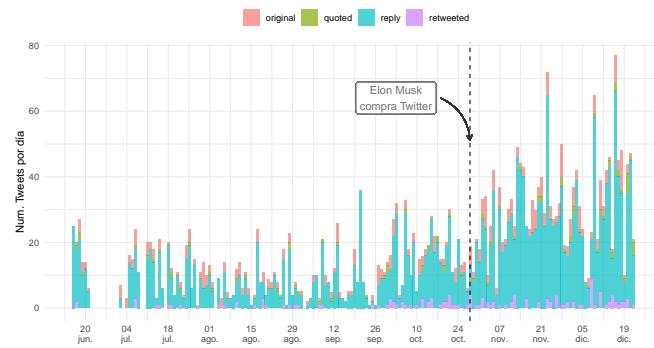


Figura 50.4: Publicación de Tweets por día de Elon Musk

50.2.5. ¿Cuál es el impacto de sus tweets?

Para comparar los *tweets* propios publicados (sin *retweets*) y el impacto que reciben (*retweets* recibidos), se utilizará una gráfica de doble escala. Dado que ambas variables tienen diferentes órdenes de magnitud, este tipo de gráfica permitirá una mejor comparación. Además, se incluirá una anotación con la fecha de compra de Twitter para distinguir los cambios antes y después de este evento.

En esta gráfica se podrá ver cómo se van superponiendo capas de dibujo.

Se preparan los datos en dos `data.frames` y se calcula la relación de escala:

- `tweets_propios_day` con los *tweets* propios por día y los mensajes originales/hora:

```
tweets_propios_day <- tweets_user |>
  # Creamos una columna con el tiempo en días
  mutate(time_in_days = floor_date(created_at, "1 day")) |>
  filter(tipo_tweet != "RT") |>  # elimina los retweets
  group_by(time_in_days) |>  # agrupa los tweets por día
  summarise( num_tweets = n(), .groups = "drop" ) |>
  ungroup()

# print (tweets_propios_day) # Descomentar para ver el resultado
```

- `tweets_RT_day` con los *retweets* recibidos por día:

50.2. Análisis visual de los tweets de Elon Musk

833

```

tweets_RT_day <- tweets_user |>
  # Creamos una columna con el tiempo en días
  mutate(time_in_days = floor_date(created_at, "1 day")) |>
  filter(tipo_tweet != "RT") |>
  group_by(time_in_days, tipo_tweet) |>
  summarise( num_tweets = sum(retweet_count), .groups = "drop" ) |>
  ungroup()

# print(tweets_RT_day) # descomentar para ver el resultado

```

- Se calculan las escalas:

```

# Máximo número de tweets propios
max_tweets <- max(tweets_propios_day$num_tweets, na.rm = TRUE)
# Máximo número de retweets recibidos
max_RT <- max(tweets_RT_day$num_tweets, na.rm = TRUE)
ajuste_escala <- max_RT / max_tweets # Ajuste de escala
# print (ajuste_escala) # descomentar para ver el ajuste
my_color <- c("Num. original tweets" = "steelblue4", "RTs" = "red4")

```

La Fig. 50.5 muestra un incremento masivo de los *retweets* recibidos desde la compra de Twitter, siendo el día que tomó posesión, el que generó el mayor pico: 800K RTs.

```

ggplot() +
  # Pinta la evolución de los tweets propios/día
  # Pinta el área que representa los tweets propios por día
  geom_area(
    data = tweets_propios_day,
    aes(x = time_in_days, y = num_tweets), fill = "steelblue4",
    alpha = 0.5
  ) +
  # Pinta el borde del área por estética
  geom_line( data = tweets_propios_day,
    aes(x = time_in_days, y = num_tweets, color = "Num. original tweets"))
  ) +
  # Pinta la evolución de los RTs/día
  geom_line(
    data = tweets_RT_day,
    aes(x = time_in_days, y = num_tweets / ajuste_escala, color = "RTs"))
  ) +
  # Marcamos la linea de la compra de Twitter por Elon Musk
  geom_vline(aes(xintercept = compra_twitter), linetype = 2) +
  # Anotamos el evento
  geom_label(
    data = tweets_propios_day,
    aes(
      x = compra_twitter - (60 * 60 * 24 * 25),

```

```

    y = max(num_tweets) * .95,
    label = "Elon Musk\ncompra Twitter"),
    color = "grey50"
) +
# Dibuja una flecha con curva para señalar el evento
geom_curve(
  data = tweets_propios_day,
  aes(
    x = compra_twitter - (60 * 60 * 24 * 10),
    y = max(num_tweets),
    xend = compra_twitter,
    yend = max(num_tweets) * 0.90
  ),
  arrow = arrow(length = unit(0.08, "inch")), linewidth = 0.5,
  color = "gray20", curvature = -0.3
) +
# Ajusta la escala de tiempo y su formato
scale_x_datetime(
  date_labels = "%d\n%b",
  date_breaks = "2 week"
) +
# doble escala, derecha: tweets propios, izquierda: retweets
scale_y_continuous(
  name = "Num. Original tweets por día",
  labels = label_number(scale_cut = cut_short_scale()),
  sec.axis = sec_axis(
    trans = (~ .* ajuste_escala), name = "RTs por día",
    labels = label_number(scale_cut = cut_short_scale())
  )
) +
scale_color_manual(values = my_color) +
labs(x = "", color = "") +
theme_minimal(base_family = "sans") +
theme(
  legend.position = "top",
  axis.title.y = element_text(color = "steelblue4", size = 12),
  axis.title.y.right = element_text(color = "red4", size = 12),
  axis.text.y = element_text(color = "steelblue4"),
  axis.text.y.right = element_text(color = "red4")
)

```

50.2. Análisis visual de los tweets de Elon Musk

835

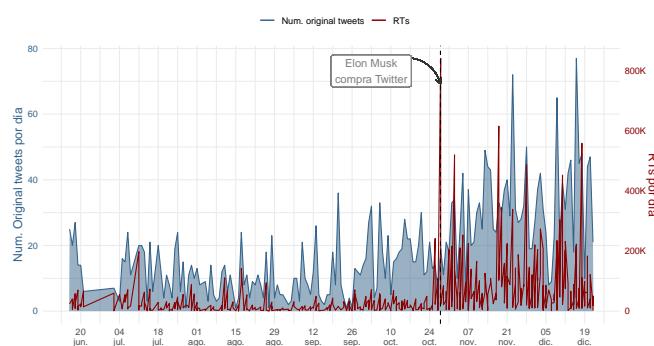


Figura 50.5: Tweets vs. retweets de Elon Musk

Capítulo 51

Análisis electoral: de Rstudio a su periódico

Borja Andrino Turón

EL PAÍS

51.1. Motivación

El uso de **R** en el entorno profesional ha llegado también a los periódicos. Cada vez es más habitual encontrar en los medios analistas de datos que lo utilizan en su día a día. En EL PAÍS, muchos de los contenidos que se publican en la Unidad de Datos surgen de un notebook de RStudio. A continuación, se muestra un análisis sobre las últimas elecciones andaluzas, de RStudio a su periódico favorito.

51.2. Obtención de los datos

Los datos electorales no siempre son igual de accesibles. Los de las elecciones que dependen del Ministerio del Interior se publican en el portal [Infoelectoral](#). En el caso de las elecciones andaluzas, los resultados a nivel de mesa se han publicado en los portales de cada convocatoria, aunque pueden encontrarse entre los contenidos del libro para replicar estos análisis.

En primer lugar se compondrá un diccionario de municipios que se usará para filtrar y agrupar los resultados por provincia. Primero se escraperá de la web del INE la relación de códigos de provincia con la librería rvest. Se lee el código html de la página y se buscan los elementos table con clase miTabla. A continuación, se usa la función html_table para convertir las tres tablas en un objeto tibble. La información con los nombres de municipios y provincias se leerá en la web del INE.

```

pacman::p_load(CDR, ggplot2, dplyr, rvest, lubridate, sf, ggtext,
                rio, janitor, here, purrr, stringr, scales)

url_provincias <-
  "https://www.ine.es/daco/daco42/codmun/cod_provincia.htm"

cod_provincias <-
  read_html(url_provincias) |>
  html_nodes("table.miTabla") |>
  html_table() |>
  map_df(as_tibble) |>
  rename(codigo_prov = 1, name_prov = 2) |>
  mutate(codigo_prov = str_pad(codigo_prov, width = 2, pad = "0", side = "left"))

url_municipios <-
  "https://www.ine.es/daco/daco42/codmun/codmun20/20codmun.xlsx"

cod_municipios <-
  import(url_municipios, skip = 1) |>
  clean_names() |>
  transmute(codigo_prov = cpro,
            codigo_mun = str_glue("{codigo_prov}{cmun}"),
            name_mun = nombre) |>
  left_join(cod_provincias) |>
  select(codigo_mun, name_mun, name_prov)

```

Se añade la información sobre municipios al dataset de elecciones que tiene los datos de cada sección censal.

```

datos_elecciones <-
  datos_elecciones |>
  left_join(cod_municipios) |>
  select(codigo_secc, codigo_mun, name_mun, name_prov, convocatoria, everything())

```

51.3. Transformación y primeros gráficos

En el primer gráfico se mostrará la evolución de los votos a partidos de izquierda y de derecha en toda Andalucía desde 2015. Primero se calculan los votos válidos en cada convocatoria. Como en la estructura de datos ese dato está repetido para cada combinación de convocatoria-sección-partido se usará la función distinct antes de agrupar y sumar los votos validos de todas las secciones.

```

datos_elecciones_validos_total <-
  datos_elecciones |>
  distinct(convocatoria, codigo_secc, .keep_all = T) |>

```

51.3. Transformación y primeros gráficos

839

```

  mutate(region = "Andalucía") |>
  group_by(convocatoria, region) |>
  summarise(validos = sum(validos), .groups = "drop")

datos_elecciones_validos_provs <-
  datos_elecciones |>
  distinct(convocatoria, codigo_secc, .keep_all = T) |>
  mutate(region = name_prov) |>
  group_by(convocatoria, region) |>
  summarise(validos = sum(validos), .groups = "drop")

datos_elecciones_validos <-
  datos_elecciones_validos_total |>
  bind_rows(datos_elecciones_validos_provs)

```

Ahora se calcula la suma de votos de cada bloque en cada convocatoria. En este caso, como cada fila tiene el dato de votos de un partido distinto no es necesaria la función distinct.

```

datos_bloques_total <-
  datos_elecciones |>
  mutate(region = "Andalucía") |>
  group_by(convocatoria, region, bloque) |>
  summarise(votos_bloque = sum(votos_partido), .groups = "drop")

datos_bloques_provs <-
  datos_elecciones |>
  mutate(region = name_prov) |>
  group_by(convocatoria, region, bloque) |>
  summarise(votos_bloque = sum(votos_partido), .groups = "drop")

datos_bloques <-
  datos_bloques_total |>
  bind_rows(datos_bloques_provs) |>
  left_join(datos_elecciones_validos) |>
  mutate(votos_bloque_pc = votos_bloque / validos)

```

A continuación, se realiza el gráfico con los datos que se han calculado antes. Se definen los colores que representan a cada bloque, las fechas para poder etiquetar en el gráfico los ticks del eje x y se programa el gráfico.

```

title <-
  "Evolución de voto a partidos de <b style='color:#457b9d;'>derecha</b>, <b
  ↵  style='color:#e63946;'>izquierda</b><br>y <b style='color:#676767;'>otros</b>
  ↵  desde 2015"

convocatorias_dates <-
  datos_bloques |>

```

```

distinct(convocatoria) |>
  pull(convocatoria)

datos_blocques |>
  filter(region == "Andalucía") |>
  ggplot(aes(x = convocatoria, y = votos_bloque_pc,
             color = bloque, group = bloque)) +
  geom_line(size = 1) +
  geom_point(size = 2) +
  geom_hline(yintercept = 0, width = 0.2) +
  scale_color_manual(values = colors_blocques) +
  scale_x_date(breaks = convocatorias_dates, date_labels = "%Y") +
  scale_y_continuous(labels = percent) +
  labs(title = title, x = "Convocatoria", y = "Votos bloque",
       caption = "Fuente: Junta de Andalucía") +
  theme_minimal() +
  theme(legend.position = "none",
        plot.title = element_markdown(margin=margin(0,0,30,0)))

```

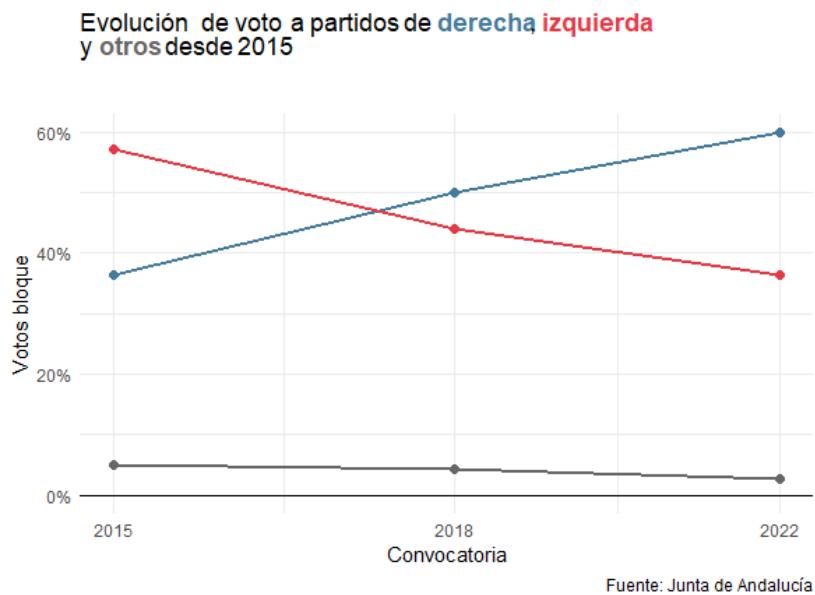


Figura 51.1: Evolución del voto en Andalucía

Replicar la Fig. 51.1 para cada provincia no es complicado. Sólo se descartarán los datos de toda Andalucía y se usará la función `facet_wrap()` que realizará el mismo gráfico con el mismo estilo para cada provincia.

51.3. Transformación y primeros gráficos

841

```
datos_bloques |>
  filter(region != "Andalucía") |>
  ggplot(aes(x = convocatoria, y = votos_bloque_pc,
             color = bloque, group = bloque)) +
  geom_line(size = 1) +
  geom_point(size = 2) +
  geom_hline(yintercept = 0, width = 0.2) +
  scale_color_manual(values = colors_bloques) +
  scale_x_date(breaks = convocatorias_dates, date_labels = "%Y") +
  scale_y_continuous(labels = percent) +
  labs(title = title, x = "Convocatoria", y = "Votos bloque",
       caption = "Fuente: Junta de Andalucía") +
  facet_wrap(~region, ncol = 4) +
  theme_minimal() +
  theme(legend.position = "none",
        plot.title = element_markdown(margin=margin(0,0,10,0)))
```

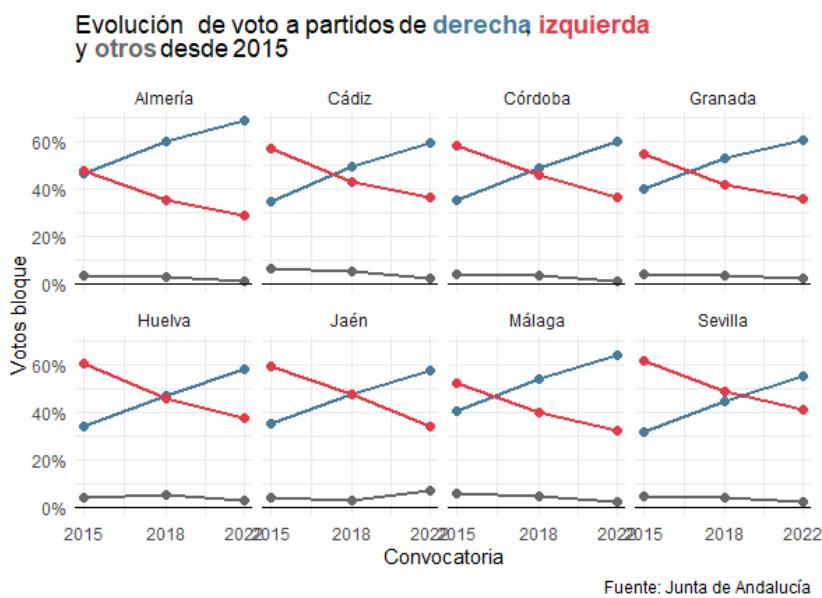


Figura 51.2: Evolución del voto provincial

La Fig. 51.2 cuenta una historia complementaria al primero. El giro no se ha producido igual en toda Andalucía, no es igual el de Almería que el de Sevilla. Para intentar buscar nuevas diferencias territoriales se explorarán los mapas de ganadores a nivel municipal. Se procede de igual manera que con los datos de provincias, salvo que en este caso se agrega a partir de la columna `codigo_mun`. Para calcular el ganador se agrupa por esta columna y se usa la función `slice_max()`, que tomará para cada municipio la fila del partido con el mayor número de votos.

```

datos_elecciones_validos_muns <-
  datos_elecciones |>
  distinct(convocatoria, codigo_secc, .keep_all = T) |>
  group_by(convocatoria, codigo_mun) |>
  summarise(validos = sum(validos),
             .groups = "drop")

datos_bloques_muns <-
  datos_elecciones |>
  group_by(convocatoria, codigo_mun, bloque) |>
  summarise(votos_bloque = sum(votos_partido), .groups = "drop") |>
  left_join(datos_elecciones_validos_muns) |>
  mutate(votos_bloque_pc = votos_bloque / validos, 1)

# Ahora calculamos los ganadores
datos_winners_muns <-
  datos_bloques_muns |>
  group_by(convocatoria, codigo_mun) |>
  slice_max(votos_bloque, n = 1, with_ties = F) |>
  select(convocatoria, codigo_mun,
         winner = bloque, votos_bloque_pc)

```

Para realizar el gráfico se tomará el objeto `sf` con los recintos de los municipios andaluces y se les añadirá los datos de ganadores calculados anteriormente con la función `left_join()`. Se usa el color del bloque para el relleno y el porcentaje de votos que suma el bloque ganador para la transparencia, de forma que de un vistazo se pueden encontrar feudos de uno u otro bloque.

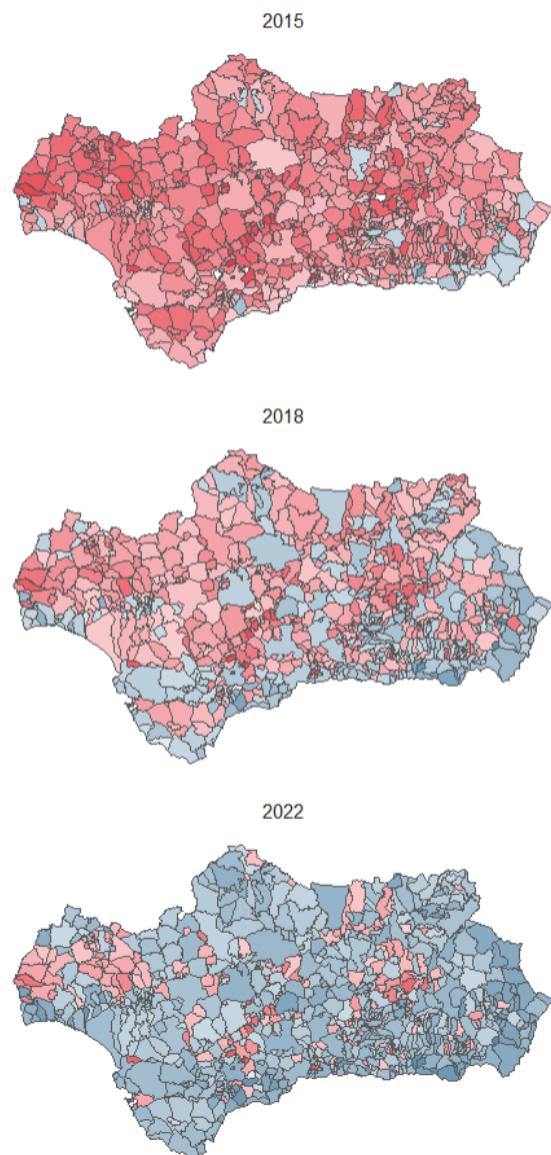
```

map_munis |>
  left_join(datos_winners_muns) |>
  mutate(convocatoria = year(convocatoria)) |>
  ggplot() +
  geom_sf(aes(fill = winner, alpha = votos_bloque_pc),
         size = 0.01) +
  scale_fill_manual(values = colors_bloques) +
  facet_wrap(~convocatoria, ncol = 1) +
  labs(title = title,
       caption = "Fuente: Junta de Andalucía") +
  coord_sf(label_graticule = "", ndiscr=0) +
  theme_minimal() +
  theme(legend.position = "none",
        plot.title = element_markdown(margin=margin(0,0,10,0)))

```

En los mapas se encuentran nuevas historias. En 2015 la derecha era fuerte en la costa de Almería y Málaga. Su presencia creció en 2018, aunque la izquierdas seguía ganando el interior de la comunidad. En 2018 el dominio del bloque de derechas se extiende por casi todo el territorio, en especial en las zonas donde ya era fuerte en 2015.

Evolución de voto a partidos de **derecha**, **izquierda** y otros desde 2015



Fuente: Junta de Andalucía

Figura 51.3: Resultados de las elecciones andaluzas

Capítulo 52

Crisis: impacto en el paro de Castilla-La Mancha

Isidro Hidalgo Arellano^a y Ángel Jiménez Rojas^a

^aObservatorio del Mercado de Trabajo de Castilla-La Mancha

52.1. Planteamiento

En los últimos 15 años el mundo ha sufrido dos grandes períodos de **crisis económica**: en **2008**, de tipo financiero; y en **2020**, a causa de la pandemia de **COVID-19**. Uno de los parámetros socioeconómicos que se ven más afectados por este tipo de procesos es el paro registrado. El paro registrado se define como el conjunto de los demandantes inscritos en las oficinas de empleo, una vez excluidos los inscritos sin disponibilidad para trabajar y los demandantes no parados, tales como estudiantes, desempleados en formación, etc. (Toharia, 2012). **Castilla-La Mancha**, comunidad autónoma interior de España, no ha sido ajena a las crisis económicas mencionadas, por lo que en este trabajo se quiere analizar el impacto de las mismas en la estructura del **paro registrado** de la región. Para ello, se utilizan las siguientes variables explicativas: **sexo** y **edad** de la persona desempleada, **sector de actividad económica de procedencia** y **tiempo de búsqueda de empleo**. El conjunto de datos utilizado comprende la **media anual del paro registrado en la comunidad autónoma de Castilla-La Mancha** desagregado según estas variables, a lo largo de los años que van desde 2007 a 2022.

Para el análisis se usan las librerías y objetos (paletas de colores para los gráficos) siguientes:

```
library("CDR")
library("tidyverse")
library("ggpubr")
paleta_heatmaps <- c(rgb(.7,.1,.0,.5),   rgb(.13,.22,.58,1))
paleta_lineas <- c("blue4", "orange", "darkgreen")
```

Para cargar el conjunto de datos, `parados_clm`, incluido en el paquete CDR, y mostrar la estructura de la `tibble` se usa:

```
data("parados_clm")
parados_clm
# A tibble: 92,215 × 8
#   anyo    sexo    edad sector t_bus_e    tramo_edad t_bus_e_agr parados
# <dbl> <fct> <dbl> <fct> <dbl> <fct> <dbl> <dbl>
# 2007 hombre 16 agricu t<=7 días <30 años t<=6 meses 0.66666667
# 2018 mujer 36 sinact t<=7 días 30-44 años t<=6 meses 1.66666667
# 2012 mujer 30 agricu t<=7 días 30-44 años t<=6 meses 5.33333333
# 2022 mujer 49 constr t<=7 días >44 años t<=6 meses 0.75000000
# 2007 mujer 54 indust t<=7 días >44 años t<=6 meses 1.50000000
# ... with 92,210 more rows
```

52.2. Evolución del paro medio anual en Castilla-La Mancha

Para ver el paro medio anual en función del tiempo, se construye un gráfico de evolución. Para ello, representamos el paro medio por año, marcando los años que suponen un máximo o mínimo en la serie:

```
resumen <- parados_clm |>
  group_by(anyo) |>
  summarise(parados = sum(parados)) |>
  mutate(anyo = as.numeric(as.character(anyo)))
anyos <- c(2007, 2013, 2019, 2020, 2022)
paro_anyos <- resumen |>
  filter(anyo %in% anyos) |>
  select(parados) |>
  mutate(parados = round(parados, 0))
puntos <- data.frame(anyos, paro_anyos)

graf <- ggplot(resumen, aes(anyo, parados)) +
  geom_line(linewidth = 2, col = paleta_lineas[1], alpha = 0.5) +
  xlab("") + ylab("número de parados") +
  geom_point(puntos, mapping = aes(x = anyos, y = parados,
    shape = "circle filled", size = 1, fill = paleta_lineas[1],
    alpha = 0.5)) +
  theme(legend.position = "none",
    axis.title = element_text(face="bold", size = 10),
    axis.text = element_text(face="bold", size = 10),
    strip.text = element_text(size = 9, face = "bold")) +
  scale_y_continuous(labels = function(x) format(x, big.mark = ".",
    scientific = FALSE))
graf
```

52.3. Evolución del paro medio anual en función de la edad y el sexo

847

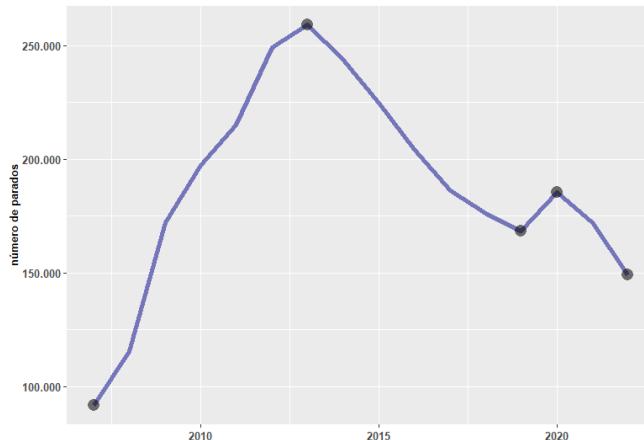


Figura 52.1: Evolución del paro medio anual en CLM

De un primer análisis visual de la Fig. 52.1 se toman como puntos de referencia los años previos a las crisis: 2007 y 2019, y el último año, 2022. Se puede observar que, si bien la crisis de la **COVID-19** ha tenido profundos efectos sectoriales, principalmente en turismo, comercio y restauración, la **crisis de 2008** tuvo un impacto enorme y generalizado en toda la economía, por lo que su efecto en el paro registrado fue devastador, multiplicando por un factor mayor de 3 la cifra total de paro en la región desde 2007. Sin embargo, a partir del año 2013 el paro registrado inicia una tendencia a la baja muy pronunciada que aún hoy continúa, después de haber repuntado ligeramente por la crisis de la COVID-19.

52.3. Evolución del paro medio anual en función de la edad y el sexo

Para ver cómo ha cambiado la estructura del paro registrado en función de la **edad** y el **sexo** de los parados se pueden utilizar diferentes gráficos. En este análisis, se usan mapas de calor y gráficos de distribución de densidad. Para hacer un mapa de calor que permita comparar dos variables simultáneamente, se construye la siguiente función:

```
heatmap_anyos <- function(var1, var2, inicio = 2007, intermedio = 2019,
                           fin = 2022){
  tabla <- select(parados_clm, anyo, var1, var2, parados) |>
    filter(anyo %in% c(inicio, intermedio, fin))
  names(tabla) <- c("anyo", "var1", "var2", "parados")
  tabla <- tabla |>
    group_by(anyo, var1, var2) |>
    summarise(parados = sum(parados))
  graf <- ggplot(tabla, aes(x = var1, y= var2, fill = parados)) +
```

```
geom_raster() +
scale_fill_gradientn(colours = paleta_heatmaps) +
facet_wrap(~ anyo) +
labs(x = "", y = "") +
theme(axis.text = element_text(size = 10, face = "bold"),
      axis.title = element_text(size = 10, face = "bold"),
      strip.text = element_text(size = 10, face = "bold"))
return(graf)}
```

Si se lanza la función `heatmap_anyos()` para las variables `edad` y `sexo`, tomando como años comparativos 2007, 2019 y 2022, se obtiene:

```
heatmap_anyos("sexo", "edad")
```

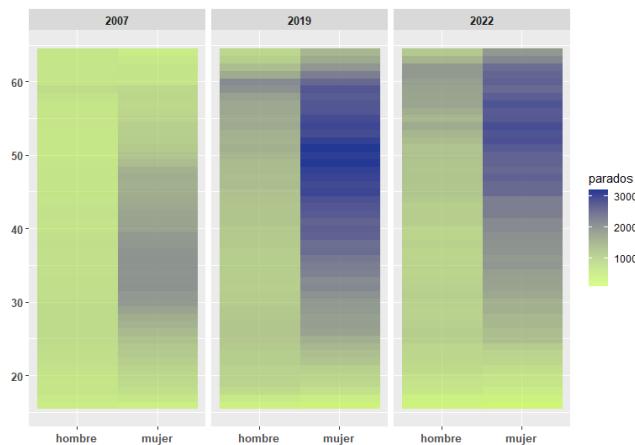


Figura 52.2: Paro medio anual según edad y sexo en 2007, 2019 y 2022

En la Fig. 52.2 se puede apreciar que en los dos procesos críticos se ha producido un **desplazamiento del paro hacia los intervalos de mayor edad**, siendo este cambio más pronunciado en las **mujeres**.

El mapa de calor es muy útil para una primera impresión de estos cambios, pero si se desea observar detalladamente cómo ha cambiado la distribución del paro según el `sexo` y la `edad`, es mejor programar la función `densidad_compara()`, que proporciona mayor nivel de detalle: produce un cuadro de gráficos comparando la distribución de la edad, para cada estrato de la variable elegida, para tres años diferentes (2007, 2019 y 2022 por defecto). Los parámetros `alpha` y `size` permiten ajustar tamaño y opacidad de las líneas, mejorando la apariencia general del gráfico.

```
densidad_compara <- function(variab, inicio = 2007, medio = 2019,
                                fin = 2022){
```

52.3. Evolución del paro medio anual en función de la edad y el sexo

849

```

tabla <- select(parados_clm, anyo, variab, edad, parados) |>
  filter(anyo %in% c(inicio, medio, fin))
names(tabla) <- c("anyo", "variable", "edad", "parados")
tabla <- tabla |>
  group_by(anyo, edad, variable) |>
  summarise(parados = sum(parados))# />

graf <- ggplot(tabla, aes(x = edad, y = parados, color = anyo,
                           fill = anyo)) + geom_line(alpha=0.6, size = 1) +
  facet_wrap(~ variable, ncol = dim(table(tabla$variable))[1]) +
  ylab("número de parados") + labs(color="año") +
  scale_color_manual(values = paleta_lineas) +
  scale_y_continuous(labels = function(x) format(x,
                                                 big.mark = ".",
                                                 scientific = FALSE)) +
  theme(strip.text = element_text(size = 10, face = "bold"),
        axis.title = element_text(size = 10, face = "bold"),
        axis.text = element_text(size = 10, face = "bold"))
return(graf)

```

Ejecutando esta función para la variable `sexo` se obtiene:

```
densidad_compara("sexo")
```

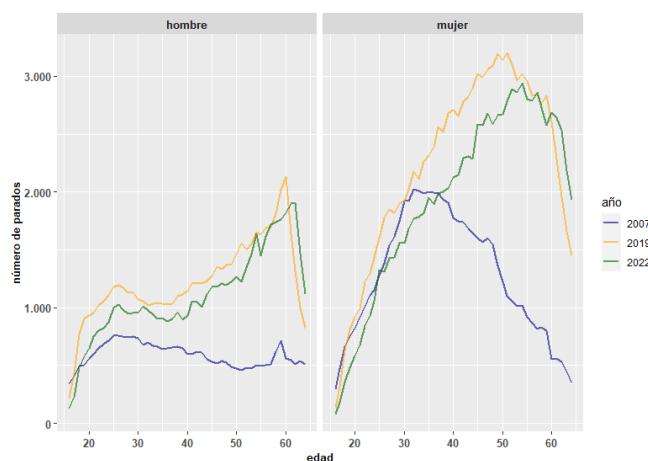


Figura 52.3: Distribución del paro medio anual por edad y sexo (2007, 2019 y 2022)

En la Fig. 52.3 se observa que en 2007, antes de ambas crisis, los hombres parados presentan **dos máximos**, en torno a 25 y 60 años, mientras que las mujeres desempleadas tienen una distribución bastante centrada entre 30 y 40 años. En cambio, en 2022 se aprecia el desplazamiento de la distribución de los parados de ambos sexos hacia los estratos de edad **mayores**

de 50 años. Este desplazamiento es algo más intenso en las mujeres.

Se observa igualmente que comparando las distribuciones de las mujeres de 2019 y 2022, la crisis de la COVID-19 ha incrementado entre 5 y 10 años la distribución de la edad de las mujeres paradas. Este desplazamiento es inferior en los hombres, donde supone menos de 5 años.

52.4. Evolución del paro medio anual según el tiempo de búsqueda de empleo

Se define el **tiempo de búsqueda de empleo** como el tiempo transcurrido ininterrumpidamente desde la última inscripción de la persona en el paro registrado (Pérez Infante, 2006).

Si, para simplificar, se agregan los doce intervalos que considera la estadística de paro registrado para el tiempo de búsqueda de empleo en tan solo cuatro, ejecutando la función `densidad_compara()` para la variable `t_bus_e_agr` se obtiene:

```
densidad_compara("t_bus_e_agr")
```



Figura 52.4: Distribución del paro medio anual por edad y tiempo de búsqueda de empleo

En la Fig. 52.4 se pone de manifiesto que el tramo con mayor incremento de número de parados es el correspondiente a más de 24 meses de búsqueda de empleo (**paro de muy larga duración**), ya que la crisis financiera de 2008 les redujo su probabilidad de encontrar empleo. Se puede afirmar también que los dos períodos de crisis han provocado la creación de un **paro estructural de larga duración**.

Se deja al lector ejecutar la función `heatmap_anyos()` para las variables `sexo` y `t_bus_e`, tomando como años comparativos 2007, 2019 y 2022. Observará en el gráfico resultante que el incremento en el paro de muy larga duración es más intenso en el colectivo de las mujeres. El código a utilizar es:

52.5. Evolución del paro medio anual según sexo, edad y sector de procedencia 851

```
heatmap_anyos("sexo", "t_bus_e")
```

52.5. Evolución del paro medio anual según sexo, edad y sector de procedencia

La variable **sector de procedencia** es un tanto particular, ya que, cuando un parado lleva mucho tiempo buscando empleo ininterrumpidamente, “ pierde” el sector de procedencia y se clasifica automáticamente en la rúbrica “sin actividad”. A la hora de analizar esta variable, por tanto, es importante tener en cuenta que una parte de los parados ubicados en la rúbrica “sin actividad”, realmente tuvieron un trabajo hace mucho tiempo.

La visualización de los cambios producidos en estas variables con un mapa de calor, se puede llevar a cabo ejecutando de nuevo la función `heatmap_anyos()` obteniendo la Fig. 52.5:

```
heatmap_anyos("sexo", "sector")
```

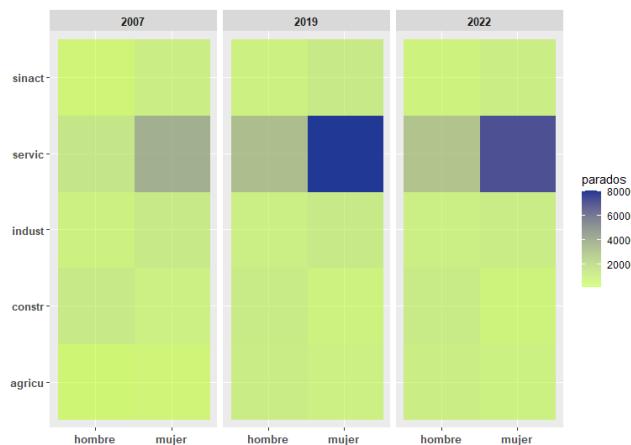


Figura 52.5: Paro medio anual según sexo y sector de procedencia

En la Fig. 52.5 se aprecia el incremento del paro registrado en el sector **servicios**, especialmente en el colectivo femenino.

Ejecutando la función `densidad_compara()` para la variable **sector** se obtiene:

```
densidad_compara("sector")
```

Como se observa en la Fig. 52.6, las diferencias a lo largo del tiempo del número de parados por sector de actividad económica revelan algunas particularidades interesantes. **Industria** y **construcción** se comportan de modo similar: hay un fuerte desplazamiento en edad desde 2007,

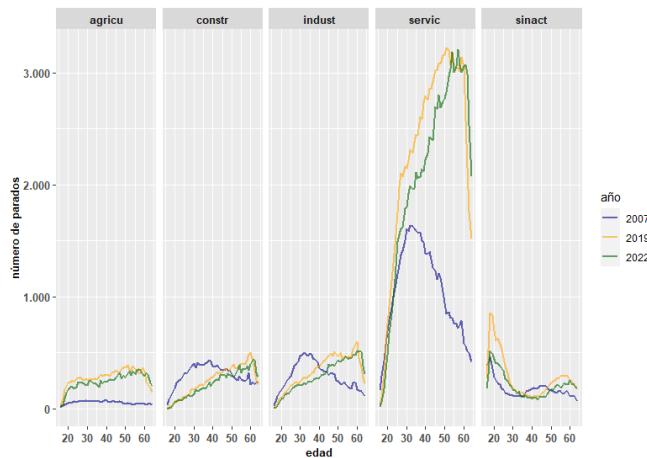


Figura 52.6: Distribución del paro medio anual por edad y tiempo de búsqueda de empleo

pero se mantiene el volumen de paro en ambos sectores a lo largo de los 15 años de estudio. El paro en el sector **agropecuario** y en el sector **servicios** también presenta desplazamiento en edad, pero además se ha incrementado notablemente en estos 15 años; ambos efectos son mucho más evidentes en el sector **servicios**. Finalmente, en el colectivo **sin actividad** se aprecian dos características: en primer lugar, los parados menores de 30 años suponen el mayor volumen en este colectivo, como era de esperar, ya que la población joven que accede al mercado laboral por primera vez, no cuenta con experiencia previa; en segundo lugar, desde 2007 a 2019, y algo menos desde 2019 a 2022, hay un incremento de volumen de paro en los **mayores de 45 años** que, con toda probabilidad, corresponde a los parados de larga duración de mayor edad.

En todos los sectores se aprecia el descenso del volumen total de paro registrado desde 2019 a 2022, a pesar de la crisis sanitaria de la COVID-19.

52.6. Conclusiones

La crisis de 2008 tuvo un gran impacto en el paro registrado de Castilla-La Mancha, multiplicándolo por un factor mayor de 3 desde 2007. Sin embargo, a partir del año 2013 el paro registrado inicia una tendencia a la baja muy pronunciada que aún hoy continúa, después de haber sufrido un rebote debido a la crisis de la COVID-19.

La estructura interna de la población parada en la región ha cambiado sustancialmente atendiendo a las variables analizadas. En efecto, la población mayor de 45 años, las mujeres, los parados de larga duración y el sector servicios son los grandes perjudicados por ambos procesos de crisis.

Capítulo 53

Segmentación de clientes en el comercio minorista

Jaime Fierro Martín^a, Rocío González Martínez^a y Cristina Sánchez Figueroa^b

^aAnalyticae, SL, ^bUniversidad Nacional a Distancia

53.1. Motivación y conceptos clave

Los comercios minoristas (*retailers*) se mueven en un entorno turbulento y necesitan acercarse a sus clientes para asegurar su supervivencia. Su producto, o servicio, es nexo clave en dicho proceso. En este contexto, conocer el **perfil de los clientes** permitirá detectar en qué momento de su ciclo de vida con la empresa se encuentran y desarrollar propuestas de valor que convengan en cada momento.

Segmentar se define como el proceso de dividir a los clientes actuales o potenciales, en diferentes grupos o segmentos consistentes en individuos con características y niveles similares de interés (véase el Cap. 31 para una explicación detallada de las técnicas del cluster no jerárquico). Es un proceso creativo e iterativo con el fin de satisfacer con mayor acierto las necesidades de los clientes, proporcionando una ventaja competitiva y sostenible a la compañía. La segmentación viene dada por las necesidades de los clientes, no de la compañía, y debería ser revisada periódicamente.

Este caso práctico de negocio está basado en un proyecto real impulsado por el departamento de marketing de una empresa del sector *retail* que necesitaba mejorar el conocimiento de sus clientes, agrupándolos en función de su comportamiento de compra. Los resultados obtenidos fueron clave para definir la estrategia de marketing relacional de la compañía.

53.2. El modelo *Recency, frequency, monetary* tradicional

El **modelo RFM** es una técnica popular que se utiliza para analizar el comportamiento de compra de los clientes: cómo compran, su frecuencia de compra y cuánto gastan. Es un método útil para enriquecer la segmentación de los clientes en varios grupos que permitan la personalización e identificación de los clientes más proclives a responder a las promociones. El análisis RFM depende de las medidas de actualidad (*recency*) (R), frecuencia (*frequency*) (F) y valor monetario (*monetary*) (M), que son tres importantes variables relacionadas con la compra que influyen en las posibilidades de compra futura de los clientes.

El **modelo RFM tradicional** categoriza el valor de las variables dividiéndolas en quintiles, a partir de los cuales se calcula una puntuación única que representa el valor del cliente. Sin embargo, no es muy preciso. Si el intervalo de frecuencia de compras se fija entre 0 y 20, en términos de negocio podría interpretarse como que un cliente con una sola compra será igual que otro que tenga 20. Por ello, los enfoques de conjuntos clásicos pueden resultar poco funcionales (Martínez et al., 2019). En este caso práctico, se propone una mejora en la definición de los intervalos mediante la aplicación del **ranking de percentiles**. Este método, que se ha denominado modelo RFM extendido, proporciona un método robusto para tratar los valores atípicos (*outliers*), y además normaliza las variables entre 0 y 1 para evitar la diferencias de peso entre las variables, permitiendo así la correcta implementación del **algoritmo de segmentación**.

53.3. El modelo *Recency, frequency, monetary* extendido

Los autores de este caso práctico recomiendan seguir una metodología de gestión de proyectos. La **metodología CRISP-DM** (Chapman et al., 2000b), presentada en el Cap. @ref(metodología) es un estándar ampliamente utilizado en los proyectos de ciencia de datos.

Una vez definido el problema (mejorar el conocimiento que una empresa de comercio minorista tiene de sus clientes, agrupándolos en función de su comportamiento de compra), la recopilación y comprensión de los datos (primera etapa del modelo CRISP-DM) se establece como etapa esencial para el desarrollo del proyecto.

53.3.1. Recopilación y comprensión de los datos

Hoy en día, la mayoría de las empresas de *e-commerce* y comercio minorista tradicional cuentan con sistemas que permiten registrar los datos básicos de cada una de sus ventas (fecha, artículo, cantidad e importe), asociados a un código único de cliente. La información contenida en estos datos de compra atesora gran valor, ya que carecen del sesgo y subjetividad propias de otras informaciones obtenidas mediante encuestas de opinión, estudios de mercado, entrevistas y grupos de discusión, etc. Estos datos suelen encontrarse en las plataformas ERP (*Enterprise Resource Planning*) de gestión de pedidos y ventas, o CRM (*Customer Relationship Management*) de las empresas.

53.3. El modelo Recency, frequency, monetary extendido

855

El lector es, o será, consciente de que la fase de extracción, carga y limpieza de los datos es la más exigente del proyecto, y donde se empleará gran parte de los recursos y tiempo de todo el proyecto. **R** cuenta con gran cantidad de paquetes y recursos que facilitan la extracción desde diferentes tipos de bases de datos.

Para este caso práctico serán necesarias las siguientes librerías:

```
library("tidyverse")
library("lubridate")
library("factoextra")
library("ggpubr")
library("CDR")
data("datos_retail")
```

Cualquier tipo de estudio o proyecto de ciencia de datos requiere familiarizarse con los datos y determinar si presentan suficiente exactitud, completitud, consistencia, credibilidad y actualidad (Muñoz-Reja et al., 2018). Los datos de transacciones de venta registrados por las empresas pueden contener datos atípicos (p.ej. valores perdidos, inexactos, outliers, etc.). Para determinar la acción a tomar, o no, de limpieza o corrección de los datos de partida, es esencial conocer el negocio y las consecuencias que éstas operaciones tendrán en el resultado final de la segmentación.

El conjunto de datos de muestra contiene 200.000 observaciones correspondientes a transacciones de compra. Las siguientes variables iniciales están explicadas en el set de datos:

```
head(datos_retail)
#> # A tibble: 6 x 4
#>   id_ticket fecha importe_venta codigo_socio
#>   <chr>     <date>    <dbl>           <chr>
#> 1 num_1646673 2021-10-30 12.4 id_1076134
#> 2 num_2762559 2021-12-03 38.8 id_0552641
#> 3 num_0309422 2022-01-07 67.8 id_0537369
```

53.3.2. Cálculo de las variables del modelo RFM

Identificadas las variables iniciales, es necesario calcular los factores clave del Modelo RFM:

- La variable actualidad, *recency* (R), es el intervalo de tiempo transcurrido desde la última compra de un cliente hasta la fecha de elaboración del modelo RFM.
- La variable frecuencia, *frequency* (F), se obtiene agrupando las compras por cliente y contando el número total de tickets únicos.
- La variable valor monetario, *monetary* (M), se calcula sumando todos los importes de venta por cliente.

```

fecha_estudio_rfm <- ymd("2022-08-01")

rfm <- datos_retail |>
  group_by(codigo_socio) |>
  summarise(
    frecuencia = n_distinct(id_ticket),
    monetario = sum(importe_venta, na.rm = TRUE),
    fecha_transaccion_reciente = first(fecha, order_by = desc(fecha))
  ) |>
  mutate(actualidad = time_length(interval(start = fecha_transaccion_reciente, end =
  fecha_estudio_rfm), unit = "days"), .keep = "unused")

head(rfm) # el lector puede ver las variables del Modelo RFM

```

53.3.3. Breve análisis exploratorio de las variables del modelo RFM

Del análisis puede concluirse que:

- 107.929 clientes han realizado una media de 1,85 compras, con un importe medio total de 70,56€ y 450,4 días de media desde la última compra hasta la fecha de realización del estudio, con una fuerte asimetría positiva de los valores *frequency* y *monetary* (ver Fig. 53.1).
- Se detecta una gran estacionalidad de las compras, como se puede apreciar en la agrupación de las observaciones de *recency*. Teniendo en cuenta la fecha en la que se realiza el análisis, los valores obtenidos en la variable *recency*, se pueden interpretar como el periodo de ventas de la campaña navideña.

```

set.seed(12345)
plot_data <- rfm |>
  slice_sample(n = 2000) |>
  pivot_longer(!codigo_socio, names_to = "variable", values_to = "valor")
plot_data |>
  ggplot(aes(x = variable, y = valor)) +
  geom_boxplot(outlier.shape = NA, color = "red") +
  geom_jitter(alpha = 1 / 10) +
  facet_wrap(~variable, ncol = 6, scales = "free") +
  theme(strip.text.x = element_blank(), text = element_text(size = 9))

```

53.3.4. Cálculo del ranking de percentiles

Los valores de ranking son relativos entre clientes y no pueden ser utilizados para objetivos de negocio, basados en valores absolutos de puntuación por cliente.

53.3. El modelo Recency, frequency, monetary extendido

857

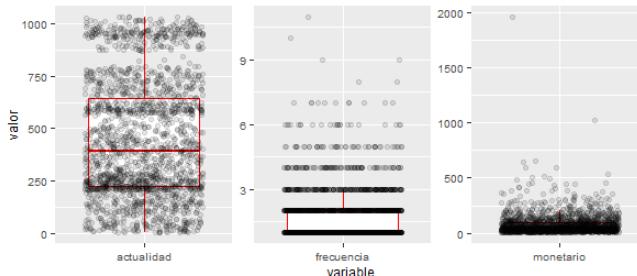


Figura 53.1: Box-plot

```
rfm_rank <- rfm |>
  mutate(across(.cols = c("frecuencia", "monetario"), percent_rank, .names =
    ~ "rank_{.col}")) |>
  mutate(across(.cols = c("actualidad"), ~ percent_rank(-.x), .names = "rank_{.col}"))
  # menor recency indica mayor puntuación en rank
```

Se podría decir que el análisis RFM combina tres atributos clave de los clientes para construir un ranking que permite agruparlos de forma útil para el negocio. Así, a si un cliente que compró en una fecha reciente (Recency) se le otorgan más puntos. Si compró muchas veces (Frequency), también se le coloca más arriba en el ranking. Finalmente, si gastó más en el total de sus compras (Monetary), también puntúa más alto. Combinando estos tres parámetros, se obtiene un ranking RFM. Para la elaboración de este ranking se parte del concepto de percentil . La idea es asignarle a cada cliente una puntuación según las tres variables o factores clave del modelo RFM, de modo que los mejores clientes serán los que tengan una puntuación mayor.

```
head(rfm_rank) # el lector puede ver la puntuación del ranking
```

Una vez que se tienen los rankings de percentiles en las tres variables para todos los clientes, se procede a su clusterización mediante el método k-means.

53.3.5. Modelado: RFM mediante k-means

El modelo establecido debe proporcionar una segmentación de clientes con sentido de negocio. En este caso práctico se opta por el algoritmo de clustering estándar que presenta la ventaja de ser muy intuitivo y permite trabajar con grandes conjuntos de datos. Como el lector ha podido comprobar, existen otros muchos algoritmos de aprendizaje no supervisado que pueden ser empleados.

El número óptimo de *clusters* (o segmentos, en la jerga del marketing) es uno de los retos a la hora de aplicar técnicas de *clustering*. No existe una manera exclusiva de encontrar el número adecuado de clusters. Se trata de un proceso subjetivo que depende de los datos, del tipo de *clustering* empleado y, en este caso, de que el número elegido tenga sentido y utilidad en el

negocio. Existen numerosos métodos para facilitar la elección del número de *clusters*; entre ellos destacan el *Elbow method*, el *Average silhouette method* y el *Gap statistic method*, que gracias a la función `fviz_nbclust()` del paquete `factoextra`, se pueden calcular con facilidad para realizar una buena elección.

Con el método Elbow, número óptimo de *clusters* se calcula como sigue:

```
set.seed(123)
muestra_clusters <- rfm_rank |>
  slice_sample(n = 5000) |>
  dplyr::select(matches("rank"))

fviz_nbclust(x = muestra_clusters, FUNcluster = kmeans, method = "wss", k.max = 10)
```

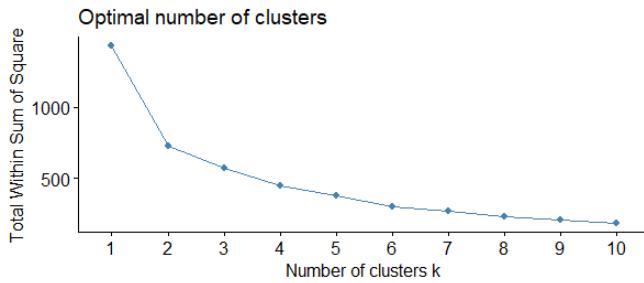


Figura 53.2: Número óptimo de clusters

En la Fig. 53.2 se observa que la varianza total *intra-cluster* apenas mejora a partir del cuarto *cluster*.

El algoritmo de clustering k-means se entrena con las variables R-F-M normalizadas con el ranking. La salida de la función `kmeans()` del paquete base `stats` es un objeto que, entre otros componentes, ofrece un vector numérico indicativo del *clusters* al que pertenece cada uno de los clientes.

```
set.seed(123)
km_fit <- kmeans(x = rfm_rank[, 5:7], centers = 4, nstart = 10)

clientes_segmentos <- rfm_rank |>
  mutate(segmento = km_fit$cluster)

head(clientes_segmentos) # el lector puede ver el segmento al que pertenece el cliente
```

53.3.6. Descriptivos e interpretación de los segmentos

Los segmentos obtenidos deben tener sentido y utilidad de negocio. Para ello es imprescindible proporcionar los estadísticos descriptivos de cada segmento y proceder a su interpretación de

53.3. El modelo Recency, frequency, monetary extendido

859

perfil de cliente.

```
descriptivo_segmentos <- clientes_segmentos |>
  group_by(segmento) |>
  summarise(across(c("monetario", "frecuencia", "actualidad"),
    .fns = mean, .names = "md_{.col}"))
  ), n_clientes = n() |>
  ungroup() |>
  relocate(segmento, n_clientes)

head(descriptivo_segmentos)
#> # A tibble: 4 x 5
#>   segmento n_clientes md_monetario md_frecuencia md_actualidad
#>   <int> <int> <dbl> <dbl> <dbl>
#> 1 1 23551 36.2 1.07 239.
#> 2 2 23632 77.0 2.26 567.
#> 3 3 28809 128. 3.11 188.
#> 4 4 31937 39.3 1.00 757.
```

Interpretación de los segmentos:

- 1-Nuevos probando: segmento que agrupa nuevos clientes que están realizando compras desde hace poco tiempo y tienen un gran potencial de desarrollo. Es un segmento de clientes con interés para la empresa.
- 2-No podemos perder: se trata de los clientes 'churn'¹ que fueron buenos clientes en términos monetarios y de frecuencia pero que hace tiempo que no realizan nuevas compras. La compañía debe hacer un esfuerzo en recuperar estos clientes para convertirlos al segmento TOP.
- 3-Top: reúne a los mejores clientes de la empresa. Son clientes que compran con frecuencia, están activos y aportan ventas a la compañía. Es el segmento de clientes con mayor interés para la empresa.
- 4-Una compra: segmento formado por aquellos clientes que han realizado una sola compra hace tiempo. Presentan frecuencia, actualidad y valor monetario bajo. Se trata de un segmento de clientes con escaso interés para la compañía.

```
segmentos_descriptivo <- clientes_segmentos |>
  mutate(segmento = case_when(
    segmento == 1 ~ "1_Nuevos probando",
    segmento == 2 ~ "2_No perder",
    segmento == 3 ~ "3_Top",
    segmento == 4 ~ "4_Una compra"
  )) |>
  group_by(segmento) |>
  summarise(
```

¹El término *customer churn* se suele traducir como perdida de clientes o rotación de clientes. Se compone de las palabras inglesas *change* (en castellano cambio) y *turn* (en castellano abandonar).

```

across(
  .cols = where(is.numeric),
  .fns = mean
),
n_clientes = n()
) |>
ungroup() |>
relocate(segmento, n_clientes)

table_dot_plot <- segmentos_descriptivo |>
# select(starts_with("rank")) |>
pivot_longer(cols = c("rank_monetario", "rank_frecuencia", "rank_actualidad"),
  ~ names_to = "Variable RFM", values_to = "Puntuación")

ggdotchart(
  table_dot_plot,
  x = "Variable RFM", y = "Puntuación",
  group = "segmento", color = "segmento", palette = "jco",
  add = "segment", position = position_dodge(0.3),
  sorting = "none", facet.by = "segmento", dot.size = 5,
  rotate = TRUE, legend = "none"
)
)

```

La Fig 53.3 muestra cada uno de los segmentos indicados. El ranking obtenido ayuda a identificar las diferencias en los tipos de clientes y es útil para decidir a qué segmentos enfocarse y qué estrategias usar para cada uno.

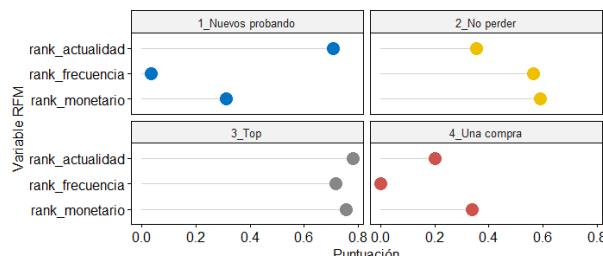


Figura 53.3: Lollipop de variables RFM

53.3.7. Puesta en producción

Calculado el modelo RFM k-means, la compañía puede incorporar periódicamente los datos de los clientes nuevos, o actualizados. De este modo, los segmentos de clientes se actualizarán y, más allá de las acciones de *marketing mix* que realicen las compañías gracias a la segmentación, podrán analizarse las migraciones de clientes entre los diferentes segmentos en el periodo estudiado. La función `cl_predict()` facilita la actualización periódica de los segmentos con el modelo entrenado.

Capítulo 54

Análisis de datos en medicina

Alberto M. Borobia^a y María Jiménez-González^a

^aHospital Universitario La Paz - IdiPAZ

54.1. Justificación

La aplicación de la estadística en la investigación clínica ha sido una de las herramientas clave en los últimos dos años. La pandemia mundial causada por la enfermedad por coronavirus (COVID-19) es una enfermedad infecciosa provocada por el virus SARS-CoV-2. Drante el año 2020, más de 13 millones de casos diagnosticados en España arrojaban un diagnóstico claro: se necesita más investigación.

El primer apartado de este capítulo señala la importancia de la identificación de los sesgos (en concreto, del sesgo de selección) que aparecen en los estudios de investigación no aleatorizados. Tras ello, se abordará un ejemplo práctico de una de las aplicaciones más significativas de la bioestadística: el análisis de supervivencia, con el que se resuelven preguntas tan importantes cómo: ¿qué factores de riesgos están asociados a la mortalidad provocada por coronavirus?.

54.2. Introducción al uso de datos en investigación clínica y ensayos clínicos

En este capítulo, y a modo ilustrativo del ámbito de la investigación clínica, se abordarán tres análisis a partir de los datos:

- Un análisis relativo a la eliminación de sesgos, o más concretamente, a la eliminación del sesgo de selección.
- Un análisis relativo a la estimación e interpretación de las curvas de supervivencia.
- Un análisis relativo a la estimación e interpretación de la Regresión de COX.

54.2.1. ¿Qué es un ensayo clínico?

En la investigación clínica existen dos tipos de estudios: *estudios observacionales* y *ensayos clínicos*.

Los ensayos clínicos aleatorios se definen como el diseño experimental óptimo para proporcionar evidencia, eficacia y seguridad de una intervención (Liu et al., 2020). Los tratamientos estudiados o investigados son asignados aleatoriamente en grupos que garantizan que las diferencias en los resultados después del tratamiento reflejen los efectos del mismo (Rosenbaum, 2005). Cuando estas condiciones ideales no son posibles (falta de recursos, financiación, tiempo, etc), se definen como estudios observacionales.

Previo a la puesta en marcha de un ensayo clínico, es imprescindible la redacción de un **Protocolo** y un **Plan de Análisis Estadístico (PAE)**.

- El protocolo, elaborado por los investigadores del estudio, precisa y justifica los métodos y planes del proceso que se llevará a cabo en el ensayo clínico (Rivera et al., 2020).
- El PAE detalla las características principales del eventual análisis estadístico de los datos, que deben describirse en la sección estadística del protocolo (Lewis, 1999).

Los documentos anteriormente mencionados, y el resto de directrices necesarias para un ensayo clínico, están regulados por la “Conferencia Internacional sobre armonización de requisitos técnicos para el registro de productos farmacéuticos para uso humano” (sus siglas ICH en inglés).

54.2.2. Limitaciones de los estudios observacionales

En el apartado anterior, se puso de manifiesto la importancia de los ensayos clínicos aleatorizados. Sin embargo, la posible falta de recursos, financiación, tiempo o materiales, dificultan la puesta en marcha y realización de los mismos.

En consecuencia, la puesta en práctica de la investigación puede no ser la ideal. Los estudios observacionales, sin embargo, son una herramienta elemental en circunstancias no tan óptimas, ya que permiten analizar e investigar (contra viento y marea).

La limitación principal de los estudios observacionales es que introducen sesgos en el análisis. Los ensayos clínicos tienen como principal objetivo eliminar el **sesgo de selección**: cuando los sujetos no son asignados aleatoriamente, por ejemplo, los resultados diferentes pueden reflejar estas diferencias iniciales en lugar de los efectos de los tratamientos (Rosenbaum, 2005).

54.2.3. Índice de propensión (*propensity score*)

Una solución aconsejable y recomendable ante los sesgos “escondidos” en los estudios observacionales es la técnica *propensity score* o índice de propensión. Esta técnica de emparejamiento equilibra las covariables observadas sesgadas ajustando por su índice de propensión, eliminando presumiblemente el sesgo. Habitualmente, el índice de propensión se obtiene a partir de un

54.2. Introducción al uso de datos en investigación clínica y ensayos clínicos 863

modelo de regresión cuya variable dependiente corresponde a la intervención o el resultado principal (por ejemplo, la muerte) y las variables independientes o covariables corresponden a las variables que puedan tener un efecto confusor en la variable dependiente [molina2015indices].

54.2.4. Ejemplo práctico en R de un estudio observacional

El *dataset* sintético `datos_observacional` reproduce los datos de un hipotético estudio observacional sobre una enfermedad **X**. El objetivo del estudio es estudiar los factores de riesgo asociados a la mortalidad causada por esa enfermedad.

```
#> # A tibble: 5 x 8
#>   ID fecha_hospitalizacion sexo   edad comorbilidades      fecha~1 exitus
#>   <dbl> <dttm>          <chr> <dbl> <chr>           <chr>    <dbl>
#> 1     1 2015-04-17 00:00:00 Mujer     76 1 o más comorbilidades 17/04/~     1
#> 2     2 2015-03-21 00:00:00 Mujer     64 1 o más comorbilidades 31/03/~     0
#> 3     3 2015-04-09 00:00:00 Hombre    65 1 o más comorbilidades 16/04/~     0
#> 4     4 2015-04-04 00:00:00 Hombre    77 1 o más comorbilidades 13/04/~     0
#> 5     5 2015-03-24 00:00:00 Mujer     66 1 o más comorbilidades 27/03/~     0
#> # ... with 1 more variable: fecha_exitus <dttm>, and abbreviated variable name
#> #   1: fecha_alta
```

En la literatura, se ha evidenciado que las mujeres de mayor edad y con una o más comorbilidades tienen más riesgo de fallecer (`exitus`) por la enfermedad **XX**. En investigación, la estructura de los resultados en un paper o en un informe estadístico, independientemente de la revista o PAE, comienza en el mismo punto: una tabla resumen de las características basales de la población objeto de estudio. El paquete `tableone` (sencillo juego de palabras) integra funciones específicas para la creación de dichas tablas. La función principal de este paquete es `CreateTableOne()`.

```
library("tableone")
my_vars <- c("sexo", "edad", "comorbilidades")
nonnormal <- c("edad")
factor_vars <- c("sexo", "comorbilidades")

# crea la tabla
tab1 <- CreateTableOne(
  vars = my_vars, factorVars = factor_vars,
  strata = "exitus", data = datos_observacional
)

# imprime la tabla
tab1 <- print(tab1,
  showAllLevels = TRUE, formatOptions = list(big.mark = ","),
  exact = "stage", nonnormal = nonnormal
)
```

Tabla 54.1: Características basales de la población

	level	Vivo	Exitus	p-valor	test
n		79	21		
sexo (%)	Hombre	28 (35.4)	2 (9.5)	0.042	
	Mujer	51 (64.6)	19 (90.5)		
edad (median [IQR])		64.00 [53.00, 73.00]	82.00 [72.00, 85.00]	<0.001	nonnorm
comorbilidades (%)	1 o más comorbilidades	43 (54.4)	18 (85.7)	0.018	
	No	36 (45.6)	3 (14.3)		

Para presentar la tabla de resultados formateada basta con usar la función `kable()`:

```
knitr::kable(tab1,
  caption = "Características basales de la población",
  col.names = c("level", "Vivo", "Exitus", "p-valor", "test")
)
```

En la Tabla 54.1 y acorde a la bibliografía existente, se confirma el sesgo de selección a través del desequilibrio de la variable principal (`exitus`) en las variables `sexo`, `edad` y `comorbilidades`, evidenciado a través de la significación de éstas. Un argumento que motiva la aplicación, en este caso, de la técnica *propensity score* se fundamenta en la viabilidad de, por ejemplo, un modelo multivariante (como puede ser un modelo de predicción). La recogida de datos de un estudio observacional, como el de este ejemplo, normalmente viene dada por la disponibilidad de la población: sujetos ingresados en el Hospital por la enfermedad (en nuestro caso, coronavirus). Por tanto, esta muestra seleccionada recogerá pacientes con pronóstico más grave que la población general (mujer de mayor edad con una o más comorbilidades).

El paquete `MatchIt` integra las funciones principales para el ajuste de la técnica *propensity score*, concretamente la función `matchit()` integra la teoría de (Ho et al., 2007) para el emparejamiento óptimo de los grupos estudiados. Los argumentos más importantes de esta función son: - `formula`: modelo de regresión que estudia la relación entre la variable principal de estudio (`exitus`) con las variables sesgadas (`sexo`, `edad` y `comorbilidades`). - `method`: especifica el método de *matching*. - `distance`: especifica el método para la estimación del índice de propensión.

La función `get_matches()` empareja, posteriormente, las coincidencias que resultan del `MatchIt`.

Nota: es imprescindible que los casos del `dataset` estén completos.

```
library("MatchIt")
match <- matchit(exitus ~ edad + as.factor(sexo) + as.factor(comorbilidades),
  method = "nearest", distance = "mahalanobis",
  data = datos_observacional
```

54.2. Introducción al uso de datos en investigación clínica y ensayos clínicos 865

Tabla 54.2: Características basales de la población aplicando la técnica de propensity score

	level	Vivo	Exitus	p-valor	test
n		21	21		
sexo (%)	Hombre	2 (9.5)	2 (9.5)	1.000	
	Mujer	19 (90.5)	19 (90.5)		
edad (median [IQR])		72.00 [69.00, 84.00]	82.00 [72.00, 85.00]	0.182	nonnorm
comorbilidades (%)	1 o más comorbilidades	18 (85.7)	18 (85.7)	1.000	
	No	3 (14.3)	3 (14.3)		

```
)
datos_observacional_match <- get_matches(match, datos_observacional)
```

Para comprobar que el sesgo evidenciado en estudios anteriores ha desaparecido, se reproduce la tabla anterior.

```
tab1_corregida <- CreateTableOne(
  vars = my_vars, factorVars = factor_vars,
  strata = "exitus", data = datos_observacional_match
)

# se imprime en el objeto tab1_corregida
tab1_corregida <- print(tab1_corregida,
  showAllLevels = TRUE, formatOptions = list(big.mark = ","),
  exact = "stage", nonnormal = nonnormal
)
```

Se formatea la salida de la tabla:

```
knitr::kable(tab1_corregida,
  caption = "Características basales de la población aplicando la técnica de propensity
  ← score",
  col.names = c("level", "Vivo", "Exitus", "p-valor", "test")
)
```

En la Tabla @ref(tab:tab1_corregida) se observa que el sesgo de selección existente en la muestra se ha resuelto equilibrando las variables (aunque reduciendo la muestra). Tras este paso previo, podría realizarse un análisis estándar de esta muestra intentando aproximarse lo máximo posible a un estudio aleatorizado.

54.3. Análisis de supervivencia

Durante la pandemia ocasionada por el SARS-CoV-2, la pregunta principal de los investigadores clínicos se centró en un mismo objetivo: factores de riesgo asociados a la mortalidad causada por COVID-19. El análisis de supervivencia ha permitido a los investigadores intentar explicar las causas más factibles que producen esa mayor probabilidad de fallecer. El análisis de supervivencia permite estudiar los factores de riesgo asociados a la mortalidad. La ventaja principal de este análisis frente a un análisis estándar (como puede ser una regresión logística) se centra en la integración en la variable respuesta del evento y del tiempo hasta el evento, que tiene como consecuencia la interpretación de “riesgo” y no de “probabilidad” en los resultados.

El *dataset* utilizado, `datos_supervivencia` está incluido en el paquete `CDR` y está formado por 301 pacientes, 101 diagnosticados con infección por SARS-CoV-2 y 100 exitus.

```
head(datos_supervivencia, 5)
#> # A tibble: 5 x 7
#>   id EXITUS_TIME DIAG_COVID EXITUS N_COMORBIDITIES SEX      EDAD
#>   <dbl>        <dbl>     <dbl>    <dbl>            <dbl> <chr>    <dbl>
#> 1 262          0         1       1             5 Hombre    83
#> 2 236          1         1       1             5 Hombre    72
#> 3 170          11        0       0             2 Mujer     65
#> 4 204          11        1       1             4 Hombre    80
#> 5 46           14        1       1             5 Hombre    90
```

54.3.1. Estimación y comparación de curvas de supervivencia

La **función (o curva) de supervivencia** estuda la probabilidad de que el paciente o sujeto, sobreviva a un tiempo **X**. El estimador más común utilizado para el ajuste de la función de supervivencia es el estimador no paramétrico **Kaplan-Meier** y su función escalonada. Una vez generadas estas curvas de supervivencia, existen diferentes métodos (paramétricos y no paramétricos) para su comparación. En este apartado, se utiliza la prueba de **Mantel-Cox** (o test **Log-Rank**) para el contraste de funciones.

Los paquetes `survival` y `survminer` integran las funciones principales de la técnica:

- La función `Surv()`, de la librería `survival`, crea un objeto de supervivencia formado por el evento (exitus) y el tiempo hasta la ocurrencia del evento.
- La función `survfit()`, de la librería `survival`, estima la función de supervivencia mediante el método *Kaplan-Meier* del objeto `Surv` y los factores de riesgo asociados.
- La función `ggsurvplot()`, genera el gráfico de la curva de supervivencia (basada en la librería `ggplot2`). El argumento principal de la función es la función de supervivencia estimada, `survfit()`. Los argumentos más importantes (y recomendables) a la hora de graficar la función de supervivencia son:
 - `pval`: muestra el p-valor correspondiente a la comparación a través del test *Log-Rank*.

54.3. Análisis de supervivencia

867

- `conf.int`: muestra los intervalos de confianza de la(s) curva(s) de supervivencia.
- `risk.table`: añade el número de sujetos (absoluto o relativo) en riesgo en cada momento del periodo objeto de estudio.

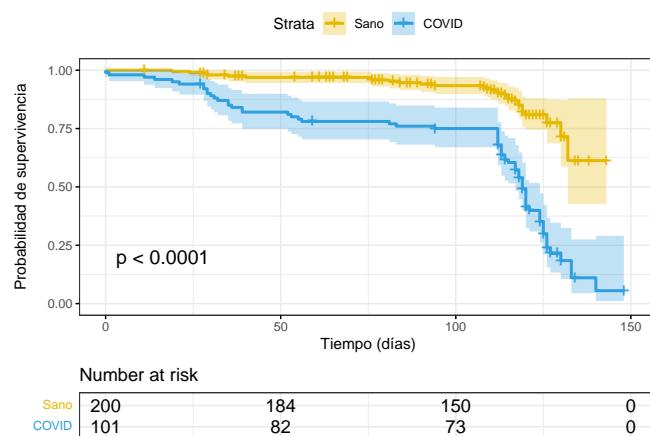
Se cragan los paquetes

```
library("survival")
library("survminer")
```

Se ajusta el modelo y, posteriormente, se representa:

```
fit <- survfit(Surv(EXITUS_TIME, EXITUS) ~ DIAG_COVID,
  data = datos_supervivencia
)

ggsurvplot(fit,
  data = datos_supervivencia,
  pval = TRUE,
  conf.int = TRUE,
  ggtheme = theme_bw(),
  palette = c("#E7B800", "#2E9FDF"),
  xlab = "Tiempo (días)",
  ylab = "Probabilidad de supervivencia",
  legend.labs = c("Sano", "COVID"),
  # añade tabla de supervivencia
  risk.table = TRUE,
  tables.height = 0.2,
  tables.theme = theme_cleantable()
)
```



En la Fig. ??, dónde el eje X corresponde al tiempo en días y el eje Y a la probabilidad de supervivencia, se observa que la probabilidad de supervivencia de las personas expuestas a COVID es significativamente menor (p -valor < 0.001) a las personas sanas. La mediana de supervivencia (línea trazada desde el 0.5 del eje Y, correspondiente al 50 % de la probabilidad de supervivencia) corresponde a los 120 días, es decir, el 50 % de los sujetos diagnosticados por COVID y objeto de estudio sobrevivieron, al menos, 120 días.

Por tanto, se puede concluir que se ha encontrado evidencia sobre el aumento de mortalidad asociada a la enfermedad COVID-19.

54.4. Regresión de COX

La regresión de Cox¹ o modelo de riesgos proporcionales es una técnica utilizada para el estudio del efecto de covariables sobre el tiempo hasta la ocurrencia de un evento (exitus, recaída, progresión, etc). La regresión de Cox es realmente una Regresión en la que la variable dependiente es siempre una función de riesgo o supervivencia (están íntimamente relacionadas) y los predictores son una función del tiempo y una función de las variables consideradas como explicativas. En general, se suele expresar así:

$$h(t, x_1, x_2, \dots, x_p) = h_0(t) + g(x_1, x_2, \dots, x_p),$$

y más concretamente,

$$h(t, x_1, x_2, \dots, x_p) = h_0(t) + e^g(x_1, x_2, \dots, x_p),$$

donde g , normalmente, indica una combinación lineal de las covariables o variables explicativas. Es, por tanto, una técnica semi-paramétrica.

La función principal para el ajuste de un modelo de regresión de Cox es `coxph()`. Esta función, al igual que la función `survfit()`, está formada por un objeto `Surv` y las covariables del modelo.

```
fit_cox <- coxph(Surv(EXITUS_TIME, EXITUS) ~ DIAG_COVID + EDAD + SEX + N_COMORBIDITIES,
                    data = datos_supervivencia
)
```

El *output* principal de una regresión de Cox,

$$h(t, x_1, x_2, \dots, x_p),$$

¹Es importante distinguir la regresión de Cox de la regresión logística (véase el Cap. 16). La Regresión logística relaciona la variable dependiente dicotómica con un conjunto de variables independientes sin contemplar el tiempo o contemplándolo sólo de forma estática, viendo en un punto fijo del tiempo si el suceso estudiado ha acontecido o no, pero no teniendo en consideración en qué momento ha sucedido. La Regresión de Cox proporciona un análisis más fino. No analiza, en un instante de tiempo dado, si un acontecimiento de interés ha sucedido o no, sino cuándo ha sucedido, si es que ha sucedido, y lo hace teniendo en cuenta el comportamiento de una o varias variables independientes. Es por ello que la regresión logística trabaja con la *odds ratio* y la regresión de Cox con la *hazard ratio*.

54.4. Regresión de COX

869

son las **razones de riesgos** o ***hazard ratios (HR)**. Es decir, la relación entre las dos funciones de riesgo en función de los cambios operados en las variables explicativas. En concreto, la exponencial del coeficiente estimado para la va

El *output* principal de una regresión de Cox , $h(t, x_1, x_2, \dots, x_p)$, son las razones de riesgos o hazard ratios (HR). Es decir, la relación entre las dos funciones de riesgo en función de los cambios operados en las variables explicativas. En concreto, la exponencial del coeficiente estimado para la variable explicativa X_i indica el incremento en el riesgo de fallecer cuando la variable explicativa aumenta en una unidad y las demás permanecen constantes. Esta razón de riesgos oscila entre 0 a ∞ , siendo el intervalo [0,1] una relación de riesgo bajo y [1, ∞] una relación de riesgo alto.

Nota

- Los HR localizados entre 1 y 2 se interpretan en porcentaje. Es decir, HR = 1.5 indica a un aumento del riesgo del 50 %.
- Los HR localizados entre 2 e ∞ se interpretan en “veces”. Es decir, HR = 3 indica a un aumento del riesgo de 3 veces.
- Los HR localizados entre 0 y 1 se interpretan como una reducción del riesgo del $(1 - HR) \times 100\%$. Es decir, HR = 0.8 indica a una disminución del riesgo del 20 %.

```
summary(fit_cox)
#> Call:
#> coxph(formula = Surv(EXITUS_TIME, EXITUS) ~ DIAG_COVID + EDAD +
#>       SEX + N_COMORBIDITIES, data = datos_supervivencia)
#>
#>     n= 271, number of events= 100
#>     (30 observations deleted due to missingness)
#>
#>                 coef  exp(coef)    se(coef)      z Pr(>|z|)    
#> DIAG_COVID     1.3023581  3.6779594  0.5184547  2.512   0.0120 *  
#> EDAD          0.0006006  1.0006008  0.0116113  0.052   0.9587    
#> SEXMujer      -1.1256901  0.3244285  0.2360183 -4.770  1.85e-06 *** 
#> N_COMORBIDITIES 0.1643743  1.1786554  0.0774043  2.124   0.0337 *  
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#>                 exp(coef) exp(-coef) lower .95 upper .95    
#> DIAG_COVID      3.6780      0.2719     1.3314    10.1605  
#> EDAD            1.0006      0.9994     0.9781     1.0236  
#> SEXMujer        0.3244      3.0823     0.2043     0.5153  
#> N_COMORBIDITIES 1.1787      0.8484     1.0127     1.3717  
#>
#> Concordance= 0.815  (se = 0.025 )
#> Likelihood ratio test= 130.1  on 4 df,   p=<2e-16
```

```
#> Wald test          = 117.9  on 4 df,   p=<2e-16
#> Score (logrank) test = 165.9  on 4 df,   p=<2e-16
```

De la compleja salida del modelo, deben resaltarse y explicarse los siguientes puntos:

- Apartado 1: Fórmula del modelo, tamaño muestral y número de eventos.
- Apartado 2: Tabla con los coeficientes del modelo y su p-valor.
- Apartado 3: Tabla con los Hazard Ratio (exponencial de los coeficientes de la tabla anterior) y sus intervalos de confianza (lower .95 y upper .95).
- Apartado 4: parámetros de bondad de ajuste del modelo.

Por tanto, de este modelo se pueden concluir las siguientes interpretaciones:

- Un paciente diagnosticado de COVID-19 tiene 3.6 veces más riesgo de fallecer que un paciente sano.
- Una mujer tiene un 67.6 % menos riesgo de fallecer que un hombre.
- Por cada comorbilidad, el riesgo de fallecer aumenta un 17.9 %.

54.5. Conclusión

Ha sido necesaria una pandemia mundial para que la sociedad empiece a dar visibilidad y reconocimiento no sólo a la bioestadística, sino a la investigación clínica y a la necesidad de gestionar el uso masivo de datos en salud. A pesar de los múltiples estudios y experiencias pasadas que llamaban a la prudencia y a la acción concreta si se daba una situación similar, el mundo ha sido incapaz de actuar convenientemente. Esto último se ve reflejado en el mínimo aumento de inversión, reconocimiento y notoriedad no sólo en investigación o desarrollo, sino en el apoyo a la ciencia.

Es, quizás, la paradoja más extraña pero que representa el dicho popular:

La sociedad no avanzará si la ciencia no lo hace primero.

Capítulo 55

Messi y Ronaldo: dos ídolos desde la perspectiva de los datos

Borja Andrino Turón

EL PAÍS

55.1. Motivación

El uso de estadísticas avanzadas en los deportes, especialmente en el fútbol, ha despegado en los últimos años. Una buena señal de su irrupción es la apuesta de algunos medios deportivos — como FiveThirtyEight o The Athletic — por contenidos basados en el análisis y la visualización de estas estadísticas para explicar las fortalezas y debilidades de jugadores y equipos. Además, la generación de estadísticas avanzadas, como los goles esperados, la amenaza o el valor con balón están comenzando a sustituir a las métricas tradicionales en la narración y las crónicas de los encuentros.

55.2. Las estadísticas y el fútbol

En el presente capítulo se usarán estadísticas de la web especializada Fbref.com para visualizar el dominio de Cristiano Ronaldo y Lionel Messi durante más de 15 años. Para usar estos datos podríamos usar técnicas de *web scraping* esta página web o usar la librería `worldfootballR`, desarrollada por Jason Zivkovic. La librería permite obtener datos de diferentes plataformas.

La publicación y explotación de estadísticas avanzadas es reciente, de las últimas seis temporadas, con lo que para analizar las carreras completas de estos dos jugadores tendremos que conformarnos, de momento, con métricas tradicionales.

En la Fig. 55.1 se ve la evolución de las principales cifras que definen a un atacante: los goles y las asistencias. Esta estandarización nos permite poder comparar ambos jugadores independientemente del número de minutos, aunque se ha añadido un filtro de al menos 1.000 minutos jugados en la temporada para evitar ruido.

Para realizar el gráfico, se toman los datos originales y se filtran para que solo aparezcan los jugadores seleccionados, en las temporadas con muestra suficiente. A continuación, se seleccionan las columnas que se usarán en el *plot* y se giran las dos métricas para poder añadirlas en un único `geom_line()`.

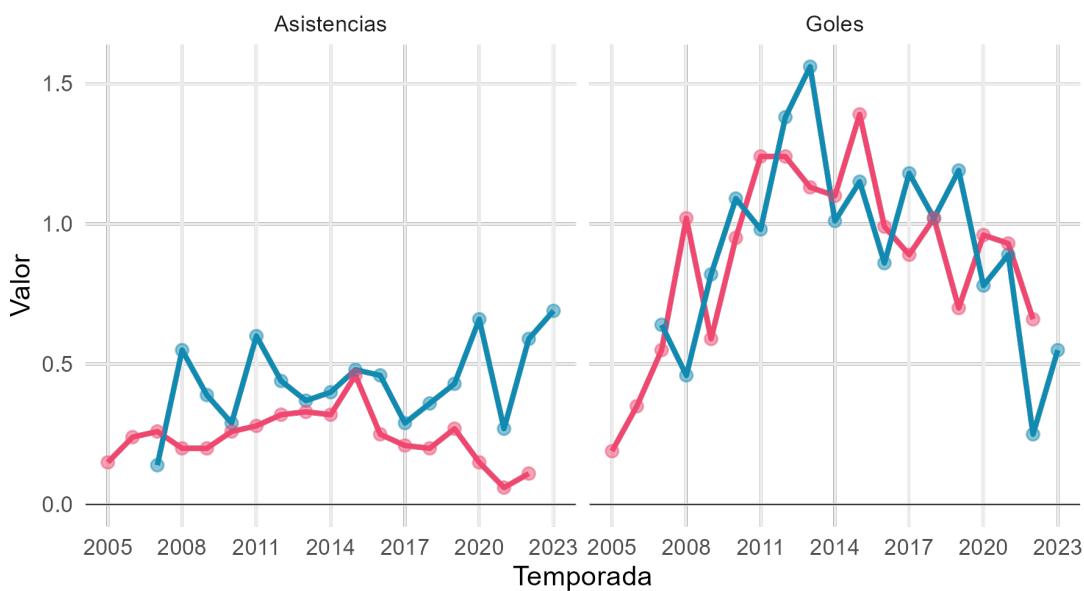
```
pacman::p_load(CDR, tidyverse, janitor, ggbeeswarm, here,
                 patchwork, ggtext, ggrepel)

datos_players |>
  filter(player %in% c("Cristiano Ronaldo", "Lionel Messi"),
         min_playing > 1000) |>
  select(season_end_year, player, Goles = gls_per, Asistencias = ast_per) |>
  pivot_longer(c(Goles, Asistencias), names_to = "metric", values_to = "value") |>
  ggplot(aes(x = season_end_year, y = value, color = player)) +
  geom_line(size = 1) +
  geom_point(size = 2, alpha = 0.5) +
  scale_color_manual(values = c("Lionel Messi" = "#118ab2",
                                "Cristiano Ronaldo" = "#ef476f")) +
  scale_x_continuous(breaks = seq(2005, 2023, 3)) +
  geom_hline(yintercept = 0, size = 0.2) +
  facet_wrap(~metric) +
  labs(title = "Evolución de los goles y asistencias por cada 90 minutos de<br><b>Cristiano Ronaldo</b> y <b>Lionel Messi</b>",
       x = "Temporada", y = "Valor", caption = "Fuente: Fbref.com") +
  theme_minimal() +
  theme(legend.position = "none",
        panel.grid.minor = element_blank(),
        plot.title = element_markdown(margin=margin(0,0,10,-30),
                                       size=12))
```

La Fig. 55.1 arroja un dato increíble, durante 10 años, tener a Messi o Cristiano en el campo significaba contar en ese partido con un gol y casi media asistencia.

Pero la visualización solo nos habla de estos dos futbolistas. Para compararlos con otros jugadores se puede calcular el percentil de goles y asistencias por 90 minutos, temporada a temporada, de los jugadores que hayan jugado más de 1.000 minutos (véase Fig. 55.2). El resultado de nuevo es impactante: durante 13 temporadas Messi y Cristiano han estado entre el 1% de jugadores con más goles. Además, el argentino ha terminado la temporada entre el 1% con más asistencias en 9 ocasiones.

Evolución de los goles y asistencias por cada 90 minutos de Cristiano Ronaldo y Lionel Messi



Fuente: Fbref.com

Figura 55.1: Evolución de goles y asistencias por 90 minutos de Cristiano y Messi desde 2005

```

percentiles_to_plot <-
  datos_players |>
  clean_names() |>
  filter(min_playing > 1000) |>
  select(season_end_year, player, min_playing, gls_per, ast_per) |>
  group_by(season_end_year) |>
  mutate(across(c(gls_per, ast_per), ntile, 100,
                .names = "{.col}_centil")) |>
  ungroup() |>
  mutate(highlighted_player = if_else(player %in%
                                         c("Cristiano Ronaldo", "Lionel Messi"),
                                         T,
                                         F)) |>
  select(season_end_year, player, highlighted_player,
         Goles = gls_per_centil, Asistencias = ast_per_centil)

percentiles_to_plot |>
  pivot_longer(c(Goles, Asistencias), names_to = "metric", values_to = "value") |>
  ggplot(aes(x = season_end_year, y = value, group = season_end_year)) +
  geom_jitter(aes(alpha = highlighted_player, color = player)) +
  scale_color_manual(values = c("Lionel Messi" = "#118ab2",
                                "Cristiano Ronaldo" = "#ef476f")) +
  geom_hline(yintercept = 0, size = 0.1) +
  labs(title = "Percentil de goles y asistencias por cada 90<br>minutos de <b>Temporada</b> para <b>Cristiano Ronaldo</b> y <b>Lionel Messi</b>",
       x = "Temporada", y = "Percentil", caption = "Fuente: Fbref.com") +
  facet_wrap(~metric, scales = "free") +
  scale_x_continuous(breaks = seq(2005, 2023)) +
  scale_alpha_manual(values = c(0.01, 1)) +
  coord_flip() +
  guides(alpha = "none") +
  theme_minimal() +
  theme(legend.position = "none",
        panel.grid.minor = element_blank(),
        plot.title = element_markdown(margin=margin(0,0,0,-30), size=12))

```

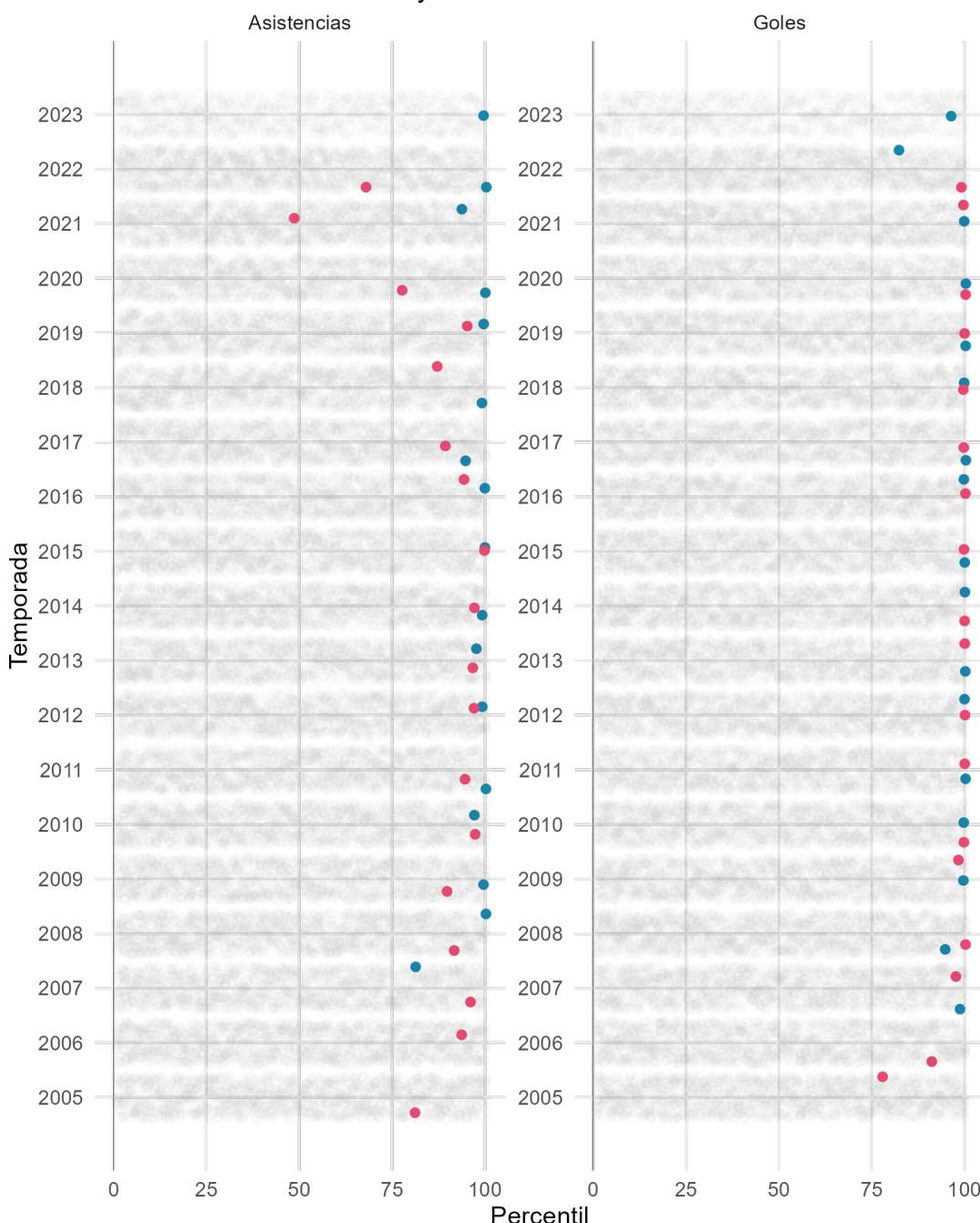
Desde la temporada 2017/18 en esta web publican estadísticas avanzadas de jugadores por partido y temporada. En la Fig. 55.3 se representan los goles esperados (miden cómo de probable es el gol dado un disparo) y las asistencias esperadas (suma de los goles esperados que suman los pases que desembocan en un tiro) por 90 minutos de los jugadores con más de 1.000 minutos. De nuevo el gráfico destaca a nuestros dos protagonistas, que se sitúan en el arco más alejado del origen de coordenadas, donde se juntan los jugadores con mejores números de asistencias y goles esperados.

```

expected_data <-
  datos_players |>
  clean_names() |>

```

Percentil de goles y asistencias por cada 90 minutos de Cristiano Ronaldo y Lionel Messi



Fuente: Fbref.com

Figura 55.2: Percentil de goles y asistencias por 90 minutos cada temporada desde 2005

```

filter(season_end_year >= 2018,
       min_playing > 1000,
       x_g_per > 0 | x_ag_per > 0) |>
mutate(highlighted_player = if_else(player %in% c("Cristiano Ronaldo", "Lionel
→ Messi"),
                                     T,
                                     F),
       label = if_else(player %in% c("Cristiano Ronaldo", "Lionel Messi"),
                      as.character(season_end_year),
                      NA_character_))

expected_data |>
  select(season_end_year, player, highlighted_player, label,
         Goles = x_g_per, Asistencias = x_ag_per) |>
  ggplot(aes(x = Asistencias, y = Goles)) +
  geom_point(aes(alpha = highlighted_player,
                 color = player)) +
  geom_text_repel(aes(label = str_sub(label, 3, 4))) +
  scale_color_manual(values = c("Lionel Messi" = "#118ab2",
                                "Cristiano Ronaldo" = "#ef476f")) +
  geom_hline(yintercept = 0, size = 0.2) +
  geom_vline(xintercept = 0, size = 0.2) +
  scale_alpha_manual(values = c(0.1, 1)) +
  labs(title = "Goles y asistencias esperadas por 90 minutos cada temporada de<br><b>Cristiano Ronaldo</b>, <b>Lionel Messi</b> y el resto de jugadores",
       x = "Asistencias esperadas", y = "Goles esperados", caption = "Fuente:
       → Fbref.com") +
  guides(alpha = "none") +
  theme_minimal() +
  theme(legend.position = "none",
        plot.title = element_markdown(margin=margin(0,0,10,-30),
                                      size=12),
        legend.title = element_blank())

```

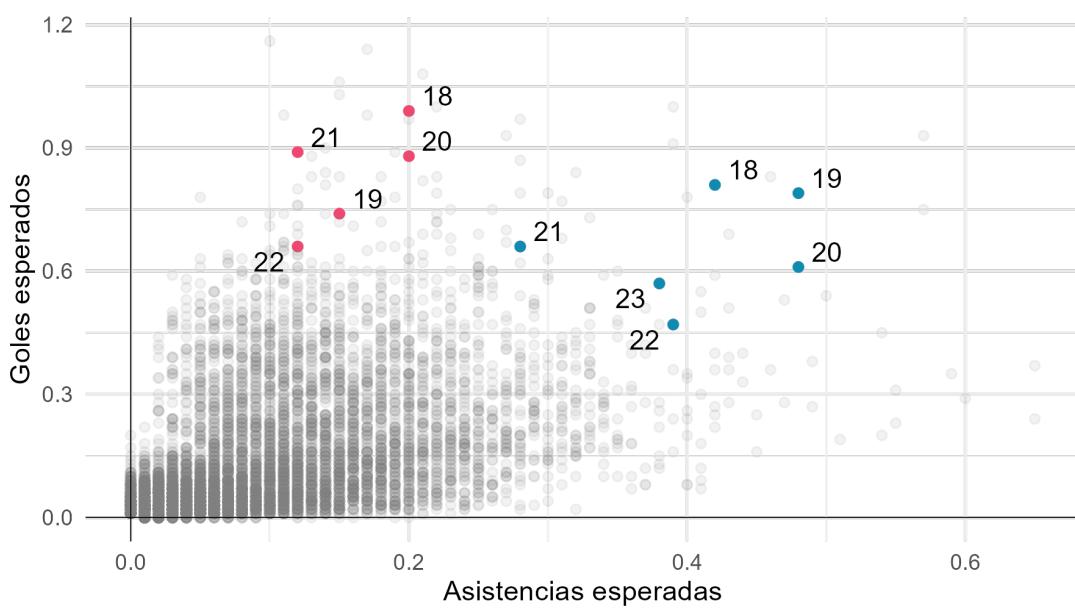
La métrica de goles esperados permite también hablar de efectividad. Cuando un jugador suma más goles con sus disparos de lo que era esperable su efectividad es alta; cuando por el contrario el jugador termina anotando menos goles de los que se preveían por sus disparos su efectividad es baja. En la Fig. ?? se muestra para cada jugador y temporada esta relación. Se vuelve a observar cómo Cristiano y Messi destacan en la generación de goles esperados, aunque hay una ligera diferencia: entre 2018 y 2021 la efectividad del argentino fue mayor que la del portugués. Los puntos de Cristiano se sitúan sobre la línea que representa lo esperado: mismo número de goles que probabilidad de que los disparos acaben en gol. Los de Messi se sitúan por encima, ha anotado más goles que los que sus disparos hacían prever.

```

expected_data |>
  select(season_end_year, player, highlighted_player, label,

```

Goles y asistencias esperadas por 90 minutos cada temporada de **Cristiano Ronaldo**, **Lionel Messi** y el resto de jugadores



Fuente: Fbref.com

Figura 55.3: Goles y asistencias por jugador y temporada

```

Goles = gls_per, `Goles esperados` = x_g_per) |>
ggplot(aes(x = `Goles esperados`, y = Goles)) +
geom_point(aes(alpha = highlighted_player,
               color = player)) +
geom_text_repel(aes(label = str_sub(label, 3, 4))) +
scale_color_manual(values = c("Lionel Messi" = "#118ab2",
                             "Cristiano Ronaldo" = "#ef476f")) +
geom_hline(yintercept = 0, size = 0.2) +
geom_vline(xintercept = 0, size = 0.2) +
geom_abline(slope = 1) +
geom_text(x = 1, y = 1.4,
          label = "Por encima de la línea\nlos jugadores más efectivos",
          size = 3, hjust = 1, vjust = 0.5) +
geom_curve(x = 1.01, y = 1.4, xend = 1.2, yend = 1.2,
           size = 0.2, curvature = -0.25, arrow = arrow(length = unit(0.02, "npc")))
           +
scale_alpha_manual(values = c(0.1, 1)) +
scale_x_continuous(limits = c(0, 1.5)) +
scale_y_continuous(limits = c(0, 1.5)) +
labs(title = "Goles esperados y conseguidos por 90 minutos cada temporada de<br><b>Cristiano Ronaldo</b>, <b>Lionel Messi</b> y el resto de jugadores",
     x = "Goles esperados", y = "Goles", caption = "Fuente: Fbref.com") +
guides(alpha = "none") +
theme_minimal() +
theme(legend.position = "none",
      plot.title = element_markdown(size=12),
      legend.title = element_blank())

```

Con estos gráficos se puede hacer una primera evaluación de los datos de estos dos grandes jugadores (y de cualquier otro) y quizás no logremos contestar a la pregunta de quién ha sido el mejor, aunque para algunos con esto ya esté claro.

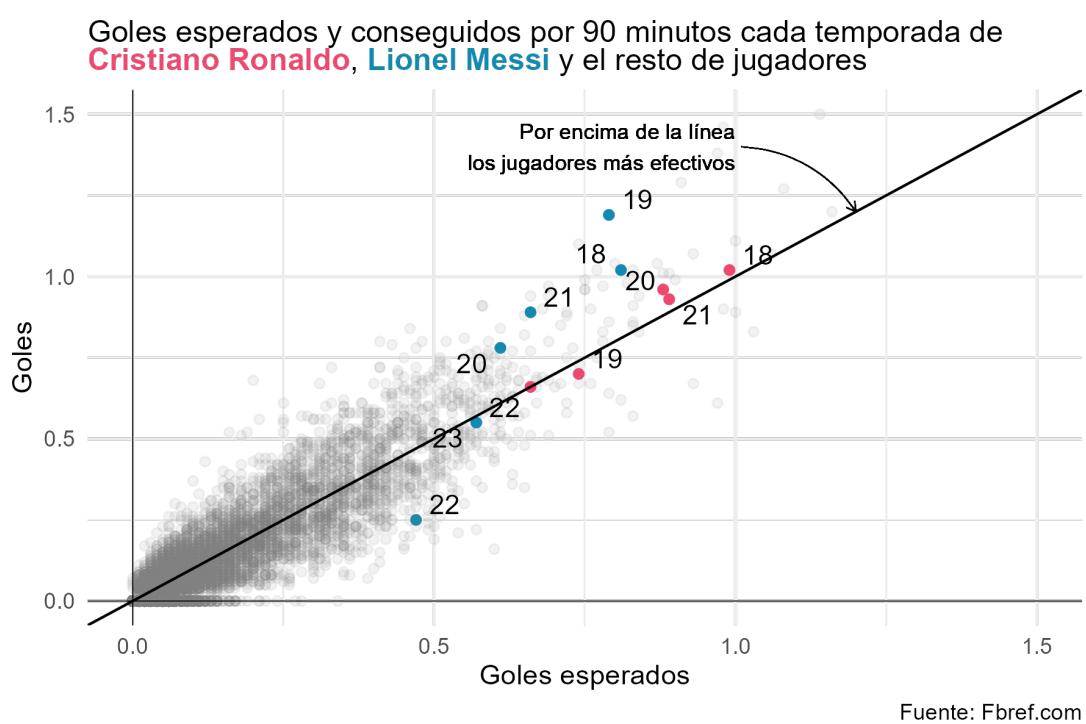


Figura 55.4: Goles esperados y anotados por jugador y temporada

Capítulo 56

Un dato sobre el cambio climático

Dominic Royé

Fundación de la Investigación del Clima

56.1. Introducción

La temperatura media global en la superficie ha aumentado en 1.1 °C desde la era preindustrial (1880-1900). A pesar de parecer un leve incremento en la temperatura, implica un aumento significativo en el calor acumulado del sistema tierra. Cuando se combinan el aumento de la temperatura con respecto a la superficie terrestre y el océano, la tasa promedio es de 0,08 °C por década desde 1880. Sin embargo, la tasa promedio de aumento desde 1981 ha sido más del doble: con 0,18°C. Los océanos se caracterizan por una menor tasa de calentamiento debido a su capacidad calorífica. No obstante, son los océanos los que absorben la mayoría del calor adicional del planeta debido al cambio climático¹.

Entre todas las regiones, la región mediterránea se está calentando un 20% más rápido que el promedio mundial. Este lugar representa actualmente el punto crítico más importante del cambio climático, donde se percibe un significativo aumento de las vulnerabilidades. La temperatura de las aguas superficiales en el Mediterráneo ha estado subiendo 0,34°C cada década desde principios de los 80 ([Cramer et al. \(2020\)](#)).

En este caso práctico con datos sobre el cambio climático se tratan las anomalías de la temperatura superficial del mar Mediterráneo en los meses estivales desde 1982 a 2022. Se hará uso del dataset con el nombre “NOAA CDR OISST v02r01”, una interpolación óptima diaria de temperatura superficial del mar (OISST, por sus siglas en inglés) con una resolución de 1/4 grados (27 km). Los datos los proporciona la *National Oceanic and Atmospheric Administration*

¹<https://www.ncei.noaa.gov/news/global-climate-202112>

(NOAA) con campos completos de temperatura del océano construidos mediante la combinación de observaciones ajustadas por sesgo de diferentes plataformas (satélites, barcos, boyas) en una cuadrícula global regular, con lagunas estimadas por interpolación (https://developers.google.com/earth-engine/datasets/catalog/NOAA_CDR_OISST_V2_1). El geoprocесamiento en nube está explicado en el Cap. @ref{geoprocесes}.

56.2. Consideraciones iniciales

La información espacio-temporal es clave en muchas disciplinas, especialmente en la climatología o la meteorología, y ello hace necesario disponer de un formato que permita una estructura multidimensional. Además, es importante que ese formato tenga un alto grado de compatibilidad de intercambio y pueda almacenar un elevado número de datos. Estas características llevaron al desarrollo del estándar abierto netCDF (*Network Common Data Form*). El formato netCDF es un estándar abierto de intercambio de datos científicos multidimensionales que se utiliza con datos de observaciones o modelos, principalmente en disciplinas como la climatología, la meteorología y la oceanografía. Se trata de un formato espacio-temporal con una cuadrícula regular o irregular. La estructura multidimensional en forma de matriz (array) permite usar no sólo datos espacio-temporales, sino también multidimensionales. Los datos multidimensionales en formato *geotiff* son menos comunes, pero también se pueden llegar a usar. Además, es posible crear objetos multidimensionales importando múltiples archivos ráster.

56.3. Paquetes

El manejo de datos en formato netCDF o múltiples archivos ráster es posible a través de varios paquetes de forma directa o indirecta. Destaca el paquete **ncdf4**, específicamente diseñado para esto, del que hacen uso también otros paquetes de forma oculta. El manejo con **ncdf4** es algo complejo, particularmente por la necesidad de gestionar la memoria RAM cuando se tratan con grandes conjuntos de datos o también por la forma de manejar la clase *array*. Otro paquete muy potente es **terra**, clave en el trabajo con datos ráster y permite usar sus funciones también para el manejo del formato netCDF.

```
# paquetes
library("tidyverse")
library("sf")
library("terra")
library("lubridate")
library("fs")
library("patchwork")
library("giscoR")
library("scales")
library("rmapshaper")
library("RColorBrewer")
library("CDR")
```

56.4. Visualización de mapas “pequeños múltiples”

Una forma muy efectiva para mostrar cambios espacio-temporales son los mapas de pequeños múltiples, donde se representan en una rejilla para cada año las anomalías observadas, lo que permite una comparación sencilla.

56.4.1. Datos

Se importa el polígono del Mar Mediterráneo para limitar los datos al área de interés.

```
data("med_limit")
```

A continuación, se importan todos los años empleando la función `dir_ls()` del paquete `fs` y la función `rast()` de `terra`. La primera función crea un vector de todos los archivos ubicados en la carpeta “data”. Finalmente se renombran todas las capas con los correspondientes años. Es importante que se garantice el correcto orden de los archivos. Siempre que se haya realizado el geoprocесamiento en nube de las anomalías (Cap. @ref{geoproc}) se puede usar la alternativa: `anom <- dir_ls("data-cc", regexp = "tif") |> rast()`.

```
# importar
anom <- dir_ls(system.file("external/data-cc/", package = "CDR"), regexp = "tif") |>
  rast()

# renombrar las capas
names(anom) <- 1982:2022
```

56.4.2. Preparación de los datos

Después de importar el polígono del Mar Mediterráneo, es necesario recortar y enmascarar el área de interés. Para ello, se usa primero la función `crop()` con los límites del Mar Mediterraneo y se pasa el resultado a la función `mask()`. El paquete `terra` necesita los datos vectoriales en su propia clase `SpatVector`, por eso, se pasa con la función `vect()`, que lo convierte de la clase `sf` a `SpatVector`. Finalmente, se reproyectan los rásters a *ETRS89-extended / LAEA Europe* con el código EPSG:3035.

```
anom <- crop(anom, med_limit) |> mask(vect(med_limit))
anom <- project(anom, "EPSG:3035")
```

En la Fig. 56.1 se puede ver el resultados de los primeros años.

```
plot(anom)
```

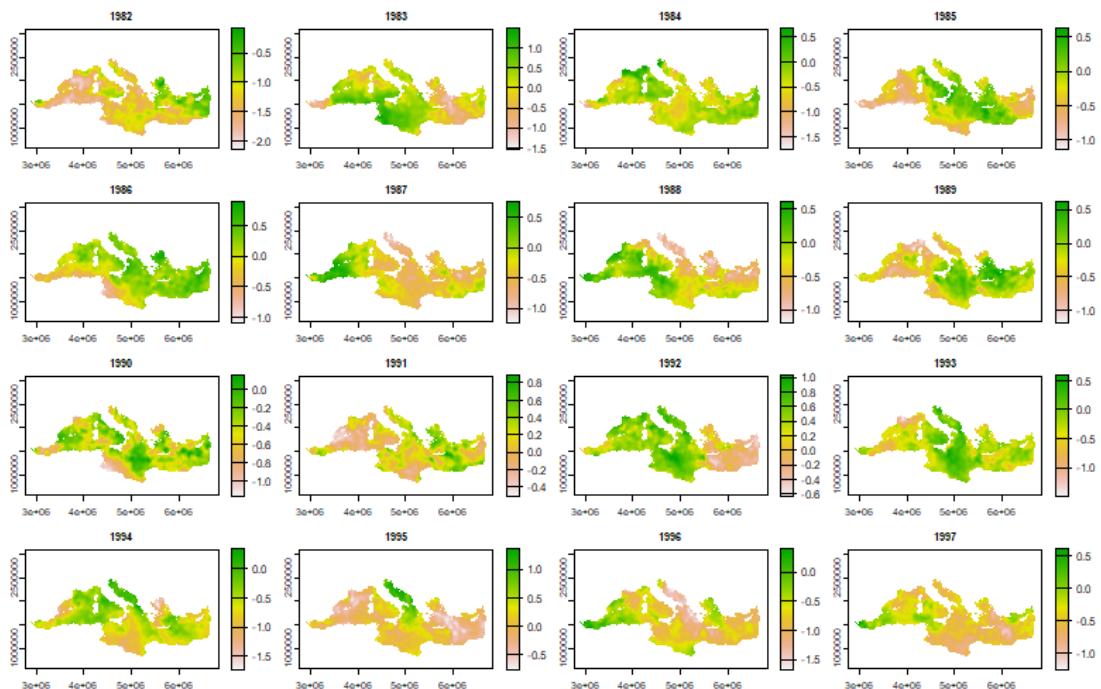


Figura 56.1: Selección de anomalías 1982 a 1997 de los datos brutos.

56.4. Visualización de mapas “pequeños múltiples”

885

Un ráster consiste en latitud, longitud y un único o múltiples valores, también llamados capas. Para poder visualizarlo en `ggplot2`, es necesario convertirlo en un `data.frame`. En este caso, se obtienen 41 columnas para las anomalías, además de las primeras dos que corresponden a la longitud y latitud.

No obstante, es necesario realizar cambios en las distribución de las variables. Ahora mismo se tiene la misma variable, la anomalía, distribuida en muchas columnas, no obstante la estructura adecuada debe ser un conjunto total de dos columnas: una que represente las anomalías y una segunda que contenga los años. Para conseguirlo, se hace uso de `pivot_longer()`, indicando el total de columnas que deben ser fusionadas y los nombres de las dos columnas resultantes.

```
df <- as.data.frame(anom, xy = TRUE)
df <- pivot_longer(df, 3:length(df),
                   names_to = "yr",
                   values_to = "anom")
```

Se añaden los años de la década de los 80 que faltan (1980, 1981) y se limitan las anomalías a un rango entre -2 y +2.

```
df <- bind_rows(df, filter(df, yr == "1982") |>
                  mutate(yr = "1981", anom = NA),
                  filter(df, yr == "1982") |>
                  mutate(yr = "1980", anom = NA)
                ) |>
  mutate(anom2 = case_when(anom > 2 ~ 2,
                           anom < -2 ~ -2,
                           TRUE ~ anom))
```

Previo a la construcción del gráfico, se estima la media de la anomalía global para toda la cuenca mediterránea de cada año. Estos datos se añadirán como texto a cada mapa. Con el objetivo de obtener las coordenadas de la posición en la proyección EPSG:3035, se fija un punto vectorial que reprojyectamos.

```
# media global
med_anom <- global(anom, fun = "mean", na.rm = TRUE)
med_anom <- rownames_to_column(med_anom, "yr")

# posición
pos_global <- st_point(c(34.24, 41.5)) |>
  st_sfc(crs = 4326) |>
  st_transform(3035) |>
  st_coordinates()
```

56.4.3. Construcción del gráfico de múltiples mapas

En el primer paso se definen los estilos partiendo de `theme_void()`, configurando los títulos, la leyenda y el color de fondo en `theme()`.

```
theme_SST_facet <- function(base_family = "Bahnschrift",
  base_size = 11,
  base_line_size = base_size/22,
  base_rect_size = base_size/22) {

  theme_void(base_family = base_family, base_size = base_size,
    base_line_size = base_line_size, base_rect_size = base_rect_size) +
    theme(strip.text = element_text(colour = "white",
      face = "bold",
      size = 12,
      margin = margin(b = 15)),
    legend.text = element_text(colour = "white"),
    legend.position = "top",
    legend.justification = .48,
    plot.margin = margin(20, 20, 20, 20),
    plot.title = element_text(colour = "white", size = 30, hjust = .5),
    plot.subtitle = element_text(colour = "white", size = 15, hjust = .5,
      margin = margin(t = 5, b = 5)),
    plot.caption = element_text(colour = "white", size = 10, hjust = 0),
    plot.background = element_rect(fill = "grey10", colour = NA),
    panel.spacing = unit(2, "lines"),
    panel.background = element_rect(fill = "grey10", colour = NA))
}
```

Para representar datos ráster en forma de xyz se utiliza `geom_tile()` o `geom_raster()` en `ggplot2`. La última geometría requiere una rejilla regular. Para este primer ensayo, se filtra sólo el año 2003, y además se añade, con `geom_sf()`, el límite del Mar Mediterráneo. La función `coord_sf()` permite fijar una proyección para objetos `sf`, y por último, se cambia el estilo definido anteriormente.

```
filter(df, yr == "2003") |>
  ggplot() +
  geom_tile(aes(x, y, fill = anom2)) +
  geom_sf(data = med_limit,
    fill = NA, colour = "white", size = .1) +
  coord_sf(crs = 3035) +
  theme_SST_facet()
```

Siguiendo el ejemplo, se modifica la gama de colores con `scale_fill_gradientn()`, en la que se pasa la paleta de colores, los extremos de valores, se reajustan los valores a una escala divergente se definen las etiquetas y sus posiciones. Dentro de la función `guides()`, se cambia el ancho y altura de la barra colores empleando la función `guide_colorbar()`.

Las geometrías `geom_point()` y `geom_text()` añadirán la información de la anomalía global. La posición se pasa de forma directa en `aes()`; además, se definen el color y el tamaño de texto. El objetivo es situar el texto a la derecha del punto. Por esa razón, es necesario un ajuste en longitud indicando un valor correspondiente en el argumento `nudge_x` en la unidad del sistema de coordenadas (SC). Recuérdese que el SC está en metros.

56.4. Visualización de mapas “pequeños múltiples”

887

La función `number()` del paquete `scales` facilita formatear las cifras con un decimal y los símbolos negativo y positivo.

```
# gama de colores
rdbu_pal <- rev(brewer.pal(11, "RdBu"))

# mapa 2003
filter(df, yr == "2003") |>
ggplot() +
  geom_tile(aes(x, y, fill = anom2)) +
  geom_sf(data = med_limit,
          fill = NA,
          colour = "white",
          size = .1) +
  geom_point(data = filter(med_anom, yr == "2003"),
             aes(x = pos_global[1,1], y = pos_global[1,2], fill = mean),
             size = 3.5, shape = 21, colour = "white") +
  geom_text(data = filter(med_anom, yr == "2003"),
            aes(x = pos_global[1,1], y = pos_global[1,2],
                label = number(mean, .1, style_positive = "plus")),
            size = 3.5, nudge_x = 700000, colour = "white") +
  scale_fill_gradientn(colours = rdbu_pal, na.value = NA,
                        values = rescale(c(-2, 0, 2)),
                        limits = c(-2, 2),
                        breaks = c(-2, -1.5, -1, -0.5, 0,
                                   .5, 1, 1.5, 2),
                        labels = c("< -2.0", "-1.5", "-1.0", "-0.5", "0.0",
                                   "0.5", "1.0", "1.5", "> 2.0")) +
  guides(fill = guide_colorbar(barwidth = 20,
                               barheight = .5)) +
  coord_sf(crs = 3035) +
  theme_SST_facet()
```

En los datos se ha añadido dos años con valores perdidos (1980 y 1981) con el objetivo de obtener por cada fila 10 años, evitando que el *facet grid* empiece por 1982 sin posibilidad de mantener en cada fila la década correspondiente. No obstante, para que los límites de la cuenca mediterránea no aparezca en las facetas de los años 1980/81, se debe repetir la geometría para todos los años.

```
med <- slice(med_limit, rep(1, 41)) |>
  dplyr::select(geometry) |>
  mutate(yr = as.character(1982:2022))
```

A continuación, se construye todo el gráfico con todas las facetas de mapas. Lo único nuevo es la función `facet_wrap()`, en la que se indica la variable por la que se crean las facetas. A diferencia de `facet_grid()`, esta variante permite fijar el número de filas y/o columnas. Además, se pasa una función menor en la función `labeller()` en el mismo argumento. Esta

función permite modificar las etiquetas de las facetas (aquí únicamente el texto de los años 1980/81).

```
# paso 1
g <- ggplot(df) +
  geom_tile(aes(x, y, fill = anom2)) +
  geom_sf(data = med, fill = NA, colour = "white", size = .1) +
  geom_point(data = med_anom,
             aes(x = pos_global[1,1], y = pos_global[1,2], fill = mean),
             size = 3.5, shape = 21, colour = "white") +
  geom_text(data = med_anom,
            aes(x = pos_global[1,1], y = pos_global[1,2],
                label = number(mean, .1, style_positive = "plus")),
            size = 3.5, nudge_x = 700000, colour = "white") +
  scale_fill_gradientn(colours = rdbu_pal,
                        na.value = NA, values = rescale(c(-2, 0, 2)),
                        limits = c(-2, 2),
                        breaks = c(-2, -1.5, -1, -0.5, 0,
                                   .5, 1, 1.5, 2),
                        labels = c("< -2.0", "-1.5", "-1.0", "-0.5", "0.0",
                                   "0.5", "1.0", "1.5", "> 2.0")) +
  guides(fill = guide_colorbar(barwidth = 20,
                               barheight = .5)) +
  facet_wrap(yr ~ .,
             ncol = 10,
             labeller = labeller(yr = function(lab){
               ifelse(lab %in% c("1980", "1981"), "", lab)}))
```

Finalmente, se combinan el objeto con definiciones finales, como los títulos, el sistema de coordenadas y el estilo. Es importante indicar *clip = “off”*, dado que en caso contrario se cortan visualmente los valores de las anomalías globales al encontrarse fuera de los límites de cada mapa.

```
# paso 2
g <- g + labs(title = "ANOMALÍA ESTIVAL DE LA TEMPERATURA DE SUPERFICIE DEL\nMar
→ Mediterráneo", subtitle = "Periodo de referencia 1982-2010.", fill = "") +
  coord_sf(crs = 3035, clip = "off") +
  theme_SST_facet()
```

A priori, no sería necesario una ampliación del resultado. No obstante, en ocasiones se requiere un mapa de orientación.

56.4.4. Mapa de orientación

A través del paquete **giscoR** se obtienen los límites administrativos, de los que únicamente se queda con una selección. También se limita la extensión a aproxidamente la de la cuenca

56.4. Visualización de mapas “pequeños múltiples”

889

mediterránea. La función `ms_innerlines()` del paquete `rmapshaper` facilita la obtención de los límites compartidos o interiores de los países seleccionados. Los nombres de los países, en forma de código ISO-3, se incluyen con ayuda de `geom_sf_text()`.

```
# límites de países
countries_med <- gisco_get_countries() |>
  filter(ISO3_CODE %in% c("ESP", "MAR", "FRA", "ITA",
    "GRC", "TUR", "DZA", "TUN",
    "LBY", "EGY", "ALB")) |>
  st_crop(xmin = -6, xmax = 36, ymin = 28, ymax = 45)

# límites internos
innerlimit <- ms_innerlines(countries_med)

# mapa
insetp <- ggplot() +
  geom_sf(data = med, size = .4, colour = NA, fill = "grey90") +
  geom_sf(data = innerlimit, size = .2, colour = "white") +
  geom_sf_text(data = countries_med,
    aes(label = ISO3_CODE),
    size = 2, colour = "white", fontface = "bold", nudge_y = .1) +
  coord_sf(crs = 3035, expand = FALSE) +
  theme_void() +
  theme(plot.background = element_blank(),
    panel.background = element_blank())
```

56.4.5. Exportar mapa final

El mapa de orientación se insertará en el gráfico, como elemento adicional, en la esquina derecha-arriba. El paquete `patchwork` puede ayudar a crear composiciones de distintos gráficos. La función empleada `inset_element()` indica la posición relativa en `xmin`, `ymin`, `xmax`, y `ymax`. Es importante recordar que cualquier modificación del tamaño de impresión (veáse `height` y `width` en `ggsave()`), puede llevar a ajustes en la posición. El argumento `align_to` = “`full`” permite posicionar sobre todo el lienzo.

```
# paso 3
p_final <- g + inset_element(insetp, 0, .75, .25, .95,
  align_to = "full")

ggsave("sst_anom_med2.png",
  p_final,
  bg = "grey10",
  height = 10,
  width = 20,
  unit = "in",
  type = "cairo-png",
  dpi = 400)
```

El resultado final, como gráfico de múltiples mapas, puede verse en la Fig. 56.2. Los mapas muestran claramente el efecto del calentamiento global, siendo los año 2003 y 2022 de mayor anomalía positiva. Destaca el hecho de que no ha habido un año con temperaturas más bajas de lo normal desde el año 1997.

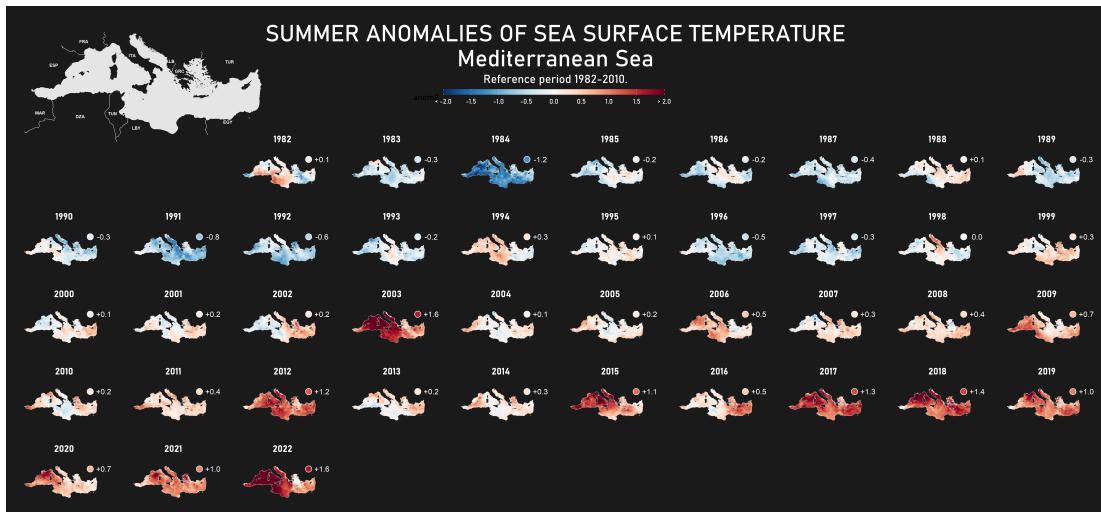


Figura 56.2: Anomalía estival de la temperatura de superficie del mar

Capítulo 57

Predicción de consumo eléctrico con redes neuronales

Jose Manuel Sanz Candales

Red Eléctrica de España

57.1. Introducción

Red Eléctrica, como Operador del Sistema, tiene como principal misión garantizar la continuidad del suministro eléctrico en España. Para ello, entre otras muchas tareas, se desarrollan, evolucionan y mantienen algoritmos de previsión del consumo eléctrico y de la producción con las principales energías renovables (eólica y solar) para distintos horizontes (próximas horas, días, meses, años, etc.) y a distintas escalas temporales (anual, horario, quinceminal).

Este caso de uso se sitúa en el departamento de Ciencia de Datos del Operador del Sistema. Es el principio del año 2018, y el área de planificación de la empresa solicita una **predicción de el consumo eléctrico en España** para el año actual y el siguiente (2018 y 2019).

IMPORTANTE: Este desarrollo no está previsto en el presupuesto del año, por lo que tanto el software como los datos de entrada deben ser, a ser posible, gratuitos.

57.2. Datos de entrada

Respecto a los datos de entrada para el modelo, se requiere tanto una serie histórica de la variable a predecir así como de otras variables que sean capaces de explicar adecuadamente el comportamiento del consumo eléctrico. En este caso es necesario utilizar un modelo de aprendizaje supervisado de regresión, dado que el consumo eléctrico es una variable numérica continua.

En cuanto a la serie histórica de consumo eléctrico anual, Red Eléctrica, en su Web corporativa, publica datos estadísticos accesibles de forma abierta. Sin embargo, el histórico publicado comienza en 2012 y sería conveniente tener un periodo de tiempo más amplio para un entrenamiento adecuado de los modelos potencialmente candidatos a ser utilizados. Se realiza una búsqueda de otras fuentes y, afortunadamente, se encuentra que el Instituto para la Diversificación y Ahorro de la Energía (en adelante IDAE) publica datos desde 1990, con agregación anual, del consumo final de energía eléctrica en miles de toneladas equivalentes de petróleo -ktep-.

Como los datos se deben entregar en MWh, las unidades de la predicción resultante se tendrán que convertir con el coeficiente que indican en la Web del IDAE (1 MWh = 0,086 tep), pero en los modelos se utilizarán las unidades originales porque más adelante se comprobará que dichas unidades resultan muy útiles para ver cómo se relaciona el consumo eléctrico con la variable predictora que se va a utilizar.

De este modo se consigue, por tanto, una parte de los datos necesarios para entrenar los modelos predictivos: **la serie histórica de nuestra variable target (o variable a predecir)**.

Para completar el conjunto de datos del modelo se necesitan, además, **las features o variables explicativas**. Se sabe que, históricamente, las variaciones interanuales de el consumo eléctrico dependen del comportamiento de la economía de una forma directa: si la economía crece, también crece el consumo eléctrico. Como indicador del comportamiento de la economía se decide tomar el PIB per cápita que se puede encontrar en el siguiente enlace, disponible de forma pública en Expansión - datos macro: <https://datosmacro.expansion.com/pib/espana>.

Adicionalmente, se utilizarán otras dos variables explicativas, relacionadas con el mercado inmobiliario y con el empleo, respectivamente. Dado que no son públicas, están anonimizadas y escaladas entre 0 y 1 (dividiendo todos los valores de cada variable entre el mayor de su serie). Debido a que se dispone de datos de estas dos variables desde el año 2000, el dataset comienza en este año.

Una vez definido el dataset de entrada para los modelos (como se verá a continuación, es un conjunto de datos muy pequeño y sencillo), se puede comenzar a construir el modelo en **R**.

57.3. Modelización

En la siguiente celda de código se lee el conjunto de datos, `consumoelectricoanual_2`, del paquete CDR, se convierte al formato `data.table` a `data.frame` y se visualizan sus primeras 3 filas:

```
library(CDR)
df <- CDR::consumoelectricoanual_2
class(df) <- class(as.data.frame(df))
head(df, 3)
#>   Año    PIB Consumo     Inmob     Empleo
#> 1 2000 15.97 16.205 0.7525093 0.6561190
#> 2 2001 17.20 17.279 0.7687356 0.6832698
#> 3 2002 18.09 17.671 0.7836143 0.7104178
```

En este caso, ya se dispone de los datos reales de 2018 y 2019 (esto permitirá validar la precisión del modelo), pero en un caso real, a principios de 2018 el PIB per cápita de 2018 y 2019 será una predicción. Lógicamente, el consumo eléctrico anual también será desconocido, ya que es lo que se necesita predecir. Es decir, se supone que los datos de consumo eléctrico de 2018 y 2019 para el modelo que se va a construir no existen, y no se pueden utilizar ni para entrenar ni para evaluar la precisión del modelo (para ello habrá que utilizar datos pasados).

La siguiente celda de código proporciona la matriz de varianzas-covarianzas de las variables PIB, Consumo, Inmob, Empleo:

```
cormat <- round(cor(df[c("PIB", "Consumo", "Inmob", "Empleo")]), 2)
head(cormat)
#>           PIB Consumo Inmob Empleo
#> PIB     1.00   0.81  0.90  0.93
#> Consumo 0.81   1.00  0.64  0.71
#> Inmob    0.90   0.64  1.00  0.99
#> Empleo   0.93   0.71  0.99  1.00
```

Como se puede observar en la matriz anterior, la correlación entre la variable a predecir y las distintas variables explicativas es fuerte y positiva (es decir, cuando crece una también crece la otra). Si, además, se visualiza la gráfica entre PIB y Consumo en el tiempo, se apreciará de forma aún más clara esta intensa correlación:

```
library("ggplot2")
library("reshape2")
df_m <- melt(df[c("Año", "PIB", "Consumo")], id.vars = "Año")
options(repr.plot.width = 15, repr.plot.height = 8)
ggplot(df_m, aes(Año, value, col = variable)) +
  geom_line(size = 2.5)
```

En la Fig. 57.1 se observa que las curvas que representan la evolución temporal de ambas variables están prácticamente superpuestas, pero desde 2006 líneas se separan. ¿A qué puede deberse? Uno de los principales motivos probablemente sean las medidas de eficiencia energética que se han ido introduciendo en los últimos lustros (iluminación led, electrodomésticos, dispositivos con menor consumo, etc.).

Una vez se han explorado los datos (en este caso ha sido muy breve, pero es muy habitual en proyectos reales que la exploración y limpieza de los datos requiera en torno al 80 % del tiempo), se procederá a dividir el conjunto de datos (desde 2000 hasta 2017, ya que 2018 y 2019 son los años a los que se pretende dar respuesta, por lo que se supone que no es conocido todavía) en dos partes:

- (I) entrenamiento+validación (90 % de las filas)
- (II) test (10 % restante)

Previamente a esto, se debe escalar también la variable PIB a valores entre 0 y 1, para que la red neuronal funcione de forma correcta:

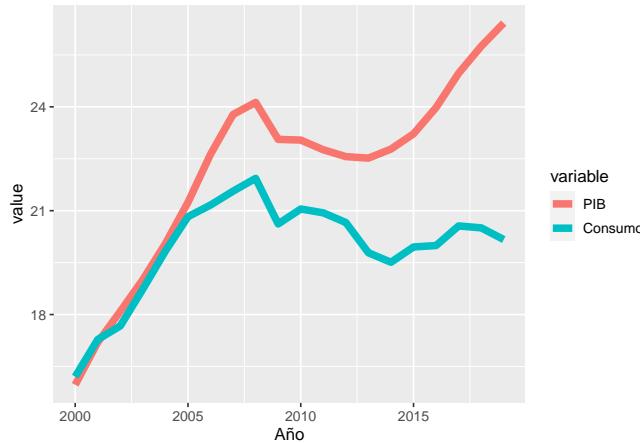


Figura 57.1: Evolución del PIB y el Consumo

```
# estandarizar la variable explicativa "PIB" entre 0 y 1
df$PIB = df$PIB/max(df$PIB)
set.seed(123)
df_aux <- df[df$Año < 2018, ]
n <- nrow(df_aux)
trainIndex <- sample(1:n, size = round(0.85 * n), replace = FALSE)
df_train <- df_aux[trainIndex, ]
df_test <- df_aux[-trainIndex, ]
df_test
```

Ahora se deberían probar distintos modelos de machine learning y comparar sus resultados para determinar cuál es el más preciso para este conjunto de datos. En este ejemplo, por simplicidad no se incluye este proceso de prueba y comparación entre distintos modelos, que en el caso de uso real da lugar elegir una red neuronal simple o perceptrón multicapa (véase Cap. 36), también conocido por su acrónimo en inglés MLP (*Multi Layer Perceptron*), que se utilizará más adelante en este capítulo al obtener los mejores resultados.

Para elegir los hiperparámetros que mejor resultado obtienen para el modelo se van a utilizar dos técnicas que son *grid search* (para probar distintas combinaciones de hiperparámetros) y *cross validation* (para entrenar y validar aprovechando todos los registros del conjunto de entrenamiento-validación).

En este caso de uso, se van a hacer distintas pruebas combinando el número de neuronas por cada capa oculta. En concreto, en la primera y la segunda capa oculta se va a dejar un número constante de neuronas (5), y es en la tercera capa oculta donde se va a probar con 4 neuronas y 6 neuronas. Es decir, se entrenará un modelo con 5 neuronas en cada capa oculta y otro con 5 neuronas en las dos primeras capas y 6 neuronas en la tercera capa.

En la siguiente celda, se importan los paquetes necesarios (neuralnet y caret), se construye la estructura de la red en la variable 'grid' y se define el número de *folds* (en cuántas partes se

divide en conjunto de entrenamiento para entrenar y validar con todos los datos del conjunto) de la validación cruzada. Por último, se entrena el modelo.

El proceso está muy simplificado para que sea fácil de entender. No obstante, lo habitual en la práctica es probar más opciones de *grid search* y hacer una división mayor del conjunto de datos para *cross validation* (es bastante habitual entre 5 y 10 folds):

```
# lee paquetes
library("neuralnet")
library("caret")
# define la estructura de la red
grid <- expand.grid(layer1 = c(5), layer2 = c(5), layer3 = c(4,5))
# establece semilla para que los resultados del entrenamiento sean siempre los mismos
set.seed(123)
# define el número de folds en validación cruzada
train_control <- trainControl(method = "cv",
                                 number = 2,
                                 verbose = TRUE)
# entrenar el modelo
model <- train(Consumo ~ PIB+Inmob+Empleo,
                data = df_train,
                trControl = train_control,
                method = "neuralnet",
                tuneGrid = grid)
```

Para mostrar los resultados se aplica la función ‘print’ sobre la variable ‘model’, cuya salida es el texto comentado debajo de la línea de ‘print’.

```
print(model)
#> Neural Network
#>
#> 15 samples
#> 3 predictor
#>
#> No pre-processing
#> Resampling: Cross-Validated (2 fold)
#> Summary of sample sizes: 8, 7
#> Resampling results across tuning parameters:
#>
#>   layer3  RMSE      Rsquared  MAE
#>   4       0.9713547  0.7631649 0.8521702
#>   5       0.9452518  0.72532010.8165410
#>
#> Tuning parameter 'layer1' was held constant at a value of 5
#> Tuning parameter 'layer2' was held constant at a value of 5
#> RMSE was used to select the optimal model using the smallest value.
#> The final values used for the model were layer1 = 5, layer2 = 5 and layer3 = 5.
```

El modelo con mejor resultado (menor error en el conjunto de entrenamiento / validación) es

el que tiene 3 capas con 5 neuronas cada una. Con este modelo, se predice el consumo eléctrico con el modelo entrenado para la parte del conjunto de datos que se habían reservado para test (años 2006 y 2007), para comprobar que el modelo generaliza bien (es decir, para datos nuevos los resultados de las predicciones tienen un error del orden de los que resultan del entrenamiento del modelo). Para ello, se calcula, por ejemplo, el MAE de las predicciones para dicho conjunto de test:

```
df_test[c("Año", "Consumo")]
#> Año      Consumo
#> 7 2006 21.163
#> 8 2007 21.564
#> 18 2017 20.559

predict(model, df_test)
#>       7       8       18
#> 21.50816 21.83445 20.12596

MAE_test = (abs(21.163-21.50816)+abs(21.564-21.83445)+abs(20.559-20.12596))/3
MAE_test
```

El modelo seleccionado ya está listo para realizar predicciones de consumo eléctrico para los años solicitados (2018 y 2019). Como se avanzó anteriormente, el objetivo del conjunto de test es comprobar la precisión del modelo con datos totalmente desconocidos para él (es decir, no utilizados en la fase de entrenamiento-validación), principalmente para asegurar que el modelo funciona bien para datos distintos a los utilizados en el entrenamiento.

Para predecir el consumo eléctrico anual para 2018 y 2019, que es el dato que solicitaron desde el área de planificación de la empresa. Simplemente se utiliza la función `predict()` del modelo. Previamente, añade una columna en el conjunto de datos original -df- que contendrá los valores predichos, para más adelante comprobar gráficamente el valor predicho frente al real que, en este caso ficticio, ya es conocido:

```
df["Prediccion_MLP"] <- NA
```

Ahora se hace la predicción del año 2018 y se añade el resultado a esta nueva columna del conjunto de datos:

```
df_pred_2018 <- df[df$Año == 2018, ]
df$Prediccion_MLP[df$Año == 2018] <- predict(model,
→ df_pred_2018[c("PIB", "Inmob", "Empleo")])
```

Se hace también la predicción para el año 2019 y se visualizan las predicciones añadidas al conjunto de datos para ambos años:

57.3. Modelización

897

```
df_pred_2019 <- df[df$Año == 2019, ]
df$Prediccion_MLP[df$Año == 2019] <- predict(model,
  ~ df_pred_2019[c("PIB","Inmob","Empleo")])
tail(df, 3)
#>   Año Consumo Prediccion_MLP
#> 18 2017 20.559      NA
#> 19 2018 20.504  20.22787
#> 20 2019 20.166  20.16409
```

Estos son los datos que se entregarían como resultado de la petición de información (convertidos a MWh aplicando el coeficiente que se mencionó en la Secc. 57.2).

Claro está, en el momento que se entregan las predicciones para 2018 y 2019 todavía no se sabría cómo de precisas han sido, pero a principios de 2020 sí es posible calcular la bondad del modelo seleccionado, y es lo que se hará en las siguientes celdas:

```
df_m_mlp <- melt(df[c("Año","Consumo","Prediccion_MLP")], id.vars = "Año")
ggplot(df_m_mlp, aes(Año, value, col = variable)) +
  geom_point(size = 2) +
  geom_line()
```

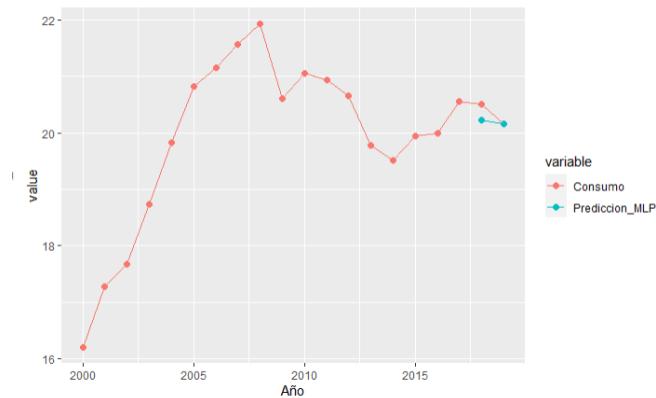


Figura 57.2: Consumo y predicción del modelo de red neuronal MLP

En la Fig. 57.2, las predicciones (puntos azules) tienen unos errores del orden de los que se habían visto en el conjunto de test cuando se hizo el entrenamiento de los modelos, por lo que parece que no hay sobreentrenamiento en el modelo.

Capítulo 58

Implementación de un sistema experto en el ámbito pediátrico

Arturo Peralta^{a,b}, José Ángel Olivas^a y Eusebio Angulo^a

^aUniversidad Internacional de Valencia, ^bUniversidad Internacional de la Rioja

58.1. Introducción

Sin lugar a duda, el análisis de situaciones complejas para la evaluación y toma de decisiones es un proceso para el que tradicionalmente se requiere el apoyo de un especialista dispuesto a poner en uso todo su conocimiento. Sin embargo, el desarrollo de sistemas automáticos capaces de modelar el conocimiento que un experto podría tener sobre un ámbito concreto, y de procesarlo para alcanzar una respuesta adecuada a una consulta relacionada, resulta cada día más extendido como mecanismo de ayuda. A este tipo de herramientas se les denomina Sistemas Expertos (SE).

En este capítulo se introducen los conceptos teóricos fundamentales de la Ingeniería del Conocimiento, los componentes y el funcionamiento de los SE para, posteriormente, presentar cómo su aplicación puede apoyar en el proceso de evaluación clínica en el ámbito pediátrico de atención primaria. Finalmente, se incluye una sencilla implementación en **R** del SE enfocado a la ayuda en esta problemática.

58.2. Marco teórico

Un **Sistema Experto (SE)** es un programa de ordenador que trata de emular el comportamiento de una persona experta en un dominio de conocimiento específico ante un problema que se plantee en dicho dominio y cómo llega a su solución.

900 Capítulo 58. Implementación de un sistema experto en el ámbito pediátrico

La **Ingeniería del Conocimiento** se ocupa, entre otras cosas, del proceso de especificación, análisis y desarrollo de un sistema experto ([Martínez R., 2005](#)).

Los principales componentes de un SE son: (i) La **Base de Hechos (BH)**, que contiene la definición del entorno sobre el que se van a resolver problemas. Hace el papel de “ojos” del SE. (ii) La **Base de Conocimientos (BC)**, que contiene la información del dominio específico, convenientemente representado, capaz de resolver problemas. También puede considerar la representación de incertidumbre. (iii) El **Motor de Inferencias**, que es el proceso de razonamiento que usa el SE. Combina hechos y conocimiento para emitir una conclusión. (iv) El **Interfaz de entrada/salida** para comunicarse con los usuarios y/o expertos.

En la Fig. 58.1 se muestra un diagrama con los principales componentes de un SE.

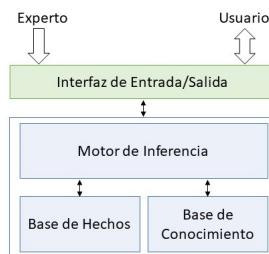


Figura 58.1: Componentes de un SE

Las principales limitaciones en la construcción de SE vienen dadas porque el conocimiento experto humano es experiencia compilada, es heurístico, esto es, basado en experiencia y en reglas prácticas. Es incompleto, impreciso e incierto, y a veces inconsistente y con errores o imprecisiones. Es por ello que las limitaciones de todo SE pueden ser que no conoce lo que conoce ni por qué, carece de imaginación, emociones, inteligencia innata, sentido común, etc. Tiene poco conocimiento de sí mismo, del usuario y del contexto de cada interacción y capacidad de razonamiento limitada por su estrategia de construcción.

En este contexto, los **sistemas de producción** son modelos de cálculo que han probado su eficiencia en la Ingeniería del Conocimiento tanto en el desarrollo de algoritmos de búsqueda como en el modelado de problemas del dominio humano. Sus componentes principales son:

- 1) Las **reglas de producción**: son la forma más extendida de representar el conocimiento, constan de Condiciones (Hipótesis) y Acciones (Conclusiones) y tienen la forma:

```

Si (If)
  Condición 1
  y Condición 2
  ...
  y Condición n
Entonces (Then)
  Conclusión 1
  Conclusión 2
  
```

...
y Conclusión m

Es la forma más extendida de representar el conocimiento. Ejemplo de Regla de producción:

```
Si
  ha fallado la bombilla
  y hay una de repuesto
  y está útil
Entonces
  cambiar la bombilla por la de repuesto
  y seguir trabajando
```

- 2) **Memoria de trabajo:** contiene una descripción del estado actual del mundo o entorno de la aplicación en cada paso del proceso de razonamiento. Esta descripción es un modelo que servirá para asociar los antecedentes de las reglas con las observaciones del mundo, con el objetivo de seleccionar o producir las acciones apropiadas. En el momento en que se cumplen todas las condiciones de una regla se produce el “disparo” de la misma, ejecutándose la acción. Esta operación alterará el contenido de la memoria de trabajo.
- 3) **Ciclo de reconocimiento y actuación:** es el procedimiento de control de un sistema de producción. Es un procedimiento de feedforward o hacia adelante. La memoria de trabajo se inicializa con la descripción del problema. Los modelos almacenados en la memoria de trabajo se tratan de superponer en las condiciones de las producciones. Tras ello, se crea un conjunto “conflicto”, es decir, un subconjunto de producciones cuyas condiciones se cumplen. Se escoge una producción y se “dispara” o se activa. La acción de la regla es “disparada” cambiando el contenido de la memoria de trabajo. Se repite todo el proceso descrito con la memoria de trabajo modificada. El proceso continúa hasta que no haya condiciones en las reglas que cumplan el contenido de los modelos de la memoria de trabajo.

Una de las principales ventajas de los sistemas de producción en los SE es la separación del conocimiento y del control. Se pueden hacer cambios fáciles de reglas sin cambiar el control y viceversa. Otra es la modularidad de las reglas de producción y la independencia del lenguaje de programación usado.

58.2.1. Razonamiento

El razonamiento se define como el proceso de obtención de inferencias o conclusiones a partir de unos hechos u observaciones reales o asumidos y de un conocimiento previo. La inferencia es el proceso por el que a partir de unos hechos conocidos se obtienen conclusiones acerca de otros desconocidos ([Fleitas, 2017](#)) y Begu04. La realización de este tipo de procesamiento es llevada a cabo por el denominado **motor de inferencia**.

902 Capítulo 58. Implementación de un sistema experto en el ámbito pediátrico

El razonamiento automático ya se utilizaba en los 50 en juegos. En 1963 se presentó el sistema “General Problem Solver” capaz de hacer inferencias lógicas (Newel y Simon).

Tipos de razonamiento en SE:

- **Forward chaining** (encadenamiento hacia delante, deductivo, progresivo, dirigido por datos o hechos): Síntomas → Causas
- **Backward chaining** (encadenamiento hacia atrás, inductivo, regresivo, dirigido por metas u objetivos): Síntomas ← Causas

Pasos del motor de inferencia:

1. Elaboración de un **conjunto conflicto** con todas las reglas cuyas condiciones se cumplen.
2. **Detección** (filtro) de reglas pertinentes o **selección** de reglas a partir de unos hechos. Se trata de obtener de la Base de Conocimiento (BC) el conjunto de reglas aplicables en una situación determinada o estado de la Base de Hechos (BH).
3. **Aplicación de reglas o resolución del conflicto.** Consiste en seleccionar una regla del conjunto conflicto y dispararla (ejecutar su conclusión). Se altera la BH o memoria de trabajo incluyendo el consecuente de la regla “disparada”.
4. **Vuelta a 1** hasta que el conjunto **conflicto** esté **vacío**.

Ciclo de “razonamiento hacia delante”:

1. Parte de unas observaciones (**hechos**).
2. A partir de los hechos observados, se **seleccionan las reglas** cuyas condiciones están relacionadas con estos.
3. Las reglas seleccionadas son examinadas para ver si cumplen todas sus condiciones. Aquellas que las verifican constituyen el “**conjunto conflicto**”.
4. Del total de reglas que forman el conjunto conflicto se selecciona una sola y se activa (se “dispara”). La selección de una regla del conjunto conflicto se denomina “**resolución del conflicto**”
5. La activación de la regla provocará la **aparición de otros hechos** que se añaden a los observados y se **actualiza** la base de hechos.
6. Volver al **paso 2** hasta analizar todos los hechos observados y deducidos.

En la Fig. 58.2 se muestra el algoritmo correspondiente al ciclo de inferencia hacia adelante.

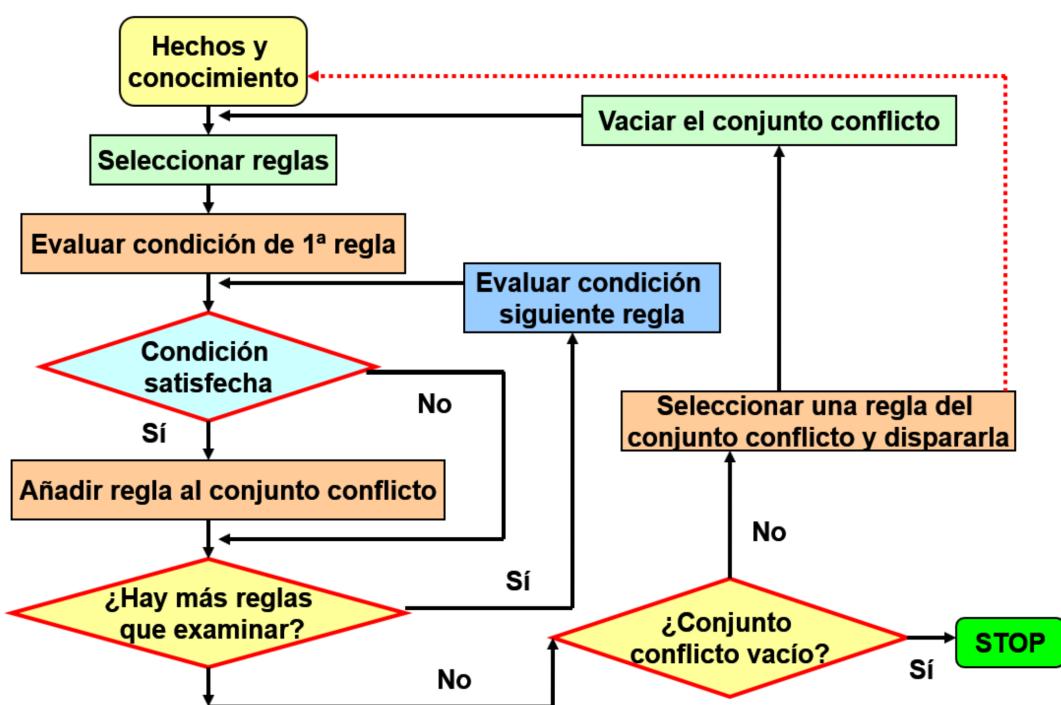


Figura 58.2: Ciclo de razonamiento hacia delante

58.3. Sistema experto para el ámbito pediátrico en atención primaria

En la actualidad, uno de los principales problemas a los que se enfrentan los profesionales de la sanidad en el ámbito pediátrico de atención primaria en España es la falta del tiempo suficiente para realizar una evaluación clínica del estado del paciente. La necesidad de un mayor número de médicos especialistas y la aparición de picos de demanda motivados fenómenos como el COVID, o de modo estacional por otras enfermedades recurrentes, favorecen esta situación.

Ante esta problemática, los centros de salud tratan de optimizar sus recursos mediante diferentes vías, poniendo especial interés en realizar procesos de triaje que les permitan priorizar la atención a los pacientes según su nivel de urgencia. Para ello, en España se utilizan escalas como el MTS (Manchester Triage System), el SET (Sistema Español de Triage) y el CPTAS (Canadian Pediatric Triage and Acuity Scale) [Soler2010] para establecer el tiempo que un paciente puede esperar para recibir atención médica en base a sus síntomas y evolución.

Sin embargo, realizar un correcto proceso de triaje, además de requerir de un gran conocimiento experto, hace necesario un tiempo para una evaluación clínica que a veces resulta difícil dedicar. En este contexto, se plantea el desarrollo de un SE capaz de ayudar en el proceso de evaluación médica, con el objetivo de facilitar el proceso de triaje.

El primer paso para el desarrollo de un SE es la definición de un conjunto de reglas que modelen el conocimiento con el que se nutrirá. Para ello, es posible recurrir al apoyo de expertos, capaces de definir su conocimiento como reglas, o la aplicación de mecanismos de extracción de conocimiento a partir del procesamiento de conjuntos de datos y sucesos.

En este ejemplo, se recopilaron y procesaron un conjunto de datos relativos a los motivos de consulta pediátrica en un centro de salud, donde la escala de triaje utilizada fue el CPTAS. Los datos fueron anonimizados, seleccionando únicamente aquellos campos que pudieran resultar clave para la extracción de conocimiento y la conformación de reglas. Adicionalmente, se contó con el apoyo de profesionales especialistas del ámbito pediátrico para revisar y complementar algunas de las reglas extraídas. Un extracto de los datos utilizados para el proceso de extracción de reglas se muestra en la Tabla 58.1.

Tabla 58.1: . Ejemplo de datos de motivos de consulta y triaje.

Sexo	Edad	Tiempo Evolución	Causa	Triaje
Hombre	1-3 años	25-72 horas	Dermatológica	5 (120 minutos)
Mujer	1-3 años	25-72 horas	Respiratoria	3 (30 minutos)
Hombre	4-6 años	7-12 horas	Gastrointestinal	5 (120 minutos)
Hombre	4-6 años	2-6 horas	Ocular	4 (60 minutos)
Mujer	4-6 años	13-24 horas	Fiebre	5 (120 minutos)

58.3. Sistema experto para el ámbito pediátrico en atención primaria

905

A partir del procesamiento de un total de 400 visitas médicas mediante un algoritmo de extracción de reglas de asociación como “Magnum Opus”, basado en la definición original de [Webb \(2011\)](#), y mediante la aplicación del conocimiento de experto proporcionado por un panel de pediatras, se extrajeron un conjunto de reglas con suficiente calidad. A continuación, se muestra un ejemplo de un conjunto de reglas con 10 de ellas.

- R1: Si Causa Ginecologica o Edad mayor de 12 años → Tiempo de Evolución mayor de 73h
- R2: Si Causa Ginecologica → Sexo Mujer
- R3: Si Edad menor de 7 días o Causa Fiebre → Tiempo de evolución de 1h
- R4: Si Edad mayor de 12 años y Tiempo de Evolución mayor de 73h → Causa Respiratoria
- R5: Si Tiempo Evolución mayor de 73h y Causa Ocular → Triaje 1
- R6: Si Causa Respiratoria y Sexo Mujer → Triaje 3
- R7: Si Tiempo de Evolución es 2-6h o Causa Neurológica → Triaje 2
- R8: Si Causa Respiratoria y Tiempo de Evolución es 2-6h → Triaje 4
- R9: Si Tiempo de Evolución es 13-24h y Causa Ginecológica y Sexo Mujer → Triaje 5
- R10: Si Tiempo de Evolución mayor de 73h → Triaje 4

A continuación, se muestra el código en **R** para la implementación de un SE capaz de procesar reglas como las anteriores, realizando una ejecución para obtener el valor de triaje correspondiente. Para este ejemplo se considera una niña mayor de 12 años de edad, cuyo motivo de consulta es ginecológico con un tiempo de evolución superior a 73 horas.

Es importante señalar que, habitualmente, un SE partirá de una base de hechos compuesta por decenas o cientos de reglas. No obstante, para simplificar el siguiente fragmento de código, se considera únicamente la carga de 10, reglas a modo de ejemplo.

Se declaran las reglas que conformarán la base conocimiento del SE:

1. La BC del SE contiene 10 reglas.
2. Cada regla se modela como una lista de antecedentes y un consecuente.
3. La relación entre los consecuentes se modela con el atributo “operador” del siguiente modo: 1 = Y lógico, 0 = O lógico, -1 = no hay operaciones

```
r1 <- list(
  antecedentes = list("Causa_Ginecologica", "Edad_>12a"),
  consecuente = list("TiempoEvolucion_>73h"), operador = 0
)
r2 <- list(
```

906 Capítulo 58. Implementación de un sistema experto en el ámbito pediátrico

```

    antecedentes = list("Causa_Ginecologica"),
    consecuente = list("Sexo_Mujer"), operador = -1
)
r3 <- list(
    antecedentes = list("Edad_<7d", "Causa_Fiebre"),
    consecuente = list("TiempoEvolucion_1h"), operador = 0
)
r4 <- list(
    antecedentes = list("Edad_>12a", "TiempoEvolucion_>73h"),
    consecuente = list("Causa_Respiratoria"), operador = 1
)
r5 <- list(
    antecedentes = list("TiempoEvolucion_>73h", "Causa_Ocular"),
    consecuente = list("Triaje_1"), operador = 1
)
r6 <- list(
    antecedentes = list("Causa_Respiratoria", "Sexo_Mujer"),
    consecuente = list("Triaje_3"), operador = 1
)
r7 <- list(
    antecedentes = list("TiempoEvolucion_2-6h", "Causa_Neurologica"),
    consecuente = list("Triaje_2"), operador = 0
)
r8 <- list(
    antecedentes = list("Causa_Respiratoria", "TiempoEvolucion_2-6h"),
    consecuente = list("Triaje_4"), operador = 1
)
r9 <- list(
    antecedentes = list("TiempoEvolucion_13-24h", "Causa_Ginecologica", "Sexo_Mujer"),
    consecuente = list("Triaje_5"), operador = 1
)
r10 <- list(
    antecedentes = list("TiempoEvolucion_>73h"),
    consecuente = list("Triaje_4"), operador = -1
)

# r1 regla completa
# r1[1] lista antecedentes
# r1[[1]] [1] primero de los antecedentes
# r1[2] lista consecuentes
# r1[[2]] [1] primero de los consecuentes

b_hechos <- list("Causa_Ginecologica", "Edad_>12a", "TiempoEvolucion_>73h")

```

Se inicializa la BC con el conjunto de Reglas y la BH con la circunstancia a evaluar.

```

# Se añaden todas las reglas a la Base de Conocimiento
b_conocimiento <- list(r1, r2, r3, r4, r5, r6, r7, r8, r9, r10)

```

58.3. Sistema experto para el ámbito pediátrico en atención primaria

907

Se define una función para comprobar la existencia de un número en una lista. Esta función será usada por el motor del SE.

```
# Función para comprobar si una lista contiene un número

contiene <- function(numero, lista) {
  existe <- FALSE
  if (length(lista) > 0) {
    for (i in 1:length(lista)) {
      if (numero == lista[[i]]) existe <- TRUE
    }
  }
  return(existe)
}
```

Se implementa el motor del SE.

El algoritmo ejecuta un bucle en el que, en cada iteración, evalúa las reglas disponibles contenidas en la *BC*. Considerando los items de la *BH*, si una regla puede ser “disparada”, se añade al *Conjunto Conflicto*. La regla “disparada” en una iteración será la primera disponible en el *Conjunto Conflicto*. El *Conjunto Conflicto* se inicializa en cada iteración. El consecuente de la regla “disparada” se añade a la *BH*. Cada regla solo puede “dispararse” una vez, por lo que se actualiza una lista de reglas “disparadas”. El algoritmo finaliza cuando el *Conjunto Conflicto* queda vacío, al haber sido “disparadas” todas las reglas o no existir más candidatas a ser “disparadas”.

```
# Motor del Sistema Experto

AlgoritmoSE <- function(b_hechos, b_conocimiento) {
  c_conflicto <- list()
  r_disparadas <- list()
  condicion <- TRUE
  iteracion <- 0
  while (condicion) {
    c_conflicto <- list()
    cat("Iteración: ", iteracion, "\n")
    iteracion <- iteracion + 1
    for (i in 1:length(b_conocimiento)) {
      if (!contiene(i, r_disparadas)) {
        if (b_conocimiento[[i]][[3]][[1]] == 1) {
          r_disparada <- TRUE
        } else {
          r_disparada <- FALSE
        }
        for (j in 1:length(b_conocimiento[[i]][[1]])) {
          antecedente <- FALSE
          for (k in 1:length(b_hechos)) {
            if (b_conocimiento[[i]][[1]][[j]] == b_hechos[[k]]) {
```

```

    if (b_conocimiento[[i]][[3]][[1]] == 0 || b_conocimiento[[i]][[3]][[1]]
    ↵ == -1) r_disparada <- TRUE
    antecedente <- TRUE
  }
}
if (b_conocimiento[[i]][[3]][[1]] == 1) {
  if (!antecedente) r_disparada <- FALSE
}
}
if (r_disparada) {
  cat("Regla", i, "añadida a Conjunto conflicto\n")
  c_conflicto[length(c_conflicto) + 1] <- i
}
}
if (length(c_conflicto) > 0) {
  # str("Conjunto conflicto:")
  # str(c_conflicto)
  r_disparadas[length(r_disparadas) + 1] <- c_conflicto[1]
  cat("Regla", r_disparadas[[length(r_disparadas)]], "disparada\n")
  b_hechos[length(b_hechos) + 1] <-
← b_conocimiento[[r_disparadas[[length(r_disparadas)]]]][[2]][[1]]
  str("Base de hechos:")
  str(b_hechos)
  cat("Consecuente:",
  ↵ b_conocimiento[[r_disparadas[[length(r_disparadas)]]]][[2]][[1]], "\n")
} else {
  condicion <- FALSE
}
}
}

```

Ahora considerese, por ejemplo, una posible paciente con más de 12 años de Edad, que acude a consulta por causa Ginecológica con un Tiempo de Evolución de los síntomas mayor de 73h. ¿Cuál será el triaje correspondiente? Para conocerlo, se inicializa la *BH* con la situación propuesta (*Causa_Ginecologica*, *Edad_>12a* y *TiempoEvolucion_>73h*) y se ejecuta el motor del SE.

```

# Se inicializa la base de hechos con la situación propuesta (paciente de más de 12
→ años, por causa ginecológica con tiempo de evolución mayor de 73h)
b_hechos <- list("Causa_Ginecologica", "Edad_>12a", "TiempoEvolucion_>73h")

# Se lanza la ejecución del motor del Sistema Experto
AlgoritmoSE(b_hechos, b_conocimiento)

```

El resultado del algoritmo, en este caso (motivo de consulta ginecológico y mayor de 12 años y con un tiempo de evolución de los síntomas mayor de 73h), es un triaje de valor 4. Es decir,

58.3. Sistema experto para el ámbito pediátrico en atención primaria

909

el tiempo de espera máximo para recibir atención médica debería ser inferior a 60 minutos. La Tabla 58.3 muestra el proceso realizado por el algoritmo en cada iteración.

IT	Base de Hechos	Conjunto Conflicto	Regla disparada	Consecuente
0	Causa_Ginecologica Edad_>12a TiempoEvolucion_>73h	R1, R2, R4, R10	R1	TiempoEvolucion_>73h
1	Causa_Ginecologica Edad_>12a TiempoEvolucion_>73h Sexo_Mujer"	R2, R4, R10	R2	Sexo_Mujer
2	Causa_Ginecologica Edad_>12a TiempoEvolucion_>73h Sexo_Mujer Causa_Respiratoria	R4, R10	R4	Causa_Respiratoria
3	Causa_Ginecologica Edad_>12a TiempoEvolucion_>73h Sexo_Mujer Causa_Respiratoria Triage_3	R6, R10	R6	Triage_3
4	Causa_Ginecologica Edad_>12a TiempoEvolucion_>73h Sexo_Mujer Causa_Respiratoria Triage_3 Triage_4	R10	R10	Triage_4
5	Causa_Ginecologica Edad_>12a TiempoEvolucion_>73h Sexo_Mujer Causa_Respiratoria Triage_3 Triage_4	Ø		

Table: . Proceso de ejecución del Sistema Experto.

Como puede observarse, el algoritmo finaliza tras 5 iteraciones, al alcanzar un conjunto conflicto vacío, dando como resultado un valor de 4 para el triaje. A continuación, se describe el proceso ejecutado en cada una de las iteraciones.

- **Iteración 0:** La Base de Hechos se inicializa con las condiciones establecidas en el ejemplo de consulta médica considerado, es decir, **Causa_Ginecologica**, **Edad_>12a** y **TiempoEvolucion_>73h**. En el Conjunto Conflicto se incluyen aquellas reglas que podrían ser lanzadas con los elementos contenidos en la BH es decir, las reglas R1, R2, R4, R10. Se dispara la regla R1, al ser la primera de la lista de reglas del Conjunto Conflicto. El consecuente de la regla disparada (**TiempoEvolucion_>73h**) se añade a la BH para la siguiente iteración, aunque en este caso no es necesario porque ya lo contiene. La iteración finaliza estableciendo como conclusión el consecuente de la regla disparada, es decir, **TiempoEvolucion_>73h**.
- **Iteración 1:** El Conjunto Conflicto se inicializa con las reglas que podrían ser lanzadas, excluyendo las ya ejecutadas (R1), a partir de los elementos incluidos en la Base de Hechos tras la iteración anterior, es decir, las reglas R2, R4 y R10. Se lanza la primera de las reglas del Conjunto Conflicto, es decir, R2. El consecuente de la regla disparada (**Sexo_Mujer**) se añade a la BH. La iteración finaliza estableciendo como conclusión el consecuente de la regla disparada, es decir, **Sexo_Mujer**.

910 *Capítulo 58. Implementación de un sistema experto en el ámbito pediátrico*

- **Iteración 2:** El Conjunto Conflicto se inicializa con las reglas que podrían ser lanzadas, excluyendo las ya ejecutadas (R1 y R2), a partir de los elementos contenidos en la BH tras la iteración anterior, es decir, las reglas R4 y R10. Se lanza la primera de las reglas del Conjunto Conflicto, es decir, R4. El consecuente de la regla disparada (*Causa_Respiratoria*) se añade a la BH. La iteración finaliza estableciendo como conclusión el consecuente de la regla disparada, es decir, *Causa_Respiratoria*.
- **Iteración 3:** El Conjunto Conflicto se inicializa con las reglas que podrían ser lanzadas, excluyendo las ya ejecutadas (R1, R2 y R4), a partir de los elementos contenidos en la BH tras la iteración anterior, es decir, las reglas R6 y R10. Se lanza la primera de las reglas del Conjunto Conflicto, es decir, R6. El consecuente de la regla disparada (*Triaje_3*) se añade a la BH. La iteración finaliza estableciendo como conclusión el consecuente de la regla disparada, es decir, *Triaje_3*.
- **Iteración 4:** El Conjunto Conflicto se inicializa con las reglas que podrían ser lanzadas, excluyendo las ya ejecutadas (R1, R2, R4 y R6), a partir de los elementos contenidos en la BH tras la iteración anterior, en este caso únicamente R10. Se lanza por tanto la regla R10. El consecuente de la regla disparada (*Triaje_4*) se añade a la BH. La iteración finaliza estableciendo como conclusión el consecuente de la regla disparada, es decir, *Triaje_4*.
- **Iteración 5:** El Conjunto Conflicto queda vacío, al no ser posible incluir en él reglas no disparadas aún y que pudieran ser ejecutadas a partir de los elementos contenidos en la BH. Por tanto, el algoritmo finaliza concluyendo como resultado el consecuente de la última regla lanzadas, es decir, *Triaje_4*.

Sin lugar a duda, la implementación del algoritmo, las reglas y el modelado considerado para este caso de estudio resulta una simplificación intencionada del problema, con el único objetivo de facilitar su comprensión desde una perspectiva académica y docente.

En un escenario de aplicación real, el volumen y la complejidad de las reglas debería ser mayor, considerando además la posibilidad de incorporar otras características que permitieran modelar de un modo más completo el estado de salud del paciente y su motivo de consulta.

Actualmente resultan innumerables los ámbitos donde la aplicación de SE puede ser de ayuda. En este caso de estudio, el objetivo ha sido facilitar y mejorar el proceso de triaje del nivel de urgencia en el ámbito pediátrico en atención primaria. Pero, en este mismo contexto, podría valorarse su uso como herramienta de apoyo para, por ejemplo, el diagnóstico de enfermedades o la prescripción de tratamientos.

La tecnología actual y lenguajes de programación como R facilitan la implementación de SE de un modo rápido y sencillo. Por tanto, pueden ser considerados como herramienta de ayuda para situaciones en las que aplicar conocimiento experto resulte clave para la solución de problemas.

58.3. Sistema experto para el ámbito pediátrico en atención primaria

911

Tabla 58.2: Proceso de ejecución del Sistema Experto

It	Base Hechos	Conjunto conflicto	Regla disparada	Conclusión
0	<i>Causa_Ginecologica</i> <i>Edad_ < 12a</i> <i>TiempoEvoluc_ > 73h</i>	R1, R2, R4, R10	R1	<i>TiempoEvoluc_ > 73h</i>
1	<i>Causa_Ginecologica</i> <i>Edad_ < 12a</i> <i>TiempoEvoluc_ > 73h</i> <i>Sexo_Mujer</i>	R2, R4 R10	R2	<i>Sexo_Mujer</i>
2	<i>Causa_Ginecologica</i> <i>Edad_ > 12a</i> <i>TiempoEvoluc_ > 73h</i> <i>Sexo_Mujer</i> <i>Causa_Respiratoria</i>	R4, R10	R4	<i>Causa_Respiratoria</i>
3	<i>Causa_Ginecologica</i> <i>Edad_ > 12a</i> <i>TiempoEvoluc_ > 73h</i> <i>Sexo_Mujer</i> <i>Causa_Respiratoria</i> <i>Triaje_3</i>	R6, R10	R6	<i>Triaje_3</i>
4	<i>Causa_Ginecologica</i> <i>Edad_ > 12a</i> <i>TiempoEvoluc_ > 73h</i> <i>Sexo_Mujer</i> <i>Causa_Respiratoria</i> <i>Triaje_3</i> <i>Triaje_4</i>	R10	R10	<i>Triaje_4</i>
5	<i>Causa_Ginecologica</i> <i>Edad_ > 12a</i> <i>TiempoEvoluc_ > 73h</i> <i>Sexo_Mujer</i> <i>Causa_Respiratoria</i> <i>Triaje_3</i> <i>Triaje_4</i>	Ø		

Capítulo 59

El procesamiento del lenguaje natural para tendencias de moda en textil

Ambrosio Nguema Ansue

59.1. Introducción

El Procesamiento del Lenguaje Natural (NLP, por sus siglas en inglés), abarca una amplia gama de técnicas y algoritmos, entre los que se encuentra el modelado de temas. El modelado de temas no es un modelo de predicción en sí mismo. En cambio, es una técnica de aprendizaje no supervisado que tiene como objetivo descubrir estructuras ocultas (temas) dentro de un conjunto de documentos o textos aunque está relacionado con el NLP, no son lo mismo, el modelado de temas es una de las muchas técnicas que forman parte del NLP. La relación entre ambos radica en que el modelado de temas utiliza enfoques del NLP para analizar y procesar el lenguaje en los textos, pero se enfoca en una tarea específica: extraer temas. En este capítulo, exploraremos cómo el modelado de temas y otras técnicas de NLP pueden aplicarse al análisis de tendencias en el mundo de la moda. El modelado de temas aplicado a la industria textil puede proporcionar información valiosa sobre las preferencias y opiniones de los clientes, lo que puede mejorar la toma de decisiones y la experiencia del cliente en el ámbito del comercio electrónico de ropa.

59.2. Análisis de tendencias de moda en textil

El conjunto `clothes` de datos incluido en el paquete `CDR` de reseñas y calificaciones de ropa de comercio electrónico para mujeres contiene 23.486 entradas relacionadas con la edad y la

914 Capítulo 59. El procesamiento del lenguaje natural para tendencias de moda en textil

revisión dada por el cliente y sus opiniones sobre la ropa de mujer de varios minoristas.

```
library("CDR")
library("readr")
library("tidyverse")
library("tidytext")
```

Las variables incluidas pueden verse con la ejecutando `names(clothes)` y una descripción de las variables con el comando `??clothes`. El primer registro presenta la siguiente estructura la información:

```
head(clothes)[1, ]
#>   ID Age Title                               Review Rating
#> 1 767 33 <NA> Absolutely wonderful - silky and sexy and comfortable 4
#>   Recommend Liked Division      Dept      Class
#> 1           1          0 Initmates Intimate Intimates
```

El conjunto de datos consta de 23.486 entradas que incluyen información acerca de la edad del cliente, las calificaciones otorgadas y las opiniones sobre la ropa comprada en comercios electrónicos para mujeres. Los datos se organizan en columnas, algunas de las cuales contienen valores enteros y otras almacenan caracteres. Todas las columnas con valores enteros están completas, mientras que algunas columnas de caracteres presentan valores faltantes (NA). La variable con la mayor cantidad de valores NA Título.

En el presente capítulo, se explora la aplicación de técnicas de análisis de texto en un conjunto de datos de reseñas y calificaciones de ropa de comercio electrónico para mujeres. En primer lugar, se realiza un análisis del porcentaje de reseñas y calificaciones en cada departamento, destacando los departamentos con mayor y menor porcentaje. Además, se lleva a cabo un análisis de bigramas para identificar las frases más comunes asociadas con diferentes calificaciones. Finalmente, se utiliza el modelado de temas con *Latent Dirichlet Allocation* (LDA) para explorar las características clave de las revisiones en el departamento de Tendencias. Los resultados del análisis proporcionan información valiosa para las empresas sobre el grupo demográfico objetivo, las preferencias de los clientes y las características clave de las prendas.

```
library("ggplot2")
clothes |>
  dplyr::count(Dept) |>
  dplyr::mutate(prop = n / sum(n)) |>
  ggplot(aes(x = Dept, y = prop * 100)) +
  geom_bar(stat = "identity", fill="blue") +
  xlab("Department Name") +
  ylab("Percentage of Reviews/Ratings (%)") +
  geom_text(aes(label = round(prop * 100, 2)), vjust = -0.25)
```

Los tops y vestidos son los departamentos que cuentan con la mayoría de las reseñas y calificaciones en el conjunto de datos, mientras que las chaquetas y la sección de tendencias tienen

59.2. Análisis de tendencias de moda en textil

915

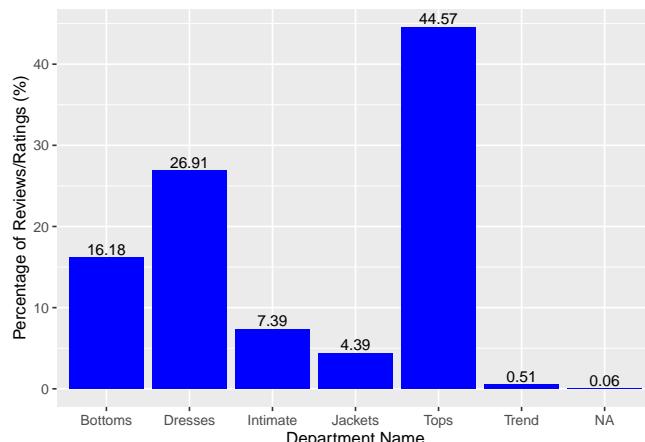


Figura 59.1: Percentage of Reviews by Department

la menor cantidad. Dado que la sección de tendencias presenta una mezcla de ropa que puede pertenecer a otros departamentos, y solo representa un 0,51% del conjunto de datos, se ha decidido excluir esta sección del análisis.

```
clothes |>
  filter(!is.na(Dept), Dept != "Trend") |>
  mutate(Dept = factor(Dept)) |>
  group_by(Dept, Rating) |>
  summarize(n = n()) |>
  mutate(perc = n / sum(n)) |>
  ggplot(aes(x = Rating, y = perc * 100, fill = Dept)) +
  geom_bar(stat = "identity", show.legend = FALSE) +
  facet_wrap(~Dept) +
  ylab("Percentage of reviews (%)") +
  geom_text(aes(label = round(perc * 100, 2)), vjust = -.2) +
  scale_y_continuous(limits = c(0, 65))
```

Se ha observado que en todos los departamentos, la calificación de 5 estrellas es la más común. A pesar de tener una menor cantidad de reseñas en general, las chaquetas tienen la mayor proporción de calificaciones de 5 estrellas en su categoría. Una posible razón de esto es que las chaquetas suelen ser más fáciles de ajustar a diferentes formas corporales en comparación con vestidos y blusas, que pueden ser más difíciles de adaptarse correctamente, especialmente cuando se compran en línea.

```
clothes |>
  filter(!is.na(Age), !is.na(Dept), Dept != "Trend") |>
  select(ID, Age, Dept) |>
  mutate(Age_group = cut(Age, breaks = c(18, 29, 39, 49, 59, 69, 79, 89, 99))) |>
  mutate(Age_group = as.character(Age_group)) |>
```

916 Capítulo 59. El procesamiento del lenguaje natural para tendencias de moda en textil

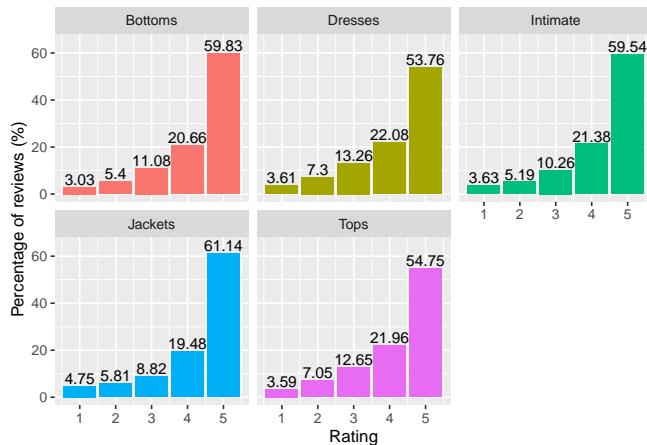


Figura 59.2: Percentage of reviews in each department

```

mutate(Age_group = factor(Age_group, levels = c("18-29", "30-39", "40-49", "50-59",
  ↪ "60-69", "70-79", "80-89", "90-99"))) |>
  mutate(Dept = factor(Dept, levels = rev(c("Tops", "Dresses", "Bottoms", "Intimate",
  ↪ "Jackets")))) |>
  filter(Age < 80) |>
  group_by(Age_group, Dept) |>
  summarize(n = n()) |>
  ggplot(aes(Dept, n, fill = Age_group)) +
  geom_bar(stat = "identity", fill="red") +
  facet_wrap(~Age_group, scales = "free") +
  xlab("Department") +
  ylab("Number of Reviews") +
  geom_text(aes(label = n), hjust = -0.1) +
  scale_y_continuous(expand = c(.1, 0)) +
  coord_flip() +
  scale_fill_manual(values = hcl.colors(8))

```

Se ha observado que la tendencia en la distribución de reseñas por departamento (es decir, tops con el mayor número de reseñas y vestidos con el segundo mayor número) es similar en la mayoría de los grupos de edad. Esto indica que la popularidad de los diferentes tipos de ropa se mantiene en gran medida constante entre los grupos de edad más jóvenes y de mediana edad.

Análisis de bigramas

El análisis de bigramas es una técnica útil para identificar patrones y tendencias en el lenguaje utilizado en las reseñas de productos. Un bigrama es un par consecutivo de palabras en un texto y puede proporcionar información valiosa sobre la frecuencia con la que ciertas palabras aparecen juntas y las combinaciones de palabras que son relevantes en las opiniones de los clientes. Al utilizar el análisis de bigramas, se espera comprender mejor las opiniones de los clientes sobre los productos de ropa en el conjunto de datos.

59.2. Análisis de tendencias de moda en textil

917

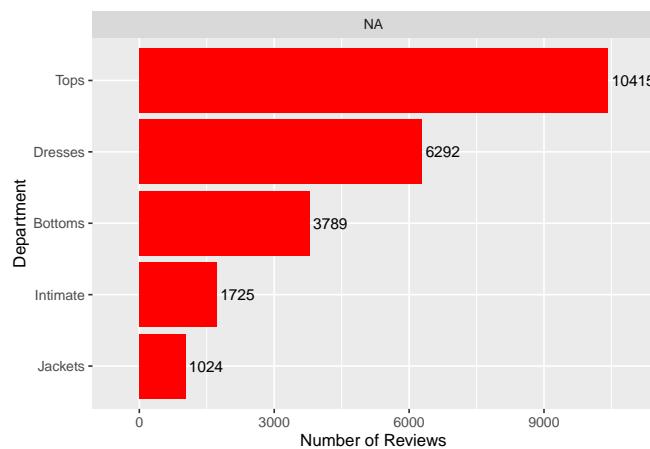


Figura 59.3: Number of Reviews by Department

```

clothesr <- clothes |> filter(!is.na(Review))
notitle <- clothesr |>
  filter(is.na>Title) |>
  select(-Title)
wtitle <- clothesr |>
  filter(!is.na>Title) |>
  unite(Review, c>Title, Review), sep = " "
main <- bind_rows(notitle, wtitle)

```

Para llevar a cabo el análisis de bigramas, se procede a procesar las palabras de las reseñas eliminando las palabras vacías (también conocidas como *stop words*), que son palabras comunes sin un significado contextual importante, y los dígitos que representan la calificación de las reseñas. Una vez procesadas las palabras, se agrupan según sus calificaciones y se representan gráficamente los 10 bigramas más comunes para cada nivel de calificación. De esta forma, se puede identificar y comprender mejor las combinaciones de palabras que son relevantes para las opiniones de los clientes y para cada nivel de calificación.

```

bigramming <- function(data) {
  cbigram <- data |> unnest_tokens(bigram, Review, token = "ngrams", n = 2)
  cbigram_sep <- cbigram |> separate(bigram, c("first", "second"), sep = " ")
  cbigram2 <- cbigram_sep |>
    filter(!first %in% stop_words$word, !second %in% stop_words$word,
      !str_detect(first, "\\\d"), !str_detect(second, "\\\d")) |>
    unite(bigram, c(first, second), sep = " ")
  return(cbigram2)
}

```

918 Capítulo 59. El procesamiento del lenguaje natural para tendencias de moda en textil

```
top_bigrams <- bigramming(main) |>
  mutate(Rating = factor(Rating, levels <- c(5:1))) |>
  mutate(bigram = factor(bigram, levels = rev(unique(bigram)))) |>
  group_by(Rating) |>
  count(bigram, sort = TRUE) |>
  top_n(10, n) |>
  ungroup()

top_bigrams |> ggplot(aes(bigram, n, fill = Rating)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~Rating, ncol = 3, scales = "free") +
  labs(x = NULL, y = "frequency") +
  coord_flip()
```

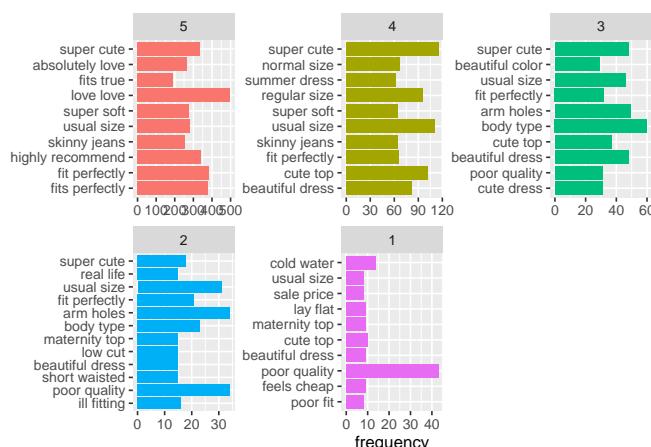


Figura 59.4: Most Common Bigrams (By Ratings)

Modelado de temas con Latent Dirichlet Allocation

El enfoque de modelado de temas de Latent Dirichlet Allocation (LDA) es una técnica ampliamente utilizada en NLP para extraer temas latentes de un corpus de texto. LDA es un algoritmo no supervisado que utiliza el aprendizaje automático para identificar patrones en grandes conjuntos de datos de texto, agrupando palabras similares en temas y asignando probabilidades a cada tema en cada documento.

En este estudio, se ha utilizado el enfoque de modelado de temas de LDA para explorar las 118 revisiones del Departamento de Tendencias. Se ajustó un modelo LDA utilizando muestreo de Gibbs y se eligió $k = 5$ para los departamentos de Bottoms, Dresses, Intimate, Jackets y Tops. A través del análisis de los resultados, se pudieron identificar las 5 palabras principales de cada tema y obtener una mejor comprensión de las características clave de las revisiones en cada departamento. De esta forma, se pudo obtener información valiosa sobre las preferencias y opiniones de los clientes en diferentes departamentos de ropa en el conjunto de datos.

59.2. Análisis de tendencias de moda en textil

919

```

library("topicmodels")
library("tm")
library("LDAdavis")

trend_count <- main |>
  filter(Dept == "Trend") |>
  unnest_tokens(word, Review) |>
  anti_join(stop_words, by = "word") |>
  filter(!str_detect(word, "\\d")) |>
  count(ID, word, sort = TRUE) |>
  ungroup()

trend_dtm <- trend_count |> cast_dtm(ID, word, n)
trendy <- tidy(LDA(trend_dtm, k = 5, method = "GIBBS", control = list(seed = 4444,
                     alpha = 1)), matrix = "beta")
top_trendy <- trendy |>
  group_by(topic) |>
  top_n(5, beta) |>
  ungroup() |>
  arrange(topic, desc(beta))

top_trendy |>
  mutate(term = reorder(term, beta)) |>
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~topic, scales = "free") +
  coord_flip()

```

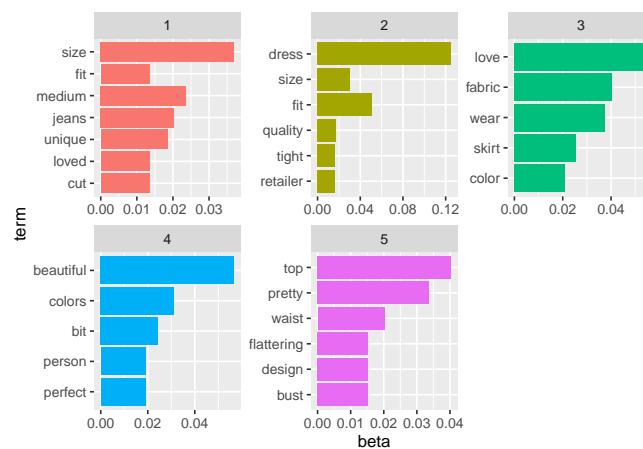


Figura 59.5: Modelo LDA (K=5)

En el modelo LDA, cada tema se representa por un conjunto de palabras que aparecen juntas con mayor frecuencia en las revisiones. Por ejemplo, al observar el tema 3, se puede identificar

920 *Capítulo 59. El procesamiento del lenguaje natural para tendencias de moda en textil*

que está caracterizado por palabras como “colors”, “wear”, “bit”, “jacket” y “price”, lo que sugiere que los clientes pueden estar comentando sobre la variedad de colores disponibles, la durabilidad de la prenda y su precio. Por otro lado, el tema 1 se caracteriza por palabras como “love”, “fit”, “fabric”, “wear” y “length”, lo que sugiere que los clientes pueden estar hablando sobre su experiencia con la prenda en términos de comodidad, ajuste y calidad de la tela. Al identificar estos temas, se pueden obtener ideas valiosas sobre las opiniones y preferencias de los clientes para mejorar la calidad de la ropa y satisfacer sus necesidades y deseos. Esto permite a las empresas tomar decisiones informadas para satisfacer las necesidades de sus clientes y mejorar la experiencia del usuario en el ámbito del comercio electrónico de ropa.

Como conclusión, destacar que el análisis de este conjunto de datos proporciona información valiosa sobre las preferencias y opiniones de los clientes en cuanto a la ropa femenina. Las reseñas de 5 estrellas son dominantes en cada departamento, y las chaquetas son las prendas que obtienen la proporción más alta de calificaciones positivas. Además, se ha observado que los clientes de entre 30 y 40 años dejan la mayoría de las reseñas y que factores como el ajuste, la comodidad/calidad del material y la estética de la prenda influyen en la calificación. La realización de análisis de datos exploratorios y de bigramas puede ayudar a las empresas a comprender mejor lo que funciona y lo que no, y seleccionar artículos con telas flexibles y cómodas puede conducir a una mayor satisfacción del cliente y mayores ventas. Por último, el modelado de temas con LDA es una herramienta útil en situaciones en las que se tienen reseñas sin marcar y puede proporcionar información valiosa sobre las características clave de las revisiones. En general, estos análisis pueden ayudar a las empresas a tomar decisiones informadas y mejorar la experiencia del usuario en el ámbito del comercio electrónico de ropa.

Capítulo 60

Detección de fraude de tarjetas de crédito

Pedro Albarracín García

60.1. Introducción

En un informe publicado en diciembre de 2021 por Nilson Report (https://nilsonreport.com/upload/content_promo/NilsonReport_Issue1209.pdf), se informó de que los emisores de tarjetas de crédito, comerciantes y consumidores sufrieron un total de 28.580 millones de dólares de pérdidas por fraude en 2020, es decir, 6,8 centavos por cada 100 dólares en volumen de compras. El fraude sólo en USA representa el 35,83 % del total mundial.

En Europa la situación no es más alentadora. Según un informe del Banco Central Europeo publicado en 2020 (<https://www.ecb.europa.eu/pub/cardfraud/html/ecb.cardfraudreport202008~521edb602b.en.html#toc2>), el valor total de las transacciones con tarjeta en la zona SEPA ascendieron a 4.84 billones de euros en 2018, de los cuales 1.800 millones correspondieron a operaciones fraudulentas.

Las entidades financieras trabajan a diario en el desarrollo de modelos de machine learning y deep learning que les permitan detectar, con la mayor precisión posible, aquellas operaciones de compra con tarjetas de crédito, débito o prepago que puedan ser sospechosas de fraude, o que al menos puedan ser identificadas como anómalas. En este sentido, es importante destacar que no existe una única solución posible, ya que el problema presenta, en la mayoría de los casos, múltiples variantes que hacen de éste, un problema complejo y que puede y debe ser abordado desde múltiples perspectivas y con diferentes enfoques.

En primer lugar, es posible identificar dos tipos de fraude. Por un lado, el que se comete físicamente, como, por ejemplo, la compra o la retirada de efectivo con tarjetas robadas o falsas. Por otro lado, están aquellas transacciones fraudulentas que se cometan online, en las

que no es necesaria la tarjeta física, y en las que se utilizan los datos de las tarjetas obtenidas por los delincuentes mediante técnicas como el phishing y que son utilizados posteriormente para realizar pagos online.

Otro hándicap asociado a este tipo de escenarios es el derivado de la gran diversidad de fuentes de datos que forman parte de una transacción y que pueden dar lugar a divergencias metodológicas, tanto en la recogida y transmisión de los datos, como en su posterior almacenaje, lo que ocasiona que, en muchos casos, la calidad de los datos disponibles no sea la esperada, o simplemente nos encontrremos ante datasets inconsistentes. Los datos requeridos para este caso de uso pueden categorizarse en variables relativas a:

- Cliente
- Transacción
- Geolocalización
- Comercio
- Tarjetas
- Hábitos de compra

Cada una de estas categorías, y otras que puedan aparecer aportan información que permite abordar el problema desde diferentes ángulos. Por un lado es posible enfocar el problema desde el punto de vista del cliente y sus hábitos de compra para ver si existe alguna característica anómala en una transacción, tal vez la hora de la compra, o quizás analizar los datos de geolocalización junto con los del comercio para analizar si es una compra en un comercio habitual y desde una localización conocida, etc.

60.2. Modelización del fraude en la compra con tarjetas de crédito

El objetivo de este caso de uso es la construcción de un modelo que permita detectar si una operación de compra realizada con tarjeta de crédito es fraudulenta o no. Para ello, se utilizará un dataset anonimizado de operaciones con tarjeta de crédito ya etiquetadas disponible desde la web de Kaggle en el siguiente enlace: <https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud>. No obstante, los datos se han incorporado la paquete CDR del libro con el nombre `creditcard`. El conjunto de datos consta de 284.807 transacciones, de las cuales 492 están etiquetadas como fraudulentas, es decir, sólo un 0,172 % del total de las transacciones. Es un dataset por lo tanto muy desequilibrado, lo que añade cierto grado de dificultad. El dataset `creditcard` tiene un conjunto de 31 variables, de las cuales 28 están identificadas como `V1`, ..., `V28`, una variable `Time` que registra los segundos transcurridos entre esa transacción y la primera, una variable `Amount` que registra el importe de la transacción, y la variable dependiente `class` que indica, con valor 0, que la operación es “no fraudulenta”, y con valor 1 las operaciones fraudulentas.

Para concluir esta breve descripción del dataset, es necesario recordar que todos los valores de entrada son numéricos y que ya han sufrido algunas transformaciones. Por motivos de confiden-

60.2. Modelización del fraude en la compra con tarjetas de crédito

923

cialidad, las variables V1 a V28 no incluyen sus nombres originales ni se añade más información de contexto.

Carga de los datos y obtención de algunos descriptivos

```
creditcard <- CDR::creditcard
head(creditcard)
#>   Time      V1      V2      V3      V4      V5      V6
#> 1 0 -1.3598071 -0.07278117 2.5363467 1.3781552 -0.33832077 0.46238778
#> 2 0 1.1918571 0.26615071 0.1664801 0.4481541 0.06001765 -0.08236081
#> 3 1 -1.3583541 -1.34016307 1.7732093 0.3797796 -0.50319813 1.80049938
#> 4 1 -0.9662717 -0.18522601 1.7929933 -0.8632913 -0.01030888 1.24720317
#> 5 2 -1.1582331 0.87773675 1.5487178 0.4030339 -0.40719338 0.09592146
#> 6 2 -0.4259659 0.96052304 1.1411093 -0.1682521 0.42098688 -0.02972755
#>          V7      V8      V9      V10     V11     V12
#> 1 0.23959855 0.09869790 0.3637870 0.09079417 -0.5515995 -0.61780086
#> 2 -0.07880298 0.08510165 -0.2554251 -0.16697441 1.6127267 1.06523531
#> 3 0.79146096 0.24767579 -1.5146543 0.20764287 0.6245015 0.06608369
#> 4 0.23760894 0.37743587 -1.3870241 -0.05495192 -0.2264873 0.17822823
#> 5 0.59294075 -0.27053268 0.8177393 0.75307443 -0.8228429 0.53819555
#> 6 0.47620095 0.26031433 -0.5686714 -0.37140720 1.3412620 0.35989384
#>          V13     V14     V15     V16     V17     V18
#> 1 -0.9913898 -0.3111694 1.4681770 -0.4704005 0.20797124 0.02579058
#> 2 0.4890950 -0.1437723 0.6355581 0.4639170 -0.11480466 -0.18336127
#> 3 0.7172927 -0.1659459 2.3458649 -2.8900832 1.10996938 -0.12135931
#> 4 0.5077569 -0.2879237 -0.6314181 -1.0596472 -0.68409279 1.96577500
#> 5 1.3458516 -1.1196698 0.1751211 -0.4514492 -0.23703324 -0.03819479
#> 6 -0.3580907 -0.1371337 0.5176168 0.4017259 -0.05813282 0.06865315
#>          V19     V20     V21     V22     V23     V24
#> 1 0.40399296 0.25141210 -0.018306778 0.277837576 -0.11047391 0.06692807
#> 2 -0.14578304 -0.06908314 -0.225775248 -0.638671953 0.10128802 -0.33984648
#> 3 -2.26185710 0.52497973 0.247998153 0.771679402 0.90941226 -0.68928096
#> 4 -1.23262197 -0.20803778 -0.108300452 0.005273597 -0.19032052 -1.17557533
#> 5 0.80348692 0.40854236 -0.009430697 0.798278495 -0.13745808 0.14126698
#> 6 -0.03319379 0.08496767 -0.208253515 -0.559824796 -0.02639767 -0.37142658
#>          V25     V26     V27     V28 Amount Class
#> 1 0.1285394 -0.1891148 0.133558377 -0.02105305 149.62 0
#> 2 0.1671704 0.1258945 -0.008983099 0.01472417 2.69 0
#> 3 -0.3276418 -0.1390966 -0.055352794 -0.05975184 378.66 0
#> 4 0.6473760 -0.2219288 0.062722849 0.06145763 123.50 0
#> 5 -0.2060096 0.5022922 0.219422230 0.21515315 69.99 0
#> 6 -0.2327938 0.1059148 0.253844225 0.08108026 3.67 0
# psych::describe(creditcard) # descomentar para ver los descriptivos
```

División de los datos

A continuación es necesario dividir los datos en dos dataframes que denominados `creditcard_X` y `creditcard_y`, de esta forma se separan las variables independientes de la variable dependiente o `Class`.

```
# Se dividen los datos
creditcard_X <- creditcard[,-31]
creditcard_y <- creditcard$Class
```

Tratamiento de datos desequilibrados

Uno de los principales problemas a la hora de abordar este tipo de escenarios, es lo que se conoce como “datos desequilibrados”. Se dice que un dataset está desequilibrado cuando la variable dependiente presenta más observaciones de una clase que de otra. En el caso de transacciones fraudulentas con tarjeta de crédito, es evidente que la mayoría de las operaciones son legítimas o benignas, y que sólo un pequeño porcentaje resultan ser maliciosas. ¿Cuál es el problema?

Por lo general, los modelos entrenados con datasets desequilibrados no se comportan bien cuando tienen que generalizar, es decir, cuando tienen que realizar predicciones sobre conjuntos de datos que no han sido vistos anteriormente por el modelo. El desequilibrio de los datos es un sesgo hacia la clase mayoritaria, por lo que, en última instancia, muestra una tendencia al sobreajuste u overfitting hacia esa clase. Existen diversas técnicas que permiten corregir esta situación:

- ***Undersampling*** o Submuestreo. Esta técnica consiste en reducir el número de observaciones de la clase mayoritaria, estableciendo quizás una ratio de 60/40. Esta técnica resulta efectiva si se respetan los grupos naturales que existen en los datos, así como el resto de las características presentes en la clase mayoritaria.
- ***Oversampling*** o Sobremuestreo. Esta técnica consiste en aumentar el número de observaciones de la clase minoritaria mediante la creación de datos sintéticos que, al igual que la técnica anterior, respeten las características de esa clase.

Para la creación de datos sintéticos en escenarios de oversampling existen varios algoritmos que proporcionan buenos resultados. Quizás el más conocido y utilizado sea *Synthetic Minority Oversampling TEchnique* (SMOTE). SMOTE no realiza una copia de las observaciones del dataset, sino que en su lugar genera nuevos datos de forma sintética utilizando los vecinos más cercanos de esos casos, respetando las características estadísticas de la clase. Además, los ejemplos de la clase mayoritaria también son submuestreados, lo que da lugar a un conjunto de datos más equilibrado. En R, el algoritmo SMOTE pertenece al paquete **DMwR**.

Para este caso particular, se utilizará una técnica simple de submuestreo basada en el paquete **unbalanced**, que actualmente no se encuentra disponible en el repositorio de CRAN por lo que, para su instalación, se debe ejecutar el siguiente código:

```
#install.packages("devtools") # descomentar para instalar
library("devtools")
devtools::install_github("dalpozz/unbalanced")
library("unbalanced")
```

Una vez instalado, al igual que todas sus dependencias, se realiza el submuestreo del dataset siguiendo los siguientes pasos:

60.2. Modelización del fraude en la compra con tarjetas de crédito

925

- 1.- Convertir la variable dependiente “Class” en factor:

```
creditcard$Class <- as.factor(creditcard$Class)
levels(creditcard$Class) <- c('0', '1')
```

- 2.- A continuación, ejecutar la función de submuestreo:

```
undersampled_creditcard <- ubBalance(creditcard_X, creditcard$Class, type='ubUnder',
→ verbose = TRUE)
#> Proportion of positives after ubUnder : 50% of 984 observations
undersampled_combined <- cbind(undersampled_creditcard$X,
                                 undersampled_creditcard$Y)

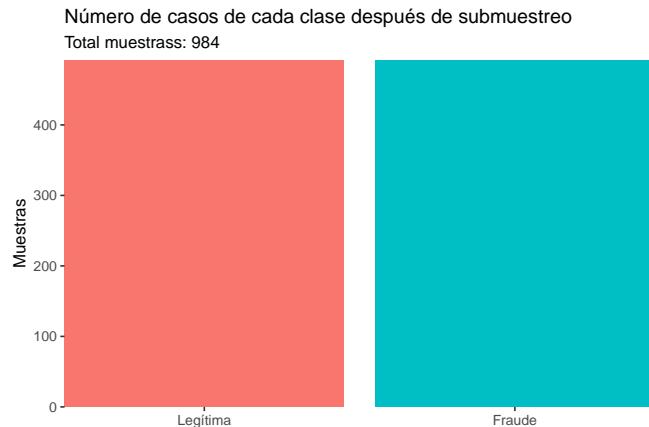
names(undersampled_combined)[names(undersampled_combined) ==
→ "undersampled_creditcard$Y"] <- "Class"
levels(undersampled_combined$Class) <- c('Legítima', 'Fraude')
```

- 3.- Comprobar el número de casos en el dataset sobre el que se ha ejecutado la función de submuestreo

```
creditcard$Class <- as.factor(creditcard$Class)
levels(creditcard$Class) <- c('0', '1')
```

- 4.- Realizar la gráfica para visualizar el número de elementos de cada clase después de realizar el submuestreo

```
library("ggplot2")
ggplot(data = undersampled_combined, aes(fill = Class))+
  geom_bar(aes(x = Class))+
  ggtitle("Número de casos de cada clase después de submuestreo",
          subtitle="Total muestras: 984")+
  xlab("")+
  ylab("Muestras")+
  scale_y_continuous(expand = c(0,0))+
  scale_x_discrete(expand = c(0,0))+
  theme(legend.position = "ninguna",
        legend.title = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.background = element_blank())
```



Modelo de clasificación mediante regresión lógística

A continuación se procederá a la construcción de un modelo de regresión logística (véase Cap. 16) para una clasificación binaria en relación al fraude en transacciones con tarjeta de crédito a partir de los datos equilibrados obtenidos anteriormente. El dataframe que se utilizará, por lo tanto, será “undersampled_combined” que contiene 984 observaciones, un 50 % de las cuales son transacciones identificadas como fraude.

Lo primero, será realizar un par de pequeños cambios en el dataset, es decir, eliminar las variables `Time` y `Amount`, ya que no van a ser relevantes para el modelo, y cambiar por 0 y 1 las etiquetas “Legítima” y “Fraude”, respectivamente.

```
undersampled_combined <- subset(undersampled_combined, select = -c(Time, Amount) )  
undersampled_combined$Class <- ifelse(undersampled_combined$Class == "Fraude", 1, 0)
```

Lo siguiente será dividir el conjunto de datos en los datasets de entrenamiento y test, para lo cual se aplicará la función “`split()`” con un `SplitRatio` de 0.80, es decir, un 80 % de los datos irán de forma aleatoria al dataset de entrenamiento, 788 observaciones, frente a las 196 observaciones que formarán el dataset de testing.

```
#install.packages("caTools") # descomentar para instalar  
library("caTools")  
set.seed(123)  
split = sample.split(undersampled_combined$Class, SplitRatio = 0.80)  
training = subset(undersampled_combined, split == TRUE)  
test = subset(undersampled_combined, split == FALSE)
```

Con los datasets necesarios ya disponibles, el siguiente paso es entrenar el modelo de regresión logística que clasificará las transacciones en legítimas o fraudulentas. Para ello se utilizará el algoritmo GLM, creando un clasificador que se identificará como `undersampledModely` al que se le pasarán los parámetros siguientes:

60.2. Modelización del fraude en la compra con tarjetas de crédito

927

- **formula:** con este parámetro se indica la variable dependiente, class, seguida del simbolo \sim y un punto (con el punto se hace referencia al resto de variables del dataset, es decir, V1 a V28).
- **data:** el dataset con los datos de entrenamiento.
- **family:** al ser un clasificador con dos valores posibles (0, 1), se indica que será de tipo "binomial".

```
undersampledModel = glm(Class ~ ., data = training, family = binomial())
```

Para ver,

```
summary(undersampledModel)
#>
#> Call:
#> glm(formula = Class ~ ., family = binomial(), data = training)
#>
#> Deviance Residuals:
#>    Min      1Q   Median      3Q     Max 
#> -2.6674 -0.1929  0.0000  0.0000  3.2276 
#>
#> Coefficients:
#>             Estimate Std. Error z value Pr(>|z|)    
#> (Intercept)  1.5825    1.7930  0.883 0.377453    
#> V1          -14.0476   4.7706 -2.945 0.003234 **  
#> V2           13.1052   4.6302  2.830 0.004649 **  
#> V3          -32.4741  10.8948 -2.981 0.002876 **  
#> V4           20.8645   6.6486  3.138 0.001700 **  
#> V5          -22.4265   7.5065 -2.988 0.002812 **  
#> V6           -7.6024   2.5551 -2.975 0.002927 **  
#> V7          -42.2560  14.2772 -2.960 0.003080 **  
#> V8            8.1928   3.1192  2.627 0.008624 **  
#> V9          -21.4464   7.1228 -3.011 0.002604 **  
#> V10         -50.4194  16.6190 -3.034 0.002415 **  
#> V11         35.3283  11.6378  3.036 0.002400 **  
#> V12         -63.4810  20.8708 -3.042 0.002353 **  
#> V13            0.2399   0.3072  0.781 0.434925    
#> V14         -66.1152  21.6160 -3.059 0.002224 **  
#> V15          -1.2975   0.4755 -2.729 0.006358 **  
#> V16         -58.6246  19.4866 -3.008 0.002626 **  
#> V17        -106.7505  35.4673 -3.010 0.002614 **  
#> V18         -39.8971  13.3043 -2.999 0.002710 **  
#> V19          13.4405   4.3901  3.062 0.002202 **  
#> V20            6.3284   2.4134  2.622 0.008738 **  
#> V21            8.8149   2.9250  3.014 0.002581 **  
#> V22            0.6016   0.4076  1.476 0.139920    
#> V23          -1.3771   0.3905 -3.526 0.000422 ***  
#> V24          -0.3582   0.5053 -0.709 0.478390    
#> V25            2.3926   0.9927  2.410 0.015947 *
```

```
#> V26          0.4518    0.5333   0.847 0.396896
#> V27         11.6609   3.7447   3.114 0.001846 **
#> V28         9.1116   2.7690   3.291 0.001000 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> (Dispersion parameter for binomial family taken to be 1)
#>
#> Null deviance: 1092.40 on 787 degrees of freedom
#> Residual deviance: 177.04 on 759 degrees of freedom
#> AIC: 235.04
#>
#> Number of Fisher Scoring iterations: 19
```

Con el modelo ya entrenado, se realizan las predicciones para los datos del conjunto de testing, utilizando para ello la función “predict()”. Los parámetros son simples: el primero es el modelo o clasificador que se va a utilizar y que será “undersampledModel”; a continuación el tipo de dato que devolverá, en este caso “response”, el cual indica que el algoritmo devolverá las probabilidades de fraude listadas en un único vector, el cual estará disponible a partir de la variable “fraud_prob”; y por último, el parámetro “newdata” que hace referencia al dataset en el que se descarta la última columna por ser la que representa la variable dependiente.

```
fraud_prob = predict(undersampledModel, type = "response", newdata = test[,-29])
head(fraud_prob)
#>      634        1726        3141        4686        6109        6259
#> 7.896759e-01 5.440909e-02 4.441147e-05 1.390099e-01 1.000000e+00 7.486356e-02
```

La visualización del vector con las predicciones puede parecer algo confusa, por lo que, a menudo, es preciso realizar una conversión de esas predicciones en valores 0 y 1, dependiendo del rango de valores a partir del cual se estime que una transacción es fraudulenta: por ejemplo, a partir del 60 % de probabilidad, la transacción será etiquetada como “1”; en caso contrario será etiquetada como “0”. Para ello se utiliza el siguiente código “ifelse”:

```
y_pred = ifelse(fraud_prob > 0.6, 1, 0)
```

La matriz de confusión

Como último paso del ejercicio se crea la matriz de confusión con el fin de visualizar qué tal se ha comportado el algoritmo, es decir, cuántos positivos y negativos ha logrado predecir correctamente. Para ello se creará la variable **confusionMatrix**, en la cual se almacenará el resultado de la comparación entre el vector del dataset de testing, es decir, los datos etiquetados originalmente, y el vector de sus traducciones a 0 y 1, resultado del algoritmo. El resultado, como se puede comprobar es que ha evaluado correctamente 186 de las 196 observaciones.

60.2. Modelización del fraude en la compra con tarjetas de crédito

929

```
confusionMatrix = table(test[, 29], y_pred)
confusionMatrix
#>     y_pred
#>     0  1
#>   0 94  4
#>   1  7 91
```


Apéndice A

Información de la sesión

```
sessionInfo()
#> R version 4.2.1 (2022-06-23 ucrt)
#> Platform: x86_64-w64-mingw32/x64 (64-bit)
#> Running under: Windows 10 x64 (build 19045)
#>
#> Matrix products: default
#>
#> locale:
#> [1] LC_COLLATE=Spanish_Spain.utf8  LC_CTYPE=Spanish_Spain.utf8
#> [3] LC_MONETARY=Spanish_Spain.utf8 LC_NUMERIC=C
#> [5] LC_TIME=Spanish_Spain.utf8
#>
#> attached base packages:
#> [1] stats      graphics   grDevices  utils      datasets   methods    base
#>
#> other attached packages:
#> [1] flextable_0.8.1   fontawesome_0.4.0
#>
#> loaded via a namespace (and not attached):
#> [1] zip_2.2.1        Rcpp_1.0.9       pillar_1.8.1     compiler_4.2.1
#> [5] R.methodsS3_1.8.2 R.utils_2.12.2   base64enc_0.1-3  tools_4.2.1
#> [9] digest_0.6.30    uuid_1.1-0      tibble_3.1.8    evaluate_0.16
#> [13] lifecycle_1.0.3  gtable_0.3.1   R.cache_0.16.0  pkgconfig_2.0.3
#> [17] rlang_1.0.6     DBI_1.1.3      cli_3.4.1      rstudioapi_0.14
#> [21] yaml_2.3.5      xfun_0.35     fastmap_1.1.0  stringr_1.4.1
#> [25] dplyr_1.0.10    officer_0.4.4  styler_1.8.1   xml2_1.3.3
#> [29] knitr_1.39     generics_0.1.3  gdtools_0.2.4  vctrs_0.5.1
#> [33] systemfonts_1.0.4 tidyselect_1.2.0  grid_4.2.1     glue_1.6.2
#> [37] data.table_1.14.6 R6_2.5.1      fansi_1.0.3    rmarkdown_2.14
#> [41] bookdown_0.28    purrrr_0.3.5   ggplot2_3.4.0  magrittr_2.0.3
#> [45] scales_1.2.1    htmltools_0.5.4 assertthat_0.2.1 colorspace_2.0-3
```

```
#> [49] utf8_1.2.2      stringi_1.7.8      munsell_0.5.0      R.oo_1.25.0
```

Bibliografía

- Aas, K., Jullum, M., and Løland, A. (2021). Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *Artificial Intelligence*, 298:103502.
- Abedjan, Z., Golab, L., and Naumann, F. (2015). Profiling relational data: a survey. *The VLDB Journal*, 24(4):557–581.
- Abraham, R., Schneider, J., and vom Brocke, J. (2019). Data governance: A conceptual framework, structured review, and research agenda. *International Journal of Information Management*, 49:424–438.
- acens.com (2014). White paper: Bbdd nosql. Report.
- Al-Ruithe, M., Benkhelifa, E., and Hameed, K. (2019). A systematic literature review of data governance and cloud data governance. *Personal and Ubiquitous Computing*, 23(5-6):839–859.
- Allaire, J. (2022). *quarto: R Interface to 'Quarto' Markdown Publishing System*. R package version 1.2.
- Almudevar, A. (2021). *Theory of Statistical Inference*. Texts in Statistical Science. Chapman & Hall/CRC.
- Amat Rodrigo, J. (2017). *Clustering y heatmaps: aprendizaje no supervisado*.
- Amazon Web Services (2018). ¿qué es nosql? Report.
- Anderberg, M. R. (1973). *Cluster analysis for applications: probability and mathematical statistics*. Academic press, New Yoor, USA.
- Ang, Q. W., Baddeley, A., and Nair, G. (2012). Geometrically corrected second order analysis of events on a linear network, with applications to ecology and criminology. *Scandinavian Journal of Statistics*, 39(4):591–617.
- Anscombe, F. J. (1973). Graphs in statistical analysis. *American Statistician*.
- Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Studies in Operational Regional Science. Springer Netherlands.

- Anselin, L. (1996). *The Moran scatterplot as an ESDA tool to assess local instability in spatial association*. Routledge. Num Pages: 16.
- Anselin, L. (2013). *Spatial econometrics: methods and models*, volume 4. Springer Science & Business Media.
- Anselin, L. (2017). A Local Indicator of Multivariate Spatial Association: Extending Geary's c. *Center for Spatial Data Science Working papers. University of Chicago*.
- Arnab, R. (2017). *Survey Sampling Theory and Applications*. Elsevier.
- Astigarraga, J. and Cruz-Alonso, V. (2022). ¡se puede entender cómo funcionan git y github! *Ecosistemas*, 31(1):2332.
- Azevedo, A. and Santos, M. F. (2008). Kdd, semma and crisp-dm: a parallel overview. *IADS-DM*.
- Baddeley, A., Davies, T. M., Rakshit, S., Nair, G., and McSwiggan, G. (2022). Diffusion smoothing for spatial point patterns. *Statistical Science*, 37(1):123–142.
- Baddeley, A., Nair, G., Rakshit, S., McSwiggan, G., and Davies, T. M. (2021). Analysing point patterns on networks-a review. *Spatial Statistics*, 42:100435.
- Baddeley, A., Rubak, E., and Turner, R. (2015). *Spatial Point Patterns: Methodology and Applications with R*. CRC Press.
- Baddeley, A. and Turner, R. (2005). spatstat: an R package for analyzing spatial point patterns. *Journal of Statistical Software*, 12:1–42.
- Balakrishnan, N., Koutras, M. V., and Politis, K. G. (2019). *Introduction to Probability: Models and Applications*. Wiley Series in Probability and Statistics. John Wiley & Sons.
- Barr, C. D. and Schoenberg, F. P. (2010). On the Voronoi estimator for the intensity of an inhomogeneous planar Poisson process. *Biometrika*, 97(4):977–984.
- Batini, C., Scannapieco, M., et al. (2016). Data and information quality. *Cham, Switzerland: Springer International Publishing*.
- Baumer, B., Kaplan, D., and Horton, N. (2021). *Modern Data Science with R*. Texts in statistical science. Chapman & Hall/CRC, Boca Raton.
- Beh, E. J. and Lombardo, R. (2014). *Correspondence Analysis: Theory, Practice and New Strategies*. Wiley Series in Probability and Statistics. Wiley.
- Berlusconi, G., Calderoni, F., Parolini, N., Verani, M., and Piccardi, C. (2016). Link prediction in criminal networks: A tool for criminal intelligence analysis. *PLOS ONE*, 11(4):e0154244.
- Bernaards, C. and Jennrich, R. (2005). Gradient projection algorithms and software for arbitrary rotation criteria in factor analysis. *Educational and Psychological Measurement*, 65:676–696.
- Biecek, P. (2018). Dalex: Explainers for complex predictive models in r. *The Journal of Machine Learning Research*, 19(1):3245–3249.

- Bivand, R. (2020). *classInt: Choose Univariate Class Intervals*. R package version 0.4-3.
- Bivand, R. (2022). *spdep: Spatial Dependence: Weighting Schemes, Statistics*. R package version 1.2-2.
- Blais, B. (2020). *Statistical Inference for everyone*. Save the broccoli publishing.
- Blaug, M. (1980). *La metodología de la economía o cómo explican los economistas*. Alianza Editorial.
- Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.
- Bock, H. H. (1974). *Automatische Classifikation*. Studia Mathematica. Vandenhoeck and Ruprecht, Göttingen, Germany.
- Boehmke, B. and Greenwell, B. M. (2019). *Hands-on machine learning with R*. CRC press.
- Boehmke, B. y Greenwell, B. (2020). *Hands-On Machine Learning with R*. Chapman and Hall/CRC. Chapman and Hall.
- Borgatti, S. P. (2022). *Analyzing Social Networks Using R: Your Essential Guide*. SAGE Publications Ltd, Estados Unidos.
- Borji, A. (2022). Generated faces in the wild: Quantitative comparison of stable diffusion, midjourney and dall-e 2. *arXiv preprint arXiv:2210.00586*.
- Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152.
- Boskovitz, A., Goré, R., and Hegland, M. (2003). A logical formalisation of the fellegi-holt method of data cleaning. In *Advances in Intelligent Data Analysis V: 5th International Symposium on Intelligent Data Analysis, IDA 2003, Berlin, Germany, August 28-30, 2003. Proceedings 5*, pages 554–565. Springer.
- Brachman, R. J. and Anand, T. (1994). The process of knowledge discovery in databases: A first sketch. In *KDD workshop*, volume 3, pages 1–12.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.
- Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). Classification and regression trees. 1st edition.
- Brian, S. (1993). Cluster analysis 3rd ed. *Edward Arnold, London*, 169.
- Brockwell, P. J. and Davis, R. A. (2016). *Introduction to Time Series and Forecasting*. Springer Texts in Statistics. Springer International Publishing, Switzerland.
- Brous, P., Herder, P., and Janssen, M. (2016). Governing Asset Management Data Infrastructures. *Complex Adaptive Systems Los Angeles, CA November 2-4, 2016*, 95:303–310.

- Brownlee, J. (2020). *Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python*. Machine Learning Mastery.
- Bunge, M. (2004). *La investigación científica. Su estrategia y su filosofía*. SIGLO XXI Editores, Buenos Aires.
- Bunge, M. (2018). *La ciencia: su método y su filosofía*, volume 1. Laetoli.
- Burkov, A. (2019). *The hundred-page machine learning book*, volume 1. Andriy Burkov Quebec City, QC, Canada.
- Caballero, I., Gualo, F., Rodríguez, M., and Piattini, M. (2022a). Br4dq: A methodology for grouping business rules for data quality evaluation. *Information Systems*, 109:102058.
- Caballero, I., Piattini, M., and Gualo, F. (2022b). Marco metodológico para la creación, implantación y mantenimiento de Sistemas de Gobierno de Datos. In Goñi Sarriguren, A., editor, *JISBD2022*. SISTEDES.
- Caballero, I., Piattini, M., and Rodríguez, M. (2023). Modelo alarcos de madurez de datos v4.0: Un modelo de referencia de procesos basado en estándares internacionales abiertos para la gestión de los datos, gestión de la calidad de los datos y el gobierno de los datos. Technical report, DQTeam, UCLM, AQCLab.
- Cancelo, J. R. (1997). *Proyecto Docente e Investigador. Concurso de acceso al cuerpo docente de Catedráticos de Universidad*. Universidad de Castilla-La Mancha.
- Carrasco-Oberto, G. I. (2020). *Cluster no jerárquicos versus cart y biplot*. PhD thesis, Universidad de Salamanca.
- Carruthers, C. and Jackson, P. (2020). *The chief data officer's playbook*. Facet Publishing.
- Casella, G. and Berger, R. L. (2007). *Statistical inference, 2nd Ed.* Cengage Learning.
- Chacon, S. (2009). *Pro Git*. Apress.
- Chalmers, A. F., Villate, J. A. P., Máñez, P. L., and Sedeño, E. P. (2000). ¿qué es esa cosa llamada ciencia?
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R., et al. (2000a). Crisp-dm 1.0: Step-by-step data mining guide. *SPSS inc*, 9(13):1–73.
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R., et al. (2000b). Crisp-dm 1.0: Step-by-step data mining guide. *SPSS inc*, 9(13):1–73.
- Chatfield, C. and Collins, A. (1980). *Introduction to multivariate analysis*. Chapman&Hall/CRC.
- Chaudhuri, A. and Stenger, H. (2005). *Survey Sampling. Theory and Methods, 2nd Ed.* STATISTICS: a series of TEXTBOOKS and MONOGRAPHS. Chapman & Hall/CRC.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

- Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., et al. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4.
- Chilès, J. P. and Delfiner, P. (1999). *Geostatistics: Modeling Spatial Uncertainty*. John Wiley and Sons, Ltd, Chichester, UK.
- Codd, E. F. (1970a). A relational model of data for large shared data banks. 13(6).
- Codd, E. F. (1970b). A relational model of data for large shared data banks. *Communications of the ACM*, 13(6):377–387.
- Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.
- Council, E. (2020). The Data Capability Assessment Model (DCAM) Framework v2.2 Overview.
- Cramer, W., Guiot, J., Marini, K., Secretariat, M., and Bleu, P. (2020). Climate and environmental change in the mediterranean basin—current situation and risks for the future. *First Mediterranean Assessment Report. Union for the Mediterranean, Plan Bleu, UNEP/MAP*.
- Cressie, N. and Wikle, C. K. (2015). *Statistics for Spatio-Temporal Data*. John Wiley & Sons.
- Cressie, N. A. C. (1993). *Statistics for Spatial Data*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA.
- Cronie, O. and Van Lieshout, M. N. M. (2018). A non-model-based approach to bandwidth selection for kernel estimators of spatial intensity functions. *Biometrika*, 105(2):455–462.
- Cryer, J. D. and Chan, K.-S. (2010). *Time Series Analysis with Applications in R*. Springer texts in Statistics. Springer, Iowa, USA.
- Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*:1695.
- Cuadras, C. M. (2007). *Nuevos Métodos de Análisis Multivariante*. CMC Editions, Barcelona, Spain.
- Cutler, A. and Zhao, G. (1999). Fast classification using perfect random trees. *Utah State University*.
- DAMA (2017). *DAMA-DMBOK: data management body of knowledge*. Technics Publications, LLC.
- DANE (2019). Proyecciones de población departamentales y municipales por área 2005-2020. www.dane.gov.co.
- Davenport, T. and Harris, J. (2017). *Competing on analytics: Updated, with a new introduction: The new science of winning*. Harvard Business Press.
- Davenport, T. H. and Patil, D. (2012). Data scientist. *Harvard business review*, 90(5):70–76.
- Davies, T. M. and Baddeley, A. (2018). Fast computation of spatially adaptive kernel estimates. *Statistics and Computing*, 28(4):937–956.

- De Boor, C. (2001). *A practical guide to splines*. Applied Mathematical Sciences. Springer-Verlag, New York.
- de Finetti, B. (2017). *Theory of Probability: A Critical Introductory Treatment*. Wiley Series in Probability and Statistics. Wiley.
- De la Fuente, S. (2011). *Ánálisis Factorial*. Madrid, España.
- de la Real Academia Española, D. (2023). Inteligencia artificial.
- de Leeuw, J. and Mair, P. (2009). Multidimensional scaling using majorization: SMACOF in R. *Journal of Statistical Software*, 31(3):1–30.
- Dembla, G. (2020). Intuition behind log-loss score. *Towards Data Science*.
- Deng, J., Berg, A., Satheesh, S., Su, H., Khosla, A., and Fei-Fei, L. (2012). Imagenet large scale visual recognition competition 2012 (ilsvrc2012). *See net. org/challenges/LSVRC*, 41.
- Diday, E. (1971). Une nouvelle méthode en classification automatique et reconnaissance des formes la méthode des nuées dynamiques. *Revue de statistique appliquée*, 19(2):19–33.
- Diday, E. (1973). The dynamic clusters method in nonhierarchical clustering. *International Journal of Computer & Information Sciences*, 2(1):61–88.
- Diggle, P. (1985). A kernel method for smoothing point process data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 34(2):138–147.
- Diggle, P. (2013). *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*. CRC press.
- Diggle, P. and Giorgi, E. (2019). *Model-based Geostatistics for Global Public Health: Methods and Applications*. Chapman and Hall/CRC.
- Díez, J. A. and Moulines, C. U. (2008). *Fundamentos de Filosofía de la Ciencia*. Editorial Ariel, Barcelona.
- Easley, David; Kleinberg, J. (2010). *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, Estados Unidos.
- Edelbrock, C. (1979). Mixture model tests of hierarchical clustering algorithms: The problem of classifying everybody. *Multivariate Behavioral Research*, 14(3):367–384.
- Efron, B. and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical science*, pages 54–75.
- Eilers, P. H. and Marx, B. D. (2010). Splines, knots, and penalties. *Wiley interdisciplinary reviews: Computational statistics*, 2(6):637–653.
- Eilers, P. H., Marx, B. D., and Durbán, M. (2015). Twenty years of p-splines. *SORT: statistics and operations research transactions*, 39(2):0149–186.
- Elhorst, J. P. (2010). Applied Spatial Econometrics: Raising the Bar. *Spatial Economic Analysis*, 5(1):9–28.

- Engels, B. (2019). Data Governance as the Enabler of the Data Economy. *Intereconomics*, 54(4):216–222.
- Eryurek, E., Gilad, U., Laksmanan, V., Kibunguchy-Grant, A., and Ashdown, J. (2021). *Data Governance: The Definitive Guide*. O'Reilly Media, Inc.
- Facebook Marketing Science (2021). *Robyn*.
- Facebook Research AI (2019). *Nevergrad - A gradient-free optimization platform*. is a Python 3.6+ library.
- Faraway, J. J. (2002). *Practical regression and ANOVA using R.*, volume 168. University of Bath Bath.
- Fay, C., Rochette, S., Guyader, V., and Girard, C. (2021). *Engineering Production-Grade Shiny Apps*. Chapman and Hall/CRC.
- Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., et al. (1996). Knowledge discovery and data mining: Towards a unifying framework. In *KDD*, volume 96, pages 82–88.
- Firth, J. (1957). A synopsis of linguistic theory, 1930–1955. In *Selected Papers of J.R. Firth 1952–1959*, ed. Frank Palmer, 168–205. Londres: Longman.
- Fleitas, F. (2017). *La Inteligencia Artificial e Ingeniería del Conocimiento: Guía de la Inteligencia Artificial e Ingeniería del Conocimiento con ejemplos de Sistemas Expertos en el Lenguaje CLIPS*. Editorial Académica Española.
- Forgy, E. (1965). Cluster analysis of multivariate data: Efficiency vs. interpretability of classification. *Biometrics*, 21(3):768–769.
- Fradejas Rueda, J. M. (2022). *Cuentapalabras. Estilometría y análisis de datos con R para filólogos*. <http://www.aic.uva.es/cuentapalabras/>.
- Fruchterman, T. M. and Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and experience*, 21(11):1129–1164.
- Fukushima, K. and Miyake, S. (1982). Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets*, pages 267–285. Springer.
- Gallardo San-Salvador, J. A. (2022). *Introducción al Análisis Cluster*.
- Gallardo-San Salvador, J. A. and Vera-Vera, J. F. (2004). *Técnicas aplicadas de análisis de datos multivariantes*. Universidad de Granada, Granada, Spain.
- Garcia, G. B., Suarez, O. D., Aranda, J. L. E., Tercero, J. S., and Gracia, I. S. (2015). *Learning Image Processing with OpenCV*. Packt Publishing.
- García Abad, J. et al. (2021). Comparativa de técnicas de balanceo de datos. aplicación a un caso real para la predicción de fuga de clientes.
- García-Alsina, M. (2017). *Big data: gestión y explotación de grandes volúmenes de datos*. UOC.

- Gentile, C. and Warmuth, M. K. (1998). Linear hinge loss and average margin. *Advances in neural information processing systems*, 11.
- Getis, A. (1999). *Spatial statistics*. Longley, P., Goodchild, M., Maguire, D. y Rhind, D. (Eds.) Geographical Information Systems.
- Gilmore, R., Hutchins, S., Pastoor, D., Attali, D., Singham, L., Raja, A. M., Trimarchi, L., Khanal, K., Columbus, A., Howard, P., and Zhang, L. (2017). Awesome R Shiny.
- Giraud, T. (2022). *mapsf: Thematic Cartography*. R package version 0.4.0.
- Gohel, D. and Skintzos, P. (2022). *flextable: Functions for Tabular Reporting*. R package version 0.8.3.
- Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. Adaptive computation and machine learning. MIT Press.
- Greenacre, M. (2008). *La práctica del análisis de correspondencias*. Fundación BBVA.
- Gualo, F., Rodríguez, M., Verdugo, J., Caballero, I., and Piattini, M. (2021). Data quality certification using ISO/IEC 25012: Industrial experiences. *J. Syst. Softw.*, 176:110938.
- Guerry, A.-M. (1833). *Essai Sur La Statistique Morale de La France*. Crochard.
- Gómez García, J. L. and Conesa i Caralt, J. (2015). *Introducción al big data*. Oberta Barcelona, UOC Publishing.
- Hajek, A. and Hitchcock, C. (2016). *The Oxford Handbook of Probability and Philosophy*. Oxford University Press.
- Hamilton (1994). *Time Series Analysis*. Statistics. Princeton University Press, Princeton, NJ, USA.
- Harman, H. H. (1976). *Modern Factor Analysis (Third Edition Revised)*. Chicago, USA.
- Harrison, T., F. Luna-Reyes, L., Pardo, T., De Paula, N., Najafabadi, M., and Palmer, J. (2019). The Data Firehose and AI in Government: Why Data Management is a Key to Value and Ethics. In *Proceedings of the 20th Annual International Conference on Digital Government Research*, dg.o 2019, pages 171–176, New York, NY, USA. Association for Computing Machinery. event-place: Dubai, United Arab Emirates.
- Hartigan, J. and Wong, M. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. series c (applied statistics)*, 28(1):100–108.
- Hartigan, J. A. (1975). *Clustering algorithms*. John Wiley & Sons, Inc.
- Hastie, T. and Tibshirani, R. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Monographs on Statistics & Applied Probability. Chapman and Hall/CRC.
- Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation*. Prentice Hall.
- Hecht, R. and Jablonski, S. (2011). Nosql evaluation: A use case oriented survey. In *Cloud and Service Computing (CSC), 2011 International Conference on*, pages 336–341. IEEE.

- Hempel, C. (2005). *La explicación científica: Estudios sobre la filosofía de la ciencia*. Ediciones Paidós, Barcelona.
- Henderson, C. (1953). Estimation of variance and covariance components. *Biometrics*, 9:226–252.
- Hernández-Orallo, J., Flach, P. A., and Ramirez, C. F. (2011). Brier curves: a new cost-based visualisation of classifier performance. In *Icml*, pages 585–592.
- Hernangómez, D. and Fernández-Avilés, G. (2022). *Visualización y geolocalización de datos con R*. Netlify, Online.
- Hester, J. and Wickham, H. (2023). *odbc: Connect to ODBC Compatible Databases (using the DBI Interface)*. R package version 1.3.4.
- Hijmans, R. J. (2022). *raster: Geographic Data Analysis and Modeling*.
- Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis*, 15(3):199–236.
- Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE.
- Hothorn, T. and Everitt, B. (2014). *A Handbook of Statistical Analyses using R*. Routledge.
- Ihaka, R. and Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of computational and graphical statistics*, 5(3):299–314.
- IIC, I. d. I. d. C. (2016). Las 7 v del big data: Características más importantes. Report.
- Illian, J., Penttinen, A., Stoyan, H., and Stoyan, D. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*, volume 70. John Wiley & Sons.
- Ilyas, I. F. and Chu, X. (2019). *Data cleaning*. Morgan & Claypool.
- ISACA (2019). COBIT | Control Objectives for Information Technologies.
- ISO (2016). ISO 8000-61:2016.
- ISO (2017). ISO/IEC 38505-1:2017 Information technology — Governance of IT — Governance of data — Part 1: Application of ISO/IEC 38500 to the governance of data.
- ISO (2018). ISO 8000-62:2018.
- ISO (2018). ISO/IEC TR 38505-2:2018 Information technology — Governance of IT — Governance of data — Part 2: Implications of ISO/IEC 38505-1 for data management.
- ISO/IEC (2008a). Iso/iec 25012: Software engineering-software product quality requierements and evaluation (square) - data quality model. Report.
- ISO/IEC (2008b). Iso/iec 25024: Software engineering-software product quality requierements and evaluation (square) - data quality model. Report.

- ISO/IEC (2011). So/iec 25040 -systems and software engineering — systems and software quality requirements and evaluation (square) — evaluation process. Report.
- ISO/IEC (2015). Iso 8000-8: Data quality — part 8: Information and data quality: Concepts and measuring. Report.
- Jackson, P. and Carruthers, C. (2019). *Data Driven Business Transformation: How to Disrupt, Innovate and Stay Ahead of the Competition*. John Wiley & Sons.
- James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer. <https://www.statlearning.com/>.
- Janssen, M., Brous, P., Estevez, E., Barbosa, L. S., and Janowski, T. (2020). Data governance: Organizing data for trustworthy artificial intelligence. *Government Information Quarterly*, 37(3):101493.
- Jenny Bryan, the STAT 545 TAs, J. H. (2021). *Happy Git and GitHub for the useR*.
- Jobson, J. D. (1992). *Applied Multivariate Data Analysis. Vol. II*. Springer test inStatistics. Springer-Verlag, NeW York, USA.
- Jockers, M. (2014). *Text analysis with R for students of literature*. Nueva York: Springer.
- Jockers, M. (2017). Introduction to the syuzhet package. <https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html>.
- Johnson, N. L., Kemp, A. W., and Kotz, S. (2008). *Univariate Discrete Distributions, 3e Set*. Wiley Series in Probability and Statistics. Wi.
- Jones, M. C. (1993). Simple boundary correction for kernel density estimation. *Statistics and computing*, 3(3):135–146.
- Journel, A. G. and Huijbregts, C. H. J. (1978). *Mining Geostatistics*. Academic Press, New York, USA.
- Kalyvas, J. R. and Overly, M. R. (2014). *Big data: a business and legal guide*. CRC Press.
- Kassambara, A. (2017). *Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning (Multivariate Analysis) 1st Ed*. sthada.com.
- Kaufman, L. and Rousseeuw, P. J. (1990). Divisive analysis (program diana). In Kaufman, L. and Rousseeuw, P. J., editors, *Finding Groups in Data: An Introduction to Cluster Analysis*, pages 68–125. John Wiley and Sons, Inc., Hoboken.
- Kelejian, H. H. and Prucha, I. R. (2010). Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances. *Journal of Econometrics*, 157(1):53–67.
- Khatri, V. and Brown, C. V. (2010). Designing Data Governance. *Commun. ACM*, 53(1):148–152. Place: New York, NY, USA Publisher: Association for Computing Machinery.
- Kiefer, J. and Wolfowitz, J. (1952). Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, pages 462–466.

- Kim, J.-H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational statistics & data analysis*, 53(11):3735–3745.
- Knuth, D. E. (1984). Literate programming. *The Computer Journal*, 27(2):97–111.
- Kuhn, M. (2008). Building predictive models in r using the caret package. *Journal of statistical software*, 28:1–26.
- Kuhn, M. (2019). CRAN Task View: Reproducible Research. R Task View.
- Kuhn, M., Johnson, K., et al. (2013). *Applied predictive modeling*, volume 26. Springer.
- Ladley, J. (2019). *Data governance: How to design, deploy, and sustain an effective data governance program*. Academic Press.
- Lê, S., Josse, J., and Husson, F. (2008). FactoMineR: A package for multivariate analysis. *Journal of Statistical Software*, 25(1):1–18.
- LeCun, Y., Bengio, Y., et al. (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995.
- LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.
- Lee, C.-P. and Lin, C.-J. (2013). A study on l2-loss (squared hinge-loss) multiclass svm. *Neural computation*, 25(5):1302–1323.
- Leisch, F. (2002). Sweave: Dynamic generation of statistical reports using literate data analysis. In *Compstat*, pages 575–580. Springer.
- LeSage, J. and Pace, R. K. (2009). *Introduction to Spatial Econometrics*. Chapman and Hall/CRC.
- Lewis, J. A. (1999). Statistical principles for clinical trials (ich e9): an introductory note on an international guideline. *Statistics in medicine*, 18(15):1903–1942.
- Lillie, T. and Eybers, S. (2019). *Identifying the constructs and agile capabilities of data governance and data management: A review of the literature*, volume 933 of *Communications in Computer and Information Science*. Pages: 326.
- Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons.
- Liu, B. (2015). *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press.
- Liu, X., Rivera, S. C., Moher, D., Calvert, M. J., and Denniston, A. K. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the consort-ai extension. *BMJ*, 370.
- Lo, F. (2017). Big data technology. Report, DataJobs.com.

- Loshin, D. (2002). Rule-based data quality. In *Proceedings of the eleventh international conference on Information and knowledge management*, pages 614–616.
- Loshin, D. (2011). Master data management and data quality bt-the practitioner’s guide to data quality improvement. In *MK Series on Business Intelligence*, pages 327–350. Morgan Kaufmann.
- Lovelace, R., Nowosad, J., and Münchow, J. (2019). *Geocomputation with R*. CRC Press, Taylor and Francis Group, Boca Raton.
- Lozano Zahonero, M. (2020). Una nueva visión de la supuesta influencia de *Madame Bovary* en *La Regenta* a través de la estilometría y el análisis de sentimientos basados en lenguaje R. *Orillas: rivista d’ispanistica*, 9:573–607.
- López, D. (2012). Análisis de las posibilidades de uso de big data en las organizaciones. *Universidad de Cantabria, Santander, España*.
- López-González, E. and Hidalgo-Sánchez, R. (2010). Escalamiento multidimensional no métrico. un ejemplo con r empleando el algoritmo smacof. *Estudios sobre educación*, 8(1):9–35.
- MacNaughton-Smith, P., Williams, W. T., Dale, M. B., Mockett, L. G., and Dunn, C. (1964). Dissimilarity analysis: a new technique of hierarchical sub-division. *Nature*, 202:1034–1035.
- MacQueen, J. (1967). Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability*, pages 281–297.
- Mahanti, R. (2019). *Data quality: dimensions, measurement, strategy, management, and governance*. Quality Press.
- Mair, P., Groenen, P. J. F., and de Leeuw, J. (2022). More on multidimensional scaling in R: smacof version 2. *Journal of Statistical Software*, 102(10):1–47.
- Mardia, K., Kent, J., and Bibby, J. (1979a). *Multivariate Analysis*. London, UK.
- Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979b). *Multivariate Analysis*. Academic Press, London, UK.
- Martínez, R. G., Carrasco, R. A., García-Madariaga, J., Gallego, C. P., and Herrera-Viedma, E. (2019). A comparison between fuzzy linguistic rfm model and traditional rfm model applied to campaign management. case study of retail business. *Procedia Computer Science*, 162:281–289.
- Martínez R., Molina J. M., C. J. (2005). *Desarrollo de sistemas basados en el conocimiento*. Sanz y Torres S. L.
- Martori, J. C., Hoberg, K., and Madariaga, R. (2008). La incorporación del espacio en los métodos estadísticos: Autocorrelación espacial y segregación. *Actas del X Coloquio Internacional de Geocrítica*.
- Matejka, J. and Fitzmaurice, G. (2017). Same Stats, Different Graphs. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 1290–1294, New York, NY, USA. ACM.

- Matheron, G. (1962). *Traité de Geostatistique Appliquée. vol I.* Éditions Technip, Paris, France.
- Matloff, N. and Zhang, W. (2022). A novel regularization approach to fair ml. *arXiv preprint arXiv:2208.06557*.
- McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5:115–133.
- McNicholas, P. D. (2016). Model-based clustering. *Journal of Classification*, 33(3):331–373.
- Mecca, M., Young, R., and Halcomb, J. (2014). *Data Management Maturity (DMM) Model*. CMMI Institute, first edition.
- Mella, J. M. and Chasco, C. (2006). *Urban Growth and Territorial Dynamics: A Spatial-Econometric Analysis of Spain*. Edward Elgar Publishing.
- Minsky, M. and Papert, S. (1969). An introduction to computational geometry. *Cambridge tiass.*, HIT, 479:480.
- Missouri, Rokia; Sarr, I. (2015). *Social Network Analysis - Community Detection and Evolution*. Springer, Estados Unidos.
- Molinaro, A. M., Simon, R., and Pfeiffer, R. M. (2005). Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, 21(15):3301–3307.
- Møller, J. and Waagepetersen, R. (2003). *Statistical Inference and Simulation for Spatial Point Processes*. CRC Press.
- Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.
- Monsalve, C. (2019). Medellín necesita 2000 uniformados más para reforzar seguridad. <https://www.bluradio.com/blu360/antioquia/medellin-necesita-2-000-uniformados-mas-para-reforzar-seguridad-policia>.
- Montero, J. M. (1997). *Proyecto Docente e Investigador. Concurso de acceso al cuerpo docente de Catedráticos de Universidad*. Universidad de Castilla-La Mancha.
- Montero, J. M. (2002). Una propuesta de corrección de continuidad asimétrica para tablas de contingencia (2x2) con totales marginales fijos. *Estadística Española*, 44(149):29–46.
- Montero, J.-M. (2007). *Estadística descriptiva*. ALFA CENTAURO.
- Montero, J.-M., Fernández-Avilés, G., and Mateu, J. (2015). *Spatial and spatio-temporal geostatistical modeling and kriging*. John Wiley & Sons.
- Montero, J. M. and Larraz, B. (2008). *Introducción a la Geoestadística Lineal*. Metodología y Análisis de Datos en Ciencias Sociales. Netbiblo, A Coruña, España.
- Montero Lorenzo, J. (2007). Estadística descriptiva, editorial thomson-paraninfo.
- Moradi, M. (2018). *Spatial and Spatio-Temporal Point Patterns on Linear Networks*. PhD Dissertation, University Jaume I.

- Moradi, M., Cronie, O., Rubak, E., Lachieze-Rey, R., Mateu, J., and Baddeley, A. (2019). Resample-smoothing of voronoi intensity estimators. *Statistics and computing*, 29(5):995–1010.
- Morgan, A. (2015). Joins and other aggregation enhancements in mongodb 3.2. Report, <http://www.clusterdb.com>.
- Morin, D. J. (2016). *Probability: For the Enthusiastic Beginner*. CreateSpace Independent Publishing Platform.
- Morrison, D. F. (1976). *Multivariate Statistical Methods-2*. New York, NY (USA) McGraw-Hill.
- Muñoz-Reja, I. C., Carretero, A. I. G., and Cejudo, F. G. (2018). *Calidad de datos*. RA-MA Editorial.
- Mínguez Salido, R. and García Centeno, M. C. (2011). *Modelos de series temporales aplicados a rendimientos financieros*. Estadística económica y finanzas. Netbiblo, España.
- Müller, K., Ooms, J., James, D., DebRoy, S., Wickham, H., and Horner, J. (2022a). *RMariaDB: Database Interface and MariaDB Driver*. R package version 1.2.2.
- Müller, K., Wickham, H., James, D. A., and Falcon, S. (2022b). *RSQlite: SQlite Interface for R*. R package version 2.2.20.
- Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *ICML 2010*, pages 807–814.
- Ng, R. T. and Han, J. (2002). Clarans: A method for clustering objects for spatial data mining. *IEEE transactions on knowledge and data engineering*, 14(5):1003–1016.
- Novikoff, A. B. (1962). On convergence proofs on perceptrons. In *Proceedings of the Symposium on the Mathematical Theory of Automata*, volume 12, pages 615–622, New York, NY, USA. Polytechnic Institute of Brooklyn.
- OECD (2019). *The Path to Becoming a Data-Driven Public Sector*. OECD Digital Government Studies. OECD Publishing, Paris.
- Okabe, A. and Sugihara, K. (2012). *Spatial Analysis Along Networks: Statistical and Computational Methods*. John Wiley & Sons.
- Olmeda, M. V. and Ibáñez, J. C. (2022). *Manual de ética aplicada en inteligencia artificial*. ANAYA MULTIMEDIA.
- O’neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway books.
- Ooms, J., James, D., DebRoy, S., Wickham, H., and Horner, J. (2022). *RMySQL: Database Interface and ‘MySQL’ Driver for R*. R package version 0.10.24.
- OpenAI (2022). Chatgpt.
- Osimo, D., Mureddu, F., Peristeras, V., Cioffi, A., Moise, C., and van, C. (2020). Data Strategies, Policies and Agenda. page 17.

- Osorio-Sanabria, M. A., Amaya-Fernández, F., and González-Zabala, M. P. (2020). Developing a model to readiness assessment of open government data in public institutions in Colombia. pages 334–340.
- Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135.
- Paul Euler, L. (1736). *Solutio problematis ad geometriam situs pertinentis*. Commentarii Academice Scientiarum Imperialis Petropolitane 8.
- Pearson, K. (1904). On the theory of contingency and its relation to association and normal correlation. In Department of Applied Mathematics, University College, U. o. L., editor, *Mathematical Contributions to the Theory of Evolution*, pages Chapter XXX, 1–34. Dulau and CO., London, UK.
- Pebesma, E. (2022). *sf: Simple Features for R*. R package version 1.0-7.
- Pebesma, E. and Bivand, R. (2022). Spatial data science: With applications in r.
- Pebesma, E. J. et al. (2018). Simple features for r: standardized support for spatial vector data. *R J.*, 10(1):439.
- Pemberton, J. (2011). Time series analysis with applications in r, second edition. *Journal of Applied Statistics*, 38(6):1311–1332.
- Perez Sola, C. (2021). *Análisis de datos de redes sociales*. Editorial UOC, España.
- Piattini, M., Marcos, E., Calero, C., and Vela, B. (2006). Tecnología y diseño de bases de datos. *Editorial Ra-Ma*.
- Pizarro, M., Hernangómez, D., and Fernández-Avilés, G. (2021). *climaemet: Climate AEMET Tools*.
- Plotkin, D. (2020). *Data stewardship: An actionable guide to effective data management and data governance*. Academic Press.
- Price, R. and Shanks, G. (2004). A semiotic information quality framework. In *Proceedings of the International Conference on Decision Support Systems DSS04*, pages 658–672.
- Pérez-Gil, J. A., Chacón, S., and Moreno, R. (2000). Validez de constructo: el uso de análisis factorial exploratorio-confirmatorio para obtener evidencias de validez. *Psicothema*, 12(Su2):442–446.
- Pérez Infante, J. I. (2006). *Las estadísticas del mercado de trabajo en España*. Ministerio de Empleo y Seguridad Social. Subdirección General de Información Administrativa y Publicaciones.
- R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- R Special Interest Group on Databases (R-SIG-DB), Wickham, H., and Müller, K. (2022). *DBI: R Database Interface*. R package version 1.1.3.

- Rakshit, S., Baddeley, A., and Nair, G. (2019a). Efficient code for second order analysis of events on a linear network. *Journal of Statistical Software*, 90:1–37.
- Rakshit, S., Davies, T., Moradi, M., McSwiggan, G., Nair, G., Mateu, J., and Baddeley, A. (2019b). Fast kernel smoothing of point patterns on a large network using two - dimensional convolution. *International Statistical Review*, 87(3):531–556.
- Rakshit, S., Nair, G., and Baddeley, A. (2017). Second-order analysis of point patterns on a network using any distance metric. *Spatial Statistics*, 22:129–154.
- Redman, T. C. (2016). *Getting in front on data: who does what*. Technics Publications.
- Restrepo, V. (2019). Qué tan segura se siente la gente en medellín? <https://www.elcolombiano.com/antioquia/seguridad/percepcion-de-seguridad-en-medellin-encuesta-de-victimizacion-PC10033581>.
- Revelle, W. (2022). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Evanston, Illinois. R package version 2.2.5.
- Reynolds, H. T. (1984). *Analysis of Nominal Data (2nd edition)*. Quantitative Applications in the Social Sciences. Sage Publication, London, UK.
- Reynolds, R. W., Banzon, V. F., and Program, N. C. (2008). Noaa optimum interpolation 1/4 degree daily sea surface temperature (oisst) analysis, version 2. *NOAA National Centers for Environmental Information*.
- Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.
- Rivera, S. C., Liu, X., Chan, A.-W., Denniston, A. K., Calvert, M. J., Ashrafi, H., Beam, A. L., Collins, G. S., Darzi, A., Deeks, J. J., et al. (2020). Guidelines for clinical trial protocols for interventions involving artificial intelligence: the spirit-ai extension. *The Lancet Digital Health*, 2(10):e549–e560.
- Romanski, P., Kotthoff, L., and Kotthoff, M. L. (2013). Package fselector. URL <http://cran/r-project.org/web/packages/FSelector/index.html>.
- Rosenbaum, P. R. (2005). Observational study. *Encyclopedia of statistics in behavioral science*.
- Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386.
- Ross, S. (2012). *A First Course in Probability 9th ed.* Prentice Hall, Upper Saddle River, NJ 07458.
- Rubenfa (2014). Mongodb: empezando por el principio. insertando datos. Report, Genbeta Dev.
- Ruiz-Mayo, L., Martin-Pliego, J., Montero, J. M., and Uríz, P. (1995). *Análisis Estadístico de Encuestas: Datos Cualitativos*. Alpha Centauro, Madrid, España.

- Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, 323(6088):533–536.
- Ryu, C. (2022). *dlookr: Tools for Data Diagnosis, Exploration, Transformation*. R package version 0.6.1.
- Saeys, Y., Inza, I., and Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *bioinformatics*, 23(19):2507–2517.
- Sakia, R. M. (1992). The box-cox transformation technique: a review. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 41(2):169–178.
- Sanabria, A. M. F., Castañeda, M. P. B., Ramos, R. R. R., and Mateu, J. (2022). Identification of patterns for space-time event networks. *Applied Network Science*, 7(1):1–24.
- Sarkar, D. (2008). *Lattice: Multivariate Data Visualization with R*. Springer, New York. ISBN 978-0-387-75968-5.
- SAS Institute Inc. (2017). Big data. what it is and why it matters.
- Schabenberger, O. and Gotway, C. A. (2005). *Statistical methods for spatial data analysis*. Texts in statistical science. Chapman & Hall/CRC, Boca Raton.
- Schapire, R. E. and Freund, Y. (2012). *Boosting: Foundations and Algorithms*. The MIT Press.
- Schloerke, B., Cook, D., Larmarange, J., Briatte, F., Marbach, M., Thoen, E., Elberg, A., Toomet, O., Crowley, J., Hofmann, H., et al. (2021). Ggally: Extension to ggplot2.
- Schölkopf, B., Simard, P., Smola, A., and Vapnik, V. (1997). Prior knowledge in support vector kernels. *Advances in neural information processing systems*, 10.
- Scholkopf, B., Sung, K.-K., Burges, C. J., Girosi, F., Niyogi, P., Poggio, T., and Vapnik, V. (1997). Comparing support vector machines with gaussian kernels to radial basis function classifiers. *IEEE transactions on Signal Processing*, 45(11):2758–2765.
- Schubert, E. and Rousseeuw, P. J. (2021). Fast and eager k-medoids clustering: O(k) runtime improvement of the pam, clara, and clarans algorithms. *Information Systems*, 101:101804.
- Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. New York: John Wiley & Sons.
- Shafique, U. and Qaiser, H. (2014). A comparative study of data mining process models (kdd, crisp-dm and semma). *International Journal of Innovation and Scientific Research*, 12(1):217–222.
- Shorten, C. and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48.
- Shumway, R. H. and Stoffer, D. S. (2017). *Time Series Analysis and Its Applications: With R Examples*. Springer Texts in Statistics. Springer International Publishing, Cham.
- Silge, J. and Robinson, D. (2017). *Text Mining with R: A Tidy Approach*. O'Reilly Media, Inc. <https://www.tidytextmining.com>.

- Silverman, B. W. (1982). Kernel density estimation using the fast Fourier transform. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 31(1):93–99.
- Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Routledge.
- Singer, J. and Willet, J. (2003). *Applied Longitudinal Data Analysis*. Oxford University Press.
- Snijders, T. (2003). Fixed and random effects. 2:664–665.
- Soares, S. (2010). *The IBM data governance unified process: driving business value with IBM software and best practices*. MC Press, LLC.
- Soares, S. (2015). *The chief data officer handbook for data governance*. Mc Press New York.
- Sokal, R. R. and Rolf, F. J. (2012). *Biometry. The Principles and Practice in Statistics in Biological Research, 4th edition*. W.H. Freeman and Company, New York, USA.
- Späth, H. (1975). Cluster-analyse-algorithmen zur objektklassifizierung und datenreduktion. *Verfahren der Datenverarbeitung*.
- Staniak, M. and Biecek, P. (2019). The landscape of r packages for automated exploratory data analysis. *arXiv preprint arXiv:1904.02101*.
- Strong, D., Lee, Y., and Wang, R. (1997a). 10 potholes in the road to information quality. *Computer*, 30(8):38–46.
- Strong, D. M., Lee, Y. W., and Wang, R. Y. (1997b). Data quality in context. *Communications of the ACM*, 40(5):103–110.
- Strozzi, C. (1998). Nosql-a relational database management system. Report.
- Tennekes, M. (2018). tmap: Thematic maps in R. *Journal of Statistical Software*, 84(6):1–39.
- Tetko, I. V., Livingstone, D. J., and Luik, A. I. (1995). Neural network studies. 1. comparison of overfitting and overtraining. *Journal of chemical information and computer sciences*, 35(5):826–833.
- Theobald, O. (2017). *Machine learning for absolute beginners: a plain English introduction*, volume 157. Scatterplot press.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, 58:267–288.
- Tierney, N. J. and Cook, D. H. (2018). Expanding tidy data principles to facilitate missing data exploration, visualization and assessment of imputations. *arXiv preprint arXiv:1809.02264*.
- Tobler, W. R. (1970a). A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46:234–240.
- Tobler, W. R. (1970b). A computer movie simulating urban growth in the Detroit region. *Economic geography*, 46(sup1):234–240.

- Toharia, L. (2012). *El mercado de trabajo en la obra de Luis Toharia*. Ministerio de Empleo y Seguridad Social.
- Treder, M. (2020). *The Chief Data Officer Management Handbook*. Apress, Berkeley, CA.
- Tribunale di Lecco, 2a Sezione Penale (2009). Sentenza nei confronti di Alcaro Luigi + 56 (Operazione Oversize). Procedimento n. 31149/01 + 10309/03 + 42859/03 + 3885/05 RGNR.
- Tribunale di Milano, Ufficio del giudice per le indagini preliminari (2006). Ordinanza di applicazione della misura della custodia cautelare in carcere nei confronti di Alcaro Luigi + 56 (Operazione Oversize). Procedimento n. 31149/01 + 10309/03 + 42859/03 + 3885/05 RGNR.
- Tukey, J. W. (1962). The future of data analysis. *The annals of mathematical statistics*, 33(1):1–67.
- Uriel Jiménez, E. and Peiro Giménez, A. (2000). *Introducción al análisis de series temporales*. Estadística. Paraninfo, Madrid, España.
- Ushey, K., Allaire, J., and Tang, Y. (2022). *reticulate: Interface to 'Python'*. R package version 1.26.
- Vallone, A. and Chasco, C. (2020). Spatiotemporal methods for analysis of urban system dynamics: an application to chile. *The Annals of Regional Science*, 64(2):421–454.
- van der Loo, M. P. and de Jonge, E. (2019). Data validation infrastructure for r. *arXiv preprint arXiv:1912.09759*.
- Vapnik, V. N. (1997). The support vector method. In *International Conference on Artificial Neural Networks*, pages 261–271. Springer.
- Venables, W. and Ripley, B. (2002). *Modern applied statistics with S*. Statistics and computing. Springer.
- Vergara, J. R. and Estévez, P. A. (2014). A review of feature selection methods based on mutual information. *Neural computing and applications*, 24:175–186.
- Vujović, Ž. et al. (2021). Classification model evaluation metrics. *International Journal of Advanced Computer Science and Applications*, 12(6):599–606.
- Wade, C. (2020). *Hands-On Gradient Boosting with XGBoost and scikit-learn: Perform accessible machine learning and extreme gradient boosting with Python*. Packt Publishing Ltd.
- Walker, K. (2022). *crsuggest: Obtain Suggested Coordinate Reference System Information for Spatial Data*. R package version 0.4.
- Wang, R. Y. (1998). A product perspective on total data quality management. *Communications of the ACM*, 41(2):58–65.
- Wasserman, S. (1995). *Social Network Analysis: Methods and Applications*. Cambridge University Press, Estados Unidos.

- Webb, G. I. (2011). Filtered-top-k association discovery. *WIREs Data Mining and Knowledge Discovery*, 1(3):183–192.
- Weber, K., Otto, B., and Österle, H. (2009). One Size Does Not Fit All—A Contingency Approach to Data Governance. *J. Data and Information Quality*, 1(1). Place: New York, NY, USA Publisher: Association for Computing Machinery.
- Wei, T., Simko, V., Levy, M., Xie, Y., Jin, Y., Zemla, J., et al. (2017). Package corrplot. *Statistician*, 56(316):e24.
- Wickham, H. (2015). *Advanced R*. Chapman & Hall/CRC The R Series. CRC Press.
- Wickham, H. (2016). *ggplot2*. Use R! Springer International Publishing, Cham, second edi edition.
- Wickham, H. (2021). *Mastering shiny*. O'Reilly Media, Inc.™.
- Wickham, H. and Grolemund, G. (2016). *R for Data Science*. O'Reilly Media.
- Wickham, H., Ooms, J., and Müller, K. (2023). *RPostgres: Rcpp Interface to PostgreSQL*. R package version 1.4.5.
- Wikle, C. K., Zammit-Mangion, A., and Cressie, N. (2019). *Spatio-temporal Statistics with R*. Chapman and Hall/CRC.
- Wilks, S. S. (1935). The likelihood test of independence in contingency tables. *Ann. Math. Statist.*, 6(4):190–196.
- Winston, C., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2020). *Shiny: Web Application Framework for R*.
- Wismüller, A., Verleysen, M., Lee, J. A., and Aupetit, M. (2010). Recent advances in nonlinear dimensionality reduction, manifold and topological learning. Conference: *ESANN 2010, 18th European Symposium on Artificial Neural Networks, Bruges, Belgium, April 28-30, 2010, Proceedings*, pages 71–80.
- Wood, S. N. (2006). *Generalized Additive Models - An introduction with R*. Texts in Statistical Science. Chapman & Hall.
- Wu, C. and Thompson, M. E. (2020). *Sampling Theory and Practice*. ICSA Book Series in Statistics. Springer.
- Xiao, N. (2018). Awesome Shiny Extensions.
- Xie, Y. (2017). *Dynamic Documents with R and knitr*. Chapman & Hall/CRC The R Series. CRC Press.
- Xie, Y., Allaire, J., and Grolemund, G. (2019). *R Markdown: The Definitive Guide*. Chapman and Hall/CRC the R Series. Taylor & Francis, CRC Press.
- Zhou, Z.-H. (2012). *Ensemble methods: foundations and algorithms*. CRC press.
- Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B*, 67:301–320.

Índice alfabético

- *bagging*, 467
- ‘fluidPage()’, 767
- accesibilidad, 57
- adstock, 813
- agregación, 298
- agrupación, 501
- ajuste automático, 422, 441, 443, 478, 490, 496
- ajuste semivariográfico, 707
 - automático, 707
 - manual, 707
- alcance, 701
- algoritmo
 - de árbol, 410
 - algoritmo de caja negra, 61
 - alpha de Cronbach, 557
 - ancho de silueta promedio, 521
 - anomalías, 881
 - ANOVA, 246, 258
- análisis
 - geoespacial, 795
 - cluster, 501
 - de colocaciones, 643
 - de componentes principales, 535, 538, 545
 - de correspondencias, 581
 - de datos, 26
 - de redes, 656
 - de sentimientos, 644–646, 651
 - de supervivencia, 864
 - de texto, 912
 - de textos, 641
 - discriminante, 359
 - discriminante cuadrático, 360
 - discriminante lineal, 360
 - factorial, 547, 548, 550, 552, 554, 558, 566
 - factorial confirmatorio, 548
 - factorial exploratorio, 548
- imagen, 557
- análisis conjunto, 375
- análisis de la varianza, 139
- análisis exploratorio de datos, 169
- apilar datos, 51
- aplicaciones web interactivas, 765
- aprendizaje
 - múltiple, 535, 546
 - no supervisado, 535, 536
 - supervisado, 535, 536
- aprendizaje ensamblado, 465, 484
- arista, 412, 660
- arquitectura del dato, 102
- asimetría, 176
- asistencias, 870
- asociación
 - negativa, 392, 399
 - perfecta e implícita de tipo 1, 399
 - perfecta e implícita de tipo 2, 399
 - perfecta y estricta, 399
 - positiva, 392, 399
- association for computing machinery, ACM, 56
- atributo, 67, 173, 387
- Autocorrelación espacial, 713, 717
- autonomía, 57
- autovalor, 537–541, 546, 553–556, 564
- autovector, 537–540, 545, 555
- bagging, 465, 467, 470, 471, 484, 487
- base, 282
- base de conocimiento, 897
- base de datos, 68
 - integración, 93
 - NoSQL, 86, 87, 89
 - relacionales, 70, 86
 - SQL, 70, 86
- batch, 608, 613

- big data, 37, 83
 - V's, 84
 - bigrama, 914, 915
 - bigramas, 912
 - binning, 139, 147
 - bioestadística, 859
 - blog, 764
 - bondad del ajuste, 247
 - boosting, 465, 484, 487, 493
 - bootstrap, 237, 467, 471, 473
 - bootstrapping, 156–158
 - botón circular, 771
 - buenas prácticas, 56
 - c-medias, 533
 - cadenas de caracteres, 42
 - caja negra, 445
 - camino, 663
 - campo aleatorio espacial, 696
 - característica de calidad del dato, 105
 - características de segundo orden, 742
 - cargas, 542
 - cargas factoriales, 549, 555–561, 566
 - categoría, 387
 - catálogos de datos, 102
 - centralidad, 667
 - cheetsheet, 756
 - chunk, 754
 - ciencia, 25
 - empírica, 28
 - formal, 28
 - ciencia de datos, 26
 - clasificación, 263, 501, 596, 599–601, 607, 609, 620, 622, 627, 635, 638
 - multiclas, 440
 - clasificación binaria, 439
 - cluster de variables, 503
 - clusterización, 501
 - jerárquica, 501
 - no jerárquica, 501
 - codificación, 147
 - de etiquetas, 147
 - one-hot, 147
 - coeficiente
 - de correlación lineal
 - cofenético, 519
 - divisivo, 517
 - de asimetría, 176
 - de coincidencias simple, 507
 - de congruencia, 506
 - de correlación, 186
 - de Kendall, 506
 - de Pearson, 139, 506
 - de Spearman, 139, 506
 - de Czekanowski, 507
 - de Gower, 508
 - de Jacard, 507
 - de Rogers-Tanimoto, 507
 - de Russell y Rao, 507
 - de Sokal y Sneath, 507
 - de variación, 176
- coeficiente de
- determinación, 161
 - determinación ajustado, 161
- coeficiente de correlación parcial, 553
- coeficientes de regresión, 246
- coeficinte
- de apuntamiento, 176
- colecciones, 798
- colinealidad, 137
- combinación lineal, 137
- comercio, 845
- complejidad, 417, 439, 472, 473, 493
- completitud, 550
- componentes principales, 318, 503, 535, 538–540, 542, 544–546, 549, 554, 566
- estimación, 540
- interpretación, 542
- número de componentes a retener, 540
- obtención, 535
- componentes reactivos, 776
- comunalidad, 548, 550, 553–559, 564
- comunalidades, 669
- construcción, 849
- consumo eléctrico, 889
- contingencia, 387
- contraste
- de Breusch-Pagan, 253
 - de esfericidad de Bartlett, 552
 - de razón de verosimilitudes, 564
 - de Shapiro-Wilk, 253
- contraste de hipótesis, 213

Índice alfabetico

955

- contraste de independencia, 390
 - bilateral, 391
 - contraste Chi-cuadrado, 397
 - contraste razón de verosimilitudes, 398
 - en tablas 2x2, 390
 - contraste aproximado, 390, 394, 396
 - contraste aproximado con corrección de continuidad, 390, 395, 396
 - contraste de razón de verosimilitudes, 396
 - contraste exacto, 390, 392, 396
 - test exacto de Fisher, 392, 393
 - total muestral fijo, 396
 - totales marginales de ambos factores fijos, 392
 - totales marginales de un factor fijos, 396
- en tablas multidimensionales, 405
- en tablas RxC, 396
 - contraste aproximado, 397
 - contraste aproximado con corrección de continuidad, 399
 - unilateral, 391
- control calidad, 129
- control de versiones, 782
- control deslizante, 770
- coordenadas, 681
- coordinate reference system, 681
- coronavirus, 859
- corpus, 642
- corrección de continuidad
 - de Yates, 395
- correlación, 136, 186
- covariables, 261
- covarianza, 186
- COVID-19, 843
- COVID-19, 859
- criterio
 - de Catell, 541
 - de Kaiser, 541
 - de Kaiser, 554
 - del bastón roto, 541, 542
- criterio del gap, 521
- cross validation, 893
- CRS, 681
 - geográficos, 683
 - proyectados, 684
- cuarteto de Anscombe, 169
- cuasicombinación lineal, 137
- cultura del dato, 102
- curtosis, 176
- curva ROC, 265
- curvilinear component analysis, 545
- código abierto, 782
- DAMA, 104
- Daniel G. Krige, 709
- dashboard, 763
- data augmentation, 629, 630
- data.frame, 41
- datos
 - sintéticos, 154
 - atípicos, 124
 - brutos, 115
 - de patrones de puntos, 680
 - desequilibrados, 922
 - duplicados, 123
 - espaciales, 679
 - estructurados, 85
 - faltantes, 121
 - geo-referenciados, 679
 - geoestadísticos, 680
 - geográficos, 679
 - integración, 115, 119
 - jerárquicos, 298
 - lattice, 680
 - limpieza, 119
 - longitudinales, 299
 - missing, 126
 - no equilibrados, 151, 154, 165
 - no estructurados, 85
 - problemas de calidad, 119
 - raster, 687
 - repositorio, 68
 - reticulares, 680
 - semiestructurados, 85
 - sintéticos, 922
 - vector, 685
- DBI, 73
- DBSCAN, 530, 531
- DCAM, 104
- deep learning, 593–597, 608, 629
- DENCLUE, 530

- dendrograma, 511, 520
- deontología, 56
- Dependencia espacial, 713
- dependencia espacial, 695, 696, 699, 700, 702, 709, 713
 - análisis estructural, 696
- desagregación, 298
- descenso del gradiente, 595, 607, 624
- desviación
 - absoluta mediana, 176
 - típica, 176
- detección, 596
- detección de emociones, 645, 652
- deviance, 161, 283, 318
- diagrama
 - de dispersión, 250
 - diagrama de doble escala, 830
 - Diagrama de Moran, 717
 - diagrama de Shepard, 573, 577
 - diagrama lineal, 827
- DIANA, 517
- diccionarios de datos, 102
- dimensión de calidad del dato, 105
- diseño experimental, 389
 - total muestral fijo, 390
 - totales marginales de ambos factores fijos, 389
 - totales marginales de uno de los factores fijos, 390
- disimilaridad, 503
- Distancia
 - reproducida, 570
- distancia, 451, 503, 567–571
 - angular, 683
 - Chebychev, 506
 - coseno, 506
 - de Chebychev, 506
 - de Gower, 451
 - de Mahalanobis, 506
 - de Minkowski, 505
 - entre centroides, 511
 - euclídea, 451, 504
 - euclídea al cuadrado, 504
 - inter-grupos, 514, 521
 - intra-grupo, 529
 - intra-grupos, 521
- Manhattan, 505, 528, 529
- manhattan, 451
- media, 511
- Minkowski, 505
- distribución, 846
- distribución Normal, 245
- distribución normal, 202
- dlookr, 112
- DMBoK, 104
- dominios irregulares, 738
- downsampling, 154
- Dropout, 628
- EAM, 161
- early stopping, 628
- ECM, 158, 159, 161
- ecuaciones de Henderson, 301
- edad, 843
- efecto pepita, 702
 - puro, 702
- efectos
 - aditivos, 261
 - aleatorios, 299
 - fijos, 299
 - multiplicativos, 261
- elastic net, 333
- elecciones Andaluzas, 835
- elemento
 - atípico, 531
 - central, 531
 - denso-alcanzable, 531
 - denso-alcanzable directamente, 531
 - denso-conectado, 531
 - frontera, 531
- EM, 530, 533
- emociones, 645
- ensayo clínico, 860
- entidad, 67
- entropía, 415, 471, 474, 535
- epsilon-neighborhood, 531
- equidad, 57
- equilibrados, 924
- ergodicidad, 337, 697
- error
 - de continuización, 395
 - del modelo, 246

Índice alfabetico

957

- estructural, 121
 - a nivel de conjunto de datos, 121
 - a nivel de variable, 122
- error de predicción, 710
 - desviación típica, 711
- error muestral, 231
- escala, 449
- escalamiento
 - multidimensional, 567, 569
 - métrico, 571, 572, 574, 579
 - no métrico, 571, 575, 579
- escalamiento multidimensional, 535
- especificidad, 265, 550
- estacionariedad, 337
- estadística, 28
 - estadística descriptiva, 169, 170
 - estadística espacial, 731
 - estadístico Chi-cuadrado, 394, 397
 - ajustado, 394
 - corregido de continuidad de Yates, 395
 - corregido de continuidad de Yates, 395
 - estadístico G, 398
 - estandarización, 325
 - estilometría, 644
 - estimación
 - por intervalos, 213
 - puntual, 213
 - estimación de coeficientes, 263
 - estimador
 - de mínimos cuadrados, 217
 - máximo verosímil, 217
 - estimador de difusión, 738
 - estimador de remuestreo-suavizado, 741
 - estructura de dependencia espacial, 742
 - estructura factorial, 551, 558, 560, 563
 - de referencia, 560
 - oblicua, 560
 - estructura multidimensional, 880
 - estructura simple, 556, 558, 560
 - oblicua, 560
 - ortogonal, 560
 - estudio observacional, 860
 - evaluación de calidad del dato, 109
 - evaluación de modelos, 151
 - exactitud, 152, 164
 - EXP (Holdout Testing o Experiments), 812
 - expecificidad, 164
 - experimentos aleatorios, 196
 - explicabilidad, 57
 - extender datos, 51
 - extraer datos, 48
 - extreme gradient boosting, 493
 - facebook, 660
 - facetas (gráficos), 184
 - factor, 42, 173, 387
 - factores, 261
 - comunes, 548–551, 553–559, 562–564, 566
 - de referencia, 560
 - específicos, 549, 564
 - subyacentes, 547, 565
 - únicos, 549, 550, 564
 - falso negativo, 164
 - falso positivo, 164
 - feature engineering, 133, 147
 - feature selection, 133, 134
 - fechas, 42
 - fenómeno regionalizado, 695
 - filograma, 513
 - filosofía, 55
 - filtrar datos, 47
 - Fisher-Jenks, 690
 - Forgy, 526
 - frecuencia
 - esperada, 392, 397
 - frequency, 852
 - frontera
 - de decisión, 437
 - fuentes de asociación, 399, 402, 404
 - funciones núcleo, 733
 - función
 - de autocorrelación (ACF), 337
 - de autocorrelación parcial (PACF), 337
 - de autocovarianza, 337
 - de enlace, 260, 261
 - de pérdida, 440
 - discriminante de Fisher, 363
 - función aleatoria
 - espacial, 696
 - estacionaria de segundo orden, 696
 - estacionaria en sentido estricto, 696
 - intrínsecamente estacionaria, 696

- no estacionaria, 696
- función de covarianza, 700
- función núcleo-calor, 738
- fútbol, 869
- geoestadística, 695
- geoprocесamiento, 795, 881
- gestión de calidad del dato, 99
- gestión del dato, 99
- git, 781
- github, 781
- glosario de negocio, 102
- gobierno del dato, 99
- goles, 870
- goles esperados, 872
- google earth engine, gee, 795
- gradient boosting, 472, 484, 487, 493, 494
- grado, 663
- grados de libertad efectivos, 282
- grafo, 659, 660
 - betweenness, 670
 - eigenvector, 670
 - walktrap, 670
- gráfico, 172
 - de barras, 174
 - de cajas, 181
 - de densidad, 179
 - de dispersión, 186
 - de violín, 181
- gráfico de dispersión, 825
- gráfico de sedimentación, 520, 541, 554
- gráfico de silueta, 528
- hard margin, 440
- heatmap, 505
- hiperparámetro, 151, 159, 422, 441, 443, 450, 472, 473, 486, 487, 493
- hiperparámetros, 530, 892
- hiperplano, 440
- histograma, 178
- hoja, 412
- hoja de cálculo, 37
- huella lingüística, 644
- I de Morna, 719
- identificación de los factores, 551
- impactos negativos, 56
- importancia, 445, 446, 470
- importar datos, 43
- impureza
 - de Gini, 412, 474
- imputación del dato, 113
- incendios forestales, 735
- incertidumbre, 195
- indentación, 756
- independencia
 - condicional, 405
 - global, 405
 - parcial, 405
- independencia condicional, 459
- índice
 - de complejidad de Hoffmann, 556
 - KMO, 552, 553
- índice de Dunn, 521
- industria, 849
- INE, 340
- inferencia estadística, 169
- información espacio-temporal, 880
- información redundante, 535
- informes, 751
 - bibliografía, 756
 - gráficos, 761
 - parámetros, 759
 - plantilla, 756
 - referencias cruzadas, 761
- informática, 28
- instalar git, 783
- inteligencia artificial, 593
- intensidad, 733
- intercambiabilidad, 299
- interfaz de usuario, 766
- internet of things, 37
- interpolation, 709
- interpretación de los factores, 552, 558, 562
- Interpretación de modelos espaciales, 725
- intervalos de confianza, 213
- IPC, 340
- ISO/IEC 25012, 107
- ISO/IEC 25024, 109
- John Snow, 679
- justicia, 57
- k-fold cross-validation, 329

Índice alfabético

959

- k-medianas, 529
- k-medias, 526
 - armónicas, 527
 - difuso, 527
 - recortadas, 527
 - sparse, 527
 - sparse robusto, 527
- k-medoides
 - CLARA, 528
 - CLARANS, 529
 - PAM, 527
- k-vecinos, 449, 451
- kernel, 441, 442, 447
- kriging, 709
 - ordinario, 709
 - simple, 709
 - universal, 709
- Laplacian eigenmaps, 545
- latent dirichlet allocation, LDA, 912
- latex, 763
- lematización, 644
- lexicón, 645
- LibreOffice, 755
- limpieza del dato, 112
- lista, 42
- literate programming, 752
- lm, 249
- loadings, 542, 543, 556
- machine learning, 445, 545, 593–595, 892
- MAMD, Modelo Alarcos de Madurez de Datos, 105
- manifold learning, 535, 546
- mapa, 688
 - de coropletas, 689
 - espacio-temporal, 691
 - interactivo, 692
- mapa de calor, 845
- mapas, 881
- mapping, 710
- Mar Mediterráneo, 881
- margen, 437, 439, 440
- markdown, 751
- marketing, 811
- Marvel, 672
- matemáticas, 28
- matriz, 41
 - de cargas, 543
 - de correlaciones, 539, 540, 546
 - de covarianzas, 536, 538, 540, 546
 - de puntuaciones, 544
 - cofenética, 519
 - de cargas, 551, 555, 558
 - de correlaciones, 186, 548, 551, 552, 564, 565
 - de correlaciones anti-imagen, 552
 - de correlaciones de los residuos, 553
 - de correlaciones reducida, 555
 - de correlaciones reproducida, 553
 - de covarianzas, 557, 565
 - de disimilaridad, 568, 571, 572, 576
 - de distancia, 568
 - de distancias, 503
 - de proximidad, 567–569, 572
 - de similaridad, 568, 569
 - de transformación, 558, 560
 - de varianzas-covarianzas, 186
- matriz de adyacencia, 662
- matriz de confusión, 163
- matriz de documentos, 646, 653
- maximum variance unfolding, 545
- mayores de 45 años, 850
- MCLUST, 530
- media, 175
- mediana, 175
- medición de calidad del dato, 109
- medida de adecuación muestral, 552
- medidas de asociación, 139
 - para tablas 2×2 , 507
 - en tablas 2x2, 399
 - cuadrado medio de la contingencia, 401
 - odds ratio, 401
 - Q de Yule, 399
 - riesgo relativo, 401
 - V de Cramer, 401
 - en tablas RxC, 402
 - basadas en la reducción proporcional del error, 403
 - coeficiente de contingencia, 401, 403
 - cuadrado medio de la contingencia, 403

- derivadas del estadístico Chi-cuadrado, [402](#)
- lambda de Goodman y Kruskal, [403](#)
- T de Tschuprow, [403](#)
- V de Cramer, [403](#)
- mejora del dato, [112](#)
- meseta, [701](#)
 - parcial, [709](#)
- Messi, [869](#)
- meta-pakage, [143](#)
- metadato, [102](#)
- metapaqute, [143](#)
- metodología, [31](#)
 - CRISP-DM, [32](#)
 - KDD, [32](#)
 - SEMMA, [32](#)
 - Metodología CRISP-DM, [852](#)
- Microsoft Word, [755](#)
- minería
 - de opinión, [645](#)
 - de textos, [641](#)
- missing, [52](#)
- missing values, [535](#)
- MMM (Marketing Mix Modeling), [811](#)
- moda, [175](#)
- modalidad, [387](#)
- modelado de temas, [644](#)
- modelo
 - aditivo, [279](#)
 - ARIMA, [339](#)
 - de regresión, [245](#)
 - de regresión lineal múltiple, [246](#)
 - lineal, [246](#)
 - lineal generalizado, GLM, [245, 260](#)
 - logarítmico lineal, [406](#)
 - no lineal, [246](#)
 - sparse, [330](#)
- modelo de calidad del dato, [107](#)
- modelo RFM, [852](#)
- Modelos económicos espaciales, [719](#)
- modelos mixtos lineales, [301](#)
- monetary, [852](#)
- mongoDB, [93](#)
- moral, [55](#)
- motor de inferencia, [897](#)
- MTA (Multitouch Attribution), [811](#)
- muestra, [169, 170](#)
- muestra aleatoria simple, [152, 214](#)
- muestra bootstrap, [158](#)
- muestreo, [229](#)
 - aleatorio estratificado, [153](#)
 - aleatorio simple, [152](#)
 - estratificado, [234](#)
 - por conglomerados, [236](#)
 - sistemático, [237](#)
- multicolinealidad, [137, 248, 535](#)
 - consecuencias, [137](#)
 - fuentes, [137](#)
- máxima verosimilitud, [708](#)
- máxima verosimilitud compuesta, [708](#)
- máxima verosimilitud restringida, [708](#)
- máximo, [175](#)
- método, [25](#)
 - alpha, [557](#)
 - CHAID, [519](#)
 - científico, [25](#)
 - de Anderson-Rubin, [565](#)
 - de Barlett, [565](#)
 - de componentes principales, [554, 555, 557](#)
 - de Fortin, [530](#)
 - de la descomposición triangular, [557](#)
 - de la distancia entre centroides, [511](#)
 - de la distancia media, [511](#)
 - de la distancia promedio, [515](#)
 - de la mediana, [512](#)
 - de Lance y Williams, [514](#)
 - de las combinaciones de Wolf, [530](#)
 - de los factores principales, [555–557, 562, 565, 566](#)
 - de máxima verosimilitud, [260, 557, 564](#)
 - de mínimos cuadrados, [247, 249](#)
 - de mínimos cuadrados generalizados, [558](#)
 - de mínimos cuadrados no ponderados, [558](#)
 - de Thompson, [565](#)
 - de Ward, [512](#)
 - del análisis de la asociación, [517](#)
 - del análisis imagen, [557](#)
 - del centroide, [557](#)
 - del detector automático de interacciones, [518](#)
 - del encadenamiento intra-grupos, [513](#)
 - del vecino más cercano, [510](#)

Índice alfabético

961

- del vecino más lejano, 510
- minres, 557
- modal de Wishart, 530
- TaxMap, 530
- método de información mutua, 139
- métodos basados en máxima verosimilitud, 708
- mínimo, 175
- mínimos cuadrados generalizados, 708
- mínimos cuadrados ordinarios, 708
- mínimos cuadrados ponderados, 708
- n-gramas, 643, 654
- Naive Bayes, 458, 459
- neurona, 598, 599, 601, 607, 620, 624
- nivel, 387
- nodo, 412, 660
- nodo raíz, 412
- normalización, 146
 - z-score, 146
 - min-max, 146
- nube, 795
- nube de palabras, 822, 824
- nubes de palabras, 646, 650, 653
- nubes dinámicas, 526
- nugget effect, 702
- número
 - de árboles, 473, 487
- número óptimo de clusters, 520, 528
- observaciones independientes, 246
- océanos, 879
- odd ratio, OR, 264
- operaciones CRUD, 69
- OPTICS, 530
- ordenar datos, 48
- outlier, 124
- outliers, 541
- overfitting, 156, 159
- oversampling, 154
- p-valor, 219, 247, 263
- padding, 625
- palabras vacías, 642, 649, 655
- pancarta, 517
- pandoc, 755
- paquete, 38
- parada temprana, 417
- paro de muy larga duración, 848
- paro registrado, 843
- partial least squares, 535, 536, 546
- partición, 412, 428, 474
- partición del conjunto de datos, 151
- parámetro de escala, 704
- parámetro de suavizado, 280, 734
- pat, 785
- patrones espaciales, 731
- patrones espaciales de puntos, 731
- patrón factorial, 550, 551, 558, 559, 561, 564
- penalización shrinkage, 325
- pequeños múltiples, 881
- perceptrón, 595, 598–600, 605, 607, 620, 624
 - multicapa, 605, 620, 638
- perfil, 376
- perfil de los clientes, 851
- perfilado del dato, 111
- periódico, 835
- población, 169, 170
- poda, 418
- política del dato, 102
- pooling, 626, 627
- poyeción
 - Robinson, 684
- precisión, 164
- predicción
 - del modelo, 261
 - del modelo logístico, 263
- predicción del modelo lineal, 248
- predicción kriguada, 696
- preprocesamiento, 115, 632
- probabilidad, 195
- procedimiento de muestreo, 389
- procesamiento del lenguaje natural, NLP, 911
- Procesamiento del Lenguaje Natural, PLN, 641
- proceso estocástico, 337
 - espacial, 696
- profundidad
 - del árbol, 417, 431, 473, 487
- programación, 28
- proyectos, 40
- punto
 - de decisión, 412
- punto de corte, 272

- puntuaciones factoriales, 551, 564, 565
- python, 753
- quarto, 751
- R, 37
- R cuadrado, 247
- R markdown, 751
- rama, 412
- random forest, 471–474, 484, 487
- rango, 701
- rango intercuartílico, 176
- ranking de percentiles, 852
- razón de riesgo, 866
- reactividad, 776
- recency, 852
- RECM, 161
- reconocimiento, 596
- recuento, 260
- red neuronal, 889, 892
 - artificial, 598, 620–622
 - convolucional, 596, 621, 638
- red social, 660
- redes lineales, 743
- reducción de la dimensionalidad, 535
- referencias cruzadas, 762
- regionalización, 696, 700, 709
- región de tolerancia, 706
- región mediterránea, 879
- regla, 897
- regresión
 - de Cox, 866
 - de Poisson, 266
 - logística, 262
- regresión lasso, 330
- regresión logística, 139
- regresión ridge, 324
- regularización, 493
- RELCM, 161
- remuestreo, 151
- repositorio local, 792
- repositorio remoto, 792
- reproducibilidad, 751
- residuos, 248
 - de Haberman, 404
 - estandarizados, 303, 404
- ajustados, 402, 404
- parciales, 289
- studentizados, 303
- restauración, 845
- riesgo, 864
- riesgo relativo, RR, 265
- RMySQL, 73
- ROI (*Return On Investment*), 811
- Ronaldo, 869
- rotaciones
 - directas, 560
 - indirectas, 560
 - oblicuas, 558, 560–562
 - ortogonales, 552, 558, 559, 561, 562
- rotación
 - BINORMALMIN, 561
 - BIQUARTIMAX, 560
 - BIQUARTIMIN, 561
 - COVARIMIN, 561
 - EQUAMAX, 560
 - OBLIMAX, 560, 561
 - OBLIMIN, 561
 - ORTOBPLICUA, 561
 - ORTOMAX, 559
 - PROMAX, 561
 - QUARTIMAX, 559, 560
 - QUARTIMIN, 561
 - VARIMAX, 559–562, 565
 - VARIMAX normalizada, 559, 562
- Rstudio, 782
- S-Stress, 570
- saber, 25
- Sammon's mapping, 545
- script, 37, 759
- sector, 843
- segmentación, 596
- segmentar, 851
- seleccionar columnas, 48
- selección
 - tipo envoltura (wrapper), 134, 141
 - tipo filtro, 134, 138
 - tipo intrínseco (embedded), 134, 142
- selección de variables, 133, 134, 317
- selección stepwise, 321
 - backward, 321

Índice alfabético

963

- forward, 321
- semivariograma, 698–700, 702
 - comportamiento, 701
 - con meseta, 703
 - con meseta y efecto hoyo, 704
 - empírico, 706
 - J-Bessel, 704
 - logarítmico, 705
 - potencial, 705
 - sin meseta, 705
 - anisotrópico, 700
 - comportamiento, 702
 - efecto pepita puro, 704
 - esférico, 703
 - exponencial, 703
 - gausiano, 704
 - isotrópico, 700, 703
 - válido, 702, 707
- semivariograma empírico
 - direccional, 706
 - omnidireccional, 706
- sensibilidad, 164, 265
- serie temporal, 337
- serie temporal (gráfico), 187
- servidor, 766
- sesgo, 158
 - de adquisición, 58
 - de implementación, 58
 - de medida, 58
 - de representación, 58
 - histórico, 58
- sesgo de selección, 860
- sexo, 843
- Shapiro-Wilks, 226
- shiny, 763
- Shrinkage, 318
- shrinkage, 324
- siete puentes de Königsberg, 659
- significativo, 247, 263
- similaridad, 507
- similitud, 504
- sistema de gobierno del dato, 101
- sistema de referencia de coordenadas, 681
- sistema experto, 897
- sistema gestor de bases de datos, 68
 - NoSQL, 68
 - relacionales, 68
- sobreajuste, 417, 467, 484, 487, 493, 494, 628
- sobreentrenamiento, 895
- sobremuestreo, 922
- soft margin, 440
- solución factorial, 563
 - completa, 551, 566
- sparse PCA, 545
- spline, 286
- splines, 709
- SQL, 70
- ssh, 784
- stemming, 644
- stopwords, 642, 649, 655
- Stress de Kruskal, 570, 573
- stride, 625
- strings, 42
- sub-muestreo, 741
- subconjunto
 - de entrenamiento, 152, 156
 - de entrenamiento propiamente dicho, 156
 - de test, 152
 - de validación, 156
 - detest, 156
 - de validación, 155
 - testing, 323
 - training, 323
- submuestreo, 922
- SVM, 439, 447
- t-distributed stochastic neighbor embedding, 545
- t-SNE, 545, 546
- tabla
 - de contingencia, 181, 387
 - de frecuencias, 173, 178
- tabla de contingencia, 388
- tabla de contingencia 2×2 , 518
- tablas
 - de contingencia 2x2, 389
 - de contingencia multidimensionales, 389
 - de contingencia RxC, 389
 - formateadas, 761
- tablas de contingencia, 139
- tabPanel(), 769
- tabsetPanel(), 769

- tanglegrama, 512
- tasa
 - de aprendizaje, 487
- tasa de riqueza léxica, TTR, 643
- temperatura superficial, 797, 879
- tendencia, 337
- teorema
 - de Bayes, 457–459
- test de esfericidad, 542
- tibble, 844
- tidyverse, 45
- tiempo de búsqueda de empleo, 843
- token, 643
- tokenización, 643
- trade off
 - sesgo-varianza, 151, 158
- transformación de variables, 133
- transparencia, 57
- triaje, 897
- truco
 - del kernel, 440
- Tukey, 26
- turismo, 845
- twitter, 821
- técnicas de agrupación, 501
 - jerárquicas, 508, 510
 - aglomerativas, 510
 - divisivas, 510, 515
 - no jerárquicas, 508, 510
 - basadas en la densidad de elementos, 510, 525
 - basadas en mixturas de modelos, 525
 - bi-cluster, 525, 532
 - block-cluster, 525, 532
 - clustering basado en mixturas de modelos, 533
 - clustering difuso, 533
 - co-cluster, 525, 532
 - de clusterización difusa, 525
 - de reasignación, 510, 525
 - de reducción de la dimensionalidad, 525, 532
 - directas, 525, 532
 - otras, 510
 - Q-cluster, 525, 532
 - R-cluster, 525, 532
- two mode-cluster, 525, 532
- técnicas híbridas, 535, 536, 544, 546
- término independiente, 246
- UMA (Unified Measurement Approaches), 812
- underfitting, 156, 159
- undersampling, 154
- unicidad, 550, 555, 556, 558, 566
- uniones
 - filtrado, 116
- upsampling, 154
- utilidades parciales, 375
- validación, 129, 155
 - cruzada con repetición, 156
 - cruzada k-grupos, 156
 - cruzada, 156
 - dejando uno fuera, 157
 - validación cruzada, 542
 - validación cruzada k-grupos, 329
- valor, 55
- valores perdidos, 52
- valores regionalizados, 696
- valores semivariográficos, 706, 707
- variabilidad, 535–538, 541, 546, 547, 550
- Variable
 - cuantitativa, 175
- variable
 - continua, 175
 - cualitativa, 173
 - discreta, 175
 - explicativa, 245
 - irrelevante, 134
 - redundante, 134
 - respuesta, 245, 259
- variable aleatoria, 198
- variable regionalizada, 696
- variables estandarizadas, 539, 547
- varianza, 176
 - intra-cluster, 513
 - compartida, 548, 566
 - común, 548, 555, 566
 - específica, 548
 - inter-grupos, 525
 - intra-grupos, 525, 526
 - no compartida, 549

Índice alfabético

965

- residual, 548
- única, 548, 550, 555, 557
- varianza a priori, 701
- varianza cercana a cero, 135
- varianza cero, 135
- varianza de predicción, 158
- vector, 41
- verdadero negativo, 164
- verdadero positivo, 164
- virus SARS-CoV-2, 859
- Visualización, 881
- visualización, 124, 169, 821
- Voronoi, 740
- WAVECLUSTER, 530
- web scraping, 869
- weight decay, 629
- YAML, 753
- ángulo de tolerancia, 706
- árbol
 - de clasificación, 409, 412, 420, 465, 473, 474
 - de decisión, 409, 465, 467, 470, 471, 473, 474, 484, 487, 493
 - de regresión, 410, 427, 432, 465, 473, 474
- área bajo la curva ROC, 165
- ética, 55
- índice de propensión, 860