eman ta zabal zazu

Universidad del País Vasco

Euskal Herriko Unibertsitatea

TAREA 9

EJERCICIOS

1. Los datos para este ejercicio proceden del sitio Web idealista.com[1], dedicado a anuncios inmobiliarios. Tienes un facsímil de la información correspondiente a una vivienda en el Apéndice C. Están en una *dataframe* de nombre pisos.dge, en el lugar habitual. Son una muestra formada por 1031 inmuebles del Gran Bilbao y alrededores, puestos a la venta durante el periodo 2009-05-08 a 2010-11-30.

   Puedes ver un resumen de las variables recogidas en el Anexo A.1.

   El mercado de la vivienda presenta características peculiares: cada bien objeto de transacción en el mercado es único e irrepetible. Cada vivienda tiene una situación, altura, orientación, estado de conservación y calidad de acabado que no puede reproducir exactamente ninguna otra.

   Con todo, hay que esperar que el mercado valore de forma más o menos coherente los diferentes equipamientos y ubicaciones, de manera que un modelo ajustando el precio unitario (o el precio por m^2) de cada vivienda podría dar cuenta de una parte apreciable de la dispersión de dichas variables.

   El resultado final de tu trabajo debe ser un modelo de valoración. Siéntete libre de hacer lo que mejor sirva a tus propósitos, sin necesidad de seguir las directrices que, como simple orientación, se te dan más abajo.

   a) Obtén el precio medio por m^2 construido en las diferentes zonas. Mira entre los comentarios de ayuda el 1*c*, con un fichero fuente de ilustración que te puede orientar en la forma de hacer tus propios análisis. ¿Varían los precios por área?

   b) Puedes plantearte un modelo cuya variable respuesta sea `Precio`, o quizá `Precio/M2`, o incluso los logaritmos respectivos. Explica las implicaciones de emplear una especificación u otra: ¿Cómo varía la interpretación de los β's?

   c) Busca un modelo que te parezca adecuado. En el contexto del mismo, responde a las siguientes cuestiones:

      1) ¿Hay alguna evidencia de que los pisos "altos" se cotizan más?
      2) ¿Cuál es el efecto *ajustado por características de las viviendas* de las diferentes ubicaciones sobre el precio del m^2 construido? (Compara con tu respuesta a la pregunta 1*a* más arriba).
      3) ¿Son casas grandes proporcionalmente más baratas que las más pequeñas?
      4) ¿Influye el número de habitaciones de forma adicional a la superficie en la valoración de las viviendas?
      5) ¿Cómo influye la antigüedad de las viviendas?

[1] En `http://www.idealista.com`.

6) ¿Cuál es la valoración de una plaza de garaje? ¿Parece uniforme en todas las zonas?
7) ¿Cuál es la repercusión sobre el precio unitario (o por m^2) de los diferentes equipamientos de calefacción? ¿De la existencia o no de ascensor?
8) ¿Hay alguna evidencia de cambio en los precios a lo largo del tiempo?

d) Haz un análisis de residuos. ¿Hay observaciones anómalas? ¿Muy influyentes? Explica lo que observes.

2. La *dataframe* `diamonds` forma parte del paquete `ggplot2`; puedes obtenerla haciendo

```
> library(ggplot2)
> data(diamonds)
```

Verás que tiene 53940 observaciones. Un descripción sumaria aparece en el Apéndice A.2.

Como en el ejercicio precedente, tu trabajo consiste en elaborar un modelo de valoración, que de aproximadamente el precio de una gema en función de sus características físicas.

AYUDAS, SUGERENCIAS Y COMPLEMENTOS

1. En relación a primer ejercicio:

a) *No intentes* hacer uso en este ejercicio de métodos automáticos de ayuda a la especificación de modelos, tipo *stepwise* o búsqueda sobre todos los subconjuntos. Nota que hay muchas variables con valores perdidos, y dependiendo de los regresores que incluyas tu muestra será muy diferente en tamaño, invalidando los criterios que presuponen una muestra de tamaño constante.

b) El sistema de calefacción y de agua caliente están codificados de la manera que lo hacen los anunciantes. Puedes simplificar dicha clasificación recodificando algunas variables.

c) Tienes, en el lugar habitual, un fichero de nombre descrip.R que realiza algunos análisis descriptivos de las variables objeto de estudio.

d) El código indicado en descrip.R codifica también la variable `Planta` convirtiéndola en un "factor ordenado" (en la terminología de R, una variable cualitativa cuyos valores o *niveles* tienen un orden natural). Podrías en efecto pensar que una vez que una vivienda es "alta", da lo mismo que sea un octavo o un décimo segundo piso. Adopta tú la codificación que desees utilizando como modelo la que se te facilita.

e) Hay muchos `NA`. Cabe pensar que cuando un oferente no señala sistema de calefacción, es que la vivienda no lo tiene[2]: puedes recodificar los `NA` de las variables `CA` y `AC` a "No". Por el mismo motivo, cabe imaginar que cuando no se señala la existencia de plaza de garaje, es que no la hay: puedes recodificar los `NA` de dicha variable a cero. Puedes plantearte hacer algo similar con la variable `Ascensor`.

f) La variable `Precio` está expresada en €. Es precio demandado por el oferente. Es lo normal que la compraventa, si finalmente se realiza, se efectúe a un precio inferior.

g) La superficie (variable `M2cons`) se especifica en m^2 construidos; sólo para algunos inmuebles se dispone además de la variable `M2util`. Como en unos casos se señala una variable y en otros otra, para componer una variable de superficie que abarque el mayor número posible de observaciones, puedes hacer uso de la regla según la cual $\texttt{M2util} \approx 0{,}85 \times \texttt{M2cons}$.

h) En la *data frame* encontrarás columnas dos columnas (`UTMX` y `UTMY`) dando la ubicación de cada inmueble en coordenadas UTM. Tienes también coordenadas geográficas (`lat` y `lon`). Tienen por objeto permitirte situar los inmuebles y representar, si quieres, sobre un mapa residuos, valores estimados, etc.

[2] Pues es un argumento de venta que nadie dejaría de dar.

i) Las coordenadas de cada inmueble son las del portal correspondiente, y han sido obtenidas de la web del Ayuntamiento de Bilbao. En Bizkaia.net[3], en el apartado "Cartografía y planes urbanísticos", encontrarás aplicaciones permitiendo localizar cualquier dirección en un callejero. La Diputación Foral de Vizcaya ha creado también un CD ROM[4] con una aplicación basada en Windows que permite hacer lo propio, quizá con más comodidad, en tu PC. Puedes no obstante ver la muestra geolocalizada en esta página. ¡Ten paciencia, tarda un poquito!

j) Los modelos como el que se espera que especifiques y ajustes, se conocen en Economía con el nombre genérico de modelos hedónicos *(hedonic models)*. Puedes obtener alguna información sobre los mismos en el capítulo correspondiente de la Wikipedia[5]. Tu profesor(a) de Microeconomía te proporcionará más información y referencias si se las pides.

k) Hay una variable (`FechaAnuncio`) con formato de fecha[6]. Si quisieras obtener una variable "tiempo" para ajustar una tendencia en, e.g., días a partir del 1-1-2009, sería fácil de generar:

```
> tiempo <- pisos$FechaAnuncio - as.Date("2009-01-01")
```

2. En relación al segundo ejercicio:

a) Puedes obtener información sobre los diamantes y su talla en sitios como la Wikipedia: mira en http://en.wikipedia.org/wiki/Diamond_cut. La figura que se reproduce en el Apéndice A.2 procede de http://es.bluenile.co.uk/. Si buscas en Google tecleando "diamond cut table" obtendrás más información de la que puedes desear.

b) Ten presente que el precio de una gema aumenta con su rareza y un diamante "el doble" de grande es *mucho* más de dos veces más raro. Te puede interesar encontrar un modelo lineal para el logaritmo (o alguna potencia menor que la unidad) del precio.

c) Manejando muestras tan grandes, muchos plots pueden resultar de escasa utilidad; sólo verás una mancha negra, sin distinguir apenas puntos individuales, sea lo que fuere lo que dibujes.

Puedes recurrir a emplear una submuestra[7].

3. En relación a ambos ejercicios:

a) Este es la primera tarea en que te encuentras con datos no "de libro", con todos los inconvenientes que presenta la información en el mundo real: ¡invertirás más tiempo en el manejo de los datos y su adaptación a tus deseos que en el análisis propiamente dicho! El código en descrip.R se te proporciona como ilustración; no tienes que ceñirte servilmente a lo que allí hay, es sólo un ejemplo.

Puedes también querer servirte de libros como [16].

b) Otros libros sobre R que pueden servirte son los tantas veces citados [18], [3], [10] y [9], [2]. Sobre el uso de R para estimar modelos de regresión, [6], [8] y [5]. Manuales generales sobre regresión y modelos lineales son: [4], [17] [14] (nueva edición [15]). y [11] entre otros.

[3] En `http://www.bizkaia.net`.
[4] Pídelo prestado a tu profesor, si quieres utilizarlo.
[5] En `http://en.wikipedia.org/wiki/Hedonic`.
[6] Formato `Date`; hay muchos modos de representar fechas en R. Mira por ejemplo [7] o [13].
[7] Otra alternativa (que te requerirá estudiar con algún detalle el paquete `ggplot2`, [19], u otros similares) es emplear gráficos especificando un grado de transparencia.

A. Resumen de los datos

A.1. Inmuebles

```
> summary(pisos)
```

```
      lat             lon          TipoInmueble     Precio
 Min.   :42.73   Min.   :-5.507   ático  : 67   Min.   :  80000
 1st Qu.:43.25   1st Qu.:-2.945   chalet :  4   1st Qu.: 220000
 Median :43.26   Median :-2.933   dúplex :  9   Median : 287885
 Mean   :43.26   Mean   :-2.934   estudio:  7   Mean   : 317820
 3rd Qu.:43.26   3rd Qu.:-2.922   piso   :944   3rd Qu.: 385000
 Max.   :43.36   Max.   :-1.647                 Max.   :1260000
 NA's   :12.00   NA's   :12.000
     Planta         Dormitorios          WC              CP
 Min.   :  1.000   Min.   :1.000   Min.   :1.000   48003  :142
 1st Qu.:  2.000   1st Qu.:2.000   1st Qu.:1.000   48007  :127
 Median :  4.000   Median :2.000   Median :1.000   48012  :115
 Mean   :  4.244   Mean   :2.538   Mean   :1.409   48002  :105
 3rd Qu.:  6.000   3rd Qu.:3.000   3rd Qu.:2.000   48004  :100
 Max.   : 19.000   Max.   :7.000   Max.   :4.000   48006  : 90
 NA's   :309.000   NA's   :8.000                   (Other):352
     M2cons               Estado                          CA
 Min.   : 23.00   A reformar : 98   central                :172
 1st Qu.: 65.00   Buen estado:923   central gasoil         :  1
 Median : 80.00   Obra nueva :  9   colectiva              : 12
 Mean   : 85.56   NA's       :  1   individual             :673
 3rd Qu.:100.00                     individual gas natural:  2
 Max.   :285.00                     NA's                   :171
 NA's   :  4.00
       CACombus                         AC             ACCombus   Ascensor
 gas natural:427   central                :127   gas natural:404   NO  :300
           :208   central gasoil         :  1   eléctrica  :191   SI  :727
 eléctrica  :160   colectiva              : 17              :133   NA's:  4
 gasoil     : 50   individual             :643   gasoil     : 40
 gas propano:  7   individual gas natural:  2   gas butano : 13
 (Other)    :  8   NA's                   :241   (Other)    :  9
 NA's       :171                                NA's       :241
     Garaje        FechaAnuncio             TipoVia       Calle
 Min.   :  1.00   Min.   :2009-05-08   AVENIDA  : 79   Length:1031
 1st Qu.:  1.00   1st Qu.:2009-09-10   CALLE    :895   Class :character
 Median :  1.00   Median :2009-12-24   CAMINO   : 14   Mode  :character
 Mean   :  1.05   Mean   :2010-01-13   CARRETERA: 12
 3rd Qu.:  1.00   3rd Qu.:2010-05-17   GRUPO    :  2
 Max.   :  2.00   Max.   :2010-11-30   PLAZA    : 22
 NA's   :851.00                        VIA      :  7
      Num          M2util                  Antiguedad              Fachada
 2      : 62   Min.   : 30.0   entre 10 y 20 años: 30   ladrillo        :   0
 1      : 59   1st Qu.: 60.0   entre 20 y 30 años:111   cemento/hormigón:   0
 3      : 56   Median : 72.0   entre 5 y 10 años : 55   piedra          :   0
 6      : 52   Mean   : 77.6   más de 30 años    :499   cristal         :   0
 4      : 42   3rd Qu.: 90.0   menos de 5 años   : 50   NA's            :1031
 9      : 38   Max.   :215.0   Obra nueva        :  9
 (Other):722   NA's   :381.0   NA's              :277
```

```
   Comunidad        Orientacion  Conserje        UTMX             UTMY
 Min.   : 10.00   sur     :159   SI  :112   Min.   :502121   Min.   :4787669
 1st Qu.: 30.00   suroeste: 64   NA's:919   1st Qu.:504562   1st Qu.:4789351
 Median : 40.00   este    : 46              Median :505565   Median :4789656
 Mean   : 48.21   sudeste : 45              Mean   :505508   Mean   :4789826
 3rd Qu.: 54.75   nordeste: 37              3rd Qu.:506457   3rd Qu.:4790253
 Max.   :220.00   (Other) : 85              Max.   :508554   Max.   :4792523
 NA's   :413.00   NA's    :595
```

A.2. Diamantes

```
> summary(diamonds)
```

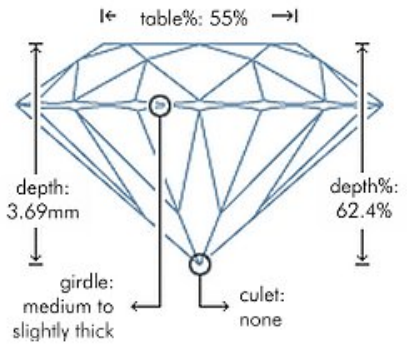
```
     carat               cut        color        clarity          depth
 Min.   :0.200   Fair      : 1610   D: 6775   SI1     :13065   Min.   :43.00
 1st Qu.:0.400   Good      : 4906   E: 9797   VS2     :12258   1st Qu.:61.00
 Median :0.700   Very Good:12082   F: 9542   SI2     : 9194   Median :61.80
 Mean   :0.798   Premium   :13791   G:11292   VS1     : 8171   Mean   :61.75
 3rd Qu.:1.040   Ideal     :21551   H: 8304   VVS2    : 5066   3rd Qu.:62.50
 Max.   :5.010                     I: 5422   VVS1    : 3655   Max.   :79.00
                                   J: 2808   (Other): 2531
     table           price             x                y
 Min.   :43.00   Min.   :  326   Min.   : 0.000   Min.   : 0.000
 1st Qu.:56.00   1st Qu.:  950   1st Qu.: 4.710   1st Qu.: 4.720
 Median :57.00   Median : 2401   Median : 5.700   Median : 5.710
 Mean   :57.46   Mean   : 3933   Mean   : 5.731   Mean   : 5.735
 3rd Qu.:59.00   3rd Qu.: 5324   3rd Qu.: 6.540   3rd Qu.: 6.540
 Max.   :95.00   Max.   :18823   Max.   :10.740   Max.   :58.900

       z
 Min.   : 0.000
 1st Qu.: 2.910
 Median : 3.530
 Mean   : 3.539
 3rd Qu.: 4.040
 Max.   :31.800
```

Los significados de las variables aparecen en el Cuadro 1. Un esquema en la Figura 1.

Cuadro 1: Variables en la dataframe `diamonds`

Variable	Descripción
`carat`	Peso de la piedra en quilates.
`cut`	Calidad de la talla (escala cualitativa)
`color`	Coloración de la piedra (escala cualitativa).
`clarity`	Claridad de la piedra (escala cualitativa).
`depth`	Profundidad (ver esquema).
`table`	"Mesa"; ver esquema.
`price`	Precio (en dólares).
`x,y,z`	Dimensiones.

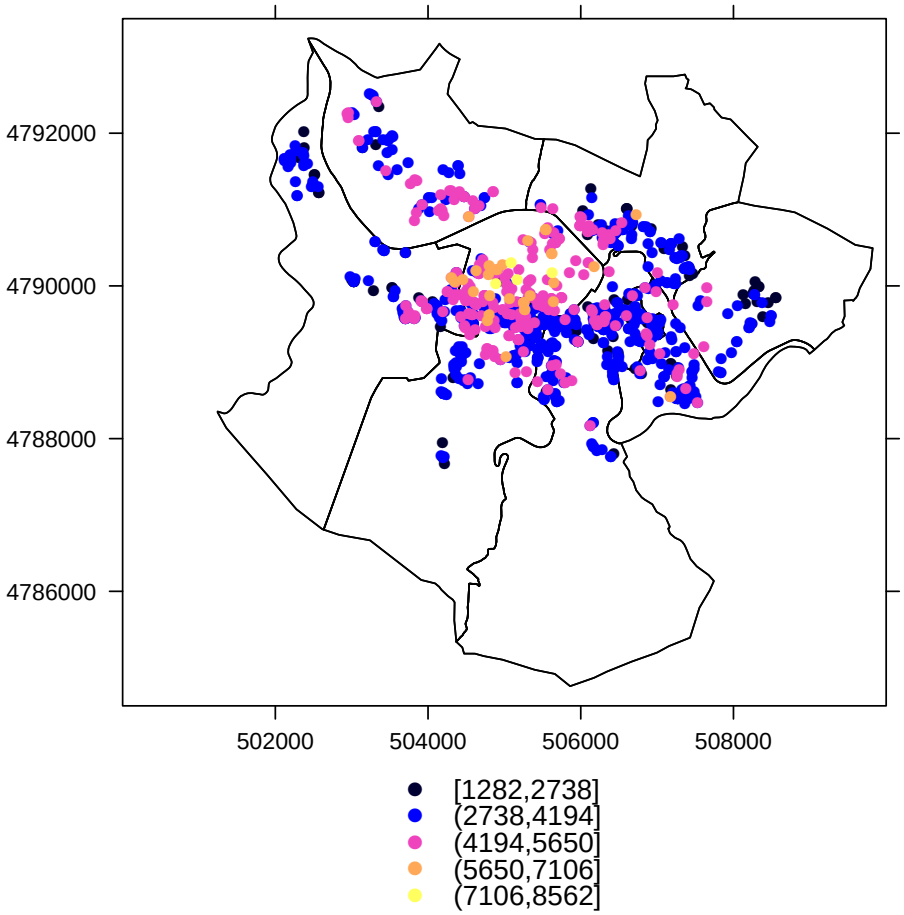
Figura 1: Esquema de dimensiones relevantes en la talla de un diamante



B. Cómo representar geográficamente datos o resultados

El siguiente código ilustra un modo simple de hacerlo; has de cargar las bibliotecas `sp` y `maptools` antes. Tienes las instrucciones que componen este fragmento de programa al final del fichero descrip.R

```
> library(sp)
> library(maptools)
> #
> #  Señalar las columnas de coordenadas
> #
> pisos <- as.data.frame(pisos)
> pisos <- cbind(pisos,PrecioM2= pisos[,"Precio"] / pisos[,"M2cons"])
> coordinates(pisos) <- ~ UTMX + UTMY
> fich <- "BilbaoDistritos.shp"            #  Señala camino completo si no está en
>                                          #         tu carpeta de trabajo.
> xx <- readShapePoly(fich)                #  Lee el fichero de polígonos.
> #
> #   Representamos la variable "Precio"; igual podríamos hacer con
> #   cualquier otra (numérica), residuos, etc.
> #
> fig <- spplot(pisos, c("PrecioM2") , col.regions=bpy.colors(10),
+        sp.layout=list("sp.polygons",xx,col="black"),
+        scales=list(draw=TRUE),
+        xlim=c(500000,510000),
+        ylim=c(4784500,4793500))
> print(fig)
```



Puedes emplear cualquier cantidad de tiempo en adaptar el mapa a tus preferencias, pero con algo relativamente básico como esto te bastará para diagnosticar residuos, casos raros, etc.

Puedes también mostrar tus datos sobre la cartografía que ofrece Google (no muy aconsejable, pero útil para algunos usos). Mira este ejemplo:

```
> library(plotGoogleMaps)
> proj4string(pisos) <- CRS("+proj=utm +zone=30 +ellps=intl +units=m +no_defs")
> mapPoints <- plotGoogleMaps(pisos, mapTypeId = "TERRAIN", filename = "mapa.htm")
```

La segunda línea establece información sobre la proyección que se está empleando (no necesita preocuparte; varía de ejemplo a ejemplo). La tercera crea en tu directorio de trabajo un fichero que puedes abrir en cualquier navegador[8] para ver un mapa "clickable" con tus observaciones. Sobre el uso de datos espaciales en R puede interesarte ver [1] y [12].

[8] En una máquina con conexión a Internet y Java instalado.

C. Facsímil de ficha de datos

piso en c. puente de deusto, 5, bilbao http://www.idealista.com/pagina/inmueble?codigoinmue...

idealista.com

idealista.com > venta, vivienda, vizcaya > gran bilbao > bilbao > pisos en deusto > **piso en c. puente de deusto, 5**

720.000 euros, 119.797.920 pts
piso de 180 m² exterior
planta 2
4 dormitorios
2 baños

c. puente de deusto, 5
distrito deusto
48014 bilbao, vizcaya

645 716 282 - tardes
645 716 243 - tardes
jaime - particular - contactar
anuncio vw292247

actualizado el 29 de octubre
anuncio visto 2.966 veces
enviado a amigos 1 vez
anunciante contactado 5 veces

4.000 euros/m², 665.544 pts/m²

hipoteca 3.437 euros
personalizar cálculo

características específicas

publicidad

180 m² construidos, 150 m² útiles
segunda mano / buen estado
calefacción central gasoil
agua caliente central gasoil
orientación sur
planta 2, edificio de 9 o más plantas con ascensor
antiguedad entre 20 y 30 años, fachada de piedra
3 vecinos por planta
90 euros al mes de gastos de comunidad

distribución y materiales

4 dormitorios
2 baños
cocina independiente equipada
8 armarios empotrados
suelos de tarima flotante

equipamiento

1 plaza de garaje incluida en el precio
trastero
la casa está dotada de antena parabólica colectiva
conserje y puerta blindada

observaciones

inmejorable situación y orientación, vistas ría, muy buen estado, 4 amplísimas habitaciones, principal con baño y vestidor, otro baño, amplio distribuidor (ideal para zona de despacho), armarios empotrados vestidos, cocina-office con despensa, garaje y trastero con acceso directo.

modificaciones a tu anuncio

pincha aquí si eres el anunciante y quieres hacer algún cambio a tu anuncio

1 de 2 15/11/06 17:35

Referencias

[1] Roger S. Bivand, Edzer J. Pebesma, and Virgilio Gómez-Rubio. *Applied Spatial Data Analysis with R.* Springer Verlag, 2008.

[2] M.J. Crawley. *The R Book.* Wiley, 2007. Signatura: 519.682 CRA.

[3] P. Dalgaard. *Introductory Statistics with R.* Statistics and Computing. Springer-Verlag, 2002. Signatura: 519.682 DAL.

[4] N. R. Draper and H. Smith. *Applied Regression Analysis.* Wiley, third edition, 1998. Signatura: 519.233.5 DRA.

[5] J. J. Faraway. *Linear Models with R.* Chapman & Hall/CRC, 2005. Signatura: 519.233 FAR.

[6] J. Fox. *An R and S-Plus Companion to Applied Regression.* Sage Pub., 2002.

[7] Gabor Grothendieck and Thomas Petzoldt. R help desk: Date and time classes in r. *R News*, 4(1):29–32, June 2004.

[8] Frank E. Harrell. *Regression Modeling Strategies (With Applications To Linear Models, Logistic Regression, And Survival Analysis).* Springer, 2006.

[9] P. Kuhnert and W. Venables. *An Introduction to R: Software for Statistical Modelling and Computing.* CSIRO Mathematical and Information Sciences, Cleveland, Australia, 2005.

[10] J. H. Maindonald. Data analysis and graphics using R - An introduction. January 2000.

[11] R. H. Myers. *Classical and Modern Regression with Applications.* PWS-KENT Pub. Co., Boston, 1990.

[12] Edzer J. Pebesma and Roger S. Bivand. Classes and methods for spatial data in r. *R News*, 5(2):9–13, November 2005.

[13] Brian D. Ripley and Kurt Hornik. Date-time classes. *R News*, 1(2):8–11, June 2001.

[14] G. A. F. Seber. *Linear Regression Analysis.* Wiley, New York, 1977.

[15] G. A. F. Seber and A. J. Lee. *Linear Regression Analysis.* Wiley, 1998.

[16] Phil Spector. *Data Manipulation with R.* Springer, 2008.

[17] A. Fdez. Trocóniz. *Modelos Lineales.* Serv. Editorial UPV/EHU, Bilbao, 1987.

[18] W.N. Venables and B.D. Ripley. *Modern Applied Statistics with S-Plus.* Springer-Verlag, New York, third edition, 1999.

[19] H. Wickham. *ggplot2 : elegant graphics for data analysis.* Springer-Verlag, 2009.