

Combinación de métodos factoriales y de análisis de conglomerados en R: el paquete FactoClass

Combination of Factorial Methods and Cluster Analysis in R: The Package FactoClass

CAMPO ELÍAS PARDO^a, PEDRO CÉSAR DEL CAMPO^b

DEPARTAMENTO DE ESTADÍSTICA, FACULTAD DE CIENCIAS, UNIVERSIDAD NACIONAL DE COLOMBIA, BOGOTÁ, COLOMBIA

Resumen

Se presenta el paquete de R `FactoClass`, donde se implementa la estrategia descrita en Lebart et al. (1995), que combina métodos factoriales con análisis de conglomerados, en la exploración multivariada de tablas de datos. Se utilizan funciones de `ade4` (Chessel et al. 2004) para realizar el análisis factorial de los datos y de `stats` para el análisis de conglomerados. Se crean funciones para tareas específicas y se modifican algunas de las existentes. Se describen los pasos para crear `FactoClass` en ambiente `Windows` y se ilustra el uso del paquete con un ejemplo.

Palabras clave: *software* estadístico, análisis multivariado, análisis en componentes principales, análisis de correspondencias, K -medias, clasificación jerárquica, L^AT_EX.

Abstract

The new R package `FactoClass` to combine factorial methods and cluster analysis is presented. This package is implemented in order to perform a multivariate exploration of a data table according to Lebart et al. (1995). We use some `ade4` functions (Chessel et al. 2004) to perform the factorial analysis of the data and some `stats` functions in R to perform cluster methods. Some new functions are programmed to make specific tasks and another old ones are modified. We describe the implementation of `FactoClass` in the `Windows` environment and illustrate its use with an example.

Key words: Statistical software, Multivariate analysis, Principal components analysis, Correspondence analysis, K-means clustering, Hierarchical clustering, L^AT_EX.

^aProfesor asociado. E-mail: cepardot@unal.edu.co

^bEstadístico. E-mail: pcdelcampon@unal.edu.co

1. Introducción

En este documento el término clasificación se utiliza como sinónimo de análisis o formación de conglomerados o clasificación no supervisada. En ningún momento hace referencia a la clasificación supervisada o discriminación, la cual no está incluida en el paquete construido.

Para el análisis de una tabla de datos haciendo uso de métodos multivariados, Lebart et al. (1995) presentan una estrategia que consiste en realizar primero un análisis factorial según la naturaleza de los datos y luego una clasificación basada en un algoritmo mixto: clasificación jerárquica con el método de Ward y agregación alrededor de centros móviles (K -medias). Finalmente se obtiene una partición del conjunto de datos y la caracterización de cada una de las clases, según las variables activas e ilustrativas, ya sean cuantitativas o cualitativas.

Para la caracterización de las clases se utilizan los *valores test*, que son índices descriptivos construidos siguiendo la metodología de pruebas de hipótesis, pero sin el objetivo de hacer inferencias. La ordenación de los valores test dentro de cada clase permite obtener las variables continuas que la caracterizan positivamente, en el sentido de que la media de la clase es suficientemente mayor de la media global, o negativamente cuando la media de la clase es inferior. Para las categorías de variables nominales, la ordenación permite obtener aquellas categorías cuya proporción dentro de la clase se diferencia lo suficiente de la proporción global, ya sea porque es mayor (valor test positivo) o menor (valor test negativo).

Para la puesta en práctica de la estrategia mencionada utilizando el lenguaje R (R Development Core Team 2007a) se programa el paquete denominado `FactoClass`, el cual utiliza funciones de `ade4` (Chessel et al. 2004) para realizar el análisis factorial de los datos y de `stats` para los métodos de clasificación. Se programan las funciones complementarias que se requieren, incluyendo algunas para producir salidas en formato \LaTeX^1 , utilizando el paquete `xtable` (Dahl 2006). `FactoClass` permite obtener salidas similares a las que aparecen en los programas estadísticos SPAD (Lebart et al. 1999) y DTM (Lebart 2007).

La creación de paquetes portables facilita la labor académica cuando se utiliza R como lenguaje para la ejecución de los métodos estadísticos. Un paquete es útil aun cuando no se requiera la programación de nuevas funciones, ya que se pueden incluir tablas de datos y líneas de comandos para los talleres de un curso. Por ejemplo el texto de Dalgaard (2002) tiene asociado el paquete `ISwR` (Dalgaard 2005), que incluye las tablas de datos y las líneas de programación para los ejemplos incluidos en el texto.

De nuestra experiencia en la creación del paquete en ambiente Windows, describimos los aspectos en los que tuvimos especial dificultad, ya que nos parece útil para quienes deseen hacer paquetes portables (ver apéndice A).

En la sección 2 se resume la estrategia de análisis combinando los métodos, luego, en la sección 3 se describe el paquete, y finalmente se ilustra su utilización con un ejemplo, en la sección 4.

¹Mi \LaTeX project page: <http://www.miktex.org/> Ver por ejemplo De Castro (2003).

2. Estrategia combinada de un método factorial y formación de conglomerados

La estrategia descrita en Lebart et al. (1995) sigue los pasos que se muestran en la figura 1. En esta sección aparecen las fórmulas que sustentan alguna modificación de una función existente o la programación de una función nueva para `FactoClass`.

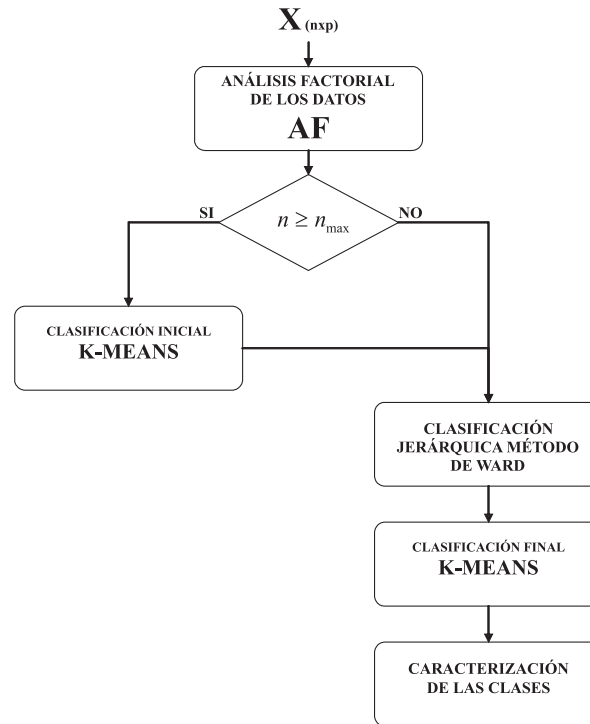


FIGURA 1: Diagrama de flujo de la estrategia combinada de análisis factorial y métodos de clasificación. Si el número n de elementos por clasificar sobrepasan el umbral n_{\max} , se realiza un agrupamiento antes de la clasificación jerárquica.

2.1. Análisis factorial

`FactoClass` utiliza el paquete estadístico `ade4` (Chessel et al. 2004) para realizar el análisis factorial de los datos, en particular las funciones:

`dudi.pca`: análisis de componentes principales (ACP),

`dudi.coa`: análisis de correspondencias simples (ACS),

`dudi.acm`: análisis de correspondencias múltiples (ACM),

`witwit.coa`: análisis de correspondencias interno (ACI) (Cazes et al. 1988).

Estas funciones retornan un objeto de tipo `dudi` con los valores y vectores propios y las coordenadas factoriales de las filas y columnas. Las demás ayudas a la interpretación se obtienen con la función `inertia.dudi`.

Las funciones `dudi` de `ade4` reciben los datos en un objeto `data.frame` y utilizan todas las columnas como activas.

El paquete `ade4` tiene varias funciones para obtener los planos factoriales; sin embargo en `FactoClass` se incluye la función `planfac` que recibe un objeto `dudi` y produce un plano factorial similar a los del paquete `FactoMineR` (Husson et al. 2007) o a los de `ade4`. Para la programación de esta función se tomaron partes de las funciones de estos dos paquetes.

2.2. Clasificación a partir de los factores

La utilización de las coordenadas factoriales permite tener un marco común en el proceso de formación de conglomerados. Para el proceso de clasificación el análisis factorial previo se constituye en un pretratamiento, que transforma los datos originales en variables continuas no correlacionadas. Tomar todos los factores para la formación de conglomerados es equivalente a efectuar una clasificación de las filas de la tabla de datos utilizando las variables originales. Tomar menos factores implica realizar un filtrado: se supone que los ejes utilizados para la clasificación tienen la *información* relevante y que los desechados se deben a las fluctuaciones aleatorias que constituyen el *ruido*. El diagrama de valores propios orienta la decisión del número de ejes que se utilizan en la clasificación. Algunas veces, sobre todo en tablas pequeñas, se usan todos los ejes.

2.3. El método de Ward

El método de Ward utiliza la distancia entre clases que cumple con el objetivo de unir, en cada paso del proceso de aglomeración, las dos clases que incrementen menos la inercia intraclases.

Sean A y B dos clases no vacías y disjuntas y sean p_A, p_B y $\mathbf{g}_A, \mathbf{g}_B$ sus pesos y centros de gravedad, respectivamente. La distancia de Ward entre los dos grupos, en función de la distancia euclidiana canónica d , viene dada por:

$$W(A, B) = \frac{p_A p_B}{p_A + p_B} d^2(\mathbf{g}_A - \mathbf{g}_B) \quad (1)$$

(la deducción se puede ver en Pardo 1992).

En particular para dos individuos i y l , con pesos p_i y p_l , la distancia de Ward es:

$$W(i, l) = \frac{p_i p_l}{p_i + p_l} d^2(i, l) \quad (2)$$

En `FactoClass` se crea la función `ward.cluster`, la cual transforma la distancia euclidiana en distancia de Ward utilizando la fórmula (2) y llama la función `hclust` del paquete básico `stats`. En `ward.cluster` se incluye una gráfica de

los índices de nivel para facilitar la decisión de cuántas clases seleccionar para la partición.

2.4. El algoritmo *K*-medias

Este algoritmo para la obtención de una partición directa de un conjunto de “individuos” por variables cuantitativas requiere el número de clases por obtener y de puntos iniciales para cada una de ellas. La propuesta de Lebart et al. (1995) es utilizarlo para obtener una partición que minimice la inercia intraclases. Esto se logra localmente (depende de los puntos iniciales) usando la distancia euclidiana canónica entre los individuos y los centros móviles utilizados para la agregación. En cada paso del algoritmo se actualizan los centros móviles calculando los centros de gravedad de la partición obtenida del paso anterior.

Para una clase k , conformada por el conjunto de individuos I_k con pesos p_i y coordenadas sobre el eje s notadas $F_s(i)$, el término general de la coordenada de su centro de gravedad sobre un eje factorial s es:

$$g_s(k) = \sum_{i \in I_k} p_i F_s(i) \quad (3)$$

y su inercia intra en el subespacio de los S primeros ejes factoriales es:

$$\text{InerciaIntra}(k) = \sum_{i \in I_k} p_i \sum_{s=1}^S (F_s(i) - g_s(k))^2 \quad (4)$$

La función `kmeans` de `stats` no maneja pesos distintos para las filas. Estos pesos influyen en los centros de gravedad (3) y en las inercias intra de las clases (4). Se modifica entonces esta función para incluir los pesos de las filas y obtener las inercias intra clases; se nombra `kmeansw`. En su opción por defecto la función `kmeans` utiliza el algoritmo de Hartigan & Wong (1979) programado en Fortran con el nombre `kmns.f`, función que se modifica dándole el nombre de `kmnsw.f`. Las líneas de programación modificadas cambian los cálculos de centros de gravedad y suma de cuadrados intra-clase por los de las fórmulas (3) y (4).

2.5. Caracterización de las clases

Para seleccionar las variables continuas o las categorías de las variables nominales más características de cada clase, se mide la desviación entre los valores relativos a la clase y los valores globales, utilizando los *valores test* (Lebart et al. 1995, pp.181-184):

Variable continua en una clase. Para una variable continua X , con media general \bar{X} , el valor test asociado a la media \bar{X}_k de la clase k es:

$$t_k(X) = \frac{\bar{X}_k - \bar{X}}{\sqrt{\frac{n - n_k}{(n - 1)n_k} S_X}} \quad (5)$$

donde S_X es la desviación estandar de la variable X en todo el conjunto de datos, n el número de “individuos” clasificados y n_k el número de “individuos” dentro de la clase k .

Categoría en una clase. En una clase k conformada por n_k individuos, de los n clasificados, n_{kj} tienen la modalidad j . El valor test para j en la clase k se obtiene con un modelo hipergeométrico: de una “urna” con n “bolas”, de las cuales n_j son “bolas negras”, se extrae una muestra de n_k “bolas” y se obtienen n_{kj} “bolas negras”. Si N es la variable aleatoria que designa la “cantidad de bolas negras en una muestra de tamaño n_k ”, el valor p asociado al supuesto de extracción aleatoria, cuando la frecuencia relativa de la categoría j en la clase es mayor que la frecuencia global, es:

$$Valorp_k(j) = Prob(N \geq n_{kj}) = \sum_{x=n_{kj}}^{\min(n_k, n_j)} h(x; n, n_j, n_k) \quad (6)$$

donde $h(x; n, n_j, n_k)$ es la distribución de probabilidad hipergeométrica de parámetros n, n_j y n_k calculada en x . Si la frecuencia relativa de la categoría j dentro de la clase k es menor que la global, entonces el valor p es:

$$Valorp_k(j) = Prob(N \leq n_{kj}) = H(n_{kj}; n, n_j, n_k) \quad (7)$$

donde $H(n_{kj}; n, n_j, n_k)$ es la función de distribución de probabilidad hipergeométrica acumulada, evaluada en n_{kj} .

El valor test se obtiene buscando el cuantil de la normal estándar correspondiente al valor p calculado mediante (6) o (7).

Para la caracterización de las clases, se incluye la función `cluster.carac`, donde se calculan los valores test con las fórmulas descritas en esta sección y se retorna un objeto `list` con información similar a la que se obtiene en los programas SPAD (Lebart et al. 1999) y DTM (Lebart 2007).

2.6. Algoritmo de clasificación mixta

La propuesta descrita en Lebart et al. (1995) aprovecha las ventajas del método de Ward y las del K -medias, combinándolos de la manera siguiente:

1. *Clasificación inicial.* Si la cantidad de individuos por clasificar es muy alta, es probable que la clasificación jerárquica no se pueda ejecutar directamente. Entonces se efectúa esta primera etapa, la cual busca obtener rápidamente y a bajo costo una partición de los individuos en s clases homogéneas, donde s es mucho mayor que el número de clases deseado en la población, y menor que la cantidad de individuos. Se emplea el algoritmo de agregación alrededor de centros móviles (K -medias).

Se utiliza la función modificada `kmeansW`, dando el número de clases, con lo cual los centros iniciales se establecen al azar.

2. *Agregación jerárquica con el método de Ward.* Se efectúa una clasificación ascendente jerárquica donde los elementos terminales del árbol son las s clases de la partición inicial o los individuos directamente. El árbol correspondiente se construye según el criterio de Ward, el cual une en cada paso de agregación las dos clases que incrementen lo menos posible la inercia intraclases.

Se hace con la nueva función `cluster.ward` luego de calcular las distancias entre filas o entre las clases previas obtenidas en 1.

3. *Corte del árbol.* El árbol o dendrograma que resume el procedimiento de clasificación permite ver la estructura de clases de los *individuos* que son objeto de análisis. En el gráfico de índices de nivel es más fácil observar los cambios de inercia más grandes (saltos) y decidir el número de clases K .

Se consigue con la función `cutree` de `stats`. Para el paso siguiente es necesario calcular los pesos y centros de gravedad de las clases obtenidas.

4. *Consolidación de la clasificación.* La partición obtenida en el paso anterior no es óptima siempre, debido a la estructura de particiones anidadas del dendrograma obtenido. Para mejorarla se utiliza de nuevo un procedimiento de agregación alrededor de centros móviles (K -medias), utilizando los centros de gravedad de las clases obtenidas al cortar el árbol como centros iniciales.

De nuevo se utiliza `kmeansW`, pero dando los centros de gravedad del paso anterior como puntos iniciales.

2.7. Identificación de las clases sobre los planos factoriales

Los centros de gravedad de las clases se pueden proyectar sobre los planos factoriales, y los individuos de cada clase se pueden diferenciar mediante signos o colores.

La función construida `plot.FactoClass` recibe un objeto de tipo `FactoClass` y produce el plano factorial solicitado.

3. Descripción del paquete `FactoClass`

La función `FactoClass` realiza la estrategia completa sin incluir el plano factorial con las clases, ni las salidas en formato \LaTeX . Sin embargo algunas funciones se pueden utilizar separadamente. Por ejemplo la función `cluster.carac` se puede usar para caracterizar la partición asociada a una variable nominal con variables continuas y nominales presentes en una tabla de datos.

Las funciones del paquete `FactoClass` se presentan en la tabla 1 y los datos de ejemplos incluidos se relacionan en la tabla 2.

No es necesario extenderse más en la descripción del paquete, ya que luego de instalado y cargado se puede obtener la ayuda para todas las funciones utilizando `?` y el nombre de la función; por ejemplo `?cluster.carac`.

TABLA 1: Funciones y su descripción del paquete `FactoClass`.

Función	Descripción
Fac.Num	División de variables cualitativas y cuantitativas.
list.to.data	<code>list</code> a <code>data.frame</code> .
planfac	Gráfico de planos factoriales.
dudi.tex	Tabla de coordenadas y ayudas para la interpretación de los ejes principales de un objeto <code>dudi</code> en formato \LaTeX utilizando <code>xtable</code> (Dahl 2006).
ward.cluster	Tabla de coordenadas, ayudas de interpretación de los ejes principales. Clasificación jerárquica por método de Ward (§2.3).
kmeansW	Realiza K -medias teniendo en cuenta las ponderaciones de los individuos (§2.4).
cluster.carac	Caracterización de las clases según variables (§2.5).
FactoClass	Combinación de métodos factoriales y de clasificación no supervisada (§2).
FactoClass.tex	Tabla de coordenadas, ayudas para la interpretación de los ejes principales y métodos de clasificación en \LaTeX .
plot.FactoClass	Gráfico de planos factoriales con clasificación para objetos <code>FactoClass</code> .

TABLA 2: Tablas de `FactoClass`.

Tabla	Descripción
Bogota	Localidades por estrato en la ciudad de Bogotá
colores	Asociaciones entre colores y adjetivos
perros	Razas de perros
Vietnam	Opinión de los estudiantes estadounidenses sobre la guerra de Vietnam

4. Ejemplo: descripción de las manzanas de Bogotá según estratos socioeconómicos

Objetivo. Comparar las 19 localidades de Bogotá según la distribución de sus manzanas en los seis estratos socioeconómicos. En Colombia se estratifican las manzanas de las ciudades en seis estratos, siendo el primero el de menor nivel socioeconómico.

Los datos. La tabla `Bogotá` es una tabla de contingencia que clasifica las manzanas según localidades y estratos en el año de 1997 (DAPD 1997). Se incluye una columna `SinEstrato` para contabilizar las manzanas no residenciales.

Procedimiento de análisis. Se realiza un análisis de correspondencias simples (ACS) tomando como columnas activas los estratos 1 a 6 y como columna ilustrativa `SinEstrato`.

Se sigue la estrategia descrita en la sección 2 para obtener una clasificación de las 19 localidades mediante su agrupación por su parecido en el perfil de distribución de manzanas según estratos.

Número de ejes utilizados para la clasificación. En tablas pequeñas como esta no es necesario hacer un filtrado, y por tanto se utilizan los cinco ejes factoriales del ACS para la clasificación. En este caso el ACS sirve como procedimiento de cuantificación de las categorías, al remplazar las frecuencias por coordenadas factoriales.

Cálculos en R utilizando el paquete FactoClass

1. *División de las columnas en activas e ilustrativas.* Los procedimientos de `ade4` requieren que los elementos activos e ilustrativos estén en tablas separadas:

```
library(FactoClass)
data(Bogota)
Bog.act <- Bogota[-1]
Bog.ilu <- Bogota[ 1]
```

2. *FactoClass en forma interactiva*

```
FC.bogota<-FactoClass(Bog.act, dudi.coa,Bog.ilu)
```

En la pantalla aparece el diagrama de valores propios, que permite escoger el número de ejes para el análisis. En este ejemplo pequeño son suficientes dos ejes que retienen 67.5% de la inercia (este valor se puede verificar una vez se complete la función `FactoClass` con: `inertia.dudi(FC.bogota$dudi)`). En la consola se tecldea `2` y *Enter*; luego se introduce el número de ejes que se utilizan en el proceso de clasificación (cinco ejes) y finalmente el número de clases deseados para la partición, decisión que se toma observando el diagrama de índices de nivel y el dendrograma (figura 2). El procedimiento termina y el objeto `FC.bogota` contiene todos los resultados: análisis factorial, lista de índices para crear el dendrograma, coordenadas de los centros de las clases, descripción de la consolidación de las clases, descripción de las clases según las variables cualitativas, cuantitativas y de frecuencia.

3. Tablas de `FC.bogota` en formato \LaTeX :

```
FactoClass.tex( FC.bogota, job="Bogota", append=TRUE,dir=getwd(),
               p.clust=FALSE )
```

4. Clases sobre el primer plano factorial factorial (figura 3):

```
plot(FC.bogota,col.row=c("maroon2","orchid4","darkgoldenrod2",
                        "dark red","aquamarine4"),cframe=1)
```

5. Columna ilustrativa *SinEstrato* (SE) sobre el primer plano factorial (figura 3):

```
SE.coor <- supcol(FC.bogota$dudi,Bog.ilu)$cosup
points(SE.coor,pch=23)
text(SE.coor,"SE",pos=3)
```

La función `supcol` es de `ade4`; `points` y `text` son funciones gráficas de R.

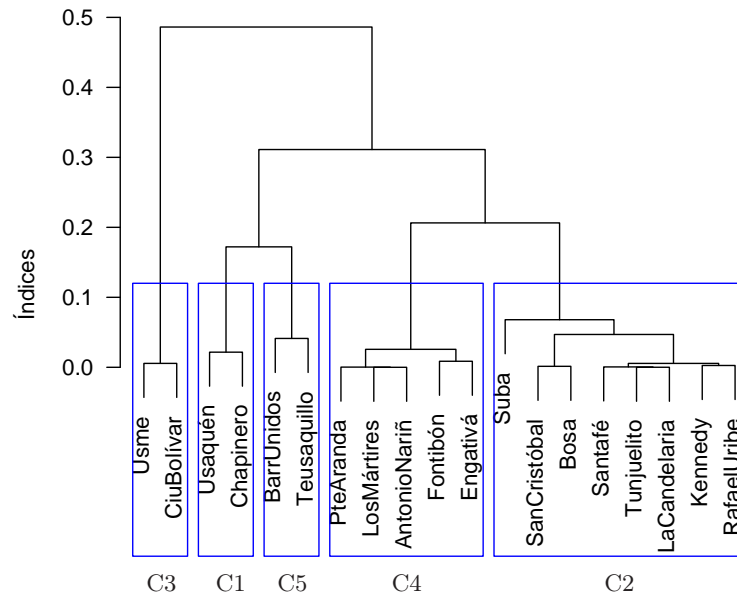


FIGURA 2: Dendrograma.

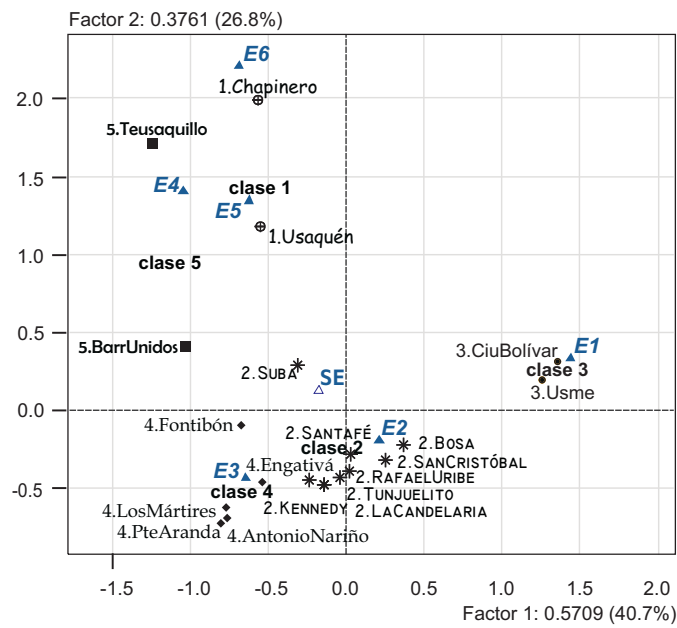


FIGURA 3: Primer plano factorial del ACS de la tabla de contingencia Bogotá. Se destacan las clases mostrando sus centros de gravedad y la clase a la que pertenece cada localidad. Se incluye SE como columna ilustrativa.

Análisis de resultados. El lector podrá verificar los resultados escribiendo en la consola de R el comando que se muestra dentro del paréntesis en cada caso.

El primer plano factorial de este análisis se muestra en la figura 3. Se observa el ordenamiento de las localidades en el plano según su perfil de estratos: al lado derecho las localidades *Usme* y *Ciudad Bolívar*, asociadas al *estrato 1*; al centro se encuentran las localidades con mayor porcentaje de *estrato 2* que el promedio; a la izquierda, las asociadas al *estrato 3*, y arriba las que incrementan el porcentaje de *estratos 4, 5 y 6* con respecto al promedio. La posición del perfil de manzanas *SinEstrato*, cerca al origen del primer plano factorial, permite concluir que su distribución según localidades es similar a la distribución global de las manzanas estratificadas.

Las localidades *Usme* y *Ciudad Bolívar* contribuyen con el 60.2% de la inercia del primer eje, mientras que *Chapinero*, *Usaquén* y *Teusaquillo* son las que más contribuyen en el segundo eje (en conjunto 63.7% de la inercia del eje) (`inertia.dudi(FC.bogota$dudi,T)$row.abs`). Las 19 localidades están bien representadas en el plano factorial (`inertia.dudi(FC.bogota$dudi,T)$row.cum`).

El proceso de consolidación no introduce cambios a la partición obtenida al cortar el árbol en cinco clases (figura 2). La clase 2 es la de más localidades y tiene casi la mitad de las manzanas. La clase cinco es la de mayor inercia intra y la más alejada del origen (`FC.bogota$clus.summ`).

TABLA 3: Caracterización de las clases según estratos.

Clase	Est.	V.test	Porcentaje			Manzanas por estrato
			clase/estrato	mod/clas	Global	
C1	E6	Inf	85.7	23.5	2.2	785
	E5	31.5	44.7	15.1	2.4	969
	E4	21.6	22.0	17.5	6.4	2276
	SE	5.9	10.8	13.3	10.2	4088
C2	E2	Inf	73.3	62.4	39.9	14282
	E5	2.3	49.8	2.6	2.4	969
	SE	-7.6	40.5	9.0	10.2	4088
C3	E1	Inf	79.5	67.6	15.9	5711
	SE	-6.8	14.5	8.1	10.2	4088
C4	E3	Inf	50.3	78.4	32.9	11786
	SE	10.0	28.1	13.2	10.2	4088
C5	E4	Inf	44.5	53.3	6.4	2276
	E3	10.1	7.1	43.7	32.9	11786

La tabla 3 de caracterización de las clases es una parte de `FC.bogota$carac.frec`. La clase asignada a cada localidad se obtiene con `FC.bogota$cluster`. Con las siguientes instrucciones se logra un `mosaicplot` (Friendly 1994) para una caracterización gráfica de clases (se requiere la función `gsummary` del paquete `nlme` (Pinheiro et al. 2007)):

```
library(nlme)
gsummary(Bog.act,groups=FC.bogota$cluster,FUN=sum)->clasEst
mosaicplot(clasEst,color=TRUE,main="")
```

En la caracterización de las clases (tabla 3) se incluye la categoría *SinEstrato* (SE), que corresponde a la columna ilustrativa. Se observa un incremento de manzanas sin estratificar en las clases C1 (seguramente por mayor cantidad de zonas verdes) y C4 (zonas industriales), y una disminución en las zonas más deprimidas (C2 y C3).

Conclusiones del ejemplo. El primer plano factorial del ACS de la tabla de contingencia que clasifica las manzanas de Bogotá según estratos y localidades (figura 3), junto con la partición en cinco clases (figura 2), permite corroborar que Bogotá es una ciudad fuertemente estratificada geográficamente. Las localidades más pobres (clase 3: Usme y Ciudad Bolívar) son las de estratos 1 y 2, prácticamente sin estrato 3 y sin manzanas en estratos superiores, se ubican en la periferia sur de la ciudad. En contraste las clases 1 (Chapinero y Usaquén) y 5 (Teusaquillo y Barrios Unidos), que contienen en conjunto el 11.2% de las manzanas de Bogotá, son las que más estratos 4, 5 y 6 tienen y se ubican en el centro-norte. Las clases 2 y 4, en el centro del plano factorial, son las mayoritarias y tienen sobre todo manzanas en estrato 2 y 3.

Obtención del paquete FactoClass

El paquete *FactoClass* se instala en R (versión 2.4.1 en adelante) a partir del *zip* disponible en la página:

<http://www.docentes.unal.edu.co/cepardot/docs/>

Agradecimientos

Agradecemos al profesor Jorge Ortiz por su labor como editor de este artículo y a los dos árbitros anónimos, quienes hicieron valiosos comentarios que ayudaron a mejorarlo. También a los estudiantes del curso de Métodos multivariados de datos de los dos semestres de 2007, de la carrera de Estadística, que utilizaron el paquete y detectaron algunos errores.

[Recibido: agosto de 2007 — Aceptado: noviembre de 2007]

Referencias

- Cazes, P., Chessel, D. & Doledec, S. (1988), 'L'analyse des correspondances internes d'un tableau partitionné. Son usage en hydrobiologie', *Revue de Statistique Appliquée* **36**(1), 39–54.
- Chessel, D., Dufour, A. B. & Thioulouse, J. (2004), 'The ade4 Package - I: One-table Methods', *R News* **4**(1), 5–10.

- Dahl, D. B. (2006), *xtable: Export Tables to LaTeX or HTML*. David B. Dahl with contributions from many others. R package version 1.4-2.
- Dalgaard, P. (2002), *Introductory Statistics with R*, Springer, New York.
- Dalgaard, P. (2005), *ISwR: Introductory Statistics with R*. R package version 1.0-6.
- DAPD (1997), *Población, estratificación y aspectos socioeconómicos de Santa Fe de Bogotá*, Departamento Administrativo de Planeación Distrital, Bogotá.
- De Castro, R. (2003), *El universo L^AT_EX*, segunda edn, Unibiblos, Universidad Nacional de Colombia, Bogotá.
- Friendly, M. (1994), 'Mosaic Displays for Multi-Way Contingency Tables', *Journal of the American Statistical Association* **89**(425), 190–200.
- Hartigan, J. A. & Wong, M. A. (1979), 'A K-means Clustering Algorithm', *Applied Statistics* **28**(100–108).
- Husson, F., Lê, S. & Mazet, J. (2007), *FactoMineR: Factor Analysis and Data Mining with R*. R package version 1.05.
*<http://factominer.free.fr>, <http://www.agrocampus-rennes.fr/math/>
- Lebart, L. (2007), 'DTM. Data and Text Mining', Software.
*<http://ses.enst.fr/lebart/>
- Lebart, L., Morineau, A., Lambert, T. & Pleuvret, P. (1999), *SPAD. Système Pour l'Analyse des Données*, Paris.
*<http://www.spad.eu>
- Lebart, L., Morineau, A. & Piron, M. (1995), *Statistique exploratoire multidimensionnelle*, Dunod, Paris.
- Ligges, U. (2006), 'R Help Desk: Accessing the Sources', *R News* **6**(4), 43–45.
- Ligges, U. & Murdoch, D. (2005), 'R Help Desk: Make 'R CMD' Work under Windows – an Example', *R News* **5**(2), 27–28.
- Pardo, C. E. (1992), Análisis de la aplicación del método de Ward de clasificación jerárquica en el caso de variables cualitativas, Tesis de Maestría, Estadística, Universidad Nacional de Colombia, Facultad de Ciencias, Departamento de Matemáticas y Estadística, Bogotá.
- Pinheiro, J., Bates, D., DebRoy, S. & the R Core team., D. S. (2007), *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-83.
- R Development Core Team (2007a), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
*<http://www.R-project.org>
- R Development Core Team (2007b), *Writing R Extensions*, R Foundation for Statistical Computing, Vienna, Austria.

Apéndice A.

Construcción de FactoClass en ambiente Windows

La versión portable del paquete se construye en ambiente Windows, siguiendo los pasos recomendados por Ligges & Murdoch (2005). La única variación que se tiene es la ubicación de MikTeX, ya que no se modificó la instalación por defecto que se tenía. La estructura de carpetas luego de terminados los pasos recomendados se muestra en la figura 4. Para obtener los archivos fuente de las funciones, específicamente el código Fortran `kmns.f`, se utiliza la guía dada por Ligges (2006). La ruta para acceder a los programas desde los comandos, asociada a la configuración de carpetas mostrada en la figura 4, es (archivo `rpath.bat`):

```
PATH=.;c:\devel\tools\bin;c:\devel\gnwin32\bin;c:\devel\MinGW\bin;
      c:\devel\R-2.5.1\bin;c:\devel\HtmlHelp;c:\devel\Perl\bin;c:\devel\Tlc\bin;
      c:\texmf\miktex\bin;%SystemRoot%\system32;%SystemRoot%
SET TMPDIR=C:\temp
```

La carpeta `FactoClass` corresponde a la librería o paquete que se quiere hacer portable; el contenido de las subcarpetas es:

R: funciones de R que se crean o modifican,

man: documentación de cada una de las funciones,

scr: subrutina `kmnsw.f` en Fortran 77, que es una modificación de `kmns.f` para introducir los pesos de los “individuos” (§2.4),

data: tablas de datos de los ejemplos,

doc: (subcarpeta de `inst`) este documento en PDF para hacerlo disponible en la subcarpeta `doc` del paquete luego de instalado.

El archivo `DESCRIPTION` está en la carpeta `FactoClass`. La descripción de las carpetas y los archivos necesarios y opcionales para la creación de un paquete se encuentra en el manual para escribir extensiones (R Development Core Team 2007b), disponible directamente en la ayuda de R.

Luego de entrar al *Símbolo del sistema*, se escribe el comando `cd \devel` y aparece: `C:\devel>`; luego se escribe `C:\devel>rpath`.

Para instalar el paquete en R y producir el `*.zip` portable para Windows usamos el comando:

```
C:\devel>Rcmd install FactoClass --build
```

y para revisar el paquete:

```
C:\devel>Rcmd check FactoClass
```

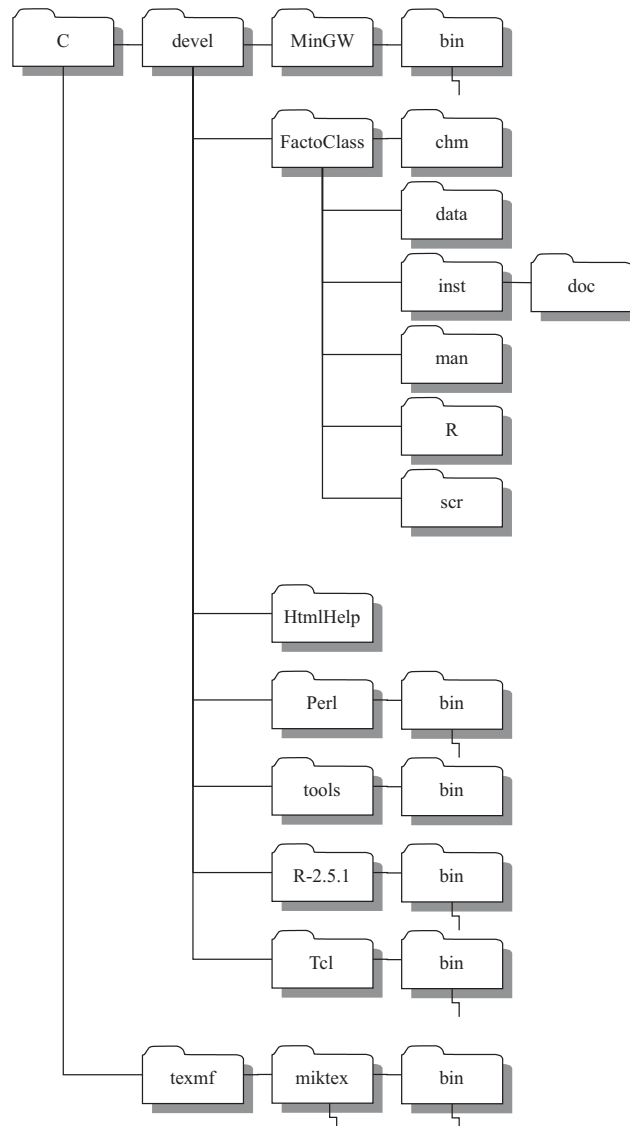


FIGURA 4: Carpetas y subcarpetas de los programas necesarios para la creación de paquetes en ambiente Windows.

Los archivos de la subcarpeta `chm` se crean en el proceso de `install`, al igual que el archivo `FactoClass.dll` en la subcarpeta `scr`, que es el ejecutable de la subrutina `kmnsw.f`. Para hacer disponible la subrutina al cargar el paquete con `library(FactoClass)` es necesario crear en la subcarpeta `R` de `FactoClass` el archivo `zzz.R` con, por lo menos, el contenido siguiente:

```
.First.lib <-function(lib, pkg) {
  library.dynam("FactoClass", "FactoClass")
}
```