Universidad Nacional de Colombia, Sede Bogotá
Facultad de Ciencias, Carrera de Estadística
Trabajos de grado - 50 años
Bogotá, Diciembre de 2007

Fundamentos y aplicaciones del análisis de correspondencias difuso (ACD)

Nurys Edilia Gámez[a]
Autor

Campo Elías Pardo[b]
Director

Departamento de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, Bogotá, Colombia

Resumen

El análisis de correspondencias difuso es un método bastante útil, pero poco conocido actualmente en nuestro medio. En este trabajo se hace una revisión en la literatura, se resume el método y se buscan algunos contextos de aplicación del mismo. Finalmente, se presenta un ejemplo para analizar los perfiles socioeconómicos de un grupo de estudiantes, comparando el análisis de correspondencias múltiples con el análisis de correspondencias difuso. De esta forma, se muestra la utilidad del ACD cuando se tiene información resumida y se resaltan las principales diferencias y similitudes entre los dos métodos con respecto al conjunto de datos trabajado.

***Palabras clave*:** Datos difusos, análisis de correspondencias, análisis multivariado, tablas de contingencia.

Abstract

The fuzzy correspondence analysis is a method very useful, but actually is not knowing in our environment. In this paper a revision is done in the literature, the method is summarize and some application contexts are search. Finally, an example for the socio-economical profile of a group of students is shown, comparing the multiple correspondence analysis with the fuzzy correspondence analysis. In this way, the utility of the ACD is shown, when has the summarize information and the principal differences and similarities between the two methods in respect to the worked data group.

***Key words*:** Fuzzy data, correspondence analysis, multivariate analysis,contingency tables

1. Introducción

En la actualidad, las técnicas de análisis de datos tales como el análisis de correspondencias simples (ACS) y el análisis de correspondencias múltiples (ACM) son bastante utilizadas. Dichas técnicas presentan salidas gráficas fáciles de interpretar y revelan relaciones no lineales entre las modalidades de diferentes variables que otros métodos no permiten observar. Sin embargo, en diferentes campos de aplicación, en especial en la ecología, es frecuente encontrar datos que son difusos por su naturaleza, es decir, se encuentra información sobre especies (o individuos) que asumen varias categorías de una misma variable con diferente grado de asociación (Dolédec & Chevenet 1994). De manera similar, las tablas de

[a]Estudiante de estadística. E-mail: negamezl@unal.edu.co
[b]Profesor asociado. E-mail: cepardot@unal.edu.co

datos que son el resultado de la yuxtaposición de varias tablas de contingencia, también se pueden ver como tablas de código difuso.

El análisis de correspondencias difuso (ACD) propuesto por Chevenet et al. (1994) es una extensión del ACM y se utiliza principalmente para analizar las posibles tendencias y relaciones entre los rasgos de especies (o subpoblaciones).

En este trabajo se hace una revisión en la literatura, se resume el método y se buscan algunos contextos de aplicación, con el fin de resaltar las principales características y su utilidad. Así mismo se ilustra el uso del método con la aplicación a una tabla de código difuso, obtenida de los datos de la encuesta de caracterización de los estudiantes que se aplica a los admitidos a la Universidad Nacional en el primer semestre de 2006. Finalmente, se comparan los resultados del ACD con el ACM de la tabla de individuos por variables.

2. Notación

Sea **A** una tabla de código difuso, con n filas (individuos, especies o subpoblaciones) y m columnas que corresponden a las categorías de las p variables, cada una con m_j categorías, en donde un individuo i puede asumir varias categorías de la variable j con diferente grado de asociación. La forma general de una tabla de datos con codificación difusa se presenta en la figura 1.

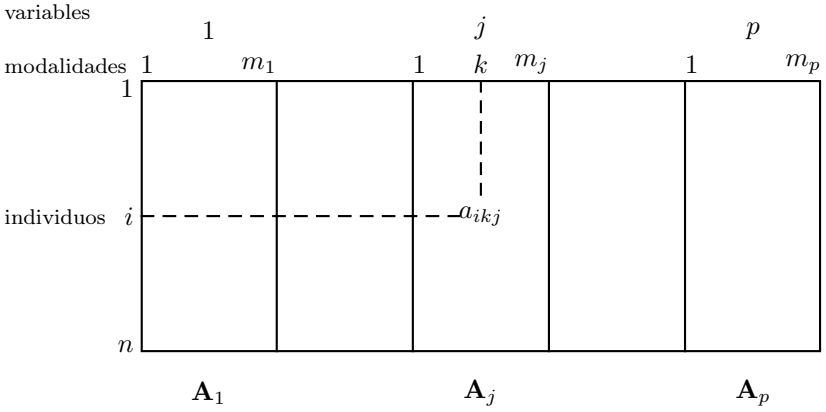


FIGURA 1: Tabla de código difuso

De esta forma, cada entrada de una tabla de codificación difusa, a_{ikj}, registra un puntaje positivo que describe la afinidad, o grado de asociación, del i-ésimo individuo con respecto a la k-ésima categoría de la j-ésima variable ($i = 1, \ldots, n; j = 1, \ldots, p; k = 1, \ldots, m_j$), tomando valores de 0: sin afinidad con la categoría hasta X: alta afinidad con la categoría (Chevenet et al. 1994).

Se tiene que:

$$m = \sum_{j=1}^{p} m_j \tag{1}$$

Partiendo de la tabla **A**, se define una nueva tabla de porcentajes o proporciones por variable (**P**) de término general:

$$p_{ikj} = \frac{a_{ikj}}{a_{i \cdot j}} \tag{2}$$

donde $a_{i \cdot j} = \sum_{k=1}^{m_j} a_{ikj}$.

En la tabla **P** se tienen las siguientes propiedades:

$$p_{i\cdot j} = \sum_{k=1}^{m_j} p_{ikj} = 1 \tag{3}$$

$$p_{i\cdot\cdot} = \sum_{j=1}^{p}\sum_{k=1}^{m_j} p_{ikj} = p \tag{4}$$

$$p_{\cdot\cdot\cdot} = \sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{k=1}^{m_j} p_{ikj} = np \tag{5}$$

$$\sum_{k=1}^{m_j} \overline{p}_{kj} = 1 \tag{6}$$

donde $\overline{p}_{kj} = \dfrac{1}{n}\sum_{i=1}^{n} p_{ikj}$ es el promedio de las proporciones que asumen los n individuos en la *k-ésima* categoría de la *j-ésima* variable.

3. Análisis de correspondencias difuso

Como se mencionó en secciones anteriores, el ACD es una extensión del ACM. A su vez, definiendo matrices de métrica y de pesos en los espacios de individuos y variables, el ACM se puede ver como un análisis en componentes principales ponderado (ACP(**X**,**M**,**D**)). Como consecuencia, el ACD también se puede ver como un ACP(**X**,**M**,**D**) y sus fórmulas se pueden deducir de éste método de análisis, como se muestra en esta sección.

3.1. El análisis en componentes principales ponderado

Los métodos en ejes principales tienen una presentación común bajo el nombre de diagrama de dualidad o análisis en componentes principales ponderado (Tenenhaus & Young 1985).

La notación $ACP(\mathbf{X}, \mathbf{M}, \mathbf{D})$ se utiliza para indicar que la tabla de datos **X** (generalmente centrada) será analizada mediante análisis en componentes principales (ACP) utilizando como pesos y métricas las matrices **M** y **D**. De esta forma, **M** es la matriz de métrica en el espacio de las filas y de pesos en el espacio de las columnas y **D**, la matriz de métrica en el espacio de las columnas y de pesos en el espacio de las filas (Escofier & Pagès 1992, Pagès 2004).

3.2. El ACM visto como un ACP ponderado

Haciendo las modificaciones convenientes, el ACM se puede ver como un análisis en componentes principales de la *tabla disyuntiva completa* (**Z**). En esta tabla con n filas y m columnas que corresponden a p variables, cada una com m_j categorías, cada entrada z_{ikj} indica si el i-ésimo individuo asume o no la k-ésima categoría de la j-ésima variable, mediante un código binario. Así, la tabla **Z** es la concatenación de p subtablas (Lebart et al. 1995):

$$\mathbf{Z} = [\mathbf{Z}_1, \ldots, \mathbf{Z}_j, \ldots, \mathbf{Z}_p]$$

Se tiene que el número total de categorías de las p variables es $m = \sum_{j=1}^{p} m_j$, igual a lo que se tiene en la tabla **A** (fórmula (1)).

Los marginales de fila de la tabla disyuntiva completa (notados como $z_{i\cdot\cdot}$) son constantes e iguales al número de variables, mientras que los marginales columna ($z_{\cdot kj}$) representan el número de individuos que asumen la categoría k de la variable j, como se muestra a continuación:

$$z_{i\cdot\cdot} = \sum_{jk} z_{ikj} = p$$

$$z_{\cdot kj} = \sum_{i=1}^{n} z_{ikj}$$

La frecuencia total, z, de la tabla disyuntiva completa es la suma de los márgenes:

$$z = \sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{k=1}^{m_j} z_{ikj} = np$$

Encontrando el mismo resultado que en la tabla $\mathbf{P}$ (ver fórmula (5)).

Se definen entonces las matrices de pesos, métricas y datos que ingresan al análisis de la siguiente forma:

- $\mathbf{X}$ es la matriz de término general: $x_{ij} = \dfrac{nz_{ikj}}{z_{\cdot kj}} - 1$
- $\mathbf{M} = \dfrac{1}{nm} diag(z_{\cdot kj})$
- $\mathbf{D} = \dfrac{1}{n}\mathbf{I}_n$

Con $\mathbf{I}_n$ la matriz identidad de tamaño $n \times n$.

3.2.1. Distancias

En el ACM todos los individuos están afectados por el mismo peso: $m_i = \dfrac{1}{n}$ y la ponderación de cada categoría es proporcional a su frecuencia: $m_k = \dfrac{z_{\cdot kj}}{np}$.

De esta forma, la distancia entre dos categorías es:

$$d^2(kj, k'j') = \sum_{i=1}^{n} n\left(\frac{z_{ikj}}{z_{\cdot kj}} - \frac{z_{ik'j'}}{z_{\cdot k'j'}}\right)^2 \tag{7}$$

Encontrando que dos categorías seleccionadas más o menos por los mismos individuos son parecidas.

Y la distancia entre dos individuos i e i'es:

$$d^2(i, i') = \frac{1}{p}\sum_{jk} \frac{n}{z_{\cdot kj}} \left(z_{ij} - z_{i'kj}\right)^2 \tag{8}$$

Es decir, que dos individuos están cerca si han seleccionado más o menos las mismas categorías de las variables.

3.2.2. Factores y relaciones cuasi-baricéntricas

La notación utilizada para los valores propios y las coordenada factoriales es:

- s: número del eje,
- λ_s: valor propio s,

- ψ_{si}: coordenada del individuo i sobre el eje s,
- φ_{skj}: coordenada de la categoría k de la variable j sobre el eje s.

La proyección de un individuo i sobre el eje factorial s en función de las coordenadas de las categorías es:

$$\psi_{si} = \frac{1}{\sqrt{\lambda_s}} \sum_{jk} \frac{z_{ikj}}{z_{i\cdot\cdot}} \varphi_{skj} = \frac{1}{p\sqrt{\lambda_s}} \sum_{(j,k)\epsilon\Im_{(j,k)}} \varphi_{skj} \tag{9}$$

donde $\Im_{(j,k)}$ es el conjunto de individuos que asumen la categoría k de la variable j.

Es decir que, omitiendo el cociente $\frac{1}{\sqrt{\lambda_s}}$, el individuo se encuentra en el punto medio de las coordenadas de las categorías que asume.

Y, de manera similar, se observa que la categoría k de la variable j se encuentra en el punto medio de los individuos que la asumen, dilatado por $\frac{1}{\sqrt{\lambda_s}}$:

$$\varphi_{skj} = \frac{1}{\sqrt{\lambda_s}} \sum_{i=1}^{n} \frac{z_{ikj}}{z_{\cdot kj}} \psi_{si} \tag{10}$$

3.2.3. Razón de correlación

En el ACM la razón de correlación entre un factor s y una variable j se define como la suma de las contribuciones de las categorías sobre ese factor (Escofier & Pagès 1992, pp. 61–62):

$$\eta_j^2 = p \sum_{k=1}^{m_j} \quad \text{Inercia de las categorías de la variable } j \text{ proyectadas sobre el eje } s$$

Este es un indicador de qué tanto contribuye la variable j a la formación del eje s y es equivalente al cociente entre la inercia inter y la inercia total:

$$\eta_j^2 = \frac{p \sum\limits_{k\epsilon J(k)} \frac{n_{kj}}{np} \lambda_s \varphi_{skj}^2}{\frac{1}{n} \sum\limits_{i=1}^{n} \psi_{si}^2} = \frac{\sum\limits_{k\epsilon J(k)} \frac{n_{kj}}{n} \lambda_s \varphi_{skj}^2}{\lambda_s} = \sum_{k\epsilon J(k)} \frac{n_{kj}}{n} \varphi_{skj}^2 \tag{11}$$

donde $J(k)$ es el conjunto de categorías que pertenecen a la variable j y n_{kj} es el número de individuos que asumen la modalidad k de la variable j. De esta forma, la razón de correlación entre un factor s y una variable j es la suma ponderada de las coordenadas de las categorías sobre el eje factorials.

Por otro lado, se tiene que la contribución absoluta de una categoría al eje s (Ca_{skj}) es:

$$Ca_{skj} = \frac{\frac{n_{kj}}{np} \varphi_{skj}^2}{\lambda_s}$$

De forma tal que $\lambda_s Ca_{skj} = \frac{n_{kj}}{np} \varphi_{skj}^2$. Y al sumar sobre todas las categorías de la variable j se obtiene que

$$\sum_{k\epsilon J(k)} \lambda_s Ca_{skj} = \frac{\eta_j^2}{p}$$

Es decir, que la razón de correlación está asociada a la contribución de la variable j al eje s, mediante la fórmula:

$$\eta_j^2 = p \lambda_s Ca_{skj} \tag{12}$$

3.3. ACD visto como una extensión del ACM

Al ser el ACD una extensión del ACM, y de acuerdo a lo planteado anteriormente, la tabla **P** se puede analizar mediante ACP(**X**,**M**,**D**) donde **M** es una matriz cuyos elementos en la diagonal son los pesos de las columnas y los valores fuera de la diagonal son cero y **D** es una matriz diagonal cuyos elementos son los pesos de las filas. Formalmente, **X**, **M** y **D** se definen como:

- **X** es la matriz de término general: $x_{ij} = \dfrac{p_{ikj}}{\overline{p}_{kj}} - 1$ $\quad$ ($\overline{p}_{kj}$ está definido en (6))
- $\mathbf{M} = Diag\left(\dfrac{1}{p}\overline{p}_{kj}\right) = \dfrac{1}{p}diag(\overline{p}_{kj})$
- $\mathbf{D} = Diag\left(\dfrac{1}{n}\right) = \dfrac{1}{n}\mathbf{I}_n$

3.4. Propiedades del ACD

Las siguientes propiedades del ACD se deducen utilizando las definiciones y conceptos propios del análisis en componentes principales ponderado. (Escofier & Pagès (1992)).

3.4.1. Distancias entre individuos y categorías

La distancia entre dos individuos i y l es:

$$d^2(i,l) = \sum_{j=1}^{p}\sum_{k=1}^{m_j} \frac{np}{\sum_{i=1}^{n} p_{ikj}} \left(\frac{p_{ikj}}{p} - \frac{p_{lkj}}{p}\right)^2 = \frac{1}{p}\sum_{j=1}^{p}\sum_{k=1}^{m(j)} \frac{1}{\overline{p}_{kj}}(p_{ikj} - p_{lkj})^2 \tag{13}$$

De lo anterior se puede concluir que dos individuos son similares si asumen más o menos la misma proporción en cada una de las categorías k de la variable j. De la misma forma que en el ACM, ésta distancia favorece a categorías "raras".

Y la distancia entre dos categorías k y c de las variables j y q, donde $k\epsilon j$ y $c\epsilon q$ es:

$$d^2(kj,cq) = \sum_{i=1}^{n} n\left(\frac{p_{ikj}}{\sum_{i=1}^{n} p_{ikj}} - \frac{p_{icq}}{\sum_{i=1}^{n} p_{icq}}\right)^2 = \sum_{i=1}^{n} n\left(\frac{p_{ikj}}{n\overline{p}_{kj}} - \frac{p_{icq}}{n\overline{p}_{cq}}\right)^2 = \sum_{i=1}^{n} \frac{1}{n}\left(\frac{p_{ikj}}{\overline{p}_{kj}} - \frac{p_{icq}}{\overline{p}_{cq}}\right)^2 \tag{14}$$

De forma tal que dos categorías k y c son cercanas, si los grupos de individuos que las han seleccionado tienen más o menos los mismos perfiles.

3.4.2. Centros de gravedad e inercia total

Al igual que en el ACM, las nubes de puntos fila y de categorías están centradas, como se muestra a continuación:

$$G_{kj} = \sum_{i=1}^{n} \frac{1}{n}\left(\frac{p_{ikj}}{\overline{p}_{kj}} - 1\right) = \frac{1}{n}\left(\sum_{i=1}^{n} \frac{p_{ikj}}{\overline{p}_{kj}} - n\right) = \frac{1}{n}\left(\frac{n\overline{p}_{kj}}{\overline{p}_{kj}} - n\right) = \frac{1}{n}(n-n) = 0$$

$$G_i = \sum_{j=1}^{p}\sum_{k=1}^{m_j} \overline{p}_{kj}\left(\frac{p_{ikj}}{\overline{p}_{kj}} - 1\right) = \sum_{j=1}^{p}\sum_{k=1}^{m_j}\left(p_{ikj} - \overline{p}_{kj}\right) = \sum_{j=1}^{p} 1 - \frac{1}{n}\sum_{j=1}^{p}\sum_{k=1}^{m_j}\sum_{i=1}^{n} p_{ikj} = p - \frac{np}{n} = 0$$

Y la inercia total es:

$$I = \frac{1}{np}\sum_{j=1}^{p}\sum_{k=1}^{m_j}\overline{p}_{kj}\sum_{i=1}^{n}\left(\frac{p_{ikj}}{\overline{p}_{kj}}-1\right)^2 = \frac{1}{p}\sum_{j=1}^{p}\sum_{k=1}^{m_j}\sum_{i=1}^{n}\frac{\left(p_{ikj}-\overline{p}_{kj}\right)^2}{\overline{p}_{kj}} \tag{15}$$

En (15) se puede observar que la inercia no depende del número de variables ni del número de categorías sino de las diferencias entre las proporciones observadas y las teóricas, definidas como las promedio para todos los individuos. A diferencia del ACM, en el ACD la inercia es interpretable de forma similar a lo que sucede en el análisis de correspondencias simples.

3.4.3. Factores y relaciones de transición o cuasibaricéntricas

A partir de las fórmulas del ACP($\mathbf{X}$,$\mathbf{M}$, $\mathbf{D}$) (Escofier & Pagès 1992), se deducen las siguientes fórmulas para el ACD.

La coordenada factorial de un individuo i sobre el eje s es:

$$\psi_{si} = \frac{1}{\sqrt{\lambda_s}}\sum_{j=1}^{p}\sum_{k=1}^{m_j}\frac{p_{ikj}}{p}\varphi_{sk} = \frac{1}{\sqrt{\lambda_s}}\frac{1}{p}\sum_{j=1}^{p}\sum_{k=1}^{m_j}p_{ikj}\varphi_{sk}$$

Análogamente, la coordenada sobre el eje factorial s de una modalidad k es:

$$\varphi_{sk} = \frac{1}{\sqrt{\lambda_s}}\sum_{i=1}^{n}\frac{p_{ikj}}{n\overline{p}_{kj}}\psi_{si} = \frac{1}{\sqrt{\lambda_s}}\frac{1}{n\overline{p}_{kj}}\sum_{i=1}^{n}p_{ikj}\psi_{si}$$

Las fórmulas de las coordenadas factoriales obtenidas para el ACD tienen la misma forma de las que se obtiene en el ACM (fórmulas (9) y (10)). En el ACD, las coordenadas factoriales de una nube de puntos (individuo o columna) son los baricentros de las coordenadas factoriales de la otra nube de puntos dilatada por $\frac{1}{\sqrt{\lambda_s}}$.

3.4.4. Razón de correlación

La varianza de las coordenadas factoriales de las categorías asociadas a la variable j en el eje s es igual a (Chevenet et al. 1994):

$$var_j = \sum_{k=1}^{m_j}\overline{p}_{kj}\varphi_{sk}^2$$

La razón de correlación η_j^2 de la variable j para la coordenada factorial ψ_{si} cuantifica qué tan separadas están las categorías unas de otras, a la vez que indica que tanta relación hay entre la variable j y el eje factorial s:

$$\eta_j^2 = \frac{var_j}{\lambda_s}$$

Se observa que al igual que en el ACM, la razón de correlación se calcula como el cociente entre la varianza (inercia) de la variable j y la varianza total asociada al eje s. Y análogo a lo que se obtiene en el ACM, la razón de correlación del ACD se asocia con las contribuciones relativas de cada variable al eje s mediante la misma relación presentada en (12).

4. Aplicaciones del ACD

4.1. Ecología

Las aplicaciones más comunes del ACD se encuentran en Ecología, ciencia en la que éste método se utiliza principalmente para analizar la estructura de la información de las especies o para encontrar las diferentes asociaciones entre éstas y sus rasgos biológicos. Algunas aplicaciones encontradas en la literatura en las que se ve reflejado lo anterior son:

- **Análisis de datos ecológicos recolectados sobre un largo periodo de tiempo.**
  Como un mecanismo para estructurar la información biológica y ambiental de la flora y la fauna del río Upher Rhone de Francia, Chevenet et al. (1994) tomaron información sobre los rasgos de las especies en esa zona, recolectada durante un largo periodo de tiempo. De esta forma, mediante el uso de la codificación difusa, obtienen una tabla en la que se describe la afinidad de 110 especies de Coleóptera acuática por 32 categorías asociadas a nueve variables ambientales y tres variables ecológicas (por ejemplo resistencia a la polución, alimentación y tolerancia a ciertos niveles de temperatura), usando puntajes positivos de 0: no afinidad de la especie con la categoría de la variable ambiental y/o ecológica a 6: alta afinidad de la especie con la categoría.

- **Análisis de patrones espaciales.**
  Santoul et al. (2005) llevaron a cabo un estudio con el fin de evaluar la variación de los rasgos biológicos de las comunidades de peces de agua dulce como la fecundidad, la edad máxima y el tamaño potencial de la especie, de acuerdo al sitio geográfico en el que se encuentran. Para esto, codificaron los rasgos de los peces, muestreados en 554 puntos diferentes a lo largo del Río Garonna, utilizando la codificación difusa para formar la tabla de especies según rasgos biológicos de acuerdo con lo que observan en los puntos de muestreo. Cada entrada de ésta tabla, indica la afinidad de la especie con el rasgo biológico en una escala de 0: no afinidad a 3: alta afinidad de la especie con la modalidad del rasgo biológico dado.

- **Representación del conocimiento usando codificación difusa.**
  Castella & Speight (1996) encontraron en el ACD una herramienta apropiada y práctica de representar el conocimiento disponible de los investigadores y la literatura en valores numéricos para caracterizar las diferentes especies mediante tablas de codificación difusa. Esta tabla presenta la información de las especies a través de nueve variables que las describen biológica y ecológicamente. La relación entre la especie y la categoría se codifica en cuatro valores: 0: no asociación; 1: asociación débil; 2: asociación significante; 3: asociación fuerte. Este grado de asociación de la especie con las características se asigna a partir del conocimiento de los expertos y la información de diferentes fuentes literarias.

4.2. Análisis armónico cualitativo

El análisis armónico cualitativo (AAC) es un método exploratorio de análisis longitudinal para variables categóricas propuesto por Deville & Saporta (1980). Es una herramienta para el análisis de una variable que describe el paso de un individuo por diferentes estados (categorías) de una variable cualitativa. Por ejemplo, el cambio de lugar de residencia, el cambio de actividad profesional o el cambio de estado civil. El objetivo del AAC es identificar tipologías compuestas por grupos de individuos con trayectorias similares, manteniendo el orden y la cronología de los eventos (Saporta 1996).

Para la implementación de éste método el periodo de análisis se dividide en p subperiodos o intervalos de tiempo. Una vez se realiza esta división, la tabla de frecuencias se puede ver como una tabla de código difuso. Se construye la tabla de tal forma que cada entrada represente el número de veces o frecuencia (según la escala de tiempo que se maneje) con la que un individuo i asume el estado k en el tiempo (o intervalo de tiempo) j, en proporción a la duración de cada intervalo (Barbary 1996).

La tabla de frecuencias que ingresa al análisis consta de n individuos por p intervalos (subperiodos) de tiempo con m estados (modalidades de la variable cualitativa) cada uno.

Al aplicar el ACD a la tabla de datos que se utiliza para el AAC y que presenta codificación difusa, se pierde la información al interior del periodo, pero no se pierde la proporción del tiempo que un individuo permanece en un estado para determinado periodo.

4.3. Concatenación de tablas de contingencia

Cuando se cuenta con la información de las categorías que asume cada individuo a través de las diferentes variables de interés, el ACM es bastante útil y apropiado para analizar éste conjunto de datos. Sin embargo, hay muchos casos en los que sólo se cuenta con la información resumida. Si se tienen datos de ésta clase, el ACD puede ser una buena aproximación del ACM, utilizando las tablas de individuos por variables categóricas concatenadas. Partiendo de esta idea, en la siguiente sección se presenta la aplicación del ACD a una tabla con codificación difusa originada de la concatenación de tablas de contingencia.

5. Perfil socioeconómico de los estudiantes admitidos a la UN, sede Bogotá, según carreras

Los datos utilizados en este ejemplo de aplicación corresponden a la información suministrada por los 2927 estudiantes que fueron admitidos a la Universidad Nacional de Colombia, sede Bogotá, en el primer semestre de 2006, facilitada por el Departamento Nacional de Admisiones (DNA). El objetivo es analizar las características socioeconómicas de éstos estudiantes.

La información y las variables utilizadas se tomaron de la encuesta de caracterización de estudiantes. Este instrumento consta de 18 preguntas, que deben ser diligenciadas por el estudiante, mediante las cuales se indaga por aspectos tales como el carácter del colegio en el que estudió el último año, la jornada en la que estudiaba, el número de personas con las que convive, acceso a Internet y/o computador y otros países que ha conocido, entre otras.

Después de seleccionar las preguntas relacionadas con características socioeconómicas de los estudiantes, algunas variables se recodificaron en menos categorías de acuerdo con el contexto de las preguntas y las frecuencias de las diferentes opciones de respuesta. Por ejemplo, la variable estrato se recodificó en bajo (estratos 1 y 2), medio (estrato 3) y alto (estratos 4, 5 y 6), al igual que la variable que hace referencia al tamaño del núcleo familiar del admitido que se recodificó en *Per1_3*, *Per4*, *Per5_6* y *Per7Mas* según si el estudiante convive sólo o con dos personas, con tres personas, con cuatro o cinco personas, o con seis personas o más, respectivamente. Después de realizada esta recodificación, se encuentra que la no respuesta a las variables carácter del colegio y jornada forman ejes en el ACM, por lo que se decide no tener en cuenta a esos 131 estudiantes en el análisis.

La tabla de datos final, presenta la información de 2796 estudiantes por las siete variables categóricas que se presentan en la tabla 1. La descripción de las categorías de las variables y los histogramas de frecuencia se presentan en el apéndice (tabla 2)

5.1. Análisis de correspondencias múltiples

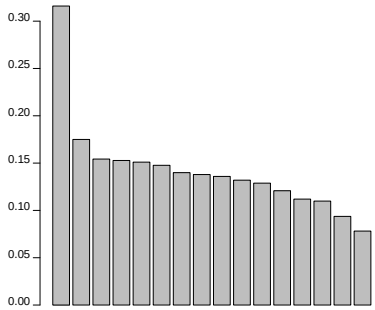
La tabla de datos que ingresa al ACM consta de 2796 estudiantes en las filas y siete variables categóricas en las columnas y la variable carrera se utiliza como ilustrativa.

5.1.1. Resultados del ACM

Valores propios: en el histograma de los valores propios (figura 2) se observa que la información se resume prácticamente sobre el primer eje. A partir del tercer eje la variabilidad se puede asumir como "ruido".

Variable	Descripción	Etiqueta	Número de modalidades
Carrera	Carreras ofrecidas por la Universidad Nacional en la Sede Bogotá.	Carrera	47
Carácter	Carácter del colegio en el que estudió el último año el admitido.	Caract	4
Jornada	Jornada del colegio en la que estudió el último año el admitido.	Jorn	4
Estrato	Estrato de la vivienda en la que habita el estudiante.	Estr	3
Tamaño	Tamaño del núcleo familiar con el que convive el estudiante incluyéndolo a él.	Tam	4
Tecnología	Acceso a computador y/o Internet.	Tecno	3
Ocupación	Actividad a la que se dedica el estudiante en el momento de ser admitido.	Ocup	4
Viaje	Viajes a otros países que ha realizado el estudiante.	Viajes	3

Tabla 1: Descripción de variables categóricas utilizadas.



	Valor propio	Proporción acumulada de la inercia
1	0.32	0.14
2	0.18	0.21
3	0.15	0.28
4	0.15	0.35
5	0.15	0.42
6	0.15	0.48
7	0.14	0.54
8	0.14	0.60

Figura 2: Histograma y tabla de valores propios del ACM

Ejes factoriales: en la tabla 3 del apéndice se presentan las contribuciones absolutas y relativas de las categorías sobre los dos primeros ejes factoriales. Con relación al primer eje se observa que está conformado principalmente por las categorías *estrato alto*, el *acceso a computador e Internet* y los *viajes a otras regiones fuera de Latinoamérica*, las cuales agrupan el 40.4 % de inercia del primer eje. Este eje contrapone las categorías *colegio académico técnico*, la *jornada nocturna o en la tarde*, los *estratos bajos*, el *no acceso a tecnología* así como el *no haber viajado a otros países además de Latinoamérica versus* las categorías *académico, completa, estrato alto, acceso a computador e Internet, estudia* y *conocimiento de otros países diferentes a los latinoamericanos*. Análogamente contrapone las carreras *medicina*, *artes plásticas*, *economía*, *cine y televisión*, *literatura*, *medicina* y las *ingenierías civil*, *química*, *industrial*, *mecatrónica* y *electrónica versus* las carreras *español y filología clásica*, *trabajo social*, *fonoaudiología*, *enfermería*, *contaduría*, *lingüística*, *estadística*, *zootecnia* e *historia*.

En el segundo contrapone las categorías *jornada mañana*, *estrato medio*, *acceso a computador* y *estudia* (abajo) versus categorías como *jornada nocturna o en la tarde*, *estrato bajo*, *no acceso a tecnología* y *trabaja*. Las carreras *farmacia* y *diseño industrial* se proyectan abajo separadas de las demás carreras, asociadas a nivel socioeconómico medio.

Observando las razones de correlación (figura 3) se corrobora que las variables que más influyen en la formación tanto del primero como del segundo eje, son el estrato del admitido y el acceso que tiene a tecnología (computador e Internet).

Plano factorial: se observa en el plano de las categorías activas del ACM (figura 4) un efecto *Guttman*, mostrando el ordenamiento de los estudiantes según su nivel socioeconómico: los de menor nivel ubicados arriba a la izquierda, los de nivel medio abajo y los de nivel más alto arriba a la derecha.

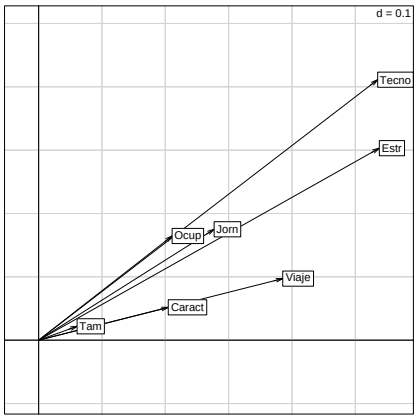


FIGURA 3: Razones de correlación del ACM

En las carreras se observa un ordenamiento similar sobretodo en el primer eje (figura 5), asociado al nivel socioeconómico de sus estudiantes:

- a la izquierda: *trabajo social*, *español y filología clásica*, *enfermería*, *contaduría*, *terapia ocupacional*, *fonoaudiología* y *lingüística*;
- abajo: *farmacia* y *diseño industrial*
- a la derecha: *artes plásticas*, *economía* y algunas ingenierías como *electrónica*, *industrial* y *mecatrónica*.

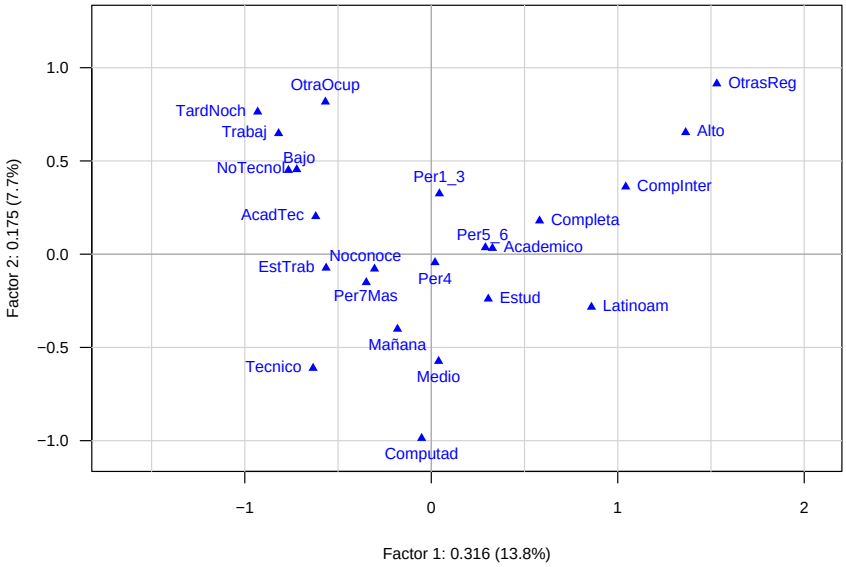


FIGURA 4: Variables activas en el Primer plano del ACM

5.2. Análisis de correspondencias difuso

Se analiza la tabla, de dimensión 47×23, que concatena siete tablas de contingencia (tabla 6 en el apéndice). Cada una cruza 47 carreras con las categorías de la variable respectiva. El valor de una celda es el número de estudiantes de la carrera i que asumen la categoría k de la variable j.

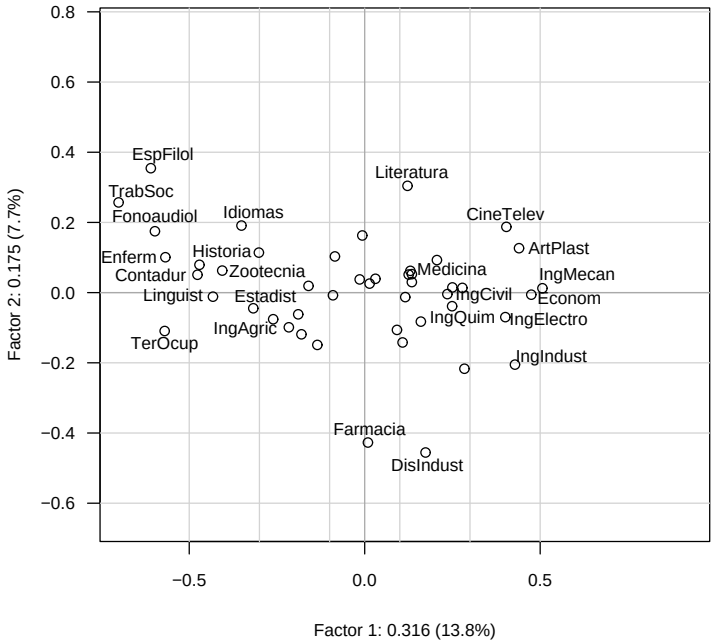


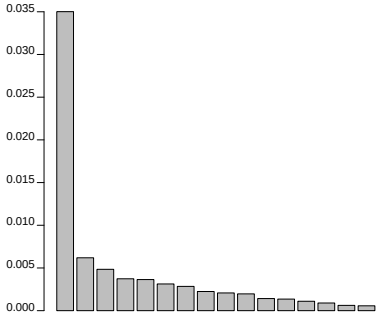
FIGURA 5: Carreras como ilustrativas en el primer plano del ACM. Se etiquetan las que tienen un valor test superior a dos en valor absoluto en alguno de los dos ejes (tabla 4 del apéndice) y *cine y televisión* (valor test 1.76) y *literatura* (1.86).

Para la aplicación del ACD, se usan como pesos las proporciones de estudiantes en cada carrera en lugar del peso constante $\frac{1}{n}$, ya que cada fila representa al grupo de estudiantes que se admitieron a una carrera.

Los pasos para realizar el análisis de correspondencias difuso al igual que el código utilizado en R (R Development Core Team 2007) se presentan en el apéndice.

5.2.1. Resultados del ACD

Valores propios: se observa que la información está en el primer eje y por lo tanto es suficiente analizar el primer plano factorial (figura 6). Los dos primeros ejes factoriales retienen el 57.5 % de la inercia total.



	Valor propio	Proporción acumulada de la inercia
1	0.035	0.49
2	0.006	0.58
3	0.005	0.64
4	0.004	0.69
5	0.004	0.75
6	0.003	0.79
7	0.003	0.83
8	0.002	0.86

FIGURA 6: Histograma y tabla de valores propios del ACD

Ejes factoriales: con base en las contribuciones absolutas de las categorías a cada uno de los ejes (tabla 5 en el apéndice) se encuentra que el primer eje está conformado principalmente por los *estratos altos* y *bajos*, el *acceso a computador e Internet* y las carreras *enfermería*, *economía*, *ingeniería mecánica*, *trabajo social* y *contaduría*. Análogamente, el segundo eje factorial está conformado principalmente por las categorías *trabaja*, *conoce países de Latinoamérica, estudia* y las carreras *contaduría*, *idiomas* y *cine y televisión*.

Las razones de correlación, presentadas en la figura 7, son muy pequeñas. Sin embargo las variables que más influyen en la construcción del primer eje son casi las mismas que las de ACM (figura 3). En el segundo eje se puede observar alguna correlación con ocupación del estudiante y la variable *viajes*.

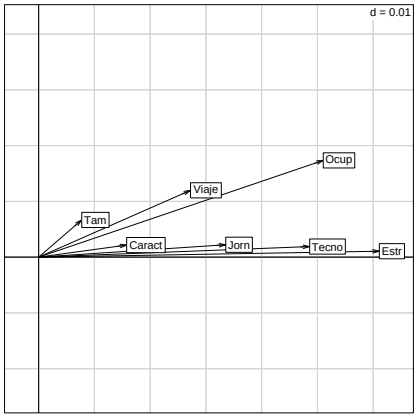


Figura 7: Razones de correlación del ACD

Plano factorial: en el gráfico 8 se observa un ordenamiento socioeconómico de los grupos de estudiantes de izquierda a derecha en el primer eje, que permite ver la contraposición de dos grupos de carreras:

- *Lingüística*, *contaduría*, *enfermería*, *terapia ocupacional*, *fonoaudiología*, *zootecnia*, *trabajo social*, *geografía* y *estadística*, asociadas a las categorías colegios de *carácter académico técnico*, estudiaban en la *tarde o en la noche*, son de *estratos bajos* y *no disponen de un computador ni acceso a Internet*.
- *Economía*, *artes plásticas* y algunas ingenierías como *electrónica*, *mecánica* e *industrial* que se asocian a estudiantes que pertenecen a *estratos altos*, disponen de un *computador* y *tiene acceso a Internet*, provienen de colegios de *jornada completa* y *carácter académico*.

Adicionalmente, en el segundo eje la carrera *cine y televisión* se encuentra arriba a la derecha, indicando que en esta carrera hay un aumento de la proporción de estudiantes que han tenido la posibilidad de realizar *viajes a países de Latinoamérica*.

5.3. Comparación entre el ACM y el ACD

En el análisis de este conjunto de datos los planos factoriales de los dos métodos (figuras 4 y 8) son muy similares y permiten mas o menos las mismas conclusiones.

El ACD de las tablas de contingencia concatenadas permite ver mejor las asociaciones entre categorías y carreras. Esto se debe a que en este análisis las carreras son activas, mientras que en el ACM son ilustrativas.

La diferencia técnica más importante se da en la disminución drástica de las razones de correlación en el ACD con respecto a las del ACM (ver figuras 3 y 7). Esto puede estar sucediendo porque en el ACD cada una de las carreras está agrupando varios estudiantes que asumen diferentes categorías de

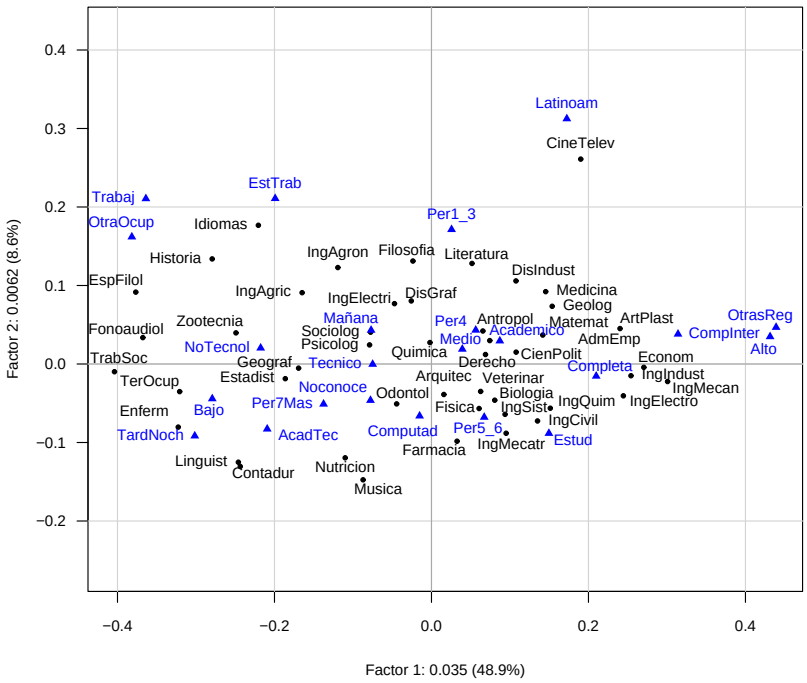


FIGURA 8: Primer plano factorial del ACD.

las variables, por lo que la varianza entre variables es pequeña. No obstante, en el ACD se conserva la estructura de correlación que se tiene en el ACM.

6. Conclusiones

- Las fórmulas que se obtienen para el ACD son análogas a las del ACM. La forma que tienen las fórmulas son equivalentes y se interpretan de manera similar, a excepción de la inercia, que en el ACD depende de los valores de la tabla no del número de variables ni del número de categorías.

- El ACD realizado sobre varias tablas de contingencia concatenadas parece ser una buena aproximación al ACM que se podría hacer si se tuviera la información de las categorías que asumen cada uno de los individuos. Sin embargo, para tener certeza sobre esta afirmación, se debe verificar con otros ejemplos de aplicación o mediante simulaciones.

Referencias

Barbary, O. (1996), Una aplicación del análisis armónico cualitativo: la tipología de trayectorias individuales, *in* 'Memorias del seminario de capacitación e investigación: Recolección y análisis de datos logitudinales', Universidad Nacional de Colombia, Bogotá, pp. 121–144.

Castella, E. & Speight, M. (1996), 'Knowledge representation using fuzzy coded variables: an example based on the use of syrphidae (insecta, diptera) in the assessment of riverine wetlands', *Ecological modelling* **85**, 13–25.

Chessel, D., Dufour, A.-B. & Thioulouse, J. (2004), 'The ade4 package-I- One-table methods', *R News* **4**, 5–10.

Chevenet, F., Dolédec, S. & Chessel, D. (1994), 'A fuzzy coding approach for the analysis of long-term ecological data', *Freshwater Biology* **31**, 295–309.

Deville, J. & Saporta, G. (1980), *Analyse harmonique qualitative,in Data Analysis and Informatics*, E. DIDAY et al.éditeurs, North Holland Publishing Compagny.

Dolédec, S. & Chevenet, F. (1994), 'Fuzzy correspondence analysis', *ADE-4, Fiche thématique 2.5* pp. 1–20.
*http://pbil.univ-lyon1.fr/R/themaold/thema25.pdf

Escofier, B. & Pagès, J. (1992), *Análisis factoriales simples y múltiples. Objetivos, métodos e interpretación*, Universidad del Pais Vasco, Bilbao.

Lebart, L., Morineau, A. & Piron, M. (1995), *Statisitique exploratoire multidimensionnelle*, Dunod, Paris.

Pagès, J. (2004), 'Multiple factor analysis: Main features and application to sensory data', *Revista Colombiana de Estadística* **27**(1), 1–26.

Pardo, C. E. & Del-Campo, P. C. (2007), 'Combinación de métodos factoriales y de análisis de conglomerados en R: el paquete FactoClass', *Revista Colombiana de Estadística* **30**(2), 231–245.

R Development Core Team (2007), *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

Santoul, Cayrou, Mastrorillo & Céréghino (2005), 'Spatial patterns of biological traits of freshwater fish communities in south-west france', *Journal of Fish Biology* **66**, 301–314.

Saporta, G. (1996), L'analyse harmonique qualitative, une synthèse de la théorie, *in* 'Memorias del seminario de capacitación e investigación: Recolección y análisis de datos logitudinales', Universidad Nacional de Colombia, Bogotá, pp. 111–120.

Tenenhaus, M. & Young, F. (1985), 'An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data', *Psychometrika* **50**, 91–119.

Apéndice

Se presentan: la descripción de las variables y sus categoría (tabla 2), el procedimiento para realizar un ACD junto con el código en R, las ayudas para la interpretación de los planos factoriales del ACM (tablas 3 y 4) y del ACD (tabla 5) y la tabla de datos para el ACD (tabla 6).

Pasos para realizar el ACD en R.

Para realizar el análisis de correspondencias difuso, se utilizan las funciones `prep.fuzzy.var` y `dudi.fca` implementada en el paquete `ade4` (Chessel et al. 2004) de R. Los pasos a seguir para su aplicación se listan a continuación:

- Ingresar la tabla con codificación difusa (**A**) a R.
- Cargar la librería `ade4`.
- Crear un vector que contenga el número de categorías de cada una de las variables difusas (`A.blo`).
- Crear un vector que contenga el peso de las filas (`row.w`). El peso de filas utilizado es la proporción de estudiantes en cada carrera.

Tabla 2: Descripción de las categorías de cada una de las variables utilizadas en los análisis.

Variable	Categorías	Frecuencia	Porcentaje	Histograma
	Academico	1833	65.6	*******
Caract	AcadTec[1]	651	23.3	**
Carácter del colegio	Tecnico	312	11.2	*
	Completa	1082	38.7	****
Jorn	Mañana	1290	46.1	*****
Jornada del colegio	TardNoch	424	15.2	*
	EstAlto[2]	503	18.0	*
Estr	EstMedio[3]	1329	47.5	****
Estrato	EstBajo[4]	964	34.5	***
Tam	Per1_3	383	13.7	*
Número de personas con	Per4	762	27.3	***
las que convive el	Per5_6	851	30.4	***
estudiante	Per7mas	800	28.6	***
Tecno	CompInter	829	29.7	***
Acceso a tecnología	Computad	828	29.6	***
	NoTecnol[5]	1139	40.7	****
	EstTrab	213	7.6	*
Ocup	Estud	1923	68.8	*******
Ocupación del estudiante	OtraOcup[6]	289	10.3	*
	Trabaj	371	13.3	*
	Latinoam	288	10.3	*
Viaje	Noconoce	2227	79.7	********
Viajes a otros países	OtrasReg[7]	281	10.1	*

[1] Colegio académico técnico o normalista; [2] Estratos 4, 5 ó 6; [3] Estrato 3; [4] Estratos 1, 2 o vivienda no estratificada;

[5] No acceso a computador ni Internet.; [6] No estudia y no trabaja o no reporta la ocupación.;

[7] Regiones fuera de Latinoamérica (Estados Unidos, Europa, otros países.)

- Obtener la tabla de proporciones **P** mediante la función `prep.fuzzy.var`. Si se tienen datos faltantes, ésta función reemplaza estos datos por el perfil medio de la variable, donde el perfil es calculado como $\frac{a_{ikj}}{a_{i\cdot j}}$
- Con esta tabla **P** se realiza el análisis de correspondencias difuso, mediante la función `dudi.fca`, eligiendo el número de ejes que se quieren retener.
- Las gráficas de los planos factoriales se realizan utilizando el paquete `planfac` implementado en la librería `FactoClass` (Pardo & Del-Campo 2007).

Código utilizado para la aplicación del ACD en R.

```
#-------------------------------------------------------------------------------
a<-read.table("TablaDifusa.txt", header=T, row.names=1) # Importa la tabla de datos
A<-data.frame(a) # Convierte la tabla de datos en una data frame
#-------------------------------------------------------------------------------
library(ade4)   # Carga la librería ade4
# Declara un vector con el número de categorías de  cada variable difusa
A.blo <- c(3,3,3,4,3,4,3)
# Especifica el vector con el nombre de las variables.
names(A.blo)<-c("Caract","Jorn","Estr","Tam","Tecno","Ocup","Viaje")
# Asignar los pesos de las filas
n <- apply(A,1,sum)
p <- n/2796
P<-prep.fuzzy.var(A,A.blo,row.w= p)
```

```
#-------------------------------------------------------------------------------
acd <- dudi.fca(P,scannf=F,nf=2) # Realiza el análisis de correspondencias difuso
#coordenadas factoriales, inercia, cosenos cuadrados y valores propios
acd$li
acd$co
inercia<-inertia.dudi(acd,T,T)
e <- acd$eig
barplot(acd$eig)
#Razon de correlacion=contribuciones absolutas de las variables
acd$cr
s.arrow(acd$cr)
#-------------------------------------------------------------------------------
library(FactoClass) # Carga libreria FactoClass
windows()
planfac(acd,cframe=1.0, all.point=T) # Gráfica del primer plano factorial del ACD
```

Tabla 3: Contribuciones relativas y absolutas de las categorías en el ACM

	Contribuciones absolutas[1]		**Contribuciones relativas**[1]	
Categoría	**Comp1**	**Comp2**	**Comp1**	**Comp2**
Academico	319	5	2047	20
AcadTec	404	78	-1165	125
Tecnico	203	340	-505	-469
Completa	590	102	2129	203
Mañana	68	607	-281	-1379
TardNoch	594	721	-1549	1042
Alto	1514	626	4082	935
Bajo	915	571	-3089	1067
Medio	3	1278	14	-2983
Per1_3	1	118	3	167
Per4	0	4	2	-7
Per5_6	116	3	368	6
Per7Mas	157	54	-487	-92
CompInter	1458	316	4584	551
Computad	4	2354	-11	-4098
NoTecnol	959	686	-3578	1419
EstTrab	110	3	-262	-4
Estud	291	322	2059	-1264
OtraOcup	151	563	-372	769
Trabaj	402	454	-1025	642
Latinoam	344	68	847	-92
Noconoce	333	40	-3624	-242
OtrasReg	1065	686	2620	934

[1] los valores están multiplicados por 10000.

Tabla 4: Valores test de las carreras en el ACM

	Valores test		Coord. factoriales			Valores test		Coord. factoriales	
Carrera	**Eje 1**	**Eje 2**	**Eje 1**	**Eje 2**	**Carrera**	**Eje 1**	**Eje 2**	**Eje 1**	**Eje 2**
AdmEmp	1.40	0.67	0.13	0.06	IngAgron	-1.72	-0.79	-0.22	-0.10
Antropol	0.91	0.36	0.13	0.05	IngCivil	2.56	-0.05	0.24	0.00
Arquitec	0.28	0.36	0.03	0.04	IngElectri	-0.53	0.64	-0.08	0.10
ArtPlast	2.46	0.71	0.44	0.13	IngElectro	3.99	-0.70	0.40	-0.07
Biologia	1.14	0.25	0.13	0.03	IngIndust	2.96	-1.42	0.43	-0.21
CienPolit	1.90	0.86	0.21	0.09	IngMecan	4.54	0.11	0.51	0.01
CineTelev	1.86	0.86	0.40	0.19	IngMecatr	0.79	-1.03	0.11	-0.14
Contadur	-5.19	0.55	-0.48	0.05	IngQuim	2.42	-0.38	0.25	-0.04
Derecho	1.15	0.47	0.13	0.05	IngSist	1.52	-0.78	0.16	-0.08
DisGraf	-0.07	0.20	-0.01	0.04	Linguist	-2.75	-0.07	-0.43	-0.01
DisIndust	1.08	-2.83	0.17	-0.46	Literatura	0.70	1.76	0.12	0.30
Econom	5.08	-0.06	0.47	-0.01	Matemat	1.67	0.10	0.25	0.01
Enferm	-5.84	1.04	-0.57	0.10	Medicina	3.25	0.16	0.28	0.01
EspFilol	-3.88	2.26	-0.61	0.35	Musica	-0.85	-0.56	-0.18	-0.12
Estadist	-2.17	-0.31	-0.32	-0.04	Nutricion	-1.22	-0.40	-0.19	-0.06
Farmacia	0.07	-3.43	0.01	-0.43	Odontol	-0.65	-0.05	-0.09	-0.01
Filosofia	-0.04	0.87	-0.01	0.16	Psicolog	-1.18	-1.31	-0.13	-0.15
Fisica	0.74	-0.85	0.09	-0.11	Quimica	0.10	0.18	0.01	0.03
Fonoaudiol	-4.04	1.18	-0.60	0.17	Sociolog	-1.02	0.12	-0.16	0.02
Geograf	-1.35	0.51	-0.30	0.11	TerOcup	-3.81	-0.73	-0.57	-0.11
Geolog	1.67	-1.27	0.28	-0.22	TrabSoc	-4.58	1.68	-0.70	0.26
Historia	-2.41	0.41	-0.47	0.08	Veterinar	0.82	-0.09	0.12	-0.01
Idiomas	-3.08	1.68	-0.35	0.19	Zootecnia	-2.78	0.43	-0.41	0.06
IngAgric	-2.19	-0.64	-0.26	-0.08					

Tabla 5: Contribuciones relativas y absolutas de las categorías en el ACD.

	Contribuciones absolutas[1]		Contribuciones relativas[1]	
Categoría	**Comp1**	**Comp2**	**Comp1**	**Comp2**
Academico	203	131	5117	586
AcadTec	416	368	-4520	-708
Tecnico	26	0	-467	0
Completa	695	21	6733	-36
Mañana	111	197	-1815	569
TardNoch	563	294	-5484	-506
Alto	1366	50	7989	52
Bajo	1097	156	-7956	-200
Medio	30	39	836	193
Per1_3	4	928	53	2375
Per4	35	117	625	366
Per5_6	56	323	1208	-1229
Per7Mas	221	172	-3479	-480
CompInter	1193	99	8627	127
Computad	3	299	-66	-1270
NoTecnol	787	39	-7439	65
EstTrab	123	781	-1364	1529
Estud	628	1238	6633	-2312
OtraOcup	615	625	-4745	854
Trabaj	717	1358	-4729	1585
Latinoam	125	2319	1188	3901
Noconoce	196	394	-5925	-2104
OtrasReg	790	51	6464	73

[1] los valores están multiplicados por 10000.

Tabla 6: Datos para el ACD: concatenación de tablas de contingencia

	Caract			Jorn			Estr			Tam				Tecnol			Ocup				Viajes		
Carrera	1	2	3	1	2	3	1	2	3	1	2	3	4	1	2	3	1	2	3	4	1	2	3
AdmEmp	86	18	8	44	49	19	23	34	55	17	38	29	28	42	27	43	9	80	11	12	11	90	11
Antropol	34	7	4	24	13	8	8	15	22	2	18	14	11	17	15	13	4	27	5	9	6	35	4
Arquitec	54	17	11	33	38	11	16	31	35	7	18	26	31	27	22	33	8	57	8	9	6	65	11
ArtPlast	21	6	4	17	12	2	11	7	13	3	10	11	7	15	8	8	3	21	3	4	3	23	5
Biologia	43	20	7	32	30	8	16	26	28	7	24	20	19	24	21	25	6	53	7	4	7	54	9
CienPolit	56	20	7	39	32	12	23	24	36	6	28	26	23	28	24	31	7	58	9	9	12	60	11
CineTelev	16	3	2	9	11	1	4	7	10	6	7	3	5	8	6	7	0	14	2	5	5	9	7
Contadur	61	36	17	36	48	30	10	59	45	13	28	27	46	15	32	67	9	81	8	16	6	105	3
Derecho	61	16	5	33	40	9	15	30	37	12	27	23	20	25	23	34	4	61	6	11	8	61	13
DisGraf	21	2	4	11	9	7	4	9	14	2	4	11	10	9	8	10	3	15	1	8	5	20	2
DisIndust	25	5	8	21	16	1	4	8	26	4	14	10	10	14	16	8	5	23	4	6	6	29	3
Econom	88	14	8	58	38	14	34	21	55	16	27	35	32	43	34	33	9	90	4	7	13	76	21
Enferm	57	33	12	26	50	26	3	54	45	14	22	30	36	13	32	57	7	58	19	18	2	96	4
EspFilol	23	14	3	13	16	11	3	20	17	9	5	7	19	5	11	24	7	14	9	10	3	35	2
Estadist	28	15	3	16	22	8	4	18	24	6	10	13	17	9	16	21	5	26	6	9	3	41	2
Farmacia	38	15	10	19	36	8	4	19	40	8	19	21	15	21	25	17	4	51	4	4	3	53	7
Filosofia	20	5	3	12	10	6	5	6	17	5	8	7	8	9	7	12	3	14	6	5	3	22	3
Fisica	40	19	4	23	33	7	12	15	36	9	14	27	13	20	18	25	6	47	4	6	3	53	7
Fonoaudiol	22	16	7	8	30	7	4	25	16	4	10	16	15	6	11	28	4	20	13	8	4	40	1
Geograf	12	5	3	5	9	6	2	9	9	2	8	7	3	5	6	9	1	12	3	4	2	17	1
Geolog	26	6	2	10	18	6	8	7	19	9	8	13	4	10	15	9	3	26	2	3	6	23	5
Historia	12	11	3	7	14	5	4	12	10	6	6	7	7	3	9	14	4	12	2	8	5	21	0
Idiomas	46	21	8	18	46	11	11	26	38	18	22	15	20	13	19	43	5	32	13	25	6	62	7
IngAgric	41	18	10	20	39	10	5	32	32	6	16	21	26	20	18	31	11	38	9	11	14	53	2
IngAgron	43	8	11	15	37	10	6	22	34	10	15	17	20	15	15	32	13	34	5	10	7	49	6
IngCivil	79	24	10	47	53	13	28	37	48	13	22	41	37	42	31	40	5	94	5	9	11	86	16
IngElectri	26	7	5	15	16	7	6	17	15	8	13	6	11	14	10	14	6	23	5	4	4	32	2
IngElectro	70	17	9	44	42	10	29	20	47	13	33	30	20	37	28	31	3	84	2	7	9	72	15
IngIndust	32	6	9	28	18	1	14	11	22	6	14	17	10	17	18	12	2	38	5	2	6	35	6
IngMecan	57	9	12	43	29	6	24	14	40	9	19	29	21	35	17	26	4	66	1	7	7	55	16
IngMecatr	29	13	10	26	19	7	12	11	29	7	16	14	15	17	16	19	0	42	3	7	3	45	4
IngQuim	62	23	6	47	35	9	20	23	48	16	19	30	26	33	24	34	4	76	5	6	9	72	10
IngSist	55	21	11	36	37	14	19	22	46	11	19	38	19	30	30	27	4	65	9	9	7	70	10
Linguist	24	12	4	14	14	12	3	20	17	2	9	12	17	8	18	14	4	22	4	10	1	38	1
Literatura	18	11	4	18	8	7	7	8	18	6	10	11	6	12	7	14	5	18	6	4	6	22	5
Matemat	33	8	3	17	24	3	11	13	20	5	18	11	10	18	11	15	1	33	2	8	4	34	6
Medicina	94	23	13	55	64	11	36	41	53	25	34	40	31	50	31	49	6	101	14	9	30	91	9
Musica	10	10	2	10	7	5	4	10	8	4	3	8	7	2	11	9	1	18	0	3	4	17	1
Nutricion	27	11	3	12	20	9	2	20	19	3	13	15	10	12	15	14	2	29	7	3	2	36	3
Odontol	32	10	8	15	24	11	7	22	21	8	12	12	18	13	17	20	4	37	5	4	4	38	8
Psicolog	47	20	8	22	44	9	8	29	38	15	22	23	15	18	27	30	5	50	12	8	8	62	5
Quimica	29	16	5	16	26	8	9	15	26	7	16	12	15	17	12	21	1	36	9	4	8	37	5
Sociolog	23	10	7	14	19	7	5	13	22	7	13	10	10	11	10	19	3	25	4	8	3	33	4
TerOcup	23	14	7	10	26	8	2	21	21	2	16	12	14	3	15	26	2	25	10	7	4	40	0
TrabSoc	28	9	5	12	21	9	1	25	16	4	12	11	15	3	10	29	2	19	8	13	0	42	0
Veterinar	33	14	2	18	26	5	13	15	21	7	15	15	12	14	17	18	3	36	5	5	3	40	6
Zootecnia	28	13	5	14	22	10	4	21	21	4	8	18	16	7	15	24	6	22	5	13	6	38	2

Las categorías están en la tabla 2