

## CAPÍTULO 6

# INTRODUCCIÓN AL ANÁLISIS DE CORRESPONDENCIAS

*Denis Baranger y Fernanda Niño*

En base a la obra fundamental de Jean-Paul Benzécri (1973) tuvo su origen la denominada ‘escuela francesa de análisis de datos’ (*analyse des données*), que ha desarrollado una serie de métodos de análisis multivariados como el análisis factorial de las correspondencias (AFC) y el análisis de correspondencias múltiples (ACM), y renovado otros como el análisis en componentes principales (ACP) y las técnicas de clasificación automática.<sup>1</sup>

En particular el análisis de correspondencias es una técnica cuyo uso se ha ido popularizando en los últimos años. En este proceso tuvo mucho que ver la utilización que hizo Pierre Bourdieu del ACM en *La distinction* (1979) y en su producción posterior.<sup>2</sup> Aunque sin llegar al extremo de lo que sucede en Francia, donde es habitual que las revistas de actualidad publiquen planos factoriales para ilustrar los resultados de encuestas, la técnica del ACM se ha ido abriendo un camino y su uso se ha ido generalizando en otros países tanto en el campo de la investigación básica en ciencias sociales como en los estudios de *marketing* y de opinión pública.

No nos referiremos aquí a los fundamentos del análisis factorial de correspondencias (AFC), el que puede ser considerado como una generalización del análisis en componentes principales (ACP) adaptado al procesamiento de datos cualitativos.<sup>3</sup> Nuestro propósito en este capítulo no es reemplazar a una indispensable formación específica en esta materia;<sup>4</sup> más bien nos proponemos realizar una presentación que ilustre acerca de las potencialidades de las técnicas multivariadas de análisis de datos, brindando al lego algunos elementos para una mejor interpretación de un modo de presentación de los resultados de la investigación al que se está viendo crecientemente confrontado.

Examinaremos aquí algunos ejemplos de usos de estas técnicas multivariadas. Comenzaremos con el análisis de correspondencias a partir de la presentación de dos ejemplos de tablas de contingencia. Luego abordaremos el ACM mostrando cómo puede ser visto simplemente como una forma más de reducción de un espacio de propiedades basado en variables cualitativas, y cómo es posible utilizarlo como una alternativa al procedimiento numérico corrientemente utilizado en la construcción de índices.<sup>5</sup> Por último ejemplificaremos cómo se puede combinar el ACM con técnicas de clasificación automática en el análisis de encuestas.

### 1. EL ANÁLISIS FACTORIAL DE CORRESPONDENCIAS

Si bien el Análisis Factorial de Correspondencias (AFC) es un método especialmente adecuado para explorar tablas de grandes dimensiones, comenzaremos analizando una tabla de 4x4 cuyo significado es relativamente simple de aprehender, pero que servirá para introducir los principios

---

<sup>1</sup> Todos éstas constituyen, según Lebart, “métodos de estadística descriptiva multidimensional”, los que comprenden dos grandes familias: los métodos factoriales y los métodos de clasificación; ambas familias de técnicas son corrientemente utilizadas de modo complementario en el análisis de datos de encuesta (Lebart; 1989: 23).

<sup>2</sup> Tanto en la revista que dirige Bourdieu, *Actes de la Recherche*, como en las obras en que da cuenta de sus investigaciones empíricas y en aquellas provenientes de su grupo de colaboradores es recurrente la presentación de resultados mediante el uso de planos factoriales.

<sup>3</sup> Cf. Tenenhaus (1994: 145).

<sup>4</sup> Durante los últimos años el Programa PRESTA <<http://www.ulb.ac.be/assoc/presta/>>, desarrollado desde la U. Libre de Bruselas con el apoyo de la Comunidad Europea, ha encarado esta tarea a nivel de América del Sur.

<sup>5</sup> Esta aplicación del ACM se inscribe dentro de la alternativa que Lazarsfeld (1993) y Barton (1969) denominaban ‘reducción funcional de un espacio de propiedades’.

básicos de esta técnica.

Tabla 6.1: Buenos Aires, 1982 - Ocupación del hijo y ocupación del padre

| Ocupación del padre | Ocupación del hijo |             |                |                | Total |
|---------------------|--------------------|-------------|----------------|----------------|-------|
|                     | Manual Baja        | Manual Alta | No manual Baja | No manual Alta |       |
| No manual Alta      | 2                  | 5           | 14             | 72             | 93    |
| No manual Baja      | 18                 | 28          | 119            | 56             | 221   |
| Manual Alta         | 23                 | 58          | 50             | 17             | 148   |
| Manual Baja         | 69                 | 51          | 52             | 18             | 190   |
| Total               | 112                | 142         | 235            | 163            | 652   |

Fuente: datos adaptados de Jorrat, 1997: 102.

Para determinar la existencia de una situación de independencia estadística entre dos variables, disponemos del coeficiente de asociación  $\varphi^2$  (fi cuadrado). Empero  $\varphi^2$ , como cualquier coeficiente, es una medida sintética que no da cuenta de la estructura de la relación entre las variables. Si, en cambio, estamos interesados en analizar esa estructura, será necesario descomponer esa relación y describir su *forma*, identificando cuáles son los valores de cada variable que se asocian más entre sí. En la práctica, éste es el tipo de análisis que presenta un mayor interés.

En el ejemplo, el  $\varphi^2$  podría ser alto, tanto porque los hijos tendieran a mantener el mismo tipo de ocupación de sus padres, como porque la cambiaran en una determinada dirección.<sup>6</sup> Por otro lado, también podría ocurrir que el valor del  $\varphi^2$  fuera bajo, y que aún así existieran ciertas asociaciones entre modalidades, y que éstas –aún siendo débiles– resultaran de interés para el problema estudiado.

En la Tabla 6.1 el estudio de la relación entre estas dos variables podría traducirse en las siguientes preguntas: ¿es idéntica la distribución en ocupaciones de los hijos de padres de diferentes categorías ocupacionales (o hay una tendencia de los hijos a repetir las ocupaciones de sus padres)?<sup>7</sup> Y por otra parte: ¿para los distintos grupos de hijos difiere la repartición de sus padres en las diferentes ocupaciones?

En una tabla de dimensiones reducidas, es posible analizar estas cuestiones mediante el cálculo de porcentajes en fila o en columna, según cuál sea la pregunta a la que se busque responder, y comparando las filas o las columnas entre ellas o con respecto al marginal (cf. *supra*, capítulo 4).

Sabemos que si existiera *independencia* entre las ocupaciones de padres e hijos, debería observarse similitud en la repartición proporcional de las ocupaciones de los hijos (el perfil ocupacional de los hijos), para cada grupo de padres, pero también coincidencia en el perfil ocupacional de los padres para cada categoría ocupacional de los hijos. En efecto, si recordamos que los marginales constituyen promedios ponderados de las distribuciones en el interior de las filas (o de las columnas), entonces en caso de independencia, esos *perfiles* deberían coincidir con la distribución en el total de las observaciones (perfil marginal). Así el análisis de una tabla se reduce a analizar en qué medida las distribuciones proporcionales en fila (o en columna) se diferencian del *perfil medio*. Consideremos la comparación de los perfiles fila:

<sup>6</sup> Para la tabla 6.1, el coeficiente V de Cramer –derivado de fi cuadrado– arroja un valor de 0,38; en cambio, gamma, que saca partido del carácter ordinal de ambas variables nos da 0,59, indicando ya una asociación positiva.

<sup>7</sup> Para un análisis detallado de estas cuestiones –que escapan a nuestro objetivo en esta presentación– recomendamos la lectura del artículo de Jorrat (1997), en el que además se introduce toda una serie de técnicas específicamente adaptadas al estudio de problemas de movilidad.

Tabla 6.1.1: Tabla de perfiles-fila

| Ocupación del padre | Ocupación del hijo |              |                |                | Total |
|---------------------|--------------------|--------------|----------------|----------------|-------|
|                     | Manual Baja        | Manual Alta  | No manual Baja | No manual Alta |       |
| No manual Alta      | 0,022              | 0,054        | 0,151          | <b>0,774</b>   | 1,000 |
| No manual Baja      | 0,081              | 0,127        | <b>0,538</b>   | 0,253          | 1,000 |
| Manual Alta         | 0,155              | <b>0,392</b> | 0,338          | 0,115          | 1,000 |
| Manual Baja         | <b>0,363</b>       | 0,268        | 0,274          | 0,095          | 1,000 |
| Total               | 0,172              | 0,218        | 0,360          | 0,250          | 1,000 |

Fuente: Tabla 6.1

En las celdas de la diagonal, que corresponden a los hijos cuya categoría ocupacional coincide con la de su padre, las proporciones son siempre mayores a la proporción marginal correspondiente, lo que pone en evidencia que las ocupaciones de padres e hijos tienden a coincidir.<sup>8</sup> Sin embargo, no todos los hijos pertenecen a la misma categoría ocupacional de su padre.

Si en la tabla se comparan las filas entre sí, se observa que la distribución proporcional de las ocupaciones de los hijos (perfil) para cada una de las categorías de ocupación de los padres, difiere del perfil ocupacional de los hijos en la población.

De la misma manera, podemos comparar las columnas para determinar si entre los individuos con diferentes posiciones ocupacionales, se observan diferencias en la ocupación de sus padres (ver Tabla 6.1.2). Así, entre los hijos que tienen ocupaciones manuales hay mayor frecuencia de padres con ocupaciones también manuales. En general, los hijos con ocupaciones manuales presentan un perfil ocupacional de sus padres muy distinto al de los hijos con ocupaciones no manuales.

Tabla 6.1.2: Tabla de perfiles-columna

| Ocupación del padre | Ocupación del hijo |              |                |                | Total |
|---------------------|--------------------|--------------|----------------|----------------|-------|
|                     | Manual Baja        | Manual Alta  | No manual Baja | No manual Alta |       |
| No manual Alta      | 0,018              | 0,035        | 0,060          | <b>0,442</b>   | 0,143 |
| No manual Baja      | 0,161              | 0,197        | <b>0,506</b>   | 0,344          | 0,339 |
| Manual Alta         | 0,205              | <b>0,408</b> | 0,213          | 0,104          | 0,227 |
| Manual Baja         | <b>0,616</b>       | 0,359        | 0,221          | 0,110          | 0,291 |
| Total               | 1,000              | 1,000        | 1,000          | 1,000          | 1,000 |

Fuente: Tabla 6.1

Tratándose de una tabla de pequeño tamaño, el análisis mediante el procedimiento de comparar porcentajes no presenta mayores dificultades. Distinto es el caso cuando la tabla es de grandes dimensiones, situación en la que el AFC se revelará de la mayor utilidad. Lo que se logra mediante

<sup>8</sup> Como lo expresa Jorrat: «...la concentración de los casos en la diagonal principal es un indicador de la tendencia a la heredad y autoreclutamiento ocupacional» (1997: 105).

el AFC es transformar la tabla de datos original en un gráfico que permite *visualizar* las asociaciones existentes entre las modalidades.<sup>9</sup> En el gráfico producido por el AFC cada perfil estará representado por un punto, de manera tal que será posible comparar la similitud entre los perfiles-fila (o columna)<sup>10</sup> y también la diferencia de cada uno de éstos con respecto al perfil global, de un modo en todo análogo al procedimiento utilizado en el análisis numérico.

Si la relación entre las variables puede estudiarse como la diferencia entre los perfiles fila/columna con respecto a los marginales fila/columna (los perfiles medios), es posible representar gráficamente esa dispersión mostrando a cada uno de los perfiles como un punto en un espacio cuyo origen de coordenadas está dado por el perfil medio. De este modo la diferencia entre perfiles se traduce en una distancia entre los puntos correspondientes; asimismo, la diferencia entre un perfil-fila (o columna) con respecto al perfil medio se traduce en la distancia entre cada punto perfil y el origen del sistema de coordenadas.

Las coordenadas de cada punto en el espacio de representación quedan determinadas por sus frecuencias relativas en la tabla de puntos-perfiles. En nuestro ejemplo, los puntos-perfiles *ocupación del padre*, tendrán cuatro coordenadas y por lo tanto deberán representarse en un espacio de cuatro dimensiones. Así, las coordenadas para el punto correspondiente a los padres con ocupación manual alta serían: 0,155; 0,392; 0,338; y 0,115. De la misma manera se representarían los puntos *ocupación del hijo*. Con este criterio de representación, la proximidad de los puntos se deberá interpretar como similitud entre perfiles: así, en nuestro ejemplo, esperamos que el punto 'Padre no manual alta' se sitúe muy lejos del punto 'Padre manual baja'.

Para obtener un indicador de la diferencia entre perfiles o de la variabilidad de los puntos con respecto al perfil medio, se puede sumar las distancias de todos los puntos al centro de gravedad (origen del sistema), pero si se quiere que esa medida de la dispersión tome en cuenta la parte de la población representada en cada punto, entonces las distancias deben ponderarse por el peso relativo de cada categoría en la población.<sup>11</sup> Así se obtiene un coeficiente indicador de la dispersión de los perfiles con respecto al perfil medio y que por lo tanto mide la relación entre las variables. Se puede demostrar que esa cantidad conocida como *inercia de la nube de puntos*<sup>12</sup> es igual al  $\varphi^2$  de la tabla.

Empero, se plantea el problema de la imposibilidad de *visualizar* las distancias entre los perfiles (filas/columnas) en los espacios originales, dado que estamos tratando con espacios de más de tres dimensiones. De lo que se trata entonces, es de generar un nuevo espacio de representación que permita esa visualización, a partir de sucesivas representaciones gráficas planas, de modo tal que estas visualizaciones en el plano conserven lo más posible las distancias originales entre los puntos.<sup>13</sup>

Para construir ese nuevo espacio de representación, se busca un primer eje tal que las coordenadas de cada punto perfil en él traduzcan lo mejor posible la distancia de los individuos en el espacio original; luego, un segundo eje que dé cuenta de las diferencias que el primero no pudo resumir, y así sucesivamente.

De esta manera el nuevo espacio de representación –que supone un cambio de coordenadas– si bien sigue siendo *multidimensional*, permite establecer una jerarquía entre los ejes, ya que se construye de manera tal que el primer eje resume la mayor parte de la inercia y por lo tanto es la mejor aproximación a la nube de puntos original, que el segundo eje da cuenta de la mayor parte

<sup>9</sup> Según lo define Bry, «el análisis factorial es un principio geométrico que permite convertir automáticamente una gran tabla de datos en imágenes sintéticas que hacen visibles sus principales estructuras» (1995: 3).

<sup>10</sup> En realidad se trata de dos representaciones gráficas –una para cada tabla de perfiles (fila y columna)– que permitirán lecturas por comparación semejantes a las realizadas sobre las tablas.

<sup>11</sup> En realidad, se hace la suma ponderada de los *cuadrados* de las distancias de cada punto al centro de gravedad.

<sup>12</sup> «El término inercia (o más específicamente 'momento de inercia') es tomado de la mecánica (...) La inercia tiene una interpretación geométrica como una medida de la dispersión de los perfiles en el espacio multidimensional. A mayor inercia, mayor dispersión de los perfiles» (Greenacre; 1994: 12).

<sup>13</sup> Como señala J. Rovin: «De acuerdo al objetivo del análisis de correspondencias, las primeras dimensiones representan habitualmente la mayor parte de la variabilidad de la matriz de datos original. Si este es el caso, la posición de las proyecciones de los puntos-perfiles en un subespacio de menores dimensiones es una buena aproximación de la posición de los perfiles en el espacio original multidimensional» (en Greenacre y Blasius; 1994: 210).

posible de la variabilidad *residual*, etc.

Se descompone así la *inercia total* o *fi cuadrado de la tabla*, en una suma de variancias proyectadas sobre cada eje, con la particularidad de que esas variancias decrecen del primero al último eje. Con esta forma de construcción del nuevo espacio, el análisis de una nube de puntos en más de tres dimensiones, puede reducirse habitualmente a las dos o tres primeras, y por lo tanto a uno o dos planos, obteniendo una buena aproximación de la nube de puntos original.

Aunque es frecuente que un AFC se resuma en un único gráfico cuando se trata de presentar los resultados de una investigación, lo cierto es que el plano factorial en sí mismo es sólo una parte del producto del AFC, el que además comprende toda una serie de tablas y de medidas numéricas que resultan indispensables para la correcta interpretación de esos resultados.

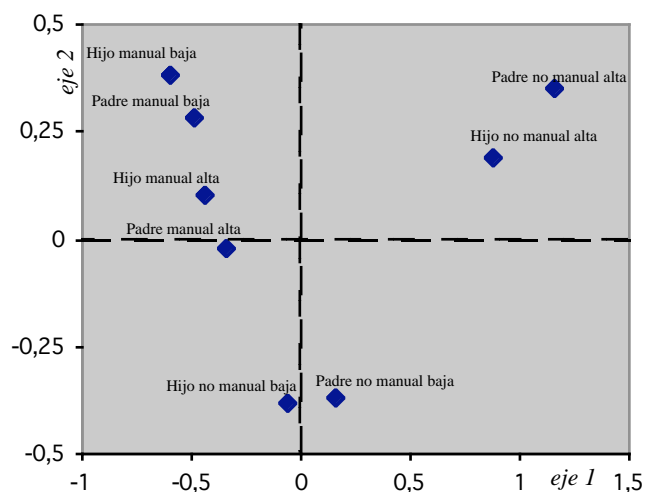
Figura 6.1: Histograma de los valores propios

| Factor<br>Nº | Valor<br>propio | %     | %<br>acum. |  |
|--------------|-----------------|-------|------------|--|
| 1            | .2978           | 71.24 | 71.24      |  |
| 2            | .0870           | 20.82 | 92.06      |  |
| 3            | 0.332           | 7.94  | 100.0      |  |

Suma de los valores propios = 0,4180

En la Figura 6.1 se presenta la descomposición de la inercia, o  $\varphi^2$  de la tabla, en tres ejes o factores. A cada eje le corresponde un valor propio que indica la cantidad de inercia o variabilidad de la que da cuenta. Así, el primer eje, con un valor propio, o *eigenvalue*,<sup>14</sup> de 0,2978, resume un 71,24% de la inercia total, mientras que el segundo eje da cuenta de un 20,82%. De este manera el plano conformado por los dos primeros factores concentra un 92,06% de la inercia, y por lo tanto constituye una buena aproximación gráfica de la tabla original. La suma de los valores propios - 0,4180- es igual a  $\varphi^2$ .

Figura 6.2: Posición de las ocupaciones de padres e hijos en el plano (1,2)



En una lectura inmediata del plano, teniendo en cuenta que el origen de las coordenadas representa el perfil medio, las categorías más diferentes de ese perfil medio aparecerán más

<sup>14</sup> Este es el término de uso habitual en la literatura anglosajona sobre análisis factorial, referido a la raíz característica, que es una propiedad matemática de una matriz (cf. Kim y Mueller; 1978).

alejadas, esto es, con coordenadas más altas –positivas o negativas-. Así en el primer eje se oponen las ocupaciones manuales y las no manuales altas tanto para los hijos como para los padres, modalidades que se alejan en direcciones opuestas respecto al centro, lo que traduce una gran diferencia en sus perfiles. Por otro lado, las categorías manuales altas y bajas aparecen con coordenadas próximas porque sus perfiles son semejantes. Así, en el primer factor -eje horizontal- los perfiles de padres e hijos se ordenan consistentemente desde las ocupaciones manuales bajas hasta las no manuales altas, mientras que el segundo eje aparece oponiendo las ocupaciones bajas no manuales al resto.

Como regla general:

*La proximidad entre dos puntos-perfiles -entre dos categorías de una misma variable- indica una similitud entre sus perfiles. En cambio, la proximidad entre elementos de diferentes variables indica asociación entre las modalidades.*

Así, por ejemplo, para la variable ‘ocupación del hijo’ la proximidad entre las categorías *Ocupación manual baja* y *Ocupación manual alta* se explica porque son semejantes en sus perfiles, ambos caracterizados por un excedente de padres con ocupación manual baja: mientras en el total de hijos hay un 29.1% con padres de ocupación manual baja, en esas categorías los porcentajes son de 61,6 y 35,9 respectivamente (cf. tabla 6.1.2).

En el primer eje las modalidades de ocupación no manual baja de padres e hijos se ubican prácticamente en el origen, lo que significa que en ese eje no están representadas sus diferencias respecto al perfil medio. Es el segundo eje el que da cuenta de las diferencias de esos perfiles con respecto a sus correspondientes marginales, y que marca la diferencia entre estos perfiles y los restantes.

**Tabla 6.2: Coordenadas, contribuciones y cosenos cuadrados de los puntos-perfiles o modalidades**

| Puntos-<br>perfiles | Dist. | COORDENADAS |      |      | CONTRIBUCIONES |      |      | COSENOS CUADRADOS |     |     |
|---------------------|-------|-------------|------|------|----------------|------|------|-------------------|-----|-----|
|                     |       | 1           | 2    | 3    | 1              | 2    | 3    | 1                 | 2   | 3   |
| HNoA                | .81   | .88         | .19  | .01  | 64.6           | 10.3 | .1   | .96               | .04 | .00 |
| HNoB                | .15   | -.06        | -.38 | -.07 | .4             | 58.4 | 5.2  | .02               | .95 | .03 |
| HMaA                | .30   | -.44        | .10  | .31  | 14.4           | 2.7  | 61.1 | .66               | .04 | .31 |
| HMaB                | .57   | -.60        | .38  | -.25 | 20.6           | 28.7 | 33.6 | .63               | .23 | .11 |
| PNoA                | 1.48  | 1.16        | .35  | .04  | 64.7           | 20.3 | .7   | .91               | .08 | .00 |
| PNoB                | .17   | .16         | -.37 | -.09 | 2.9            | 54.2 | 9.0  | .15               | .80 | .05 |
| PMaA                | .22   | -.34        | -.02 | .32  | 8.7            | .1   | 68.5 | .53               | .00 | .47 |
| PMaB                | .34   | -.49        | .28  | -.16 | 23.6           | 25.5 | 21.8 | .71               | .22 | .07 |

Fuente: Tabla 6.1.

Dentro del producto de un AFC se incluye generalmente una tabla en la que figuran distintos indicadores que permiten afinar la interpretación de los resultados visibles en el plano factorial. Así la *distancia al origen* (‘dist.’) está indicando para cada punto cuanto se aleja globalmente del perfil medio dentro del espacio original. Las *coordenadas* en los tres ejes indican la posición de cada punto en cada uno de los factores, y cuando asumen valores extremos es porque se trata de un perfil muy diferente al del perfil medio.

Las *contribuciones* nos informan para cada punto acerca de su aporte a la inercia resumida en cada eje, y por lo tanto sobre la intervención del punto en cuestión en la variabilidad explicada por ese eje. Este aporte está dependiendo tanto de la diferencia de ese perfil al marginal o perfil medio, como de su peso relativo en la población total. De esta manera, una contribución importante de un elemento puede devenir de una gran diferencia respecto al perfil medio, aunque su peso no sea tan grande. Es lo que ocurre con la modalidad ‘PNoA’,<sup>15</sup> la que más contribuye al primer factor (64,7%), porque su coordenada en el primer eje es muy alta. Este primer eje está *construido*

<sup>15</sup> Es corriente recurrir al uso de etiquetas abreviadas para representar a las modalidades tanto en los planos factoriales como en las tablas. Así, ‘HNoA’ deberá leerse: ‘hijos con ocupación no manual alta’, etc.

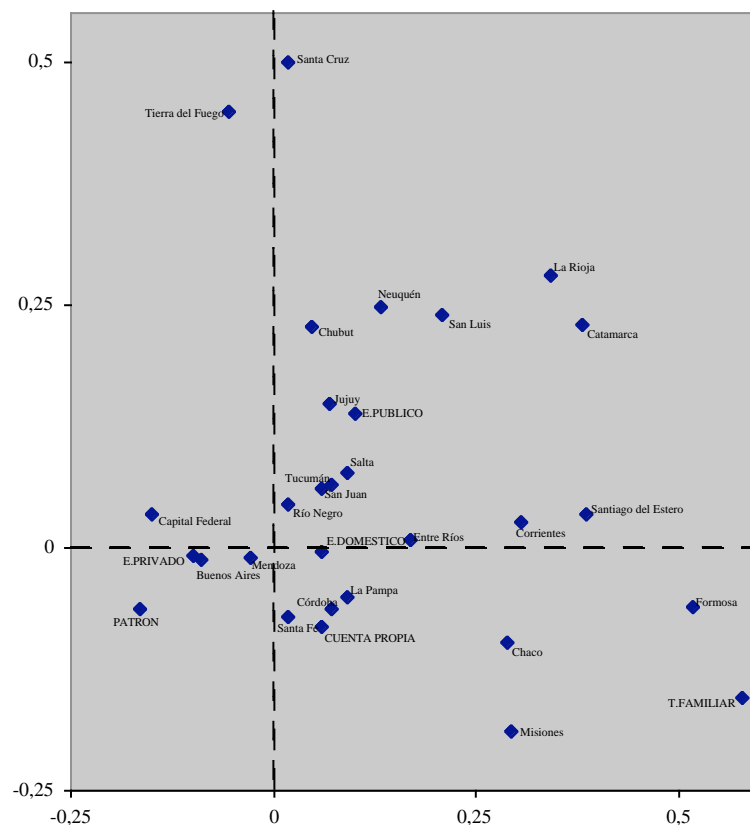
especialmente por ese punto y el punto ‘PMaB’ (contribución 23.6%).

La parte de la tabla bajo el título *cosenos cuadrados*, indica para cada punto su calidad de representación en todos los ejes –expresada en porcentajes–, lo que significa en qué medida el eje traduce la diferencia de ese punto-perfil con respecto al perfil medio. Por otro lado, como un punto estará totalmente representado en todos los ejes, la suma de esos índices a lo largo de todos los ejes da 100. Por ejemplo, la diferencia al perfil medio del punto HNoA está casi totalmente representada en el primer eje (96%), y totalmente representada en el primer plano (96% + 4% = 100%). En cambio en el primer eje la diferencia al perfil medio del punto HNoB no está bien traducida, puesto que su coseno cuadrado es de 2%; sin embargo sus características son traducidas por el segundo eje, donde su calidad de representación es del 95%.

El AFC es esencialmente una técnica de exploración de los datos que nos permite llegar a los mismos resultados que el análisis numérico convencional. Empero, sus ventajas son más evidentes cuando se trabaja sobre tablas de grandes dimensiones, al permitir una visualización de conjunto de la estructura de los datos.

El siguiente ejemplo basado en datos del Censo de 1980 nos permitirá ilustrar rápidamente este punto. Se trata de una tabla de contingencia en la que están dispuestas en filas las 24 provincias argentinas, y en columnas seis categorías ocupacionales: ‘patrón o socio’, ‘cuenta propia’, ‘trabajador familiar sin remuneración’, ‘empleado privado’, empleado público’ y ‘empleado en servicio doméstico’.

Figura 6.3: Argentina, 1980 - Provincias y posiciones ocupacionales



Inmediatamente salta a la vista como las distintas regiones del país se ubican en distintas porciones del plano: la región nordeste en la zona inferior derecha, el noroeste en la superior derecha, etc. Este ordenamiento es producto de las posiciones similares que ocupan en los dos primeros factores las provincias de una misma región. El primer eje opone un polo “moderno” representado por Capital Federal y la Provincia de Buenos Aires a otro “tradicional” o “subdesarrollado”. En lo que hace a las posiciones ocupacionales, este factor diferencia los distritos con mayor proporción de “empleados en el sector privado” –significando una mayor desarrollo de las relaciones de producción propiamente capitalistas– de aquellos en los que predominan formas familiares de producción evidenciadas en la mayor participación de la categoría “trabajador familiar sin remuneración” –Misiones, Chaco y Formosa.

En el segundo eje es muy buena la calidad de la representación de las provincias patagónicas: Santa Cruz (.97 de coseno cuadrado), Chubut (.90), Tierra del Fuego (.87), Neuquén (.77). Estas cuatro provincias en conjunto contribuyen en un 38% a ese segundo eje. En lo que hace a las posiciones ocupacionales el eje discrimina a los distritos con mayor peso del empleo público – región patagónica y noroeste- de aquellos donde hay predominio de los “cuenta propia”.

Algunas provincias están mal representadas en el plano (1,2). Así, La Pampa, Tucumán y Mendoza en realidad sólo aparecerían bien representadas en el eje 3, y Río Negro en el 4 (coseno cuadrado = .55). En estos casos el gráfico no está mostrando de modo adecuado cual es la distancia que mantienen respecto al perfil medio, y en consecuencia no se puede afirmar, por ejemplo, que Río Negro y San Juan presenten perfiles similares.

## 2. ALGUNAS APLICACIONES DEL ACM

Empero, la popularidad creciente del análisis de correspondencias no obedece tanto al AFC propiamente dicho, sino a la utilidad que presta el ACM en el análisis de encuestas.<sup>16</sup> Para explicar la naturaleza de esta técnica comenzaremos retomando el problema de construir un índice a partir de un conjunto de indicadores.

### 2.1 Construcción de un índice por medio del ACM

Los datos se refieren a la población de dos establecimientos de enseñanza técnica agropecuaria ubicados en una localidad de la zona centro de la Provincia de Misiones.<sup>17</sup> En esta población nos proponemos distinguir estratos que reflejen las diferencias existentes en el nivel de vida de sus miembros. Para ello, disponemos de una serie de indicadores referidos a los artefactos poseídos y a los servicios a los que tienen acceso las unidades domésticas a las que pertenecen los alumnos.<sup>18</sup> Mediante estos indicadores es posible construir un índice de equipamiento doméstico y acceso a servicios (IES).

Tabla 6.3: Valores, puntajes y modalidades de las variables seleccionadas para el IES

|  | <i>Variables</i>                | <i>Valores</i> | <i>Puntos<br/>Índice</i> | <i>Modalidades<br/>ACM</i> |
|--|---------------------------------|----------------|--------------------------|----------------------------|
|  | <i>Conexión de electricidad</i> | Tiene          | 1                        | LUZ1                       |
|  |                                 | No tiene       | 0                        | LUZ0                       |
|  | <i>Teléfono</i>                 | Tiene          | 1                        | TEL1                       |
|  |                                 | No tiene       | 0                        | TEL0                       |
|  | <i>TV Color</i>                 | Tiene          | 1                        | TVC1                       |
|  |                                 | No tiene       | 0                        | TVC0                       |
|  | <i>Videocasetera</i>            | Tiene          | 1                        | VID1                       |
|  |                                 | No tiene       | 0                        | VID0                       |
|  | <i>Canal de cable</i>           | Tiene          | 1                        | CAB1                       |
|  |                                 | No tiene       | 0                        | CAB0                       |
|  | <i>Heladera</i>                 | Tiene          | 1                        | HEL1                       |
|  |                                 | No tiene       | 0                        | HEL0                       |
|  | <i>Cocina a Gas</i>             | Tiene          | 1                        | GAS1                       |
|  |                                 | No tiene       | 0                        | GAS0                       |
|  | <i>Lavarropas</i>               | Tiene          | 1                        | LAV1                       |
|  |                                 | No tiene       | 0                        | LAV0                       |

<sup>16</sup> El ACM puede ser presentado como una simple generalización del AFC (cf. Crivisqui, 1993: 207).

<sup>17</sup> El relevamiento consistió en un censo de alumnos que tuvo lugar en 1996 y estuvo a cargo de estudiantes de la Licenciatura en Trabajo Social con la coordinación de María Rosa Fogeler.

<sup>18</sup> Muchos de estos indicadores son tomados en cuenta habitualmente para la construcción de índices de nivel económico-social, aunque para la elaboración de esta medida es usual combinarlos con otros indicadores referidos a las dimensiones ocupacional y cultural.



La combinación de estas ocho variables dicotómicas genera un espacio de propiedades que se podría diagramar en una tabla de  $2^8 = 256$  celdas. Empero, conforme a la lógica del procedimiento explicitado en el capítulo 5, el IES como índice sumatorio puede obtenerse por el simple expediente de asignar un puntaje a cada uno de los valores de las variables incluidas en la Tabla 6.3, y de sumar estos puntajes para cada unidad de análisis.<sup>19</sup> Así, una unidad de análisis, en vez de ser descrita por los valores que presenta en cada uno de los ocho indicadores puede ser caracterizada mediante un único valor numérico comprendido entre 0 y 8, que está resumiendo su posición en el espacio de propiedades. A su vez ese puntaje puede ser recodificado en cuatro categorías; cortando ‘por el lado de la variable’, damos lugar a una variable de nivel ordinal:

|                |            |
|----------------|------------|
| IES muy bajo   | 0-2 puntos |
| IES bajo       | 3-4 puntos |
| IES medio      | 5-6 puntos |
| IES medio alto | 7-8 puntos |

De este modo el puntaje IES constituye una medida que sintetiza las ocho dimensiones del espacio de propiedades original en una sola. Ahora bien, en cualquier técnica de análisis factorial el propósito es básicamente análogo: se trata siempre de resumir un conjunto de variables en un nuevo conjunto de variables más reducido.<sup>20</sup> El objetivo es reducir el espacio de propiedades generando variables-resumen, los denominados *factores*, que permitan poner en evidencia las diferencias entre las unidades de análisis de acuerdo a las combinaciones de características que presentan.<sup>21</sup>

En este sentido, el ACM permite visualizar similitudes y diferencias en los individuos -las unidades de análisis- a través de representaciones gráficas construidas a partir de un análisis de la estructura que surge de la interrelación entre las características observadas. Esta estructura llevará a identificar grupos de individuos que presentan características semejantes.

Para proceder a construir nuestro índice por medio del ACM partimos de las mismas variables con las *modalidades*<sup>22</sup> cuyas etiquetas identificadoras están incluidas en la última columna de la Tabla 6.3. El análisis de correspondencias permitirá generar gráficos en los que se localizará cada una de las unidades de análisis (en nuestro ejemplo, los 240 estudiantes censados) en el espacio de las modalidades, de manera que individuos con características semejantes aparecerán próximos los unos de los otros en ese espacio. Simultáneamente cada una de las modalidades se localizará en el espacio de los individuos, de modo tal que las modalidades asociadas presentarán coordenadas similares. Ambos espacios, el de los individuos y el de las modalidades, constituyen lecturas complementarias del mismo conjunto de datos. De este modo, es posible asignar a cada individuo sus coordenadas factoriales, las que constituirán variables numéricas que podrán ser utilizadas para la construcción de un índice.

Evidentemente el conjunto de las modalidades que intervienen de modo activo en la generación del espacio factorial ha de ser homogéneo, para que el análisis pueda tener un significado interpretable. En nuestro ejemplo, las dieciseis modalidades de las ocho variables del IES son tomadas como activas, y su homogeneidad deviene de que todas pueden considerarse como

<sup>19</sup> En la práctica bastará, en el programa en el que estemos operando, con generar una nueva columna de la matriz -el puntaje IES- mediante una simple fórmula que consista en sumar las columnas de los puntajes correspondientes a cada variable-indicador.

<sup>20</sup> Para una sugerente introducción, no técnica, al análisis factorial clásico, ver Gould (1997, cap. 6).

<sup>21</sup> Así, de querer llevar a cabo el procedimiento numérico de un modo más ortodoxo, tratándose de variables intervalares, se puede aplicar un análisis de componentes principales (ACP) sobre las variables-indicadores previamente estandarizadas (es decir, sometidas a una transformación lineal de modo que todas presenten una distribución semejante con una media aritmética de cero y una varianza igual a uno). En ese caso, ante una estructura claramente unidimensional, la coordenada de cada unidad de análisis en el primer factor se tomará como su puntaje en el índice. Cf., por ejemplo, Weller y Romney (1990: 26-31).

<sup>22</sup> Se utiliza el término *modalidad*, como sinónimo de ‘categoría’, para referirse a los valores de una variable cualitativa.

indicadores del nivel de vida.<sup>23</sup>

Si las variables están relacionadas unas con otras, la información que aporte cada una será redundante, y por lo tanto será posible reducir el espacio de propiedades. Así, por ejemplo, en una situación en la que todos los que tuvieran teléfono estuvieran conectados a un canal de cable, y viceversa, con una sola de estas dos variables se tendría la misma información que con las dos, llegándose a la misma clasificación de los individuos; y en ese caso la variabilidad entre los individuos estaría explicada por una sola variable.<sup>24</sup> Sin embargo, en la práctica la situación es bastante más compleja, ya que la variabilidad depende de más de dos variables, y además éstas no se encuentran perfectamente correlacionadas.

La llamada ‘tabla de Burt’ permite analizar las relaciones que mantienen todas las variables entre sí,<sup>25</sup> y consiste en una presentación conjunta de todas las tablas de contingencia generables a partir de un conjunto de variables tomadas de a dos. En cada línea como en cada columna de esta tabla figuran las dieciseis modalidades que componen las ocho variables consideradas, y en la diagonal se ve la relación de cada variable consigo misma (vale decir su distribución de frecuencias). La tabla de Burt resume la información que se quiere analizar y sobre ella operan los algoritmos de cálculo del ACM.<sup>26</sup>

**Tabla 6.4: Tabla de Burt para los ocho indicadores del IES**

|      |     |    |     |     |     |    |     |     |     |     |     |    |     |    |     |    |
|------|-----|----|-----|-----|-----|----|-----|-----|-----|-----|-----|----|-----|----|-----|----|
| LUZ1 | 206 | 0  | 35  | 171 | 155 | 51 | 58  | 148 | 43  | 163 | 190 | 16 | 156 | 50 | 175 | 31 |
| LUZ0 | 0   | 34 | 0   | 34  | 3   | 31 | 0   | 34  | 0   | 34  | 13  | 21 | 12  | 22 | 0   | 34 |
| TEL1 | 35  | 0  | 35  | 0   | 31  | 4  | 17  | 18  | 26  | 9   | 35  | 0  | 35  | 0  | 35  | 0  |
| TEL0 | 171 | 34 | 0   | 205 | 127 | 78 | 41  | 164 | 17  | 188 | 168 | 37 | 133 | 72 | 140 | 65 |
| TVC1 | 155 | 3  | 31  | 127 | 158 | 0  | 57  | 101 | 43  | 115 | 152 | 6  | 128 | 30 | 139 | 19 |
| TVC0 | 51  | 31 | 4   | 78  | 0   | 82 | 1   | 81  | 0   | 82  | 51  | 31 | 40  | 42 | 36  | 46 |
| VID1 | 58  | 0  | 17  | 41  | 57  | 1  | 58  | 0   | 21  | 37  | 56  | 2  | 50  | 8  | 55  | 3  |
| VID0 | 148 | 34 | 18  | 164 | 101 | 81 | 0   | 182 | 22  | 160 | 147 | 35 | 118 | 64 | 120 | 62 |
| CAB1 | 43  | 0  | 26  | 17  | 43  | 0  | 21  | 22  | 43  | 0   | 43  | 0  | 43  | 0  | 43  | 0  |
| CAB0 | 163 | 34 | 9   | 188 | 115 | 82 | 37  | 160 | 0   | 197 | 160 | 37 | 125 | 72 | 132 | 65 |
| HEL1 | 190 | 13 | 35  | 168 | 152 | 51 | 56  | 147 | 43  | 160 | 203 | 0  | 156 | 47 | 168 | 35 |
| HEL0 | 16  | 21 | 0   | 37  | 6   | 31 | 2   | 35  | 0   | 37  | 0   | 37 | 12  | 25 | 7   | 30 |
| GAS1 | 156 | 12 | 35  | 133 | 128 | 40 | 50  | 118 | 43  | 125 | 156 | 12 | 168 | 0  | 145 | 23 |
| GAS0 | 50  | 22 | 0   | 72  | 30  | 42 | 8   | 64  | 0   | 72  | 47  | 25 | 0   | 72 | 30  | 42 |
| LAV1 | 175 | 0  | 35  | 140 | 139 | 36 | 55  | 120 | 43  | 132 | 168 | 7  | 145 | 30 | 175 | 0  |
| LAV0 | 31  | 34 | 0   | 65  | 19  | 46 | 3   | 62  | 0   | 65  | 35  | 30 | 23  | 42 | 0   | 65 |
|      | 1   | 0  | 1   | 0   | 1   | 0  | 1   | 0   | 1   | 0   | 1   | 0  | 1   | 0  | 1   | 0  |
|      | LUZ |    | TEL |     | TVC |    | VID |     | CAB |     | HEL |    | GAS |    | LAV |    |

El estudio de la asociación entre las variables conduce a analizar la asociación entre sus modalidades, donde cada modalidad representa la clase de los individuos que la poseen. Cada línea de la tabla de Burt permite estudiar la relación entre una modalidad y el conjunto de las modalidades restantes. En un plano factorial dos modalidades estarán más próximas entre sí cuanto más se comporten en forma similar en relación a las restantes modalidades. Así, ‘CAB1’ y ‘TEL1’

<sup>23</sup> Nótese que la homogeneidad demandada es exactamente la misma que requiere el procedimiento numérico, en el cual los indicadores también deben ser seleccionados en base a su significado común. Del mismo modo que la edad o el sexo no pueden ser tomados como indicadores para la construcción de un IES, tampoco tendría sentido que sus modalidades intervinieran activamente en la generación de este espacio factorial.

<sup>24</sup> Según lo expresaba Lazarsfeld: «En el caso de la reducción funcional, algunas combinaciones son eliminadas en vista de las relaciones existentes entre las variables» (1993: 161)

<sup>25</sup> Obsérvese que, si se tratara de variables numéricas, como es el caso en un ACP, este papel de la tabla de Burt sería suplido simplemente por la matriz de correlaciones (cf. *supra*: capítulo 4).

<sup>26</sup> La tabla de Burt resulta generada automáticamente a partir de la matriz de datos original -también denominada ‘tabla de códigos condensados’- por cualquier *software* adecuado. La utilización práctica de estas técnicas, dado el volumen de los cálculos demandados, sólo ha sido posible con el advenimiento de la computación. En Francia, dos programas muy conocidos son el SPAD y el ADDAD. También estas técnicas han sido incorporadas -con éxito dispar- a los principales paquetes estadísticos: el BMDP en 1988, el SAS y el SPSS en 1990 (cf. Greenacre y Blasius, 1994).

representan clases de individuos que se comportan de manera similar respecto al resto de las modalidades. El análisis de correspondencias va a producir una representación geométrica de estas asociaciones.

El ACM puede realizarse igualmente sobre una tabla ‘lógica’ o ‘booleana’, también denominada ‘disyuntiva completa’.<sup>27</sup>

Tabla 6.5: Versión abreviada de la tabla disyuntiva completa para 240 estudiantes y 16 modalidades del IES

| UA    | Luz |    | Tel |     | TVC |    | Vid |     | Cab |     | Hel |    | Gas |    | Lav |    |   |
|-------|-----|----|-----|-----|-----|----|-----|-----|-----|-----|-----|----|-----|----|-----|----|---|
|       | sí  | no | sí  | no  | sí  | no | sí  | no  | sí  | no  | sí  | no | sí  | no | sí  | no |   |
| 1     | 0   | 1  | 0   | 1   | 0   | 1  | 0   | 1   | 0   | 1   | 0   | 1  | 0   | 1  | 0   | 1  | 8 |
| 2     | 1   | 0  | 0   | 1   | 1   | 0  | 0   | 1   | 0   | 1   | 1   | 0  | 1   | 0  | 0   | 1  | 8 |
| 3     | 1   | 0  | 0   | 1   | 0   | 1  | 0   | 1   | 0   | 1   | 0   | 1  | 0   | 1  | 0   | 1  | 8 |
| 4     | 1   | 0  | 0   | 1   | 1   | 0  | 0   | 1   | 0   | 1   | 1   | 0  | 0   | 1  | 0   | 1  | 8 |
| 5     | 1   | 0  | 0   | 1   | 1   | 0  | 0   | 1   | 0   | 1   | 1   | 0  | 0   | 1  | 0   | 1  | 8 |
| .     | .   | .  | .   | .   | .   | .  | .   | .   | .   | .   | .   | .  | .   | .  | .   | .  | 8 |
| i     | .   | .  | .   | .   | .   | .  | .   | .   | .   | .   | .   | .  | .   | .  | .   | .  | 8 |
| .     | .   | .  | .   | .   | .   | .  | .   | .   | .   | .   | .   | .  | .   | .  | .   | .  | 8 |
| 239   | 1   | 0  | 1   | 0   | 1   | 0  | 1   | 0   | 1   | 0   | 1   | 0  | 1   | 0  | 1   | 0  | 8 |
| 240   | 1   | 0  | 1   | 0   | 1   | 0  | 0   | 1   | 1   | 0   | 1   | 0  | 1   | 0  | 1   | 0  | 8 |
| Total | 206 | 34 | 35  | 205 | 158 | 82 | 58  | 182 | 43  | 197 | 203 | 37 | 168 | 72 | 175 | 65 |   |

En una tabla booleana cada renglón corresponde a una unidad de análisis. A diferencia de la matriz de datos, cada columna se corresponde con una modalidad (y no ya con una variable). Para cada modalidad su presencia-ausencia en un individuo es simbolizada respectivamente por un ‘1’ o por un ‘0’. Así la suma de todos los números de cada renglón es siempre la misma y resulta igual al número de las variables (ocho, en este ejemplo). En cambio la suma de cada columna nos informa acerca del número de unidades de análisis que presentan una modalidad en particular.

La tabla disyuntiva puede ser vista simultáneamente como un conjunto de columnas y como un conjunto de hileras. Así las unidades de análisis serán más o menos parecidas unas a otras de acuerdo al perfil que presenten: por ejemplo, el individuo 1 y el 239 presentan perfiles opuestos, son totalmente disímiles, mientras que el 4 y el 5 tienen perfiles idénticos.

Una representación geométrica adecuada de la variabilidad observada entre los individuos debe posicionarlos de tal modo que aparezcan más próximos en el espacio cuanto más semejantes sean en las modalidades que presentan (así, el 1 y el 239 deberán aparecer muy alejados el uno del otro, mientras que el 4 y el 5 deberán ocupar el mismo punto).

Ahora bien, es posible realizar el mismo análisis con las columnas. Tratándose de variables dicotómicas, como es el caso, los perfiles de sus modalidades son exactamente opuestos (hay una correlación negativa perfecta entre ambas modalidades). Igualmente en este caso es posible observar columnas que son semejantes, significando con ello que las modalidades en cuestión tienden a estar presentes -o ausentes- en los mismos individuos: es lo que ocurre en nuestro ejemplo con ‘Tel sí’ y ‘Cab sí’. De este modo, también las modalidades pueden ser consideradas en su perfil como puntos localizables en el espacio de los individuos.<sup>28</sup> Estos puntos aparecerán más próximos unos a otros cuanto más semejantes sean las modalidades entre sí, esto es, cuanto mayor

<sup>27</sup> «El origen de la terminología ‘tabla disyuntiva completa’ es el siguiente: el conjunto de los valores  $x_{jk}$  de un mismo individuo para las modalidades de una misma variable, comporta el valor 1 una vez (completa) y sólo una vez (disyuntiva)» (Escofier y Pagès; 1990: 48)

<sup>28</sup> El ACM hace jugar a las columnas y los renglones de la tabla el mismo papel, en el sentido de que no otorga preeminencia en el análisis a uno de estos elementos sobre el otro. Desde este punto de vista, en los términos en que se planteaba la estructura del dato en el capítulo 1, es indiferente afirmar «el auto es rojo» o «el rojo es auto» (cf. Fénelon; 1981). Es necesario pasar de una visión en la que el énfasis está puesto exclusivamente en las variables y en sus relaciones, para las cuáles las unidades de análisis funcionan apenas como soportes, a un punto de vista en el que las variables pueden jugar igualmente el rol de soportes con relación a los individuos.









sea el número de individuos en que coincidan en su presencia-ausencia.

En el caso del ACM, al igual que en el AFC, la técnica va a operar generando un espacio de representación que permita visualizar adecuadamente las distancias entre modalidades y entre individuos en los espacios originales. Ello supone generar nuevas variables -los *factores*- que sinteticen el conjunto de las variables originales; tales nuevas variables deberán mantener la más alta correlación posible con las variables originales.

El nuevo sistema de ejes -factores- se obtiene a partir de una operación matemática llamada 'diagonalización de una matriz', que se realiza sobre la tabla de Burt o la tabla lógica. Así, es posible localizar los individuos y las modalidades en el nuevo sistema de coordenadas, en el que el primer eje da cuenta de la mayor proporción de la inercia susceptible de ser resumida, el segundo eje explica la mayor parte posible de la variabilidad residual, etc.

Al igual que en caso del AFC, el ACM no se limita a la presentación de un plano factorial para dar cuenta de los resultados de una investigación, sino que incluye también todas las denominadas 'ayudas a la interpretación', esto es, las tablas en que se presentan medidas numéricas indispensables para una correcta interpretación de los datos.

**Figura 6.4: Histograma de los valores propios**

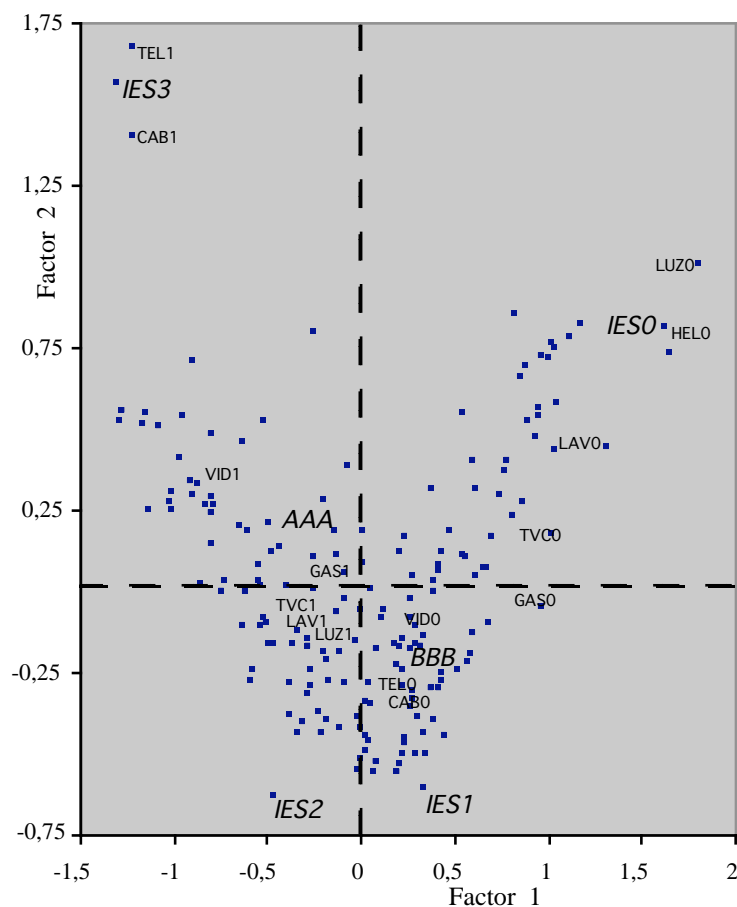
| Factor<br>Nº | Valor<br>propio | %     | %<br>acum. |   |
|--------------|-----------------|-------|------------|---|
| 1            | .4237           | 42.37 | 42.37      |   |
| 2            | .1656           | 16.56 | 58.94      |    |
| 3            | .1092           | 10.92 | 69.86      |    |
| 4            | .0860           | 8.60  | 78.46      |  |
| 5            | .0690           | 6.90  | 85.36      |  |
| 6            | .0636           | 6.36  | 91.72      |  |
| 7            | .0452           | 4.52  | 96.24      |  |
| 8            | .0376           | 3.76  | 100.00     |  |

Del análisis del histograma surge que el primer factor con un valor propio de .4237 concentra un 42,37 % de la inercia -o la variancia- original, en tanto que el segundo factor sólo representa un 16,56%. Esta baja pronunciada de la inercia que exhibe el segundo factor en comparación con el primero es compatible con la idea de una estructura básicamente unidimensional. Por otra parte, desde el tercer factor en adelante la disminución de la inercia de los sucesivos factores tiene lugar muy paulatinamente. En suma, en la combinación de los factores 1 y 2 que define el primer plano factorial se concentra un 58,94% de la inercia original, por lo que pragmáticamente podemos decidir limitar nuestro análisis a estos dos factores.<sup>29</sup>

En el ACM la inercia total es una cantidad fija independiente de cual sea la estructura de los datos.<sup>30</sup> Esta cantidad es igual a  $(k/p)-1$ , en la que  $k$  es el número total de modalidades y  $p$  el número de variables. En este caso particular, con 16 modalidades que corresponden a 8 variables, la inercia total es igual a 1, lo que ocurrirá siempre, cualquiera sea el número de variables, si éstas son todas dicotómicas. Asimismo, tratándose de variables dicotómicas el número máximo de factores, también es igual a  $p$  (en general el número de ejes es igual a  $k - p$ ).

<sup>29</sup> La decisión de cuántos factores hay que tomar en cuenta en un ACM no es algo que se pueda resumir en un conjunto de reglas preestablecidas.

<sup>30</sup> «La inercia total depende únicamente del número de variables y de modalidades y no de las relaciones entre las variables» (Lebart, Morineau, Piron; 1995: 120). Por ende, en ACM, el tamaño absoluto del valor propio por sí mismo no indica nada, puesto que puede corresponder a una inercia de cualquier magnitud.

**Figura 6.5: Ubicación de los individuos y de las modalidades activas y suplementarias en el plano factorial (1,2)**

Tanto los individuos como las modalidades aparecen representados como puntos en el plano factorial determinado por la combinación de los dos primeros factores (en realidad, los puntos están localizados en el espacio de ocho factores, y por ende lo que se observa en el plano 1,2 es simplemente la mejor proyección posible de aquella localización en un plano).

En este plano factorial, el primer factor está representado en el eje horizontal y el segundo en la ordenada. Los puntos representan la proyección de los 240 alumnos en este espacio, mientras que las modalidades, también simbolizadas por puntos, están identificadas cada una por su etiqueta. Si dos individuos aparecen con coordenadas similares en ambos ejes es porque presentan globalmente las mismas características. Si dos modalidades aparecen próximas en el plano, es por encontrarse asociadas en general con las mismas modalidades de las restantes variables (como lo indica la tabla de Burt) y porque se encuentran presentes globalmente en el mismo conjunto de individuos (lo que se hace visible en la tabla lógica).

Al considerar la proyección en el eje 1 horizontal de las dieciseis modalidades de la Tabla 6.1 que corresponden a las ocho variables consideradas para el índice, se observa que éstas se ordenan de izquierda a derecha desde 'TEL1' hasta 'LUZ0' y 'HELO'. En los extremos se ubican aquellas modalidades (clases de individuos) más diferentes del total de la población (perfil medio); así, este eje opone los que poseen los bienes y servicios menos elementales (cable, teléfono y video), al grupo de los que no acceden a los equipamientos y servicios básicos (luz y heladera),<sup>31</sup> con lo que la coordenada en el primer factor resulta efectivamente interpretable en términos del nivel de equipamiento doméstico de las familias de los estudiantes. Por su parte, el segundo eje da cuenta

<sup>31</sup> El que los individuos más ricos aparezcan en el sector izquierdo del plano, es un efecto producido arbitrariamente por el algoritmo actuante en la reducción del espacio original: el signo -negativo o positivo- de los valores en un eje carece por completo de importancia, y en este caso habrá que interpretar que los valores positivos más altos corresponden a niveles mayores de pobreza.

de una diferencia residual entre los individuos, separando las situaciones extremas en uno u otro sentido de las intermedias.

De este modo se distinguen distintas zonas en el plano: en el cuadrante superior izquierdo están los individuos en situación más favorable; a medida que nos desplazamos hacia abajo y hacia la derecha aparecen individuos en situaciones menos favorables; finalmente, ya del lado derecho se ubican individuos cada vez más pobres cuanto más elevada es su coordenada en el segundo eje.<sup>32</sup>

En todas estas técnicas de análisis multivariado, tanto los individuos como las modalidades pueden jugar dos papeles: activo e ilustrativo. Los individuos y las modalidades **activos** son los que intervienen en la generación de los factores. En cambio, las modalidades e individuos **ilustrativos** (a veces también denominados ‘suplementarios’) no contribuyen a la construcción del espacio factorial y simplemente resultan proyectados en éste.<sup>33</sup> La proyección de estas características suplementarias o ilustrativas tiene por objetivo estudiar su relación con los factores.

Al proyectar las modalidades de la variable ‘Escuela’ se observa que mientras que la Escuela A –‘AAA’- está del lado izquierdo del diagrama, la B se localiza en la zona correspondiente a los individuos con menor IES. Esto parece indicar una diferencia en la extracción social del alumnado de estas instituciones: en promedio los alumnos de la Escuela A tienden a presentar un más alto nivel de equipamientos y servicios que los de la B.

En base a las coordenadas de los individuos en el primer factor hemos generado mediante un procedimiento de clasificación ascendente jerárquica una variable ordinal, distinguiendo cuatro niveles de IES. Al proyectar también en el plano las modalidades de esta variable observamos que se despliegan siguiendo la forma de una parábola o de una herradura. Mientras el primer eje ordena las modalidades desde la más baja (‘IES0’) a la más alta (‘IES3’), el segundo factor opone las modalidades extremas ubicadas en la zona superior, a las no-extremas (‘medio’ y ‘bajo’) del IES en la parte inferior,<sup>34</sup> como era esperable dado que estas modalidades se construyeron en base el ACM anteriormente realizado.

Finalmente, podemos preguntarnos qué relación existe entre este IES generado mediante el ACM y el que habíamos construido antes por el procedimiento sumatorio más habitual.

**Tabla 6.6: Relación entre el IESacm y el IES sumatorio**

| IESacm | IES sumatorio |      |       |            | Total |
|--------|---------------|------|-------|------------|-------|
|        | Muy bajo      | Bajo | Medio | Medio-alto |       |
| IES3   | 0             | 0    | 0     | 39         | 39    |
| IES2   | 0             | 0    | 93    | 0          | 93    |
| IES1   | 0             | 61   | 0     | 0          | 61    |
| IES0   | 47            | 0    | 0     | 0          | 47    |
| Total  | 47            | 61   | 93    | 39         | 240   |

Como se puede observar, hay una correlación positiva perfecta ( $V = 1$ ) entre los resultados de las dos técnicas.<sup>35</sup> ¡Por algo el procedimiento sumatorio convencional -tan sencillo como robusto- ha sido utilizado con éxito durante tanto tiempo!

Es importante aclarar cómo ambas formas de reducción –la numérica y la factorial (o *funcional*)- resultan conceptualmente diferenciables. Según Lazarsfeld, «en la reducción funcional se da una *relación real* entre dos de los atributos que reduce el número de combinaciones» (1993: 160; *itálicas nuestras*), y se trabaja a partir de esas relaciones comprobadas. En cambio la reducción numérica, cuando su propósito es clasificar a las unidades de análisis en una escala

<sup>32</sup> Así, es posible entender que lo que está representado en el plano factorial es el *espacio social* en el sentido de Bourdieu: «las distancias espaciales sobre el papel equivalen a las distancias sociales» (Bourdieu; 1997: 30).

<sup>33</sup> Se calcula las coordenadas de los individuos ilustrativos como media ponderada de las coordenadas de las modalidades que presentan, y análogamente las coordenadas de las modalidades suplementarias como media ponderada de las coordenadas de los individuos que las presentan.

<sup>34</sup> Se trata de una configuración típica a propósito de la cual se suele hablar de “efecto Guttman”. Según explican Escofier y Pagès, «Cuando un efecto de escala es muy fuerte, influencia varios ejes de acuerdo a la siguiente propiedad: el factor de rango  $s$  es una función polinómica de grado  $s$  del primero» (1990: 240).

<sup>35</sup> La correlación entre los puntajes del IES sumatorio, tomados como una variable intervalar, y las coordenadas en el Factor 1 arroja  $r = 0,996$ .

unidimensional (como es el caso, por ejemplo, en un índice de nivel económico-social), se basa en una suerte de apuesta implícita acerca de la existencia de esas relaciones.

## 2.2 Construcción de una tipología

Para finalizar presentamos un ejemplo más en el que exponemos los resultados de la aplicación de un ACM en combinación con el uso de un procedimiento de clasificación para la construcción de una tipología.

Entre los objetivos de la investigación realizada se planteaba determinar si los dos establecimientos de enseñanza secundaria diferían en cuanto al origen social de su alumnado. Teniendo en cuenta la información relevada, nos propusimos generar una tipología de los estudiantes dividiéndolos en clases que resultaran ser lo más homogéneas posibles internamente a la vez que lo más heterogéneas unas con respecto a las otras. Lo que esperábamos era que dichos tipos se diferenciaron unos de otros en sus conductas, expectativas y opiniones.<sup>36</sup> Las variables que consideramos significativas a este efecto fueron las siguientes:

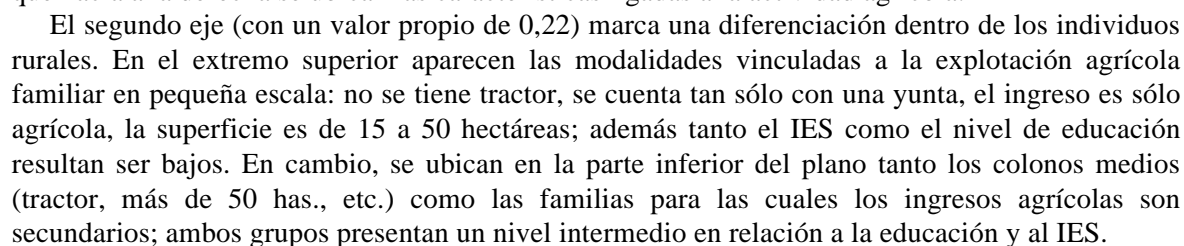
**Tabla 6.7: Variables y modalidades para generar tipos sociales**

| Variables                 | Valores                   | Etiquetas ACM |
|---------------------------|---------------------------|---------------|
| Residencia habitual       | Urbana                    | URB           |
|                           | Rural                     | RUR           |
| Tenencia de chacra        | Sí                        | CHSI          |
|                           | No                        | CHNO          |
| Tenencia de tractor       | Sí                        | TRSI          |
|                           | No                        | TRNO          |
|                           | No corresponde            | TRNC          |
| Nº de yuntas de bueyes    | No tiene                  | YUN0          |
|                           | Una yunta                 | YUN1          |
|                           | Dos o más yuntas          | YUN2          |
|                           | No corresponde            | YUNC          |
|                           | Sin dato                  | YU??          |
| Nº de hectáreas           | Sin chacra                | HTNC          |
|                           | < de 15 has.              | HCT1          |
|                           | 15-<30 has.               | HCT2          |
|                           | 30-<50                    | HCT3          |
|                           | 50 o más                  | HCT4          |
|                           | Sin dato                  | HC??          |
| Fuente de ingreso         | Sólo agrícola             | \$SAg         |
|                           | Agrícola principal        | \$AgP         |
|                           | Agrícola Secundario       | \$AgS         |
|                           | No agrícola               | \$NoA         |
| Nivel de educación        | No completó Primaria      | EDU0          |
|                           | Primaria Completa         | EDU1          |
|                           | Secundaria completa (y +) | EDU2          |
| I. Equipamiento doméstico | Muy bajo                  | IES0          |
|                           | Bajo                      | IES1          |
|                           | Medio                     | IES2          |
|                           | Medio-alto                | IES3          |

Nuevamente se trata de realizar la reducción de un espacio de propiedades mediante un ACM, con la salvedad que ya no se presume que se trata de una estructura unidimensional como era el caso con el IES. Empero, hay todavía una homogeneidad de significado: todas estas variables se refieren a características que pueden ser consideradas como pertinentes al proponernos generar

<sup>36</sup> En términos de Bourdieu (1980), se puede pensar que cada una de estas clases así construidas tenderá a agrupar a individuos con similares *habitus*.

**Figura 6.6: Origen social de los estudiantes**

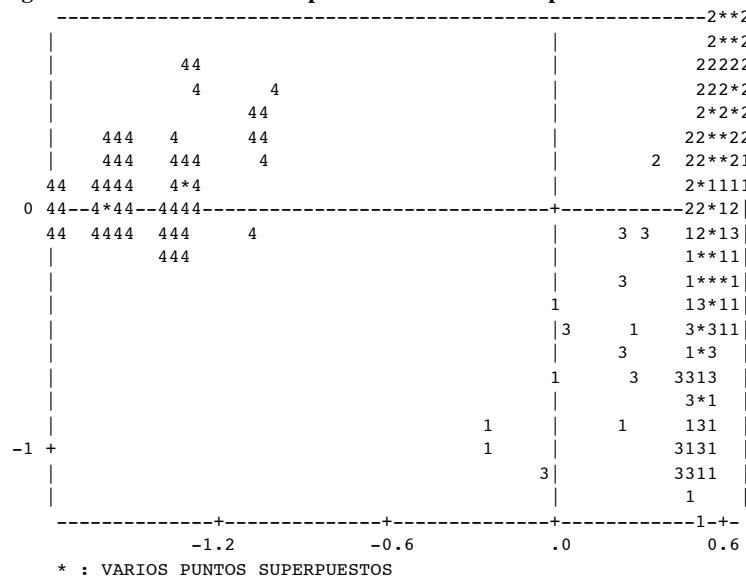


<sup>38</sup> Como la asociación entre estas modalidades es perfecta, sus coordenadas son idénticas, con lo que los puntos aparecen exactamente superpuestos.



Una clasificación automática en base a los factores producidos por el ACM<sup>39</sup> arroja los resultados que se observan en la figura 6.7, en la que cada individuo en vez de estar representado por un punto, lo está por un número que identifica la clase a la que pertenece.

**Figura 6.7: Localización en el plano de los individuos pertenecientes a cada tipo**



En el sector izquierdo, la clase 4 abarca a un 27% de las familias; es la de los “*no agricultores*”<sup>40</sup>: los ingresos son exclusivamente no agrícolas, y el 89 % vive en la ciudad. Una mitad tiene el máximo nivel de IES (“medio-alto”), y además esta categoría concentra un 77% de los jefes de familia con instrucción superior a la primaria.

En las tres clases restantes el 100% de las familias tienen chacra. Así, la clase 1 es la de los “*colonos medios*”, que son un 25% del total. Un 85% de ellos cuenta con tractor, un 51% tiene al menos dos yuntas; y un 59% opera explotaciones de más de 50 hectáreas. El ingreso es sólo agrícola para un 69% de ellos. Por otra parte, 58% alcanza un nivel de IES medio.

La Clase 3 (15% de la población) agrupa a los “*agricultores part-time*”. Aunque entre éstos un 56% tienen menos de 15 hectáreas, todos los que tienen esta superficie caen en esta clase. En un 69% se trata de familias para las cuales la actividad agrícola provee sólo de un ingreso secundario. El 92 % carece de tractor y el 47% no tiene ni una yunta. El 66% de los jefes cuenta con primaria completa.

Por último, en el sector superior derecho se ubican los miembros de la clase 2, la que totaliza un 34% de los casos. Se trata de los “*colonos pobres*”: en un 99% son rurales, en un 93% el ingreso es solamente agrícola, y el 98% no tiene tractor; en esta clase un 77% de los jefes no completó la escuela primaria.

<sup>39</sup> Este modo de proceder que combina los resultados de un análisis factorial con técnicas de clasificación automática es típico de la escuela francesa. El método de clasificación ascendente jerárquica compara a los individuos a través de sus coordenadas factoriales y los agrupa de tal manera que las clases sean lo más homogéneas *dentro* de ellas y lo más heterogéneas *entre* ellas. En realidad lo que se obtiene es una jerarquía de particiones, a partir de cuyo análisis se decide el número de clases que interesan. Estos métodos se utilizan frecuentemente como complemento del análisis factorial ya que permiten distinguir grupos de individuos similares más allá de los primeros factores que se consideran en el análisis factorial. Así, los agrupamientos resultan de comparaciones sobre todas las dimensiones en que se descompuso la inercia, o bien sobre una aproximación de ese espacio total, dejando de lado los últimos ejes que en general dan cuenta de variaciones aleatorias.

<sup>40</sup> Según recomienda Lebart, es conveniente que las etiquetas elegidas sean lo más neutras posibles (1986:7); de hecho, las clases inevitablemente incluirán algunos individuos para los cuales las etiquetas no resultarán del todo adaptadas.

**Tabla 6.8: Extracción social de los estudiantes en dos establecimientos de enseñanza agropecuaria (%)**

|           | Colonos medios | Colonos pobres | Agricultores part-time | No agricultores | Total (100%) |
|-----------|----------------|----------------|------------------------|-----------------|--------------|
| Escuela A | 25             | 16             | 13                     | 46              | (123)        |
| Escuela B | 24             | 52             | 17                     | 7               | (117)        |

$$\chi^2 = 57,00; p < 0,0001; V = 0,44.$$

Fuente: Encuesta a establecimientos de enseñanza agropecuaria, 1996.

Basándonos en la tipología que hemos construido podemos sintetizar sin dificultad las diferencias entre las dos escuelas en lo que hace a la extracción social de sus estudiantes. De este modo el establecimiento A exhibe casi una mitad (46%) de sus estudiantes que son hijos de no agricultores, mientras que la escuela B registra 52% de alumnos procedentes de familias de colonos pobres, en tanto que en ambos institutos hay la misma proporción -una cuarta parte de los estudiantes- que proviene de hogares de colonos medios.

Este ha sido solamente un último ejemplo en este capítulo donde apenas nos hemos propuesto introducir algunas aplicaciones del ACM, a costa de una simplificación que muchos podrán juzgar excesiva. En la presentación que hemos hecho del tema nos hemos preocupado por enfatizar la continuidad entre este nuevo enfoque del análisis de encuestas y el más tradicional: se trata de un conjunto de herramientas alternativo para lograr resultados que no deberían ser sustancialmente distintos de los alcanzados por otros vías. En definitiva, el uso provechoso que se pueda hacer de estas técnicas va a depender tanto de la característica del problema específico que se aborde, como del estilo de trabajo con el cual el investigador se encuentre más a gusto. Aspectos ambos que se encuentran obviamente ligados entre sí, además de relacionarse, claro está, con las preferencias teóricas y las posturas epistemológicas. Por nuestra parte, si supiéramos que este capítulo ha servido al menos para despertar el interés del lector, estimulándolo a perfeccionarse en esta temática,<sup>41</sup> nos consideraríamos plenamente satisfechos.

<sup>41</sup> La bibliografía es frondosa, y contantemente se van agregando nuevos aportes. Algunos textos en inglés, todavía no tan numerosos, son los de Blasius y Greenacre (1998), Greenacre y Blasius (1994), Weller y Romney (1990), y Phillips (1998). Pero las referencias más importantes todavía siguen siendo producidas en la lengua francesa. Así, existe una revista especializada -*Cahiers de l'analyse des données*- donde se han publicado regularmente artículos a cargo de Benzécri y de sus discípulos. El poco convencional texto de Fénelon (1981) contiene sin duda muchos hallazgos de gran valor pedagógico para el neófito. En cuanto a la obra de Escofier y Pagès (1990), puede considerarse como un manual clásico de introducción a las técnicas factoriales francesas. El libro de Lebart, Morineau y Piron (1995) es sumamente recomendable, ya que logra una presentación elegante y concisa de todas las técnicas, y a la vez brinda muy útiles indicaciones acerca de su complementariedad con la versión de la estadística predominante en los países anglosajones. En castellano, aparte de la traducción del texto de Escofier y Pagès, se puede encontrar una presentación rigurosa y accesible del análisis de correspondencias simples y múltiples en el libro de Crivisqui (1993).