EL ANÁLISIS DE COMPONENTES PRINCIPALES: APLICACIÓN AL ANÁLISIS DE DATOS SECUNDARIOS

Carlos Lozares Colina
Pedro López Roldán
Departament de Sociologia
Universitat Autònoma de Barcelona

Resumen

El interés del artículo se centra en presentar el Análisis de Componentes Principales como ejemplo práctico de aplicación, dejando de lado las consideraciones algebraicas y estadísticas que no sean imprescindibles. Se insiste en los procesos de elección del campo de las variables, de los criterios de elección del número de componentes y de la interpretación de los ejes. El Análisis de Componentes Principales se aplica como un primer paso de la construcción de una muestra estratificada, en particular con la utilización de los datos censales. En este sentido se trata de un análisis de datos que va más allá de lo que las estadísticas oficiales procuran.

Resum

L'interès de l'article està orientat a presentar l'Anàlisi de Components Principals com a exemple pràctic d'aplicació, deixant de banda les no indispensables consideracions algebraïques i estadístiques. En efecte, per un costat, s'insisteix en els processos d'elecció del camp de les variables, dels criteris d'elecció del nombre de components i de la interpretació dels eixos; per l'altre, l'Anàlisi de Components Principals s'aplica com a primer pas de la construcció d'una mostra estratificada, en particular amb la utilització de les dades censals.

Abstract

The purpose of this article is to present the Principal Components Analisys, as a practical example of its aplication, leaving aside the always necessary algebraic and statistical considerations. Indeed, on one hand, we emphasize on the choice processes of the number of component and axis interpretations. On the other hand, the Principal Components Analisys is applied as the first step towards the construction of an stratified sample, particulary using census data.

INTRODUCCIÓN

El objetivo de este artículo no se centra tanto en desarrollar el método factorial de Análisis de Componentes Principales (ACP) desde una perspectiva exclusivamente teórica, algebraica y/o estadística, sino sobre todo en aplicarlo a un análisis concreto a fin de ver más cercanamente la eficacia, las exigencias y los límites de tal método.

Por otro lado, lo que ahora presentamos es una parte de otra investigación más amplia realizada en el marco de la *Enquesta Metropolitana*[1], esto es, la construcción de una muestra estratificada por métodos multivariables para dicha encuesta. En el diseño de la muestra, el ACP se ha concebido como instrumento parcial pero indispensable y con entidad propia: siendo una etapa intermedia en la construcción de la muestra posee un interés en sí mismo que va más allá de los objetivos propios de toda muestra para convertirse en un subproducto que por los criterios, las decisiones y las interpretaciones que se manejan, calificamos de específicamente sociológico. Así mismo, el caso presentado es un ejemplo de utilización de datos secundarios para su análisis descriptivo y de dimensionalización de la realidad social investigada, como en nuestro caso son los del *Padró d'Habitants* del año 1986.

La metodología propuesta en la construcción de la muestra estratificada está basada en la complementariedad entre las técnicas de dimensionalización y de clasificación. En este contexto los momentos básicos del proceso de construcción de la muestra se concretan, elegido el universo poblacional y el campo de las variables/criterio, primero, en la factorialización por medio del ACP, segundo, la estratificación en sentido estricto a partir de la realidad dimensionalizada y mediante la agrupación de la población en estratos homogéneos resultantes de la aplicación del análisis de *Cluster*, por último, se aplican los necesarios pasos de determinación del tamaño muestral y de afijación de la muestra.

En estas páginas pues desarrollaremos con mayor profundidad la etapa destinada exclusivamente a la utilización del ACP para la dimensionalización del campo de variables empleado, señalando que este estudio es continua-

1. La *Enquesta Metropolitana* es una investigación que fue proyectada en el año 1984 para analizar las actividades y formas de vida de la población del Área Metropolitana de Barcelona (un total de 27 municipios) ante la ausencia de datos sistemáticos de este contenido y para convertirla en un instrumento periódico de recogida de información. Encargado por la antigua Corporació Metropolitana de Barcelona, la primera edición de la misma (1986) abarcó el periodo 1985-89 con diversos informes publicados. En la actualidad, con la financiación de la Mancomunitat de Municipis de l'Àrea Metropolitana a través del Institut d'Estudis Metropolitans de Barcelona y en colaboración con la Diputació, el INEM y la UAB, se inicia la segunda edición extendiéndose a la totalidad de la Regió I en la división territorial de Catalunya que abarca un total de 250 municipios.

ción, como allí se anunció, de otro trabajo aparecido en esta misma revista con el título de *La tipología en Sociología, mas allá de la simple taxonomía: conceptualización y cálculo*[2], con vistas a la estratificación y tipificación del universo de los individuos. Por ello el presente artículo estaría incompleto si no se siguiera de un tercero, que aparecerá en próximos números, sobre dicha estratificación.

Después de una introducción sobre la definición, objetivos y modelo de análisis inherente al método (1), se tratarán sucesivamente: la Elección del campo de las variables (2) y el Proceso detallado del ACP (3).

1. DEFINICIÓN, OBJETIVOS Y MODELO DE ANÁLISIS EN ACP

1.1. Definición

El ACP es un método algebraico/estadístico que trata de sintetizar y dar una estructura a la información contenida en una matriz de datos. El procedimiento consiste en homologar dicha matriz a un espacio vectorial tratando de encontrar en él unos ejes o dimensiones que, siendo combinación lineal de las variables introducidas,

— no pierdan la información inicial al conservar la varianza total,
— no tengan correlación entre ellos, esto es, sean linealmente independientes, lo que asegura la estructuración de las variables iniciales,
— tengan una importancia diferencial y conocida en la explicación de la varianza total.

Realizadas estas exigencias, el objetivo básico consiste en reducir el número de variables introducidas. Para ello se toman como nuevas variables los ejes o componentes hallados, eligiendo un número y peso de los mismos suficiente para que la pérdida de varianza total sea sólo la conveniente, llenando así las finalidades del método, esto es, las de simplificar, reducir y estructurar la información inicial.

Para el sociólogo el análisis ni empieza por el tratamiento descrito ni se acaba en él. Previamente habrá de definir las variables que entran en juego en el análisis, validarlas y seguramente habrá propuesto un modelo, aunque sea elemental, de hipótesis, descriptivas o explicativas, que dé cuenta de la problemática considerada. Simultánea y posteriormente al proceso habrá de

2. Lozares, C. «La tipología en Sociología, mas allá de la simple taxonomía: conceptualización y cálculo» en *Papers. Revista de Sociología*, 34, Edicions 62, UAB, Barcelona, 1990, pp. 139-164.

tener criterios estadísticos pero sobre todo sociológicos, para elegir el número de ejes y fundamentalmente para dar identidad a los mismos, así como para interpretar y proponer una estructura del conjunto de las variables y de las relaciones o agrupaciones entre ellas.

1.2. Objetivos

A este método se acostumbra a clasificarlo entre los descriptivos. Nuestro punto de vista no es exactamente este. El método, y su técnica correspondiente, permite ir más allá y llenar otras finalidades que las que se derivan de la naturaleza empirista y de rango menor y subsidiario que, como nivel de análisis, es habitualmente atribuido a este método. Con todo somos de la opinión de que esta técnica es un valioso complemento a otras multivariables.

Veamos pues alguno de los efectos de dicho análisis, no todos ellos puestos en evidencia en este artículo aunque han sido llenados en otras investigaciones paralelas.

La utilización del método llena, sobre todo si es previo a la estratificación que conlleva la construcción de una muestra como es el caso, diversas finalidades de interés:

a) Reducir el espacio tal y como se ha anunciado.

De las variables introducidas en el método se retendrán unas combinaciones lineales de las mismas salvando en ellas las puntuaciones de los individuos (secciones censales en nuestro caso).

b) Eliminación de la información redundante.

De las variables introducidas en este análisis se disminuye la repercusión que en el cálculo de muestra tiene la redundancia informativa pues no se tiene en cuenta la acumulación de covarianza de las variables primitivas entre sí.

c) Captar en los nuevos ejes parte de la varianza total.

El efecto más decisivo consiste en que siendo dichos ejes linealmente independientes y los que más acumulan la varianza del campo de variables introducido y en el grado decidido por el sociólogo, se asegura doblemente:

— la incorrelación entre estas dimensiones, lo que era una condición decisiva de las establecidas en el apartado primero,
— el máximo poder discriminante establecido además jerárquicamente según dichos ejes: la estratificación saldrá así beneficiada.

d) Los ejes factoriales proveen también de una triple ventaja:

— por un lado estructuran la realidad introducida, que es la única existente en los datos disponibles.
— por otro, que aunque las variables introducidas en el análisis factorial estén elegidas como identificativas o explicativas de las que se utilicen en la encuesta, el método puede atestarnos sobre la validez de la elección.
— además, y *a posteriori*, estos ejes guardarán un poder extraordinario a la hora de validar precisamente variables importantes de la misma encuesta puesto que, en el caso examinado, se trata de variables censales.

El ACP aparece como un complemento necesario a otras técnicas de categorización de individuos ya que su lógica precisamente es la de agrupar variables.

1.3. El modelo subyacente

El modelo matemático que está en la base del método descansa sobre tres pivotes.

1) El primero consiste en el tratamiento de la matriz de datos (E_i. x $X._j$) como si fuera un espacio vectorial euclídeo de p variables, tantas como variables $X._j$ (j=1,2,...p) existan. En dicho espacio los individuos, entidades o unidades son posiciones puntuales E_i. (i=1,2,...n). Dichos puntos forman una nube N^n_p. Este conjunto de variables engendra pues un espacio vectorial del que, a partir de una métrica euclídea, se extrae su base o factores. Así pues las operaciones y sus propiedades definidas en el espacio vectorial son de aplicación en el presupuesto adoptado.

2) El método de maximización de la varianza, como condición para determinar cada uno de los ejes, se realiza haciendo que cada uno, gradual y progresivamente, vaya conteniendo o proyectando sobre él el máximo de la *inercia* de todos los puntos/posición una vez definida la métrica.

3) El modelo básico de dependencia en la definición de los ejes, en las condiciones dadas, es un modelo lineal, fundamental a la hora misma de hallar la base del sistema vectorial como dimensiones linealmente independientes. A su vez en la expresión de las variables primitivas en función de los ejes a través de la matriz de saturación se reproduce la dependencia lineal tal,

$$X=YA'$$

donde X, matriz (nxp) de datos inicial de Unidades/Variables.
Y, matriz (nxp ó m si hay reducción con m<p) de Unidades/Ejes.
A, matriz (pxp ó m si hay reducción con m<p) de Variables/Ejes que evidencia el modelo lineal comentado.

1.4. El ACP dentro de los análisis factoriales

La expresión X=YA' es clave a la hora de diferenciar el método aquí propuesto de otros factoriales. Concretamente:

— El *Análisis de Componentes Principales* (ACP). En la definición de las componentes Y interviene el conjunto de las variables X contribuyendo cada una en su totalidad, sin diferenciar en cada variable una parte común y otra específica que no intervenga en la creación de los ejes. Idénticamente para dar cuenta de cada variable no se supone la existencia de una parte de la misma explicada por la comunalidad conjunta de las variables y otra parte inexplicada por ella, o sea específica de la variable. Todas las variables contribuyen a dar cuenta de todas.
— El *Análisis Factorial Confirmatorio* (AFC) en el que se realiza una hipótesis y distinción entre la parte con que cada variable, y cuáles entre ellas por consiguiente, contribuye a la creación de la explicación común y en consecuencia a generar los ejes y la parte específica no explicada por la totalidad, según modelo o diseño previo.
— El *Análisis de Correspondencias Múltiples* (ACM) que partiendo de una métrica definida en la matriz de partida -sea de contingencia, sea lógica o bien cuantitativa- trabaja equivalentemente en los dos espacios, el de las variables como dimensiones con nubes de unidades como puntos o en el de los individuos como dimensiones con nubes de variables como puntos, consiguiendo una sola proyección de las variables/individuos en un espacio de componentes únicas. La fecundidad de tal técnica para el análisis en sociología parece evidente.

1.5. La clasificación del ACP

Aunque no se trate de algo fundamental, pues sobre todo afecta a la manipulación de los datos iniciales, conviene dar cuenta de alguna de las clasificaciones que aparecen en la literatura, si bien se refieren solamente a la naturaleza de las variables puestas en juego al inicio del análisis:

— El *ACP Canónico* en el que a todas las variables se da el mismo peso en la definición de la métrica, expresándose al mismo tiempo en la misma unidad; el criterio adoptado para la misma es,

$$d^2 (E_{i\cdot},E_{i'\cdot}) = \sum_{j=1}^{p} \alpha_j (X_{ij} - X_{i'j})^2, \text{ siendo } \alpha_1=\alpha_2=...=\alpha_p=1$$

— El *ACP Normado* en el que las variables no intervienen con igual peso en la métrica adoptada puesto que, aunque se expresen en la misma unidad, al dividir cada variable por su módulo, su variabilidad es diferente. En la expresión de la métrica anterior

$$d^2 (E_{i\cdot},E_{i'\cdot}) = \sum_{j=1}^{p} \alpha_j (X_{ij} - X_{i'j})^2$$

siendo las a_j diferentes con $\alpha_j = 1/\sigma_j$
por consiguiente la métrica queda,

$$d^2 (E_{i\cdot},E_{i'\cdot}) = \sum_{j=1}^{p} \frac{(X_{ij} - X_{i'j})^2}{\sigma_j}$$

— El *ACP* sobre *Variables Reducidas* en el que las variables son previamente centradas y estandarizadas lo que permitirá trabajar sobre la matriz de correlaciones en lugar de la matriz de varianza/covarianza; por lo demás es la transformación con la que habitualmente se trabaja.

1.6. El programa utilizado

En estas páginas se reproducirán básicamente los resultados que ofrece el paquete estadístico SPSSC+, con algunos comentarios adicionales que faciliten su lectura e interpretación.

El programa que hemos utilizado para llevar a cabo el análisis recoge las siguientes instrucciones:

```
FACTOR VARIABLES=P1 TO P23
/ANALYSIS=P1,P2,P4,P6 TO P9,P11,P12,P14,P15 TO P20,P23
/FORMAT=SORT
/PRINT=ALL
/CRITERIA=FACTORS(4)
/EXTRACTION=PC
```

```
/ROTATION=NOROTATE
/PLOT=EIGEN ROTATION (1,2) (1,3) (2,3) (1,4) (2,4) (3,4)
/ROTATION=VARIMAX
/PLOT=EIGEN ROTATION (1,2) (1,3) (2,3) (1,4) (2,4) (3,4)
/SAVE=REG (ALL FSC)
```

2. ELECCIÓN DEL CAMPO DE LAS VARIABLES

2.1. Condiciones y exigencias previas

Las variables definen el campo de la problemática de la investigación. Por consiguiente se supone que existe una elección previa de las mismas según la pertinencia con relación al objeto investigado.

Aparte de esta primera exigencia sobre las variables previamente consideradas se ha de realizar, si es el caso, una segunda pesquisa para decidir la conveniencia de llevar a cabo el ACP,

— por un lado, si las variables fueran manifiestamente independientes, lo que haría inútil el análisis: es difícil *a priori* saber si se cumple esta condición pues es precisamente lo que se busca; de todas maneras existen índices que, en cierto grado, nos aseguran de tal conveniencia o no.
— por otro, y es lo más habitual, dicha pesquisa es importante a fin de llevar a cabo una segunda depuración de variables para eliminar las informaciones redundantes al ser entre ellas combinación lineal.

2.2. El universo poblacional y el campo de las variables

La población objeto del ACP es la de los individuos mayores de 18 años de la *Regió I* en la división territorial de Catalunya. Se trata de una muestra de individuos y no de unidades familiares. El número es de 3.177.765 personas.

La extensión geográfica de dicha población coresponde a las comarcas del Barcelonès, Maresme, Vallès Oriental, Vallès Occidental y Baix Llobregat, zona que concentra los 2/3 de la población de Catalunya. Los municipios correspondientes son en número de 250.

La base de sondeo inicial o base de sondeo/matriz es el *Padró d'Habitants* del año 1986, archivo del que se han obtenido los datos de las variables/criterio agregadas en secciones censales. Así pues la unidad elemental a estratificar no son los individuos del padrón sino la agregación de éstos en *sec-*

TABLA 1

Media, desviación y descripción de las variables utilizadas en la muestra estratificada

Variable		*Media*	*Desviación*	*Descripción*
P1	*	20.18	6.36	Jóvenes de menos de 15 años
P2	*	13.17	6.83	Viejos mayores de 65 años
P3		86.10	114.70	Índice envejecimiento 65/15
P4	*	37.39	11.47	Inmigración fuera Catalunya
P5		2.00	1.54	Extranjeros
P6	*	5.84	3.21	Nuevos residentes municipio 81-86
P7	*	5.18	4.40	Analfabetos mayores de 10 años
P8	*	9.67	10.22	Titulados medio-superiores >20 años
P9	*	47.65	14.09	Escolarización 14-24 años
P10		40.70	17.36	Escolarización 2-5 años
P11	*	14.89	5.05	Parados antes ocupados
P12	*	9.17	4.76	Paro busca primer trabajo
P13		24.06	8.56	Paro total
P14	*	31.78	5.61	Mujeres activas mayores de 15 años
P15	*	18.12	13.26	Profesiones altas
P16	*	4.21	3.71	Profesiones bajas
P17	*	12.40	4.40	Terciario medio/comercio/hostelería
P18	*	4.29	3.35	Terciario alto/finanzas
P19	*	.99	2.97	Agropecuario
P20	*	37.66	12.32	Vehículo privado trabajo
P21		5.93	7.40	Vehículo privado estudio
P22		23.58	8.80	Vehículo privado trabajo-estudio
A1		1159.35	570.70	Población Sección Censall
P23	*	3.68	13.25	Población Sección/Municipio

* Variables utilizadas en la factorialización

ciones censales que alcanzan el número de 3.509 extendidas en el área geográfica mencionada.

En cuanto a las *variables* utilizadas en primera instancia como criterio para la construcción de la muestra estratificada se han tenido en cuenta diferentes componentes. En la Tabla 1 las variables se expresan en porcentajes de población de la sección censal que posee las características mencionadas sobre la población total de la sección censal.

Como se observa las variables extraídas son de diferentes tipos, tal y como aparece en la Tabla 1: poblacionales, cultural-educativas, ocupacionales, ca-

tegorías y sectores profesionales, movilidad/consumo. Por lo demás dicha tabla no necesita explicación complementaria.

La elección inicial de estas variables está justificada desde varios puntos de vista:

— desde la disponibilidad de los datos del Padrón'86 y el nivel de agregación realizado en los servicios estadísticos no necesariamente sometido a las exigencias temporalmente impulsivas de este tipo de encuestas,
— pasando por el hecho de que en la *Enquesta Metropolitana* del 86 la elección de variables equivalentes a las presentadas consiguió buenos resultados,
— como también por el hecho de que dichas variables se hayan revelado altamente eficaces en análisis multivariables realizados *a posteriori* como resultado de un examen atento de las variables/encuesta del 86 con variables de más poder de categorización en ella.

2.3. Las variables utilizadas en la factorialización

No todas las variables precedentes han sido utilizadas en el análisis factorial previo y, posteriormente, en el de la estratificación. Solamente 17 de las 23, tal y como aparece remarcado en la Tabla 1.

Los estadísticos computados expresan medias y desviaciones de variables cuya métrica en nuestro caso se refiere al porcentaje de una característica dentro de cada sección censal.

Los criterios de eliminación de las variables han sido coherentes con la naturaleza propia del ACP y con los objetivos que cumple dentro del proceso de construcción de la muestra:

— manifiesta combinación lineal y consiguiente correlación entre las variables: la simple comparación en la Tabla 1 patentiza este criterio, y el
— escaso valor y/o dispersión en algunas variables.

Para la elección dos son los criterios básicos:

— la certidumbre de que guardan correlación con las variables/problema, esto es, las de identificación de la propia encuesta, aseverada con los resultados de la EM'86, y
— la observación de la relación entre magnitud y dispersión de las variables.

3. EL PROCESO DEL ACP

El proceso del ACP pasa por cinco momentos básicos:

1) Cálculo de los ejes factoriales o componentes.
2) Cálculo de los valores propios o de las varianzas incorporadas a cada uno de los ejes y del número de los mismos a retener.
3) Recomposición de la matriz de individuos en los nuevos ejes retenidos.
4) El cálculo de la correlación de las componentes con las variables primitivas, comunalidades, recomposición de la matriz de correlaciones.
5) Interpretación de las componentes, rotados o no los ejes.

3.1. Los ejes factoriales

De lo que se trata es de encontrar la dimensiones latentes del campo de variables considerado con relación a los individuos introducidos.

3.1.1. El cálculo de los ejes

Dichos ejes o vectores se hallan en el espacio de referencia R, esto es, en el espacio vectorial engendrado por las variables iniciales y en donde los individuos, en nuestro caso las secciones censales, configuran una nube N de puntos. Las condiciones impuestas a dichos vectores o dimensiones son varias y han estado en buena medida expresadas anteriormente:

— que sean base del sistema vectorial, esto es, sean linealmente independientes,
— que acumulen o expliquen la máxima varianza de la inercia total del sistema,
— que dicha varianza extraída en cada uno de los ejes se realice de manera jerárquica,
— y, como condición complementaria, que sean unitarios.

Las dos primeras condiciones se llenan simultáneamente. En efecto la condición de que cada eje U_k (vector en la dirección de los ejes) maximice la suma de las proyecciones al cuadrado sobre dicho eje de todos los vectores/puntos de la nube N tomando como origen el centro de inercia de la nube (o lo que es equivalente: maximice la inercia con relación a un hiperplano perpendicular a dicho eje, o se haga mínima la suma de las distancias al cuadrado de los puntos al eje, o mínima también la inercia con relación al eje) es lo mismo que hallar los *vectores propios* de la matriz de

Varianzas/Covarianzas (o de Correlaciones con variables estandarizadas) de las variables iniciales que engendran el espacio vectorial (si las variables iniciales están reducidas lo indicado será con la matriz de Correlaciones). Los *valores propios* de la matriz son las sucesivas varianzas incorporados a cada uno de los ejes.

El problema se reduce a hallar $U_{.1}$, vector de dichos ejes, de tal forma que,

$$\sum_{i=1}^{n} d^2(E_{pi},0)$$

sea máxima, siendo E_{pi} la proyección de los puntos en dichos ejes. Hacer máxima la expresión anterior equivale a hacer máxima,

$$S = n\, U'_{.1}\, V\, U_{.1}$$

donde V es la matriz de Varianzas/Covarianzas (o R si se trabaja con la matriz X estandarizada).

Los $U_{.1}$ son vector propio de V (o R con X estandarizada). Si λ_1 es el valor propio correspondiente a $U_{.1}$ se demuestra que,

$$I/n = Tr(V) = \sum_{s=1}^{p} U'_{.s}\, V\, U_{.s} = \sum_{s=1}^{p} \lambda_s = Tr(U'VU)$$

$$\text{y que } \lambda_s = U'_{.1}\, V\, U_{.1}$$

donde I es la Inercia total del sistema y en el que sumatorio se extiende de s=1...p.

La tercera de las condiciones se llena en el proceso mismo de extracción al maximizar sucesivamente sobre cada eje la varianza residual.

En cuanto a la cuarta se trata de una de las condiciones que eliminan la indeterminación del sistema homogéneo de ecuaciones al calcular los vectores propios, condición por lo demás lógica ya que unifica, normalizando, los vectores propios correspondientes a los ejes.

El cálculo de los ejes conlleva pues como elemento de partida la matriz de correlaciones, si se trata de variables estandarizadas. El programa nos suministra dicha matriz junto con otros resultados e índices. Es conveniente detenerse siempre en analizar previamente la matriz de correlaciones y sus secuelas, como la significación de los coeficientes, la matriz inversa, etc., y es necesario además consultar los índices sobre la conveniencia de utilizar el método de ACP. Así pues los analizaremos, para el caso presentado, dentro de este apartado. Dada la extensión impuesta al artículo pasaremos por alto

algunos de los análisis laterales anunciados para centrarnos en los contenidos más específicos del ACP.

3.1.2. La utilización de otros índices

Estos índices son de dos tipos:

El *Test de Barlett* es un indicador del grado de esfericidad de las variables iniciales dado por la expresión,

$$X^2 = -(n-p-1/2)/\ln r^3$$

que sigue la distribución de Chi-cuadrado con los grados de libertad:

$$g.l. = 1/2(m-l+2)(m-l-1)$$

donde p= número de variables.
m= número de componentes comprendidas en el test.
l= número de componentes principales no comprendidas en test.
n= número de casos.

$$r = \frac{(\prod_{k=1}^{m} k)^{1/m}}{\sum_{k=1}^{m} k/p}$$

El valor en el caso analizado de este test llamado también de esfericidad es de 41,770.15. La significación da la probabilidad de que se cumpla la hipótesis de no existencia de correlación entre las variables introducidas, o lo que es equivalente de independencia entre ellas, lo que supone evidentemente la no conveniencia de realizar el ACP. Dicha probabilidad es en el caso mucho menor que 0.05.

El segundo tipo de índices se basa precisamente en la relación existente entre el conjunto de coeficientes de correlación y de correlación parcial. Se entiende que si el conjunto de los coeficientes de correlación parcial son de valores reducidos, la parte específica de las variables será menor con relación a la parte común, lo que hará más pertinente la realización del ACP.

El índice definido para el conjunto de los coeficientes es el de Kaiser-Meyer-Olkin (KMO) que compara el sumatorio de los coeficientes de correlación simples ($r_{jj'}$) con los de correlación parcial ($a_{jj'}$) para las variables j,j' según la expresión:

$$KMO = \frac{\sum\sum r_{jj'}^{2}}{\sum\sum r_{jj'}^{2} + \sum\sum a_{jj'}^{2}}$$ extendiendo el sumatorio a $j = 1,2,..p$ $j' = 1,2,..p$ con $j \neq j'$

Este índice oscila entre 0 y 1; la tendencia a 1 es indicativa de la validez de la aplicación al caso del ACP. En el caso analizado el valor del índice de KMO obtenido es de 0.85, que se puede catalogar de muy óptimo.

Si se desea ver dicha validación para una variable únicamente, por ejemplo la $X_{.j}$ se tendría como expresión:

$$MSAj = \frac{\sum r_{jj'}^{2}}{\sum r_{jj'}^{2} + \sum a_{jj'}^{2}}$$ con $j,j' = 1,2,..p$ y $j' \neq j$

que son los valores que aparecen en la diagonal de la matriz anti-imagen de correlaciones. Para una buena aplicación del ACP el conjunto de estos

TABLA 2

Estadísticos iniciales de comunalidad
y valores propios de las variables utilizadas en el ACP

Variable	*Comunalidad*	*	*Factor*	*Valor Propio*	*% de Varianza*	*% Acumulado*
P1	1.00000	*	1	6.11177	36.0	36.0
P2	1.00000	*	2	2.94263	17.3	53.3
P4	1.00000	*	3	1.89366	11.1	64.4
P6	1.00000	*	4	1.15287	6.8	71.2
P7	1.00000	*	5	.79328	4.7	75.8
P8	1.00000	*	6	.63969	3.8	79.6
P9	1.00000	*	7	.60860	3.6	83.2
P11	1.00000	*	8	.49491	2.9	86.1
P12	1.00000	*	9	.43712	2.6	88.7
P14	1.00000	*	10	.40398	2.4	91.1
P15	1.00000	*	11	.34859	2.1	93.1
P16	1.00000	*	12	.31279	1.8	94.9
P17	1.00000	*	13	.29592	1.7	96.7
P18	1.00000	*	14	.26240	1.5	98.2
P19	1.00000	*	15	.14077	.8	99.1
P20	1.00000	*	16	.11602	.7	99.7
P23	1.00000	*	17	.04501	.3	100.0

valores para todas las variables ha de ser elevado, como es el caso. Puede servirnos de criterio para introducir alguna variables o no en el análisis.

3.2. Valores propios o varianza incorporada a cada eje y la elección del numero de ejes a retener

3.2.1. Valores propios o varianza incorporada a cada eje

Si la matriz de correlaciones es no singular, el número de vectores propios que procura coincide con el número de variables introducidas. Retenerlos todos, aunque tengan propiedades interesantes no nos arrienda la ganancia. Precisamente uno de los objetivos del ACP consiste en reducir el espacio de atributos inicial. Esta reducción implicará normalmente una pérdida de la inercia total o, dicho con otras palabras, si se extraen menos dimensiones de las que configuran el espacio inicial no se podrá dar cuenta de la inercia total del sistema por ejes linealmente independientes. Pero la pérdida por este lado puede implicar ganancia en cuanto a sencillez de la estructura adoptada para el sistema sociológico y en cuanto a la interpretación de los ejes.

El programa nos ofrece los valores propios de la matriz correspondientes a cada uno de los ejes tal y como aparece en la Tabla 2. De la observación de dicha tabla se deduce el rápido decrecimiento de dichos valores y su acumulación en los primeros valores, lo que facilitará la reducción.

3.2.2. Los ejes retenidos

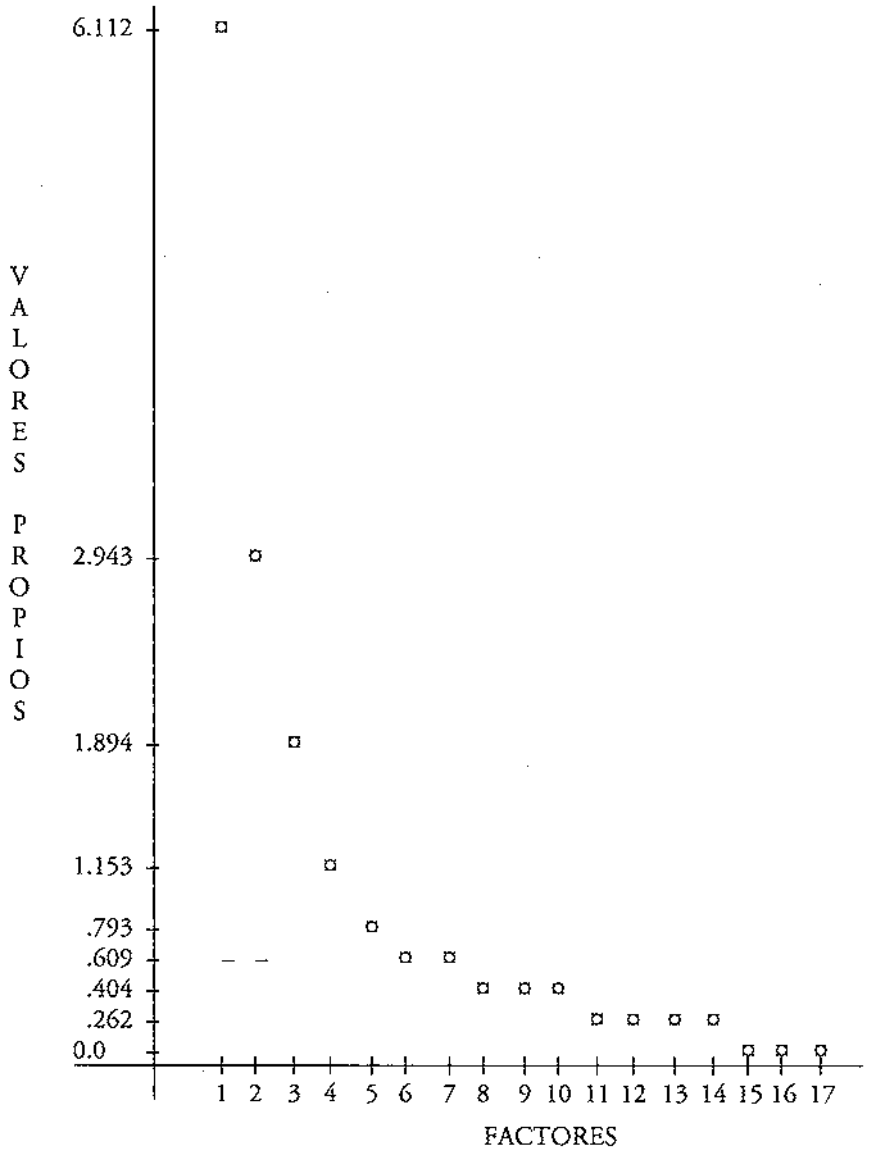
Todos los criterios que convencionalmente se toman a la hora de decidir sobre el número de ejes a retener funcionan aquí al unísono. La gran correlación encontrada anunciaba ya que con relativamente pocos factores se llegaría a dar cuenta de una buena parte de la varianza.

Sólo cuatro de los ejes acumulan el 71.2% de la misma como muestra la Tabla 2. Precisamente uno de los criterios convenidos consiste en tomar tantos ejes cuantos entre ellos acumulen al menos el 70% de la varianza en sus valores propios. Además los cuatro son los que tienen un valor propio superior a la unidad, que suele ser otro de los criterios tomados. Si se toma el Gráfico 1, en el que los valores propios aparecen representados en vertical en función de su orden, se observa para uno de los órdenes, el 4, y para su valor propio, una ruptura de la continuidad, codo o quebrada de valle, que es el valor y orden adecuados para ser tomado como criterio.

Dado que la importancia de cada eje es desigual se habrá de relativizar en la interpretación el peso de las diferentes dimensiones como factores que estructuran el conjunto, pero aún no estamos en dicho momento. Además,

GRÁFICO 1

Representación gráfica de los valores propios
de los factores resultantes del ACP



como se verá posteriormente, los ejes hallados inicialmente sufrirán rotaciones con lo que cambiarán las partes correspondientes de explicación de cada uno si bien el porcentaje total de varianza explicado, 71.2 %, queda inalterable para el conjunto de los cuatro ejes, que serán los finalmente retenidos.

3.3. Recomposición de la matriz de individuos en los nuevos ejes retenidos

Los puntos de la nube, secciones censales en el caso analizado, tendrán otra proyección o componentes sobre los nuevos ejes. De aquí precisamente la expresión de componentes a los ejes factoriales.

La expresión matemática de dichas componentes viene dada por

para una unidad i, en la componente K, $Y_{ik} = X_{i}.U._{k}$

para una unidad i, en componente genérica, $Y_{i}. = X_{i}.U$

para componente k, en unidad genérica, $Y._{k} = X.U._{k}$

en general $Y = XU$

Una vez reducidas las variables a un número considerablemente inferior de factores es posible calcular las puntuaciones factoriales para cada caso, individuo o unidad considerada. El interés de dichas componentes no es necesariamente inmediato dentro del proceso del ACP, en particular si las unidades representadas son intercambiables y por consiguiente no cualificadas o significativas en tanto que entidades o categorías sociológicas. Los factores, como nuevas variables que son, tendrán un valor concreto para cada individuo que podrán ser reutilizados posteriormente en otros tipos de análisis. Estas componentes quedan guardadas y pueden recuperarse en tanto que matriz resituando así los individuos en el nuevo espacio. Concretamente en el caso analizado dichas componentes, para los cuatro ejes mencionados, fueron las variables utilizadas para realizar el análisis de *cluster* con vistas a constituir los estratos de población homogéneos de donde extraer una muestra aleatoria. La matriz que nos permite obtener estas puntuaciones a partir de la matriz de datos original aparece a continuación. Se reproduce en la Tabla 4.

3.4. El cálculo de la correlación de las variables primitivas con las componentes, la comunalidad y la recomposición de la matriz de correlaciones

Interesa encontrar la relación que las variables primitivas tienen con las componentes. Dicha relación permite cubrir varios objetivos simultáneamente:

TABLA 4

Matriz de coeficientes de las puntuaciones factoriales

	Factor 1	*Factor 2*	*Factor 3*	*Factor 4*
P1	-.07150	.20522	-.04003	.23236
P2	.06360	-.25642	.07799	-.17980
P4	-.12293	.10133	-.15328	.00843
P6	-.01469	-.15429	.09771	.48516
P7	-.12359	-.02959	.00468	-.01522
P8	.15518	.10075	-.04867	-.06603
P9	.15685	.12277	-.06273	-.11293
P11	-.13167	-.15824	-.11457	.07338
P12	-.10249	.11942	-.03874	-.15666
P14	.03049	-.07773	-.17150	.45439
P15	.15764	.07292	-.04466	-.06366
P16	-.08787	.13156	-.00113	-.10742
P17	-.06394	-.47855	.03148	.29889
P18	.11815	-.06948	-.09754	.08374
P19	-.00782	-.03712	.43420	-.06281
P20	.03525	.25672	.15586	.02312
P23	-.00136	.00301	.42685	-.04591

a) recomponer las variables originales en los nuevos ejes lo que nos mostrará la estructura del primer espacio de atributos, y

b) dar identidad a las componentes.

Pero antes procederemos por pasos para mejor llenar estos objetivos finales.

3.4.1. La correlación de las variables con las componentes, coeficientes y matriz de saturaciones.

La expresión de las componentes en función de las variables iniciales está dada por la expresión,

$$Y_o = XU$$

donde la Y_o es la Y precedente que está sin estandarizar.

Realizando la transformación

$Y = Y_0 D^{-1/2}$ para estandarizar la Y se obtiene,

$$X = YA'$$

con $A = UD^{1/2}$ y D matriz diagonal de los valores propios.

La A es una matriz (pxm), o (pxp si no hay reducción), y se denomina matriz de saturación o matriz factorial (*Factor Matrix*).

A partir de la relación X=YA' se tienen para situaciones particulares:

$$X_{ij} = \sum Y_{ik}\, a_{jk}$$

$$X_{i\cdot} = Y_{i\cdot}\, A'$$

$$X_{\cdot j} = Y\, A'_{j\cdot}$$

En el caso propuesto se han elegido previamente 4 factores. A partir de estos cuatro factores, los coeficientes que relacionan las variables con éstos aparecen en la matriz factorial. Cada fila contiene los coeficientes que expresan la variable estandarizada en términos de los factores. Así, por ejemplo, la variable P15 se puede expresar como:

$$P15 = .90640F_1 - .01627F_2 - .12385F_3 - .11479F_4$$

Estos coeficientes son los llamados *factores de carga* (*Factor Loading*) y nos indican el peso que tiene cada factor en cada una de las variables. El conjunto de todos ellos forman la matriz del modelo factorial (*Factor Pattern Matrix*). Como los factores son ortogonales, no correlacionados, los factores de carga, de hecho, son las correlaciones entre las variables y los factores, y coincide con la denominada matriz de la estructura factorial (*Factor Structure Matrix*), por eso aparece titulada a continuación como matriz factorial, englobando esta coincidencia.

Los valores a_{jk} de la matriz de saturación, o también de pesos, se interpretan como los coeficientes de correlación entre cada una de las componentes y las variables o, lo que es equivalente, los cosenos del ángulo, o cosenos directores, entre ambas variables en el espacio en que los ejes son los individuos y siempre que las componentes estén estandarizadas. Dicha interpretación nos da las pistas suficientes, como veremos más adelante, para identificar las componentes.

En la Tabla 5 aparece dicha matriz teniendo en horizontal las variables y en vertical las componentes.

Tabla 5

Matriz factorial o de saturaciones del ACP

	Factor 1	*Factor 2*	*Factor 3*	*Factor 4*
P15	.90640	-.01627	-.12385	-.11479
P8	.88204	.03255	-.15307	-.13383
P9	.86330	.01656	-.18238	-.18980
P4	-.78195	.03011	-.33984	-.01868
P18	.77294	-.13849	-.13239	.12743
P7	-.75053	-.07560	.04087	.01619
P12	-.73340	-.05406	-.11868	-.20670
P11	-.72773	-.35246	-.07559	.20223
P16	-.62295	.09971	-.07957	-.17585
P20	.15946	.77161	.03977	-.17136
P1	-.38869	.75623	-.29906	.10174
P2	.38193	-.72727	.39148	-.02520
P6	.18621	.59867	.16153	.53995
P19	-.04254	.38354	.77675	-.10708
P23	-.00735	.48473	.72612	-.11618
P17	-.10995	-.46279	.37325	.58730
P14	.41016	.35677	-.34719	.50257

Otra interpretación que se puede hacer de esta matriz es considerar a los factores de carga como los coeficientes resultantes de una ecuación de regresión múltiple, donde la variable original sería la variable dependiente y los factores las independientes. Como éstos están incorrelacionados, los coeficientes no dependerán el uno del otro y representarán la contribución única de cada factor o la correlación entre ambos, factor y variable. De esta forma podemos calcular la proporción de varianza de cada variable explicada por el modelo de 4 factores, que será la suma de la proporción de varianza explicada por cada factor. Por ejemplo, si consideramos la anterior variable P15 la varianza explicada por los 4 factores será:

$$h^2 = (.90640)^2 + (-.01627)^2 + (-.12385)^2 + (-.11479)^2 =$$

$$= .8215609 + .00026 + .0153388 + .0131767 =$$

$$= .85034$$

que se denominará comunalidad, tal y cono la definiremos en el apartado siguiente.

Se puede comprobar que la expresión,

$$\lambda_1 = \sum a^2_{j1} \text{ para el factor 1 y en general}$$

$$\lambda_k = \sum a^2_{jk}$$

nos da los valores propios o varianza explicada por cada componente en función de los coeficientes de saturación. Dichos valores coinciden con los primitivos mientras no se realice rotación de los ejes.

El interés mayor de la matriz de saturaciones estriba, a partir del significado de sus coeficientes, en que puede interpretarse como un nuevo espacio vectorial en el que las variables primitivas son puntos en las cuatro dimensiones elegidas. La proximidad a uno de los ejes de una de las variables significará la gran correlación positiva (o negativa) de la variable con dicho eje. La proximidad entre variables indicará la correlación positiva entre ellas en el espacio de las cuatro dimensiones, o de otras dimensiones tomadas si es el caso. Estas reflexiones dan criterios sencillos para la interpretación de los ejes.

3.4.2. La comunalidad

La expresión y definición de la comunalidad para la variable j es la siguiente:

$$h^2_j = \sum_{k=1}^{P} a^2_{jk}$$

Como se observa el sumatorio se extiende a todas las componentes, tantas como variables iniciales. Se demuestra que tal sumatorio coincide con el índice de correlación de cada variable consigo misma, esto es, con 1. Geométricamente se interpreta como la longitud de dicho vector o variable en el espacio de las componentes. Tomadas todas las componentes, sin reducciones, la comunalidad de cada variable es la misma e igual a 1, y pone de manifiesto el hecho de que todas intervienen «comunmente» en la estructuración de la totalidad o contribuyen igualmente a la inercia total: aparecen en la Tabla 2 como comunalidades iniciales. En ello se diferencia el ACP de otros factoriales, por ejemplo el confirmatorio. En este último la comunalidad inicial no es idéntica para todas las variables puesto que se parte de hipótesis previas en las que la parte común de cada variable que interviene en la explicación de la totalidad, y por tanto su parte específica, es diferente.

TABLA 6

Estadísticos finales de comunalidad
y valores propios de las variables utilizadas en el ACP

Variable	*Comunalidad*	*	*Factor*	*Valor propio*	*% de Varianza*	*% Acumulado*
P1	.82275	*	1	6.11177	36.0	36.0
P2	.82869	*	2	2.94263	17.3	53.3
P4	.72819	*	3	1.89366	11.1	64.4
P6	.71072	*	4	1.15287	6.8	71.2
P7	.57094	*				
P8	.82039	*				
P9	.81484	*				
P11	.70043	*				
P12	.59760	*				
P14	.66864	*				
P15	.85034	*				
P16	.43527	*				
P17	.71050	*				
P18	.65038	*				
P19	.76372	*				
P20	.65176	*				
P23	.77576	*				

En el caso de reducción de componentes, que es el fin perseguido, el sumatorio se extiende solamente a los ejes tomados lo que da como expresión,

$$h^2_j = \sum_{k=1}^{m} a^2_{jk}$$

siendo en nuestro caso m=4.

La Tabla 6 muestra dichas comunalidades, siendo la intepretación idéntica a la anterior pero ahora sobre un espacio de cuatro ejes: cada valor es la expresión de la longitud al cuadrado de la variable en el espacio de las componentes, ahora 4, lo que es equivalente a afirmar que la comunalidad es la parte con que el conjunto de los cuatro ejes contribuye a la varianza de dicha variable o, y es lo más interesante desde el punto de vista de la interpretación, es la parte de la contribución de dicha variable a estructurar el sistema de

los cuatro ejes. Por consiguiente comunalidades altas serán interesantes en el sentido de que tiene importancia en la «creación» de los cuatro ejes independientemente de que puedan tenerlo en alguno en particular. Dichas variables estarán alejadas en la representación del centro de la misma. Variables con comunalidad baja contribuyen poco a la formación del sistema o a estructurar el espacio, lo que significa que son variables que contribuyen poco a dispersar la nube de puntos: se situarán en puntos próximos al centro de masas del sistema, contrariamente a las primeras que intervienen fuertemente en la dispersión de la población.

La rotación de los ejes no afectará a la comunalidad de cada variable con relación a las cuatro componentes que se toman. Por ello las reflexiones que se hacen aquí continuarán siendo válidas una vez rotados los ejes iniciales por el procedimiento *varimax*.

La mayor parte de las variables contribuyen fuertemente a la creación de los factores elegidos. Solamente una de ellas, la P16, tiene una comunalidad inferior a 0.50, concretamente 0.43, encontrándose 11 con una comunalidad superior a 0.70, lo que ya se suponía al analizar las diferentes correlaciones. Por ejemplo, las variables que más comunalidad tienen, P15, P2, P1, P8, P9 (todas ellas con comunalidad superior a 0.80) son entre las que más correlación tienen entre sí.

3.5. Interpretación de las componentes

El objetivo último del ACP es, como anteriormente se indicaba, doble: primero, recomponer las variables originales en los nuevos ejes lo que mostrará la estructura del primer espacio de atributos y, segundo, interpretar las componentes, esto es, dar carta de identidad a la estructura emergida.

3.5.1. La rotación de los ejes

La interpretación puede hacerse a partir del análisis de la Tabla 5 precedente, pero habitualmente la extracción inicial de los factores no nos permite identificar con toda claridad la relación o el modelo subyacente que se establece entre los factores y las variables. Con el objeto de evidenciar esta relación se procede a la llamada rotación de los factores que consiste en una transformación de la matriz factorial original en otra más simple que adecúa mejor los ejes al aproximarlos a las variables correlacionadas. Facilitando la interpretación de la estructura de los datos no se altera la bondad de ajuste de la solución factorial, las comunalidades y los porcentajes de varianza explicada se mantienen inalterados, simplemente se redistribuye la varianza explicada entre los factores.

Los procedimientos empleados son de dos tipos:

a) La rotación rectangular

La rotación se efectúa haciendo que los ejes permanezcan perpendiculares. Dicha transformación, al conservar las distancias, deja inalterable la comunalidad de cada variable lo que hace que las interpretaciones encontradas a partir de ellas sean las mismas. No así el valor propio de cada componente, siendo diferente su importancia relativa en la explicación de la varianza total.

Hay que recordar que las rotaciones son más utilizadas para análisis factoriales confirmatorios, pues para estos análisis fueron ideadas, pero es ya práctica habitual utilizarlas también para el ACP, en particular la técnica denominada *Varimax*.

Tres son, a su vez, los procedimientos utilizados en rotación rectangular: unos con más interés que otros desde el punto de vista del ACP.

a.1) Varimax:

Es la técnica comunmente usada y lo es en este estudio. Minimiza el número de variables que tienen un factor o componente de saturación sobre una variable, acentuando los que lo tienen más elevado. Las componentes quedan más limpias al tener sobre ellas las variables que más peso tienen, eliminando sobre dicha componente las intermedias.

La expresión a maximizar es, para un eje:

$$V(Y_{\cdot k}) = \sum_{j=1}^{p} (a^2_{jk} - a^2_{0k})^2/p = (p \sum_{j=1}^{p} a^4_{jk} - (\sum a^2_{jk})^2)/p^2$$

$$\text{para el conjunto, } V = \sum_{j=1}^{p} V(Y_{\cdot k})$$

donde a^2_{0k} es la media de las a^2_{jk} del conjunto de valores $j=1,\ldots,p$.

a.2) Quartimax

Es una técnica que minimiza el número de factores que corresponden a una variable. Se trata de que cada variable se proyecte al máximo sobre factores o componentes diferentes, dentro evidentemente de los límites del método. La lógica de la reducción buscada en un análisis básicamente exploratorio no va en esta dirección. El supuesto más adecuado para su aplicación consistiría en que las variables introducidas poseyeran determinados grados de independencia supuesta.

La expresión a maximizar es, para un eje:

$$V(Y_j.) = \sum_{k=1}^{m} (a^2_{jk}-a^2_{j0})^2/m = (m \sum_{k=1}^{m} a^4_{jk}-(\sum a^2_{jk})^2)/m^2$$

$$\text{para el conjunto } Q = \sum_{k=1}^{m} V(Y_j.)$$

donde a^2_{j0} es la media de las a^2_{jk} del conjunto de valores k=1,...,m.

a.3) Equimax, biquartimax, etc.

Se trata de una combinación de las dos técnicas precedentes. La expresión a maximizar es:

$$E = \alpha Q + ß V$$

donde los valores de α,ß son diferentes según la técnica.

b) La rotación oblicua.

Los ejes rotados a partir de los primeros factores no conservan la ortogonalidad, lo que tiene como consecuencia que tampoco conserven la comunalidad de cada variable, y desde luego rompe con uno de los objetivos que consiste en buscar la incorrelación de los ejes. La técnica es más útil y utilizada con modelos previos, esto es con análisis factoriales confirmatorios, por ello no insistimos.

Evidentemente tanto para tener una idea de la envergadura de la rotación realizada, como para saber en qué se han transformado los valores propios o la parte explicada de cada eje, interesa conocer la proyección de los factores o componentes primitivos sobre los factores rotados es decir, la matriz de transformación entre ellos tal y como aparece en la Tabla 7.

3.5.2. La matriz de saturaciones y los pesos de cada componente en los ejes rotados

La Tabla 8 muestra la matriz de saturaciones para las componentes rotadas según el procedimiento *varimax*. Aparecen además los coeficientes distribuidos en grupo según el orden de importancia en cada uno de los ejes, para facilitar la interpretación.

Como se ha anunciado anteriormente la rotación *varimax* redistribuye diferentemente la varianza entre los ejes aunque la cantidad global de los cuatro quede invariante e igual a 12.1 siendo el porcentaje acumulado de 71.2%.

TABLA 7

Matriz de transformación factorial
después de la rotación varimax del ACP

	Factor 1	*Factor 2*	*Factor 3*	*Factor 4*
Factor 1	.99594	-.04736	.00886	.07610
Factor 2	-.01274	.70576	.44054	.55468
Factor 3	-.01789	-.44072	.88595	-.14330
Factor 4	-.08735	-.55266	-.14468	.81609

TABLA 8

Matriz factorial o de saturaciones del ACP
después de la rotación varimax

	Factor 1	*Factor 2*	*Factor 3*	*Factor 4*
P15	.91517	.06361	-.09225	-.01597
P8	.89247	.12262	-.09409	-.00210
P9	.87984	.13270	-.13376	-.07225
P4	-.77144	.21838	-.29205	-.00935
P18	.76280	-.14643	-.18989	.10497
P7	-.74866	-.04477	-.00609	-.09169
P11	-.73660	-.29273	-.25795	-.07501
P12	-.70955	.16312	-.10555	-.23748
P16	-.60491	.23213	-.00665	-.12421
P17	-.16159	-.81048	.04086	.16074
P2	.38484	-.68997	.03347	-.45100
P1	-.40028	.62770	.05004	.51577
P20	.16323	.61420	.40137	.29459
P23	-.01634	.08664	.87359	.06945
P19	-.05180	-.01044	.87225	.01081
P6	.12776	.04410	.33038	.76374
P14	.36626	.10763	-.21950	.68901

Recalculando de nuevo los valores propios, teniendo en cuenta que los procentajes finales están calculados tomando el 71.2% de la varianza total como 100%, aparece la distribución siguiente:

1er. eje	v.p.= 6.07	% varianza = 50.20%
2do. eje	v.p.= 2.20	% varianza = 18.20%
3er. eje	v.p.= 2.07	% varianza = 17.20%
4to. eje	v.p.= 1.74	% varianza = 14.40%

La distribución es algo diferente a la dada inicialmente al disminuir mínimamente el primer eje, algo más el segundo y los dos últimos. Dejando pues la importancia del primero casi inalterable (conlleva la mitad) se consigue reequilibrar los otros tres teniendo entre ellos un peso más equilibrado no llegando cada uno a 1/5 del total.

3.5.3. La identidad de los ejes

La interpretación fundamental se realiza a partir de la matriz de saturaciones, es decir, por el análisis de la Tabla 8, que da las correlaciones de las variables primitivas con las componentes, esto es, la proyeción de las variables sobre las componentes una vez están estandarizadas ambas. De ayuda inestimable son las representaciones gráficas en las que por pares de ejes se van proyectando las variables

Los criterios seguidos para encontrar la identidad buscada y agrupar las variables (verdadero *cluster* de variables) son varios:

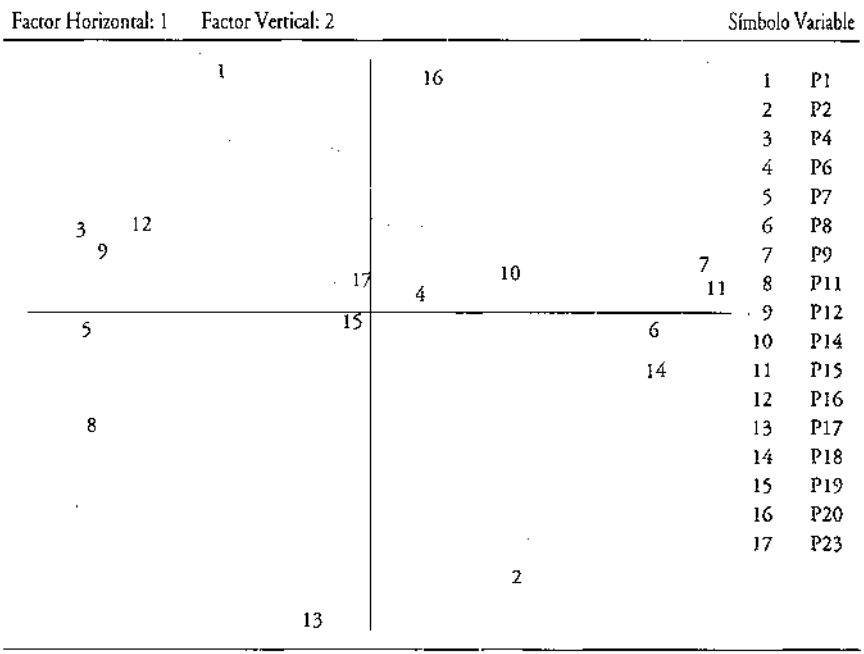
a) Comenzar a identificar el eje de más peso teniendo este dato presente para dar a dicho eje la importancia que le corresponde.

b) Dentro de las variables de máxima comunalidad elegir las de mayor valor sobre el eje analizado y mínima en los otros, lo que equivale a elegir, del eje estudiado, aquellas variables que mayor coeficiente de saturación tienen sobre él y mínimo en los otros ejes y ello en ambos lados: coeficientes positivos y negativos.

c) Establecer una escala de variables que recorran todo el eje y ello independientemente de sus comunalidades: primero, evidentemente para las variables más próximas a él, y luego para otras siempre que se mantengan constantes con relación a los otros ejes. Este criterio es importante ya que esta

GRÁFICO 2.1

Representación gráfica de los factores 1, 2
y las variables utilizadas en el ACP después de la rotación varimax



Las coordenadas son los valores que aparecen en la matriz factorial rotada (Tabla 8) para los factores 1 y 2.

escala de variables puede marcar o manifestar progresivamente la idea latente de la dimensión.

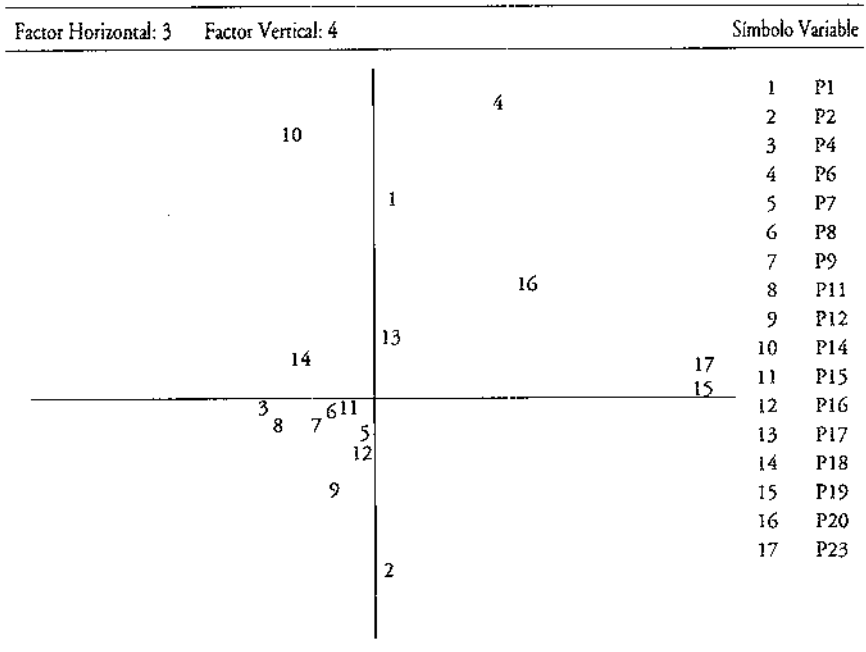
d) Agrupar las variables por proximidades, sea en todos los ejes, sea en pares de ellos. Ello podrá ofrecernos los *cluster* de las mismas. Si al mismo tiempo se sitúan dichos grupos en los ejes se contribuye mejor a comprender la estructura del sistema social tenido en cuenta, poniéndonos en el camino de crear o corroborar tipologías.

e) Establecer agrupaciones de variables según comunalidades.

Para la interpretación se ha tenido en cuenta la Tabla 8 y los *gráficos* resultantes de representar las coordenadas de cada par de ejes. Ahorrando el

GRÁFICO 2.2

Representación gráfica de los factores 3, 4
y las variables utilizadas en el ACP después de la rotación varimax



Las coordenadas son los valores que aparecen en la matriz factorial rotada (Tabla 8) para los factores 3 y 4.

detalle de un análisis minucioso solamente se apuntan las tendencias de los ejes. El lector tiene datos suficientes a partir de los gráficos que se presentan a continuación, Gráficos 2.1 y 2.2, de ir más lejos en la interpretación.

La identidad de los ejes apunta hacia dimensiones que nos aseguran en la buena elección de las variables iniciales y en la justeza del procedimiento en general, al mismo tiempo que nos procura el objetivo buscado: identificar las dimensiones básicas que en el porcentaje elegido dan cuenta de la diversidad de la *Regió I.*

El primer eje

En el conjunto de las variables este eje acapara más del 30% de la varianza y, teniendo en cuenta solamente la estructura elegida (cuatro ejes), la mitad de la varianza descrita por los cuatro. Tiene, por consiguiente, un peso con-

siderable y la dimensión que representa será decisiva a la hora de configurar la estructura de las variables elegidas.

En uno de los polos del eje aparecen,

Categorías profesionales y niveles educativos altos:

— profesiones altas,
— títulos medios y superiores,
— escolarización entre 14-24 años,
— terciario, altas finanzas.

y en el otro polo se tiene,

Categorías profesionales y niveles educativos bajos junto a elevados índices migratorios y de paro:

— analfabetos con más de 10 años,
— presencia migratoria,
— parados antes ocupados,
— profesiones bajas.

Ambos polos dan contenido a una dimensión cargada por variables que van en la dirección de categoría socioprofesional y nivel formativo lo que haría intuir que se trata de una dimensión definida como *categoría sociocultural.*

El hecho de aparecer las variables migratorias o de origen y las ocupacionales hace perfilar mejor aún la naturaleza de este eje: aparecen pues particularmente correlativas dichas variables con las de categorías bajas. Por ello se puede afirmar que se trata de un eje que marca también *la integración en el mundo laboral y el origen,* fenómenos en estrecha relación con la categoría sociocultural.

Se trata pues de una dimensión que atraviesa y estructura la población considerada según *la categoría social entendiéndola como un compuesto de integración laboral ocupacional, categoría socioprofesional y cultural y origen inmigrante.*

El segundo eje

Tomando solamente como referencia los cuatro ejes considerados, éste no llega al 20% de la varianza de los mismos, lo que es importante tener en cuenta a la hora del darle el peso explicativo correspondiente. Por ello y por el hecho de que algunas de las variables que más se proyectan en él aparecen también sobre otros ejes, su identidad es algo más difusa.

Básicamente se trata de un eje que va de un polo en el que aparecen variables como:

— jóvenes menores de 15 años,
— utilización del vehículo privado para ir al trabajo,

al otro en el que se tienen variables,

— los mayores,
— terciario medio, comercio y hostelería.

La identidad de esta dimensión está dada básicamente por la Edad y, por consiguiente, aunque más parcialmente por profesiones/ocupaciones a ella vinculadas.

Se trata de una dimensión que va en la dirección del *ciclo vital y/o de la edad* con la correspondiente y parcial repercusión de una cierta *estagnación profesional* en uno de los polos.

El tercer eje

Conlleva, como el segundo, un peso que no llega al 20% del conjunto de los cuatro considerados, lo que nos orienta acerca de la importancia relativa de esta dimensión. Aunque el peso de la varianza no es considerable, sin embargo, la naturaleza del eje aparece bien definida ya que las variables que sobre al menos uno de los polos se proyectan tienen poca incidencia sobre otros ejes.

De uno de los lados de la polaridad aparecen nítidamente,

— los municipios pequeños,
— el sector agropecuario,

del otro no se da un proyección neta de determinadas variables, de alguna manera se puede hablar del

— resto poblacional.

Se trata de una dimensión de *identidad metropolitana o de metropolización* que diferencia el primer grupo de variables del resto. La consecuencia es de interés pues este eje marca la diferencia entre la población de la encuesta del 86 de la actual.

El cuarto eje

Como se ha anunciado este eje tiene poco peso de varianza en el conjunto de los cuatro elegidos: significa bastante menos del 20% atribuido a los dos anteriores. Uno de los polos aparece más claramente definido que el otro.

Así sobre el primero aparecen claramente variables como:

— nuevos residentes venidos a la sección entre los años 81-86,
— mujeres activas de más de 15 años,
— jóvenes (aunque también con proyección sobre el segundo eje)

y, sobre el otro, aunque más difusas,

— parados antes ocupados (aunque con proyección importante sobre el primero y el segundo de los ejes),
— mayores en edad (aunque con repercusión importante en el segundo eje).

Se trata de una dimensión que indica:

La sedimentación-dinamismo poblacional/residencial, ligada a la actividad de la mujer y al Índice ocupacional.

A partir del cuerpo de variables elegidas la estructura que se nos aparece como resultado de la dimensionalización de la realidad de la *Regió Metropolitana de Barcelona* es la de una población atravesada básicamente por la:

Categoría social entendida en sentido amplio (categorías profesionales + niveles educativos+ origen+ ocupación)

Edad con las consecuencias sobre aspectos de dinamismo profesional

La metropolización

Movilidad geográfica y ocupacional.

COMENTARIOS FINALES

El objetivo de este artículo se centraba en dar cuenta de uno de los métodos multivariables que en nuestra opinión tienen gran importancia y utilidad en el análisis sociológico, a partir de un ejemplo y sin cargar las tintas en los aspectos exclusivamente matemáticos y técnicos. Pensamos que con el nivel de explicación dado será suficiente para que puedan ser comprendidas por el profano las líneas básicas de esta técnica. En la selección bibliográfica que aparece al final de la revista se incluyen libros generales y de carácter específico que desarrollan este método factorial.

Otro de los objetivos que nos planteábamos consistía en la utilización de datos secundarios provenientes de otras fuentes que la de la encuesta a fin de mejor validarla y diseñarla. Aunque esta técnica ha sido utilizada formando parte de un proceso para construir la muestra estratificada, los resultados

nos han sido y serán preciosos, por un lado, como estudio en sí mismo que permite el conocimiento del territorio desde su caracterización social, mostrándose asimismo como ejemplo de análisis de datos secundarios que va más allá de lo que las estadísticas oficiales procuran, y, por otro, para contrastar y validar los resultados que se obtengan de la *Enquesta Metropolitana*'90 así como otras posibles construcciones de variables sobre las que la estructura aquí presentada actuará de importante referente.

Este artículo estaría incompleto si, a partir de los resultados aquí obtenidos, no tratáramos de estratificar la población identificando una tipología poblacional, en el sentido de que además de categorizarla la sitúe en la estructura del campo de las variables que la definen, permitiéndonos obtener unas zonas sociológicas en el territorio que vayan más allá de las divisiones geográficas y administrativas. Dicha tipificación aparecerá en un próximo artículo.