LOS METODOS MULTIVARIANTES DE REPRESENTACION SIMULTANEA COMO TECNICAS DE INSPECCION DE DATOS ECONOMICOS

J. L. VICENTE VILLARDON*, M. P. GALINDO**,S. VICENTE TAVERA***, A. MARTIN* , I. BARRERA**, M. J. FERNANDEZ**

* Facultad de C.C. Económicas y Empresariales. ** Facultad de Medicina. ***E. U. de E. Empresariales.

Departamento de Estadística y Matemática Aplicadas. Universidad de Salamanca.

RESUMEN

El análisis de muchos problemas económicos requiere el manejo de grandes matrices de datos, cuya inspección directa hace inviable el conocimiento de los **"patterns"** de variación . Los métodos multivariantes abordan este problema pero, en la mayoría de los casos, estos métodos se reducen a Análisis de Clusters, Análisis Discriminante etc, que se caracterizan por tener restricciones en las hipótesis de base, necesarias para su fiabilidad, pero que limitan su éxito.

Las técnicas que reducen la dimensionalidad descomponiendo la variabilidad global según direcciones principales de inercia, son claramente una herramienta de gran utilidad ya que, permiten presentar líneas de la matriz en un subespacio de dimensión reducida donde es posible interpretar sus posiciones relativas.

Particularmente interesantes son las técnicas de representación simultanea (Análisis Factorial de Correspondencias (BENZECRI, 1976), Análisis Biplot (GABRIEL, 1971, 1990; GALINDO, 1986)) las cuales permiten interpretar la relación entre variables en términos de covariación, la relación entre poblaciones o individuos en términos de similitud y la relación entre ambos conjuntos en términos de preponderancia.

En este trabajo se lleva a cabo un estudio comparativo de los métodos y se pone de manifiesto como los Métodos "Biplot", cuya aplicación tiene una gran difusión en otros campos de la Ciencia, resultan útiles en la resolución de problemas económicos. El objetivo es encuadrar la Comunidad de Castilla y León en el ámbito nacional en base a la información suministrada por variables de producción agrícola y ganadera.

INTRODUCCION

Tradicionalmente, los datos económicos se han tratado con técnicas descriptivas unidimensionales que presentan sus resultados en forma gráfica, sin embargo, estas técnicas producen sólo una visión parcial de la información debido a que no tienen en cuenta las relaciones entre todas las variables que intervienen en el proceso. Parece entonces, que el análisis multivariante tiene importante futuro dentro de esta ciencia, a pesar de que hasta el momento no se ha utilizado mucho.

En los últimos años, debido al avance de la informática, es posible realizar tratamientos simultáneos de un gran número de variables utilizando las técnicas del Análisis Multivariante. Muchos de estos métodos tienen su resultado expresado en forma gráfica por lo que resultan de fácil interpretación.

Partimos de un conjunto de n individuos(o poblaciones) sobre el que se han medido p variables, las observaciones x_{ij} del individuo i para la variable j se ordenan en forma de una matriz con n filas y p columnas $X_{(nxp)}$. El Análisis Multivariante trabaja, entonces, con matrices de datos.

Pueden encontrarse diversos tipos de técnicas, de acuerdo con varios criterios; uno muy general podría ser el siguiente:

- **Técnicas descriptivas** (Gráficos Multivariantes, Análisis de Componentes Principales, Análisis Factorial de Correspondencias, Biplot, Análisis de Clusters, Multidimensional Scaling, etc ...)
- **Técnicas inferenciales** (Análisis Discriminante, Correlación Canónica, Análisis de Compoentes Principales, Análisis Factorial, Regresión Multivariante, Modelos de Ecuaciones Simultáneas, etc...)

En Economía se han utilizado, fundamentalmente, las técnicas del segundo grupo, concretamente las técnicas de regresión, que muchos autores, ni siquiera consideran dentro de las disciplinas propias del Análisis Multivariante sino de la Estadística Clásica.

El resto de los métodos de este grupo pueden ser muy útiles en determinados casos, pero las restricciones que conllevan en las hipótesis básicas (Normalidad, Igualdad de Varianzas) de los modelos que utilizan, hacen que su aplicabilidad sea bastante limitada.

Nos centraremos, entonces, en el primer tipo de técnicas: los Métodos Descriptivos del Análisis Multivariante, ya que éstos no tienen practicamente ninguna restricción en cuanto a las hipótesis que deben verificar las matrices de datos a tratar.

Podemos utilizar los métodos deseados desde dos puntos de vista:

- **Análisis Confirmatorio.**
- **Análisis Exploratorio.**

En el primer caso se trata, simplemente, de verificar algunas hipótesis supuestas de antemano. En el segundo, de intentar ver la estructura de los datos sin hipótesis preconcebidas. En ambos casos las técnicas estadísticas son las mismas.

Dividiremos, una vez más, los métodos en dos tipos:

- **Métodos Gráficos Multivariantes.**
- **Métodos Multivariantes que presentan sus resultados en forma gráfica.**

Los primeros son, simplemente, una transcripción directa de los datos en un gráfico, mientras que los segundos son técnicas estadísticas complejas cuyo resultado final es un gráfico (representaciones cartesianas, demdogramas, árboles aditivos, etc...)

METODOS GRAFICOS MULTIVARIANTES

Intentan, mediante la transcripción directa de los datos, resumir la información que producen varias variables de forma que sea posible distinguir individuos o grupos de individuos mediante observación rápida de los gráficos.

En sólo dos dimensiones el problema es muy simple, ya que un diagrama de dispersión contiene toda la información.

Cuando se tienen más de dos variables, la representación gráfica directa es dificil, por lo que es necesario recurrir a figuras o representaciones que, a través de un dibujo alusivo, resuman toda la infomación.

Uno de los métodos más conocidos fue propuesto por ANDREWS (1972). Cada uno de los puntos (individuos, poblaciones) del espacio p dimensional, se representa en un espacio de dos dimensiones mediante una curva sinusoidal, de forma que individuos similares están representados por curvas similares.

Las curvas utilizadas son series de Fourier finitas. Un vector p dimensional $x' = (x_1, x_2, \ldots, x_p)$ vendría representado por la curva

$$f_x(t) = (x_1 / \sqrt{2}) + x_2 \operatorname{sen} t + x_3 \cos t + x_4 \operatorname{sen} 2t + x_5 \cos 2t + \ldots$$

que se dibuja para unos pocos valores de t en el rango $-\pi < t < \pi$.

Una solución más simple del problema consiste en representar los perfiles de las distintas observaciones. Un gráfico de perfiles (BERTIN, 1967) es una representación

bidimensional en la que a cada variable se le asigna una barra vertical cuya altura es proporcional al valor que toma en la variable. El método puede resultar útil para encontrar grupos con perfiles similares.

Otro tipo de gráficos serían aquellos en que cada observación está representada mediante una figura, la diferencia en las figuras pone de manifiesto las diferencias en las observaciones. Podemos incluir en este grupo por ejemplo los "glyph" y "metroglyph" (ANDERSON. 1957), triángulos (PICKETT and WHITE, 1966), polígonos (SIEGD et al, 1971), estrellas (WELSCH, 1976), caras de Chernoff (CHERNOFF, 1973).

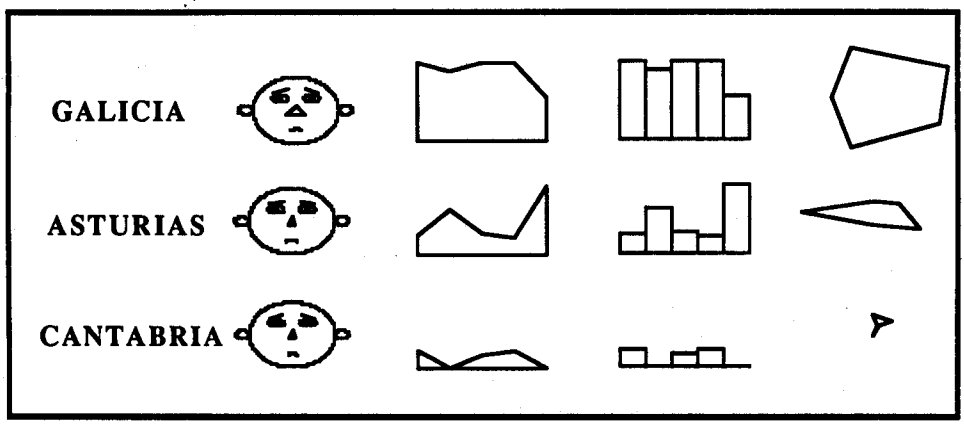

Figura 1:Algunos ejemplos de gráficos multivariantes (caras de Chernoff, perfiles y estrellas).

Un "glyph" es un círculo de radio fijo con p radios de varias longitudes que representan a cada una de las p coordenadas o variables que pueden ser cualitativas o cuantitativas.

Las estrellas (STAR) son muy similares a los anteriores, se diferencian en que los radios se colocan igualmente espaciados y se unen los puntos extremos de éstos. (fig. 1).

El método de las caras consiste en asignar un rasgo facial a cada una de las variables sobre una cara esquematizada. El método intenta aprovechar la capacidad humana para reconocer y diferenciar los rostros. Naturalmente la asignación de los rasgos a la variables puede ser arbitraria por lo que pueden obtenerse representaciones diferentes para el mismo conjunto de datos. Así mismo, varios autores dan diferentes esquematizaciones del rostro (CHERNOFF and RIZVI, 1975; FIENBERG, 1979; EVERITT, 1978).

Aunque se trata de un método bastante reciente, algunos programas para ordenador ya lo incluyen (SYSTAT, etc.).

La única limitación está en el número de variables que se pueden utilizar

(aproximadamente 20) aunque utilizando caras asimétricas es posible doblarlo.

En general, todas estas reperesentaciones gráficas son artificiales por lo que, en muchos casos, la interpretación no es inmediata y es necesaria bastante práctica.

Para evitar las diferencias en las escalas de medida conviene estandarizar los datos (restar la media y dividir por la desviación típica) en aquellos casos en que las variables están medidas en escalas muy diferentes.

TECNICAS MULTIVARIANTES CON RESULTADOS GRAFICOS

A diferencia de los métodos anteriores, este conjunto de técnicas no realiza una transcripción directa de las observaciones a un gráfico, sino que el gráfico es el resultado de un procedimiento matemático complejo que es necesario realizar con un ordenador.

Es por esta razón por la que este tipo de métodos ha comenzado a desarrollarse en los últimos años a pesar de que sus bases matemáticas se encuentran en la bibliografía desde principios de siglo (PEARSON, 1901).

Los gráficos resultantes en este caso son fundamentalmente de dos tipos:

- **Dendogramas y árboles de clasificación.**
- **Gráficos cartesianos.**

Los dendogramas son árboles que resultan tras sucesivos procesos de agrupamiento. Los procedimientos de este tipo son conocidos como Análisis de Clusters y están incluidos en la mayor parte de los programas para ordenador, por lo que su utilización está bastante extendida. Son muy útiles en la observación de la similaridad entre los individuos pero no es posible conocer las variables responsables de la clasificación.

En Economía se han utilizado para la clasificación de diferentes zonas geográficas de acuerdo con sus características económicas y en marketing para clasificar diferentes grupos de consumidores.

En cuanto a las técnicas cuyo resultado final es una representación cartesiana de puntos que representan a los entes estudiados, el procedimiento básico es el Análisis de Componentes Principales cuyas bases matemáticas fueron propuestas por PEARSON (1901) y que fue desarrollado por HOTELLING (1933), RAO(1964), etc...

Desde su aparición ha habido diferentes enfoques del método. El enfoque probabilístico, en el que se pretende la estimación de las componentes principales poblacionales, y el

enfoque geométrico o descriptivo, en el que se intenta reducir la dimensión; es este último punto de vista el que nos interesa, ya que no tiene, en principio, restricciones en las distribuciones de variables a estudiar.

ANALISIS DE COMPONENTES PRINCIPALES (ACP)

Una componente principal no es más una combinación lineal de las variables originales, más generalmente, el análisis busca combinaciones lineales que resuman los datos con la mínima pérdida de información posible. Además las componentes principales son incorreladas entre si y de varianza máxima, es decir, cada una de ellas se situa en la dirección de máxima variabilidad que además sea ortogonal a todas las componentes anteriores. Naturalmente la primera componente principal se situa en la dirección de máxima variabilidad global y habrá a lo sumo tantas Componentes Principales como variables observables.

La situación en dos dimensiones podría resumirse en la figura 2

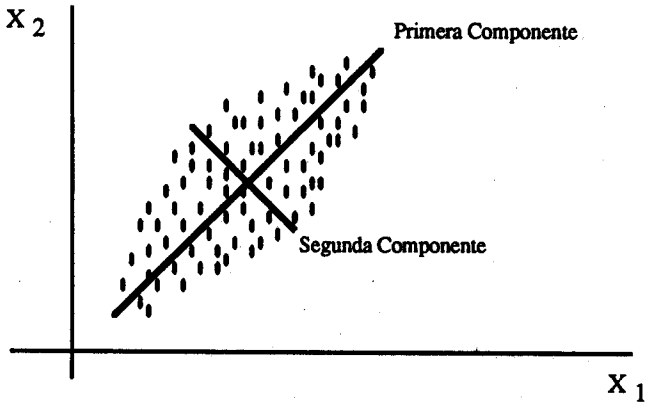

Figura 2: Esquema de las Componentes Principales en dos dimensiones.

Las CP resumen la estructura de las relaciones entre las variables ya que cuanto mayores sean las diferencias en sus longitudes (varianzas) más fuertemente relacionadas están las variables originales.

Por otra parte, si proyectamos las observaciones sobre la primera CP, la pérdida de información es mínima y la distancia entre dos observaciones sobre las proyecciones reproduce casi perfectamente la distancia original.

Está claro que la verdadera importancia de las proyecciones está en más de dos

dimensiones. Buscaremos la proyección en un subespacio de dimensionn reducida de forma que las proyecciones de las observaciones sobre este subespacio reproduzcan casi perfectamente las configuraciones del espacio multidimensional.

En el espacio de dimensión reducida será posible interpretar la proximidad entre los puntos como similaridad en sus características.

La solución matemática se obtiene de la minimización de la suma de cuadrados de las distancias de cada punto a la CP medidas perpendicularmente sobre ésta. El resultado final viene dado por la descomposición espectral de la matriz de dispersión que puede estar representada por $X'X$, la matriz de varianzas-covarianzas o la matriz de correlaciones.

Entonces, podemos descomponer la matriz como

$$X'X = V \Lambda V'$$

donde Λ es la matriz de vectores propios y V es la matriz de vectores propios normalizados de $X'X$. Los coeficientes de cada componente principal vienen dados por las columnas de V. Los valores propios están asociados con la longitud(varianza o dispersión) de las Componentes.

El porcentaje de la información que tiene cada Componente (% de varianza absorbida) se calcula dividiendo el valor propio asociado a cada componente por la suma de todos ellos.

El método descrito se extendido mucho entre todos los investigadores debido a que funciona en casi todos los programas de ordenador que existen en el mercado aunque sus aplicaciones se confunden, en muchos casos con las del Análisis Factorial cuyos planteamientos sos distintos.

La principal limitación del ACP está en que solamente representa individuos o variables por separado, siendo dificil la interpretación de las relaciones entre ambos conjuntos; además, su rendimiento óptimo sè produce para variables cuantitativas siendo su funcionamiento más deficiente con variables cualitativas (en forma de variables ficticias, codificadas, etc...).

Una completa descripción de las Componentes principales puede encontrarse por ejemplo en CUADRAS(1981), MARDIA et al (1979), JOLLIFFE(1986) y SEBER(1984).

Otro tipo de técnicas que sólo representan individuos o variables son las conocidas como Análisis de Proximidades (Análisis de Coordenadas Principales y 'Multidimensional Scaling'). La idea básica consiste en trabajar con una matriz de similaridades entre objetos intentando repreducir, sobre un espacio euclídeo, una configuración que reproduzca las

distancias de los objetos de partida.

El Análisis de Proximidades procede fundamentalmente de la Psicología y tiene su origen en los artículos de TORGERSON(1952), SHEPARD(1962), KRUSKAL(1964), GUTTMAN(1968).

En Economía se ha utilizado en el campo del 'Marketing'.

TECNICAS DE REPRESENTACION SIMULTANEA

La idea principal es similar a la del Análisis de Componentes Principales (Reducción en la dimensión de los datos), la diferencia se encuentra que ahora los individuos y las variables se representan conjuntamente sobre un espacio de dimensión reducida de forma que se tenga una visión más general de la matriz de datos.

Ahora se podrán interpretar las posiciones relativas de individuos y variables por separado y las conexiones entre ambos conjuntos.

La técnica más extendida dentro de este grupo es el denominado Análisis Factorial de Correspondencias (AFC) (BENZECRI, 1972, GREENACRE, 1984) que describiremos más tarde ya que puede considerarse como un caso particular de los conocidos como Métodos Biplot (GABRIEL, 1971).

Según la definición que el propio autor da en algunos de sus artículos

Un Biplot es una representación gráfica de una matriz $\mathbf{X}_{(n \times p)}$ por medio de unos marcadores $\mathbf{g}_1, \ldots\ldots, \mathbf{g}_n$ para sus filas y $\mathbf{h}_1, \ldots\ldots, \mathbf{h}_n$ para sus columnas, elegidos de tal forma que el producto interno $\mathbf{g}_i'\mathbf{h}_j$ represente el elemento x_{ij} de la matriz $\mathbf{X}$.

Si consideramos los marcadores $\mathbf{g}$ como filas de una matriz $\mathbf{G}$ y los $\mathbf{h}$ como filas de una matriz $\mathbf{H}$, entonces $\mathbf{GH'}$ representa la matriz de partida.

El Biplot se diferencia del resto de las técnicas de repesentación de datos en que dibuja simultáneamente filas y columas mientras que el resto dibuja sólo filas o sólo columnas.

Cualquier matriz de rango r puede factorizarse de la forma $\mathbf{X} = \mathbf{GH'}$ con $\mathbf{G}_{(n \times r)}$ y $\mathbf{H}_{(p \times r)}$ ambas de rando r .

Si la matriz de partida es de rango 2 puede realizarse una factorización exacta en dos dimensiones, si es de rango 3 en tres dimensiones, etc ...

Si la matriz es de rango superior a tres y queremos hacer una representación bi o tridimensional necesitamos recurrir a la aproximación a bajo rango por mínimos cuadrados (GABRIEL, 1971).

GOLUB y REINSCH (1970) hacen aproximaciones a bajo rango partiendo de la

descomposición en valores singulares propuesta por ECKART y YOUNG (1936).

Si **X** es una matriz de números reales se puede descomponer de la forma

$$\mathbf{X} = \mathbf{U}\,\Sigma\,\mathbf{V}'$$

donde

U son los vectores propios de **XX'**; **V** son los vectores propios de **X'X** y Σ es la matriz diagonal que contiene los valores singulares de **X** (que coinciden con la raiz cuadrada de los valores propios de **X'X** o **XX'**).

La mejor aproximación en rango dos se consigue utilizando solamente las dos primeras columnas de **U** y **V** en la descomposición en valores singulares.

$$\mathbf{X}_{(2)} = \mathbf{U}_{(2)}\,\Sigma_{(2)}\,\mathbf{V}_{(2)}$$

La descomposición Biplot puede hacerse tomando

$$\mathbf{G}_{(2)} = \mathbf{U}_{(2)}\,\Sigma_{(2)}^{\alpha} \qquad \mathbf{H}_{(2)} = \mathbf{V}_{(2)}\,\Sigma_{(2)}^{1-\alpha}$$

donde α puede tomar distintos valores entre 0 y 1.

Si $\alpha = 0$ se obtiene el denominado **GH'-Biplot** que produce una representación óptima para las columnas pero pobre para las filas y si $\alpha = 1$ obtenemos el **JK'-Biplot** que produce una representación óptima para las filas pero muy pobre para las columnas.(GABRIEL, 1971).

El autor realiza entonces una representación conjunta de las filas y columnas de la matriz de datos aunque no pueda considerarse una representación simultánea en sentido estricto ya que no consigue la misma bondad de representación para las filas que para las columnas.

Los métodos Biplot y el Análisis de Componentes principales están íntimamente relacionados ya que ambos son dos variantes de la misma idea (la reducción de la dimensión de los datos con pérdida de información mínima) y ambos proceden de la diagonalización de las matrices de dispersión. (Vease por ejemplo GABRIEL, 1971, 1981; GALINDO, 1986; GALINDO y CUADRAS, 1986).

Otra técnica de representación simultánea muy extendida es la conocida como Análisis Factorial de Correspondencias (AFC) que produce una representación geométrica de las filas y las columnas de una tabla de contingencia. El AFC representa los perfiles (frecuencias relativas) de las líneas de una matriz de datos positivos, generalmente

frecuencias. Puede considerarse como una representación Biplot de la siguiente matriz

$$Y = R^{-1/2} X C^{-1/2}$$

donde **R** y **C** son matrices diagonales en las que aparecen los correspondientes totales de filas y columnas.

Esta técnica se ha desarrollado de forma autónoma en la literatura científica francesa y ha sido aplicada en muchos campos de la Ciencia entre los que se incluye la Economía (KONTSANDREAS y DOSSON, 1985; ARBACHE, 1982; HAGGAG, 1983; BENZECRI, 1983, 1985; ALAWIEH, 1986; DARONKH, 1989).

GALINDO (1986), utilizando las ideas básicas de los métodos Biplot y del AFC propone una nueva técnica que denomina **HJ-Biplot** y que utiliza como marcadores para filas $G=U\Sigma$ y como marcadores para las columnas $H=V\Sigma$.

Se demuestra que en el HJ-Biplot se consigue la misma calidad de representación para filas que para columnas y que, además, es más alta que en los Biplots clásicos de Gabriel. Se demuestra tambien que ambos marcadores pueden representarse en el mismo sistema de referencia ya que cada coordenada para una fila puede expresarse en función de las de las columnas y viceversa. El HJ-Biplot proporciona tambien las mejores representaciones β-baricéntricas en el sentido propuesto por LEBART y col.(1983).

En principio se interpreta la proximidad entre individuos como similaridad, los ángulos entre dos ventores que unen dos variables con el origen como correlación y la proximidad de un grupo de marcadores fila a un marcador columna en términos de preponderancia. Más concretamente, dos variables separadas por un ángulo pequeño están fuertemente relacionadas y un grupo de individuos próximo a una variable indica que los individuos han tomado valores preponderantes para esta variable.

Proyectando los individuos sobre los vectores que representan a las variables es posible determinar la distribución aproximada que los individuos toman sobre esa variable (GABRIEL, 1990).

A pesar de que la interpretación parece sencilla, hay que hacerla con cuidado ya que las posiciones de los puntos sobre los planos principales puede ser sólo aparente; para evitar los posibles errores de interpretación GALINDO y CUADRAS (1986) proponen varias medidas que serán útiles en la interpretación de los gráficos resultantes en un HJ-Biplot. Las más importantes son:

-Contribuciones del elemento al factor(eje): Son las partes tomadas por cada

elemento en la varianza explicada por el factor.

-Contribuciones a la traza: Parte tomada por cada elemento en la dispersión total.

-Contribuciones del factor al elemnto: Parte de la dispersión de un elemento explicada por el factor.

-Calidad de representación: Parte de la información de un elemento contenida en un eje, plano, etc...

Describen, tambien, la utilización de estas medidas en la interpretación de los resultados.

La más importante de todas ellas es, quizás, la Contribución del factor al elemento ya que permite determinar las variables responsables de las configuraciones de individuos resultantes al proyectar los puntos sobre los ejes.

Evidentemente, la interpretación sobre los distintos ejes y planos factoriales se realiza con mayor fiabilidad para aquellos elementos bien representados.

El mayor problema que puede plantearse a la hora de aplicar los métodos Biplot es el de las diferentes escalas de medida en las variables; por esta razón, en muchos casos es necesaria la utilización de centrados y estandarizaciones de las variables antes de la aplicación de la representación simultánea.

Muchos autores han estudiado los efectos de las transformaciones realizadas en los datos de partida (LEFEBVRE, 1976; OKAMOTO, 1972; CUADRAS, 1982; CHARD, 1976).

A pesar de esta aparente restricción, fácilmente resoluble, el método HJ-Biplot puede utilizarse para resumir la información contenida en cualquier matriz de datos, por tanto este método extiende el AFC en el sentido de que no se restringe a matrices de datos positivos o más concretamente a matrices de frecuencias.

Dado el caracter multidimensional de los datos económicos parece que la técnica puede ser útil en la descripción e interpretación de datos en Economía.

Los Biplots se han aplicado en muchas situaciones:

-Clasificación de paises según el consumo de proteinas (GABRIEL, 1982).

-Clasificación de especies en Biología (PEREZ MELLADO y GALINDO, 1986).

-Diagnóstico de modelos que pueden ajustarse a una matriz de datos (BRADU y GABRIEL, 1978; TSIANCO, 1984).

-Meteorología (TSIANCO, 1984).

-Medicina (GABRIEL, 1990; ORFAO y col.1988, PEDRAZ, C. et al. ,1985).

-Diagnóstico en regresión y Análisis Discriminante (GALINDO y col., 1990; VICENTE-VILLARDON y col., 1990).

-Diagnóstico en Análisis Multivariante de la Varianza (TSIANCO, 1981).

-Control de calidad en alimentación (VICENTE y col., 1990; SANTOS y col., 1990).

APLICACION PRACTICA

Con el fin de ilustrar los métodos descritos hemos realizado una aplicaión sencilla.

El objetivo principal consiste en encuadrar la Comunidad Autónoma de Castilla-León dentro del resto de las comunidades, analizando variables de producción agrícola.

Para el estudio se han considerado las 17 comunidades autónomas como poblaciones y la superficie dedicada a diversas formas de aprovechamiento del suelo como variables. Las definiciones precisas de cada una de dichas variables estudiadas pueden encontrarse en el Anuario de Estadística Agraria publicado por el Ministerio de Agricultura, Pesca y Alimentación; todas ellas corresponden al último año del que se dispone información : 1987.

Se consideran únicamente 5 variables procedentes de una clasificación muy amplia del territorio de cada comunidad.

1: Total de tierras cultivadas. 2: Prados y pastizales 3: Superficie forestal.

4: Asociación de monte abierto con diferentes cultivos. 5 : Otras (Terrenos improductivos, no agrícolas, superficie de rios y lagos, etc...

Como técnica de representación simultánea de comunidades y superficies se ha utilizado el HJ-Biplot. En la figura 3 aparece la representación obtenida para los tres primeros planos principales. (I-II, I-III, II-III).

Los porcentajes de varianza explicados por los tres primeros ejes es de 54.54%, 32.57% y 12.81% respectivamente por lo que latasa de inercia en el espacio es del 99.92% de la varianza total, se ha conseguido reducir la dimensión de 5 a solo 3.

La tabla 1 muestra las contribuciones relativas del factor al elemento, mediante las cuales es posible conocer cuales son las variables responsables de la posición de los ejes y por tanto de la configuración obtenida sobre ellos.

El plano I-II (Figura 3a) está definido principalmente por las superficies cultivadas y las forestales (eje I) y las no productivas (eje II).

Se ve claramente como a la derecha del eje I se situan aquellas comunidades en las que se dedican porcentajes más altos a la tierras de cultivo (Baleares , Murcia, Castilla-La Mancha, Andalucía, Castilla-León, etc) mientras que en el otro extremo se concentran aquellas comunidades que dedican un porcentaje más alto de sus tierras a la producción forestal (Galicia, Asturias, Cantabria, Pais Vasco, etc...).

AUTONOMIAS

	EJE1	EJE2	EJE3		EJE1	EJE2	EJE3
GALICIA:	670	230	110	ASTURIAS:	790	190	19
CANTABRIA:	970	7	19	P. VASCO:	850	78	69
NAVARRA:	0	210	780	RIOJA:	140	32	830
ARAGON:	420	550	38	CATALUÑA:	150	320	530
BALEARES:	700	130	170	CAST-LEON:	710	44	240
MADRID:	300	510	190	LA MANCHA:	850	150	7
VALENCIA:	100	180	720	MURCIA:	840	79	82
EXTREMADURA:	0	200	770	ANDALUCIA:	700	300	4
CANARIAS:	27	930	45				

VARIABLES

	EJE1	EJE2	EJE3		EJE1	EJE2	EJE3
CULTIVOS:	840	160	1	PRADOS:	310	4	690
FORESTAL:	690	200	110	ASOCIAC.:	52	40	120
OTRAS:	22	920	56				

Tabla 1: Contribuciones relativas de los tres primeros ejes a las 17 autonomías y a las 5 variables.

El eje II está definido por las tierras que no se dedican a la producción agrícola y separa Canarias del resto de las comunidaes ya que tiene un mayor porcentaje de tierras improductivas (Fig. 3a).

Si proyectamos los puntos que representan a las poblaciones sobre cada uno de los vectores que representan a las variables obtendremos un gradiente aproximado de la importancia de las superficies en las distintas comunidades. Por ejemplo, proyectando sobre el vector que representa a la superficie Forestal vemos como Galicia, Pais Vasco, Cantabria y Asturias tienen mayor número de Bosques seguidas por Cataluña, Navarra, Extremadura, Valencia, Andalucía, La Rioja, Castilla-León, Baleares, Murcia, Castilla-La Mancha, Aragón, Madrid y por último Canarias que sería la comunidad con menor porcentaje de terreno forestal.

La tabla 2 muestra las comunidades ordenadas de acuerdo con los porcentajes dedicados a cada una de las categorías estudiadas.

Obsevamos como efectivamente el gráfico proporciona una ordenación aproximada de las comunidades con respecto a la variable que está bien representada en el plano factorial. Las ordenaciones de las poblaciones son más claras en los puntos extremos que en los puntos intermedios ya que generalmente son aquellos los que tienen calidades de representación más altas (GALINDO Y CUADRAS, 1986).

T. cultivo	Prados y Pastizal	T. Forestal	Otros
MURCIA	CANTABRIA	GALICIA	CANARIAS
CAST-LA MANCHA	NAVARRA	PAIS VASCO	MADRID
BALEARES	ASTURIAS	CANTABRIA	ASTURIAS
ANDALUCIA	LA RIOJA	ASTURIAS	ARAGON
CASTILLA-LEON	EXTREMADURA	CATALUÑA	C. VALENCIANA
C. VALENCIANA	MADRID	C. VALENCIANA	MURCIA
NAVARRA	CASTILLA-LEON	EXTREMADURA	LA RIOJA
LA RIOJA	PAIS VASCO	BALEARES	BALEARES
ARAGON	ARAGON	NAVARRA	CATALUÑA
MADRID	GALICIA	ANDALUCIA	CANTABRIA
EXTREMADURA	CATALUÑA	MURCIA	CASTILLA-LEON
CATALUÑA	CAST-LA MANCHA	ARAGON	ANDALUCIA
GALICIA	ANDALUCIA	CASTILLA-LEON	CAST-LA MANCHA
CANARIAS	CANARIAS	LA RIOJA	PAIS VASCO
PAIS VASCO	MURCIA	CAST-LA MANCHA	EXTREMADURA
CANTABRIA	C. VALENCIANA	MADRID	NAVARRA
ASTURIAS	BALEARES	CANARIAS	GALICIA

Tabla 2: Ordenación de los porcentajes dedicados a cada categoria.

La contribución relativa del factor al elemento más alta en el eje III se obtiene para los Prados y pastizales por lo que es esta variable la que define a este eje (Fig. 3a y 3b).

Según lo expuesto a partir de la observación de los gráficos, la Comunidad de Castilla-León se encuentra situada entre aquellas comunidades con altos porcentajes de tierra cultivada, bajos de terreno forestal, intermedios en cuanto a otras superficies e intermedios en la zona superior para los prados y pastizales.

Castilla-León se encuentra próxima en la representación del plno I-II a comunidades como Castilla-La Mancha, Extremadura, Valencia, Andalucia etc.. por lo que podemos inferir que los porcentajes de dedicación a pastos, Terreno Forestal y Superficie Improductiva son parecidos en todas estas comunidades. Se diferencia de Cananarias y Madrid porque éstas son las comunidades con mayor porcentaje de Otros terrenos y de las comunidades de la cornisa cantábrica por el Terreno Forestal.

Sobre el plano I-III (Fig. 3b) vemos como Castilla-León se situa en una posición intermedia con respecto a los Prados y Pastizales.

La variable 4 (Asociación de monte con plantas herbaceas) aparece representada sobre el origen en todos los planos considerados por lo que su calidad de representación es muy baja y no se ha tenido en cuenta a la hora de realizar la interpretación.

Por tanto hemos logrado el objetivo de encuadrar la Comunidad de Castilla-León dentro del ámbito Nacional en base a la información suministrada por variables de

aprovechamiento del suelo.

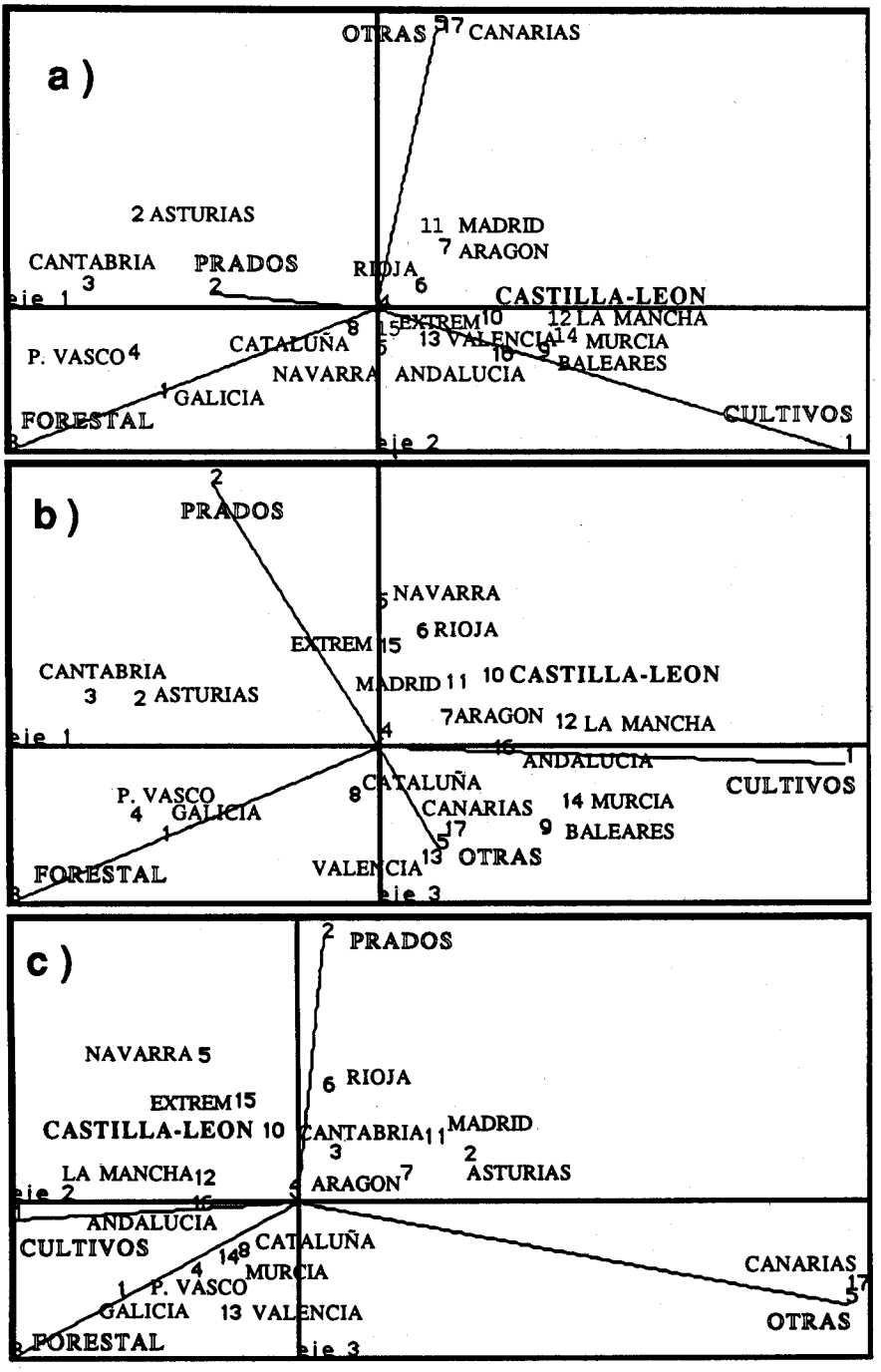

Figura 3: Proyección sobre los tres ejes principales en el HJ-Biplot.

REFERENCIAS

-ALAWIEH, A.A. (1986). Le Marché du Riz en Europe Occidentale". *Cahiers de L'Analyse des Données*. Vol XI, nº 3, 355-362.

-ANDERSON,E. (1974) A semi-graphical method for the analysis of complex problems. *Proc. Nat. Acad. Sci. U.S.A.* **43**, 923-927.

-ANDREWS, D. F. (1972) Plots of high dimensional data. *Biometrics* **28**, 125-136.

-ARBACHE, Ch.(1982) Evolution des productions de L'Agriculture Syrienne par regions, de 1960 a 1977. *Cahiers de L'Analyse des Données*. Vol VII, nº 1, 67-91.

-BARNETT, V. (ed.) (1980) *"Interpreting Multivariate Data"*. Wiley.

-BENZECRI, J. P. et al. (1973) *"L'Analyse des Données"* Dunod.

-BENZECRI, J. P. et al. (19864) "Pratique de *L'Analyse des Données en Economie* " Dunod.

-BENZECRI, J. P. (1983) "Definition et mesure de l'élasticité des consommations" *Cahiers de L'Analyse des Données*. Vol VIII, nº 1, 73-88.

-BERTIN, T. (1967) *Semiologie Graphique*. Gauthier Villars.

-BRADU, D. and GABRIEL, K. R.(1978) The Biplot as diagnostic Tool for Models of Two-Tables. *Technometrics* **20**, 47-68.

-CHARDY, P. et al. (1976). Application of Inertia Methods to Benthic Marine Ecology. Practical Implications of the Basic Options. *Estuarine and Coastal Marine Science*. **4**, 179-205.

-CHERNOFF, H.(1973) Using Faces torepresent points in k-dimensional space graphically. *J. Amer. Statist. Assoc.*, **68**, 361-368.

-CHERNOFF, H. and RIZVI, M.H.(1975) Effect on classification error of dandom per4mutations of Features in representing Multivariate Data by Faces. *J. Amer. Statist. Assoc.* , **70**, 548-554.

-CUADRAS, C.M. (1983) *Métodos de Análisis Multivariante*. Eunibar.

-DARONKH, D. (1989) Marché immobilier des cinquante premières villes de province et rentabilite locative de l'ancien Paris. *Cahiers de L'Analyse des Données*. Vol XIV, nº 4, 393-400.

-EVERITT, B. (1978) *Graphical Techniques for Multivariate Data*. Heinemann Educational Books.

-FIENBERG, S.E. (1979) Graphical Methods in Statistics. *The American Statistician*, **33**, 165-177.

-GABRIEL, K. R. (1971) 'The Biplot Graphic Display of Matrices with Application to Principal Components Analysis'. *Biometrika*. Vol 58,3. Pgs 453-467.

-GABRIEL, K. R. and ODOROFF, C. L. (1990) 'Biplots in Biomedical Research'. *Statistics in Medicine*. Vol. 9, 469-485.

-GALINDO, M. P. (1986) 'Una alternativa de representación simultánea: HJ-BIPLOT'. *QUESTIIO*. V.10, nº 1. Pgs: 13-23.

-GALINDO, M. P.; CUADRAS, C. M. (1986) 'Una extensión del método Biplot y su relación con otras técnicas'. *Publicaciones de Bioestadística y Biomatemática*. Universidad de Barcelona.

GALINDO, M.P.; VICENTE-VILLARDON, J.L.; BARRERA, I., GARCIA, A.(1990) Una contribución al dignóstico de la colinealidad basada en los métodos Biplot. *Cuadernos de Bioestadística y sus Aplicaciones Informáticas*. **8(1)**, 53-61.

-GOLUB, G.H. and REINSCH, (1970) Singular Value Descomposition and Least Squares Solution. *Numer. Math.* **14**, 403-420.

-GREENACRE, M. J. (1984) *Theory and Application of Correspondence Analisys*. Academic Press. London.

-GUTTMAN, L.A. (1968) A General Nonmetric Technique for finding the Smallest Coordinate space for a Configuration of Points. *Psychometrika*. **33**, 469-506.

-HAGGAG, A. (1983) L'Analyse des données boursières. *Cahiers de L'Analyse des Données*. Vol VIII, nº 2, 205-220.

-HOTELLING, H. (1933) Analysis of a Complex of Statistical variables into Principal Components. *J. Educ. Psychol.* **24(6)**, 417-441.

-JOLLIFFE, I. T.(1986) *Principal Component Analysis*. Springer-Verlag. New York.

-KONTSANDREAS, K. et DOSSON, F. (1985) Une enquête sur les crêateurs d'enterprises. *Cahiers de L'Analyse des Données*. Vol X, nº 4, 425-436.

-KRUSKAL, J.B. (1964) Multidimensional Scaling by Optimizing Goodness-of-fit to a Nonmetric Hypothesis. *Psychometrika.* **29**, 1-27.
-LEBART, L.; MORINEAU, A. y FENELON, J. P. (1983) *Tratamiento estadístico de datos.* Marcombo.
-LEFEBRE, J. (1976) Introduction aux Analyses Statistiques Multidimensionelles. Masson.
MARDIA, K.V.; KENT, J.T. and BIBBY, J.M. (1979) *Multivariate Analysis.* Academic Press.
-OKAMOTO, M. (1972) Four Techniques of Principal Component Analysis. *J. Japan Stat. Soc.* **2**, 63-69.
-ORFAO, A. et al. (1988) "Clinical and Inmunological Findings in Large B-Cells Chronic Lymphocityc Leukaemia". *Clinical Inmunology and Inmunopathology.* **46**, 177-185.
-ORLOCI, (1967) Data Centering: E review and Evaluation with reference to Component Analysis. *Sist. Zool.* **3(16)**, 208-212.
-PEARSON, K. (1901) On Lines and Planes of Closest Fit to Systems of Points in Space. *Phil. May. Ser.* **6, 2 (11)**, 559-572.
-PEDRAZ, C. et al. (1985) "Estudio de los factores Socio-culturales que influyen en la elección de la lactancia natural". *Archivos de Pediatría.* **36**, 469-477.
-PEREZ-MELLADO, V.; GALINDO, M. P. (1986) 'Biplot Graphic Display of Iberian and North African Populations of Podarcis'. *Studies in Herpetology, Rocek Z.* (Ed.) 197-200.
-PICKETT, R. and WHITE, B.W. (1966) Constructing Data Pictures*proc. 7th Nat. Symp. Inform. Display*, 75-81.
-RAO, C.R. (1964) The Use and Interpretation of Principal Component Analysis in Applied Research. *Sankya. Ser. A.* **26**, 329-358.
-SANTOS, C.; MUÑOZ,S.S.; GUTIERREZ, Y.; HEBRERO, E.; VICENTE-VILLARDON, J.L.; GALINDO, P. and RIVAS, J.C. (1990) Craracterization of young red wines by application of HJ-Biplot Analysis to anthocyanin profiles. *J. Agric. Food Chem.* (en prensa).
-SEBER, G.A.F. (1984) *Multivariate Observations.* Wiley.
-SHEPARD, R.N. (1962) The Analysis of Proximities: Multidimensional Scaling with a unknown distance function. *Psychometrika.* **27**, 219-246.
-SIEGEL, J. H.; GOLDWYN, R.M. and FRIEDMAN, H.P. (1971) Pattern and process of the evolution of human septic shock. *Surgery*, **70**, 232-245.
-TORGERSON, W.S. (1952) Multidimensional Scaling, I: Theory and Method. *Psichometrika.* **17**, 401-419.
-TSIANKO, M.C. (1981) Modelling Temperature Data: An Ilustration of the use of Biplots in Nonlinear Modelling. *University of Rochester, Statist. Tech. Rep.* 81/15
-TSIANKO, M.C. and GABRIEL, K.R. (1984) Modelling Temperature Data: An Ilustration of the use of Biplots and Bimodels in Nonlinear Modelling. *Journal of Climate and Applied Meteorology.* **23**, 787-799.
-VICENTE-VILLARDON, J.L. et al. (1990).HJ-Biplot Analysis and Logistic Modelling: Complementary use in the Selection of Variables and Discrimination among Groups. XVth International Biometric Conference. Budapest.
-WELSCH, R.E. (1976) Graphic for data analysis. *Comput. Graphics*, **2**, 31-37.