IV Congreso de Economía Regional de Castilla y León **Burgos 1994, Comunicaciones 3:** 1518-1533

UN ESTUDIO DE LAS DISPARIDADES ECONOMICO-SOCIALES DE LAS PROVINCIAS ESPAÑOLAS DESDE EL ANALISIS DE CONGLOMERADOS

FERNANDEZ DE LA MORA, Julia
FERNANDEZ-ABASCAL TEIRA, Hermenegildo
Profesores del departamento de Economía Aplicada (Estadística y Econometría) de la Univesidad de Valladolid.

En este trabajo se estudian las provincias españolas a partir de un conjunto de variables que tratan de recoger una cierta idea del bienestar social que disfrutan sus habitantes. La metodología utilizada en este análisis de datos multivariantes es el análisis de conglomerados. De acuerdo con esto, se clasificarán las provincias por su grado de bienestar, creando grupos más o menos homogéneos en cuanto a la calidad de vida.

1.- EL ESTUDIO DEL BIENESTAR SOCIAL

Concretar la idea de bienestar individual y, por agregación, de bienestar social es tarea difícil, cuando no imposible (en Zarzosa (1992) y Pena (1977) se puede encontrar un resumen de las distintas interpretaciones del concepto de bienestar social). El bienestar no es un estado material y objetivo, sino, más bien, una sensación personal y subjetiva de satisfacción. Las causas de este bienestar hay que buscarlas, por un lado, en unas determinadas condiciones materiales (ingresos, situación laboral, nivel educativo, dotaciones de la vivienda, etc.) y, por otro lado, en apreciaciones subjetivas del individuo. Mientras que las primeras pueden ser observadas y medidas, no ocurre lo mismo con las consideraciones subjetivas. Este es uno de los motivos, no el único, que dificulta cualquier intento de medición del bienestar.

La idea de bienestar puede concretarse más si la comparamos con la idea de desarrollo. Durante las décadas posteriores a la Segunda Guerra Mundial, los economistas y políticos pusieron énfasis en el desarrollo económico como vía para conseguir el bienestar individual y, por ende, colectivo. La experiencia de aquellos años no fue todo lo positiva que se esperaba: el crecimiento económico por el crecimiento económico no se tradujo siempre en una mejora en la calidad de vida de los individuos. Por el contrario, en muchos casos se agrandaron las diferencias sociales dentro de un mismo país y entre países, se agredió el medio ambiente con políticas industriales "sucias", pero económicamente rentables, se concentró a la población en grandes centros

urbanos inhabitables y se introdujo al individuo en situaciones altamente competitivas. A partir de los 70, tanto economistas como políticos quisieron poner freno a la situación, tratando de buscar desarrollos más racionales y menos agresivos, que incidieran más en las condiciones de vida de los individuos que en el incremento del PIB.

Mientras que la medición y valoración del desarrollo económico no presentaba problemas notables (siempre se podía echar mano de las grandes cifras macroeconómicas), la complejidad del concepto de bienestar hace difícil su medición, teniéndonos que conformar con modestas aproximaciones.Pero esta dificultad de medición no sólo es fruto del citado aspecto subjetivo que lleva consigo la idea de bienestar. Además, el sujeto del bienestar es, en último término, el individuo, considerándose el bienestar social como el agregado del bienestar de los distintos individuos. No se puede ocultar el problema que esta agregación implica.

Asimismo, mientras cualquier país genera una vasta información estadística sobre datos macroeconómicos que nos proporcionan una idea de su nivel de desarrollo, el panorama no es igual en cuanto a datos que pongan en evidencia su nivel de bienestar. Por un lado, existen dificultades para determinar qué variables miden este bienestar; por otro, para comprender los aspectos subjetivos que conlleva el bienestar harían falta encuestaciones individuales, siempre complejas y costosas.

En este afán de medir el bienestar se han desarrollado tres distintas metodologías o enfoques: el de las funciones de utilidad, el contable y el de los indicadores sociales. Es en este último enfoque donde se situa este trabajo.

Podemos entender los indicadores sociales como «*compendios de datos básicos que dan una medida concisa de la situación y cambios relativos a aspectos de las condiciones de vida de la población que son objeto de preocupación social... El procedimiento usual de elaboración de un indicador es resumir en forma de porcentaje, índice, tasa, valor medio o cualquier otra medida sintética la información extraída de dos o más datos estadísticos de menor grado de concisión*»(1). Los indicadores ponen en evidencia variables teóricas que subyacen en los fenómenos sociales, de los que la variable observada, el indicador, no sería sino un síntoma.

Este carácter de los indicadores, reflejos de una realidad subyacente no medible, puede constituir uno de los principales problemas de su utilización e interpretación: la ambigüedad.

Otras limitaciones de los indicadores sociales, según Zarzosa (1992), son:

- La escasez de datos estadísticos.

- La heterogeneidad de las fuentes.
- La ausencia de indicadores de percepción.
- El carácter desagregado de los indicadores sociales.

Cabe decir que, de todas estas limitaciones, la ausencia de indicadores de percepción constituye un escollo, a veces insalvable, para afrontar la medición del bienestar social desde una perspectiva "moderna".

En toda investigación sobre bienestar social por medio de indicadores sociales se plantea un problema previo: la selección de indicadores. Para facilitar este proceso de selección, se suele dividir el objeto de nuestro estudio, en este caso el bienestar, en parcelas o componentes que recogen ciertos aspectos de un todo. En el caso del bienestar social, existen diversos criterios empíricos de clasificación según el organismo que realice el estudio (Naciones Unidas, OCDE, ONU, INE, ...), si bien todas ellas recogen, de un modo u otro, las mismas componentes: salud, educación, vivienda, trabajo, ocio,... En un segundo paso se eligen indicadores referidos a cada una de las componentes del bienestar teniendo en cuenta unos determinados criterios técnicos y metodológicos.

Por último, cabe hacer un breve comentario sobre una cuestión de cierta relevancia que se esconde detrás de este trabajo y de todos aquellos que tratan de medir el bienestar entre provincias, regiones o paises. Como se dijo anteriormente, el sujeto del bienestar es el individuo, o bien, por motivos técnicos y metodológicos, la unidad familiar. Ahora bien, aquí no se va a comparar el bienestar entre distintas familias, sino el bienestar entre provincias, haciendo una primera simplificación que consiste en suponer que todas las familias de una misma provincia son idénticas en cuanto a sus condiciones de bienestar. (Es obvio que muchos de los indicadores agregados de nivel de desarrollo pueden esconder grandes desigualdades que aumentan la percepción subjetiva del malestar. También es verdad que estas distorsiones, patentes e intrínsecas en un enfoque contable y desarrollista del bienestar, pueden aminorarse con la utilización de indicadores no puramente monetaristas).

Por otra parte, la elección del territorio, en nuestro caso la provincia, como unidad de estudio no plantea únicamente el problema de homogeneización de todos los individuos o familias que en él habitan. Dado que el territorio es el lugar en el que se "vive" el bienestar, la población tiende a concentrarse en los focos más dinámicos de actividad al encontrar en ellos mayores oportunidades de bienestar. Esa concentración de los individuos buscando optimizar su bienestar puede producir un resultado no deseado: zonas urbanas industriales, degradadas medioambientalmente, con paro crónico, delincuencia,...

Para terminar, hay que incidir en el interés, no sólo

académico, de los estudios sobre el bienestar. Cualquier política económica y social que tenga como meta reducir las desigualdades sociales, deberá partir de un conocimiento de esa realidad desigual, actuando sobre los factores más relevantes de la misma.

2.- BREVE DESCRIPCION DEL ANALISIS DE CONGLOMERADOS Y OTRAS TECNICAS MULTIVARIANTES

En el estudio de los problemas económicos o de otra índole, nos encontramos normalmente con grandes masas de datos que nos proporcionan una información muchas veces excesiva y redundante. Las distintas técnicas de análisis multivariante tratan, básicamente, de reducir el número de datos, eliminando la información no relevante y repetida, y de detectar relaciones entre las variables o los individuos.

Una de las técnicas más utilizadas, y con menos sofisticación desde el punto de vista teórico, es el análisis de conglomerados, también denominado análisis *cluster* o taxonomía numérica.

El objetivo del análisis de conglomerados es clasificar n individuos, caracterizados por los valores que en ellos toman k variables, $X_1,\ldots,X_k$, en grupos de acuerdo con la mayor o menor semejanza entre los mismos. Cada grupo, en número no determinado, estará formado por individuos "parecidos" entre sí y "distintos" a los individuos de los otros grupos. Hay que dar una medida de similitud entre los individuos, que refleje globalmente la cercanía o lejanía entre todas las variables que los caracterizan: esta medida es la **distancia** (si las variables son cualitativas la medida utilizada es la **semejanza**).

Dos individuos, w_i y w_j $(i,j=1,\ldots,n)$, quedan caracterizados por $(x_{i1},\ldots,x_{ik})$ y $(x_{j1},\ldots,x_{jk})$, pudiéndose valorar de distinta forma la distancia entre los mismos. Lo más usual es la distancia euclídea,

$$d(w_i,w_j)=\sqrt{(x_{i1}-x_{j1})^2+\ldots+(x_{ik}-x_{jk})^2}\ ,$$

que presenta el problema de verse afectada por cambios de origen y de escala en las unidades de las variables.

Esto se puede solucionar de forma sencilla si se supone que las variables son independientes. En este caso, una simple tipificación de éstas, sirve para hacer válida "estadísticamente" la distancia euclídea, que quedará definida como:

$$d_T(w_i,w_j)=\sqrt{\left(\frac{x_{i1}-x_{j1}}{S_{x_1}}\right)^2+\ldots+\left(\frac{x_{ik}-x_{jk}}{S_{x_k}}\right)^2}\ .$$

Ahora bien, si las variables son dependientes, más en concreto, si están correlacionadas, lo anterior no es suficiente para dar una buena distancia "estadística". En este caso hay que corregir la distancia teniendo en cuenta la covarianza, que nos mide el grado de relación lineal entre las variables. En este sentido, Mahalanobis propuso en 1936 la distancia estadística general,

$$d_M(w_i, w_j) = [(\mathbf{X}_i - \bar{\mathbf{X}})' S^{-1} (\mathbf{X}_j - \bar{\mathbf{X}})]^{\frac{1}{2}},$$

donde $\mathbf{X}_i$ y $\mathbf{X}_j$ son los vectores columna con las coordenadas de w_i y w_j, $\bar{\mathbf{X}}$ es el vector columna con las medias de las variables y S es la matriz de varianzas-covarianzas.

Una vez elegida la distancia más adecuada, obtenemos la matriz de distancia n×n. A partir de esta matriz de distancias existen dos tipos de métodos para formar los conglomerados: los métodos jerárquicos y los no jerárquicos.

Los métodos jerárquicos parten de considerar tantos *clusters* como individuos, es decir, n conglomerados. En un primer paso se rebaja el número de *clusters* a n-1 formando un *cluster* con dos individuos, los dos más próximos, aquellos que presenten una menor distancia. A continuación, se vuelve a calcular la matriz de distancias entre n-1 individuos, n-2 *clusters* con un único individuo y uno con dos individuos. El problema está en calcular la distancia entre este *cluster* y el resto (las otras distancias son, obviamente, las de la matriz original). Para ello existen diversos métodos: distancia media (*average linkage*), distancia mínima (*single linkage*), distancia máxima (*complete linkage*), distancia mediana y distancia entre centroides, según se calcule la distancia entre dos *clusters* como la distancia media, mínima, máxima, o mediana entre los puntos de los dos *clusters*, o la distancia entre sus puntos "centrales", respectivamente.

Utilizando uno de estos métodos, se halla la matriz de distancias entre los n-1 *clusters*. A continuación se rebaja el número de *clusters* por unificación de los dos más próximos. El proceso es iterativo, calculando de nuevo la matriz de distancias entre las nuevas agrupaciones. En última instancia, se conseguiría agrupar los n individuos en un único *cluster*. El proceso se describe gráficamente mediante un dendograma, una representación arborescente que nos hace ver el proceso de formación de los *clusters*. Obviamente, el investigador tiene que saber parar el proceso en un momento determinado, con un número de conglomerados razonable.

En los métodos no jerárquicos se fija previamente un número determinado de conglomerados, p. Entonces, o bien se arranca de una partición aleatoria de los n individuos en p agrupaciones y progresivamente se van intercambiando puntos de una agrupación a otra por el criterio de buscar agrupaciones homogéneas, o bien, se elige un individuo para

cada *cluster* (*seed point*) y el resto de los individuos se clasifica en el *cluster* cuyo *seed point* sea más cercano.

Los métodos no jerárquicos no necesitan recalcular en cada paso la matriz de distancias, cosa que sí ocurría con los métodos jerárquicos. Esto hace que su utilización sea muy conveniente cuando se tienen grandes conjuntos de datos, pues su proceso de cálculo es mucho más sencillo y rápido que el de los métodos jerárquicos. Ahora bien, según Johnson y Wichern (1982), el hecho de que prefijemos el número de *clusters* y elijamos los *seed points* de forma arbitraria, puede generar *clusters* poco diferenciados, muy dispersos o disparatados.

El análisis de conglomerados no exige ninguna condición especial a las variables, dado que en lo fundamental es un método métrico y no estadístico. La única precaución que hay que tener es que la información no sea redundante, es decir, que dos variables distintas no aporten una misma información. Si dos o más variables tienen información común, esta estaría sobrevalorada en el cálculo de las distancias.

Por tanto, previamente a la realización de un análisis *cluster,* tenemos que depurar nuestros datos para evitar la información redundante. En Estadística existen diversos métodos para eliminar esta información superflua. Recordaremos aquí, brevemente, los tres métodos utilizados en este trabajo.

El **análisis de correlación** trata de averiguar el grado de relación lineal entre las variables. Dadas dos variables, X e Y, el coeficiente de correlación lineal de Pearson,

$$r=\frac{S_{xy}}{S_x\,S_y},$$

nos proporciona una medida de su correlación. Valores de r cercanos a 1 ó -1 nos indican la existencia de una cierta relación lineal creciente o decreciente, respectivamente, entre las variables; valores cercanos a cero señalan ausencia de correlación. De esta forma, si tenemos dos variables con un r^2 próximo a la unidad podemos eliminar una de ellas para el posterior tratamiento.

El análisis se puede afinar si consideramos r, coeficiente de correlación muestral, como un estimador del coeficiente de correlación poblacional, ρ. A partir del estadístico

$$Z=\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right),$$

de distribución aproximadamente normal, podemos contrastar

H_0: $\rho=0$ (incorrelación) contra H_1: $\neq 0$ (correlación), que nos llevará a aceptar o rechazar la hipótesis de incorrelación de acuerdo con el valor muestral del estadístico r.

El **análisis de componentes principales** es una de las principales técnicas multivariantes que tiene como finalidad reducir la dimensionalidad de un conjunto de datos. Si partimos de k variables, $X_1,\ldots,X_k$, referidas a n individuos, buscaremos k nuevas variables, $Y_1,\ldots,Y_k$, combinación lineal de las anteriores, cumpliendo las siguientes condiciones:

- La primera variable, Y_1, explicará la mayor varianza posible.
- La segunda, Y_2, explicará la mayor varianza posible no explicada por Y_1 y, además, estará incorrelacionada con ella (no tendrán información común).
- La tercera, Y_3, explicará la mayor varianza posible no explicada por Y_1 e Y_2 y, además, estará incorrelacionada con ellas.
- Así, sucesivamente.

De esta forma, las variables $Y_1,\ldots,Y_k$, componentes principales, explicarán toda la varianza del conjunto de datos originales. Ahora bien, la forma en que han sido construidas permite que, tomando un número reducido de ellas, por ejemplo s, $Y_1,\ldots,Y_s$, con s<k, éstas expliquen un porcentaje aceptable de la varianza. De esta forma, reducimos la dimensionalidad de los datos a cambio de una pequeña pérdida de información. Normalmente, las primeras componentes explicarán un alto porcentaje de la variabilidad total (entre el 80% y el 90%), despreciando el resto de componentes que no aportarán una cantidad significativa de información.

Otra técnica multivariante que utilizaremos en nuestra aplicación para evitar la información redundante es el **análisis factorial**. Su objetivo es explicar, según un modelo lineal, un conjunto grande de variables observables mediante un número más reducido de variables no observables. La idea que está detrás de este planteamiento es que existen una serie de factores subyacentes a los fenómenos en estudio, factores de los cuales no podemos obtener observación directa; simplemente observaremos variables que son síntomas de estos factores subyacentes.

Básicamente, se trata de explicar k variables observables, $X_1,\ldots,X_k$, por m+k (m<k) variables incorrelacionadas no observables, $Y_1,\ldots,Y_m,E_1,\ldots,E_k$, variables que llamaremos factores, de acuerdo con el siguiente modelo lineal:

$$X_1 = a_{11}\,Y_1 + \ldots + a_{1m}\,Y_m + b_1\,E_1$$
$$\ldots\ldots\ldots\ldots\ldots\ldots$$
$$X_k = a_{k1}\,Y_1 + \ldots + a_{km}\,Y_m + b_k\,E_k$$

o bien, en forma matricial,

$$X = AY+BE.$$

Las variables $Y_1,\ldots,Y_m$ son los factores comunes, pues

influyen, en mayor o menor medida, en todas las variables. Las variables $E_1,\ldots,E_k$ son los factores específicos pues cada uno influye únicamente en una variable. Los factores comunes son los que deben explicar la mayor parte de la variabilidad, despreciando la influencia de los factores específicos.

Calculada la matriz factorial, A, se interpretarán los factores de acuerdo con los pesos, a_{ij}, de las variables en los mismos. El proceso siempre complicado de interpretación de los factores se puede facilitar rotando la matriz factorial de cara a conseguir una nueva matriz cuyas componentes sean próximas a +1, -1 ó 0. No comentaremos los distintos métodos de rotación, pues no va a ser nuestro interés interpretar los factores, sino, sencillamente, reducir la dimensionalidad de los datos trabajando con m factores, en vez de con las k variables.

3.- PLANTEAMIENTO GENERAL DEL ESTUDIO SOBRE DISPARIDADES SOCIO-ECONOMICAS PROVINCIALES

El objetivo de este trabajo es clasificar las provincias españolas en grupos más o menos homogéneos, de acuerdo con su nivel de bienestar, utilizando para ello la técnica multivariante de análisis *cluster*. Obsérvese que, a diferencia de los trabajos que le han servido de referencia (Zarzosa (1992) e INE (1991)), no tratamos de ordenar las provincias de acuerdo a un indicador sintético de bienestar, sino clasificarlas en grupos homogéneos, creando un "mapa" de zonas de bienestar. Ahora bien, la observación de estos grupos, así como la comparación con los resultados de los citados trabajos, permitirá ordenar los grupos obtenidos de acuerdo con su nivel de vida.

Las pretensiones de este trabajo no han llevado a plantearse una previa selección de indicadores. Hemos obviado el problema utilizando básicamente los indicadores propuestos en Zarzosa (1992) actualizados, en la medida de lo posible, con la base de datos de INE (1991). La mayor parte de las variables se refieren al año 1986, salvo algunas que corresponden a otros periodos temporales anteriores por falta de información para 1986 (2).

Las variables utilizadas fueron, en concreto, las siguientes:

Componente "Empleo"

X1 Tasa de ocupación: porcentaje de ocupados sobre población de 16 a 64 años. 1986.
X2 Tasa de paro (por 100 activos) ambos sexos. 1986.
X3 Tasa de paro (por 100 activos) varones. 1986.
X4 Tasa de paro (por 100 activos) mujeres. 1986.
X5 Tasa de paro (por 100 activos) menores de 25 años. 1986.
X6 Tasa de paro (por 100 activos) de 25 o más años. 1986.

Componente "Salud"

X7 Tasa de mortalidad infantil por 1.000 nacidos vivos. 1985.
X8 Enfermos dados de alta según provincia de residencia por 100.000 habitantes. 1986/87.
X9 Accidentes de trabajo por 1.000 ocupados. 1986.

Componente "Educación"

X10 Tasa bruta de escolaridad por 100 habitantes en preescolar. 1986.
X11 Tasa bruta de escolaridad por 100 habitantes en EGB. 1986.
X12 Tasa bruta de escolaridad por 100 habitantes en BUP y COU. 1986.
X13 Tasa bruta de escolaridad por 100 habitantes en FP. 1986.
X14 Tasa de escolaridad de la población de 16 a 35 años (por 1.000 hab.). 1986.
X15 Porcentaje de población de 16 y más años de edad con estudios primarios. 1986.
X16 Porcentaje de población de 16 y más años de edad con estudios medios. 1986.
X17 Porcentaje de población de 16 y más años de edad con estudios de nivel anterior al superior. 1986.
X18 Porcentaje de población de 16 y más años de edad con estudios superiores. 1986.
X19 Porcentaje de población de 16 y más años de edad sin estudios. 1986.
X20 Porcentaje de población de 16 y más años de edad analfabetos. 1986.

Componente "Ingreso y Consumo"

X21 Ingreso medio anual de los hogares por persona (en miles de pesetas). 1986.
X22 Gasto anual medio de los hogares por persona (en miles de pesetas). 1986.
X23 Porcentaje de gastos en alimentación respecto del consumo total de los hogares. 1986.

Componente "Vivienda"

X24 Superficie útil de la vivienda por miembro del hogar. 1981.
X25 Número de miembros del hogar por habitación. 1981.
X26 Porcentaje de hogares sin agua corriente en la vivienda principal. 1981.
X27 Porcentaje de hogares con agua fría y caliente general en la vivienda principal. 1981.
X28 Porcentaje de hogares sin servicios de higiene en la vivienda principal. 1981.
X29 Porcentaje de hogares con al menos un cuarto de baño en la vivienda principal. 1981.
X30 Porcentaje de hogares sin calefacción en la vivienda principal. 1981.
X31 Porcentaje de hogares con calefacción central en la vivienda principal. 1981.

X32 Porcentaje de hogares que tienen teléfono. 1981.
X33 Porcentaje de hogares que tienen automóvil. 1981.
X34 Porcentaje de hogares que tienen lavadora automática. 1981.
X35 Porcentaje de hogares que tienen cámara fotográfica. 1981.
X36 Porcentaje de hogares que tienen magnetófono o radiocassette o tocadiscos. 1981.
X37 Consumo de energía eléctrica para usos domésticos (kwh por hab.). 1986.

No se nos escapan las limitaciones que presenta esta relación de indicadores. Por un lado, se echa en falta indicadores de percepción; por otro, tampoco se cuenta con indicadores que recojan ciertas componentes sumamente importantes del bienestar: ocio, cultura, delincuencia, condiciones medioambientales, etc. Además, no se puede ignorar la ambigüedad manifiesta de algunos indicadores de la relación anterior; así, por ejemplo, el consumo de energía eléctrica para usos domésticos, X37, está, por una parte, muy influenciado por la climatología y, por otra, valores altos pueden interpretarse como un despilfarro y una mala utilización de las siempre sucias fuentes de energía.

Con esta base de datos nos planteamos, en primer lugar, eliminar las variables que presentaban fuertes correlaciones. Se siguió el criterio de considerar aquellas cuyo coeficiente de correlación fuese mayor que 0,8 o menor que -0,8; en ese caso, se eliminaba la variable que presentase menos variabilidad, es decir, la que tuviese el coeficiente de variación más bajo. Obsérvese que el criterio es muy poco restrictivo pues, a un nivel de significación de 0,05, sólo se admite la incorrelación, $\rho=0$, de las variables cuyo coeficiente de correlación, r, esté comprendido entre -0,286 y 0,286. De aplicar este criterio, estimamos que se hubiera perdido mucha información lo que no se compensaría con las ventajas de trabajar con variables incorrelacionadas.

De esta forma se eliminaron las variables X3, X5, X6, X22, X27, X28, X29, X32, X33, X34 y X36; asimismo, se prescindió de X18 por formar combinación lineal con las variables X15, X16, X17, X19 y X20.

A.- Primera fase del análisis

Con las 25 variables restantes se ejecutó el procedimiento de análisis *cluster* de STATGRAPHICS (3). Dicho procedimiento tiene como opción seis métodos a la hora de construir conglomerados: cinco jerárquicos (distancia media, mínima, máxima, mediana y distancia entre centroides) y uno no jerárquico (*seeded*).

Dado que los métodos jerárquicos daban lugar a conglomerados muy descompensados y dificilmente interpretables, se utilizó el método *seeded*, no jerárquico.

Este método exige determinar previamente el número de conglomerados, p, en que se desea clasificar a los n individuos. Determinado este número, se asigna un individuo a cada uno de los *clusters* (estos individuos son los *seed points*, puntos preseleccionados o puntos "semilla"). El resto de los individuos se va clasificando en el *cluster* donde está el *seed point* más cercano. De esta forma sólo se necesita calcular la matriz de distancias entre las n observaciones originales (de hecho no se necesita toda la información de esta matriz, sino únicamente las distancias entre los n-p puntos no preseleccionados y los p puntos preseleccionados).

Evidentemente, se presentan dos problemas: la elección del número de conglomerados y la asignación a cada uno de ellos de los *seed points*. En cuanto al número de conglomerados, se tomó la decisión de utilizar cinco. Por un lado, es un valor intermedio que discrimina más que valores más pequeños (2 ó 3 grupos) y va a ser más fácil de interpretar que un número grande (10 o más niveles de bienestar). Por otro lado, nos va a permitir interpretar los resultados en cinco niveles de bienestar, alto, medio-alto, medio, medio-bajo y bajo, algo común a muchas mediciones (escala de Likert), con lo que se facilita la posterior interpretación.

En cuanto a la elección de los *seed points*, es decir, la elección de cinco provincias que representasen cada uno de los niveles de bienestar antes citados, se utilizó la información proporcionada por la matriz de distancias así como los resultados obtenidos en los trabajos de Zarzosa (1992) y del INE (1991), resultados similares a pesar de que parten de fechas y metodologías distintas. Se tomaron cinco provincias, una situada en la parte alta de la clasificación, otra en la media-alta, etc. (de hecho se probó con distintas posibilidades de elección, escogiendo la más estable y razonable).

Las provincias elegidas fueron:
- Navarra: bienestar alto.
- Valencia: bienestar medio-alto.
- León: bienestar medio.
- Cuenca: bienestar medio-bajo.
- Badajoz: bienestar bajo.

Preseleccionados los cinco puntos de arranque, se ejecutó el procedimiento *Cluster Analysis* de STATGRAPHICS (Método: *Seeded*, número de clusters: 5, variables tipificadas, distancia: euclídea) obteniéndose los siguientes resultados:

Provincias con bienestar alto: Zaragoza, Madrid, Navarra, Alava, Guipúzcoa y Vizcaya.

Provincias con bienestar medio-alto: Huesca, Baleares, Valladolid, Guadalajara, Barcelona, Gerona, Lérida, Tarragona, Alicante, Castellón, Valencia y Murcia.

Provincias con bienestar medio: Teruel, Asturias, Cantabria, Burgos, León, Palencia, Salamanca, Segovia, Soria, Zamora, La Coruña, Pontevedra y La Rioja.

Provincias con bienestar medio-bajo: Avila, Albacete,

Ciudad Real, Cuenca, Toledo, Cáceres, Lugo y Orense.

Provincias con bienestar bajo: Almería, Cádiz, Córdoba, Granada, Huelva, Jaén, Málaga, Sevilla, Las Palmas, Sta. Cruz de Tenerife y Badajoz.

B.- Segunda fase del análisis

Posteriormente, se trató de eliminar la información redundante mediante el método de componentes principales aplicado a las 25 variables, es decir, la base de datos de trabajo, excluídas las variables fuertemente correlacionadas. Con estas 25 variables se ejecutó el procedimiento *Principal Components Analysis* de STATGRAPHICS. Los resultados obtenidos fueron:

Principal Components Analysis

Component Number	Percent of Variance	Cumulative Percentage
1	36.31750	36.31750
2	15.72874	52.04624
3	9.92606	61.97230
4	6.92781	68.90011
5	5.16336	74.06347
6	4.30687	78.37034
7	3.59433	81.96468
8	2.92512	84.88980
9	2.76044	87.65024
10	1.91273	89.56298
11	1.66250	91.22547
12	1.44444	92.66992
13	1.27800	93.94792
14	1.12568	95.07360
15	1.06826	96.14187
16	.84093	96.98280
17	.78116	97.76396
18	.61192	98.37588
19	.42931	98.80518
20	.35372	99.15890
21	.28888	99.44778
22	.25183	99.69961
23	.16795	99.86756
24	.13045	99.99801
25	.00199	100.00000

Dado que las seis primeras componentes explican un porcentaje aceptable de la varianza total (78,37%), se optó por hacer análisis *cluster* con estas componentes. La utilización de métodos jerárquicos condujo, de nuevo, a agrupaciones descompensadas, por lo cual se decidió utilizar también el método *seeded*. Lo razonable hubiera sido elegir unos nuevos *seed points* teniendo en cuenta los resultados de los trabajos de referencia y la nueva matriz de distancias. Aun así, para poder hacer la comparación con los resultados obtenidos en la primera fase, se mantuvieron los mismos *seed points*, es decir, Navarra, Valencia, León, Cuenca y Badajoz, representando los cinco niveles decrecientes de bienestar.

Ejecutando bajo estas condiciones el procedimiento *Cluster Analysis* se obtuvieron los siguientes resultados:

Provincias con bienestar alto: Huesca, Zaragoza, Madrid, Navarra, Alava, Guipúzcoa y Vizcaya.

Provincias con bienestar medio-alto: Teruel, Baleares, Palencia, Valladolid, Guadalajara, Toledo, Barcelona, Gerona, Lérida, Tarragona, Alicante, Castellón, Valencia y Murcia.

Provincias con bienestar medio: Asturias, Las Palmas, Sta. Cruz de Tenerife, Cantabria, Avila, Burgos, León, Salamanca, Segovia, Soria, Zamora, La Coruña, Lugo, Orense, Pontevedra y La Rioja.

Provincias con bienestar medio-bajo: Albacete y Cuenca.

Provincias con bienestar bajo: Almería, Cádiz, Córdoba, Granada, Huelva, Jaén, Málaga, Sevilla, Ciudad Real, Badajoz y Cáceres.

C.- Tercera fase del análisis

Por último, se trató de evitar la información redundante y reducir la dimensionalidad utilizando el procedimiento de análisis factorial aplicado a las 25 variables. En concreto, se ejecutó este procedimiento de STATGRAPHICS con las variables tipificadas, resultando:

Variable	Communality	Factor	Eigenvalue	Percent Var	Cum Percent
X2	0.93330	1	8.94579	42.3	42.3
X4	0.87755	2	3.79006	17.9	60.3
X7	0.50593	3	2.33772	11.1	71.4
X8	0.73105	4	1.54016	7.3	78.6
X9	0.85067	5	1.11311	5.3	83.9
X10	0.61333	6	.79950	3.8	87.7
X11	0.51339	7	.55170	2.6	90.3
X12	0.86497	8	.46697	2.2	92.5
X13	0.79557	9	.44494	2.1	94.6
X14	0.75169	10	.27723	1.3	95.9
X15	0.99894	11	.24775	1.2	97.1
X16	0.99614	12	.22860	1.1	98.2
X17	0.93131	13	.12906	.6	98.8
X19	0.99847	14	.10683	.5	99.3
X20	0.99554	15	.08942	.4	99.7
X21	0.95997	16	.05207	.2	100.0
X23	0.82328	17	.00442	.0	100.0
X24	0.75159	18	-.00066	.0	100.0
X25	0.81974	19	-.01130	.0	100.0
X26	0.78981	20	-.03437	.0	100.0
X30	0.70198	21	-.05818	.0	100.0
X31	0.77921	22	-.06581	.0	100.0
X35	0.89221	23	-.09548	.0	100.0
X37	0.87491	24	-.10998	.0	100.0
X1	0.87886	25	-.12013	.0	100.0

Los cinco primeros factores explican el 83,9% de la varianza total (no entraremos en la interpretación de cada uno de estos factores pues no es el propósito de este trabajo). Así, caracterizadas las 50 provincias por las puntuaciones de esos cinco factores, se ejecutó el procedimiento *Cluster Analysis* manteniendo las mismas especificaciones que en las dos fases anteriores (5 *seed points*: Navarra, Valencia, León, Cuenca y Badajoz). Los resultados obtenidos fueron los siguientes:

Provincias con bienestar alto: Huesca, Zaragoza, Madrid, Navarra y Vizcaya.

Provincias con bienestar medio-alto: Teruel, Baleares, Palencia, Valladolid, Guadalajara, Toledo, Barcelona, Gerona, Lérida, Tarragona, Alicante Castellón, Valencia, Murcia, Alava y Guipúzcoa.

Provincias con bienestar medio: Asturias, Las Palmas, Sta. Cruz de Tenerife, Cantabria, Avila, Burgos, León, Salamanca, Segovia, Soria, Zamora, La Coruña, Lugo, Orense, Pontevedra y La Rioja.

Provincias con bienestar medio-bajo: Albacete y Cuenca.

Provincias con bienestar bajo: Almería, Cádiz, Córdoba, Granada, Huelva, Jaén, Málaga, Sevilla, Ciudad Real, Badajoz y Cáceres.

IV.- RESULTADOS Y CONCLUSIONES

A pesar de que el análisis *cluster* tiene como única finalidad clasificar individuos en grupos homogéneos, el método utilizado (la elección, más o menos arbitraria, pero intencionada, de los *seed points* y la comparación con otros trabajos similares) permite una ordenación de los conglomerados, y la consiguiente interpretación.

En rasgos generales, las tres fases del análisis proporcionan resultados similares. Así, el grupo de provincias de alto bienestar está básicamente formado por Zaragoza, Madrid, Navarra y el País Vasco (en la tercera fase no se incluyen Alava y Guipúzcoa, que caen al grupo siguiente). Puede ser más sorprendente que Huesca aparezca en dos de los análisis en este grupo elegido de provincias de alto bienestar.

En el grupo de provincias de bienestar medio-alto, se encuentra la franja mediterránea (Cataluña, País Valenciano y Murcia), incluidas las Islas Baleares y alguna provincia del interior (Valladolid y Palencia, Guadalajara y Toledo, y Teruel). Puede ser rara la presencia de Barcelona, que en otros estudios ocupa posiciones superiores, y de las provincias mesetarias de Palencia, Guadalajara y Toledo.

El grupo intermedio está compuesto por la franja cantábrica (Galicia, Asturias y Cantabria), Castilla y León (salvo Palencia y Valladolid), La Rioja y, en dos de los análisis, las provincias canarias. Hay que hacer notar que estas dos provincias aparecen en el análisis de la primera

fase (aplicación directa del análisis *cluster*) en el último grupo; ello puede ser debido a que en esa fase estén muy sobrevaloradas ciertas variables (en concreto, las relativas a hogares con calefacción, X125 y X126) que, por motivos climáticos, toman valores que se pueden interpretar falsamente como síntomas de malestar.

El cuarto grupo, provincias de bienestar medio-bajo, no es susceptible de una clara interpretación dado que, en dos de las fases, sólo está integrado por Albacete y Cuenca; aún así, podemos situar, más o menos, a este grupo en Castilla-La Mancha.

Por último, el grupo de cola, provincias de bajo bienestar, está formado por las provincias del sur de la península (Andalucía, Extremadura y Ciudad Real).

Aunque los trabajos de Zarzosa (1992) e INE (1991) ordenan las provincias según su nivel de bienestar utilizando metodologías diferentes y bases de datos distintas en tamaño y fecha, los resultados obtenidos son coincidentes en lo fundamental.

No interpretaremos aquí los resultados desde el punto de vista de la geografía económica. Solamente recogeremos ciertas conclusiones generales que, a pesar de la antigüedad de los datos, se desprenden de los resultados.

Por un lado, el bienestar sigue concentrándose en los focos industriales históricos (Madrid y País Vasco, a pesar de la recesión industrial sufrida por la región norteña en los últimos años) y en los más recientes focos interiores (Zaragoza y Valladolid). Por otro lado, se observa el creciente empuje del eje mediterráneo; el manejo de datos posteriores pudiera confirmar la colocación de algunas de las provincias ribereñas en la cabeza del bienestar.

En niveles medios de bienestar se encuentra, en descenso, parte de la cornisa cantábrica y, en ascenso, las provincias castellanas. Por último, en las situaciones de déficit de bienestar están las provincias que representan el histórico retraso del sur español, Andalucía y Extremadura, si bien hay que señalar que la fecha de los datos, 1981-1986, impide que se recoja la hipotética mejoría en el bienestar provocada por las grandes inversiones públicas de los últimos años.

NOTAS

(1) INE: *Indicadores Sociales*, INE, Madrid, 1991; pág. 13.

(2) Obviamente, se podían haber actualizado algunas de las variables. Ahora bien, la no disponibilidad de los datos de la última Encuesta de Presupuestos Familiares (1991) desaconsejó esta actualización pues la fuerte disparidad temporal hubiera provocado desajustes en el análisis.

(3) El programa utilizado en el tratamiento de los datos ha sido el STATGRAPHICS, versión 6.0. Entre otras limitaciones

hay que citar que el procedimiento *Cluster Analysis* no incorpora la distancia de Mahalanobis, pudiéndose utilizar únicamente la distancia euclídea. En la primera fase del análisis esto puede ocasionar un problema dado que algunas de las 25 variables presentan correlaciones, si bien débiles; en la segunda y tercera fase, se soluciona este problema pues trabajaremos con variables incorrelacionadas por definición.

BIBLIOGRAFIA

ABASCAL, E. y GRANDE, I.: *Métodos multivariantes para la investigación comercial. Teoría, aplicaciones y programación BASIC*, Ariel, Barcelona, 1989.

CUADRAS, C.M.: *Métodos de análisis multivariante*, Eunibar, Barcelona, 1981.

GARCIA-DURAN, J. y PUIG, P.: *La calidad de la vida en España. Hacia un estudio de indicadores sociales*, Moneda y Crédito, Madrid, 1980.

G. BEZARES, F.: *Cómo utilizar e interpretar la Estadística (Con aplicaciones a la gestión de la empresa y a la investigación social)*, Ibérico Europea de Ediciones, Madrid, 1983.

INE: *Indicadores Sociales*, INE, Madrid, 1991.

JOHNSON, R.A. y WICHERN, D.W.: *Applied Multivariate Statistical Analysis*, Prentice-Hall, Englewood Cliffs, 1982.

MARTIN-GUZMAN, M.P.: *"Métodos estadísticos en el análisis regional"*, Estudios Regionales nº 22, 1988, págs. 149-170.

MARTIN-GUZMAN, M.P. y MARTIN, J.: *Curso básico de Estadística Económica*, AC, Madrid, 1989.

MILLIGAN, G.W.: *"An examination of the effect of six types of error perturbation of fifteen clustering algorithms"*, Psychometrika, nº 45, 1980, págs. 325-342.

PENA, B.: *Problema de la medición del bienestar y conceptos afines. (Una aplicación al caso español)*, INE, Madrid, 1977.

STATGRAPHICS: *Quickstart Guide, User Manual, Reference Manual, Examples Manual*, Manugistics Inc., Cambridge, 1992.

ZARZOSA, P.: *Aproximación a la medición del bienestar social: Estudio de la idoneidad del Indicador Sintético "Distancia-P_2"*, tesis doctoral (inédita), Valladolid, 1992.