Facultad
de
Ciencias

Métodos cuantitativos y Geografía: Aplicación del Análisis Multivariante al estudio de los desequilibrios en materia de Bienestar en España

Trabajo de Fin de Grado
para acceder al

GRADO EN MATEMÁTICAS

Autor: Juan Antonio Romillo Barquín

Director: Juan Antonio Cuesta Albertos y Pedro Reques Velasco

Octubre-2013

Índice.

Resumen. 3
1. El bienestar social. Una aproximación teórica. 4
2. Metodología. 8
2.1. Análisis de componentes principales. 8
2.1.1. Cambios de base. Rotación Varimax……..……..11
2.2. Análisis Cluster. 11
2.3. Correlación. 13
3. Resultados. 15
3.1. Preliminares. Selección de variables. 15
3.2. Análisis de componentes principales. 18
3.2.1. Interpretación de los factores. …….……..23
3.3. Análisis Cluster. 27
4. Conclusiones. 32
5. Agradecimientos. 333
6. Fuentes y bibliografía. 334
ANEXO I. Figuras y tablas. 36
ANEXO II. Programa. 37

Resumen.

El trabajo se plantea como objetivo fundamental determinar las estructuras territoriales del bienestar social en España, haciendo uso de técnicas del Análisis Multivariante. Estas técnicas, aunque no han sido utilizadas en la Geografía clásica, han servido ampliamente en las últimas décadas a la Geografía cuantitativa. El trabajo que se presenta, pues, se sitúa entre la Geografía Humana y la Estadística y pretende ser un puente entre ambas disciplinas.

Con el fin de alcanzar el objetivo central del trabajo se seleccionaron variables inequívocamente ligadas al bienestar social, relacionadas con la actividad económica, el desarrollo cultural, la modernización demográfica, los movimientos migratorios y el desarrollo turístico. Estas variables, una vez reducidas o transformadas en indicadores, se agruparon posteriormente en unos pocos factores que las resumen y explican. Finalmente, y tomando las provincias como unidad de análisis, se llevaron a cabo agrupaciones espaciales atendiendo a su similitud interna, con el fin de definir y cartografiar las estructuras territoriales del bienestar en España.

Summary.

Te goal of this research is to determine the territorial structures of social welfare in Spain, using Multivariate Analysis techniques. These techniques, which had not been used before in classical Geography, have been extensively employed over the past decades in quantitative Geography. Therefore the research presented can be considered as a bridge between the two disciplines, Human Geography and Statistics.

In order to achieve the main objective of the research, we selected several variables, directly related to welfare, which measured aspects of the economic activity, the socio cultural development, the demographic modernization, the population displacements and the tourism development. These variables were reduced or transformed into indicators, and then, they were grouped in a few factors that summarize and explain them. Taking the provinces as the unit of analysis, we conducted a spatial clustering process based on the internal similarity of each group of provinces, in order to define and map the territorial structures of welfare in Spain.

Capítulo 1. El bienestar social. Una aproximación teórica.

Por bienestar cabe entenderse "*la situación en la que se encuentra una persona cuando se satisfacen todas sus necesidades, y se prevé que se seguirán satisfaciendo*" (Morales, 1994). Esta definición permite apreciar que el bienestar no sólo tiene relación con la situación en la que una persona se encuentra en el presente sino que a su vez se ha de tener en cuenta en qué situación se espera encontrar en el futuro. Esto puede parecer no ser importante, pero es algo que nos diferencia de otros seres vivos que tan sólo aspiran a un bienestar momentáneo, a satisfacer sus necesidades al instante (por instinto).

De otra parte, el bienestar es un concepto que para cada persona puede ser entendido de un modo distinto, bien porque se tengan distintas necesidades, bien porque cada persona se sienta realizada de distintas formas. A pesar de las grandes diferencias entre los individuos, todos ellos tienen una base común: la necesidad de encontrarse en una situación en la que las necesidades básicas estén satisfechas, pues no hay lugar para el bienestar y la felicidad cuando faltan los elementos para alcanzarlo.

El bienestar es un concepto complejo ligado a las necesidades, percepciones y valores de las personas, pero lo que hace incluso más complicado a este término es el hecho de que éstas no son conscientes muchas veces del porqué, del motivo que les mueve. Una persona sana mentalmente debería preguntarse las razones que lo impulsan a necesitar o querer una cosa y no dejarse llevar por los motivos de los que le rodean.

La tendencia que tiene el ser humano en su consecución del Bienestar no es algo que surja en sociedad, no es algo impuesto por la sociedad sino que es el propio ser humano el que ve la sociedad como el medio para alcanzar el bienestar. Otra cosa son las necesidades que puedan surgir debidas al hecho de que el ser humano conviva en sociedad. De hecho, las necesidades del ser humano no son fijas sino que varían a lo largo de las distintas épocas a medida que la sociedad evoluciona. Debido a la invención surgen nuevas necesidades que satisfacer y gracias a la técnica es capaz de satisfacer necesidades que antes no podía.

Surge entonces un nuevo término: el ***bienestar social***, que lo entenderemos como todas aquellas circunstancias o bienes que inciden, de un modo positivo en que los individuos de una sociedad puedan disfrutar de una buena calidad de vida. Por lo tanto, un individuo en sociedad precisa de un empleo, de manera que pueda disponer de los recursos económicos suficientes para satisfacer sus necesidades, necesita también educación, una vivienda, buena salud, tiempo para disfrutar...

Cómo se siente una persona en el seno de la sociedad, lo cual fue expuesto por Keyes (1998), es algo que también se ha de tener muy en cuenta: si una persona se siente feliz; si está satisfecha con su vida; si siente que forma parte de la sociedad; si la aceptan y confían en ella; si se siente útil para la sociedad; si confía en que la sociedad

se encamina hacia el bienestar; si entiende cómo funciona y por qué motivos la sociedad se desenvuelve de un modo u otro...

Y, por último, como síntesis de lo que se acaba de exponer cabe decir que el bienestar no es un término simple, no es algo que pueda ser explicado con facilidad, que cada persona tiene un concepto distinto del término, que hay muchos tipos de enfoques distintos (económico, social, geográfico, antropológico, psicológico ...). En definitiva, el bienestar es un concepto poliédrico que requiere de muchos matices para lograr explicarlo correctamente, pero lo esencial es la base sobre la que uno puede construir.

El bienestar, pues, tiene muchas dimensiones: la psicológica y social, a la que hemos aludido, la económica y, la que aquí más nos interesa: la territorial.

David M. Smith (1997), padre de la Geografía del Bienestar, en sugerente definición considera a esta corriente geográfica "*como el estudio de quién consigue qué, dónde y cómo",* dado que proporciona un enfoque distinto que permite facilitar el análisis de todas las variables y las estructuras de la vida humana con relación al bienestar. Cada una de las palabras quién, qué, dónde y cómo plantea su propio conjunto de problemas.

Cuando preguntamos '*¿quién*?' estudiamos la manera adecuada de elegir a la población que estamos investigando, como basándonos en características comunes como la clase social, la posición económica, etc. Esto nos preocupa dado que en numerosas ocasiones es impracticable estudiar a cada individuo por separado.

Cuando preguntamos '*¿qué*?' planteamos la cuestión de saber qué es lo que da satisfacción, felicidad o una mayor, o menor, calidad de vida a los seres humanos.

Cuando se pregunta '*¿dónde*?' se plantea la cuestión geográfica básica de cómo subdividir el territorio para poder investigarlo: se trata de un problema de una importancia especial en la investigación sobre el bienestar social.

Cuando se pregunta '*¿cómo*?' se ha de identificar y comprender la estructura, el proceso y la causa por la que actúa una sociedad y que desemboca en una estructura concreta de quién consigue qué y dónde.

Hasta mediados de los años 50, la Geografía Humana era más un arte que una ciencia. A menudo, los hechos se descubrían por observaciones casuales y no por mediciones cuidadosas; las relaciones causales que emitían los geógrafos estaban basadas en su intuición; y la labor del geógrafo era más bien la de escribir una obra sobre un territorio concreto que una obra a partir de la cual se pudiese generalizar. Aunque, también es cierto que los geógrafos de la primera mitad del siglo XX tenían una perspectiva más cercana a la realidad de los habitantes que los actuales debido a que estos se sirven de métodos más convencionales haciendo uso de la estadística. Y para que la Geografía Humana responda a las necesidades de una sociedad ha de concentrarse directamente en los mismos problemas con que se enfrenta la gente cada día.

Según la definición de Mishan (1964), otro de teóricos de la geografía del bienestar, «*La geografía del bienestar social es la disciplina que trata de formular*

proposiciones que nos permitan ordenar en una escala de mejor a peor las situaciones geográficas alternativas que se le plantean a la sociedad.» Una situación geográfica designa cualquier aspecto de la disposición espacial de la existencia humana. Puede referirse a la asignación de recursos, a la renta o a cualquier otro factor que influya en el bienestar humano. Puede referirse a la pobreza o a cualquier problema social. Puede utilizarse para describir la localización industrial, la distribución de la población, la localización de los servicios sociales, las redes de transporte, los movimientos migratorios o de bienes, y en general cualquier disposición espacial que repercuta sobre la calidad de la vida de forma variable con la disposición espacial. Y debajo de todo hallamos las estructuras económicas, sociales y políticas que generan todas las variables anteriores.

Pero los geógrafos del bienestar van más allá y plantean que en el tema del bienestar social han de tratarse cinco tareas básicas que han definido durante algún tiempo el ámbito de la geografía humana. Se trata de la descripción, la explicación, la valoración, la prescripción y la realización.

1) La *descripción* trata de modo previo y primero la identificación empírica de los niveles territoriales de la prosperidad social o de cualquier término que deseemos considerar en la condición humana.

2) La *explicación* cubre el cómo de nuestra anterior definición. Exige la identificación de las relaciones de causa a efecto entre las diversas actividades que se llevan a cabo dentro de la sociedad y que contribuyen a determinar quién consigue qué y dónde.

3) La *valoración* se ocupa de la formulación de juicios sobre el carácter deseable de estados geográficos alternativos, y las estructuras de la sociedad que los originan. Decir que una estructura espacial del bienestar humano es preferible a otra equivale a decir que le corresponde un nivel más elevado de bienestar social.

4) La *prescripción* exige concretar los estados geográficos alternativos y las estructuras alternativas de la sociedad destinadas a producirlos. Este proceso consiste en planificar la organización espacial de la actividad humana, es decir, la reorganización espacial. La prescripción implica una respuesta a la cuestión ética de quién debería conseguir qué y dónde.

5) La *actuación* constituye el proceso final de sustituir un estado que se considera indeseable por algo superior. Se refiere a la cuestión del cómo, una vez decidido quién debería conseguir qué y dónde.

Los geógrafos dan una importancia transcendental al territorio, de hecho, son los únicos estudiosos que dan una preferencia al espacio y la localización. Son los geógrafos los que poseen la capacidad para explicar los asuntos humanos tanto políticos, como económicos, como sociales... partiendo de las relaciones espaciales.

Para resaltar el valor de este trabajo cabe mencionar que los geógrafos parten del territorio para lograr explicar las variables y las estructuras políticas, sociales, económicas, etc., es decir, la distribución espacial es en sí misma una variable a tener en cuenta. Sin embargo, lo ideal no es esto sino que partiendo de ciertas variables (como la renta *per cápita*, indicadores de pobreza, indicadores sociales...) que tienen una relación, positiva o negativa, con el bienestar se llegue al territorio, es decir, se llegue a apreciar la existencia de una distribución territorial en función del bienestar.

Capítulo 2. Metodología.

En este capítulo se exponen las técnicas y métodos que hemos seleccionado para la resolución del problema que nos ocupa. Estas técnicas provienen del campo de la Estadística Multivariante.

La Estadística Multivariante se ocupa de analizar simultáneamente conjuntos de datos multivariantes, en el sentido de que hay varias variables medidas para cada individuo u objeto estudiado, con el propósito de lograr un mejor entendimiento del problema bajo consideración, obteniendo información que los métodos estadísticos univariantes son incapaces de conseguir de modo que facilite la labor a los expertos en la materia a la hora de tomar decisiones.

El contenido del capítulo es el siguiente: En la Sección 2.1 presentamos el Análisis de Componentes Principales que es una técnica pensada para reducir el número de variables disponibles con la menor pérdida posible de información. La Sección 2.2 presenta el Análisis Cluster que permite realizar agrupaciones de individuos. Y por último, en la Sección 2.3 nos ocuparemos de la correlación que permite medir cómo de fuerte es la relación lineal entre variables.

Las ideas básicas aquí expuestas han sido tomadas de Cuesta Albertos (2012).

2.1. Análisis de componentes principales.

Supongamos que se dispone de n datos medidos sobre p variables. Como hemos señalado, el Análisis de Componentes Principales es una técnica para reducir el número de variables perdiendo la menor cantidad posible de información. Sobre todo, en los casos en los que p es grande, existen fuertes relaciones entre las variables, lo que hace que contengan información redundante y, en consecuencia, que sea posible la eliminación de algunas de ellas sin gran pérdida de información.

No es éste el lugar para discutir lo que se entiende por información. En su lugar, incluimos algunos ejemplos que permiten aclarar el objetivo que persigue esta técnica:

Por ejemplo, si una variable tiene los mismos valores que una segunda multiplicada por dos, entonces parecería razonable el eliminar cualquiera de las dos y quedarnos con la otra, ya que, en este caso, no perderíamos información porque podríamos recuperar la otra variable íntegramente conociendo de qué forma una explica a la otra. Otro ejemplo parecido sería que dispusiésemos de tres variables y una de ellas

fuese combinación lineal de las otras dos, en este caso, parece razonable el quitar una de ellas y quedarnos con las otras dos.

Un ejemplo algo más complicado es el de los datos reflejados en la siguiente tabla:

Nº del dato	Variable 1	Variable 2	Variable 3	Variable 4
1	100	200	300	400
2	100	201	302	405
3	98	205	295	387

Tabla 2.1. *Ejemplo simulado de variables fuertemente relacionadas.*

Como se puede apreciar en la Tabla 2.1, la primera variable explica razonablemente bien las otras tres, aunque hay algunas diferencias que pueden considerarse pequeñas frente al valor de las variables. En este caso, podríamos quedarnos con todas las variables sin perder información o podríamos despreciar las tres últimas y quedarnos tan sólo con la primera, ya que perdiendo tan sólo un poco de información se reduciría en gran medida el número de variables utilizadas para explicar los datos.

Lo que podemos sacar en claro a partir de estos ejemplos es que, si llamamos p a la dimensión inicial de los datos, en todos ellos, se puede seleccionar una variedad lineal de dimensión q menor que p de forma que, con tan sólo la información contenida en los datos q-dimensionales se pueden reconstruir nuestros datos iniciales, o casi.

La idea de que la variedad q-dimensional casi contenga a los datos dependerá del problema concreto en que nos encontrarnos y qué sea lo que interese. Por ejemplo, ocurrirá que con 3 variables se expliquen relativamente bien determinados datos, a pesar de que con 7 variables se conserve más información. Claro está que optando por la segunda opción habremos aceptado hacer un análisis de los datos más complejo.

El Análisis de Componentes Principales es una técnica que pretende encontrar una variedad lineal de $\mathbb{R}^p$de dimensión lo menor posible, llamémosla q, que cumpla que las proyecciones de los n puntos en esta variedad estén próximas a las posiciones de los puntos. Por lo tanto, podemos considerar que, a efectos prácticos, los puntos iniciales son q-dimensionales. El interés de esta técnica se basa en que normalmente permite obtener reducciones muy importantes en la dimensión inicial debido a las relaciones existentes entre las variables involucradas en el estudio.

Definición 2.2 *Supongamos que se dispone de p variables, entonces la matriz de covarianzas muestrales Σ es una matriz cuadrada p x p que tiene en la posición (i,i) la varianza σ_x de la variable i y en la posición (i,j) la covarianza σ_{xy} entre la variable i y la j. Donde*

$$\sigma_{xy} = \sum(x_i - \bar{x}) \cdot (y_i - \bar{y}) \quad , \qquad \sigma_x = \sqrt{\sum(x_i - \bar{x})^2}$$

Resulta que, si llamamos Σ a la matriz de covarianzas muestrales, entonces, la variedad lineal de dimensión q que minimiza las distancias de las proyecciones sobre ella a los puntos, es la variedad que contiene a la media de los puntos y está generada por q vectores propios ortogonales de Σ asociados a los q mayores propios de esta matriz. La demonstración puede encontrarse en Cuesta Albertos (2012).

Definición 2.2 *Se llaman Componentes Principales (CPs) a las q variables que acabamos de seleccionar.*

Como la distancia de los puntos a la variedad es la suma de los valores propios desde el $q+1$ al final, la demostración puede encontrarse en Cuesta Albertos (2012), se puede considerar que la información explicada por la variedad es $\sum_{j=1}^{q} \lambda_j \Big/ \sum_{j=1}^{p} \lambda_j$ que es la proporción de varianza explicada por las q primeras componentes principales. Por lo tanto, se busca un equilibrio entre maximizar esta expresión y reducir el número de variables empleadas.

En este trabajo partiremos de una serie de variables relacionadas con el desarrollo social y económico de nuestro país. Al aplicar esta técnica, obtendremos unas nuevas variables que resumirán la situación de las diferentes provincias según su nivel de desarrollo socio-económico. Nuestra intención es utilizarlas como índices del bienestar a nivel provincial. Pero, para ello, es imprescindible buscar una interpretación razonable a cada una de ellas. En otras palabras, determinar qué tipo de desarrollo (social, económico,…) mide cada una de ellas.

Este problema se resuelve mirando la relación existente entre las nuevas variables y las iniciales. Por un lado, las variables iniciales tienen una interpretación clara (la renta per-cápita, el índice de analfabetismo,…). Por otro lado, podemos considerar estas variables como una base de $\mathbb{R}^p$ y podemos calcular las coordenadas de los q vectores unitarios propios seleccionados en esta base. En algunas ocasiones, los valores relativos de estas coordenadas nos permitirán realizar la interpretación buscada[1].

Esta interpretación se suele hacer buscando aquellas coordenadas con valores, en módulo, relativos altos y desestimando las que presentan valores bajos. Sin embargo, puede suceder que (y, de hecho, es la situación que vamos a encontrar aquí) que no abundan los valores altos. Una posible solución es recurrir a elegir otra base del subespacio determinado por las Componentes Principales seleccionadas, cuyos integrantes sean más fácilmente interpretables. Ese problema se analiza en la siguiente subsección.

[1] Normalmente se consideran los coeficientes de correlación entre las variables mencionadas. Pero estos coeficientes vienen a ser una normalización de las coordenadas de que venimos hablando.

2.1.1. Cambios de base. Rotación Varimax.

Como hemos explicado, la interpretación de las nuevas variables se basa en el análisis de las coordenadas de los nuevos vectores en términos de las variables iniciales. Para realizar la interpretación, lo ideal sería disponer de unas componentes de forma que su proyección sobre las variables iniciales fuese lo más radical posible, es decir, que cada componente estuviese 'casi' íntegramente representada por un grupo de variables y, al mismo tiempo que esas variables no tuviesen importancia para el resto de las componentes.

En caso de que esto no suceda, siempre es posible cambiar la base del subespacio elegido de modo que las nuevas variables se acerquen a la situación descrita en el párrafo anterior. En la literatura se han propuesto diferentes posibilidades para realizar este cambio de base. Algunas de ellas mantienen el requerimiento de que la nueva base esté formada por vectores ortogonales (Varimax, Quartimax...) y otras prescinden de la ortogonalidad (Promax, Oblimin...).

En este trabajo hemos elegido el método Varimax porque da resultados mejores que las otras posibilidades consideradas. Esta rotación mantiene la ortogonalidad de las componentes.

Se puede encontrar más información sobre esta rotación y las otras mencionadas en Mardia, K.V. et al. (1979).

2.2. Análisis Cluster.

La segunda parte de este trabajo es tratar de encontrar agrupaciones de provincias de acuerdo con sus niveles de desarrollo de bienestar social. Las técnicas disponibles para ello se agrupan dentro del Análisis Cluster.

Estas técnicas pretenden encontrar grupos naturales en un conjunto de datos. El problema general es el siguiente: Se dispone de n datos p-dimensionales y se pretende dividirlos en cierto número de grupos de modo que dos datos dentro del mismo grupo sean similares entre sí y, además, estén claramente diferenciados de los miembros de otros grupos. Obviamente, en la práctica no suele presentarse una situación tan idílica como la descrita.

Hay dos tipos de técnicas de clustering:

1. Análisis no-jerárquico: inicialmente se establece el número de grupos que se desean formar y cada dato se asigna a uno de ellos tratando de conseguir la mayor homogeneidad posible.
2. Análisis jerárquico: Pueden ser aglomerativos o divisivos. Las técnicas incluidas en el primer tipo, comienzan por partir los datos en n grupos, porque cada dato comienza siendo su propio grupo. Sucesivamente el número de grupos se va reduciendo de uno en uno, ya que, en cada paso, se forman grupos más grandes

juntando los dos grupos existentes en el paso anterior que sean más parecidos entre sí. El proceso termina cuando todos los datos forman un único grupo. El análisis divisivo sigue un proceso inverso: En el paso inicial se dispone de un único grupo que contiene a todos los individuos. En cada uno de los pasos sucesivos, se incrementa en una unidad el número de grupos debido a la partición en dos de uno de los grupos existentes en ese momento. En ambos casos, el proceso finaliza cuando se consigue un nivel de homogeneidad adecuado.

A continuación se explica con un poco más de detalle el cluster jerárquico, que ha sido el empleado en la realización de este trabajo:

El punto clave en el proceso es la medida de la homogeneidad. Para ello, necesitamos disponer de algún tipo de distancia (aunque con una disimilitud es suficiente), entre elementos (los datos) y otra distancia (o disimilitud) entre los grupos de elementos.

Vamos a definir distancia y disimilitud:

Definición 2.2 *Sea X un conjunto no vacío. Diremos que una aplicación d: X x X -> ℝ es una disimilitud en X, si cumple que:*

1. *d(x, y) = d(y, x), para todo x, y en X.*
2. *d(x, y) ≥ 0, para todo x, y en X.*
3. *d(x, x) = 0, para todo x en X.*

Definición 2.3 *Sea X un conjunto no vacío. Diremos que una aplicación d: X x X -> ℝ es una distancia en X, si es una disimilitud que, además, verifica que*

4. *d(x, y) = 0 si y sólo si x = y.*
5. *d(x, y) ≤ d(x, z) + d(z, y), para todo x, y, z en X.*

Por lo tanto, las disimilitudes son ciertas relajaciones de las distancias que pueden ser útiles en determinado tipo de problemas. Por otro lado, existen muchos tipos de distancias en $\mathbb{R}^p$ (la euclídea, la de Manhattan o L_1,...), pero hay que tener en cuenta que la distancia o la disimilitud empleada tenga sentido en el problema en cuestión. En este trabajo utilizaremos una distancia euclídea asociada a una matriz diagonal. Los elementos de la diagonal servirán para tener en cuenta la importancia relativa de las variables involucradas.

Una vez que está fijada la disimilitud entre puntos, con el fin de poder realizar agrupaciones, debemos elegir una discrepancia entre grupos. Para ello existen muchas posibilidades. A continuación presentamos algunas de las más utilizadas. En estas definiciones, el símbolo "*dist*" representa la disimilitud elegida entre puntos.

1. <u>*Entorno próximo*</u>. La discrepancia entre dos grupos se define como la menor distancia que existe entre los elementos que forman cada grupo. Sean *A, B* dos grupos la discrepancia entre ellos se define como:

$$d(A,B) = inf\{dist(a,b): a \in A, b \in B\}.$$

2. *Entorno lejano.* La discrepancia entre dos grupos se define como la mayor disimilitud que existe entre los elementos que forman cada grupo. Sean A, B dos grupos la discrepancia entre ellos se define como:

$$\mathrm{d(A,B)} = \sup\{\mathrm{dist(a,b)}: \mathrm{a} \in \mathrm{A}, \mathrm{b} \in \mathrm{B}\}.$$

3. *Discrepancia promedio.* La discrepancia entre dos grupos es la media de las disimilitudes entre los elementos que forman cada grupo. Sean *A, B* dos grupos, con *n* y *m* el número de elementos respectivamente. La discrepancia entre ellos se define como:

$$d(A,B) = \sum dist(a,b)/(n \cdot m), \text{ donde } a \in A, b \in B.$$

4. *Método de Ward.* Este método es un poco diferente de los anteriores. En cada paso del análisis, se considera la posibilidad de la unión todos los posibles pares de grupos y se opta por la fusión de aquellos dos grupos que menos incrementen la suma de los cuadrados de las desviaciones al unirse, por desviación se entiende la distancia del elemento a la media del grupo. Este método es bastante costoso computacionalmente dado que, en cada paso, han de comprobarse todas las posibles combinaciones de grupos. A pesar de ello, es uno de los más usados en la práctica por los buenos resultados que proporciona.

No es muy difícil comprobar que, aunque "*dist*" sea una distancia, no todas las discrepancias son, ni siquiera, discrepancias. Por ello, hemos introducido el término discrepancia para referirnos a los conceptos introducidos.

Tanto la elección de la disimilitud entre puntos como la discrepancia entre grupos influyen de modo decisivo en la agrupación final obtenida y puede suceder que, dependiendo de la elección concreta que se haga, se obtengan agrupaciones completamente diferentes. Por ello es imprescindible disponer de un buen criterio a la hora de elegir un método u otro.

2.3. Correlación.

La correlación se usa para medir la fuerza de la relación entre dos variables. En general, se basa en un *coeficiente de correlación,* cuyo valor se encuentra entre -1 y +1, de modo que una correlación de 0 indica la ausencia de relación, mientras que +1 y -1 se corresponden con la mayor correlación posible, teniendo en cuenta que el signo determina la dirección de la relación: +1 indica que la relación es creciente y -1 que es decreciente.

El coeficiente de correlación lineal mide cuál es la intensidad de una la relación lineal entre las variables implicadas. Por tanto, el hecho de obtener un valor cercano a 0 del coeficiente de correlación no implica que no sean variables relacionadas, sino que no lo son linealmente.

El coeficiente de correlación lineal r se calcula del siguiente modo:

$$r = \sigma_{xy}/(\sigma_x \cdot \sigma_y) \ ,$$

donde σ_{xy} representa la covarianza entre x e y, σ_x denota la desviación típica de la variable x, y σ_y denota la desviación típica de la variable y. Estos términos se calculan como se explicó en la Sección 2.1.

Los valores que puede tomar el coeficiente de correlación r varían entre -1 y 1. Hablando un poco imprecisamente[2], podríamos decir que, dadas las variables x e y, este coeficiente mide el grado de similitud que puede obtenerse entre y y una transformación lineal de x (y viceversa). En este caso, un coeficiente de 0 querría decir que el coeficiente de x en la mejor transformación lineal sería 0, mientras que un coeficiente de +1 ó -1 se correspondería con una situación en la que y es, exactamente, una transformación lineal de x. La interpretación de este coeficiente sería la siguiente:

1. Si $r > 0$, la correlación lineal es positiva, es decir, si aumenta el valor de una variable también lo hace el de la otra. La correlación es tanto más intensa cuanto más se aproxime a 1.
2. Si $r < 0$, la correlación lineal es negativa, es decir, si aumenta el valor de una variable disminuye el de la otra. La correlación negativa es tanto más fuerte cuanto más se aproxime a -1.
3. Si $r = 0$, no existe correlación lineal entre las variables.

Hay que tener en cuenta que el hecho de que el coeficiente de correlación sea alto no implica la existencia de una relación de causa-efecto entre las variables en cuestión, lo que complica el análisis de la dependencia entre variables. Como ejemplo de esto último, puede pensarse en estas dos variables cuantitativas: el número de helados que se venden y el número de personas que se ahogan; es claro que no existe una relación causal entre ellas, pero sí que ocurre que cuando una de ellas aumenta también lo hace la otra (siempre y cuando no haya habido una catástrofe natural) porque ambos aumentos están provocados por un incremento de la temperatura.

[2] Con un poco más de precisión, la situación es la siguiente: Disponemos de n mediciones de cada variable. Por lo tanto, podemos entender que las p variables que estamos considerando son elementos del espacio Euclídeo n-dimensional usual con las coordenadas determinadas por las n mediciones mencionadas. En primer lugar, procedemos a centrar las variables restando a las puntuaciones de los individuos la media de las puntuaciones de los restantes individuos. Ahora, sucede que el coeficiente de correlación entre dos variables es igual al coseno del ángulo que forman los dos vectores n-dimensionales asociados a las variables centradas. Los casos de coeficientes ±1 se corresponden con los casos de colinealidad. En cambio el coeficiente 0 indicaría ortogonalidad entre los vectores considerados.

Capítulo 3. Resultados.

Este capítulo constituye la parte central de la memoria por cuanto en él se presentan los resultados de todo el estudio. Su contenido se estructura de la manera siguiente: En la Sección 3.1 se comentan algunos aspectos previos al estudio tales como la selección de las variables que hemos utilizado y el tratamiento que hemos hecho de las mismas antes de comenzar el estudio propiamente dicho. Los resultados del Análisis de Componentes Principales aparecen en la Sección 3.2. Los del Análisis Cluster se muestran en la Sección 3.3.

Los cálculos han sido realizados utilizando el lenguaje de programación R. En concreto, hemos utilizado la función *princomp* para el cálculo de las Componentes Principales, la función *principal* que precisa de la librería *psych* y la librería *GPArotation* para la rotación de las componentes y la función *agnes* de la librería *cluster* para el análisis cluster.

3.1. Preliminares. Selección y estandarización de variables.

Como hemos señalado anteriormente, uno de los objetivos de este trabajo es la confección de uno (o varios) índice(s) del bienestar en las provincias españolas utilizando para ello la información disponible en el Instituto Nacional de Estadística (INE) a nivel provincial.

Obviamente, en el INE se ofrece una cantidad ingente de información a este nivel. El primer paso es tratar de seleccionar adecuadamente las variables que vamos a considerar.

A continuación se exponen, en el orden que nosotros mismos seguimos, los motivos por los cuales descartamos los grupos de variables que se indican. Evidentemente, esta selección supone un grado alto de subjetividad, que creemos inevitable en un estudio de estas características.

A) Variables no disponibles a nivel provincial. Existen muchas variables que, en la actualidad, no se miden a nivel provincial sino al de comunidad autónoma.

B) Variables que no tienen una relación tan clara como otras con el bienestar social o variables que explican una misma dimensión del bienestar social que otras, o incluso, una misma variable explicada con distintos indicadores.

C) Variables cuya calidad podría no ser la adecuada, por problemas de subestimación, de significación estadística o de excesiva oscilación interanual. Dentro de este apartado, hemos eliminado variables como la tasa de suicidios, que no es una variable fiable, ya que muchas muertes de este tipo aparecen subestimadas y se muestran al resto de la sociedad como muertes naturales y, por lo tanto, no aparecen recogidas apropiadamente.

También incluimos en este capítulo otras variables que no presentan el suficiente número de datos para ser considerada una variable significativa. Por ejemplo, variables con una frecuencia de alrededor de diez casos por millón de habitantes que presentan unas fluctuaciones excesivas de un año a otro. Entre ellas se encuentran las muertes por homicidio.

D) Variables que consideramos como menos relevantes. Por ejemplo: el número de personas que van al cine por cada 100.000 habitantes, el índice de precios de consumo, el número de detenidos, la producción total de energía eléctrica,…

E) Variables que ya están explicadas por otras. Para identificar estos casos hemos utilizado la matriz de correlaciones entre las variables que quedaban después de ejecutar las eliminaciones descritas en los pasos anteriores. Dada una lista de variables, de acuerdo con lo que se explicó en la Sección 2.3 esta matriz es cuadrada, su dimensión coincide con el número de variables y en cada celda, contiene la correlación existente entre las variables que correspondan. Tal y como hemos señalado en la Sección 2.3, el hecho de que la correlación entre dos variables sea alta no implica que exista una relación causal, por lo tanto, no se pueden descartar variables por este motivo y hemos de analizar caso por caso a qué es debida esa alta correlación. Por ejemplo, tanto la tasa de juventud (el porcentaje de personas menores de 15 años en la población total) como la tasa de envejecimiento (el porcentaje de personas mayores de 65 años en la población total) presentaban altas correlaciones con la tasa de dependencia (la suma de las dos tasas anteriores). Debido a que no queríamos variables redundantes que pudiesen afectar sobremanera a la elección de las componentes principales, descartamos las dos primeras.

Finalmente, después de descartar unas 200 variables de acuerdo con lo expuesto anteriormente, quedaron 22 variables, que a nuestro entender, son representativas del bienestar en la nación. A continuación incluimos su lista junto con una breve descripción de cada una de ellas.

1) Inmigración extranjera (*InmExtr10*): Inmigrantes procedentes del extranjero por cada 100.000 habitantes durante el año 2010. No se incluyen personas procedentes de otras provincias españolas.

2) Potencial turístico (*Turismo10*): Pernoctaciones en establecimientos hoteleros por cada 100 habitantes durante el año 2010.

3) Analfabetismo funcional. (*SinEst10*): Proporción de población de 16 y más años analfabeta/sin estudios en el año 2010.

4) Divorcios. (*TDIV11*): Número de divorcios por cada 100 matrimonios durante el año 2011.

5) Tasa de paro. (*PARO10*): Proporción de parados durante el año 2010.

6) Tasa de Mortalidad infantil. (*MOR06-10*): Tasas de mortalidad infantil desde el año 2006 hasta el 2010. Se consideran los nacidos vivos que mueren en el primer año de vida por cada 1.000 nacidos vivos.

7) *PIB per cápita (PIBC10*): Producto Interior Bruto *per cápita* en el año 2010.

8) Índice de fecundidad *(IFec10*): Indicador coyuntural de fecundidad durante el año 2010. Es el número medio de hijos por mujer en edad fértil (de 15 a 49 años).

9) Edad media de la maternidad (*Mat10*): Edad media a la maternidad en el año 2010.

10) Tasa de dependencia (*TDep10*): Número de personas entre 0 y 14 y mayores de 65 años dividido entre el número de personas entre 15 y 64 años.

11) Edad media (*EM10*): Edad media de la población en el año 2010.

12) Tasa de prestación por desempleo (*Tpresta10*): Proporción de beneficiarios de prestaciones económicas por desempleo durante el año 2010.

13) Prestaciones de renta activa de inserción (*PRenAdm10*): Proporción de beneficiarios de prestaciones económicas por desempleo de renta activa de inserción durante el año 2010.

14) Porcentaje de población activa en la agricultura (*TAgri11*): Proporción de población activa que se dedica a la agricultura en el año 2011.

15) Porcentaje de población activa en la industria (*TInd11*): Proporción de población activa que se dedica a la industria en el año 2011.

16) Porcentaje de población activa en la construcción (*TCons11*): Proporción de población activa que se dedica a la construcción en el año 2011.

17) Porcentaje de población activa en los servicios *(Tserv11*): Proporción de población activa que se dedica al sector servicios en el año 2011.

18) Abandono (*TAband11*): Proporción de población activa que en el año 2011 habían dejado su último empleo hacía tres años o más.

19) Primer empleo (*TPrEmpleo11*): Proporción de población activa que estando parada buscan el primer empleo en el año 2011.

20) Esperanza de vida a partir de 65 años (*EspVida65_10*): Esperanza de vida a partir de los 65 años en el año 2010.

21) Esperanza de vida (*EspVida10*): Esperanza de vida al nacer en el año 2010.

22) Desarrollo de infraestructura de transporte terrestre (*KM10*): Kilómetros de autopistas y autovías por 1000 Km^2.

La mayoría de las variables están referidas al año 2010 por no existir datos posteriores. Sin embargo, hay algunas variables que no estaban disponibles para dicho año. En estos casos, la variable en cuestión ha sido reemplazada por la correspondiente del año 2011, de modo que el estudio fuese lo más reciente posible.

Una vez realizada la selección anterior, se procedió a la estandarización de las variables involucradas para que las unidades de medida no tuvieran influencia en el resultado final. Nótese que las componentes principales están totalmente determinadas por los valores propios de la matriz de covarianzas, y, estos valores, a su vez, están muy influidos por la escala utilizada en las variables.

Por otro lado, la estandarización, también lleva al centrado de las variables. Aunque no es muy importante, con ello se consigue que la variedad determinada por las componentes principales incluya al origen. Es decir, tenemos un subespacio vectorial lo cual simplifica algo el tratamiento del problema.

Hay que mencionar también el hecho de que para este estudio se han descartado Ceuta y Melilla, debido a que su reducido tamaño en comparación con el resto, introducía un ruido que desvirtuaba el estudio.

3.2. Análisis de componentes principales.

Una vez se han seleccionado las variables representativas del bienestar con las cuales se va a trabajar, nos ocuparemos de resumir la información que se esconde detrás de estas variables utilizando el Análisis de Componentes Principales y se tratará de dar una explicación, si es posible, a estas componentes.

La siguiente tabla contiene la desviación típica, la proporción de varianza explicada por cada una de las 7 primeras componentes principales así como la proporción de varianza explicada acumulada de las 7 primeras componentes principales:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7
Desviación típica	2.711	2.216	1.717	1.254	1.057	0.857	0.788
Proporción de varianza explicada	0.334	0.223	0.134	0.072	0.051	0.033	0.028

Proporción de varianza acu.	0.334	0.557	0.691	0.763	0.813	0.847	0.875

Tabla 3.1. *Resultados del Análisis de Componentes Principales para las 7 primeras componentes.*

En la Tabla 3.1 llama la atención el hecho de que las tres primeras componentes principales contienen el 69% de la información que se encuentra en nuestras variables, lo cual es muy satisfactorio porque sacrificando el 31% se reduciría nuestro problema inicial de 22 variables a tan sólo 3 nuevas variables.

Sin embargo, una "rule of thumb" muy popular en la literatura recomienda seleccionar las 5 primeras componentes. Esto es así porque cualquiera de esas componentes explica más proporción de varianza que cualquiera de nuestras variables iniciales (recuérdese que la estandarización hacía que las variables iniciales tuvieran varianza igual a uno), dado que su desviación típica es mayor que 1.

Por otro lado, en la literatura también se recomienda analizar como decrece la proporción de varianza explicada por cada componente y buscar el punto a partir del cual este decrecimiento es lineal porque se considera que, a partir de este momento, las componentes únicamente explican un ruido aleatorio. De acuerdo con este procedimiento, habría que elegir cinco componentes. Pues es a partir de la secta cuando el decrecimiento es lineal.

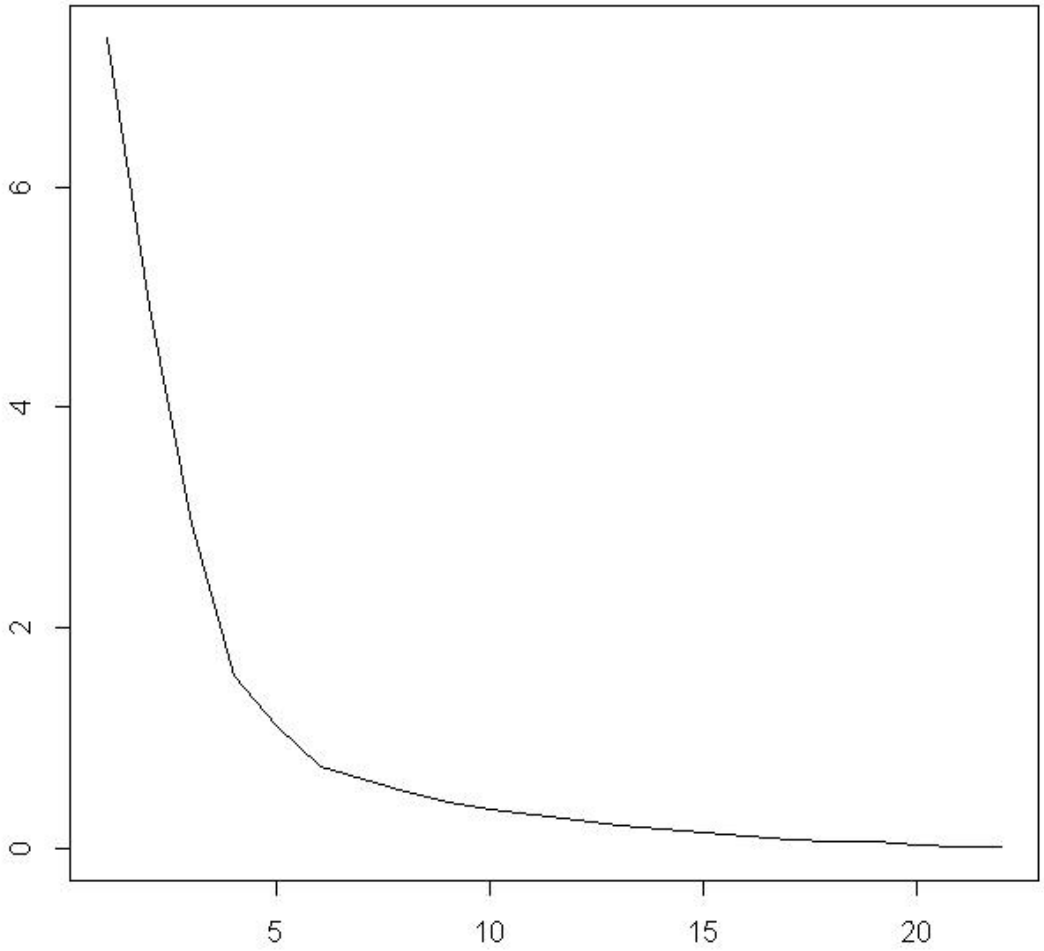

Figura 3.1. *En el eje y se muestra la varianza explicada por cada una de las 10 primeras componentes.*

De acuerdo con lo indicado en la Sección 2.1, el paso siguiente es buscar una interpretación a las componentes obtenidas, para lo que hay que basarse en los coeficientes de estas componentes en términos de las variables iniciales. Pueden verse estos coeficientes en la Tabla 3.2.

Loadings:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
TDIV11	-0.161	0.214	0.318	-0.223	
SinEst10	-0.263	-0.146	-0.170	0.104	0.193
PARO10	-0.342				0.101
MORO6.10	-0.191				-0.645
KM10		0.167	0.405		
PibC10	0.247	0.267	0.121		
IFec10		0.328		0.328	0.110
Mat10	0.293	-0.120	0.254	0.112	
TDep10	0.184	-0.287	-0.257		-0.102
EM10	0.261	-0.278		-0.135	-0.209
Tpresta10	-0.338				
PRenAdm10	-0.280	-0.184	0.155		0.134
TAgri11		-0.316	-0.273		-0.140
TInd11	0.266	0.103		0.225	-0.319
TCons11		-0.285		-0.421	
TServ11	0.141	-0.181	0.338	-0.302	
TAband11		-0.209	0.431	0.107	0.146
TPrEmpleo11		-0.319	0.259		0.216
EspVida65_10	0.319		-0.139		0.274
EspVida10	0.290		-0.157		0.364
Turismo10	-0.109	0.109	-0.140	-0.607	-0.136
InmExtr10		0.358	-0.100	-0.244	0.106

Tabla 3.2. *Coeficientes de las 5 primeras componentes principales en términos de las variables iniciales. No se muestran los coeficientes menores que 0.1 en valor absoluto.*

Como se puede observar, los coeficientes toman valores relativamente bajos. Salvo unas pocas excepciones, todos ellos presentan valores absolutos por debajo de 0'4, lo que dificulta considerablemente la interpretación de las variables.

Por ello, recurrimos a buscar una nueva base del subespacio encontrado cuyos vectores tengan coeficientes con valores mayores en módulo. Entre las técnicas manejadas, resultó que la rotación Varimax (ver Sección 2.1.1) fue la que ofrecía resultados más satisfactorios. A continuación se muestran los resultados obtenidos. Hay que indicar que a la quinta componente se le ha cambiado el signo para la construcción de esta tabla. Con ello se consigue que valores mayores de esta variable se correspondan con mejores condiciones de bienestar y que sea más sencilla su interpretación. Este cambio de signo es irrelevante pues genera la misma variedad lineal.

```
Loadings:
              Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
TDIV11        -0.161  0.819        -0.151 -0.280
SinEst10      -0.812 -0.261 -0.108        -0.138
PARO10        -0.845  0.181 -0.134        -0.334
MORO6.10      -0.172                      -0.845
KM10           0.361  0.672         0.283
PibC10         0.787  0.378 -0.137 -0.131  0.241
IFec10                0.319 -0.827
Mat10          0.693         0.358  0.513  0.201
TDep10         0.262 -0.845  0.236  0.117  0.133
EM10           0.529 -0.586  0.567  0.111
Tpresta10     -0.761  0.202 -0.199        -0.442
PRenAdm10     -0.775  0.106  0.210  0.291 -0.305
TAgri11       -0.210 -0.804  0.163        -0.153
TInd11         0.849        -0.160  0.201
TCons11       -0.337 -0.280  0.703 -0.142
TServ11        0.278  0.172  0.781  0.248
TAband11      -0.137  0.253  0.472  0.705
TPrEmpleo11   -0.336         0.535  0.608
EspVida65_10   0.580 -0.278                 0.688
EspVida10      0.503 -0.208                 0.743
Turismo10     -0.188  0.172  0.223 -0.821 -0.120
InmExtr10             0.457 -0.364 -0.618  0.204
```

Tabla 3.3. *Coeficientes de las 5 primeras componentes principales, después de aplicar la rotación Varimax, en términos de las variables iniciales. No se muestran los coeficientes menores que 0.1 en valor absoluto.*

La Tabla 3.3 muestra la misma información que la Tabla 3.1, pero referida a los vectores que componen la nueva base.

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Proporción de varianza explicada	0.269	0.171	0.145	0.117	0.112
Proporción de varianza acumulada	0.269	0.440	0.585	702	0.813

Tabla 3.4. *Proporciones explicadas y acumuladas por las 5 primeras componentes después de aplicar la rotación Varimax.*

La Tabla 3.2 muestra que el resultado de la rotación es similar a lo que se esperaba. Las variables iniciales, mayoritariamente, están representadas en una sola componente: En general, dada una variable inicial, tenemos 5 coeficientes asociados a

ella y sólo uno es alto (digamos que superior en módulo a 0.7), siendo los módulos de los demás inferiores a 0.2. Este hecho facilita bastante la interpretación de estas componentes o factores.

Por otro lado, según se desprende de la Tabla 3.3, la proporción de varianza explicada con 5 componentes rotadas es la misma que con las 5 componentes principales, aunque el reparto de esta varianza entre ellas difiera.

Presentamos la interpretación de las componentes rotadas en la siguiente subsección.

3.2.1. Interpretación de los factores.

En este apartado presentamos la interpretación de los factores obtenidos como resultado del análisis descrito. Para ello se va a tener en cuenta los coeficientes de mayor valor absoluto que son aquellos que más importancia tienen:

1. **Factor 1.** Está relacionado positivamente con que haya una gran proporción de la población que se dedique a la industria, y que el PIB *per cápita* y la edad de la maternidad sean altas. En cambio, está relacionado negativamente con que el paro sea alto; que haya mucha gente que no tenga estudios o sea analfabeta; y, que la proporción de la población que percibe prestación por desempleo y que recibe prestaciones por renta activa de inserción sean altas.

   Por lo tanto, este factor parece estar muy vinculado al **desarrollo económico** de las provincias.

2. **Factor 2**. Esta componente está caracterizada por su relación con la tasa de divorcios y con el número de kilómetros de autovías y autopistas por cada 1000 kilómetros cuadrados. Y está relacionada negativamente con la tasa de dependencia, relacionada positivamente con que la proporción de la población se dedica a la agricultura sea alta y con una edad media de la población alta. Si nos fijamos, en las zonas urbanas es donde se producen más divorcios y donde más kilómetros de carreteras hay, además en las zonas rurales la tasa de dependencia es mayor debido a una población envejecida, que aumenta la edad media, y también se dedican más a la agricultura.

   Por todo ello, vamos a llamar a este factor el **factor urbanización.**

3. **Factor 3**. Para este factor no hemos encontrado un nombre apropiado, pero sí sabemos qué es lo valora por encima de las demás variables. Tienen mucha importancia en este factor la proporción de la población que se dedica a la construcción y la que se dedica al sector servicios, en el sentido positivo. También la tiene aunque en el sentido negativo, el que haya un alto índice de fecundidad.

4. **<u>Factor 4</u>**. En este factor pesan mucho (positivamente) las siguientes variables: la proporción de la población activa que ha abandonado su empleo hace 3 años o más y la que está en el paro y no encuentra su primer empleo. Además influyen en él negativamente el turismo y la inmigración extranjera.

   Por lo que, en resumen, este factor está vinculado a que haya un grave **deterioro del mercado laboral** y a la **influencia extranjera** (ya sea en forma de turismo o extranjeros que residen la provincia).

5. **<u>Factor 5</u>**. Este factor será el de **desarrollo demográfico** porque está vinculado a la esperanza de vida al nacer, a la esperanza de vida a partir de los 65 años y penaliza, es decir, está influido negativamente por la mortalidad infantil.

Una vez se han encontrado unos factores bastante razonables que resumen aceptablemente bien las variables originales, se va a tratar de observar si existe alguna relación entre ellos. Para ello, una vez se han estandarizado los datos, se ha procedido a realizar un cambio de base y escribir las coordenadas de cada una de las provincias que estaban representadas por las variables originales, en $\mathbb{R}^{22}$, en una la base cuyos cinco primeros elementos son los factores encontrados generada por los factores en $\mathbb{R}^5$.

Entre todas las posibles parejas de gráficos, a continuación mostramos dos que comparan los factores 1, 2 y 5 por ser aquellos en los aparece recogida una relación más clara entre las diferentes provincias.

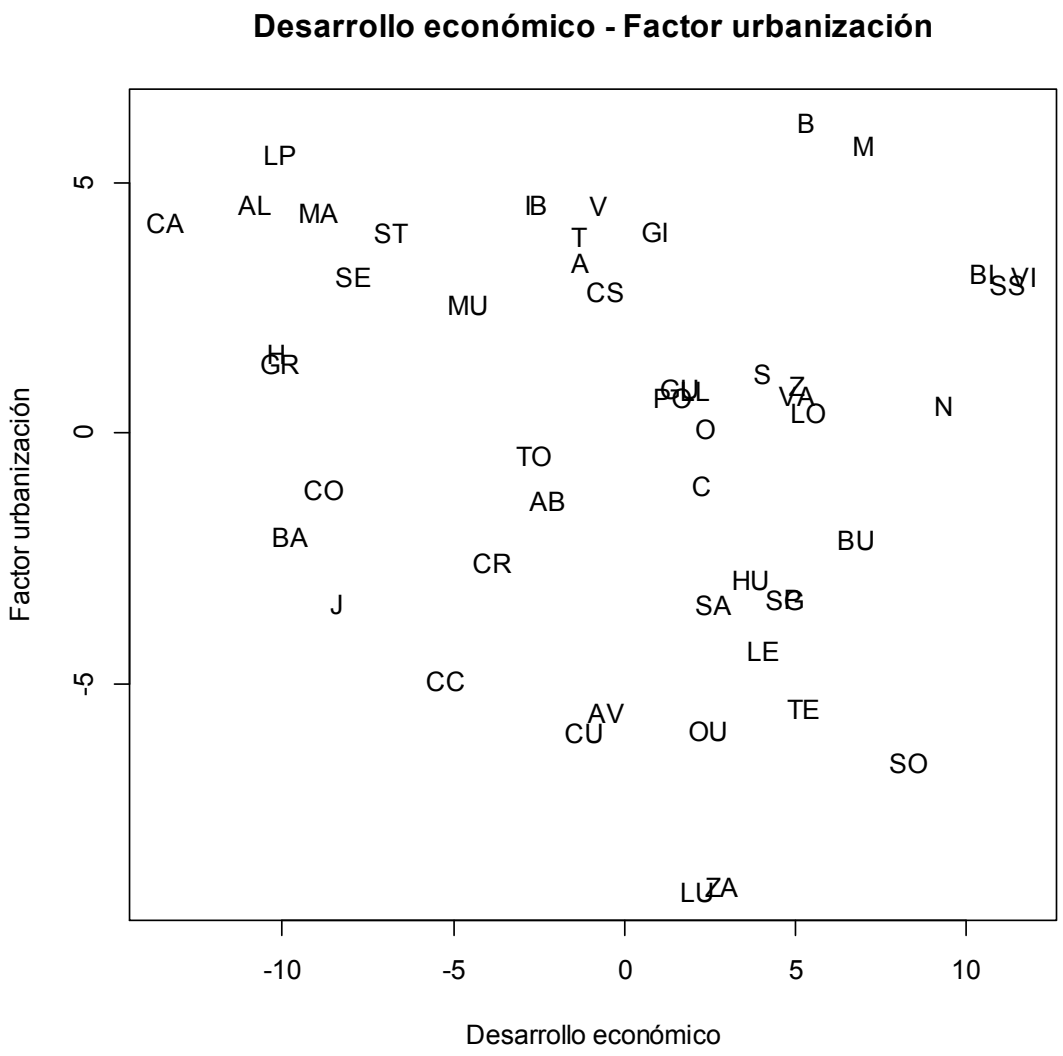

Figura 3.2. S*e muestran las 50 provincias en el plano determinado por los factores 'urbanización' (en el eje y) y 'desarrollo económico' (en el eje x). Las provincias se identifican utilizando la simbología de las antiguas matrículas de automóvil.*

Se observa a partir de la figura que podríamos tener una ordenación de las provincias bastante coincidente con la imagen que de ellas se tiene si proyectásemos sobre la diagonal principal del plano determinado por estas dos componentes rotadas. En la esquina superior derecha se encontrarían: Barcelona, Madrid y las provincias vascas, en la esquina inferior izquierda se hallarían Jaén, Badajoz, Cáceres, Córdoba...

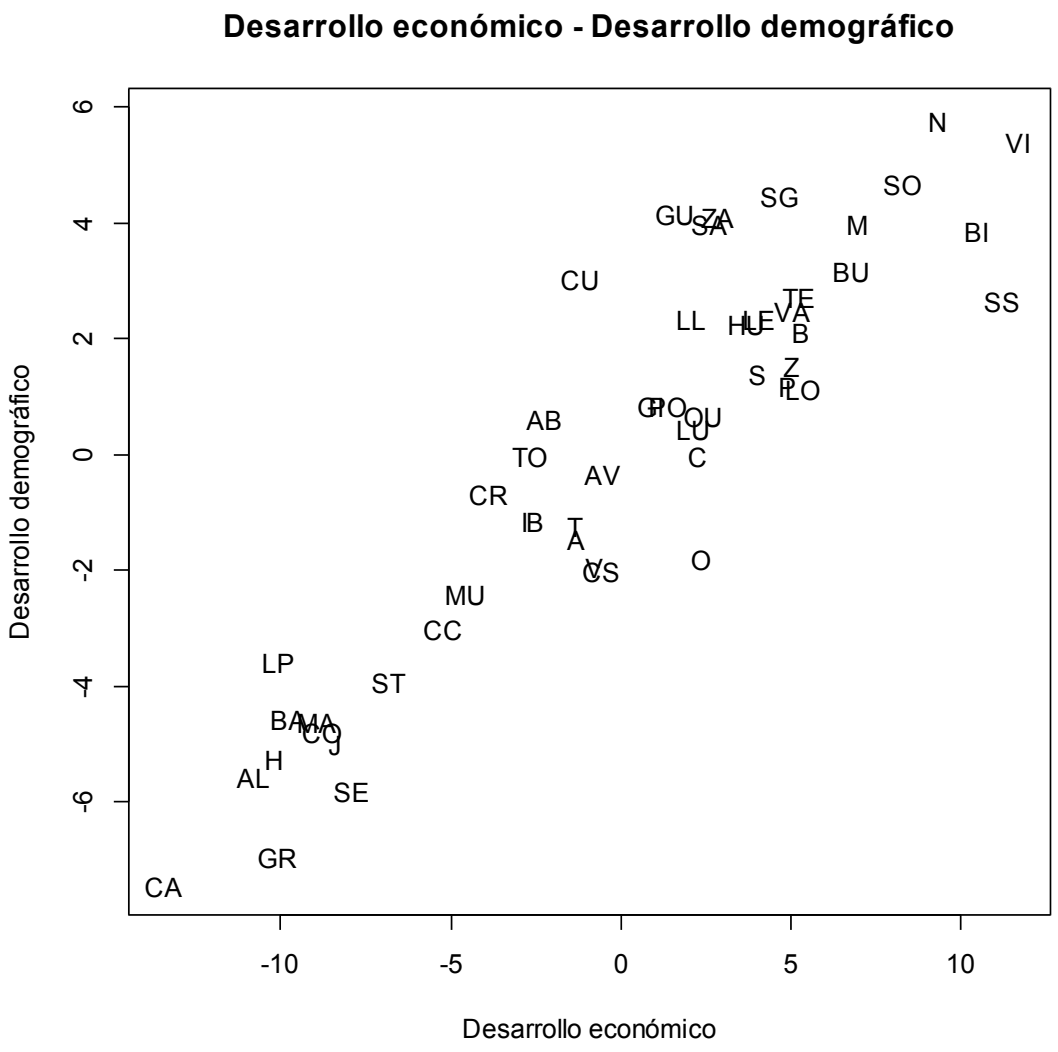

Figura 3.3. *Se muestran las 50 provincias en el plano determinado por los factores 'desarrollo demográfico' (en el eje y) y 'desarrollo económico' (en el eje x). Las provincias se identifican utilizando la simbología de las antiguas matrículas de automóvil.*

Se observa en la figura que existe una relación muy grande entre el factor 'desarrollo económico' y el factor 'desarrollo demográfico', de hecho la correlación lineal es de 0.91. No es muy sorprendente la existencia de esta relación, sin embargo sí que es llamativa la fuerza de la misma.

3.3. Análisis Cluster.

En esta sección nos ocupamos de tratar de encontrar agrupaciones de las provincias españolas teniendo en cuenta su nivel de desarrollo socio-económico. Para ello, utilizaremos el Análisis Cluster. De acuerdo con lo indicado en la Sección 2.3, antes de nada tenemos que determinar la distancia o disimilitud a utilizar entre individuos y, también, el método a emplear para el cálculo de discrepancias entre grupos.

Pero, antes de nada, tenemos que decidir las variables que vamos a utilizar. Lo ideal para hacerlo sería disponer de la mayor cantidad de información posible, lo cual nos conduciría irremediablemente a considerar todas las variables. Pero, el uso de todas las variables, puede implicar el utilizar varias veces la misma información debido a las relaciones existentes entre ellas. Por ello, es aconsejable utilizar menos variables con la menor relación entre ellas. Cuando se realizó el análisis de CPs se obtuvo que con las cinco primeras componentes se resumía el 81% de la información, así que parece sensato aplicar el análisis para ver si con esas cinco componentes se obtienen unos grupos razonables.

Para ello, inicialmente, utilizamos la métrica euclídea usual aplicada a las proyecciones de las provincias en el subespacio de dimensión 5 determinado por las cinco primeras CPs. Respecto de la distancia entre grupos, el procedimiento que mejores resultados produjo fue el método de Ward. Se puede ver el resultado en la Figura 3.4.

Esta figura contiene el "dendograma" del análisis. Su interpretación es la siguiente: De cada grupo sale una línea horizontal que termina en el punto (de la escala "Height") en que ese grupo se une con otro. De este modo, como en el paso inicial cada provincia es un grupo, tenemos que de cada una de ellas sale una línea. A continuación, resulta que una de las parejas más homogéneas la forman Toledo y Albacete (hacia la mitad de la figura) porque son las que antes se unen y lo hacen a una altura de, aproximadamente, 1. Bastante homogéneas también son las parejas Almería-Huelva y Alicante-Valencia (en la parte superior). En cambio, Cádiz es, quizás la que más tiempo tarda en unirse a otro grupo (formado por Córdoba, Badajoz, Jaén, Granada y Sevilla).

Para determinar la presencia de grupos, se recomienda buscar una zona amplia en la que no se producen agrupamientos y "cortar" por ahí. En la figura que nos ocupa, esta zona se presenta en el intervalo de alturas del 6 al 8 y, por lo tanto, obtendríamos cinco grupos.

A pesar de que los grupos que resultan son razonables, no son totalmente satisfactorios ya que buscamos una clara distribución territorial en España. Una razón para ello, puede ser que la proporción de varianza explicada por cada componente es distinta, por ejemplo, la primera explica el 33.4%, la segunda el 22.3% y la tercera el 13.4%, por lo tanto, la primera variable contiene dos veces y media la cantidad de información que la tercera. En cambio, si utilizamos la métrica euclídea usual, estamos dando la misma importancia a todas ellas.

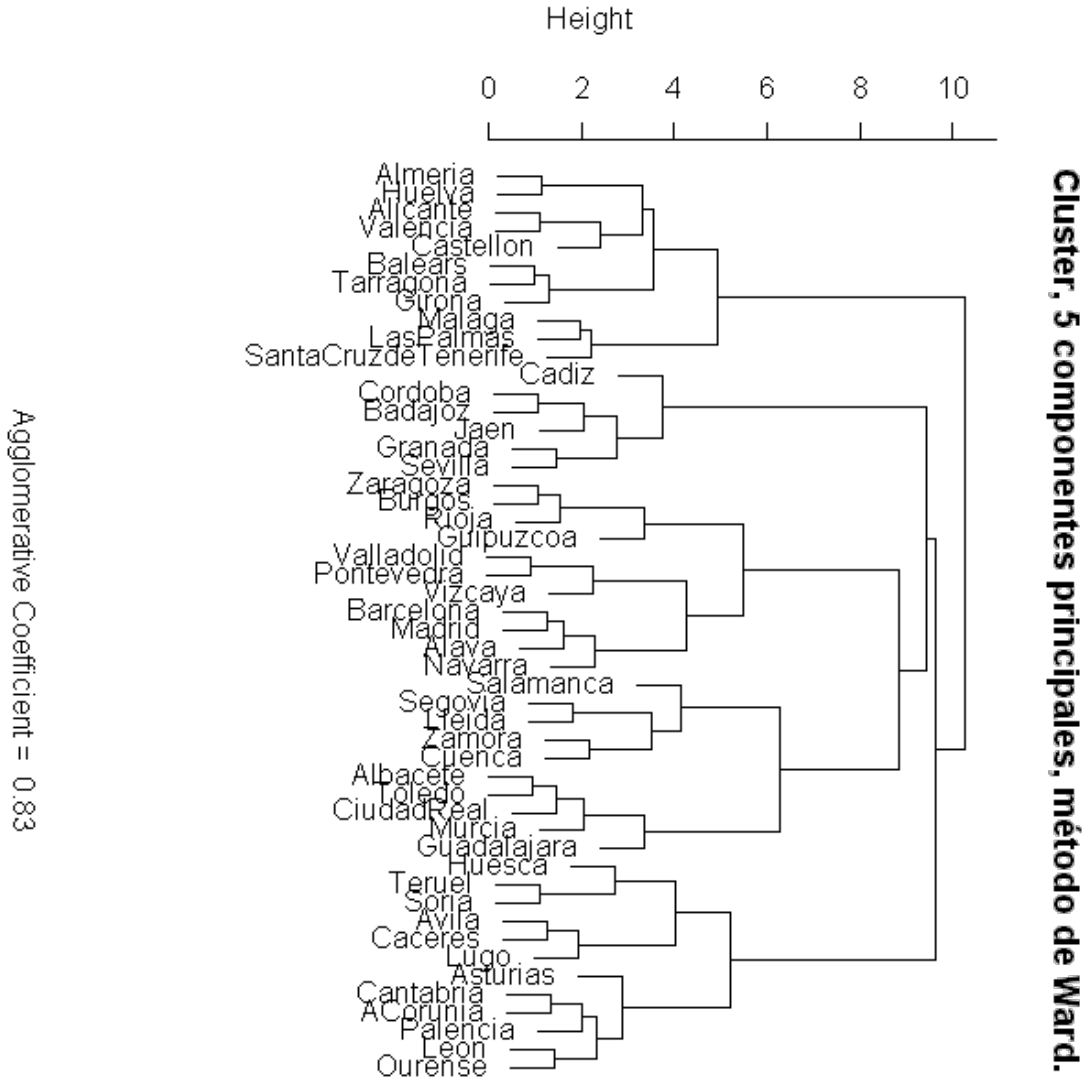

Figura 3.4 *Agrupaciones de las provincias realizadas considerando las cinco primeras componentes principales y el método de Ward. En el eje y se muestra la distancia a la que se forman los distintos grupos. La figura aparece girada para facilitar la visualización de los grupos.*

Para solucionar este problema pensamos en utilizar una distancia euclídea con pesos que reflejasen la importancia de cada coordenada. De este modo, si *x, y* son las coordenadas de dos provincias cualesquiera en la base determinada por las 22 Componentes Principales, podemos definir la distancia entre ellas del siguiente modo:

$$dist(x,y) = \sqrt{\sum_{i=1}^{22} qi \cdot (x_i - y_i)^2} \ ,$$

donde qi es la proporción de varianza asociada a la componente i. Nótese que ahora no es necesario limitarse a considerar las proyecciones en el subespacio determinado por las primeras 5 componentes porque el problema de doble medición de que hablábamos

con anterioridad, desaparece porque los pesos adjudicados a cada componente son proporcionales a su importancia.

Realizando un análisis cluster, que utilice esta distancia definida de este modo, se obtiene el siguiente dendograma:

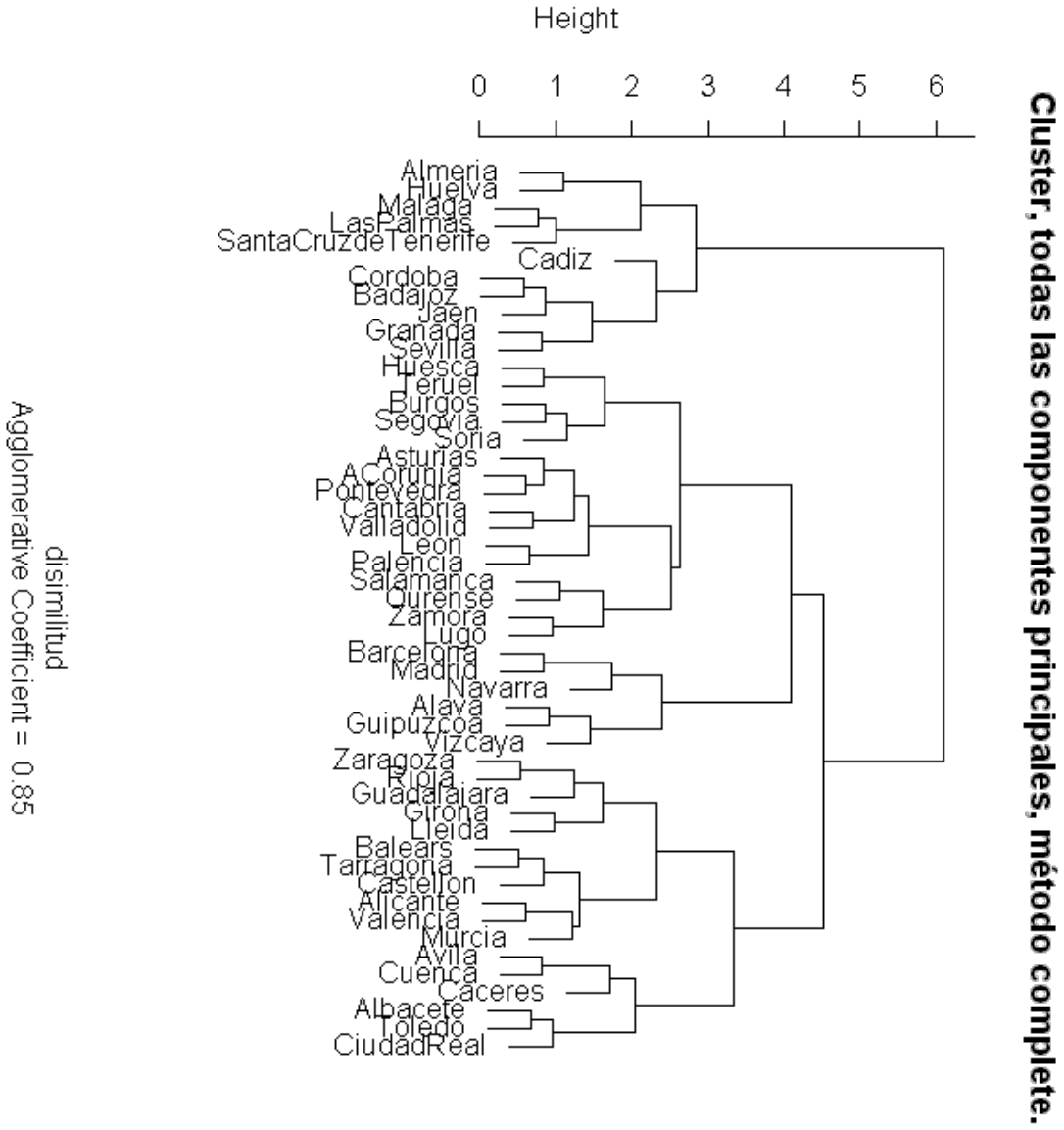

Figura 3.5. *Agrupaciones de las provincias realizadas considerando todas las componentes principales con pesos y el método del entorno lejano. En el eje y se muestra la distancia a la que se forman los distintos grupos. La figura ha sido girada.*

En la Figura A.1 del anexo puede verse cómo con otro método (el de Ward) se obtienen unas agrupaciones muy similares. Este hecho puede tomarse como un indicio de la solidez de la agrupación encontrada.

En este caso, la zona de corte no está tan claro como en el anterior. Sin embargo existen tres zonas posibles de corte: A la altura del 5 (con lo que tendríamos dos grupos:

El primero formado por Almería, Huelva, Málaga, Las Palmas y Tenerife y el otro por el resto). La siguiente está en el 3.25 lo que proporcionaría un total de 5 grupos. Y la tercera estaría a la altura del 2.5 y produce una partición en 11 grupos que, a su vez, son subgrupos de los anteriores. Esta última partición no es muy clara en 2 de los 5 grupos anteriores debido a Cáceres en uno de ellos y a Cádiz en otro, por ello no se ha escogido como la partición principal sino como la secundaria. No obstante, la hemos incluido porque puede servir para matizar un poco la principal.

Finalmente, pareció razonable coger dos particiones: La obtenida a nivel del 3.25 como principal, que se subdivide de acuerdo con la siguiente (2.5).

A continuación, se muestra un mapa de España en el que aparecen coloreadas las provincias en función de a qué grupos pertenezcan. Los 5 grupos principales se diferencian por sus colores. Los subgrupos aparecen definidos por medio de mallas. Hay que advertir que provincias con la misma malla, pero diferente color no tienen nada en común.

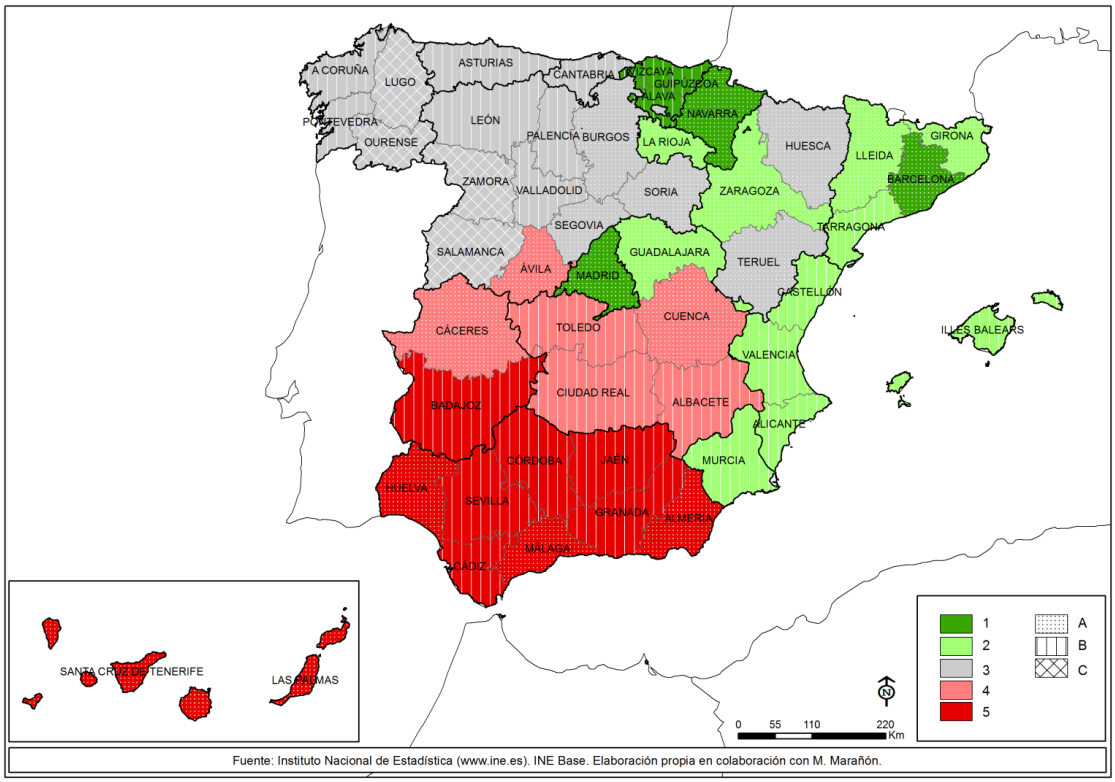

Figura 3.6. *Partición en grupos de las provincias españolas según su índice de bienestar.*

El resultado final son estos cinco grupos de provincias:

El grupo 1 que engloba a las tres provincias vascas (Guipúzcoa, Vizcaya y Álava), además de Navarra, Madrid y Barcelona, tiene en común ser provincias muy

desarrolladas económicamente, altamente urbanizadas, y, por tanto, con un alto grado de bienestar social.

El grupo 2, que abarca las provincias del valle del Ebro y todo el litoral mediterráneo, presenta, el segundo nivel de desarrollo económico y de urbanización (favorecida en el caso de las provincias costeras mediterráneas por el desarrollo turístico) y conforman con el grupo anterior lo que algunos economistas han llamado "la lambda invertida del desarrollo económico español", que lo es también del bienestar social, y por tanto un estructura territorial nítidamente definida.

Los grupos 5 y 4, por el contrario, conformados por las provincias andaluzas y canarias, así como la provincia de Badajoz, de una parte, y las provincias castellano-manchegas y Cáceres y Ávila de otra, se opone al primer y segundo grupo: presentan las más bajas rentas *per cápita*, los más bajos niveles de desarrollo social, mayor dependencia del sector primario, y por tanto el grado de bienestar social más bajo del país, más el 5º grupo que el 4º.

El grupo 3, finalmente, que queda conformado a partir de las provincias del cuadrante noroccidental, al que se suman provincias de Teruel y Huesca, ocupa posiciones intermedias en cuanto a bienestar social. Se trata de provincias con un fuerte componente rural, envejecidas, tradicionalmente emigratorias, pero que presentan esperanza de vida, un desarrollo demográfico y socioeducativo altos, y un nivel de bienestar social medio.

Capítulo 4. Conclusiones.

Como los aspectos más representativos de este trabajo se van a resaltar cuatro resultados:

1) El hecho de haber encontrado una relación tan fuerte entre el factor demográfico y el factor desarrollo económico, que hace pensar que el incremento del todas esas variables que están relacionadas positivamente con el factor económico (*PIB per cápita*, desarrollo industrial …) y la disminución de todas las que se encontraban relacionadas negativamente (el paro, la tasa de prestaciones por desempleo…), supondrían un aumento en la esperanza de vida y una disminución en la mortalidad infantil.

2) La clara distribución territorial de las provincias españolas en materia de bienestar social. Habiendo partido de variables representativas del bienestar social se ha llegado a una estructura territorial nítida y a unos grupos muy bien definidos.

3) Hay que resaltar que ha sido la utilización de los pesos, tomados de la proporción de varianza explicada de las componentes principales, al definir la distancia entre las provincias lo que ha conducido a unos resultados tan satisfactorios en el Análisis Cluster.

4) Y por último, cabe comentar que si este trabajo se hiciera con indicadores de décadas anteriores tal vez las diferencias no fueran excesivas respecto al momento actual, lo que probaría el carácter estructural que las desigualdades socio-territoriales, respecto al bienestar social, presentan en España. Desde las ciencias sociales (Geografía, Economía, Sociología…) (Reques, 2011) ya se había llegado a esta conclusión, pero tal vez nunca se había hecho a partir de los métodos del Análisis Multivariante, como los que en esta trabajo hemos utilizado.

5.Agradecimientos.

A Pedro Reques por enseñarme los conocimientos necesarios sobre Geografía para poder llevar a cabo este trabajo y por toda su ayuda.

A Juan Antonio Cuesta por ayudarme y guiarme en la resolución del problema.

A María Marañón por la realización del mapa de España.

6. Fuentes y bibliografía.

Fuentes:

Instituto Nacional de Estadística. *INEbase,* 2013.

Lenguaje de programación:

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL *http://www.R-project.org.*

Bibliografía:

Blanco, A. y Darío Díaz, D. (2005). "El bienestar social: su concepto y medición". *Psicothema 2005. Vol. 17, nº 4, pp. www.psicothema.com*

Cuesta Albertos, J. A. (2012). *Análisis Multivariante*. Universidad de Cantabria, Santander, España.

Keyes, C. (1998). "Social well-being. *Social". Psychology Quarterly, 61*, 121-140.

Keyes, C., Shmotkin, D. y Ryff, C. (2002). Optimizing well-being: the empirical encounter of two traditions. *Journal of Personality and Social Psychology, 82*, 1.007-1.022.

Mardia, K.V., Kent, J.T. y Bibby, J.M. (1979). *Multivariate Analysis.* Academic Press. Londres.

Smith, D. (1997). *Geografía del Bienestar*. Barcelona, Oikos-Tau.

Reques Velasco, P. (1986): "El análisis factorial como técnica de regionalización de los espacios agrarios" *Geografía Teórica y Cuantitativa: conceptos y métodos*. Oviedo, Universidad de Oviedo, 1986, págs 113-127.

Reques Velasco, P (1986): "Localización de centros funcionales y delimitación de áreas deprimidas en los espacios rurales a partir del análisis factorial". *Métodos Cuantitativos en Geografía: Enseñanza, Investigación y Planeamiento*. Asociación de Geógrafos Españoles. Grupo de Métodos Cuantitativos. Madrid, Universidad Complutense, 1986, págs. 222-237.

Reques Velasco, P. (2011): *El factor D: Los nueve retos demográficos de la España actual.* Bilbao, Cinco Días.

Páginas web:

-Bienestar:

http://es.wikipedia.org/wiki/Bienestar_social
http://www.definicionabc.com/social/bienestar.php
http://definicion.de/bienestar/
http://www.psicothema.com/pdf/3149.pdf

-Métodos estadísticos

http://www.uv.es/ceaces/multivari/cluster/met_ward.htm
http://www.ditutor.com/estadistica_2/correlacion_estadistica.html
http://es.wikipedia.org/wiki/Correlaci%C3%B3n
http://www.aulafacil.com/CursoEstadistica/Lecc-12-est.htm

ANEXO I. Figuras y tablas.

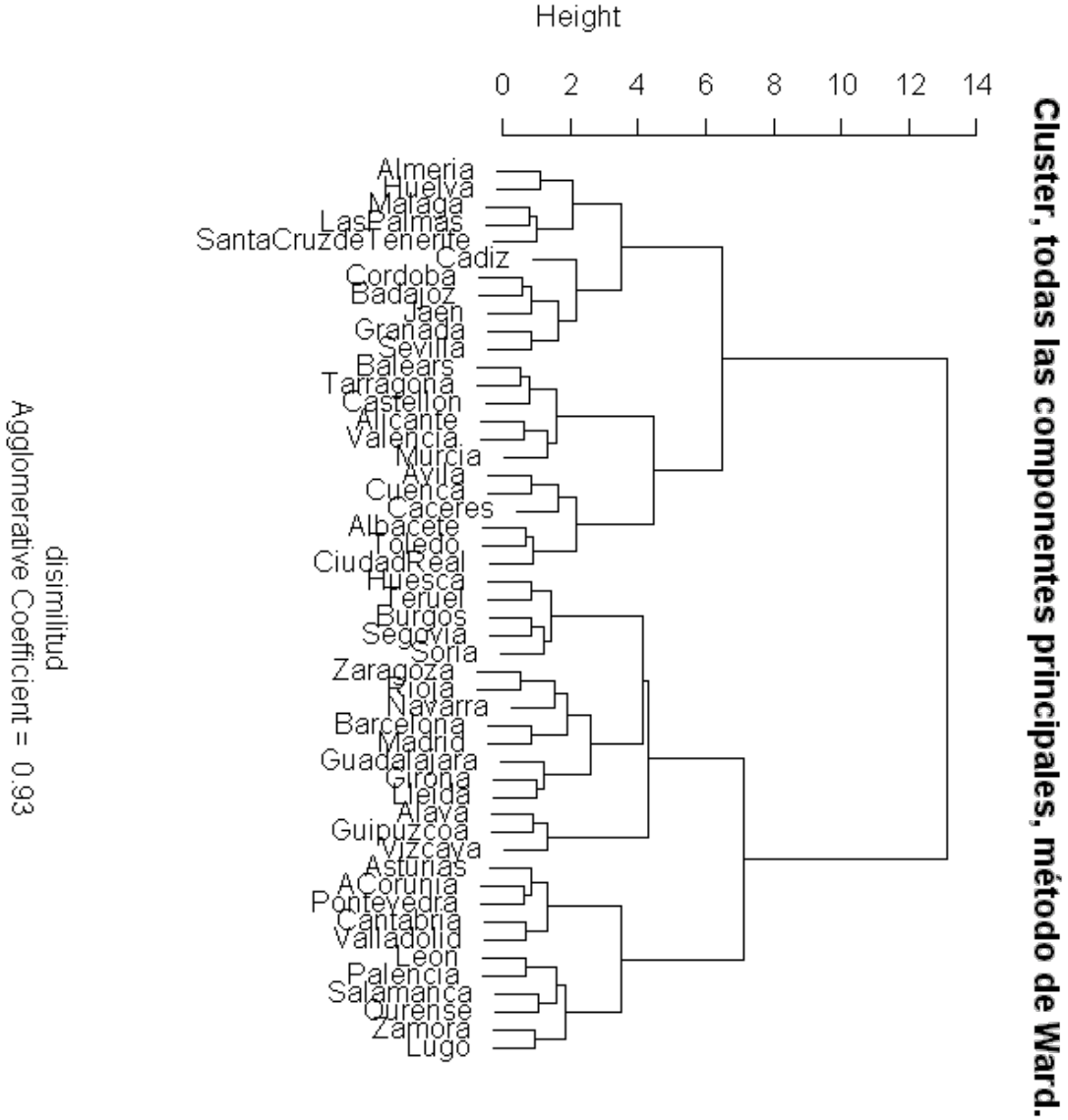

Figura A.1. *Agrupaciones de las provincias realizadas considerando todas las componentes principales y el método de Ward. En el eje y se muestra la distancia a la que se forman los distintos grupos. La figura ha sido girada.*

ANEXO II. Programa.

A continuación, se expone un *script* resumido de lo que se ha implementado para resolver el problema. Los comentarios aparecen después de '#'.

```
# Carga de datos.

datos3 <- read.csv("Resumen_turismo1.csv", header=TRUE, sep=";", dec=",",
stringsAsFactors=FALSE) # cuidado con el archivo resumen3.csv

b=datos3[-c(1,2,11,17,21,31,37,42,46,49,57,62,63),]

provincias_ultimo_2010=b[,c("Provincia","Provincia1","Provincia2","Matriculas","TDI
V11","SinEst10","PARO10","MOR06.10","KM10","PibC10","IFec10","Mat10","TDep
10","EM10","Tpresta10","PRenAdm10","TAgri11","TInd11","TCons11","TServ11","T
Aband11","TPrEmpleo11","EspVida65_10","EspVida10","Turismo10","InmExtr10")]

matriculas=4


# Cargamos las librerías.

library("psych")
library("GPArotation")
library(cluster)


# Hallamos las componentes principales.

bienestar.cp=princomp(provincias_ultimo_2010[,-c(1,2,3,4)],cor=T)
summary(bienestar.cp)
screeplot(bienestar.cp,type="lines",main="Varianza explicada")
bienestar.cp$loadings
biplot(bienestar.cp,xlabs=provincias_ultimo_2010[,matriculas],cex=0.8)


# Hallamos las componentes rotadas.

cor_matrix_1 <- cor(provincias_ultimo_2010[,-c(1,2,3,4)])
pca.5 <- principal(r = cor_matrix_1, nfactors = 5, residuals = FALSE, rotate =
"varimax")
pca.5$loadings


# Estandarizamos las variables y hallamos la matriz de disimilitudes.
```

```
num_comp <- 22
provincias <- provincias_ultimo_2010[,-c(1,2,3,4)]

for(i in 1:dim(provincias)[2]){
  media <- mean(provincias[,i])
  desv <- sd(provincias[,i])
  provincias[,i] <- (provincias[,i]-media)/desv
}

proyecciones <- matrix(NA,50,num_comp)
for(i in 1:dim(provincias)[1]){
  for(j in 1:num_comp){
    proyecciones[i,j] <- sum(provincias[i,]*bienestar.cp$loadings[,j])
  }
}

k <-
c(0.3340215,0.2231874,0.1339122,0.07152428,0.05073945,0.03340621,0.02819305,0.
02325967,0.01914003,0.01578455,0.01388933,0.01192802,0.00961104,0.008015421,0
.006249165,0.005249445,0.003730813,0.002855189,0.00245949,0.001645578,0.00090
33741,0.0002947893)
k <- k[1:num_comp]
k <- k/sum(k)

# métrica euclídea:
disimilitud <- matrix(NA,50,50)
for(i in 1:dim(provincias)[1]){
  for(j in 1:dim(provincias)[1]){
    disimilitud[i,j] <- sqrt(sum(k*(proyecciones[i,]-proyecciones[j,])^2))
  }
}


# Realizamos las agrupaciones con el método del entorno lejano (que en este programa
# se denomina “complete”) y el de Ward.

bienestar.cluster.ave.pesos =  agnes(disimilitud, diss = T,metric = "euclidian", stand =
T, method = "complete")
plot(bienestar.cluster.ave.pesos,labels=provincias_ultimo_2010[,2],main="Cluster,
todas las componentes principales, método complete.")

bienestar.cluster.ave.pesos =  agnes(disimilitud, diss = T,metric = "euclidean", stand =
T, method = "ward")
plot(bienestar.cluster.ave.pesos,labels=provincias_ultimo_2010[,2],main="Cluster,
todas las componentes principales, método de Ward.")


# Hallamos las proyecciones de los datos sobre nuestros factores.
```

```
proyecciones <- matrix(NA,50,5)
for(i in 1:dim(provincias)[1]){
  for(j in 1:dim(pca.5$loadings)[2]){
    proyecciones[i,j] <- sum(provincias[i,]*pca.5$loadings[,j])
  }
}


# Dibujamos en 2-D nuestros factores (1 vs 1).

# Económico vs urbanización.

plot(proyecciones[,1],proyecciones[,2],xlab="",ylab="",type="n")
text(proyecciones[,1],proyecciones[,2],labels=provincias_ultimo_2010[,matriculas])
title("Desarrollo económico - Factor urbanización", xlab="Desarrollo
económico",ylab="Factor urbanización")

# Económico vs demográfico.

plot(proyecciones[,1],-proyecciones[,5],xlab="",ylab="",type="n")
text(proyecciones[,1],-proyecciones[,5],labels=provincias_ultimo_2010[,matriculas])
title("Desarrollo económico - Desarrollo demográfico", xlab="Desarrollo
económico",ylab="Desarrollo demográfico")
```