III Congreso de Economía Regional de Castilla y León Segovia 1992, Comunicaciones 3: 1624-1637

AGRUPACION DE AREAS HOMOGENEAS EN CASTILLA Y LEON MEDIANTE EL ANALISIS CLUSTER

Luis César HERRERO PRIETO
Departamento de Economía Aplicada
Universidad de Valladolid

1.- EL ANALISIS CLUSTER

El nombre de *Análisis Cluster* se utiliza para definir una gran variedad de técnicas que tienen por objetivo la búsqueda de grupos en un conjunto de individuos. Junto a este nombre, también se utilizan otros para definir el mismo procedimiento, como, por ejemplo, análisis de conglomerados, taxonomía numérica, reconocimiento de patrones o, simplemente, creación de tipologías. A diferencia del resto de las técnicas multivariantes, el contenido propiamente estadístico del Análisis Cluster es mucho menor, limitándose con frecuencia a la especificación de distancias y a un posible tratamiento previo de los datos para su simplificación. Por otra parte, no suele ser su pretensión última la síntesis de la información, sino el establecimiento de agrupaciones entre las observaciones, en función de criterios de homogeneidad. No es sorprendente, pues, que constituya una de las técnicas de investigación actualmente más utilizadas en el análisis regional.

El esquema del procedimiento es el siguiente[1]. Si partimos de una matriz de datos de n individuos por p variables, típica del análisis multivariable, las dos preguntas que nos podemos hacer son las siguientes: ¿son las variables similares?, ¿son los individuos similares?. A partir de la matriz de correlaciones, obtenida de los datos (p x p), al análisis factorial da la solución a la primera pregunta, pues nos

[1]Exponemos tan sólo una idea intuitiva del análisis, que puede ampliarse en Martínez Ramos (1984) y Arnáiz (Ed.) (1987).

permite sintetizar la información en grupos de variables. Calculando una matriz de proximidades de las observaciones (n x n), el Análisis Cluster resuelve la segunda, pues nos proporciona una clasificación de los individuos en base a la información disponible. Básicamente podemos decir, entonces, que el Análisis Cluster consiste en poner límites o barreras a un conjunto de individuos (objetos). Si representamos a los individuos en un espacio euclídeo, donde los ejes son las variables, la posición de los individuos dependerá de los valores que tomen en las variables. El análisis tratará de agrupar a los individuos en función de su similaridad en todas las variables consideradas simultáneamente[2].

Para proceder a la clasificación de las observaciones, el análisis sigue dos etapas: medición de la similaridad entre las mismas y la clasificación propiamente dicha. A partir de la matriz de datos inicial, podemos valorar la similaridad de los individuos en función de los valores que en cada uno de ellos tomen las variables introducidas y midiendo la *distancia* entre los mismos; es decir, considerando a los individuos como vectores en un espacio p-dimensional, siendo p el número de variables[3]. El concepto de distancia más utilizado es el de distancia euclidiana[4], que es la suma de las diferencias al cuadrado de los valores de cada variable, para cada par de individuos. Según esto, dos individuos que tomen valores próximos para todo el conjunto de variables tendrán una distancia pequeña; es decir, son semejantes.

Una vez elegido el criterio de similaridad o distancia podemos construir la matriz de similaridad entre los sujetos: en cada casilla tendremos un número que es reflejo de la medida de semejanza entre ellos. Esta matriz, por ser simétrica, sólo utiliza la mitad de las posiciones y, además, la diagonal no tiene sentido ya que nos expresa la similaridad de cada sujeto consigo mismo.

---

[2]Existe una cierta confusión que se produce al decir que los individuos se asignan a uno u otro grupo, puesto que puede dar a entender que los objetos se mueven en el espacio. Los objetos permanecen estáticos y lo que se mueve son los límites de los conglomerados con el fin de dar cabida a los individuos que les pertenecen. Cf. Sánchez Carrión (1984).

[3]Existen también otros criterios de similaridad basados en coeficientes de correlación y en tablas de posesión y no posesión de atributos. Ver Martínez Ramos (1984).

[4]Otros conceptos de distancia, como la Mahalanobis, permite el tratamiento de variables correlacionadas.

Ahora deberemos estudiar la manera de formar los grupos de individuos. Para ello existen diversos procedimientos, pero vamos a centrarnos en los métodos jerárquicos ascendentes[5]. Estos proceden iterativamente de la forma siguiente: a partir de los *n* elementos básicos o conglomerados iniciales se agrupan los dos más próximos, con lo que el conjunto primitivo de *n* conglomerados se ha transformado en otro de *n*-1. De entre ellos se fusionan a su vez los dos más cercanos, dando lugar a un nuevo conjunto de *n*-2 agrupaciones. Procediendo así se llega, al cabo de *n*-1 etapas, a un núcleo único.

Evidentemente, la definición de distancia entre conglomerados va a tener una influencia decisiva sobre las agrupaciones resultantes. En este sentido las más utilizadas son las siguientes:

- La del vecino más próximo (*single linkage* o *nearest neighbour*). Aquí la distancia entre dos conglomerados C y C' se define como la más pequeña posible entre un elemento de C y otro de C'. En base a ella se agrupan en cada etapa los conglomerados que la minoran.

- La del vecino más lejano, en la que la distancia entre los conglomerados C y C' es la mayor posible entre los elementos de cada uno de ellos.

- La del centroide (*average linkage*). En este método la distancia que se computa es la que existe entre los centroides de los grupos; y se unirán aquellos grupos que tengan sus centroides más próximos[6].

- De la media del grupo, en la que la distancia se define como la media de las distancias de todos los pares formados con un elemento del primero y otro del segundo.

Un problema característico de los métodos de partición es el de la fijación del número de conglomerados finales a obtener, que puede hacerse *a priori* o puede

[5]Explicamos tan sólo el método empleado en esta investigación pero pueden verse algunos procedimientos más en Martínez Ramos (1984) y Arnáiz (Ed.) (1987).

[6]Es el método empleado en este trabajo.

determinarse en algún momento del proceso de agrupación. El primer caso es el más frecuente[7]. Cuando esto ocurre, lo que se hace es elegir en el espacio *p*-dimensional (siendo *p* el número de características de cada elemento) unos puntos *G*, en número *k* igual al de conglomerados buscados, y que actúen como núcleos o centroides de los mismos. Los elementos iniciales se van agregando a éstos en función de su proximidad y van modificando a su vez el centroide de cada grupo con cada nueva anexión. Más concretamente, en el procedimiento de computación que hemos utilizado en esta investigación[8], se eligen en una primera etapa los casos que tienen una mayor distancia entre ellos y se toman temporalmente como los centros de los cluster predeterminados. Posteriormente, se calculan los valores medios de las variables para los casos de cada conglomerado, y estos datos se usan para una nueva reclasificación. De esta forma se mejora el proceso de agrupación, y así sucesivamente.

2.- APLICACION DEL ANALISIS CLUSTER A LA REALIDAD ECONOMICA MUNICIPAL DE CASTILLA Y LEON

La aplicación de la metodología Cluster al entorno económico de Castilla y León exige definir previamente la matriz de datos inicial de la investigación, lo cual se concreta en justificar convenientemente la elección de los elementos objeto de estudio y las variables que van a caracterizarlos. En cuanto a los primeros, se han tomado como observaciones la totalidad de los municipios de Castilla y León que son, en número, 2.248[9]. La selección de las variables se ha efectuado teniendo en cuanta que deben reunir las peculiaridades socioeconómicas de cada unidad de análisis. Y, en este sentido, se han considerado las siguientes:

---

[7]También ha sido una condición impuesta por el método de computación empleado en este estudio, debido al desmesurado tamaño de la matriz de proximidades considerada: el cuadrado de 2.248 municipios existentes en Castilla y León.

[8]QUICK CLUSTER de SPSS/PC+ Advanced Statistics V2.0.

[9]Hay que señalar que algunos municipios presentaban valores ausentes en determinadas variables (sobre todo en la cifra de Presupuestos Municipales), pero la aplicación del Análisis Cluster se ha hecho de manera que el proceso de agrupación se mantenía para las variables con valores concretos, y ello ha permitido conservar la totalidad de los municipios.

CUADRO 1.- VARIABLESDE CARACTERIZACION

1.- MUPOBH86: Población de Hecho de 1986
2.- MUPOB89: Población de Hecho de 1989
3.- PARPRO: Participación de la población municipal en el total de la provincia (1981)
4.- DENSI89: Densidad de población en 1989
5.- CREC70: Crecimiento anual acumulativo de la población entre 1950 y 1970
6.- CREC89: Idem 1970-1989
7.- PORAGR: Porcentaje de población activa agraria
8.- PORIND: Porcentaje de población activa industrial
9.- PORSER: Porcentaje de población activa en servicios
10.- PARAGR: Participación de los activos agrarios municipales sobre el total provincial del sector
11.- PARIND: Idem industria
12.- PARSER: Idem servicios
13.- RRELAGR: Cociente entre activos agrarios/no agrarios
14.- PORAS: Porcentaje de activos asalariados
15.- PORNOAS: Porcentaje de activos no asalariados
16.- PORMIN: Porcentaje de población menor de 16 años
17.- PORMAX: Porcentaje de población mayor de 65 años
18.- TASPAR: Tasa de paro en 1986
19.- POREFAGR: Porcentaje de viviendas familiares ocupadas con usos agrarios
20.- POREFA2: Porcentaje de viviendas familiares ocupadas de dos o más viviendas
21.- INVTOT: Inversión industrial acumulada (1980-1988) en miles de pesetas constantes de 1987
22.- EMPTOT: Empleo industrial acumulado (1980-1988)
23.- LINOC: Líneas telefónicas ocupadas
24.- PRESUP: Presupuesto de ingresos municipales 1988

Como las variables están expresadas en diferentes unidades, resulta conveniente *normalizarlas* (llevarlas a una métrica común), conscientes de que tal procedimiento evita la incidencia de la unidad de medida en el análisis y, por tanto, tiende a diluir las diferencias entre los grupos. La forma más habitual de normalizar una variable es restarla su media y dividir todo por la desviación típica; de esta forma, las nuevas variables tendrán media cero y varianza uno. Estas son, pues, las variables consideradas en el Análisis Cluster, aun cuando posteriormente hayamos utilizado las variables sin estandarizar para caracterizar cada conglomerado.

El procedimiento específico aplicado a esta matriz de datos inicial, ha consistido en la construcción de la matriz de proximidades mediante la distancia euclídea, y, posteriormente, un proceso de agrupación jerárquico ascendente, según el criterio del centroide. No se ha impuesto ningún criterio de contigüidad de los municipios para que las áreas homogéneas se revelasen por sí mismas.

Se ha realizado, por otra parte, diversos ensayos de agrupación imponiendo *a priori* sucesivos números de conglomerados, hasta que se ha llegado a la clasificación considerada más conveniente. En realidad, todos los ensayos han

discernido siempre con claridad la dislocación entre municipios desarrollados y áreas atrasadas, característica de la realidad regional. A lo sumo se conseguían un número mayor de agrupaciones entre los municipios subdesarrollados recalcando algunas diferencias de matiz debidos a la mayor o menor talla demográfica. Se ha optado por obviar esos agrupamientos para determinar una clasificación más simple y sencilla.

El resultado definitivo se presenta en el Cuadro 2, en el que aparecen los ocho grupos que se han determinado al final, junto con el número de observaciones (municipios) correspondiente a cada conglomerado y las medias de determinadas variables para caracterizar a los mismos. La valoración del resultado es bastante aceptable pues, tal y como se ve en el análisis de la varianza reflejada en el Cuadro 3, la varianza entre los cluster (*between-cluster*) es muy alta, mientras que la varianza para las observaciones de cada cluster (*within-cluster*) es pequeña. El ratio entre ambas varianzas es grande, y esto quiere decir, entonces, que los grupos resultantes distan mucho entre sí, pero los individuos que agrupan son bastante homogéneos.

La caracterización de los conglomerados no presenta demasiadas dificultades, pues, en realidad, refleja la dislocación característica del sistema urbano castellano-leonés entre las capitales de provincia y cabeceras de comarca, por un lado; y, por otro, el resto de los municipios con un tamaño reducido. De esta forma tenemos la jerarquía siguiente:

- Cluster 1.- Valladolid

- Cluster 2.- Burgos

- Cluster 3.- León

- Cluster 4.- Salamanca

- Cluster 5.- Avila, Palencia, Segovia, Soria, Zamora

- Cluster 6.- Aranda de Duero, Béjar, Miranda de Ebro, Villamuriel de Cerrato

CUADRO 2.- ANÁLISIS CLUSTER Y CARACTERIZACIÓN DE CONGLOMERADOS

	CLUSTER 1	CLUSTER 2	CLUSTER 3	CLUSTER 4	CLUSTER 5	CLUSTER 6	CLUSTER 7	CLUSTER 8
CASOS	1	1	1	1	5	4	394	1.840
%S/REGIÓN	0,04%	0,04%	0,04%	0,04%	0,22%	0,18%	17,53%	81,85%
MUPOB89	333.230	161.538	137.261	160.522	54.661	21.689	2.038	358
DENSI89	1.982,00	1.495,72	3.431,53	4.115,95	369,35	270,13	40,67	11,54
CREC70	0,03	0,02	0,03	0,02	0,02	0,01	-0,01	-0,02
CREC89	0,02	0,02	0,01	0,01	0,02	0,02	-0,01	-0,02
PORMIN	30,31	28,06	26,20	26,67	26,97	27,20	22,43	15,92
PORMAX	8,55	9,54	10,94	11,28	11,38	9,49	15,97	22,30
PORAGR	1,80	1,90	1,60	2,60	3,40	7,14	27,88	69,50
PORIND	38,00	34,40	15,50	15,60	18,52	42,26	20,85	6,16
PORSER	52,00	54,10	73,30	71,00	66,58	37,64	33,62	17,07
PORAS	85,40	84,60	79,70	79,30	81,36	81,74	65,01	29,73
PORNOAS	11,50	12,10	16,70	16,90	14,92	15,36	28,84	54,98
INVTOT	90.496.555	40.879.441	15.558.072	8.673.846	6.292.479	30.031.330	248.637	17.311
EMPTOT	312.111,00	11.084,00	5.362,00	4.430,00	2.731,00	4.662,00	91,70	5,84

Fuente: Elaboración propia

CUADRO 3.- ANALISIS DE LA VARIANZADE LOS CONGLOMERADOS

Variable	Varianza "entre"	Varianza "intra"	Ratio E/I	Prob
ZPARPRO	267.9454	.0462	5799.9441	.000
ZLINOC	270.7396	.0310	8727.6984	.000
ZPOB86	270.7600	.0361	7491.7487	.000
ZPARIND	259.3363	.0724	3581.9459	.000
ZPARSER	271.3737	.0292	9300.3252	.000
ZPARAGR	95.0966	.6621	143.6222	.000
ZDENSI	264.1485	.0555	4755.5031	.000
ZPORAGR	111.0284	.6049	183.5407	.000
ZPORIND	71.5362	.7467	95.7996	.000
ZPORSER	49.8851	.8245	60.5056	.000
ZPORAS	90.1150	.6800	132.5185	.000
ZPORNOAS	61.6837	.7821	78.8688	.000
ZRRELAGR	12.4397	.9591	12.9698	.000
ZPORMAX	15.0712	.9495	15.8732	.000
ZPORMIN	25.7885	.9110	28.3082	.000
ZCREC70	64.2757	.7728	83.1726	.000
ZCREC89	54.0785	.8092	66.8262	.000
ZTASPAR	48.7655	.8285	58.8607	.000
ZPOREFAG	37.0835	.8704	42.6063	.000
ZPOREFA2	106.4676	.6211	171.4092	.000
ZINVTOT	267.7436	.0469	5706.5110	.000
ZEMPTOT	268.5083	.0442	6076.6940	.000
ZPRESUP	199.4949	.0358	5565.0717	.000

- Cluster 7.- Se incluyen aquí los municipios de la región con un tamaño apreciable (2.038 vecinos en media) y una densidad de habitantes por km^2 (40.6) superior a la media regional. La estructura de la población activa está más o menos equilibrada entre los sectores productivos, lo cual, en un entorno fundamentalmente agrario, hace resaltar su característica industrial y de servicios. Esto viene confirmado en el hecho de que el 65 % de la población activa sea asalariada.

- Cluster 8.- Aparecen aquí, la gran extensión de municipios agrarios y en atonía demográfica de la región; son el 81.5 % del total de entidades de población, y se caracterizan por tener un tamaño muy reducido (357,5 habitantes en media), una densidad de población inferior a la media, en pleno proceso de vaciamiento demográfico (descensos del 2 % anual acumulativo), una población envejecida (el 22 % tiene más de 65 años) y una población activa absolutamente agraria (el 69 %); de ahí que la tasa de paro sea la más pequeña y generan, tan sólo 17 millones de pesetas de inversión industrial durante los años ochenta.

Con el fin de simplificar la ordenación, podemos considerar que se han obtenido tres niveles jerárquicos: (1) las capitales de provincia y grandes cabeceras de comarcales; (2) municipios desarrollados y de tamaño intermedio; y, (3) gran extensión de municipios periféricos en plena atonía económica y demográfica. Con esta clasificación se ha procedido a la representación cartográfica de todos los municipios en los mapas adjuntos para poder observar la localización geográfica de los mismos.

3.- CONCLUSIONES

El Análisis Cluster resulta ser una técnica estadística útil para resumir la información de un número muy grande de observaciones, mediante la clasificación en grupos homogéneos. Con este fin se ha aplicado a la realidad económica municipal de Castilla y León, proporcionando una descripción puramente estática de la jerarquía regional de núcleos.

Las características que se deducen del análisis son, en primer lugar, la atonía

generalizada, tanto económica como demográfica, de los municipios castellano-leoneses; fenómeno del que sólo se salvan las capitales de provincia y determinadas cabeceras de comarca con un tamaño apreciable. Resalta, por otra parte, la ausencia de núcleos intermedios, con lo que el sistema urbano aparece desarticulado. En este sentido, los municipios incluidos en el Cluster 7, resultado de esta investigación, pueden servir como sistema de referencia para una política de ordenación del territorio que pretenda racionalizar la articulación regional y su sistema de asentamientos.

4.- BIBLIOGRAFIA

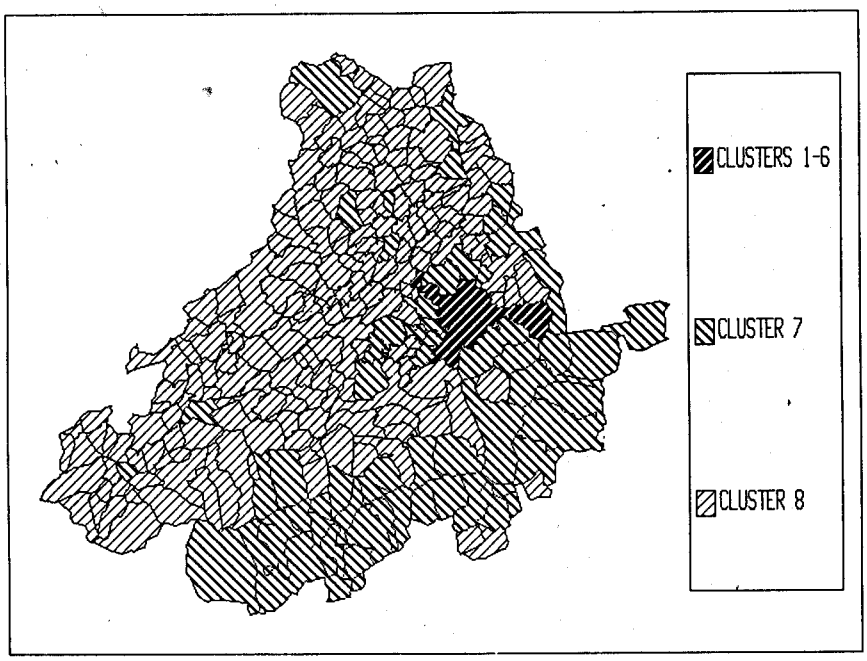
ARNAIZ, G.; MARTIN GUZMAN, M.P.; MARTIN, T. y TOLEDO, I. (1987) *Discriminación y Clasificación de las Regiones Fiscales en España*, Instituto de Estudios Fiscales, Madrid.

HERRERO PRIETO, L.C. (1992) *Criterios Multivariantes para el Estudio del Desarrollo Económico y la Organización del Espacio en Castilla y León*, MIMEO, Departamento de Economía Aplicada, Universidad de Valladolid, Valladolid.

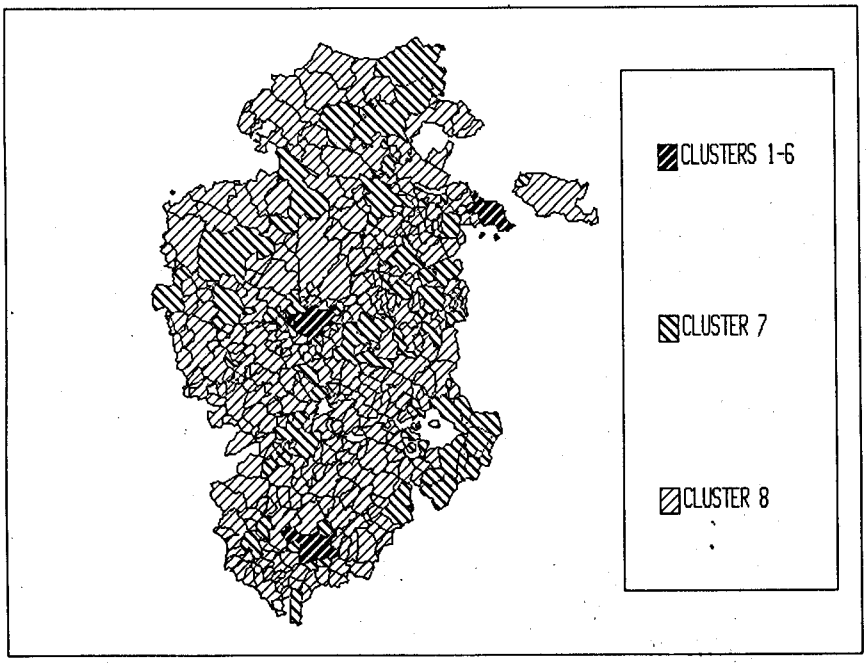
MARTINEZ RAMOS, E. (1984) "Aspectos teóricos del análisis cluster y aplicación a la caracterización del electorado potencial de un partido", en SANCHEZ CARRION, J.J. (1984).

SANCHEZ CARRION, JJ. (ED.) (1984) *Introducción a las Técnicas de Análisis Multivariante aplicadas a las Ciencias Sociales*, Centro de Investigaciones Sociológicas, Madrid.

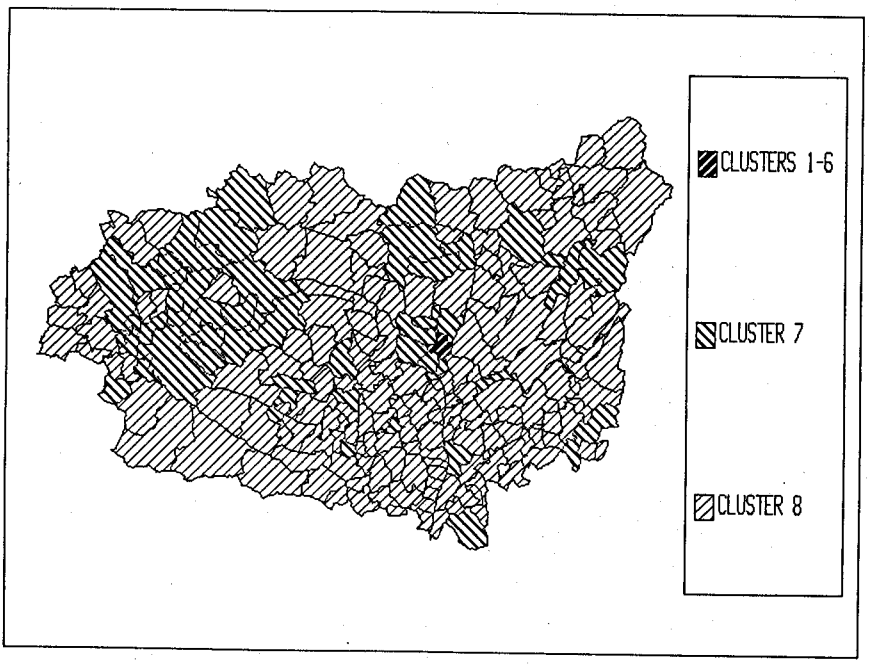
AVILA



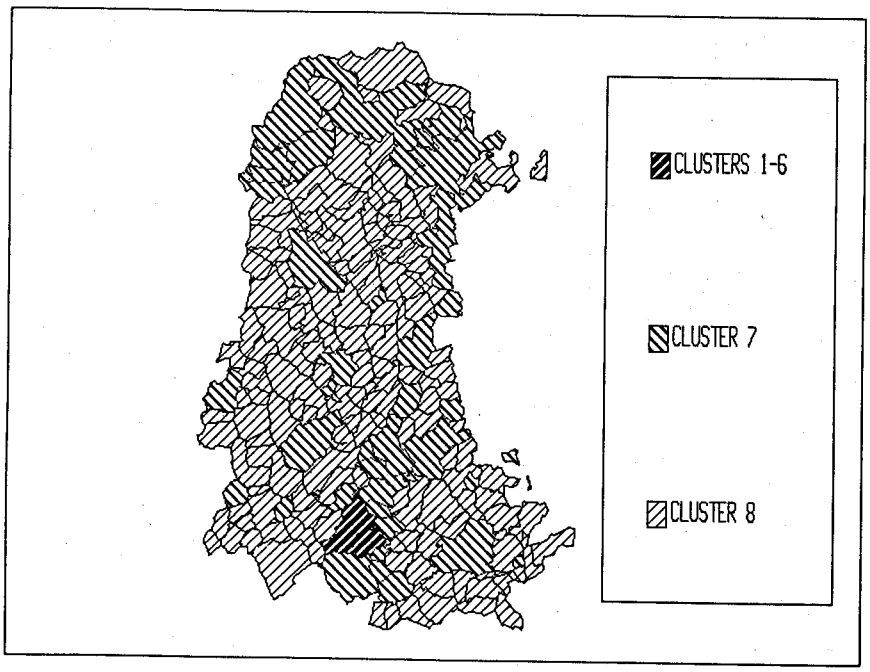
BURGOS



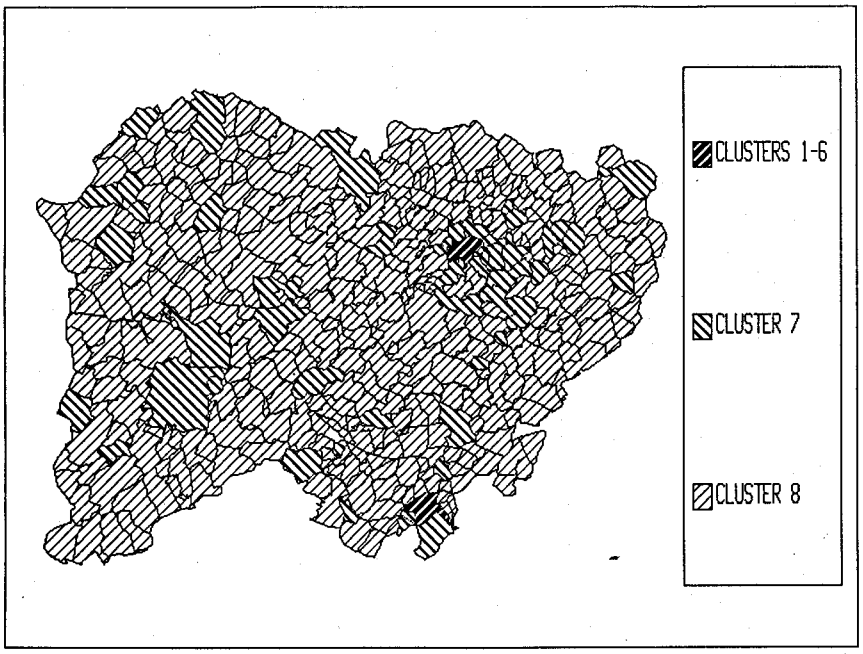
LEON



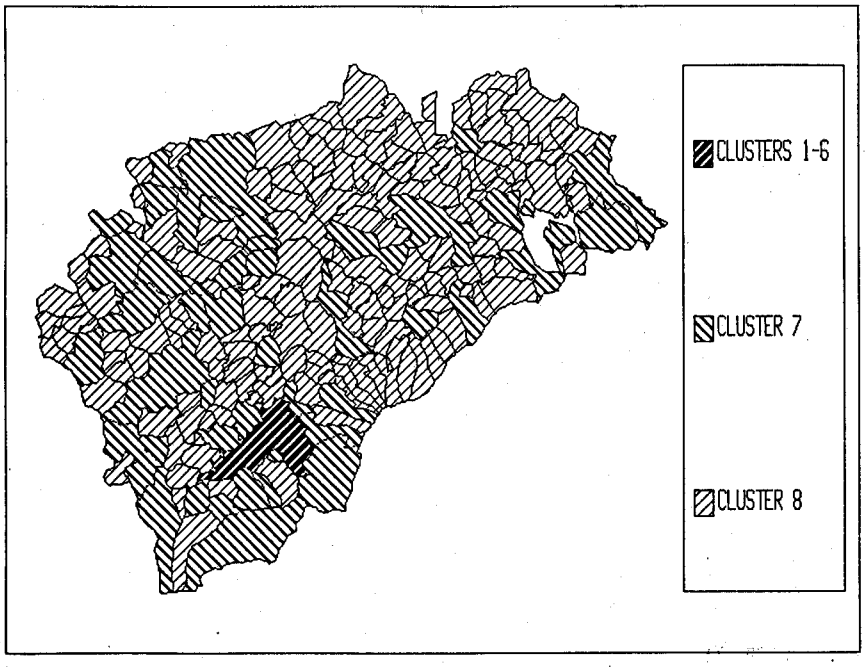
PALENCIA



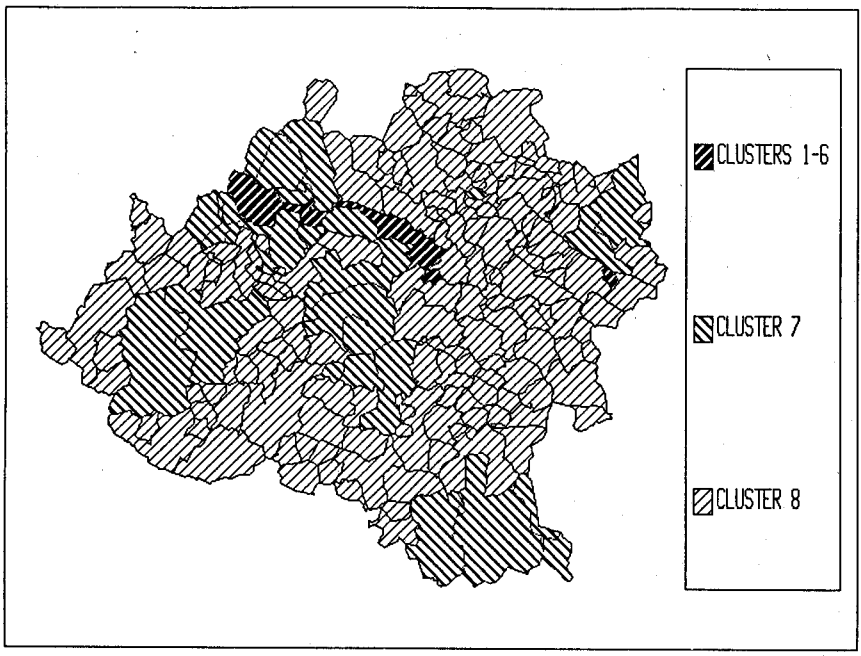
SALAMANCA



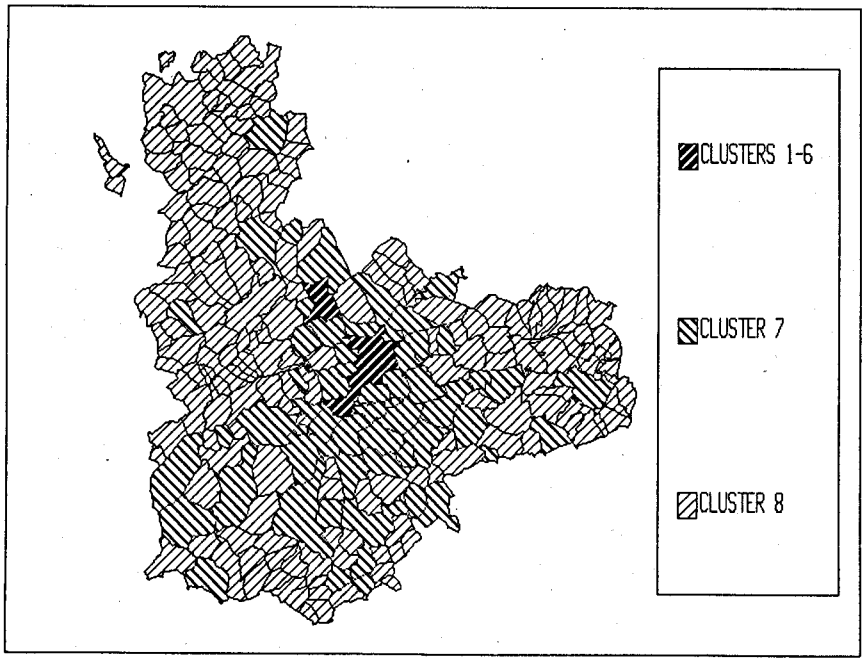
SEGOVIA

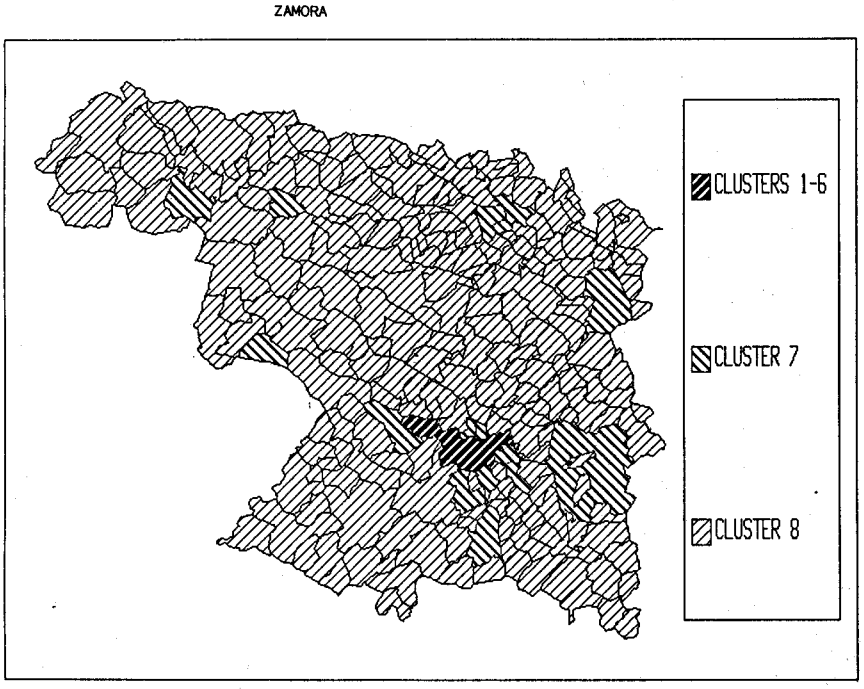


SORIA



VALLADOLID




ZAMORA
CLUSTERS 1-6
CLUSTER 7
CLUSTER 8