QÜESTIIÓ, Vol 20, 3 pp. 449-482, 1996

LA COMBINACIÓ DE TÈCNIQUES DE GEOMETRIA DIFERENCIAL AMB ANÀLISI MULTIVARIANT CLÀSSICA: UNA APLICACIÓ A LA CARACTERITZACIÓ DE LES COMARQUES CATALANES

SERGI VIVES*

ÀNGEL VILLARROYA*

Universitat de Barcelona

En aquest treball s'estudien les característiques i les relacions de les 41 comarques de Catalunya basant-se en diverses variables socio-econòmiques. La metodologia utilitzada combina tècniques clàssiques d'anàlisi de dades amb un nou mètode per a la representació simultània de poblacions i variables, basat en tècniques de geometria diferencial, anomenat IDA (Intrinsic Data Analysis). Els resultats obtinguts han permès classificar les comarques de Catalunya en 9 grups que es poden reagrupar en 4 grans blocs (agrícola, turístic, administratiu i industrial), que presenten una clara gradació geogràfica.

Combining geometric-differential techniques with classic multivariate analysis: an application to characterize catalan regions

Keywords: Intrinsic Data Analysis; distància de Rao; Anàlisi de Correspondències; Anàlisi de conglomerats.

Classificació AMS: 62H25, 62H30.

*S. Vives i À. Villarroya. Dept. d'Estadística. Universitat de Barcelona. Av. Diagonal, 645. 08028 Barcelona. Espanya.

- Article rebut el gener de 1995.
- Acceptat el setembre de 1996.

1. INTRODUCCIÓ

Des del treball pioner de Rao (1945), molts autors han desenvolupat una branca de l'estadística basada en aplicacions de la geometria diferencial (Efron, 1975; Atkinson i Mitchell, 1981; Burbea i Rao, 1982, 1984). Malauradament, molts d'aquests treballs són molt teòrics i s'aprecia una mancança d'aplicacions pràctiques. En aquest treball es mostra un exemple de com algunes d'aquestes tècniques, per si mateixes o combinades amb altres de *clàssiques*, constitueixen unes eines molt potents de fàcil aplicació i interpretació.

Els objectes estadístics d'aquest treball els formen les *comarques* de Catalunya. La divisió territorial de Catalunya en comarques va ser proposada per primera vegada el 1936 per la Generalitat de Catalunya, encara que aquesta divisió no va gaudir de poders administratius fins als anys 80. Els criteris originals en què es va basar aquesta divisió van ser fonamentalment comercials (proximitat al mercat més proper), sense deixar de banda aspectes geogràfics i històrics, importants, però de segon ordre. No cal dir que l'origen d'aquestes comarques va comportar problemes, ja que algunes entitats geogràfiques van quedar fora de la divisió original dividides entre les comarques oficials, com és el cas del Lluçanès i el Moianès, que es coneixen com a subcomarques. Amb la nova ampliació a les 41 comarques actuals, l'any 1988, s'han solucionat alguns d'aquests problemes.

En aquest treball ens plantegem la pregunta de quines són les característiques que defineixen les comarques de Catalunya i les seves relacions. Com que des d'un punt de vista antropològic un dels factors que més influeixen en el caràcter d'una població és l'activitat a què es dediquen els seus habitants, serà aquesta variable la que farem servir de base en el nostre estudi.

2. METODOLOGIA UTILITZADA

Com ja s'ha dit, les poblacions objecte del nostre estudi són les 41 comarques en què actualment es troba dividida Catalunya, i la variable principal del nostre estudi és la X: *Població activa per grups professionals* segons el cens de 1991 (dades facilitades per l'Institut d'Estadística de Catalunya), dividida en vuit categories:

X_1: Professionals i tècnics	X_2: Personal directiu
X_3: Serveis administratius	X_4: Comerciants i venedors
X_5: Hoteleria i altres	X_6: Agricultura i pesca
X_7: Indústria	X_8: Forces armades

L'elecció d'aquesta variable es fonamenta en el fet que quan es vol determinar el caràcter o, fins i tot, el grau de desenvolupament d'una població des d'un punt de vista antropològic, un primer factor a estudiar és com es reparteixen el recursos humans entre les diferents tasques que cal realitzar.

Una primera aproximació per determinar les semblances i les diferències entre les comarques podria ser la seva representació gràfica en un espai de dimensió reduïda segons el vector X, mitjançant el mètode denominat IDA, Instrinsic Data Analysis (Ríos *et al.*, 1994), com s'explica a la secció següent, però del qual volem destacar que, a diferència de la majoria de mètodes de reducció de la dimensió, es tracta d'una metodologia no específica d'un model estadístic, sinó que es pot aplicar a qualsevol amb moments de segon ordre finits.

L'IDA ens permet visualitzar les analogies entre les diferents comarques, però si el que volem és realitzar-ne una classificació en una sèrie de grups cal utilitzar alguna de les tècniques conegudes amb el nom d'anàlisi de conglomerats (*cluster analysis*). Sota aquest nom s'engloben un conjunt de tècniques que tenen per finalitat, a partir d'un conjunt inicial, obtenir una classificació dels objectes que el formen en un conjunt de grups que, dintre de cadascun, tinguin la màxima homogeneïtat possible, mentre que entre si l'heterogeneïtat sigui màxima respecte a les variables estudiades. En el nostre cas vam triar una tècnica aglomerativa jeràrquica denominada *average linkage clustering*, coneguda originalment com a UPGMA (Sokal i Sneath, 1963), que és una de les més utilitzades en la pràctica. L'UPGMA parteix d'una matriu d'interdistàncies entre els objectes a classificar i dóna com a resultat una classificació jeràrquica en forma de dendrograma. En el nostre cas, i d'acord amb la natura de les dades, vàrem triar la distància de *Bhattacharyya*, que presenta l'avantatge addicional de coincidir, excepte constants, amb la distància que utilitza l'IDA per al model multinomial.

Així doncs, el següent pas va consistir a calcular la matriu D d'interdistàncies entre totes les comarques. De manera que l'element d_{ij} de la matriu D representa la distància de Bhattacharyya entre la i-èsima i la j-èsima comarca calculada segons l'expressió:

$$(1) \qquad d_{ij} = d(p^{(i)}, p^{(j)}) = \operatorname{acos}\left(\sum_{k=1}^{8} \sqrt{p_{ik} p_{jk}}\right) \qquad i, j = 1, \ldots, 41$$

on p_{ik} representa la probabilitat que un individu qualsevol de la i-èsima comarca pertany al k-èsim grup professional ($k = 1, \ldots, 8$). Sobre la matriu D es va aplicar l'UPGMA per tal d'obtenir una classificació jeràrquica de les comarques.

Finalment, segons el dendrograma obtingut per l'UPGMA i la representació gràfica mitjançant l'IDA es van agrupar les 41 comarques en una sèrie de *clusters*.

L'últim pas d'aquest treball va consistir en l'estudi de les característiques de la classificació obtinguda i les relacions entre els diferents *clusters* obtinguts, tant per la

variable original X com per altres variables d'interès social (procedents del cens del 1991 o de l'any més proper disponible): nivell d'estudis, saldo migratori, densitat, renda per càpita, distribució per edats de la població activa, etc. A l'Apèndix es troben les dades originals, algunes de les quals procedeixen de l'Anuari Estadístic de Catalunya de 1992, i d'altres, encara no publicades en el moment d'elaborar aquest treball, ens han estat facilitades per l'Institut d'Estadística de Catalunya.

3. REPRESENTACIÓ MITJANÇANT L'IDA

3.1. Generalitats

Qualsevol mètode de representació gràfica pressuposa una determinada mètrica. Així, per exemple, l'Anàlisi de Components Principals (ACP) es basa en la distància euclidiana i l'Anàlisi de Correspondències (AC) en la distància khi-quadrat. L'elecció de la mètrica és arbitraria i dependrà de les propietats que es considerin més importants. L'IDA es basa en una mètrica riemanniana segons un enfocament que sintetitzem a continuació.

Donat un model paramètric amb moments de segon ordre finits, podem dotar-lo amb una estructura de varietat riemanniana, V, on el tensor mètric ve donat per la matriu d'informació de Fisher. Els punts d'aquesta varietat (V) representen poblacions estadístiques (mesures de probabilitat) que tenen com a coordenades els valors dels seus paràmetres (per exemple en el cas del model normal univariant, $N(\mu,\sigma)$ $\mu \in \mathbb{R}$, $\sigma > 0$, les coordenades de les poblacions a la varietat corresponent serien els parells (μ,σ)) i on la mesura de dissimilaritat entre poblacions ve donada per la distància riemanniana corresponent (coneguda en aquest cas com a distància de Rao). Aquesta aproximació permet utilitzar les eines de la geometria diferencial amb les seves propietats d'invariància versus canvis de coordenades (és a dir, versus transformacions admissibles dels paràmetres i de les variables). Una descripció molt més detallada d'aquest enfocament pot trobar-se a Amari (1985), Atkinson i Mitchell (1981), Bandorff-Nielsen (1984) Burbea i Rao (1982) i Oller (1989) entre d'altres.

Donades r poblacions pertanyents a un model probabilístic amb n paràmetres, l'IDA considera els passos següents:

- Representació de les poblacions $p^{(1)},\ldots,p^{(r)}$ a la varietat riemanniana V.
- Determinació de q, l'origen de la representació final, que es recomana que sigui el baricentre (centre de masses riemannià) de les poblacions.
- Mapatge de les poblacions des de V fins a V_q, l'espai tangent a la varietat al punt q, amb la mínima distorsió possible. Aquest pas es justifica pel fet que

V_q és un espai euclidià i, per tant, la posterior representació a un espai de dimensió reduïda mitjançant les tècniques clàssiques, com l'ACP, ja és factible. El mapatge des de V fins a V_q es realitza mitjançant la denominada *inversa de la funció exponencial al punt q* ($\exp_q^-$) que presenta la propietat de conservar les interdistàncies entre q i les poblacions.

- Finalment apliquem l'ACP als punts $m^{(1)},\ldots,m^{(r)}$ que representen les poblacions a l'espai V_q, tenint en compte que la matriu de la mètrica a V_q ve donada per G_q, la matriu d'informació de Fisher a q. Per tant, l'ACP es resumeix en sol·lucionar la següent diagonalització:

$$Cu_i = \lambda_i G_q^- u_i \tag{2}$$

on C representa la matriu de covariàncies (o qualsevol múltiple d'aquesta matriu) de les coordenades de les poblacions a V_q i u_i és el vector propi (ortonormal) associat al i-èsim valor propi λ_i. De manera que la projecció de V_q a $\mathbb{R}^s$ $(s \leq n)$ es realitza mitjançant:

$$X = MU \tag{3}$$

on M és la matriu $r \times n$ que té a les seves files les coordenades de les poblacions a V_q $(m^{(i)})$, U és la matriu $n \times s$ que té a les seves columnes els s primers vectors propis i X és la matriu $r \times s$ que té a les seves files les coordenades de les poblacions a $\mathbb{R}^s$ $(x^{(i)})$.

Esquema de la representació de les poblacions mitjançant l'IDA.

Varietat V	$\xrightarrow{\exp_q^-}$	Espai tangent V_q	$\xrightarrow{ACP}$	$\mathbb{R}^s$
$p^{(1)},\ldots,p^{(r)}$		$m^{(1)},\ldots,m^{(r)}$		$x^{(1)},\ldots,x^{(r)}$

Les variables es representen per les corbes de màxima variació de les seves esperances. Així, donada la variable X (amb esperança finita), podem definir el camp vectorial $\mathrm{grad}(E(x))$ (el gradient de l'esperança d'X) i associat a aquest tenim el flux integral tal que les seves corbes indiquen, localment, les direccions de màxima variació de l'esperança d'X. Amb això tenim les variables representades com a corbes a la varietat V i ja només cal repetir els passos indicats per a les poblacions: transport a V_q mitjançant $\exp_q^-$ i finalment projecció a $\mathbb{R}^s$.

Una descripció molt més acurada de l'IDA (incloent la representació de regions confidencials, efecte dels individus sobre l'estimació, etc.) pot trobar-se a Ríos *et al.* (1994).

3.2. Aplicació de l'IDA al nostre exemple

Tenim $r = 41$ poblacions (comarques) i considerem el model de Bernoulli multivariant (multinomial de mida $N = 1$) amb funció de densitat:

$$f_i(x_1,\ldots,x_8) = p_{i1}^{x_1} \cdot \ldots \cdot p_{i8}^{x_8} \qquad i = 1,\ldots,41 \tag{4}$$

$$x_j = \{0,1\} \qquad p_{ij} \in [0,1] \quad \text{i} \quad \sum_{j=1}^{8} p_{ij} = 1$$

on x_j $(j = 1,\ldots,8)$ són les variables indicadores de cada classe (grup professional) i p_{ij} és la probabilitat que un individu qualsevol de la i-èsima població pertany a la j-èsima classe. Les p_{ij} són, per tant, els paràmetres del nostre model. La geometria d'aquest model és ben coneguda i pot trobar-se a Atkison i Mitchell (1981), Oller (1982).

a) Representació de les poblacions

Pas a1: Representem les 41 poblacions (comarques) $p^{(1)},\ldots,p^{(41)}$ a la varietat V. Les coordenades estan donades pels paràmetres del model multinomial (les probabilitats de cada classe), és dir, les proporcions (p_{ij}) de cada grup professional a la població corresponent:

$$p^{(i)} = (p_{i1},\ldots,p_{i8}) \qquad i = 1,\ldots,41 \tag{5}$$

Pas a2: Determinem q, el baricentre de les poblacions, és a dir, el punt q de la varietat $(q \in V)$ que minimitza la suma de distàncies al quadrat a les poblacions:

$$\sum_{k=1}^{41} \left(d\left(p^{(i)},q\right)\right)^2 \leq \sum_{k=1}^{41} \left(d\left(p^{(i)},v\right)\right)^2 \qquad \forall v \in V \tag{6}$$

on $d\left(p^{(i)},q\right)$ representa la distància riemanniana (distància de Rao) entre $p^{(i)}$ i q, que en el model multinomial té la forma:

$$d\left(p^{(i)},q\right) = 2 \operatorname{acos}\left(\sum_{k=1}^{8} \sqrt{p_{ik}q_k}\right) \tag{7}$$

Les coordenades de q es resolen (6) numèricament i en el nostre cas s'obtingué:

$$q = (0.104, 0.020, 0.113, 0.118, 0.098, 0.107, 0.437, 0.003) \tag{8}$$

Pas a3: Transportem els punts $p^{(i)}$ a l'espai tangent a la varietat en el punt q, és dir a V_q, mitjançant la denominada *inversa de la funció exponencial* al punt q

($\exp_q^-(\cdot)$). En aquest model, si tenim el punt $p^{(i)}$ de la varietat, les seves coordenades a l'espai tangent ($m^{(i)} = (m_{i1}, \ldots, m_{i7}) \in V_q$) es calculen com:

$$m_{i\alpha} = \rho(i) \frac{\sqrt{p_{i\alpha} q_\alpha} - q_\alpha \cos(\rho(i)/2)}{\sin(\rho(i)/2)} \qquad \alpha = 1, \ldots, 7 \tag{9}$$

on $\rho(i)$ simbolitza la distància entre $p^{(i)}$ i q, és a dir $\rho(i) = d(p^{(i)}, q)$. Cal fer notar que la dimensió de V_q és 7, això és una conseqüència que la vuitena coordenada a V és redundant, ja que la suma de totes elles ha de ser 1.

Pas a4: Projecció de les poblacions des de V_q fins a l'espai de dimensió reduïda, en el nostre cas $\mathbb{R}^2$. Primer calculem la matriu d'informació de Fisher al punt q (G_q), que en aquest model té la forma:

$$G_q = g_{ij} = \left(\frac{\delta_{ij}}{q_i} - \frac{1}{q_8} \right) \qquad i, j = 1, \ldots, 7 \tag{10}$$

Definim com a M la matriu 41×7 que té a les files les coordenades de les poblacions i calculem C la matriu de covariàncies de les poblacions a V_q. Efectuem la diagonalització: $Cu_i = \lambda_i G_q^- u_i$ tal com es va indicar a (2). Com que estem interessats en una representació bidimensional, construïm la matriu U amb els dos primers vectors propis (u_1, u_2) i obtenim:

(11)

$$U' = \begin{pmatrix} 0.767 & 1.171 & 1.181 & 0.940 & 0.941 & -2.375 & 0.605 \\ 2.948 & 2.645 & 3.301 & 3.265 & 2.062 & 3.392 & 4.844 \end{pmatrix}$$

Finalment, les coordenades de la i-èsima població a $\mathbb{R}^2$ ($x^{(i)}$) es correspondrà amb la i-èsima fila de la matriu X calculada com: $X = MU$.

Si prenem com a exemple la comarca del Barcelonès ($p^{(13)}$), hem seguit els passos següents:

$$p^{(13)} = (0.171, 0.029, 0.214, 0.148, 0.112, 0.004, 0.320, 0,001)$$

$$m^{(13)} = \exp_q^-(p^{(13)}) = (0.071, 0.011, 0.098, 0.042, 0.024, -0.164, -0.080)$$

$$x^{(13)} = m^{(13)} U = (0.587, -0.196)$$

b) Representació de les variables

Pas b1: La corba integral, associada amb la variable X_i $i = 1, \ldots, i = 1, \ldots, 8$ que s'origina a q ve donada per:

$$\pi_i^{\alpha}(t) = \frac{q_\alpha(1+\delta_{i\alpha}(e^t-1))}{1+q_i(e^t-1)} \qquad \alpha = 1,\ldots,7 \tag{12}$$

on δ_{ij} representen les deltes de Kronecker. Per tant, pot transportar-se a l'espai tangent V_q mitjançant exp_q^-, repetint l'explicat per a les poblacions.

4. RESULTATS

La representació gràfica obtinguda mitjançant l'IDA es mostra a la figura 1, en què s'ha pres com a origen el baricentre de les poblacions, que es podria interpretar com la *comarca mitjana*. De la figura 1 es desprèn que el factor més important en la diferenciació de les comarques és l'agrícola (X_6), ja que la fletxa corresponent a aquesta variable és pràcticament paral·lela al primer component principal (que explica el 72.03 % de la variabilitat total) i és la de major longitud. Les comarques situades a l'esquerra són les que, proporcionalment, més recursos humans dediquen a les tasques agràries i de pesca. D'altra banda, la fletxa que correspon al component industrial (X_7) és pràcticament paral·lela al segon component principal (que explica el 17.20 % de la variabilitat total) i la segona de major longitud, indicant-nos que aquesta variable és la segona en importància en la discriminació entre les comarques segons els seus recursos humans. Les comarques representades en la part superior del gràfic són les que tenen una proporció més gran d'habitants dedicats a tasques industrials. El component *forces armades* (X_8) creix en la mateixa direcció, però en sentit oposat al component industrial, encara que qualsevol possible interpretació d'aquest fet seria qüestionable per l'escassa importància relativa d'aquest component. La resta de variables ($X_1,\ldots,X_5$), que corresponen majoritàriament a les activitats de tipus *serveis*, donen una informació aparentment redundant, ja que totes creixen segons una direcció i sentit molt semblant (inferior dreta de la fig. 1), sent la variable X_5 la de major longitud d'aquest grup de variables.

L'excel·lent relació entre les dades originals i la representació gràfica a través de l'IDA (fig. 1) es demostra pel fet que aquesta explica pràcticament el 90 % de la variabilitat original i pel coeficient de correlació entre les 820 interdistàncies originals de les 41 comarques i les interdistàncies de la figura 1, que té un valor de 0.9888.

La figura 1 ja ens permet visualitzar la formació de grups, però per establir una agrupació més objectiva vàrem utilitzar l'UPGMA. El dendrograma resultant es mostra a la figura 2. Es va tallar a un nivell de $d = 0.294$, ja que d'aquesta manera vàrem obtenir un nombre de *clusters* o *grups* raonable, nou, que s'han identificat a la fig. 1, mostrant la bona concordança entre els resultats de l'UPGMA i de l'IDA. A la taula 1 es detalla la classificació obtinguda i la numeració dels nou *clusters*; a la figura 3 es representen les mitjanes de la variable X per a cada *grup*.

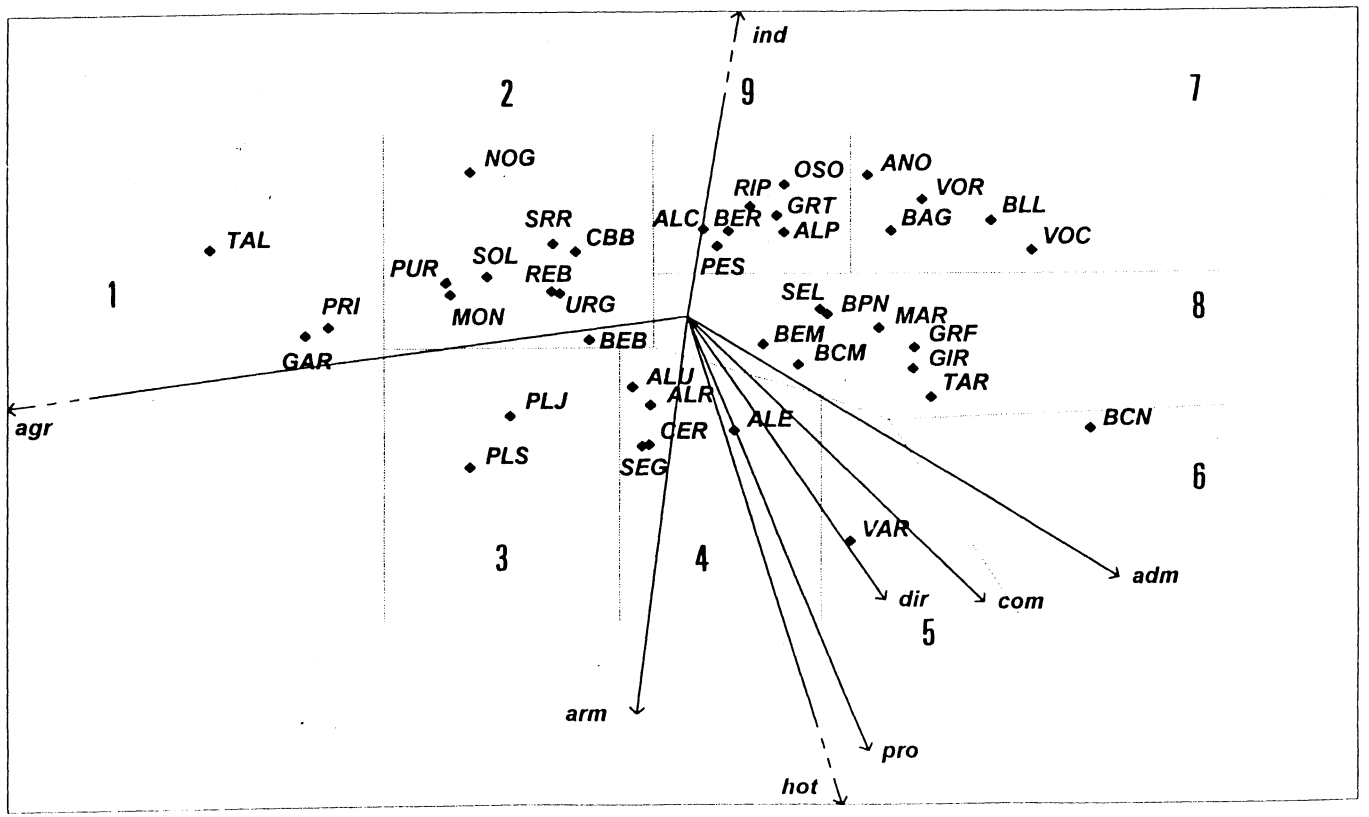


Figura 1

Representació bidimensional de les 41 comarques de Catalunya i de les 8 variables (grups professionals) obtinguda mitjançant l'IDA, corresponent al cens de 1991. L'origen de la representació correspon al baricentre de les poblacions. Cada comarca s'identifica per un codi de tres lletres (majúscules) que es mostra a la Taula 1. Les fletxes representen els grups professionals (identificats per les tres primeres lletres) i les línies discontínues separen els 9 grups en què es classifiquen les comarques.

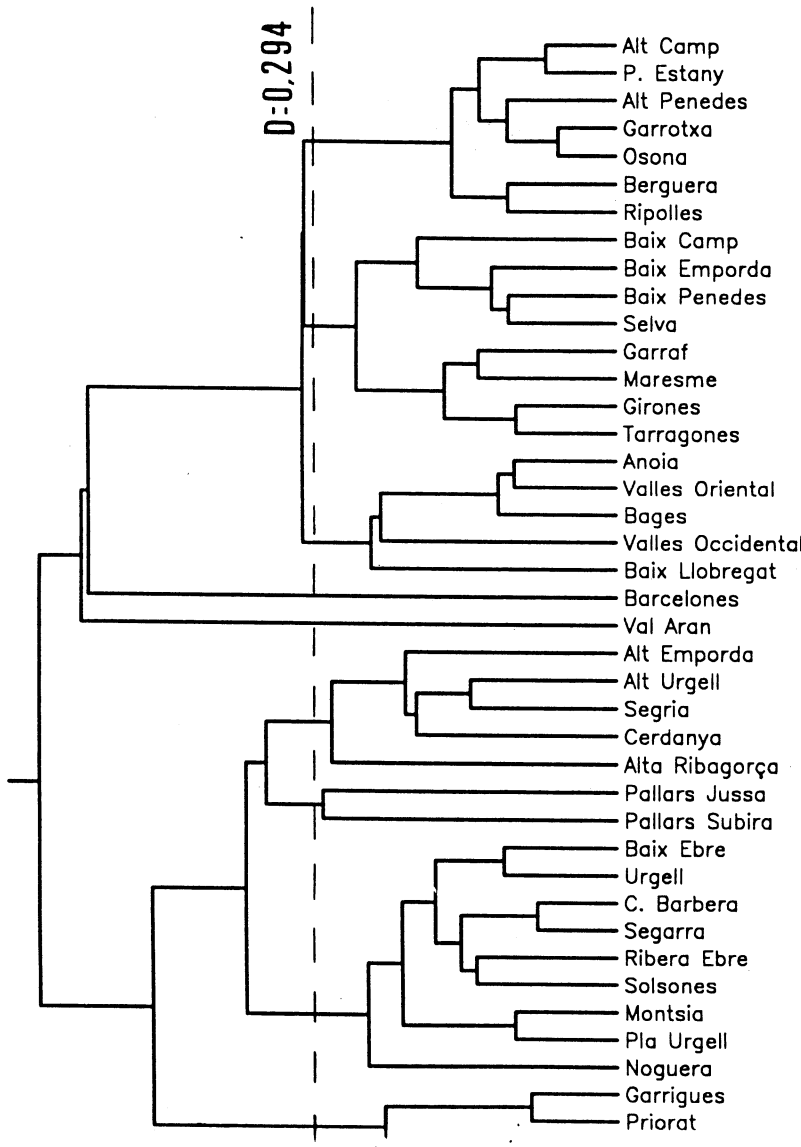


Figura 2

Dendrograma obtingut per l'UPGMA per les 41 comarques utilitzant la distància de Bhattacharyya. S'ha indicat la línia de tall, situada a d = 0.294, que dona els nou grups en què hem agrupat les comarques.

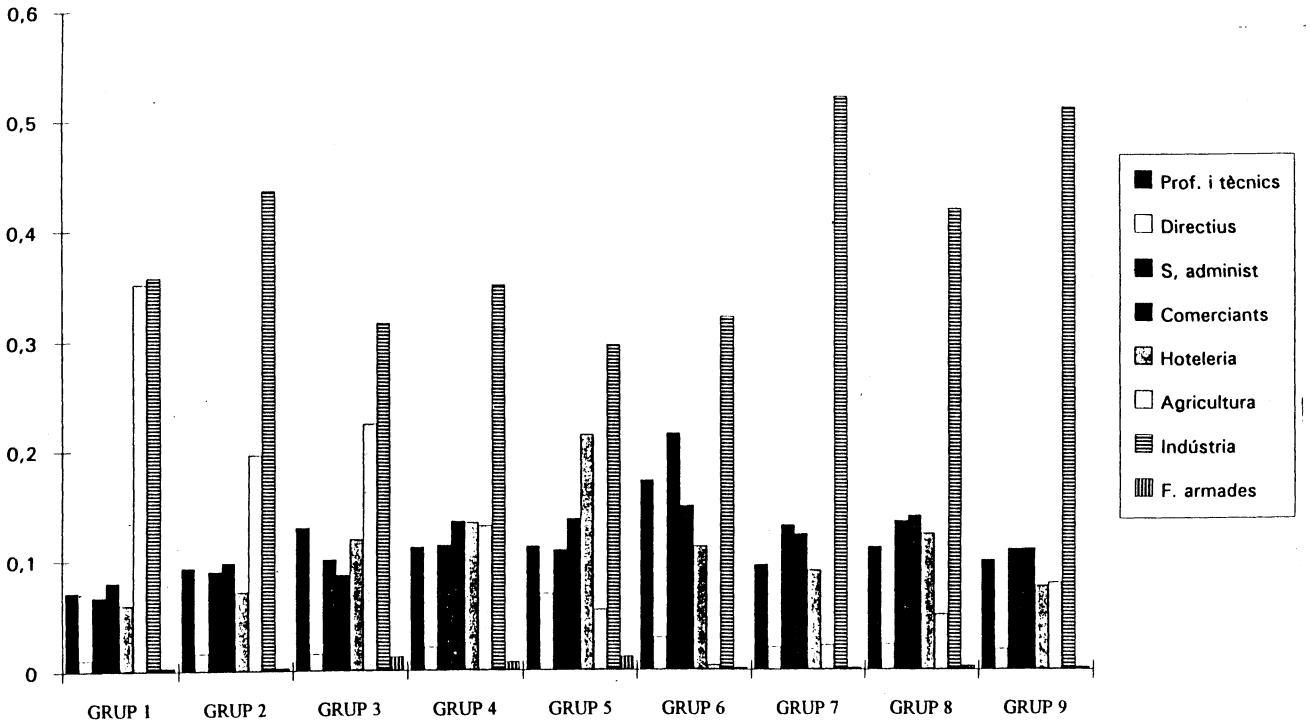


Figura 3

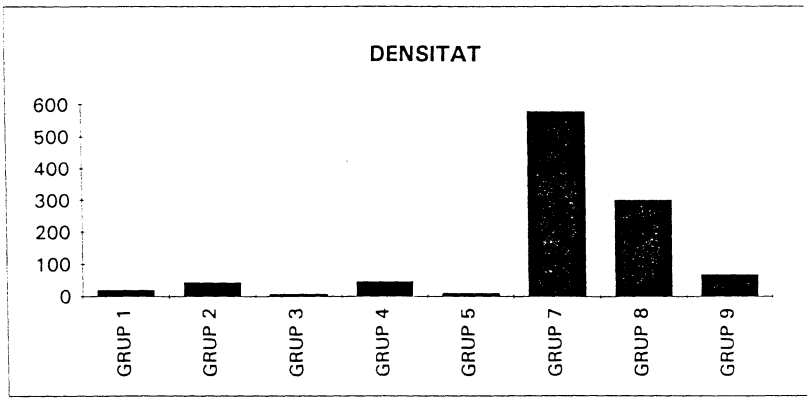
Proporcions mitges dels vuit grups professionals, variable **X**, pels nou grups en què s'han agrupat les quaranta-i-una comarques.

L'última part d'aquest treball va consistir a estudiar les característiques dels nou grups obtinguts. De les diverses variables estudiades, les que finalment van resultar més informatives van ser: la densitat (h./km^2), el saldo migratori, la renda familiar per càpita i la distribució per grups d'edat de la població activa. Els valors d'aquestes variables per a cada *grup* es troben a la figura 4.

5. DISCUSSIÓ I CONCLUSIONS

Amb vista a altres interpretacions i generalitzacions, cal no oblidar que en el nostre treball hem volgut donar el mateix pes a totes les comarques, ja que ens interessava, justament, estudiar-ne les relacions i característiques independentment del nombre d'habitants. És evident que si haguéssim treballat en nombres absoluts la comarca del Barcelonès hauria disfressat la pràctica totalitat dels resultats, ja que hi viuen el 38% dels habitants de Catalunya. En aquest sentit, una alternativa a la representació mitjançant l'IDA hauria estat l'Anàlisi de Correspondències (AC). És ben conegut que l'AC, per fer servir la distància khi quadrat, depèn de les mides mostrals. Quan es vol evitar aquesta dependència alguns autors (Greenacre, 1984; Greenacre,1993) recomanen realitzar l'AC sobre la taula de freqüències relatives (obtinguda dividint cada fila pel nombre total d'individus de la comarca corresponent). Els resultats, així obtinguts, es mostren a la figura 5 segons la denominada forma biplot en la que les files (comarques) es representen segons les seves coordenades principals i les fletxes representen les columnes (grups professionals) segons la direcció dels seus vèrtexs, però reescalades mitjançant la multiplicació de les seves coordenades estàndards per les freqüències relatives de cada grup professional (veure els capítols 10 i 13 de Greenacre, 1993, per a una excel·lent explicació). Una altra característica de la distància khi quadrat és que la distància entre dues poblacions també depèn de la resta de poblacions que considerem, ja que a la fórmula de la distància khi quadrat intervenen les sumes parcials tant per files com per columnes. Això pot provocar que la configuració obtinguda amb l'AC per a un conjunt de poblacions es vegi modificada si s'afegeix un segon grup de poblacions, tal com es mostra a l'exemple 4.2.1 de Rao (1995). En canvi, l'IDA utilitza la distància de Rao que no depèn de mides mostrals i la distància entre dues poblacions només depèn dels valors observats per a aquestes, però no de la resta de poblacions considerades. De fet, en estar basat en tècniques de geometria diferencial, l'IDA és invariant versus transformacions admissibles de les variables i/o els paràmetres. Així mateix, com per a la segona part del treball, l'elaboració d'una classificació jeràrquica de les comarques, van triar la distància de Bhattacharyya, semblava més coherent triar l'IDA com a mètode de representació, encara que, evidentment, l'AC hauria estat una alternativa (la similitud entre les figures 1 i 5 és molt clara), com també ho hauria estat el mètode descrit a Rao (1995) i la distància d'Hellinger.

a)



b)

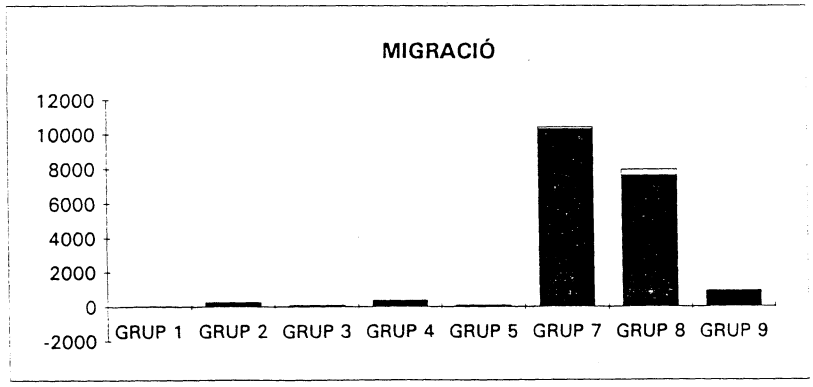


Figura 4

Algunes característiques socio-demogràfiques pels 9 grups en què hem classificat les comarques de Catalunya:

a) Densitat (h./km^2),

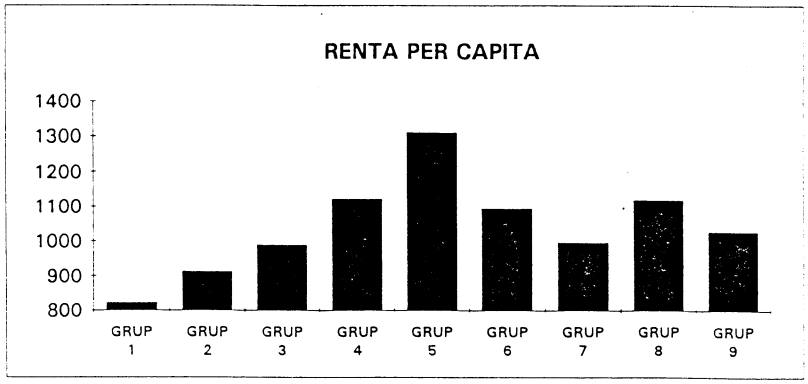
b) Saldo migratori total,

c) Renda familiar disponible per càpita, i

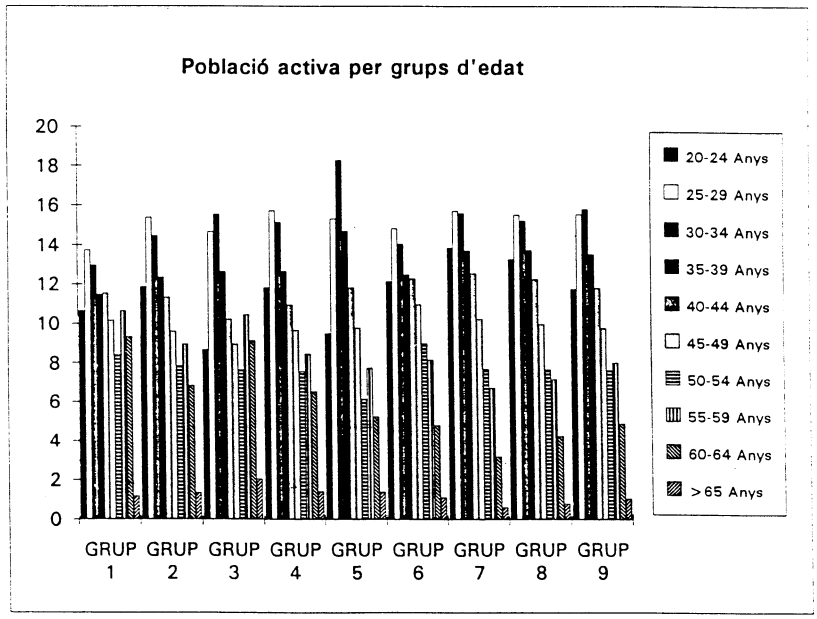
d) Distribució de la població activa, en percentatge, per grups d'edat.

En *a)* i *b)* s'ha omès el grup 6, corresponent al Barcelonès, pels seus valors atípics: densitat de 16.091 hab/Km^2 i saldo migratori negatiu de 21121 habitants, dividit en 19330 cap a Catalunya i 1791 cap a la resta d'Espanya.

c)


RENTA PER CAPITA
1400
1300
1200
1100
1000
900
800
GRUP 1
GRUP 2
GRUP 3
GRUP 4
GRUP 5
GRUP 6
GRUP 7
GRUP 8
GRUP 9

d)


Població activa per grups d'edat
20
18
16
14
12
10
8
6
4
2
0
GRUP 1
GRUP 2
GRUP 3
GRUP 4
GRUP 5
GRUP 6
GRUP 7
GRUP 8
GRUP 9
20-24 Anys
25-29 Anys
30-34 Anys
35-39 Anys
40-44 Anys
45-49 Anys
50-54 Anys
55-59 Anys
60-64 Anys
>65 Anys

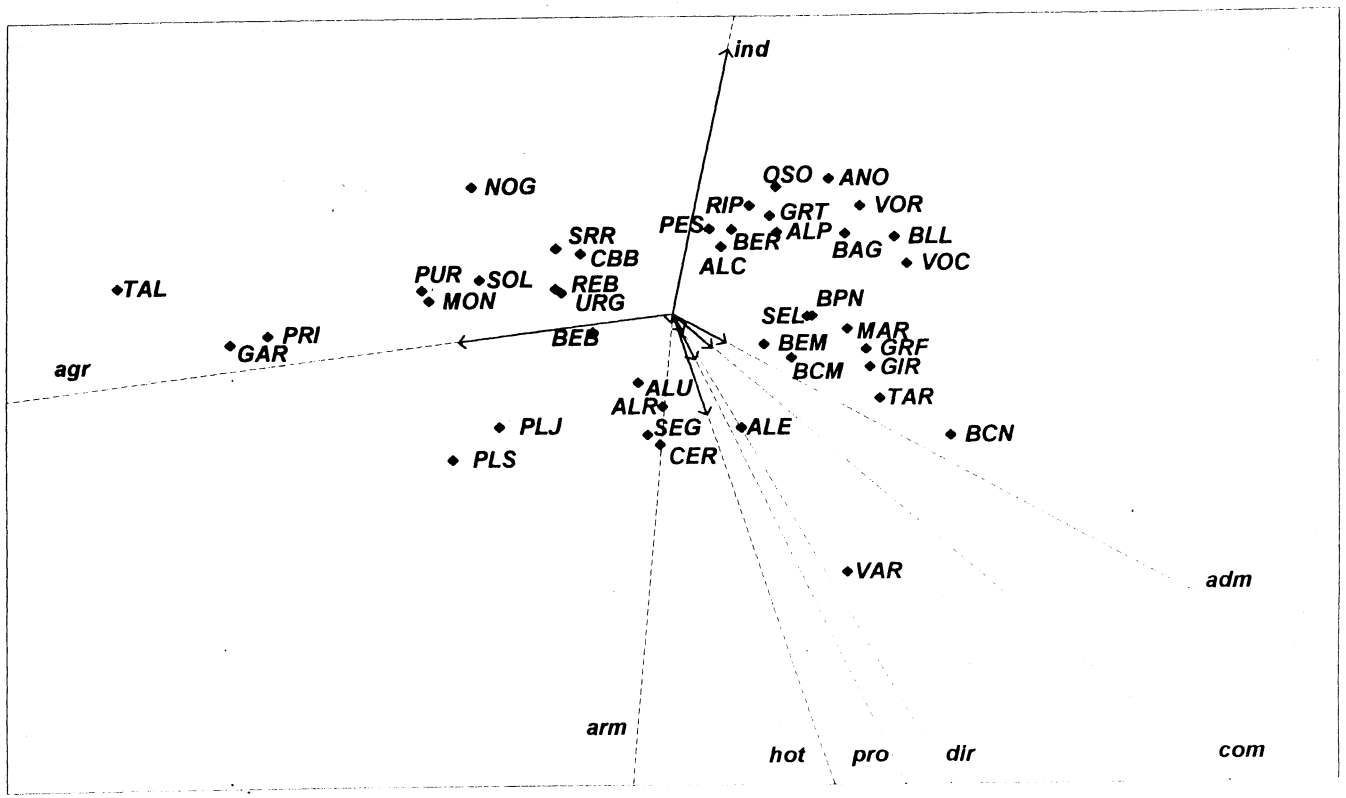


Figura 5

Representació bidimensional de les 41 comarques de Catalunya i de les 8 variables (grups professionals) obtinguda mitjançant l'Anàlisi Factorial de Correspondències (AC) aplicat a les mateixes dades que la figura 1. L'AC es mostra en la seva forma biplot (veure el text per més detalls) aplicat sobre les freqüències relatives. Les fletxes representen els grups professionals i s'han prolongat amb línies discontínues per tal de visualitzar amb claredat les direccions.

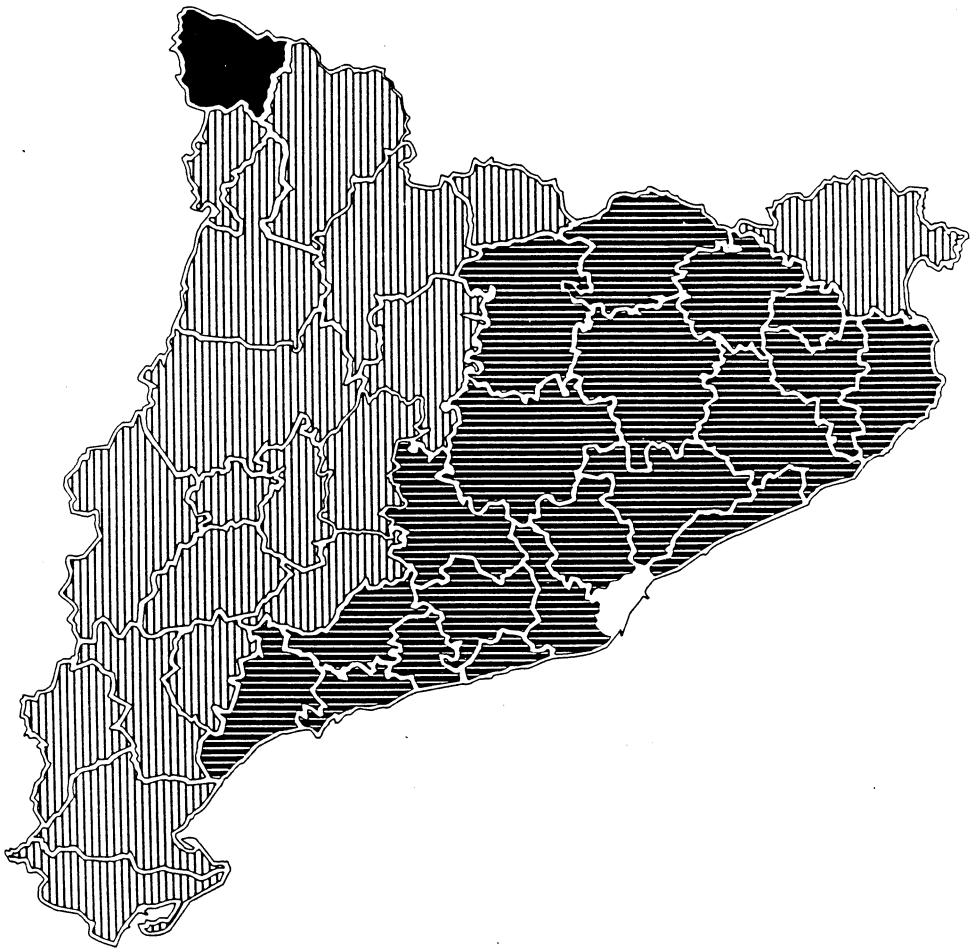


Figura 6

Distribució geogràfica dels blocs en que hem classificat les comarques de Catalunya:

a) BLOC AGRARI, identificat amb línies verticals,
c) BLOC TURÍSTIC, en negre
b) BLOC ADMINISTRATIU, en blanc
c) BLOC INDUSTRIAL, amb línies horitzontals

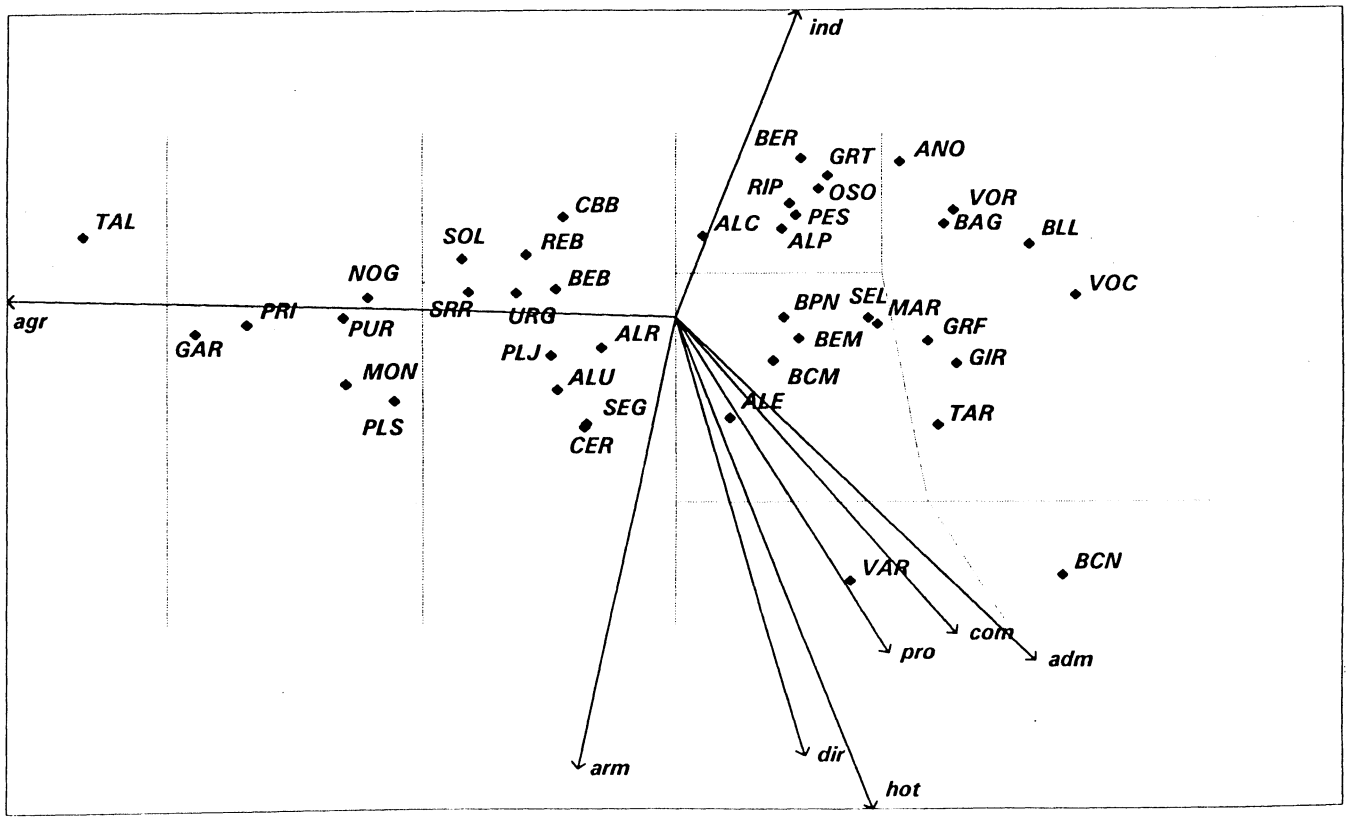


Figura 7

Representació bidimensional de les 41 comarques de Catalunya i de les 8 variables (grups professionals) obtinguda mitjançant l'IDA, corresponent al cens de 1986. La terminologia utilitzada és la mateixa que la de la fig. 1.

Taula 1

Classificació de les comarques de Catalunya obtinguda segons l'IDA (model multinomial) i l'UPGMA (distancia de Bhattacharyya).

	Grup 1		**Grup 6**
GAR:	Garrigues	BCN:	Barcelonès
PRI:	Priorat		
TAL:	Terra Alta	**Grup 7**	
		ANO:	Anoia
	Grup 2	BAG:	Bages
BEB:	Baix Ebre	BLL:	Baix Llobregat
CBB:	Conca de Barberà	VOC:	Vallès Occidental
MON:	Montsià	VOR:	Vallès Oriental
NOG:	Noguera		
PUR:	Pla d'Urgell		**Grup 8**
REB:	Ribera d'Ebre	BCM:	Baix Camp
SRR:	Segarra	BEM:	Baix Empordà
SOL:	Solsonès	GRF:	Garraf
URG:	Urgell	GIR:	Gironès
	Grup 3	MAR:	Maresme
PLJ:	Pallars Jussà	SEL:	Selva
PLS:	Pallars Sobirà	TAR:	Tarragonès
	Grup 4		**Grup 9**
ALE:	Alt Empordà	ALC:	Alt Camp
ALR:	Alta Ribagorça	ALP:	Alt Penedès
ALU:	Alt Urgell	BER:	Berguedà
CER:	Cerdanya	GRT:	Garrotxa
SEG:	Segrià	OSO:	Osona
		PES:	Pla de l'Estany
	Grup 5	RIP:	Ripollès
VAR:	Val d'Aran		

Un resultat important del present estudi és que la divisió comarcal de Catalunya es pot interpretar en termes de les variables aquí estudiades (a més del factor geogràfic que comentarem posteriorment) i d'acord amb altres estudis basats en diferents tipus de dades (Calapell i Hernández, 1993). Així mateix, també és remarcable que sigui el component agrícola (X_6) el més important en la seva diferenciació (figura 1), ja que cal no oblidar que l'existència de mercats agrícoles va ser precisament un dels factors claus en la divisió original del territori de Catalunya en comarques. Per tant, a pesar de l'evolució soferta i de la creixent industrialització del seu territori (hem vist

que en tots els grups era sempre el component industrial el majoritari), el component agrícola segueix sent el que millor explica les diferències intercomarcals.

Com hem vist a la secció de *Resultats*, hem classificat les comarques de Catalunya en nou *grups* que, a la vista dels resultats anteriors, es poden agrupar en quatre grans blocs amb les característiques essencials següents:

- **Bloc agrícola**: Comprèn els *clusters* 1, 2, 3 (situats a l'esquerra de la fig 1) i 4 (part central inferior) en què el component X_6 (agricultura i pesca) té més importància (figures 1 i 3), encara que convé no oblidar que el sector a què més recursos humans dediquen els nou *grups* és l'industrial. Aquest bloc també es caracteritza, en general, per un saldo migratori molt baix (positiu o negatiu), una densitat baixa i una renda familiar per càpita baixa. Cal assenyalar que el *grup* 4, encara que d'acord amb l'UPGMA s'engloba clarament en aquest bloc, presenta certes característiques particulars: els components d'X corresponents al sector serveis presenten unes freqüències relativament altes (figura 3) i la renda familiar per càpita és superior a la de la resta dels grups (figura 4c). Aquestes peculiaritats concorden perfectament amb la seva posició a la figura 1, molt menys extrema que la de la resta de grups d'aquest bloc, i podrien justificar-se per la importància del sector turístic en aquestes comarques.

- **Bloc turístic**: Format pel *cluster* 5 (part inferior central de la fig. 1) que inclou únicament la Vall d'Aran. Es caracteritza per la importància del sector *serveis* (professionals (X_1), administratius (X_3), comerciants (X_4) i, especialment, hotelers (X_5)), que s'explicaria per la importància del turisme en aquesta comarca. Així mateix, es caracteritza per una elevada renda per càpita (fig. 4c) i una densitat baixa.

- **Bloc administratiu**: Inclou el *cluster* 6 (part inferior dreta de la fig. 1), format únicament pel Barcelonès (la comarca que inclou Barcelona), amb unes connotacions molt especials. Els sectors professionals i comercials estan àmpliament representats, però és sobretot la importància del sector administratiu el que més el diferencia de la resta. Aquest fet s'explica fàcilment pel seu paper de capital administrativa del principat. Altres característiques d'aquest bloc són: densitat molt elevada (16.901 h./km^2), saldo migratori molt negatiu dirigit principalment a la resta de Catalunya i pràctica absència del component agricultura i pesca. També resulta destacable el fet que, en contra de les idees preestablertes que es poden tenir, en el Barcelonès el percentatge de la població activa que es dedica a la indústria és inferior al dels *grups* que l'envolten geogràficament, probablement a causa de la descentralització creixent d'aquest sector i la seva importància administrativa.

- **Bloc industrial**: Inclou els *grups* 7, 8, i 9 (part superior dreta de la fig. 1), caracteritzats per un component industrial molt alt, mentre que el component

agrícola és poc important. El saldo migratori és positiu (provinent de la resta d'Espanya i de Catalunya), la densitat elevada (encara que no comparable a la del Barcelonès) i la renda per càpita intermèdia.

Respecte a la distribució per grups d'edat (fig. 4d), cal indicar que les diferències no són tant clares com a les variables ja comentades, encara que s'observen algunes peculiaritats. Així, en el *bloc agrícola* s'observa una distribució bimodal amb dos màxims: el primer corresponent als grups de 25-29 i 30-34 anys, i el segon situat als 55-59 anys. En el *bloc turístic* el segon màxim gairebé no s'aprecia, i es confirma aquesta tendència en els *blocs administratiu i industrial*, en què el segon màxim ha desaparegut totalment. També cal tenir en compte que aquesta divisió en blocs no és absoluta, sinó que presenta una gradació, de manera que els grups (i per tant les comarques incloses) que a la fig. 1 es troben pròxims, encara que pertanyin a blocs diferents, poden presentar certes similituds. Així, prenent com a exemple el *bloc* agrícola, queda clar que el *cluster* 1 (el més extrem del bloc) presenta totes les característiques que defineixen aquest bloc, mentre que el *cluster* 4 (de posició més intermèdia) ja hem vist que presenta certes similituds amb les d'altres blocs.

Es destacable que aquests quatre *blocs* presenten una clara gradació geogràfica, tal com es mostra a la figura 6. Així, el *bloc industrial* correspon a les comarques orientals, de manera que observem clarament una línia divisòria amb les comarques de ponent que corresponen al *bloc agrícola* amb l'única excepció de l'Alt Empordà. Els dos, formats només per una comarca *turística i administrativa*, es troben intercalats entre els blocs anteriors.

Un aspecte interessant, amb vista a estudis posteriors, és el seguiment de l'evolució de la variable X al llarg del temps. En aquest sentit es va repetir aquest estudi amb les dades de l'any 1986 (Anuari Estadístic de Catalunya, 1990), ja que no es disposa de dades per comarques de censos anteriors. A la figura 7 se sintetitzen els resultats obtinguts per l'IDA i l'UPGMA (tallant també per d = 0.294), i s'obté una classificació en 8 *grups*. Com era d'esperar, la interpretació de la figura 6 és molt semblant a la de la figura 1, de manera que continua sent la variable X_6 (agricultura i pesca) la que millor discrimina les relacions entre les comarques, seguida d'X_7 (indústria). Sintetitzant els resultats podem dir que la classificació en blocs s'ha mantingut pràcticament inalterable i només una comarca, l'Alt Empordà, ha canviat de bloc. Respecte als grups, l'any 1986, el Barcelonès i la Vall d'Aran també es presentaven en blocs individuals; els grups del *bloc industrial* presenten una gran estabilitat. Per contra, els grups agrícoles són els que més han evolucionat durant aquest període, de manera que a l'any 1986 només es distingien tres grups de composició clarament diferent a la dels quatre actuals. Resultarà interessant comprovar si la tendència del grup 4 cap al sector serveis es confirma en propers censos.

Per últim, respecte a la metodologia utilitzada, volem destacar la fàcil interpretació de la representació obtinguda pel mètode IDA, com també la seva concordança amb els

resultats obtinguts a partir de l'UPGMA. Tot això mostra com les tècniques basades en l'aplicació de la geometria diferencial a l'estadística poden ser utilitzades sense gaires dificultats i obtenir resultats òptims en casos pràctics. Amb aquest treball esperem contribuir a disminuir la reticència a la utilització d'aquestes tècniques en el camp aplicat, justificada, en part, pel caràcter tan teòric que solen tenir les publicacions d'aquesta branca de l'estadística.

AGRAÏMENTS

Els autors volen agrair al Sr. Jordi Oliveras i Prats, director de l'Institut d'Estadística de Catalunya de la Generalitat de Catalunya, el fet d'haver-nos facilitat moltes de les dades utilitzades en aquest treball corresponents al cens de 1991 i que encara no han estat publicades.

APÈNDIX

Dades originals empleades

Població ocupada per grups professionals (1991).

Dades facilitades per l'Institut d'Estadística de Catalunya.

Comarca	Professionals i tècnics	Personal directiu	Serveis administratius	Comerciants i venedors	Hoteleria i altres	Agricultura i pesca	Indústria	Forces armades
Alt Camp	1231	243	1446	1420	875	1265	6286	25
Alt Empordà	2948	793	5040	5510	4823	3509	12083	317
Alt Penedès	2419	502	3667	3077	2000	1827	13118	36
Alt Urgell	778	135	835	1020	798	1068	2777	79
Alta Ribagorça	175	23	98	131	199	163	469	1
Anoia	2764	614	3462	3556	2408	1124	17472	43
Bages	6274	1022	6485	7095	4570	1755	28255	171
Baix Camp	5699	989	6165	7029	5221	3270	18436	110
Baix Ebre	2446	383	2311	2808	1994	3682	8846	65
Baix Empordà	2810	737	3716	4900	4635	2747	14519	127
Baix Llobregat	12371	4009	31296	26849	24955	2605	110826	274
Baix Penedès	1116	320	1705	1997	1762	785	6305	49
Barcelonès	146521	24845	182813	126740	95496	3462	274395	1258
Berguerà	1373	164	1207	1555	1131	1129	6910	78
Cerdanya	492	116	462	679	786	670	1695	38
Conca de Barberà	563	124	636	631	488	1068	3018	7
Garraf	3484	549	3419	3875	3559	836	11448	43
Garrigues	539	79	524	619	424	2338	2286	13
Garrotxa	1909	390	2064	2037	1420	1264	9712	32
Gironès	7315	1187	8884	7173	5127	1727	19917	269
Maresma	12837	3475	15056	15560	10867	4504	45818	189
Montsià	1329	282	1600	2046	1394	4588	7716	77
Noguera	1131	185	931	1226	824	3215	7911	35
Osona	4901	901	5277	5423	3238	3076	26436	50
Pallars Jussà	567	79	479	465	410	955	1530	101
Pallars Sobirà	280	27	200	148	307	497	620	6
Pla d'Urgell	863	169	1019	1020	597	2570	4200	24
Pla de l'Estany	923	187	1036	881	587	804	4004	8
Priorat	287	34	245	255	232	1063	1179	10
Ribera d'Ebre	936	75	684	657	592	1318	3263	27
Ripollès	1012	193	905	1106	1006	801	5908	27
Segarra	654	125	653	560	415	1152	3023	6
Segrià	7841	1279	8280	8294	6253	8678	18970	577
Selva	2776	744	4106	4720	5758	2149	17562	66
Solsonès	431	61	330	315	348	900	1854	6
Tarragonès	8047	1201	9403	7294	7309	1640	21352	348
Terra Alta	217	41	220	324	209	1757	1710	16
Urgell	1020	235	1099	1431	758	1991	4699	31
Val d'Aran	295	182	286	360	562	143	779	32
Vallès Occidental	28614	5383	34772	31343	21310	1610	114191	231
Vallès Oriental	9550	2250	13548	11619	8395	2499	54530	122

Densitat de població (any 1991).

Dades reproduïdes de l'Anuari Estadístic de Catalunya de 1992.

Comarca	hab./Km^2
Alt Camp	62.5
Alt Empordà	67.6
Alt Penedès	117.9
Alt Urgell	13.0
Alta Ribagorça	8.2
Anoia	95.1
Bages	117.5
Baix Camp	189.3
Baix Ebre	65.4
Baix Empordà	128.4
Baix Llobregat	1254.2
Baix Penedès	128.9
Barcelonès	16091.0
Berguerà	33.0
Cerdanya	22.7
Conca de Barberà	27.7
Garraf	417.8
Garrigues	24.3
Garrotxa	62.7
Gironès	218.7
Maresma	738.5
Montsià	76.6
Noguera	20.1
Osona	92.9
Pallars Jussà	10.0
Pallars Sobirà	4.0
Pla d'Urgell	94.6
Pla de l'Estany	80.2
Priorat	19.1
Ribera d'Ebre	27.9
Ripollès	28.3
Segarra	23.6
Segrià	176.9
Selva	98.7
Solsonès	10.8
Tarragonès	491.5
Terra Alta	17.5
Urgell	50.8
Val d'Aran	10.0
Vallès Occidental	1118.9
Vallès Oriental	308.2

Moviments migratoris (any 1990).

Dades reproduïdes de l'Anuari Estadístic de Catalunya de 1992.

Comarca	Amb la resta de Catalunya	Amb la resta d'Espanya	Saldo total
Alt Camp	116	11	127
Alt Empordà	174	237	411
Alt Penedès	489	30	519
Alt Urgell	67	120	187
Alta Ribagorça	10	10	20
Anoia	407	89	496
Bages	150	−44	106
Baix Camp	536	385	921
Baix Ebre	42	−5	37
Baix Empordà	444	333	777
Baix Llobregat	3236	186	3422
Baix Penedès	786	55	841
Barcelonès	−19330	−1791	−21121
Berguerà	−79	−2	−81
Cerdanya	47	24	71
Conca de Barberà	25	−30	−5
Garraf	1064	97	1161
Garrigues	−7	−9	−16
Garrotxa	52	58	110
Gironès	515	295	810
Maresma	3137	276	3413
Montsià	91	50	141
Noguera	1	13	14
Osona	152	41	193
Pallars Jussà	35	−10	25
Pallars Sobirà	39	−8	31
Pla d'Urgell	−16	−6	−22
Pla de l'Estany	186	−6	180
Priorat	−4	−2	−6
Ribera d'Ebre	−64	−20	−84
Ripollès	−76	−22	−98
Segarra	49	19	68
Segrià	7	35	42
Selva	722	472	1194
Solsonès	72	−15	57
Tarragonès	386	597	983
Terra Alta	0	−17	−17
Urgell	44	32	76
Val d'Aran	0	70	70
Vallès Occidental	3351	312	3663
Vallès Oriental	3144	395	3539

Renda familiar disponible per càpita (any 1989).

Dades reproduïdes de l'Anuari Estadístic de Catalunya de 1992.

Comarca	Milers de PTA
Alt Camp	990.3
Alt Empordà	1221.6
Alt Penedès	1063.1
Alt Urgell	978.6
Alta Ribagorça	1123.1
Anoia	1009.3
Bages	961.5
Baix Camp	1016.9
Baix Ebre	993.8
Baix Empordà	1202.2
Baix Llobregat	959.1
Baix Penedès	1153.7
Barcelonès	1095.6
Berguerà	895.1
Cerdanya	1340.2
Conca de Barberà	916.0
Garraf	1103.0
Garrigues	776.7
Garrotxa	1090.2
Gironès	1103.1
Maresma	1126.6
Montsià	1001.7
Noguera	874.2
Osona	1060.2
Pallars Jussà	953.8
Pallars Sobirà	1026.0
Pla d'Urgell	908.0
Pla de l'Estany	1111.7
Priorat	834.5
Ribera d'Ebre	918.3
Ripollès	994.6
Segarra	811.2
Segrià	945.8
Selva	1171.3
Solsonès	910.6
Tarragonès	1084.4
Terra Alta	853.5
Urgell	873.3
Val d'Aran	1311.8
Vallès Occidental	981.7
Vallès Oriental	1078.4

Població activa per grups d'edat (any 1991).

Dades facilitades per l'Institut d'Estadística de Catalunya.

Comarca	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	> 65
Alt Camp	1530	1936	2012	1622	1385	1253	900	888	607	111
Alt Empordà	4274	5162	4704	4216	3906	3268	2648	2755	1902	366
Alt Penedès	3356	3958	3847	3334	2925	2442	1779	1847	1356	237
Alt Urgell	863	1064	1088	891	814	696	500	622	523	138
Alta Ribagorça	129	208	210	163	110	110	94	104	81	21
Anoia	3859	4673	4597	4046	3788	3131	2324	2236	1054	213
Bages	6629	8348	8687	7280	6520	5321	3939	3963	1949	434
Baix Camp	5878	7014	6880	5888	5480	4446	3351	3153	2071	369
Baix Ebre	2748	3156	2908	2520	2613	2207	1884	1926	1269	258
Baix Empordà	4423	5049	4705	4194	3750	3201	2510	2551	1679	290
Baix Llobregat	31717	32662	30854	28056	27066	22686	17570	13218	5682	870
Baix Penedès	1902	2108	1895	1742	1637	1302	1065	987	575	117
Barcelonès	100357	122492	116025	103243	101586	90528	74228	67333	39678	9326
Berguerà	1425	2122	2257	1857	1572	1232	915	891	554	174
Cerdanya	592	699	693	630	544	491	352	383	267	79
Conca de Barberà	743	963	881	742	694	632	477	584	495	88
Garraf	3488	3970	4067	3698	3184	2566	2086	1747	882	169
Garrigues	744	958	857	788	726	684	518	707	538	62
Garrotxa	1881	2621	2821	2510	2243	1796	1492	1700	916	198
Gironès	6284	7488	7661	7080	6037	4810	3536	3527	2166	490
Maresma	13019	15634	15642	14736	13548	10879	8136	7222	3652	748
Montsià	2355	2635	2396	2128	2027	1834	1584	1664	1302	200
Noguera	1350	1873	1693	1458	1311	1108	917	1122	989	147
Osona	5767	7501	7599	6578	5577	4385	3480	3594	1827	470
Pallars Jussà	400	670	617	532	496	412	374	517	399	94
Pallars Sobirà	173	294	354	275	192	179	143	192	192	41
Pla d'Urgell	920	1278	1258	1112	931	767	587	653	475	118
Pla de l'Estany	1330	1646	1437	1201	1052	919	768	884	651	113
Priorat	324	448	430	359	371	322	258	348	303	46
Ribera d'Ebre	701	1057	1134	951	853	750	615	715	506	66
Ripollès	1036	1492	1644	1506	1384	1121	931	946	444	102
Segarra	779	1093	944	782	710	523	422	570	469	88
Segrià	6973	9248	8892	7633	6820	5671	4444	4443	2994	557
Selva	4912	5760	5599	4836	4238	3553	2739	2568	1492	242
Solsonès	477	623	601	472	467	370	349	352	290	87
Tarragonès	6879	8273	8412	7722	6818	5501	4037	3773	2348	391
Terra Alta	453	547	537	485	518	433	402	445	448	50
Urgell	1354	1762	1531	1274	1154	1031	782	1042	791	125
Val d'Aran	245	395	471	379	305	252	158	200	135	36
Vallès Occidental	31836	35907	35347	32083	28660	22408	16614	13421	6426	1220
Vallès Oriental	14194	15082	15229	13231	11841	9578	7326	6226	3110	575

Població ocupada per grups professionals (1986).

Dades reproduïdes de l'Anuari Estadístic de Catalunya de 1991.

Comarca	Professionals i tècnics	Personal directiu	Serveis administratius	Comerciants i venedors	Hoteleria i altres	Agricultura i pesca	Indústria	Forces armades	Altres
Alt Camp	1028	280	913	1137	654	1417	5455	15	334
Alt Empordà	2153	564	3418	4757	3673	3485	10448	287	871
Alt Penedès	1785	435	2369	2086	1219	1906	10757	10	707
Alt Urgell	649	238	593	580	515	1345	2528	98	447
Alta Ribagorça	124	20	65	91	96	180	427	4	69
Anoia	2032	301	2432	2831	1661	1226	15092	12	496
Bages	4704	903	4519	5924	3758	1738	25562	111	633
Baix Camp	4554	1030	4128	5014	3823	3920	14590	128	564
Baix Ebre	1975	377	1518	2248	1407	4392	7788	47	366
Baix Empordà	2229	721	2658	4193	3999	2722	11844	65	278
Baix Llobregat	15002	3059	18646	17558	16478	2775	86143	224	2118
Baix Penedès	798	228	985	1456	1111	973	4724	24	322
Barcelonès	109438	40542	122925	79951	65488	2466	209762	1499	83063
Berguedà	1098	128	842	1220	778	996	7802	57	399
Cerdanya	324	153	291	478	468	761	1411	50	214
Conca de Barberà	457	137	464	484	388	1304	2896	8	144
Garraf	2326	340	1995	2888	2416	783	9452	16	813
Garrigues	437	94	341	415	300	3011	1760	21	349
Garrotxa	1496	774	1510	1616	960	1321	9940	8	253
Gironès	6128	1364	6015	4815	3615	1387	17713	272	802
Maresme	8747	3122	9632	10882	8190	5001	38721	105	1960
Montsià	921	246	978	1343	862	5089	5142	63	2291
Noguera	908	157	699	1055	604	3848	3826	40	378
Osona	3904	763	3611	4293	2392	3108	24107	19	1465
Pallars Jussà	426	166	338	378	305	903	1552	124	77
Pallars Sobirà	191	68	115	136	178	613	591	9	100
Pla d'Urgell	602	149	494	679	355	2917	3128	32	889
Pla de l'Estany	783	278	793	874	457	737	4014	1	98
Priorat	217	32	163	182	151	1260	919	12	206
Ribera d'Ebre	772	81	459	560	379	1547	2870	18	218
Ripollès	925	212	575	1032	635	803	5536	25	350
Segarra	501	95	506	418	352	1527	2090	11	171
Segrià	6116	1581	5848	5490	3899	9758	15699	470	1707
Selva	2143	769	2959	4289	5673	2177	15148	32	370
Solsonès	305	83	228	246	239	1000	1504	11	107
Tarragonès	6503	1267	6663	4968	5950	1806	16696	216	800
Terra Alta	169	46	131	208	155	2413	1210	15	112
Urgell	783	219	734	1207	544	2341	3721	37	250
Vall d'Aran	237	116	189	399	460	137	622	36	77
Vallès Occidental	22082	3824	20615	20135	14845	1217	87170	121	5564
Vallès Oriental	6644	1893	7985	8050	5657	2503	42332	72	1341

REFERÈNCIES

[1] **Amari,S.** (1985). «Differential–Geometrical Methods in Statistics». *Lecture notes on statistics*, **28**. Springer-Verlag, New York.

[2] **Atkinson, C.** and **Mitchell, A. F. S.** (1981). «Rao's Distance Measure». *Sankhyā*, **43, A**, 345–365.

[3] **Bandorff-Nielsen, O.E.** (1984). «Differential geomtry in statistical inference». Capater Differential and integral geometry in statistical inference Volume 10 of *Lecture Notes - Monographs Series* Institute of Mathematical Statistics, Hayward, California.

[4] **Burbea, J.** and **Rao, C. R.** (1982). «Entropy differential metric, distance and divergence measures in probability spaces: a unified approach». *J. Multivarate Anal.*, **12**, 575–596.

[5] **Burbea, J.** and **Rao, C. R.** (1984). «Differential metrics in probability spaces». *Probability Math. Statist.*, **3**, 241–258.

[6] **Calafell, F.** and **Hernández, M.** (1993). «Multivariate approach to matrimonial mobility in Catalonia». *Human Biology*, **65**, 731–742.

[7] **Efron** (1975). «Defining the curvature of a statistical problem (with application to second order efficiency) (with discussion)». *Annals of Statistics*, **3**, 1189–1242.

[8] **Generalitat de Catalunya.** *Anuari Estadístic de Catalunya de 1990.*

[9] **Generalitat de Catalunya.** *Anuari Estadístic de Catalunya de 1992.*

[10] **Greenacre, M.J.** (1984). *Theory and applications of Correspondence Analysis.* Academic Press, London.

[11] **Greenacre, M.J.** (1993). *Correspondence analyisis in practice.* London: Academic Press.

[12] **Oller, J. M.** (1982). *Utilización de métricas riemannianas en análisis de datos multidimensionales y su aplicación a la Biología.* Barcelona: Publicaciones de Bioestadística.

[13] **Oller, J.M.** (1989). «Statistical data analysis and inference». Chapter: *Some geometrical aspects of data analysis and statistics.* Elsevier science publishers B.V. North Holland, Amsterdam.

[14] **Rao, C. R.** (1945). «Information and accuracy attainable in the estimation of parameters». *Bull. Calcurra Math. Soc.*, **37**, 81–91.

[15] **Rao, C.R.** (1995). «A review of canonical coordinates and an alternative to correspondence analysis using hellinger distance». *Qüestiió*, **19**, 23–64.

[16] **Ríos, M., Villarroya, A.** and **Oller, J. M.** (1994). «Intrinsic Data Analysis: a new method for simultaneous representation of populations and variables». *Mathematical Preprint Series*, **160**, Universitat de Barcelona.

[17] **Sokal, R.** and **Sneath, P.H.A.** (1963). *Principles of Numerical Taxonomy*. W. H. Freeman and Co.

ENGLISH SUMMARY

COMBINING GEOMETRIC-DIFFERENTIAL TECHNIQUES WITH CLASSIC MULTIVARIATE ANALYSIS: AN APPLICATION TO CHARACTERIZE CATALAN REGIONS

SERGI VIVES*

ÀNGEL VILLARROYA*

Universitat de Barcelona

In this work the characteristics and relationships, from a social-economic point of view, of the 41 **comarques** *which compose Catalonia are studied. The methodology used is a combination of classical data analysis techniques and a new method for graphical display of statistical models, called IDA (Intrinsic Data Analysis), based on differential geometry tools. This approach has allowed us to classify the* **comarques** *of Catalonia into 9 clusters which, in their turn, can be grouped into four blocks: agricultural, tourist, administrative and industrial.*

Keywords: Intrinsic Data Analysis; distància de Rao; Anàlisi de Correspondències; Anàlisi de conglomerats.

AMS Classification: 62H25, 62H30.

*S. Vives i À. Villarroya. Dept. d'Estadística. Universitat de Barcelona. Av. Diagonal, 645. 08028 Barcelona. Espanya.

-Received january 1995.

-Accepted septembre de 1996.

The use of differential geometry in statistics is a fruitful branch of this science but is hardly used in practical applications, probably due to the theoretical character of published papers about this subject. This paper illustrates how such techniques, alone or together with classical ones, can be used to solve practical problems easily.

The populations of this study are the 41 *comarques* (geographic and administrative divisions) of Catalonia. Originally the criteria used to divide the region were mainly commercial ones, based on proximity to agricultural markets. The main variable of our paper is the random variable X: *Active population in professional groups*, according to the census of 1991, where the groups are:

X_1: Professionals and technicians	X_2: Managing directors
X_3: Civil servants	X_4: Traders and salesmen
X_5: Hotels and others	X_6: Agriculture and fishing
X_7: Industry	X_8: Army

This variable was chosen because, from an anthropological point of view, the way in which human resources are distributed is one of the most important distinctive traits of a population.

To study the relationships between *comarques* the first step was their graphical representation in a plane by means of a new method called intrinsic data anaylis (IDA; Ríos *et al.*, 1994) based on the application of differential geometry tools to statistics. IDA allows the joint representation of populations and variables and can be applied to any statistical model; in our case, according to the nature of variable X, the multinomial model was used. The result is shown in figure 1, which explains 90% of original variability. From this display we can deduce that the most important factor in the differentiation of *comarques* is X_6 (the agricultural component), followed by X_7 (the industrial component). In figures 5a and 5b displays are shown of the same data using correspondence analysis.

The next step consisted of grouping the *comarques* into a reasonable number of meaningful groups. To do it we used the cluster analysis technique known as *average linkage clustering* (also called UPGMA), applied to the interdistance matrix calculated by means of the Bhattacharyya distance, every *comarca* being identified by the value of X. The selection of the Bhattacharyya distance was due to its good properties and to the fact that it coincides with the Rao distance for the multinomial model, the same that IDA uses. UPGMA results are given in figure 2, showing that the dendogram has been cut at level $d = 0.294$ reulting in nine clusters. In table 1 the *comarca* nomenclature and the group composition is shown. In figure 3 the mean values of vector X for each group can be seen.

At this point we proceeded to study the characteristics of the nine clusters. Several variables were studied, but the most informative turned out to be: density, migratory balance, *per capita* family income and age distribution of the active population. The observed mean values for each cluster are displayed in figure 4.

From previous results some conclusions can be made. First of all, when professional groups of the active population are considered, the agrarian component (X_6) is the most important one for the *comarques* differenciation, although the industry is the major component in all the groups (see figures 1 and 3). Another important result is that the 41 *comarques* can be classified into 9 clusters which, in their turn, can be grouped into the following four blocks:

- **Agricultural block**: Composed of clusters 1, 2, 3 (placed on the left side of figure 1) and 4 (central-down placed) characterized by the great importance of X_6, a low migratory balance (positive or negative), a low density and low *per capita* family income, although cluster 4 presents some characteristics similar to those of the other blocks.

- **Touristic block**: Composed of cluster 5, which only includes the *Val d'Aran comarca*, with great importance of service components and with the higher *per capita* family income.

- **Administrative block**: Composed of cluster 6, Barcelones, the *comarca* which includes Barcelona city) with very special characteristics: it includes 38 % of Catalonia's citizens, very high population density, very negative migratory balance and a very small agricultural component.

- **Industrial Block** : Composed of clusters 7, 8 and 9 (upper right side of figure 1) with a very high industrial component and an absence of the agrarian component. Migratory balance is positive, population density high and family income intermediate.

Till this point we have not considered the geographic location of the *comarques*, but when the four previous blocks are placed on a map of Catalonia (figure 6) it is observed that the agricultural block is placed on the west side (with only one exception) and clearly separated from the *comarques* of the industrial block situated on the east side. Service and administrative blocks are situated inside the agricultural and industrial blocks, respectively.

Finally, results corresponding to census of 1991 were compared with those from the census of 1986, the only other census with the same data available (figure 7). The industrial block is the most stable during this period, while the agricultural block changes the most.