MÈTODES DE REPRESENTACIÓ DE DADES I LA SEVA APLICACIÓ A LA BIOLOGIA

per CARLES M. CUADRAS

Unitat Docent de Bioestadística. Facultat de Biologia.
Universitat de Barcelona

ABSTRACT

Methods of data representation and their application in Biology

Several multivariate methods of data representation widely used in Biology are presented and compared. Their utilization in the analysis of biological data is discussed and several examples of application are given.

INTRODUCCIÓ

En aquest treball es comenten les tècniques per a representar dades, d'aplicació més corrent a Biologia. Aquestes tècniques multivariades formen un important capítol de l'anomenada «Geometria de l'Estadística», el qual té per finalitat analitzar i interpretar el material estadístic mitjançant representacions gràfiques.

La informació d'entrada en una anàlisi de dades és una matriu de dades X

	Variables		
	X_1	$X_2 \ldots$	X_n
Individus			
1	x_{11}	$x_{12} \ldots$	x_{1n}
2	x_{21}	$x_{22} \ldots$	x_{2n}
⋮		....	
R	x_{k1}	$x_{k2} \ldots$	x_{kn}

$$X = (x_{ij})$$

on x_{ij} és el valor observat de la variable X_j sobre l'individu i.

Normalment, (X_j) són mesures biomètriques assimilades a variables aleatòries i els individus són k representants d'una espècie, k races geogràfiques o k espècies (etc.). Però de vegades la distinció entre espècies i variables no és prou clara. Els individus poden ésser espècies de zooplàncton i les «variables» els llocs geogràfics on han estat trobades, essent aleshores x_{ij} l'abundància (freqüència) de l'espècie i en la localitat j.

La informació de sortida en una anàlisi de dades és una representació euclídia dels individus en $d = 2$ dimensions que descriu les analogies i diferències entre ells. Les representacions en $d = 3$ dimensions són menys freqüents.

En taxonomia numèrica, la informació de sortida és una representació gràfica anomenada dendrograma que permet establir una classificació jeràrquica entre els individus.

Distàncies estadístiques

Donada la matriu de dades X, cada individu i el podem representar com un punt P_i de l'espai euclidià R^n, de coordenades

$$P_i : (x_{i1}, x_{i2}, \ldots, x_{in})$$

Les distàncies entre aquests punts $P_1,\ldots,P_k$ informen sobre les esmentades analogies i diferències entre els individus. Cal doncs introduir una distància entre els individus. Hi ha moltes maneres de fer-ho. La més corrent és definir la distància² euclídia.

$$d^2(i,j) = d^2(P_i, P_j) = \sum_{h=1}^{n} (x_{ih} - x_{jh})^2$$

Però aquesta distància té alguns inconvenients: pressuposa que les variables són independents i queda alterada per canvis d'escala en les variables. Això no s'esdevé, en canvi, per a la distància de Mahalanobis,

$$D^2(i,j) = (P_i - P_j)' \cdot C^{-1} \cdot (P_i - P_j)$$

on C^{-1} és la inversa de la matriu C de covariàncies entre les variables. D^2 té en compte les correlacions entre les variables i és independent de l'escala de mesura (metres, centímetres, etc.) de cada variable. Quan existeixen relacions lineals entre les variables, es pot prendre la distància:

$$D^2(i, j) = (P_i - P_j)' \cdot C^{-} \cdot (P_i - P_j)$$

on C^- és una g- inversa de C (la inversa de C no existeix), que té les mateixes propietats i no depèn de la g- inversa calculada.

Tant d com D són distàncies que provenen d'un producte intern entre les variables.

En efecte,

$$d^2\ (i,\ j) = s_{ii} + s_{jj} - 2\ s_{ij} \text{ essent } s_{ij} = \sum_{h=1}^{n} x_{ih}\ x_{jh}$$

$$D^2\ (i,j) = P_i' \cdot C \cdot \bar{P}_i + P_j' \cdot C \cdot \bar{P}_j - 2\ P_i' \cdot C \cdot \bar{P}_j$$

és a dir, són distàncies associades a productes interns de matrius I (identitat) i C^- respectivament, entre les variables $X_1,...,X_n$.

Aquestes dues distàncies són les que posen menys problemes de representació en dimensió reduïda. La distància D formulada de diferents maneres, ha estat emprada en genètica. Veure PREVOSTI (1974), OCAÑA (1975).

De vegades té interès utilitzar distàncies del tipus

$$\bar{d}\ (i,j) = \sum_{h=1}^{n} |x_{ih} - x_{jh}|$$

que no provenen d'un producte intern. Aleshores, la representació euclídia de les dades es fa seguint un camí diferent.

ANÀLISI DE COMPONENTS PRINCIPALS

Adoptem ara la distància euclídia d. Si T és una matriu ortogonal d'ordre n x n, la transformació

$$Y = X\ T$$

dóna una nova matriu de dades Y, que defineix una configuració de punts

$$Q_i : (y_{i1},...,y_{in})$$

tal que $d^2\ (P_i,\ P_j) = d^2\ (Q_i,\ Q_j)$.

Els k individus estan aleshores també representats pels punts Q_i. La distància entre dos individus en relació a les d primeres dimensions és

$$d^2\ (Q_i,\ Q_j)_d = \sum_{h=1}^{n} (y_{ih} - y_{jh})^2$$

Doncs bé, l'anàlisi de components principals obté la matriu T que fa que aquestes d dimensions siguin les més rellevants possibles. Es demostra que si T conté els n vectors propis de la matriu de covariàncies C

$$C = T\ D\ T' \qquad D = \text{diag}\ (\lambda_1,\ldots,\lambda_n)$$

ordenats de forma creixent pels valors propis de C, aleshores

$$\sum_{i,j=1}^{n} d^2\ (Q_i, Q_j)_d = \text{màxima}$$

és a dir, la dispersió, mesurada per la suma de distàncies entre els punts, és màxima en dimensió d.

La dispersió global és proporcional a $(\lambda_1 + \ldots + \lambda_n) = \text{Traça}\ (C)$. Si $(\lambda_1 + \ldots + \lambda_d)$ absorbeix una part important (el 75 % per exemple) la representació en dimensió d és adequada.

La matriu de dades $Y = X\ T$

		Components			
		Y_1	Y_2...		Y_n
Individus	1	y_{11}	y_{12}...		y_{1n}
	2	y_{21}	y_{22}...		y_{2n}
	⋮		.....		
	k	y_{k1}	k_{k2}...		y_{kn}

és tal que la variabilitat de les columnes decreix d'esquerra a dreta. Si prenem $d = 2$ agafarem les 2 primeres columnes, o sigui, les coordenades

$$(y_{11},\ y_{12}),\ (y_{21},\ y_{22}),\ \ldots,\ (y_{k1},\ y_{k2})$$

i farem la representació respecte a dos eixos ortogonals (Fig. 1).

A més a més, les noves variables

$$Y_j = t_{j1}\ X_1 + \ldots + t_{jn}\ X_n \quad j = 1, \ldots, n$$

anomenades components principals, verifiquen: 1) estan incorrelacionades dues a dues, 2) tenen variàncies respectivament màximes

$$\text{Var}\ (Y_1) = \lambda_1 \geqq \text{Var}\ (Y_2) = \lambda_2 \geqq \ldots \geqq \text{Var}\ (Y_n) = \lambda_n$$

De vegades les components tenen interessants interpretacions biològiques.

Aquest tipus d'anàlisi es pot fer també diagonalitzant la matriu de correlacions *R*. Els resultats són diferents. Cal utilitzar *C* quan les variables són raonablement comparables (exemple: talles biomètriques mesurades en centímetres). En cas contrari (variables de naturalesa diferent) cal utilitzar R, que significa treballar amb variables sense dimensió física.

L'anàlisi de components principals va ésser introduït per HOTELLING (1933). RAO (1964) va escriure un esclaridor treball sobre la seva utilització i aplicació. S'han fet moltes aplicacions a la Biologia. Vegeu, per exemple, el treball de CHARDY, GLEMAREC i LAUREC (1976) aplicat a l'ordenació de comunitats bentòniques. Els autors comparen els resultats que dóna l'anàlisi de components amb les anàlisis de coordenades principals i de correspondències.

Exemple: En un treball realitzat per ROMERO (1978) es pretèn tipificar diverses comunitats vegetals del massís de Collcerola (Barcelona). Es tenen $n = 14$ formes biològiques, que fan el paper de variables (1 = macrofaneròfits laurifolis, 2 = macrofaneròfits escleròfils, ..., 9 = camèfits suculents, ..., 13 = geòfits, 14 = hidròfits), i $k = 33$ inventaris d'espècies, amb les característiques generals següents: comunitats ruderals (24,25), prats secs (3,8,9,21,32,33), alzinars (1,2,6,12,17,19,29), matolls (15,30,31), diverses comunitats de caducifolis, relacionats amb la presència de microclimas humits (14,18,5,10,26), i carritxars (a la vora d'un estany, 16,17).

El valor x_{ij} que cada variable (forma biològica) fa correspondre a cada individu (inventari) és el nombre d'espècies de la forma biològica trobades a l'inventari que es feia sobre superfícies de 10×10 m^2. Per exemple, a l'inventari 6 (efectuat en un alzinar) vàrem trobar 10 espècies de la forma 3, és a dir, $X_{63} = 10$.

La fig. 1 és la representació per anàlisi de components principals amb dimensió $d = 2$. Els 2 primers eixos expliquen el 66,8 % de la variància total. Les conclusions que es treuen són:

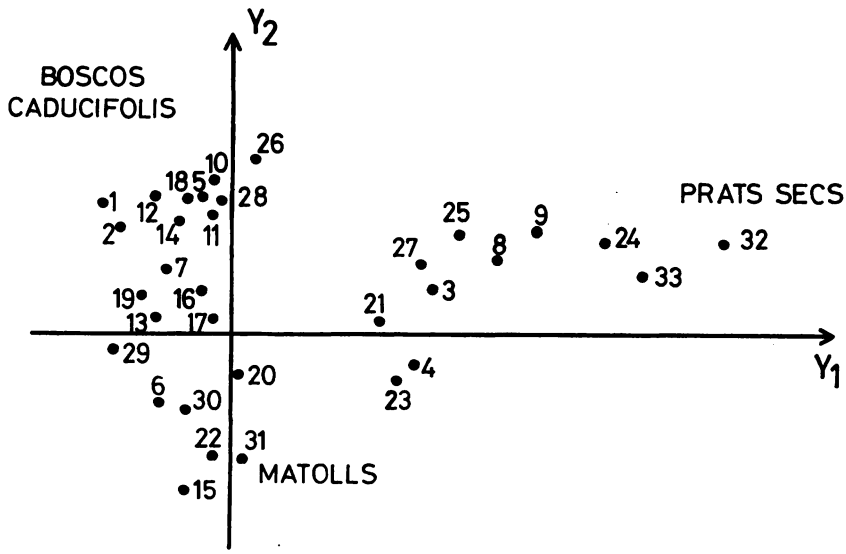

FIG. 1. — Representació per anàlisi de components principals de 33 inventaris d'espècies vegetals del massís de Collcerola (Barcelona)

1) La fisonomia dels inventaris ben definits queda ben reflectida; 2) Les comunitats intermèdies queden ben identificades. Aquests híbrids apareixen també en lloc intermedi de la gràfica; *3*) Observant els alzinars, es veu que 1 i 2 tenen forta influència de caducifolis, mentre que 29 i 6 són alzinars degradats, amb un estrat arburtiu cada cop més important, en detriment de l'estrat arbori.

La interpretació dels eixos s'obté buscant les correlacions entre les variables inicials i les components principals. La correlació entre la variable X_i i la compent Y_j és

$$\rho\ (X_i,\ Y_j) = \frac{t_{ij}}{\sigma_i}\sqrt{\lambda_j}$$

essent σ_i la desviació típica de X_i. Així, destaca la correlació 0,841 entre la primera component i la forma 12 (teròfits: plantes de cicle vital no superior a un any i que travessen l'estació desfavorable en forma de llavor). El primer eix s'interpreta com una dimensió que representa la riquesa de la comunitat en plantes anuals. La segona component correlaciona amb les formes 5 i 6 (nanofaneròfits planifolis i aciculifolis) i el segon eix sembla expressar la importància de l'estrat arbustiu.

Anàlisi de coordenades principals

La tècnica per trobar les components principals és particularment adequada quan les variables són contínues.

Suposem ara que les variables X_i són dicotòmiques, basades en absència (—) o en presència (+) de caràcters qualitatius. Un individu queda aleshores caracteritzat per les presències o les absències del n caràcters.

	X_1	X_2	X_3	...	X_n
i	+	—	+	...	+

La informació útil és aleshores el nombre de caràcters presents sobre els n caràcters estudiats. Per representar els k individus, amb aquests tipus de dades, és molt útil l'anàlisi de coordenades principals introduïda per Gower (1966).

L'associació entre els individus i, j, s'obté de la taula de freqüències

		i: +	i: —
j: +		a	b
j: —		c	d

$$n = a + b + c + d$$

que conté el nombre de caràcters comuns a, el nombre de caràcters no comuns, d, etc. L'associació es mesura mitjançant un coeficient de similaritat s_{ij} que verifica, en general, les propietats

$$0 \leq s_{ij} \leq 1$$

$s_{ij} = 0$ si $c + b = n$ (discrepància total)

$s_{ij} = 1$ si $a + b = n$ (concordància total)

La similaritat s_{ij} dóna el grau de semblança entre i, j, en relació als n caràcters. Exemples de coeficients de similaritat són

SOKAL i MICHENER: $\dfrac{a+d}{n}$

SOKAL i SNEATH: $\dfrac{a}{a+2\,(b+c)}$

JACCARD: $\dfrac{a}{a+b+c}$

RUSSELL i RAO: $\dfrac{a}{n}$

Un cop el biòleg ha escollit el coeficient que millor expressa la similaritat entre els k individus, es forma la matriu d'associacions.

$$S = \begin{pmatrix} s_{11} & \ldots\ldots & s_{1k} \\ s_{k1} & \ldots\ldots & s_{kk} \end{pmatrix}$$

Definim ara la distància[2] entre i, j

$$d^2_{ij} = d^2\ (i, j) = s_{ii} + s_{jj} - 2s_{ij}$$

Quan s_{ij} és una similaritat del tipus descrit, és

$$d^2_{ij} = 2\ (1 - s_{ij})$$

Aquesta distància és adequada ja que

$$d_{ii} = 0 \qquad i = i, \ldots, k$$
$$d_{ij} = 0 \quad \text{si} \quad s_{ij} = 1 \quad \text{(similaritat total entre i, j)}$$
$$d_{ij} = 2 \ \text{(màxima distància) si} \ s_{ij} = 0 \ \text{(similaritat nulla).}$$

El mètode de GOWER consisteix a trobar una matriu de dades

$$\begin{array}{c|c} & \\ 1 & y_{11}\ y_{12}\ \ldots\ y_{1k} \\ 2 & y_{21}\ y_{22}\ \ldots\ y_{2k} \\ \text{Individu} \ \vdots & \ldots\ldots\ldots \\ k & y_{k1}\ y_{k2}\ \ldots\ y_{kk} \end{array} \qquad Y = (y_{ij})$$

que defineixi una configuració de punts en R^k, de manera que la seva distància euclídia coincideixi amb d_{ij}

$$d^2_{ij} = \sum_{h=1}^{k} (y_{ih} - y_{jh})^2$$

Agafant aleshores les dues primeres coordenades, tindrem una representació dels individus en dimensió $d=2$. Per obtenir aquestes coordenades es forma la matriu $T=(t_{ij})$ essent

$$t_{ij} = s_{ij} - \bar{s_i} - \bar{s_j} + s_{ij}$$

on s_i és la mitjana de la i -èsima fila de S, $\bar{s_j}$ és la mitjana de la f -èsima columna, $\bar{s}$ és la mitjana de tots els elements de S. T verifica

$$\text{rang (T)} = \text{rang (S)} - 1$$

i per tant té almenys, un valor propi nul. Les columnes de la matriu Y són els vectors propis de T, calculats de manera que

$$\sum_{h=1}^{k} y^2_{hj} = \lambda_j \qquad \lambda_j \text{ valor propi de T}$$

és a dir, que la seva norma sigui el valor propi corresponent. Si els vectors propis s'ordenen segons l'ordre decreixent dels valors propis, la dispersió de les d primeres columnes és màxima, en el sentit de la secció anterior.

Exemple: CANTON i SANCHO (1976) apliquen aquesta tècnica per a classificar 57 soques del gènere *Pseudomonas* en relació a 39 proves (glucosa, manitol..., producció de fluoresceïna), utilitzant el coeficient de similaritat de Jaccard.

En aquest cas, $k=57$, $n=39$. La representació en dimensió $d=3$ es troba a la Fig. 2. Els autors formen 7 grups de soques, que identifiquen totalment o parcialment, amb diverses espècies (*aeruginosa, mendocina, putrida*, etc.).

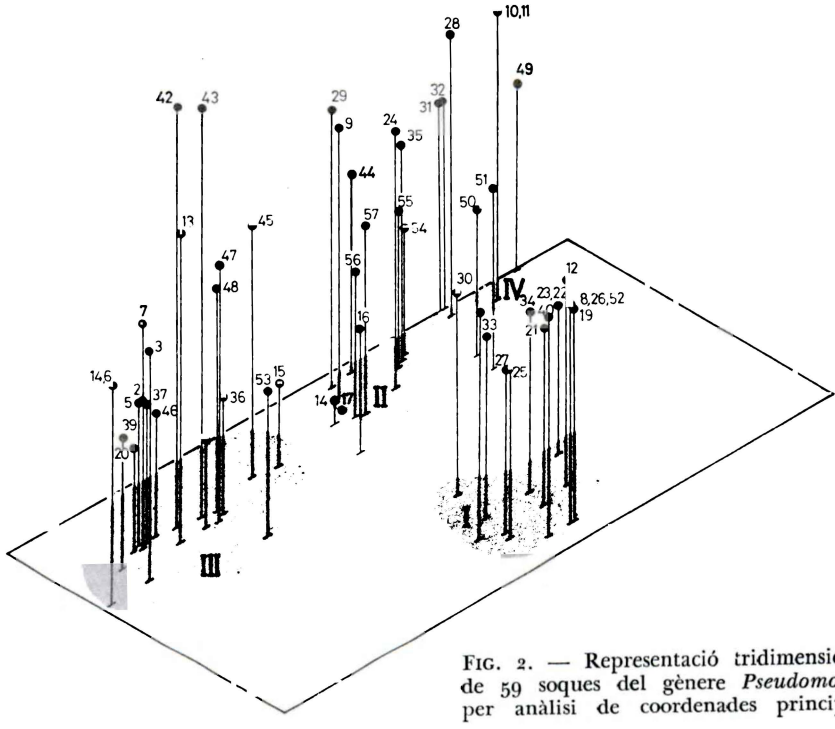

FIG. 2. — Representació tridimensional de 59 soques del gènere *Pseudomonas* per anàlisi de coordenades principals

ANÀLISI DE CORRESPONDÈNCIES

Suposem, tot seguit, que les nostres dades responen a dos criteris de classificació, els quals convencionalment podem anomenar de caràcters i poblacions, i que tenim una taula de contingència de freqüències

		Caràcters				
		A_1	A_2	...	A_n	
Poblacions	1	f_{11}	f_{12}	...	f_{1n}	$f_{1.}$
	2	f_{21}	f_{22}	...	f_{2n}	$f_{2.}$
	⋮	.....				
	k	f_{k1}	f_{k2}	...	f_{kn}	$f_{k.}$
		$f_{.1}$	$f_{.2}$	...	$f_{.n}$	

$$f_{i.} = \sum_h f_{ih} \qquad f_{j.} = \sum_h f_{hj}$$

on f_{ij} és la freqüència de la població *i*, caràcter A_j. Les diferències entre dues poblacions queden reflectades per les diferències entre les distribucions de freqüència dels seus caràcters. Assignem, aleshores, a les poblacions les coordenades

$$\text{Població } i: \quad \left(\frac{f_{i1}}{f_{i.}}, \frac{f_{i2}}{f_{i.}}, \ldots, \frac{f_{in}}{f_{i.}}\right) \quad i = 1, \ldots, k$$

que donen la distribució de probabilitats dels caràcters dins cada població *i*.

Per comparar dues poblacions, es defineix la distància xi-quadrat

$$d^2(i, j) = \sum_{h=1}^{n} \frac{1}{f_{.h}} \left(\frac{f_{ih}}{f_{i.}} - \frac{f_{jh}}{f_{j.}}\right)^2$$

Aquesta distància és també igual a

$$d^2(i, j) = \sum_{h=1}^{n} \left(\frac{f_{ih}}{f_{i.}\sqrt{f_{.h}}} - \frac{f_{jh}}{f_{j.}\sqrt{f_{.h}}}\right)^2$$

i per tant, si formem la nova taula de dades X

	A_1	A_2	...	A_n
1	$\frac{f_{11}}{f_{1.}\sqrt{f_{.1}}}$	$\frac{f_{12}}{f_{1.}\sqrt{f_{.2}}}$	...	$\frac{f_{1n}}{f_{1.}\sqrt{f_{.n}}}$
⋮		. . . . .		
k	$\frac{f_{k1}}{f_{k.}\sqrt{f_{.1}}}$	$\frac{f_{k2}}{f_{k.}\sqrt{f_{.2}}}$	...	$\frac{f_{kn}}{f_{k.}\sqrt{f_{.n}}}$

veiem que la distància Xi-quadrat es una distància euclídia ordinària entre *k* punts de R^n, representats per la taula X.

La distància xi-quadrat reflecteix la diferència de les proporcions relatives dels caràcters a través de cada població. El divisor $f_{.h}$ pondera la presència de petites desviacions dels caràcters de petita freqüència, i permet augmentar la seva importància.

L'anàlisi factorial de correspondències consisteix a aplicar una anàlisi de components principals a la matriu de dades X, tal com expliquem a la secció 3. La nova matriu de dades $Y=XT$ defineix les coordenades de les k poblacions. Agafant les d primeres, tindrem una representació que tindrà màxima dispersió en dimensió d.

Fins aquí, aquest tipus d'anàlisi no és més que aplicar una anàlisi de components principals per a representar les poblacions, però agafant la distància xi-quadrat en lloc de la distància euclídia. L'interès d'aquest mètode resideix en el fet que també podem representar els caràcters amb referència a les seves diferències de distribució respecte a les poblacions. En efecte, es defineix la distància xi-quadrat entre els caràcters A_i, A_j

$$d^2(A_i, A_j) = \sum_{h=1}^{k} \frac{1}{f_{h.}} \left(\frac{f_{hi}}{f_{.i}} - \frac{f_{hj}}{f_{.j}} \right)^2$$

que reflecteix la diferència de les proporcions relatives de les poblacions a través de cada caràcter. De manera anàloga, doncs, per anàlisi de components principals trobarem una matriu de dades Z que ens donarà les coordenades dels n caràcters. Agafant les d primeres, tindrem una representació que tindrà màxima dispersió en dimensió d.

Un altre avantatge és la possibilitat de representar simultàniament caràcters i poblacions, amb referència a uns mateixos eixos de coordenades. En efecte, si les coordenades en dimensió d són

$$\text{Població } i : (y_{i1}, y_{i2}, \ldots, y_{id})$$

$$\text{Caràcter } A_j : (Z_{j1}, Z_{j2}, \ldots, Z_{jd})$$

es demostra la següent relació entre la h -èsima coordenada de la població i, i les h-èsimes coordenades dels n caràcters

$$y_{ih} = \frac{1}{\sqrt{\lambda_h}} \left(\frac{f_{i1}}{f_{i.}} Z_{1h} + \ldots + \frac{f_{in}}{f_{i.}} Z_{nh} \right) \quad i = 1, \ldots, k$$

on λ_h és el valor propi número h. Amb altres paraules, el punt que representa la població i és un terme mitjà dels punts que representen els caràcters, ponderat per les probabilitats de presència en la població i. La proximitat, en la representació simultània, d'una població a un determinat grup de caràcters, indica que predominen en aquesta població.

La representació simultània de caràcters i poblacions havia estat esmentada per KENDALL i estudiada per COOMBS, BENNETT i HAYS, entre els

anys 1950 i 1960. Però no arribaria a ésser popular fins que BENZECRÍ va introduir l'anàlisi de correspondències l'any 1963.

Exemple: ALONSO (1975) fa un ampli estudi de la distribució geogràfica del polimorfisme cromosòmic de *Drosophila subobscura* utilitzant anàlisi factorial de correspondències. Sobre la taula de freqüències de 66 poblacions i 8+3+7+23+11 ordenacions dels 5 cromosomes A, J, E, U i O respectivament, fa una anàlisi de correspondències global, i diverses anàlisis parcials. Utilitzarem com ilHustració una de les anàlisis parcials, concretament la que agafa 13 poblacions i 3 inversions del cromosoma A. Les dades es donen en forma de percentatge a la taula 1.

TAULA 1. — Percentatges de freqüència de tres inversions del cromosoma A per a 13 poblacions de *Drosophila subobscura*

	Inversions		
Poblacions	A-ST	A-1	A-1
HELSINKI	96.0	4.0	0.0
DROBACK	78.4	16,2	5.4
HERIOT	100.0	0.0	0.0
DALKEITH	100.0	0.0	0.0
GRONINGEN	80.0	16.0	4.0
FONTAINEBLEAU . . .	88.5	7.7	3.8
VIENA	56.9	35.8	7.4
ZURICH	67.8	24.4	7.8
FRUSKA-GORA (YUG.) . .	36.1	55.6	8.3
LAGRASSE	72.5	17.5	10.0
MONTPELLIER	60.2	24.3	15.5
CARASCO	50.0	31.8	18.2

L'anàlisi s'ha de fer, naturalment, sobre les freqüències originals. El resultat és la fig. 3.

Observem que Heriot i Dalkeith queden representades en un mateix punt, donada la seva distribució idèntica. També queda reflectida la similaritat entre Drobak i Fontainebleau. La població Frusca-Gora és la que queda més al marge, degut a la influència de l'ordenació A-1, menys freqüent a les altres poblacions. Les proporcions de les 3 ordenacions queden molt ben reflectides a la gràfica; es veu que les poblacions queden més pròximes de les ordenacions que s'hi presenten més.

Comparacions entre les anàlisis de components principals, coordenades principals i correspondències.

En aquesta secció es comenta, prenent com exemple situacions concretes, quin és el mètode més adequat per a representar unes dades determinades.

El punt fonamental que cal tenir ben present és que la representació gràfica dels objectes (individus, espècies, etc.) no és més que una imatge

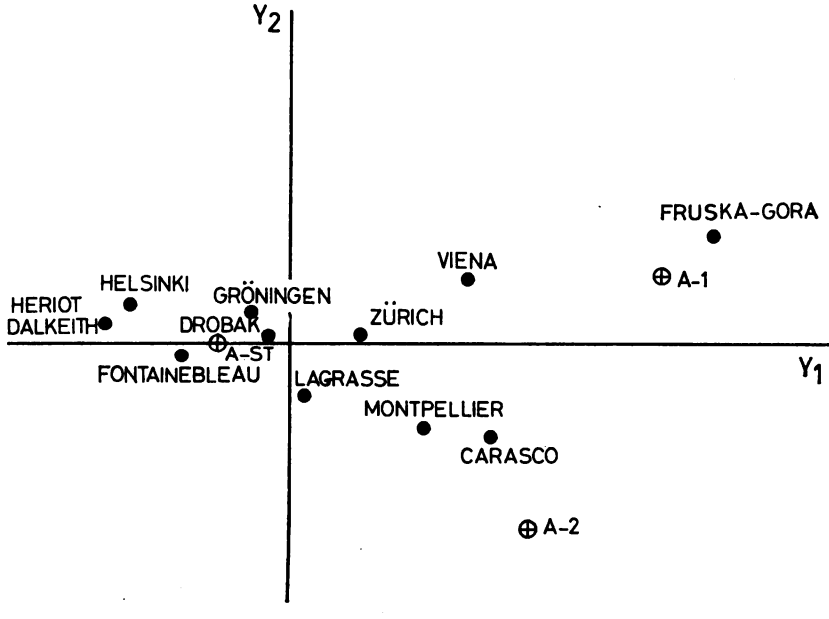

FIG. 3. — Anàlisi de correspondències sobre 13 poblacions europees de *Drosophila subobscura*, en relació a tres inversions del cromosoma A. Les inversions hi són també representades; reflecteixen la incidència que tenen a cada població

de la distància que definim entre objectes. Distàncies diferents donaran lloc a imatges diferents.

Suposem que volem comparar diverses races humanes amb referència a la distribució dels grups sanguinis. Si les freqüències per dues races són

	O	A	B	AB
Raça i	58	56	11	5
Raça j	68	64	12	6

la distància euclídica és 15,3, mentre que la distància xi-quadrat val O. Aleshores, si apliquem l'anàlisi de components principals, les dues races sortiran més o menys diferenciades. Però, en canvi, obtindrem un mateix punt utilitzant l'anàlisi de correspondències. La diferència consisteix que en el primer cas es té en compte el nombre d'individus (130 i 150), és a dir, la grandària de la mostra influencia la discriminació. En el segon cas es compara la distribució de percentatges dels grups sanguinis, que és idèntica per a cada raça. En general, quan les dades s'assimilen a una taula de contingència o classificació doble de dos classes de característiques, és més apropiat l'anàlisi de correspondències. L'exemple donat a la secció anterior (poblacions X ordenacions cromosòmiques) és una correcta ilustració d'això. L'anàlisi de components seria en canvi menys adequada perquè reflectiria el nombre de mosques de les poblacions, que és arbitrari ja que depèn en general del material experimental que aporten els diferents investigadors que han estudiat cada població.

Cal utilitzar l'anàlisi de components quan es volen comparar individus o espècies amb referència a caràcters biomètrics quantitatius (llargada de cos, extremitats, pes, etc.). També és aconsellable aquesta anàlisi quan es volen comparar espècies respecte al nombre d'individus trobats en quadrats o zones de la mateixa extensió, obtinguts, de les mostres preses d'una gran extensió. Per exemple: nombre d'individus de cada una de k espècies de fitoplàncton trobats en n quadrats en el transcurs d'una campanya oceanogràfica. Si l'abundància de cada espècie és una dada important, l'anàlisi de components serà més adequat que el de correspondències. La primera component reflectirà, en general, la grandària o abundància de les espècies. Per poder donar aquesta interpretació cal que la primera component correlacioni positivament amb tots els quadrats (que fan el paper de variables). La forma de les poblacions es representa prenent la segona i tercera components.

L'exemple de representació d'inventaris de la secció 3 és una aplicació correcta d'anàlisi de components. Però en aquest cas la primera component no es podrà interpretar com un factor de grandària, ja que correlaciona negativament amb algunes formes. (Tinguem en compte que en l'exemple es representen els inventaris i no les formes.)

És important observar, també, que l'anàlisi de correspondències, donat el pes que es dóna a cada caràcter (o a cada quantitat), tendeix a destacar les poblacions que tenen una presència més singular.

Hi ha una altra manera de representar poblacions, fent una interpretació factorial de l'anàlisi de components principals. Suposem, per exemple, que tenim n espècies E_1, ..., E_n i prenem com a dades el nom-

bre d'individus per espècie trobats en N mostres. Sigui R la matriu de correlacions entre les espècies i sigui

$$R = T\ D\ T' = A \cdot A' \qquad A = D^{1/2}$$

(T = matriu ortogonal amb els vectors propis, D = matriu diagonal amb els valors propis) una descomposició factorial de R. La matriu factorial $A = (a_{ij})$

	C_1	C_2	...	C_n
E_1	a_{11}	a_{12}	...	a_{1n}
E_2	a_{21}	a_{22}	...	a_{2n}
⋮		. . . . .		
E_n	a_{n1}	a_{n2}	...	a_{nn}

dóna les saturacions entre les espècies (E_i) i les components principals (C_i). La saturació a_{ij} coincideix amb la correlació entre E_i i C_j. Aleshores es representen les poblacions agafant, com a coordenades de cada espècie, les fileres de A

$$E_i: \quad (a_{i1},\ a_{i2},\ \ldots,\ a_{in}) \qquad i = 1,\ \ldots,\ n$$

Estudiem ara la distància² entre espècies:

$$d^2(E_i, E_j) = \sum_{h=1}^{n} (a_{ih} - a_{jh})^2 = \sum_{h=1}^{n} a^2_{ih} + \sum_{j=1}^{n} a^2_{jh} - 2 \sum_{h=1}^{n} a_{ih} a_{jh}$$

segons el teorema de THURSTONE, la correlació entre E_i, E_j és

$$\gamma_{ij} = \sum_{h=1}^{n} a_{ih}\ a_{jh}$$

i per tant,

$$d^2(E_i, E_j) = 2\ (1 - \gamma_{ij})$$

que coincideix amb la distància de GOWER construïda a partir d'un coeficient d'associació. Podem afirmar: l'anàlisi de components principals

que representan les espècies que utilitzen la matriu factorial, dóna les mateixes distàncies que l'anàlisi de coordenades principals.

Però, si bé les representacions amb totes les dimensions coincideixen en els dos mètodes, les representacions amb dimensió reduïda no donen el mateix resultat. Per exemple, sigui la matriu de correlacions

$$R = \begin{pmatrix} 1 & 0.9 & 0 \\ 0.9 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Els resultats, aplicant components principals i coordenades principals són, respectivament,

$$A = \begin{pmatrix} 0{,}975 & 0 & 0{,}22 \\ 0{,}975 & 0 & -0{,}22 \\ 0 & 1 & 0 \end{pmatrix} \qquad Y = \begin{pmatrix} 1{,}18 & 0{,}31 \\ 1{,}18 & -0{,}31 \\ -2{,}36 & 0 \end{pmatrix}$$

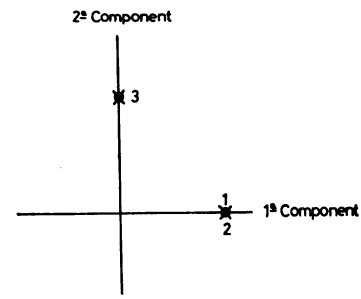

FIG. 4. — Representació de tres espècies per anàlisi de components sobre la matriu de correlacions

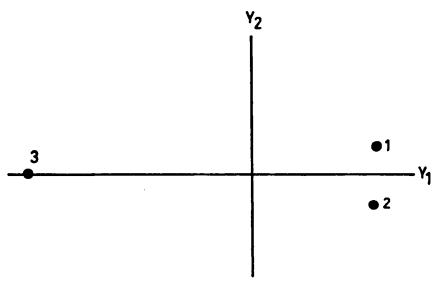

FIG. 5. — Representació de les mateixes espècies per anàlisis de coordenades principals

Les figures 4 i 5 són les corresponents representacions en dimensió 2.

De la Fig. 4 podem concloure que les espècies 1 i 2 estan molt lligades a la primera component i, per tant, molt relacionades entre sí;

l'espècie 3 defineix íntegrament la segona component. De la Fig. 5 podem concloure que les espècies 1 i 2 són molt pròximes entre sí, però allunyades de l'espècie 3. La Fig. 6 relaciona aquestes dues representacions.

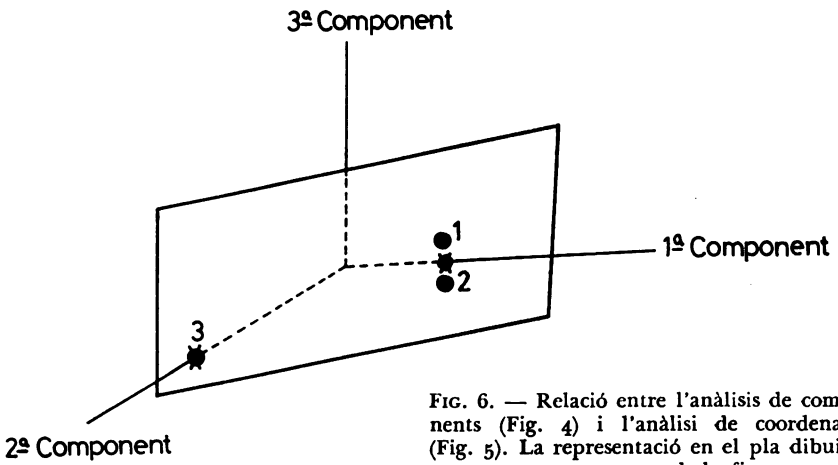

FIG. 6. — Relació entre l'anàlisis de components (Fig. 4) i l'anàlisi de coordenades (Fig. 5). La representació en el pla dibuixat es correspon amb la fig. 5

L'anàlisi de components tendirà a agrupar en direcció a un eix les variables fortament correlacionades amb una component, i en direcció oposada les variables amb correlació negativa. Unes variables poc correlacionades amb unes altres (3 respecte a 1 i 2, per exemple) quedaran representades ortogonalment. Com que la distància està basada en la correlació, no influeix la grandària (nombre d'individus) de l'espècie. Aquesta representació s'ha de interpretar segons els criteris de l'anàlisi factorial, que agrupen les variables (espècies) segons la seva relació amb uns determinats factors (per exemple: conjunts d'aigua definits per llur salinitat, temperatura, etc., en el cas de fitoplàncton), de la mateixa manera que les manifestacions de la personalitat es relacionen amb factors (neuroticisme, estabilitat, introversió, extroversió), fent un símil amb la Psicologia.

Encara que la distància global és la mateixa, l'anàlisi de coordenades representa distàncies amb màxima resolució en dimensió reduïda, en comptes de representar direccions. Les distàncies oscil·laran entre O (correlació 1) 2 (correlació O), 4 (correlació -1). Si el criteri d'aquesta anàlisi és maximitzar la suma de quadrats de les distàncies, el criteri de l'anàlisi de components és maximitzar la variància explicada per cada component. Les representacions en dimensió reduïda seran, en general, sem-

blants. De fet, l'anàlisi de coordenades principals ens aporta novetat metodològica quan la comparació entre espècies la fem sobre una matriu de similaritats calculada sobre variables dicotòmiques, ja que una anàlisi de components sobre aquesta matriu (que no és de correlacions) seria poc adequada.

Com a aplicació d'aquesta versió de l'anàlisi de components, vegeu ESTRADA (1975).

Estudiem ara la representació de dades de presència i d'absència de caràcters. Com hem vist a la secció 4, l'anàlisi de coordenades principals ens permet representar objectes en referència a una distància deduïda d'un coeficient de similaritat. Si codifiquem les dades en 0 (absència), en 1 (presència), la distància² euclídia és

$$d^2_{ij} = b + c$$

Si escollim el coeficient de similaritat de SOKAL i MICHENER, la distància² deduïda d'aquest coeficient és

$$d^2_{ij} = 2\left(1 - \frac{a+d}{n}\right) = \frac{2}{n}(b + c)$$

Tenim, en conseqüència, que l'anàlisi de components principals amb variables dicotòmiques és equivalent a una anàlisi de coordenades principals utilitzant el coeficient de SOKAL i MICHENER, llevat de la constant $2/n$, que no l'afecta, perquè tota representació de distàncies és relativa.

L'anàlisi de correspondències dóna també el mateix resultat si cada caràcter es presenta el mateix nombre t de vegades, perquè és fàcil veure que la distància xi-quadrat és

$$d^2\ (i,\ j) = \frac{1}{tn^2}(b + c)$$

Quan el nombre de presències del caràcter és variable, hi ha un factor de ponderació que, com ja hem esmentat, tendeix a destacar els objectes que tenen unes característiques que la majoria no tenen. Aleshores surten en llocs extrems de la gràfica.

Per concloure aquesta secció, tornem a repetir que el resultat de la representació gràfica depèn de la distància que definim, si bé és possible, en alguns casos, obtenir resultats semblants o àdhuc idèntics. Vegeu amb profit l'esmentat treball metodològic de CHARDY, GLEMAREC i LAUREC

(1976) per ampliar aquesta comparació dels tres mètodes de representació de dades en Biologia.

Taxonomia numèrica

Donar una classificació ordenada dels fenòmens naturals és una de les tasques fonamentals de la ciència. Només fent una bona classificació és possible establir relacions entre la gran varietat de resultats de l'observació científica.

Tots els biòlegs coneixen molt bé el sistema de classificació sistemàtica de C. Linneus. En aquest sistema cada ser vivent té assignat un nom llatí amb el gènere i l'espècie.

El sistema taxonòmic de Linneus es pot descriure com una jerarquia organitzada a nivells, on les classes disjuntes a cada nivell constitueixen les anomenades *taxes*. A un nivell donat, les taxes constitueixen les categories. Així es parla de les categories «espècies», «gèneres», «famílies», «ordres», «classes», etc. La categoria gènere, per exemple, té diverses taxes: els gèneres corresponents a una família donada.

La taxonomia numèrica és un intent de construir classificacions naturals sobre la base de la semblança fenotípica entre els individus. Aquesta semblança fenotípica es mesura agafant un conjunt de caràcters significatius i calculant una matriu de similaritats S, per algun dels procediments explicats a la secció anterior. Però en taxonomia numèrica és preferible treballar amb dissimilaritats. Una dissimilaritat és una distància que en general no verifica la propietat triangular. S'obté una dissimilaritat només posant

$$d_{ij} = 1 - s_{ij}$$

Però hi ha també altres maneres de definir dissimilaritats.

La dissimilaritat mesura les diferències fenotípiques entre dos individus o dues classes. Doncs bé, conegudes les dissimilaritats entre k individus, el resultat final d'una taxonomia numèrica és una jerarquia indexada o dendrograma, que és una representació gràfica de la classificació.

Els termes família, gènere, espècie, etc., tenen aleshores un significat més precís, ja que es parla de d-taxes o classes amb distància fenètica. Per exemple, la similitud entre les espècies 2 i 3 (d=0,3) és més gran que entre les espècies 4 i 5 (d=0,5). A partir de d=0,6 es parla de gèneres i a partir de d=0,8 es parla de famílies. Hi ha dues famílies (1,2,3)+(4,5)

i tres gèneres (1)+(2,3)+(4,5). La distància fenètica entre el gènere (1) i el gènere (2,3) és 0,8, etc. La distància fenètica *d* és l'índex de la jerarquia, com ara explicarem.

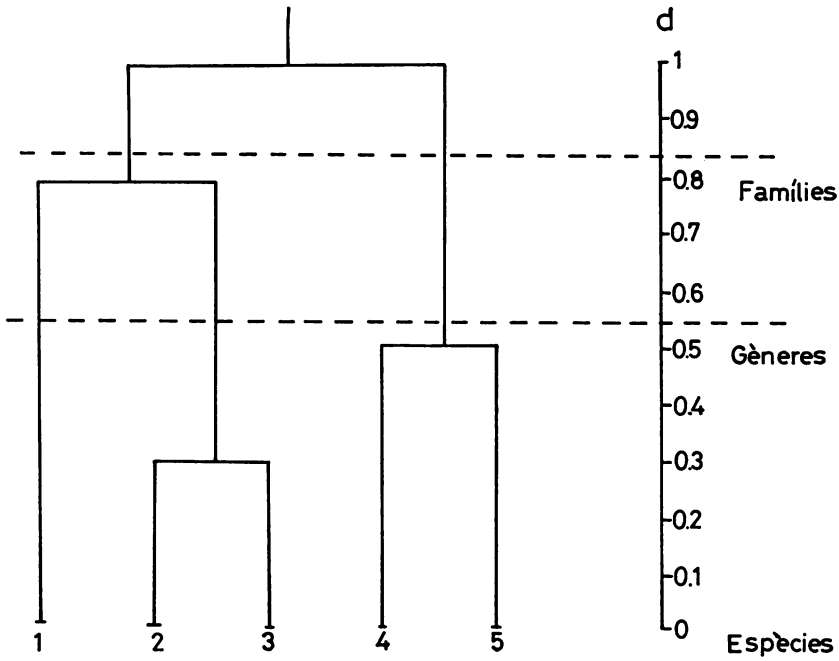

FIG. 7. — Exemple hipotètic de dendrograma; *d*=distància fenètica

Passem ara a donar els conceptes generals. Sigui E = (1, 2, ..., k) un conjunt d'objectes (individus, espècies, etc.). Es diu que H, conjunt de parts de E, és una jerarquia, si:

1) Per tot h, h' ε H o bé h C h', o h' C h o h $\cap$ h' = $\varnothing$
2) Tot h és reunió dels h' inclosos en h
3) {i} ε H per tot i ε E; E ε H

Els elements de H s'anomenen classes.

Per exemple:

$$H = \{ \{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{2,3\}, \{4,5\}, \{1,2,3\}, \{1,2,3,4,5\} \}$$

és una jerarquia sobre E = {1,2,3,4,5}, representada en la fig. 7.

Una jerarquia H es diu que és indexada si existeix una aplicació d: $H \rightarrow R^+$ verificant

$$d(\{i\}) = 0 \qquad \text{per tot } i \in E$$
$$h \subset h' \text{ implica } d(h) < d(h')$$

essent la inclusió estricta.

Per exemple $d(\{2,3\}) = 0{,}3$, $d(\{4,5\}) = 0{,}5$, $d(\{1,2,3,\}) = 0{,}8$, etc., és un índex sobre H que coincideix amb l'esmentada distància fenètica.

La representació d'una jerarquia indexada és el dendrograma, que permet establir una classificació jeràrquica entre els objectes de E.

Per trobar una tal jerarquia cal construir sobre E un tipus especial de dissimilaritat anomenada ultramètrica. Una ultramètrica verifica el següent axioma

$$u(i,j) \leq \sup \{u(i,t), u(j,t)\} \qquad \text{per tot } i,j,t \in E$$

Es demostra fàcilment que aquest axioma és equivalent al fet que tot triangle (conjunt format per tres objectes) és isòsceles, i la base és el costat més petit.

Doncs, bé, coneguda una ultramètrica *u*, es pot construir una jerarquia indexada i recíprocament. La jerarquia *H* donat *u*, s'obté per un algorisme que va ajuntant els objectes més pròxims per la dissimilaritat *u*. Recíprocament, es construeix *u* a partir d'una jerarquia indexada posant

$$u(i, j) = d_h$$

si h és la classe més petita que conté i,j. Per exemple, de la fig. 3 tindríem: $u(2,3) = 0.3$, $u(1,3) = 0.8$, $u(1.2) = 0{,}8$. Fixem-nos que $\{1,2,3\}$ és un triangle isòsceles amb base $\{2,3\}$.

Així doncs, podrem construir una classificació jeràrquica si tenim una dissimilaritat ultramètrica definida sobre E. Aquest seria l'ideal del taxonomista. Però, per desgràcia, només trindrem, en general, una matriu de dissimilaritats $\Delta = (\delta_{ij})$ on δ_{ij} (que moltes vegades es calcula a partir d'una similaritat s_{ij}) que no és pas una ultramètrica. Un algorisme de classificació consisteix en obtenir una distància ultramètrica partint d'una dissimilaritat δ_{ij}. Aleshores la distància ultramètrica es representa mitjançant un dendrograma.

$$\Delta = \begin{pmatrix} \delta_{11} & \dots & \delta_{1k} \\ & \dots & \\ \delta_{k1} & \dots & \delta_{kk} \end{pmatrix} \xrightarrow{\textit{Algorisme}\ U} \begin{pmatrix} u_{11} & .. & u_{1k} \\ & \dots & \\ u_{k1} & \dots & u_{kk} \end{pmatrix} \longleftrightarrow \text{Dendrograma}$$

Matriu dissimilaritats Matriu ultramètrica

Cal que la ultramèrica no deformi gaire la dissimilaritat inicial. Vegem alguns d'aquests algorismes.

JOHNSON (1967) proposa obtenir la més gran de totes les ultramètriques inferiors a

$$u_{iM}\ (i,j) = \sup\{u\ (i,j) \ / \ u\ (i,j) \leq \delta_{ij}\}$$

u_{iM} és la ultramètrica inferior màxima. Equival a deformar tot triangle (i,j,t) prenent com a base el costat més petit i com a longitud dels costats iguals (tot triangle és isòsceles) la del següent costat més petit. La ultramètrica u_{iM} és única i és sistemàticament inferior a δ. Com algorisme de classificació, havia estat introduït per Sneath l'any 1957.

JOHNSON (1967) proposa també la més petita de totes les ultramètriques superiors a δ

$$u_{Sm}\ (i,j) = \inf\{u\ (i,j) \ / \ u\ (i,j) \geqq \delta_{ij}\}$$

u_{Sm} és una ultramètrica superior mínima i n'hi poden haver diverses. Aquesta ultramètrica és sistemàticament superior a δ. El corresponent algorisme de classificació havia estat introduït per SÖRENSEN l'any 1948.

Les dos ultramètriques de JOHNSON són inferiors i superiors respectivament a la dissimilaritat inicial. Altres autors han proposat ultramètriques que siguin un terme mitjà entre aquestes dues. Vegeu ROHLF (1970).

El UPGMA (*Unweighted pair group method using method averages*) de SOKAL i SNEATH (1963) és un dels més utilitzats.

La taxonomia numèrica es va desenvolupar a partir del 1957 amb la publicació de diversos articles, en defensa d'aquest mètode, escrits pel microbiòleg britànic P. H. SNEATH i els entomòlegs americans C. D. MICHENER i R. K. SOKAL. Però, sobretot, es va desenvolupar després de l'obra *Principles of Numerical Taxonomy*, publicada l'any 1963, escrita per

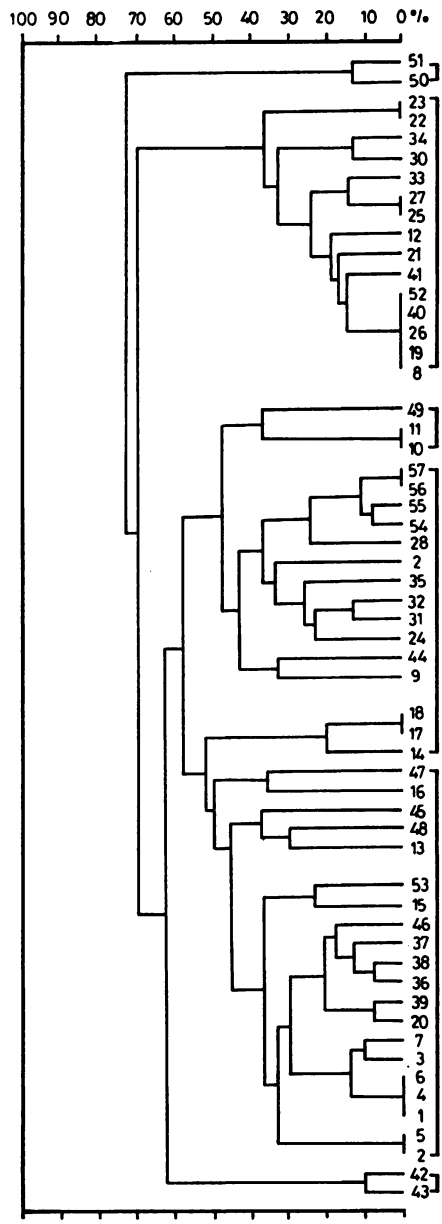

FIG. 8. — Dendrograma de 59 soques del gènere *Pseudomonas* obtingut pel algorisme UPGMA. L'índex és la ditància fenètica, que té les propietats de distància ultramètrica

SOKAL i SNEATH, on s'exposa l'estudi teòric de la classificació, incloent les seves bases, principis, procediments, i regles.

Exemple: Continuant l'exemple de la secció 4, E. CANTON i J. SANCHO (1976) realitzen també una taxonomia numèrica de les 59 soques del gènere *Pseudomonas*. El coeficient de similaritat de Jaccard el transformen en dissimilaritat posant

$$\delta_{ij} = 100\ (1 - s_{ij})$$

Aquesta dissimilaritat oscilla entre 0 (soques fenotípicament idèntiques) i 100 (soques completament oposades). El resultat d'aplicar el algorisme UPGMA a la matriu de dissimilaritats és el dendrograma de la fig. 8. Els autors comenten aleshores, els diferents grups trobats. (En realitat els autors treballen amb l'índex $100.s_{ij}$, que és complementari de δ_{ij}; el resultat és idèntic en interpretació.)

Per estudiar el grau de distorsió de la distància ultramèrica respecte la dissimilaritat inicial, es correlacionen els $k\ (k-1)/2$ parells de distàncies (δ_{ij}, u_{ij}). El coeficient de correlació obtingut s'anomena correlació cofenètica. Els autors troben una correlació cofenètica de 0.96. El dendograma interpreta molt bé les dissimilaritats inicials.

ANÀLISI DE PROXIMITATS

Donats k objectes (individus, espècies, estímuls) i coneguda una determinada informació Δ sobre les diferències entre els objectes, l'anàlisi de proximitats («multidimensional scaling» en anglès) és una tècnica que obté una configuració euclídia, formada per k punts $P_1, \ldots, P_k$, de manera que les seves distàncies euclídies concordin amb la informació Δ.

Aquesta informació pot venir donada de diferents maneres: ordenació entre les distàncies dels objectes, matriu de dissimilaritats o distàncies no euclídies, distàncies afectades d'errors aleatoris o d'aproximació numèrica, etc.

SHEPARD (1962 a,b) va formular la següent conjectura: Coneguda només l'ordenació entre les distàncies dels objectes

$$d\ (i_1, i_2) \leq d\ (i_3, i_4) \leq \ldots \leq d\ (i_{m-1}, i_m) \qquad (m = k\ (k-1)/2)$$

existeix una configuració euclídia tal que les seves distàncies verifiquen aproximadament aquesta ordenació.

SHEPARD no va donar una prova rigorosa de la seva conjectura, però la va ilustrar amb nombrosos exemples i va formular un algorisme nu-

mèric que trobava aquesta configuració. Altres autors varen afirmar i desenvolupar la idea de SHEPARD, proposant mètodes de construcció (KRUSKAL, GUTTMAN, LINGOES, YOUNG, BENCECRI, CARROLL, etc.).

Vegem-ne tres exemples. Siguin quatre objectes A,B,C,D, tals que les ordenacions entre les seves distàncies desconegudes siguin (poden haver-hi ranges iguals):

	A	B	C	D
A	0	1	2	3
B		0	1	2
C			0	1
D				0

(1)

	A	B	C	D
A	0	1	1	1
B		0	1	1
C			0	0
D				0

(2)

	A	B	C	D
A	0	1	1	1
B		0	1	1
C			0	1
D				0

(3)

Aleshores estarien representades per configuracions euclídies en 1, 2 i 3 dimensions respectivament.

L'interès del mètode de SHEPARD és trobar una configuració en un espai de dimensió reduïda que reprodueixi aproximadament l'ordenació inicial. Aleshores podrem representar els k objectes gràficament, conservant les relacions de proximitat.

El sol coneixement de l'ordenació és el cas d'informació més pobra. Normalment es coneix una matriu de distàncies no euclídies o dissimilaritats $\Delta = (\delta_{ij})$. El resultat final d'una anàlisi de proximitats és una matriu de dades Y,

	1	2	...	d
1	y^{11}	y^{12}	...	y^{1d}
2	y^{21}	y^{22}	...	y^{d2}
⋮	. . . . .			
k	y^{k1}	y^{k2}	...	y_{kd}

P_i: $(y_{i1}, \ldots, y_{id})$

que defineix k punts de R^d tals que les seves distàncies[2]

$$d^2(P_i, P_j) = d^2_{ij} = \sum_{h=1}^{d} (y_{ih} - y_{jh})^2$$

siguin semblants a les dissimilaritats s_{ij}. La matriu Y es pot calcular trobant primer la matriu $S=(s_{ij})$ essent

$$s_{ij} = -\frac{1}{2}(\delta^2_{ij})$$

i diagonalitzant a continuació la matriu T, pel procediment explicat en l'anàlisi de coordenades principals (vegeu GOWER, 1966). Però la matriu

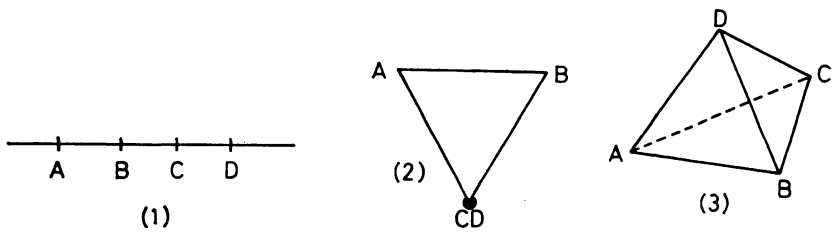

FIG. 9. — Representació mètrica en 1, 2 i 3 dimensions de 4 objectes, coneguts només els rangs (1), (2) i (3) entre les seves proximitats

T pot tenir valors propis negatius i obtenir, doncs, una configuració poc concordant amb $\Delta=(\delta_{ij})$.

El mètode de KRUSKAL (1964) amplia el mètode de SHEPARD i dóna solucions a aquest problema. Consisteix a construir una nova dissimilaritat, anomenada disparitat, que sigui una funció monòtona creixent de δ_{ij}, és a dir,

$$d_{ij} = f(\delta_{ij}) \qquad \delta_{ij} \leq \delta_{i'j'} \rightarrow \hat{d}_{ij} \leq \hat{d}_{i'j'}$$

Aquesta transformació de δ_{ij} conserva l'ordenació inicial. Utilitzant d_{ij} es troba una matriu de dades Y (pel procediment indicat) que dóna unes distàncies euclídies d_{ij}. Com a mesura d'ajust entre d_i i d_{ij}, Kruskal defineix la quantitat

$$S = \left(\frac{\sum_{i<j} (d_{ij} - \hat{d}_{ij})^2}{\sum_{i<j} d^2_{ij}} \right)^{1/2}$$

anomenada *stress*. Verifica $0 \leq S \leq 1$, però es dóna en forma de percentatge. KRUSKAL desenvolupa un mètode adient per trobar d_{ij}, Y i $\hat{d}_{ij}$ de manera que quedi minimitzat el *stress*. Considera que la representació és bona si no supera el 5 %.

Agafant les coordenades que conté Y, tindrem una representació dels objectes que conserva aproximadament les relacions de proximitat inicials, malgrat que la distància euclídia sigui una deformació de la dissimilaritat inicial.

El *stress* disminueix a mida que augmenta la dimensió *d*. Cal agafar *d* que doni un *stress* acceptable. Cal també calcular el coeficient de correlació ordinal de KENDALL entre les distàncies euclídies i les dissimilaritats originals, per mesurar el grau de concordància entre l'ordenació inicial i l'ordenació entre les distàncies dels punts representats.

Exemple: PREVOSTI (1974) proposa una distància genètica basada en les diferències d'ordenacions cromosòmiques. La distància entre dues poblacions és

$$D\ (1,2) = \frac{1}{2r} \sum_{j=1}^{r} \sum_{h=1}^{s_j} \left| p^1_{jh} - p^2_{jh} \right|$$

on r és el nombre de diferents cromosomes, s_j és el nombre de diferents ordenacions en el cromosoma *j* (ordenacions dels seus gens), p_{1jh} i p_{2jh} són les freqüències de l'ordenació h del cromosoma j en les poblacions 1 i 2 respectivament. Aquesta distància verifica $0 \leq D\ (1,2) \leq 1$. Els seus avantatges com a distància genètica han estat discutits per l'autor.

Però aquesta distància no prové d'un producte escalar. Una configuració euclídia entre poblacions, que reprodueixi les distàncies originals, es pot obtenir per anàlisi de proximitats.

Per a il·lustrar-ho, farem servir part de les dades del treball de PREVOSTI, OCAÑA i ALONSO (1975), que aplica aquesta distància per estudiar el polimorfisme cromosòmic de *Drosophila subobscura*. Les distàncies genètiques entre k = 6 poblacions del nord i centre d'Europa són:

	1 Dröbak	2 Dalkeith	3 Gröningen	4 Fontainebleau	5 Viena	5 Zürich
1	0	0,307	0,152	0,271	0,260	0,235
2		0	0,276	0,225	0,370	0,300
3			0	0,150	0,187	0,112
4				0	0,195	0,120
5					0	0,128
6						0

Realitzant una anàlisi de proximitats amb dimensió d=2, per un algorisme iteratiu desenvolupat per CUADRAS (1979), s'obté que la funció monòtona creixent

$$f(\delta_{ij}) = d_{ij} = 1.062\ \delta_{ij} - 0.04 \qquad (\delta_{ij} = D(i,j))$$

dóna un *stress* del 4,4 %. La fig. 10 és una representació de la configuració euclídia obtinguda.

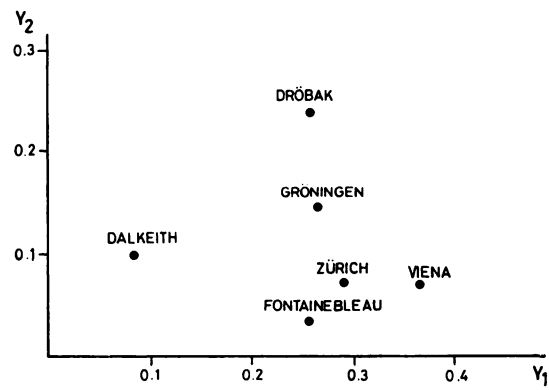

FIG. 10. — Representació en dues dimensions de 6 poblacions de *Drosophila subobscura* per anàlisi de proximitats, amb un «stress» del 4,4 %. Les distàncies representades es diferencien molt poc de la distància genètica inicial. Compareu amb la fig. 3

L'evolució dels *stress* en diferents dimensions va ésser:

Dimensió:	1	2	3	4	5
Stress:	40,0	4,4	1,27	0,24	0

El coeficient de correlació ordinal de KENDALL entre les distàncies dels punts de la fig. 2 i les distàncies originals va donar 0,9929. Les ordenacions entre ambdós conjunts de distàncies pràcticament coincideixen.

Cal observar que la matriu T calculada pels 6 punts no tenia cap valor propi negatiu. Això vol dir que existeix una configuració euclídia que reprodueix exactament les distàncies originals. Aquesta configuració que té dimensió 5, es pot obtenir per anàlisi de coordenades principals, amb resultats similars als obtinguts. La utilitat de l'anàlisi de proximitats es fa patent quan agafem moltes poblacions. Aleshores es presenta una influència notable dels valors propis negatius, i no existeix cap configuració euclídia, a menys que deformem monotònicament les distàncies inicials.

Anàlisi canònica de poblacions

Malgrat que la distància D de Mahalanobis és la més perfecta per a representar dades, encara no l'hem utilitzada. El problema que significa utilitzar D resideix en el fet que, generalment, no es coneix la matriu de covariàncies C entre les variables. Si poguéssim conèixer C amb independència de la matriu de covariàncies de la taula de dades X, podríem representar els k individus aplicant la distància D, amb resultats més escaients que l'anàlisi de components principals, la qual utilitza la distància euclídia ordinària.

En l'anàlisi canònica de poblacions és possible fer-ho. Suposem que cada espècie (o raça, o grup, etc.), està representat per més d'un individu, i que la taula de dades és (k poblacions i p variables)

Variables

	X_1	X_2	...	X_p	
Població 1	X_{111}	X_{121}	...	X_{1p1}	
	. . . . .				n_1
	X_{11n_1}	X_{12n_1}	...	X_{1pn_1}	
⋮		. . . . .			
	X_{k11}	X_{k21}	...	X_{kp1}	
Població k		. . . . .			n_k
	X_{k1n_k}	X_{k2n_k}	...	X_{kpn_k}	

on n_t és el nombre d'individus de la població t. Prenem com a representant de la població t l'individu mitjà de coordenades

$$M_t: \ \bar{x}_{t1}, \ \bar{x}_{t2}, \ \ldots, \ \bar{x}_{tp}) \qquad t = 1,2, \ \ldots, \ k$$

on x_{ti} és la mitjana de la variable i dins de la població t. Aleshores podem estimar la matriu de covariàncies calculant

$$C = \frac{1}{N-k} \sum_{t=1}^{k} n_t S_t \qquad (N = \sum_{t=1}^{k} n_t)$$

on S_t és la matriu de covariàncies dins la població t,

$$S_t = (s_{ij}) \text{ essent } s_{ij} = \frac{1}{n_t} \sum (x_{tih} - \bar{x}_{ti}) \ (x_{tjh} - \bar{x}_{tj})$$

La distància de Mahalanobis entre les poblacions *i*, *j* és

$$D^2 \ (i,j) = (M_i - M_j) \ ' \ C^{-1} \ (M_i - M_j)$$

El problema a resoldre és trobar una configuració euclídia de *k* punts, deduïda d'una taula de dades Y,

Eixos canònics

	1	...	m
Població 1	y_{11}	...	y_{1m}
...		...	
Població k	y_{k1}	...	y_{km}

de manera que la distància euclídia coincideixi amb D

$$\sum_{h=1}^{n} (y_{ih} - y_{jh})^2 = D^2 \ (i,j)$$

S'anomenen eixos canònics els utilitzats per a representar les *k* poblacions. La dimensió màxima és m = min (k-1, p).

Per trobar la taula de coordenades canòniques Y, cal obtenir la matriu $B = (b_{ij})$ essent

$$b_{ij} = \frac{1}{k} \sum_{h=1}^{k} (\bar{x}_{ih} - \bar{x}_i) \cdot (\bar{x}_{ih} - \bar{x}_j)$$

la «covariància» entre les variables X_i, X_j, agafant només les k mitjanes. Aleshores es busquen els *m* vectors propis de B respecte de C resolent l'equació matricial

$$B \ V_i = \lambda_i \ C \ V_i \qquad (\lambda_i = \text{valor propi de B respecte de C})$$

Si V és la matriu dels vectors propis, aleshores

$$Y = MV$$

essent M la matriu que conté les k mitjanes per les p variables.

Finalment, prenent les d primeres coordenades, tindrem una representació en màxima dispersió en dimensió d. Es verifica

$$\sum_{i,j=1}^{k} \sum_{h=1}^{d} (y_{ih} - y_{jh})^2 = 2k\ (\lambda_1 + \ldots + \lambda_d)$$

L'anàlisi canònica ha d'anar acompanyada d'un test multivariant de comparació de mitjanes i d'un test d'homogeneïtat entre les k matrius de covariàncies de les poblacions. Perquè l'anàlisi tingui sentit, cal que el primer test sigui significatiu i el segon, en canvi, no ho sigui. Vegeu CUADRAS (1979).

Com que les mitjanes de les poblacions tenen fluctuacions degudes al mostratge, cal trobar una regió confidencial per la representació canònica de cada individu mitjà M_i. La regió confidencial és un cercle ($d=2$) o una esfera ($d=3$), de radi

$$\frac{R_\alpha}{\sqrt{n_i}} \text{ essent } R^2_\alpha = F_\alpha \frac{(N-k)}{(N-k-p+1)}$$

F_α és el valor de la F de SNEDECOR tal que $P\ (F > F_\alpha) = \alpha$ amb p i $(N-k-p+1)$ graus de llibertat. Aleshores el coeficient de confiança és $1-\alpha$. Generalment es pren $1-\alpha = 0{,}9$, o sigui, $\alpha = 0{,}1$.

Exemple: PETITPIERRE i CUADRAS (1977) fan una classificació sistemàtica de 32 poblacions (trobades en localitats geogràfiques diferents) de coleòpters del gènere *Timarcha*. Per diferenciar els individus mascles, es prenen 5 mesures biomètriques del pretòrax i dels èlitres, i 3 mesures del diàmetre dels tarses de I, II i III parells de potes. Per les femelles es prenen només les primeres mesures.

La fig. 11 és una representació canònica de 32 poblacions de mascles. la fig. 12 és una representació de 29 poblacions de femelles (no es disposava de femelles per a tres de les poblacions). Hi havia uns quaranta exemplars per població.

Les distàncies representades en dimensió 2 són projecció de les distàncies de Mahalanobis amb dimensió $m=8$. Observant les dues gràfi-

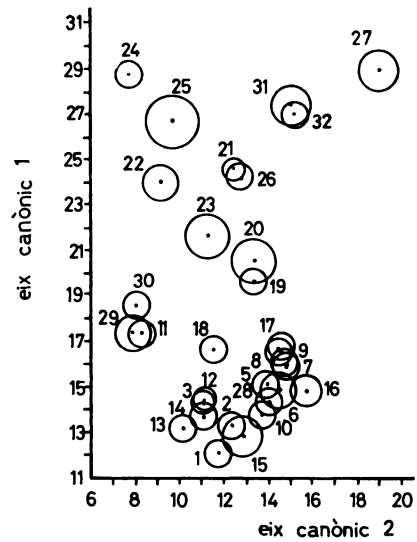

FIG. 11. — Representació canònica de 32 poblacions de coleòpters mascles del gènere *Timarcha*. Les regions confidencials dels individus mitjans estan construïdes al 90 %

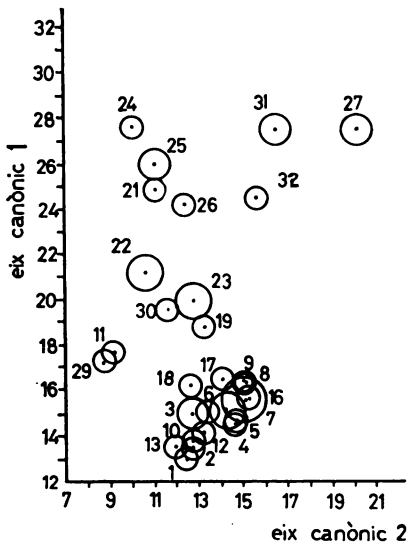

FIG. 12. — Representació canònica de 29 poblacions de coleòpters femelles del gènere *Timarcha*. Les regions confidencials pels individus mitjans estan construïdes al 90 %

ques, es pot veure que les poblacions 11, 20...32 es distingeixen com espècies diferenciades. En canvi les poblacions 1, 2, 10, 15,... formen un conglomerat que les fa interpretar com a races geogràfiques d'una mateixa espècie.

L'anàlisi canònica, en representar les espècies (o races, inventaris, etc.), amb referència a la distància de Mahalanobis, invariable per canvis d'escala i que té en compte les correlacions entre les variables, és la forma més objectiva i correcta de representar dades. Però, perquè es pugui aplicar, cal que tinguem una mostra d'individus representants de cada espècie.

Anàlisi canònica generalitzada

Podem mirar l'anàlisi canònica de poblacions com una representació gràfica dels nivells d'un disseny d'un sol factor. Quan les nostres dades depenguin de dos o més factors, haurem de realitzar tantes anàlisis canòniques dels nivells d'un factor determinat com nivells tenen els altres factors. En l'exemple de la secció anterior, hi ha dos factors: sexe i població, amb 2 i 29 nivells (ignorem ara les tres poblacions de mascles on no hi ha les corresponents femelles). Aleshores hem fet dues representacions canòniques: una per els mascles i l'altra per a les femelles. Però això ens condueix a algunes contradiccions. Per exemple, les poblacions 31 i 32 coincideixen en els mascles, però queden diferenciades en les femelles. Seria absurd distingir com a espècies diferents les femelles i com espècies iguals els mascles. Com es pot resoldre això?

Interpretem-ho com un problema d'anàlisi de la variància. La variabilitat total d'un disseny es descomposa en la variabilitat deguda als nivells de cada factor, la deguda a les interaccions i la variabilitat residual. Quan intervenen $p > 1$ variables observades, aleshores hi ha una matriu de covariàncies total que es descomposa anàlogament amb suma de diverses matrius de covariàncies. Aleshores, si busquem els vectors propis de la matriu de covariàncies que dóna la covariabilitat dels nivells d'un factor, respecte a la matriu de covariàncies residual (que és

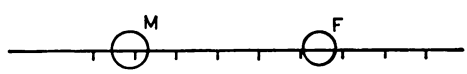

Fig. 13. — Representació canònica de mascles i femelles. Es pot interpretar com una estimació del grau de dimorfisme sexual

la que dóna una estimació de la matriu de covariàncies C entre les variables), tindrem, per analogia amb l'anàlisi canònica de poblacions, una forma de representar els nivells d'un factor, eliminant la variabilitat deguda als altres factors i les interaccions. Àdhuc podrem representar també els altres factors.

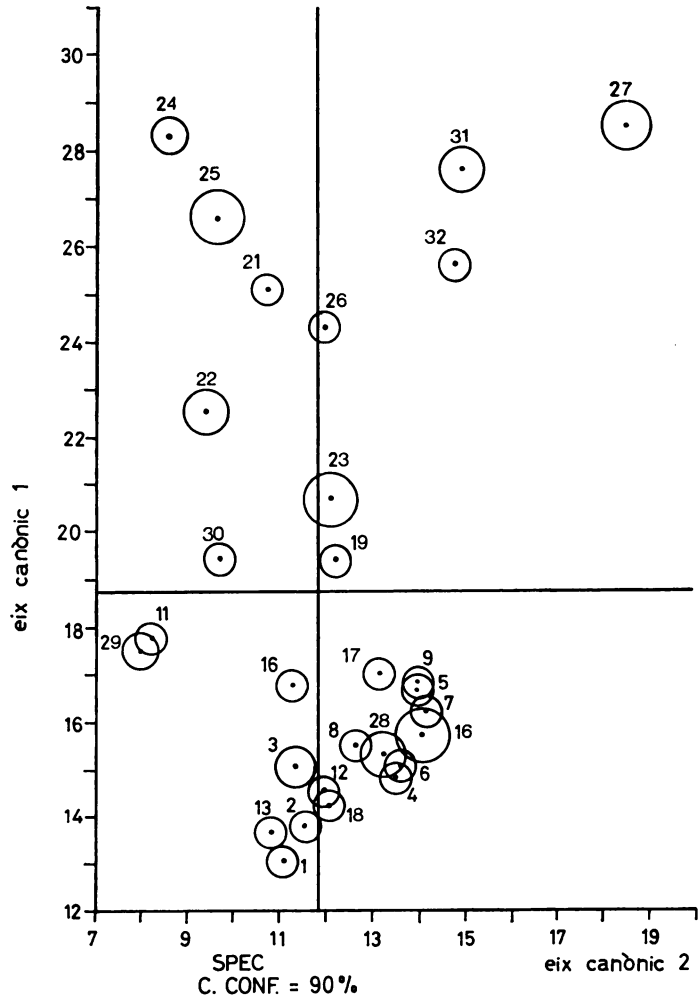

FIG. 14. — Representació canònica de 29 poblacions de coleòpters del gènere *Timarcha*. És una representació conjunta de mascles i femelles que elimina la covariabilitat deguda al dimorfisme sexual i a la interacció i resumeix en una sola les representacions fetes independentment per a mascles (Fig. 11) i femelles (Fig. 12)

El desenvolupament matemàtic d'aquesta generalització forma part de la teoria multivariant de funcions paramètriques estimables, i és massa complicat per poder-ho resumir aquí. Vegeu CUADRAS (1974).

Com una aplicació, vegem el resultat de representar les 2×29 poblacions de mascles i femelles del gènere *Timarcha*. El primer factor,

o factor sexe, té 2 nivells. La fig. 13 és una representació canònica (unidimensional naturalment) dels mascles i de les femelles.

Aquesta figura es pot interpretar com una representació gràfica del grau de dimorfisme sexual. Les femelles serien, en general, més grans que els mascles, mentre que l'eix seria una dimensió de grandària.

La fig. 14 és una representació canònica dels 29 nivells dels segons factors, o sigui, de les 29 poblacions, havent eliminat la variabilitat deguda al sexe i a la interacció. Tenim així una representació única i objectiva de les 29 races geogràfiques o espècies. Així, les poblacions 31 i 32 deuen ésser considerades definitivament diferents.

Cal dir, també, que les distàncies representades són projecció d'una distància que és generalització de la distància de Mahalanobis. Té, doncs, les propietats esmentades d'invariabilitat per canvis d'escala en la mesura de les variables biomètriques, etc.

Altres aplicacions de l'anàlisi canònica generalitzada a la Biologia, i també a la Psicologia i a la Medicina, es poden veure a CUADRAS, PETITPIERRE i COLL (1977).

BIBLIOGRAFIA

1. ALONSO, G.: *Estudio de la distribución geográfica del polimorfismo cromosómico en* Drosophila subobscura. Tesina Fac. Biologia. Univ. Barcelona. 275 pp. (1975).
2. CANTON, E. i SANCHO, J.: *Andlisis Numérico de un grupo de* Pseudomonas *aeróbicos.* Microbiol. Española. 29, 59-73 (1976).
3. CHARDY, P.; GLEMAREC, M. i LAUREC, A.: *Application of inertia methods to benthic marine ecology. Practical Implications of the basic options.* Estuarine coastal marine science. 4: 179-205 (1976).
4. CUADRAS, C. M.: *Andlisis discriminante de funciones paramétricas estimables.* Trab. Estad. Inv. Oper. 25: 3-31 (1974).
5. CUADRAS, C. M.: *Métodos de andlisis multivariante.* Ed. Eunibar. Barcelona (en premsa) (1979).

5 bis. CUADRAS, C. M.; PETITPIERRE, E. i COLL, M. D.: *Generalized discriminant canonical analysis: Applications to Biology, Psychology and Medicine.* First World Conf. on Mat. at the Serv. of Man, Barcelona, Preprints (1977).

6. ESTRADA, M.: *Statistical consideration of some limnological parameters in Spanish reservoirs. Ver.* Internat. Verein. Limnol. 19: 1849-1859 (1975).
7. GOWER, J. C.: *Some distance properties of latent root and vector methods used in multivariate analysis.* Biometrika. 53: 325-038 (1966).
8. HOTTELLING, H.: *Analysis of a complex of statistical variables into principal components.* J. educ. Psychol. 24: 417-441 (1933).
9. JOHNSON, S. C.: *Hierarchical clustering schemes.* Psychometrika. 32: 241-254 (1967).
10. KRUSKAL, J. B.: *Non metric multidimensional scaling: a numerical method.* Psychometrika. 29: 115-129 (1964).
11. OCAÑA, J.: *Sobre la distancia genética.* Tesina Fac. Biologia. Univ. Barcelona. 130 pp. (1975).

12. PETITPIERRE, E. i CUADRAS, C. M.: *The canonical analysis applied to the taxonomy and evolution of the genus* Timarcha Latr. (Coleoptera, Chrysomelidae). Mediterranea. 2: 13-28 (1977).
13. PREVOSTI, A.: *La distancia genética entre poblaciones.* Miscelánea Alcobé. 109-118 (1974).
14. PREVOSTI, A.; OCAÑA, J. i ALONSO, G.: *Distance between populations of* Drosophila subobscura *based on chromosome arrangement frequencies.* Theor. Appl. Genetics. 45: 231-241 (1975).
15. RAO, C. R.: *The use and interpretation of principal components analysis in applied research.* Sanhkya A, 26: 328-358 (1964).
16. ROMERO, J.: *Análisis de componentes principales y una aplicación a la fitosociología.* Trabajo práctico de Bioestadística. Univ. Barcelona (inèdit) (1978).
17. ROHLF, F. J.: *Adaptative hierarchical clustering scheme.* Syst. Zool. 19 (1): 58-82 (1970).
18. SHEPARD, R. N.: *The analysis of proximities: multidimensional scaling with an unknown distance function. I.* Psychometrika. 27: 125-139 (1962 a).
19. SHEPARD, R. N.: *The analysis of proximities: multidimensional scaling with an unknown distance function. II.* Psychometrika. 27: 219-246 (1962 b).
20. SNEATH, P. H. A. i SOKAL, R. R.: *Numerical taxonomy.* W. H. Freeman Co. San Francisco. 573 pp. (1973).
21. SOKAL, R. R. i SNEATH, P. H. A.: *Principles of numerical taxonomy.* W. H. Freeman Co. San Francisco. 359 pp. (1963).

DISCUSSIÓ

FLÓS

Moltes vegades, quan es mesurèn unes variables, no es té encara cap model fet. *A posteriori* se n'adopta un, però existeix una interacció contínua que porta a replantejar-lo sobre la marxa.

J. WAGENSBERG

El problema és abans de l'estadística i després de la biologia. Consisteix en filtrar les dades en funció del model que es busca. S'han d'eliminar interaccions a fi que apareixin factors simples.

FLÓS

Això pot ser molt difícil a la realitat. Pensem, per exemple, en el fitoplàncton. Tenim el moviment de l'aigua, l'absorció de nutrients, etc. Un model es complica de seguida. Molts paràmetres fisiològics no es coneixen i cal fer suposicions.

J. WAGENSBERG

Qualsevol fenomen té un grau de determinisme. Depèn del nivell que es consideri.

VALLESPINÓS

No cal ser nihilista. Sovint, el que es busca es només un model que doni resultats orientatius.

ALONSO

De fet, el model es tria quan es fa l'experiència. Per què no plantejar d'entrada el problema a un matemàtic?

MARGALEF

El nombre de dades no és mai prou important per a estimar les variables que entren en el model. El mostratge es pot simplificar, però el sentit biològic dels factors és sempre molt complicat. La tria dels paràmetres que es volen mesurar es fa en general per intuïció.

ESTRADA

D'altra banda, moltes vegades es mesuren paràmetres sense una adequada dimensió temporal. Per exemple, una mesura de la concentració de nutrients associada a una densitat de població de plàncton en un moment determinat, té poc valor per sí mateixa. És com un fòssil al qual cal buscar un context històric.

ESCARRÉ

És bo que les dades siguin cada vegada de tipus diferent. Així s'estableix un diàleg que obliga el matemàtic a espabilar-se.

CUADRAS

El matemàtic utilitza models que pugui elaborar. En FISHER, per exemple, havia d'imaginar quadrats llatins.

J. WAGENSBERG

No es pot admetre un model diferent cada vegada.

MARGALEF

El biòleg tendeix a uniformitzar la variabilitat. On és més alta pren més mostres. En oceanografia, per exemple, els punts de mostratge són més pròxims en sentit perpendicular a la costa. Falta una discussió crítica de com una teoria basada en la uniformitat pot servir de guia quan es tendeix a uniformitzar la variació.