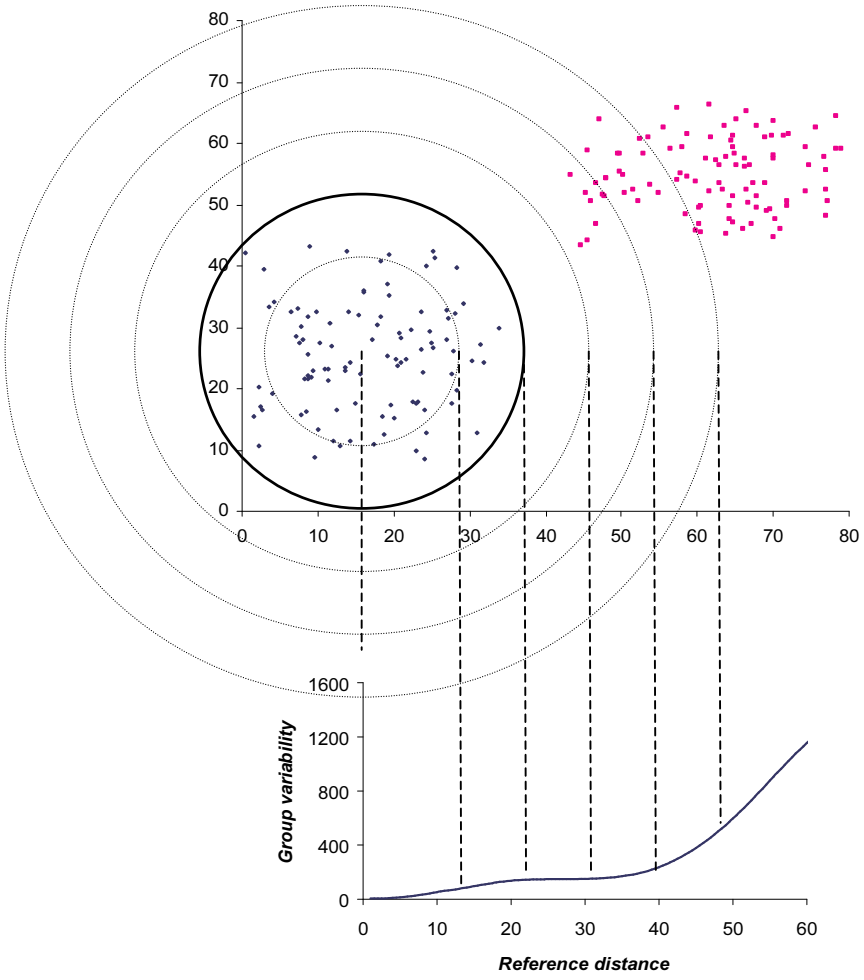
III – ASPECTES METODOLÒGICS DE LA CLASSIFICACIÓ NUMÈRICA DE COMUNITATS DE VEGETALS



Capítol 3.1: Models i mètodes de classificació numèrica de la vegetació

3.1.1 Introducció a la classificació numèrica

3.1.1.1 Sintaxonomia numèrica

La sintaxonomia numèrica (*numerical syntaxonomy*) fou definida el 1981 per van der Maarel com '*the phytosociological discipline employed to establish a syntaxonomical system of plant communities with the help of numerical methods*' (Mucina & van der Maarel 1989). La sintaxonomia numèrica s'inclou dins del que podem anomenar "ecologia numèrica". En aquest capítol abordarem un subconjunt dels mètodes multivariants que s'apliquen a la sintaxonomia numèrica. Utilitzarem únicament aquelles tècniques relacionades amb la classificació de les comunitats de vegetals. Per a tenir una visió més àmplia del tema, el lector es pot adreçar a les nombroses revisions del tema existents a la literatura científica (per exemple, James & McCulloch 1990, Feoli & Orlóci 1991, Escudero *et al.* 1994). D'altra banda, existeixen manuals d'ús estès sobre mètodes multivariants aplicats a l'ecologia com, per exemple, Whittaker (1973), Legendre & Legendre (1998) o McCune & Grace (2000). Han avaluat l'impacte d'aquestes tècniques en la classificació de comunitats de vegetals Goodall (1986), Kent & Ballard (1988), Mucina & van der Maarel (1989) i, més recentment, Mucina (1997).

3.1.1.2 Classificació: Anàlisi discriminant i anàlisi de clústers (*clustering*)

La necessitat d'establir grups en un conjunt d'elements és motivada per la creença *a priori* de que la variabilitat observada en les dades permet dividir els elements en categories que siguin útils per a simplificar el seu estudi ulterior. Tanmateix, la manera d'abordar el problema de la classificació a nivell matemàtic és diferent, segons el grau de coneixement de que partim. Si disposem d'unes dades classificades *a priori* que ens permetin entrenar-nos en la identificació de grups, el problema esdevé matèria de l'anàlisi discriminant (*discriminant analysis*). Els mètodes d'anàlisi discriminant tracten de construir, a partir de les dades classificades *a priori*, una o un conjunt de regles o criteris numèrics de classificació. Aquestes regles són posteriorment aplicades a noves observacions de classificació desconeguda. La qualitat de les regles construïdes dependrà de la quantitat i qualitat de la informació que tinguem *a priori* sobre els grups. La millor situació es dóna quan no tan sols coneixem la classificació sinó també la funció de distribució multivariant de les variables en cada un dels grups i els paràmetres de la mateixa. En aquest cas ideal hom podria aplicar la regla òptima o de Bayes per a classificar nous

elements. Si coneixem el tipus distribució teòrica de les variables però no els seus paràmetres, podem intentar estimar-los a través d'estadístics mostrals. No obstant, és força freqüent desconèixer també la distribució teòrica de les dades. Alguns manuals de referència enfocats a l'ús de l'anàlisi discriminant són McLachlan (1992), Mardia *et al.* (1994), Huberty (1994) o Duda *et al*. (2001).

Si la classificació de les dades és també desconeguda, hom s'allunya de les tècniques relacionades amb l'anàlisi discriminant per acostar-se al camp d'estudi de l'anàlisi de clústers o anàlisi de conglomerats (*cluster analysis*). L'anàlisi de clústers és l'estudi formal dels algorismes i mètodes per agrupar i classificar objectes. Anomenarem mètodes de *clustering* els mètodes de classificació de l'anàlisi de clústers, per tal de diferenciar-los dels mètodes d'anàlisi discriminant que també són mètodes de classificació. A la literatura especialitzada sobre reconeixement de patrons i intel·ligència artificial, l'anàlisi de clústers s'anomena també aprenentatge no supervisat (*unsupervised learning*). L'objectiu dels algorismes de *clustering* és, precisament, determinar l'estructura de grups de les dades quan aquesta és desconeguda. En comparació a l'anàlisi discriminant, la manca de coneixement previ fa que els mètodes de *clustering* siguin força heurístics en la seva aproximació a la classificació i no permetin contribuir massa a testar hipòtesis. Hartigan (1975) remarcava, en referència a les tècniques d'anàlisi de clústers: '*they are not yet an accepted inhabitant of the statistical world*'. Manuals de referència especialitzats en *cluster analysis* són, entre d'altres, Hartigan (1989), Jain & Dubes (1988), Everitt (1993) o Gordon (1999). Les àrees de coneixement on s'aplica l'anàlisi de clústers són molt variades: psicologia, ecologia, reconeixement d'imatges, marketing... Per aquest motiu, la bibliografia referent a la metodologia de *cluster analysis* es troba no només en revistes de caire estadístic sinó que apareix repartida en camps molt dispars.

En el cas de les nostres dades de vegetació (els sintàxons de base de *Brometalia* i *Quercetea*), la informació de que disposem *a priori* és bastant pobre des del punt de vista estadístic. La funció de distribució teòrica de les variables (els tàxons) en els nostres grups (sintàxons) ens és desconeguda. Sortosament, disposem d'una classificació de partida, la que ens proporciona la fitosociologia tradicional. Tanmateix, la confiança en aquesta classificació inicial no és del tot plena: no tots els grups establerts per la metodologia tradicional són igualment vàlids per les diverses escoles de fitosociòlegs. A més, cal diferenciar l'acceptació de la validesa d'un sintàxon per part d'un o més experts de vegetació de la validesa des del punt de vista numèric (veure capítol 2.3). Per aquests motius, tot i podent encarar el present treball vers l'anàlisi discriminant, no pressuposarem que els inventaris de *Brometalia* i *Quercetea* ja estan correctament classificats sinó que prendrem una actitud més cautelosa i mirarem d'explorar les dades amb mètodes propis de l'anàlisi de clústers. Caldrà discutir *a posteriori* si els grups trobats són vàlids fitosociològicament i comprovar si coincideixen amb els resultats proporcionats per la classificació sintaxonòmica tradicional.

3.1.1.3 Definició de clúster i models de classificació

Un dels problemes fonamentals de l'anàlisi de clústers és la manca d'una definició precisa del concepte de clúster. Aquesta mancança ha portat al desenvolupament d'un gran nombre de mètodes de *clustering*, basats en conceptes de clúster diferents (de vegades *ad hoc*) i un nombre d'algorismes encara més gran.

L'anàlisi de clústers es basa sovint en les relacions entre els individus (objectes) a classificar. És freqüent que la distància geomètrica entre individus en un espai multivariant jugui un paper essencial en molts mètodes de *clustering*. Hi ha, a partir d'aquí, dos criteris geomètrics en la definició de clúster que permeten veure la transcendència de la definició: 1) Si hom considera que un clúster és una regió de punts interconnectats, la forma del conjunt resultant pot ser molt variable. Ens trobarem davant d'un conjunt de punts que descriurà un patró. 2) Si hom considera un clúster com a una poblacions estadística amb vector de característiques mitjanes i una certa variabilitat, la forma dels clústers pot ser hiperesfèrica o hiperel·lipsoidal, però altres formes tenen menys sentit. A més, en aquest cas no tots els punts han d'estar "connectats" a l'estructura. A la figura 3.1.1 hem reproduït, a partir de Jain & Dubes (1988), exemples de patrons espacials, alguns dels quals admeten també la consideració de clústers segons el criteri (2). L'exemple més clar de la diferencia entre els dos criteris exposats rau en el cas A: Hom pot considerar el cercle exterior com a un patró (1), però no un clúster segons el criteri (2).

És important esmentar que existeixen molts mètodes de classificació que no es basen en les relacions espacials entre individus, sinó que utilitzen altres criteris d'agrupament (vegeu, per exemple, els mètodes de l'apartat 3.1.4).

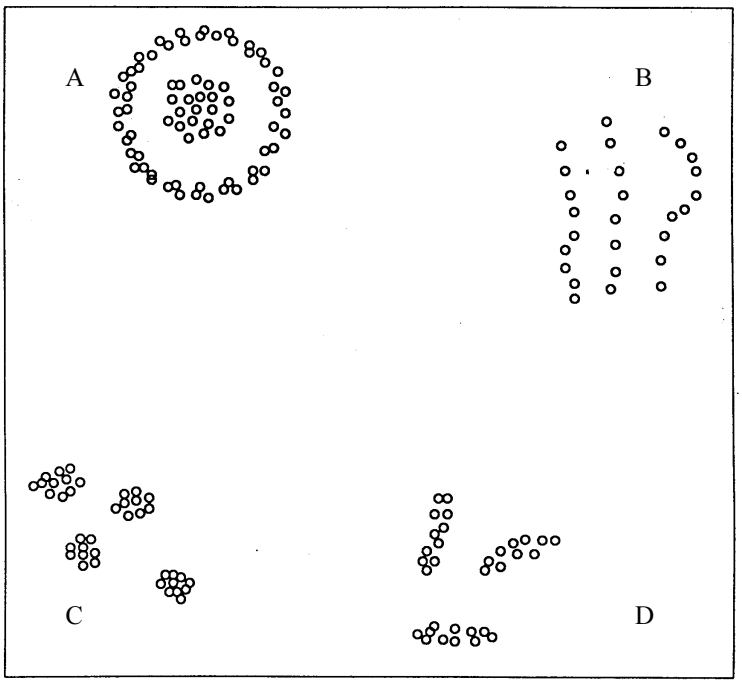


Figura 3.1.1: Exemples de patrons espacials (extret de Jain & Dubes, 1988). Els 4 grups principals es poden subdividir en subgrups. Alguns d'ells són senzills de establir numèricament (C o D) mentre que en d'altres la majoria d'algorismes són totalment ineficaços (A o B).

Quina definició de clúster s'adapten millor a les unitats de vegetació abstractes de la fitosociologia? Els fitosociòlegs semblen no estar totalment d'acord entre ells sobre la naturalesa dels tipus de vegetació que cerquen. Bruelheide & Flintrop (1994) opinen que no hi ha una definició consistent, reconeguda generalment, del terme "tipus de vegetació". Els fitosociòlegs són notablement reticents a donar una definició formal, confiant en l'aprenentatge de la manipulació de taules d'inventaris. La reorganització manual de taules ha estat força mitificada com a mètode eficient per a determinar grups d'inventaris (Mucina 1997). És, per tant, freqüent la promoció i ús de mètodes numèrics que simplement automatitzin el procés d'ordenació de files i columnes. Noteu que aquests mètodes no estan basats en conceptes de clúster geomètrics.

En un àmbit més numèric d'estudi de la vegetació, Gauch & Whittaker (1981) proposaren que les dades per elles mateixes havien de suggerir categories "naturals": *Els punts de mostratge en l'espai dels atributs (tàxons en el cas de comunitats) poden ser agrupats de manera 'natural', tenint els elements d'un mateix grups dissimilaritats petites entre ells i grans amb els elements d'altres grups. Només així els grups apareixen "naturals" a les dades. Si les mostres d'un mateix grup estan relativament molt separades, és l'algorisme de clustering que arbitràriament imposa la seva estructura. Els límits dels grups no són llavors defensables ni importants. El mateix nombre de grups desitjat pot acceptar un grau de variació. En aquest darrer cas, els grups no són 'naturals' (per referència a l'espai de dades) sinó arbitraris* (Gauch & Whittaker 1981). Aquesta definició de clúster sí és geomètrica. La cerca de "naturalitat" dels clústers pot conduir-nos a descartar el concepte de clúster com a patró de punts connectats, però constitueix una definició encara força imprecisa. On Gauch & Whittaker encerten plenament és en que moltes vegades en l'anàlisi de comunitats vegetals busquem grups naturals però acabem determinant grups arbitraris. Per a Dale (1995) hi ha dos problemes bàsics: 1) La mostra reflecteix adequadament l'estructura de la població? 2) Quina és la naturalesa dels clústers que busquem? Segons aquest autor, la definició de tipus de vegetació és pragmàtica, reflectint una apreciació de la variació interna exhibida pels clústers proposats, nascuda d'un coneixement de la vegetació real observada.

En l'estadi de traduir un criteri de clúster en un mètode de *clustering* hi ha un altre punt important a decidir: Quin model de classificació és millor per a les nostres dades? Entendrem com a model de classificació el tipus d'estructura que representarà formalment el resultat de la classificació. En els propers apartats descriurem tres models diferents: les jerarquies (3.1.2), les particions (3.1.3) i els blocs d'espècies/inventaris (3.1.4); i presentarem alguns mètodes de *clustering* basats en aquests tres models. En aquest capítol pretenem comparar models i/o mètodes concrets de classificació de la vegetació. Reservarem l'estudi d'un quart model de classificació per al capítol 3.3.

3.1.2 Model de classificació jeràrquic

3.1.2.1 Definició del model jeràrquic

En un model de classificació jeràrquic la pertinença a un grup o clúster d'un nivell de la jerarquia condiciona la pertinença a grups de nivells superiors. L'estructura jeràrquica resultant aporta un coneixement de les dades a diversos nivells, però en detriment de la flexibilitat. Una bona introducció a aquest model de classificació es pot trobar a Gordon (1996).

Aplicat a la fitosociologia, el model jeràrquic presenta problemes teòrics (Dale 1988, Mirkin 1989). En efecte, tal i com indica Mucina (1997), només si assumim que el procés que genera la variabilitat de les dades té una component jeràrquica (com es compleix, a grans trets, en la taxonomia) aquest model es pot justificar com a "natural". Ja hem comentat, a la introducció de la memòria (capítol 1.1) alguns problemes d'aplicar a les relacions entre comunitats una estructura jeràrquica. Tanmateix, hom ha d'admetre que la mateixa sintaxonomia no deixa de tenir un caràcter convencional (Mirkin 1989), i que el propi ús per part dels fitosociòlegs fa palesa una funcionalitat del model jeràrquic.

Els mètodes jeràrquics van ser molt utilitzats durant els primers anys de la sintaxonomia numèrica (veure referències més endavant) i és per aquest motiu que els tractarem aquí. Es divideixen en aglomeratius i divisius, segons si construeixen les jerarquies agrupant objectes (*bottom-up*) o dividint seqüencialment les dades (*top-down*).

3.1.2.2 Algorismes jeràrquics aglomeratius

Els algorismes jeràrquics aglomeratius s'anomenen freqüentment *SAHN*: *sequentially adding hierarchical algorithms*. Parteixen, normalment, d'una matriu de proximitats (similaritats o distàncies/dissimilaritats) simètrica de dimensions NxN. Sobre aquesta matriu, els algorismes van agrupant elements progressivament. El nombre d'agrupacions realitzades al final és $N-1$. Com a sortida gràfica, generen estructures jeràrquiques, tècnicament anomenades arbres ultramètrics o dendrogrames, on es compleix la propietat ultramètrica. La ultrametricitat implica que, per a qualsevol triplet d'objectes (ω_1, ω_2 i ω_3), dos de les tres distàncies són idèntiques i no més petites que la tercera. Formalment, en el cas de distàncies, la propietat ultramètrica s'enuncia:

$$d(\omega_1,\omega_2) \leq \max\left[d(\omega_1,\omega_3), d(\omega_2,\omega_3)\right].$$

Les diferències entre uns i altres algorismes rau en el criteri emprat per a definir la proximitat dels clústers que es van creant respecte als preexistents. A continuació descrivim breument les variants més conegudes. Per a una explicació detallada dels algorismes *SAHN* vegeu Sneath & Sokal (1973), Jain & Dubes (1988) o Legendre & Legendre (1998).

Pel text que segueix, un node es refereix a un objecte o a un clúster d'objectes ja format.

- ***Single linkage*** (o *nearest neighbour linkage*): La dissimilaritat del node creat amb un node extern s'agafa com la mínima entre les dissimilaritats dels elements que formen el nou node i el node extern. Això produeix sovint un efecte d'encadenament (*chaining*) perquè com més elements tingui un node més fàcil és que se n'hi afegeixin. Dit d'altra manera, *single linkage* contrau l'espai de referència al voltant del clúster. En conseqüència, és molt sensible a l'existència de continuïtats a les dades. *Single linkage* està relacionat amb la construcció d'arbres de mínima extensió (*minimum spanning trees*, Gower & Ross 1969). Jardine & Sibson (1968) generalitzaren *single linkage* proposant l'algorisme B_k, que permet un cert nombre d'objectes (k - 1) compartits entre clústers.
- ***Complete linkage*** (o *furthest neighbour linkage*). Contràriament a l'anterior, la dissimilaritat del node creat amb un node extern s'agafa com la màxima entre les dissimilaritats dels elements que formen el nou node i el node extern. Com més elements té un node més difícil esdevé afegir-ne més. Per tant *complete linkage,* contràriament a l'anterior, expandeix l'espai de referència al voltant del clúster. Els propers cinc mètodes conserven l'espai de referència.
- ***Unweighted arithmetic average clustering*** (*UPGMA*). En aquest algorisme, la dissimilaritat del node creat amb un node extern s'agafa com la mitjana aritmètica de dissimilaritats entre els elements que formen el nou node i el node extern. Tots els objectes reben el mateix pes en el càlcul.
- ***Weighted arithmetic average clustering*** (*WPGMA*). És igual que l'anterior, excepte que dóna un pes igual als dos nodes que compara. Alguns nodes tenen més elements que altres. Per tant, donar igual de pes a dos nodes equival a donar pesos diferents als elements que formen els nodes (és a dir, donar menys pes als elements de nodes nombrosos).
- ***Unweighted centroid clustering*** (*UPGMC*). La dissimilaritat del node creat amb un node extern s'agafa com la dissimilaritat entre el centroide dels elements que formen el nou node i el node extern. Com a *UPGMA*, tots els objectes reben el mateix pes en el càlcul del centroide.
- ***Weighted centroid clustering*** (*WPGMC*). És l'anàleg a *WPGMA* en centroides. El mètode *WPGMC* és igual que *UPGMC*, excepte que dóna un pes igual a les branques que formen un node per a obtenir el centroide. Com a resultat, els elements originals reben diferent pes depenent de la mida de la branca on pertanyen.
- ***Ward's method*** o ***minimum sum of squares method*** (Ward 1963, Orloci, 1967) Està relacionat amb els dos anteriors en el sentit de que els centroides hi juguen un paper. Per formar grups, el mètode de Ward's minimitza, a cada pas, la suma d'errors quadràtics: fusiona els nodes que produeixen un increment de l'error quadràtic menor. Aquesta funció objectiu és equivalent al funcional del mètode partitiu *K-means* (veure 3.1.3.2). Cal notar que

aquesta aproximació es diferencia respecte a la minimització de la variància (Diday *et al*. 1982) en que té una tendència major a fer grups d'igual mida. No obstant, Podani (1989) indica que és freqüent la denominació errònia del mètode de Ward com a mètode de mínima variància.

- ***β-Flexible method***. Lance & Willams (1967) proposaren un mètode jeràrquic generalitzat per a englobar tots els mètodes anteriors. Siguin ω_i i ω_j els objectes d'un grup recentment format, i ω_k un objecte extern, la distància d'aquest al nou grup (ij) es calcula mitjançant l'equació: $d_{k(ij)} = \alpha_i \cdot d_{ki} + \alpha_j \cdot d_{kj} + \beta \cdot d_{ij} + \lambda \cdot \left| d_{ki} - d_{kj} \right|$. Segons els valors dels 4 paràmetres del model es poden obtenir algorismes completament equivalents als anteriors. A la vegada Lance & Willams (1967) proposaren variar un dels paràmetres, β, entre -1.0 i 1.0, per tal d'obtenir una sèrie de solucions intermèdies entre *single linkage* i *complete linkage*. Aquest mètode és anomenat per autors posteriors *β-flexible clustering*. Les propietats de l'espai es conserven per valors de $\beta \approx -0.25$, valor recomanat per Milligan (1989a). Cal esmentar, que el nombre d'algorismes jeràrquics aglomeratius pot eixamplar-se més afegint nous termes a l'equació de Lance & Williams (veure, p.e. Podani 1989c).

Un desavantatge d' *UPGMC* i *WPGMC* és que poden produir reversions en l'arbre ultramètric. Una reversió es produeix quan en l'agrupació de dos nodes, la distància que hi ha entre el node resultant i un tercer element és inferior a la distància que hi havia entre els nodes agrupats. Per altra banda, *single linkage* i *complete linkage* són els únics algorismes, d'entre els citats, que romanen invariants a transformacions monòtones de la matriu de proximitats inicial (Milligan 1979).

Els mètodes jeràrquics aglomeratius foren emprats sovint en els primers treballs de sintaxonomia numèrica (Orlóci 1967, Jancey 1980, Gauch & Whittaker 1981). Entre els algorismes jeràrquics més utilitzats cal destacar *complete linkage* (p.e. Mucina 1989, Šeffer et al. 1989) i el mètode de Ward's (p.e. van Speybroeck *et al*. 1989). Actualment, els mètodes jeràrquics aglomeratius no són massa populars, en benefici de *TWINSPAN* (veure 3.1.2.3) i dels mètodes partitius. Segons Hill *et al.* (1974) els inconvenients dels mètodes jeràrquics aglomeratius són que no permeten classificar nous elements d'una manera fàcil i es tornen impracticables per volums de dades molt grans. De tota manera, encara actualment es troben treballs de vegetació on s'utilitzen aquest tipus d'algorismes (Torres *et al.* 1995, Smith & Steenkamp 2001).

Darrerament, Podani *et al.* (2000) han proposat emprar un model jeràrquic aglomeratiu no basat en la propietat ultramètrica. Concretament, proposen estudiar les comunitats amb arbres additius generats a partir de l'aplicació de l'algorisme *neighbor-joining* (Saitou & Nei 1987). Els

arbres additius, força emprats en filogènia, compleixen la propietat de ser mètriques de quatre punts (*four-point metrics*), és a dir:

$$d(\omega_1,\omega_2)+d(\omega_3,\omega_4)\leq \max\left[d(\omega_1,\omega_3)+d(\omega_2,\omega_4), d(\omega_1,\omega_4)+d(\omega_2,\omega_3)\right]$$

Com a restricció matemàtica, aquesta propietat és més feble que la ultramètrica. Això implica que qualsevol ultramètrica és una mètrica de quatre punts a la vegada. Podani *et al*. (2000) recomanen aquest mètode per a estudis de seguiment de processos de revegetació i, en general, en processos de successió. També presenten un exemple fitosociològic. Segons el nostre punt de vista, creiem la utilitat dels arbres additius en sintaxonomia numèrica presenta les mateixes dificultats teòriques esmentades per Hill *et al.* (1974) en referència als arbres ultramètrics.

3.1.2.3 Algorismes jeràrquics divisius. TWINSPAN

Els mètodes de classificació jeràrquics divisius parteixen del conjunt sencer dels objectes i el divideixen en dos o més subgrups. A continuació consideren cada un dels subgrups i els divideixen altra vegada. Aquest procés continua fins que algun criteri d'aturada (*stopping rule*) ho exigeix. Els mètodes divisius poden ser monotètics o politètics. Els primers basen la divisió en un sol caràcter cada vegada. Un exemple n'és l'anàlisi d'associacions (Williams & Lambert 1959), que fou el primer mètode de sintaxonomia numèrica. Els mètodes politètics per contra, prenen en consideració tots (o molts) els descriptors en les divisions que efectuen. En aquest sentit els algorismes presentats a l'apartat anterior són tots politètics, com també ho són els algorismes partitius de la secció 3.1.3.

Hom pot obtenir una divisió politètica eficient mitjançant els eixos sorgits d'ordenacions. Emprant una anàlisi de components principals (*principal components analysis*, PCA) podem dividir el conjunt d'objectes en dos grups, depenent de si tenen valors positius o negatius en la primera component principal. Podem repetir una PCA per cada un dels grups obtinguts i d'aquesta manera obtenir una estructura jeràrquica (veure Legendre & Legendre 1998). La mateixa estratègia la podem emprar valent-nos d'una anàlisi de coordenades principals (ACoP, Gower 1966). Lefkovitch (1976) proposà emprar una ACoP per a generar una classificació jeràrquica divisiva. No obstant, i a diferència del mètode que acabem de descriure, la proposta de Lefkovitch no implica refer l'anàlisi per a cada subgrup sinó que proposa prendre successivament els altres eixos de l'ACoP.

La divisió polítètica basada en mètodes d'ordenació ens serveix per a introduir l'algorisme *two-way indicator species analysis*, *TWINSPAN* (Hill *et al*. 1974, Hill 1979). TWINSPAN és un algorisme jeràrquic politètic divisiu basat en l'anàlisi factorial de correspondències. A continuació, en resumim els passos (Hill *et al.* 1974):

(i) Les dades són primer ordenades mitjançant una ordenació per *reciprocal averaging* (RA, Hill 1973a) equivalent a la primera component d'una anàlisi factorial de correspondències (*correspondence analysis*, CA).

(ii) Les mostres són dividides en dos grups, dividint l'eix per el centre de gravetat.

(iii) Es seleccionen les espècies millor indicadores de la divisió.

(iv) Es calculen uns *scores* indicadors per als inventaris a partir del nombre d'espècies indicadores que presenten.

(v) Permetent una zona d'indiferència per casos límits, es divideix de nou la ordenació original en un punt que coincideix, tant com es pugui, amb un llindar de *score* indicador.

(vi) El procés es repeteix per a cada grup per separat, generant una classificació jeràrquica, on cada nivell porta associades unes espècies indicadores.

(vii) El criteri d'aturada de les divisions es especificat per l'usuari: Es pot establir un nombre màxim de divisions o el nombre mínim d'inventaris d'un grup per a permetre la seva subdivisió.

Des de la seva concepció original, *TWINSPAN* ha estat objecte d'algunes propostes de millora. Per una banda Carleton *et al*. (1996) proposen substituir el paper de CA per l'anàlisi canònica de correspondències (CCA), donant lloc a *COINSPAN* (c*onstrained indicator species analysis*). Per l'altra, Oksanen & Minchin (1997) proposen modificar *TWINSPAN* per a corregir alguns defectes d'estabilitat sota canvis en l'ordre d'entrada de les dades.

TWINSPAN ha estat un dels algorismes d'anàlisi de clústers més utilitzats per a classificar inventaris de vegetació (veure per exemple, Westman 1983, Retuerto & Carballeira 1991, Zhang & Oxley 1994, i Kazmierczak *et al.* 1995). Tanmateix, ja que *TWINSPAN* es basa amb l'anàlisi de correspondències (CA), alguns dels inconvenients que es poden argumentar contra aquesta tècnica d'ordenació poden ser igualment aplicables a l'algorisme de *clustering*. Concretament, alguns autors, com van Groenewoud (1992) o Bruelheide & Chytry (2000), han criticat la inestabilitat dels resultats de *TWINSPAN* en base a la seva dependència de CA.

3.1.3 Model de classificació partitiu

3.1.3.1 Concepte de partició

Una partició és una divisió d'un conjunt de objectes en un nombre predefinit de subconjunts, anomenats grups o clústers, sense relacions jeràrquiques entre ells. En una partició dura o clàssica (*hard partition*), cada objecte pertany a un grup de la partició i només a un. Emprant la teoria de la lògica difusa (Zadeh 1965, veure capítol 1.1), hom pot generalitzar la partició clàssica en una partició difusa o borrosa (*fuzzy partition*). En aquest segon cas, els objectes poden pertànyer a més d'un grup, sempre i quan la suma de les pertinences a tots els grups sigui igual a 1 (figura 3.1.2). Per tant, la pertinença total en els dos tipus de particions és 1.

A continuació definim, formalment, una partició de K grups o clústers sobre el conjunt d'N objectes. Donat un enter K comprès entre 1 i N, la matriu de pertinença (*membership matrix*) $\mathbf{U}_{NxK} = \{u_{i(k)}\}$ és una partició difusa no-degenerada si compleix ($k = 1,\ldots,K$, $i=1,\ldots,N$):

(a) $u_{i(k)} \in [0,1] \forall\, i,k$ **(b)** $\sum_{k=1}^{K} u_{i(k)} = 1 \quad \forall\, i$ **(c)** $0 < \sum_{i=1}^{N} u_{i(k)} < N \;\forall k$

on $u_{i(k)}$ indica la pertinença de l'objecte ω_i al clúster Ω_k ($u_{i(k)} = I[\omega_i \subset \Omega_k]$). La única diferència formal entre una partició difusa i una partició dura és que la condició **(a)** es transforma en : **(a')** $u_{i(k)} \in \{0,1\} \forall\, i,k$. La condició (**b**) és fonamental en la definició de una partició. Indica que en les particions, tan dures com difuses, els clústers no s'estableixen independentment sinó que es defineixen per exclusió d'altres clústers (vegeu capítol 3.3).

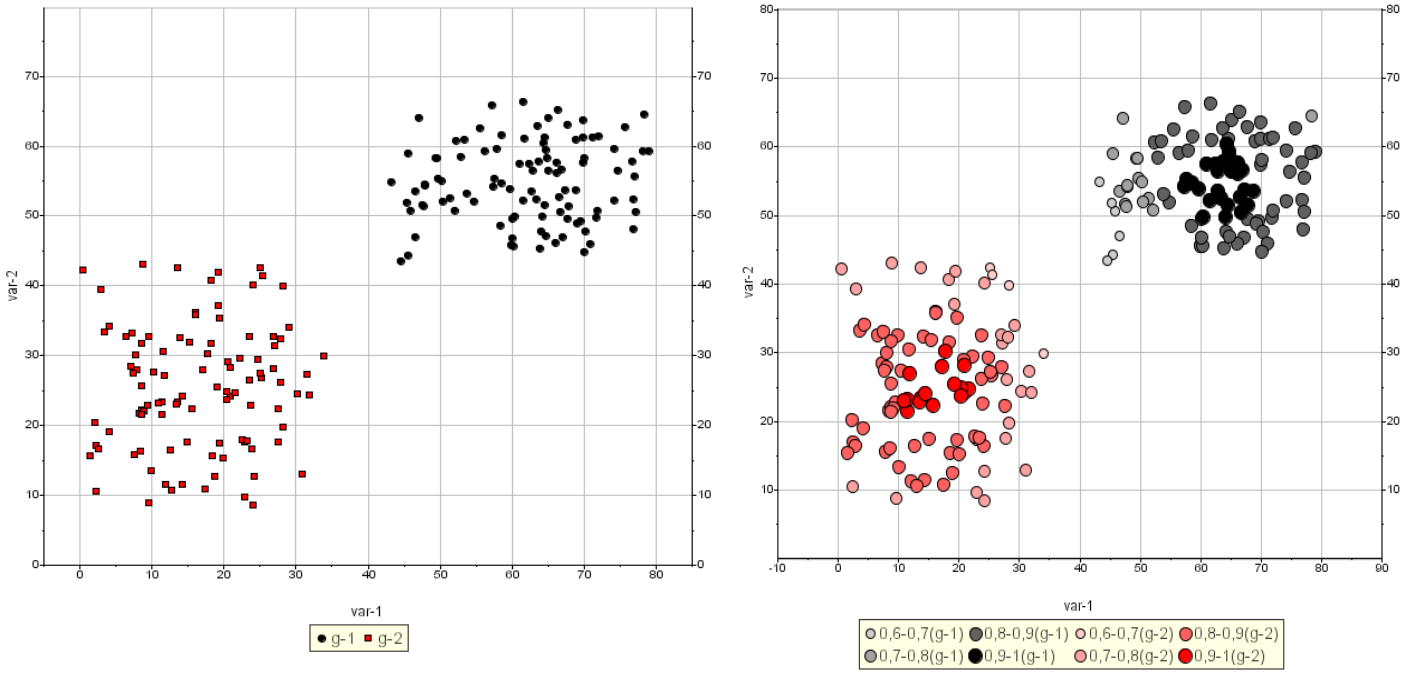


Figura 3.1.2: Exemple d'una partició de dos *clústers* (K=2), determinats amb lògica clàssica (a l'esquerra) o amb lògica difusa (a la dreta) sobre un espai simulat bidimensional. En el cas difús es representa per a cada objecte la pertinença més alta. La pertinença a l'altre grup és la complementària del valor mostrat.

3.1.3.2 *K-means* (KM)

K-means (MacQueen 1967) és un algorisme iteratiu que minimitza a cada iteració la dispersió interna dels clústers. En altres paraules, el funcional a minimitzar és la suma total dels quadrats dels errors (*TESS*):

$$TESS_K = \sum_{k=1}^{K} E_{(k)}^2 = \sum_{k=1}^{K}\sum_{i=1}^{N} I[\omega_i \subset \Omega_k]\, e_{i(k)}^2$$

essent $E_{(k)}^2$ la suma de quadrats dels errors (*ESS*) per al clúster Ω_k, $I[\omega_i \subset \Omega_k]=1$ si l'objecte ω_i ha estat assignat a Ω_k, $I[\omega_i \subset \Omega_k]=0$ en cas contrari. $e_{i(k)}^2$ és la distància al quadrat de cada objecte al centroide de Ω_k. En altres paraules, $e_{i(k)}^2$ és la dispersió de l'objecte respecte al centroide:

$$e_{i(k)}^2 = d^2(\omega_i, \Omega_k) = \sum_{j=1}^{P} (x_{ij} - \bar{x}_{(k)j})^2 = (\mathbf{x}_i - \bar{\mathbf{x}}_{(k)})'(\mathbf{x}_i - \bar{\mathbf{x}}_{(k)})$$

on $\bar{x}_{(k)j} = \frac{1}{N_k}\sum_{i=1}^{N} I[\omega_i \subset \Omega_k]\, x_{ij}$, $N_k = \sum_{i=1}^{N} I[\omega_i \subset \Omega_k]$. En general, $e_{i(k)}^2$ pot ésser qualsevol distància calculada com una mètrica induïda per una norma: $e_{i(k)}^2 = (\mathbf{x}_i - \bar{\mathbf{x}}_{(k)})'\,\mathbf{A}\,(\mathbf{x}_i - \bar{\mathbf{x}}_{(k)})$

on $\mathbf{A}_{PxP}$ és una matriu inductora de la norma. Si $\mathbf{A}_{PxP} = \mathbf{I}_P$, la norma resultant seria la Distància Pitagòrica (Euclidiana). Si $\mathbf{A}_{PxP}$ és una matriu diagonal amb les inverses de les variàncies de cada variable, la norma seria la Diagonal. Finalment, si $\mathbf{A}_{PxP}$ és la inversa de la matriu de variàncies-covariàncies, la norma seria la de Mahalanobis.

Coneguda aquesta definició de funcional, hom es podria plantejar comprovar exhaustivament totes les particions possibles fins a trobar aquella que resultés en un mínim de *TESS*. Malauradament, el nombre de casos a comprovar pot ser molt gran. El nombre de particions possibles d'N objectes en K grups, $S(N, K)$, és (vegeu Jain & Dubes 1988):

$$S(N,K) = \frac{1}{K!}\cdot\sum_{k=1}^{K}\left[(-1)^{K-k}\cdot\binom{K}{k}\cdot k^N\right]$$

Per exemple, $S(10,2) = 511$, $S(10,4) = 34105$ i $S(19,4) = 11259666000$! És palès, doncs, que no podem testar exhaustivament totes les particions possibles quan N es gran. És necessari emprar algorismes que optimitzin la funció anterior.

Donada una partició inicial en K clústers, la tècnica clàssica de *K-means* es basa en el següent senzill algorisme iteratiu:

1. Calcular de les posicions dels centroides $\bar{\mathbf{x}}_{(k)}$ dels K clústers.
2. Per a cada objecte, calcular la seva distància als K centroides , $e_{i(k)}^2$.
3. Reassignació de cada objecte al clúster més proper.

L'algorisme que acabem de descriure correspon concretament a la variant exposada per Forgy's (Forgy 1965), també anomenada *H-means* (Belacel *et al.* 2002), enfront a la estratègia original (MacQueen 1967) d'assignar cada objecte per separat actualitzant a cada pas les posicions dels centroides. En el seu treball, Belacel *et al.* (2002) també proposen una nova variant algorísmica, anomenada *J-Means*, del mateix mètode. La diferència entre unes i altres variants no és massa rellevant pel problema que ens ocupa, pel que aquí ens referirem en tot moment a *K-means* per conveniència, malgrat estiguem emprant en realitat la variant algorísmica *H-means*.

3.1.3.3 *LOO K-means*: correcció de l' efecte "atractor"

L'algorisme de reassignació dels objectes a *K-means* presenta un inconvenient: la distància d'un objecte al clúster al que ha estat prèviament assignat té 'biaix'. Aquest serà major com menys objectes tingui el grup. Efectivament, hom pot comprovar a la figura 3.1.3 com la distància es redueix degut a l'efecte d'atracció que realitza l'objecte sobre el centroide per la seva pertinença *a priori* al clúster. En el cas límit, un clúster que en una determinada iteració contingui un únic objecte, aquest quedarà immobilitzat.

Amb l'objectiu d'eliminar aquest efecte no desitjat, proposem (ja publicat a Oliva *et al*. 2001) calcular la distància d'un element a cada centroide eliminant-ne la influència, de manera equivalent a una extracció momentània. La nostra proposta té una evident analogia amb una validació creuada o *leave-one-out* (*LOO*). L'expressió de $e^2_{i(k)}$ corregida és la següent:

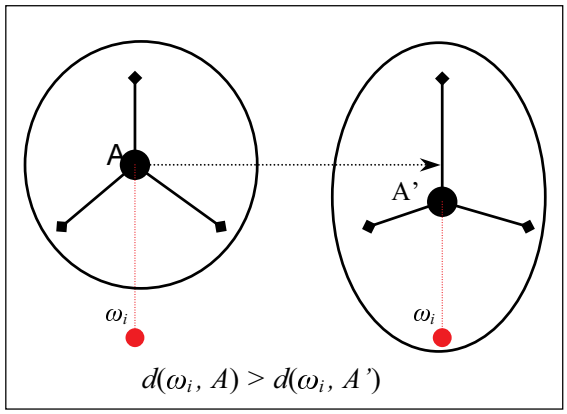


Figura 3.1.3: Atracció que genera l'objecte ω sobre el centroide A en ésser inclòs.

$$e^{2\,(LOO)}_{i(k)} = \left(\frac{n_k}{n_k - I[O_i \subset C_k]} \right)^2 e^2_{i(k)}$$

És evident que, amb aquesta correcció, la suma d'errors quadràtics serà més gran, degut a l'augment de la distància entre els objectes i el centroides. En el cas *crisp*, la relació entre *TESS* i $TESS^{(LOO)}$ és ben senzilla:

$$TESS^{\,(LOO)}_K = \sum_{k=1}^{K} E^{2\,(LOO)}_{(k)} = \sum_{k=1}^{K} \left(\frac{n_k}{n_k - 1} \right)^2 E^2_{(k)}$$

En definitiva, doncs, tots dos valors , *TESS* i $TESS^{(LOO)}$, no poden comparar-se directament, car obeeixen a criteris diferents d'avaluació de les distancies dels elements als clústers. A més, cal ressaltar que la correcció comporta la desaparició d'un clúster quan tingui un sol objecte. En aquest cas, seria recomanable un estudi detallat de l'individu per detectar si es tracta d'un *outlier*

(per exemple, comprovar l'efecte que tindria la seva eliminació sobre el *ESS* del clúster al que finalment hagi estat assignat). La correcció que proposem és aproximada en el cas de proximitats que necessitin paràmetres per a ésser calculades (com per exemple la Distància χ^2 o la distància de Mahalanobis), ja que l'extracció de l'individu comportaria calcular els paràmetres de nou.

3.1.3.4 *Fuzzy C-Means* (FCM)

Fuzzy C-means (*FCM*, Bezdek 1981) és una generalització de l'algorisme *K-means* a l'àmbit de la lògica difusa (Zadeh 1965). La idea original d'utilitzar conjunts difusos en tècniques d'anàlisi de clústers fou una proposta de Ruspini (1969, 1970), desenvolupada posteriorment per Dunn (1974) i Bezdek (1981) entre d'altres autors. La manera com *FCM* introdueix la lògica difusa al funcional de partició de *K-means* és (Bezdek 1981, 1987):

$$FTESS_{K,m} = \sum_{k=1}^{K} J_{(k),m}^{2} = \sum_{k=1}^{K} \sum_{i=1}^{N} u_{i(k)}^{m} \, e_{i(k)}^{2}$$

essent $u_{i(k)}$ la pertinença de l'element ω_i al conjunt difús Ω_k i $m \in (1, \infty)$ un exponent de *fuzziness* ("borrositat") que determina la incidència de las pertinences difuses a la partició. $e_{i(k)}^{2}$ té el mateix significat que per a *KM*, però el càlcul de la posició del centroide ara és:

$$\overline{x}_{(k)j} = \frac{\sum_{i=1}^{N} u_{i(k)}^{m} \cdot x_{ij}}{\sum_{i=1}^{N} u_{i(k)}^{m}}$$

La funció del grau de pertinença $u_{i(k)}$ s'obté mitjançant la optimització del funcional, utilitzant multiplicadors de Lagrange. L'equació resultant (Bezdek 1987) és:

$$u_{i(k)} = \frac{1}{\sum_{l=1}^{K} \left[\frac{e_{i(k)}}{e_{i(l)}} \right]^{2/(m-1)}}$$

que compleix la restricció **(b)**. En el cas indeterminat $e_{i(l)} = 0$, assignem arbitràriament la pertinença $u_{i(l)} = 1$ i $u_{i(k)} = 0$ per a tot $k \neq l$.

Observeu que, en el cas límit $m = 1$ s'obté de nou l'algorisme partitiu *hard* (*crisp*), és a dir, *K-Means:*

$$u_{i(k)} = \begin{cases} 1 & si \; e_{i(k)} = \min_{l} \{e_{i(l)}\} \\ 0 & si \; e_{i(k)} \neq \min_{l} \{e_{i(l)}\} \end{cases}$$

A la figura 3.1.4 mostrem l'efecte que provoca l'augment del paràmetre de *fuzziness* (m) sobre el resultat de *FCM*. Com més alt sigui l'exponent m més intermedis seran els valors de pertinença i, per tant, més difusa serà la partició. L'augment de m presenta dues avantatges i un inconvenient. En primer lloc, les pertinences per a valors m relativament alts indiquen per elles mateixes si els objectes són més propers al centroides o són més perifèrics en el clúster. En segon lloc, l'increment d'm també presenta una avantatge algorísmica, donat que el trànsit d'una solució a una altra es més suau que en el cas *crisp* (*a priori* aquest fet no resulta evident). El gran inconvenient en l'augment de m és que es pot arribar en un cas extrem a una situació totalment difusa en que totes les pertinences siguin $1/K$ per a tots els grups. En aquest cas els centroides de tots els grups s'apropen al centroide global de les dades.

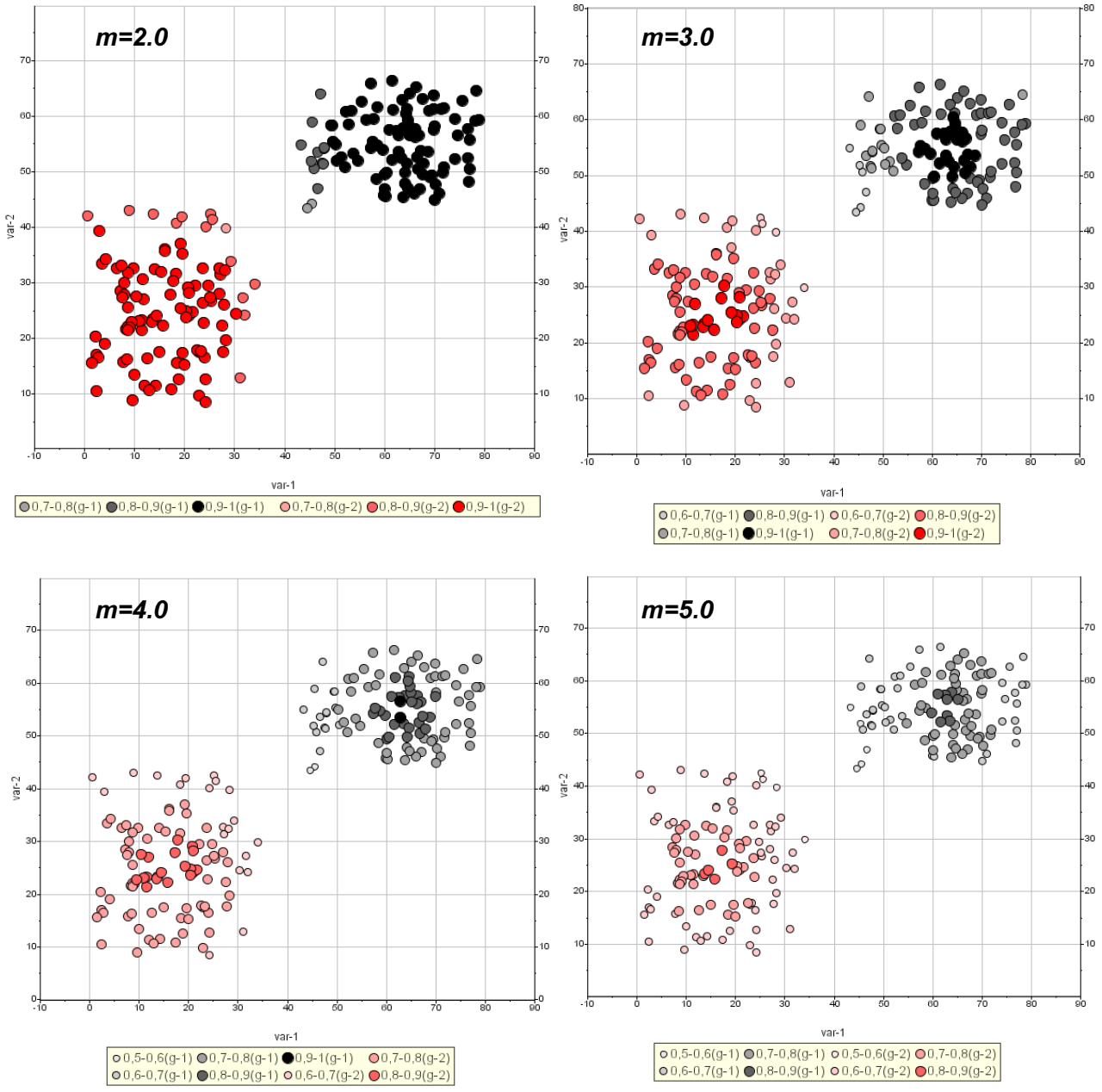


Figura 3.1.4: Els dos clústers de la figura 3.1.2 classificats amb l'algorisme *FCM*, emprant diferents nivells de m.

S'han proposat diverses extensions de *FCM* per a poder detectar estructures no esfèriques, donat que *FCM* té importants aplicacions en el reconeixement automatitzat de patrons. Per exemple, si emprem per al càlcul de la distància al centroide la mètrica de Mahalanobis en comptes de la pitagòrica (Bezdek & Pal 1992), podrem detectar millor el·lipses i línies (vegeu els exemples B i D de la figura 3.1.1). Per altra banda una mètrica radial ens permetria detectar estructures circulars (Krishnapuram 1992, vegeu l'exemple A de la figura 3.1.1).

FCM fou inicialment aplicat a l'anàlisi de comunitats vegetals per Marsili-Libelli (1989) i Equihua (1990) i la seva utilització es troba en expansió (Mucina 1997). Actualment, són força nombroses les aplicacions d'aquest algorisme (per exemple, Olano *et al.* 1998a, Podani 1990, Escudero & Pajarón 1994). Fins i tot seria possible que *FCM* acabés prenent el lloc de *TWINSPAN* com a algorisme més popular en l'anàlisi de clústers de comunitats vegetals. És important remarcar que, per tal d'aplicar convenientment *FCM* en dades de vegetació, cal emprar valors de l'exponent m propers a 1, normalment m=1.1 o m=1.2.

Dale (1995) troba dos inconvenients a l'algorisme *FCM*: 1) Es pot veure afectat per la contaminació de tipus de vegetació no ben representats en les dades mostrejades i amb centres externs a les mateixes. 2) L'algorisme estàndard no considera la possibilitat de variacions en les matrius de variàncies-covariàncies dels grups, pel que això imposa una restricció en la forma dels clústers.

3.1.3.5 Estratègies d'inicialització de *KM* i *FCM*

Una propietat important de *K-means* i *Fuzzy C-means* és que la solució (partició) final depèn de la configuració inicial (partició) dels clústers escollida, o, en altres paraules, dels centroides inicials escollits. L'algorisme pot convergir en molts casos a mínims locals del funcional. La configuració inicial dels clústers es pot establir seguint diferents criteris o estratègies:

- Subministrar una hipòtesi de partida a partir d'un coneixement *a priori* de les dades.
- Executar en primer lloc un algorisme de classificació jeràrquic. En general, és més convenient emprar un algorisme de clúster jeràrquic que no contragui o expandeixi l'espai de dades (*UPGMA*, *UPGMC*, Ward's, ...). L'arbre ultramètric resultant hauria d'ésser 'tallat' a un nivell de proximitat tal que en resultés una partició amb el nombre de clústers K desitjat (Milligan & Sokol 1980). Un problema d'aquesta estratègia és la excessiva dependència que el mètode partitiu pot presentar vers la solució jeràrquica de partida.

- Realitzar una partició en un nombre elevat de clústers. Seguidament, agrupar aquells dos grups més propers i tornar a repetir l'anàlisi de clúster partitiu. Continuar agrupant i executant l'algorisme fins a arribar al nombre de grups desitjat. Aquesta estratègia pot ser més costosa que l'anterior, però és menys depenent d'una estructura jeràrquica.
- Generar aleatòriament particions inicials diverses vegades. La millor partició final en termes de *TESS* és la solució a conservar. Aquesta estratègia conduirà en molts casos a generar centroides propers al centre de dades, pel que no sembla massa adient.
- Escollir aleatòriament objectes com a centroides inicials diverses vegades. Altra vegada, la millor partició final en termes de *TESS* és la solució a conservar. En general, com a exploració aleatòria, aquesta estratègia és més eficient que l'anterior.
- S'han proposat estratègies més complexes d'inicialització que permeten una exploració de l'espai de solucions del funcional més eficient. Aquestes inclouen tècniques com *simulated annealing* (Al-Sultan & Selim 1993), algorismes genètics, *tabu search* (Al-Sultan & Fedjki 1997), *variable neighbourhood search* (Hansen & Mladenovic 2001, Belacel *et al.* 2002) o xarxes neurals (Pedrycz 1998).

3.1.4 Model de classificació per blocs d'espècies/inventaris

Les taules fitosociològiques ordenades per grups d'espècies i inventaris foren introduïdes a la ciència de la vegetació pels fundadors de l'escola sigmatista. El mètode d'ordenació manual es pot trobar breument descrit a la introducció de la memòria. Amb l'arribada dels ordinadors es desenvoluparen força algorismes de classificació que intentaven emular aquest reordenament manual (per exemple Ceska & Roemer 1971, Moore 1973 in: Mucina & van der Maarel 1989, Wildi 1989). Una diferència fonamental entre aquests algorismes i els que hem descrit fins ara és la doble classificació de espècies i inventaris. Així, a la vegada es realitza una anàlisi de tipus Q (anàlisi d'objectes) i una anàlisi de tipus R (anàlisi d'atributs). Per altra banda, el concepte de clúster d'aquests mètodes no es basa en les relacions geomètriques dels inventaris. A continuació descrivim alguns mètodes d'aquest model.

3.1.4.1 *ESPRESSO* i *COCKTAIL*

L'algorisme *ESPRESSO* (Bruelheide & Flintrop 1994) es basa en la determinació de grups d'espècies i inventaris (blocs) que maximitzin la densitat d'aparicions dins dels blocs. El procediment utilitzat consisteix a eliminar les files/columnes menys denses fins a tenir una densitat interna de tots els elements de la submatriu resultant igual o més gran que un determinat llindar. Els blocs d'espècies/inventaris es generen un darrera l'altre. En conseqüència, la generació d'un bloc depèn dels blocs establerts prèviament. *COCKTAIL* (Bruelheide 2000) es un algorisme conceptualment semblant a *ESPRESSO*, però basat en la maximització del nombre de tàxons fidels. *COCKTAIL* produeix grups d'espècies la concurrència dels quals és més freqüent que l'esperada en cas de distribució aleatòria en el conjunt d'inventaris. Una característica important de *COCKTAIL* és que l'usuari preselecciona una espècie o un conjunt d'espècies per a actuar com a grup inicial, cosa que determina la composició final. A continuació, l'algorisme construeix un blocs d'espècies i inventaris alternant la determinació del grup d'inventaris i del grup d'espècies que formen el bloc. En les iteracions, noves espècies poden entrar a formar part del grup o en poden veure's excloses, en base a la fidelitat mesurada amb l'estadístic u (Bruelheide 2000, vegeu capítol 2.2). La mida del grup d'espècies es determina a través d'un llindar de fidelitat. Els inventaris que pertanyen al bloc s'estableixen comparant el nombre de tàxons del grup que presenten amb el que haurien de presentar si les espècies del grup es distribuïssin aleatòriament entre els inventaris. Una avantatge de *COCKTAIL* és la facilitat per tractar grans volums de dades. Es troba implementat al programa *JUICE* (Tichý 2002). En aquest estudi, no compararem aquí *COCKTAIL* amb els altres mètodes numèrics de classificació, degut a considerar que la selecció inicial del grup inicial de tàxons és un factor de subjectivitat que és determinant pel resultat de l'algorisme i posa en desavantatge els altres mètodes, els quals no utilitzen aquest coneixement *a priori* de les dades.

3.1.4.2 REBLOCK

REBLOCK (Podani & Feoli 1991) és un algorisme de classificació que realitza una doble partició en grups d'espècies i grups d'inventaris. Permet optimitzar tres funcions diferents:

1) Maximitzar la contingència entre blocs mitjançant l'estadístic χ^2 per a dades binàries:

$$\chi^2_{(K_1,K_2)} = \sum_{i=1}^{K_1}\sum_{j=1}^{K_2} \frac{\left(f_{ij} - \frac{f_{i\cdot}\cdot f_{\cdot j}}{f_{\cdot\cdot}}\right)^2}{\frac{f_{i\cdot}\cdot f_{\cdot j}}{f_{\cdot\cdot}}}$$

on K_1 i K_2 són el nombre de grups d'inventaris i tàxons respectivament, f_{ij} és el nombre de presències en el bloc ij. L'estadístic mesura la divergència de la distribució de blocs respecte la distribució esperada.

2) Minimitzar un estadístic d'informació per a caràcters nominals (qualitatius no ordinals):

$$H_{(K_1,K_2)} = \sum_{i=1}^{K_1}\sum_{j=1}^{K_2}\left(k_i \cdot k_j \cdot \log(k_i \cdot k_j) - \sum_{h=1}^{S}\log(f_{hij})\right)$$

on k_1 i k_2 són el nombre de variables del grup i i el nombre d'objectes del grup j, S és el nombre d'estats de caràcter i f_{hij} és la freqüència de l'estat h en el bloc ij. Cal remarcar que si apliquem aquest funcional a dades binàries, els estats presència i absència són tractats simètricament, mentre que no és així amb l'estadístic χ^2.

3) Minimitzar la suma de quadrats dins dels blocs per a variables quantitatives:

$$TESS_{(K_1,K_2)} = \sum_{i=1}^{N}\sum_{j=1}^{P}\left(x_{ij} - \bar{x}_{(K_1,K_2)ij}\right)^2$$

on $\bar{x}_{(K_1,K_2)ij}$ denota la mitjana dels valors en el bloc que conté l'element de la matriu ij. Noteu que, en el cas de un sol grup de variables, aquest funcional equival al de *K-means*.

Un inconvenient important de *REBLOCK* és la necessitat de definir a priori el nombre de grups, tan d'inventaris com de tàxons. L'algorisme *REBLOCK* parteix d'una doble partició inicial, sobre la qual a cada iteració recol·loca aquella variable o objecte que provoqui una millor optimització del funcional. Com en el cas dels algorismes partitius, la solució final de *REBLOCK* depèn de la doble partició inicial, pel que és necessari realitzar diferents execucions de l'algorisme partint de dobles particions generades aleatòriament.

REBLOCK ha estat utilitzat satisfactòriament per Pausas & Feoli (1996) per a estudiar les relacions indirectes entre la vegetació dels boscos pirinencs de *Pinus sylvestris* i variables químico-físiques ambientals.

3.1.5 Comparació i avaluació de classificacions

Un cop realitzada una anàlisi de clústers, o una anàlisi multivariant en general, és essencial disposar de mètodes per avaluar els resultats obtinguts i comparar els seus mèrits en relació amb altres anàlisis. Com en el cas dels mètodes de classificació, hom troba bibliografia sobre aquest tema publicada en molts camps diferents, essent el grau de formalisme també variable. Com a introducció a la comparació de resultats de mètodes multivariants en el camp de l'ecologia numèrica hom pot referir-se a les revisions de Rohlf (1974) i Podani (1989a).

3.1.5.1 Comparació de classificacions: mesures d'acord i classificacions consens

Com que les classificacions resultants de l'anàlisi de clústers estan influenciades per les diferents decisions que l'investigador ha pres en les etapes de mostratge i anàlisi, la comparació entre classificacions alternatives dels mateixos inventaris hauria de ser part integral dels estudis de vegetació. En aquest sentit són especialment útils les mesures d'acord entre parelles de classificacions, tema al que dedicarem els propers apartats.

D'altra banda, si hom està interessat en examinar l'acord general entre diferents classificacions alternatives de les mateixes dades, la construcció d'una classificació consens pot donar informació molt útil. En general les classificacions consens ens ajuden a no interpretar en excés els resultats d'una sola classificació. L'aplicació de mètodes consens no es restringeix a la comparació de diferents alternatives algorísmiques de classificació, sinó que també pot servir crear una classificació consens quan hom disposa de classificacions generades amb subconjunts de variables diferents.

En la comparació de dendrogrames, hi ha diferents aproximacions per a generar arbres consens. És usual la construcció d'arbres de consens estricte (*strict consensus trees*, Sokal & Rohlf 1981), on només es conserven aquells clústers comuns a tots els arbres originals. Leftkovich (1985) proposa una estratègia de construcció d'arbres consens a partir de les coordenades principals derivades de les matrius ultramètriques dels arbres originals.

Degut a la seva complexitat, no revisarem aquí la generació de particions consens. El lector interessat pot adreçar-se als treballs de Neuman & Norton (1986), Podani (1986, 1989b), per a particions *hard*, i Podani (1990) per a particions difuses.

3.1.5.2 Mesures d'acord entre particions *crisp*

Siguin $\mathbf{U}_{NxK}$ i $\mathbf{V}_{NxK'}$ dues particions del mateix conjunt d'objectes en un nombre de clústers K i K' respectivament, l'estudi de la semblança entre dues particions permet diferents aproximacions.

En una primera aproximació, hom genera una taula de classificacions creuades (*cross classification table*) o taula de confusió $\mathbf{T}_{KxK'}$ on es creuen les dues particions. Cada element $t_{kk'}$ de la taula conté el nombre d'objectes classificats al grup k d' $\mathbf{U}$ i al grup k' de $\mathbf{V}$. La simple inspecció de la matriu $\mathbf{T}$ ja proporciona molta informació de la relació entre les particions. A partir de la matriu $\mathbf{T}$ hom pot calcular coeficients de contingència, com l'estadístic χ^2:

$$\chi^2(\mathbf{U},\mathbf{V}) = \sum_{k=1}^{K}\sum_{k'=1}^{K'}\left[\frac{\left(t_{kk'} - \left(\frac{t_{k.}\cdot t_{.k'}}{t_{..}}\right)\right)^2}{\frac{t_{k.}\cdot t_{.k'}}{t_{..}}}\right]$$

que permet testar la independència de les particions. A la pràctica, testar la independència entre particions té relativament poc interès, perquè l'objectiu principal de la comparació és l'avaluació de la semblança o diferència. Alternativament, hom pot calcular mesures d'informació teòrica, com la divergència entre les dues classificacions (Feoli *et al*. 1984):

$$I(\mathbf{U},\mathbf{V}) = \sum_{k=1}^{K}\sum_{k'=1}^{K'} t_{kk'}\cdot\ln\left[\frac{t_{kk'}\cdot t_{..}}{t_{k.}\cdot t_{.k'}}\right]$$

La comparació de la correspondència es pot aplicar a la correlació de parelles de clústers, pel que hom pot, finalment obtenir una taula de mesures de correspondència entre els grups de les dues classificacions. Per exemple, Bruelheide & Chytry (2000) empren el coeficient Φ (vegeu el capítol 2.2), equivalent a una mesura de correlació per a dades binàries, per tal d' avaluar la correspondència entre classificacions.

Una altra aproximació per a mesurar l'acord entre particions implica el pas intermedi de descriure cada classificació a partir d'una matriu d'incidència $\mathbf{C}_{NxN}$. En una matriu d'incidència, cada casella c_{ij} pren el valor 1 si els objectes ω_i i ω_j pertanyen al mateix clúster, i 0 en cas contrari. A partir d'aquí, hom pot estudiar la relació entre les dues particions comparant les caselles de les respectives matrius d'incidència. Una de les mesures d'acord entre particions més emprades és l'índex de Rand (1971). Aquest índex calcula la probabilitat de que dos objectes agafats aleatòriament tinguin el mateix tractament en les dues particions.

Formalment, l'índex de Rand es defineix com a:

$$Rand(\mathbf{U},\mathbf{V}) = \frac{\sum_{i}^{N}\sum_{j<i}^{N}\gamma_{ij}}{\binom{N}{2}} \quad \text{, on}$$

a. $\gamma_{ij} = 1$ si ω_i i ω_j pertanyen a un mateix grup a **U** i **V**: $c(\mathbf{U})_{ij} = 1$ i $c(\mathbf{V})_{ij} = 1$.

b. $\gamma_{ij} = 0$ si ω_i i ω_j pertanyen a un grup diferent a **U** i a un mateix grup a **V**: $c(\mathbf{U})_{ij}=0$ i $c(\mathbf{V})_{ij} =1$.

c. $\gamma_{ij} = 0$ si ω_i i ω_j pertanyen a un mateix grup a **U** i a un grup diferent a **V**: $c(\mathbf{U})_{ij}=1$ i $c(\mathbf{V})_{ij}=0$.

d. $\gamma_{ij} = 1$ si ω_i i ω_j pertanyen a un grup diferent, tant a **U** com a **V**: $c(\mathbf{U})_{ij}=0$ i $c(\mathbf{V})_{ij}=0$.

L'índex de Rand equival al càlcul del *simple matching coefficient* (Sokal & Michener 1958) sobre una taula de contingència dos per dos, taula on cada casella conté el nombre de casos en que es dóna (a), (b), (c) o (d). Per tant, una altra notació de l'índex de Rand és:

$$Rand(\mathbf{U},\mathbf{V}) = (a+d)/(a+b+c+d)\,.$$

Tal i com indica Podani (1986), hom podria calcular altres índexs binaris, basant-se en la mateixa taula 2x2 com, per exemple, l'índex de Jaccard (1901) o el de Sørensen (1948).

Es pot demostrar amb certa facilitat que l'índex de Rand es pot expressar també a partir de la taula de classificacions creuades **T**:

$$Rand(\mathbf{T}(\mathbf{U},\mathbf{V})) = \frac{\binom{N}{2} + \sum_{i=1}^{K}\sum_{j}^{K'} t_{ij}^2 - \frac{1}{2}\left(\sum_{i=1}^{K} t_{i\cdot}^2 + \sum_{j=1}^{K'} t_{\cdot j}^2\right)}{\binom{N}{2}} \quad \text{, on } \binom{x}{2} = x\cdot(x-1)/2\,.$$

A partir d'aquesta darrera notació, Hubert & Arabie (1985) proposaren corregir l'índex de Rand per treure'n l'acord entre particions degut únicament a l'atzar, considerant una distribució hipergeomètrica de les particions. L'índex de Rand corregit és:

$$CorrectedRand(\mathbf{T}(\mathbf{U},\mathbf{V})) = \frac{Rand - Expected(Rand)}{Maximum(Rand) - Expected(Rand)}$$

$$= \frac{\sum_{i=1}^{K}\sum_{j}^{K'}\binom{t_{ij}}{2} - \sum_{i=1}^{K}\binom{t_{i\cdot}}{2}\cdot\sum_{j=1}^{K'}\binom{t_{\cdot j}}{2}\Big/\binom{N}{2}}{\frac{1}{2}\left(\sum_{i=1}^{K}\binom{t_{i\cdot}}{2} + \sum_{j=1}^{K'}\binom{t_{\cdot j}}{2}\right) - \sum_{i=1}^{K}\binom{t_{i\cdot}}{2}\cdot\sum_{j=1}^{K'}\binom{t_{\cdot j}}{2}\Big/\binom{N}{2}} \quad \text{, on } \binom{x}{2} = x\cdot(x-1)/2\,.$$

L'índex de Rand corregit ateny un valor 0 quan el nombre de coincidències equival a les esperades únicament per atzar (model nul). Un valor 1 implica, com en el cas no corregit, una correspondència perfecte entre els grups de les dues particions comparades. L'índex de Rand, corregit o no, presenta l'avantatge, respecte altres estratègies de comparació, de poder ésser calculat quan el nombre de clústers d'una partició és diferent del nombre de clústers de l'altra (K pot ésser diferent de K').

Una darrera aproximació a l'avaluació de la proximitat entre dues particions consisteix en expressar la distància entre les particions en termes del nombre mínim de transformacions admisibles necessari per transformar una partició en l'altra (*MINDMT: minimum number of divisions, mergences and transfers,* Day 1981, Podani 1986).

3.1.5.3 Mesures d'acord entre particions difuses. Generalització de l'índex de Rand.

Per tal de comparar particions difuses, una primera aproximació consisteix en aplicar els mètodes *crisp* després de convertir les particions difuses en particions *crisp*, procés que, en anglès, s'anomena *defuzzification*. La *defuzzification* es pot realitzar de dues maneres bàsiques: 1) Per a cada objecte escollir el grup amb més pertinença, al qual s'hi assigna un valor 1 i a la resta de grups zero. 2) A partir d'un "tall α" (α-*cut*), és a dir, definint una matriu de pertinença *crisp* on el valor d'una casella és 1 si a la corresponent casella de la matriu difusa la pertinença és igual o superior a α (Bodjanova 1999). El segon mètode pot donar lloc a particions degenerades (particions que no compleixen la condició **b** de l'apartat 3.1.3.1). En conseqüència, l'estratègia de *defuzzification* més freqüent és la primera.

Si hom vol valorar l'acord entre particions difuses sense haver de perdre informació en la *defuzzification*, cal disposar de mesures d'acord que puguin tractar valors de pertinença difusos. En aquest sentit, Podani (1990) proposa estendre *MINDMT* al cas difús. Concretament, aquest autor proposa comprovar totes les permutacions possibles de les columnes de $\mathbf{U}$ respecte $\mathbf{V}$ per trobar aquella permutació que minimitzi la distància:

$$d^2(\mathbf{U},\mathbf{V}) = \sum_{i=1}^{N}\sum_{k=1}^{K}(u_{ij} - v_{ij})^2$$

L'aproximació *MINDMT* presenta l'inconvenient de requerir el mateix nombre de classes a $\mathbf{U}$ i $\mathbf{V}$. Podani (1990) argumenta que les altres mesures d'acord per a particions *crisp* no són aplicables a particions difuses ja que requereixen classes determinístiques. No obstant, segons proposem a continuació, la generalització és possible si expressem $\mathbf{T}$ i $\mathbf{C}$ en notació matricial. Efectivament, el càlcul de la taula de contingència, $\mathbf{T}$, es pot expressar en notació matricial a partir del producte de les matrius de pertinença: $\mathbf{T}(\mathbf{U},\mathbf{V}) = \mathbf{U}'\cdot\mathbf{V}$. Anàlogament, el càlcul de la matriu d'incidència, $\mathbf{C}$, es pot expressar: $\mathbf{C}(\mathbf{U}) = \mathbf{U}\cdot\mathbf{U}'$. Aquest notació matricial del càlcul de $\mathbf{C}$ i $\mathbf{T}$, obre la porta a calcular estadístics de comparació de particions *crisp* sobre particions difuses no degenerades. Especialment interessant és l'extensió de l'índex de Rand al cas difús, on el càlcul de les caselles de la taula de contingència dos per dos seria:

$$a = \sum_{i=1}^{N}\sum_{j=i+1}^{N} c(\mathbf{U})_{ij}\cdot c(\mathbf{V})_{ij} \qquad b = \sum_{i=1}^{N}\sum_{j=i+1}^{N}\left(1 - c(\mathbf{U})_{ij}\right)\cdot c(\mathbf{V})_{ij}$$

$$c = \sum_{i=1}^{N}\sum_{j=i+1}^{N} c(\mathbf{U})_{ij}\cdot\left(1 - c(\mathbf{V})_{ij}\right) \qquad d = \sum_{i=1}^{N}\sum_{j=i+1}^{N}\left(1 - c(\mathbf{U})_{ij}\right)\cdot\left(1 - c(\mathbf{V})_{ij}\right)$$

És fàcil comprovar que aquesta definició dels valors de les caselles de la taula de contingència dos per dos generalitza el cas *crisp* i compleix, en el cas de particions no degenerades, que $(a+b+c+d) = (N \cdot (N - 1)) / 2$. Aquesta generalització de l'índex de Rand també es pot estendre a l'índex de Rand corregit per l'atzar de Hubert & Arabie (1985). A les figures 3.1.6 i 3.1.7 mostrem dos casos senzills per exemplificar de les generalitzacions proposades.

	Partition U			**Partition V**	
	$\mathbf{U}_1$	$\mathbf{U}_2$		$\mathbf{V}_1$	$\mathbf{V}_2$
ω_1	0	1	ω_1	0	1
ω_2	0	1	ω_2	0	1
ω_3	1	0	ω_3	1	0
ω_4	1	0	ω_4	1	0
ω_5	0	1	ω_5	0	1
ω_6	0	1	ω_6	**1 - p**	**p**

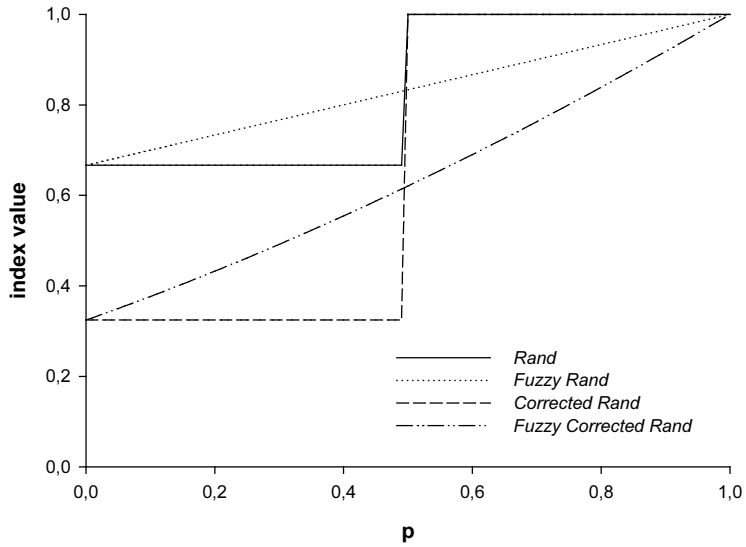


Figura 3.1.6: Exemple de generalització de l'índex de Rand al cas difús. A l'esquerra hi ha dues particions (**U** i **V**) de dimensions 6x2. Les particions difereixen tan sols en la classificació, a la partició **V**, del darrer objecte , ω_6, en que la pertinença a ambdós grups depèn del paràmetre **p**. El diagrama de la dreta mostra els valors obtinguts per a l'índex de Rand amb i sense correcció, en les versions *crisp* i *fuzzy*. En els extrems (p = 0 o p = 1), la versions *fuzzy* dels indexos coincideix amb les versions *crisp*. D'altra banda, per valors de pertinença difusa de ω_6, el càlcul *fuzzy* dels indexos permet un trànsit gradual entre les situacions extremes. Les versions de l'índex corregides per l'atzar, presenten lògicament, valors inferiors que les no corregides.

	Partition U			**Partition V**	
	$\mathbf{U}_1$	$\mathbf{U}_2$		$\mathbf{V}_1$	$\mathbf{V}_2$
ω_1	0	1	ω_1	p	1.0 - p
...	...	...	...	...	...
$\omega_{N/2-1}$	0	1	$\omega_{N/2-1}$	p	1.0 - p
$\omega_{N/2}$	1	0	$\omega_{N/2}$	1.0 - p	p
...	...	...	...	...	...
ω_N	1	0	ω_N	1.0 - p	p

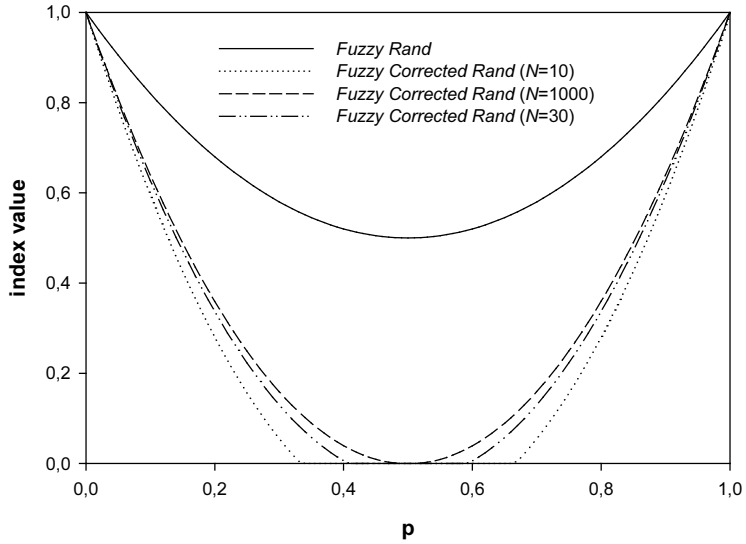


Figura 3.1.7: Exemple de generalització de l'índex de Rand al cas difús. A l'esquerra hi ha dues particions (**U** i **V**) de dimensions *N*x2. **U**, és en una partició en dos grups de mida *N*/2. A la partició **V**, la pertinença a ambdós grups depèn del paràmetre 'p'. El diagrama de la dreta mostra els valors obtinguts per a l'índex de Rand *fuzzy* amb i sense correcció. En els extrems (p = 0 o p = 1) les particions coincideixen i l'índex és 1.0. Per a la zona més confusa d'**V** (p =0 .5) l'índex sense corregir esdevé 0.5 i el corregit 0.0, independentment del nombre d'objectes de la partició. Aquest fet reflexa la indeterminació de la matriu **V**, que des d'un punt de vista *crisp* correspondria a l'assignació aleatòria de cada objecte a un dels grups. El resultat 0.0 de l'índex de Rand difús corregit és acceptable i desitjable.

3.1.5.4 La comparació de dendrogrames

Els primers assaigs de comparació entre dendrogrames consistien en estudiar la correlació entre les matrius ultramètriques associades a cada arbre (*cophenetic correlation*, vegeu Robertson 1979). La mesura de correlació pot ser el coeficient de correlació lineal de Pearson o el coeficient de correlació per rangs de Spearman. Un dels inconvenients d'aquest mètode de comparació, aplicable a la correlació de matrius simètriques en general, és que tracta els elements de les matrius com si fossin independents quan en realitat no ho són. A més, en el cas de la comparació de matrius ultramètriques, la comparació per correlació pondera en excés la coincidència dels clústers de més alt rang, ja que aquests estan representat per més elements a les matrius (Rolhf 1982). Malgrat reconèixer aquests inconvenients, Podani & Dickinson (1984) precisen que les correlacions entre matrius ultramètriques poden ésser realitzades a partir de cinc "descriptors" dels dendrogrames, que reflecteixen diferents aspectes de la estructura dels mateixos (per a més informació sobre aquests descriptors consulteu Podani 2000).

Alternativament a la comparació de matrius ultramètriques, hom pot realitzar "talls" en un determinat nivell de similaritat als dos arbres jeràrquics a comparar, i aplicar a continuació mesures d'acord a les particions resultants. Aquestes línies de "tall" s'anomenen també *phenon lines*. Fowlkes & Mallows (1983) proposaren comparar dos arbres jeràrquics realitzant "talls" a diferents nivells ($K = 2, 3, \ldots, N - 1$), calculant a continuació un índex de semblança entre particions. Com a resultat Fowlkes & Mallows obtingueren perfils de semblança entre els dos arbres. Malgrat Fowlkes & Mallows (1983) proposaren un índex propi de comparació de particions, altres índexs de comparació de particions serien igualment aplicables en la generació d'aquests perfils. D'altra banda, si la mesura d'acord emprada permet un nombre de grups diferent entre les dues classificacions, com amb l'índex de Rand o el de Fowlkes & Mallows, aquesta estratègia de comparació també fa possible la comparació entre una partició i un dendrograma. Sovint el que més interessa en estudis aplicats és la comparació d'aquest "tall" del dendrograma a un nivell determinat, mentre que l'estructura sencera del dendrograma pot tenir un interès menor.

És important remarcar que la validesa de les *phenon lines* en algorismes jeràrquics que, com *single linkage* o *complete linkage*, contraguin o dilatin l'espai de dades, és dubtosa degut, precisament, a la deformació de les relacions al voltant dels clústers.

3.1.5.5 Avaluació de classificacions

L'estructura de classificació generada per un algorisme de *clustering* no té perquè ser aquella que les dades exhibeixen de manera "natural" per elles mateixes. Els mètodes de *clustering* estan, doncs, limitats pel model de classificació que generen i pels paràmetres utilitzats en l'anàlisi. Per exemple, en una execució de *K-means* amb el paràmetre K=3 grups, la partició final serà una partició en tres grups independentment de si les dades exhibeixen tres grups de forma natural o n'exhibeixen només dos. Una classificació resultarà més adequada quan la estructura de grups reflecteixi una informació veritable de les dades. Per tant, és molt important poder avaluar la significació de les estructures generades en l'anàlisi de clústers.

Segons Jain & Dubes (1988), podem emprar tres tipus de criteris, que expressen la nostra estratègia de validació:

1. Els ***criteris externs*** utilitzen una estructura coneguda *a priori* que s'empra com a patró de comparació. Aquesta estructura pot provenir de la aplicació de mètodes de *clustering* amb un conjunt de variables diferent o del coneixement previ dels objectes a classificar. La validació, es fa emprant mètodes de comparació, com els esmentats als apartats anteriors. Alternativament, en el cas de la classificació de comunitats de vegetals hom pot utilitzar una anàlisi discriminant amb variables ambientals com a criteri extern (p.e. Hakes 1994, Pausas & Feoli 1996).
2. Els **criteris *relatius*** decideixen quina, d'entre dues o més estructures, és millor en algun sentit, tal com ésser més estable o més apropiada per a les dades. Un exemple aplicat d'aquest tipus de raonament es pot trobar a Anderson & Clements (2000).
3. Els ***criteris interns*** estimen l'acord entre l'estructura obtinguda i les dades emprant només les dades amb que s'ha obtingut la classificació. En els treballs que estudien tests de significació de clúster cal establir distribucions de referència amb les que comparar els clústers. Com que hom sovint desconeix la distribució teòrica de les variables, les distribucions de referència s'obtenen mitjançant mètodes d'aleatorització de les dades (mètodes de Montecarlo, *bootstrap*,...). En aquest sentit, destaquem els treballs de Strauss (1982), Bock (1985), Nemec & Brinkhurst (1988), Pillar & Orlóci (1996) i Pillar (1999a, 1999b).

L'estratègia d'avaluació de classificacions serà diferent depenent del model de classificació emprat en generar les classificacions que es pretenen comparar. Ja hem revisat en anteriors apartats les mesures d'acord entre classificacions. Utilitzarem aquestes mesures quan el criteri d'avaluació d'una partició sigui una classificació externa. En els propers apartats presentarem alguns estadístics que poden ésser emprats com a criteri intern per a avaluar classificacions dels mètodes *K-means*, *FCM*, i REBLOCK. Concretament, aquests estadístics permeten determinar quin nombre de grups és el més idoni a cercar en les dades de que es disposa.

3.1.5.6 Determinació del nombre de grups a *K-means*

Un dels paràmetres de qualsevol algorisme partitiu és la selecció del nombre de grups que contenen les dades, K. Rarament en una situació d'anàlisi de clústers es coneix *a priori* el nombre de grups a cercar. La estratègia a seguir és provar diferents valors de K i comparar, mitjançant algun criteri intern, la idoneïtat de les particions resultants. Quan els estadístics de selecció del nombre de grups s'apliquen als resultats de mètodes jeràrquics, s'anomenen regles d'aturada (*stopping rules*). En el cas dels algorismes jeràrquics divisius com *association analysis* (Williams & Lambert 1959), la "regla d'aturada" era sovint la significació de l'estadístic χ^2.

L'estratègia per a determinar el nombre de grups està molt lligada al mètode de *clustering* utilitzat i al model de clúster subjacent. No pretenem aquí revisar totes les propostes, que són nombroses, sinó centrar-nos en aquelles aplicables als resultats de *K-means* i, per extensió, a *FCM*. Un bon resum de les estratègies possibles es pot trobar a Gordon (1999).

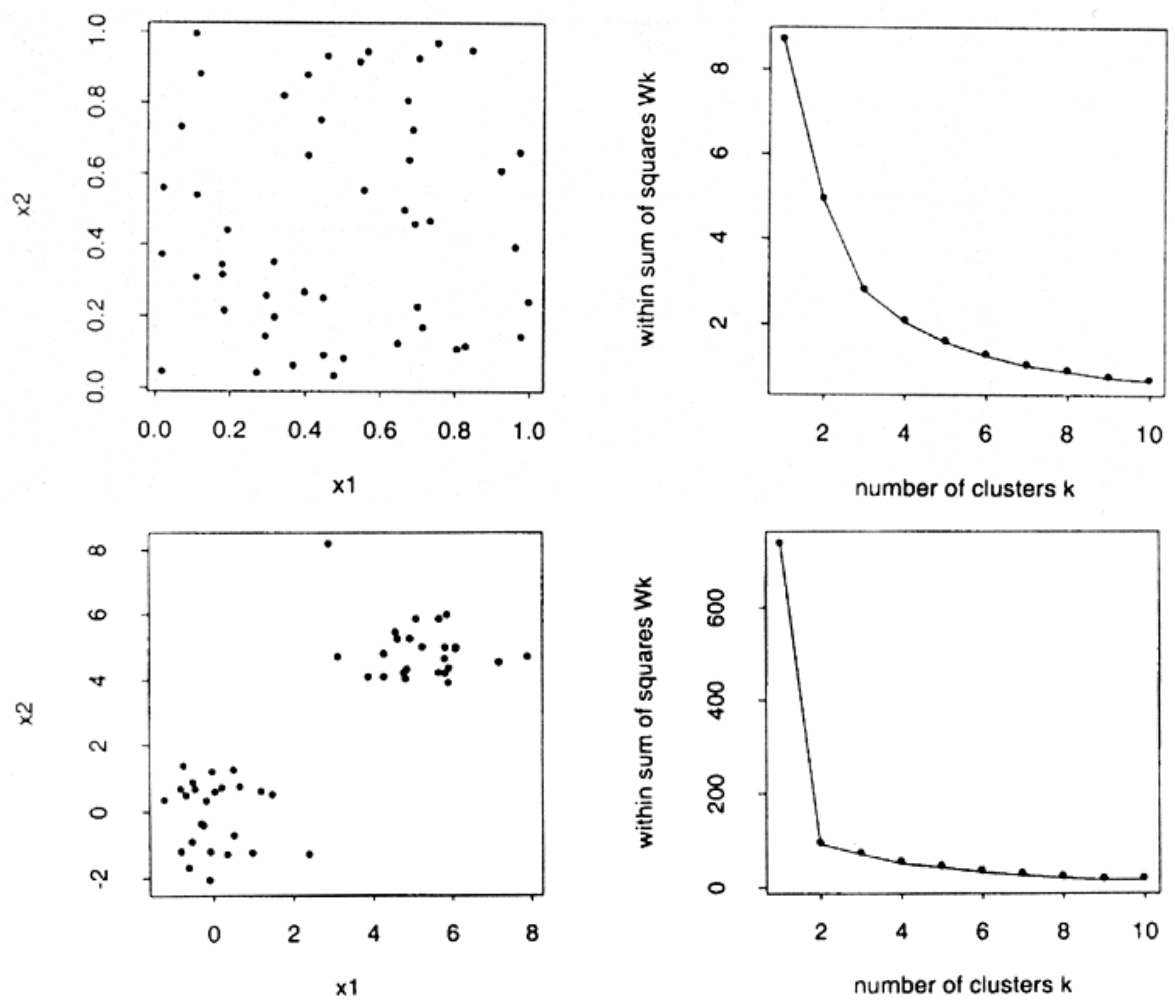


Figura 3.1.5: Distribució de punts aleatòria (superior esquerre) i en 2 *clusters* (superior dreta). Corba de decreixement de la suma d'errors quadràtics total (*TESS*) per a les corresponents situacions inferior (extret de Tibshirani *et al.* 2001).

A la figura 3.1.5 reproduït un exemple de Tibshirani *et al.* (2001), on es mostra l'evolució de *TESS* (W_k) per a diferent nombre de clústers en una situació de distribució aleatòria (a dalt) i una distribució en 2 clústers (a baix). La suma d'errors quadràtics (*TESS*) disminueix invariablement en augmentar el paràmetre K, tant si l'estructura existeix com si no. No obstant, en el cas amb estructura la davallada inicial és més forta i després la caiguda és molt suau. Aquest "colze" és el que convé detectar amb un estadístic apropiat (*elbow problem*). Així doncs, la majoria de estadístics convencionals de detecció del nombre de grups adequat es basen d'una manera o l'altra amb la variabilitat dels clusters.

Un dels estadístics de selecció de K més freqüentment implementat en programes d'estadística multivariant és el criteri de maximització de pseudo-F (Calinsky-Harabasz 1974):

$$F(K) = \frac{\mathbf{A}_K /(K-1)}{\mathbf{W}_K /(N-K)}$$

D'altra banda, Hartigan (1975), proposà minimitzar l'estadístic:

$$H(K) = \left[\frac{\mathbf{W}_K}{\mathbf{W}_{K+1}} - 1 \right] \Big/ (N-K-1)$$

$H(K)$ presenta l'avantatge de poder ésser calculat per a K=1, permetent concloure que les dades no expressen cap clúster. En un important estudi de simulació Milligan & Cooper (1985) examinaren l'efectivitat de 30 criteris diferents per determinar el nombre de clústers idoni. Entre els seus resultats recomanaren l'ús de $F(K)$ de Calinsky-Harabasz. Cal remarcar que l'estudi de simulació generava clústers amb distribucions normals multivariants, tal i com es descriu a Milligan (1985), cosa que afavoreix estadístics basats en la normalitat de les dades, tal i com els mateixos autors advertien.

Posteriorment a l'estudi de Milligan & Cooper, Krzanowski & Lai (1988) proposaren un altre estadístic, més elaborat que els comparats per Milligan & Cooper (1985):

$$KL(K) = \left| \frac{DIFF(K)}{DIFF(K+1)} \right| \text{ on } DIFF(K) = (K-1)^{2/P} \cdot \mathbf{W}_{K-1} - K^{2/P} \cdot \mathbf{W}_K$$

Cal remarcar que aquest índex requereix conèixer el nombre de variables, P, per a ésser calculat. Si la variància estigués calculada a partir de distàncies entre objectes (vegeu capítol 3.3) aquest estadístic no es podria calcular. Recentment, Tibshirani *et al.* (2001) proposen un nou estadístic, anomenat *gap*:

$$Gap(K) = E^*\{\log(TESS)\} - \log(TESS)$$

on $E^*\{\log(TESS)\}$ és el valor esperat de $\log(TESS)$ segons una distribució de referència generada per mètodes de Montecarlo.

Els estadístics esmentats fins ara utilitzen tots la dispersió de les dades (*TESS*) com a font de informació primària i tots funcionen millor si les poblacions mostrejades tenen una distribució

aproximadament normal multivariant. Per desgràcia acostumen a presentar-se ineficaços per al *clustering* de dades de distribució multivariant diferent o desconeguda. Altres aproximacions, més *ad hoc*, però encara basades en la dispersió "dins" i "entre" grups, es troben als treballs d'Orloci (1967), Hogeweg (1976), Ratcliff & Pieper (1981), Popma *et al*. (1983) i Mucina & Hauser (1993). Una possible avantatge d'aquestes aproximacions és que sovint restringeixen les comparacions "entre" grups als veïns més propers, donant més importància a l'aïllament local dels grups.

Una possibilitat no paramètrica, però sí geomètrica, és fer una representació gràfica de la partició que indiqui, per a cada objecte si es troba geomètricament a prop del centroide del clúster al qual ha estat assignat o es troba més proper a altres centroides. Aquesta és la definició de les siluetes (*silhouettes*), introduïdes per Rousseeuw (1987) i emprades en utilitzades en els algorismes descrits a Kaufman & Rousseeuw (1990). Per a cada element ω_i hom calcula la seva silueta, s_i, de la següent manera:

1) Denotem Ω_a el clúster al que ha estat assignat ω_i. Definim a_i com la dissimilaritat mitjana entre ω_i i els altres elements d' Ω_a (suposant que n'hi ha més).
2) Definim $d_i(k)$ com la dissimilaritat mitjana entre ω_i i els elements d'Ω_k $(k \neq a)$.
3) Seleccionem el mínim de les dissimilaritats mitges anteriors: $b_i = \min_{k \neq a} d_i(k)$. El clúster corresponent a aquest mínim és el clúster veí al que ha estat assignat l'objecte, Ω_b.
4) s_i s'obté, finalment, combinant a_i i b_i:

$$s_i = \frac{b_i - a_i}{max\{a_i, b_i\}}$$

De la definició anterior, de seguida veiem que $-1 \leq s_i \leq 1$. Valors propers a zero ens indiquen que no és clar si l'objecte hauria d'ésser assignat a Ω_a o a Ω_b. Valors positius indiquen una bona classificació, mentre que valors negatius probablement indiquen que l'objecte ha estat mal classificat (segons aquest criteri). La representació gràfica de la silueta d'un clúster es dibuixaria considerant els elements com files i estenent una línia per cada element proporcional al seu valor s_i. La mitjana aritmètica d' s_i per a tots els elements d'un clúster és la silueta mitjana del clúster. Anàlogament, la mitjana aritmètica d' s_i per a tots els elements és la silueta mitjana de la partició.

3.1.5.7 Avaluació d'una partició difusa

Paral·lelament al desenvolupament d'algorismes partitius difusos es proposaren alguns índexs per avaluar a "borrositat" (*fuzziness*) de la partició resultant (Dunn 1976). Per a una revisió recent del tema, vegeu Kim *et al*. (2004). El coeficient de partició (*partition coefficient* o *PC,* Bezdek 1974, 1981) també anomenat coeficient de Dunn, mesura el grau de 'duresa' o 'borrositat' d'una partició:

$$PC(K,\mathbf{U}) = \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{N} u_{i(k)}^2$$

Per a una partició completament difusa $PC(K,\mathbf{U}) = 1/K$, mentre que per a una partició *crisp* $PC(K,\mathbf{U}) = 1$. Una versió normalitzada del coeficient de partició o de Dunn és:

$$PC_n(K,\mathbf{U}) = \frac{PC(K,\mathbf{U}) - (1/K)}{1-(1/K)} = \frac{K\cdot PC(K,\mathbf{U}) - 1}{K-1}$$

Una altra mesura interna difusa és l'entropia de la partició (Ruspini 1969), mesura de la informació continguda en una partició difusa:

$$H(K,\mathbf{U}) = -\frac{1}{N\cdot \log_a(K)}\sum_{k=1}^{K}\sum_{i=1}^{N} u_{i(k)}\cdot \log_a(u_{i(k)})$$

Anàlogament, si normalitzem l'entropia de la partició dividint pel valor màxim, obtenim l'entropia normalitzada o eficiència de la partició (Dunn 1976):

$$H_n(K,\mathbf{U}) = \frac{H(K,\mathbf{U})}{(1-K/N)}$$

Noteu que la eficiència de la partició només és significativament diferent de l'entropia quan K s'aproxima a N. Tant el coeficient de partició normalitzat com l'entropia normalitzada són estadístics que podem utilitzar per decidir el nombre de grups en particions difuses (Dunn 1976, Marsili-Libelli 1989, Equihua 1990). Escollirem aquell valor K que generi particions amb un coeficient de partició més elevat o una entropia més baixa.

Noteu, però, que aquests coeficients de "borrositat" només utilitzen els valors de pertinença com a font d'avaluació, però no l'estructura geomètrica de les dades. Per altra banda, és relativament senzill estendre al cas difús els estadístics de determinació del nombre de grups basats en la partició de la variància. Finalment, voldríem remarcar que és possible generalitzar al cas difús les siluetes, simplement redefinint a_i i b_i del final de l'apartat anterior:

$$a_i = \frac{\sum_{j=1}^{N} u_{j(a)}\cdot d_{ij}}{\sum_{j=1}^{N} u_{j(a)}} \quad \text{i} \quad b_i = \min_{k\neq a}\left[\frac{\sum_{j=1}^{N} u_{j(k)}\cdot d_{ij}}{\sum_{j=1}^{N} u_{j(k)}}\right].$$

En referència a aquesta possibilitat no hem trobat a la bibliografia cap aplicació d'aquesta generalització de la silueta com a criteri intern de determinació del nombre de clústers idoni.

3.1.5.8 Avaluació i comparació dels resultats de *REBLOCK*

Podani & Feoli (1991) proposen comparar les dobles particions que genera *REBLOCK* mitjançant una extensió de *MINDMT* (vegeu apartat 3.1.5.2). La avaluació de les particions es faria, segons aquests autors, amb una anàlisi de concentracions (*Analysis Of Concentrations*, *AOC*, Feoli & Orlóci 1979). En la presentació d'aquesta tècnica (Feoli & Orlóci 1979), l'estadístic $\chi^2_{(K_1,K_2)}$ fou emprat per a avaluar la agudesa (*sharpness*) de dobles particions de taules fitosociològiques. Feoli & Orlóci (1979) també proposaren una mesura d'agudesa relativa:

$$C = \chi^2_{(K_1,K_2)} / \left[f_{..} \cdot \min(K_1 - 1, K_2 - 1) \right].$$

Creiem que la maximització d'aquesta mesura de divergència relativa pot ser un bon punt de partida per a establir un criteri intern d'avaluació dels resultats de *REBLOCK* per a les diferents combinacions de K_1 i K_2 . No obstant, l'estadístic C presenta l'inconvenient de ser mínim quan $K_1=K_2$, pel que aquestes combinacions quedaran excloses. Per aquest motiu creiem que és millor utilitzar:

$$C = \chi^2_{(K_1,K_2)} \Big/ \left[f_{..} \cdot \sqrt{(K_1 - 1) \cdot (K_2 - 1)} \right].$$

Aquest segon criteri, per contra, té la tendència a presentar un màxim quan $K_1=K_2$, però considerem que aquesta biaix està d'acord amb el concepte de clúster del mètode *REBLOCK*.

3.1.5.9 Estudis de comparació i avaluació de mètodes de *clustering* de comunitats

És important remarcar el baix consens metodològic que existeix a l'hora d'avaluar i comparar els resultats produïts per algorismes d'anàlisi de clústers. Amb l'objectiu de descriure la bibliografia consultada sobre el tema, hem dividit els treballs que incloïen comparacions i/o avaluacions en dos blocs: (A) Els que comparen els resultats de diferents mètodes *clustering* amb l'objectiu determinar quin d'ells és el més idoni per a l'anàlisi de la vegetació. (B) Els que comparen la classificació obtinguda per un mètode numèric amb la classificació tradicional sigmatista.

A. Comparació dels resultats de diferents mètodes numèrics.

La comparació de l'eficiència relativa entre mètodes numèrics de *clustering* es pot fer en base a la seva aplicació en dades simulades o en exemples reals. Robertson (1979) comparà tres mètodes jeràrquics - el mètode de Ward (1963), *MINFO* (un mètode partitiu que fa ús de la teoria de la informació, vegeu Orloci 1978) i *UPGMA* - a partir de la simulació de gradients ecològics amb campanes de Gauss (Gauch & Whittaker 1972). Robertson estimà *UPGMA* com a menys eficient que els altres dos.

Gauch & Whittaker (1981) compararen diferents algorismes jeràrquics de classificació concloent que *TWINSPAN* era la millor eina per dades complexes o amb soroll. Entre les raons que donaren hi havia: a) Utilitza la matriu original d'espècies/inventaris en comptes d'una matriu secundària. b) Ordena la seqüència de mostres en un dendrograma. c) Agrupa tant inventaris com espècies. d) Produeix una matriu de dades reordenada. e) Té uns requeriments computacionals baixos que augmenten tan sols linealment amb el volum de dades.

Belbin & McDonald (1993) compararen *TWINSPAN* amb *UPGMA* i el mètode partitiu *ALOC*, mitjançant dades simulades amb el programa *COMPAS* (Minchin 1987). En el model de simulació de *COMPAS* la resposta de les espècies als gradients ecològics s'aproxima amb funcions beta (Austin *et al.* 1994). Belbin & McDonald conclogueren que *TWINSPAN* produïa pitjors resultats que els altres dos models. Segons aquests autors, l'algorisme *TWINSPAN* té dos problemes principals, la dependència d'un gradient primari predominant i la dicotomització errònia en aquest. Cal esmentar aquí que la validesa de l'estudi de Belbin & McDonald (1993) ha estat posteriorment força discutida (Dale 1995, Cao *et al*. 1997).

És important remarcar que els estudis amb dades simulades, malgrat permeten comparar els resultats amb la classificació estimada 'veritable', no asseguren el realisme de les dades simulades. Concretament, en el cas de la simulació amb *COMPAS*, el realisme de les dades simulades passa per la idoneïtat de diferents factors: la teoria del nínxol sobre la que es basa la resposta de les espècies, la modelització del tipus de resposta fisiològica de les plantes als gradients ecològiques mitjançant funcions beta (Oksanen 1997), la modelització de les interaccions entre plantes, el mostratge aplicat als gradients teòrics generats... En suma, és palès que, malauradament, el coneixement teòric actual de la distribució dels tàxons en comunitats és probablement insuficient per a generar dades simulades amb un cert grau de confiança respecte al seu realisme.

Més recentment, Torres *et al*. (1995) comparen els resultats de classificació d'un gran nombre de combinacions de mesures de proximitat i mètodes jeràrquics aglomeratius a partir d'exemples de vegetació reals. Torres *et al*. comparen l'eficiència dels algorismes mitjançant mesures pròpies d'avaluació dels resultats i conclouen que el mètode de Ward's és el que donava millors resultats, en combinació amb la mesura de proximitat *similarity ratio*.

Cao *et al*. (1997) comparen l'habilitat de quatre mètodes jeràrquics (UPGMA, Ward, *complete linkage* i TWINSPAN) per a reconèixer estructures en tres conjunts de dades de comunitats de macroinvertebrats diferents. Aquests autors avaluen els resultats comparant-los amb classificacions pre-definides en base a les característiques biològiques i de l'hàbitat, suportades per mètodes d'ordenacions. Conclouen, com en el cas anterior, que el mètode de Ward és el més efectiu, seguit de *TWINSPAN* i deixant en darrer lloc *UPGMA*.

Finalment, Bruelheide & Chytrý (2000) comparen *TWINSPAN* amb l'algorisme *COCKTAIL*. Concretament estudiaren la capacitat dels dos algorismes de produir classificacions semblants en aplicació a sets de dades diferents però amb una estructura semblant. Bruelheide & Chytrý (2000) conclouen que *TWINSPAN* és massa depenent de les primeres divisions de l'anàlisi de correspondències i, per tant, *COCKTAIL* genera classificacions més comparables.

B. Comparació entre la classificació obtinguda per mètodes numèrics i la classificació tradicional sigmatista.

Els estudis que comparen la sintaxonomia tradicional i la sintaxonomia numèrica tenen un interès especial pel tema que ens ocupa en aquesta memòria. No obstant, presenten el problema de que les dues aproximacions sovint tenen bases epistemològiques diferents (Hakes 1994). Trobem força estudis comparatius d'aquesta mena (Moore *et al.* 1970, Stanek & Orloci 1973, Feoli & Gerdol 1979, Mucina 1982, Wildi 1989 i Hakes 1994).

Stanek & Orloci (1973) compararen la classificació tradicional amb la proporcionada pel mètode de Ward (1963) "tallant" els dendrogrames a 2 nivells de proximitat. Estimaren els resultats d'ambdues aproximacions com a molt similars. Mucina (1982) compara la sintaxonomia tradicional amb els resultats de diversos mètodes jeràrquics aplicats a l'anàlisi de taules sintètiques. Conclou que els mètodes més propers a la sintaxonomia són el de Ward, *complete linkage* i *WPGMA*.

Uns anys més tard que Mucina, Wildi (1989) comparà el seu mètode de reordenació de taules amb el mètode manual emprant exemples de dades conegudes i remarcà la obtenció de resultats similars. Hakes (1994) comparà la classificació tradicional de dades pròpies amb la obtinguda pel mètode de Wildi (1989). Hakes avaluà els resultats dels dos mètodes mitjançant l'anàlisi de concentracions (Feoli & Orloci 1979) i l'anàlisi discriminant amb variables ambientals. Com els autors anteriors, conclogué que els resultats d'ambdues aproximacions són similars, sobretot pel que fa a comunitats de condicions ambientals marginals. No obstant, Hakes (1994) estima que la capacitat de predicció de la classificació numèrica és més alta que la de l'aproximació tradicional.

3.1.6 Estudi comparatiu de diferents models de classificació

3.1.6.1 Objectius i metodologia general

L'objectiu principal d'aquesta secció 3.1.6 és fer paleses les diferències existents entre les aproximacions numèriques a la classificació que hem introduït en els anteriors apartats. Alhora, voldríem discutir alguns dels seus avantatges i inconvenients per a l'anàlisi de la vegetació.

Hem inclòs en la comparació algorismes força diversos, per tal de posar l'èmfasi en la importància del model de classificació i en el concepte de clúster. Al quadre 3.1.1 enumerem els mètodes seleccionats, agrupats pel model de classificació que generen. Cal esmentar que, en el cas de *K-means* hem testat la variant correctora de l'efecte "atractor" (*KM-loo*) proposada a l'apartat 3.1.3.3. Per a l'execució d'aquest i els altres mètodes de *clustering* hem utilitzat el programa *GINKGO* (vegeu capítol 4.2), amb l'excepció de *TWINSPAN*, mètode pel qual hem utilitzat el programa PC-ORD (McCune & Mefford 1999).

Quadre 3.1.1: Mètodes de classificació a comparar i abreviacions utilitzades.

Model	*Mètode de classificació*	*Abreviació*
Jeràrquic	1.Mètodes aglomeratius (*SAHN*) :	
	Single linkage	SL
	Complete linkage	CL
	UPGMA.	UPGMA
	Mètode de *Ward*	Wards
	β-flexible	B-flex
	2.TWINSPAN	TW
Partitiu	3. *K*-means	KM
	4. *Fuzzy C-means*	FCM
Blocs d'especies/inventaris	5. REBLOCK	RBLCK

Tant els mètodes *SAHN* com els mètodes partitius (*KM* i *FCM*) cerquen grups en un espai de proximitats entre inventaris. En executar aquests mètodes d'anàlisi de clústers, hom escull de manera explícita o implícita una mesura de proximitat. En aquest capítol hem escollit utilitzar, com al capítol 2.3, la distància de la corda (Orlóci 1967). Al proper capítol (3.2) estudiarem l'efecte que l'elecció d'altres proximitats produeix en el resultats dels mètodes de classificació.

Per a l'avaluació dels resultats dels mètodes partitius i de *REBLOCK* hem utilitzat criteris interns i externs. També, per al mètode jeràrquic de Ward és possible emprar els mateixos criteris d'avaluació interna perquè la funció que optimitza és semblant a la de *KM*. Malauradament, per a la resta d'algorismes jeràrquics aglomeratius i per a *TWINSPAN* no disposàvem d'un criteri intern d'avaluació dels resultats. Per tant, els resultats d'aquests mètodes han estat avaluats exclusivament per criteris externs.

Com a estadístics d'avaluació interna hem utilitzat aquells que són útils per a determinar quina estructura sorgeix de les dades de manera més "natural". Un dels objectius que ens plantegem és determinar quin d'aquests estadístics presenta una major sensibilitat a l'hora d'assenyalar particions naturals de les dades de vegetació. Concretament, els estadístics escollits per a avaluar les particions de *K-means* i el mètode de Ward han estat: *pseudo-F* (Calinsky-Harabasz 1974), siluetes (*silhouettes*, Rousseeuw 1987). Per a *FCM*, a més dels anteriors, s'han utilitzat dos estadístics d'avaluació de particions difuses: El coeficient de Dunn o coeficient de partició normalitzat (Bezdek 1974, 1981) i la entropia normalitzada o eficiència de la partició (Dunn 1976). Finalment, per tal d'avaluar els resultats de *REBLOCK* s'ha utilitzat la divergència relativa, estadístic proposat a l'apartat 3.1.5.8.

L'avaluació dels resultats amb criteris externs s'ha fet mitjançant la comparació amb la partició proporcionada per la bibliografia o per l'assignació sintaxonòmica dels inventaris. La mesura d'acord entre particions utilitzada ha estat l'índex de Rand (Rand 1971) corregit per l'atzar (Hubert & Arabie 1985). Aquest índex ens ha permès comparar també les diferents aproximacions numèriques entre elles i determinar quines generaven resultats semblants. Addicionalment, hem realitzat comparacions entre clústers individuals mitjançant el coeficient de correlació Φ.

Finalment hem inclòs un darrer criteri d'avaluació exclusivament encarat a la fitosociologia. Concretament hem calculat la mitjana per clúster del nombre de tàxons que presentaven una fidelitat Φ superior a 0.3 (vegeu capítol 2.2). Cal notar que la maximització del nombre de tàxons fidels és un criteri que no utilitzen la majoria dels algorismes que comparem, amb l'excepció de *REBLOCK* i *TWINSPAN*. Per tant, sembla lògic esperar un major nombre de tàxons fidels per als resultats d'aquests dos algorismes. Per aquest mateix motiu, els únics mètodes de *clustering* on tindria sentit emprar l'estadístic com a criteri d'avaluació intern seria en *REBLOCK* i *TWINSPAN*.

3.1.6.2 Anàlisi de clústers amb els inventaris de Bowman & Wilson (1986)

El primer cas d'anàlisi de l'estructura en clústers que estudiarem aplega 41 inventaris i 33 espècies, de la plana al·luvial del riu Adelaide a Austràlia. Es tracta d'un conjunt de dades recol·lectades originalment per Bowman & Wilson (1986). Les localitzacions dels inventaris foren escollides subjectivament, en dos indrets geogràficament disjunts. A la publicació original, Bowman & Wilson realitzaven una anàlisi de correspondències (CA) i una interpretació ambiental del gradient resultant. Aquest conjunt d'inventaris fou també l'exemple triat per Dale (1988a) en el seu estudi sobre aproximacions difuses a la fitosociologia. A les dades de Bowman & Wilson, malgrat hi apareixen divisions que es permeten reconèixer grups, presenten també una clara estructura en gradient que provoca que els grups delimitats no siguin del tot clars. La classificació que feu Dale (1988a), i que nosaltres utilitzarem com a criteri "extern", fou obtinguda

aplicant la mètrica de Canberra (Williams & Lance 1967), i el mètode *SAHN* β-flexible (β = -0.25). Dale escollí un tall a tres grups, però indicà que una partició en quatre grups també era possible. A la taula 3.1.1 presentem les dades de Bowman & Wilson, amb la classificació dels inventaris en tres grups que proposà Dale (1988a).

		2	5	6	7	8	9	11	14	15	17	19	22	25	28	33	1	3	4	10	23	16	18	24	26	27	29	30	31	32	12	13	20	21
Grup A	5	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	1	3	.	.	2	.	4	2	.	1	.	.	.	.	.
	8	.	.	.	.	.	.	.	.	.	.	.	1	.	1	.	.	.	.	1	2	.	.	1	.	2	.	.	1	.	.	.	.	.	.
	13	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	1	1	.	1	1	.	5	.	.	1	.	.	.	.	.	.
	4	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	.	.	.	.	5	.	.	1	.	.	.	.	.	.
	17	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	3	.	.	.	1	3	.	.	1	.	.	.	.	.	.
	3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	2	.	1	4	.	.	.	.	.	.
	9	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	2	.	.	1	6	.	.	.	.	.
	21	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	3	1	.	.	.	.	.	.	.	.
	16	.	.	.	.	.	.	.	.	.	.	.	2	.	.	.	.	.	.	2	1	.	.	.	1	3	3	2	1	.	.	.	.	.	.
	14	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	4	1	3	1	.	.	.	.	.	.
	2	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	5	1	1	1	1	.	.	.	.	.
	15	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	4	.	2	1	3	.	.	.	.	.
	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	5	2	.	1	.	.	.	.	.	.
	7	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	1	.	.	2	2	4	1	.	.	.	.	.	.
	10	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	1	2	5	2	.	.	.	.	.	.
Grup B	40	.	.	.	.	.	1	.	1	1	.	.	.	.	.	.	.	.	.	.	.	5	.	.	.	1	1	.	.	.	.	.	.	.	.
	23	.	.	1	1	.	1	.	1	.	.	.	3	.	.	1	.	.	.	.	.	4	.	2	.	3	.	.	.	.	.	.	.	.	.
	25	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	4	.	1	.	3	1	.	.	.	.	.	.	.	.
	22	.	.	.	1	.	1	.	1	1	.	.	.	.	.	.	.	.	.	.	.	3	.	1	.	4	.	.	.	.	.	.	.	.	.
	20	.	.	2	.	.	.	.	.	1	.	.	.	.	.	2	.	.	.	.	.	4	1	1	.	1	1	.	.	.	1	1	.	.
	6	.	.	.	.	.	1	.	.	.	.	.	.	.	.	1	.	.	.	.	2	4	2	1	.	.	1	.	.	.	.	.	.	.	.
	18	.	.	.	.	.	1	.	.	.	.	.	2	.	.	.	.	.	.	.	.	5	2	1	.	.	1	.	.	.	.	.	.	.	.
	12	.	.	.	.	.	1	.	.	.	.	.	1	.	.	.	.	.	.	.	.	5	1	1	1	.	1	1	.	.	.	.	.	.	.
	39	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1	2	.	1	.	.	1	.	.	.	.	.	.	.	.
	19	.	.	1	1	.	1	.	.	.	.	1	.	.	.	1	.	.	.	.	1	5	.	2	2	2	2	.	.	.	.	.	.	.	.
	11	.	.	.	.	.	1	.	.	1	.	.	.	.	.	1	.	.	.	.	.	3	3	1	1	.	1	1	.	.	.	.	.	.	.
Grup C	30	.	1	.	1	.	1	.	1	.	.	.	.	.	.	.	4	1	.	1	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.
	34	1	.	.	.	.	1	.	1	1	.	.	.	.	.	.	3	.	4	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.
	28	.	.	.	1	.	1	.	1	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
	31	.	.	.	.	.	1	.	.	1	.	.	.	.	.	.	2	1	4	2	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.
	26	.	.	.	.	.	.	.	1	1	.	.	.	.	.	.	4	.	.	1	1	3	.	.	.	.	.	.	.	.	.	.	.	.	.
	29	.	.	.	1	.	1	.	.	.	.	.	.	.	.	.	1	1	1	4	1	3	.	1	.	.	.	.	.	.	.	.	.	.	.
	33	.	.	.	1	1	1	.	.	.	.	.	.	.	.	.	1	.	.	.	1	4	.	.	.	.	.	.	.	.	.	.	.	.	.
	24	.	.	.	1	.	1	.	1	.	1	.	.	.	.	.	.	.	3	.	1	1	.	.	.	.	.	.	.	.	.	.	.	.	.
	36	.	.	.	1	.	1	.	1	.	.	.	.	.	.	1	.	.	4	.	1	2	1	1	1	.	.	.	.	.	.	.	.	.	.
	37	.	1	.	1	.	2	.	1	.	.	.	.	.	.	.	.	.	4	.	2	.	2	.	1	.	.	.	.	.	.	.	.	.	.
	41	.	.	.	3	.	1	.	1	.	.	.	.	.	.	.	.	.	3	.	.	.	1	.	.	2	.	.	.	.	.	.	.	.	.
	27	.	2	.	6	1	2	.	1	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
	32	.	.	.	4	.	2	1	1	.	.	.	.	.	.	.	.	.	.	1	1	.	2	.	.	.	.	.	.	.	.	.	.	.	.
	35	.	.	.	1	.	1	.	1	1	.	.	.	1	.	.	.	.	.	.	.	.	1	2	.	2	.	.	.	.	.	.	.	.	.
	38	.	.	.	1	.	1	.	2	1	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	4	.	.	.	.	.	.	1	1

1.*Cyperus rotundus* 2.*Abelmoschus ficulneus* 3.*Ipomoea coptica* 4.*Cynodon arcuatus* 5.*Merremia hederacea* 6.*Alysicarpus vaginalis* 7.*Panicum cambodiense* 8.*Abelmoschus moschatus* 9.*Melochia corcholifolia* 10.*Waltheria indica* 11.*Ludwigia octovalvis* 12.*Poaceae* sp.1 13.*Heliotropum crispatum* 14.*Euphorbia vachellii* 15.*Echinochloa colona* 16.*Phyla nodiflora* 17.*Paspalum scrobiculatum* 18.*Echinochloa elliptica* 19.*Poaceae* sp.2 20.*Phyllantus* sp. 21.*Goodenia purpurescens* 22.*Cardiospermum halicacabum* 23.*Sesbania* sp. 24.*Heliotropum indicum* 25.*Dentalla dioica* 26.*Ipomoea aquatica* 27.*Oryza* sp. 28.*Cassia obtusifolia* 29.*Eleocharis* sp. 30.*Pseudoraphis spinescens* 31.*Ludwigia adscendens* 32.*Polygonum attenuatum* 33.*Aeschynomene indica*

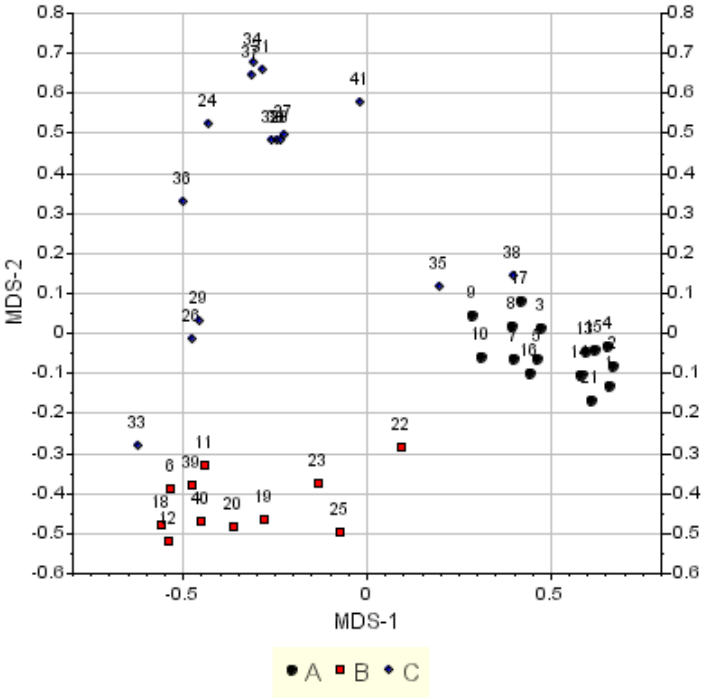
Taula 3.1.1: Dades de Bowman & Wilson (1986). Els inventaris corresponen a les files de la taula i els tàxons a les columnes. S'indiquen els grups determinats per Dale (1988a) a la columna de l'esquerra.

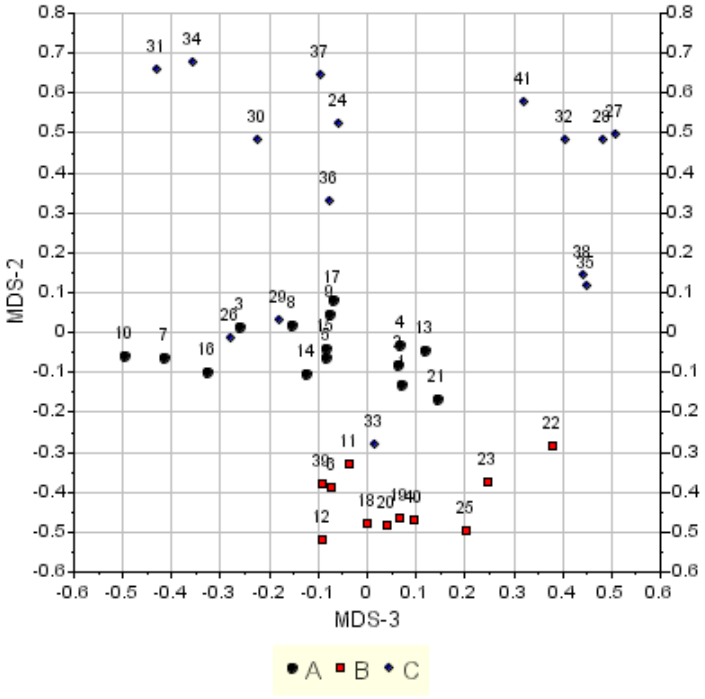
Per tal de conèixer millor els atributs fitosociològics d'aquesta classificació "externa", hem calculat la fidelitat dels tàxons de la taula als tres grups de Dale, mitjançant el coeficient Φ. A la taula inferior (3.1.2) mostrem aquells tàxons amb un valor de Φ superior a 0.4 per a algun dels grups. Segons aquest criteri, el grup A presenta només tres tàxons fidels, però un d'ells, *Ludwigia adscendens* es pot considerar característic. Entre els altres dos, *Oryza* sp és present també en els grups B i C, i *Pseudoraphis spinescens* és força selectiva però relativament poc present al grup A. Els grups B i C presenten més tàxons fidels, però amb valors més baixos. *Phyla nodiflora* sembla diferenciar B i C del grup A, tot i que és més present a B. El cas contrari el trobem amb *Euphorbia vachelii*, que també diferencia B i C del grup d'A, però és més present a C. *Aeschyomene indica* i *Alysicarpus vaginalis* són exclusius del grup B però amb una baixa presència. El mateix succeeix amb *Cynodon arcuatus* i *Cyperus rotundus* per al grup C. En conjunt, hom pot afirmar que, els grups B i C tenen una caracterització florística més feble que el grup A.

Grup A		**Grup B**		**Grup C**	
Espècie	**Φ**	**Espècie**	**Φ**	**Espècie**	**Φ**
31. *Ludwigia adscendens*	0.948	16. *Phyla nodiflora*	0.757	14 . *Euphorbia vachellii*	0.685
27. *Oryza* sp.	0.639	33. *Aeschynomene indica*	0.528	7. *Panicum cambodiencee*	0.638
30. *Pseudoraphis spinescens*	0.454	24. *Heliotropum indicum*	0.510	4. *Cynodon arcuatus*	0.597
		29. *Eleocharis* sp.	0.496	1. *Cyperus rotundus*	0.545
		6. *Alysicarpus vaginalis*	0.464	9. *Melochia corcholifolia*	0.472

Taula 3.1.2: Tàxons fidels als grups de Dale (1988) determinats per nosaltres amb el coeficient de fidelitat Φ.

Per a finalitzar la presentació del conjunt de dades de Bowman & Wilson, mostrem als diagrames de la figura 3.1.6 els tres primers eixos d'una anàlisi de coordenades principals, en que la mesura de proximitat entre inventaris és la distància de la corda (Orlóci 1967). El percentatge de variabilitat mostrat entre els tres eixos és del 57%. Al diagrama de les coordenades principals 1 i 2, que mostra el 47.5% de la variabilitat, els grups de Dale es mostren força aïllats, a excepció d'alguns inventaris del grup C. No obstant, aquest aïllament no es repeteix als altres dos diagrames.





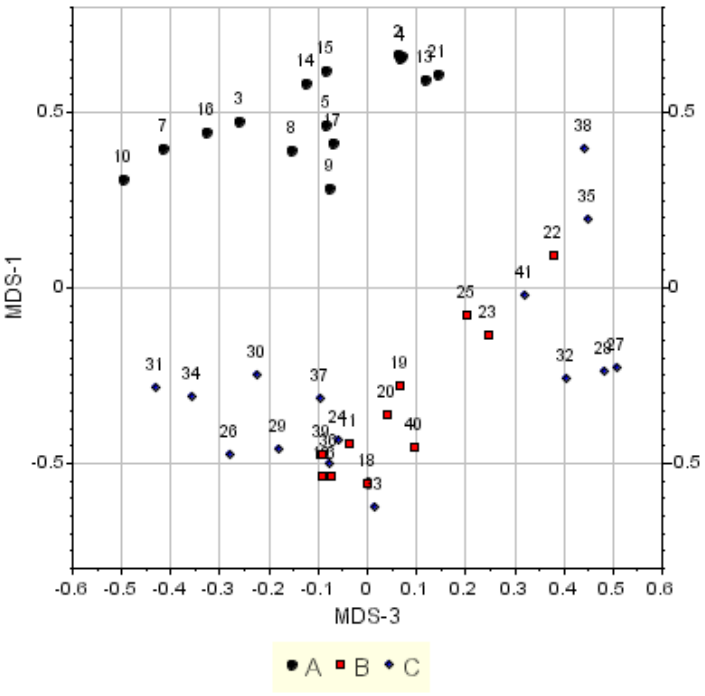


Figura 3.1.6: Anàlisi de coordenades principals (*MDS* mètric) de les dades de Bowman & Wilson (1986) emprant la mesura de dissimilaritat de la corda. Percentatge de variabilitat mostrada: 28.6% + 18.9% + 9.5% = 57%.

Mètodes jeràrquics aglomeratius

Els dendrogrames que produeixen els mètodes *SAHN* poden ser molt diferents segons el mètode utilitzat. Com a un exemple de tres mètodes de sortida dispar, a la figura 3.1.8 mostrem els dendrogrames obtinguts de l'aplicació de s*ingle linkage* (esquerra), *complete linkage* (centre) i el mètode de Ward de mínimització de la suma de quadrats (dreta). El dendrograma de *single linkage* presenta un marcat encadenament, posant de relleu l'estructura en gradient de les dades. Tan sols un subconjunt d'inventaris de C es separa clarament de la resta. *Single linkage* implementa un concepte de clúster proper a un "patró de punts interconnectats" (vegeu apartat 3.1.1.3). Com a mètode d'anàlisi de clústers ens ajuda a fer palesa la continuïtat en les dades però no permet establir divisions fàcilment. En segon lloc, el mètode del veí més llunyà o *complete linkage*, no presenta encadenament i si talléssim el dendrograma en tres grups la partició resultant seria propera a la classificació de Dale. No obstant, la longitud de les branques del dendrograma no motiva una partició en tres grups. Finalment, l'algorisme de minimització de la suma de quadrats o mètode de Ward, proporciona un dendrograma amb una clara estructura de 3 grups que es manté en un rang de dissimilaritats ampli i que coincideix, a grans trets, amb la classificació de Dale.

A banda de l'apreciació visual de la longitud de les branques, és difícil proporcionar criteris interns de validació dels resultats dels mètodes jeràrquics aglomeratius. Per a donar una idea de l'eficiència relativa entre els diferents algorismes, a la figura 3.1.7 mostrem l'ajust amb l'índex de Rand corregit, entre els talls dels dendrogrames a diferent nombre de grups i el criteri "extern" de la classificació de Dale (1988) de tres grups. El mètode que, clarament, produeix classificacions més allunyades del criteri extern és *single linkage*. D'entre els restants, el mètode de Ward sembla destacar en la comparació al nivell de $K = 3$.

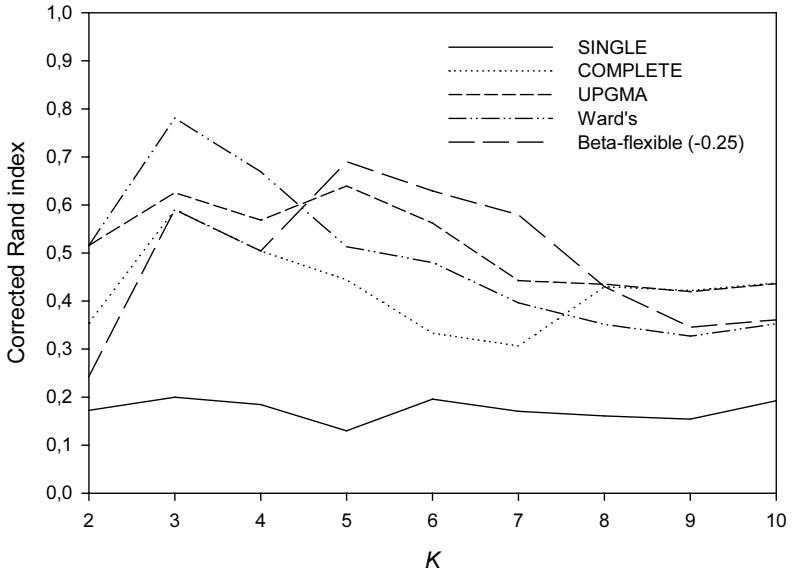


Figura 3.1.7: Comparació, mitjançant l'índex de Rand corregit per l'atzar, entre la classificació de Dale (1988a) de tres grups i els talls, desde K=2 a K=10, dels dendrogrames trobats pels mètodes jeràrquics.

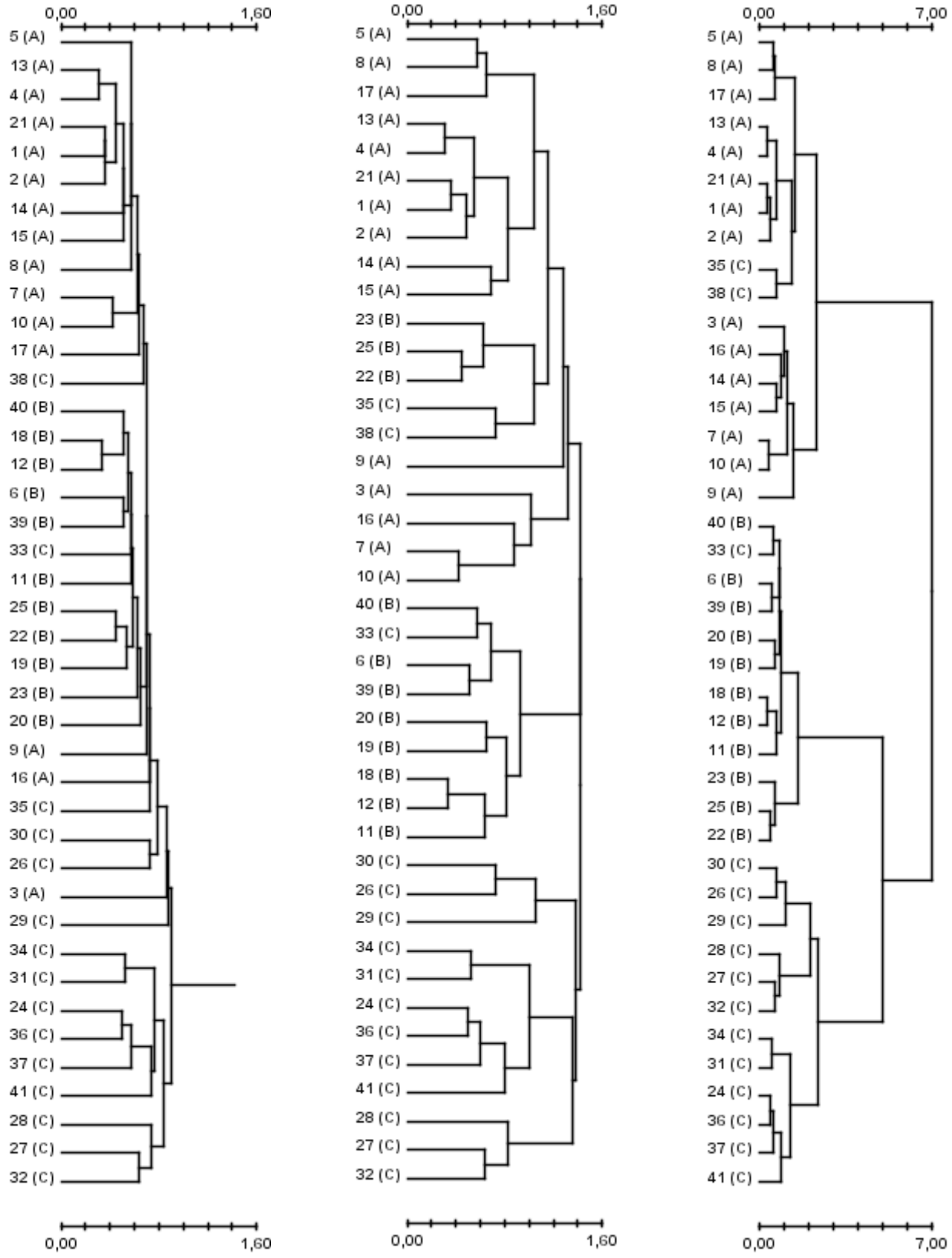


Figura 3.1.8: Dendrogrames resultants de la classificació jeràrquica aglomerativa de *single linkage* (esquerra), *complete linkage* (centre) i *minimum sum of squares method* (*Ward's*, dreta). Entre parèntesi es mostren els grups de la classificació de Dale (1988a).

TWINSPAN

La implementació de *TWINSPAN* del programa *PC-ORD* (McCune & Mefford 1999) ha estat executada de dues maneres: (a) sense considerar l'abundància dels tàxons (P/A) i (b) incloent l'abundància en forma de pseudo-espècies, un concepte propi d'aquest algorisme. En ambdós casos, però, l'algorisme ha produït classificacions semblants. *TWINSPAN* han realitzat dues divisions de les dades. A la primera, *TWINSPAN*, identifica 31.*Ludwigia adscendens* i 27.*Oryza* sp. com a tàxons indicadors positius; mentre que 9.*Melochia corcholifolia* i 7.*Panicum cambodiencee* són considerats indicadors negatius. Com a resultat de la posterior ponderació i divisió dels inventaris, l'algorisme separa el grup A (grup '1') de la resta (grup '0'). A continuació, la segona divisió separa raonablement bé B i C (grups '01' i '00', respectivament) alhora que divideix A en dos nous grups, '10' i '11', dels quals en l'anàlisi amb pseudo-espècies el grup '10' conserva la major part d'inventaris d'A. Entre els indicadors positius per a '01', *TWINSPAN* identifica els tàxons 29.*Eleocharis* sp., 16.*Phyla nodiflora*, 24.*Heliotropum indicum*, i altre cop 27.*Oryza sp.*; mentre que estableix com a sol indicador negatiu (identificador de C), el tàxon 4.*Cynodon arcuatus*. És en aquest darrer grup C, en el que *TWINSPAN* reconeix menys tàxons dels considerats com a fidels *a priori*. Observeu com TWINSPAN utilitza el tàxon *Oryza sp.* com a indicadors per a una divisió i el torna a utilitzar en següents subdivisions. Per tant, la noció de fidelitat que utilitza aquest algorisme és fonamentalment de tipus diferencial.

La taula 3.1.3 conté les correlacions entre els grups trobats per *TWINSPAN* en dues divisions de les dades i els tres grups de Dale (1988a). Hom observa que, per aquestes dades, l'ús de pseudo-espècies té poca transcendència en els resultats de la classificació. Tan sols difereixen en la divisió del grup '1' (A). Exceptuant la tendència de *TWINSPAN* a produir un nombre de clústers potencia de dos, hom pot concloure que els resultats d'aquest mètode són força satisfactoris per a aquest conjunt de dades.

P/A		**1ª divisió**		**2ª divisió**			
		0	**1**	**00**	**01**	**10**	**11**
	N_k	26	25	12	14	6	9
A	15	-1.000	1.000	-0.489	-0.547	0.545	0.698
B	11	0.460	-0.460	-0.390	0.841	-0.251	-0.321
C	15	0.577	-0.577	0.847	-0.227	-0.314	-0.403
Pseudo-species		**1ª divisió**		**2ª divisió**			
		0	**1**	**00**	**01**	**10**	**11**
	N_k	26	25	12	14	12	3
A	15	-1.000	1.000	-0.489	-0.547	0.847	0.370
B	11	0.460	-0.460	-0.390	0.841	-0.390	-0.170
C	15	0.577	-0.577	0.847	-0.227	-0.489	-0.213

Taula 3.1.3: Resultats de la classificació *TWINSPAN* de les dades de Bowman & Wilson (1986) emprant només presències i absències (P/A, a dalt), o 6 nivells de pseudo-espècies. Es mostren els valors de l'índex de Rand corregit entre les particions corresponents als diferents nivells de divisió de *TWINSPAN* i la classificació de Dale. N_k: nombre d'inventaris de cada grup.

Mètodes partitius: K-means i Fuzzy C-means

Hem executat els algorismes *KM*, amb la correcció per *leave one out*, i *FCM*, amb cinc nivells de *fuzziness* (m=1.1, m=1.2, m=1.3, m=1.5 i m=2.0), sobre la matriu de distàncies de la corda entre inventaris. Per tal d'evitar escollir *a priori* el nombre de grups a cercar, hem executat els algorismes des de K=2 fins a K=10. L'estratègia d'inicialització dels algorismes ha estat triar inventaris a l'atzar per utilitzar com a centroides inicials. S'han aturat la execució dels algorismes quan el millor funcional trobat es repetia 20 vegades. Per tant, el nombre de execucions aleatòries ha estat diferent en cada cas, depenent de la complexitat de la situació (determinada pel nombre de clústers a cercar i exponent de *fuzziness*).

A continuació, hem avaluat la 'bondat' de les particions resultants mitjançant quatre estadístics per a la determinació del nombre de grups: *pseudo-F*, la silueta mitjana i el coeficient de partició de Dunn i l'entropia normalitzada. La figura 3.1.9 mostra els valors obtinguts pels estadístics en les diferents situacions. A les gràfiques de *pseudo-F* i la silueta mitjana hi hem inclòs els valors resultants de realitzar talls al dendrograma de l'algorisme jeràrquic de Ward. Hem aplicat aquests criteris d'avaluació de l'aïllament amb l'algorisme de Ward i no amb els altres algorismes jeràrquics perquè aquest és el que més s'adiu, per construcció, a generar clústers aïllats (vegeu pàg. 104).

L'estadístic *pseudo-F* presenta el seu valor màxim a la partició en K=3 grups en totes les variants algorísmiques a excepció de *FCM* m=1.5, on la partició recomanada és K=2. El criteri de la silueta mitjana assenyala que la millor opció per a particions *KM* seria K=4. Per al tall del dendrograma de Ward indica igualment K=4, o K=6. En el cas de *FCM*, el criteri de la silueta proposa K=3 o K=4 i K=7 o K=8 depenent del nivell de *fuzziness*. Els perfils del coeficient de partició de Dunn i l'entropia normalitzada varien, obviament, segons el nivell de *fuzziness* emprat a *FCM*. Per a exponents molt baixos, m=1.1 o m=1.2, el coeficient de partició és molt proper a 1.0 i la entropia propera a zero. Aquest fet emmascara la possibilitat de decidir el nombre de grups idoni. Per a exponents m= 1.3 i m =1.5 presente valors intermitjos i una major resolució respecte a la comparació entre nombre de grups diferents. Assenyalen com a particions més adequades les particions difuses de K=3 i K=7. Per a valors alts de l'exponent - m=2.0 - les particions són progressivament més i més difuses i la distinció entre el nombre de grups es torna a perdre.

En la comparació entre els estadístics de determinació del nombre de grups idoni, creiem que l'estadístic que presenta una sensibilitat més alta de detecció d'estructures és la silueta mitjana. El nombre de grups més plausibles, suportats per força estadístics, serien K=3 o K=4. D'altra banda observem que els dos estadístics d'avaluació de particions borroses tendeixen a donar respostes molt semblants com a criteris interns. Per tant, considerem que emprar qualsevol dels dos criteris és indiferent.

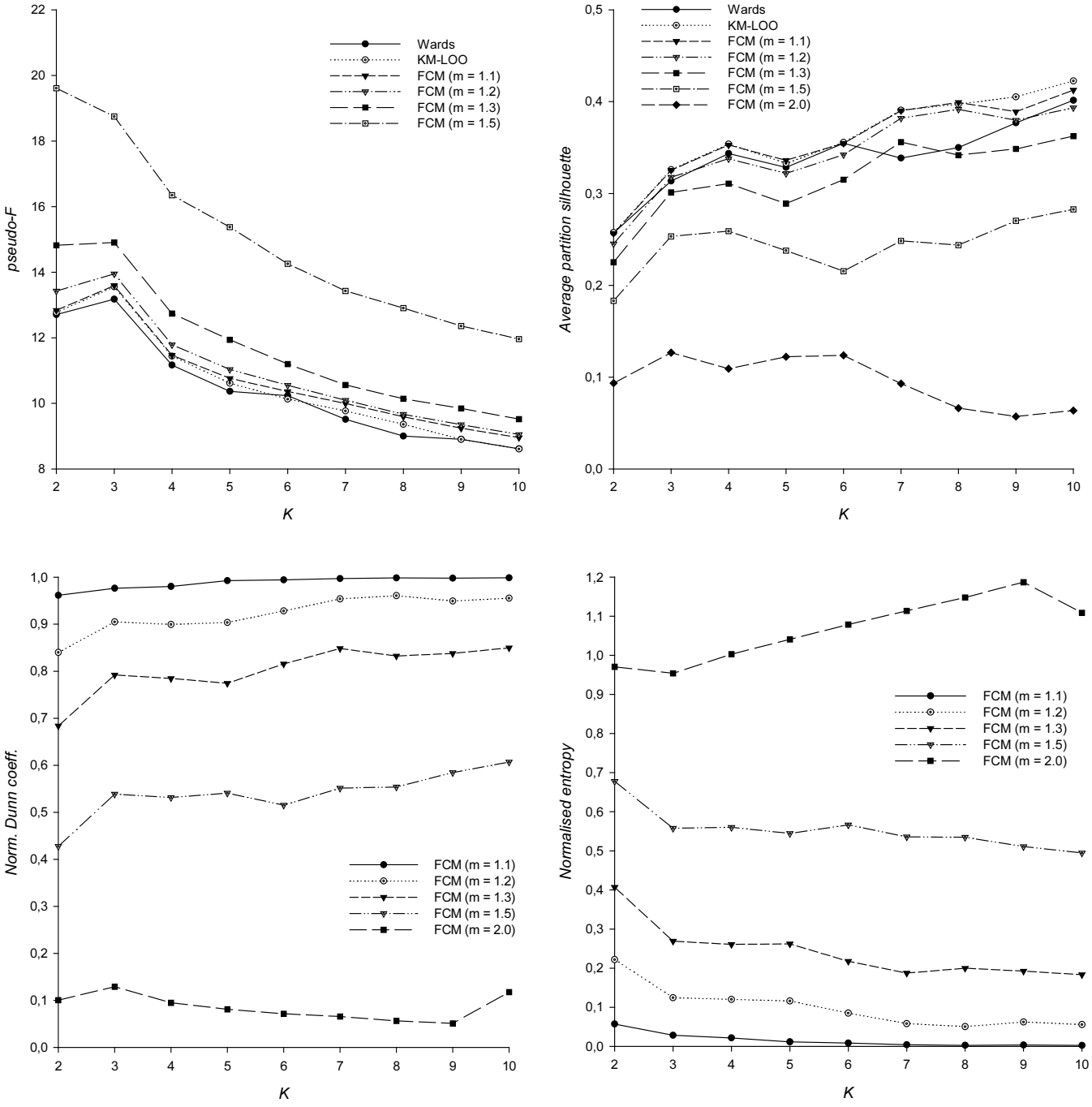


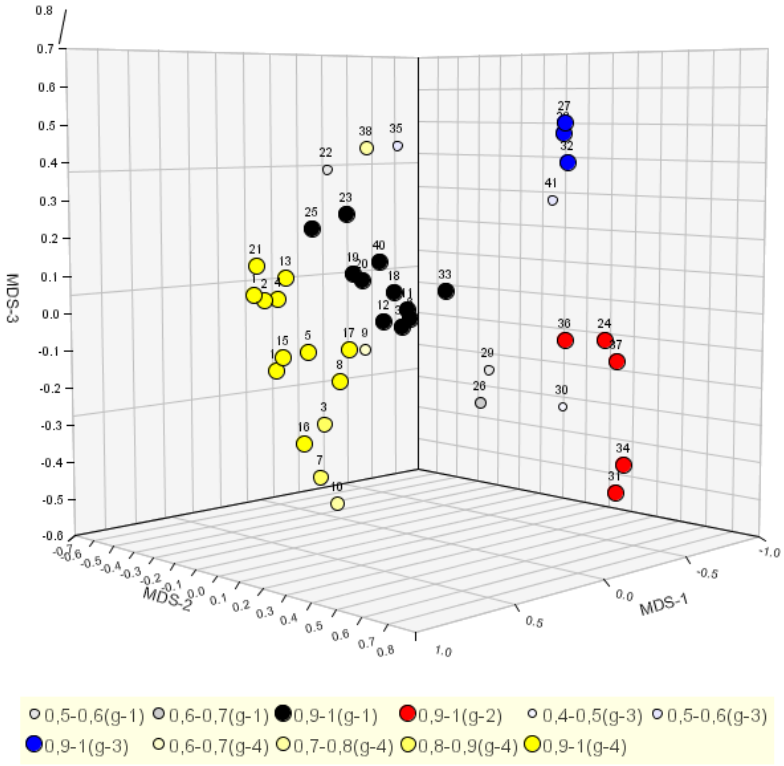
Figura 3.1.9: Estadístics de determinació del nombre de grups idoni: pseudo-F (a dalt, a l'esquerra), siluetes (a dalt, a la dreta), coeficient de Dunn normalitzat (a baix, a l'esquerra) i entropia normalitzada (a baix, a la dreta). Valors entre $K = 2$ i $K = 10$.

A la taula 3.1.4 mostrem les particions difuses obtingudes amb *FCM* K=3 i *FCM* K=4, emprant en ambdós casos m=1.3. Les particions *crisp* de *K-means* són equivalents a les mostrades. A la partició K=3, els inventaris dels grups originals A i B es presenten agrupats en els respectius grups f-1 i f-3. El grup original que ha resultat ser més problemàtic sembla ser el C, del que *FCM* en classifica dos inventaris amb f-1 (A), tres amb f-3 (B) i els restants amb f-2.

A la partició *FCM* K=4, es subdivideix el grup f-2 en 2 nous subgrups (g-2 i g-3), mentre que els inventaris C que es classificaven a f-1 i f-3 ara es classifiquen en els corresponents (g-1 i g-4). A la figura 3.1.10 mostrem la partició en K=4 grups sobre una anàlisi de coordenades principals, la qual sembla donar-hi suport.

	f-1	f-2	f-3	g-1	g-2	g-3	g-4
5	0.98	0.01	0.01	0.01	0.01	0.01	0.98
8	0.93	0.04	0.04	0.03	0.03	0.03	0.91
13	1.00	0.00	0.00	0.00	0.00	0.00	0.99
4	1.00	0.00	0.00	0.00	0.00	0.00	1.00
17	0.94	0.03	0.02	0.02	0.02	0.03	0.93
3	0.91	0.05	0.04	0.04	0.03	0.05	0.88
9	0.68	0.17	0.14	0.13	0.11	0.16	0.60
21	0.99	0.00	0.00	0.00	0.00	0.00	0.99
16	0.95	0.02	0.03	0.02	0.01	0.02	0.95
14	0.99	0.00	0.00	0.00	0.00	0.00	0.99
2	1.00	0.00	0.00	0.00	0.00	0.00	1.00
15	0.98	0.01	0.01	0.01	0.01	0.01	0.98
1	1.00	0.00	0.00	0.00	0.00	0.00	1.00
7	0.90	0.05	0.05	0.04	0.03	0.04	0.89
10	0.78	0.11	0.12	0.11	0.07	0.10	0.73
40	0.00	0.00	1.00	1.00	0.00	0.00	0.00
23	0.06	0.03	0.92	0.90	0.01	0.03	0.05
25	0.04	0.01	0.96	0.95	0.00	0.01	0.03
22	0.42	0.05	0.53	0.52	0.02	0.08	0.38
20	0.01	0.01	0.98	0.98	0.01	0.01	0.01
6	0.00	0.00	0.99	0.99	0.00	0.00	0.00
18	0.00	0.00	1.00	1.00	0.00	0.00	0.00
12	0.00	0.00	1.00	1.00	0.00	0.00	0.00
39	0.00	0.01	0.99	0.98	0.01	0.01	0.00
19	0.00	0.00	0.99	0.99	0.00	0.00	0.00
11	0.02	0.03	0.95	0.94	0.02	0.02	0.01
30	0.09	0.81	0.10	0.09	0.39	0.44	0.08
34	0.02	0.97	0.02	0.00	0.99	0.01	0.00
28	0.05	0.88	0.07	0.01	0.01	0.98	0.00
31	0.02	0.96	0.02	0.00	0.99	0.01	0.00
26	0.06	0.25	0.70	0.64	0.20	0.10	0.05
29	0.05	0.34	0.61	0.56	0.26	0.13	0.05
33	0.00	0.01	0.99	0.98	0.01	0.01	0.00
24	0.01	0.98	0.01	0.00	0.99	0.01	0.00
36	0.01	0.93	0.06	0.01	0.98	0.01	0.00
37	0.01	0.99	0.01	0.00	0.98	0.01	0.00
41	0.02	0.97	0.01	0.01	0.37	0.58	0.03
27	0.05	0.88	0.06	0.00	0.01	0.99	0.00
32	0.04	0.91	0.05	0.00	0.01	0.99	0.00
35	0.63	0.25	0.12	0.07	0.05	0.58	0.30
38	0.89	0.08	0.03	0.03	0.03	0.22	0.72

Figura 3.1.10: Diagrama de dispersió dels 3 primers eixos de l'anàlisi de coordenades principals (Var. = 56.5%) amb la partició difusa de *FCM* (m = 1.3, K = 4).



Taula 3.1.4: Particions de *FCM* (m=1.3) K=3 i K=4. S'han ressaltat aquelles pertinences majors de 0.5.

Si tornem a mirar la taula original de dades, es fàcil adonar-se que la divisió de C en dos grups es possible, car aquest conjunt d'inventaris és el que visualment presenta més variabilitat dels tres. D'altra banda, els tres inventaris del grup C que *KM* i *FCM* classifiquen al grup B, 26,

29 i 33, presenten tots abundàncies relativament importants de l'espècie *Phyla nodiflora* (16), mentre que hi són pràcticament absents els tàxons *Euphorbia vachelii* (14) i *Cynodon arcuatus* (4), que són dels que més caracteritzen el grup C. Finalment, als inventaris 35 i 38, que *KM* i *FCM* classifiquen al grup A, hi manquen les espècies 4 i 16, i en canvi presenten els tàxons 14, fidel al grup C, i *Oryza sp*. (27) que els acosta al grup A. Per tant, la seva classificació intermèdia a la taula 3.1.4 sembla prou justificada.

REBLOCK

Hem executat l'algorisme *REBLOCK* (Podani & Feoli 1991) amb diferents combinacions del nombre de grups d'objectes (K_1) i nombre de grups de variables (K_2). Per tal d'avaluar les diferents classificacions obtingudes i escollir-ne la més adequada hem emprat el criteri de la maximització de la divergència relativa. Les combinacions que s'haurien de veure afavorides segons la equació de la divergència relativa, són aquelles en que $K_1 \approx K_2$. A la pràctica, veiem a la figura 3.1.11 com la decisió d'escollir K_1 i K_2 es pot reduir a algunes combinacions: Per a un nombre de grups de variables baix (K_2=2 o K_2=3), la millors opció de K_1 és K_1=2. Per a un nombre de grups de variables més gran (K_2=4, K_2=5, o K_2=6) el millor nombre de grups d'objectes és K_1=3. Per a valors alts del nombre de grups de variables l'estadístic sembla afavorir un nombre de grups major: [K_1=6, K_2=7]. Finalment, hem escollit la doble partició [K_1=3, K_2=6] que correspon a un dels màxims esmentats.

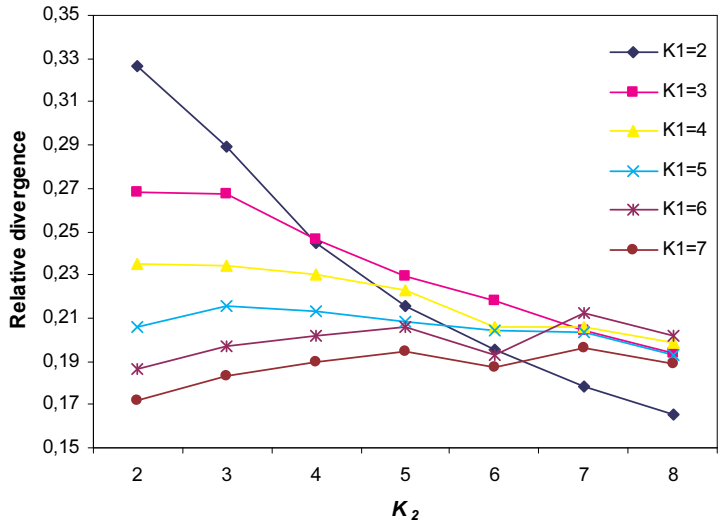


Figura 3.1.11: Estadístic de divergència relativa avaluada amb les dobles particions obtingudes amb REBLOCK emprant diferents combinacions de K_1 i K_2.

A la taula 3.1.6 apareix de nou la matriu de dades de Bowman & Wilson, aquest cop reordenades segons la doble partició obtinguda amb *REBLOCK* [K_1 = 3, K_2 = 6]. Per altra banda,

mostrem a la taula 3.1.5 les contribucions relatives a l'estadístic $\chi^2_{(K_1,K_2)}$ dels diferents blocs, i les contribucions marginals de cadascun dels grups d'inventaris i grups de tàxons. En ambdues taules hem ressaltat aquells blocs que més contribueixen a l'estadístic de divergència. A les classificacions de *REBLOCK*, una contribució alta a la divergència per part d'un bloc pot venir tant per una fidelitat positiva com una fidelitat negativa del grup de tàxons que hi intervé. Observeu com el bloc OG-3/DG-4 es composa de tàxons absents. No obstant, hom pot esperar que aquest últim cas sigui menys freqüent en dades amb molts inventaris i molts tàxons ja que els tàxons constants són molt infreqüents.

Els grups d'inventaris de la doble partició [$K_1 = 3$, $K_2 = 6$] són gairebé idèntics a la classificació de Dale (l'índex de Rand corregit és 0.807). La diferència rau en tres inventaris del grup C (35, 36 i 38) que a *REBLOCK* apareixen juntament amb els del grup B per formar el grup OG-1. Curiosament, el grup d'inventaris amb menys contribució a la divergència, i per tant, amb menys entitat és aquest.

A la taula 3.1.6 hem assenyalat aquells tàxons considerats fidels a la classificació de Dale segons el coeficient Φ (vegeu grups A, B i C entre parèntesi). Observeu que en la majoria de grups de tàxons de *REBLOCK* els tàxons fidels originals que hi apareixen són fidels al mateix grup. L'excepció la conformen els tàxons del grup DG-4, en que hi apareixen els tàxons que diferencien el grup C dels grups A i B. Per tant, la tendència de *REBLOCK* és a classificar grups d'espècies relacionades entre elles, pel que hom pot esperar que, de retruc, mostri una tendència a agrupar espècies fidels comunes d'un mateix grups d'inventaris.

		DG-1	**DG-2**	**DG-3**	**DG-4**	**DG-5**	**DG-6**	
	n	10	2	8	5	4	4	
OG-1	14	6.2%	2.0%	2.9%	0.4%	2.7%	0.1%	**14.2%**
OG-2	15	3.5%	0.4%	2.6%	3.3%	13.1%	4.5%	**27.4%**
OG-3	12	1.1%	1.0%	12.9%	6.3%	3.0%	3.0%	**27.3%**
		10.8%	**3.4%**	**18.4%**	**10.0%**	**18.8%**	**7.6%**	**100%**

Taula 3.1.5: Contribucions relatives dels diferents blocs (cel·les) i grups (sumes marginals) de la doble partició [K_1=3, K_2=6] a l'estadístic de divergència $\chi^2_{(K_1,K_2)}$.

		DG-1											DG-2		DG-3									DG-4					DG-5				DG-6			
		(B) 6	19	25	(B) 33	(B) 16	18	12	13	20	21		10	23	2	5	8	11	17	(C) 1	3	(C) 4		22	(B) 24	26	(A) 27	(B) 29		28	(A) 30	(A) 31	32	(C) 7	(C) 9	(C) 14	15
OG-1	40.B					5																				1	1							1	1	1
	23.B	1			1	4																		3	2		3						1	1	1	
	25.B					4																			1		3	1						1			
	22.B					3																			1		4							1	1	1	1
	20.B	2			2	4	1	1	1																1		1	1								1
	6.B				1	4	2							2											1			1							1		
	18.B					5	2																	2	1			1							1		
	12.B					5	1																	1	1	1		1			1				1		
	39.B					2								1											1			1							1		
	19.B	1	1		1	5								1											2	2	2	2						1	1		
	11.B				1	3	3																		1	1		1			1				1		1
	36.C				1	2	1							1								4			1	1								1	1	1	
	35.C			1			1																		2		2							1	1	1	1
	38.C						1			1	1																4							1	1	2	1
OG-2	5.A												1	3											2		4	2				1				1
	8.A												1	2										1	1		2			1		1				
	13.A						1						1	1											1		5					1				1
	4.A												1	1													5					1				
	17.A													3												1	3					1		1		
	3.A																									1	2			1	4					
	9.A																								1	1	2					1	6			
	21.A																								1		3	1								
	16.A												2	1										2		1	3	3			2	1				
	14.A																								1	1	4	1			3	1			1	
	2.A																										5	1			1	1	1		1	
	15.A																										4				2	1	3			
	1.A																										5	2				1				
	7.A						1							1													2	2			4	1				
	10.A																								1		1	2			5	2				
OG-3	30.C												1	1		1				4	1													1	1	1	
	34.C													1	1					3		4													1	1	1
	28.C																																	1	1	1	1
	31.C												2	1						2	1	4														1		1
	26.C					3							1	1						4																1	1
	29.C					3							4	1						1	1	1			1									1	1		
	33.C					4								1			1			1														1	1		
	24.C					1								1					1			3												1	1	1	
	37.C						2							2		1						4				1								1	2	1	
	41.C						1															3					2							3	1	1	
	27.C															2	1																	6	2	1	1
	32.C						2						1	1				1																4	2	1	

Taula 3.1.6: Dades de Bowman & Wilson (1986) reordenades segons la doble partició de *REBLOCK* [K_1=3, K_2=6]. Els inventaris corresponen a les files de la taula i els tàxons a les columnes. Els grups d'inventaris són: OG-1, OG-2 i OG-3 i els grups de tàxons van des de DG-1 a DG-6. S'han marcat aquells blocs d'inventaris i espècies que més contribueixen a l'estadístic de divergència. La numeració dels inventaris inclou la classificació original de Dale (1988a). Addicionalment, s'assenyala aquells tàxons considerats fidels als grups originals amb el grup entre parèntesi.

Comparació entre mètodes

És important saber quins mètodes de classificació són més susceptibles de produir resultats semblants i en quins és probable que produeixin resultats més dispars. Com a resum de l'estudi de les dades de Bowman & Wilson, hem calculat l'índex de Rand corregit entre parelles de particions derivades dels diferents mètodes de classificació. Per als dendrogrames i els algorismes partitius s'ha escollit la partició $K = 3$, per permetre una millor comparació amb la classificació de Dale (1988a). En el cas de *TWINSPAN* s'han comparat dues particions, la generada per la segona divisió i aquella sorgida de subdividir només el grup '0' i no el grup '1'. Finalment, per a *REBLOCK* s'ha utilitzat la doble partició [$K_1 = 3$, $K_2 = 6$].

La figura 3.1.12 dibuixa els dos primers eixos d'una anàlisi de coordenades principals realitzada a partir de la dissimilaritat que s'obté com a complement de l'índex de Rand corregit. El diagrama de dispersió resultant mostra clarament que el mètode més dispar d'entre els comparats és *single linkage*. Els mètodes jeràrquics restants (*complete linkage*, el mètode de Ward, *UPGMA* i *β-flexible*) implementen un concepte de clúster que aproximadament és pot assimilar al dels mètodes partitius basats en la minimització de la dispersió. D'altra banda, *TWINSPAN* i *REBLOCK* provenen d'un concepte de clúster força diferent, en el que s'espera una dependència entre variables i objectes. És lògic, per tant, que aquests darrers proporcionin solucions separades de l'anterior grup de mètodes.

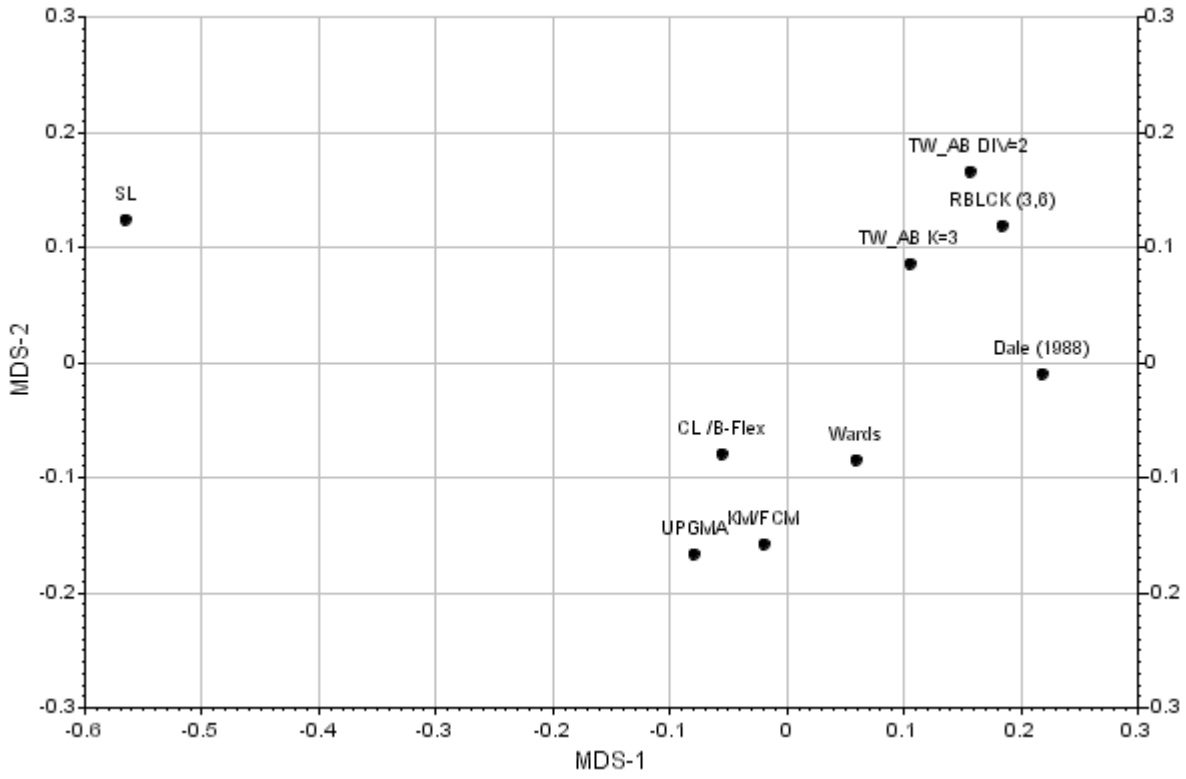


Figura 3.1.12: Anàlisi de coordenades principals de la matriu de distàncies complement de l'índex de Rand corregit, calculat entre particions derivades dels diferents mètodes de classificació La variabilitat representada al diagrama és del 83.3%.

A la taula 3.1.7 es mostra l'índex de Rand corregit entre els diferents mètodes i la classificació de Dale (1988a). Noteu que els mètode que generen classificacions més allunyades de la de Dale són *single linkage* i, després, *complete linkage* i *β-flexible*. Com que Dale classificà les dades de Bowman & Wilson amb l'algorisme *β-flexible* i a partir de l'espai generat per la mètrica de Canberra, veiem ara que la diferència que provoca utilitzar una mètrica diferent. Aquesta pot ser la raó per la qual els algorismes geomètrics han produït classificacions força diferents a la de Dale, i en canvi els mètodes no geomètrics han produït classificacions més properes. Per tant, cal interpretar amb molta cautela la comparació amb aquest criteri "extern". També mostrem a la figura 3.1.7 la mitjana entre clústers del nombre de tàxons fidels. Aquesta darrera columna fa palès que els mètodes de classificació que més tendeixen a produir grups d'inventaris suportats per tàxons fidels són aquells que incorporen el concepte de fidelitat al seu concepte de clúster: *REBLOCK* i *TWINSPAN*.

Mètode i paràmetres	Rand *vs.* Dale	Mitjana tàxons Φ>0.3
Dale (1988) - Canberra metric i algorisme *Beta-flexible*	-	5.33
Single linkage *K*=3 (SL)	0.199	4.33
Complete linkage / Beta-flexible *K*=3 (CL/ B-Flex)	0.590	5.33
UPGMA K=3	0.625	4.33
Ward's *K*=3	0.780	5.66
TWINSPAN abundàncies, 2 divisions (TW_AB DIV=2)	0.696	5.25
TWINSPAN abundàncies, *K*=3 (TW_AB K=3)	0.807	6.00
REBLOCK K_1=3, K_2=6 (RBLCK)	0.807	6.67
K-means/FCM *K*=3 (KM/FCM)	0.671	4.66

Taula 3.1.7: Avaluació dels resultats del diferents mètodes d'anàlisi de clúster en les dades de Bowman & Wilson. Valor de l'índex de Rand en comparació amb la classificació de Dale (1988a). Valor mitjà del nombre de tàxons amb fidelitat Φ més gran que 0.3.

3.1.6.3 Anàlisi de clústers dels sintàxons de *Brometalia* i *Quercetea*

A l'hora d'escollir els inventaris a analitzar en aquest apartat hem de tenir en compte que en capítols anteriors vam comprovar que no tots els sintàxons de base eren igualment vàlids. Alguns presentaven un bon grup de tàxons fidels i la seva disposició en l'espai multivariant és compacte i aïllada. En d'altres casos, per contra, la caracterització del sintàxon de base es basa en unes poques espècies diferencials difícils de suportar numèricament i la representació en l'espai multivariant del sintàxon es solapa àmpliament amb la d'altres grups. A banda de la possibilitat construir un espai de dades més adequat per a la descripció de la vegetació és palès que hi ha sintàxons de base amb problemes de discriminabilitat estadística en l'espai de la corda.

Per aquestes raons, hem escollit analitzar aquí un subconjunt dels inventaris. Concretarem hem escollit els inventaris dels sintàxons per als quals considerem la classificació sigmatista com a prou vàlida per ésser emprada com a patró extern de comparació. En el cas de les dades de *Brometalia erecti*, hem seleccionat l'aliança *Xerobromion erecti*. En aquesta aliança els problemes de discriminabilitat són molt menors que a *Mesobromion*, implicant tan sols el nivell de subassociacions (d'*Irido-Brometum* i *Teucrio-Brometum*, principalment). Per a les dades de *Quercetea ilicis* hem exclòs la subaliança *Quercenion ilicis* (alzinars litorals i muntanyencs, principalment), que és un grup de dades que, al nostre parer, presenta molts trànsits i una baixa discriminabilitat dels seus sintàxons. Pel que fa a la resta d'inventaris de la classe, els sintàxons de base es presenten força aïllats, amb l'excepció de les diferents varietats de garrigars, els quals presenten una baixa discriminabilitat, juntament amb l'associació *Rhamno-Quercetum*. Les matrius de dades a analitzar han estat, finalment:

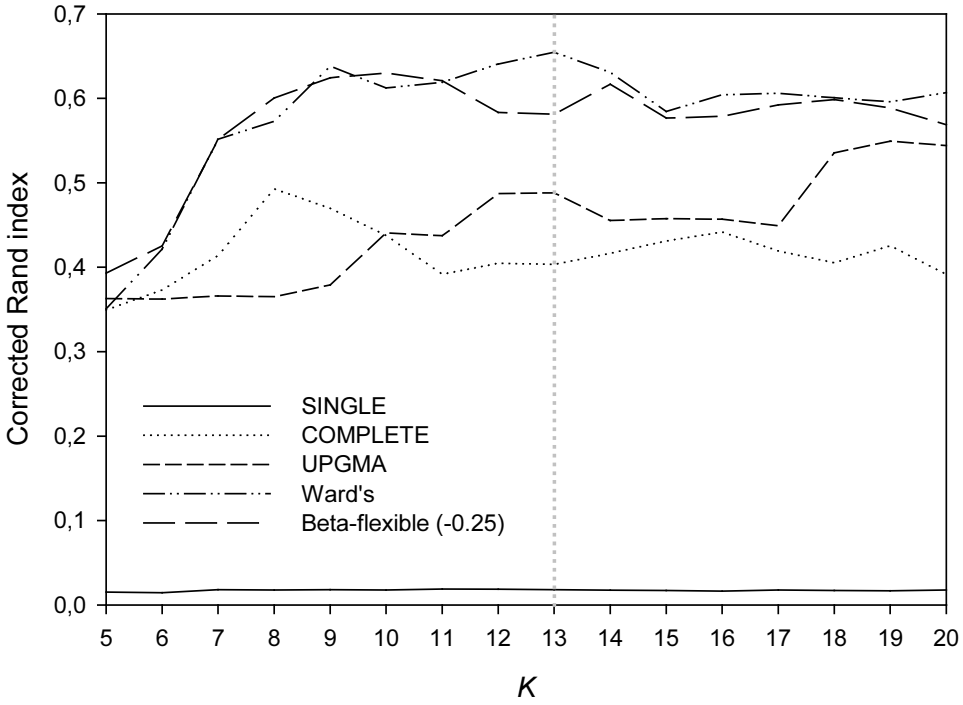
A) *Xerobromion erecti*: 248 inventaris, 548 tàxons. Classificació tradicional en 13 sintàxons de base (5 associacions i 8 subassociacions).
B) *Quercetea ilicis* sense *Quercenion:* 376 inventaris, 493 tàxons. Classificació tradicional en 16 sintàxons de base (8 associacions i 8 subassociacions).

Com en el capítol 2.3, la transformació de l'escala ordinal de Braun-Blanquet emprada ha estat la de van der Maarel (1979). La resta d'opcions d'anàlisi (la distància de la corda, els mètodes de classificació a comparar, i els mètodes d'avaluació dels resultats) han estat similars als descrits per estudi de les dades de Bowman & Wilson.

Mètodes jeràrquics aglomeratius

Ja hem esmentat a l'apartat 3.1.2.2 que els algorismes jeràrquics aglomeratius són inadequats per a volums de dades grans (Hill *et al.* 1974). Tanmateix, però, hem executat aquests algorismes de *clustering* a partir de la matriu de distàncies de la corda entre els inventaris dels dos conjunts de dades. No presentem, per resultar massa grans, els dendrogrames que en resulten. A les figures 3.1.13.A i 3.1.13.B mostrem l'ajust, mesurat mitjançant l'índex de Rand corregit, entre els talls realitzats sobre els dendrogrames i la classificació tradicional sintaxonòmica presa com a criteri "extern". Els mètodes que millor reprodueixen els resultats de l'escola tradicional són el de Ward i el *β-flexible*. En una posició intermèdia quedarien *complete linkage* i UPGMA, i en darrer lloc roman, el mètode del veí més proper o *single linkage*. En general, es repeteix el mateix ordre que en el cas d'estudi de l'anterior apartat. Noteu, però, que per a les dades de *Quercetea ilicis* el màxim d'acord entre el mètode de Ward i l'escola tradicional no es produeix al nombre de grups que aquesta escola determina (K=16) sinó en un nombre inferior (K=11) cosa que indica que potser la classificació fitosociològica ha afinat més alguns sintàxons de base que altres.

A. *Xerobromion erecti*



B. *Quercetea ilicis* sense *Quercenion*

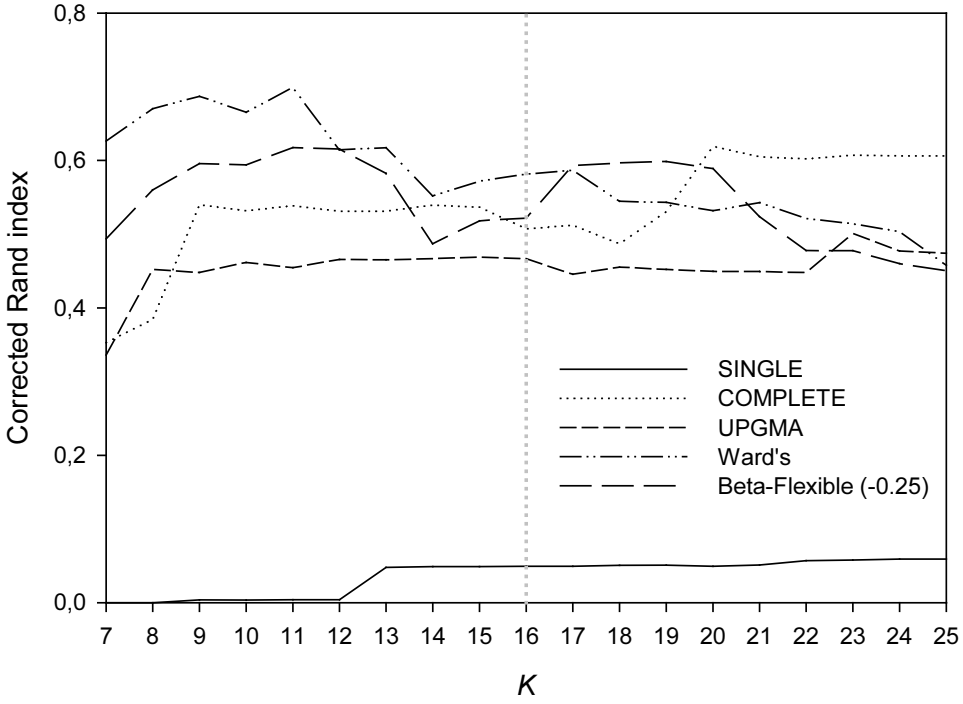


Figura 3.1.13: Comparació, amb l'índex de Rand corregit per l'atzar, entre la classificació sintaxonòmica i les línies de tall dels dendrogrames generats pels algorismes jeràrquics aglomeratius. Les línies puntejades verticals indiquen el nombre de grups presents a la classificació tradicional.

TWINSPAN

En executar el mètode *TWINSPAN* sobre les matrius A i B hem emprat diferents tractaments. En primer lloc, hem volgut testar l'efecte de treure les espècies més rares de (aquelles amb menys de tres aparicions a la taula). En segon lloc, hem considerat dos nivells pel que fa al tractament de les abundàncies, com en conjunt de dades precedent: P/A i pseudoespècies (nivells 1,2,3...9). Per tant, hem analitzat un total de quatre situacions "experimentals" de classificació. El mínim nombre d'inventaris que havia de tenir un grup per a que *TWINSPAN* el pogués subdividir és de 5.

L'avaluació de les jerarquies resultants de l'anàlisi de cada tractament s'ha fet exclusivament amb el criteri extern tradicional. Concretament, hem comparat la partició tradicional en sintàxons de base amb cada una de les particions corresponents a les quatre darreres divisions de *TWINSPAN* -2ª, 3ª, 4ª, i 5ª -. A continuació hem escollit el tractament que ha generat una partició amb un ajust més elevat. La jerarquia completa resultant de l'anàlisi d'aquest tractament ha estat emprada per a analitzar les correlacions per grups i poder interpretar els clústers de *TWINSPAN*.

La taula 3.1.8 mostra els valors de l'índex de Rand corregit per a totes les comparacions esmentades. Dins de cada divisió, les dades de *Xerobromion erecti* presenten un ajust força semblant independentment dels quatre tractaments d'anàlisi. La partició amb un ajust més elevat és la sorgida de la quarta divisió en el tractament d'emprar només presències i absències i incloure els tàxons rars (en negreta). Pel que fa a *Quercetea ilicis* sense *Quercenion*, l'ajust dels darrers és clarament millor en utilitzar les abundàncies. Això probablement denota que en la obtenció de la classificació tradicional les abundàncies dels tàxons han tingut un paper important. La partició amb un índex més alt correspon a la quarta divisió del tractament amb pseudo-espècies i sense tàxons rars.

A. Xerobromion erecti

	PA		Ab (9 pseudosp.)	
	Rare	**No Rare**	**Rare**	**No Rare**
2nd div.	0.184	0.200	0.170	0.158
3rd div.	0.360	0.352	0.299	0.291
4th div.	**0.489**	0.473	0.420	0.431
5th div.	0.460	0.460	0.457	0.467

B. Quercetea ilicis* sense *Quercenion

	PA		Ab (9 pseudosp.)	
	Rare	**No Rare**	**Rare**	**No Rare**
2nd div.	0.230	0.225	0.303	0.280
3rd div.	0.279	0.279	0.464	0.437
4th div.	0.279	0.294	0.433	**0.608**
5th div.	0.401	0.329	0.490	0.557

Taula 3.1.8: Resultats de la classificació *TWINSPAN* de les dades de *Xerobromion erecti* (A) i *Quercetea ilicis* sense *Quercenion* (B) amb la combinació d'emprar pseudospècies o P/A i incloure o no les espècies de baixa presència. Es mostren els valors de l'índex de Rand corregit entre les particions corresponents als diferents nivells de divisió de *TWINSPAN* i la partició tradicional en sintàxons de base. Hem ressaltat la casella amb el valor més alt de cada fila. En negreta es marca la partició de *TWINSPAN* amb un ajust més alt de cada taula.

A les taules 3.1.9.A i 3.1.9.B mostrem la correlació, mesurada amb el coeficient Φ, entre els diferents nivells de divisions de *TWINSPAN* que contenien la partició amb un ajust més elevat i la classificació tradicional. Per facilitar la lectura de la taula, mostrem només aquelles correlacions positives, entre les quals hem ressaltat aquelles $\Phi > 0.5$. Els diferents nodes de la jerarquia de grups de *TWINSPAN* es corresponen força bé amb els sintàxons establerts per l'escola de classificació tradicional, però aquesta correspondència es troba a diferents nivells de la jerarquia de divisions. Per exemple, a les dades de *Xerobromion*, *TWINSPAN* identifica un grup a la segona divisió, que anomena '00', el qual es correspon totalment a *Teucrio-Brometum* subass. *festucetosum fallacis* (TBF). A la següent divisió, el grup '00' és dividit en dos, pel que la correspondència amb TBF es perd. El mateix succeix amb *Teucrio-Festucetum* (TF) i el grup '11'. D'altra banda *Lino-Brometum* (LB) queda ben força ben representat a la 3ª divisió amb '101', i *Achilleo-Dichantietum* (AD) a la 4ª amb '0100'. Finalment, hi ha sintàxons que encara necessitarien una 5ª divisió (no mostrada) de les dades per a ésser identificats per *TWINSPAN*. És el cas del grup '0110' de TWINSPAN, que engloba tres sintàxons de base: la associació *Teucrio-Avenuletum* (TAV) i les dues subassociacions d' *Irido-Brometum* (IBT i IBL).

			LB	AB	KG	TBT	TBF	TBH	AD	KAT	KAA	TAV	IBT	IBL	TF	TB sl.	Ka sl.	IB sl.
		N	32	21	19	29	13	9	23	31	15	13	10	6	27	51	46	16
2nd div.	00	13					**1.000**									0.462		
	01	133			0.207				0.269	0.351	0.236	0.219	0.150	0.146			0.444	0.211
	10	78	0.465	0.106		0.429		0.240								0.364		
	11	24													**0.936**			
3rd div.	000	10					**0.872**									0.403		
	001	3					0.470									0.217		
	010	72							0.469	**0.591**	0.322		0.004				**0.700**	
	011	61		0.129	0.399							0.412	0.169	0.276				0.307
	100	50		0.208		**0.536**		0.332								**0.515**		
	101	28	**0.813**															
	110	22													**0.893**			
	111	2													0.258			
4th div.	0000	4					**0.544**									0.252		
	0001	6					**0.669**									0.309		
	0100	26							**0.843**				0.131					0.071
	0101	46								**0.792**	0.445						**0.947**	
	0110	44		0.200								**0.506**	0.227	0.339				0.394
	0111	17			**0.882**													
	1000	12		**0.606**	0.006			0.057										
	1001	38				**0.646**		0.336								**0.614**		
	1010	22	**0.726**															
	1011	6	0.331										0.101					0.065
	1100	15													**0.726**			
	1101	7													0.488			

Taula 3.1.9.A: Resultats de la classificació *TWINSPAN* dels inventaris de *Xerobromion*. Es mostren els valors de correlació Φ entre els grups obtinguts per TWINSPAN en les divisions 2ª, 3ª i 4ª amb la sintaxonomia tradicional al nivell d'associació i subassociació. Hem exclòs de la taula aquelles correlacions negatives per tal de facilitar la lectura de la mateixa. Aquelles correlacions $\Phi > 0.500$ es troben ressaltades.

A la classificació que *TWINSPAN* realitza de les dades de *Quercetea* es repeteix la mateixa situació d'identificació de sintàxons a diferents nivells de la jerarquia de divisions. Les associacions *Buxo-Juniperetum* (BJ) i *Clematido-Osyrietum* (CO) són representades pels grups '00' i '11' a la 2ª divisió, però posteriorment es perden. Pel que fa a *Quercetum rotundifoliae*, la subassociació *buxetosum* (QRB) es correspon al grup '010' del 3er nivell, i la subassociació *ulicetosum* (QRU) al 4rt, parcialment confosa amb la subassociació *rhamnetosum* (QRR). D'altra banda, les diferents subassociacions de *Quercetum cocciferae* encara al 4rt nivell romanen barrejades, juntament amb l'associació *Rhamno-Quercetum* (RQ). El mateix succeeix amb les associacions *Calicotomo-Myrtetum* (CM) i *Myrto-Juniperetum* (MJ). Per tant, a diferència del cas anterior hi ha sintàxons que fins i tot a la darrera divisió romanen junts. Aquest fa palès de nou la manca de tàxons fidels (diferencials o característics) d'alguns sintàxons de base de *Quercetea*.

			OL	QL	CM	MJ	RJ	RQ	QCR	QCBR	QCC	QCBT	QCT	BJ	CO	QRB	QRR	QRU	QC (s.l.)	QR (s.l.)
		N	16	55	11	24	10	12	23	5	6	26	8	41	38	70	9	22	68	101
2nd div.	00	34												**0.901**						
	01	124										0.121	0.171			**0.682**	0.223	0.307		**0.838**
	10	182	0.218	0.427	0.116	0.270	0.171	0.157	0.241	0.120	0.131								0.181	
	11	36													**0.910**					
3rd div.	000	14												**0.562**						
	001	20												**0.678**						
	010	78										0.016				**0.918**				**0.726**
	011	46										0.154	0.339				0.366	**0.599**	0.162	0.305
	100	128		**0.529**			0.230	0.221	0.332	0.162	0.177	0.070							0.333	
	101	54	**0.515**		0.334	**0.638**														
	110	20													**0.707**					
	111	16			0.120										**0.541**					
4th div.	0000	9												0.448						
	0001	5												0.332						
	0010	6												0.364						
	0011	14												**0.562**						
	0100	6										0.467							0.271	
	0101	72														**0.965**				**0.773**
	0110	20						0.024				0.309	**0.540**				0.118		0.381	
	0111	26															0.369	**0.825**		0.450
	1000	71		**0.800**			0.343	0.144												
	1001	57						0.134	**0.542**	0.275	0.301	0.236							**0.668**	
	1010	48	0.314		0.359	**0.683**														
	1011	6	**0.604**																	
	1100	15													**0.608**					
	1101	5													0.346					
	1110	14													**0.587**					
	1111	2			0.421															

Taula 3.1.9.B: Resultats de la classificació TWINSPAN dels inventaris de *Quercetea ilicis* sense *Quercenion*. Es mostren els valors de correlació del coeficient Φ entre els grups obtinguts per TWINSPAN en les divisions 2ª, 3ª i 4ª amb la sintaxonomia tradicional al nivell d'associació i subassociació. Hem exclòs de la taula aquelles correlacions negatives per tal de facilitar la lectura de la mateixa. Aquelles correlacions Φ > 0.500 es troben ressaltades.

Mètodes partitius: K-means i *Fuzzy C-means*

Estratègies d'inicialització

A l'hora d'utilitzar mètodes amb optimització local, és important proporcionar una bona estratègia d'inicialització, ja que d'això en depenen els resultats. El perill d'arribar a mínims locals del funcional és més acusat quan el conjunt de dades a analitzar és gran i/o complex. És per aquest motiu que l'anàlisi *K-means* i *FCM* sobre les matrius de dades de *Brometalia* i *Quercetea* necessita un estudi previ sobre estratègies d'inicialització en aquests algorismes.

Amb l'objectiu de trobar l'estratègia més eficient d'anàlisi d'aquest tipus de dades, hem comparat la inicialització de *K-means* a partir del tall d'un arbre ultramètric de Ward amb la inicialització a partir d'inventaris escollits a l'atzar com a centroides inicials. Hem combinat les estratègies d'inicialització amb l'ús o no de la correcció *leave one out* (*loo*) degut a que l'ús aquesta correcció pot influir en la capacitat d'arribar a mínims globals del mètode.

Per a presentar els resultats d'aquest estudi, hem pres la inicialització amb el tall del dendrograma de Ward sense emprar la correcció *loo* com a estratègia de referència. Així, hem calculat la diferència entre el valor final de $TESS$ obtingut per *K-means* amb les altres opcions d'anàlisi i l'obtingut amb aquesta estratègia.

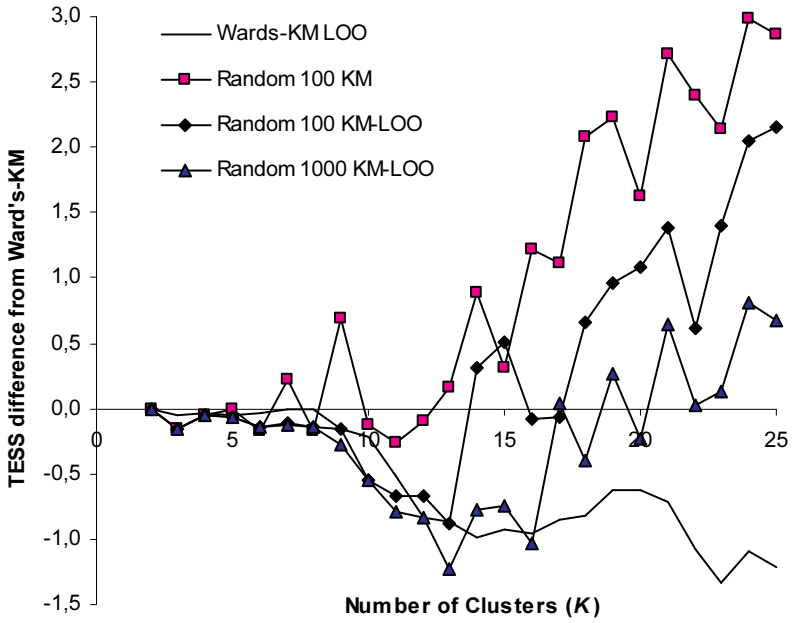
A les figures 3.1.14.A i 3.1.14.B A es mostren les diferències de TESS que s'obtenen en augmentar el nombre de grups (K). Per a un nombre de grups petit totes les estratègies arriben al mínim global del funcional. Com que, en aquesta escala d'anàlisi, els grups tenen molts inventaris, això els confereix més mobilitat per a afegir o treure elements. En conseqüència, és relativament senzill trobar el mínim absolut de funcional. Tot i això, són més recomanables les d'inicialització a l'atzar, que en aquest estadi de complexitat són relativament poc costoses. En incrementar el nombre de grups les estratègies d'inicialització a l'atzar es tornen cada cop menys eficients, requerint un major nombre d'execucions de l'algorisme per a trobar millors solucions. En aquesta situació la inicialització a partir del tall d'un dendrograma pot resultar molt més eficient, raó per la qual les diferències amb la inicialització a l'atzar augmenten a mida que el nombre de grups augmenta.

Pel que fa a la correcció *leave one out*, veiem que aquesta augmenta notablement la flexibilitat de *K-means*, independentment de quina sigui l'estratègia d'inicialització. Fins i tot pot permetre que grups amb pocs objectes puguin perdre'n un si pertoca. Globalment, La correcció *loo* proporciona un benefici en eficiència important a *K-means*, perquè augmenta la capacitat de l'algorisme a l'hora d'escapar de mínims locals.

Entre les estratègies estudiades, considerem que la més eficient en conjunt és la d'emprar una partició generada tallant l'arbre ultramètric de Ward i utilitzar la correcció *loo*. Tanmateix,

hom no pot assegurar mai haver trobat la millor solució amb una sola execució de *K-means*. Les inicialitzacions a l'atzar sempre tenen la possibilitat de donar una solució millor, tot i que aquesta pot requerir un nombre d'intents realment gran. Les mateixes conclusions s'aplicarien a *FCM*, exceptuant que la correcció *loo* és, en aquest cas, innecessària donat el caràcter més flexible del propi algorisme de *clustering*.

A. *Xerobromion*



B. *Quercetea ilicis* sense *Quercenion*

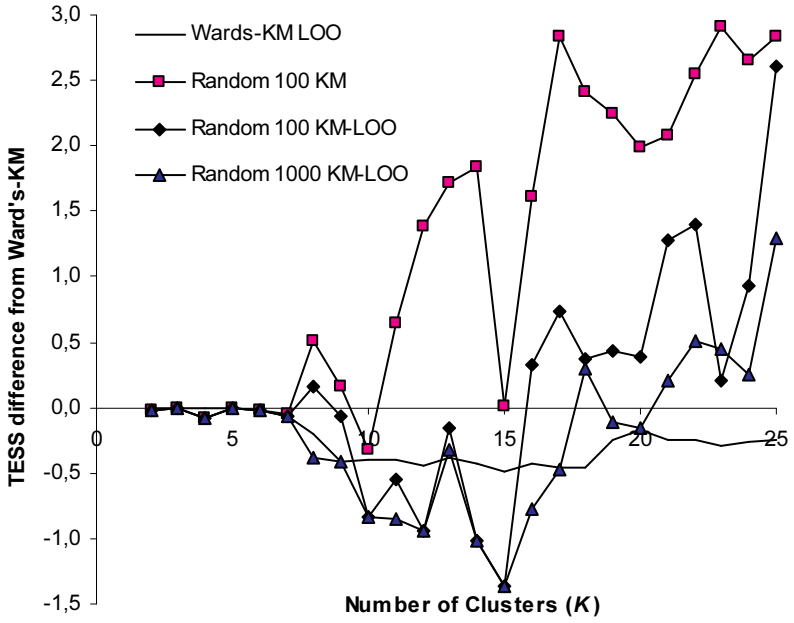


Figura 3.1.14: Diferència entre els valors finals del funcional final de *K-means* amb la inicialització amb el mètode de Ward i els mateixos valors amb la inicialització escollint llavors a l'atzar (en 100 o 1000 execucions), emprant o no la correcció *leave one out* (vegeu text). Evolució d'aquestes diferències en augmentar el nombre de clústers a cercar.

Avaluació per criteris interns

Un cop escollida una estratègia d'inicialització per als algorismes partitius, cal saber quin nombre de grups és el que les dades presenten de manera més "natural" en sentit estadístic. En primer lloc hem executat els algorismes *KM* i *FCM* incrementant el nombre de grups (K) des de K=2 fins a K=20. Per a *FCM* hem assajat alhora diferents valors del paràmetre de *fuzziness* (m=1.1, m=1.15 i m=1.2). A continuació, i com en el cas de les dades de Bowman & Wilson, hem aplicat criteris interns per a decidir el nombre de grups idoni. Concretament, hem calculat els estadístics *pseudo-F* i la silueta mitjana. A més per a les particions de *FCM* hem calculat també el coeficient de Dunn normalitzat. Hem exclòs el càlcul de l'entropia normalitzada per les raons exposades al l'apartat anterior.

A la figura 3.1.15 mostrem els perfils dels tres estadístics escollits, calculats per a les particions *K-means* obtingudes de les dades de *Xerobromion erecti* (A, esquerra) i *Quercetea ilicis* sense *Quercenion* (B, dreta). Per a les particions derivades de *K-means*, l'estadístic *pseudo-F* es mostra com el més ineficient alhora d'assenyalar una estructura concreta. Només en cas de (B) sembla indicar feblement una possible estructura a $K = 3$ o $K = 5$. La silueta mitjana, sembla indicar per a (A) les particions $K = 5$ i $K = 9$, tot i que presenta una tendència a l'aguments. Per a (B) assenyala clarament $K = 5$, mentre que una segona possibilitat de partició podria ser $K = 17$.

Pel que fa al coeficient de Dunn normalitzat, aquest estadístic mostra que les particions difuses de *FCM* de les dades (A) presenten un grau de "borrositat" més gran que les particions difuses de (B). Això fa que les execucions de *FCM* sobre el conjunt de dades (A) tinguin més mobilitat. Per tant, l'estructura de *FCM* que els estadístics silueta i Dunn s'assenyalen com a idònia és diferent segons l'exponent de *fuzziness* utilitzat. Concretament, amb m=1.1 la partició difusa que recomanen la silueta mitjana i el coeficient de Dunn normalitzat és semblant a la que s'assenyala per a *K-means*: K=5 i K=9; però per a m=1.15 el nombre de grups recomanat és ara K=5 o K=10; i per a m=1.2 la recomanació torna a canviar: K=7, K=12 o K=15. En el cas de les dades de *Quercetea ilicis* (B), l'augment de l'exponent de *fuzziness* també provoca un augment de la "borrositat" de les partició difuses però no canvia sensiblement l'estructura trobada. La recomanació que fan els estadístics de selecció del nombre de grups és similar a la que fan per a *K-means*. No obstant, així com la silueta mitjana continua assenyalant K=5, i K=17, aquest segon pic no és detectat pel coeficient de Dunn normalitzat.

Com a conclusió d'aquest estudi del nombre de grups naturals de les dades, hem decidit finalment estudiar amb més detall les següents particions de *K-means*:

- K=5 i K=9 per a *Xerobromion erecti*.
- K=5 i K=17 per a *Quercetea ilicis* sense *Quercenion*.

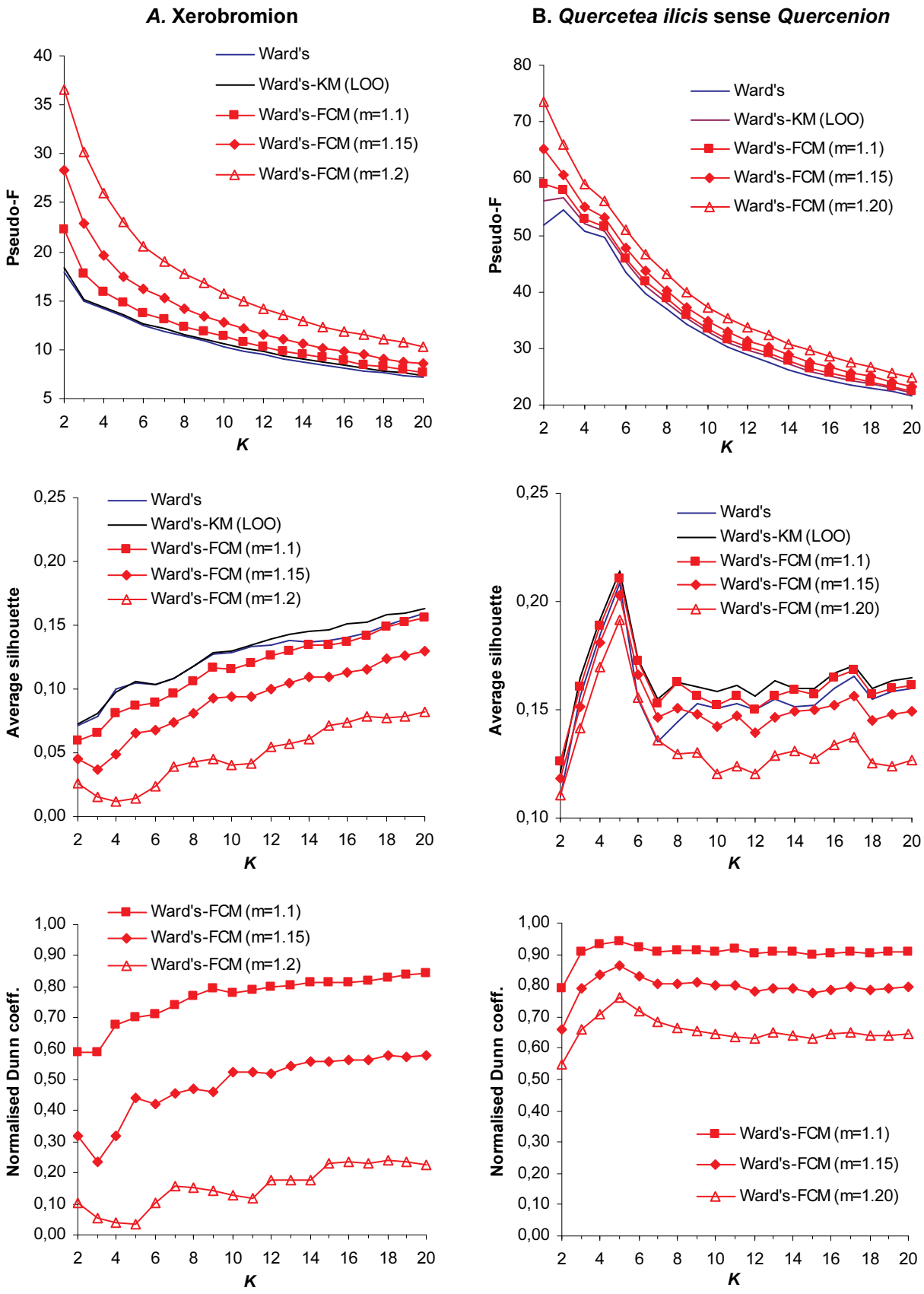


Figura 3.1.15: Estadístics de determinació del nombre de grups idoni per *K-means* i *FCM*: *pseudo-F* (a dalt), *silhouettes* (al mitg) i coeficient de Dunn normalitzat (a baix). Valors obtinguts entre $K = 2$ i $K = 20$.

Avaluació amb el criteri extern de la fitosociologia tradicional

En darrer lloc hem comparat cada un dels grups de les particions *K-means* escollides amb els grups de la sintaxonomia tradicional a nivells d'associació i subassociació. Les comparacions han estat realitzades mitjançant, com abans, el coeficient de correlació Φ.

A les taules 3.1.10.A i 3.1.10.B mostrem els valors de correlació entre els grups *K-means* i els sintàxons determinats pel mètode tradicional. A continuació en comentem els resultats que considerem més importants, desglossats en conjunts de dades i particions:

A.*Xerobromion erecti*

K = 5: Els dos primers grups, G1 i G2, engloben diversos sintàxons de base, i G3 barreja KG i TAV. En canvi, els dos darrers grups identifiquen, ja a aquest nivell, els sintàxons de base TBF i TF.

K = 9: Alguns grups encara podrien ésser subdividits: G6 inclou KAA i les dues subassociacions d'*Irido-Brometum* (IBT i IBL), mentre que G4 barreja LB amb dues subassociacions de *Teucrio-Brometum* (TBT iTBL). La resta dels 7 sintàxons de base de *Xerobromion* estan ben representats per cada un dels restants grups de la partició.

		LB	AB	KG	TBT	TBF	TBH	AD	KAT	KAA	TAV	IBT	IBL	TF	TB (s.l.)	KA (s.l.)	IB (s.l.)
	N	32	21	19	29	13	9	23	31	15	13	10	6	27	51	46	16
G1	87							0.377	**0.514**	0.345		0.236	0.159			**0.649**	0.289
G2	90	0.485	0.343		0.430		0.212								0.342		
G3	33			**0.691**							**0.600**						
G4	26													**0.979**			
G5	12					**0.959**									0.443		
G1	20			0.973													
G2	13										**1.000**						
G3	26													**0.979**			
G4	65	**0.618**			0.411		0.277								0.377		
G5	25		**0.860**		0.087												
G6	33									0.399		**0.523**	0.325			0.210	**0.622**
G7	19							**0.901**									
G8	11					**0.916**									0.423		
G9	36					0.006			**0.813**	0.136						**0.775**	

Taula 3.1.10.A: Comparació amb el coeficient de correlació Φ entre els grups obtinguts per K-means (particions *K*=5 i *K*=9) i els sintàxons de la classificació tradicional de *Xerobromion erecti*. S'han exclòs de la taula aquelles correlacions negatives per tal de facilitar la lectura de la mateixa. Aquelles correlacions Φ > 0.500 es troben ressaltades.

B. *Quercetea ilicis* sense *Quercenion*:

K = 5: Els grups G1 i G2 engloben comunitats de garrigues (subassociacions de *Quercetum cocciferae* i l'associació *Rhamno-Quercetum*) i matollars (OL, CM i MJ) respectivament. G3 identifica l'associació *Clematido-Osyrietum* (CO) i G4 els diferents carrascars (subassociacions de *Quercetum rotundifoliae*). Finalment, G5 agrupa les associacions *Rhamno-Juniperetum* i *Buxo-Juniperetum*.

K = 17: En primer lloc, hi ha grups de la partició amb una clara correspondència als sintàxons de base: G1-CO, G4-MJ, G8-CM, G15-RJ. En segon lloc els carrascars (QR) queden dividits en 4 grups (G3, G7, G10 i G11) que no es corresponen massa bé amb les tres subassociacions. En canvi, la majoria de garrigues queden encara barrejades en el grup G13, excepte QCR i QCBT que s'assemblen als grups G6 i G12, respectivament. Finalment hi ha 3 associacions que queden dividides en dos grups cada una: *Querco-Lentiscetum* (QL) entre G2 i G17; *Oleo-Lentiscetum* (OL) entre G9 i G16 i *Buxo-Juniperetum* (BJ) entre G5 i G14.

		OL	QL	CM	MJ	RJ	RQ	QCR	QCBR	QCC	QCBT	QCT	BJ	CO	QRB	QRR	QRU	QC (s.l.)	QR (s.l.)
	N	16	55	11	24	10	12	23	5	6	26	8	41	38	70	9	22	68	101
G1	119		0.317				0.234	0.375	0.171	0.187	0.401	0.217						**0.691**	
G2	71	0.437	0.146	0.360	0.541														
G3	37													**0.985**					
G4	100														**0.795**	0.181	0.389		**0.966**
G5	49					0.427							**0.828**						
G1	37													**0.985**					
G2	42		**0.737**				0.176												
G3	27															0.361	**0.852**		0.459
G4	22				**0.908**														
G5	23					0.027							**0.694**						
G6	19						0.234	**0.600**				0.050						0.333	
G7	19														0.389	0.123			0.381
G8	11			**1.000**															
G9	6	**0.604**																	
G10	37														**0.691**				**0.545**
G11	16														0.441				0.348
G12	19										**0.799**							0.459	
G13	38						0.040	0.283	0.346	0.309	0.187	0.379						**0.645**	
G14	18												**0.601**						
G15	9					**0.947**													
G16	9	**0.570**														0.089			
G17	24	0.107	0.385		0.065		0.015			0.054									

Taula 3.1.10.B: Comparació amb el coeficient de correlació Φ entre els grups obtinguts per K-means (particions *K*=5 i *K*=9) i els sintàxons de la classificació tradicional de *Quercetea ilicis* sense *Quercenion*. S'han exclòs de la taula aquelles correlacions negatives per tal de facilitar la lectura de la mateixa. Aquelles correlacions Φ > 0.500 es troben ressaltades.

REBLOCK

Per als conjunts de dades de *Xerobromion* i *Quercetea ilicis* sense *Quercenion*, l'execució de REBLOCK es fa extremadament lenta. El motiu és l'elevat nombre de recol·locacions de files i columnes que s'han d'assajar a cada iteració. Tot i que considerem d'interès el funcional de REBLOCK, la implementació algorísmica que proposen Podani & Feoli (1991) es torna impracticable per a volums de dades moderadament grans, fins i tot tenint en compte l'elevada potència de càlcul dels ordinadors actuals. És per aquest motiu que, malauradament no presentem els resultats de l'aplicació d'aquest mètode en l'anàlisi de clústers d'aquests conjunts de dades.

Comparació entre mètodes

Per tal de donar una visió global de les semblances i diferències entre els algorismes assajats. Hem mesurat l'acord entre particions derivades de les diferents classificacions obtingudes. Per a permetre la comparació amb el criteri extern tradicional, hem escollit comparar les classificacions obtingudes al nombre de grups determinat pels sintàxons de base de la sintaxonomia tradicional, és a dir, K=13 grups per a *Xerobromion* i K=16 grups per a *Quercetea ilicis*. Per als resultats de l'algorisme *TWINSPAN* hem utilitzat les particions amb un valor més alt d'ajust a la sintaxonomia, malgrat que això potser introdueixi un element d'afavoriment de *TWINSPAN* respecte la comparació amb la partició extern.

La classificació que ha proporcionat un ajust més baix ha estat, altra vegada, la produïda per l'algorisme *single linkage*. Amb l'objectiu d'incrementar la proporció de variabilitat mostrada, hem exclòs aquest algorisme de la representació de les figures 3.1.16.A i B. Així, podem visualitzar millor les relacions proximitat entre els altres mètodes de classificació. Entre aquests, *TWINSPAN* es situa en una banda dels diagrames d'ordenació (a la dreta per a *Xerobromion* i a baix per a *Quercetea ilicis*), mentre que la resta ho fan a l'altra. Aquest fet fa palesa, altre cop, la divergència de criteris entre *TWINSPAN* i els altres mètodes comparats. Entre els mètodes jeràrquics, els mètodes de Ward i *β-flexible* (β =-0.25) proporcionen particions properes a les que s'obtenen amb els mètodes partitius. En canvi, les solucions de *complete linkage* i *UPGMA* no sempre s'hi assemblen.

Finalment, presentem a les taules 3.1.11.A i 3.1.11.B els valors de l'índex de Rand corregit sorgits de la comparació de cada mètode amb la sintaxonomia tradicional, així com la mitjana i variança del nombre de tàxons fidels de cada classificació.

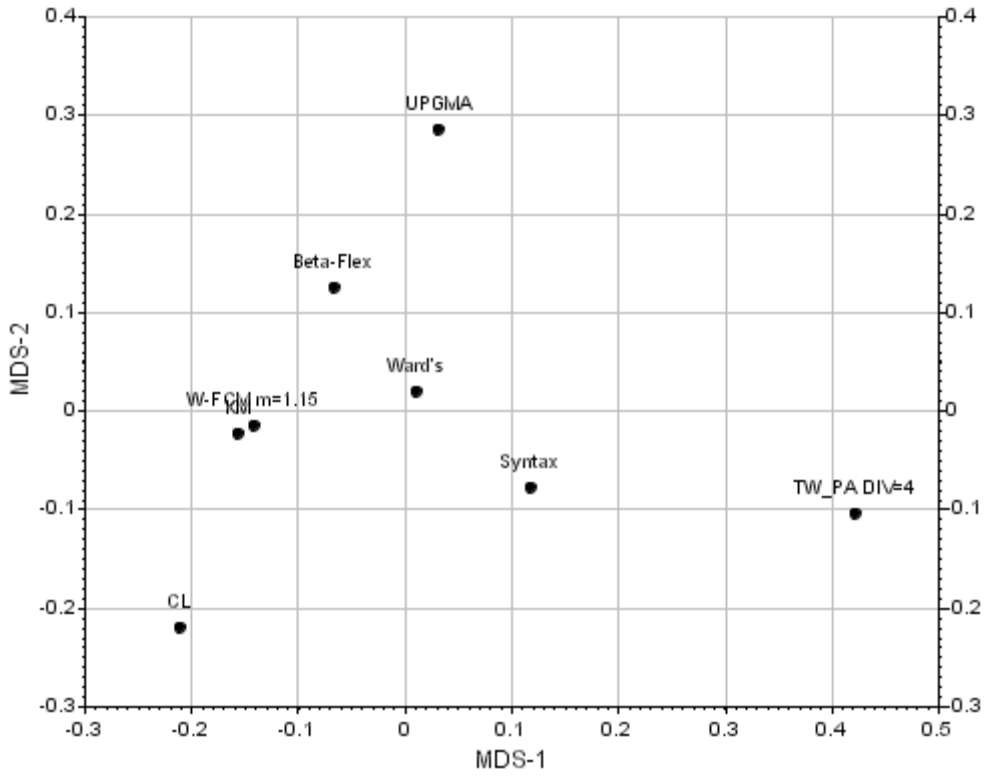


Figura 3.1.16.A: Anàlisi de Coordenades Principals de la matriu de distàncies complement de l'índex de Rand corregit, calculat entre particions derivades dels diferents mètodes de classificació aplicats a les dades de *Xerobromion erecti* (Variabilitat representada = 58.7%).

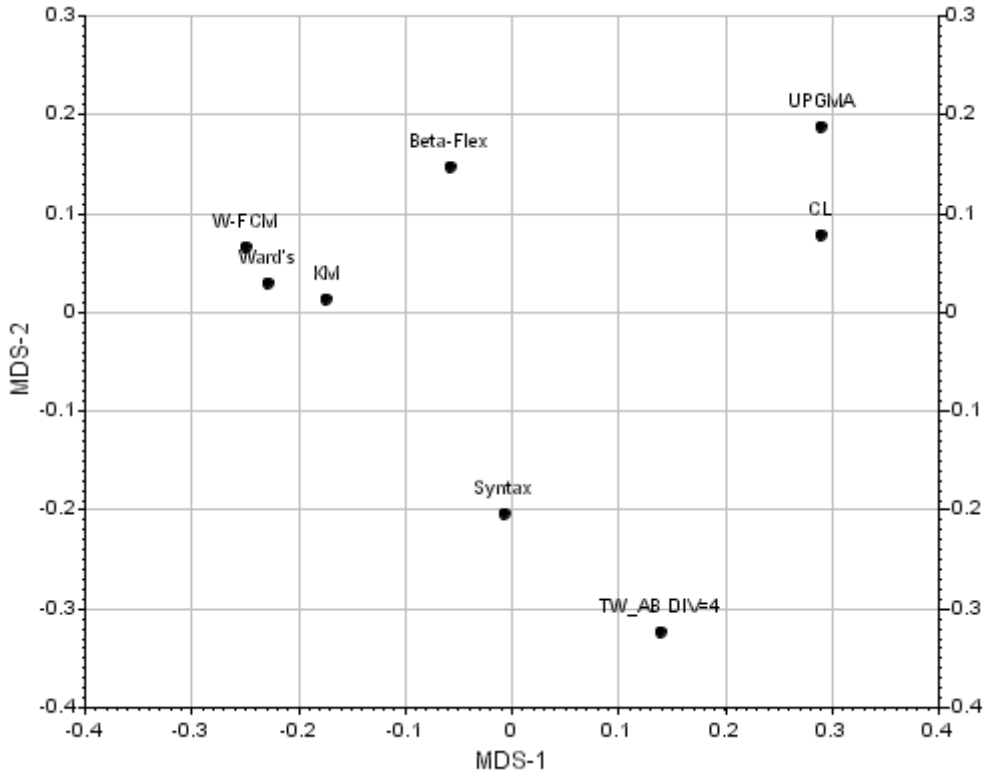


Figura 3.1.16.B: Anàlisi de Coordenades Principals de la matriu de distàncies complement de l'índex de Rand corregit, calculat entre particions derivades dels diferents mètodes de classificació aplicats a les dades de *Quercetea ilicis* sense *Quercenion* (Variabilitat representada = 73.3%).

Els mètodes més propers a la classificació tradicional són el de *Ward* per a *Xerobromion erecti* i *K-means* per a *Quercetea ilicis* (encara que al diagrama 3.1.16.B aparegui *TWINSPAN* com a més proper). Tanmateix, cal prendre aquests resultats amb cautela, per raons que discutirem en el proper apartat. Pel que fa al criteri del nombre de tàxons fidels, les variàncies són elevades, i en alguns casos molt elevades, pel que la descripció amb la mitjana és, potser, insuficient. En general, els mètodes numèrics donen valors mitjans força semblants. Per a *Xerobromion*, destaca el valor baix que presenta la classificació *single linkage*. En canvi, per a *Quercetea* sorprenentment el valor més baix l'obté la classificació tradicional. Aquest fet reforça, de nou, la necessitat de revisar la sintaxonomia de les dades, sobretot pel que fa a les subassociacions de *Quercetum cocciferae*, ja que qualsevol dels mètodes numèrics obté un millor resultat pel que fa a tàxons fidels.

Mètode i paràmetres	Rand *vs.* *Syntax*	Mitjana tàxons Φ>0.3	Variància tàxons Φ>0.3
Sintaxonomia	-	17.54	89.60
Single linkage *K*=13 (SL)	0.018	8.38	38.92
Complete linkage *K*=13 (CL)	0.440	17.69	119.56
β-flexible *K*=13 (B-Flex)	0.581	**19.31**	84.56
UPGMA *K*=13	0.488	15.54	113.60
Ward's *K*=13	**0.655**	18.46	119.94
TWINSPAN P/A, 4º division (TW_PA DIV=4)	0.489	17.14	54.59
K-means *K*=13 (KM)	0.547	17.46	88.93
FCM *m*=1.15 *K*=13 (FCM)	0.559	17.31	96.73

Taula 3.1.11.A: Avaluació dels resultats del diferents mètodes d'anàlisi de clúster en les dades de *Xerobromion erecti*. Valor de l'índex de Rand en comparació amb la classificació tradicional sintaxonòmica (*K*=13). Mitjana i variança del nombre de tàxons amb fidelitat Φ més gran que 0.3.

Mètode i paràmetres	Rand *vs.* *Syntax*	Mitjana tàxons Φ>0.3	Variància tàxons Φ>0.3
Sintaxonomia	-	8.44	39.73
Single linkage *K*=16 (SL)	0.050	**10.88**	110.78
Complete linkage *K*=16 (CL)	0.507	10.13	127.85
β-flexible *K*=16 (B-Flex)	0.522	10.00	38.40
UPGMA *K*=16	0.466	10.06	48.46
Ward's *K*=16	0.581	10.00	34.13
TWINSPAN pseudo-espècies, 4ª divisió (TW_AB DIV=4)	0.608	**10.88**	79.45
K-means *K*=16 (KM)	**0.630**	10.13	45.05
FCM *m*=1.15 *K*=16 (FCM)	0.547	10.44	56.13

Taula 3.1.11.B: Avaluació dels resultats del diferents mètodes d'anàlisi de clúster en les dades de *Quercetea ilicis* sense *Quercenion*. Valor de l'índex de Rand en comparació amb la classificació tradicional sintaxonòmica (*K*=16). Mitjana i variança del nombre de tàxons amb fidelitat Φ més gran que 0.3.

3.1.6.4 Discussió

Als països europeus, el mètode de Braun-Blanquet ha estat i és l'aproximació preeminent en l'estudi i descripció de la vegetació. En general, els investigadors en aquesta disciplina no han incorporat encara els mètodes numèrics de classificació (Mucina & van der Maarel 1989, Daniëls *et al*. 2004). La utilització de mètodes numèrics en fitosociologia és permet formalitzar millor el procés de classificació de la vegetació. No afegeix, però, molta més objectivitat al procés perquè, com en l'anàlisi tradicional, l'investigador ha de prendre moltes decisions que condicionen el resultat de l'anàlisi (Mucina 1997). A més, les decisions d'anàlisi de l'investigador sovint estan mediatitzades per la disponibilitat d'un programa informàtic, independentment de les qualitats dels mètodes que aquest implementi. L'elevada confiança en el reordenament manual de taules d'inventaris ha fet que el desenvolupament de mètodes numèrics d'anàlisi de clústers de la vegetació hagi tingut sempre la tendència a voler reproduir els resultats del mètode tradicional. Aquest desig és, certament, quelcom utòpic. El motiu és que l'anàlisi tradicional inclou moltes premisses de context que porten a la ponderació diferencial dels tàxons (Mucina 1997) i que difícilment poden ésser incloses en un mètode numèric d'anàlisi de clústers. Els algorismes *TWINSPAN* (Hill *et al*. 1974), *REBLOCK* (Podani & Feoli 1991) i COCKTAIL (Bruelheide 2000) són un exemple d'aquesta tendència d'imitació del mètode tradicional.

La resta d'algorismes que hem comparat aquí han estat desenvolupats en camps d'aplicació diferents. Les seves propietats són en general més conegudes i són aplicables a un ventall més ampli de situacions d'anàlisi de grups. A continuació discutim els avantatges i inconvenients de les diferents alternatives algorísmiques que hem presentat i comparat en aquest capítol. Alguns dels arguments ja han estat esmentat a les seccions introductòries del capítol i altres són resultats del nostre propi estudi.

Algorismes jeràrquics aglomeratius

Ja hem esmentat diverses vegades els problemes d'aquests mètodes quan són aplicats a grans volums de dades. No obstant voldríem ressaltar aquí la utilitat del mètode de minimització de la suma d'errors quadràtics, o mètode de Ward (1963), en l'anàlisi de la vegetació. Aquesta preferència ha estat expressada anteriorment també per Mucina (1982), Torres *et al*. (1995) i Cao *et al.* (1997).

En referència als altres mètodes, ja hem vist la poca utilitat de *single linkage* per a analitzar l'estructura de grups en fitosociologia. Pel que fa a *complete linkage* la deformació de l'espai de referència és un mal atribut teòric. Finalment, *UPGMA* i sobretot i *β*-flexible es presenten com a més adequats que els anteriors, però el concepte de clúster d'ambdós algorismes és menys elegant que al mètode de Ward.

Incorporant un criteri intern (com la silueta mitjana o similar) que permetés "tallar" adequadament els dendrogrames per a produir particions faria que el mètode de Ward fos una eina potent i ràpida d'anàlisi de les dades (*data mining*). En aquest sentit hem d'admetre que hauríem d'haver dedicat un major esforç a estudiar estadístics d'aquesta mena. Hem esmentat aquí els treballs de Hogeweg (1976), Popma *et al*. (1983) i Mucina & Hauser (1993), entre d'altres. Els estadístics que proposen aquests autors es basen en la dispersió "dins" (*within*) i "entre" (*between*) grups. Com a criteri intern tenen validesa per a mètodes de classificació basats en la variabilitat de les dades i els grups. Aquesta és l'aproximació del mètode de Ward, però no la dels altres mètodes jeràrquics aglomeratius. En aquest sentit desaconsellem utilitzar criteris de decisió d'aquesta mena sobre els resultats d'altres mètodes jeràrquics (com fa, per exemple, Mucina 1982).

REBLOCK

La nostra experiència amb aquest algorisme ens impedeix recomanar-lo per a l'anàlisi de la vegetació. Els nostres motius són principalment dos:

- No hi ha un procediment clar de decisió del nombre de clústers d'inventaris i d'espècies. En aquest sentit hem intentat aquí fer alguna modesta aportació, però caldria realitzar estudis més aprofundits del tema.
- L'algorisme és molt poc eficient computacionalment i, per tant, resulta inadequat per a volums de dades grans. Per a solucionar aquest problema, creiem que fóra necessari modificar el pas referent a la modificació dels blocs d'espècies i inventaris d'una determinada iteració. Concretament, caldria evitar assajar un nombre tan elevat de recol·locacions de tàxons i inventaris i proposar una estudiar una estratègia que permetés optimitzar el funcional d'una manera menys costosa.

TWINSPAN

Tal i com indica Dale (1995), l'algorisme *TWINSPAN* no cerca formes hiperesfèriques ni de variància mínima. Tot i això, es basa parcialment en l'estadística multivariant. *TWINSPAN* parteix de l'anàlisi de correspondències i, per tant, de l'espai multivariant proporcionat per la distància χ^2. No obstant, a cada divisió *TWINSPAN* escull solament la primera component principal i posteriorment ordena els inventaris mitjançant unes poques espècies fidels. Per tant, es fa difícil comparar els resultats d'aquest algorisme amb els produïts pels mètodes que prenen l'espai multivariant sencer com a font per a analitzar l'estructura de les dades. En aquest sentit, voldríem comentar el treball d'Equihua (1990). Aquest autor compara *TWINSPAN* i *FCM* utilitzant el darrer en l'espai resultant de la ordenació de l'anàlisi de correspondències (CA). Equihua

restringeix l'aplicació de *FCM* a les primeres (d'una a tres) components de l'espai de la ordenació. La inicialització de *FCM* que aquest autor utilitza consisteix en dividir el primer eix de CA en tants segments com clústers a cercar, i posteriorment assignar valors alts de pertinença als inventaris que es troben dins de cada grup. És força evident, que aquesta estratègia d'anàlisi de clústers està encarada a reproduir l'estratègia de *clustering* de *TWINSPAN* en l'àmbit de la lògica difusa. A parer nostre hi ha almenys dues objeccions a fer a l'aproximació d'Equihua: **a.**Escollir les primeres components de CA pot resultar en una disminució del soroll però introdueix més elements subjectius de decisió en el procés de *clustering*. En un principi no hi ha raons per descartar la informació de les altres components. **b.**L'espai de dades proporcionat per la distància χ^2 pot no ser el més adequat per algorismes partitius com *K-means* o *FCM* (vegeu el proper capítol 3.2). Per tant, la dependència de TWINSPAN respecte la distància χ^2 existeix però això no el fa gaire comparable a altres algorismes que operin sobre aquest espai de dades.

Pel que fa a l'algorisme *TWINSPAN* mateix, els inconvenients d'aquest mètode de classificació són principalment dos:

1) *TWINSPAN* no contempla cap mesura d'ajust al model que proposa que pogués servir de criteri intern per aturar les divisions. L'únic criteri d'aturada disponible a les implementacions del mètode és el nombre d'inventaris mínim per realitzar una divisió. És per aquest motiu que presenta una tendència a obtenir un nombre de grups igual a una potencia de dos. Aquesta manca d'un criteri intern (com ara l'estadístic χ^2) per a aturar les divisions, fa que grups que es troben en una divisió desapareguin en la següent, quan potser no hi ha massa espècies diferencials que justifiquin tal divisió. L'ús de criteris externs (vegeu per exemple, Westman 1983) és la solució pragmàtica a aquest problema.

2) *TWINSPAN* admet la classificació de nous inventaris en la seva jerarquia. Per un inventari a classificar hom pot calcular la suma dels tàxons indicadors positius (+1) i negatius (-1) que ha establert *TWINSPAN* i comparar el resultat amb el llindar de classificació (Bruelheide & Chytrý 2000). Aquesta senzilla funció discriminant es pot anomenar 'oligotètica' en el sentit que són uns pocs tàxons els que finalment decideixen l'assignació. Com que l'establiment dels tàxons indicadors depèn del primer eix de anàlisi de correspondències (CA) enlloc del conjunt del l'espai, petits canvis en l'establiment d'aquest eix tenen un efecte profund en la classificació resultant i en la determinació del grup per a nous elements. És, principalment, per aquest motiu que Bruelheide & Chytrý (2000) el desaconsellen per ésser massa depenent de l'estructura de les dades, per exemple amb la presència de petits grups d'inventaris de condicions extremes. Segons indiquen Bruelheide & Chytrý problema és més fàcil que aparegui quan els dos primers eixos de CA tinguin valors propis semblants. Per a solucionar aquesta inestabilitat potser caldria basar un algorisme semblant a *TWINSPAN* però que utilitzés un mètode d'ordenació diferent de CA.

Mètodes partitius: K-means* i *Fuzzy C-means

Els mètodes partitius com *K-means* (*KM*, MacQueen 1965) o *Fuzzy C-Means* (*FCM* Bezdek 1981), malgrat disposar d'una bibliografia teòrica molt més extensa que *TWINSPAN* tenen una aplicació a la sintaxonomia numèrica relativament modesta. Tot i això, *FCM* sembla que pot guanyar adeptes per acostar el factor de l'indeterminisme a la fitosociologia (Mucina 1997).

La principal desavantatge de *KM* és la tendència a trobar mínims locals. És per aquest motiu que existeix força literatura proposant algorismes alternatius d'optimització del mateix funcional. En espera d'una possible millora en aquest aspecte, les aplicacions de *KM* poden fer ús de la correcció *leave-one-out*. Aquesta correcció, malgrat augmentar sensiblement la mobilitat de l'algorisme, no assegura trobar el màxim global del funcional. Per la seva banda, l'algorisme *FCM* presenta més mobilitat com més alt sigui l'exponent de *fuzziness*. L'increment de "borrositat" en les solucions en augmentar l'exponent m constitueix una contrapartida a l'increment de la mobilitat. Dale (1988a), examina diferents aproximacions *fuzzy* a la fitosociologia i conclou que *FCM* es massa sensible als valors de l'exponent m. En el nostre cas, creiem que *FCM* es capaç de trobar estructures de grup interessants per a valors de *m* intermedis. Els valors massa petits donen solucions molt semblants a *KM* i, per tant, no justifiquen l'enfoc difús. D'altra banda, els valors massa alts proporcionen particions completament borroses i, per tant, sense informació. Els principals avantatges de *FCM* respecte *KM* són la possibilitat d'obtenir graus pertinença intermedis i una major facilitat per trobar òptims absoluts.

En quant a les estratègies d'inicialització, l'ús d'una partició derivada del mètode jeràrquic aglomeratiu de minimització de l'increment de la suma de quadrats (o mètode de Ward) és recomanable quan el nombre de grups a cercar faci impracticable estratègies basades en l'exploració aleatòria de l'espai de solucions. No obstant, i sobretot per a *K-means*, cal ésser conscients que aquesta aproximació imposa una constricció molt forta en la solució.

Una gran avantatge dels mètodes partitius és la possibilitat d'emprar estadístics com a criteris interns. És important que els usuaris de mètodes de *clustering* puguin determinar el nombre de clústers presents a les seves dades sense utilitzar coneixements externs. Desgraciadament, els treballs d'aplicació fitosociològica en fan un ús molt limitat, probablement per la manca de disponibilitat d'aquests estadístics en programes d'ús comú. En el nostre estudi hem vist com un dels estadístics recomanats en altres camps com és *pseudo-F* (Milligan & Cooper 1985), té una utilitat molt limitada a l'hora de posar de manifest estructures fitosociològiques. La raó és que la distribució dels inventaris en l'espai dels tàxons s'adiu a la distribució normal multivariant. Pel que fa al coeficient de partició o coeficient de partició (o coeficient de Dunn) normalitzat i l'entropia normalitzada s'han mostrat com a criteris força

semblants. Dale (1988c) diu, en referència al primer estadístic: '*It is certainly not clear why, if you accept the need for a fuzzy solution, the desirable solution should be chosen as that which is more crisp!*'. Aquesta aparent contradicció d'avaluar un mètode *fuzzy* amb una eina que mesura la rigidesa de la partició resultant no creiem que sigui vàlida, sempre i quan les comparacions de l'estadístic es facin entre solucions obtingudes amb el mateix exponent de *fuzziness* (m). D'altra banda, sí que és cert que ambdós estadístics tenen una tendència intrínseca a afavorir un nombre de clústers alt (Equihua 1990). La raó és que es basen exclusivament en els valors de pertinença de la partició difusa i no en la geometria de les dades. Finalment, el criteri de la silueta mitjana malgrat tenir una definició força *ad hoc* es mostra més sensible que els criteris anteriors. Presenta, també, un biaix cap a afavorir un nombre de grups gran (sobretot quan hi ha pocs inventaris a cada clúster). No obstant, considerem que podria ser utilitzat de manera freqüent com a eina de validació interna per a algorismes basats en la comparació de distàncies a centroides.

Una darrera propietat avantatjosa dels mètodes partitius és que si un grup que apareix en una partició és prou aïllat i compacte, no té perquè subdividir-se o canviar de composició en una partició amb un nombre de grups més gran. Aquest és un inconvenient que sí hem observat en l'algorisme *TWINSPAN*.

Atenint-nos als mètodes actualment disponibles, creiem que sense un coneixement aprofundit del criteri de clúster que cada mètode o algorisme determinat utilitzen, i sense un coneixement previ del concepte concret de clúster que un camp d'aplicació necessita, l'aplicació l'anàlisi de clústers pot conduir a resultats poc adequats al coneixement de l'espai de dades d'estudi. En aquest sentit, considerem que caldria aclarir si la classificació numèrica de la vegetació ha de reproduir els resultats de la classificació tradicional o pot incorporar conceptes de clúster més fàcilment formalitzables.

Com a apreciació subjectiva sobre aquest tema, creiem que la existència de estructures molt variades en l'espai de relacions entre inventaris de vegetació hauria de portar al desenvolupament i ús de mètodes d'anàlisi que permetéssin detectar alhora grups i gradients (no necessàriament linears) de variació. Aquest enfoc portaria a una descripció més complerta de l'espai multivariant de la vegetació i permetria prendre decisions fitosociològiques amb un major fonament numèric.

3.1.6.5 Conclusions

En aquest capítol hem intentat descriure alguns dels mètodes de *clustering* utilitzats per al tractament d'inventaris de vegetació. Seguidament resumim, breument, les principals conclusions que es poden extreure de la lectura del present capítol:

- Els algorismes jeràrquics aglomeratius són ràpids però incòmodes per a volums grans de dades. El mètode jeràrquic més eficient fitosociològicament és el de Ward, però resultaria més útil si hom disposés d'un criteri fàcil d'establiment de punts de tall. Proposem aquí emprar l'estadístic silueta mitjana (Rousseeuw 1987).
- *TWINSPAN* (i possiblement *REBLOCK*) són mètodes de *clustering* que tendeixen a proporcionar més tàxons fidels als grups generats.
- *REBLOCK* necessita escollir el nombre de grups de variables i no disposa d'un criteri intern de decisió fàcilment interpretable. A més, és ineficient per a volums de dades grans.
- *TWINSPAN* presenta una tendència a produir un nombre de clústers potencia de dos. Al no disposar d'un criteri intern d'aturada de les divisions alguns grups ben determinats a les primeres divisions són destruïts sense massa fonament en divisions successives. A més és massa depenent a canvis en l'estructura de les dades.
- Els mètodes partitius necessiten bones estratègies d'inicialització per tal d'evitar mínims locals a l'hora d'analitzar volums grans de dades. Utilitzar el tall d'un dendrograma generat amb el mètode de Ward pot ésser una solució ràpida al problema.
- L'ús de la correcció *leave-one-out* a *K-means* augmenta la mobilitat de l'algorisme i permet incrementar la eficiència a l'hora de trobar mínims globals de funcional.
- Entre els criteris interns comparats, creiem que l'estadístic més sensible per a detectar estructures de grup en dades de vegetació és la silueta mitjana de la partició. En general, però, constatem la necessitat de disposar de criteris interns d'avaluació eficients per a aquest tipus de dades.
- *FCM* té més mobilitat que *K-means* per a trobar mínims de funcional. Aquesta mobilitat s'incrementa en augmentar l'exponent de *fuzziness*. Tanmateix, si aquest exponent és molt alt *FCM* produeix particions completament difuses i, per tant, no informatives. Globalment, el criteri de clúster de *FCM* no difereix prou del de *K-means* com per a proporcionar solucions millors.
- L'anàlisi del nombre mitjà de tàxons fidels ha posat de manifest, de nou, la manca de tàxons fidels de la classificació tradicionalment acceptada dels sintàxons de *Quercetea ilicis*, en aquest cas sense incloure l'ordre *Quercenion ilicis*.
- Els tres mètodes sintaxonomia numèrica de la vegetació més propers a l'aproximació tradicional són, sense importar l'ordre: el mètode jeràrquic de *Ward*, els mètodes partitius (*KM* i *FCM*) i *TWINSPAN*.

Capítol 3.2: Transformacions, mesures de proximitat i classificació

3.2.1 L'abundància dels tàxons: Escales i transformacions

3.2.1.1 Escales de mesura de l'abundància d'un tàxon en la recollida de dades.

Una de les decisions que cal prendre a l'hora d'estudiar les espècies d'una comunitat és la de com mesurar llur abundància. A primer cop d'ull, hom pot pensar que mesurar el nombre d'individus, la biomassa o el recobriment són estimacions d'abundància no gaire diferents. No obstant, hi ha espècies amb molts individus però de recobriment o biomassa molt baix, i a l'inrevés, a vegades uns pocs individus d'una espècie tenen un recobriment molt alt. Si el cost de mesura no és un problema, el mètode de punts donarà una estimació amb molta precisió de l'abundància dels tàxons. Sovint, però, cal estalviar esforços i s'empren mètodes de mostratge menys precisos i més subjectius però molt més ràpids, com l'escala d'abundància/dominància de Braun-Blanquet. Aquesta escala (Taula 3.2.1) combina la mesura del recobriment i l'abundància en set estats principals.

Escala	Definició dels estats	Interval de cobertura	Equivalència numèrica Braun-Blanquet (1964)
r	Planta molt rara, molt pocs individus.	?	( )
+	Planta escassa; recobriment molt baix (<5%)	5%	0.1
1	Planta abundant però de recobriment molt baix o planta escassa amb recobriment entre el 5% i el 10%	5%	5.0
2	Planta molt abundant però de recobriment molt baix o bé recobriment comprès entre el 11% i el 25%	15%	17.5
3	Recobriment comprès entre el 26% i el 50%	25%	37.5
4	Recobriment comprès entre el 51% i el 75%	25%	62.5
5	Recobriment superior al 75%	25%	87.5

Taula 3.2.1: Escala d'abundància/dominància de Braun-Blanquet (de Westhoff & van der Maarel, 1973).

Alguns autors (vegeu Westhoff & van der Maarel 1973: 640) proposaren subdividir l'estat '2' de l'escala en tres subestats: '2m' = cobertura fins al 5%, '2a' = cobertura 6-12% i '2b' = cobertura 12-25%. Aquesta subdivisió augmenta la dificultat del mostratge, raó per la qual no és una pràctica seguida per tots els fitosociòlegs.

La definició dels quatre primers estats de l'escala de Braun-Blanquet ('r','+','1' i '2') implica la mesura combinada de dos conceptes; l'abundància en nombre d'individus del tàxon i la seva dominància o recobriment. Els darrers tres estats de l'escala estan definits exclusivament a partir del recobriment, assumint que si el recobriment és alt s'entén que la planta és abundant en nombre d'individus. Aquesta combinació de criteris fa que l'amplada dels intervals de cobertura dels diferents estats no sigui constant: si bé els tres darrers tenen una amplada del 25% els primers són molt més estrets. Per tant, l'escala de Braun-Blanquet no és lineal pel que fa a l'estimació de la cobertura.

Una altra escala emprada en la recollida de dades és la de *Domin* (Currall 1987), que mostrem a la taula 3.2.2. L'escala *Domin* es basa en la dominància, entesa com el percentatge de cobertura d'una espècie a la localitat inventariada. Es tracta d'una escala amb més estats que l'escala de Braun-Blanquet, i per tant, més difícil d'aprendre a utilitzar. Altra vegada, els intervals entre estats de l'escala són variables. L'escala *Domin* no sembla haver assolit gaire acceptació per part dels fitosociòlegs.

Domin	**% de cobertura**	**Interval**	**Valor mig**
10	95-100	15%	97.5
9	75-94	19%	84.5
8	50-74	24%	62.0
7	33-49	16%	41.0
6	25-32	7%	28.5
5	10-24	14%	17.0
4	5-9	4%	7.0
3	1-4	3%	2.5
2	<1		
1	<1	1%	0.5
+	<1		

Taula 3.2.2: Escala *Domin*. Modificat de Currall (1987)

El fet que les dues escales de mesura tinguin una precisió més alta en les classes d'abundància baixa que en les classes d'abundància alta denota una tendència a sobrevalorar la presencia del tàxon respecte a la seva dominància. Segons Currall (1987), aquesta és una pràctica comú en la observació de la natura. El debat sobre la importància de les classes d'abundància baixes respecte a les altes es manté a l'hora de parlar de transformacions numèriques de l'escala per a finalitats analítiques.

3.2.1.2 Transformacions escalars

Les escales de Braun-Blanquet i *Domin* són escales ordinals alfanumèriques. Per tal de poder operar aritmèticament amb les abundàncies dels tàxons cal transformar els valors ordinals a una escala enterament numèrica. D'altra banda, és interessant la proposta de Dale (1989), que suggereix l'ús de mètriques basades en escales ordinals.

La taula 3.2.3 (modificat de van der Maarel 1979) recull la majoria de les transformacions proposades. La diferència principal entre unes transformacions i les altres rau en la importància que hom dóna a valors alts d'abundància respecte als valors baixos: és important quantificar adequadament la relació entre l'absència/presència respecte els diferents valors d'abundància un cop l'espècie ja és present. La raó entre la traducció numèrica dels estats '+' i '5' varia molt segons la transformació: des de valors 1 per a l'aproximació de presencies i absències fins a valors propers a zero per a les transformacions de Braun-Blanquet (1964) i Tüxen-Ellenberg (1937). Aquestes darreres transformacions equivalen aproximadament a traduir els estats ordinals en els valors mitjans de cobertura de cada classe.

	r	+	1	2	3	4	5
Tüxen-Ellenberg 1937	( )	0.1	2.5	15	37.5	62.5	87.5
Braun-Blanquet transf. 1964	( )	0.1	5.0	17.5	37.5	62.5	87.5
Etter 1949	0	1.0	10.0	20.0	37.5	62.5	87.5
Londo 1971	0	0	0	1	3	5	7
van der Maarel 1966	0	0	0.5	1	2	4	5
Dagnelie 1960	( )	0.1	1	2	3	4	5
Schwickerath 1931	( )	0.25	1	2	3	4	5
Coetzee & Werger 1973	1	5	10	20	30	40	50
Barkman et al. 1964	0	2	3	6	8	9	10
Barkman ms	0.5	1	2	4	8	10	10
Schmid-Kuhn 1970	2	4	6	7	8	9	9
Moore 1966	10	10	11	12	13	14	15
Jensén 1978	1.00	1.00	1.69	2.25	2.58	2.79	2.94
Combined Transform (v.d. Maarel 1979)	1	2	3	5	7	8	9
Angular transf. (v.d. Maarel 1966)	0 (1)	1	2	3	5	7	9

Taula 3.2.3: Transformacions aplicades a l'escala d'abundància/dominància de Braun-Blanquet. Modificat de van der Maarel (1979). Vegeu referències allà.

L'escurçament o allargament del rang de valors obtinguts per les diferents transformacions de la taula 3.2.3 pot ser aproximat mitjançant una forma general:

$$y = x^w$$

on x és el valor numèric sorgit de la transformació combinada (van der Maarel 1979). Anàlogament, Currall (1987) proposà transformar valors de l'escala *Domin* basant-se en una transformació similar. D'altra banda, vegeu a Noest *et al*. (1989) un exemple de transformació més complexe de l'escala de Braun-Blanquet.

A partir d'emprar diferents exponents w, hom pot generar transformacions que aproximin les de la taula 3.2.3. Des de les que donen més pes a les espècies rares - $w = 0.25$ - fins a les fan dominar completament les espècies més abundants - $w = 2.0$ o $w = 4.0$. A la taula 3.2.4 i la figura 3.2.1 mostrem el resultat de canviar l'exponent de transformació.

Br-Bl	Comb	*w*=0.25	*w*=0.5	*w*=0.75	*w*=1.0	*w*=1.5	*w*=2.0	*w*=4.0
r	1	1.00	1.00	1.00	1.00	1.00	1	1
+	2	1.19	1.41	1.68	2.00	2.83	4	16
1	3	1.32	1.73	2.28	3.00	5.20	9	81
2m	4	1.41	2.00	2.83	4.00	8.00	16	256
2a	5	1.50	2.24	3.34	5.00	11.18	25	625
2b	6	1.57	2.45	3.83	6.00	14.70	36	1296
3	7	1.63	2.65	4.30	7.00	18.52	49	2401
4	8	1.68	2.83	4.76	8.00	22.63	64	4096
5	9	1.73	3.00	5.20	9.00	27.00	81	6561
Min % of Max		57%	33%	19%	11%	4%	1.2%	0.01%

Taula 3.2.4: Transformacions obtingudes a partir d'aplicar un exponent *w* al valor de la transformació combinada de van der Maarel (1979, segona primera columna).

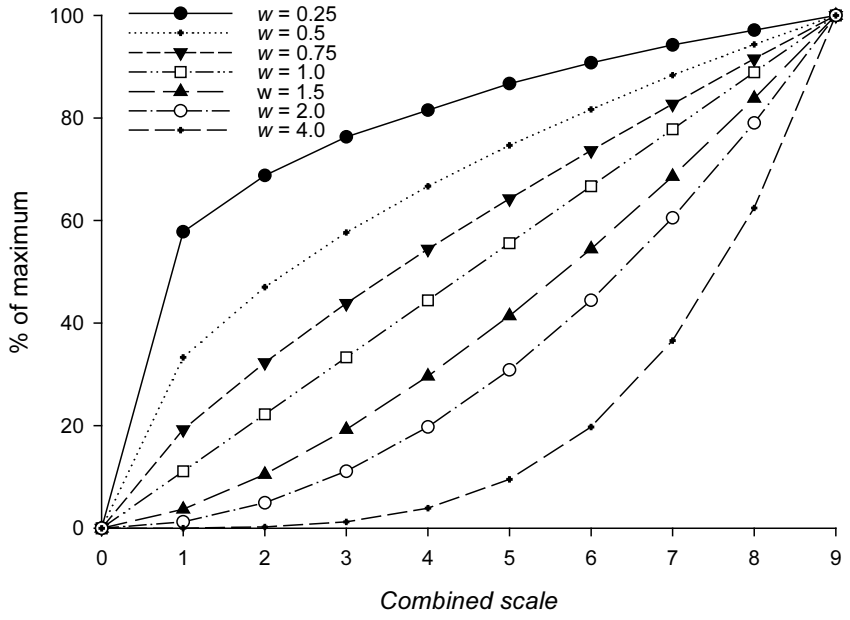


Figura 3.2.1: Efecte de l'exponent de transformació *w* sobre l'escala combinada de van der Maarel (1979). L'eix horitzontal mostra el valor de l'escala combinada. L'eix vertical indica el percentatge del valor transformat respecte el màxim que assoleix per a la transformació del valor 9.

Van der Maarel (1979) comparà l'efecte de diferents exponents en relació a la classificació per mètodes jeràrquics aglomeratius. Aquest autor arribà a la conclusió que els exponents intermedis ($w \approx 1.0$) eren els que generaven classificacions més properes al patró fitosociològic amb el que les comparava. Campbell (1978) qui també comparà algunes transformacions aplicades a mètodes jeràrquics recomanà la transformació de Coetze & Werger (1993), que també es pot considerar com una transformació amb exponent intermedi.

En treballs aplicats és habitual aplicar a les dades la transformació arrel quadrada després de la traducció a l'escala combinada (per exemple Hakes 1994). Aquesta transformació, evidentment, equival a utilitzar l'exponent w=0.5. Donat que el mètode sigmatista de classificació de la vegetació es basa en bona mesura en la presència o absència d'espècies diferencials, no és sorprenent que exponents més grans que 1.0 sobrevalorin les espècies abundants i facin que la resposta d'una classificació numèrica difereixi excessivament de la classificació tradicional.

D'altra banda, Campbell (1978) i van der Maarel (1979) recolzen que en una classificació jeràrquica que comprengui diferents nivells d'agrupament de comunitats el pes de les espècies abundants respecte a les poc abundants canviï segons el nivell d'abstracció. Per als agrupaments de més alt nivell (és a dir, classes, ordres…) la base és exclusivament la presència/absència de tàxons. En baixar a nivells més baixos (aliances, associacions i subassociacions) la diferència quantitativa entre els tàxons es fa més important. Les subdivisions més fines es caracteritzen per la dominància o preeminència d'una o poques espècies, donant-se un fort pes a la diferència de cobertura-abundància. La nostra opinió és que, en l'anàlisi de les relacions de florístiques, caldria utilitzar la mateixa transformació per a tots els nivells sintaxonòmics. Com a regla pràctica general, si hom vol treure profit de la informació proporcionada per les abundàncies, però sense sobrevalorar les espècies més abundants respecte a les poc abundants, cal escollir una transformació de les dades que mantingui els valors més baixos entre un 10% i un 60% dels valors més alts. Un percentatge inferior sobrevalorarà les espècies abundants i un percentatge superior acostarà massa les abundàncies al cas qualitatiu perdent-se la informació que puguin aportar.

Un cop escollida una transformació de les dades cal escollir una mesura de proximitat per avaluar numèricament les semblances i diferències entre inventaris. Aquesta elecció a vegades va implícita en la posterior utilització d'un determinat mètode d'anàlisi multivariant. En altres ocasions cal fer una elecció explícita de la mesura de proximitat *a priori* d'aplicar mètodes d'anàlisi. En tots els casos, però, la mètrica implícita o explícita té una importància cabdal en la estructura que resulta de les anàlisis. A la propera secció revisarem algunes de les mesures de proximitat emprades en sintaxonomia numèrica. Per a una discussió extensa sobre mesures de proximitat vegeu Legendre & Legendre (1998).

3.2.2 Mesures de proximitat: Similaritats i distàncies

3.2.2.1 Conceptes bàsics

Anomenarem mesura de proximitat (o d'associació) aquella mesura escalar que expressa el grau de semblança (similaritat) o diferència (dissimilaritat) entre dos elements, en el nostre cas entre dos inventaris. Una similaritat (s) és una mesura de proximitat entre dos objectes ω_A i ω_B, que compleix:

1) $s(\omega_A,\omega_B)\geq 0 \quad \forall(\omega_A,\omega_B)$, 2) $s(\omega_A,\omega_B)=s(\omega_B,\omega_A)$, 3) $s(\omega_A,\omega_B)\leq s(\omega_A,\omega_A)$.

D'altra banda, una dissimilaritat o distància (d) és una mesura de proximitat entre dos objectes ω_A i ω_B, que compleix:

1) $d(\omega_A,\omega_B)\geq 0 \quad \forall(\omega_A,\omega_B)$, 2) $d(\omega_A,\omega_A)=0$.

Algunes dissimilaritats poden tenir una cota superior (dissimilaritat acotada), però moltes no en tenen. D'altra banda, per a que una dissimilaritat pugui ésser considerada una mètrica ha de complir l'anomenada desigualtat triangular:

3) $d(\omega_A,\omega_C)\leq d(\omega_A,\omega_B)+d(\omega_B,\omega_C) \quad \forall(\omega_A,\omega_B,\omega_C)$

S'anomena semi-mètrica la mesura de dissimilaritat que no compleix, per a tots els casos, aquesta darrera desigualtat. Finalment, hom considera que una mètrica (o distància) és euclidiana (noteu les minúscules) o "euclidianitzable" si admet una representació complerta en un espai vectorial nodrit d'una norma Euclidiana. Les propietats geomètriques de les dissimilaritats (o de les similaritats transformades a dissimilaritats) són importants en alguns casos. Gower & Legendre (1986) revisaren les propietats mètriques i euclidianes de molts dels coeficients de dissimilaritat que aquí tractem. Les seves conclusions es poden trobar també a Legendre & Legendre (1998).

Per a una similaritat acotada en l'interval $[0,1]$, la distància corresponent es pot obtenir escollint una de les següents transformacions:

1. La distància complement: $d(\omega_1,\omega_2)=1-s(\omega_1,\omega_2)$.
2. La transformació de Gower (1966): $d(\omega_1,\omega_2)=\sqrt{s(\omega_1,\omega_1)+s(\omega_2,\omega_2)-2\cdot s(\omega_1,\omega_2)}$ que, si $s(\omega_A,\omega_A)=1$, és equivalent a: $d(\omega_1,\omega_2)=\sqrt{1-s(\omega_1,\omega_2)}$.
3. L'arrel del complement del quadrat: $d(\omega_1,\omega_2)=\sqrt{1-s(\omega_1,\omega_2)^2}$.

Per a transformar una dissimilaritat en similaritat, com que hi ha dissimilaritats no acotades hom pot, en primer lloc normalitzar les distàncies mitjançant:

$$d_{NORM}(\omega_1,\omega_2)=\frac{d(\omega_1,\omega_2)}{d_{\max}} \quad \text{o} \quad d_{NORM}(\omega_1,\omega_2)=\frac{d(\omega_1,\omega_2)-d_{\min}}{d_{\max}-d_{\min}}$$

i a continuació aplicar les equacions anteriors revertides. Una altra opció seria emprar la transformació exponencial $s(\omega_1,\omega_2)=\exp\left\{-\alpha\cdot d(\omega_1,\omega_2)^{\beta}\right\}$, on α i β serien dos paràmetres.

3.2.2.2 La distància Euclidiana i les dobles absències

La distància pitagòrica o distància Euclidiana (emprarem la inicial majúscula per diferenciar aquesta mètrica de la propietat euclidiana d'algunes dissimilaritats) es defineix com a:

$$d_{EU}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{P} \left(x_{1j} - x_{2j}\right)^2}$$

on hem substituït els objectes ω_1 i ω_2 per la seva representació en vectors de dades: $\mathbf{x}_1$ i.$\mathbf{x}_2$.

La distància Euclidiana presenta alguns problemes quan és aplicada a matrius de composició específica. El més important és el que s'anomena "problema de les dobles-absències". En ecologia s'accepta que la absència d'una espècie d'una mostra no hauria d'ésser considerada simplement com el límit zero d'una variable quantitativa, sinó que és qualitativament diferent de qualsevol valor positiu: Mentre que la presència d'una espècie és una evidència positiva de que les condicions ambientals cauen dins del rang de tolerància de la espècie; la seva absència no és una evidència positiva del contrari (Goodall 1973). Romandre fora del rang de tolerància de l'espècie pot donar-se per excés o per defecte. És per aquesta raó que, en general, és preferible excloure les dobles-absències en calcular similaritats o dissimilaritats entre inventaris. Tanmateix, encara hi ha autors que defensen l'ús de la informació de la doble absència (veure Tamás *et al*. 2001). Els coeficients de proximitat que exclouen de la comparació les dobles absències s'anomenen coeficients asimètrics. Per contra, els coeficients anomenats simètrics tracten el valor '0' exactament com qualsevol altre valor. Els emprarem quan zero, '0', representi el mateix tipus d'informació que qualsevol altre valor (per exemple 0 mgO_2/L o 8.2 mgO_2/L). La distància Euclidiana és una mesura de proximitat simètrica, però també ho són, per exemple, la distància de Manhattan o la mètrica de Canberra (vegeu més endavant).

En dades de vegetació és normal que el nombre d'inventaris a comparar i el nombre de tàxons (P) sigui elevat. Per tant, incloure les dobles absències en les comparacions pot resultar en unes relacions de semblança en que la semblança entre comunitats amb poques espècies es vegi incrementada respecte la semblança entre comunitats amb un nombre d'espècies elevat. Un altre problema de la comparació d'inventaris, associat també a la diferenciació entre absència i presència, és el tractament de l'abundància total (suma) de l'inventari. Les comparacions entre inventaris amb abundàncies totals molt diferents resultaran en distàncies més grans que les comparacions entre inventaris normalitzats per tenir la mateixa abundància total. Així, el problema dels dobles zeros s'accentua quan els inventaris a comparar tenen abundàncies totals diferents. Es pot arribar a donar el cas en que dos inventaris sense cap espècie en comú i valors d'abundància total baixos apareguin com a més similars que dos inventaris que tenen els mateixos tàxons però amb abundàncies molt diferents. Així, en general, és difícil distingir i ponderar la dissimilaritat està basada en la diferent composició específica vers la basada en una abundància diferent dels tàxons o l'inventari.

3.2.2.3 Tres coeficients de similaritat binaris: *SMC*, Jaccard i Sørensen

Si hom vol comparar dos inventaris, ω_1 i ω_2, en base als tàxons que presenten podem construir una simple taula 2x2, on :

'a' és el nombre d'espècies presents a ω_1 i ω_2 alhora.
'b' és el nombre d'espècies presents a ω_1 i absents a ω_2.
'c' és el nombre d'espècies presents a ω_2 i absents a ω_1.
'd' és el nombre d'espècies absents a ω_1 i ω_2 alhora.

	Pres. ω_1	Abs. ω_1
Pres. ω_2	***a***	***b***
Abs. ω_2	***c***	***d***

A partir d'establir aquests elements bàsics de similaritat binària, un índex simètric senzill és el *simple matching coefficient* (*SMC*):

$$s_{SMC}(\mathbf{x}_1, \mathbf{x}_2) = \frac{a+d}{a+b+c+d}.$$

Si tractem amb comunitats d'espècies, tenim dues alternatives asimètriques força comunes, l'índex de Jaccard (1901, 1908):

$$s_{JAC}(\mathbf{x}_1, \mathbf{x}_2) = \frac{a}{a+b+c},$$

o bé el de Sørensen (1948):

$$s_{SOR}(\mathbf{x}_1, \mathbf{x}_2) = \frac{2a}{2a+b+c} = \frac{a}{(2a+b+c)/2}.$$

La relació entre Jaccard i Sørensen és monotònica: són completament equivalents pel que fa a l'ordre de les similaritats. No obstant, Sørensen dóna un pes doble a les coincidències.

Emulant l'estudi de Hajdu (1981) hem dissenyar dues senzilles series de comparació (OCCAS) de presència/absència per a posar de relleu les diferències entre aquests dos índexs de similaritat qualitatius: A la primera, *OCCAS-PA1*, els diferents elements de la sèrie van perdent consecutivament elements. A la segona sèrie, *OCCAS-PA2*, hi ha una substitució d'unes espècies per unes altres. A la figura 3.2.2 hom comprova que, a *OCCAS-PA1*, l'índex de Jaccard es comporta linealment però l'índex de Sørensen no. En canvi, a *OCCAS-PA2*, l'índex amb una resposta més lineal és el de Sørensen. Per tant, Jaccard és un índex adequat quan es comparen asimètricament dues mostres en la que una és complerta i l'altre es presenta com una submostra de les espècies de la primera. En canvi, l'índex de Sørensen té un comportament numèric que afavoreix una comparació simètrica dels inventaris amb un nombre total de tàxons semblant.

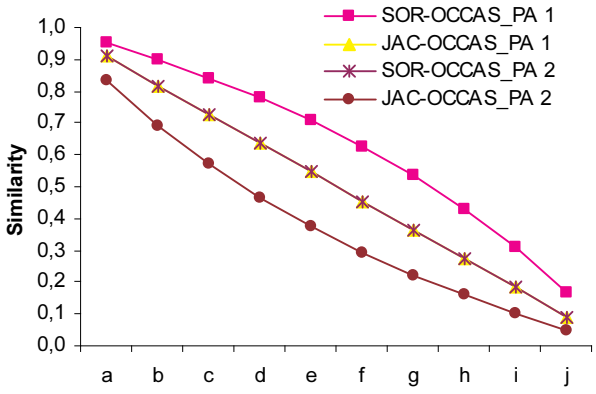


Figura 3.2.2: OCCAS-PA1 i OCCAS-PA2. Resposta dels índexs de Jaccard i Sørensen.

3.2.2.4 Generalitzacions de l'índex de Jaccard al cas quantitatiu

Alguns coeficients de similaritat quantitatius donen molt pes a la ocurrència mínima d'espècies. És el cas dels índexs de Gleason (1920) i Ellenberg (1956), que són dues generalitzacions de l'índex de Jaccard al cas quantitatiu:

$$s_{Gleason}(\mathbf{x}_1, \mathbf{x}_2) = \frac{Ma}{Ma + Mb + Mc} \quad \text{i} \quad s_{Ellenberg}(\mathbf{x}_1, \mathbf{x}_2) = \frac{\frac{1}{2} Ma}{\frac{1}{2} Ma + Mb + Mc}.$$

on Ma és la suma de valors no nuls en ambdós inventaris, Mb és la suma de valors no nuls només a ω_1 i Mc la suma de valors no nuls només a ω_2. Goodall (1973) mostra que la similaritat entre inventaris mesurada per aquests índexs és incrementada substancialment en afegir una espècie amb ocurrència baixa. Un altre inconvenient es presenta quan totes les espècies són comunes, cas en que els índexs de Ellenberg i Gleason es comporten com si fossin mesures qualitatives donant una similaritat màxima independentment dels valors d'abundància!

L'índex de Spatz és una modificació del de Gleason que intenta corregir els defectes que hem esmentat:

$$s_{Spatz}(\mathbf{x}_1, \mathbf{x}_2) \cdot \frac{1}{P} \sum_{ji=1}^{P} \frac{\max(x_{1j}, x_{2j})}{\min(x_{1j}, x_{2j})}.$$

Tanmateix, Campbell (1978) desaconsella l'ús de qualsevol dels tres índexs per a la classificació d'inventaris.

La raó de similaritat o *similarity ratio* (Wishart 1969) és una altra generalització de l'índex de Jaccard al cas quantitatiu. Aquest índex no pateix els defectes de les anteriors mesures:

$$SR(\mathbf{x}_1, \mathbf{x}_2) = \frac{\sum_{j=1}^{P} x_{1j} \cdot x_{2j}}{\sum_{j=1}^{P} x_{1j}^2 + \sum_{j=1}^{P} x_{2j}^2 - \sum_{j=1}^{P} x_{1j} \cdot x_{2j}}$$

Similarity ratio fou recomanada en els estudis de comparació de mesures de proximitat realitzats per Campbell (1978) i Hajdu (1981). Autors que han utilitzat aquesta mesura són, per exemple, Jensén (1978), Janssen (1975), van der Maarel *et al*. (1978) o Hakes (1994). És interessant la proposta que fan Janssen (1975) i van der Maarel *et al*. (1978) d'utilitzar la raó de similaritat com a mesura de comparació de la similaritat entre un inventari i un grup d'inventaris, representat pel vector **z**:

$$SR(\mathbf{x}, \mathbf{z}) = \frac{N \cdot \sum_{j=1}^{P} x_j \cdot z_j}{N^2 \cdot \sum_{j=1}^{P} x_j^2 + \sum_{j=1}^{P} z_j^2 - N \cdot \sum_{j=1}^{P} x_j \cdot z_j}$$

3.2.2.5 Mesures de proximitat relacionades amb la diferència en valor absolut

La mètrica de Manhattan o distància ciutat es calcula com a suma de les diferències en valor absolut:

$$d_{Man}(\mathbf{x}_1, \mathbf{x}_2) = \sum_{j=1}^{P} \left| x_{1j} - x_{2j} \right|$$

Una raó per a sumar les diferències en valor absolut i no elevant-les al quadrat, com en la distància Euclidiana, és que, en el segon cas, les diferències entre abundàncies grans incrementen molt més que les diferencies entre valors petits.

En l'àmbit de l'anàlisi de proximitats mètrica (*metric scaling*), Cuadras & Fortiana (1995) proposen utilitzar una modificació de la distància de Manhattan que presenta la propietat de ser euclidiana. Cuadras & Fortiana l'anomenen distància valor absolut:

$$d_{AbsVal}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{P} \left| x_{1j} - x_{2j} \right|} \, .$$

Com en el cas de la distància Euclidiana, la mètrica de Manhattan i la distància valor absolut són mesures simètriques. Una manera de fer la mètrica de Manhattan més adient a l'anàlisi de dades de composició específica és dividint els valors per l'abundància total de l'inventari (vegeu la transformació del perfil d'espècies, a l'apartat 3.2.2.6). Amb aquesta transformació hom obté la mètrica de Manhattan "relativitzada" (Faith *et al*. 1987), equivalent al doble de l'índex d'associació de Whittaker (1952, in: Legendre & Legendre 1998):

$$d_{ManRel}(\mathbf{x}_1, \mathbf{x}_2) = 2 \cdot d_{Whittaker}(\mathbf{x}_1, \mathbf{x}_2) = \sum_{j=1}^{P} \left| \frac{x_{1j}}{x_{1+}} - \frac{x_{2j}}{x_{2+}} \right| \text{, on } x_{i+} = \sum_{j=1}^{P} x_{ij} \, .$$

Una mesura de dissimilaritat relacionada és la distància de Bray-Curtis (1957):

$$d_{BC}(\mathbf{x}_1, \mathbf{x}_2) = \frac{\sum_{j=1}^{P} \left| x_{1j} - x_{2j} \right|}{\sum_{j=1}^{P} (x_{1j} + x_{2j})} = \frac{\sum_{j=1}^{P} \left| x_{1j} - x_{2j} \right|}{x_{1+} + x_{2+}}$$

La distància de Bray-Curtis va ser descrita inicialment per Odum (1950), qui la va anomenar diferència de percentatge. És el complement del coeficient de similaritat atribuït a Steinhaus per Motyka (1947).

$$d_{BC}(\mathbf{x}_1, \mathbf{x}_2) = 1 - s_{Steinhaus}(\mathbf{x}_1, \mathbf{x}_2) = 1 - \frac{2 \cdot \sum_{j=1}^{P} \min(x_{1j}, x_{2j})}{x_{1+} + x_{2+}}$$

Hom anomena sovint 'coeficient de Czekanowski' a la similaritat de Steinhaus (erròniament segons Legendre & Legendre 1998). El coeficient de similaritat de Steinhaus/Czekanowski és, a

la vegada, una generalització al cas quantitatiu de l'índex de Sørensen. Un coeficient de similaritat molt relacionat és el coeficient de Kulczynski (1927):

$$s_{Kulc}(\mathbf{x}_1,\mathbf{x}_2)=\frac{1}{2}\cdot\left[\frac{\sum_{j=1}^{P}\min(x_{1j},x_{2j})}{x_{1+}}+\frac{\sum_{j=1}^{P}\min(x_{1j},x_{2j})}{x_{2+}}\right].$$

La distància de Bray-Curtis i el complement de l'índex de similaritat de Kulczynski són mesures de dissimilaritat semi-mètriques acotades en l'interval $[0,1]$. En el cas de Bray-Curtis normalment esdevé mètrica i euclidiana en prendre l'arrel quadrada (el mateix s'aplica al cas qualitatiu de l'índex de Sørensen, veure Gower & Legendre 1986 o Legendre & Legendre 1998). D'altra banda, la metricitat de Bray-Curtis i el complement de l'índex de Kulczynski es pot obtenir també si els inventaris tenen la mateixa abundància total. És important remarcar, en aquest cas, la equivalència entre les quatre mesures presentades. És fàcil provar que si $x_{1+}=x_{2+}$, llavors:

$$0.5\cdot d_{ManRel}(\mathbf{x}_1,\mathbf{x}_2)=d_{Whittaker}(\mathbf{x}_1,\mathbf{x}_2)=d_{BC}(\mathbf{x}_1,\mathbf{x}_2)=1-d_{Steinhaus}(\mathbf{x}_1,\mathbf{x}_2)=1-d_{Kulc}(\mathbf{x}_1,\mathbf{x}_2)$$

Els australians Lance & Williams (1967) proposen altres variants de la mètrica de Manhattan. La més coneguda és la mètrica de Canberra:

$$d_{Canberra}(\mathbf{x}_1,\mathbf{x}_2)=\sum_{j=1}^{P}\frac{\left|x_{1j}-x_{2j}\right|}{(x_{1j}+x_{2j})}$$

Aplicada a dades de composició específica, aquesta mètrica necessita excloure els dobles zeros per evitar la indeterminació en el denominador. D'altra banda, la mètrica de Canberra és una mesura de dissimilaritat simètrica. Per tal de tornar-la una mesura asimètrica es recomana emprar la forma d'Adkins (Lance & Willams 1967):

$$d_{Canb-Ad}(\mathbf{x}_1,\mathbf{x}_2)=\frac{1}{(P-Z_{12})}\cdot\sum_{j=1}^{P}\frac{\left|x_{1j}-x_{2j}\right|}{(x_{1j}+x_{2j})}$$

on Z_{12} és el nombre de dobles absències. La forma d'Adkins de la mètrica de Canberra és la distància complementària d'una altra generalització de l'índex de Jaccard al cas quantitatiu. Efectivament, si $\mathbf{X}$ és una matriu de valors binaris (0 o 1):

$$d_{Canb-Ad}(\mathbf{x}_1,\mathbf{x}_2)=\frac{1}{a+b+c}\cdot\sum_{j=1}^{P}\frac{\left|x_{1j}-x_{2j}\right|}{(x_{1j}+x_{2j})}=\frac{b+c}{a+b+c}=1-\frac{a}{a+b+c}$$

La mètrica de Canberra i la distància de Bray-Curtis són equivalents en el cas univariant. La taula 3.2.5 exemplifica el comportament de les dues mesures de dissimilaritat en el cas d'una sola variable: Si la diferència entre abundàncies augmenta, però és manté la proporció entre

aquesta i la suma de valors, la dissimilaritat es manté constant. Per exemple, si doblem els valors doblem la diferència sobre una abundància total doble, i obtenim com a resultat una distància igual (cas B). Són, per tant, mesures invariants a canvis d'escala. En canvi, si augmentem les abundàncies però no la diferència (cas C) la distància es redueix. Quan l'espècie només existeix en un dels inventaris la distància és sempre 1, independentment del valor d'abundància (cas D). La taula 3.2.5 ens indica també que els valors de distància obtinguts sobre dades quantitatives respecte dades qualitatives seran sempre iguals o menors, ja que la diferència entre l'absència i la presència sempre és 1 mentre que la comparació entre dues abundàncies no nul·les (> 0) resulta en una diferència menor.

Cas	sp. 1	sp. 2	dif.	suma	BC o Cnb
A	1.5	2.5	1	4	0.25
B	3	5	2	8	0.25
C	4.5	5.5	1	10	0.1
D	0	a	a	a	1
E	a	a	0	2a	0

Taula 3.2.5: Exemples univariants per a les dissimilaritats de Bray-Curtis i Canberra.

La diferència entre la mètrica de Canberra i la distància de Bray-Curtis rau en que la primera és senzillament la suma de casos univariants, mentre que en la segona la normalització per a un tàxon concret es veu afectada per la abundància dels altres tàxons, pel que la mesura és més complexa. Aquest és el cas també de la mètrica de Manhattan relativitzada i de l'índex de Kulczynski.

Un problema de la mètrica de Canberra és que, en les comparacions entre l'absència i qualsevol altre valor la diferència és sempre màxima, independentment de si l'altre valor és petit o gran (Campbell 1978). Per a posar-hi remei, Noest & van der Maarel (1989) proposen una nova mesura: la dissimilaritat d'Uppsala (*Uppsala Dissimilarity*, *UD*):

$$d_{UD}(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{(P - Z_{12})} \cdot \left[\sum_{j=1}^{P} \frac{1}{2} \cdot \left(\frac{\left| x_{1j} - x_{2j} \right|}{(x_{1j} + x_{2j})} + \frac{\left| x_{1j} - x_{2j} \right|}{(x_{\max} - x_{\min})} \right) \right]$$

on x_{max} i x_{min} són els valors màxim i mínim de la matriu $\mathbf{X}$ (p.e. 0 i 9 amb la transformació combinada). La dissimilaritat d'Uppsala s'obté com a mitjana aritmètica de la mètrica de Canberra (en la forma d'Adkins) i la mètrica de Gower (1971). Aquesta darrera es basa en la divisió entre la diferència en valor absolut i rang de la variable:

$$d_{Gow}(\mathbf{x}_1, \mathbf{x}_2) = \sum_{j=1}^{P} \left(\frac{\left| x_{1j} - x_{2j} \right|}{(x_{\max} - x_{\min})} \right).$$

3.2.2.6 Distàncies euclidianes obtingudes per transformació de les dades

En aquest apartat presentem una sèrie de distàncies que poden ésser obtingudes mitjançant una transformació de les dades prèvia al càlcul de la distància Euclidiana. Totes elles admeten una representació complerta en un espai euclidià, i són, per tant, mesures de dissimilaritat mètriques i euclidianes. Aquestes distàncies han estat revisades recentment a Legendre & Gallagher (2001) per al seu ús en mètodes d'ordenació basats en la norma Euclidiana (l'anàlisi de components principals i l'anàlisi de les redundàncies). Comentem a continuació algunes de les transformacions- distàncies, l'efecte que tenen sobre les dades i les relacions entre elles.

Distància de la corda

Si els nostres mètodes d'anàlisi no poden evitar l'ús de la norma Euclidiana, podem eliminar fàcilment l'efecte de les dobles absències amb una simple estandardització per la longitud del vector fila.

Transformació de la corda : $y_{ij} = \dfrac{x_{ij}}{\sqrt{\sum_{j=1}^{P} x_{ij}^2}}$

És molt habitual en vegetació emprar aquesta transformació vectorial (per exemple Escudero & Pajarón 1994, Olano *et al*. 1998b). La distància obtinguda, amb la transformació anterior, respecte les dades originals és la distància de la corda (Orlóci 1967):

$$d_{chord}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{P} \left(\frac{x_{1j}}{\sqrt{\sum_{j=1}^{P} x_{1j}^2}} - \frac{x_{2j}}{\sqrt{\sum_{j=1}^{P} x_{2j}^2}} \right)^2} = \sqrt{2\left(1 - \frac{\sum_{j=1}^{P} x_{1j} x_{2j}}{\sqrt{\sum_{j=1}^{P} x_{1j}^2 \sum_{j=1}^{P} x_{2j}^2}}\right)}$$

La distància de la corda entre dos objectes és equivalent a la longitud de la corda unint dos punts en un segment d'una hiper-esfera de radi 1. La distància màxima entre dos inventaris sense espècies en comú és l'arrel de $2 \approx 1.414$. La segona expressió de la distància de la corda ens revela que aquesta distància està relacionada amb el cosinus de l'angle entre els vectors de dades, normalitzats o no:

$$d_{chord}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{2(1 - \cos\theta)}$$

La distància de la corda té un comportament semblant al complementari de *similarity ratio*. Aquesta relació és més clara si comparem la segona manera d'expressar la distància de la corda amb l'equació de la mesura de similaritat.

Perfil d'espècies

La distància perfil d'espècies no és una mesura de proximitat massa coneguda ni utilitzada. No obstant, ens és útil comentar-la aquí per la simplicitat de la transformació associada:

Transformació perfil d'espècies: $y_{ij} = \frac{x_{ij}}{x_{i+}}$

Distància perfil d'espècies: $d_{Sp\,\mathrm{Pr}\,of}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{P}\left[\frac{x_{1j}}{x_{1+}} - \frac{x_{2j}}{x_{2+}}\right]^2}$

La transformació perfil d'espècies converteix els valors d'abundància dels tàxons en proporcions de la abundància total de l'inventari. La divisió dels valors per la suma d'abundàncies de la fila fa que totes les files tinguin una abundància total igual a 1. Els inventaris amb poques espècies se'n veuen beneficiats. A més, les espècies amb una cobertura-abundància alta estaran ponderades en excés en aquests inventaris. Per aquest motiu, la distància perfil d'espècies fou considerada no desitjable per van der Maarel (1979). A diferència de la distància perfil d'espècies, a la distància de la corda la divisió es fa per la longitud del vector, i els inventaris encara conserven diferències d'abundància total després de la transformació.

És important destacar que, mentre que per una norma L1 la transformació que fa que els vectors tinguin norma 1 és el perfil d'espècies, per a la norma L2 la transformació que genera vectors de norma 1 és la transformació de la corda.

Distància de Hellinger

La transformació de Hellinger (Rao 1995) és l'arrel quadrada de la proporció de l'espècie en el conjunt de valors.

Transformació Hellinger: $y_{ij} = \sqrt{\frac{x_{ij}}{x_{i+}}}$

Distància de Hellinger (Rao 1995): $d_{Hell}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{P}\left[\sqrt{\frac{x_{1j}}{x_{1+}}} - \sqrt{\frac{x_{2j}}{x_{2+}}}\right]^2}$

A diferència de la transformació perfil d'espècies, la presència de l'arrel quadrada fa que les proporcions més grans augmentin menys que les petites, pel que les diferències entre espècies abundants i espècies rares es fa més curta. Aquesta transformació és més desitjable que la dels perfils d'espècies. La distància de Hellinger és una de les que es recomana per a ordenacions de dades d'abundàncies de espècies (Rao 1995).

Noteu que si hom eleva al quadrat els elements d'**X** abans d'aplicar la transformació de Hellinger:

$$y_{ij} = \sqrt{\frac{x_{ij}^2}{\sum_{j=1}^{P} x_{ij}^2}} = \frac{x_{ij}}{\sqrt{\sum_{j=1}^{P} x_{ij}^2}}$$

la transformació resultant és equivalent a la estandardització pel vector fila i, per tant, equivalent a la transformació de corda. Dit d'altra manera, hom pot afirmar que la distància de Hellinger és equivalent a la distància de la corda aplicada sobre l'arrel quadrada dels valors originals.

$$d_{Hell}(\mathbf{x}_1^w, \mathbf{x}_2^w) = d_{Chord}(\mathbf{x}_1^{w/2}, \mathbf{x}_2^{w/2})$$

Donat que en l'anàlisi de dades de composició específica hom sovint ha de escollir la transformació escalar les dades que proporcioni una relació entre espècies abundants i espècies poc abundants adequada, les distàncies de la corda i Hellinger es mostren equivalents sota aquest punt de vista. Aquesta relació passà desaparcebuda, aparentment, en l'estudi de Legendre & Gallagher (2001).

Distància χ^2

La transformació que dóna lloc a la distancia χ^2 és la més complexa de les que tractem ja que implica una transformació doble, per files i per columnes:

Transformació χ^2: $y_{ij} = \sqrt{x_{++}} \cdot \frac{x_{ij}}{x_{i+} \cdot \sqrt{x_{+j}}}$

Distància χ^2: $d_{\chi^2}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{P} \frac{1}{x_{+j} / x_{++}} \left(\frac{x_{1j}}{x_{1+}} - \frac{x_{2j}}{x_{2+}} \right)^2}$

La distància χ^2 és important en ecologia numèrica per que és la dissimilaritat preservada en l'anàlisi factorial de correspondències (*correspondence analysis*). No obstant, les espècies rares i els inventaris amb poques espècies es veuen sobrevalorats. Ja vam veure al cap. 2.1 que el nombre de tàxons rars depèn en gran mesura del mostratge. Per tant, la nostra opinió és que χ^2 no resulta una bona mesura de proximitat per a un *clustering* amb bases geomètriques. Faith *et al*. (1987) en un estudi comparatiu conclogueren que aquesta és una de les pitjors distàncies per a dades de composició de comunitats. Legendre & Gallagher (2001) proposen la distància de la corda o la distància de Hellinger com a alternatives a χ^2 en el context de la ordenació i regressió basada en distàncies (Legendre & Anderson 1999).

3.2.2.7 Estudis de comparació de transformacions/mesures de proximitat

Els estudis que han tractat de diferenciar teòricament o pràctica les transformacions de dades i mesures de proximitat aplicades a l'anàlisi multivariant són força nombrosos. A continuació en citem aquells treballs que hem consultat i considerem més rellevants.

Campbell (1978) estudià diversos coeficients de proximitat, tant per les seves propietats com en aplicació per a mètodes de classificació. El mètode de comparació que emprà fou a través de establir dos senzills casos de comparació formats cadascú per dos inventaris amb poques espècies, on en el segon cas modificava algun aspecte del primer. Mitjançant aquest mètode, Campbell conclogué que la distància euclidiana i els índexs de similaritat de Gleason, Ellenberg i Spatz tenien propietats poc desitjables. D'altra banda, Campbell testà sis mesures d'interès per a la classificació, comparant el dendrograma de *UPGMA* amb la classificació tradicional. Entre les sis mesures estudiades, Campbell en recomanà el que ell anomenà "índex de Czekanowski relativitzat", que en la nostra notació equival a la mètrica de Manhattan relativitzada o l'índex de Whittaker. La distància de Canberra fou desaconsellada per Campbell perquè l'estandardització en la comparació de cada tàxon provoca que les espècies amb baixa cobertura tinguin un pes excessiu.

Hajdu (1981) realitzà un extens estudi de comparació de mesures de proximitat i transformacions aplicades a la fitosociologia a través de series de casos (*OCCAS*, citats a la pàg. 174), que estenien el mètode de comparació de Campbell. Hajdu avaluà la idoneïtat de les mesures de proximitat en base a la linealitat i el poder de resolució que presentaven respecte les diferents OCCAS. Els resultats indicaren que *similarity ratio*, el coeficient de Czekanowski (complement de la distància de Bray-Curtis) i el de Kulczynski són les mesures que presentaven un millor compromís entre linealitat i poder de resolució. Hajdu (1981) destina un apartat a la comparació de transformacions de l'escala de Braun-Blanquet en el que proposa la transformació combinada de van der Maarel (1979) com la més avantatjosa per al *clustering*.

Bloom (1981) construí una nova sèrie de comparació basada en la superposició entre dues corbes normals estandarditzades. Bloom comparà les dissimilaritats de Bray-Curtis i Canberra, entre d'altres i conclou que la primera és la que es correspon més linealment a la superposició teòrica esperada. Per construcció, la sèrie de comparació de Bloom s'assembla, en el cas qualitatiu a *OCCAS-PA2*. Com que Bray-Curtis és el complement quantitatiu de l'índex de Sørensen, és lògic que apuntés a la distància de Bray-Curtis com a mesura més adient.

En un treball més teòric, Faith (1984) estudià la sensibilitat de les mesures d'associació i proposà dues classes conceptuals de mesures: aquelles 'sensibles a la separació' i aquelles 'sensibles a un valor mínim'. Alhora, Faith proposà una nova mesura intermèdia en aquests dos aspectes.

Dale (1988a) examinà diferents coeficients de semblança com a variacions de mesures de comparació de cadenes (*Levenshtein distances*) per tal de proporcionar una base racional de comparació. A més, Dale comparà les classificacions obtingudes per mètodes jeràrquics sobre diversos coeficients. Posteriorment, Dale (1989) estudià la possibilitat de definir mesures de dissimilaritat basades en escales ordinals, enlloc de les més correntment emprades escales mètriques.

Faith, *et al.* (1987) estudiaren la 'robustesa' de diferents mesures de proximitat, conjuntament amb diferents estandarditzacions prèvies de les dades, mitjançant la generació de gradients ecològics simulats amb el programa *COMPAS* (Minchin 1987). La longitud dels gradients simulats era variable i la resposta de les espècies als gradients ecològics eren simulada amb funcions beta de diversa forma. Els autors avaluaren la 'robustesa' calculant la correlació (lineal i per rangs) entre les distàncies ecològiques que proporcionava el model de simulació i les dissimilaritats sorgides de l'aplicació de cada mesura de proximitat i estandardització. Faith *et al*. conclogueren que les mesures de proximitat més robustes són Bray-Curtis, la mètrica de Manhattan relativitzada i l'index de similaritat de Kulczynski (1927). La vàlua de l'estudi de Faith *et al*. depèn, com en el cas de la construcció de sèries de comparació, de la validesa del model de simulació de gradients, o sigui, de les assumpcions teòriques de la resposta dels tàxons als gradients ecològics (vegeu 3.1.5.9). A més, està orientat principalment a avaluar la utilitat de les mesures de proximitat a l'anàlisi de proximitats (*scaling*).

Més recentment, De'ath (1999) proposa una mesura dissimilaritat estesa, per al cas de gradients amb una diversitat beta alta. De nou, el treball de De'ath està dirigit sobretot a les tècniques d'ordenació, ja que el mètode consisteix en transformar les distàncies grans per eliminar l'efecte d'arc en els resultats d'aquest tipus d'anàlisi multivariant. Donat que per a l'anàlisi de clúster són més importants les distàncies curtes que les llargues, no tractarem aquest tipus d'extensions aquí.

En el treball ja citat de Legendre and Gallagher (2001) aquests autors comparen amb un petit set de dades sintètiques la resposta lineal de diverses distàncies, recomanant les distàncies de Hellinger, Bray-Curtis i la corda.

Destinem la propera secció 3.2.3 a introduïr la classificació basada en matrius de distàncies entre objectes. Les seves conclusions ens permeten abordar, a la secció 3.2.4, la comparació de mesures de proximitat i transformacions de les dades en la seva aplicació al *clustering* de comunitats vegetals. Tanmateix, el lector que ho desitji pot saltar-se aquesta secció, més teòrica, i iniciar directament la lectura de l'estudi aplicat.

3.2.3 La classificació basada en matrius de distàncies entre objectes

3.2.3.1 Variabilitat geomètrica i proximitat d'un individu a una població

Sigui **X** un vector aleatori P-dimensional definit sobre un espai de probabilitat (Π, A, ***P***) que pren valors $S \subset \Re^P$ amb funció de densitat de probabilitat f respecte a una mesura adequada λ. Considerem $d(\cdot,\cdot)$ una funció de dissimilaritat definida sobre les parelles d'elements de Π, tal que llur quadrat sigui integrable en S. La variabilitat geomètrica d'**X** respecte a $d(\cdot,\cdot)$ (Cuadras & Fortiana 1995) ve definida per:

$$V_d(\mathbf{X}) = \frac{1}{2} E\left[d^2(\mathbf{X}_1, \mathbf{X}_2) \,\middle|\, \mathbf{X}_1, \mathbf{X}_2 \in S\right] = \frac{1}{2} \int_{S\times S} d^2(\mathbf{x}_1, \mathbf{x}_2) f(\mathbf{x}_1) f(\mathbf{x}_2)\, \lambda(d\,\mathbf{x}_1)\, \lambda(d\,\mathbf{x}_2)$$

Donat $\mathbf{x}_o \in \Re^P$, definim la proximitat d'$\mathbf{x}_o$ a la població Π respecte a $d(\cdot,\cdot)$ com a (Cuadras *et al*. 1997):

$$\phi_d^2(\mathbf{x}_o, \Pi) = \int_S d^2(\mathbf{x}_o, \mathbf{x}) f(\mathbf{x})\, \lambda(d\,\mathbf{x}) \; - V_d(\mathbf{X})$$

Si existeix una representació de $d(\cdot,\cdot)$, és a dir, si existeix una funció $\psi : R^P \to L$ (on $L, \langle\cdot,\cdot\rangle$ simbolitza un espai euclidià o Hilbert amb producte escalar $\langle\cdot,\cdot\rangle$, de manera que $\| \mathbf{u} \|^2 = \langle \mathbf{u}, \mathbf{u} \rangle$ és la norma natural per a tot $\mathbf{u} \in L$) llavors, suposant que $E(\psi(\mathbf{X}))$ i $E(\| \psi(\mathbf{X}) \|^2)$ siguin finites:

a) $V_d(\mathbf{X}) = E(\| \psi(\mathbf{X}) - E(\psi(\mathbf{X})) \|^2)$

b) $\phi_d^2(\mathbf{x}_o) = \left\| \psi((\mathbf{x}_o) - E(\psi(\mathbf{X})) \right\|^2$

En el cas mostral, donades les mostres $\mathbf{x}(k)_1, \ldots, \mathbf{x}(k)_{n_k}$ de Π_k, $k = 1, \ldots, K$, la distància d'un element $\mathbf{x}_o$ al grup Π_k serà (Cuadras *et al*. 1997):

$$\hat{\phi}_d^2(\mathbf{x}_o, \Pi_k) = \frac{1}{n_k} \sum_{h=1}^{n_k} d^2(\mathbf{x}_o, \mathbf{x}(k)_h) - \hat{V}_d(k)$$

$$\hat{V}_d(\Pi_k) = \frac{1}{2\, n_k^2} \sum_{h,l} d^2(\mathbf{x}(k)_h, \mathbf{x}(k)_l)$$

Les expressions de la variabilitat geomètrica i proximitat es poden generalitzar al cas en que el conjunt d'objectes sigui difús. Donada una partició difusa $\mathbf{U}_{\mathrm{NxK}}$ i un exponent de *fuzziness* (m):

$$\hat{V}_{fd}(\Pi_k) = \frac{1}{2\left(\sum_{i=0}^{N} u_{i(k)}^m \right)^2} \sum_{h,l} u_{h(k)}^m \cdot u_{l(k)}^m \cdot d^2(\mathbf{x}(k)_h, \mathbf{x}(k)_l)$$

$$\hat{\phi}_{fd}^2(\mathbf{x}_o, \Pi_k) = \frac{1}{\sum_{h=1}^{N} u_{h(k)}^m} \sum_{h=1}^{N} u_{h(k)}^m d^2(\mathbf{x}_o, \mathbf{x}(k)_h) - \hat{V}_{fd}(k)$$

3.2.3.2 Els algorismes *K-means* i *Fuzzy C-means* basats en matrius de dissimilaritats (DB)

El concepte de variabilitat geomètrica ha permès desenvolupar un anàlisi discriminant basat en distàncies (Cuadras *et al*. 1997). Proposem aquí, una generalització anàloga per als algorismes d'anàlisi de clústers *K-means* (MacQueen 1967) i *Fuzzy C-means* (Bezdek 1981). La generalització que proposem consisteix en considerar les relacions:

1) Entre la dispersió d'un objecte respecte el centroide i la proximitat d'un individu a una població: $e^2_{i(k)} = \hat{\phi}^2_d(\mathbf{x}_i, \Pi_k)$.

2) Entre la dispersió d'un clúster i la variabilitat geomètrica del mateix: $E^2_{(k)} = n_k \hat{V}_d(\Pi_k)$.

Acceptant aquestes relacions, hom pot calcular $e^2_{i(k)}$ i $E^2_{(k)}$ en base a matrius de dissimilaritat. En el cas *crisp* tenim (emprant la notació de l'apartat 3.1.3.2):

$$E^2_{(k)} = \frac{1}{2n_k} \sum_{i,j=1}^{N} I[\omega_i \subset \Omega_k] I[\omega_j \subset \Omega_k] d^2_{ij}$$

$$e^2_{i(k)} = \frac{1}{n_k} \sum_{h=1}^{N} I[\omega_h \subset \Omega_k] d^2_{ih} - \frac{1}{2n_k^2} \sum_{h,l=1}^{N} I[\omega_h \subset \Omega_k] I[\omega_l \subset \Omega_k] d^2_{hl}$$

Aquest darrera equació substitueix les equacions originals del càlcul de les coordenades del centroide i el càlcul de les distàncies dels objectes al centroides (vegeu l'apartat 3.1.3.2 en referència als passos originals de l'algorisme *K-means*). Cal notar que l'aproximació basada en distancies (*distance based K-means* o *DB-KM*) és computacionalment més costosa, a no ser que el nombre de variables originals sigui molt gran en relació al nombre d'objectes. En el cas difús, que anomenarem *DB-FCM*, hom pot determinar expressions equivalents. La dispersió de Ω_k és:

$$J^2_{(k),m} = \frac{\sum_{i,j=1}^{N} u^m_{i(k)} u^m_{j(k)} d^2_{ij}}{2 \sum_{i=1}^{N} u^m_{i(k)}}$$

Finalment, la distància $e^2_{i(k)}$ de l'objecte ω_i al centroide del clúster Ω_k és:

$$e^2_{i(k)} = \frac{1}{\sum_{h=1}^{N} u^m_{h(k)}} \sum_{h=1}^{N} u^m_{h(k)} d^2_{ih} - \frac{1}{2\left(\sum_{h=1}^{N} u^m_{h(k)}\right)^2} \sum_{h,l=1}^{N} u^m_{h(k)} u^m_{l(k)} d^2_{hl}$$

L'aproximació al *clustering* partitiu basat en matrius de distàncies ja fou explorada per Späth (1980) per a *KM*, i Hathaway *et al*. (1989) i Bezdek *et al*. (1991) per a *FCM*. Hathaway *et al*. (1989) l'anomenaren *Relational FCM*. Hom pot consultar també la revisió feta del tema per Hathaway *et al*. (1996).

3.2.3.3 Aproximació *DB* i euclidianitat.

És coneguda la equivalència entre executar algorismes partitius del tipus *KM/FCM* a partir de matrius rectangulars (objectes-variables) o mitjançant l'aproximació basada en distàncies (*distance based*, *DB*) amb la distància Euclidiana. D'altra banda, per a algunes distàncies euclidianes (és a dir, que tenen representació en un espai euclidià) és també equivalent l'aproximació *DB* i la dels algorismes basats en matrius rectangulars, amb les variables originals prèviament transformades adequadament (vegeu secció 3.2.2.6). Ara bé, l'aproximació *DB* és més general i permet tractar també altres distàncies euclidianes que no admeten l'enfoc clàssic (per exemple, la distància valor absolut) i distàncies no euclidianes. Però, què succeeix a la pràctica si la distància no és euclidiana? L'enfoc *DB* en la seva versió mostral és equivalent a realitzar un *classical MDS* (Gower 1966) i calcular les distàncies de cada objecte al centroid del grup o clúster. En el cas de distàncies que no admetin una representació euclidiana és fàcil demostrar que el *classical MDS* produeix alguns valors propis negatius i els vectors propis associats són aleshores components imaginàries, de manera que les distàncies entre dos punts pot definir-se com a:

$$d_{lk}^2 = \|\mathbf{a}_l - \mathbf{a}_k\|^2 + \|i\mathbf{b}_l - i\mathbf{b}_k\|^2 = \|\mathbf{a}_l - \mathbf{a}_k\|^2 + (i^2)\|\mathbf{b}_l - \mathbf{b}_k\|^2 = \|\mathbf{a}_l - \mathbf{a}_k\|^2 - \|\mathbf{b}_l - \mathbf{b}_k\|^2$$

i, per tant, $d_{lk}^2 = d_{lk(real)}^2 - d_{lk(imag)}^2$. És evident que les distàncies entre dos objectes observats no poden ser negatives (no serien distàncies), però sí les "distàncies" entre un objecte i el centroid del grup que calcula l'enfoc *DB*.

Hom pot introduïr elements de no-euclidianitat en una matriu de distàncies euclidiana modificant arbitràriament els valors de distancia respecte a un punt qualsevol. A la figura 3.2.3 s'escurcen les distàncies a un objecte central, fent que el clúster sigui més compacte. La no-euclidianitat de la figura 3.2.3 podria haver estat també en el sentit contrari: incrementant les distàncies a l'objecte central.

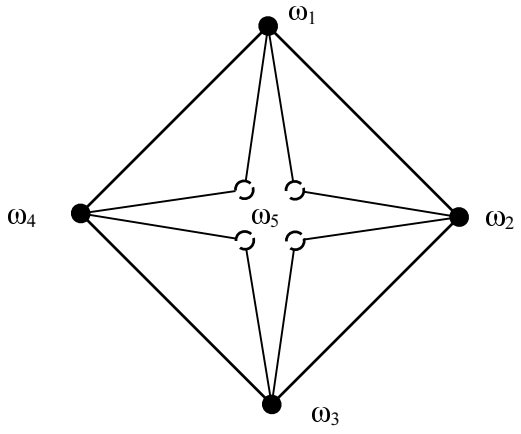


Figura 3.2.3: Exemple de no-euclidianitat entre 5 objectes. Les distàncies dels quatre punts exteriors al punt central (ω_5) han estat escurçades, de manera que es compleix la desigualtat triangular però no la propietat euclidiana de l'espai.

Tal i com hem esmentat, l'enfoc *DB* pot fins i tot portar a "distàncies" al centroide negatives (si bé a la pràctica aquest és un fet poc habitual). Aquesta "incomoditat" ha portat a proposar correccions per a assolir l'euclidianitat, com, per exemple, la proposta de Hathaway & Bezdek (1994) que donà lloc a l'algorisme anomenat *Non Euclidean Relational FCM* (*NERF-CM*). Ara bé, l'objectiu de *DB-KM* i *DB-FCM* no és la representació, de manera que la no euclidianitat no és gaire problema per a la seva aplicació (aspectes formals apart). Per a *DB-KM* un valor negatiu de "distància" quadràtica li correspon sempre una pertinença 1. Per a *DB-FCM* l'equació de pertinença produirà pertinences lleugerament majors a 1 i altres lleugerament menors que zero.

A continuació estudiem un petit conjunt de dades, extret de Hathaway & Bezdek (1994). L'objectiu d'aquest estudi és mostrar l'ús de l'aproximació *DB* en espais de relacions no euclidians i comparar l'efecte de les diferents correccions d'euclidianitat sobre els resultats del *clustering*. A les primeres dues columnes taula 3.2.6 es mostren les coordenades d'onze punts, els quals formen una estructura amb dos clústers romboïdals amb un punt intermedi (obj-6). Aquesta senzilla estructura de dos clústers es pot veure representada a la figura 3.2.4. A les columnes següents de la taula 3.2.6 es mostren els valors de pertinença del primer clúster obtinguts mitjançant *FCM* (m=2.0) a partir de diferents casos. En primer lloc apareixen les pertinences obtingudes amb la distància Euclidiana, *FCM* (L2). La matriu de pertinences difusa apareix representada a la figura 3.2.4 amb intensitats de colors i mides de cercle.

	x1	x2	FCM (L2)	DB-FCM (L1)	NERF-CM (L1)	FCM (L1-MDS)	FCM (L1-MDS*)	DB-FCM SQRT(L1)
Obj-1	-5.00	0.00	0.93	0.92	0.90	0.90	0.75	0.80
Obj-2	-3.00	2.00	0.90	0.91	0.89	0.82	0.73	0.79
Obj-3	-3.00	0.00	1.00	1.03	1.00	0.99	0.77	0.93
Obj-4	-3.00	-2.00	0.90	0.91	0.89	0.82	0.74	0.79
Obj-5	-1.00	0.00	0.81	0.79	0.76	0.80	0.61	0.69
Obj-6	0.00	0.00	0.50	0.50	0.50	0.50	0.50	0.50
Obj-7	1.00	0.00	0.19	0.21	0.24	0.20	0.38	0.31
Obj-8	3.00	2.00	0.10	0.09	0.11	0.18	0.27	0.21
Obj-9	3.00	0.00	0.00	-0.03	0.00	0.01	0.23	0.07
Obj-10	3.00	-2.00	0.10	0.09	0.11	0.18	0.27	0.21
Obj-11	5.00	0.00	0.07	0.08	0.10	0.10	0.25	0.20
	Coef. de Dunn		**0.625**	**0.636**	**0.569**	**0.532**	**0.187**	**0.349**

Taula 3.2.6: Exemple d'aplicació de *FCM* en situacions de no euclidianitat (veure text). Les primeres dues columnes mostren les coordenades dels punts. Les restants columnes mostren els valors de pertinença finals a un dels dos clústers en diferents modalitats d'execució de l'algorisme (vegeu text).

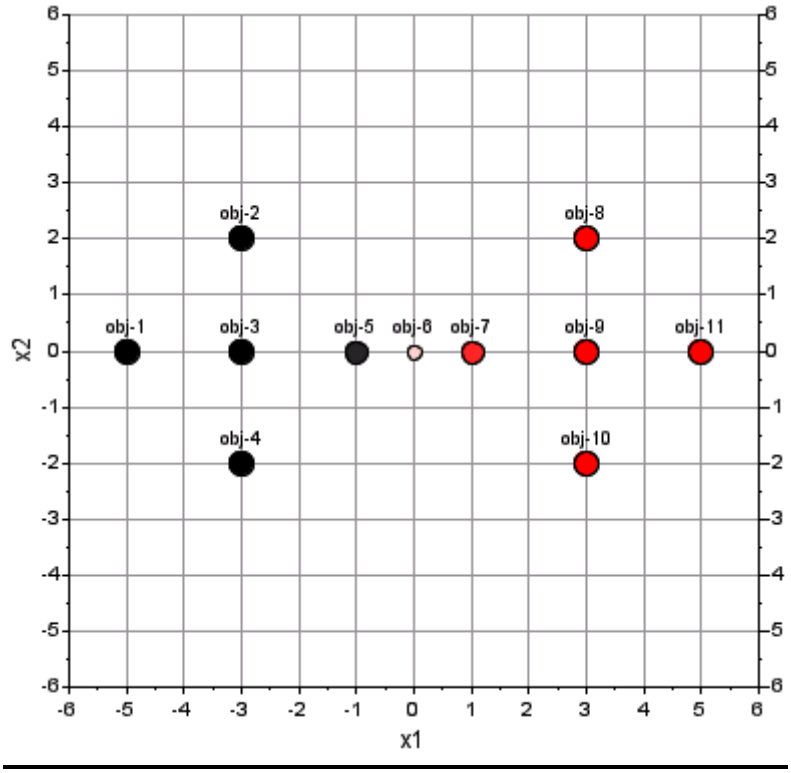


Figura 3.2.4: Exemple d'aplicació de *FCM* en situacions de no euclidianitat (veure text). Diagrama de dispersió en el que es representen els 11 punts de la taula 3.2.6. S'indiquen les pertinences obtingudes amb *FCM* (m=2.0) a través de la mida i color dels cercles.

Què succeeix si hom prefereix estudiar les relacions entre els 11 objectes mitjançant la norma L1 (mètrica de Manhattan)? La columna DB-FCM (L1) de la taula 3.2.6 mostra les pertinences dels objectes en executar el mètode *DB-FCM* sobre la matriu de dissimilaritats L1. Els dos objectes obj-3 i obj-9, que coincideixen en l'espai Euclidià amb els centroides dels clústers, en executar *DB-FCM* sobre L1 s'obtenen distancies quadràtiques negatives als centroids! Això es tradueix en valors de pertinença lleugerament superiors a 1 o inferiors a zero. No obstant, la partició resultant és correcte i el coeficient de partició (o coeficient de Dunn) normalitzat indica una "borrositat" de la partició semblant al cas Euclidià, tot i que els valors no són comparables. Aquesta aproximació tracta euclidianament un espai que no es euclidià. No obstant, no deforma l'espai de dades. Les quatre aproximacions següents introdueixen correccions a l'espai de la norma L1 per tal d'assolir la euclidianitat. El resultat és que la partició que produeix *FCM* és més difusa que en el cas sense corregir.

La columna NERF-CM (L1) mostra els resultats obtinguts amb l'algorisme *NERF-CM* de Hathaway & Bezdek (1994), en que s'introdueix una correcció per a assolir l'euclidianitat. La correcció ($\beta = 3.56$) és la mínima necessària per a evitar distàncies quadràtiques als centroides negatives. La partició resultant és correcte però més difusa que sense la correcció, tal i com fa palès el coeficient de partició.

Una possible manera de realitzar una partició amb un enfoc *DB* consisteix en descartar les components imaginàries del *classical MDS* i utilitzar les coordenades principals reals per al *clustering* amb *KM* o *FCM*. Aquest procés d'anàlisi produeix les pertinences mostrades a la

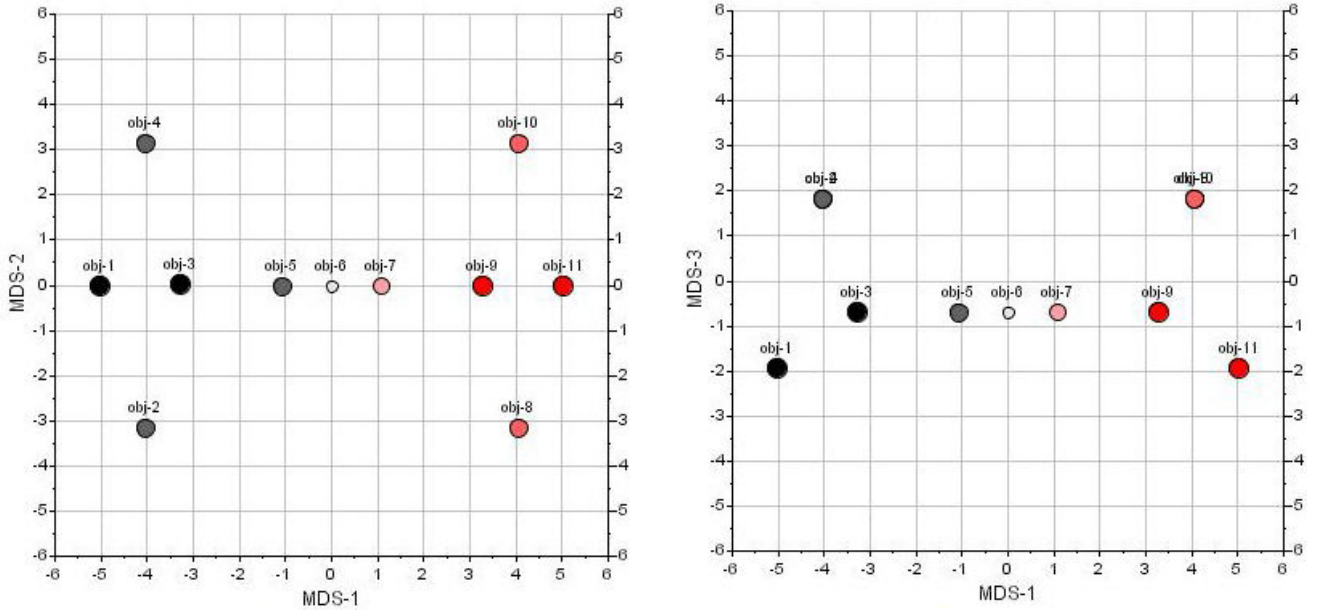
columna FCM(L1-MDS). Observeu que en aquesta anàlisi hom perd la informació de distància que aporten les components imaginàries. La partició final és més difusa que NERF-CM (L1) i DB-FCM (L1). Mostrem les primeres components reals amb la partició generada FCM(L1-MDS) a la figura 3.1.5.A.

La següent columna de la taula 3.2.6, FCM(L1-MDS*), mostra el resultat de l'aplicació *FCM* sobre l'espai creat per *MDS* amb la correcció de Lingoes (1971). La correcció de Lingoes busca una solució en dimensió $N-2$, i consisteix a sumar una constant $\beta = -2\lambda$ als valors de distancia al quadrat, on λ és el valor propi negatiu més gran. Les distàncies petites es veuen proporcionalment més afectades que les distàncies grans. A la figura 3.1.5.B apareixen les primeres coordenades principals de l'espai de dissimilaritats corregit amb la partició que s'obté de *FCM*. Noteu que l'estructura és força borrosa i el percentatge de variabilitat mostrat és baix. Com més gran sigui la correcció per euclidianitat, més "borrositat" s'introduirà en la correcció. En aquest cas, la correcció de Lingoes ($\beta = 48$) és més gran que la correcció de *NERF-CM*, i la partició resultant és molt més borrosa.

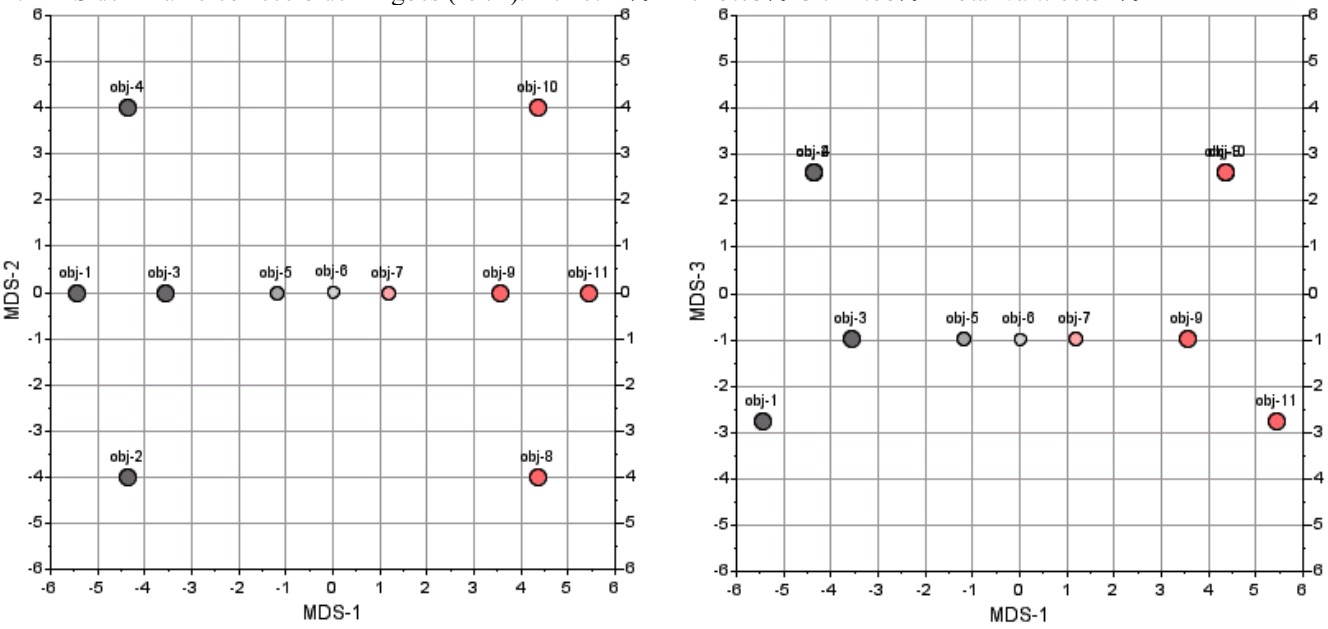
Una altra solució seria, en aquest cas, transformar *a priori* la matriu de dissimilaritats L1 calculant l'arrel quadrada de cada distància. Aquesta transformació equival a calcular la distància valor absolut sobre la matriu rectangular original. Cuadras & Fortiana (1995) demostren que les representacions de la distància valor absolut a *MDS* són sempre euclidianes. A la figura 3.2.5.C mostrem els primers eixos de *MDS* obtinguts sobre la distància valor absolut, juntament amb la corresponent partició de *FCM*. Noteu que, respecte la distància de Manhattan (3.2.5.A), en calcular l'arrel quadrada les distàncies grans disminueixen més que les petites i, de resultes, l'estructura es corba en el tercer eix. Podani & Miklos (2002) estudien l'efecte que té afegir un exponent c a una distància (d^c). Podani & Miklos observen que els valors de c menors que 1 (per exemple c=0.5) fan augmentar la euclidianitat de les distàncies però, en disminuir més les distàncies grans, l'efecte de "ferradura" de la ordenació es fa més acusat. Per contra, els valors grans de c fan augmentar la linealitat i disminuir l'efecte de "ferradura", però a costa d'augmentar el nombre de valors propis negatius. A la darrera columna de la taula 3.2.6 podem comprovar que la solució de *FCM* en aquest cas és menys difusa que FCM(L1-MDS*) però més que les altres aproximacions. Per tractar-se d'una mesura de proximitat diferent, aquest valor és poc comparable amb els altres. Tanmateix, com que la transformació de l'arrel quadrada es recomana en alguns manuals (p.e. Legendre & Legendre) com a transformació correctora (en molts casos) de la no-euclidianitat, hem cregut interessant posar de manifest l'augment de borrositat que la transformació provoca.

Figura 3.2.5 (pàgina següent): Representacions de l'espai de la mètrica de Manhattan utilitzant diverses alternatives per corregir o evitar la no-euclidianitat de la matriu de dissimilaritats. Els diagrames de dispersió de l'esquerra mostren les primeres dues coordenades principals i els de la dreta mostren la primera i tercera components. S'ha afegit en cada cas la partició difusa obtinguda amb *FCM* (m=2.0).

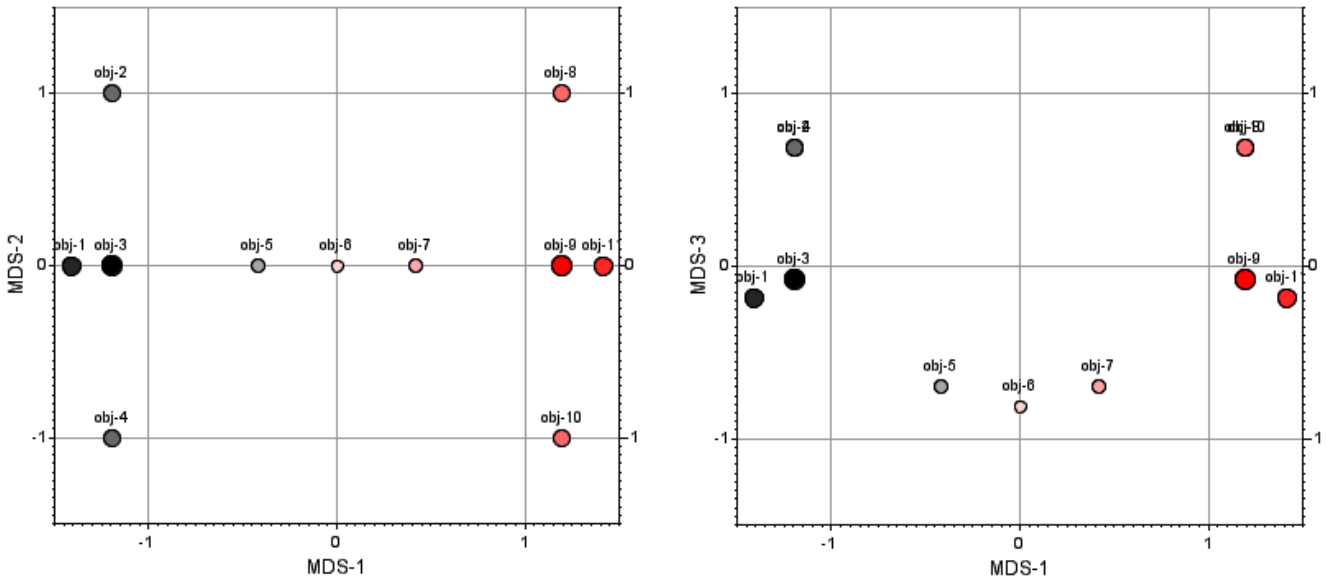
A. *MDS* de L1 seleccionant les components reals. 1ª: 57.57% 2ª: 16.52% 3ª: 9.73% Total var.: 83.82%



B. MDS de L1 amb correcció de Lingoes (1971). 1ª: 40.14% 2ª: 15.73% 3ª: 11.68% Total var.: 67.54%



C. MDS de la distància valor absolut (arrel d'L1). 1ª: 49.97% 2ª: 15.49% 3ª: 13.85% Total var.: 79.32%



En aquest petit exemple hem constatat que, en voler resoldre el problema de la euclidianitat de les dades hom introdueix una alteració de l'estructura geomètrica que es tradueix en un augment de la borrositat de la partició *FCM.* Podem imaginar que, en un context més complex, aquest augment de la "borrositat" condueixi a pèrdua de la capacitat de detecció dels clústers. D'altra banda, a la pràctica les distàncies al centroide negatives en utilitzar l'aproximació *DB* en espais no euclidians són infreqüents. Per els motius exposats, a la propera secció optarem per utilitzar l'aproximació *DB* en tots els casos, siguin euclidians o no.

Finalment voldríem citar altres tècniques d'anàlisi multivariant que poden ésser basades en matrius de distàncies: l'anàlisi de redundàncies (Legendre & Anderson 1999, McArdle & Anderson 2001), l'ANOVA multivariada (Anderson 2001) i *related multidimensional scaling* (RMDS, Cuadras & Fortiana 1998).

3.2.4 L'espai de relacions i els resultats de l'anàlisi de clústers

3.2.4.1 Objectius

En aquesta darrera secció ens proposem esbrinar experimentalment quin espai de relacions entre inventaris és més adient per a la classificació de comunitats vegetals. Entenem aquí per "espai de relacions" l'espai de proximitats entre inventaris sorgit de l'aplicació conjunta d'una transformació escalar de les dades i una mesura de proximitat. El criteri d'avaluació dels espais de relacions que adoptarem aquí es basa, d'una banda, en la "naturalitat" amb que l'espai expressa estructures de grups i, de l'altra, en l'ajust *a posteriori* d'aquests grups amb un criteri extern de comparació. Més concretament, volem abordar els punts següents:

1. Estudiar les relacions entre mesures de proximitat a partir de la correlació entre les matrius de proximitat. Esbrinar si la correlació de les matrius de proximitat implica un comportament semblant a l'hora de classificar els objectes.
2. Estudiar la capacitat de cada mesura de proximitat per a posar de manifest estructures de grups mitjançant criteris interns d'avaluació dels resultats del *clustering* (vegeu secció 3.1.5). Estudiar en quines mesures de proximitat aquesta detecció d'estructures depèn més de la transformació de les dades escollida.
3. Estudiar la semblança entre les particions resultants de cada 'espai de relacions' i comparar-los amb una partició utilitzada com a criteri extern d'avaluació.

3.2.4.2 Dades

Per tal de dur a terme la comparació entre espais de relació, ens hem basat en l'estudi de 3 sets de dades:

A. Les dades de Bowman & Wilson (1984): 41 inventaris i 33 tàxons de la plana al·luvial del riu Adelaide a Austràlia, utilitzades en els estudis de Dale (1988a, 1988b). Hom pot trobar la matriu de dades a la taula 3.1.1 del capítol precedent. L'escala de valors és compresa en l'interval [0, 6]. No utilitzarem aquí la comparació amb la classificació *crisp* que Dale (1988a) proposa sobre aquest set de dades.
B. Els inventaris de *Xerobromion erecti*: 248 inventaris i 548 tàxons de comunitats pratenses xeròfiles montanes. Els valors de l'escala ordinal de Braun-Blanquet han estat transformats a l'escala combinada de van der Maarel (1979). La classificació tradicional estableix 13 sintàxons de base (5 associacions i 8 subassociacions).
C. Els inventaris de *Quercetea ilicis* sense *Quercenion ilicis*: 376 inventaris i 493 tàxons de matollars mediterranis. Els valors de l'escala de Braun-Blanquet han estat transformats a l'escala combinada de van der Maarel (1979). La classificació tradicional estableix 16 sintàxons de base (8 associacions i 8 subassociacions).

Notació	Nom	Fórmula	Propietats
ED	Distància Euclidiana	$d_{ED}(\mathbf{x}_1,\mathbf{x}_2)=\sqrt{\sum_{j=1}^{P}(x_{1j}-x_{2j})^2}$	Mètrica i euclidiana. No acotada.
EDSP	Distància perfil d'especies	$d_{Sp\,\mathrm{Pr}\,of}(\mathbf{x}_1,\mathbf{x}_2)=\sqrt{\sum_{j=1}^{P}\left[\frac{x_{1j}}{x_{1+}}-\frac{x_{2j}}{x_{2+}}\right]^2}$	Mètrica i euclidiana. No acotada.
MAN	Mètrica de Manhattan (o *city block*)	$d_{Man}(\mathbf{x}_1,\mathbf{x}_2)=\sum_{j=1}^{P}\lvert x_{1j}-x_{2j}\rvert$	Mètrica no euclidiana. No acotada.
MANSP	Mètrica de Manhattan sobre perfils d'espècies.	$d_{ManSp\,\mathrm{Pr}\,of}(\mathbf{x}_1,\mathbf{x}_2)=\sum_{j=1}^{P}\left\lvert\frac{x_{1j}}{x_{1+}}-\frac{x_{2j}}{x_{2+}}\right\rvert$	Mètrica no euclidiana. Acotada entre 0 i 2.
CHO	Distància de la corda	$d_{Chord}(\mathbf{x}_1,\mathbf{x}_2)=\sqrt{\sum_{j=1}^{P}\left(\frac{x_{1j}}{\sqrt{\sum_{j=1}^{P}x_{1j}^2}}-\frac{x_{2j}}{\sqrt{\sum_{j=1}^{P}x_{2j}^2}}\right)^2}$	Mètrica i euclidiana. Acotada entre 0 i $\sqrt{2}$.
HELL	Distància de Hellinger	$d_{Hell}(\mathbf{x}_1,\mathbf{x}_2)=\sqrt{\sum_{j=1}^{P}\left[\sqrt{\frac{x_{1j}}{x_{1+}}}-\sqrt{\frac{x_{2j}}{x_{2+}}}\right]^2}$	Mètrica i euclidiana. Acotada entre 0 i $\sqrt{2}$.
SR	Distància complement de *similarity ratio*	$d_{SR}(\mathbf{x}_1,\mathbf{x}_2)=1-\frac{\sum_{j=1}^{P}x_{1j}\cdot x_{2j}}{\sum_{j=1}^{P}x_{1j}^2+\sum_{j=1}^{P}x_{2j}^2-\sum_{j=1}^{P}x_{1j}\cdot x_{2j}}$	Mètrica no euclidiana. Acotada entre 0 i 1.
CNB	Mètrica de Canberra (forma d'Adkins)	$d_{Canb-Ad}(\mathbf{x}_1,\mathbf{x}_2)=\frac{1}{(P-Z_{12})}\cdot\sum_{j=1}^{P}\frac{\lvert x_{1j}-x_{2j}\rvert}{(x_{1j}+x_{2j})}$	Mètrica no euclidiana. Acotada entre 0 i 1.
BC	Distància Bray-Curtis	$d_{BC}(\mathbf{x}_1,\mathbf{x}_2)=\frac{\sum_{j=1}^{P}\lvert x_{1j}-x_{2j}\rvert}{\sum_{j=1}^{P}(x_{1j}+x_{2j})}=\frac{\sum_{j=1}^{P}\lvert x_{1j}-x_{2j}\rvert}{x_{1+}+x_{2+}}$	Semi-mètrica no euclidiana. Acotada entre 0 i 1.
CHI	Distància χ^2	$d_{\chi^2}(\mathbf{x}_1,\mathbf{x}_2)=\sqrt{\sum_{j=1}^{P}\frac{1}{x_{+j}/x_{++}}\left(\frac{x_{1j}}{x_{1+}}-\frac{x_{2j}}{x_{2+}}\right)^2}$	Mètrica i euclidiana. No acotada.

Taula 3.2.7: Nom, abreviació, fórmula i propietats dels coeficients de dissimilaritat a comparar.

3.2.4.3 Transformacions i mesures de proximitat

A partir de cada set de dades, hem aplicat cinc transformacions diferents, generades mitjançant la equació $y = x^w$. Els exponents escollits han estat: w=0.0 (presència/absència), w=0.5, w=1.0, w=1.5, w=2.0. Aquests cinc exponents de transformació estableixen un gradient d'increment de la preponderància dels tàxons abundants respecte als tàxons poc abundants. Entre la gran quantitat de mesures de proximitat possibles, hem escollit comparar un subconjunt d'entre les mesures presentades a l'apartat 3.2.2. Les deu mesures de proximitat seleccionades es mostren a la taula 3.2.7. Hem exclòs algunes de les mesures de l'apartat 3.2.2 pels inconvenients teòrics que presenten, com ara els índexs de Gleason, Ellenberg i Spatz. També hem exclòs l'índex de Kulczynski degut a la seva estreta relació amb Bray-Curtis, i la distància valor absolut per l'augment de "borrositat" que provoca la transformació arrel quadrada en la distància de Manhattan (vegeu l'exemple de l'apartat 3.2.3.4).

Combinant les 10 dissimilaritats amb els cinc nivells de transformació, resulten 50 espais de relació, representats per 50 matrius de dissimilaritat, que hem anomenat mitjançant l'abreviació de la mesura de proximitat i el valor de l'exponent de transformació utilitzat (per exemple ED-0 per a la matriu de proximitats de la distància Euclidiana i la transformació amb exponent w=0).

Per tal de facilitar la representació, hem agrupat aquelles combinacions de transformació i mesura de proximitat que, per construcció, resultaven en un espai de relacions idèntic. És el cas de SR-0/CNB-0, CHO-0/HELL-0, CHO-0.5/HELL-1, CHO-1/HELL-2. Finalment, doncs, els espais de relació efectivament diferents són 46. Els índexs binaris de Jaccard i Sørensen estan representats per les combinacions SR-0/CNB-0 i BC-0, respectivament.

3.2.4.4 Metodologia de comparació

Inicialment, hem comparat les 46 matrius de dissimilaritat mitjançant la correlació de matrius simètriques amb el coeficient de Spearman (correlació per rangs). Així, les relacions que hem posat de manifest són les de monotonicitat, però aquestes no impliquen relacions lineals. Per a facilitar la interpretació de la matriu de correlació hem aplicat l'algorisme jeràrquic aglomeratiu del veí més llunyà o *complete linkage* sobre la matriu de correlacions resultant.

Per a poder avaluar la capacitat que té cada espai de relacions a l'hora de posar de manifest estructures de grup, hem executat l'algorisme *K-means* basat en distàncies (*DB-KM* apartat 3.2.3) sobre cada espai, cercant un nombre de grups des de K=2 fins a K=10 (per a les dades A) o K=20 (per a les dades B i C). La partició inicial de l'algorisme ha estat generada a partir del tall del dendrograma del mètode jeràrquic de Ward, i s'ha utilitzat la correcció *leave-one-out* (LOO)

en la classificació dels individus. Aquestes dues opcions són les que proporcionen una major efectivitat de l'algorisme en un temps de execució curt, tal i com vam concloure al capítol 3.1.

L'avaluació de la capacitat de revelar estructures s'ha fet mitjançant la silueta mitjana (Rousseeuw 1987). Cal tenir en compte que els valors de la silueta mitjana depenen de la magnitud dels valors de dissimilaritat, pel que la comparació de la silueta mitjana entre diferents espais de relacions s'ha fer en termes relatius, no absoluts. En el nostre cas, hem comparat la capacitat de l'espai de relacions per posar de manifest màxims de l'estadístic (o, en el seu defecte, punts de disminució de la pendent), i el nombre de grups al que es produïen aquests màxims.

El fet que cada espai de relacions posi de manifest estructura a escales diferents (diferent K) suposa un inconvenient a l'hora de comparar les particions resultants. És per aquest motiu que, per a comparar cada parella d'espais de relació hem calculat la mitjana d'acord a partir de la comparació a totes les escales. Concretament, per a cada parella d'espais 1 i 2, hem calculat l'índex de Rand (1971) corregit per l'atzar (Hubert & Arabie 1985) entre les particions K_1=2 amb K_2=2, K_1=3 amb K_2=3, ..., fins arribar a K_1=10 amb K_2=10, o K_1=20 amb K_2=20, segons el conjunt de dades. A continuació hem calculat la mitjana aritmètica d'aquests deu o vint valors per a tenir un sol valor de comparació per a cada parella d'espais de relacions. La matriu simètrica resultant de comparar totes les parelles d'espais de relacions mostra la proximitat dels resultats de classificació a totes les escales.

En el cas de la comparació amb la partició "externa", la classificació tradicional consta d'un nombre de grups fix (K_e=13 per a B o K_e=16 per a C). Donat que l'índex de Rand permet comparar nombres de grups diferents s'ha calculat la mitjana dels valors obtinguts entre les comparacions K=2 amb K_e=13, K=3 amb K_e=13, ..., fins a K=20 amb K_e=13 per al conjunt de dades A i entre les comparacions K=2/K_e=16, K=3/K_e=16,..., K=20/K_e=16 per a B. Hem representat les matrius simètriques resultants de les comparacions esmentades mitjançant l'anàlisi de coordenades principals i el mètode jeràrquic aglomeratiu del veí més llunyà o *complete linkage*.

Finalment, hem avaluat la capacitat de cada espai de relacions de recuperar la classificació original en l'execució d'una anàlisi discriminant basada en distàncies (Cuadras *et al.* 1997). L'esmentada capacitat ha estat avaluada comptant la freqüència d'èxit de classificació de la regla discriminant amb extracció de inventaris a classificar (validació *leave-one-out*).

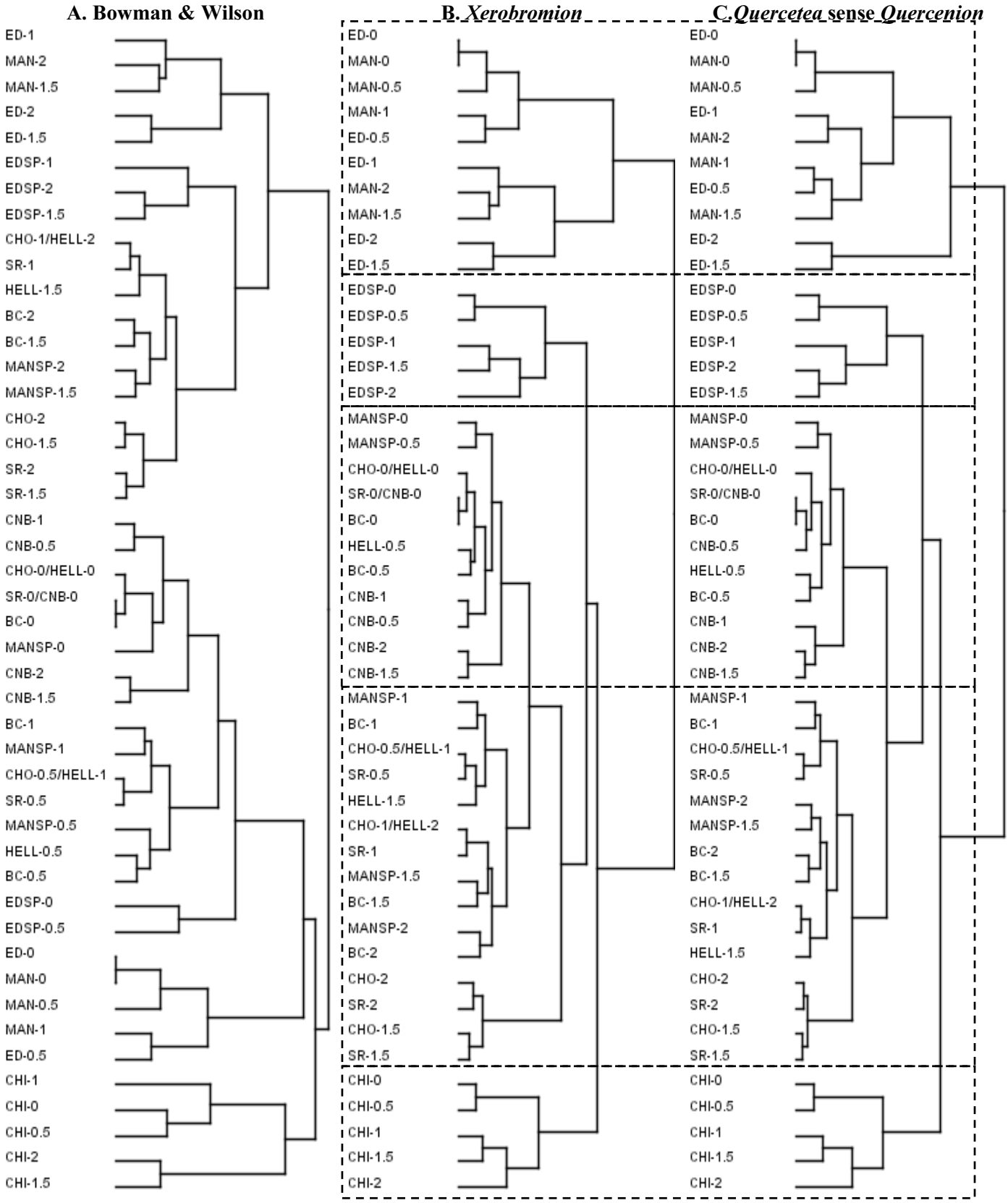


Figura 3.2.6.A-C: Dendrogrames *complete linkage* del complement de la correlació per rangs de Spearman entre matrius de dissimilaritat creades mitjançant diferents transformacions i mesures de proximitat. La correlació augmenta en incrementar-se el nombre d'objectes pel que la longitud de les branques no és significativa i l'escala ha estat omesa. Hem assenyalat amb quadres de línies discontínues els cinc grups d'espais de relacions que es repeteixen en els dendgrogrames B i C.

3.2.4.5 Resultats

Comparació de matrius de proximitat

Per a començar l'estudi, hem comparat les 46 matrius de dissimilaritat sorgides d'aplicar les transformacions i mesures de proximitat esmentades. La comparació utilitzada ha estat la correlació per rangs (Spearman) de matrius simètriques. Les figures 3.2.6.A-C mostren, per a cada un dels conjunts de dades respectius, els dendrogrames obtinguts pel mètode *complete linkage* aplicat sobre la distància arrel del complement del coeficient de correlació per rangs. En altres paraules, els dendrogrames mostren la proximitat entre els espais de relacions mesurada globalment.

Els dos dendrogrames B i C són més semblants entre ells que amb el dendrograma A, probablement perquè el conjunt de dades de Bowman & Wilson és força més petit. Per aquesta raó considerem més adequat extreure les conclusions dels dos dendrogrames B i C. Hem assenyalat amb quadres de línies discontínues els cinc grups d'espais de relacions que es repeteixen allí, que descrivim a continuació.

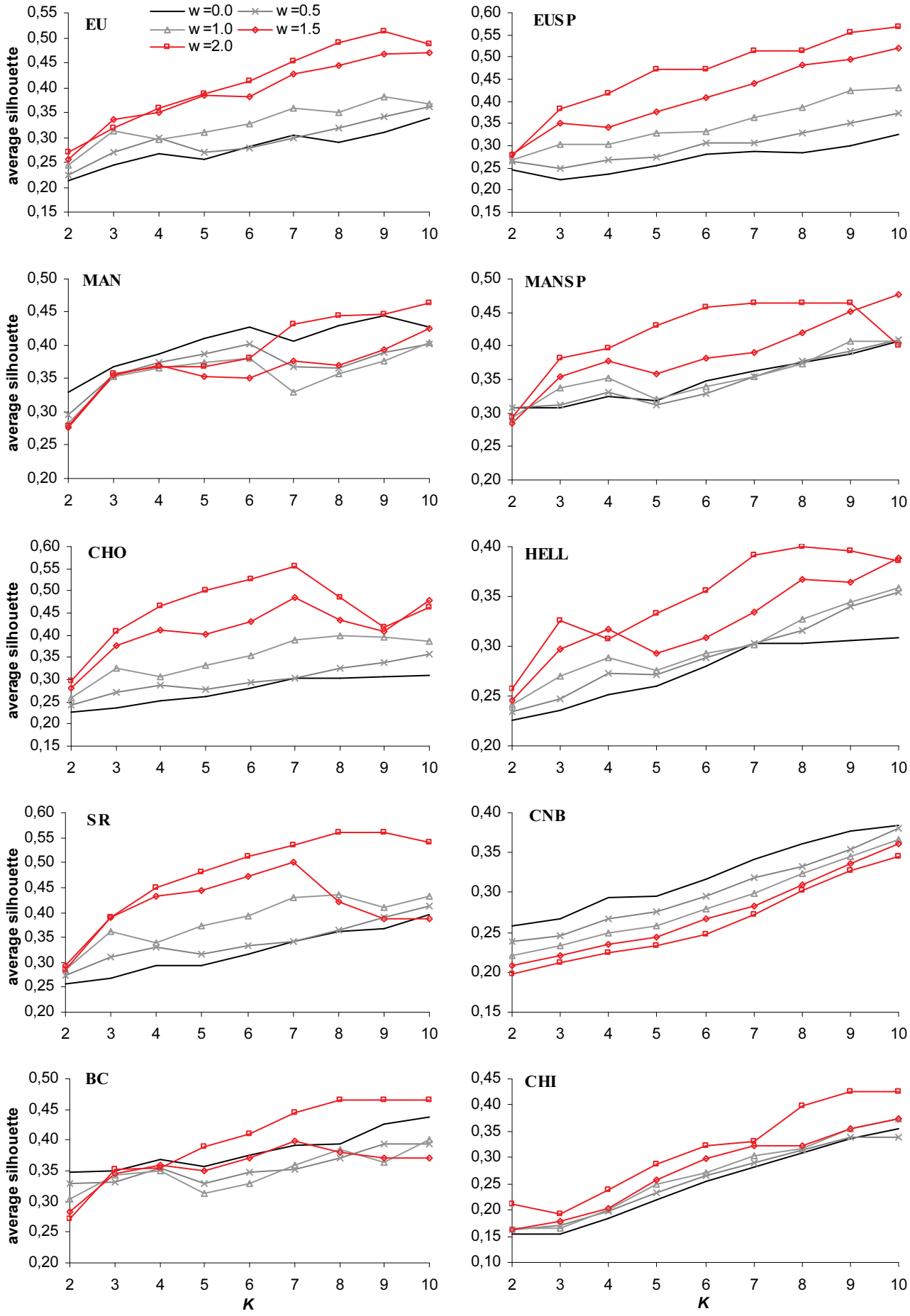
a) Les normes L1 i L2 (ED i MAN) formen un primer grup. Suggereixen un gradient en que progressivament prenen importància les diferències d'abundàncies grans: ED-0 i MAN-0 són completament monotòniques; després vindria MAN-0.5, MAN-1 i ED-0.5,... , fins ED-2.

b) Els espais de relacions generats amb el perfil d'espècies (EDSP) tenen configuracions particulars, lleugerament properes entre elles i distants de la resta. La distància χ^2 (CHI) proporciona espais de relacions que configuren també un altre grup separat.

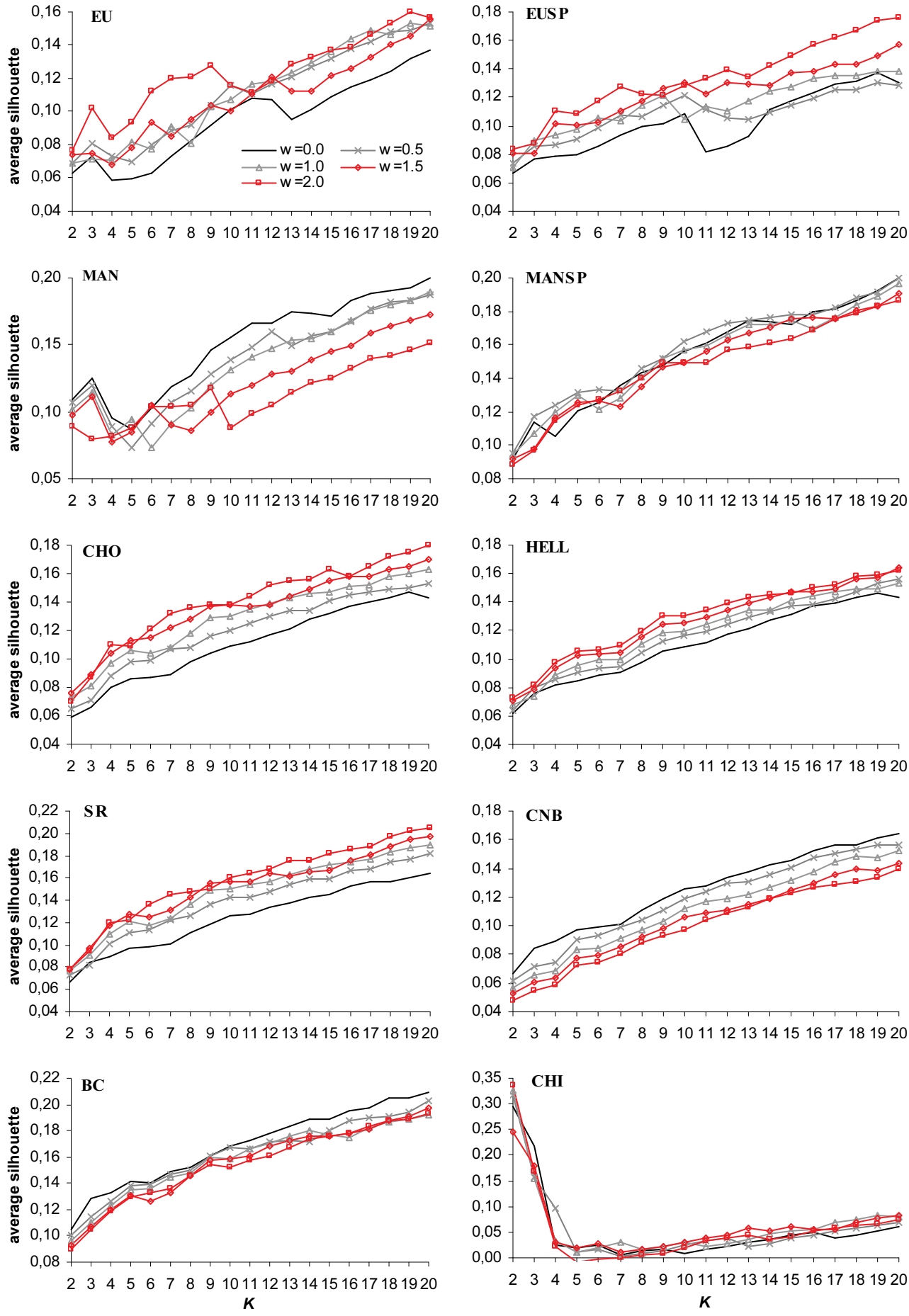
c) El tercer i quart grup de mesures assenyalats al dendrograma acullen els espais de relacions de les restants mesures de proximitat: MANSP, CHO, HELL, SR, BC i CNB. El tercer grup engloba els exponents de transformació baixos w=0 o w=0.5. Totes les matrius de proximitat de la mètrica de Canberra en la forma d'Adkins (CNB) es mostren properes entre elles i incloses dins d'aquest grup. En canvi, el quart grup agrupa els exponents de transformació mitjos i alts (w=1.0 a w=2.0). Dins d'aquest, les matrius de la distància Bray-Curtis (BC) i la mètrica de Manhattan estandarditzada (MANSP) sempre queden properes. Al seu torn, *similarity ratio* (SR) proporciona matrius de proximitat molt semblants en tots els casos a la distància de la corda (CHO). Noteu que aquestes dues mesures es separen de la resta per a exponents alts (w=1.5 i w=2.0).

Figures 3.2.7.A-C (properes tres planes): Perfils de l'estadístic silueta mitjana per a diferents espais de relació. Cada una de les 10 gràfiques correspon a una mesura de dissimilaritat i els 5 perfils que s'hi mostren es corresponen a les 5 transformacions (és a dir exponents de transformació) de les dades. L'eix de les abscisses marca el nombre de grups, entre K=2 i K=10 (A) o entre K=2 i K=20 (B i C). L'eix de les ordenades marca la silueta mitjana. La importància de l'escala d'aquest eix és secundària ja que es veu influïda pel rang de valors possibles de la mesura de dissimilaritat.

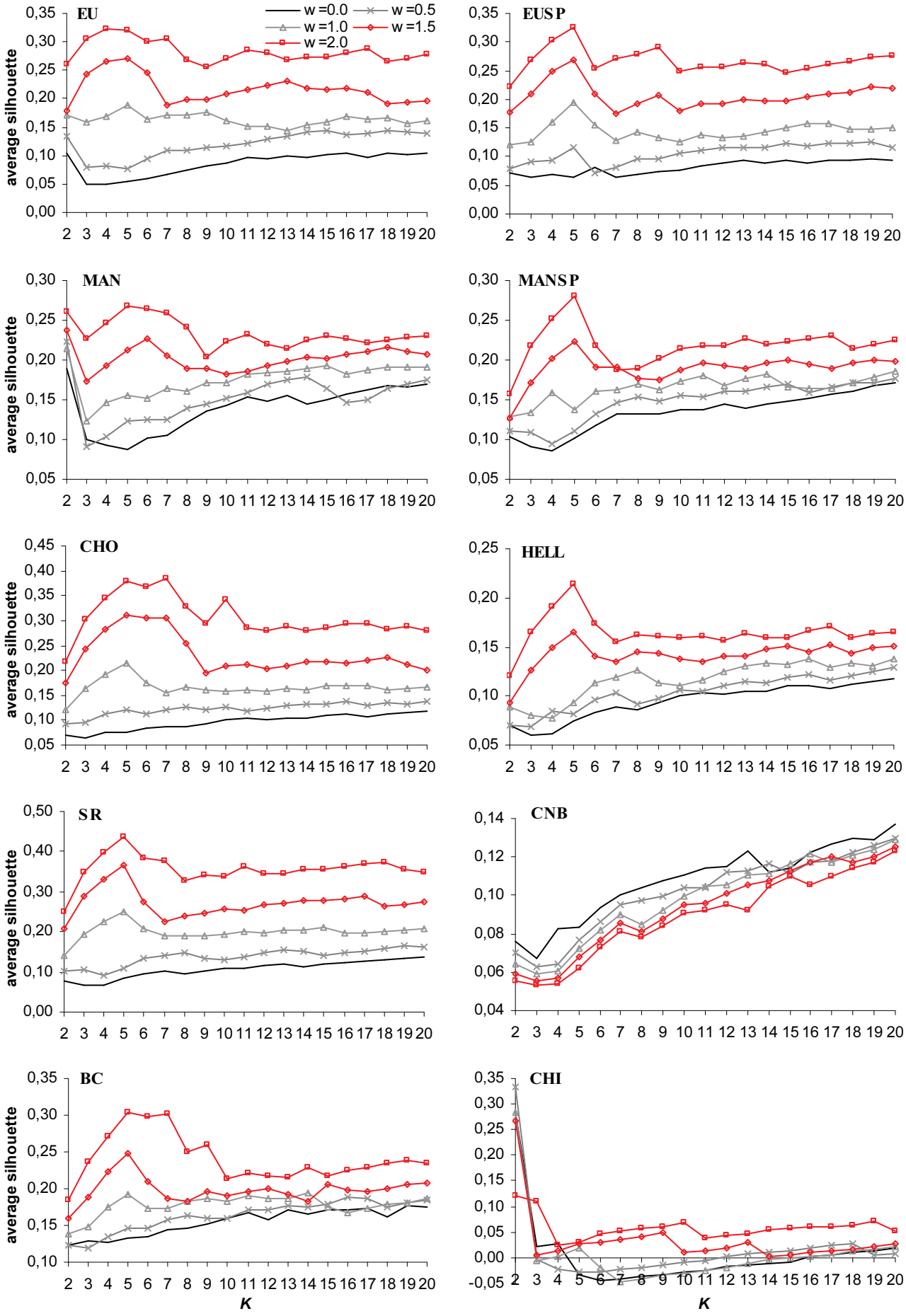
A. Bowman & Wilson



B.*Xerobromion erecti*



C.*Quercetea ilicis* sense *Quercenion*



Detecció de grups "naturals" en diferents espais de relacions

Pel que fa a la capacitat de cada mesura de proximitat de posar de manifest divisions "naturals" de les dades, a les figures 3.2.7.A-C mostrem els perfils de silueta mitjana obtinguts per cada mesura de dissimilaritat i transformació (vegeu properes tres planes). L'estadístic silueta mitjana presenta la tendència a augmentar en incrementar el nombre de grups. Aquest 'biax' es presenta quan el nombre d'objectes per grup és baix. No obstant això, sigui en termes absoluts o relatius a aquesta tendència, es poden observar a les gràfiques màxims o canvis de pendent que marquen el nombre de grups més idoni per a cada espai de relacions.

A les dades de Bowman & Wilson, Dale (1988a) establí que el nombre de grups de les dades era 3 o 4 (vegeu 3.1.6.2). Hom pot comprovar a les figures 3.2.7.A com no tots els espais de relacions expressen aquest nombre de grups de manera natural. Si comparem els perfils de la transformació neutra, w=1, BC-1 i MANSP-1 assenyalen molt clarament K=3 i K=4 com a bones opcions de tall. SR-1, CHO-1, MAN-1 i ED-1 indiquen K=3, i HELL es decanta per K=4. A la resta de dissimilaritats, CNB-1, CHI-1 i EDSP-1, el perfil és poc clar.

Els prats de *Xerobromion erecti* haurien de mostrar, segons la classificació tradicional, K=13 sintàxons de base, dels quals sabem que algunes subassociacions presenten problemes de baixa discriminabilitat. La aproximació numèrica partitiva no mostra massa punts de tall clars al voltant d'aquest valor però si a escales més grans (K inferiors). Per exemple, la norma Euclidiana (ED-1) i la mètrica de Manhattan (MAN-1) presenten un pic clar a K=3. En canvi, EDSP-1 assenyala les particions K=6 i K=9. Les mesures de dissimilaritat CHO-1, HELL-1, SR-1 i BC-1 assenyalen freqüentment K=5 i K=9 com a punts de tall més indicats. MANSP i CNB semblen ressaltar també K=5 però no K=9. Finalment, CHI-1 indica clarament una divisió en dos grups, en clara discordància amb els anteriors. Els matollars i màquies de *Quercetea* són dividits en la sintaxonomia tradicional en *K*=16 sintàxons de base, dels quals, no obstant, ja en coneixem els problemes de les subassociacions de *Quercetum cocciferae* (veure capítols 2.3 i 3.1). L'aproximació de *K-means* basada en distàncies genera aquí perfils de silueta mitjana més plans, ja que el nombre d'inventaris per grup és més gran. A més, assenyala en força espais de relació la partició més natural la corresponent a K=5. És el cas, amb w=1 de ED-1, EDSP-1, SR-1, CHO-1 i BC-1. En canvi CNB-1 assenyala *K*=7, i HELL-1/CHO-0.5 assenyala K=8. Al voltant del nombre de sintàxons de base originals (entre K=14 i K=16) hi ha algunes distàncies que marquen un possible punt de tall. De nou, la distància χ^2 marca, com en el cas de *Xerobromion*, la partició en dos grups. Per a més informació sobre les particions resultants d'aquests punts de tall (amb la distància de la corda), consulteu el capítol precendent (apartat 3.1.6.3, pp. 154-157).

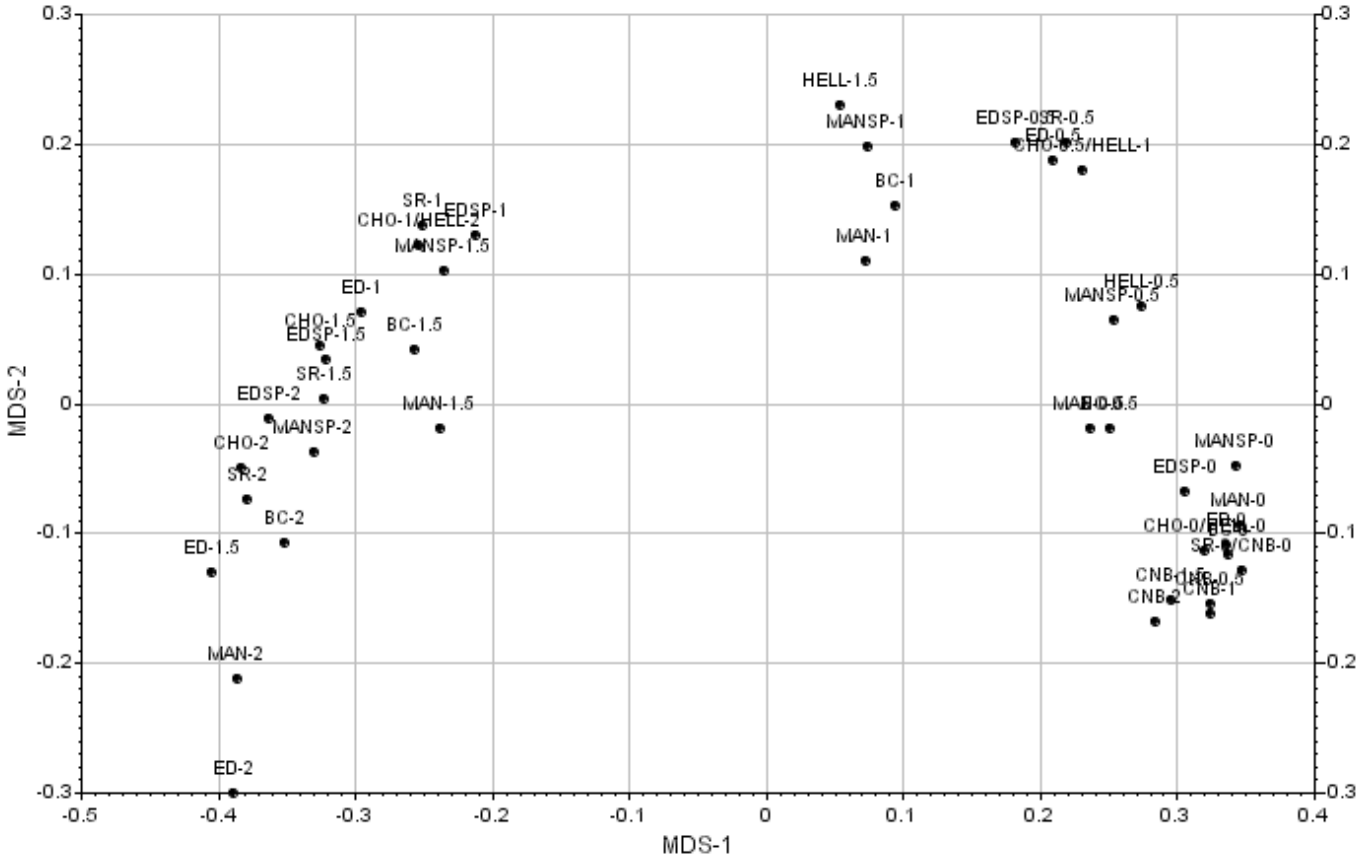
Com a corol·lari dels resultats que acabem de comentar podem observar que algunes mesures de dissimilaritat presenten més variació de la resposta que altres en relació al exponent de transformació. Aquesta estabilitat/inestabilitat de perfils es manifesta de manera semblant en

tots els sets de dades. És evident que CHO variarà més, per construcció, que HELL. SR mostra perfils molt semblants a CHO. D'altra banda, ED i MAN presenten més disparitat de solucions entre diferents exponents que no pas les seves versions transformades amb el perfil d'espècies (EDSP i MANSP). BC i CHI es mostren bastant estables per a exponents intermitjos. De llarg, la mesura que presenta perfils més estables és CNB, que a la vegada, són els menys marcats. Aquesta observació concorda amb la proximitat dels espais als dendrogrames 3.2.6.B i 3.2.6.C. Per tant, podem concloure que CNB és una mesura de proximitat amb una sensibilitat baixa.

Comparació entre particions

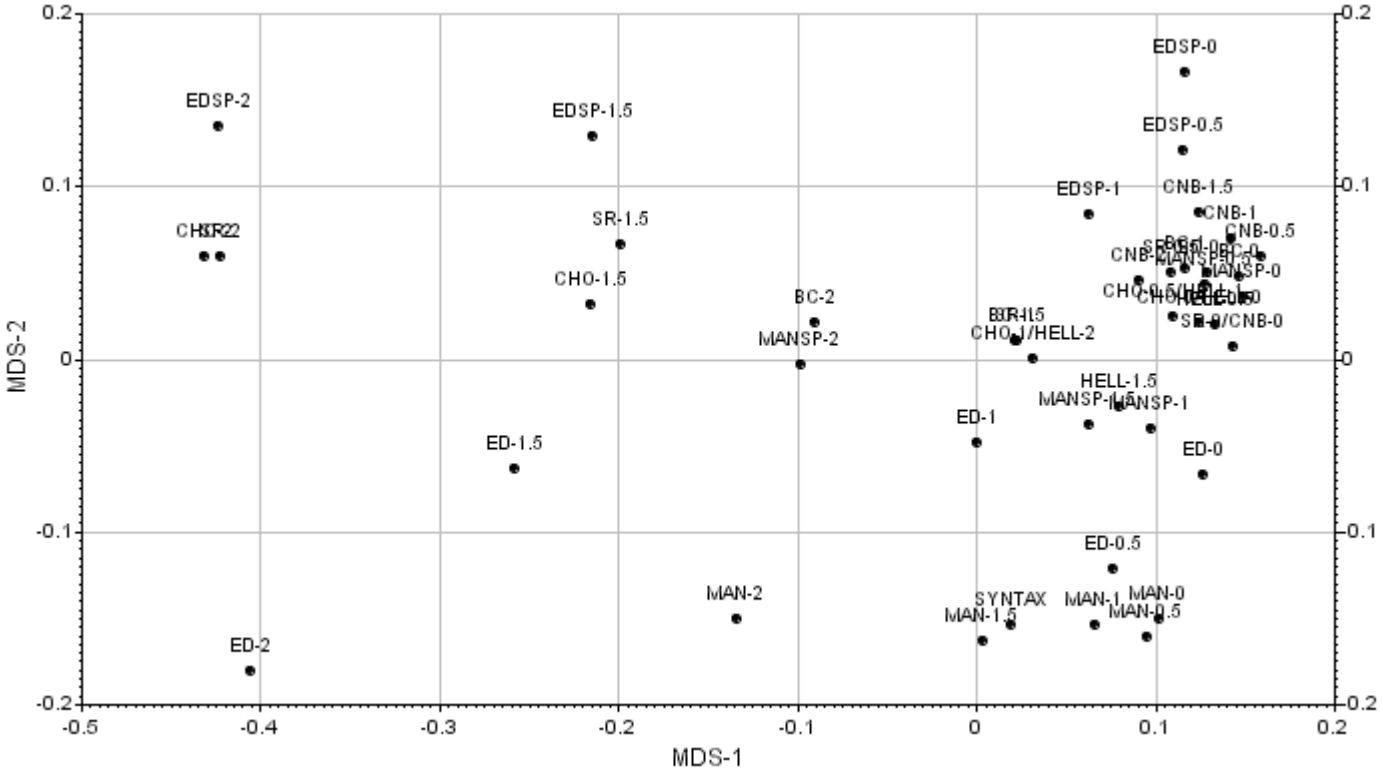
La següent manera de comparar els espais de relacions ha estat estudiant l'acord entre les particions generades a diferents escales. Per a cada parella d'espais hem calculat la mitjana aritmètica de l'índex de Rand corregit. Vegeu la pàgina 195 per a una descripció detallada del procés de comparació. A les figures 3.2.8.A-C presentem les dues primeres coordenades principals de la matriu de dissimilaritats complement de l'índex de Rand mitjà. Hem exclòs de les anàlisis d'ordenació les particions generades per la distància CHI per què, en ésser molt diferents de la resta de particions, acumulaven gran part de la variabilitat i impedien interpretar les relacions entre la resta de mesures de proximitat i transformacions.

A. Bowman & Wilson %Variabilitat = 56.17% + 11.28% = 67.45%

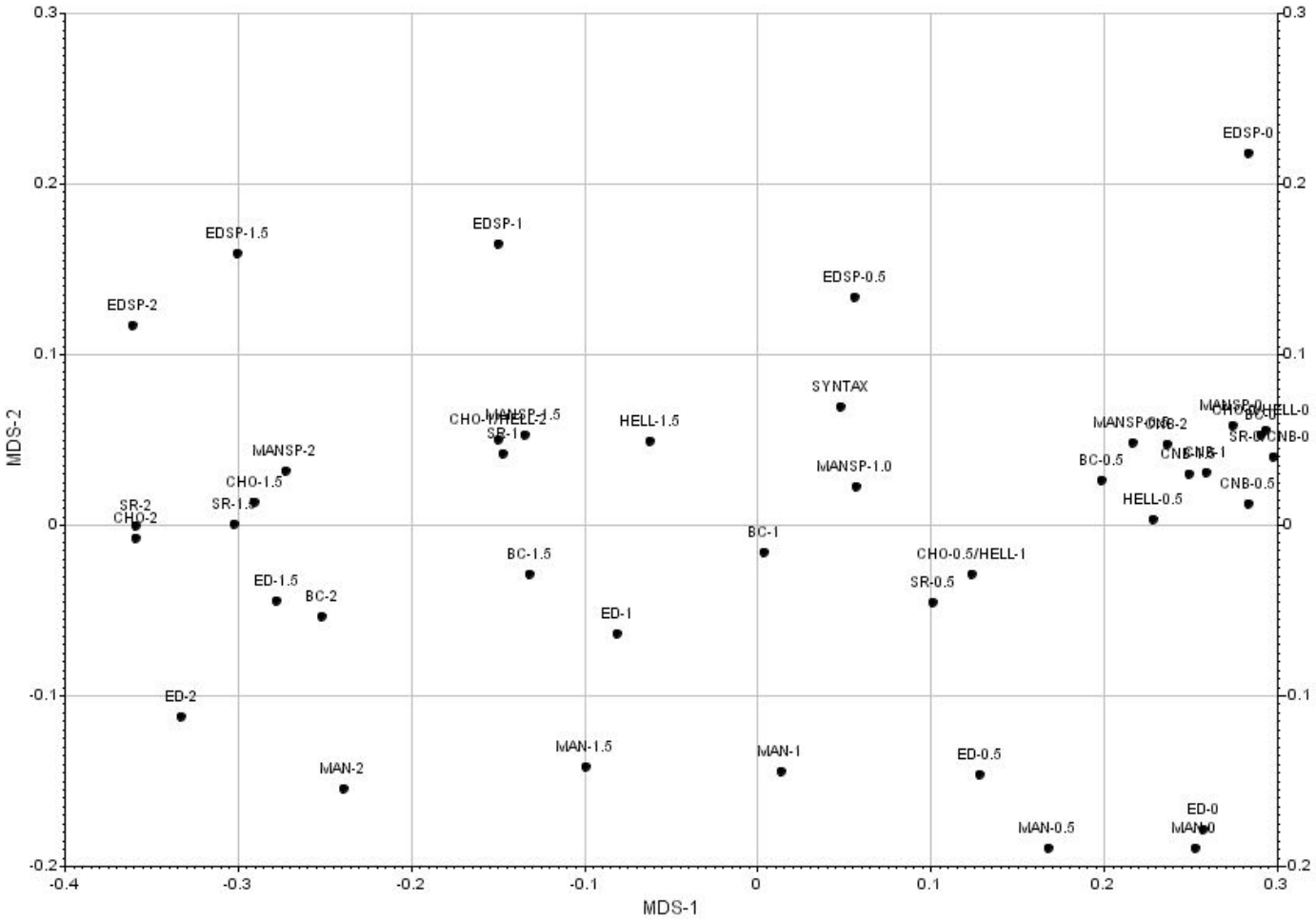


Figures 3.2.8.A-C: Diagrames de dispersió de les dues primeres components d'una anàlisi de coordenades principals realitzada a partir de la dissimilaritat complement de l'índex de Rand corregit per l'atzar promitg. S'indica a cada gràfica la proporció de variabilitat mostrada.

B.*Xerobromion erecti* %Variabilitat = 39.99% + 10.55% = 50.54%



C.*Quercetea ilicis* sense *Quercenion* %Variabilitat = 53.97% + 9.41% = 63.38%



La figura 3.2.8.A presenta un arc clar. El primer eix de la ordenació acumula gran part de la variabilitat representada (56%). Per aquest motiu serà l'únic que interpretarem. A l'extrem dret de l'eix es situen les transformacions properes a w=0 i a l'altre costat, a l'esquerra, les transformacions que donen més pes als tàxons abundants (w=1.5 i w=2). La majoria de mesures de proximitat tenen representants a les diferents parts de l'arc, a excepció de CNB, que concentra les classificacions properes al cas binari. D'altra banda, la distància de Hellinger no arriba a la part més esquerra de l'arc, perquè inclou la transformació arrel quadrada.

Les figures 3.2.8.B i 3.2.8.C no presenten aquesta estructura d'arc. Tanmateix, la interpretació del primer eix de la ordenació és semblant. En canvi, el segon eix de totes dues ordenacions té a un extrem els espais de relació de EDSP i a l'altre extrem els de ED i MAN. Els espais de MANSP també queden situats força amunt a la figura 3.2.8.C. La resta de mesures de proximitat (sense comptar CHI) es situaen a la part intermèdia. Una possible interpretació de l'eix tindria relació amb la inclusió de la abundància total de l'inventari en la mesura de les relacions de proximitat.

Les figures 3.2.9.A-C complementen la informació proporcionada per les ordenacions anteriors. Mostren els dendrogrames resultants del mètode aglomeratiu *complete linkage* aplicat sobre les matrius de similaritat valorada amb l'índex de Rand mitjà. Els dendrogrames A-C s'assemblen menys entre ells que els de les figures 3.2.6.A-C, fent palès l'increment de complexitat que suposa l'aplicació dels mètodes de *clustering*. El primer dendrograma, 3.2.9.A separa en dos grups les classificacions. Aquests dos grups es corresponen, evidentment, als grups observats en l'arc de la ordenació 3.2.8.A i concorden quasi completament amb els dos grans grups del dendrograma de la figura 3.2.6.A. En canvi, els cinc grups d'espais de relació que trobavem als dendrogrames 3.2.6.B i 3.2.6.C no es repeteixen aquí, amb l'excepció, potser, del grup determinat per la mesura de proximitat CHI. Comparant globalment els dendrogrames de les figures 3.2.6.A-C amb els de les figures 3.2.9.A-C hom es pot plantejar: *És equivalent la comparació entre espais de relació per a ordenacions que per a classificacions?* Per tal d'esbrinar-ho, hem calculat la correlació per rangs de Spearman entre la matriu d'acord entre matrius de dissimilaritat (la utilitzada per a generar els dendrogrames de les figures 3.2.6.A-C), i la matriu d'acord entre particions (la utilitzada per a generar els dendrogrames de les figures 3.2.9.A-C). Les correlacions són 0.648 per a A, 0.503 per a B i 0.500 per a C. Aquests valors de correlació no són massa alts, i demostren que el comportament semblant de dues mesures de proximitat en el conjunt de l'espai de relacions no implica un comportament semblant quan s'apliquen mètodes de *clustering*. Els darrers es veuen més afectats per les diferències a les distàncies petites. En canvi, en la correlació entre matrius de dissimilaritat intervenen totes les distàncies, grans i petites. Com que els mètodes d'ordenació donen prioritat a les distàncies grans, hom pot esperar que les relacions que mostren en el diagrames resultants siguin també lleugerament diferents de les relacions importants per als mètodes de *clustering*.

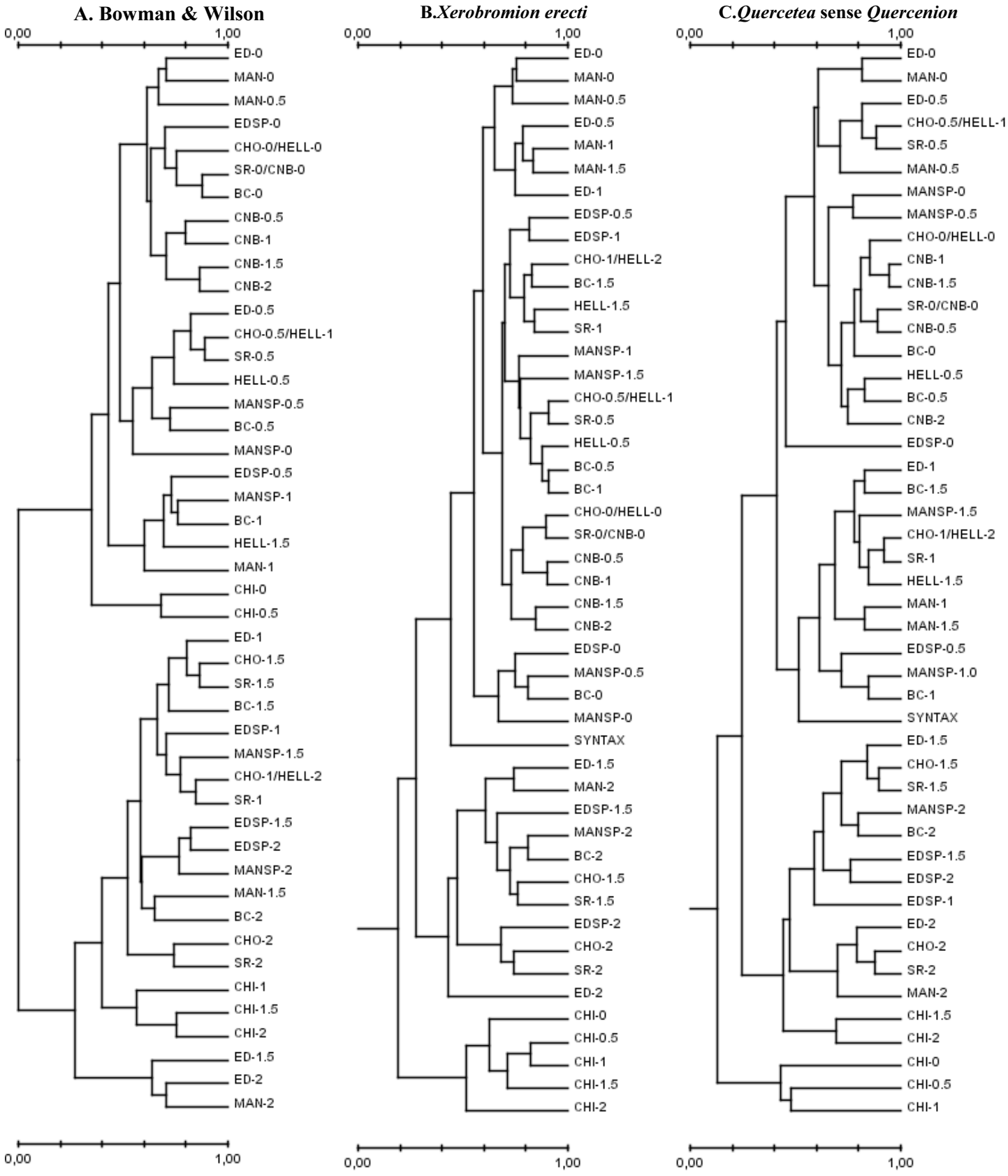


Figura 3.2.9.A-C: Dendrogrames obtinguts mitjançant *complete linkage* sobre la matriu de valors mitjans de l'índex de Rand.

Comparació amb el criteri tradicional

La partició tradicional en sintàxons de base apareix força aïllada als dos dendrogrames de les figures 3.2.9.B i 3.2.9.C. En tot cas, el criteri tradicional sembla estar relacionat sobretot amb particions obtingudes amb exponents de transformació mitjos/baixos en B (des de w=0 a w=1.5) i mitjos en C (des de w=0.5 a w=1.5). Les taules 3.2.8.B i C presenten els valors d'ajust de Rand mitjà de comparació amb la sintaxonomia. A les dades de *Xerobromion erecti* les cinc mesures de proximitat que generen classificacions més pròximes a la tradicional són, per ordre, MANSP, HELL (o CHO), ED i SR. En el cas de *Quercetea* sense *Quercenion*, les mesures de proximitat més idònies serien HELL (o CHO), SR, EDSP i BC. Recordeu que HELL, CHO i SR són mesures de proximitat amb propietats semblants, i que també ho són BC i MANSP. Sorprèn la relativa bona marca d'ED i EDSP amb exponents baixos (w=0.5). Com que els seus resultats varien molt per a diferents exponents de transformació, no considerem els resultats prou concloents com per a incloure EDSP o ED entre les mesures de proximitat habitualment més eficients.

Pel que fa als exponents de transformació, el més indicats serien w=0.5 o w=1.0 per al *Xerobriomion erecti* i lleugerament més elevats per a *Quercetea*. Aquesta diferència entre un i altre conjunt de dades vindria a fer palès que als boscos i matollars de *Quercetea* les abundàncies dels tàxons són probablement més importants per a establir els grups que als prats de *Xerobromion*. Aquest fet confirma els resultats del capítol precedent (vegeu pàg. 149).

B.*Xerobromion erecti*

	ED	EDSP	MAN	MANSP	CHO	HELL	SR	CNB	BC	CHI
0	0.488	0.447	0.499	0.443	0.503	0.503	0.502	0.502	0.472	0.238
0.5	0.543	0.482	0.510	0.490	0.544	0.539	0.527	0.512	0.519	0.293
1	0.506	0.479	0.514	**0.549**	0.505	0.544	0.525	0.497	0.512	0.288
1.5	0.428	0.431	0.518	0.512	0.415	0.528	0.424	0.486	0.525	0.318
2	0.272	0.312	0.495	0.470	0.336	0.505	0.338	0.490	0.487	0.329

C.*Quercetea ilicis* sense *Quercenion*

	ED	EDSP	MAN	MANSP	CHO	HELL	SR	CNB	BC	CHI
0	0.444	0.443	0.457	0.452	0.517	0.512	0.475	0.475	0.504	0.256
0.5	0.533	0.553	0.498	0.497	0.567	0.525	0.558	0.476	0.545	0.305
1	0.531	0.476	0.526	0.536	0.539	0.567	0.536	0.529	0.551	0.416
1.5	0.444	0.426	0.518	0.531	0.456	**0.569**	0.448	0.534	0.519	0.415
2	0.378	0.345	0.443	0.463	0.396	0.539	0.388	0.517	0.47	0.408

Taules 3.2.8.B-C: Acord entre les classificacions generades per DB-KM, emprant les diferents mesures de proximitat (columnes) i els diferents exponents de transformació (files), i la partició de la sintaxonomia tradicional en K =13 (B.*Xerobromion erecti*) i K =16 (C.*Quercetea ilicis* sense *Quercenion*). Els valors mostrats són els de l'índex de Rand corregit per l'atzar, promitjat entre les particions K =2, K =3,..., K=20. S'assenyala la transformació de cada mesura de proximitat que presenta un millor ajust. A la vegada, hem ressaltat amb negreta el valor més alt de la taula.

Per tal de completar la comparació amb el criteri extern, presentem a la taules 3.2.9.B-C la capacitat de classificació correcta d'inventaris de cada espai de relacions, valorada a partir de l'execució d'una anàlisi discriminant basada en distàncies. Noteu que l'extracció *leave-one-out* en el cas de la distància CHI no és del tot vàlida ja que caldria, a cada extracció, calcular de nou tota la matriu de distàncies. No obstant, no creiem que aquest inconvenient hagi afectat massa els resultats. Els percentatges de recuperació *leave-one-out* són força elevats en tots els casos (recordeu que no són valors de l'índex de Rand). Si comparem els valors més alts de cada columna, veurem que les mesures de proximitat amb percentatges de recuperació més elevats són CHO i HELL, seguides de CNB, BC i EDSP. Pel que fa a la transformació més idònia els resultats confirmen la tendència observada a les taules 3.2.8.B-C.

B.*Xerobromion erecti*

	ED	EDSP	MAN	MANSP	CHO	HELL	SR	CNB	BC	CHI
0	89.1	87.9	87.9	87.9	89.1	89.1	89.9	89.9	88.3	82.7
0.5	89.1	90.3	87.9	89.1	**90.7**	90.3	89.9	89.9	90.3	85.5
1	87.1	86.7	87.5	88.7	88.3	**90.7**	87.9	90.3	89.1	85.1
1.5	83.1	82.3	87.5	86.7	83.1	89.1	83.1	89.9	86.3	85.1
2	76.6	76.2	85.1	85.5	79.0	88.3	79.4	89.1	85.9	83.5

C.*Quercetea* sense *Quercenion*

	ED	EDSP	MAN	MANSP	CHO	HELL	SR	CNB	BC	CHI
0	86.2	87.8	80.1	83.8	89.6	89.6	89.1	89.1	88.0	80.9
0.5	89.6	90.7	83.8	87.8	**91.2**	90.7	89.9	90.2	89.6	84.6
1	88.3	85.4	84.8	89.4	88.6	**91.2**	86.4	91.0	89.1	87.8
1.5	83.2	77.7	84.3	88.8	83.0	90.2	81.4	90.7	87.2	87.2
2	76.3	70.2	81.4	83.2	77.7	88.6	76.1	90.7	84.3	84.8

Taules 3.2.9.B-C: Capacitat de reclassificació d'inventaris dels diferents espais de relació. Percentatges de reclassificació correcta a partir de la avaluació *leave-one-out* de la regla discriminant basada en distàncies emprant la sintaxonomia tradicional com a classificació de partida. S'assenyala la transformació de cada mesura de proximitat que presenta una millor reclassificació. A la vegada, hem ressaltat amb negreta el valor més alt de la taula.

3.2.4.4 Discussió

Són àmpliament acceptats els inconvenients de la distància Euclidiana quan és aplicada a l'estudi de les relacions entre inventaris. En el nostre estudi hem mostrat que la capacitat de la distància Euclidiana per a posar de relleu estructures és força depenent de la transformació escollida per a les dades. A la vegada, l'ajust al criteri extern és dels més pobres.

La mètrica de Manhattan sembla obtenir resultats lleugerament millors que la distància Euclidiana. Tot i això, cal recordar que en l'aproximació *DB* (*distance based*) de *K-means* i a l'anàlisi discriminant, el punt central del clúster a partir del qual es calculen les distàncies al mateix és el centroide. Per tant, els algorismes *DB* operen com si la norma fos L2, independentment de que en realitat ho sigui o no. En el cas d'una norma L1 seria més adequat, potser, el càlcul de medoids (punts central mediana). Tanmateix, pel que fa al *clustering* sembla clar que el fet que una mesura sigui simètrica (no elimini les dobles absències) perjudica els resultats de la classificació, ja que tant MAN com ED proporcionen resultats pobres en comparació amb altres dissimilaritats. No creiem, però, que aquest sigui l'únic factor de la seva falta d'eficiència.

La transformació del perfil d'espècies provoca canvis profunds en l'espai de relacions, tant de ED com a MAN. No obstant, en el cas de la norma Euclidiana l'espai resultant, EDSP, no sembla més adequat per al *clustering* que ED. En canvi, la mateixa transformació aplicada a MAN incrementa notablement tant la capacitat de detectar estructures com la semblança dels resultats amb el criteri extern fitosociològic. En aquest sentit, confirmem la recomanació de MANSP feta per Cambell (1978) i Faith *et al*. (1987). D'altra banda, la estreta relació d'aquesta dissimilaritat amb la distància de Bray-Curtis (BC, 1957) mereix un estudi més aprofundit de les relacions teòriques de les dues mesures.

Ja hem comentat que, en el context de transformacions de les dades adoptat, cal tractar les distàncies de la corda (CHO, Orlóci 1967) i Hellinger (HELL, Rao 1995) com a proximitats equivalents. L'èxit de Hellinger per damunt de la corda es deu només al rang de transformacions que hem comparat. Amb la distància de la corda, com amb ED, les transformacions w=1.5 i w=2.0 exageren massa la importància de les espècies amb abundància gran i els resultats del *clustering* es veuen deformats. La transformació de Hellinger es pot aplicar a un rang més ampli de dades, ja que l'arrel quadrada implícita en la transformació és un criteri una mica conservador pel que fa a donar massa pes a unes espècies respecte les altres. Tot i això creiem que, en conjunt, HELL és la millor opció d'entre aquelles mesures de proximitat euclidianes.

Similarity ratio (SR, Wishart 1969) fou una mesura força utilitzada durant els anys 70 i 80. Actualment, sembla haver caigut en desús en favor de CHO o HELL. Confirmem aquí que els

resultats de SR són molt propers als de CHO, tant en ordenacions com en classificacions. La no euclidianitat de SR és un arguments prou fort per a preferir CHO o HELL.

La mètrica de Canberra en la forma d'Adkins (CNB, Lance & Willams 1967) presenta percentatges de classificacions correctes a l'anàlisi discriminant força bones. Aquest fet podria ser degut a la coincidència amb el criteri tradicional a donar pes a espècies poc abundants. En fitosociologia a vegades els tàxons poc abundants tenen força importància perquè poden presentar una fidelitat alta. No obstant, la baixa capacitat de CNB per a posar de relleu estructures (almenys amb el criteri de la silueta mitjana) fa que aquesta mètrica sigui un mal candidat per a estudiar l'estructura d'inventaris de vegetació sense cap coneixement *a priori* del possible nombre de grups de les dades.

La distància de Bray-Curtis (BC, Bray & Curtis 1957) és una dissimilaritat semi-mètrica i no euclidiana. Malgrat això, sembla donar força bons resultats en combinació amb els mètodes de classificació basats en distàncies. Juntament amb MANSP, és una mesura sensible a l'hora posar de manifest estructures de classificació, i l'ajust a la classificació fitosociològica tradicional és dels millors, darrera de HELL.

D'entre les mesures de proximitat estudiades, la distància χ^2 (CHI) és la menys adequada per a l'aproximació *DB*, fins i tot per sota d'ED. La raó és la ponderació en excés d'espècies rares i inventaris amb pocs tàxons, que resulta en un espai de relacions poc apte per a cercar estructures esfèriques. En relació a aquest espai de CHI, l'aproximació *DB* tendeix a senyalar particions en $K=2$ grups. Aquest fet ens permet especular que hom podria dissenyar un algorisme que realitzés recursivament particions binàries. Aquest algorisme seria anàleg a l'algorisme *TWINSPAN* (Hill *et al*. 1974), però emprant tot l'espai proporcionat per la distància CHI i no només el primer eix de l'anàlisi de correspondències. Un criteri d'aturada de les divisions podria ser la disminució d'un estadístic d'avaluació de l'aïllament o la disminució del nombre de tàxons fidels per sota d'un determinat llindar.

3.2.4.5 Conclusions

Les aportacions fetes en aquest capítol es poden sintetitzar en els següents punts:

- Hem aportat alguns elements teòrics nous de comparació entre mesures de proximitat. Concretament, creiem interessants les relacions entre la distància de la corda i la distància de Hellinger; les relacions entre les normes L1 i L2 i les transformacions que proporcionen longituds del vector 1; i la relació entre la distància de Manhattan relativitzada i la distància de Bray-Curtis.
- Hem proposat una nova aproximació al *clustering* partitiu basada en matrius simètriques de dissimilaritat entre objectes (aproximació *DB*). Aquesta aproximació obre la porta a realitzar particions sobre qualsevol espai de dissimilaritats. També, hem estudiat els efectes que provoca la no-euclidianitat de l'espai de relacions sobre l'aproximació *DB* demostrant que les correccions per assolir euclidianitat augmenten la borrositat de l'espai de relacions.
- Entre les mesures de proximitat aplicades al *clustering*, les opcions que combinen una bona capacitat de detecció de clústers i un acord amb la classificació tradicional són les transformacions/distàncies de la corda/Hellinger per a la norma L2 i la transformació del perfil d'espècies per a la norma L1. A la taula següent resumim les dissimilaritats recomanades segons els requeriments de metricitat i euclidianitat desitjats.

Requeriments	Distàncies recomanades
Mètrica i euclidianitzable	HELL (CHO)
Mètrica	HELL (CHO) o MANSP
Semimètrica	HELL (CHO), MANSP o BC

- Pel que fa a la relació entre la transformació de les escales ordinals de Braun-Blanquet i *Domin* i els resultats dels algorismes de *clustering* basats en espais mètrics, és important la raó entre el valor que representa l'absència (0) i l'augment per als restants estats ordinals. Recomanem un exponent de transformació escalar comprès entre $w=0.5$ i $w=1$ a partir de la transformació combinada de van der Maarel, o un exponent comprès entre $w=0.25$ i $w=0.5$ a partir dels percentatges de cobertura de Braun-Blanquet. D'altra banda, els resultats semblen indicar lleugeres diferències, pel que fa a l'exponent idòni, entre els parts de *Xerobromion* i les bosquines o matollars de *Quercetea* sense *Quercenion*.

Capítol 3.3: Identificació de regions denses en l'espai multivariant de comunitats

3.3.1 Introducció

3.3.1.1 L'anàlisi de densitats per a reconèixer grups de comunitats vegetals

La majoria de treballs d'anàlisi de comunitats utilitzen sovint els mateixos mètodes multivariants de classificació i ordenació. Podani *et al*. (2000) adverteixen del risc d'aplicar còmodament, sense crítica, un cop i altre els mateixos mètodes. En general, els estudis d'anàlisi de clúster (*cluster analysis*) de vegetació es limiten a l'aplicació dels models jeràrquic (aglomeratiu o divisiu) i, més recentment, el partitiu (vegeu el capítol 3.1 per a una revisió d'aquests mètodes). En taxonomia i sintaxonomia numèrica existeixen, però, alguns precedents de cercar altres models de classificació amb menys restriccions.

Ja Wishart defensava a l'any 1968 que el *cluster analysis* hauria d'orientar-se a identificar modes a les dades. Un clúster "natural" hauria de mostrar un centre dens (de qualsevol forma), envoltat per un núvol de punts. El problema rau en aïllar correctament aquests centres. Wishart es referia als *outliers* com a les fonts de soroll de les dades, i advertia que aquest soroll pot alterar les tècniques de *clustering*. Wishart optà per eliminar aquells objectes que no arriben a tenir un nombre de connexions donat. Uns anys més tard, Wildi (1979) reprengué la cerca de regions denses o nodes (*nodal types*). Wildi proposà un mètode de classificació (GRID). Aquest mètode es basa en dividir els primers eixos d'una ordenació en segments per a cercar després aquells hipercubs multidimensionals més densos. Així, Wildi basa el seu mètode en un criteri de densitat local. Una de les limitacions del seu mètode és haver d'escollir el nombre de segments per eix. Aquest mètodex fou l'escollit per Escudero & Pajarón (1994) per a determinar aliances i subaliances d'*Asplenietalia petrarchae* (vegetació rupícola) sobre una anàlisi factorial de correspondències (*correspondence analysis*).

En aquest capítol volem reprendre la cerca de regions denses com a model de classificació. En primer lloc analitzarem alguns problemes de les particions d'objectes. En segon lloc, introduirem el model de classificació possibilístic i l'algorisme *Possibilistic C-means* (PCM, Krishnapuram & Keller 1993), que posteriorment intentarem adaptar a les necessitats de l'ecologia numèrica. A la darrera secció testarem les modificacions d'aquest algorisme sobre les dades de vegetació que ja coneixem (*Brometalia erecti* i *Quercetea ilicis*), avaluarem la seva resposta i la compararem amb la que proporcionen els models de classificació partitius.

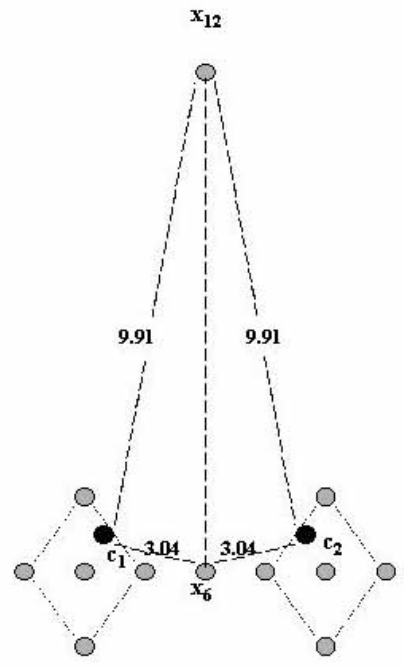
3.3.1.2 Particions i els *outliers*

En la terminologia de *clustering*, s'anomena *outlier* aquell objecte que presenta una posició extrema, distant dels altres objectes. A vegades apareix degut a un error o a soroll en el procés de mesura de les dades (Jain & Dubes 1988). El model de classificació partitiu de *K-means* (*KM*, MacQueen 1967) i *Fuzzy C-means* (*FCM*, Bezdek 1981, 1987) presenta alguns problemes quan les dades estan "contaminades" per *outliers*.

Per a fer entendre millor el problema dels objectes extrems en algorismes partitius, presentem un exemple extret de Pal *et al*. (1997). L'exemple és semblant al presentat a l'apartat 3.2.3.4 Suposem una situació de dos clústers com la que apareix a la figura 3.3.1. Un objecte $\mathbf{x}_6$ comparteix una relació amb dos clústers $\mathbf{C}_1$ i $\mathbf{C}_2$ per igual (al punt mig de la recte que uneix els dos centroides $\mathbf{C}_1$ i $\mathbf{C}_2$). L'execució de *FCM* ($K = 2$) sobre aquest set de dades portarà $\mathbf{x}_6$ a una pertinença de 0.5 als dos clústers, indicant aquesta posició intermèdia (taula 3.3.1). D'altra banda, tenim un objecte, $\mathbf{x}_{12}$, equidistant però allunyat dels dos centroides d' $\mathbf{C}_1$ i $\mathbf{C}_2$. Aquest segon objecte obtindrà, amb *FCM*, també una pertinença de 0.5 en ambdós clústers! La conseqüència més important no és la consideració de classificació errònia d'$\mathbf{x}_{12}$, sinó que, en tenir pertinença 0.5, aquest objecte exerceix un efecte d'atracció sobre els centroides, cosa que pot modificar la classificació dels altres objectes. En dades on el nombre d'*outliers* sigui considerablement gran això pot esdevenir un problema.

En resum, tal i com apunten Krishnapuram & Keller (1993), les pertinences d'una partició difusa són valors relatius, que posen de manifest el grau de compartiment (*sharing degree*) de l'objecte entre els clústers. No hi ha una superposició real dels mateixos. Sobra comentar que els *outliers* són també un problema per a establir particions *crisp*.

	Data		FCM	
pt	x	y	U1	U2
1	-5.00	0.00	0.93	0.07
2	-3.34	1.67	0.97	0.03
3	-3.34	0.00	0.99	0.01
4	-3.34	-1.67	0.90	0.10
5	-1.67	0.00	0.92	0.08
6	0.00	0.00	0.50	0.50
7	1.67	0.00	0.08	0.92
8	3.34	1.67	0.03	0.97
9	3.34	0.00	0.01	0.99
10	3.34	-1.67	0.10	0.90
11	5.00	0.00	0.06	0.94
12	0.00	10.00	0.50	0.50



Taula 3.1.1 i Figura 3.3.1: Configuració de dos clústers amb un objecte intermedi, $\mathbf{x}_6$, i un *outlier*, $\mathbf{x}_{12}$. Matriu de dades i partició *FCM* a baix i diagrama de dispersió a la dreta. Exemple extret de Pal et al. (1997).

3.3.1.3 Un model de classificació difús no partitiu

Per tal de superar l'inconvenient dels *outliers* sense abandonar el model partitiu, hom pot augmentar el nombre de grups i obtenir grups amb un sol objecte. Alternativament hom pot desfer-se del concepte de partició. Prendrem aquí a aquesta segona opció. Per a començar, recordem les 3 condicions per que una matriu $\mathbf{U}_{NxK}$ sigui una partició difusa no degenerada (apartat 3.1.3.1):

$$\text{(a) } u_{i(k)} \in [0,1] \forall\, i,k \qquad \text{(b) } \sum_{k=1}^{K} u_{i(k)} = 1 \quad \forall\, i \qquad \text{(c) } 0 < \sum_{i=1}^{N} u_{i(k)} < N \;\forall k$$

Si deixem de considerar la condició (b), la suma de les pertinences d'un objecte a cada un dels grups serà indeterminada, i, per tant, els objectes extrems podran tenir pertinences baixes per a tots els clústers. Per contra, es podran donar situacions en que objectes tinguin una alta pertinença en més d'un grup. En aquest model de classificació podrem establir cada clúster independentment dels altres i, sortosament, el nombre de clústers a trobar no serà un paràmetre a decidir per l'usuari. És important recordar, que cada problema concret de classificació ha de suggerir si aquest o un altre model de classificació és el més adequat per a les dades de que es disposa.

Per a diferenciar aquest model d'una partició, anomenarem a partir d'ara tipicalitat, *t*, al grau de pertinença en que un objecte és membre d'un clúster independent. Les restriccions algebraiques d'una tipicalitat són:

$$\text{(a) } t_{i(k)} \in [0,1] \forall\, i,k \quad \text{(b) } 0 < \sum_{i=1}^{N} t_{i(k)} < N \;\forall k \quad \text{(c) } \max_{i}(t_{i(k)}) > 0 \quad \forall i$$

La tipicalitat, o pertinença possibilística, és una mesura absoluta, per comparació amb la pertinença de la partició difusa, que és una mesura relativa (pertinença probabilística).

3.3.1.4 Funcions de tipicalitat. Algorisme *Possibilistic C-means*

Un cop definit el model de classificació possibilístic, hem de trobar una funció de tipicalitat que ens permeti desenvolupar un algorisme de classificació. Sota el criteri de Wishart (1968) i Wildi (1989), un clúster és una regió relativament densa de punts en l'espai. En aquest escenari, la classificació deriva en una cerca de regions denses o modes. Una funció de tipicalitat útil per cercar regions denses ha de complir, a més de les condicions de tipicalitat (*tip*) de l'anterior apartat, les tres condicions següents:

(d) Ésser una aplicació sobre l'espai de relacions contínua i derivable.

(e) $tip\ (d = 0) = 1.0$

(f) Ésser monotònicament decreixent en augmentar la dissimilaritat.

Krishnapuram & Keller (1993) proposaren un senzill algorisme difús, anomenat *Possibilistic C-means* (*PCM*), on la funció de pertinença compleix les condicions de funció de tipicalitat per cercar regions denses. La funció de tipicalitat que proposaren aquests autors s'entén millor si partim de la funció de pertinença de *FCM*. Recordem, de l'apartat 3.1.3.4 la pertinença $u_{i(k)}$ d'un objecte ω_i a un clúster difús Ω_k:

$$u_{i(k)} = \frac{1}{\sum_{l=1}^{K} \left[\frac{e_{i(k)}}{e_{i(l)}} \right]^{2/(m-1)}}$$

on $m \in (1, \infty)$ és l'exponent de *fuzziness* ("borrositat"). En un model de clústers independents, cal que el valor de tipicalitat per a un clúster donat es calculi sense emprar cap informació relacionada amb altres clústers. En comptes de comparar la distància al centroide del clúster que ens interessa amb la distancia a la resta de centroides, Krishnapuram & Keller feren la comparació emprant una distància de referència (η_k, *reference distance*) que en l'aproximació de *FCM* representaria tot allò que no és el clúster d'interès. Concretament, l'equació de tipicalitat que proposen és:

$$t_{i(k)} = \frac{1}{1 + \left(\frac{e_{i(k)}}{\eta_k} \right)^{2/(m-1)}}$$

on η_k seria l'esmentada distància de referència, $e_{i(k)}$ la distància de l'objecte ω_i al centroide de Ω_k, i m l'exponent de *fuzziness*. La manera en que l'equació de tipicalitat de *PCM* relaciona $e_{i(k)}$ i $t_{i(k)}$, per a diferents valors dels paràmetres, es pot entendre fàcilment mitjançant la figura 3.3.2.

El funcional minimitzat a *PCM* és:

$$PCM_{K,m,\eta} = \sum_{k=1}^{K} \sum_{i=1}^{N} t_{i(k)}^{m} \, e_{i(k)}^{2} + \sum_{k=1}^{K} \eta_k^2 \sum_{i=1}^{N} (1 - t_{i(k)})^m$$

Aquest nou funcional és un sumatori de suma de quadrats ponderada per a tots els clústers. No obstant, de fet cada clúster s'optimitza independentment, mitjançant el funcional:

$$PCM\,(k)_{m,\eta_k} = \sum_{i=1}^{N} t_{i(k)}^{m} \, e_{i(k)}^{2} + \eta_k^2 \sum_{i=1}^{N} (1 - t_{i(k)})^m$$

Tal i com es pot provar fàcilment, igualant a 0 la derivada d'aquest funcional respecte $t_{i(k)}$, podem aïllar la funció $t_{i(k)}$ que minimitza el funcional. Aquesta és, lògicament, la funció tipicalitat presentada.

L'algorisme de classificació de *PCM* és idèntic a *FCM*, amb l'excepció de que la classificació dels objectes es realitza utilitzant la funció de tipicalitat enlloc de la de pertinença. Davé & Krishnapuram (1997) anomenen N/PC1 l'algorisme anàleg que optimitza un clúster cada cop.

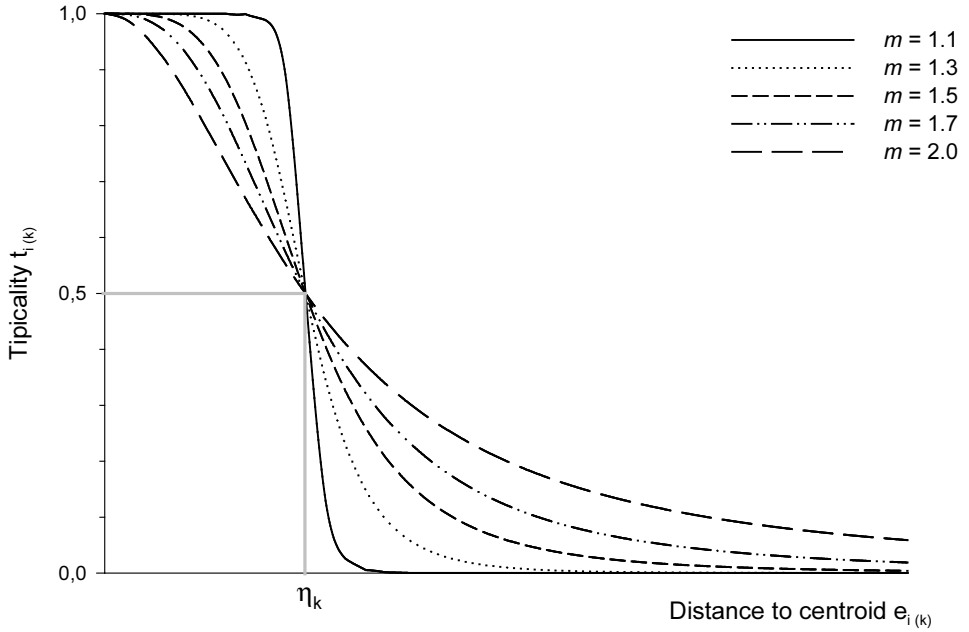


Figura 3.3.2: Corbes distància al centroide *versus* la tipicalitat de *PCM* per a diferents valors de l'exponent de *fuzziness* (*m*), essent la distància de referència constant: $\eta_k = 0.5$

És important tenir present que hom podria definir altres funcions de tipicalitat, associades a la minimització d'altres funcionals. Per exemple, una funció senzilla que compleix les condicions (a-f) és la funció exponencial (Krishnapuram & Keller 1996):

$$t_{i(k)} = \exp\left\{-\frac{e_{i(k)}^2}{\eta_k^2}\right\}$$

i el funcional seria en aquest cas: $J(k)_{m,\eta_k} = \sum_{i=1}^{N} t_{i(k)}\, e_{i(k)}^2 + \eta_k^2 \sum_{i=1}^{N} (t_{i(k)} \log t_{i(k)} - t_{i(k)})$

Com que el funcional de *PCM* s'origina de la modificació de *FCM*, hom pot arribar a pensar que són algorismes semblants. No obstant, tal i com expliquen els autors Krishnapuram & Keller (1996), hi ha diferencies fonamentals entre els dos algorismes, perquè representen models de classificació molt diferents.

PCM fou força debatut en els seus inicis (Barni *et al.* 1996, Krishnapuram & Keller 1996, Pal *et al.* 1997). Concretament, Barni *et al.* (1996) criticaren la tendència de *PCM* a generar solucions coincidents i una exagerada dependència de les condicions inicials. Degut a que el relaxament de la restricció (b), *PCM* té una inestabilitat més gran de les solucions. L'espai de solucions possibles de *PCM* és, de lluny, molt més ampli que a *FCM*.

3.3.1.5 Els paràmetres de *PCM*.

Els dos paràmetres que modulen el comportament de l'algorisme *PCM* són m i η_k. L'exponent de *fuzziness* (m) canvia la forma de la funció de tipicalitat (vegeu figura 3.3.2), mentre que η_k determina la mida del clúster (vegeu figura 3.3.3). Normalment s'utilitza un exponent de *fuzziness* adequat a la quantitat de soroll de les dades, tal i com es fa amb *FCM*. No obstant, el significat de m és diferent en els dos algorismes. Mentre que a l'algorisme *FCM* un increment de m representa incrementar el nombre de punts 'compartits' entre diferents clústers, el mateix increment d'm a *PCM* equival a augmentar la possibilitat de que tots els punts del conjunt de dades pertanyin completament a cada clúster.

El paràmetre de *PCM* més compromès d'estimar és la distància de referència (η_k). La distància de referència és la distància d'un objecte al centre del clúster per la qual la tipicalitat és 0.5. Té un significat anàleg a una amplada de banda, resolució o escala. És un paràmetre que no pot ésser canviat durant la execució de l'algorisme, altrament no està provada la optimització del funcional. Krishnapuram & Keller (1993) proposen dues equacions per a estimar η_k:

$$\eta_k = C \cdot \left(\frac{\sum_{i=1}^{N} t_{i(k)}^m \cdot e_{i(k)}^2}{\sum_{i=1}^{N} t_{i(k)}^m} \right)^{1/2} \quad \text{o bé} \quad \eta_k = \left(\frac{\sum_{x_i \in \Pi_i} e_{i(k)}^2}{\left| (\Pi_i)_\alpha \right|} \right)^{1/2},$$

on C és una constant i $(\Pi_i)_\alpha$ un α-tall del conjunt Π_i (el conjunt *crisp* resultant de la *defuzzification* del conjunt difús a un nivell α de pertinença, vegeu l'apartat 3.1.5.3).

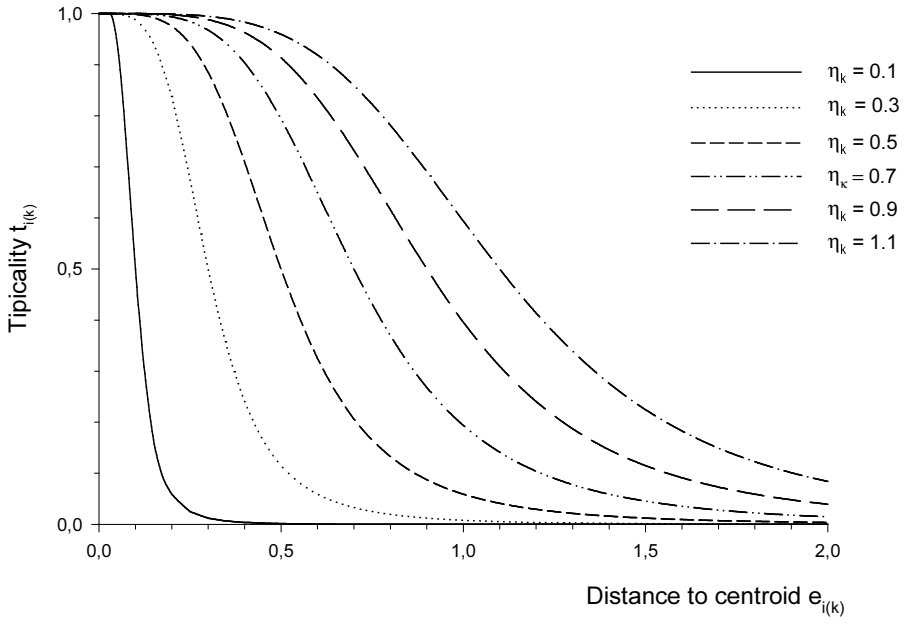


Figura 3.3.3: Corbes de distància al centroide *vs*. la tipicalitat de *PCM* per a diferents valors de la distància de referència η_k, essent l'exponent de *fuzziness* constant, $m = 1.5$.

Per a utilitzar eficientment *PCM*, Krishnapuram & Keller (1993) suggereixen inicialitzar les tipicalitats a partir de la solució de l'algorisme *FCM*, i emprar la primera equació per a estimar el paràmetre η_k. Després d'executar *PCM* caldria emprar la segona equació per tal de refinar la estimació de η_k i executar de nou *PCM*. Noteu que si la constant $C = 1$, la primera equació equival a l'arrel quadrada de la variància (o variabilitat geomètrica) del clúster. Com que conté distàncies elevades al quadrat, aquesta estimació de la distància de referència és sensible als *outliers*, que d'altra banda és possible que contingui la partició difusa no degenerada obtinguda de *FCM*.

3.3.1.6 Extensions de *PCM* i altres algorismes insensibles a *outliers*

Anàlogament al seu predecessor, *PCM* pot ésser estès per a detectar el·lipses i línies. Efectivament, podem emprar en el càlcul de les distàncies als centres, la distància de Mahalanobis escalada, utilitzant la matriu de covariàncies *fuzzy* del grup d'interès, $\mathbf{S}_k$:

$$e^2_{i(k)} = \left|\mathbf{S}_k\right|^{1/P} \cdot (x_i - \bar{x}_k)' \cdot \mathbf{S}_k^{-1} \cdot (x_i - \bar{x}_k)$$

L'ús de la inversa d' $\mathbf{S}_k$ transforma les distàncies en unitats de desviació típica, mentre que l'ús del determinant re-escala el volum original del clúster. Aquesta variant, anomenada *PGK* (Krishnapuram & Keller 1993) és equivalent a una extensió semblant de *FCM*, anomenada *FGK* que proposaren Gustafson & Kessel (1979). *PGK* ha estat emprat en l'anàlisi d'imatges per a detectar segments lineals en aplicacions d'enginyeria (Barni & Gualtieri 1999) o per classificar comunitats d'esculls en imatges multi-espectrals (Andréfouët et *al*. 2003).

Timm *et al*. (2004) han proposat una modificació de *PCM* que intenta evitar la tendència d'aquest algorisme a proporcionar solucions coincidents. Concretament, Timm *et al*. introdueixen un nou terme en el funcional de *PCM* que incorpora la repulsió entre clústers evitant les solucions coincidents, amb la contrapartida, però, d'introduir dos nou paràmetres al funcional.

L'algorisme *PCM* no és l'únic intent que s'ha fet per superar els inconvenients amb els *outliers* dels algorismes *KM* (Cheng & Milligan 1996a, 1996b) o *FCM* (Gruijter & McBratney 1988, Davé 1991) o tots dos (Wu & Yang 2002). Per exemple, Frigui & Krishnapuram (1996) combinaren *FCM* amb estimadors estadístics robustos (Huber 1981) per alleugerir el pes dels objectes *outliers*. Cal notar que el mateix algorisme *PCM* pot ésser considerat una aproximació robusta al *clustering*, ja que equival a optimitzar W-estimadors de prototipus de clústers (Nasraoui & Krishnapuram 1995). D'altra banda, Pal *et al*. (1997) proposen una variant algorísmica intermèdia entre *FCM* i *PCM*, anomenada *FPCM*, que incorpora pertinences probabilístiques (relatives) i possibilístiques (absolutes) en el seu funcional:

$$FPCM\ (k)_{m,\eta_k} = \sum_{i=1}^{N} (u^m_{i(k)} + t^{\eta}_{i(k)})\, e^2_{i(k)}$$

Creiem, però, que aquest algorisme híbrid a la pràctica no difereix gaire de *FCM*.

3.3.2 Aportacions metodològiques per a *PCM*

3.3.2.1 Proposta de correcció de *PCM* per a distàncies acotades

En l'anàlisi multivariant de dades ecològiques és freqüent l'ús de mesures de dissimilaritat amb cota superior, és a dir, amb un valor màxim de distància. Algunes de les mesures de més interès ho són, com per exemple, la distància de la corda, la distància de Hellinger o la distància de Bray-Curtis (vegeu capítol 3.2). Un inconvenient d'aplicar *PCM* en un espai de relacions donat per una mètrica com les anteriors és que la funció de tipicalitat tendeix a zero tan sols de manera asimptòtica. A la cota superior de dissimilaritat, la tipicalitat no és zero, al contrari del que el concepte de dissimilaritat màxima hauria de suggerir.

Per tal de corregir aquest inconvenient proposem, senzillament, re-escalar les tipicalitats en l'interval [0, tip(e_{max})]. Siguin e_{max} la cota superior d'una dissimilaritat i $t_{e_{max}(k)}$ el valor de tipicalitat corresponent a la cota superior:

$$t_{e_{max}(k)} = \frac{1}{1+\left(\frac{e_{max}}{\eta_k}\right)^{2/(m-1)}}$$

Definim, $t^*_{i(k)}$ la tipicalitat corregida, que s'expressa:

$$t^*_{i(k)} = \frac{t_{i(k)} - t_{e_{max}(k)}}{1 - t_{e_{max}(k)}} \quad \text{o bé} \quad t^*_{i(k)} = t_{i(k)} \cdot \left[1 - \left(\frac{e_{i(k)}}{e_{max}}\right)^{2/m-1}\right].$$

Un cop adoptem la correcció per dissimilaritats acotades, l'algorisme, que hem anomenat *PCM**, presenta el funcional (vegeu la demostració a l'apèndix 3.3.A.1 del capítol):

$$PCM^*(k)_{m,\eta_k} = \sum_{i=1}^{N} \alpha_i^m \cdot t^{*m}_{i(k)} \cdot e^2_{i(k)} + \eta_k^2 \sum_{i=1}^{N} \alpha_i^m \cdot \left(\alpha_i^{-1} - t^*_{i(k)}\right)^m ,$$

$$\text{on } \alpha_i = \left[1 - \left(\frac{e_{i(k)}}{e_{max}}\right)^{2/m-1}\right]^{-1}$$

Considerem que el funcional de *PCM** és una generalització del funcional *PCM* original, car per a dissimilaritats no acotades:

$$\lim_{e_{\max}\to\infty}\left\{PCM^*(k)_{m,\eta_k}\right\} = PCM(k)_{m,\eta_k} \text{ , a la vegada que } \lim_{e_{max}\to\infty}\left(t^*_{i(k)}\right) = t_{i(k)} .$$

Hom pot observar a la figura 3.3.4 que correcció proposada fa que, efectivament, per a una distància acotada a e_{max} la tipicalitat corresponent sigui zero.

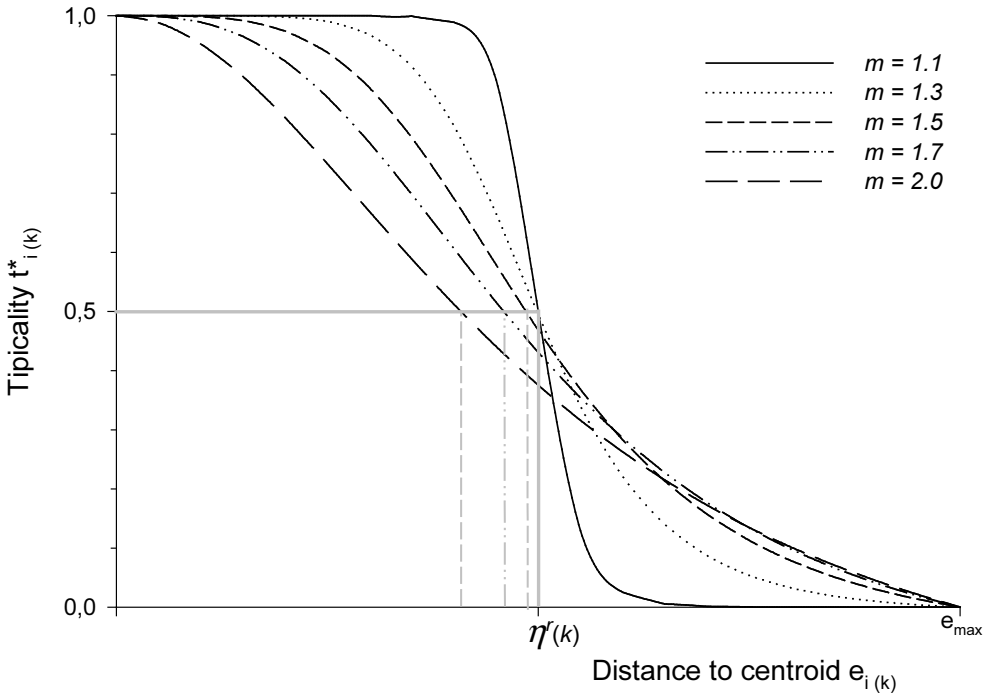


Figura 3.3.4: Corbes distància al centoid ($e_{i(k)}$) vs. tipicalitat després de la correcció per a distàncies acotades ($t^*_{i(k)}$). S'observa una disminució de la distància de referència real en augmentar l'exponent de *fuzziness*.

La transformació presentada provoca, col·lateralment, que la distància per la qual s'obté un valor de tipicalitat 0.5 es vegi desplaçada cap a l'esquerra respecte al valor emprat com a paràmetre (η_k). Per a diferenciar aquesta distància del paràmetre de *PCM* η_k, l'hem anomenada distància de referència real o η_k^r. La deformació observada en la distància de referència real és més acusada com més gran sigui el valor del paràmetre de *fuzziness*. Hom pot calcular la distància de referència real a partir de:

$$\eta_k^r = \eta_k \cdot \left[\frac{2}{1 + t_{e_{\max}(k)}} - 1 \right]^{\frac{m-1}{2}}$$

D'altra banda, el paràmetre η_k que cal emprar com a paràmetre de *PCM** per tal de tenir una distància de referència real desitjada és superior a aquesta:

$$\eta_k = \eta_k^r \cdot e_{\max} \cdot \left[e_{\max}^{2/m-1} - 2 \cdot \eta_k^{r\,2/m-1} \right]^{(1-m)/2}$$

Per exemple, si hom disposa d'una dissimilaritat acotada a 1 i desitja emprar *PCM** amb m=2.0. Per a tenir una tipicalitat de 0.5 als objectes situats a una distància 0.3 ($\eta_k^r = 0.3$), hom necessita situar el paràmetre η_k a:

$$\eta_k = 0.3 \cdot 1 \cdot \left[1^{2/2-1} - 2 \cdot 0.3^{2/2-1} \right]^{(1-2)/2} = 0.3313 .$$

3.3.2.2 Relació entre la distància de referència i la variabilitat d'un clúster.

Per a que l'algorisme *PCM* tingui utilitat com a eina d'anàlisi de clústers, és de vital importància establir distàncies de referència adequades a la mida de cada clúster a detectar. Una distància de referència massa petita impedirà reconèixer el grup en la seva totalitat. Per contra, una distància de referència excessivament gran tendirà a detectar només grans grups i, a l'extrem, el conjunt de dades sencer com a grup. Aquest paràmetre, doncs, ens marca la escala a la que cerquem estructures de grup. Fóra bo, doncs, poder estimar automàticament aquest paràmetre.

Visualment, hom estableix els límits d'un clúster quan observa al voltant una espai buit de punts o amb una densitat menor. Per tant, sembla lògic situar la distància de referència en aquell valor que provoqui una disminució de tipicalitat per a aquells objectes de la regió perifèrica en que la densitat de punts disminueix. La densitat és un concepte relacionat amb el volum. Malauradament, en una població multivariant de distribució desconeguda és difícil determinar el volum amb exactitud. D'altra banda, la variació de la densitat d'un clúster es tradueix en canvis en la seva variància (la dispersió per individu). Per aquest motiu, ens hem proposat estudiar la relació que té la variació de la distància de referència amb la variància d'un clúster ($V(k)$). La variància d'un clúster de *PCM* és la dispersió mitjana per individu:

$$\hat{V}_{fd}(k) = \frac{\sum_{i=1}^{N} t_{i(k)}^{m} \, e_{i(k)}^{2}}{\sum_{i=1}^{N} t_{i(k)}^{m}}$$

Imaginem ara una situació de dos grups com la representada a la figura 3.3.5. El nostre objectiu es determinar una distància de referència idònia per al clúster de l'esquerra. Suposarem que el centroide del clúster està correctament determinat i es manté immòbil a canvis de la distància de referència, la qual establim que inicialment té un valor petit. Si hom augmenta la distància de referència, augmentarà la tipicalitat dels objectes (o entraran nous objectes al clúster, en termes *crisp*). A la vegada, augmentarà la dispersió mitjana per individu (variància) del grup, perquè els nous objectes estaran sempre més allunyats del centroide que els que ja hi havia dins. La funció que relaciona la variància d'un clúster amb la distància de referència serà, doncs, una funció creixent.

Suposem ara que augmentem la distància de referència fins a incloure una regió al voltant del clúster que no té punts. L'augment de la distància de referència farà augmentar la pertinença dels objectes del clúster. No obstant, un cop tots tinguin pertinences altes la variància o dispersió mitjana s'estabilitzarà. En aquesta regió la variància esdevindrà asímoptòticament plana. Un cop haguem sobrepassat aquesta regió absent en punts i torni a augmentar la densitat l'augment de la distància de referència farà créixer de nou la variància del clúster.

A la part inferior de la figura 3.3.5 apareix una gràfica amb la variació de $V(k)$ en augmentar la distància de referència, tal i com hem descrit en el paràgraf anterior. Sabent que la derivada d'una funció en un punt ens dóna la pendent de la corba en aquell punt, podem deduir que un mínim en la derivada de $V(k)$ respecte η_k equival a una regió pobre en punts (poc densa). Serà, doncs, una bona regió per establir el punt de tall entre un grup i el seu exterior. Seguint aquest criteri, només ens resta ara calcular la derivada analítica $\delta V(k)/\delta\eta_k$.

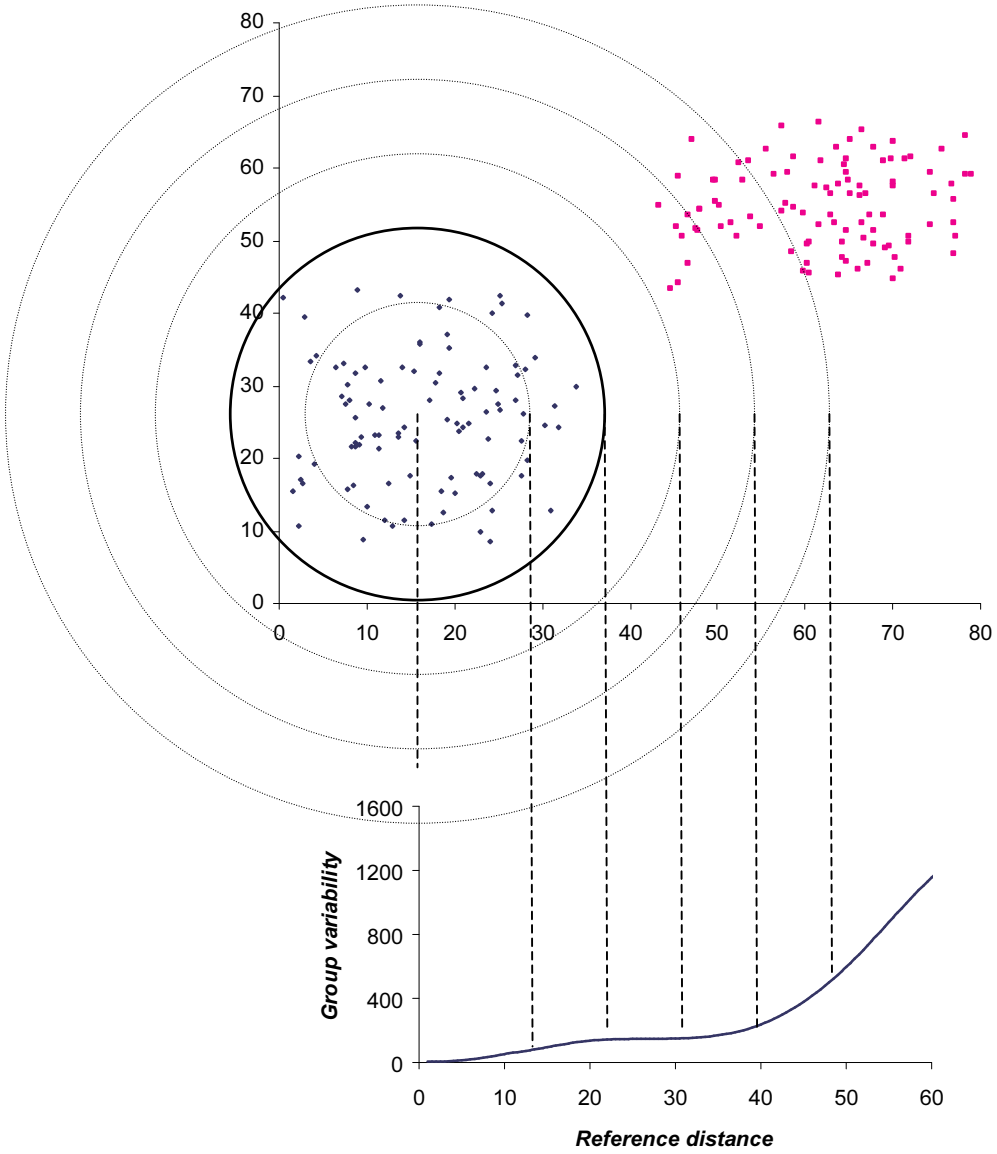


Figura 3.3.5: Exemple de dos clústers en dues dimensions. Evolució de la variabilitat del clúster (*group variability*) de l'esquerra en augmentar la distància de referència (*reference distance*).

Si, com hem suposat anteriorment, considerem el centroide com a fix, la distància de qualsevol dels objectes al centroide és constant. Per tant, podem tractar-la com a tal en el procés de derivació. D'altra banda, la tipicalitat no és constant, sinó que cal derivar la funció de tipicalitat de *PCM*. La derivada $\delta V(k)/\delta\eta_k$, obtinguda analíticament, és:

$$\frac{\partial \hat{V}_{fd}(k)}{\partial \eta_k} = \left(\frac{2\cdot m}{m-1}\right)\cdot \eta_k^{-1} \cdot \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} e_{i(k)}^2 \cdot t_{i(k)}^m \cdot t_{j(k)}^m \cdot \left(t_{i(k)} \cdot \left(\frac{e_{i(k)}}{\eta_k}\right)^{2/(m-1)} - t_{j(k)} \cdot \left(\frac{e_{j(k)}}{\eta_k}\right)^{2/(m-1)} \right)}{\sum_{i=1}^{N}\sum_{j=1}^{N} t_{i(k)}^m \cdot t_{j(k)}^m}$$

Aquesta expressió ens permet calcular per a qualsevol configuració de grup i η_k, el valor de la pendent de canvi de $V(k)$. L'equació anàloga per al cas de *PCM** és:

$$\frac{\partial \hat{V}^*_{fd}(k)}{\partial \eta_k} = \left(\frac{2\cdot m}{m-1}\right)\cdot \eta_k^{-1} \cdot \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} e_{i(k)}^2 \cdot t_{i(k)}^{*m} \cdot t_{j(k)}^{*m} \cdot \left(t_{i(k)} \cdot \left(\frac{e_{i(k)}}{\eta_k}\right)^{2/(m-1)} - t_{j(k)} \cdot \left(\frac{e_{j(k)}}{\eta_k}\right)^{2/(m-1)} \right)}{\sum_{i=1}^{N}\sum_{j=1}^{N} t_{i(k)}^{*m} \cdot t_{j(k)}^{*m}}$$

La demostració complerta d'aquests resultats es troba a l'apèndix 3.3.A.2 d'aquest capítol.

Volem, a continuació, dissenyar un algorisme que ens permeti proporcionar bones estimacions de la distància de referència a *PCM*. Per a una distància de referència inicial, disposarem d'una configuració de clúster estable segons *PCM*. Avaluarem en aquest moment $\delta V(k)/\delta\eta_k$. Com conèixer si aquest valor és un mínim o no? Necessitem comparar-lo amb valors de la derivada a distàncies de referència properes. Gràcies a disposar d'una expressió analítica, podem variar lleugerament el paràmetre η_k, calcular les tipicalitats, suposant el centroide fix, i tornar a avaluar la derivada, sense necessitat d'executar de nou *PCM*. Els valors de la derivada al voltant de la η_k inicial ens indicaran la direcció en que la que hem de moure η_k per a trobar el mínim. Si la funció creix en augmentar η_k, cal que ens moguem disminuint η_k. En canvi, si la funció decreix és necessari augmentar η_k. Aquest raonament es tradueix a estudiar la pendent de $\delta V(k)/\delta\eta_k$. És a dir, a estudiar la segona derivada, la funció analítica que ens proporciona la pendent de la primera derivada. Si aquesta és positiva caldrà disminuir η_k i si és negativa, caldrà augmentar el paràmetre. La derivada segona de la variabilitat geomètrica respecte a la distància de referència és (la demostració d'aquests resultats es poden trobar a l'apèndix 3.3.A.3):

$$\frac{\partial^2 \hat{V}_{fd}(k)}{\partial^2 \eta_k} = \frac{m\cdot\sum_{i=1}^{N}\sum_{j=1}^{N} e_{i(k)}^2 \cdot t_{i(k)}^m \cdot t_{j(k)}^m \cdot \left[(m+1)\cdot\left[\left(\alpha_{i(k)}^2 - \frac{\alpha_{i(k)}}{(m-1)\cdot\eta_k}\right) - \left(\alpha_{j(k)}^2 - \frac{\alpha_{j(k)}}{(m-1)\cdot\eta_k}\right)\right] - (\alpha_{i(k)} - \alpha_{j(k)})\cdot 2\cdot \frac{\sum_{l=1}^{N} m\cdot t_{l(k)}^m \cdot \alpha_{l(k)}}{\sum_{l=1}^{N} t_{l(k)}^m} \right]}{\left(\sum_{i=1}^{N} t_{l(k)}^m\right)^2}$$

i la corresponent expressió per a *PCM** acotat és:

$$\frac{\partial^2 \hat{V}_{fd}(k)}{\partial^2 \eta_k} = \frac{m \cdot \sum_{i=1}^{N}\sum_{j=1}^{N} e_{i(k)}^2 \cdot t_{i(k)}^{*m} \cdot t_{j(k)}^{*m} \cdot \left[(m+1) \cdot \left[\left(\alpha_{i(k)}^2 - \frac{\alpha_{i(k)}}{(m-1)\cdot \eta_k} \right) - \left(\alpha_{j(k)}^2 - \frac{\alpha_{j(k)}}{(m-1)\cdot \eta_k} \right) \right] - \left(\alpha_{i(k)} - \alpha_{j(k)} \right) \cdot 2 \cdot \frac{\sum_{l=1}^{N} m \cdot t_{l(k)}^{*m} \cdot \alpha_{l(k)}}{\sum_{l=1}^{N} t_{l(k)}^{*m}} \right]}{\left(\sum_{i=1}^{N} t_{l(k)}^{*m} \right)^2}$$

en ambdós casos, $\alpha_{i(k)} = t_{i(k)}^{-1} \cdot \frac{\partial t_{i(k)}}{\partial \eta_k} = t_{i(k)}^{*-1} \cdot \frac{\partial t_{i(k)}^{*}}{\partial \eta_k} = \left(\frac{2}{m-1} \right) \cdot \eta_k^{-1} \cdot t_{i(k)} \cdot \left(\frac{e_{i(k)}}{\eta_k} \right)^{2/(m-1)}$.

La figura 3.3.6 mostra la relació entre *V(k)* i les seves dues derivades respecte η_k. En aquest cas, es tracta d'un exemple de dades reals: l'associació *Gentiano-Potentilletum* de l'ordre *Brometalia erecti* en un espai de dissimilaritat de la distancia de Bray-Curtis, acotat en l'interval [0,1]. Hem escollit aquest sintàxon perquè sabem del capítol 2.3 que és un sintàxon de base ben aïllat.

La primera derivada de *V(k)* és més variable per a distàncies petites. Això es degut a que ens apropem a l'escala dels individus i la aparició de nous individus provoca canvis bruscs de densitat i variància. Per a escales més grans els canvis són raonablement més suaus. Els mínims de la primera derivada de la variància són els punts que ens interessen perquè es corresponen amb una disminució de la densitat de punts (punts d'inflexió de *V(k)*) i indiquen clústers a diferents escales. Com més propera a zero sigui la primera derivada, menys densitat de punts hi haurà a la regió límit del clúster d'interès i, per tant, més ben aïllat estarà el grup.

A la figura 3.3.6, el rang de valors d'η_k entre dos màxims de la primera derivada ve marcat per fletxes entre les línies verticals contínues. Si partim d'una distància de referència qualsevol en aquest interval i avaluem la derivada segona sabrem, pel seu signe, si el mínim es troba en valors més petits o més grans que la distància de referència inicial i ens mourem en el sentit de les fletxes. Independentment del valor que prenguem d'inici per a η_k, sempre convergirem a la mateixa solució (el mateix mínim). Si augmentem η_k fins a sortir d'aquest interval, arribaríem a l'àrea d'influència d'un altre clúster, a escala més gran i que contindria *Gentiano-Potentilletum*. Com més ample sigui l'espai entre els dos màxims de la primera derivada, més entitat tindrà el nostre clúster perquè serà "identificable" a un nombre major d'escales. Hem anomenat aquest rang el temps de vida (*lifetime*) del clúster. Aquest terme fou emprat per primer cop per Ling (1973) en referència a clústers d'arbres ultramètrics. Ling definí el temps de vida d'un clúster com la diferència entre els rangs del valor més baix de dissimilaritat al que un clúster existeix i el valor de dissimilaritat al que és incorporat en un clúster de mida superior Gordon (1994). En general es fa servir aquest terme en referència a les escales de validesa de clústers individuals o de particions (Nakamura & Kehtarnavaz 1998).

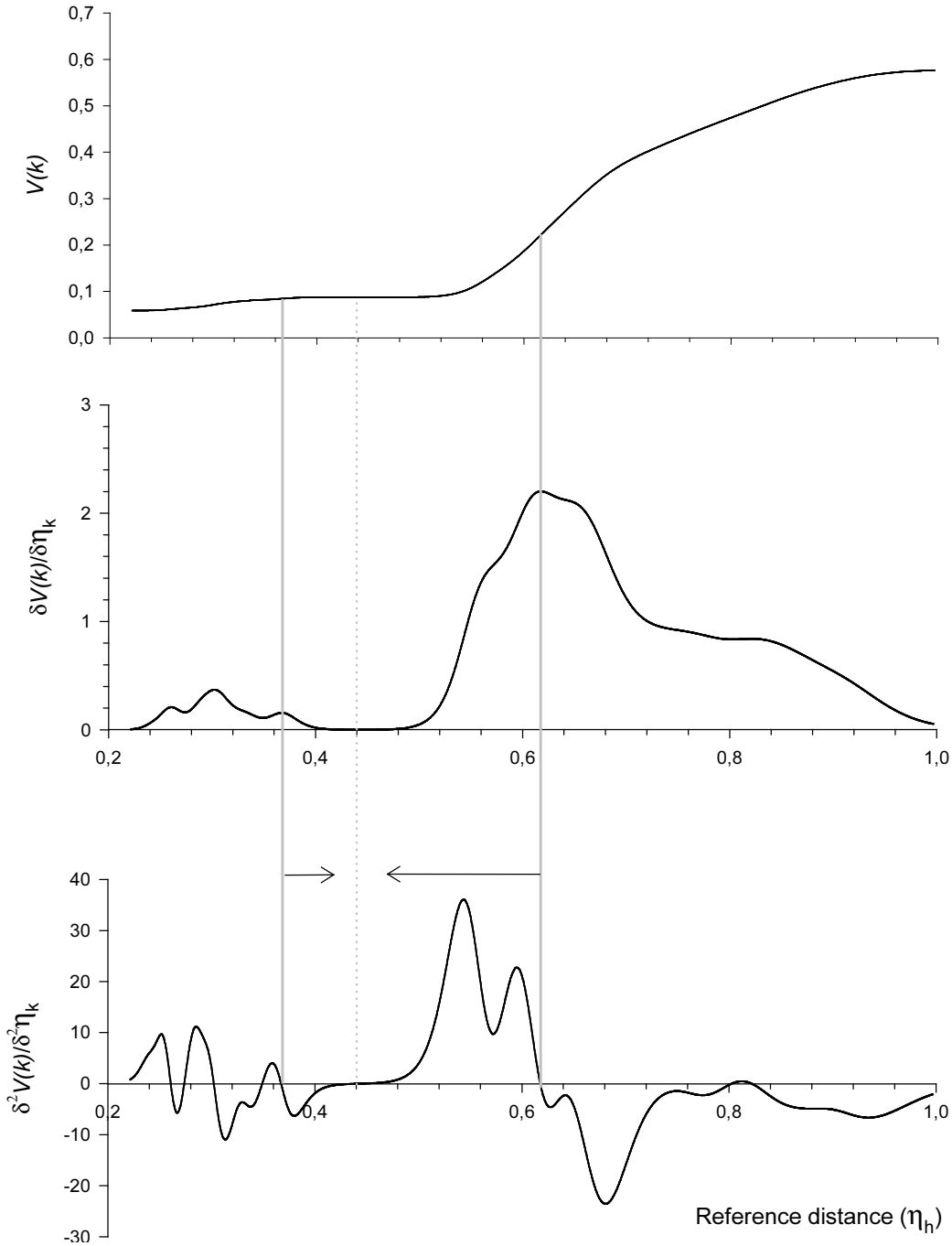


Figura 3.3.6: Evolució de la variància del clúster (a dalt), i les derivades primera (al mig) i segona (a baix), per a diferents valors de la distància de referència. Les línies verticals contínues marquen els dos màxims de la primera derivada de *V(k)* que envolten el mínim, indicat per la línia discontínua. Les fletxes horitzontals indiquen el moviment a prendre en avaluar les derivades en punts dins de l'interval marcat pels màxims (veure text).

3.3.2.3 Algorisme *meta-PCM*

La discussió del darrer apartat ens posa en condicions de proposar un senzill meta-algorisme per a detectar un clúster amb *PCM* (o *PCM**) establint, a la vegada, la distància de referència idònia del grup. Essent estrictes, caldria anomenar *N/PC1* (Davé & Krishnapuram 1997) a l'algorisme *PCM* per a un sol clúster. No obstant, i per simplicitat, utilitzarem l'acrònim *PCM* per a referir-nos a *N/PC1*.

Al quadre 3.3.1 descrivim aquest meta-algorisme, anomenat *meta-PCM*. L'execució de *PCM* i el canvi de la distància de referència s'han de realitzar alternadament i no a la vegada, perquè la convergència de *PCM* només està provada quan η_k es manté constant.

Quadre 3.3.1: Algorisme *meta-PCM*.

Partim d'una configuració inicial del clúster, una estimació de η_k, i uns valors constants per als paràmetres *prec.*-η_k i *MinCard*. En cas de que la configuració inicial de clúster estigui sense optimitzar per *meta-PCM* començarem a(1). Si el clúster del que partim ja ha estat optimitzat anteriorment per *meta-PCM* començarem fent-lo créixer a (5).

1. Executem l'algorisme *PCM* (o *PCM** si la mètrica és acotada).
2. Cerquem la distància de referència més propera en que $\delta^2V(k)/\delta^2\eta_k$ s'anul·la. Assignem aquest valor a *new*-η_k.
3. Si la cardinalitat a η_k o *new*-η_k és propera al nombre d'objectes total ens aturem perquè hem arribat al clúster representat per tots els objectes.
4. Si *new*-η_k és igual o inferior a zero, o la cardinalitat a *new*-η_k és inferior a *MinCard* anem a (5) per tal de fer créixer el clúster i superar el màxim de la primera derivada. En cas contrari anem a (9).
5. Guardem el valor de $\delta^2V(k)/\delta^2\eta_k$.
6. Incrementem η_k en *prec*-η_k.
7. Executem *PCM* (o *PCM** si la mètrica és acotada).
8. Avaluem $\delta^2V(k)/\delta^2\eta_k$. Si el valor és negatiu i el valor guardat era positiu anem a (2), en cas contrari tornem a (5).
9. Si la diferència entre η_k i *new*-η_k és menor que *prec.*-η_k/100, el clúster es considera optimitzat i passem a (10). En cas contrari, situem η_k a *new*-η_k i tornem a (1).
10. Emprant la funció $\delta^2V(k)/\delta^2\eta_k$ cerquem els màxims de la primera derivada a esquerra i dreta del mínim trobat per tal d'establir el temps de vida del grup trobat.

L'estratègia per a realitzar una anàlisi d'exploració de clústers a diverses escales seria la següent: En primer lloc, hom determinaria clústers a petita escala executant *meta-PCM* a partir de llavors inicials. A continuació, per a cercar clústers en escales superiors caldria executar de nou *meta-PCM*, aquest cop a partir dels clústers que ja havien estats optimitzats. Com que un clúster prèviament optimitzat ja es troba en un mínim de $\delta V(k)/\delta\eta_k$, és necessari, per tant, incrementar η_k per tal de passar el màxim que es troba a la dreta del mínim (passos 5-8), i quan la derivada torna a ésser decreixent (la segona derivada negativa) tornar al pas (2). Cal notar que els clústers trobats a partir de diferents llavors poden convergir a clústers iguals. Per aquest motiu, cal comprovar per a cada clúster trobat ja havia estat identificat anteriorment i en cas

afirmatiu descartar-lo. Aquesta estratègia d'exploració de dades amb l'algorisme *meta-PCM* es troba implementada al programa *GINKGO*, que descrivim al capítol 4.2.

Un cop explorades les dades a diverses escales es planteja el problema de representar aquesta estructura "multi-clúster" de manera intel·ligible. La manera més senzilla seria a través d'un 'arbre de clústers'. L'inconvenient d'una estructura jeràrquica és que els clústers no poden superposar-se entre ells o compartir nodes de clústers "fills". Com que l'exploració de les dades es susceptible de produir situacions de superposició és, doncs, necessari disposar d'algun criteri de decisió que permeti escollir un o altre clúster en aquests casos. Aquest criteri pot consistir en eliminar aquells clústers que no compleixen un determinat criteri de mida o de temps de vida. Tanmateix, per a dades complexes cal reconèixer que la decisió pot ser difícil.

Per a acabar aquesta apartat, resumim les principals propostes que hem fet fins ara:

- Disposem d'un criteri per decidir, en cada cas, els valors més adequats per a la distància de referència, un dels dos paràmetres de *PCM*. El criteri es basa en la cerca de valors mínims de derivada parcial variabilitat del clúster respecte a aquest paràmetre. D'aquesta manera, es redueixen enormement el nombre de possibilitats paramètriques a avaluar.
- Aquest criteri ens ha permès definir un meta-algorisme de classificació basat en la iteració de la estima de η_k i l'execució de *PCM*.
- La distància entre dos màxims de la derivada ens indica el rang escales en que el clúster té validesa i informa a la vegada la facilitat amb que pot ésser trobat. Hem anomenat aquest valor temps de vida (*lifetime*).
- El temps de vida o la mida dels clústers poden ser criteris per a eliminar *a posteriori* alguns dels clústers trobats, o escollir entre clústers que presenten superposició.

3.3.3 Aplicació de *meta-PCM* a la classificació de comunitats de vegetals

3.3.3.1 Objectius

En aquesta secció ens proposem estudiar el funcionament de l'algorisme *PCM* amb les modificacions metodològiques proposades a la secció anterior (3.3.2). L'objectiu és determinar quines avantatges i quins inconvenients presenta aquesta estratègia de *clustering* en relació als mètodes d'anàlisi de clústers partitius que tractarem al capítol 3.1.

3.3.3.2 Dades i construcció de l'espai de relacions

Les dades a estudiar són, com en capítols precedents:

A. Les dades de Bowman & Wilson (1984): 41 inventaris i 33 tàxons de la plana al·luvial del riu Adelaide a Austràlia, utilitzades en els estudis de Dale (1988a i 1998b). Hom pot trobar la matriu de dades a la taula 3.1.1. L'escala de valors és compresa en l'interval $[0,6]$. No utilitzarem aquí la comparació amb la classificació *crisp* que Dale (1988a) proposa sobre aquest conjunt de dades.
B. Els inventaris de *Xerobromion erecti*: 248 inventaris i 548 tàxons de comunitats pratenses xeròfiles montanes. Els valors de l'escala ordinal de Braun-Blanquet han estat transformats a l'escala combinada de van der Maarel (1979). La classificació tradicional consta de 13 sintàxons de base (5 associacions i 8 subassociacions).
C. Els inventaris de *Quercetea ilicis* sense *Quercenion ilicis*: 376 inventaris i 493 tàxons de bosquines i matollars mediterranis. Els valors de l'escala de Braun-Blanquet han estat transformats a l'escala combinada de van der Maarel (1979). La classificació tradicional consta de 16 sintàxons de base (8 associacions i 8 subassociacions).

La transformació escalar de les dades utilitzada segueix la equació $y = x^w$ (van der Maarel 1979, Currall 1987). Per a cada conjunt de dades s'ha emprat un exponent de transformació (w) diferent:

A. w=2.0 (CHO-1/HELL-2)
B. w=1.0 (CHO-0.5/HELL-1)
C. w=1.5 (CHO-0.75/HELL-1.5)

La mesura de proximitat entre inventaris emprada ha estat la de Hellinger (equivalent a la Corda per a transformacions escalars $w/2$). La selecció d'aquests exponents de transformació i aquesta mesura de proximitat sorgeix de les conclusions obtingudes en el capítol precedent (3.2) pel que fa a la capacitat de detecció d'estructures dels diferents espais de relacions.

3.3.3.3 Metodologia de *clustering* i avaluació dels resultats

Hem utilitzat l'algorisme *meta-PCM* per a explorar els espais de relacions A, B i C. En tots els casos els paràmetres *prec.*-η_k i *minCard* eren fixos (*prec.*-η_k=0.005 i *minCard*=3).

Com que la distància de Hellinger és una mètrica amb una cota superior (e_{max}=1.4142), s'ha utilitzat la correcció per mètriques acotades.

Per al conjunt de dades A, l'exponent de *fuzziness* s'ha fet variar des de m=1.05 fins a m=1.20 per tal de fer palesa la sensibilitat de l'algorisme a aquest paràmetre. Als espais B i C, l'exponent de *fuzziness* ha estat m=1.05.

Com a configuració inicial dels clústers hem emprat tríades d'inventaris. Per a cada inventari s'hem cercat els dos inventaris més propers. Junts configuren una triada d'inventaris propers. Inicialment hem definit tantes tríades com inventaris però hem eliminat aquelles tríades repetides.

La distància de referència inicial és un paràmetre important. Si aquesta és massa baixa la primera execució de *PCM* pot tendir fàcilment acabar produint un clúster d'1, 2 o 3 individus (inventaris). Els clústers tant petits són molt bruscos en els canvis i per sortir del seu mínim de $\delta V(k)/\delta\eta_k$ cal augmentar molt la distància de referència, fent que es passi d'aquests 'micro-clústers' a clústers massa grans. La solució que hem trobat eficient *ad hoc*, és la de calcular aquella distància de referència que, emprant distàncies corregides per *leave-one-out*, proporciona un clúster de mida (cardinalitat) no inferior a tres.

Meta-PCM produeix, com a resultat, clústers de diferents escales. Malgrat representar regions denses de les dades, no tots ells tenen la mateixa validesa. Hem sotmès els clústers trobats a un filtrat en el s'han eliminat aquells clústers amb un temps de vida inferior a 0.2 o que tenien un mida (cardinalitat) inferior a quatre. Abans i després d'aquest filtrat s'ha avaluat la superposició de clústers. Hem considerat que un clúster presentava encavalcament amb un altre si almenys el 15% del grup pertanyia també a l'altre clúster i no es tractava d'un clúster 'pare'. Noteu que aquesta definició de superposició és asimètrica. Per tant, cal avaluar el superposició d'una parella de clústers en els dos sentits. Addicionalment, i per tal de poder comparar els resultats de *meta-PCM* amb la sintaxonomia tradicional, per als conjunts de dades B i C hem realitzat un filtrat manual de clústers. Ha estat necessari eliminar completament els casos de superposició i decidir entre utilitzar en la comparació els clústers pares o els clústers fills en casos d'anidament. Els criteris seguits han estat els de maximització del temps de vida o de la cardinalitat.

Un cop s'ha disposat d'un grup d'inventaris sense superposició ni "anidament" cal transformar els clústers *PCM* a una partició que pugui ésser comparada mitjançant mesures d'acord entre particions (Matsakis *et al*. 2000). Concretament, hem generat particions seguint el següent protocol de comparació:

1) Com que l'algorisme *PCM* exclou *outliers*, s'han seleccionat aquells inventaris que tenien una tipicalitat a un dels grups igual o més gran que diferents valors: 0.1, 0.05, 0.01, 0.005, 0.001 i 0.0. Com més petit és aquest llindar de tipicalitat, més gran és el nombre d'inventaris per als quals compararem la classificació de *meta-PCM* amb la tradicional. El darrer cas és aquell en que s'han seleccionat tots els inventaris.
2) Per a cada grup d'inventaris seleccionat hem generat dues particions:
    i. Hem determinat, per a cada inventari, el clúster amb tipicalitat màxima (criteri possibilistic: *PCM*).
    ii. Hem determinat, per a cada inventari, el clúster més proper (criteri probabilístic: *KM/FCM*).
3) Hem comparat les particions 12 (6x2) amb la partició derivada de la sintaxonomia mitjançant l'índex de Rand (1971) corregit per l'atzar (Hubert & Arabie, 1985) i el coeficient de correlació phi (Φ).

3.3.3.4 Resultats

La taula 3.3.2 mostra el nombre de estructures trobades per l'algorisme *meta-PCM* sobre els espais de dades A, B i C. Per al conjunt de dades A (Bowman & Wilson) hem executat l'algorisme emprant diferents coeficients de *fuzziness* (*m*). La taula fa palesa la dependència del nombre i qualitat dels clústers produïts en relació a aquest paràmetre. En augmentar l'exponent de *fuzziness* el nombre d'estructures trobades disminueix. D'altra banda, la validesa de les estructures trobades, valorada pel temps de vida, és més gran, i la superposició entre elles disminueix.

Hem escollit la solució obtinguda amb m=1.12 per a representar-la sobre els eixos d'una anàlisi de coordenades principals. Els diagrames de dispersió de la figura 3.3.7 es poden comparar amb els presentats a la figura 3.1.6 del capítol 3.1. A diferència del model partitiu, amb *PCM* alguns inventaris tenen tipicalitats nul·les o gairebé nul·les per a tots els clústers. Per tant, hom pot considerar que resten sense classificar. És el cas dels inventaris 3, 7, 9 i 10, inclosos al primer grup de la classificació de Dale (1988a). Aquests inventaris són els que tenen valors més baixos d'un dels tàxons més fidels del grup i que, a més, és el més abundant: 27.*Oryza* sp. La resta d'inventaris "no classificats" (26,27,28, 29, 30 i 32) pertanyen tots al tercer grup , segons la

classificació de Dale. Ja vam determinar, al capítol 3.1, que aquest era el grup amb més problemes de classificació.

Per acabar els comentaris sobre els resultats de *meta-PCM* en aquest espai de dades, voldríem fer esment que si l'aplicació de l'algorisme es realitza amb un exponent de transformació escalar més baix (per exemple w=0.5 o w=1.0, resultats no mostrats), el tercer grup de Dale es torna massa dispers i no es possible establir-hi un clúster de *meta-PCM*. Aquest resultat està en acord amb la figura 3.2.6.A del capítol precedent, en que per a les mesures de proximitat de la corda i Hellinger, el nombre de clústers indicat era tres quan l'exponent de transformació era baix.

	N	Triades	*m*	Clústers trobats	Casos de superposició entre trobats	Clústers LT>0.2	Clústers LT>0.2 i card>4	Casos de superposició finals
A. Bowman & Wilson	41	33	1.05	23	37	13	6	5
			1.06	21	28	9	7	5
			1.07	17	19	8	8	5
			1.08	14	11	8	8	4
			1.09	12	3	7	7	2
			1.10	6	0	5	5	0
			1.11	5	0	5	5	0
			1.12	**5**	**0**	**5**	**5**	**0**
			1.13	6	1	6	6	0
			1.14	5	0	5	5	0
			1.15	5	0	5	5	0
			1.16	5	0	5	5	0
			1.17	4	0	4	4	0
			1.18	3	0	3	3	0
			1.19	3	0	3	3	0
			1.20	2	0	2	2	0
B.*Xerobromion erecti*	248	213	1.05	64	110	45	25	13
C.*Quercetea* sense *Quercenion*	376	324	1.05	106	284	55	37	22

Taula 3.3.2: Execucions de *meta-PCM* sobre les dades A, B i C. S'indiquen: el nombre d'inventaris inicials (N), el nombre de tríades d'inventaris emprades com a llavors de *clústers*, el paràmetre de *fuzziness* (m), el nombre de clústers trobats per l'algorisme, el nombre de casos de superposició calculats sobre els clústers trobats, el nombre de clústers amb temps de vida més gran que 0.2, el nombre de clústers amb temps de vida més gran que 0.2 i mida (cardinalitat) més gran que 4, i el nombre de casos de superposició en aquest darrer grup de clústers. Els resultats de la classificació amb m=1.12 es mostren a la figura 3.3.6.

En l'exploració dels espais de dades B i C hem utilitzat *meta-PCM* amb un exponent de *fuzziness* baix, m=1.05, degut a que la complexitat de les dades és més gran que per A. La contrapartida és que el nombre de clústers trobats és força elevat (vegeu taula 3.3.2). També són més nombrosos els casos de superposició entre clústers. A partir de la eliminació d'aquells clústers amb cardinalitat baixa o temps de vida baix hem seleccionat aquells clústers més vàlids.

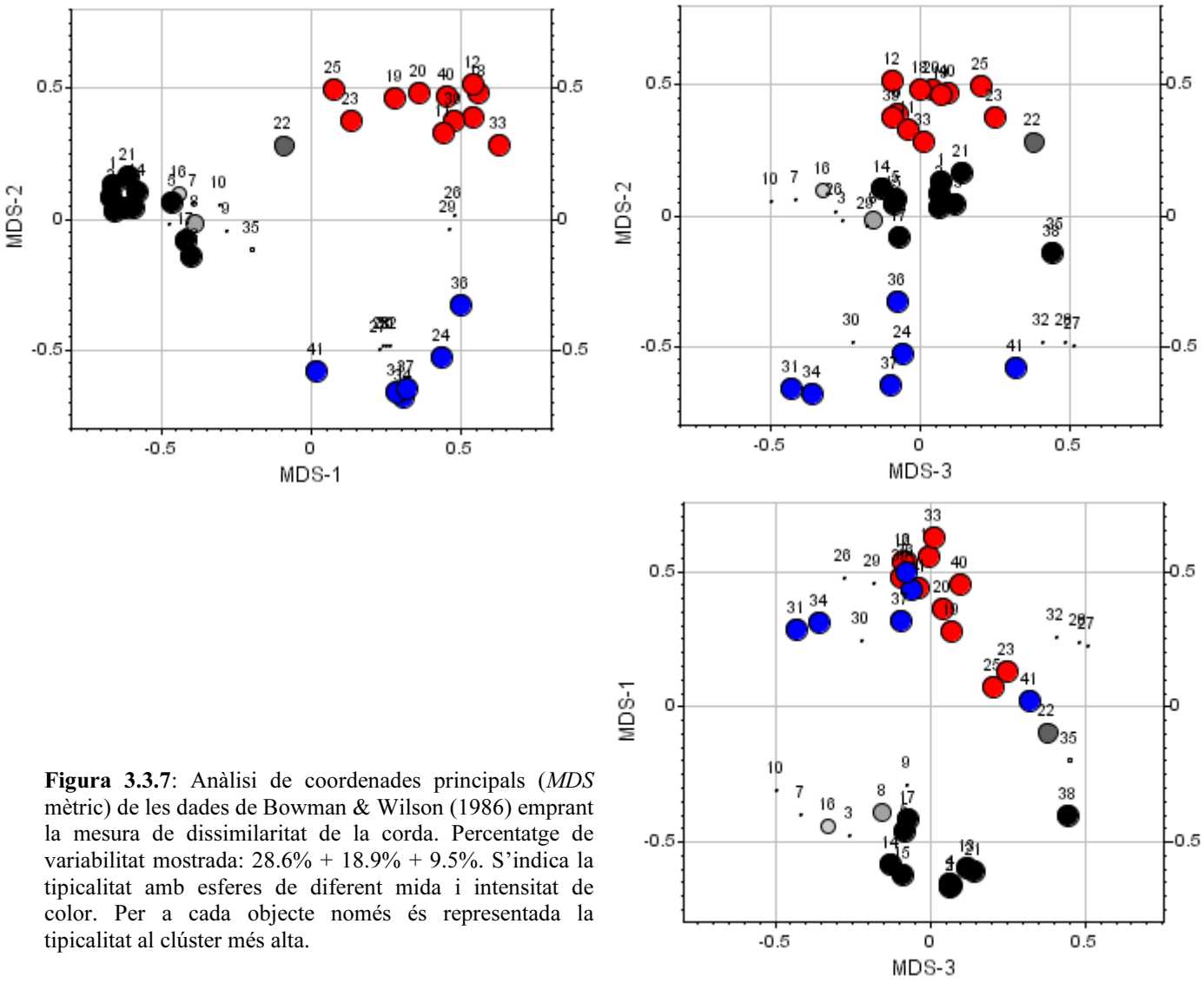


Figura 3.3.7: Anàlisi de coordenades principals (*MDS* mètric) de les dades de Bowman & Wilson (1986) emprant la mesura de dissimilaritat de la corda. Percentatge de variabilitat mostrada: 28.6% + 18.9% + 9.5%. S'indica la tipicalitat amb esferes de diferent mida i intensitat de color. Per a cada objecte només és representada la tipicalitat al clúster més alta.

Els atributs dels clústers determinats com a vàlids, i la seva correlació Φ amb els sintàxons de base de la classificació tradicional es mostren a les taules 3.3.3.B i 3.3.C. Els clústers de *meta-PCM* s'anomenen amb el número de l'inventari a partir del qual es construí la tríada original i el nombre d'execucions de l'algorisme que s'han dut a terme.

Els clústers que *meta-PCM* identifica a *Xerobromion erecti* (B) són sovint subgrups dels sintàxons que defineix la sintaxonomia tradicional. A vegades s'identifiquen diversos clústers, amb relacions d'anidament entre ells, que corresponen tots a subgrups d'un mateix sintàxon. És el cas de TBF (*Teucrio-Brometum festucetosum fallacis*), identificat pels grups G88R3, G88R1 i G88R2, els quals es corresponen a les successives execucions de *meta-PCM* sobre el primer clúster trobat (G88R1). En el cas d'aquest sintàxon, el darrer clúster (G88R3) es correspon totalment al sintàxon i, a la vegada, és el que presenta un major temps de vida dels tres. Per al sintàxon de base AB (*Adonido-Brometum erecti*) també hi ha dos clústers anidats, G29R1 i G31R1. De nou, el clúster que presenta un temps de vida més gran també és el que més correlació presenta amb el sintàxon de base. Per tant, el criteri de selecció del clúster amb valors

més grans per aquest estadístic sembla una bona estratègia per als casos d'"anidament". D'altra banda, per al sintàxon TAV succeeix a la inversa, tot i que els temps de vida dels dos clústers, G164R1 i G164R2, no són massa dispars.

Cal destacar també la tendència de *meta-PCM* a unir en un sol clúster les dues subassociacions de l'*Irido-Brometum* (IBL i IBT). El mateix succeeix per les subassociacions del *Koelerio-Avenuletum* (KAA i KAT). D'entre les subassociacions de *Teucrio-Brometum* tan sols TBF és identificat, tot i que TBH apareixia inicialment en un clúster que fou descartat per tenir un temps de vida de 0.19. En canvi, per a AD apareixen dos grups diferents i no "anidats" (cadascun doblat: G120R1/G120R2 i G103R1/G104R1). El grup G120R1/G120R2 inclou els inventaris de l'estudi de Soriano (1992) a la serra del Moixeró, mentre que l'altre grup inclou els inventaris inicialment utilitzats per a la descripció de l'associació (vegeu l'annex A de la memòria).

	RD	Card	VG	dVG	LT	LB	AB	KG	TBT	TBF	TBH	AD	KAT	KAA	TAV	IBT	IBL	TF	TB (s.l.)	KA (s.l.)	IB (s.l.)
G53R1	0.639	5.987	0.282	0.012	0.330			**0.547**													
G125R1	0.708	9.733	0.393	0.194	0.315								0.288	0.306						0.433	
G120R1	0.711	4.992	0.269	0.045	0.295							0.449									
G184R1	0.670	4.963	0.339	0.022	0.290									0.085		0.408	0.164			0.006	0.429
G120R2	0.759	5.987	0.315	0.058	0.290							0.493									
G103R1	0.655	7.919	0.321	0.057	0.285							**0.573**									
G164R1	0.646	6.986	0.288	0.010	0.280										**0.725**						
G88R3	0.881	12.96	0.478	0.077	0.275					**1.000**									0.462		
G104R1	0.695	8.946	0.337	0.067	0.270							**0.608**									
G208R1	0.776	6.948	0.388	0.156	0.265													0.490			
G20R1	0.700	6.925	0.380	0.029	0.250									0.060		0.338	0.445				**0.549**
G123R1	0.680	6.915	0.343	0.095	0.240								0.154	0.371						0.359	
G164R2	0.772	11.93	0.391	0.130	0.240										**0.962**						
G86R1	0.730	7.957	0.364	0.084	0.235					**0.779**									0.360		
G8R1	0.858	115.9	0.650	0.679	0.225		0.309	0.090	0.255		0.236						0.060		0.166		
G29R1	0.736	13.35	0.434	0.400	0.220		**0.816**														
G132R1	0.704	6.883	0.372	0.142	0.220								0.453							0.359	
G147R1	0.675	4.955	0.326	0.073	0.220								0.381							0.301	
G7R1	0.638	4.950	0.312	0.029	0.215	0.373															
G9R1	0.670	4.958	0.322	0.068	0.210	0.114			0.219										0.142		
G91R1	0.671	5.967	0.317	0.048	0.210					**0.671**									0.310		
G144R1	0.641	4.976	0.296	0.022	0.210								0.119	0.325						0.301	
G46R1	0.689	6.956	0.310	0.123	0.205			**0.594**													
G128R1	0.678	4.949	0.333	0.090	0.205								0.381							0.301	
G31R1	0.687	6.941	0.337	0.085	0.200		**0.562**														

Taula 3.3.3.B: Clústers trobats amb *meta-PCM* sobre el set de dades de *Xerobromion erecti*. Es mostren només aquells clústers amb temps de vida més gran que 0.2 i cardinalitat més gran que 4. Les columnes de l'esquerra corresponen a estadístics descriptors dels clústers: **RD** – Distància de referència; **Card.** – Cardinalitat; **VG** – Variabilitat geomètrica; **dVG** – Primera derivada de la variabilitat geomètrica; **LT** – Temps de vida (*lifetime*). Els clústers (files) estan ordenats per temps de vida decreixents. A les restants columnes es mostren els valors de correlació de phi (Φ) amb els sintàxons de base de la sintaxonomia tradicional. Només es mostren aquelles correlacions positives. S'han ressaltat aquells valors de Φ>0.5.

	RD	Card	VG	dVG	LT	OL	QL	CM	MJ	RJ	RQ	QCR	QCBR	QCC	QCBT	QCT	BJ	CO	QRB	QRR	QRU	QC(s.l.)	QR(s.l.)
G1R1	0.739	5.000	0.274	0.001	0.565	**.551**																	
G5R1	0.810	113.3	0.463	0.337	0.420		.479				.247	.384	.182	.058	.276	.066						**.519**	
G205R1	0.796	64.90	0.472	0.456	0.400														**.896**	.132			**.778**
G208R2	0.859	35.47	0.515	0.237	0.345													**.972**					
G106R1	0.469	4.000	0.088	0.000	0.295							.082	**.667**									.221	
G117R1	0.590	11.91	0.254	0.291	0.285						.129	**.607**										.321	
G69R1	0.556	4.995	0.201	0.006	0.285			.255	.255														
G196R1	0.769	20.72	0.359	0.287	0.280												**.698**						
G78R1	0.595	9.906	0.248	0.106	0.275				**.636**														
G245R1	0.585	14.89	0.239	0.092	0.270														.428				.338
G176R1	0.535	7.964	0.177	0.089	0.265												.423						
G11R1	0.731	19.64	0.349	0.446	0.250				**.912**														
G101R2	0.682	6.949	0.317	0.101	0.250					**.835**													
G8R1	0.727	4.990	0.285	0.063	0.250	**.551**																	
G214R1	0.640	4.961	0.302	0.051	0.250													.347					
G206R1	0.705	8.780	0.389	0.165	0.245												.450						
G220R1	0.606	4.978	0.264	0.020	0.245													.347					
G306R1	0.599	4.992	0.248	0.006	0.245																.466		.192
G203R1	0.700	6.933	0.350	0.101	0.240												.395						
G265R1	0.733	4.977	0.375	0.038	0.240														.243				.192
G288R1	0.600	4.989	0.237	0.019	0.240														.243				.192
G201R1	0.694	5.952	0.335	0.095	0.235												.365						
G96R1	0.747	18.82	0.338	0.404	0.225												**.668**						
G253R1	0.617	4.964	0.282	0.036	0.225															**.742**			.192
G291R1	0.518	8.960	0.184	0.035	0.220														.328				.259
G223R1	0.669	8.866	0.338	0.110	0.215													.469					
G173R1	0.570	8.981	0.191	0.038	0.215												.448						
G321R2	0.600	4.987	0.227	0.028	0.215																.466		.192
G240R1	0.631	4.974	0.286	0.035	0.215													.347					
G74R1	0.635	12.86	0.282	0.242	0.210				**.734**														
G9R1	0.690	16.80	0.319	0.210	0.210				**.838**														
G147R1	0.640	6.973	0.224	0.121	0.210										**.507**							.294	
G312R1	0.665	9.777	0.332	0.214	0.205																**.668**		.275
G255R1	0.617	4.967	0.281	0.032	0.205															**.593**	.068		.192
G213R1	0.655	5.904	0.324	0.120	0.200													.382					
G148R1	0.539	6.918	0.214	0.068	0.200										**.506**							.294	
G211R1	0.654	6.875	0.330	0.102	0.200													.413					
G54R1	0.492	5.977	0.161	0.038	0.200		.308																

Taula 3.3.3.C: Clústers trobats amb *meta-PCM* sobre el set de dades de *Quercetea ilicis* sense *Quercenion*. Es mostren només aquells clústers amb temps de vida més gran que 0.2 i cardinalitat més gran que 4. Les columnes de l'esquerra corresponen a estadístics descriptors dels clústers: **RD** – Distància de referència; **Card.** – Cardinalitat; **VG** – Variabilitat geomètrica; **dVG** – Primera derivada de la variabilitat geomètrica; **LT** – Temps de vida (*lifetime*). Els clústers (files) estan ordenats per temps de vida decreixents. A les restants columnes es mostren els valors de correlació de phi (Φ) amb els sintàxons de base de la sintaxonomia tradicional. Només es mostren aquelles correlacions positives. S'han ressaltat aquells valors de Φ>0.5.

Pel que fa a les comunitats de *Quercetea ilicis* sense *Quercenion* (taula 3.3.3.C), hi ha quatre associacions correlacionades, parcial o plenament, amb clústers de *meta-PCM*: MJ, RJ, BJ i CO. Les diferents subassociacions de *Quercetum cocciferae* (QC) queden englobades en un sol clúster, juntament amb les associacions *Rhamno-Quercetum* (RQ) i *Querco-Lentiscetum* (QL). Dins d'aquest gran clúster, *meta-PCM* identifica dos subgrups, corresponents a QCBR i QCR, respectivament. D'altra banda, les subassociacions de *Quercetum rotundifoliae* (QR) es poden considerar totes juntes en un sol clúster (G205R1) però també tenen identitat de manera aïllada. Finalment, i com en el cas d'AD, apareixen dos clústers distints per a l'associació *Oleo-Lentiscetum* (OL). Es corresponen amb els treballs fitosociològics de Bolòs *et al.* (1984) a les illes Medes i de Franquesa (1995) al cap de Creus, respectivament (vegeu l'annex A de la memòria).

La solució dels problemes d'anidament i superposició mitjançant la selecció manual de clústers ens ha dut a considerar 12 clústers per a B i 15 per a C. A continuació hem generat diferents particions segons hem explicat a l'anterior apartat.

B.*Xerobromion erecti*

Min. Tip.	Num. inv.	% inv.	PCM Det.	KM/FCM Det.	KM-LOO (K=12)	Dif.
0.1	97	39%	0.831	0.831	0.798	0.034
0.05	100	40%	0.847	0.847	0.816	0.031
0.01	108	44%	0.841	0.841	0.816	0.026
0.005	113	46%	0.815	0.831	0.804	0.028
0.001	135	54%	0.711	**0.781**	0.793	-0.012
0.0	248	100%	0.345	0.536	0.685	-0.150

C.*Quercetea* sense *Quercenion*

Min. Tip.	Num. inv.	% inv.	PCM Det.	KM/FCM Det.	KM-LOO (K=15)	Dif.
0.1	224	60%	0.856	0.877	0.585	0.292
0.05	230	61%	0.849	0.850	0.583	0.268
0.01	249	66%	0.791	0.825	0.574	0.251
0.005	257	68%	0.779	0.825	0.579	0.246
0.001	277	74%	0.748	**0.809**	0.634	0.175
0.0	376	100%	0.447	0.719	0.655	0.064

Taules 3.3.4.B-C: Generació de particions comparables a la sintaxonomia tradicional per als clústers obtinguts amb *meta-PCM* sobre els sets de dades B i C. **Min. Tip.** – Tipicalitat mínima d'un inventari per ésser inclòs a la classificació a comparar. **Num. Class.** – Nombre d'inventaris inclosos (classificats). **% Class.** – Percentatge d'inventaris classificats respecte al total. **PCM Det.** – Index de Rand corregit per l'atzar entre la sintaxonomia tradicional i la partició sorgida del valor més alt de tipicalitat per als inventaris inclosos. **KM/FCM Det.** – Índex de Rand corregit per l'atzar entre la sintaxonomia tradicional i la partició sorgida del valor més baix de distància al centroide per als inventaris inclosos. **KM-LOO (K=12/15)** – Índex de Rand corregit per l'atzar entre la sintaxonomia tradicional i la partició obtinguda per *K-means* amb correcció *leave-one-out* amb el nombre de grups indicat. **Dif.** – Diferència entre les columnes [KM/FCM Det.] i [KM-LOO (K=12/15)]. Les particions assenyalades són les que es comparen a les taules 3.3.5.B-C.

A les primeres columnes de les taules 3.3.4.B-C mostrem nombre i percentatge d'inventaris seleccionats en cada llindar de tipicalitat (0.1, 0.05, ..., 0.0). Hom pot constatar com, evidentment, el percentatge d'inventaris inclosos creix en disminuir el llindar de tipicalitat utilitzat en la selecció. A la darrera fila de les taules (llindar 0.0) hom inclou tots els inventaris, també aquells que *PCM* considera *outliers*.

A les columnes següents de les taules mostrem els valors de l'índex de Rand d'ajust al criteri tradicional obtinguts les dues estratègies de determinació: 1) basada en la tipicalitat més alta (PCM Det: Determinació possibilística) i 2) basada en la distància al centroide més curta (KM/FCM Det.: Determinació probabilística). Quan la determinació es restringeix als inventaris més propers, les dues estratègies són semblants i presenten un ajust elevat. A mida que hom inclou més inventaris en la selecció, disminueix l'acord amb el criteri tradicional. No obstant, l'acord obtingut amb la partició per la distància més curta (determinació probabilística) esdevé més alt que l'acord obtigut amb la partició per determinació possibilística.

A la darrera columna de les taules 3.3.4.B i C, mostrem l'índex de Rand de les particions obtingudes executant l'algorisme *K-means* sobre els mateixos inventaris. Quan el llindar de tipicalitat utilitzat per seleccionar els inventaris restringeix la determinació als inventaris més propers al nucli dels clústers, la solució de *meta-PCM* s'ajusta més a la tradicional que no la obtinguda per l'algorisme *K-means*. En incrementar els objectes que volem determinar en forma de partició incrementa la probabilitat d'error per a *meta-PCM*, ja que les tipicalitats inicials cada cop són més baixes. Si hom vol determinar la filiació de tots els inventaris la solució partitiva de *K-means* s'ajusta més al criteri extern.

Hem escollit una partició de cada conjunt de dades, per tal d'estudiar amb detall les correlacions de cada un dels grups determinats amb els de la sintaxonomia tradicional. Concretament hem escollit les particions corresponents a la determinació probabilística amb el llindar de tipicalitat 0.001 perquè creiem que aquest punt ofereix un bon compromís entre la capacitat de determinació correcta i el nombre d'inventaris determinats. Les particions escollides estan ressaltades a les corresponents caselles de les taules 3.3.4.B i C.

Mostrem els valors de correlació Φ calculats a les taules 3.3.5.B-C. A les primeres files de cada taula, hi apareixen el nombre i percentatge d'inventaris de cada un dels grups originals que s'han inclòs en la nova partició. Els sintàxons de base més aïllats tenen percentatges elevats d'inventaris inclosos i correlacions altes. És el cas dels sintàxons AB, TBF, TAV a *Brometalia erecti*, i MJ, RJ, CO, QRB i QRU a *Quercetea ilicis*. Per contra, per als sintàxons amb problemes d'aïllament el nombre d'inventaris no considerats *outliers* són menys i la correspondència (correlació) també és menor.

B.*Xerobromion erecti*

		LB	AB	KG	TBT	TBF	TBH	AD	KAT	KAA	TAV	IBT	IBL	TF	TB(s.l.)	KA(s.l.)	IB(s.l.)
Nº Class.		11	20	8	8	13	1	16	21	8	13	3	2	11	22	29	5
% Class.		34%	95%	42%	28%	100%	11%	70%	68%	53%	100%	30%	33%	41%	43%	63%	31%
G7R1	6.00	**0.835**			0.032		0.232										
G29R1	10.93		**0.956**		0.049		0.016										
G53R1	13.00			**1.000**													
G9R1	7.64	0.235			**0.679**										0.361		
G88R3	13.97					**1.000**									**0.740**		
G120R1	13.26							**0.588**									
G103R1	10.16							**0.770**									
G132R1	22.14								**0.775**							**0.641**	
G125R1	10.22								0.395	**0.544**						**0.661**	
G164R2	12.93										**0.998**						
G184R1	8.04									0.181		**0.677**	**0.550**			0.019	**0.880**
G208R1	6.70													**1.000**			

C.*Quercetea* sense *Quercenion*

		OL	QL	CM	MJ	RJ	RQ	QCR	QCBR	QCC	QCBT	QCT	BJ	CO	QRB	QRR	QRU	QC(s.l.)	QR (s.l.)
Nº Class.		12	13	3	24	10	6	15	3	2	17	0	37	38	69	8	20	37	97
% Class.		75%	24%	27%	100%	100%	50%	65%	60%	33%	65%	0%	90%	100%	99%	89%	91%	54%	96%
G1R1	5.00	**0.637**																	
G8R1	13.55	**0.699**																	
G54R1	20.74		**0.984**							0.107									
G11R1	14.34			0.312	**0.919**														
G101R2	5.99					**0.978**													
G117R1	66.05						**0.506**	**0.751**	0.071	0.005								0.470	
G147R1	5.02							0.029			**0.623**							0.451	
G148R1	18.02							0.023	0.343		**0.624**							**0.556**	
G206R1	9.66												**0.551**						
G196R1	37.99												**0.760**						
G208R2	14.09									0.089				**0.971**					
G205R1	28.03														**0.917**	0.003			**0.767**
G265R1	23.51														0.236				0.185
G253R1	7.99															**0.829**			0.228
G312R1	7.02									2E-04							**0.957**		0.363

Taules 3.3.5.B-C: Correlació phi (Φ) entre la sintaxonomia tradicional i la partició de clústers generada a partir de seleccionar aquells inventaris amb tipicalitat més gran que 0.001 i utilitzant les distàncies als centroides per establir la pertinença (veure taules 3.3.3.B-C i text). S'han ressaltat aquells valors de Φ>0.5. Es mostren els valors de cardinalitat de les dues particions així com el percentatge d'inventaris respecte al total de cada grup de la sintaxonomia tradicional.

3.3.3.5 Discussió

L'estratègia d'anàlisi de clústers que hem estudiat en aquest capítol presenta avantatges i inconvenients respecte als models i mètodes de classificació considerats en anteriors capítols. En primer lloc, no tots els individus (inventaris) són classificats. *Meta-PCM* pot proporcionar com a solució fins i tot l'absència total d'estructura en grups. La solució $K=1$ és més probable si l'exponent de *fuzziness* és gran. Aquesta característica de l'algorisme pot ésser considerada un avantatge o un inconvenient, depenent de si hom desitja la completa classificació de les observacions o no. Tanmateix, és una possibilitat de resposta no contemplada pels algorismes partitius (*KM, FCM*).

La selecció *a priori* del nombre de clústers a cercar ja no és una decisió a prendre per l'investigador, si no que és el mateix algorisme de classificació que serveix alhora de criteri intern i proporciona només aquelles estructures més vàlides des del seu punt de vista. Com a contrapartida, per a dades amb una gran complexitat, la multiplicitat de solucions pot forçar a una selecció *a posteriori* entre estructures anidades o encavalcades, cosa que introdueix un nou element de decisió que pot ésser considerat com un inconvenient. Tot i això, creiem que el fet de posar de manifest diferents solucions alternatives i proporcionar criteris (de vegades divergents) per a facilitar-ne la selecció pot ésser desitjable perquè constitueix una visió més realista i complerta de la complexitat en l'estructura de dades.

Pel que fa a l'exploració de l'espai de solucions, hem proposat aquí una estratègia d'inicialització de clústers i execució successiva de l'algorisme que es mostra capaç de detectar força estructures. No obstant, com en el cas dels algorismes partitius, *meta-PCM* és un algorisme iteratiu que convergeix en mínims locals. Gairebé mai hom podrà estar segur de que totes les estructures possibles han estat trobades.

La determinació per tipicalitats (criteri possibilistic) o per distàncies (criteri probabilístic) fa referència al que Oliva (1995) esmenta com a "problema de la variabilitat". Com podem discriminar dues poblacions de diferent variabilitat? Si utilitzem el criteri de la tipicalitat afavorirem el grup més variable. D'altra banda, si utilitzem el criteri de la distància al nucli afavorirem el grup menys variable. En el nostre cas, la consideració de la distància (o pertinença relativa) es presenta com a més eficient que la tipicalitat (o pertinença absoluta), pel que la millor solució al problema de la variabilitat seria la consideració de la equivariància.

Tanmateix, cal tenir en compte que hem aplicat les estratègies de determinació després de eliminar molts dels possibles inventaris *outliers* de les dades. Per tant, hem eliminat el possible efecte d'atracció d'aquests inventaris sobre el centroide, que anunciàvem a l'inici del capítol com a problema de les particions. La prova de que aquesta eliminació és positiva rau en la davallada de l'ajust quan hom inclou tots els inventaris en la determinació.

En conjunt, els resultats obtinguts en aquest estudi semblen indicar que el més adequat per a explorar l'estructura d'unes dades és:

1. Identificar cada població per separat, permetent la superposició entre grups però situant correctament cada un dels centroides de les poblacions.
2. Si es desitja una solució partitiva, cal filtrar aquells grups menys vàlids, i arribar a una conjunt de clústers de variabilitat comparable.
3. En la determinació final partitiva, considerar que tots els clústers restants tenen igual variabilitat i utilitzar una partició probabilística.

Evidentment la estructura espacial dels inventaris depèn en gran mesura de la mètrica emprada per construir les relacions entre ells. Aquí ens hem restringit exclusivament a la distància de la corda/hellinger amb les transformacions indicades. D'altra banda, la forma dels clústers que cerca *meta-PCM* és hiperesfèrica. Això limita força la quantitat d'estructures que són vàlides. Si bé és possible emprar la mètrica de Mahalanobis amb *meta-PCM* per a poder detectar clústers hiperel·lipsoidals, aquesta possibilitat és impracticable en matrius d'espècies degut a que la matriu de variàncies-covariàncies és freqüentment singular. En el proper capítol estudiarem altres maneres d'abordar l'espai de relacions que ens permeti assolir un aïllament de les estructures de classificació.

3.3.3.6 Conclusions

Les principals conclusions que hem obtingut a aquesta secció són:

- *Meta-PCM* permet identificar aquells grups que es mostren clars en les dades (grups 'naturals'), sense necessitat d'especificar un nombre de grups prefixat. Aquest és el principal avantatge que presenta vers algorismes partitius com *KM* o *FCM*.
- *Meta-PCM* identifica més grups en reduir el paràmetre de *fuzziness*. Per contra, això treu mobilitat a l'algorisme, i els clústers obtinguts tenen un temps de vida més curt.
- El temps de vida dels clústers ens informa del rang d'escales en que podem trobar els grups. Actua com a criteri intern de validació del propi algorisme, servint com a criteri de selecció de clústers en cas de superposició o anidament.
- En generar particions, la determinació per la distància al centroide més curta s'ajusta més al criteri extern de la partició fitosociològica que la determinació per la tipicalitat més alta.
- Quan s'exclouen els inventaris de baixa tipicalitat, les particions obtingudes de l'aplicació de *meta-PCM* amb la determinació per la distància més curta classifica la resta inventaris de manera més ajustada al criteri fitosociològic que els algorismes partitius.
- El mètode de classificació proposat, *Meta-PCM*, ha estat implementat al programa *GINKGO* (vegeu cap. 4.2).

3.3.A Apèndixs: Demostracions de les aportacions a *PCM*

3.3.A.1 Correcció per a dissimilaritats acotades

La demostració de que emprant l'equació de tipicalitat per a distàncies acotades:

$$t^{*}{}_{i(k)} = \frac{t_{i(k)} - t_{e_{max}(k)}}{1 - t_{e_{max}(k)}} = t_{i(k)} \cdot \left[1 - \left(\frac{e_{i(k)}}{e_{max}} \right)^{2/m-1} \right] \qquad \text{(A.1.1)}$$

s'optimitza el funcional:

$$PCM^{*}(k)_{m,\eta_k} = \sum_{i=1}^{N} \alpha_i^m \cdot t_{i(k)}^{*m} \cdot e_{i(k)}^2 + \eta_k^2 \sum_{i=1}^{N} \alpha_i^m \cdot \left(\alpha_i^{-1} - t_{i(k)}^{*} \right)^m , \qquad \text{(A.1.2)}$$

$$\text{on } \alpha_i = \left[1 - \left(\frac{e_{i(k)}}{e_{max}} \right)^{2/m-1} \right]^{-1}$$

s'obté igualant la derivada parcial del funcional A.1.2 a zero, i aïllant $t^{*}_{i(k)}$:

$$\frac{\partial (PCM^{*}(k)_{m,\eta_k})}{\partial t^{*}_{i(k)}} = \alpha_i^m \cdot m \cdot t_{i(k)}^{*(m-1)} \cdot e_{i(k)}^2 - \eta_k^2 \cdot \alpha_i^m \cdot m \cdot \left(\alpha_i^{-1} - t_{i(k)}^{*} \right)^{m-1} = 0$$

$$t_{i(k)}^{*(m-1)} \cdot e_{i(k)}^2 = \eta_k^2 \cdot \left(\alpha_i^{-1} - t_{i(k)}^{*} \right)^{m-1}$$

$$t_{i(k)}^{*} \cdot e_{i(k)}^{2/(m-1)} = \eta_k^{2/(m-1)} \cdot \left(\alpha_i^{-1} - t_{i(k)}^{*} \right)$$

$$t_{i(k)}^{*} \cdot \left(e_{i(k)}^{2/(m-1)} + \eta_k^{2/(m-1)} \right) = \alpha_i^{-1} \cdot \eta_k^{2/(m-1)}$$

$$t_{i(k)}^{*} = \frac{\alpha_i^{-1} \cdot \eta_k^{2/(m-1)}}{e_{i(k)}^{2/(m-1)} + \eta_k^{2/(m-1)}}$$

Finalment, si substituïm α_i:

$$t_{i(k)}^{*} = \frac{\left[1 - \left(\frac{e_{i(k)}}{e_{max}} \right)^{2/m-1} \right] \cdot \eta_k^{2/(m-1)}}{e_{i(k)}^{2/(m-1)} + \eta_k^{2/(m-1)}}$$

i dividim per $\eta_k^{2/(m-1)}$ obtindrem de nou l'equació de tipicalitat A.1.1:

$$t_{i(k)}^{*} = \frac{1 - \left(\frac{e_{i(k)}}{e_{max}} \right)^{2/m-1}}{1 + \left(\frac{e_{i(k)}}{\eta_k} \right)^{2/(m-1)}} = t_{i(k)} \cdot \left[1 - \left(\frac{e_{i(k)}}{e_{max}} \right)^{2/m-1} \right]$$

3.3.A.2 Derivada de la variància d'un clúster respecte la distància de referència

La variància o variabilitat geomètrica difusa d'un clúster de *PCM* és:

$$\hat{V}_{fd}(k) = \frac{\sum_{h,l} t_{h(k)}^m \cdot t_{l(k)}^m \cdot d_{h,l}^2}{2 \cdot \left(\sum_{i=1}^{N} t_{i(k)}^m \right)^2} = \frac{\sum_{i=1}^{N} t_{i(k)}^m \, e_{i(k)}^2}{\sum_{i=1}^{N} t_{i(k)}^m} \qquad \text{(A.2.1)}$$

Emprant la segona forma de la variància, podem calcular la derivada parcial d'aquesta respecte η_k. Suposant que el centroide resta immòbil en augmentar η_k i que, per tant, les distàncies al centroide són constants (però no les tipicalitats, que depenen de la distància de referència) la derivada d'A.2.1 és:

$$\frac{\partial \hat{V}_{fd}(k)}{\partial \eta_k} = \frac{\left(\sum_{i=1}^{N} e_{i(k)}^2 \cdot m \cdot t_{i(k)}^{m-1} \cdot \frac{\partial t_{i(k)}}{\partial \eta_k} \right) \cdot \left(\sum_{i=1}^{N} t_{i(k)}^m \right) - \left(\sum_{i=1}^{N} e_{i(k)}^2 \cdot t_{i(k)}^m \right) \cdot \left(\sum_{i=1}^{N} m \cdot t_{i(k)}^{m-1} \cdot \frac{\partial t_{i(k)}}{\partial \eta_k} \right)}{\left(\sum_{i=1}^{N} t_{i(k)}^m \right)^2}$$

$$\frac{\partial \hat{V}_{fd}(k)}{\partial \eta_k} = m \cdot \frac{\left(\sum_{i=1}^{N} \sum_{j=1}^{N} e_{i(k)}^2 \cdot t_{i(k)}^{m-1} \cdot \frac{\partial t_{i(k)}}{\partial \eta_k} \cdot t_{j(k)}^m \right) - \left(\sum_{i=1}^{N} \sum_{j=1}^{N} e_{i(k)}^2 \cdot t_{i(k)}^m \cdot t_{j(k)}^{m-1} \cdot \frac{\partial t_{j(k)}}{\partial \eta_k} \right)}{\left(\sum_{i=1}^{N} t_{i(k)}^m \right)^2}$$

$$\frac{\partial \hat{V}_{fd}(k)}{\partial \eta_k} = \frac{m \cdot \left(\sum_{i=1}^{N} \sum_{j=1}^{N} e_{i(k)}^2 \cdot t_{i(k)}^m \cdot t_{j(k)}^m \cdot \left(t_{i(k)}^{-1} \cdot \frac{\partial t_{i(k)}}{\partial \eta_k} - t_{j(k)}^{-1} \cdot \frac{\partial t_{j(k)}}{\partial \eta_k} \right) \right)}{\left(\sum_{i=1}^{N} t_{i(k)}^m \right)^2} \qquad \text{(A.2.2)}$$

Calculem ara la derivada de $t_{i(k)}$:

$$t_{i(k)} = \frac{1}{1 + \left(\frac{e_{i(k)}}{\eta_k} \right)^{2/(m-1)}} = \frac{1}{1 + \eta_k^{-2/(m-1)} \cdot e_{i(k)}^{2/(m-1)}}$$

$$\frac{\partial t_{i(k)}}{\partial \eta_k} = \frac{(-1) \cdot e_{i(k)}^{2/(m-1)} \cdot \left(\frac{-2}{m-1} \right) \cdot \eta_k^{(-2/m-1)-1}}{\left(1 + \left(\frac{e_{i(k)}}{\eta_k} \right)^{2/(m-1)} \right)^2}$$

$$\frac{\partial t_{i(k)}}{\partial \eta_k} = \left(\frac{2}{m-1} \right) \cdot \eta_k^{-1} \cdot \left(\frac{e_{i(k)}}{\eta_k} \right)^{2/(m-1)} \cdot t_{i(k)}^2 \qquad \text{(A.2.3)}$$

A continuació, podem realitzar el següent canvis en A.2.3:

$$t_{i(k)}^{-1} \cdot \frac{\partial t_{i(k)}}{\partial \eta_k} = \left(\frac{2}{m-1} \right) \cdot \eta_k^{-1} \cdot t_{i(k)} \cdot \left(\frac{e_{i(k)}}{\eta_k} \right)^{2/(m-1)} \tag{A.2.4}$$

$$t_{i(k)}^{-1} \cdot \frac{\partial t_{i(k)}}{\partial \eta_k} - t_{j(k)}^{-1} \cdot \frac{\partial t_{j(k)}}{\partial \eta_k} = \left(\frac{2}{m-1} \right) \cdot \eta_k^{-1} \cdot \left(t_{i(k)} \cdot \left(\frac{e_{i(k)}}{\eta_k} \right)^{2/(m-1)} - t_{j(k)} \cdot \left(\frac{e_{j(k)}}{\eta_k} \right)^{2/(m-1)} \right) \tag{A.2.5}$$

Finalment, substituint A.2.5 a A.2.2 arribem a la següent expressió de la derivada:

$$\frac{\partial \hat{V}_{fd}(k)}{\partial \eta_k} = \left(\frac{2 \cdot m}{m-1} \right) \cdot \eta_k^{-1} \cdot \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} e_{i(k)}^2 \cdot t_{i(k)}^m \cdot t_{j(k)}^m \cdot \left(t_{i(k)} \cdot \left(\frac{e_{i(k)}}{\eta_k} \right)^{2/(m-1)} - t_{j(k)} \cdot \left(\frac{e_{j(k)}}{\eta_k} \right)^{2/(m-1)} \right)}{\sum_{i=1}^{N} \sum_{j=1}^{N} t_{i(k)}^m \cdot t_{j(k)}^m} \tag{A.2.6}$$

Per a PCM*, la demostració és semblant. La tipicalitat d'una dissimilaritat acotada és (A.1.1):

$$t^*_{i(k)} = t_{i(k)} \cdot \left[1 - \left(\frac{e_{i(k)}}{e_{max}} \right)^{2/(m-1)} \right]$$

Seguint un mateix procés anàleg a A.2.1 podem demostrar que:

$$\frac{\partial \hat{V}^*_{fd}(k)}{\partial \eta_k} = \frac{\sum_{i=1}^{N} t^{*m}_{i(k)} \, e^2_{i(k)}}{\sum_{i=1}^{N} t^{*m}_{i(k)}} = \ldots = \frac{m \cdot \left(\sum_{i=1}^{N} \sum_{j=1}^{N} e^2_{i(k)} \cdot t^{*m}_{i(k)} \cdot t^{*m}_{j(k)} \cdot \left(t^{*-1}_{i(k)} \cdot \frac{\partial t^*_{i(k)}}{\partial \eta_k} - t^{*-1}_{j(k)} \cdot \frac{\partial t^*_{j(k)}}{\partial \eta_k} \right) \right)}{\left(\sum_{i=1}^{N} t^{*m}_{i(k)} \right)^2} \tag{A.2.7}$$

Com que A.2.6 és $t_{i(k)}$ multiplicada per una constant, pel que fa a la derivada respecte η_k:

$$\frac{\partial t^*_{i(k)}}{\partial \eta_k} = \frac{\partial t_{i(k)}}{\partial \eta_k} \left[1 - \left(\frac{e_{i(k)}}{e_{max}} \right)^{2/(m-1)} \right] \tag{A.2.8}$$

$$t^{*-1}_{i(k)} \cdot \frac{\partial t^*_{i(k)}}{\partial \eta_k} = t^{-1}_{i(k)} \cdot \left[1 - \left(\frac{e_{i(k)}}{e_{max}} \right)^{2/(m-1)} \right]^{-1} \cdot \frac{\partial t_{i(k)}}{\partial \eta_k} \cdot \left[1 - \left(\frac{e_{i(k)}}{e_{max}} \right)^{2/(m-1)} \right] = t^{-1}_{i(k)} \cdot \frac{\partial t_{i(k)}}{\partial \eta_k} \tag{A.2.9}$$

pel que finalment obtenim una expressió molt semblant a A.2.5:

$$\frac{\partial \hat{V}^*_{fd}(k)}{\partial \eta_k} = \left(\frac{2 \cdot m}{m-1} \right) \cdot \eta_k^{-1} \cdot \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} e^2_{i(k)} \cdot t^{*m}_{i(k)} \cdot t^{*m}_{j(k)} \cdot \left(t_{i(k)} \cdot \left(\frac{e_{i(k)}}{\eta_k} \right)^{2/(m-1)} - t_{j(k)} \cdot \left(\frac{e_{j(k)}}{\eta_k} \right)^{2/(m-1)} \right)}{\sum_{i=1}^{N} \sum_{j=1}^{N} t^{*m}_{i(k)} \cdot t^{*m}_{j(k)}} \tag{A.2.10}$$

Es compleix doncs, que: $\lim_{e_{max} \to \infty} \left[\frac{\partial \hat{V}^*_{fd}(k)}{\partial \eta_k} \right] = \frac{\partial \hat{V}_{fd}(k)}{\partial \eta_k}$

3.3.A.3 Segona derivada de la variància d'un clúster respecte la distància de referència

Per a calcular la segona derivada partirem de la primera derivada, en la forma d'A.2.2:

$$\frac{\partial \hat{V}_{fd}(k)}{\partial \eta_k} = \frac{m \cdot \left(\sum_{i=1}^{N} \sum_{j=1}^{N} e_{i(k)}^2 \cdot t_{i(k)}^m \cdot t_{j(k)}^m \cdot \left(t_{i(k)}^{-1} \cdot \frac{\partial t_{i(k)}}{\partial \eta_k} - t_{j(k)}^{-1} \cdot \frac{\partial t_{j(k)}}{\partial \eta_k} \right) \right)}{\left(\sum_{i=1}^{N} t_{i(k)}^m \right)^2}$$

A continuació per simplificar l'àlgebra fem la substitució:

$$\alpha_{i(k)} = t_{i(k)}^{-1} \cdot \frac{\partial t_{i(k)}}{\partial \eta_k} \qquad \text{(A.3.1)}$$

L'equació a derivar és ara:

$$\frac{\partial \hat{V}_{fd}(k)}{\partial \eta_k} = \frac{m \cdot \left(\sum_{i=1}^{N} \sum_{j=1}^{N} e_{i(k)}^2 \cdot t_{i(k)}^m \cdot t_{j(k)}^m \cdot \left(\alpha_{i(k)} - \alpha_{j(k)} \right) \right)}{\left(\sum_{i=1}^{N} t_{i(k)}^m \right)^2} \qquad \text{(A.3.2)}$$

Tenim una derivada d'un quocient. A continuació procedim per parts. En primer lloc calculem la derivada del numerador:

$$\frac{\partial \left[m \cdot \sum_{i=1}^{N} \sum_{j=1}^{N} e_{i(k)}^2 \cdot t_{i(k)}^m \cdot t_{j(k)}^m \cdot \left(\alpha_{i(k)} - \alpha_{j(k)} \right) \right]}{\partial \eta_k} = m \cdot \sum_{i=1}^{N} \sum_{j=1}^{N} e_{i(k)}^2 \cdot \frac{\partial \left[t_{i(k)}^m \cdot t_{j(k)}^m \cdot \left(\alpha_{i(k)} - \alpha_{j(k)} \right) \right]}{\partial \eta_k}$$

$$\frac{\partial \left[t_{i(k)}^m \cdot t_{j(k)}^m \cdot \left(\alpha_{i(k)} - \alpha_{j(k)} \right) \right]}{\partial \eta_k} =$$

$$= m \cdot t_{i(k)}^{m-1} \cdot \frac{\partial t_{i(k)}}{\partial \eta_k} \cdot t_{j(k)}^m \cdot \left(\alpha_{i(k)} - \alpha_{j(k)} \right) + t_{i(k)}^m \cdot m \cdot t_{j(k)}^{m-1} \cdot \frac{\partial t_{j(k)}}{\partial \eta_k} \cdot \left(\alpha_{i(k)} - \alpha_{j(k)} \right) + t_{i(k)}^m \cdot t_{j(k)}^m \cdot \left(\frac{\partial \alpha_{i(k)}}{\partial \eta_k} - \frac{\partial \alpha_{j(k)}}{\partial \eta_k} \right)$$

$$= m \cdot t_{i(k)}^m \cdot \alpha_{i(k)} \cdot t_{j(k)}^m \cdot \left(\alpha_{i(k)} - \alpha_{j(k)} \right) + t_{i(k)}^m \cdot m \cdot t_{j(k)}^m \cdot \alpha_{j(k)} \cdot \left(\alpha_{i(k)} - \alpha_{j(k)} \right) + t_{i(k)}^m \cdot t_{j(k)}^m \cdot \left(\frac{\partial \alpha_{i(k)}}{\partial \eta_k} - \frac{\partial \alpha_{j(k)}}{\partial \eta_k} \right)$$

$$= t_{i(k)}^m \cdot t_{j(k)}^m \cdot \left[m \cdot \alpha_{i(k)} \cdot \left(\alpha_{i(k)} - \alpha_{j(k)} \right) + m \cdot \alpha_{j(k)} \cdot \left(\alpha_{i(k)} - \alpha_{j(k)} \right) + \left(\frac{\partial \alpha_{i(k)}}{\partial \eta_k} - \frac{\partial \alpha_{j(k)}}{\partial \eta_k} \right) \right]$$

$$= t_{i(k)}^m \cdot t_{j(k)}^m \cdot \left[m \cdot \alpha_{i(k)}^2 - m \cdot \alpha_{i(k)} \cdot \alpha_{j(k)} + m \cdot \alpha_{j(k)} \cdot \alpha_{i(k)} - m \cdot \alpha_{j(k)}^2 + \left(\frac{\partial \alpha_{i(k)}}{\partial \eta_k} - \frac{\partial \alpha_{j(k)}}{\partial \eta_k} \right) \right]$$

$$= t_{i(k)}^m \cdot t_{j(k)}^m \cdot \left[m \cdot \alpha_{i(k)}^2 - m \cdot \alpha_{j(k)}^2 + \left(\frac{\partial \alpha_{i(k)}}{\partial \eta_k} - \frac{\partial \alpha_{j(k)}}{\partial \eta_k} \right) \right] = t_{i(k)}^m \cdot t_{j(k)}^m \cdot \left[\left(m \cdot \alpha_{i(k)}^2 + \frac{\partial \alpha_{i(k)}}{\partial \eta_k} \right) - \left(m \cdot \alpha_{j(k)}^2 + \frac{\partial \alpha_{j(k)}}{\partial \eta_k} \right) \right]$$

Finalment, la derivada del numerador és (A.3.3):

$$\frac{\partial \left[m \cdot \sum_{i=1}^{N} \sum_{j=1}^{N} e_{i(k)}^2 \cdot t_{i(k)}^m \cdot t_{j(k)}^m \cdot \left(\alpha_{i(k)} - \alpha_{j(k)} \right) \right]}{\partial \eta_k} = m \cdot \sum_{i=1}^{N} \sum_{j=1}^{N} e_{i(k)}^2 \cdot t_{i(k)}^m \cdot t_{j(k)}^m \cdot \left[\left(m \cdot \alpha_{i(k)}^2 + \frac{\partial \alpha_{i(k)}}{\partial \eta_k} \right) - \left(m \cdot \alpha_{j(k)}^2 + \frac{\partial \alpha_{j(k)}}{\partial \eta_k} \right) \right]$$

Podem continuar, tractem la derivada $\frac{\partial \alpha_{i(k)}}{\partial \eta_k}$, a partir de l'equació A.2.4:

$$\frac{\partial \alpha_{i(k)}}{\partial \eta_k} = \frac{\partial\left(\left(\frac{2}{m-1}\right)\cdot \eta_k^{-1}\cdot t_{i(k)}\cdot\left(\frac{e_{i(k)}}{\eta_k}\right)^{2/(m-1)}\right)}{\partial \eta_k} = \left(\frac{2}{m-1}\right)\cdot e_{i(k)}^{2/(m-1)}\cdot\frac{\partial\left(\eta_k^{-1}\cdot t_{i(k)}\cdot \eta_k^{-2/(m-1)}\right)}{\partial \eta_k} =$$

$$= \left(\frac{2}{m-1}\right)\cdot e_{i(k)}^{2/(m-1)}\cdot\left[(-1)\cdot\eta_k^{-2}\cdot t_{i(k)}\cdot\eta_k^{-2/(m-1)} + \eta_k^{-1}\cdot\frac{\partial t_{i(k)}}{\partial \eta_k}\cdot\eta_k^{-2/(m-1)} + \eta_k^{-1}\cdot t_{i(k)}\cdot\left(\frac{-2}{m-1}\right)\cdot\eta_k^{(-2/(m-1))-1}\right] =$$

$$= \left(\frac{2}{m-1}\right)\cdot e_{i(k)}^{2/(m-1)}\cdot\eta_k^{-1}\cdot t_{i(k)}\cdot\eta_k^{-2/(m-1)}\cdot\left[(-1)\cdot\eta_k^{-1} + t_{i(k)}^{-1}\cdot\frac{\partial t_{i(k)}}{\partial \eta_k} + \left(\frac{-2}{m-1}\right)\cdot\eta_k^{-1}\right] =$$

$$= \alpha_{i(k)}\cdot\left[(-1)\cdot\eta_k^{-1} + \alpha_{i(k)} + \left(\frac{-2}{m-1}\right)\cdot\eta_k^{-1}\right] = \alpha_{i(k)}\cdot\left[\left(\frac{-m-1}{m-1}\right)\cdot\eta_k^{-1} + \alpha_{i(k)}\right]$$

$$\frac{\partial \alpha_{i(k)}}{\partial \eta_k} = \alpha_{i(k)}^2 - \alpha_{i(k)}\cdot\left(\frac{m+1}{m-1}\right)\cdot\eta_k^{-1} \qquad \text{(A.3.4)}$$

I llavors,

$$m\cdot\alpha_{i(k)}^2 + \frac{\partial \alpha_{i(k)}}{\partial \eta_k} = m\cdot\alpha_{i(k)}^2 + \alpha_{i(k)}^2 - \alpha_{i(k)}\cdot\left(\frac{m+1}{m-1}\right)\cdot\eta_k^{-1} = \alpha_{i(k)}^2\cdot(m+1) - \alpha_{i(k)}\cdot\left(\frac{m+1}{m-1}\right)\cdot\eta_k^{-1}$$

$$m\cdot\alpha_{i(k)}^2 + \frac{\partial \alpha_{i(k)}}{\partial \eta_k} = m+1)\cdot\left(\alpha_{i(k)}^2 - \left(\frac{\alpha_{i(k)}}{(m-1)\cdot\eta_k}\right)\right) \qquad \text{(A.3.5)}$$

Finalment, podem expressar la derivada del numerador, substituint A.3.5 a A.3.3, com:

$$\frac{\partial\left[m\cdot\sum_{i=1}^{N}\sum_{j=1}^{N}e_{i(k)}^2\cdot t_{i(k)}^m\cdot t_{j(k)}^m\cdot\left(\alpha_{i(k)}-\alpha_{j(k)}\right)\right]}{\partial \eta_k} = m\cdot\sum_{i=1}^{N}\sum_{j=1}^{N}e_{i(k)}^2\cdot t_{i(k)}^m\cdot t_{j(k)}^m\cdot(m+1)\cdot\left[\left(\alpha_{i(k)}^2 - \left(\frac{\alpha_{i(k)}}{(m-1)\cdot\eta_k}\right)\right) - \left(\alpha_{j(k)}^2 - \left(\frac{\alpha_{j(k)}}{(m-1)\cdot\eta_k}\right)\right)\right]$$

Per altra banda, la derivada del denominador és (canviant els subíndexs per evitar confusions):

$$\frac{\partial\left[\left(\sum_{i=1}^{N}t_{l(k)}^m\right)^2\right]}{\partial \eta_k} = 2\cdot\left(\sum_{l=1}^{N}t_{l(k)}^m\right)\cdot\left(\sum_{l=1}^{N}m\cdot t_{l(k)}^{m-1}\cdot\frac{\partial t_{l(k)}}{\partial \eta_k}\right) = 2\cdot\left(\sum_{l=1}^{N}t_{l(k)}^m\right)\cdot\left(\sum_{l=1}^{N}m\cdot t_{l(k)}^m\cdot\alpha_{l(k)}\right)$$

La derivada segona es troba posant en comú els resultats obtinguts per al denominador i numerador:

$$\frac{\partial^2 \hat{V}_{fd}(k)}{\partial^2 \eta_k} = \frac{\partial}{\partial \eta_k}\left[\frac{m\cdot\left(\sum_{i=1}^{N}\sum_{j=1}^{N} e_{i(k)}^2 \cdot t_{i(k)}^m \cdot t_{j(k)}^m \cdot \left(\alpha_{i(k)} - \alpha_{j(k)}\right)\right)}{\left(\sum_{i=1}^{N} t_{i(k)}^m\right)^2}\right] =$$

$$= \frac{\frac{\partial\left[m\cdot\sum_{i=1}^{N}\sum_{j=1}^{N} e_{i(k)}^2 \cdot t_{i(k)}^m \cdot t_{j(k)}^m \cdot \left(\alpha_{i(k)} - \alpha_{j(k)}\right)\right]}{\partial \eta_k}\cdot\left(\sum_{i=1}^{N} t_{l(k)}^m\right)^2 - \left[m\cdot\sum_{i=1}^{N}\sum_{j=1}^{N} e_{i(k)}^2 \cdot t_{i(k)}^m \cdot t_{j(k)}^m \cdot \left(\alpha_{i(k)} - \alpha_{j(k)}\right)\right]\cdot\frac{\partial\left[\left(\sum_{i=1}^{N} t_{l(k)}^m\right)^2\right]}{\partial \eta_k}}{\left(\sum_{i=1}^{N} t_{l(k)}^m\right)^4}$$

$$\frac{\partial^2 \hat{V}_{fd}(k)}{\partial^2 \eta_k} = \frac{m\cdot\sum_{i=1}^{N}\sum_{j=1}^{N} e_{i(k)}^2 \cdot t_{i(k)}^m \cdot t_{j(k)}^m \cdot \left[(m+1)\cdot\left[\left(\alpha_{i(k)}^2 - \frac{\alpha_{i(k)}}{(m-1)\cdot\eta_k}\right) - \left(\alpha_{j(k)}^2 - \frac{\alpha_{j(k)}}{(m-1)\cdot\eta_k}\right)\right] - \left(\alpha_{i(k)} - \alpha_{j(k)}\right)\cdot 2\cdot\frac{\sum_{l=1}^{N} m\cdot t_{l(k)}^m \cdot \alpha_{l(k)}}{\sum_{l=1}^{N} t_{l(k)}^m}\right]}{\left(\sum_{i=1}^{N} t_{l(k)}^m\right)^2}$$

on $\alpha_{i(k)} = t_{i(k)}^{-1}\cdot\frac{\partial t_{i(k)}}{\partial \eta_k} = t_{i(k)}^{*-1}\cdot\frac{\partial t_{i(k)}^*}{\partial \eta_k} = \left(\frac{2}{m-1}\right)\cdot\eta_k^{-1}\cdot t_{i(k)}\cdot\left(\frac{e_{i(k)}}{\eta_k}\right)^{2/(m-1)}$

El procediment per a *PCM** és semblant. Partim, en aquest cas, d' A.2.7:

$$\frac{\partial \hat{V}_{fd}(k)}{\partial \eta_k} = \frac{m\cdot\left(\sum_{i=1}^{N}\sum_{j=1}^{N} e_{i(k)}^2 \cdot t_{i(k)}^{*m} \cdot t_{j(k)}^{*m} \cdot \left(t_{i(k)}^{*-1}\cdot\frac{\partial t_{i(k)}^*}{\partial \eta_k} - t_{j(k)}^{*-1}\cdot\frac{\partial t_{j(k)}^*}{\partial \eta_k}\right)\right)}{\left(\sum_{i=1}^{N} t_{i(k)}^{*m}\right)^2}$$

La substitució A.3.1 ens serveix tant per *PCM* com per *PCM**:

$$\alpha_{i(k)} = t_{i(k)}^{-1}\cdot\frac{\partial t_{i(k)}}{\partial \eta_k} = t_{i(k)}^{*-1}\cdot\frac{\partial t_{i(k)}^*}{\partial \eta_k} \qquad \text{(A.2.9)}$$

pel que la equació a derivar és:

$$\frac{\partial \hat{V}_{fd}(k)}{\partial \eta_k} = \frac{m\cdot\left(\sum_{i=1}^{N}\sum_{j=1}^{N} e_{i(k)}^2 \cdot t_{i(k)}^{*m} \cdot t_{j(k)}^{*m} \cdot \left(\alpha_{i(k)} - \alpha_{j(k)}\right)\right)}{\left(\sum_{i=1}^{N} t_{i(k)}^{*m}\right)^2}$$

Es compleix, a més: $\frac{\partial \alpha_{i(k)}}{\partial \eta_k} = -\alpha_{i(k)}^2 + t_{i(k)}^{*-1}\cdot\frac{\partial^2 t_{i(k)}^*}{\partial^2 \eta_k}$ i $t_{i(k)}^{*-1}\cdot\frac{\partial^2 t_{i(k)}^*}{\partial^2 \eta_k} = t_{i(k)}^{-1}\cdot\frac{\partial^2 t_{i(k)}}{\partial^2 \eta_k}$.

Per tant, la segona derivada per a *PCM** acotat és:

$$\frac{\partial^2 \hat{V}_{fd}(k)}{\partial^2 \eta_k} = \frac{m \cdot \sum_{i=1}^{N}\sum_{j=1}^{N} e_{i(k)}^2 \cdot t_{i(k)}^{*m} \cdot t_{j(k)}^{*m} \cdot \left[(m+1) \cdot \left[\left(\alpha_{i(k)}^2 - \frac{\alpha_{i(k)}}{(m-1)\cdot \eta_k} \right) - \left(\alpha_{j(k)}^2 - \frac{\alpha_{j(k)}}{(m-1)\cdot \eta_k} \right) \right] - (\alpha_{i(k)} - \alpha_{j(k)}) \cdot 2 \cdot \frac{\sum_{l=1}^{N} m \cdot t_{l(k)}^{*m} \cdot \alpha_{l(k)}}{\sum_{l=1}^{N} t_{l(k)}^{*m}} \right]}{\left(\sum_{i=1}^{N} t_{l(k)}^{*m} \right)^2}$$

on $\alpha_{i(k)} = t_{i(k)}^{-1} \cdot \frac{\partial t_{i(k)}}{\partial \eta_k} = t_{i(k)}^{*-1} \cdot \frac{\partial t_{i(k)}^{*}}{\partial \eta_k} = \left(\frac{2}{m-1} \right) \cdot \eta_k^{-1} \cdot t_{i(k)} \cdot \left(\frac{e_{i(k)}}{\eta_k} \right)^{2/(m-1)}$

Capítol 3.4: Sobre la ponderació de variables en la classificació de comunitats

3.4.1 Introducció

3.4.1.1 Ponderació, selecció i estandardització de variables

En l'anàlisi de dades ecològiques, és habitual recollir el màxim de descriptors possibles i considerar-los tots ells igualment importants. Aquest mode d'anàlisi fou també, durant molt de temps, una de les bases de la taxonomia numèrica (Sneath & Sokal 1973) i continua aplicant-se en el cas de la sintaxonomia numèrica. Tanmateix, la inclusió de totes les variables pot resultar una font de problemes, perquè hom pot estar introduint soroll no desitjat a les dades. Tal i com remarcà Milligan (1996): "*Only those variables which are believed to discriminate between clusters should be used*". Milligan (1980) mostrà que la sola addició d'unes poques variables irrellevants podia interferir seriosament en la recuperació de grups prèviament coneguts. Algunes variables poden ocultar la informació que proporcionen les altres, fent que l'anàlisi de grups o clústers (*cluster analysis*) sigui una tasca feixuga o condemnada a resultats erronis.

No és sempre fàcil decidir *a priori* quines variables ajuden a establir estructures de classificació i quines no. Segons Gnanadesikan *et al*. (1995) hi ha dues aproximacions possibles: la selecció i/o la ponderació de variables. En la ponderació de variables, hom estima la importància relativa de les variables segons uns pesos calculats a partir de les mateixes dades. En la selecció de variables, hom escull directament les variables que haurien d'ésser incloses a l'anàlisi de clúster i descarta les altres. Les dues aproximacions intenten facilitar la formació de clústers. En realitat, podem considerar la selecció de variables com un cas extrem de la ponderació. En l'àmbit de la vegetació, un argument per a seleccionar o ponderar variables és que com les espècies presenten diferents sensibilitats de resposta a factors ambientals, han de presentar també contribucions diferents per a generar patrons de vegetació (Shaukat 1989).

Un tema molt relacionat, però aparentment en sentit contrari, és el de l'estandardització de variables. Abans de ponderar les variables podem assumir que totes tenen igual importància i estandarditzar-les segons un mètode concret. Amb l'estandardització, no s'intenta posar l'èmfasi en la relativa importància de les variables sinó que intenta eliminar-ne les dimensions físiques.

Milligan & Cooper (1988) estudiaren vuit estandarditzacions diferents i conclogueren que la divisió per rangs era el mètode més efectiu. Tanmateix, els seus resultats no són generalitzables a qualsevol tipus de dades. L'estandardització de variables és recomanable quan aquestes tenen unitats diferents. En el cas de matrius de composició de tàxons, les espècies acostumen a estar mesurades en les mateixes unitats. Per tant des d'aquest punt de vista sembla innecessari, o fins i tot perjudicial, estandarditzar les variables a l'hora d'analitzar les relacions entre comunitats vegetals.

3.4.1.2 Ponderació i selecció de variables en mètodes de *clustering*

Gran part de la literatura aplicable al problema de la importància de les variables en mètodes d'anàlisi de clústers s'obté d'estadística multivariant clàssica per a classificacions ja conegudes. Un cop tenim una classificació en K clústers o grups, hom pot realitzar K tests de la t de Student, o bé utilitzar la T^2 de Hotelling, MANOVA, anàlisi discriminant *stepwise*, etc. Evidentment, el problema afegit en l'anàlisi de clústers de vegetació és que la classificació és *a priori* desconeguda, cosa que complica molt la possible selecció o ponderació de variables.

A partir dels anys 80, es començà a estudiar la definició de mètodes de *clustering* que incloguessin procediments de ponderació o selecció de variables. Fowlkes *et al*. (1988) proposaren un algorisme per a seleccionar variables progressivament, de manera semblant a la selecció de variables progressiva (*forward*) que s'utilitza en anàlisi discriminant lineal. Concretament, Fowlkes *et al.* aplicaren la selecció de variables al mètode jeràrquic del veí més llunyà (*complete linkage*). Per a seleccionar la primera variable, en primer lloc calculaven dendrogrames a partir de cada una de les variables disponibles; seguidament realitzaven talls als dendrogrames, en nivells successius de la jerarquia, i sobre les particions resultants, calculaven un estadístic per a mesurar la separació o aïllament entre grups. La primera variable seleccionada era la que presentava una major separació. A continuació es seleccionaven progressivament més variables cercant quina combinació de variables produïa una partició amb aïllament més significatiu.

En l'àmbit de la ponderació de variables, DeSarbo *et al*. (1984) proposaren l'algorisme *SYNCLUS* (*Synthetised clustering*), per a ésser aplicat al màrqueting. Aquest complex algorisme partitiu incloïa una generalització de l'algorisme *K-means* que calculava els pesos de les variables i l'assignació dels individus alternativament. *SYNCLUS* produïa, com a solució, els pesos de les variables i la partició dels individus. Green *et al*. (1990) estudiaren la vàlua d'aquest mètode per a posar de manifest quines variables eren importants. Arribaren a la conclusió de que, en alguns casos, la solució depenia críticament del punt de partida.

Una línia de treball diferent és la de cercar pesos que optimitzin (en termes de mínims quadrats) una determinada funció. De Soete (1986, 1988) proposà un mètode, que anomenà *OVWTRE*, per a trobar els pesos òptims de les variables en arbres ultramètrics i arbres additius. En un estudi de simulació, Milligan (1989) avaluà la capacitat de l'algorisme OVWTRE de donar pesos nuls o gairebé nuls a les variables que emmascaraven l'estructura. Milligan conclogué que *OVWTRE* era una bona opció en aquelles aplicacions de d'anàlisi de clústers on s'utilitzés la distància Euclidiana i algorismes jeràrquics agglomeratius. Més recentment, Makarenkov & Legendre (2001) estengueren les idees de De Soete de ponderació òptima de variables al cas de l'algorisme *K-means* (MacQueen 1967). Un dels problemes de l'aproximació dels pesos òptims és l'elevat cost de computació que implica trobar-los, fins i tot per a conjunts de dades de mida relativament petita.

Gnanadesikan *et al*. (1995) compararen nou aproximacions diferents a la ponderació de variables sobre dades reals i dades simulades: la ponderació igual, l'estandardització de les variables per la desviació típica i per rangs, diverses alternatives basades en la variabilitat intra- i/o entre-clústers, *SYNCLUS* i *OWVTRE*. Entre les aproximacions que consideraren més efectives hi havia aquelles basades en la estimació de la variabilitat intra- i entre- clústers. Gnanadesikan *et al*. (1995) també estudiaren el procediment de selecció de variables de Fowlkes *et al*. (1988). Malgrat que trobaren aquesta aproximació força útil en comparació amb les estratègies de ponderació de variables, no era consistentment el millor mètode i, a més, estava limitat a dades auto-escalades.

Recentment, Brusco & Cradit (2001) han treballat la selecció de variables a *K-means* proposant el mètode *HINoV* (*Heuristic identification of noisy variables*). Anàlogament a la metodologia de Fowlkes *et al*. (1988), *HINoV* realitza inicialment una partició *K-means* per a cada variable. A continuació calcula índexs de Rand (1971) corregits per l'atzar (Hubert & Arabie 1985) entre parelles de particions, obtenint una matriu simètrica. Per a cada variable, es sumen els valors d'acord entre la seva partició i cada una de les particions sorgides de les altres variables. Seguidament les variables amb valors baixos d'aquest estadístic són eliminades. Per acabar hom executa *K-means* només en base a les variables restants. Brusco & Cradit (2001) admeten que *HINoV* té tendència a fallar en la selecció quan: a) hi hagi un nivell de correlació alt entre les variables que emmascaren l'estructura, i b) hi hagi múltiples estructures de clúster en el mateix set de dades.

Segons el nostre punt de vista, les aproximacions de selecció de variables de Fowlkes *et al*. (1988) de Brusco & Cradit (2001) no sembla massa adequades quan es variables són abundàncies de tàxons. Si les abundàncies provenen de la transformació d'una escala ordinal, el nombre de valors és limitat i, per tant, no té massa sentit establir classificacions en K grups a partir d'una sola variable.

Els estudis que hem esmentat demostren que la ponderació i selecció de variables, si és fa de manera adequada, pot incrementar substancialment l'eficiència dels mètodes de classificació en la majoria de casos. No obstant, també és cert que tots els mètodes esmentats han tingut un impacte bastant limitat i a la pràctica han estat molt poc utilitzats. Abans de recórrer a la selecció o ponderació de variables en mètodes de classificació de la vegetació és necessari, al nostre parer, considerar els següents punts:

a) *Té sentit ponderar/seleccionar les variables en el nostre problema de classificació?* En els pròxims apartats suposarem que per a la classificació de comunitats la resposta és afirmativa, però al final tornarem a plantejar-nos aquesta qüestió en vista dels resultats obtinguts.

b) En cas afirmatiu, cal definir els pesos o seleccionar les variables d'acord amb el concepte de clúster del nostre problema concret. Els criteris de selecció i/o ponderació han d'ésser fàcilment interpretables.

c) És important ésser conscient de la tautologia implícita en el càlcul de pesos o la selecció de variables: Es necessari disposar de classificacions *a priori* per calcular els pesos de les variables o seleccionar aquelles que clarifiquen l'estructura. És a dir, necessitem una classificació *a priori* per a saber quines variables són importants i quines no. A la vegada, necessitem els pesos o la selecció de variables per a poder descobrir l'estructura que d'altra banda, amb totes les variables queda amagada o distorsionada. La circularitat esmentada és palesa per als algorismes de *clustering* però menys problemàtica en l'anàlisi discriminant, ja que, en aquest darrer cas la classificació d'entrenament es considera, d'inici, veritable.

En el proper apartat (3.4.2.1) estudiem la possibilitat d'emprar una ponderació de variables en conjunció amb mesures de dissimilaritats diferents de la distància Euclidiana, tema que, a la bibliografia consultada, resta fonamentalment per tractar. A continuació, exposem dues estratègies de ponderació de variables que hem cregut interessant testar en la problemàtica de la classificació d'inventaris. Noteu que el punt de vista d'aquest capítol es basa en la transformació de les columnes (variables) de la matriu de dades. Per tant, estem modificant, en el cas lineal, l'angle entre els eixos de l'espai de dades. Complementariament, al capítol 3.2 estudiavem l'ús de mesures de proximitat, i algunes d'aquestes mesures implicaven transformacions de les files (inventaris), que correspon a resituar els inventaris individualment sense alterar el conjunt dels eixos de dades. Finalment (a la secció 3.4.2) presentem un estudi de l'aplicació d'aquestes estratègies de ponderació a les nostres dades. L'estudi que realitzem no comporta l'aplicació de mètodes de *clustering* sinó que tan sols analitza la deformació geomètrica i el canvi en la discriminabilitat deguda a l'acció dels pesos. Creiem, però, que els resultats són extensibles a la incorporació d'aquests pesos en algorismes partitius.

3.4.2 Estratègies de ponderació en la classificació de comunitats vegetals

3.4.2.1. Distàncies amb variables ponderades

Sigui $\mathbf{X}_{NxP}$ una matriu de P variables (columnes) observades en N objectes (files). La ponderació de variables pot ésser aplicada de diferents maneres. Una primer aproximació consisteix en transformar $\mathbf{X}$ multiplicant-la per una matriu de pesos $\mathbf{M}_{PxP}$:

$$\widetilde{\mathbf{X}} = \mathbf{X} \cdot \mathbf{M}$$

Assumirem aquí que $\mathbf{M} = Diag(m_1, m_2, ..., m_p)$ i $m_j \in \Re^+ \; \forall j$. Per tant, els pesos seran tots positius i no considerarem la possible relació entre les variables en la definició dels pesos. Després de la ponderació. la distància Euclidiana al quadrat entre dos objectes és:

$$d_{12}^2 = (\mathbf{x}_1 - \mathbf{x}_2)' \mathbf{M}' \cdot \mathbf{M} (\mathbf{x}_1 - \mathbf{x}_2)$$

Alternativament, Gnanadesikan *et al*. (1995) consideraren l'estratègia d'aplicar la matriu $\mathbf{A}_{PxP}$ inductora d'una norma. En aquest segon cas la distància al quadrat entre els dos objectes és:

$$d_{12}^2 = (\mathbf{x}_1 - \mathbf{x}_2)' \mathbf{A} (\mathbf{x}_1 - \mathbf{x}_2)$$

Hom podria imaginar encara un tercer cas en que es combinessin la transformació associada a una norma i la ponderació de les variables:

$$d_{12}^2 = (\widetilde{\mathbf{x}}_1 - \widetilde{\mathbf{x}}_2)' \mathbf{A} (\widetilde{\mathbf{x}}_1 - \widetilde{\mathbf{x}}_{2'}) = (\mathbf{x}_1 - \mathbf{x}_2)' \mathbf{M}' \mathbf{A} \mathbf{M} (\mathbf{x}_1 - \mathbf{x}_2)$$

Noteu que si $\mathbf{M}$ no és singular (això és cert en el cas de $\mathbf{M}$ diagonal no nul·la), la nova matriu inductora de la norma, $\mathbf{M'AM}$, és encara semidefinida positiva. En el cas concret d' $\mathbf{A} = \mathbf{I}$ tenim la distància Euclidiana ponderada (al quadrat):

$$d_{ii'}^2 = \sum_{j=1}^{p} m_j^2 (x_{ij} - x_{i'j})^2$$

L'efecte de la matriu $\mathbf{M}$ en distancies que poden ésser expressades com a un producte creuat pot ésser estudiat analitzant $\mathbf{M'AM}$. Cal tenir en compte, que, en el cas de la distància de Mahalanobis és inútil aplicar una matriu de ponderació, car aquesta distància és invariant a qualsevol transformació escalar o sorgida d'una combinació lineal de variables.

Què cal fer quan hom desitja ponderar les variable en un espai que no pot ésser expressat com a una norma de productes creuats? Aquest és el cas de força mesures de distància emprades en ecologia, com la distància de la corda (Orlóci 1967) o la distància de Hellinger (Rao 1995). Aquestes proximitats es poden expressar a partir de la matriu original com a transformacions de les files (Legendre & Gallagher 2001) però no de les columnes (vegeu apartat 3.2.2.6 del capítol 3.2).

Concretament, la transformació per files que dóna lloc a la distància de Hellinger és:

$h_{ij} = \sqrt{\frac{x_{ij}}{\sum_{l=1}^{P} x_{il}}}$, i llavors la distància es calcula com a $d_{12}^2 = (\mathbf{h}_1 - \mathbf{h}_2)'(\mathbf{h}_1 - \mathbf{h}_2)$

Ja vam veure al capítol 3.2 que la distància de Hellinger era una mètrica més adequada per a la classificació d'inventaris de vegetació que la distància Euclidiana. Per tant, sembla desitjable combinar la ponderació de variables amb aquesta mètrica. Com aplicar a la vegada, la transformació de Hellinger i una transformació de ponderació? Tenim dues opcions:

1) Com que la distància de Hellinger és una mesura euclidiana, podem aplicar la transformació de Hellinger en primer lloc i després ponderar l'espai euclidià resultant:

$$d_{12}^2 = (\mathbf{h}_1 - \mathbf{h}_2)'\mathbf{M}'\cdot\mathbf{M}(\mathbf{h}_1 - \mathbf{h}_2)$$

L'inconvenient d'aquesta opció és que es perd la distància màxima entre observacions, que amb Hellinger és $\sqrt{2}$.

2) A la inversa, aplicar la distància de Hellinger sobre la matriu de dades ponderada:

$$\widetilde{h}_{ij} = \sqrt{\frac{\widetilde{x}_{ij}}{\sum_{l=1}^{P} \widetilde{x}_{il}}}, d_{12}^2 = (\widetilde{\mathbf{h}}_1 - \widetilde{\mathbf{h}}_2)'(\widetilde{\mathbf{h}}_1 - \widetilde{\mathbf{h}}_2)$$

Aquesta darrera aproximació preserva la distància màxima i els pesos serveixen per modular les proporcions de les variables dins del vector. Cal notar però, que en aquest cas la construcció dels pesos es realitza sobre la matriu **X** i no sobre la matriu **H**.

3.4.2.2. Introducció d'un exponent de transformació

Quan tractem amb dades reals, la proporció entre variables informatives i variables soroll és impredictible. Si cap de les variables és aleatòria respecte a l'estructura a detectar una ponderació excessiva pot suposar un inconvenient per a la classificació. És, doncs, aconsellable poder controlar l'efecte dels pesos sobre el procés de classificació. Si **M** és una matriu diagonal, una manera fàcil de modular el seu efecte és introduint un exponent $s \in \Re^+$ als pesos:

$$\mathbf{M}^S = Diag(m_1^S, m_2^S, ..., m_p^S)$$

Quan s=0 totes les cel·les diagonals esdevenen 1 i desapareix l'efecte de la ponderació. Com més alt sigui l'exponent més gran serà la influència dels pesos en la classificació.

3.4.2.3 Estratègia de ponderació basada en la capacitat discriminant de les variables

La intenció és reduir l'efecte de les variables considerades "soroll". És a dir, aquelles variables que, pel propòsit de la classificació en grups, són poc o gens informatives, o indueixen a classificacions errònies. Aquestes variables haurien d'ésser ponderades amb un pes baix. A la vegada, les variables que contribueixin a diferenciar grups haurien d'ésser preservades, cosa que implica concedir a aquestes un pes més elevat.

Assumint que la estructura del clúster és coneguda *a priori* a través d'una partició, hom pot emprar la coneguda descomposició de la suma de quadrats (*sum of squares*, *SS*). Així definim:

$$\mathbf{T} = \mathbf{B} + \mathbf{W}$$

on **T** és la matriu *SS* total, mentre que **B** i **W** són les matrius *SS* entre grups i dins de grups (*between* i *within*), respectivament.

La nostra definició de pesos es basa en la consideració de la diagonal de les matrius **B** i **W**, denotades per $Diag(\mathbf{B})$ i $Diag(\mathbf{W})$:

$$\mathbf{M} = Diag(\mathbf{B}) \cdot Diag(\mathbf{W})^{-1}$$

Tant **B** com **W** són en realitat estimacions, perquè són calculades a partir de particions dels objectes, que poden no correspondre amb la partició "veritable" (si és que alguna partició es pot considerar "veritable").

El càlcul dels elements diagonals d'**M** és el següent:

- Sigui $I(\omega_i \subset \Omega_k)$ una variable indicadora (delta de Kronecker) que pren el valor 1 quan l'objecte ω_i és un membre del clúster Ω_k, i 0 en qualsevol altre cas.
- Sigui $N_k = \sum_{i=1}^{n} I(\omega_i \subset \Omega_k)$ el nombre d'objectes de Ω_k.
- Llavors, la descomposició de *SS* per a una variable *j* és:

$$b_j = \sum_{k=1}^{K} N_k \cdot \left(\hat{x}_{j(k)} - \bar{x}_j\right)^2$$

$$w_j = \sum_{k=1}^{K} \sum_{i=1}^{N} \left(I(\omega_i \subset \Omega_k) \cdot \left(x_{ij} - \hat{x}_{j(k)}\right)^2\right)$$

$$t_j = w_j + b_j = \sum_{i=1}^{N} \left(x_{ij} - \bar{x}_j\right)^2$$

on $\hat{x}_{j(k)} = \left(\sum_{i=1}^{N} I(\omega_i \subset \Omega_k) \cdot x_{ij}\right) \Big/ N_k$ és la mitjana estimada de *j* al clúster Ω_k i $\bar{x}_j$ la mitjana global.

Les expressions anàlogues de **B** , **W** i **T** el cas d'una partició difusa sorgeixen de la substitució de la variable indicadora per el grau de pertinença.

- Siguin $\mathbf{U}_{NxK}$, una partició de N objectes en K clústers, i f el coeficient de *fuzziness* (Bezdek, 1981):

$$b_{fj} = \sum_{k=1}^{K}\sum_{i=1}^{N} u_{ik}^{f} \cdot \left(\hat{x}_{j(k)} - \bar{x}_{j}\right)^2$$

$$w_{fj} = \sum_{k=1}^{K}\sum_{i=1}^{N} \left(u_{ik}^{f} \cdot \left(x_{ij} - \hat{x}_{j(k)}\right)^2\right)$$

$$t_{fj} = w_{fj} + b_{fj}$$

on $\hat{x}_{j(k)} = \left(\sum_{i=1}^{N} u_{ik}^{f} \cdot x_{ij}\right) \Big/ \sum_{i=1}^{N} u_{ik}^{f}$ és la mitjana estimada de j al clúster difús Ω_k.

L'efecte esperat de la ponderació basada en les matrius **B** i **W** és ajudar a augmentar la distància entre clústers i a la vegada disminuir la seva variabilitat interna. Les cel·les de la diagonal de $\mathbf{M} = Diag(\mathbf{B}) \cdot Diag(\mathbf{W})^{-1}$ són estadístics semblants a l'estadístic F de l'anàlisi de la variància (*ANOVA*), sense els graus de llibertat:

$$m_j = b_j / w_j = F_j \cdot \frac{(N-K)}{(K-1)}$$

Com que la part dels graus de llibertat és constant per a totes les variables, podem excloure-la del còmput dels pesos.

En el seu estudi, Gnanadesikan *et al*. (1995) examinaren l'efecte de diverses matrius de ponderació **M** sobre la classificació en mètriques induïdes per una norma. Les mètriques ponderades foren aplicades a diferents conjunts de dades, i es compararen els resultats de la classificació, juntament amb els que proporcionava *SYNCLUS* i *complete linkage*. Entre les matrius de ponderació estudiades recomanaren $\mathbf{M} = Diag(\mathbf{B}) \cdot Diag(\mathbf{W})^{-1}$ pels casos en que algunes variables tinguessin una forta estructura de clústers.

El rang de variació dels pesos definits amb $\mathbf{M} = Diag(\mathbf{B}) \cdot Diag(\mathbf{W})^{-1}$ és $m_j \subset (0,\infty)$. Una aproximació semblant en quan a l'ordre dels pesos, però amb un rang $m_j \subset [0,1]$ és :

$$m_j = b_j / t_j ,$$

essent més fàcil d'entendre la relació amb l'anterior a través de la inversa: $w_j / b_j = t_j / b_j - 1$. El rang d'aquesta darrera definició de pes facilita la comparació dels seus resultats amb la estratègia de ponderació que proposem en el proper apartat.

3.4.2.4 Estratègia de ponderació basada en la correlació tàxon-grup

Com en el cas anterior, a l'estratègia que proposem aquí modifiquem l'espai de dades amb una matriu diagonal de transformació de les variables. En aquest cas, però, la transformació és diferent per a cada clúster Ω_k:

$$\widetilde{\mathbf{X}}_{k,s} = \mathbf{X} \cdot \mathbf{M}_{k,s} \text{, on } \mathbf{M}_{k,s} = diag(m_{1k}^S, m_{2k}^S, ..., m_{pk}^S)$$

on, altra vegada, $m_{jk} \in \Re^+ \ \forall j$.

Un precedents d'aquest tipus d'aproximació el trobem, per exemple, a Chernoff (1972), que utilitzà una mètrica diferent per a cada clúster. És recomanable estudiar aquesta ponderació individualitzada principalment per dues raons: En primer lloc, permet l'ús de pesos en el model de classificació de *Possibilistic C-means* (Krishnapuram & Keller 1995). En segon lloc, a les comunitats vegetals, els tàxons actuen sovint com a indicadors de certs clústers, raó per la qual els fitosociòlegs acostumen a basar la determinació d'inventaris de vegetació en la presència o absència d'aquells tàxons considerats més fidels a la comunitat.

La fidelitat és un concepte que relaciona la distribució d'un tàxon amb la d'un sintàxon (vegei capítol 2.2). En termes estadístics es tracta d'una correlació. La correlació de Pearson entre la presència d'un tàxon *j* i la pertinença a un clúster Ω_k és:

$$r_{jk}(\mathbf{z}_j, \mathbf{u}_k) = \frac{\sum_{i=1}^{N} (z_{ij} - \overline{z}_j) \cdot (u_{ik} - \overline{u}_k)}{\sum_{i=1}^{N} (z_{ij} - \overline{z}_j)^2 \cdot \sum_{i=1}^{N} (u_{ik} - \overline{u}_k)^2}$$

on $z_{ij} = I(x_{ij} > 0)$ és una variable indicadora de la presència o absència del tàxon.

Quan $\mathbf{U}$ és una partició clàssica (*crisp*), és fàcil comprovar que r equival al coeficient d'associació Φ (Sokal & Rohlf 1995:741-743) entre dues variables binàries. Seguint la notació de Chytrý *et al*. (2002) aquest coeficient és:

$$\Phi = \frac{n_k - N_k \cdot (n/N)}{\sqrt{n \cdot N_k \cdot (N-n) \cdot (N-N_k)/(N^2)}}$$

on:

- $N_k = \sum_{i=1}^{N} u_{ik}$ és el nombre d'objectes o inventaris de *k*
- $n = \sum_{i=1}^{N} b_{ij}$ és el nombre d'aparicions del tàxon en el conjunt de les dades
- $n_k = \sum_{i=1}^{N} u_{ik} \cdot b_{ij}$ és el nombre d'aparicions del tàxon al grup *k*.

El coeficient Φ s'utilitza sovint com a mesura de fidelitat de tàxons a sintàxons (Chýtrý *et al*. 2002, veure capítol 2.2). La relació entre Φ i r es compleix només quan ambdues variables - la que representa la presència del tàxon i la que representa la pertinença al clúster - són binàries, però no es compleix quan una d'elles o totes dues són contínues. Per tant, si hom desitja calcular fidelitats a partir de conjunts *fuzzy*, els valors de Φ i r seran diferents. Per tractar-se d'una mesura més correntment emprada en estadística, recomanem aquí emprar r enlloc de Φ, per a calcular fidelitats de tàxons a conjunts difusos.

La fidelitat d'un tàxon a un sintàxon, com la correlació entre dues variables, pot ser positiva (denotant una preferència) o negativa (denotant una exclusió). L'interès de la fidelitat negativa sobre una base de dades d'inventaris àmplia és molt baix (veure cap. 2.2). Per tant, en la construcció dels pesos és recomanable basar-se només en valors positius de fidelitat. Per a construir una estratègia de ponderació a partir d'una mesura de fidelitat o correlació, hem escollit transformar la mesura de fidelitat:

$$m_{jk} = 1 + \max(0, r_{jk})$$

L'addició de la constant evita donar un pes excessiu a la fidelitat, ja que $m_{jk} \in (1,2)$. Amb l'aplicació de l'exponent *s* els pesos finals són: $m_{jk}^{s} \in (1,\infty)$.

3.4.2.5 La significació dels pesos i la selecció de variables

L'estadístic *F* pot ésser testat per conèixer la seva significació. També es possible conèixer la significació de Φ transformant-lo en l'estadístic u (Bruelheide 2000). A causa de l'esperada forta desviació de la normalitat en la distribució de les dades de vegetació, la distribució d'aquests estadístics s'hauria d'estimar empíricament a partir de tests de permutació. La significació dels pesos obre la porta a la selecció de variables enlloc de la ponderació de les mateixes, per exemple, descartant aquelles variables que no arribessin a un nivell de significació preestablert.

3.4.3 La ponderació dels tàxons en la determinació d'inventaris de vegetació

3.4.3.1 Objectius

Els objectius d'aquesta secció són principalment dos. En primer lloc, estudiar la deformació geomètrica de l'espai de dades que introdueixen les dues estratègies de ponderació de variables que hem presentat en els apartats precedents. En segon lloc, determinar si aquestes estratègies de ponderació de variables augmenten la discriminabilitat numèrica dels grups establerts per l'escola fitosociològica tradicional.

3.4.3.2 Metodologia

Matrius de dades i transformacions

Les matrius de dades que estudiarem són dos:

A. Els inventaris de *Xerobromion erecti*: 248 inventaris i 548 tàxons de comunitats pratenses xeròfiles montanes. La classificació tradicional consta de 13 sintàxons de base (5 associacions i 8 subassociacions).
B. Els inventaris de *Quercetea ilicis* sense *Quercenion ilicis*: 376 inventaris i 493 tàxons de matollars mediterranis. La classificació tradicional consta de 16 sintàxons de base (8 associacions i 8 subassociacions).

Hem exclòs d'aquest estudi les dades de Bowman & Wilson (1986) perquè, per a aquest conjunt de dades, la classificació que fins ara hem considerat "criteri extern" té un origen numèric (Dale 1988a) i no sigmatista. D'altra banda, estimem que la deformació geomètrica dels pesos en l'espai de dades de Bowman & Wilson conduiria a les mateixes conclusions generals.

Els valors d'abundància, inicialment mesurats amb l'escala de cobertura-abundància de Braun-Blanquet, han estat transformats a valors numèrics amb la transformació combinada (van der Maarel 1979). A continuació, hem aplicat una transformació escalar de les abundàncies dels tàxons seguint la equació $y = x^w$ (van der Maarel 1979, Currall 1987), on x és el valor de la transformació combinada. Els exponents triats són w=1.0 per al conjunt de dades A i w=1.5 per al *set* de dades B. La selecció d'aquests exponents de transformació sorgeix de les conclusions obtingudes en el capítol 3.2 pel que fa a la capacitat de detecció d'estructures dels diferents espais de relacions.

Estratègies de ponderació

A partir de les matrius d'abundàncies transfomades $\mathbf{Y}$ hem aplicat les dues estratègies de ponderació de les variables introduïdes en apartats precedents. Siguin *s* un exponent de ponderació positiu, i $\mathbf{U}_{NxK}$ la matriu de pertanyences clàssica sorgida de la classificació sintàxonomica tradicional:

W1. Ponderació basada en la discriminabilitat:

$$\widetilde{\mathbf{Y}} = \mathbf{Y} \cdot \mathbf{M}^s \text{, on}$$

$$\mathbf{M}^s = Diag(m_1^s, m_2^s, ..., m_p^s),\ m_j = 1 + b_j / t_j,$$

$$b_j = \sum_{k=1}^{K} \cdot \sum_{i=1}^{N} u_{i(k)} \left(\bar{y}_{j(k)} - \bar{y}_j\right)^2,$$

$$t_j = \sum_{i=1}^{N} \left(y_{ij} - \bar{y}_{j(k)}\right)^2,$$

$$\bar{y}_{j(k)} = \left(\sum_{i=1}^{N} u_{ik} \cdot y_{ij}\right) \Big/ \sum_{i=1}^{N} u_{ik} \text{ i } \bar{y}_j = \left(\sum_{i=1}^{N} y_{ij}\right) \Big/ N.$$

W2. Ponderació basada en la correlació tàxon-clúster:

$$\widetilde{\mathbf{Y}}_k = \mathbf{Y}_k \cdot \mathbf{M}_k^s \text{, on}$$

$$\mathbf{M}_k^s = Diag(m_{1k}^s, m_{2k}^s, ..., m_{pk}^s),\ m_{jk} = 1 + \max(0, r_{jk}),$$

$$r_{jk} = r_{jk}(\mathbf{z}_j, \mathbf{u}_k) = \frac{\sum_{i=1}^{N} (z_{ij} - \bar{z}_j) \cdot (u_{ik} - \bar{u}_k)}{\sum_{i=1}^{N} (z_{ij} - \bar{z}_j)^2 \cdot \sum_{i=1}^{N} (u_{ik} - \bar{u}_k)^2},$$

$$z_{ij} = I(y_{ij} > 0),$$

$$\bar{z}_j = \left(\sum_{i=1}^{N} z_{ij}\right) \Big/ N \text{ i } \bar{u}_k = \left(\sum_{i=1}^{N} u_{ik}\right) \Big/ N.$$

Les dues definicions de pesos, m_j i m_{jk} tenen un rang $[1, 2]$ per tal que les dues estratègies siguin el màxim de comparables. El valor màxim es veu incrementat en augmentar l'exponent de ponderació s. Els exponents assajats han estat $s = 0, 1, 2, ..., 10$.

A partir de les matriu ponderada $\widetilde{\mathbf{Y}}$ (o qualsevol $\widetilde{\mathbf{Y}}_k$), hem aplicat la transformació que dóna lloc a la distància de Hellinger:

$$\widetilde{h}_{ij} = \sqrt{\frac{\widetilde{y}_{ij}}{\sum_{l=1}^{P} \widetilde{y}_{il}}}$$

L'avaluació de la deformació geomètrica per efecte dels pesos i la discriminabilitat dels clústers sota diferents exponents de ponderació, s, s'ha fet sobre la matriu $\widetilde{\mathbf{H}}$ que és la que conté l'espai de relacions entre inventaris definitiu. Noteu que si $s = 0$, llavors $\widetilde{\mathbf{H}}$ equival a l'espai de Hellinger sense ponderar.

Avaluació de la deformació geomètrica

La deformació de l'espai ponderat s'ha estudiat per a cada clúster per separat. Els passos que hem seguit per a avaluar la deformació són:

1. Hem calculat la nova posició del centroide, $\overline{\widetilde{\mathbf{h}}}_{(k)}$:

$$\overline{\widetilde{h}}_{j(k)} = \left(\sum_{i=1}^{N} u_{ik} \cdot \widetilde{h}_{ij} \right) \Big/ \sum_{i=1}^{N} u_{ik}$$

2. Hem calculat les distàncies dels inventaris al centroide:

$$\widetilde{d}_{i(k)} = d(\widetilde{\mathbf{h}}_i, \overline{\widetilde{\mathbf{h}}}_{(k)}) = \sqrt{\sum_{j=1}^{P} (\widetilde{h}_{ij} - \overline{\widetilde{h}}_{j(k)})^2}$$

   Noteu que la distància calculada es la Euclidiana o Pitagòrica, però la distància respecte la matriu ponderada $\widetilde{\mathbf{Y}}$ és la de Hellinger.

3. Hem calculat les mitjanes aritmètiques i els coeficients de variació (c.v.) de les distàncies al centroide per als inventaris membres del clúster i els inventaris externs. Aquesta avaluació externa i interna de la deformació geomètrica proporciona una idea del canvi en aïllament que experimenten els clústers. En una situació ideal hom esperaria una disminució de la distància mitjana per als inventaris membres i un augment de la mateixa per als no membres.

Per tal de presentar els resultats de manera més entenedora, no mostrem els valors absoluts, sinó que hem calculat la diferència de tots els estadístics respecte a la situació no ponderada (s=0). En relació als valors originals hom pot comprovar més fàcilment el sentit i la magnitud deformació que s'esdevé, tant en la posició del centroide com en els estadístics que avaluen de l'aïllament del clúster.

Avaluació de la variació de la discriminabilitat dels grups

Per a començar l'avaluació del canvi en la discriminabilitat, hem realitzar una anàlisi discriminant basada en distàncies (Cuadras *et al*. 1997). Concretament, hem calculat les pertinences difuses a partir de les noves distàncies als centroides dels diferents grups:

$$\widetilde{u}_{i(k)} = \frac{1}{\sum_{l=1}^{K}\left[\frac{\widetilde{d}_{i(k)}}{\widetilde{d}_{i(l)}}\right]^{2/(f-1)}}, \text{ on } f = 1.1.$$

El fet de realitzar una determinació difusa en lloc d'una determinació clàssica (simplement assignant l'inventari al grup pel que presenta la distància al centroide més petita) ens permet avaluar de manera més precisa els canvis en la discriminabilitat. La determinació de l'anàlisi discriminant ha estat realitzada per resubstitució dels inventaris. Si bé seria més recomanable una avaluació creuada per *leave-one-out* això implicaria calcular tot el procés de ponderació per a l'extracció de cada inventari. L'increment notable del temps de càlcul no creiem que hagués suposat un canvi substancial de les conclusions, raó per la qual hem optat per la determinació per resubstitució.

Un cop obtinguda les particions per resubstitució, $\widetilde{\mathbf{U}}_{NxK}$, l'hem comparada amb la partició original, $\mathbf{U}_{NxK}$. Concretament, hem calculat la correlació entre el clúster original i el sorgit de l'anàlisi discriminant difusa:

$$r(\mathbf{u}_k, \widetilde{\mathbf{u}}_k) = \frac{\sum_{i=1}^{N}(u_{ik} - \overline{u}_k)\cdot(\widetilde{u}_{ik} - \overline{\widetilde{u}}_k)}{\sum_{i=1}^{N}(u_{ik} - \overline{u}_k)^2 \cdot \sum_{i=1}^{N}(\widetilde{u}_{ik} - \overline{\widetilde{u}}_k)^2}$$

Cal no confondre aquesta correlació amb la emprada per definir els pesos, malgrat l'estadístic sigui el mateix. Les correlacions s'han calculat per a cada exponent de ponderació (s). Com en el cas de l'avaluació de la deformació geomètrica, hem mesurat el canvi en la discriminabilitat calculant la diferència entre el valor de correlació per a cada valor d' s respecte a la determinació sense ponderació (s=0).

L'ajust global entre les particions s'ha mesurat amb l'índex de Rand (1971) corregit per l'atzar (Hubert & Arabie 1985), en la seva versió clàssica i difusa (vegeu apartat 3.1.5.3 del capítol 3.1). L'ús de la versió difusa integra el guany o pèrdua de objectes en els clústers amb l'increment o reducció de l'aïllament dels mateixos.

Per acabar hem comparat les particions al nivell de ponderació $s = 4$, observant els canvis en la matriu de confusió soferts en relació a $s = 0$.

3.4.3.3 Resultats

Desplaçament dels centroides

A les figures 3.4.1-A i 3.4.1-B podem observar els desplaçaments que experimenten els centroides dels diferents grups en incrementar la intensitat de la ponderació de les espècies. La mesura de desplaçament és la distància entre la posició del centroid en l'espai de Hellinger sense i amb l'aplicació dels pesos. Obviament, el esplaçament és sempre creixent per a un valor creixent de l'exponent de ponderació (s).

Una primera observació és que les dues estratègies de ponderació provoquen desplaçaments de magnitud semblant. Noteu, no obstant, que la magnitud del desplaçament és variable pels diferents grups. El sintàxon de *Xerobromion* que experimenta un desplaçament més gran és IBT (*Irido-Brometum* subass. *typicum*) en les dues estratègies de ponderació. En canvi, a *Quercetea* sense *Quercenion* els sintàxons amb més desplaçament són QRU (*Quercetum rotundifoliae* subass. *ulicetosum*) i QRB (*Quercetum rotundifoliae* subass. *buxetosum*) per a l'estratègia W1, mentre que per l'estratègia W2 ho són CO (*Clematido-Osyrietum*) i QL (*Querco-Lentiscetum*). Si bé les magnituds de desplaçament són semblants, l'ordre entre sintàxons és diferent per a les dues estratègies de ponderació, fet que fa palesa la diferència entre les dues estratègies.

La distància al centroide dels inventaris membres

Les figures 3.4.2.A i 3.4.2.B mostren la variació de la mitjana (a dalt) i del coeficient de variació (a baix) de la distància al centroide, per als inventaris membres del grup. La tendència que mostren la majoria de sintàxons és la de disminuir la distància mitjana al centroide. Aquest efecte augmenta la compactació general del clúster, que és un dels resultats que persegueix l'estratègia de ponderació adoptada. En general, aquells clústers on el desplaçament del centroide era més pronunciat, també presenten una major disminució de les distàncies al centroide (IBT a *Xerobromion*; CO i QRB a *Quercetea*).

Malauradament, l'aplicació dels pesos fa créixer la variància i, evidentment, el coeficient de variació de les distàncies al centroide. Aquest resultat sembla indicar la presència d'inventaris que, en contra de la tendència mitjana, augmenten la seva distància al centroide per efecte de la ponderació. Són inventaris on els tàxons afavorits per la ponderació hi manquen. A *Xerobromion* els sintàxons de base on l'increment del c.v. és més gran són TBF (*Teucrio-Brometum* subass. *festucetosum fallacis*) i TF (*Teucrio-Festucetum*). En el cas de B (*Quercetea ilicis* sense *Quercenion*) hi ha més diferències, de nou, entre W1 i W2. Noteu també que la magnitud de l'increment del coeficient de variació a 3.4.2.B és més gran en la ponderació W1 que a W2.

W1. $m_j = 1 + b_j / t_j$ **W2.** $m_{jk} = 1 + \max(0, r_{jk})$

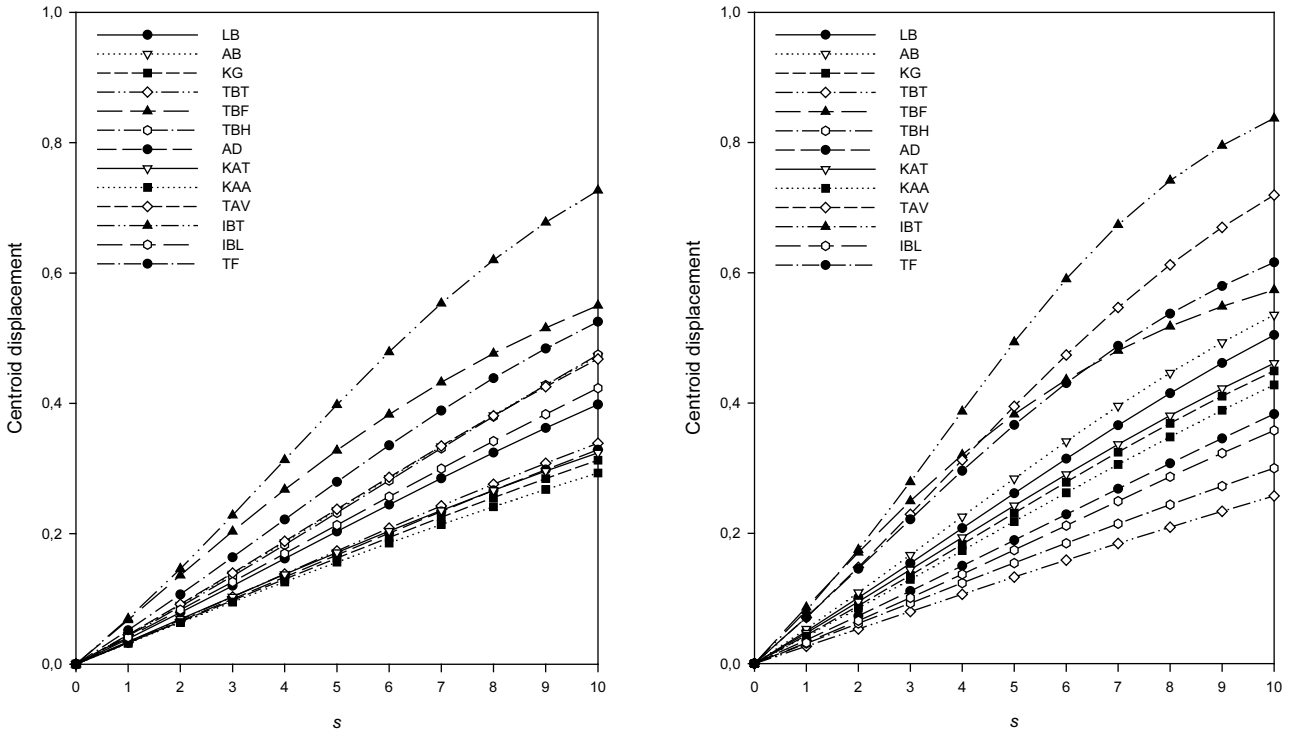


Figura 3.4.1.A: Desplaçament de la posició del centroide sota l'efecte de la ponderació de variables per a diferents exponents de ponderació (*s*). Resultats del conjunt de dades de *Xerobromion erecti*.

W1. $m_j = 1 + b_j / t_j$ **W2.** $m_{jk} = 1 + \max(0, r_{jk})$

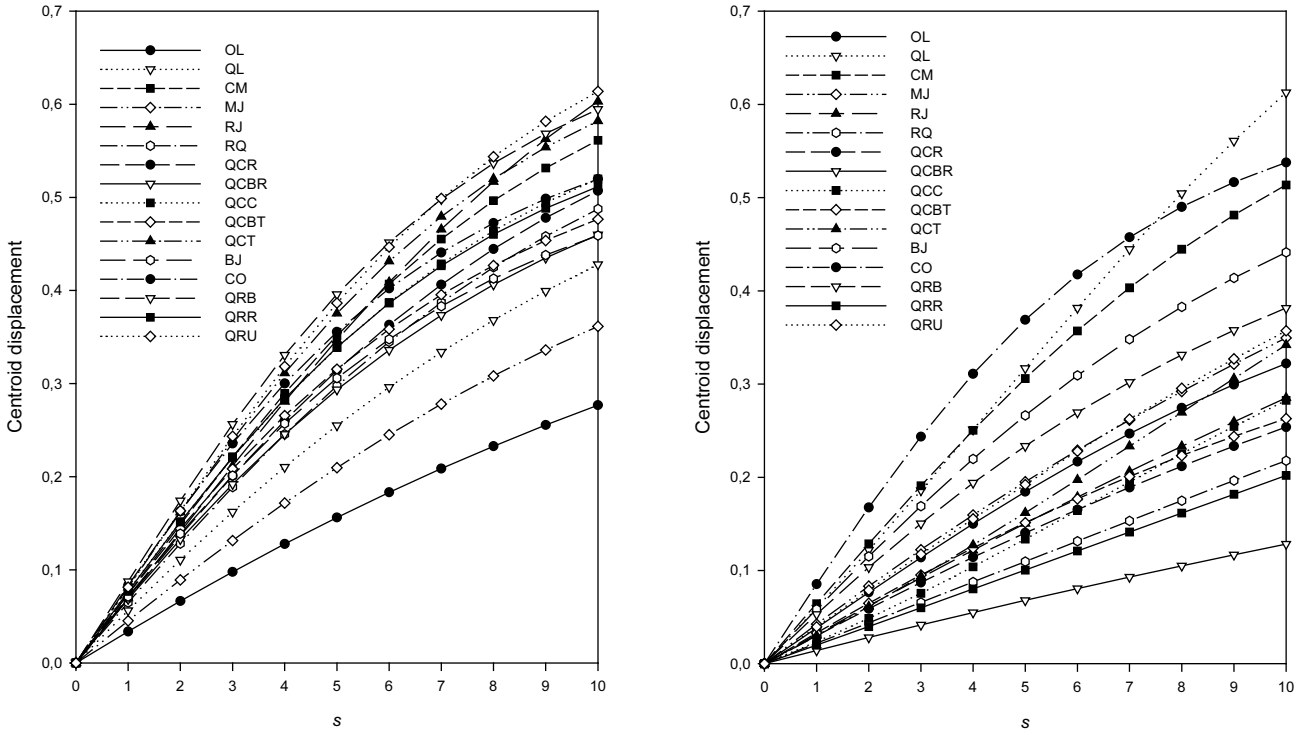


Figura 3.4.1.B: Desplaçament de la posició del centroide sota l'efecte de la ponderació de variables per a diferents exponents de ponderació (*s*). Resultats del conjunt de dades de *Quercetea ilicis* sense *Quercenion*.

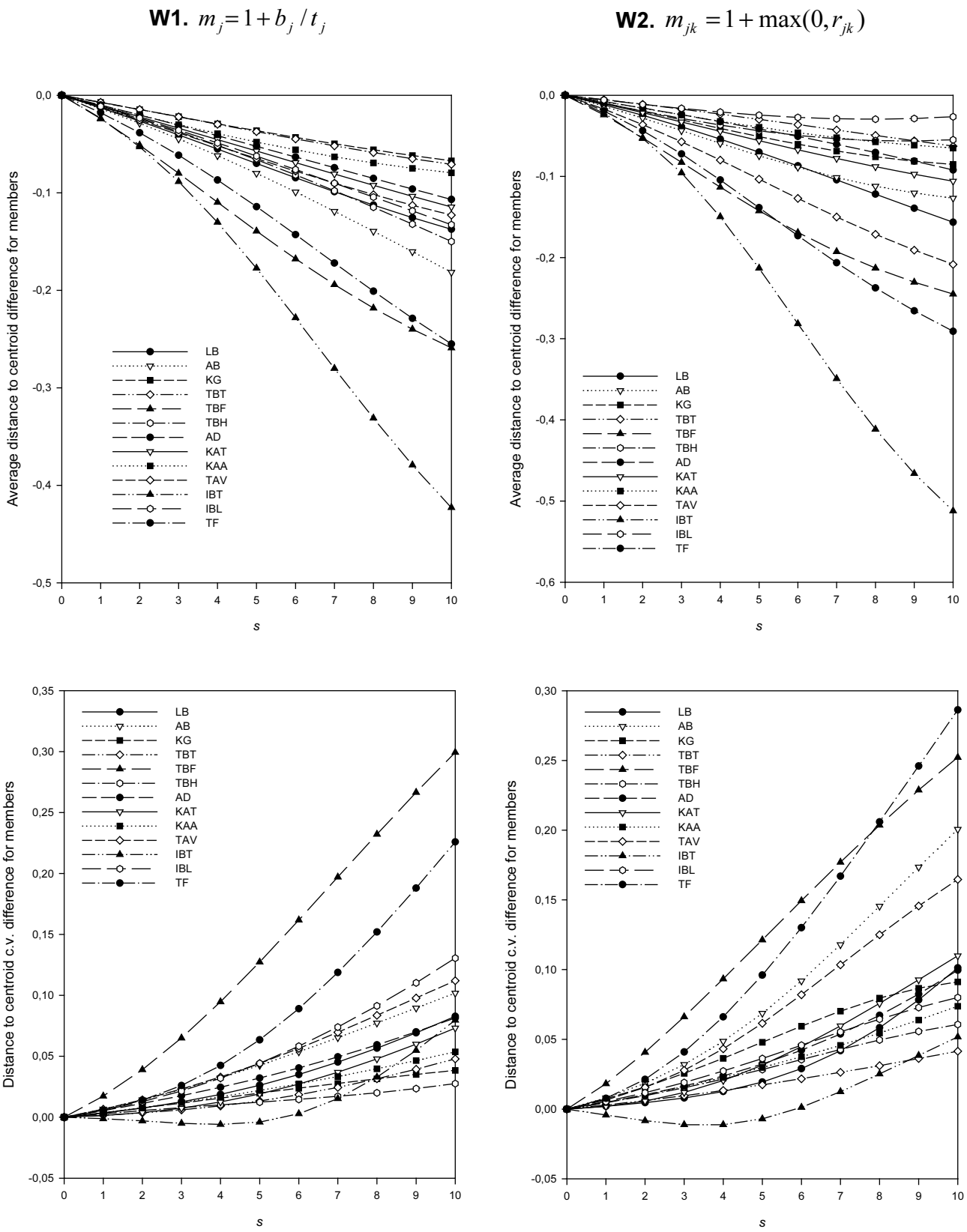


Figura 3.4.2.A: Efecte de la ponderació de variables en les distàncies de Hellinger al centroide dels **inventaris membres** dels sintàxons de base per a diferents exponents de ponderació (*s*). Diferències respecte a la situació no ponderada en la mitjana (a dalt) i el coeficient de variació (c.v, a baix) de les distàncies al centroide. Resultats del conjunt de dades de *Xerobromion erecti*.

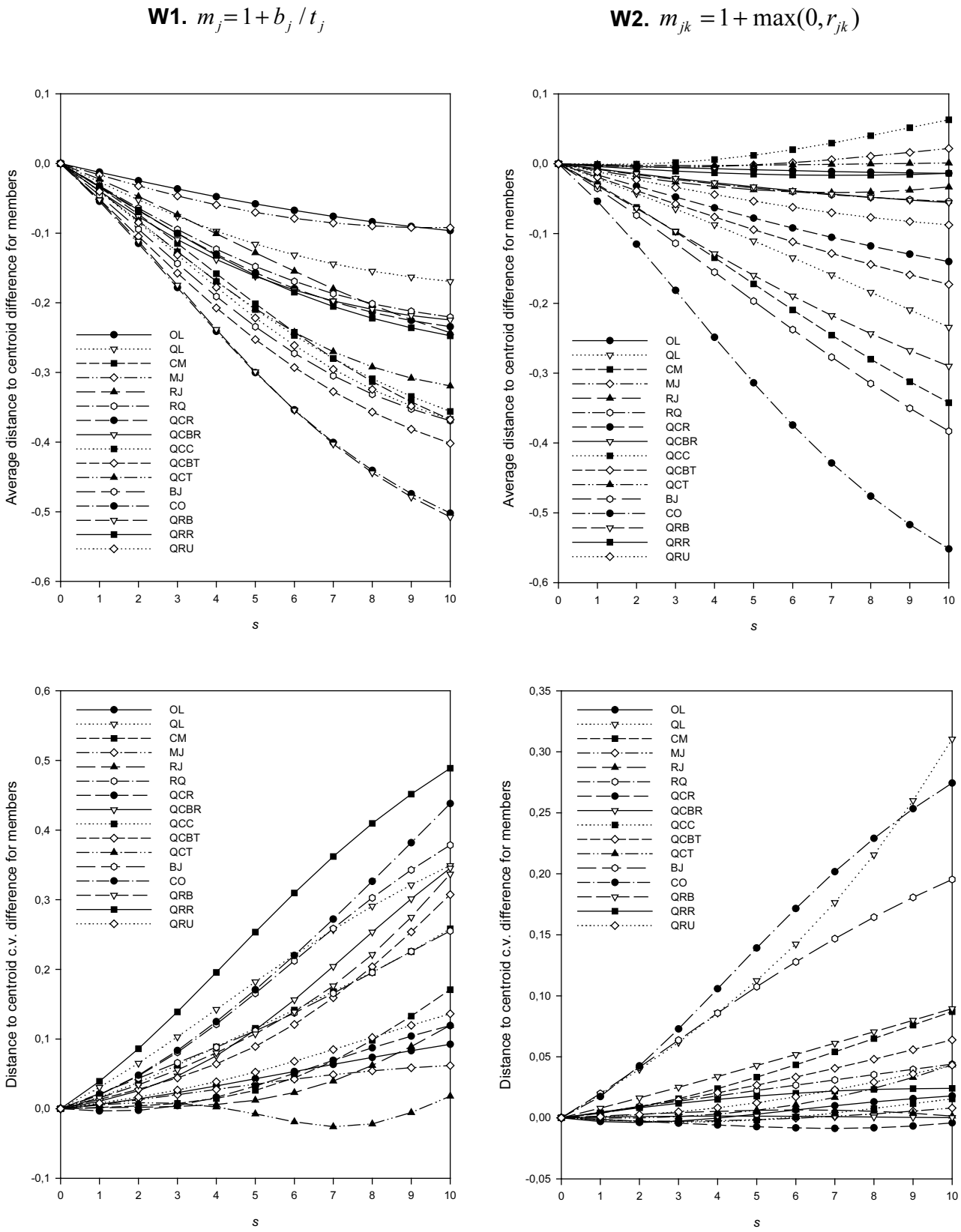


Figura 3.4.2.B: Efecte de la ponderació de variables en les distàncies de Hellinger al centroide dels **inventaris membres** dels sintàxons de base per a diferents exponents de ponderació (*s*). Diferències respecte a la situació no ponderada en la mitjana (a dalt) i el coeficient de variació (c.v, a baix) de les distàncies al centroide. Resultats del conjunt de dades de *Quercetea ilicis* sense *Quercenion*.

W1. $m_j = 1 + b_j / t_j$ **W2.** $m_{jk} = 1 + \max(0, r_{jk})$

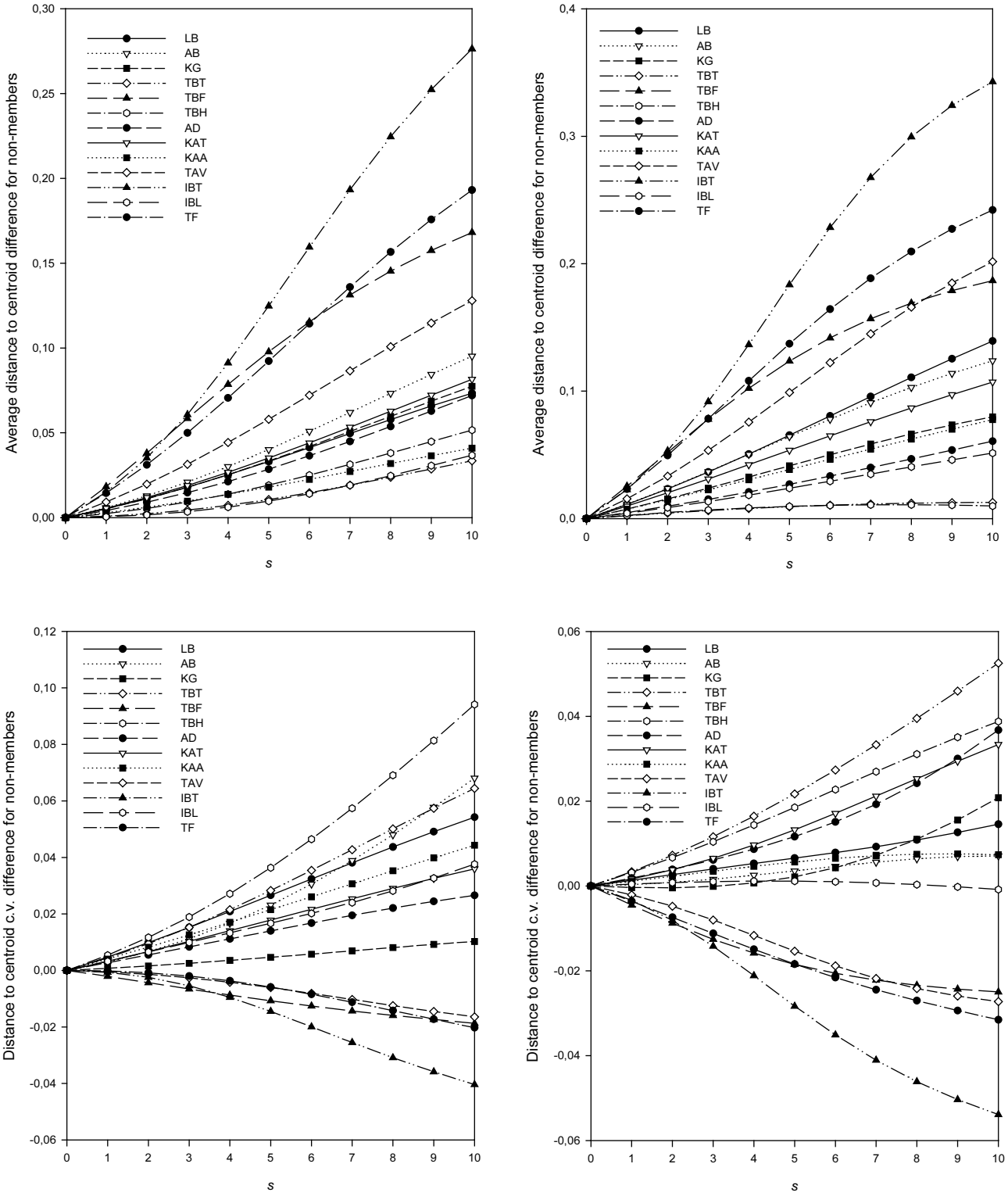


Figura 3.4.3.A: Efecte de la ponderació de variables en les distàncies de Hellinger al centroide dels **inventaris no membres** dels sintàxons de base per a diferents exponents de ponderació (*s*). Diferències respecte a la situació no ponderada en la mitjana (a dalt) i el coeficient de variació (c.v, a baix) de les distàncies al centroide. Resultats del conjunt de dades de *Xerobromion erecti*.

W1. $m_j = 1 + b_j / t_j$ **W2.** $m_{jk} = 1 + \max(0, r_{jk})$

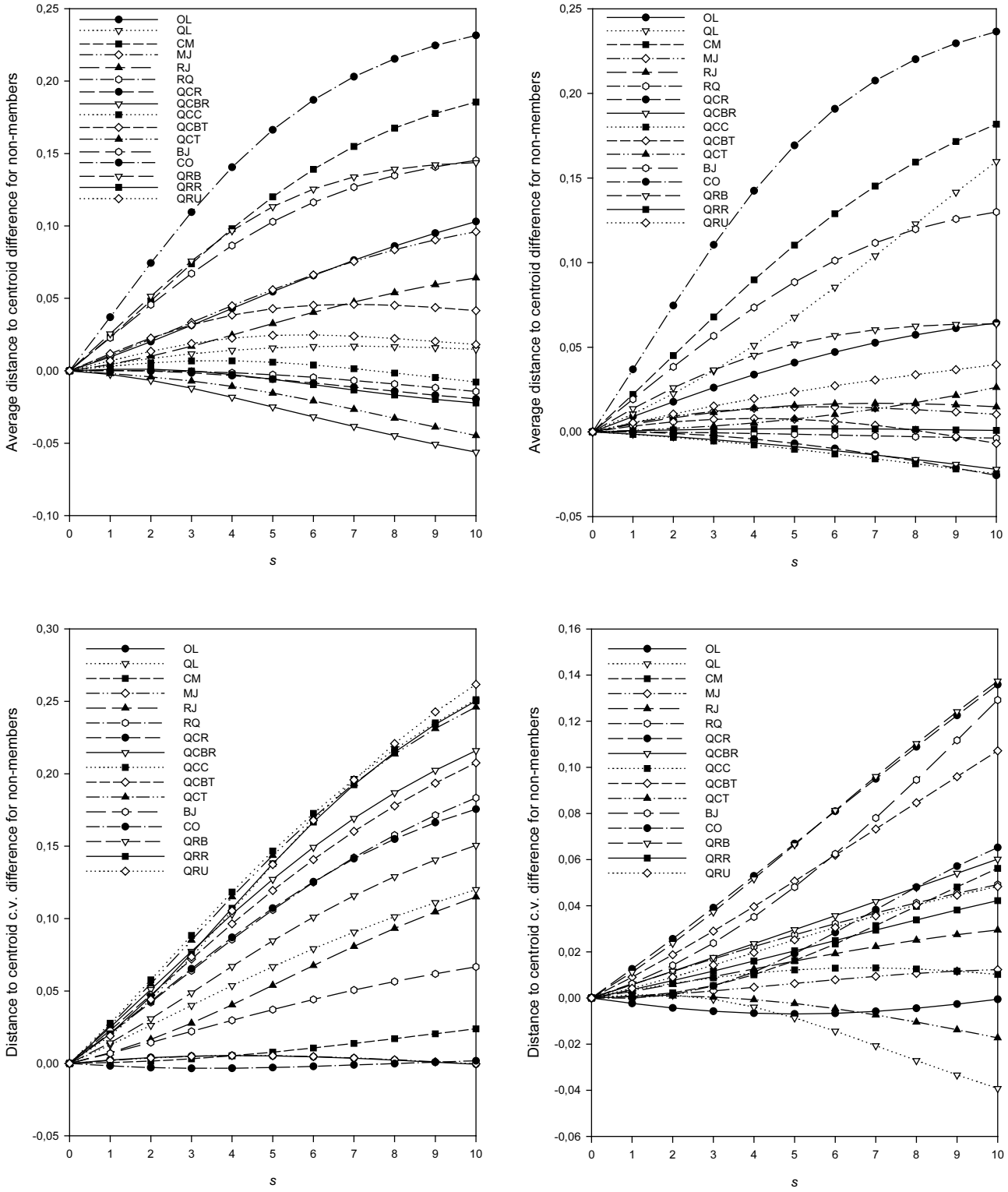


Figura 3.4.3.B: Efecte de la ponderació de variables en les distàncies de Hellinger al centroide dels **inventaris no membres** dels sintàxons de base per a diferents exponents de ponderació (*s*). Diferències respecte a la situació no ponderada en la mitjana (a dalt) i el coeficient de variació (c.v, a baix) de les distàncies al centroide. Resultats del conjunt de dades de *Quercetea ilicis* sense *Quercenion*.

La distància al centroide dels inventaris no membres

Les figures 3.4.3.A i 3.4.3.B mostren els resultats anàlegs a les figures 3.4.2.A i 3.4.2.B, avaluats en aquest cas per als inventaris externs als grups. La tendència de les distàncies al centroide mitjanes és contraria al cas anterior: L'efecte dels pesos provoca, en general, l'augment progressiu de les distàncies dels inventaris externs, incrementant l'aïllament del clúster.

Malauradament, l'augment observat del coeficient de variació en molts sintàxons fa pensar que hi ha inventaris externs que poden disminuir la distància al centroide del grup, tot i que la tendència mitjana sigui l'augment. Les excepcions on disminueix el c.v. són, per exemple, TF (*Teucrio-Festucetum*) i QL (*Querco-Lentiscetum*), aquest darrer només per a la ponderació W2.

Com a resum de les figures mostrades fins ara podem afirmar que, si bé l'aïllament general dels clústers augmenta sota l'efecte de la ponderació de les variables, l'augment paral·lel del coeficient de variació de les distàncies al centroide fa pensar que els inventaris intermedis (els problemàtics de classificar) poden veure's perjudicats per la ponderació. Per tant, els resultats posen en dubte l'efectivitat de les estratègies de ponderació pel que fa a l'augment de la discriminabilitat global de les dades.

Discriminabilitat dels sintàxons sota l'efecte de la ponderació

La variació de la correlació Φ entre els clústers originals (sintàxons tradicionals) i els sorgits de l'anàlisi discriminant basada en les distàncies ponderades, apareix a les figures 3.4.4.A i 3.4.4.B.

El descens de la majoria de perfils de les figures posa en evidència que la majoria de sintàxons disminueixen la discriminabilitat per l'efecte dels pesos. Si bé alguns sintàxons es veuen lleugerament afavorits per a exponents de ponderació baixos, per a exponents alts la correlació disminueix quasi sense excepció. Tan sols IBT i KG a *Xerobromion*, i QCBR a *Quercetea* sense *Quercenion* mantenen una lleugera millora de la correlació sostinguda a través dels exponents de ponderació creixents. Per tant, es confirmen les sospites encetades pels resultats anteriors.

Aquesta disminució de la correlació ve acompanyada lògicament d'una disminució de l'ajust global entre la partició original i la obtinguda en l'anàlisi discriminant. Les figures 3.4.5.A i 3.4.5.B mostren la variació de l'índex de Rand corregit en augmentar l'exponent *s*.

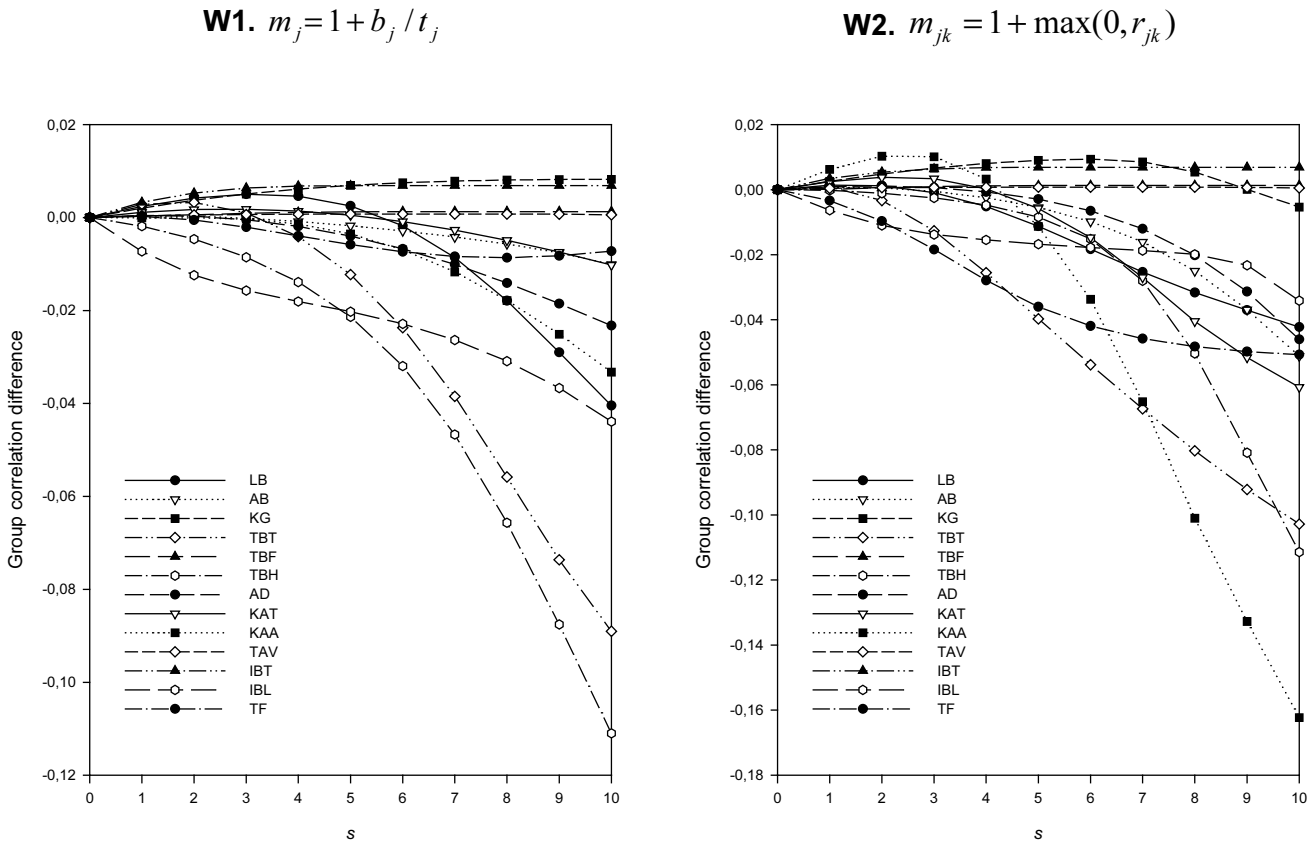


Figura 3.4.4.A: Correlació entre els grups determinats per l'anàlisi discriminant basada en les distàncies de Hellinger ponderades i els grups de la classificació tradicional per a diferents exponents de ponderació (*s*). Resultats del conjunt de dades de *Xerobromion erecti*.

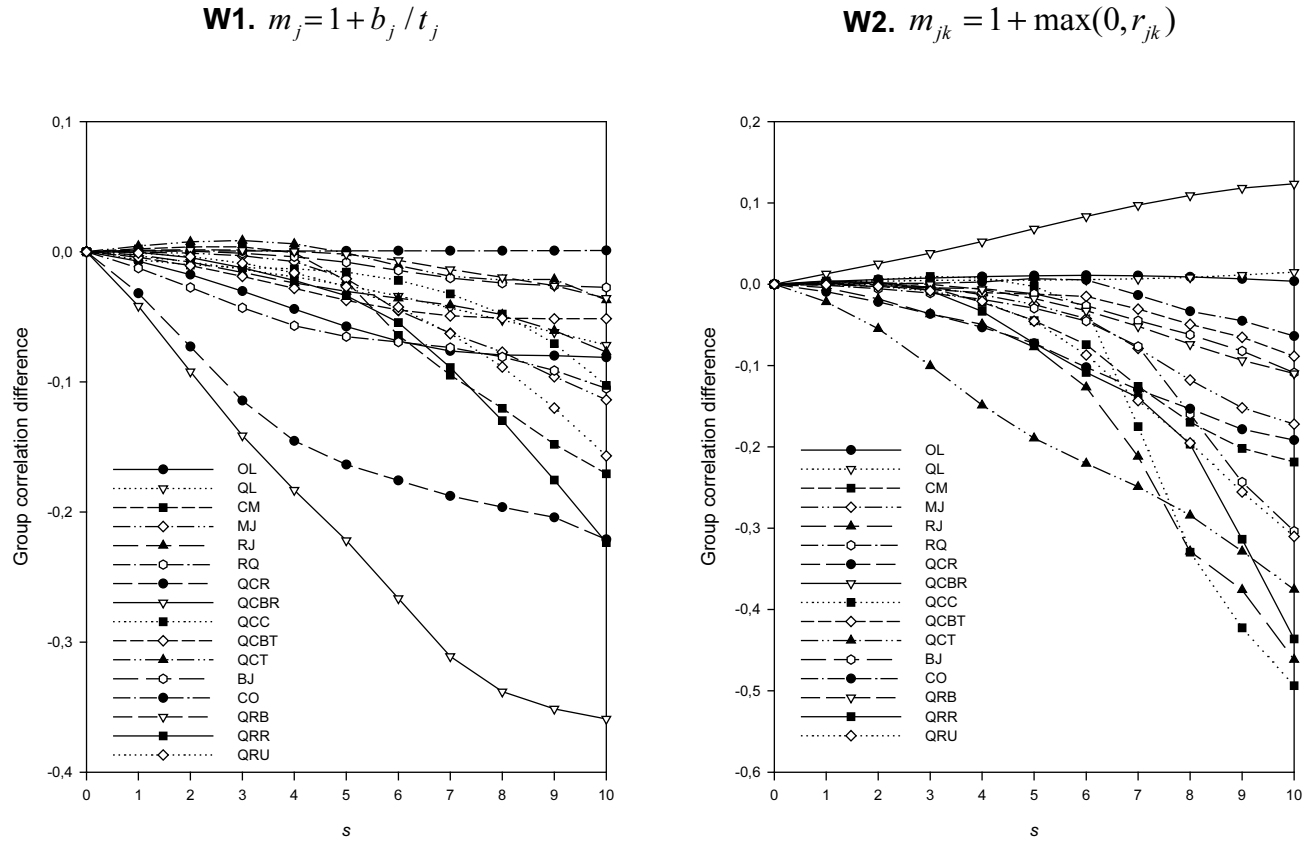
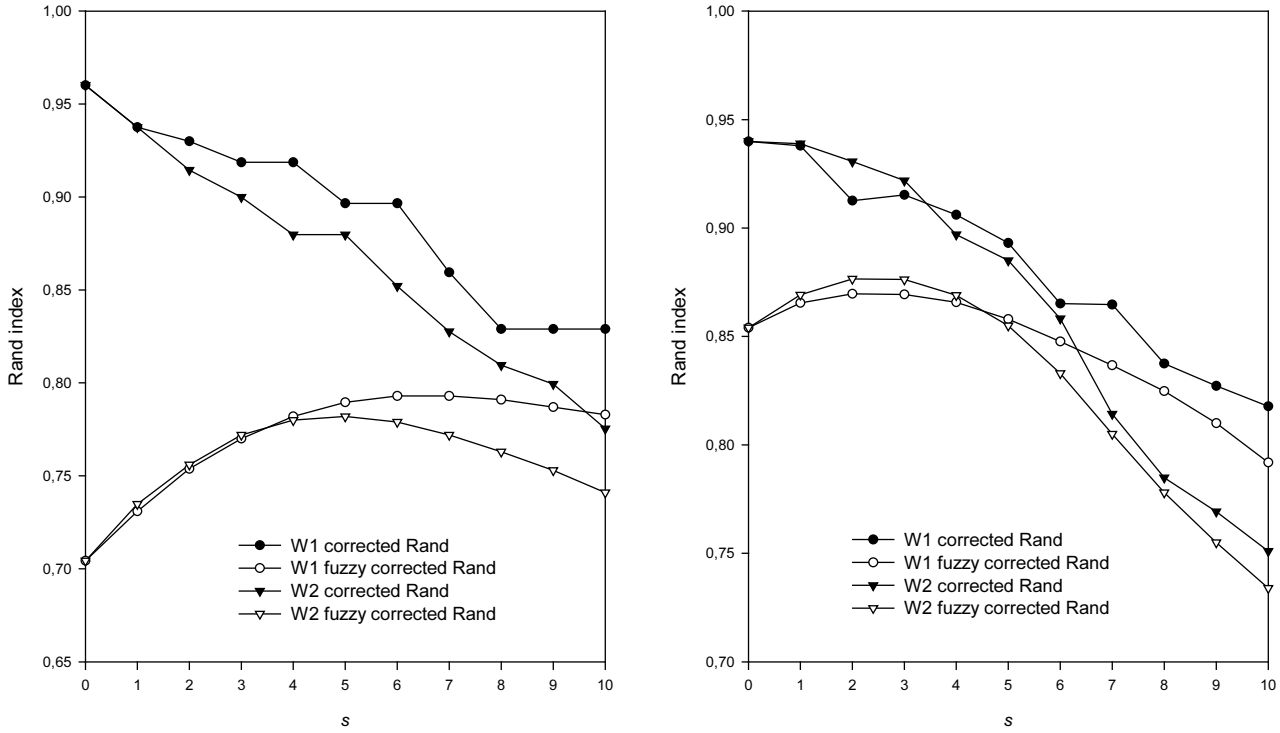


Figura 3.4.4.B: Correlació entre els grups determinats per l'anàlisi discriminant basada en les distàncies de Hellinger ponderades i els grups de la classificació tradicional per a diferents exponents de ponderació (*s*). Resultats del conjunt de dades de *Quercetea ilicis* sense *Quercenion*.

A. *Xerobromion erecti* **B. *Quercetea ilicis* sense *Quercenion***



Figures 3.4.5.A-B: Valors d'ajust dels índexs de Rand clàssic i difús corregits per l'efecte de l'atzar entre la partició fitosociològica tradicional i la obtinguda per l'anàlisi discriminant basada en distàncies de Hellinger ponderades per a diferents exponents de ponderació.

Per una banda, l'índex de Rand *crisp* mostra una tendència a disminuir en tots els casos. Per tant, la pèrdua de discriminabilitat que apuntava l'augment del coeficient de variació es veu confirmada també pels resultats de l'ajust global. Suposant, doncs, que la classificació tradicional sigui "veritable", aquesta no esdevé més discriminable numèricament amb l'ajut de les ponderacions de variables proposades.

D'altra banda, l'índex de Rand difús mostra un augment sensible per a exponents de ponderació mitjans (*s* entre 4 i 6) a *Xerobromion erecti*, i baixos (*s* al voltant de 2 i 3) a *Quercetea ilicis* sense *Quercenion*. Això és degut a que, si bé hi ha una pèrdua de discriminabilitat d'elements intermedis, aquest fet és compensat en l'índex difús per l'augment de l'aïllament general dels clústers, que fan que la "borrositat" de la partició disminueixi. Així tenim una millora de la seguretat en la determinació dels inventaris fàcils de classificar acompanyada d'una pèrdua en força casos de la discriminabilitat dels inventaris poc o molt problemàtics.

Per tal de veure més en detall quines modificacions ocorren en la classificació dels inventaris s'han calculat les diferències entre la matriu de confusió de la determinació $s = 0$ (sense pesos) i la matriu de confusió de la determinació $s = 4$. Les taules 3.4.1.A-B i 3.4.2.A-B, mostren el canvi ocorregut en la determinació dels inventaris entre les dues situacions. En una

situació de millora de la discriminabilitat amb la ponderació hom observaria valors positius en la diagonal i valors negatius en aquelles caselles on, en la situació sense ponderar, s'errava la determinació.

W1	g-1	g-2	g-3	g-4	g-5	g-6	g-7	g-8	g-9	g-10	g-11	g-12	g-13
LB	-1			1									
AB		-1		1									
KG			0										
TBT				-1		1							
TBF					0								
TBH						0							
AD							0						
KAT								0					
KAA									0				
TAV										0			
IBT											0		
IBL												0	
TF				1									-1

Taula 3.4.1.A: Variació de la taula de confusió entre la classificació tradicional original (files) i la determinació obtinguda en l'anàlisi discriminant basada en distàncies (columnes). Resultats obtinguts del *set* de dades de *Xerobromion erecti* amb l'estratègia de ponderació W1.

W2	g-1	g-2	g-3	g-4	g-5	g-6	g-7	g-8	g-9	g-10	g-11	g-12	g-13
LB	-3			3									
AB		-1		1									
KG			0										
TBT				0									
TBF					0								
TBH						0							
AD							0						
KAT								1	-1				
KAA								1	-1				
TAV										0			
IBT											0		
IBL												0	
TF				3									-3

Taula 3.4.2.A: Variació de la taula de confusió entre la classificació tradicional original (files) i la determinació obtinguda en l'anàlisi discriminant basada en distàncies (columnes). Resultats obtinguts del *set* de dades de *Xerobromion erecti* amb l'estratègia de ponderació W2.

En el cas de les pastures de *Xerobromion erecti* (taules 3.4.1.A i 3.4.2.A) partim d'una determinació sense pesos molt bona (Rand = 0.96) i pocs inventaris canvien la seva determinació per efecte dels pesos. La diagonal esta composada de valors nuls o negatius, amb algun valor positiu per al cas de KAT sota la ponderació W2. És palesa, doncs, la pèrdua de discriminabilitat general.

Les dues estratègies de ponderació augmenten la confusió en els mateixos grups: LB, AB i TF perden inventaris originalment ben classificats en pro de TBT. Recordem que TF presentava

coeficients de variació creixents en la distància al centroide dels inventaris membres. En canvi, TBT presentava coeficients de variació alts sobretot en la distància al centroide dels inventaris no membres. Són doncs, aquests inventaris que són traspassats d'un grup a l'altre per efecte de la ponderació. És important ressaltar que TBT és una comunitat propera a altres sintàxons (vegeu pàg. 85) i que és la comunitat de *Xerobromion* amb menys tàxons fidels o discriminants entre totes les estudiades (vegeu pàg 71).

W1 s=4	g-1	g-2	g-3	g-4	g-5	g-6	g-7	g-8	g-9	g-10	g-11	g-12	g-13	g-14	g-15	g-16
OL	0															
QL	2	-2	-1	1		-2		3	-1							
CM			0													
MJ				0												
RJ					-1							1				
RQ	1					-2	1									
QCR						1	-7	5	2	-1						
QCBR								0								
QCC							-1	1	0							
QCBT							1			-2	1					
QCT							-1				1					
BJ												0				
CO													0			
QRB														0		
QRR	1										-1				0	
QRU															1	-1

Taula 3.4.1.B: Variació de la taula de confusió entre la classificació tradicional original (files) i la determinació obtinguda en l'anàlisi discriminant basada en distàncies (columnes). Resultats obtinguts del *set* de dades de *Quercetea ilicis* sense *Quercenion* amb l'estratègia de ponderació W1.

W2 s=4	g-1	g-2	g-3	g-4	g-5	g-6	g-7	g-8	g-9	g-10	g-11	g-12	g-13	g-14	g-15	g-16
OL	-1			1												
QL	2	-14	1	1	1	1		1	6	1						
CM			0													
MJ			6	-6												
RJ					-1							1				
RQ	1					-5			3	1						
QCR							-8	-3	3	6						2
QCBR							1	-2		1						
QCC							-1		1							
QCBT						1	2	-1	1	-3						
QCT						1	2		1	3	-7					
BJ					4		-1					-7				4
CO										-1			1			
QRB														-4		4
QRR	1										-1			4	-7	3
QRU														6		-6

Taula 3.4.2.B: Variació de la taula de confusió entre la classificació tradicional original (files) i la determinació obtinguda en l'anàlisi discriminant basada en distàncies (columnes). Resultats obtinguts del *set* de dades de *Quercetea ilicis* sense *Quercenion* amb l'estratègia de ponderació W2.

Els matollars i garrigars de *Quercetea ilicis* (taules 3.4.1.B i 3.4.2.B), també parteixen d'una determinació sense pesos força bona (Rand = 0.94). D'altra banda, experimenten més canvis en la determinació dels inventaris que el conjunt de dades anterior.

Altra vegada hi ha força elements de la diagonal amb valors negatius. Els únics sintàxons beneficiats de la ponderació són QCT (*Quercetum cocciferae* subass. *thalictretosum*) per a W1; i QCC (*Quercetum cocciferae* subass. *callunetosum*) i CO (*Clematido-Osyrietum*) per a W2.

El sintàxon de base amb més tendència a perdre inventaris, sense guanyar-ne d'altres, és QL (*Querco-Lentiscetum*) que perd 14 inventaris amb W2. La mitjana de distàncies al centroid d'aquest sintàxon es poc variable, tan pels inventaris membres com els no membres. En canvi, el coeficient de variació augmenta per els inventaris membres i disminueix per als no membres, indicant que la tendència de QL sota els pesos és, efectivament, a perdre inventaris.

És també interessant comentar el trasvàs d'inventaris de MJ a CM. CM té força menys tàxons fidels que MJ. Aquest fet sembla indicar, com en el cas de TBT, que l'absència de tàxons fidels en un sintàxon de base provoca un efecte "atractor" per a inventaris poc típics d'altres sintàxons en un context de ponderació d'espècies. Si bé aquest efecte d'atracció ja existeix en un context sense ponderar (vegeu capítol 2.3), la ponderació d'espècies l'afavoreix.

3.4.3.4 Discussió

Els resultats obtinguts contenen aspectes que es poden considerar positius i altres negatius. D'una banda, l'aplicació de la ponderació ens fa assolir l'objectiu d'incrementar l'aïllament general dels grups. Aquest increment és cert si considerem com a descriptors les distàncies al centroide mitjanes. D'altra banda, els inventaris on manquen els tàxons considerats fidels o discriminants per a un grup, s'allunyen del centroide del seu grup i es fan més propers a altres grups. En conseqüència, observem una disminució de la capacitat de determinar correctament el grup al que pertanyen els inventaris poc típics d'un sintàxon. El grup de destí d'aquests inventaris és sovint un sintàxon amb pocs tàxons fidels.

Addicionalment, hom pot especular que sovint la presència de tàxons fidels a un inventari, va acompanyada de la semblança general que aporten la resta de tàxons. En aquest inventari, la ponderació per tàxons fidels o diferencials és probablement innecessària. La hipòtesi a formular és: Si bé la presència d'espècies discriminants i espècies fidels és un element suficient per a la determinació dels inventaris, la seva absència no és determinant del contrari. Per tant, la ponderació orientada a afavorir aquests tàxons pot perjudicar aquells inventaris que presentin un

cert ajust general de composició específica però no presentin aquests tàxons. En futurs treballs fóra interessant testar la hipòtesi que acabem de plantejar aquí.

Seguint les aproximacions a la ponderació de variables esmentades a la introducció, hem realitzat estudis de ponderació de variables sobre dades sintètiques amb l'algorisme *K-means*. Malgrat que no presentem aquí els resultats obtinguts d'aquests treballs, la nostra conclusió és que, en l'anàlisi de clústers de conjunts de dades simulades és relativament senzill aconseguir bons resultats amb la ponderació i/o selecció de variables. En canvi, l'aplicació de la ponderació i/o selecció de variables a dades reals és molt més complexa. Si sumem aquest fet a la hipòtesi esmentada en l'anterior paràgraf i la tautologia implícita en l'ús de pesos a l'anàlisi de clústers, tot plegat, fa plantejar-se la necessitat de ponderar les espècies en la classificació de comunitats. Com a mínim, si més no, posa en dubte les estratègies de ponderació assajades aquí. Queda per veure si altres estratègies de ponderació podrien ser més útils.

Dale *et al*. (1986) compararen diferents mètodes de selecció de variables en l'àmbit de la vegetació. En aquest estudi indiquen tres raons principals per desitjar seleccionar un grup d'espècies: a) La facilitat computacional. b) Reduir la distinció complexa a una distinció més simple c) Reduir l'efecte de espècies irrellevants. Dels tres arguments, el primer resulta actualment una restricció relativament poc important, i el segon argument és, si més no, allunyat de la fitosociologia. Ens centrarem, doncs, en el tercer cas, per ésser el més interessant. *Quan una espècie és irrellevant?* o *què fa una espècie interessant en un context fitosociològic?* Dale *et al*. (1986) opinen que cal distingir les espècies constants, discriminants i fidels, de les espècies "soroll". Si acceptem que la aparició de les espècies en un inventari obeeix diferents factors, alguns dels quals són fruit d'elements diferents de l'autoecologia i la interacció específica, hom podria realitzar el següent raonament: Les espècies que introdueixen més soroll són aquelles en que la complexitat de la seva dinàmica poblacional i emmascara la seva preferència ecològica. Aquest fet les converteix en poc informatives des del punt de vista de l'anàlisi de relacions sociològiques entre comunitats. Un exemple d'aquesta mena d'espècies són els briòfits, que responen sovint a condicions extremament locals, deslligades de les condicions a les que responen els cormòfits de la comunitat.

El nostre parer és que la classificació de comunitats hauria de basar-se en el conjunt dels tàxons, excloent o ponderant negativament, si de cas, aquells tàxons en que la seva presència/absència i abundància no obeís o fos fruit de causes no lligades a la ecologia predominant de la comunitat. Aquest coneixement, però, sembla difícil d'obtenir de la pròpia matriu de dades en estudi, pel que sembla inútil cercar-lo a partir de la mateixa. Fóra més pràctic realitzar estudis previs de l'autoecologia de les espècies, per tal de determinar quines d'elles haurien d'ésser excloses en l'anàlisi de la vegetació. En un àmbit més numèric, és interessant l'aproximació a la ponderació que fan, indirectament, Brisse *et al*. (1995). Aquests autors

construeixen un espai de relacions entre inventaris que es basa en les relacions de concurrència entre espècies calculades en una base de dades. Creiem que fóra interessant estudiar, en un futur, aproximacions semblants a la ponderació de tàxons.

3.4.3.5 Conclusions

Tot seguit resumim les conclusions, sobre l'efecte de la ponderació de variables en la classificació de comunitats de vegetals, a les que ens han conduït els resultats d'aquest capítol:

- És possible augmentar l'aïllament general dels clústers d'inventaris mitjançant pesos basats en la capacitat discriminant de les variables (tàxons) o la correlació entre la presència dels tàxons i la pertinença als clústers.
- Les dues estratègies de ponderació proporcionen resultats força semblants en el cas de que les variables siguin tàxons. Els conceptes de variable discriminant i variable indicadora (fidel) són similars doncs, per a aquest tipus de dades.
- Aquest augment general de l'aïllament no comporta un augment de la capacitat de determinar sintàxons de base en una anàlisi discriminant, degut a que la determinació dels inventaris amb característiques intermèdies esdevé incorrecte. Els inventaris intermedis poden tenir una semblança de composició global relativament bona però estar mancats dels tàxons més discriminants o més fidels.
- El grup de destí dels inventaris amb característiques intermèdies o poc típics és sovint un sintàxon de base poc caracteritzat florísticament.
- La classificació numèrica de comunitats de vegetals s'adiu més a la classificació tradicional quan hom pren en compte la totalitat dels tàxons per igual que mitjançant les estratègies de ponderació testades.