

CAD

Mémoires reçus

Les cahiers de l'analyse des données, tome 1, n° 3 (1976),
p. 335-338

[<http://www.numdam.org/item?id=CAD_1976__1_3_335_0>](http://www.numdam.org/item?id=CAD_1976__1_3_335_0)

© Les cahiers de l'analyse des données, Dunod, 1976, tous droits réservés.

L'accès aux archives de la revue « Les cahiers de l'analyse des données » implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
<http://www.numdam.org/>

MÉMOIRES REÇUS

. *A comparative Study of Reciprocal Averaging and Other Ordination Techniques* : H.G. Gauch, Jr., R.H. Whittaker, & T.R. Wentworth; Ecology and Systematics, Cornell University, Ithaca, N.-Y. 14853.

Dans ce mémoire, qu'ils ont bien voulu nous communiquer (47 pp; non publié) les auteurs comparent l'efficacité de l'analyse des correspondances à celle d'autres techniques communément utilisées aux E.U. pour ordonner les relevés floristiques : notamment diverses variantes de l'analyse en composantes principales (avec ou sans normalisation des lignes et des colonnes, etc ...) et la sériation bipolaire (*Polar Ordination* : technique simple quant aux calculs, qui place l'ensemble des relevés relativement à deux pôles fixes choisis *a priori* par l'utilisateur suivant certains critères). L'analyse des correspondances est appelée par eux : *reciprocal averaging*, moyennes réciproques parce qu'en bref les facteurs peuvent être caractérisés par la propriété de se reproduire par transi-tion d'un espace à l'autre (au facteur $\lambda^{1/2}$ près). Les auteurs suivent l'exposé de Hill, (dont le travail est analysé par nous au § 3.5 de notre Histoire et Préhistoire de l'analyse des données, en cours de publication dans ces cahiers); mais il semble que la *Phytosociologie* de M. Guinochet (Paris; Masson 1973), ouvrage cité par eux, soit à l'origine de leurs travaux.

Bien qu'ils soient au premier chef botanistes et non statisticiens, les auteurs ne consacrent que quelques pages aux résultats qu'ils ont obtenus en traitant des données naturelles, et insistent le plus sur les expériences auxquelles ils se sont livrés sur des tableaux engendrés suivant certains modèles. Il ne faut pas voir seulement ici une marque de l'importance prépondérante accordée aux modèles par certaines écoles statistiques : est aussi en cause une conception de l'écologie, différente de celle prônée en France par M. Guinochet dans la ligne de l'Ecole Zuricho-Montpelliéraine (dite encore sigmatiste) fondée par Braun-Blanquet. Les auteurs écrivent en effet : "In ecology, ordination refers to the arrangement of samples (of species) in relation to environmental gradients, or axes that may correspond to environmental gradients; a major purpose of such arrangement is the recognition of joint variation in community composition and environmental factors". M. Guinochet (in : *Phytosociologie*, p. 173) professe au contraire que : "Une ordonnance de parcelles de végétation, ou, plus correctement, de leurs descriptions, ne nous fournit que des rapports d'ordres entre celles-là en fonction des caractères pris en considération. Mais elle n'offre pas les mêmes possibilités de généralisation ni la même puissance de prédiction qu'une classification,..." "La grande différence entre les écoles de Wisconsin [dont les travaux sont cités par G., W., & W.] et sigmatiste est que les adeptes de la première ont un état d'esprit qui ne leur permet pas de passer à la deuxième phase, plus synthétique, de détermination explicite de groupes à partir de l'ordination". Sans méconnaître l'existence de

transitions continues en fonction e.g. de l'altitude (*op. laud.* p. 23) ni "la supériorité de la composition floristique globales sur les espèces indicatrices" (p. 98), M. Guinochet recherche à la fois des discontinuités entre les relevés (relevant d'associations différentes) et une hiérarchisation dans la valeur indicatrice des espèces (auxquelles est dévolu en écologie un rôle analogue à celui que jouent les caractères en taxinomie animale et végétale). De semblables recherches on ne trouve pas l'écho dans le mémoire analysé ici : comme données naturelles, les auteurs considèrent principalement des suites de relevés effectués sur un *transect* (ou coupe : e.g. 49 relevés de 1830 à 2140 m dans l'Arizona; etc.): et les données qu'ils simulent sont aussi des séries de relevés gradués suivant une ou deux dimensions (un segment de droite, ou un rectangle) en mettant en jeux des facteurs d'hétérogénéité et des perturbations aléatoires d'amplitude plus ou moins grande.

Ces prémisses étant posées l'objectif pour les auteurs, consiste en ce que (dans le cas unidimensionnel) : "Given this one-dimensional input, an ideal ordination should recover this structure in evenly spaced sample positions on the first ordination axis and should have no sample displacements into higher dimensions". Commentons ce programme : à des relevés afférents à des lieux équidistants, doivent correspondre des points figuratifs équidistants (evenly spaced) sur le diagramme, plus précisément sur le 1^o axe; sans effet Guttman (ligne parabolique dans le plan 1 x 2; et au delà, facteurs dépendant polynomialement du premier : sample displacements into higher dimensions); et surtout sans *involution*, par quoi les auteurs entendent : repli des extrémités de la ligne en projection sur le premier axe.

Dans l'ensemble les résultats obtenus par les auteurs sont nettement favorables à l'analyse des correspondances. Citons d'abord quelques unes de leurs conclusions générales, avant de discuter des objections de détail auxquelles une expérience approfondie de l'analyse des données permet de répondre.

"Principal components analysis in its most familiar form (centred; non-standardized or sample-standardized) is relatively unreliable for ordination ..." "We have not identified sample set properties for which PCA [composantes principales] gives better ordination than R.A.". [Reciprocal a.; i.e. a. des correspondances]...

"Polar ordination is a simple and in some respects robust technique, generally free from involution and little vulnerable to distortion by clusters and outliers" [nous reviendrons sur ces points]... "The technique's major limitation is its requirement for endpoint choice and vulnerability of the ordination to effects of different endpoint choices"...

"Reciprocal averaging or correspondence analysis employs a simultaneous double standardization and is resistant to involution... R.A. has major advantage over non standardized PCA and at least marginal advantage over standardized PCA in its lower vulnerability to distortion by sample clusters, sample errors, and outliers ... R.A. produces simultaneous sample and species ordination on axes that may be treated as coordinates... and its species ordinations are superior to those of PO [Polar ordination] and PCA".

Entrons maintenant dans quelques détails. Parmi les perturbations envisagées par les auteurs, on note outre les fluctuations ordinaires ce qu'ils appellent *clusters* et *outliers*. Parfois la densité des relevés (non tant sur le terrain; que sur l'espace typologique idéal où ils s'ordonnent suivant leurs ressemblances) n'est pas uniforme : il y a des grappes (clusters) ici; des points isolés là. A la vérité nous avons dit plus haut que pour M. Guinochet, cette discontinuité n'est pas accidentelle, mais essentielle. De plus certains relevés peuvent occuper des positions tout à fait excentriques (outliers) : dans les simulations les auteurs introduisent de tels relevés soit (a) en mêlant

au hasard des notes (abondances) effectivement attestées dans les relevés figurant déjà au tableau des données, soit (b) en choisissant au hasard une espèce dominante. Le résultat est que "RA was robust against effects of type "a" outliers : they tended to ordinate around the center of the ordination field but caused negligible displacement of the other samples. Type "b" outliers, in contrast, ordinate around the periphery of the ordination field and strongly affect positions of the other samples". A propos de quoi nous ferons deux remarques. 1°) Du point de vue de M. Guinochet confirmé par les recherches de A. Lacoste, M. Roux etc., la présence d'une espèce est un caractère essentiel à l'individu d'association; tandis que son abondance est un caractère contingent; l'on doit donc analyser plutôt des tableaux en présence-absence (tableaux booléens en 0-1) que des tableaux d'abondance : du coup, les individus aberrants du type "b" rentrent dans le type "a". 2°) Il est commun en analyse des données, que des individus excentriques fassent à eux seuls un voire plusieurs axes et brouillent la représentation de l'ensemble : mais la consultation des tableaux de contribution (et dès l'abord la seule vue des graphiques) impose de mettre ces individus en éléments supplémentaires.

Tout en reconnaissant l'intérêt éventuel d'un second facteur fourni par l'analyse des correspondances "RA will in some cases reveal a second direction of sample variation in its second axis", les auteurs trouvent dans l'interprétation des facteurs d'ordre supérieur des difficultés qu'ils attribuent principalement à l'effet Guttman. "in other cases this direction is obscured by distortion in the second axis, and deferral of secondary gradients, or sample variation into higher axes" (i.e. s'il y a deux dimensions dans les données, la première peut occuper, e.g., deux axes; et la deuxième n'apparaître que sur le troisième axe). "Given the curvilinearity and other characteristics of community data, RA can no more than PCA be trusted to produce several independent axes as ecologically significant, potentially interpretable directions of community and environmental variation". Que répondre à cela ? Reconnaissons d'abord que l'interprétation successive de nombreux facteurs bien que souvent réalisée en analyse des données, est chose délicate. Mais d'une part les résultats observés par les auteurs correspondent aux données particulières auxquelles ils s'intéressent : lesquelles sont, répétons-le, des gradations quantitatives sur le terrain exprimées par l'abondance des espèces, plutôt que des oppositions qualitatives révélées par la présence ou l'absence de caractéristiques; d'autre part l'effet Guttman ne nous apparaît pas comme étant seulement fâcheux. En effet plusieurs dimensions peuvent être requises pour exprimer adéquatement l'étalement des faits sur un continuum sous-jacent unidimensionnel : il est classique de distinguer sur le deuxième axe, parmi les individus centraux, entre ceux qui ne s'écartent en rien du centre (sommet de la parabole) et ceux qui réalisent une conjonction des tendances extrêmes (centre du croissant parabolique) : cette interprétation a été vue dès le départ par L. Guttman.

Nous concluons ce compte-rendu, en remerciant les auteurs de nous avoir communiqué, avant sa publication, un intéressant mémoire qui contribuera à mettre en lumière les présupposés propres aux différentes écoles; écoles dont la diversité accélère le progrès de la recherche et en garantit l'objectivité, mais entre lesquelles des acquisitions définitives doivent être fixées, comme un héritage commun à ceux qui, s'il plait à Dieu, nous suivront.

Notice d'un programme d'analyse des correspondances; par A. Fontaine C.E.R.I.A. Centre d'Etudes et de Réalisation en Informatique Appliquée. 25, rue Vergniaud 75013 - Paris.

A. Fontaine propose un programme d'analyse factorielle des correspondances qui présente deux particularités :

- il est écrit en langage APL
- il s'exécute sur un microordinateur

Le matériel utilisé est un micro-ordinateur de type MCM-70 fabriqué par la firme canadienne MCM. Ce microcalculateur est construit autour d'un microprocesseur du type "8008 Intel" et dispose d'un interpréteur APL microprogrammé.

La mémoire centrale est de 8 K octets et on dispose de deux dérouleurs de cassettes magnétiques permettant de stocker deux fois 110 K octets sur des cassettes type 3M.

Connectée à ce calculateur on trouve une imprimante du type "DIABLO HYTYPE" permettant une impression de 30 caractères par seconde sur 132 caractères de large. La particularité de cette imprimante est que l'on peut déplacer la tête d'impression dans les quatre directions : avant, arrière, haut, bas.

La lenteur relative et la faible capacité de stockage ainsi que quelques restrictions dues au type d'APL utilisé ne permettent pas des performances élevées et on ne pourra traiter que de petits tableaux, le maximum prévu étant 256 individus x 30 variables; mais une version en accès direct permettrait d'accroître à volonté le nombre des individus traités.

L'algorithme utilisé est classique mais son aspect conversationnel permis par l'utilisation d'APL est d'un emploi assez agréable.

Les calculs sont automatiquement effectués en double précision ceci garantissant la fiabilité de la méthode de diagonalisation utilisée (Hotelling).