LES CAHIERS DE L'ANALYSE DES DONNÉES

PH. NABHAN

Problème proposé sans sa solution

Les cahiers de l'analyse des données, tome 19, n° 3 (1994), p. 377-380

<http://www.numdam.org/item?id=CAD_1994__19_3_377_0>

Les Cahiers de l'Analyse des Données
Vol. XIX - 1994 - n°3 - pp.377 380.

PROBLÈME PROPOSÉ SANS SA SOLUTION

[PROBLÈME]

*Ph. NABHAN**

L'objet de la présente note est de proposer, en guise d'énigme, un résultat, semblant paradoxal, qui nous a arrêté quelque temps dans une application de la formule de reconstitution, à une colonne supplémentaire. Nous nous proposons de publier, prochainement, la solution que nous avons trouvée à l'énigme (cf. [PLANCTON LIBAN]); avec l'espoir que certains lecteurs auront, par eux-mêmes, retrouvé ou amélioré cette solution.

1 Rappel sur la reconstitution des données en fonction des facteurs

La formule de reconstitution des données en fonction des facteurs est l'une des premières considérées dans l'étude de l'analyse des correspondances. Elle s'écrit de diverses façons, selon qu'y figurent les facteurs F_α, de variance λ_α, ou les facteurs normalisés φ_α; et que l'on prend comme quantité à reconstituer une valeur brute k(i, j), ou une composante d'une densité, d'un profil.

Voici quelques exemples de ces écritures diverses. Convenons de noter:

$$R(i, j) = (1 + (1/\sqrt{(\lambda_1)}).F_1(i).F_1(j) + (1/\sqrt{(\lambda_2)}).F_2(i).F_2(j) + \ldots)$$
$$= (1 + \sqrt{(\lambda_1)}.\varphi_1(i).\varphi_1(j) + \sqrt{(\lambda_2)}.\varphi_2(i).\varphi_2(j) + \ldots) \quad ;$$

alors on a, en marquant par le caractère '≈' une approximation qui dépend de l'étendue de la somme R, et sur laquelle nous reviendrons dans la suite:

$$k(i, j) \approx (k(i).k(j)/k) \,.\, R(i, j) \quad ;$$
$$(k(i, j).k) / (k(i).k(j)) = f_{ij} / (f_i.f_j) \approx R(i, j) \,;$$
$$k(i, j)/k(i) = f_j^i \approx (k(j)/k) \,.\, R(i, j) = f_j \,.\, R(i, j) \,;$$
$$k(i, j)/k(j) = f_i^j \approx (k(i)/k) \,.\, R(i, j) = f_i \,.\, R(i, j) \,.$$

La première formule veut exprimer un terme, k(i, j), en fonction de la somme R, des valeurs de marges, k(i), k(j), et du total général, k, du tableau de base. La deuxième formule interprète la somme R comme la densité de la

(*) Université Libanaise: Faculté des Sciences ; Fanar, Mont-Liban, Liban.

Les cahiers de l'analyse des données 0339 3097/94/03/04/$ 4.00/© Gauthier-Villars

loi f_{IJ} par rapport au produit $f_I \times f_J$ des lois marginales; la troisième lie f_J^i, profil sur J de la ligne i, à R et à la loi marginale f_J sur J; et la quatrième fait de même pour le profil f_I^j de la colonne j et la loi f_I.

Ci-dessous, on introduit le cardinal de l'ensemble I:

$$k(i, j) = (k(i).cardI/k) \,.\, (k(j)/cardI) \,.\, R(i, j) \quad .$$

Ainsi, le terme k(i, j) apparaît comme le produit de trois termes interprétables:

A) le premier terme, (k(i).cardI/k), rapporte le total k(i) de la ligne i au quotient (k/cardI), qui serait le total de chacune des lignes si celles-ci avaient toutes le même poids: et si, du moins, les lignes ont même ordre de grandeur, le terme (k(i).cardI/k) s'écarte peu de 1;

B) le deuxième terme, (k(j)/cardI), est la moyenne arithmétique des nombres k(i', j) inscrits dans la colonne j du tableau de base; si ces nombres sont tous du même ordre de grandeur, chacun s'écarte peu de (k(j)/cardI).

C) le troisième terme R(i, j) exprime, en fonction des facteurs, dans quelle exacte mesure k(i, j) s'écarte de ce qui serait sa valeur si, en A) et B), au lieu de "même ordre de grandeur", on avait "égalité" stricte.

Précisons maintenant l'approximation obtenue.

La formule de reconstitution est une égalité stricte si i et j sont des éléments principaux, et que la somme R(i, j) s'étend à tous les facteurs extraits. Si l'on ne prend dans R(i, j) qu'une partie des facteurs, la reconstitution n'est qu'approchée. Dans les cours, figure un résultat précis qui lie la qualité de l'approximation pour l'ensemble d'une colonne j (ou, de même, pour une ligne i), à ce que les listages notent QLT(j), i.e. à la somme des $COR_\alpha(j)$ étendue aux indices α retenus dans la somme R(i, j).

Dans la pratique, la formule de reconstitution offre un intérêt particulier quand on l'applique à une colonne j_s mise en supplément dans l'analyse: car alors, en bref, la formule exprime la fonction $k(i, j_s)$ de i, comme une combinaison linéaire des facteurs $F_\alpha(i)$, multipliée par la densité marginale k(i)/k, calculée sur le tableau principal; et l'on peut dire, en un certain sens, que la colonne supplémentaire est exprimée en fonction du tableau de base. Toutefois, de façon précise, la qualité de l'approximation est limitée par la somme, QLT, des $COR_\alpha(j_s)$; somme ≤ 1; et ne valant 1 que si le profil de j_s est combinaison linéaire des profils des colonnes principales j.

2 Résultat surprenant d'un essai de reconstitution

Ceci étant rappelé, voici le résultat qui nous a arrêté.

Les 48 lignes du tableau principal sont des relevés écologiques mensuels de plancton, suivant un ensemble J de six groupes d'espèces: cardI = 48 ; cardJ = 6. La colonne supplémentaire, j_S, considérée est la salinité, sal. Avec, pour les 5 facteurs que produit l'analyse, les résultats suivants.

SIG	QLT	PDS	INR	F1	COR	CTR	F2	COR	CTR	F3	COR	CTR	F4	COR	CTR	F5	COR	CTR
sal	821	153	144	-67	266	93	-81	393	177	-11	8	8	-49	144	196	-12	10	178

D'après ces résultats, on attend une bonne représentation de sal(i)=k(i, sal), en fonction des facteurs 1, 2 et 4; avec pour qualité:

QLT = 266+393+144 = 803.

De façon précise, notons:

salm = (1/48).$\sum${sal(i) | i=1,...48};

pond(i) = 48.k(i)/k ; où k(i) et k sont calculés sur les 6 colonnes principales;

$$Rs(i) = 1 + (-0{,}067/\sqrt{(\lambda_1)}).F_1(i) + (-0{,}081/\sqrt{(\lambda_2)}).F_2(i) + (-0{,}049/\sqrt{(\lambda_4)}).F_4(i) ;$$

sal'(i) = pond(i).salm.Rs(i) ;

*sal'(i) = salm.Rs(i) ; *sal(i) = sal(i)/pond(i) .

Ceci posé, *sal' apparaît comme une très bonne approximation de *sal; on a, conformément au théorème classique (la corrélation étant calculée avec pour pondération les k(i)):

corr(*sal'(i), *sal(i) = $\sqrt{}$(.803) = .896 ;

mais, pour la salinité elle-même, l'approximation, mesurée en terme de corrélation, ne vaut rien; car on a:

corr(sal'(i), sal(i) = -.187 ;

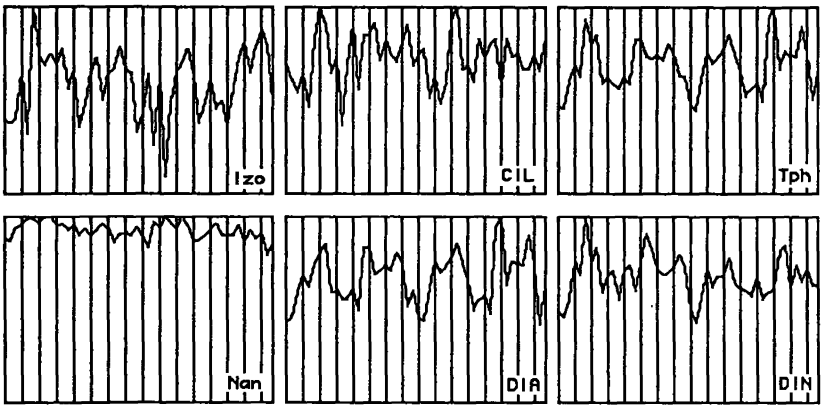
le résultat ne s'améliorant pas si l'on calcule corr avec pondération uniforme.

3 Le problème

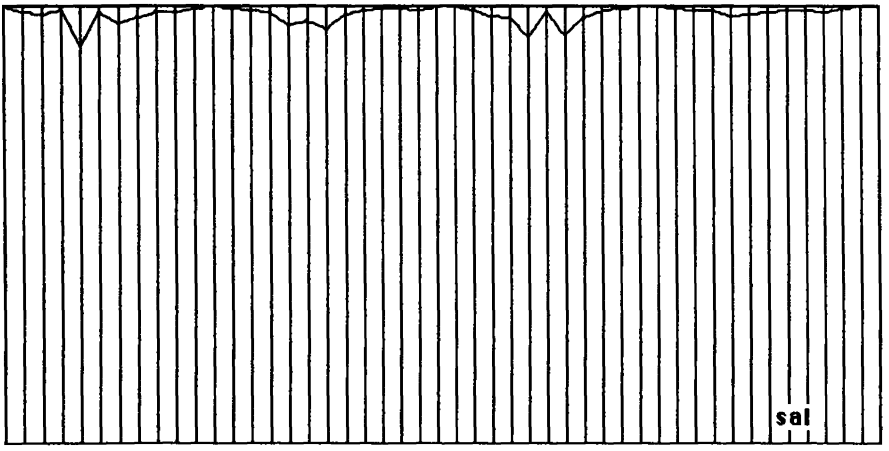
La question que nous posons est donc la suivante:

Par quelle particularité des données s'explique cette faible valeur de corr; laquelle, sans contredire le théorème classique, atteste que l'estimation de la salinité ne vaut aucunement ce que la valeur de QLT laissait espérer.

Et encore: trouver un tableau très simple, ayant trois lignes, deux colonnes principales et une supplémentaire, et présentant le même paradoxe.



Afin d'aider le lecteur dans sa recherche, nous publions, sous forme graphique, les séries chronologiques que sont les six colonnes du tableau des 48 relevés; avec, à plus grande échelle, la série des salinités.



Référence bibliographique

M. ABBOUD-ABI SAAB, Ph NABHAN: "Interrelations entre les différents groupes planctoniques dans les eaux côtières libanaises"; [PLANCTON LIBAN]; à paraître dans *CAD*, Vol.XIX, n°4.