LES CAHIERS DE L'ANALYSE DES DONNÉES

J.-P. BENZÉCRI
T. K. GOPALAN

Sur l'application des méthodes multidimensionnelles à une anthologie de données. (2) Formes vivantes dans leur milieu

Les cahiers de l'analyse des données, tome 18, n° 4 (1993), p. 447-454

<http://www.numdam.org/item?id=CAD_1993__18_4_447_0>

Les Cahiers de l'Analyse des Données
Vol. XVIII 1993 n°4 pp.447-454.

SUR L'APPLICATION DES MÉTHODES MULTIDIMENSIONNELLES À UNE ANTHOLOGIE DE DONNÉES

(2) FORMES VIVANTES DANS LEUR MILIEU

[MÉTH. ANTH. DONNÉES (2)]

J.-P. BENZÉCRI
T. K. GOPALAN

2 Les formes vivantes dans leur milieu

2.1 Répartition génique dans seize colonies d'une espèce de papillons

S.W. McKECHNIE, P.R. EHRLICH, R.R. WHITE: "Population genetics of Euphydryas butterflies; I: Genetic variation and the neurality hypothesis"; *Genetics*, Vol 81, pp.571-594; (1975).

Les données concernent 16 colonies de papillons - appartenant à l'espèce *Euphydryas editha* - réparties géographiquement sur le territoire de l'État de Californie (à l'exclusion de la partie de cette province conservée par le Mexique): depuis {UO, LO} situées à la frontière du Mexique; jusqu'à MC, située vers le Nord de l'État; et SS comprise dans l'Oregon, mais à peu de distance de la frontière Nord de la Californie.

Les données constituent deux tableaux distincts, papC et papß, concernant, respectivement, le milieu climatique et la biochimie. Quant au climat, les quatre variables {HH PP TM tm} donnent l'altitude (en pieds), la pluviométrie annuelle (en pouces), les températures Maxima et minima annuelles (dans l'échelle Farenheit).

Les données biologiques sont relatives à une enzyme: la phosphoglucose isomérase. On a, par électrophorèse, déterminé, pour chaque colonie, une répartition en six pics consécutifs (notés par nous {µ4 µ6 µ8 µa µc µd}), qui sont considérés comme les % de six allèles distincts d'un gène.

Euphydryas editha :
phosphoglucose isomerase

6	µ4	µ6	µ8	µa	µc	µd
SS	0	3	22	57	17	1
SB	0	16	20	38	13	13
WSB	0	6	28	46	17	3
JRC	0	4	19	47	27	3
JRH	0	1	8	50	35	6
SJ	0	2	19	44	32	3
CR	0	0	15	50	27	8
UO	10	21	40	25	4	0
LO	14	26	32	28	0	0
DP	0	1	6	80	12	1
PZ	1	4	34	33	22	6
MC	0	7	14	66	13	0
IF	0	9	15	47	21	8
AF	3	7	17	32	27	14
GH	0	5	7	84	4	0
GL	0	3	1	92	4	0

Euphydryas editha :
l'environnement

4	HH	PP	TM	tm
SS	50	43	98	17
SB	80	20	92	32
WSB	57	28	98	26
JRC	55	28	98	26
JRH	55	28	98	26
SJ	38	15	99	28
CR	93	21	99	28
UO	65	10	101	27
LO	60	10	101	27
DP	150	19	99	23
PZ	175	22	101	27
MC	200	58	100	18
IF	250	34	102	16
AF	200	21	105	20
GH	785	42	84	5
GL	1050	50	81	<0

On voit que les données sont restreintes. Mais, en biologie, on a, du moins, un tableau homogène de profils; qu'il s'impose de soumettre, tel quel, à l'analyse des correspondances.

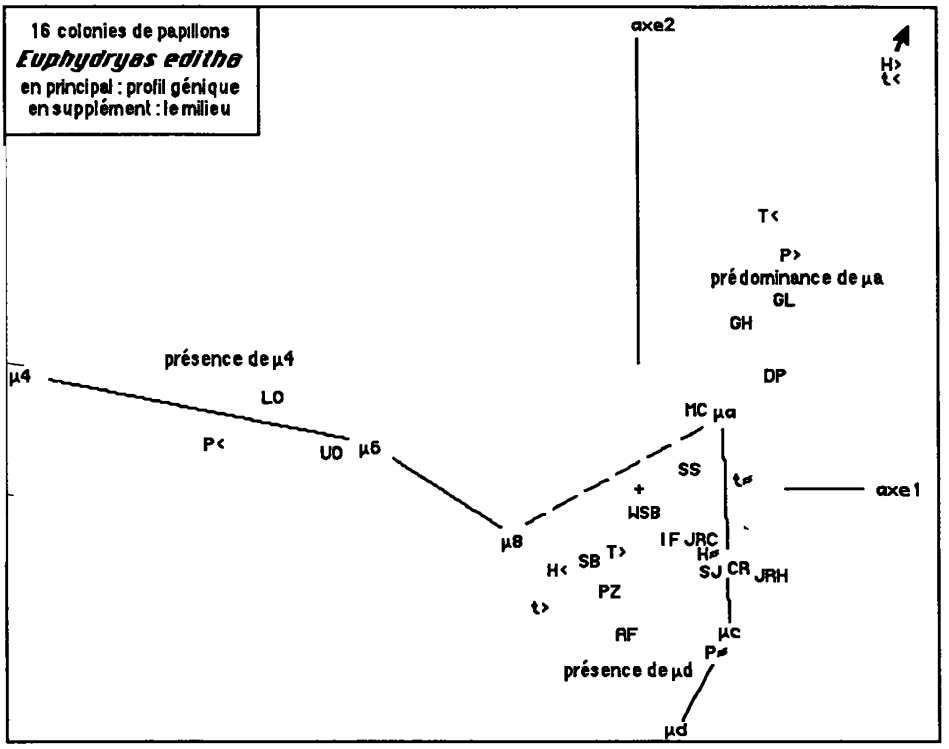
Les variables climatiques, ne couvrant même pas une région, ne peuvent fournir que des éléments supplémentaires. On les a découpées en classes, non sans hésitations car la distribution en est sporadique. Dans chaque modalité, rentre un groupe de colonies: e.g., dans P≈: pluviométrie entre 15 et 28 pouces, {DP SB CR AF PZ WSB JRC JRH}; ce niveau - ≈500mm - caractérise un climat aride, mais non désertique.

En toute rigueur, on doit associer à chaque modalité une ligne de cumul des profils biologiques de ses colonies, et d'adjoindre cette ligne, en supplément, à l'analyse du tableau papß des profils des colonies individuelles. Ici on a adjoint une colonne supplémentaire, en (0,1), caratérisant la modalité: ce qui amplifie les valeurs des facteurs dans le rapport (1/√lambda). En somme, le tableau en (0,1), à 11 colonnes, issu du découpage de papC, a été adjoint à l'analyse de papß.

Euphydryas editha : phosphoglucose isomérase

trace	:	4.095e-1				
rang	:	1	2	3	4	5	
lambda	:	2328	1189	339	187	53	e-4
taux	:	5685	2903	827	456	130	e-4
cumul	:	5685	8588	9415	9870	10000	e-4

Les résultats de l'analyse factorielle sont présentés sur le plan (1,2), où figurent les trois ensembles: J, 6 allèles; I, 16 colonies (profils sur J); Js, 11 modalités du climat (traduites par des fonctions en 0,1 sur I). De plus, on publie une CAH de I.

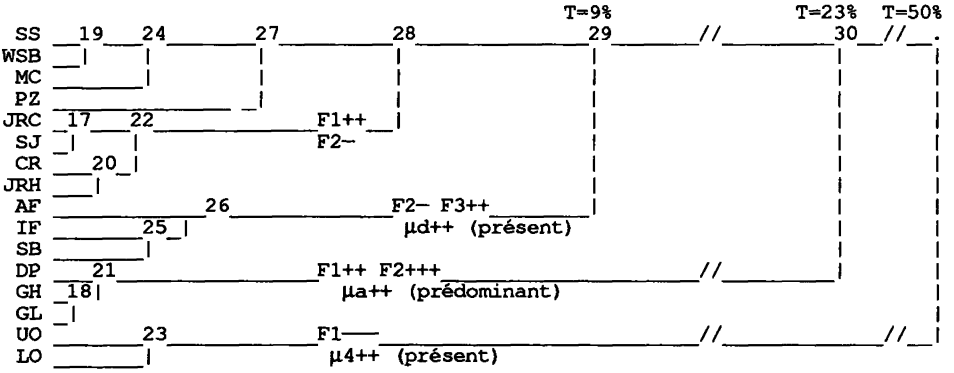


L'analyse confirme, en les ordonnant dans une vue d'ensemble, les remarques qui s'imposent à qui frappe les données, en les considérant attentivement. L'allèle µ4 n'est présent notablement que dans deux colonies, {UO LO}; µ6 a un profil voisin de µ4, mais moins contrasté. À l'extrémité opposée de l'échelle de mobilité électrophorétique, µd a un profil où prédominent {AF IF SB}. Les allèles moyens {µ8, µa, µc}, ont les plus grands poids en %; mais seul µa peut atteindre 2/3 voire 4/5, dans {DP GH GL}.

L'axe 1 est créé par l'association entre {UO LO} et µ4 (F1<0); là se projette P<, pluviométrie minima.

Sur l'axe 2, on a {DP GH GL} et µa du côté (F2>0); avec {H> t<}: {GH GL} sont en montagne, avec les températures minima les plus basses. S'écartent sur (F2<0), mais également sur (F3>0), {AF IF SB} et µd.

On notera, suivant le demi-axe (F1<0), l'alignement {µ4 µ6 µ8}; et, suivant l'axe 2, {µd µc µa}.



Euphydryas editha : phosphoglucose isomérase et climat
bornes pour le découpage des variables du milieu
le nombre des variables est 4

HH	a 3 modalités	dont les sigles	et valeurs maxima	sont	
	H<	H≈	H>	930	2500 10500
PP	a 3 modalités	dont les sigles	et valeurs maxima	sont	
	P<	P≈	P>	15	28 58
TM	a 2 modalités	dont les sigles	et valeurs maxima	sont	
	T<		T>	92	105
tm	a 3 modalités	dont les sigles	et valeurs maxima	sont	
	t<	t≈	t>	5	23 32

Le tableau ci-dessus spécifie la définition des modalités des variables. Comme en convient l'auteur du '*Primer*', la distribution géographique étant sporadique, il n'y a pas lieu d'espérer un modèle corrélationnel entre allèles et climat.

2.2 Survie des moineaux et morphométrie

H. C. BUMPUS: *The elimination of the unfit as illustrated by the introduced sparrow, Passer domesticus.* Biological lectures, Marine Biology Laboratory, Woods Hole, 11th Lecture, pp.209-226; (1898).

49 Femelles de moineaux; mortes: 22-49; en mm/10;
L= Longueur totale ; A=envergure Alaire ; T=Tête + bec ;
H=humerus ; S=carène du bréchet (Sternum);

5	*L	*A	*T	*H	*S
a	1560	2450	316	185	205
b	1540	2400	304	179	196
c	1530	2400	304	179	196

2.2.1 Présentation des données

Ayant recueilli, dans son laboratoire un groupe de moineaux fort éprouvé par une tempête, H. C. BUMPUS entreprit de chercher dans la morphométrie une caractérisation des individus qui avaient survécu. On considère ici un

ensemble I de 49 femelles, sommairement décrites par un ensemble J de cinq mesures. Un extrait du tableau, donne un exemple de ces mesures, en mm/10. Parmi les individus, 21 survécurent; 28 périrent.

```
49 Femelles de moineaux; vives 1-21 ;  mortes: 22-49;
corrélation entre col       4: *H et col       5: *S
corr(*H,*S) =  6.1600447e-1
*S -  2.0806122e+2 ≈  1.0874518e+0 * (*H -  1.8459184e+2)
*H -  1.8459184e+2 ≈  3.4894560e-1 * (*S -  2.0806122e+2)
corrélation entre col       1: *L et col       2: *A
corr(*L,*A) =  7.3493662e-1
*A -  2.4132654e+3 ≈  1.0191720e+0 * (*L -  1.5797959e+3)
*L -  1.5797959e+3 ≈  5.2997124e-1 * (*A -  2.4132654e+3)
corrélation entre col       1: *L et col       3: *T
corr(*L,*T) =  6.7332872e-1
*T -  3.1446939e+2 ≈  1.4859688e-1 * (*L -  1.5797959e+3)
*L -  1.5797959e+3 ≈  3.0510168e+0 * (*T -  3.1446939e+2)
```

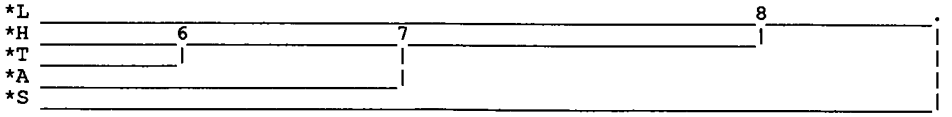
2.2.2 Calculs de corrélation

Répondant à une question soulevée par Br. F.J. M., on a calculé des corrélations entre mesures brutes. Il est a priori attendu que celles-ci soient, deux à deux, corrélées positivement: sur le listage, on ne trouve que des corr > 0,60.

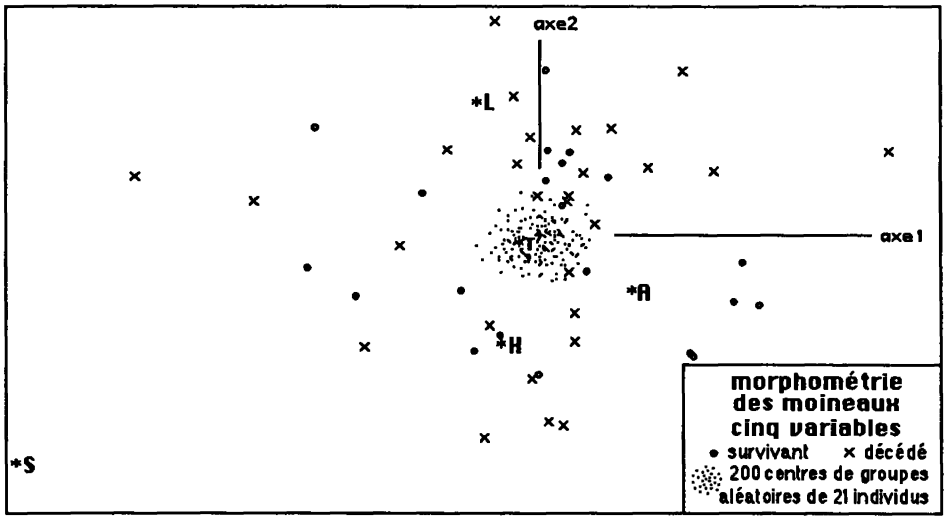
```
49 Femelles de moineaux; vives 1-21 ; mortes: 22-49;
trace  :   1.481e-4
rang   :      1     2     3     4
lambda :      1     0     0     0   e-4
taux   :   4895  3059  1487   559   e-4
cumul  :   4895  7955  9441 10000   e-4
```

2.2.3 Analyse des correspondances

SIGI	QLT	PDS	INR	F 1	CO2	CTR	F 2	CO2	CTR	F 3	CO2	CTR	F 4	CO2	CTR
ci-dessous élément(s) supplémentaire(s)															
§V	1000	428	5	0	150	1	-1	713	11	0	94	3	0	42	4
§M	1000	572	4	0	150	1	1	713	8	0	94	2	0	42	3
SIGJ	**QLT**	**PDS**	**INR**	**F 1**	**CO2**	**CTR**	**F 2**	**CO2**	**CTR**	**F 3**	**CO2**	**CTR**	**F 4**	**CO2**	**CTR**
*L	1000	336	213	-4	176	76	9	813	565	1	9	13	0	3	10
*A	1000	513	180	6	667	246	-3	270	159	2	60	72	0	3	9
*T	1000	67	116	-1	7	2	0	2	1	-15	873	684	-5	118	247
*H	1000	39	85	-2	20	3	-7	171	47	-10	336	192	12	473	718
*S	1000	44	406	-33	812	673	-15	172	228	4	14	39	-1	2	16

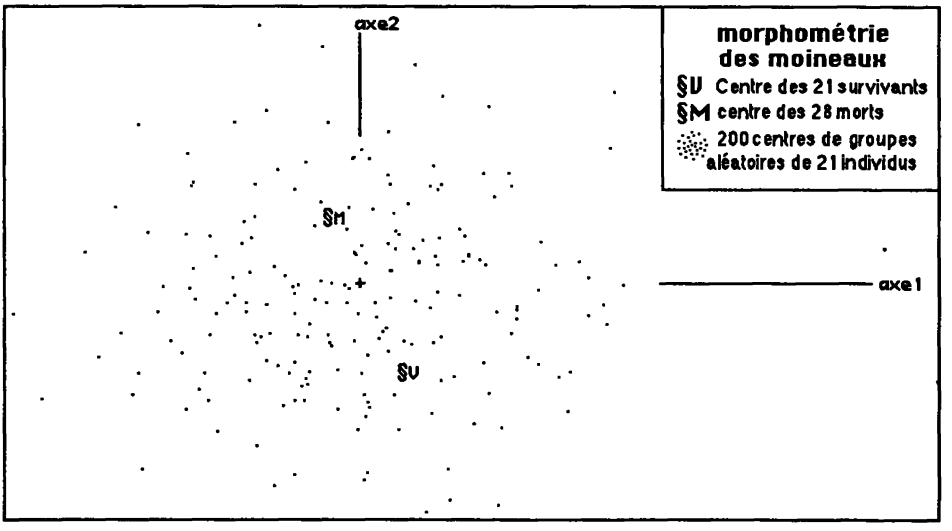


On passe donc à l'analyse des formes, en soumettant le tableau des mesures à l'analyse des correspondances: L'axe 1 est créé par l'opposition entre sternum (bréchet) et envergure alaire.



Sur l'axe 2, la longueur du corps s'oppose à l'envergure (associée ici au sternum). Les axes 3 et 4 rendent compte du développement de T et H relativement à l'ensemble: T et H semblent liés, ce que confirme la CAH.

Mais il apparaît que morts et vifs ne diffèrent pas notablement Le plus simple est de voir, dans le plan (1,2), le point §V (centre des survivants) au sein d'un nuage de 200 groupes fictifs de même effectif. En effet, c'est avec l'axe 2 que se corrèle le mieux l'opposition V≠M. On trouve §V au sein du nuage des



200 centres. Un essai d'analyse discriminante entre les deux sous-ensembles n'aboutit à rien parce que l'inertie du couple des deux centres {M,V} est très faible (1/100 de celle du nuage des individus). Voici la trace:

```
49 Femelles de moineaux; vives 1-21 ; mortes: 22-49;
trace  :   1.233e-6
```

2.2.4 Étude comparative de la dispersion des groupes

A posteriori, on considérera les intentions de l'auteur: pour lui, les morts sont plus dispersés, moins bien calibrés par la sélection que ne le sont les survivants. Les morts seraient à la périphérie du nuage I, dans toutes les directions; les survivants se grouperaient au voisinage du centre. On doit considérer la dispersion des individus d'après leurs contributions à l'inertie; globalement et sur les différents axes. Ainsi, sur l'axe 1, 1/3 de l'inertie vient de deux individus morts: l'un associé à S, l'autre opposé: ils sont à 3 écarts-type de l'origine. Peut-être la mesure de S est-elle fausse pour ces individus; et la dispersion des décédés surestimée d'autant.

total des inerties/orig par groupes

10	INR	CTR1	CTR2	CTR3	CTR4
vif	368	400	320	306	452
mor	632	600	680	694	548
µor	428	269	658	463	548

total des inerties/cdg par groupes

10	Inr	Ctr1	Ctr2	Ctr3	Ctr4
vif	363	399	309	303	448
mor	628	399	672	692	545

Le tableau ci-joint donne, dans l'espace et sur chacun des 4 axes, le total des inerties des individus relativement à l'origine, pour les trois goupes {vif mor µor}: vif (les 21 survivants), mor (les 28 morts) et µor (26 morts, sans les deux individus extrêmes); ces totaux sont à rapporter aux effectifs de chaque groupe. En toute rigueur, il convient de retrancher de cette inertie celle du centre de gravité relativement à l'origine; afin d'avoir l'inertie du groupe relativement à son centre: on obtient ainsi les valeurs notées Inr, Ctr.

Les totaux pour 'vif' sont faibles ; mais il est apparu qu'ils ne sont aucunement exceptionnels parmi ceux afférents à des groupes fictifs d'effectif 21; (même si l'on admet dans ceux-ci les individus extrêmes). À titre d'exercice, précisons comment on a procédé. Le tableau à cinq colonnes {INR CTR1 CTR2 CTR3 CTR4}, pour l'ensemble I des 29 moineaux, peut être extrait, par le programme 'coupli', du listage d'analyse des correspondances. Sur ce tableau, le programme 'cums' calcule des cumuls de lignes afférents aux 200 groupes aléatoires (de 21 moineaux chacun), déjà considérés ci-dessus: la liste définissant ces cumuls, ayant été créée par le programme 'cumhasx'. D'où un tableau à 200 lignes chacune analogue à:

```
vif   368   400   320   306   452
```

ValSup	262	282	301	320	338	356	374	392	411	427	448	466	483	502	…
126	1	1	1	2											
160	1	1													
179	1	1	1	1											
232		3	1	4	1	1									
269		1	3	4	2	1	1	1	1						
308			1	3	3	3	4	2	1						
346				1	2	3	4	3	2	3	1	1			
385			1		2	1	7	4	4	3	5	1	1		
421							2	3	3	6	5	2			
457							1	1	1	4	4	3		1	
494								1	1	3	3	7	1	3	…
531											3	2	4		…
567												1	4	4	…
596														2	…
637														1	…
678															…
703															
754															

tri croisant INR (n°1;col) et CTR1 (n°2;lignes) pour 200 groupes aléatoires de 21
247 ≤ Inr ≤ 576 ; 90 ≤ Ctr1 ≤ 754
pour le groupe des 21 survivants : Inr = 363 ; Ctr1 = 399

mais afférente à un groupe aléatoire; ce tableau a été juxtaposé (par 'juxtab') à un tableau {inr ctr1 ctr2 ctr3 ctr4} extrait du listage des centres des groupes, mis en supplément à l'analyse principale. Enfin 'cums' a servi pour retrancher des inerties des groupes, relatives à l'origine (INR CTR1...), les inerties des centres (inr ctr1...); d'où le tableau des inerties des groupes, chacun rapporté à son propre centre de gravité (Inr Ctr1...).

On considère la distribution des colonnes de ce tableau par 'zrang'. Nous publions un extrait du tri croisant Inr et Ctr1: le coin supérieur droit, encadré sur cette copie d'écran, recense de nombreux groupes aléatoires pour lesquels Inr et Ctr1 (rapportés au cdg du groupe) sont tous deux inférieurs aux valeurs trouvées pour 'vif'. La concentration des formes des survivants autour de leur centre de gravité n'est donc aucunement exceptionnelle.

N'est donc pas confirmée l'hypothèse de l'auteur, selon laquelle le groupe des survivants aurait une homogénéité exceptionnelle.