

M. ROUX

C. MONTANA

**Pondération des contributions en analyse
des correspondances quand les nombres de
modalités des variables diffèrent grandement
: application en écologie**

Les cahiers de l'analyse des données, tome 13, n° 4 (1988),
p. 459-468

<http://www.numdam.org/item?id=CAD_1988__13_4_459_0>

© Les cahiers de l'analyse des données, Dunod, 1988, 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/>

PONDÉRATION DES CONTRIBUTIONS EN ANALYSE DES CORRESPONDANCES QUAND LES NOMBRES DE MODALITÉS DES VARIABLES DIFFÈRENT GRANDEMENT: APPLICATION EN ÉCOLOGIE

[POND. CONTRI.]

M. ROUX*
C. MONTANA**

1 Introduction

Dans une étude antérieure d'écologie végétale sur la région de Mapimi, (État de Durango, Mexique) il est apparu que les contributions des variables aux axes varient notablement suivant le nombre de modalités en lesquelles ces variables sont découpées (Montana et Lepart, non publié).

Dans cette étude, qui fournit la base du présent travail, on cherchait à évaluer l'influence des variables de l'environnement sur la distribution des espèces végétales. Le tableau analysé était le suivant: en colonnes, (selon l'usage de la spécialité), les espèces les plus fréquentes, (au nombre de 125); en ligne, un ensemble de modalités des variables de l'environnement: Position topographique, Pente, Exposition, Géomorphologie, Type d'érosion,... ; au total 19, variables; à l'intersection de la colonne j et de la ligne i: le nombre de fois que l'espèce j a été vue, en même temps qu'était enregistrée la modalité i.

Parmi les 19 variables de l'environnement, figure l'Altitude, qui n'avait pas été enregistrée sous forme discrète, mais sous forme continue; et qu'il a donc fallu découper en classes en vue de l'analyse; (tandis que, e.g., la Pente, autre variable continue, était enregistrée suivant des classes établies avant le recueil des données).

Il est apparu que la contribution totale de la variable altitude variait nettement selon le nombre de modalités qu'on attribuait à celle-ci.

(*) Centre d'études phytosociologiques et écologiques (CNRS) B.P. 5051, 34033 Montpellier CEDEX, France.

(**) Instituto de Ecologia, Apartado 263-B, 35070 GOMEZ PALACIO, Mexique.

Notre but est ici de permettre à l'utilisateur de corriger les valeurs des contributions par des calculs simples, (faits, si possible, à la main), afin de tenir compte des nombres inégaux de modalités.

Les contributions sont des fractions de l'inertie des axes; la somme de ces inerties n'est autre que l'inertie totale du nuage; i.e., (au coefficient N près, total de toutes les cases), le χ^2 du tableau complet.

De nombreux auteurs ont, par le passé, observé que le χ^2 dépend des nombres de lignes et colonnes du tableau; et ils ont proposé des corrections dont, au §2, nous étudions l'effet sur des tableaux de contingence créés par tirage au hasard. Le coefficient de Kendall et Stuart nous étant apparu le plus satisfaisant, nous l'avons appliqué aux données réelles de Mapimi.

Nous examinons, au §3, les tableaux de contingence simple croisant une seule variable de l'environnement avec l'ensemble des espèces; puis, au §4, nous considérons le tableau complet soumis à l'analyse des correspondances. Enfin, au §5, nous concluons en justifiant la règle de pondération que nous avons adoptée.

2 Mesure d'association pour les tableaux de contingence simples

Pearson, (1904), a proposé le coefficient suivant, appelé par lui *coefficient de contingence* :

$$P = \sqrt{(\chi^2 / (\chi^2 + N))} ;$$

P doit son nom au fait que, lorsqu'il s'agit du croisement de deux variables continues découpées en classes, il tend vers le coefficient de corrélation usuel entre les deux variables sous les conditions suivantes: la loi conjointe des variables est une loi normale, et $N \rightarrow \infty$. Mais, (cf. A. Patris et coll., 1985) cette convergence est lente et en butte à un biais systématique qui dépend du nombre des classes. De plus, dans ce coefficient, les composantes du χ^2 n'interviennent pas de façon additive; ce qui complique l'interprétation en terme de contributions.

Cramer (1946), a proposé un autre coefficient:

$$C = (\chi^2 / N) / (\min\{\text{Card}_i, \text{Card}_j\} - 1) ;$$

(où Card_i et Card_j sont les nombres de modalités de chacune des deux variables). Comme le précédent, ce coefficient C varie entre 0 et 1; de plus, la borne 1 est atteinte s'il y a dépendance totale entre les deux variables (et que le nombre de valeurs différentes prises est effectivement $\min\{\text{Card}_i, \text{Card}_j\}$).

Enfin, un troisième coefficient, sur lequel se fixera notre choix, est celui de Kendall et Stuart (1961):

$$K = (\chi^2 / N) / \sqrt{((Cardi - 1)(Cardj - 1))} .$$

Pour comparer ces coefficients, nous procédons comme suit. Nous effectuons d'abord des tirages aléatoires suivant une loi normale bidimensionnelle de coefficient de corrélation r fixé. (Ceci se fait aisément sur ordinateur en tirant deux coordonnées normales indépendantes, puis multipliant le vecteur obtenu par une matrice convenable). Nous découpons alors les variables en classes d'égal effectif, (de diverses manières), et construisons un tableau de contingence pour lequel sont calculés les coefficients retenus ci-dessus.

Le modèle normal n'est pas réaliste, mais il est commode. Le découpage en classes d'égal effectif a été retenu selon la pratique courante dans le codage en vue de l'Analyse des Correspondances. Le coefficient de corrélation a été fixé à 0,866: valeur assez élevée, mais conforme au cas qui nous intéresse pratiquement. On a effectué 600 tirages de couples de valeurs. La première variable a été découpée en 6 classes; et la deuxième, en un nombre NC de classes variant de 2 à 14.

Nc	chi2	Pearson	Cramer	Kendall
2	323	.592	.538	.241
3	490	.671	.409	.258
4	564	.696	.313	.243
5	624	.714	.260	.233
6	697	.733	.232	.232
7	686	.730	.229	.209
8	706	.735	.235	.199
9	702	.734	.234	.185
10	726	.740	.242	.180
11	720	.739	.240	.170
12	748	.745	.249	.168
13	763	.748	.254	.164
14	755	.746	.252	.156

Tableau 1 : Comparaison des coefficients de contingence issus du chi2, sur des données simulées

Les résultats rapportés dans le tableau 1, illustré par la figure 2, montrent la relative stabilité des coefficients de Pearson et de Kendall; celui-ci l'emportant pour les petits nombres de classes.

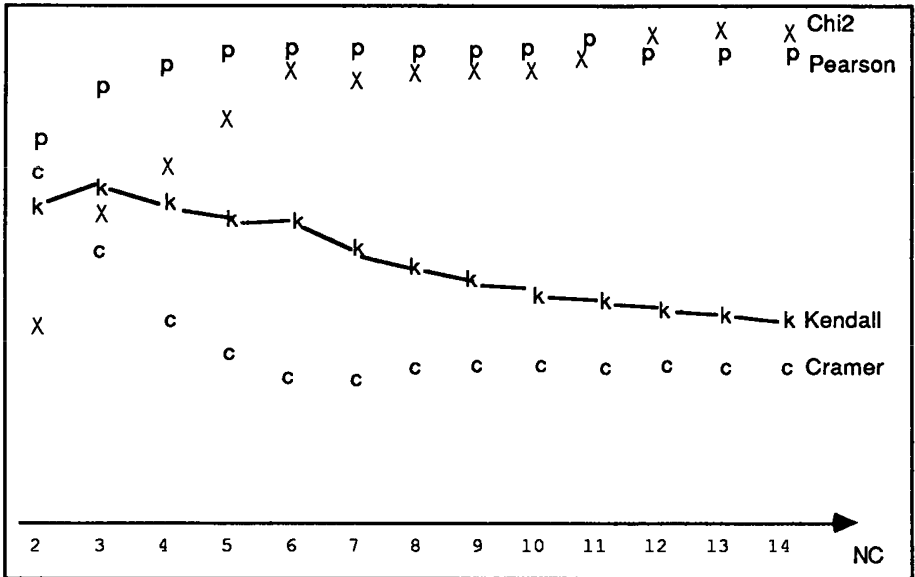


Figure 1 : Comparaison des coefficients sur des données simulées (cf Tableau 1)

3 Tableaux de contingence simples sur des données réelles

3.1 Influence du nombre des classes sur les coefficients de contingence

Nous avons construit une série de tableaux de contingence croisant l'ensemble des 125 espèces avec l'altitude découpée en un nombre de classes variant de 2 à 10, les effectifs des classes d'un même découpage étant sensiblement égaux. Comme le nombre des espèces n'est pas modifié, nous éliminons le facteur correspondant du calcul de l'indice de Kendall et Stuart; (pour celui de Cramer, le nombre d'espèces, étant supérieur au nombre des modalités de l'altitude, n'intervient pas; mais on a éliminé $1/N$).

Le tableau 2 montre une stabilité du coefficient de Kendall et Stuart encore meilleure que dans le cas des données artificielles du §2. Ce coefficient sera donc seul utilisé au §3.2.

3.2 Classement des variables selon le degré de liaison avec les espèces

Nous fixons ici à 6 le nombre des modalités de l'altitude; et comparons les rangs des variables classées soit selon le χ^2 , soit selon l'indice de Kendall et Stuart.

Nc	chi2	Pearson	N*Cramer	Kendall* $\sqrt{124}$
2	656	.458	656	656
3	971	.531	485	687
4	1126	.559	375	650
5	1334	.592	334	667
6	1555	.621	311	696
8	1931	.662	276	730
10	2141	.681	238	714

Tableau 2 : Comparaison des coefficients de contingence issus du chi2, sur des données réelles

Variable	Sigle NC	chi2	rchi	Kendall	rkend	
Position topographique	TOP	6	1935	7	865	7
Pente	PEN	3	1118	15	791	12
Couv. géo-pédologique	PED	5	1915	8	958	4
Erosion	ERO	9	1834	9	648	16
Submersion	SUB	3	1377	12	974	3
Espèce dominante 1 (*)	DM1	7	2054	5	851	9
Espèce dominante 2 (*)	DM2	11	2369	3	784	13
Recouvrement strate 3	RC3	7	1211	13	494	18
Recouvrement strate 5	RC5	3	663	19	469	19
Unité de paysage	PAY	8	2377	2	898	5
Géoforme	GFO	10	3230	1	1077	1
Lithologie	LIT	6	1969	6	881	6
Texture	TEX	6	2253	4	1007	2
Profondeur	PRO	2	761	18	761	14
Horizon diagnostique	HOR	4	980	16	566	17
Phase P-G	PPG	3	1155	14	817	10
Salinité	SAL	4	1409	11	813	11
Phase Na	PNA	2	862	17	862	8
Altitude	ALT	6	1555	10	696	15

Tableau 3 : Classement des variables selon les coefficients de contingence
 (*) Pour ces tableaux, on a tenu compte de ce que l'effectif est légèrement diminué

Si on recherche, sur le tableau 3, les variables qui occupent les 6 premiers rangs, on en trouve 4 qui subsistent quel que soit l'indice utilisé, (χ^2 , ou indice de Kendall) ; soit: Géoforme, Unité de paysage, Texture et Lithologie; mais il ya désaccord pour DM1, DM2, PED et SUB. La variable d'Altitude, quant à elle, passe du 10-ème au 15-ème rang.

4 Analyse factorielle d'un tableau de contingence multiple

En Analyse des Correspondances, la variance sur chaque axe représente une fraction du χ^2 , lequel mesure, à un coefficient près, l'inertie totale du nuage. Les contributions CTR représentent la part, additive, afférente à chaque élément, (ligne ou colonne), dans l'inertie sur l'axe considéré. Comme toutes ces grandeurs sont additives, on peut leur appliquer la correction que Kendall et Stuart appliquent au χ^2 .

Le tableau soumis à l'analyse comporte 125 espèces qui constituent l'ensemble F, (*Floristique*), des colonnes. Les modalités des variables de l'*Environnement* constituent l'ensemble E des lignes du tableau. Si l'on met à part, provisoirement, l'Altitude, il y a, au total, 99 modalités. On cherche à apprécier l'importance relative du rôle de ces variables dans la dispersion des espèces le long des différents axes. On pense à ranger les variables suivant le total des contributions CTR que leurs modalités respectives apportent à chaque axe.

Or ce total, comme le χ^2 , dépend, dans une large mesure, du nombre des modalités en lesquelles la variable a été découpée. Aussi proposons-nous d'introduire un facteur correctif, en nous inspirant du coefficient de Kendall et Stuart. Dans le cas présent, chaque variable est croisée avec l'ensemble F des modalités d'une même et unique variable *Floristique*, pour laquelle il n'y a pas lieu d'introduire de facteur correctif. Tandis qu'afin de tenir compte des nombres inégaux de modalités des variables de l'environnement, nous divisons chaque contribution $CTR_{\alpha}(me)$ par $\sqrt{(c_j - 1)}$, c_j étant le nombre des modalités de la variable j à laquelle appartient la modalité me .

Comme précédemment, on a fait varier de 2 à 10 le nombre des modalités de l'altitude. Notons respectivement A2, A3, ... , A10 les ensembles de modalités créés par ces découpages.

Dans un premier essai, les ensembles A2, A3, A4, A6, A8 et A10 sont en supplémentaire; tandis qu'est en principal l'ensemble A5, avec les ensembles de modalités des autres variables de l'environnement. Dans un deuxième essai, tous les ensembles A2 à A10 sont en principal, avec les 99 modalités des autres variables.

Les résultats sont résumés dans les tableaux 4a, 4b et 5, où les CTR sont exprimées en ‰ de l'inertie sur chaque axe. Le tableau 4a donne les valeurs de la contribution globale de l'Altitude, en fonction du découpage adopté, sur les 5 premiers axes, quand seul est en principal A5; le tableau 4b donne les mêmes résultats quand les A_n , de A2 à A10, sont tous en principal. Le tableau 5 concerne l'ensemble des variables de l'environnement, y compris ALT, découpée en 5 classes; les tableaux A_n autres que A5 étant éliminés de l'analyse.

	axe1		axe2		axe3		axe4		axe5	
Val. propre	.3078		.1277		.0561		.0302		.0191	
% d'inertie	45.45		18.85		8.29		4.46		2.62	
	CTR	CC	CTR	CC	CTR	CC	CTR	CC	CTR	CC
2 classes	38	38	2	2	14	14	7	7	0	0
3 classes	31	22	44	31	23	16	22	15	40	28
4 classes	37	22	34	20	26	15	25	14	88	51
5 classes	40	20	42	21	34	17	18	9	80	40
6 classes	45	20	43	19	50	22	30	13	89	40
8 classes	47	18	47	18	69	26	37	14	96	36
10 classes	47	16	47	16	66	22	31	10	111	37

Tableau 4a : CTR et CC, (CTR corrigées), de l'Altitude, A5 étant seul en principal

4.1 Variations des contributions de l'altitude

La comparaison des tableaux 4a et 4b montre peu de différence entre les deux façons de traiter l'altitude. Celle-ci apporte sa contribution la plus importante à l'axe 5; et, sauf correction, cette contribution croît rapidement avec le nombre de modalités. Quand on applique la correction, l'ampleur des variations est beaucoup plus faible. Sur les autres axes les fluctuations nous paraissent être trop faibles pour qu'on tente de leur attribuer une valeur typique. On notera cependant un phénomène constant: il y a des variations brutales quand on passe de 2 à 3 classes; ce qui suggère d'éviter les variables à 2 modalités, lorsqu'on a, par ailleurs, des variables à 3 modalités et davantage.

	axe1		axe2		axe3		axe4		axe5	
Val. propre	.2927		.1166		.0536		.0270		.0216	
% d'inertie	45.10		17.96		8.26		4.16		3.33	
	CTR	CC	CTR	CC	CTR	CC	CTR	CC	CTR	CC
2 classes	32	32	0	0	10	10	2	2	0	0
3 classes	29	22	38	27	19	14	28	18	17	12
4 classes	34	19	28	16	24	14	34	20	48	28
5 classes	35	18	35	18	34	17	22	11	53	27
6 classes	40	18	36	16	47	21	44	20	56	25
8 classes	41	15	40	15	61	23	49	19	89	34
10 classes	41	14	40	13	58	19	41	14	107	36

Tableau 4b : CTR et CC, (CTR corrigées), de l'Altitude, les An étant tous en principal

4.2 Importance de l'altitude par rapport aux autres variables de l'environnement

Examinons le tableau 5 qui donne les CTR et les CTR corrigées (CC) de toutes les variables aux cinq premiers facteurs, l'Altitude étant ici découpée en 5 classes. Pour chaque axe, les variables sont disposées suivant le rang, rc, des CC; le rang, ra, des CTR étant également mentionné pour faciliter les comparaisons.

En ce qui concerne l'Altitude, l'effet de la correction est sensible sur l'axe 1, où elle passe du rang 15 au rang 11; sur les autres axes, le rang de ALT varie au plus de 1. Pour les autres variables, la correction n'apporte de modification majeure au traitement que sur l'axe 1; où l'on note que la PROfondeur passe de ra = 14 à rc = 3; que la PENte passe de ra = 10 à rc = 5; tandis que l'espèce dominante en second (DM2) rétrograde de ra = 6 à rc = 16. Cependant ces changements sont à interpréter avec prudence; car on a vu que la correction proposée, (en suivant Kendall et Stuart), a tendance à favoriser les variables comptant peu de modalités, (ce qui est précisément le cas de PRO et de PEN), au détriment des variables qui, (telles DM2), comptent de nombreuses modalités.

Contributions à l'axe1						Contributions à l'axe2						Contributions à l'axe3					
VAR	NC	CTR	ra	CC	rc	VAR	NC	CTR	ra	CC	rc	VAR	NC	CTR	ra	CC	rc
PNA	2	44	9	44	1	PED	5	104	2	47	1	SAL	4	63	3	36	1
SUB	3	55	3	39	2	LIT	6	94	5	42	2	GFO	10	105	1	33	2
PRO	2	36	14	36	3	TEX	6	97	4	39	3	PPG	3	47	6	33	3
PPG	3	46	8	32	4	TOP	6	99	3	37	4	RC5	3	45	8	32	4
PEN	3	44	10	31	5	GFO	10	106	1	34	5	DM2	11	89	2	28	5
SAL	4	53	4	31	6	ERO	9	94	6	33	6	DM1	7	62	4	25	6
TEX	6	56	2	23	7	PAY	8	80	7	30	7	SUB	3	36	9	25	7
PED	5	47	5	21	8	ALT	5	35	8	18	8	PEN	3	32	12	23	8
LIT	6	46	7	20	9	RC3	7	26	9	11	9	TEX	6	51	5	21	9
GFO	10	62	1	19	10	PPG	3	14	11	10	10	PAY	8	47	7	18	10
ALT	5	35	15	18	11	PEN	3	10	13	7	11	ALT	5	34	10	17	11
DM1	7	41	13	17	12	SUB	3	10	12	7	12	HOR	4	28	15	16	12
HOR	4	29	16	17	13	PNA	2	7	15	7	13	PRO	2	14	18	14	13
PAY	8	41	12	16	14	PRO	2	5	16	5	14	TOP	6	33	11	13	14
TOP	6	42	11	16	15	DM2	11	15	10	5	15	PED	5	29	14	13	15
DM2	11	46	6	15	16	DM1	7	10	14	4	16	ERO	9	32	13	11	16
RC5	3	20	18	14	17	SAL	4	5	17	3	17	LIT	6	18	16	8	17
ERO	9	21	17	8	18	RC5	3	3	19	2	18	RC3	7	16	17	6	18
RC3	7	20	19	7	19	HOR	4	4	18	2	19	PNA	2	0	19	0	19

Tableau 5: Variables rangées suivant leurs CTR corrigées aux axes 1, 2 et 3.

Contributions à l'axe4						Contributions à l'axe5					
VAR	NC	CTR	ra	CC	rc	VAR	NC	CTR	ra	CC	rc
PAY	8	149	1	56	1	DM1	7	113	1	46	1
DM1	7	112	2	46	2	GFO	10	105	2	33	2
DM2	11	107	3	34	3	ALT	5	53	3	27	3
RC3	7	77	5	31	4	DM2	11	.81	4	26	4
GFO	10	90	4	28	5	SAL	4	36	6	21	5
HOR	4	41	7	24	6	LIT	6	44	5	20	6
PEN	3	29	9	21	7	PEN	3	25	12	18	7
SUB	3	28	10	20	8	TEX	6	36	7	15	8
ERO	9	43	6	15	9	PED	5	29	9	13	9
TOP	6	35	8	13	10	HOR	4	22	13	13	10
LIT	6	27	11	12	11	SUB	3	17	15	12	11
ALT	5	22	12	11	12	PAY	8	28	10	11	12
PED	5	14	13	6	13	TOP	6	30	8	11	13
PRO	2	5	15	5	14	ERO	9	28	11	10	14
TEX	6	12	14	5	15	RC3	7	20	14	8	15
SAL	4	5	16	3	16	PRO	2	4	16	4	16
RC5	3	2	18	2	17	PNA	2	3	17	3	17
PPG	3	3	17	2	18	PPG	3	3	19	2	18
PNA	2	0	19	0	19	RC5	3	3	18	2	19

Tableau 5: Variables rangées suivant leurs CTR corrigées (CC) aux axes 4 et 5. Le découpage A5 de l'Altitude est seul pris en compte; toutes les modalités sont en principal.

5 Conclusion

Le χ^2 de contingence dépend notablement des nombres de modalités des variables dont on étudie la liaison. Des trois corrections considérées pour comparer entre eux des χ^2 calculés pour des variables découpées en des nombres différents de modalités, nous avons retenu celui de Kendall et Stuart, bien qu'il soit un peu moins stable que celui de Pearson dans l'expérience de simulation. Car cette correction a pour elle la simplicité et le fait qu'elle conserve le caractère additif propre aux contributions CTR de l'analyse des correspondances.

Nous avons dit que ces contributions représentent une fraction du χ^2 de contingence calculé sur l'ensemble du tableau. Dans le cas où l'on croise toujours la même variable, comme nous le faisons ici pour l'espèce, avec un

ensemble d'autres variables qu'on peut appeler explicatives, la correction que nous proposons est très simple: elle consiste à diviser chaque CTR, afférente à une modalité de variable explicative, par la racine carrée du nombre, diminué de 1, des modalités de cette variable.

En traitant de cette façon des données réelles, on observe une lente décroissance des contributions corrigées avec le nombre des classes, au lieu de la croissance constatée si l'on ne fait pas de correction. Mais, sur chaque axe, le rangement des variables suivant leurs contributions apparaît peu sensible au fait que les CTR ont été, ou non, corrigées.

Références bibliographiques

Cramer H. (1946) : *Mathematical methods for Statistics*; Princeton University Press.

Kendall M.G. & Stuart A. (1961): *The advanced theory of Statistics*, Vol.I, Griffin, London.

Patris A., Chau N., Kohler F., Martin J., Demogeot F., Boitel L., Saurin S., Legras B. (1985): Analyses multidimensionnelles sur le coefficient de contingence de Pearson modifié. Application à l'analyse du profil dépressif des salariés. *Quatrièmes journées internationales Analyse des Données et Informatique*. INIRIA, Versailles, Tome II, pp. 567-572.

Pearson K. (1904): *On the theory of contingency and its relation to association and normal correlation*. Drapers Co Memoirs, Biometric series, n° 1, London.