

J.-P. BENZÉCRI

F. BENZÉCRI

Extension d'un tableau par adjonction de lignes et colonnes créées par cumul

Les cahiers de l'analyse des données, tome 15, n° 1 (1990),
p. 109-116

http://www.numdam.org/item?id=CAD_1990__15_1_109_0

© Les cahiers de l'analyse des données, Dunod, 1990, 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/>

EXTENSION D'UN TABLEAU PAR ADJONCTION DE LIGNES ET COLONNES CRÉÉES PAR CUMUL

[CUM. LI. COL.]

J.-P. & F. BENZÉCRI

1 Utilité des lignes et colonnes créées par cumul ou moyenne

À partir d'un tableau de données croisant un ensemble I de pays et un ensemble J de marchandises, on peut créer des cumuls de lignes ou de colonnes pour faire figurer sur les graphiques des points supplémentaires représentant une région ou un continent (cumul de lignes) ou une classe de produits (cumul de colonnes). Éventuellement, ce seront les cumuls qui constitueront les éléments principaux de l'analyse, les éléments de base du tableau étant, au contraire, en supplémentaire.

C'est, en particulier, ce que l'on fait en analyse discriminante. Dans le cas le plus simple, les individus de base sont répartis en deux classes, d'où deux centres de gravité; et chaque individu de base est ensuite affecté au centre dont il est le plus proche; lequel n'est pas toujours celui de sa classe d'origine: on parle alors d'erreur d'affectation. Plus généralement, on peut avoir plusieurs centres calculés comme moyenne de classes: par exemple, dans l'analyse de dossiers médicaux, chaque classe correspondra à une maladie et l'affectation sera une procédure de diagnostic, plus ou moins efficace selon les données disponibles.

Jusqu'ici nous avons employé indifféremment les termes de cumul ou de moyenne. Il est vrai qu'en analyse des correspondances une ligne (ou colonne) construite par cumul se situe au centre de gravité des éléments cumulés, chacun de ceux-ci étant affecté de sa masse propre. Mais si l'on fait l'analyse en ne gardant en principal que les éléments construits par cumul, chacun de ceux-ci reçoit pour masse la somme des masses des individus qu'il comprend.

Dans le cas très simple d'une discrimination entre deux classes, d'inégal effectif, de lignes d'un tableau codé sous forme disjonctive complète (donc où toutes les lignes ont même poids) les deux centres recevront des poids différents et l'origine de l'axe unique créé par l'analyse ne tombera pas au milieu des deux centres. Il est facile d'obtenir des lignes nouvelles toutes de même poids en

calculant non un cumul mais une moyenne; c'est-à-dire en divisant la ligne (ou colonne) cumul par son total (ou le centième de son total).

2 Programme adjoignant des cumuls à un tableau de base

Il va sans dire que les calculs de cumul ou de moyenne ne présentent, pour le programmeur, aucune difficulté fondamentale; d'ailleurs les tableurs permettent de faire ces calculs de façon plus ou moins fastidieuse. Toutefois, comme il s'agit d'opérations qui, à l'instar des juxtapositions de tableaux, extractions de sous-tableaux, ou transpositions, se répètent souvent, il nous a paru pratique de disposer, pour effectuer les cumuls, d'un programme spécial.

Ainsi, après 'soustab', 'juxtab', ..., on a écrit, pour le logiciel MacSAIF, un programme 'cums'. La difficulté, ici, n'est pas de concevoir un algorithme, mais de choisir les fonctions qu'on attribuera au programme et d'en rendre l'accès facile à l'utilisateur! Nous décrivons d'abord le programme, tel qu'il est présentement, en en montrant le fonctionnement sur un petit tableau où on a, en $k(i,j)$, les exportations en phosphates de j vers i ; puis donnons la voie suivie, pour une application à l'analyse discriminante, dans [NOUV. TEST. GREC.].

Nous décrivons successivement la création de cumul entièrement par dialogue à l'écran, puis à l'aide d'un fichier texte auxiliaire spécifiant les lignes et colonnes à cumuler.

2.1 Création de lignes et colonnes cumul par dialogue

On note qu'avant le nom du tableau à compléter, est demandé, afin de réserver l'espace, le nombre maximum de lignes qu'on envisage de créer.

```

ce programme cree des lignes et colonnes nouvelles par cumul
le nombre maximum de lignes nouvelles a creer est 20
max = 20
ce choix est il confirme 0 ou N 0
le fichier des donnees est B:data:phos
B:data:phos
ce nom est il confirme oui(O) ou non(N) 0
COMMERCE MONDIAL DES PHOSPHATES
34 x 8
  B   eBL   eUS   eJR   eMR   eSN   eTG   eTN   eCC
iBL   0  1305   0  3573   25   500  110  293
iCA   2  8335   0   0     0     8    0    0
iFR  1311 2691   70 4891  1484 2526 1697  4
IDL  1322 3808   0 1445   261   200  288 1442
iIT   42 1883   194 2881   67  195  493  0
iJP   0 4426   522 1540  239  93   0    0
iNL  299 1484   0 1853   249 1584  59  84
iSP   20 339   0 5073   2    85   11  0
iUK  122 645   2 2852   971  36  197  0
iIN   0 2559  996 1149  218  0  134  0
iBR  29 4918   0 1398   15  0  241  0
iPL   0 1284  271 3311   33  540  548 1996
iAM   0 768  483 1541   6  138  22 1206
iEE   2 201  136 1398   0  0  333 5533
COMMERCE MONDIAL DES PHOSPHATES
cari = 14   carj = 8
le sigle du tableau a creer est

```

```

iIN  0 2559  996 1149  218  0 134  0
iBR  29 4918  0 1398  15  0 241  0
iPL  0 1284  271 3311  33 540 548 1996
iRM  0 768  483 1541  6 138  22 1206
iEE  2 201  136 1398  0  0 333 5533
COMMERCE MONDIAL DES PHOSPHATES
car1 = 14  car2 = 8
le sigle du tableau a creer est **
**
ce sigle est il confirme oui(O) ou non(N) O
faut-il creer une ligne(L) une colonne(C) ou finir(F) L
ce choix est il confirme O ou N O
le sigle de la ligne a creer est itot
itot
ce sigle est il confirme oui(O) ou non(N) O
les lignes a cumuler seront elles designees
une par une(U) ou par blocs(B) B
ci dessous sont rappelles les num et sigles des lignes
 1 iBL; 2 iCA; 3 iFR; 4 iDL; 5 iIT; 6 iJP; 7 iNL; 8 iSP; 9 iUK;
10 iIN; 11 iBR; 12 iPL; 13 iRM; 14 iEE;
le nombre des blocs de lignes a cumuler sera 1
le numero de la 1ere ligne du bloc a cumuler 1 est 1
le num de la derniere ligne du bloc a cumuler 1 est 14
le choix des lignes a cumuler est il confirme O ou N O
20015 itot ligne cumul des lignes suivantes
 1 iBL ; 2 iCA ; 3 iFR ; 4 iDL ; 5 iIT ; 6 iJP ;
 7 iNL ; 8 iSP ; 9 iUK ; 10 iIN ; 11 iBR ; 12 iPL ;
13 iRM ; 14 iEE ;
faut-il creer une ligne(L) une colonne(C) ou finir(F) C
ce choix est il confirme O ou N O
le sigle de la colonne a creer est etot
etot
ce sigle est il confirme oui(O) ou non(N) O
les colonnes a cumuler seront elles designees
une par une(U) ou par blocs(B) B
ci dessous sont rappelles les num et sigles des colonnes
 1 eBL; 2 eUS; 3 eJR; 4 eMR; 5 eSN; 6 eTG; 7 eTN; 8 eCC;
le nombre des blocs de colonnes a cumuler sera 1
le numero de la 1ere colonne du bloc a cumuler 1 est 1
le num de la derniere colonne du bloc a cumuler 1 est 8
le choix des colonnes a cumuler est il confirme O ou N O
20009 etot colonne cumul des colonnes suivantes
 1 eBL ; 2 eUS ; 3 eJR ; 4 eMR ; 5 eSN ; 6 eTG ;
 7 eTN ; 8 eCC ;
faut-il creer une ligne(L) une colonne(C) ou finir(F) F
ce choix est il confirme O ou N O

```

On voit que le programme comporte, pour chaque ligne ou colonne à créer par cumul, un dialogue de définition de sous-ensemble qui est le même que celui utilisé dans 'qori', pour spécifier les éléments supplémentaires, ou dans 'soustab', pour choisir les lignes ou les colonnes à conserver.

Dans le cas présent, on s'est borné à créer, par cumul, une seule ligne et une seule colonne; dont chacune n'est autre que la marge correspondante du tableau de base. Cette opération n'offre pas d'intérêt en elle-même; mais comme le programme 'cums' crée, non seulement un tableau complété, mais aussi une liste de définition des cumuls effectués (analogue à celle que l'on voit affichée à l'écran après qu'a été confirmée la demande de création, e.g. de itot), nous nous proposons d'utiliser les listes ainsi produites des indentificateurs de lignes et de colonnes, pour composer sur éditeur de texte, un fichier spécifiant des cumuls plus intéressants.

COMMERCE MONDIAL DES PHOSPHATES													
18 x 12		eBL	eUS	eJR	eMR	eSN	eTG	eTN	eCC	eAfN	eAeq	eAfr	eAf%
iBL	0	1305	0	3573	25	500	110	293	3683	525	4208	9	0
iCA	2	8335	0	0	0	8	0	0	0	8	8	0	0
iFR	1311	2691	70	4891	1484	2526	1697	4	6588	4010	10598	23	0
iDL	1322	3808	0	1445	261	200	288	1442	1733	461	2194	5	0
iIT	42	1883	194	2881	67	195	493	0	3374	262	3636	8	0
iJP	0	4426	522	1540	239	93	0	0	1540	332	1872	4	0
iNL	299	1484	0	1853	249	1584	59	84	1912	1833	3745	8	0
iSP	20	339	0	5073	2	85	11	0	5084	87	5171	11	0
iUK	122	645	2	2852	971	36	197	0	3049	1007	4056	9	0
iIN	0	2559	996	1149	218	0	134	0	1283	218	1501	3	0
iBR	29	4918	0	1398	15	0	241	0	1639	15	1654	4	0
iPL	0	1284	271	3311	33	540	548	1996	3859	573	4432	10	0
iRM	0	768	483	1541	6	138	22	1206	1563	144	1707	4	0
iEE	2	201	136	1398	0	0	333	5533	1731	0	1731	4	0
iEst	2	2253	890	6250	39	678	903	8735	7153	717	7870	17	0
iEs%	0	11	5	32	0	3	5	44	36	4	40	0	0
iEur	3118	14408	1156	28818	3098	5804	3758	10558	32576	8902	41478	89	0
iEu%	4	20	2	41	4	8	5	15	46	13	59	0	0

2.2 Création de lignes et colonnes à l'aide d'un fichier texte auxiliaire spécifiant les éléments à cumuler.

Sans doute est-il préférable de montrer le résultat avant d'expliquer comment on a obtenu celui-ci! On reconnaît le tableau de données qui s'est affiché à l'écran au début du dialogue du §2.1. Mais, à ce tableau, se sont ajoutées quatre nouvelles colonnes et quatre nouvelles lignes.

La ligne 'iEst' est le cumul des 4 lignes {iPL, iRM, iEE} (qui concernent respectivement la Pologne, la Roumanie et l'Europe de l'Est; celle-ci déjà définie par cumul de la Bulgarie de la Hongrie et de la RDA). La ligne 'iEs%' ne diffère de la précédente que parce qu'elle est une ligne de pourcentages: de façon précise le total des éléments de cette ligne, *situés dans les 14 colonnes du tableau de départ*, est 100. On a défini, de même 'iEur' et 'iEu%' en cumulant les lignes afférentes à tous les pays européens: ceux non situés à l'Est et 'iEst'. Pour le programme 'cums', tout sigle terminé par '%' donne l'ordre de calculer des %!

Pour les colonnes, on a cumulé d'abord 'eMR' et 'eTN' (Maroc et Tunisie) pour créer 'eAfN'; puis 'eTG' et 'eSN', pour créer 'eAeq'; le cumul de 'eAfN' avec 'eAeq' a donné 'eAfr', cumul des 4 exportateurs africains présents dans le tableau de base; enfin 'eAf%' est la colonne de pourcentage associée à 'eAfr'. On vérifiera que, comme pour les lignes de cumul, les pourcentages sont calculés sur l'ensemble des éléments de base: la colonne 'eAf%' n'est autre que 'eAfr', divisée par le total de ses 8 premiers nombres, puis multipliée par 100.

Voici maintenant le listage 'B:data:phos Σ^* Σ ', créé par 'cums' afin de rappeler à l'utilisateur la définition des lignes et colonnes ajoutées au tableau de base 'B:data:phos' pour créer le tableau 'B:data:phos Σ^* ww' (le caractère '*' étant ici le sigle choisi par l'utilisateur; et le ' Σ ' servant à rappeler qu'on procède par cumul). Quant à la composition des cumuls, la lecture du listage est facile; il faut seulement expliquer que le nombre 20015 est le rang de la ligne nouvelle

```

B:data:phos $\Sigma^*\Sigma$ 
20015 iEst ligne cumul des lignes suivantes
  12 iPL ; 13 iRM ; 14 iEE ;
20016 iEs% ligne cumul des lignes suivantes
  15 iEst ;
20009 eAfN colonne cumul des colonnes suivantes
  4 eMR ; 7 eTN ;
20010 eAeq colonne cumul des colonnes suivantes
  5 eSN ; 6 eTG ;
20011 eAfr colonne cumul des colonnes suivantes
  9 eAfN ; 10 eAeq ;
20012 eAf% colonne cumul des colonnes suivantes
  11 eAfr ;
20017 iEur ligne cumul des lignes suivantes
  1 iBL ; 3 iFR ; 4 iDL ; 5 iIT ; 7 iNL ; 8 iSP ;
  9 iUK ; 15 iEst ;
20018 iEu% ligne cumul des lignes suivantes
  17 iEur ;

```

'iEst' créée après les 14 lignes du tableau de base; de même 20012, avant le sigle 'eAf%' signifie qu'il s'agit d'une 12-ème colonne; etc...

Si l'on ouvre le programme 'cums' alors qu'existe déjà, créé sur un éditeur, un listage 'B:data:phos $\Sigma^*\Sigma x$ ', identique au listage 'B:data:phos $\Sigma^*\Sigma$ ' ci-dessus, et qu'on donne pour tableau des données 'B:data:phos' avec pour sigle du tableau à créer '*', le dialogue de spécification des cumuls n'a pas lieu; car 'cums' utilise le listage ' $\Sigma^*\Sigma x$ ' pour créer précisément le tableau correspondant. Mais il n'est pas nécessaire que le listage de commande ' $\Sigma^*\Sigma x$ ' soit aussi explicite que le listage ' $\Sigma^*\Sigma$ ' créé par 'cums'. De façon précise, examinons le bref listage qui a été utilisé.

```

phos $\Sigma^*\Sigma x$  : demande de cumul lignes et colonnes pour phos
20015 iEst ligne 10012 ; 14 ;
20016 iEs% ligne 15 ;
20009 eAfN colonne 4 ; 7 ;
20010 eAeq colonne 5 ; 6 ;
20011 eAfr colonne 9 ; 10 ;
20012 eAf% colonne 11 ;
20017 iEur ligne 1 ; 10003 ; 5 ; 10007 ; 9 ; 15 ;
20018 iEu% ligne 17 ;

```

Après une ligne de titre terminée dont le contenu est indifférent, on a, pour chaque élément à créer, une ligne (ou éventuellement un bloc de lignes si la place manque) en spécifiant le contenu. Commençons par la première ligne: le nombre 20015 est indispensable, mais sa valeur précise n'est que pour la commodité de l'utilisateur: 'cums' demande seulement un nombre de 20000 à 29999. Suit le sigle de l'élément à créer (on a déjà dit que si ce sigle ce termine par '%', 'cums' calculera des pourcentages). Le mot "ligne" ou "colonne" est indispensable, pour spécifier la nature de l'élément à créer.

On entre alors dans l'énumération des éléments à cumuler. Les sigles sont omis; en revanche, les ';', entourés de deux blancs, sont indispensables comme séparateurs. Reste à expliquer l'usage de nombres supérieurs à 10000: ceux-ci

équivalent à une commande de cumul de bloc. De façon précise, le cumul commence à partir du rang correspondant au nombre diminué de 10000; et va jusqu'au nombre suivant; à condition que celui-ci existe et qu'il ait un rang supérieur au point de départ supposé du cumul. Ici, on a: 10012 ; 14 ; le cumul porte donc sur les éléments {12, 13, 14}; mais s'il y avait: 10012 ; 11 ; seraient seulement pris les éléments isolés de rang '12' et '11', mais il n'y aurait pas de cumul de bloc.

Le cumul peut porter sur des blocs et des éléments isolés à la fois: ainsi, pour 'iEur', on a: {1, {3,4,5}, {7,8,9}, 15}. On notera que, ici comme au dialogue à l'écran, les éléments déjà créés peuvent être utilisés pour faciliter la spécification d'un cumul: e.g. 15 = 'iEst', dispense de répéter, d'une manière ou d'une autre {12,13,14}.

```
B:data:phosΣ**Σ
20015 itot ligne cumul des lignes suivantes
  1 iBL ; 2 iCA ; 3 iFR ; 4 iDL ; 5 iIT ; 6 iJP ;
  7 iNL ; 8 iSP ; 9 iUK ; 10 iIN ; 11 iBR ; 12 iPL ;
  13 iRM ; 14 iEE ;
20009 etot colonne cumul des colonnes suivantes
  1 eBL ; 2 eUS ; 3 eJR ; 4 eMR ; 5 eSN ; 6 eTG ;
  7 eTN ; 8 eCC ;
```

Quant à la création du listage de spécification de blocs, nous rappelons qu'on a pu utiliser le listage 'Σ**Σ' créé après le dialogue décrit au §2.1; listage que nous publions pour plus de clarté.

3 Une application à l'analyse discriminante

```
B:Ac:9tΣ*Σ
20159 Luca ligne cumul des lignes suivantes
  1 L@A ; 2 L@B ; 3 L@C ; 4 L@D ; 5 L@E ; 6 L@F ;
  7 L@G ; 8 L@H ; 9 L@I ; 10 L@J ; 11 L@K ; 12 L@L ;
  13 L@M ; 14 L@N ; 15 L@O ; 16 L@P ; 17 L@Q ; 18 L@R ;
  19 L@S ; 20 L@T ; 21 L@U ; 22 L@V ; 23 L@W ; 24 L@X ;
20160 Jean ligne cumul des lignes suivantes
  25 J@A ; 26 J@B ; 27 J@C ; 28 J@D ; 29 J@E ; 30 J@F ;
  31 J@G ; 32 J@H ; 33 J@I ; 34 J@J ; 35 J@K ; 36 J@L ;
  37 J@M ; 38 J@N ; 39 J@O ; 40 J@P ; 41 J@Q ; 42 J@R ;
  43 J@S ; 44 J@T ; 45 J@U ;
20161 Acta ligne cumul des lignes suivantes
  46 A@A ; .....
  .....
20175 Mar% ligne cumul des lignes suivantes
  166 Marc ;
20176 ARp% ligne cumul des lignes suivantes
  167 ARlp ;
```

Dans [NOUV. TEST. GREC], le tableau des données, 'B:Ac:9t', croise 158 chapitres de 9 œuvres (8 livres du Nouveau Testament et un livre de la *Politique* d'Aristote) avec un ensemble de 96 mots (plus exactement, formes de mots déclinés ou conjugués) dont on a dénombré le nombre d'occurrences dans chaque chapitre.

```

B:Ac:9tΣ*Σx exemple de demande de création de lignes par cumul
20159 Luca ligne 1 ; 2 ; 3 ; 4 ; 5 ; 6 ; 7 ; 8 ; 9 ; 10 ; 11 ;
    12 ; 13 ; 14 ; 15 ; 16 ; 17 ; 18 ; 19 ; 20 ; 21 ; 22 ; 23 ; 24 ;
20160 Jean ligne 25 ; 26 ; 27 ; 28 ; 29 ; 30 ; 31 ; 32 ; 33 ; 34 ;
    35 ; 36 ; 37 ; 38 ; 39 ; 40 ; 41 ; 42 ; 43 ; 44 ; 45 ;
20161 Acta ligne 46 ; .....
.....
20167 ARlp 96 ; 97 ; 98 ; 99 ; 100 ; 101 ; 102 ; 103 ; 104 ; 105 ;
    106 ; 107 ; 108 ;
.....
20176 ARp% ligne 167 ;

```

Les œuvres se succèdent par blocs de lignes; on reconnaît les chapitres d'une même œuvre à ce qu'ils commencent par une ou deux lettres avec ensuite un numérotage fait par des lettres (afin de prévoir le cas où il y aurait de très nombreux chapitres).

On a créé par cumul 18 lignes: une ligne de cumul proprement dit pour chaque œuvre; et, ensuite, une ligne de pourcentages. Nous donnons trois listages, ou extraits de listage, afin de montrer les diverses manières de spécifier les cumuls à effectuer.

Le listage le plus explicite est le listage 'B:Ac:9tΣ*Σ' créé par le programme 'cums' avec les sigles mêmes des chapitres des œuvres. On notera que le sigle du tableau complété est, ici encore, tout simplement '*'. Ensuite nous donnons deux exemples possibles de listages 'B:Ac:9tΣ*Σx' pouvant commander à 'cums' la création du même tableau 'B:Ac:9tΣ*ww'. Le premier de ces listages n'utilise pas la notation par blocs: l'écriture complète en serait très

```

B:Ac:9tΣ*Σx exemple de demande de cumul
20159 Luca ligne 10001 ; 24 ;
20160 Jean ligne 10025 ; 45 ;
20161 Acta ligne 10046 ; 73 ;
20162 Apoc ligne 10074 ; 95 ;
20163 Hebr ligne 10109 ; 121 ;
20164 Roma ligne 10127 ; 142 ;
20165 lJoh ligne 10122 ; 126 ;
20166 Marc ligne 10143 ; 158 ;
20167 ARlp ligne 10096 ; 108 ;
20168 Luc% ligne 159 ;
20169 Jea% ligne 160 ;
20170 Act% ligne 161 ;
20171 Apo% ligne 162 ;
20172 Heb% ligne 163 ;
20173 Rom% ligne 164 ;
20174 lJh% ligne 165 ;
20175 Mar% ligne 166 ;
20176 ARp% ligne 167 ;

```

fastidieuse; et c'est en vue de tels cas qu'on a prévu la commande par blocs.

Au demeurant, la création du présent tableau par dialogue n'offre pas de difficulté; car il n'y a qu'à spécifier 9 blocs par début et fin.

Quant à l'utilisation du tableau 'B:Ac:9tΣ*ww' ainsi créé, il nous suffira de dire que ce tableau (176 × 96) a été transposé; puis analysé, en mettant en principal soit les 9 cumuls de chapitres d'une œuvre, soit les '%'; soit encore seulement les cumuls de chapitres d'une œuvre du Nouveau Testament, à l'exclusion du livre de la *Politique* d'Aristote; les 158 chapitres étant toujours en supplémentaires. Les intéressants résultats obtenus, pour la discrimination des œuvres, sont exposés dans [NOUV. TEST. GREC], §4.4.4.