La structure d'ordre dans un tableau croisé : degré d'association

Philippe Cibois
(Printemps, Université de Versailles - St. Quentin)

Version française de « An Order on Cross-Tabulations and Degrees of Association »

BMS - Bulletin de Méthodologie Sociologique 2013, n°119, p. 24–43

Résumé. La structure d'ordre dans un tableau croisé : degré d'association. Il est souvent affirmé qu'une structure d'ordre existe dans un tableau croisé quand les marges du tableau disposent d'une telle structure. On peut s'affranchir de ce point de vue et définir précisément une structure d'ordre du tableau lui-même. Comme l'avait déjà remarqué Louis Guttman dans le cas de la scalogram analysis, il faut souvent des déplacements alternatifs des lignes et des colonnes pour parvenir à repérer une échelle. Dans ce cas, c'est bien l'ordre du tableau structuré qui induit un ordre sur les marges et non l'inverse. Cependant Goodman et Kruskal au moment où ils présentent l'indice gamma qui permet de définir l'intensité d'une liaison dans le cas ordonné, n'utilisent que l'ordre des marges, et ils ont été suivis depuis. Il convient de revenir à l'intuition de Guttman et d'utiliser des résultats obtenus ultérieurement pour montrer qu'au moins une approximation d'une structure d'ordre est pratiquement toujours présente sur un tableau. Le tableau croisé issu de questions ordonnées n'est qu'un cas parmi d'autres et inversement un tableau disposant d'une liaison ordonnée forte induit un ordre interprétable sur les modalités des questions. En partant d'exemples réels on montrera que l'on dispose de critères pour définir un ordre sur un tableau, de méthodes formalisées pour rendre apparente la structure associée, de différents indices pour mesurer l'intensité de la liaison, de tests pour en évaluer le degré de signification.

Mots clefs :Tableau croisé, structure d'ordre, indice gamma, indice PEM

Mesurer le degré d'association entre lignes et colonnes dans un tableau croisé est une question qui a été abordée depuis plus d'un siècle. Si l'on ne retient que les méthodes encore utilisées, on trouve le coefficient de contingence de Karl Pearson (1904), le coefficient de Tschuprow (1925), le coefficient de Cramér (1946). Cette question a été traitée dans une série de quatre articles du *Journal of the American Statistical Association* par Goodman et Kruskal (1954, 1959, 1963, 1972) qui les ont ensuite rassemblés dans un ouvrage (1979). Par contre, l'idée que s'il existe une structure d'ordre sur les lignes et les colonnes, le tableau résultant a une structure particulière qui puisse être repérée a fait l'objet de peu de recherches : ce sera l'objet de notre première partie avant de reprendre le problème du degré d'association.

1. Structure d'ordre d'un tableau à marges ordonnées

Louis Guttman dans le 4^{e} volume de *l'American soldier* (1950) pose les bases de la *scalogram analysis* avec laquelle il met en ordre un tableau croisant des réponses et des individus en recherchant par des déplacements alternatifs des lignes et des colonnes à parvenir à une forme homogène dont il déduit un ordre qu'il appelle une échelle. C'est bien l'ordre du tableau structuré qui induit un ordre sur les marges et non l'inverse. Cette technique sera développée ultérieurement par Bertin (1967).

Avant de reprendre cette idée, nous montrerons à partir d'exemples, comment se pose le problème.

Supposons un tableau 2 x 2 A_iB_j dont les marges A_i et B_j possèdent une structure d'ordre définie de la façon suivante $A_1 > A_2$ et $B_1 > B_2$. Soit l'exemple suivant :

Tableau2X2	B_1	B_2	Total
A_1			140
A_2			60
Total	80	120	200

Les marges sont ordonnées mais le tableau lui-même ne possède pas nécessairement une structure d'ordre comme dans le cas suivant

Tableau2X2	B_1	B_2	Total
A_1	56	84	140
A_2	24	36	60
Total	80	120	200

En effet, pour la première case A_1B_1, on voit que le produit des marges divisé par le total est égal à 56, l'effectif observé. Nous sommes donc bien dans le cas d'indépendance entre lignes et colonnes, pourtant ordonnées.

Pour s'écarter de cette situation, on peut soit ajouter, soit retrancher un individu dans cette case A_1B_1. Si on l'ajoute, on suppose qu'il y a une attraction entre A_1 et B_1 et par conséquent, puisqu'on se situe dans l'univers fixe des marges, une opposition entre A_1 et B_2 ainsi qu'entre A_2 et B_1 et enfin une autre attraction entre A_2 et B_2. Le déplacement élémentaire est alors le suivant :

	B_1	B_2
A_1	+1	-1
A_2	-1	+1

Le tableau résultant de ce déplacement est le suivant :

	B_1	B_2	Total
A_1	57	83	140
A_2	23	37	60
Total	80	120	200

Ce tableau a maintenant une structure d'ordre qui est définie par la structure des signes des écarts à l'indépendance : avant d'envisager une définition de la structure d'ordre, on peut dire qu'un tableau 2 x 2 est ordonné quand une des diagonales porte des écarts positifs et l'autre des écarts négatifs.

On peut répéter le déplacement élémentaire un certain nombre de fois, mais le nombre maximum de déplacements est atteint quand la case A_1B_1 est égal à 80 du fait de la contrainte de marge (et la case A_2B_1 est égale de ce fait à zéro:

	B_1	B_2	Total
A_1	80	60	140
A_2	0	60	60
Total	80	120	200

On a donc opéré 24 déplacements élémentaires qui ont correspondu à la décroissance progressive de la case A_2B_1 de 24 à 0.

Le déplacement élémentaire avec les signes inverses entraine un tableau qui est dans l'ordre inverse de celui de A et B.

	B_1	B_2
A_1	-1	+1
A_2	+1	-1

et le maximum de la liaison en sens inverse de la liaison entre A et B est :

	B_1	B_2	Total
A_1	20	120	140
A_2	60	0	60
Total	80	120	200

Dans ce cas 36 déplacements élémentaires ont été nécessaires. On a donc en tout 60 déplacements auxquels il faut ajouter la situation d'indépendance soit 61 situations possibles (ce qui correspond à la plus petite des marges + une unité). Comme une seule case définit l'ensemble de ce tableau à un seul degré de liberté, on peut donc résumer l'ensemble des cas possibles de la manière suivante en prenant A_2B_2 comme référence:

60 – 59 – 58 – 57 – 56 ……37 – **36** – 35 4 – 3 – 2 – 1 – **0**

Max. de la liaison indépendance……..Max. de la liaison inverse

Si le tableau observé se trouve entre 37 et 60, le tableau est ordonné dans le sens de A et B ; si l'observé se trouve entre 35 et 0, le tableau est ordonné dans le sens inverse de A et B. Tous ces tableaux (sauf celui de l'indépendance) sont munis d'une structure d'ordre définie par les deux déplacements élémentaires et leurs répétitions possibles.

Comme la situation d'indépendance joue un rôle central, on peut, en soustrayant la valeur de l'indépendance à l'échelle précédente, examiner l'échelle des écarts à l'indépendance, toujours pour la case A_2B_2

+24 +23 +22 +21 +20 ……+1 **0** -1 -32 -33 -34 -35 **-36**

Max. de la liaison indépendance…….. Max. de la liaison inverse

Soit par exemple le tableau des Écarts pour $A_2B_2 = 50$

	B_1	B_2
A_1	+14	-14
A_2	-14	+14

et le tableau pour $A_2B_2 = 30$

	B_1	B_2
A_1	-6	+6
A_2	+6	-6

Cette échelle nous fournit également un indice d'intensité de la liaison en situant un écart par rapport au maximum.

Pour le cas A_2B_2 = 50, l'écart est de 14 par rapport à un maximum de 24, il représente donc 14 / 24 x 100 = 58,3% du maximum, indicateur que l'on nommera de ce fait PEM *Pourcentage de l'Écart Maximum* (Cibois 1993).

Pour le cas A_2B_2 = 30 le maximum de la liaison dans le sens négatif serait de -36, l'écart est de -6 qui représente -6 / -36 x 100 = 16,7% du maximum, que l'on affecte par convention du signe négatif pour indiquer qu'il s'agit d'un écart négatif.

Comme nous l'avons montré (Cibois 1993), il est possible d'étendre la procédure de recherche d'un PEM à chacune des cases du tableau : il suffit d'isoler la case dont on cherche l'intensité de la liaison et de regrouper toutes les autres lignes dans une seule ligne et toutes les autres colonnes dans une seule colonne, ce qui nous ramène au cas du tableau 2 x 2.

En conclusion, dès qu'il n'est plus dans la situation d'indépendance, un tableau 2 x2 possède toujours une structure d'ordre identifiable par un ordre existant sur les marges. Comme dans le cas général, la situation d'indépendance tombe rarement sur un effectif observable, on peut dire que la structure d'ordre est pratiquement générale.

1.1 Un exemple réel : Londres 1911

Nous allons désormais travailler sur un tableau réel pris dans Kendal & Stuart (1961 : 558) et qui donne les résultats d'une enquête faite à Londres en 1911 (noté Londres 4x6[1]).

The table (shows the distribution of 1725 school children who were classified (1) according to ther standard of clothing (Very well clad, Welle clad, Poor but passable, Very badly clad), and (2) according to their intelligence (Very able, Distinctly capable, Fairly intelligent, Slow but intelligent, Dull, Mentally deficient or slow and dull)

Londres46	VABL	DCAP	FINT	SLBI	DULL	DEFI	Total
VWEL	39	194	209	113	48	33	636
WELL	15	138	255	202	100	41	751
POOR	4	33	61	70	58	39	265
VBAD	1	10	10	22	13	17	73
Total	59	375	535	407	219	130	1725

Soit le tableau suivant à 3 lignes et 3 colonnes (noté Londres 3x3) obtenu en regroupant les colonnes deux à deux et les lignes 3 et 4 :

Londres33	Intel+	Intel=	Intel-	Total
Hab+	233	322	81	636
Hab=	153	457	141	751
Hab-	48	163	127	338
Total	434	942	349	1725

Londres33 peut se décomposer en la somme des deux tableaux correspondant à l'indépendance et aux écarts à l'indépendance :

[1] Pour *Londres* au format d'origine avec 4 lignes et 6 colonnes

Indépendance	Intel+	Intel=	Intel-	Total
Hab+	160,0	347,3	128,7	636
Hab=	188,9	410,1	151,9	751
Hab-	85,0	184,6	68,4	338
Total	434	942	349	1725

Ecarts	Intel+	Intel=	Intel-
Hab+	**73,0**	-25,3	-47,7
Hab=	-35,9	**46,9**	-10,9
Hab-	-37,0	-21,6	**58,6**

Autour de la première diagonale où les écarts sont tous positifs (marqués en gras), tous les écarts sont négatifs. Cependant la notion de diagonale doit être précisée si le nombre de lignes et de colonnes ne sont plus égaux d'une part et même dans le cas d'égalité quand la structure des marges a des effets déformants.

- inégalité du nombre de lignes et de colonne.

Formons un nouveau regroupement de Londres (Londres 2x3) où les colonnes sont regroupées comme précédemment et les lignes 2 à 4 sont regroupées. On a la décomposition suivante :

Londres23	Intel+	Intel=	Intel-	Total
HabitsSup	233	322	81	636
HabitsInf	201	620	268	1089
Total	434	942	349	1725

Indépendance	Intel+	Intel=	Intel-	Total
HabitsSup	160	347,3	128,7	636
HabitsInf	274	594,7	220,3	1089
Total	434	942	349	1725

Ecarts	Intel+	Intel=	Intel-
HabitsSup	**73**	-25,3	-47,7
HabitsInf	-73	**25,3**	**47,7**

On voit que sur la colonne 2, l'écart positif se trouve dans la 2^{e} ligne.

- contraintes de marges

On fait un nouveau tableau Londres 3 x 3 (Londres33B) en conservant les mêmes regroupements de colonnes mais en rendant les lignes moins équilibrées en ce qui concerne les marges. On regroupe les lignes 1 et 2 (désormais HabA) et on laisse identique les deux lignes restantes (POOR devient HabB et VBAD devient HabC).

On a la décomposition suivante :

Londres33B	Intel+	Intel=	Intel-	Total
HabA	386	779	222	1387
HabB	37	131	97	265
HabC	11	32	30	73
Total	434	942	349	1725

Indépendance	Intel+	Intel=	Intel-	Total
HabA	349,0	757,4	280,6	1387
HabB	66,7	144,7	53,6	265
HabC	18,4	39,9	14,8	73
Total	434	942	349	1725

Ecarts	Intel+	Intel=	Intel-
HabA	**37,0**	**21,6**	-58,6
HabB	-29,7	-13,7	**43,4**
HabC	-7,4	-7,9	**15,2**

On voit cette fois que l'effet diagonal existe toujours mais qu'il est déformé (écarts positifs en gras). On voit ici que le fort poids de la marge HabA a tiré vers la droite et le haut la diagonale des écarts positifs.

Donc, soit pour des raisons de format, soit pour des raisons de contraintes de marges, seules les cases extrêmes de la diagonale sont toujours en écart positif (pour respecter l'ordre des marges). Pour passer d'une extrémité à l'autre, le chemin des écarts positifs peut s'écarter plus ou moins de la diagonale, les écarts positifs étant toujours contigus (latéralement) ou adjacents (en diagonale). C'est l'existence de cette « ligne de crête » où se trouvent les écarts positifs qui isole tous les écarts négatifs qui sera la définition d'un tableau muni d'une structure d'ordre.

Définition : un tableau est muni d'une structure d'ordre quand la diagonale, qui relie (par contiguïté latérale ou par adjacence diagonale) les cases extrêmes définies par l'ordre sur les lignes, correspond à des écarts positifs à l'indépendance. Les écarts négatifs se situent de part et d'autre de cette diagonale.

Décomposons finalement le tableau d'origine de Londres 1911 dont les écarts à l'indépendance sont les suivants :

Londres46	VABL	DCAP	FINT	SLBI	DULL	DEFI
VWEL	**17,2**	**55,7**	**11,7**	-37,1	-32,7	-14,9
WELL	-10,7	-25,3	**22,1**	**24,8**	**4,7**	-15,6
POOR	-5,1	-24,6	-21,2	**7,5**	**24,4**	**19,0**
VBAD	-1,5	-5,9	-12,6	**4,8**	**3,7**	**11,5**

Une « ligne de crête » relie bien les deux extrémités de la diagonale, elle est plus ou moins large. Tous les écarts positifs situés sur cette ligne sont contigus et/ou adjacents, tous les écarts négatifs se trouvent de part et d'autre de la ligne de crête.

1.2 Situation réciproque

Reprenons maintenant la problématique initiée par Guttman et posons-nous maintenant la question réciproque : si on découvre une structure d'ordre dans un tableau, quel ordre cela implique-t-il pour les lignes et les colonnes ? Prenons l'exemple du tableau suivant : il s'agit d'un tableau issu d'une enquête sur les opinions politiques et syndicales d'ouvriers français en 1970 (Adam 1970) dont on extrait un tableau portant sur la confiance dans les syndicats en fonction du syndicat choisi en cas d'élection professionnelle.

Les lignes du tableau sont ordonnées (« Pour la défense de vos intérêts, faites-vous très confiance, plutôt confiance, plutôt pas confiance, pas confiance du tout à l'action des syndicats ») mais les colonnes sont des réponses à la question suivante : « en cas d'élections professionnelles dans votre entreprise, est-ce que vous voteriez plutôt pour une liste présentée par FO, CFDT, des non-syndiqués, CGT, un syndicat autonome ou indépendant, CFTC, vous ne voteriez pas ? » Voici le tableau d'observation et celui des écarts à l'indépendance

Confiance dans les syndicats	FO	CFDT	Non-Synd	CGT	Auto	CFTC	Non-Vote	Total
Très confiance	14	24	12	137	11	4	6	208
Plutôt confiance	38	43	22	137	40	12	45	337
Plutôt pas confiance	15	7	19	25	25	4	34	129
Pas conf. du tout	11	13	38	18	25	3	62	170
Total	78	87	91	317	101	23	147	844

Écarts	FO	CFDT	Non-Synd	CGT	Auto	CFTC	Non-Vote
Très confiance	-5,2	**2,6**	-10,4	**58,9**	-13,9	-1,7	-30,2
Plutôt confiance	**6,9**	**8,3**	-14,3	**10,4**	-0,3	**2,8**	-13,7
Plutôt pas confiance	**3,1**	-6,3	**5,1**	-23,5	**9,6**	**0,5**	**11,5**
Pas conf. du tout	-4,7	-4,5	**19,7**	-45,9	**4,7**	-1,6	**32,4**

En mettant en relief graphiquement les écarts à l'indépendance positifs on constate des proximités de profils : entre CGT et CFDT (écarts positifs pour les forts degrés de confiance) ; entre FO et CFTC (écarts positifs pour les degrés intermédiaires) et entre Autonomes, non-syndiqués et non-votants (écarts positifs pour les degrés les plus bas de confiance).

On peut réordonner le tableau de façon à retrouver la structure d'ordre définie préalablement où les écarts positifs partitionnent le tableau autour de la première diagonale, les écarts négatifs étant de part et d'autre :

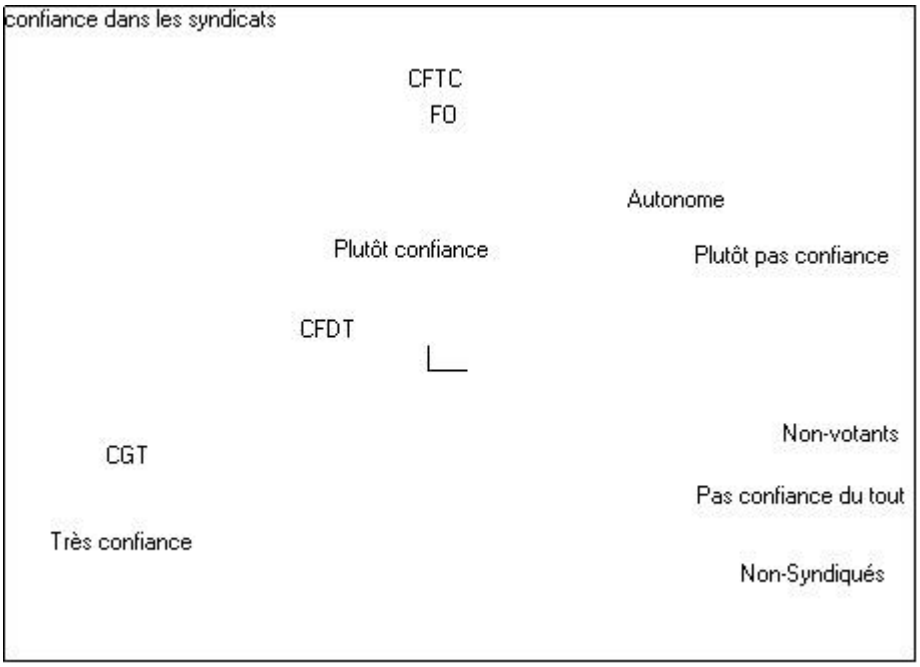
Écarts	CGT	CFDT	CFTC	FO	Auto	Non-Synd	Non-Vote
Très confiance	**58,9**	**2,6**	-1,7	-5,2	-13,9	-10,4	-30,2
Plutôt confiance	**10,4**	**8,3**	**2,8**	**6,9**	-0,3	-14,3	-13,7
Plutôt pas confiance	-23,5	-6,3	**0,5**	**3,1**	**9,6**	**5,1**	**11,5**
Pas conf. du tout	-45,9	-4,5	-1,6	-4,7	**4,7**	**19,7**	**32,4**

Il reste à définir ce qu'est l'ordre sur les colonnes : comme en France, les deux syndicats CGT et CFDT sont les plus revendicatifs alors que CFTC et FO ont des

positions moins radicales et que les syndicats autonomes sont le plus souvent des syndicats maisons mis en œuvre pour s'opposer aux syndicats contestataires, on peut réinterpréter la question posée en fonction des réponses obtenues. L'ordre sur les syndicats manifeste le degré d'opposition à l'ordre établi (Cibois 1984 : 20-21).

1.3 Recherche d'une structure d'ordre

Le problème précédent était particulièrement simple puisqu'il existait déjà une structure d'ordre sur les lignes et qu'il suffisait donc de quelques permutations sur les colonnes pour mettre à jour la structure d'ordre sur le tableau. Pour résoudre le problème dans le cas général, nous ferons appel à la technique de l'analyse des correspondances car Benzécri (1976 : 279-280) montre que s'il existe une structure d'ordre sur les lignes et les colonnes, le premier facteur d'une analyse des correspondances manifestera cet ordre. On peut le vérifier pour le tableau précédent dont le plan du premier facteur (horizontal) et du deuxième (vertical) permet de retrouver l'ordre sur les lignes et les colonnes.



On peut calculer tous les PEM positifs du tableau et relier les points par un trait d'autant plus fort que le PEM est élevé. Voici le tableau des PEM :

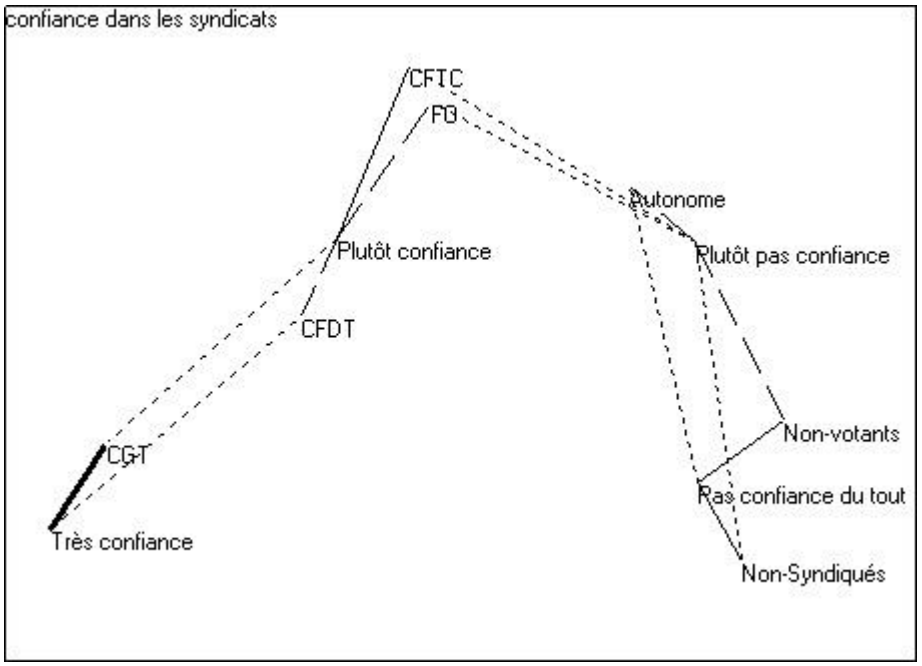
PEM	CGT	CFDT	CFTC	FO	Auto	Non-Synd	Non-Vote
Très confiance	45,3	3,9	-29,4	-27,2	-55,8	-46,5	-83,4
Plutôt confiance	5,5	15,8	20,4	14,6	-0,8	-39,5	-23,3
Plutôt pas confiance	-48,4	-47,4	2,5	4,7	11,2	6,6	10,8
Pas conf. du tout	-71,8	-25,8	-35,2	-30	5,8	27,1	27,6

Précisons sur cet exemple la procédure pour calculer un PEM sur une case (dit PEM local). Soit à chercher l'intensité de la liaison entre la ligne « très confiance » et le syndicat CGT. On réduit le tableau au tableau 2 x 2 suivant sur lequel on peut opérer comme précédemment. On procède de cette façon pour chacune des cases du tableau:

Tableau 2 x 2	CGT	Autres colonnes	Total
Très confiance	137	71	208
Autres lignes	180	456	636
Total	317	527	844

écart à l'indépendance observé = 137 – (208 x 318 / 844) = 58,9
écart à l'indépendance dans le cas du maximum = 208 - (208 x 318 / 844) = 129,9
PEM local = 58,9 / 129,9 x 100 = 45,3%

Visualisation des PEM locaux dans le plan factoriel :



La structure d'ordre du tableau suit bien le premier facteur (horizontal) de l'analyse des correspondances.

Nous disposons désormais d'une procédure pour rechercher une structure d'ordre dans tout tableau croisé : examinons maintenant la question de l'intensité de la liaison.

2. L'intensité de la liaison

Nous recherchons un indicateur nous donnant le degré d'association entre l'ordre des lignes et celui des colonnes. Nous partirons du jugement de Goodman et Kruskal (1954) qui ont repris complètement le problème et proposés des indices d'association qui ne soient plus basés sur le Khi-deux car « The fact that an excellent test of independence may be based on χ^2 does not at all mean that χ^2, or some simple function of it, is an appropriate *measure* of degree of association » (1954 : 740). Ensuite nous critiquerons cet indice et proposerons une généralisation du PEM.

2.1 Le gamma de Goodman et Kruskal

Pour présenter cet indicateur, nous réutiliserons les données de Londres d'abord sous forme du tableau 2 x 2 suivant :

Londres22	IntellSup	IntelInf	Total
HabitsSup	850	537	1387
HabitsInf	119	219	338
Total	969	756	1725

Sur un tel tableau 2 x 2, Yule (1900) définissait un coefficient d'association en utilisant les produits croisés (850 x 219 = 186 150 et 119 x 537 = 63 903) : s'ils sont égaux, il y a indépendance et le coefficient Q d'association est égal au rapport de leur différence sur leur somme :

Q = (186 150 - 63 903) / (186 150 + 63 903) = 122 247 / 250 053 = 0,489

Goodman et Kruskal reprennent cette idée d'utiliser les produits croisés : ils appellent *situation d'accord* les produits croisés de la première diagonale 850 x 219 (et le symétrique 219 x 850) où, quand on passe d'une case à l'autre, le rang s'élève tant pour les lignes que pour les colonnes. Symétriquement, ils appellent *situation de désaccord* quand le rang s'élève pour les lignes mais baisse pour les colonnes (ou inversement). C'est le cas de la deuxième diagonale où par exemple de 119 à 537 on passe de la case Habits Inf – Intelligence Sup à la case Habits Sup – Intelligence Inf : on monte dans l'ordre de l'habillement mais on descend dans l'ordre de l'intelligence. C'est donc un cas de désaccord de rang. Formellement, Goodman et Kruskal (1954 : 749) définissent les cas (en proportion) de la manière suivante :

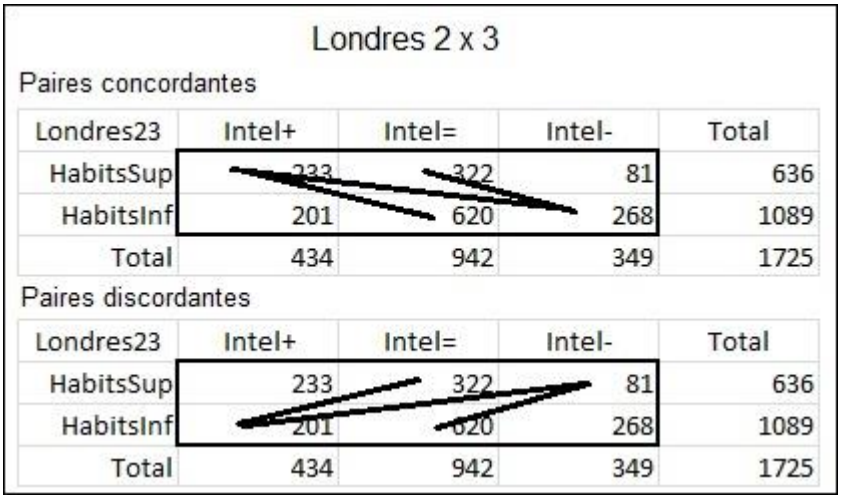
Π_s= Pr {$a_1 < a_2$ and $b_1 < b_2$; or $a_1 > a_2$ and $b_1 > b_2$}
Π_d =Pr { $a_1 < a_2$ and $b_1 > b_2$; or $a_1 > a_2$ and $b_1 < b_2$}
Π_t = Pr { $a_1 = a_2$ or $b_1 = b_2$}.

Les cas d'égalité dans le cas présent correspondent aux paires 119-850, 119-219, etc. ainsi que les paires correspondant à l'identité 119-119 etc. : elles ne sont pas pris en compte dans le calcul de gamma.

Goodman et Kruskal définissent Gamma comme $\gamma = (\Pi_s - \Pi_d) / (\Pi_s + \Pi_d)$, ce qui dans le cas du tableau 2 x 2 correspond au Q de Yule. Mais ils généralisent : pour comprendre ce qui se passe alors, reprenons les données de Londres 2 x 3.

Londres23	Intel+	Intel=	Intel-	Total
HabitsSup	233	322	81	636
HabitsInf	201	620	268	1089
Total	434	942	349	1725

Si nous prenons la paire de cases en opposition sur la première diagonale, on voit que si l'on part de 127, par rapport à 386, on monte dans l'ordre des lignes et des colonnes. Mais c'est aussi le cas pour 127 à 779 et pour 163 vers 386. Visualisons ces paires concordantes et discordantes.



Calculons gamma à partir des effectifs des produits des paires concordantes et des paires discordantes, on a :

233 x 620 =	144460
233 x 268 =	62444
322 x 268 =	86296
Paires concordantes =	293200

201 x 322 =	64722
201 x 81 =	16281
620 x 81 =	50220
Paires discordantes =	131223

Gamma : C-D / C + D =	0,382

La justification de ces calculs par Goodman et Kruskal est la suivante :

Suppose that two individuals are taken independently and at random from the population. Each falls into some (A_a, B_b) cell. (…) If there is high association one expects that the order of the *a*'s would generally be the same as that of the *b*'s.

Faire les produits des paires concordantes revient à compter les paires d'individus en situation d'ordre ; faire les produits des paires discordantes revient à compter les paires d'individus qui ne sont pas en situation d'ordre. Plus la situation d'ordre par rapport au total augmente, plus l'association est importante. Plusieurs coefficients utilisent des comptages de nombre de paires : le tau de Kendall, le tau-C de Stuart, les D asymétriques de Somers. Quand les lignes et colonnes ne sont pas munies d'un ordre, ces techniques ne peuvent plus être utilisées et on constate que souvent les utilisateurs reviennent aux indicateurs dérivés du Khi-deux, malgré leur mise en accusation par Goodman et Kruskal.

La difficulté de cette procédure est que la recherche des paires concordantes et discordantes ne tient aucun compte de la structure observée du tableau et ne se basent que sur l'ordre des lignes et des colonnes, alors qu'une structure d'ordre existe peut-être. On s'affranchit de cette difficulté en mettant en ordre les lignes et colonnes en suivant le premier facteur de l'analyse des correspondances qui donne toujours un ordre que nous utiliserons pour la mise au point d'un indice d'association dérivé du PEM.

2.2 Le PEM global

Le coefficient d'association général envisagé et une mesure d'association entre lignes et colonnes qui fait l'hypothèse que :

- si une structure d'ordre est connue pour les lignes et les colonnes, celle-ci peut être constatée empiriquement et que inversement,
- si une structure d'ordre n'a pas été repérée elle existe peut-être cependant, même si cet ordre n'est pas très prononcé,
- le coefficient peut être utilisé pour déterminer l'intensité de l'association dans une case du tableau et pour le tableau entier.
- en suivant les recommandations de Goodman et Kruskal rappelées plus haut, il n'utilisera pas le khi-deux
- sa valeur sera nulle en cas d'indépendance
- il variera entre -1 et 1 pour aller de la dépendance dans un sens à la dépendance dans l'autre (le signe étant conventionnel). Des valeurs proches de ces valeurs maximums doivent correspondre à des cas qui se rencontrent empiriquement.
- les valeurs de l'indice doivent être comparables d'un tableau à un autre, même s'ils sont différents quant aux nombres de lignes ou colonnes.
- le principe du coefficient doit être simple à comprendre, même s'il est le résultat d'opérations longues qui ne peuvent être réalisée à la main que dans les cas élémentaires.

Comme la situation d'indépendance est parfaitement définie et sert toujours pour indiquer l'absence d'association, on prendra comme principe pour mesurer l'association d'examiner (dans la logique du PEM local) le rapport entre la somme des écarts positif à l'indépendance dans le cas observé à la somme des écarts positifs dans le cas ou la liaison serait à son maximum.

Reprenons le tableau ordonné par le premier facteur de l'analyse des correspondances de l'enquête sur la confiance dans les syndicats. En reprenant le tableau des écarts à l'indépendance, la somme des écarts à l'indépendance est égale à 176,26

Il faut maintenant définir le maximum. On a un tableau ordonné dont on ne conserve que les marges : comme le tableau est ordonné, la diagonale des écarts positifs part soit de la liaison entre CGT et « très confiance », soit de la case Non-votant, « pas confiance du tout ». Le choix du point de départ est indifférent et conduit dans les deux cas au même résultat.

Partons de la case en haut à gauche. Tous les CGT qui sont 317 ne peuvent être « très confiant dans les syndicats » puis que la marge correspondante n'est que de 208 mais inversement tous les « très confiants » peuvent être mis dans la case CGT.Il restera alors 317 – 208 = 109 CGT que l'on mettra dans la case contigüe (latérale) la plus proche, « plutôt confiance ».Le tableau sera alors le suivant :

Confiance	CGT	CFDT	CFTC	FO	Auto	Non-Synd	Non-Vote	Total
Très	208							208
Plutôt	109							337
Plutôt pas								129
Pas du tout								170
Total	317	87	23	78	101	91	147	844

Tous les effectifs de la ligne 1 et de la colonne 1 sont maintenant répartis. Pour la ligne 2 il reste 337 (marge) – 109 (CGT) soit 228 « plutôt confiance ». On peut les répartir dans CFDT (87), CFTC (23) FO (78), il en reste 40 à placer qui seront mis

dans « autonomes ». Toute la 2e ligne est placée et on repart en colonne, ou il reste 101 – 40 autonomes à placer que l'on mettra dans la case contigüe « plutôt pas confiance ». Tous les autonomes sont placés, mais il reste dans les « plutôt pas », 129 – 61 = 68 qui seront placés en non-syndiqués, dont les 23 restants ainsi que tous les non-votants seront mis en « pas confiance du tout ». Ce qui donne le tableau final (qui aurait pu être obtenu avec le même algorithme en partant de la case « Non-votant » - « Pas confiance du tout ». La solution est unique et on trouvera l'algorithme programmé en annexe 1.

Confiance	CGT	CFDT	CFTC	FO	Auto	Non-Synd	Non-Vote	Total
Très	208							208
Plutôt	109	87	23	78	40			337
Plutôt pas					61	68		129
Pas du tout						23	147	170
Total	317	87	23	78	101	91	147	844

La somme des écarts positifs à l'indépendance sur ce maximum est de 464,53

Le PEM global est le rapport des deux sommes : 176,26 / 464,53 x 100 = 37,9%

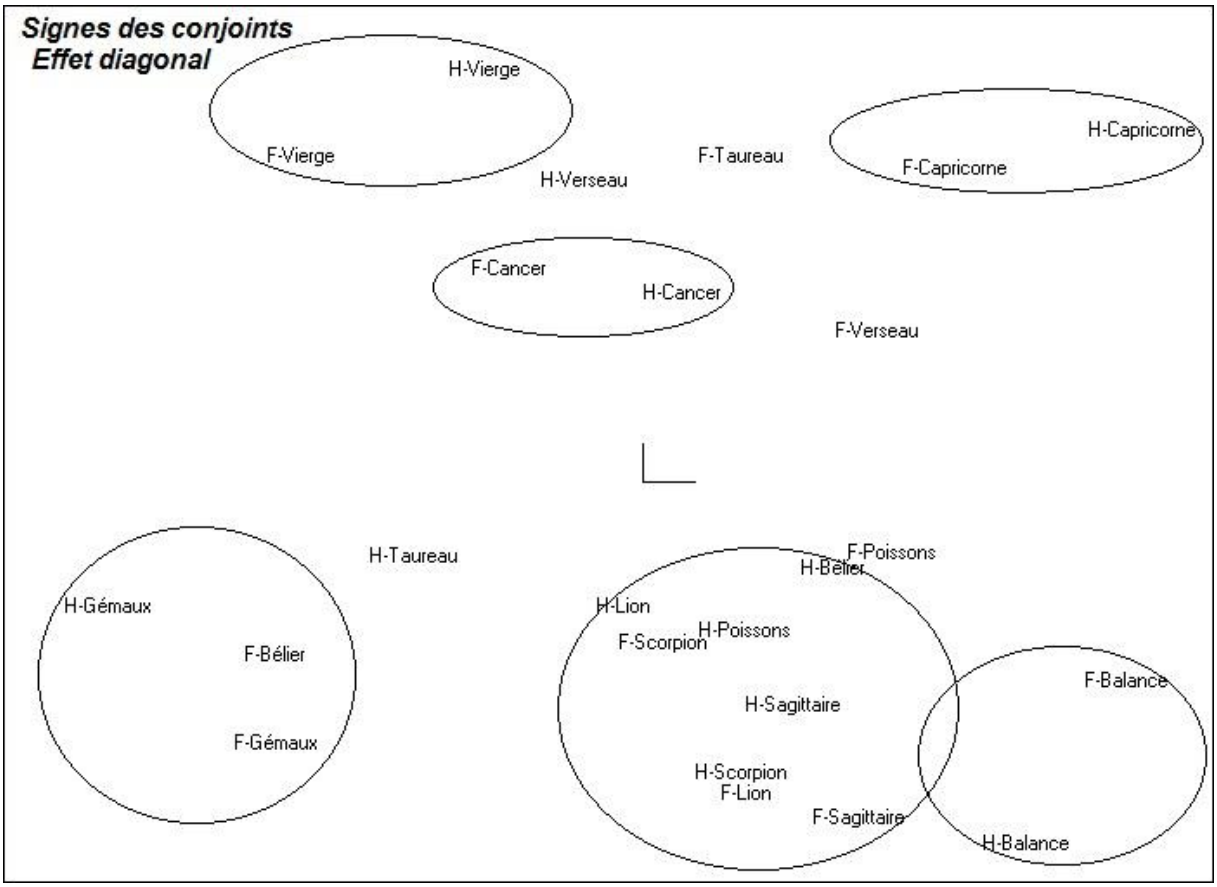
Sur des tableaux issus de données réelles, comme il est toujours possible de mettre en ordre les données selon le premier facteur de l'AFC, on peut dire qu'il existe toujours une structure d'ordre et qu'il est toujours possible de calculer un PEM global. Ce résultat peut sembler dangereux car dans certains cas, cet ordre peut être entièrement lié à une structure aléatoire de données qui n'ont pas véritablement d'ordre. Nous allons affronter cette situation dans le cas suivant où l'on sait *a priori* qu'il n'y a pas d'ordre sur les lignes et les colonnes.

2.3 Les signes des conjoints

Étudions donc un cas où la structure d'ordre est absente et soumettons-le aux procédures de recherche d'une structure d'ordre. Il s'agit d'un tableau qui avait été réalisé dans le but de montrer l'échec de l'astrologie (données présentées dans Cibois 1997). Pour une population de 68 000 couples, on construit un tableau de 12 lignes et 12 colonnes, les lignes correspondent aux signes astrologiques des hommes et les colonnes à ceux des femmes. On note à l'intersection d'une ligne et d'une colonne, le nombre de couples ayant des signes identiques.

	Femmes												
	Ver	Poi	Bel	Tau	Gém	Can	Lio	Vie	Bal	Sco	Sag	Cap	Total
H-Verseau	536	478	518	535	532	500	451	478	478	413	430	502	5851
H-Poisson	482	592	536	541	525	506	484	463	503	475	443	482	6032
H-Bélier	555	560	596	584	525	508	543	452	525	461	451	521	6281
H-Taureau	511	508	582	607	552	523	527	462	490	448	438	460	6108
H-Gémaux	488	497	557	520	577	496	469	461	433	433	421	458	5810
H-Cancer	487	508	512	530	478	504	446	436	462	397	420	456	5636
H-Lion	456	502	522	482	478	461	466	431	455	440	402	472	5567
H-Vierge	445	463	489	500	426	464	413	457	409	381	395	434	5276
H-Balance	490	494	482	493	481	450	482	406	494	392	449	440	5553
H-Scorpion	441	437	459	483	464	433	426	382	434	392	432	401	5184
H-Sagittaire	455	445	475	436	456	423	411	395	443	377	419	435	5170
HCapricorne	498	496	445	554	456	461	443	398	469	411	398	494	5523
Total	5844	5980	6173	6265	5950	5729	5561	5221	5595	5020	5098	5555	67991

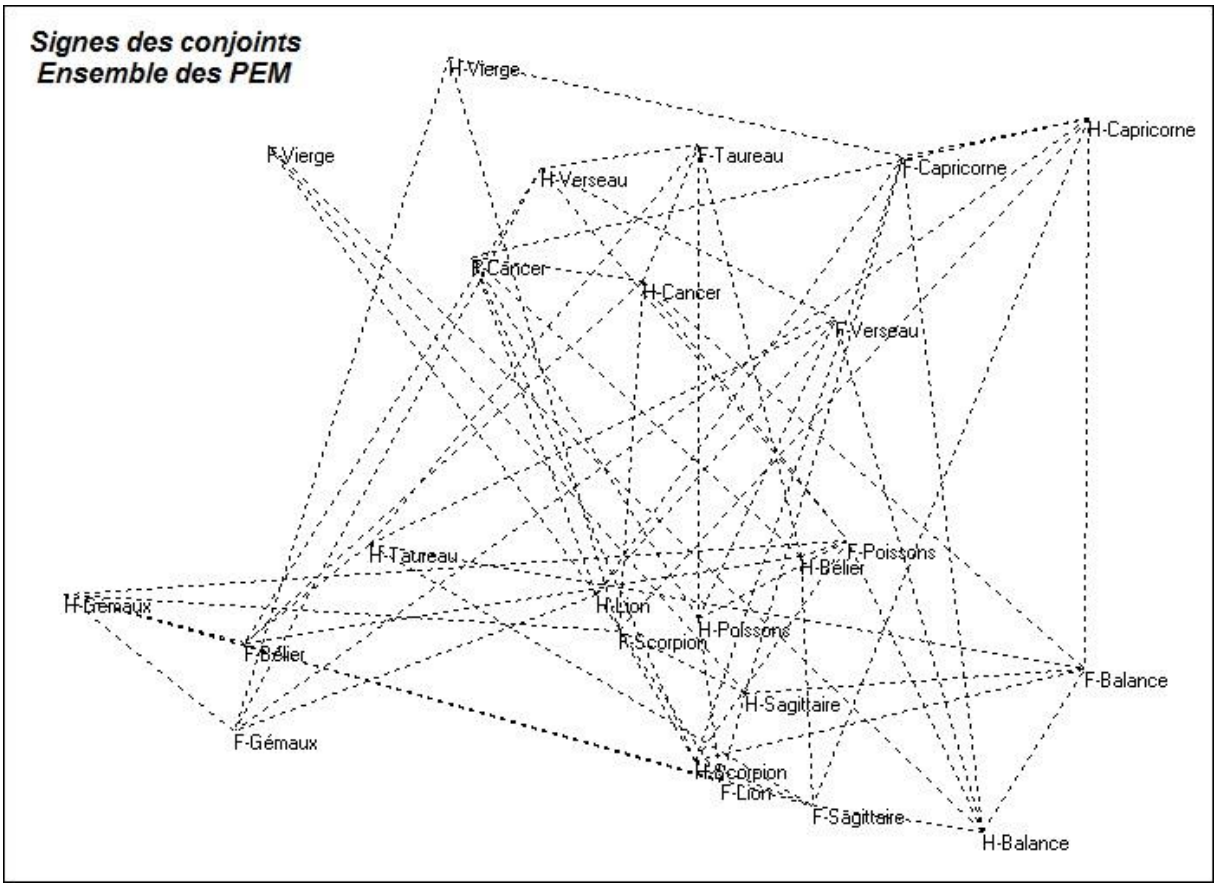
Si l'on fait l'analyse des correspondances de ce tableau, on peut être troublé par le graphique factoriel qui met en évidence des proximités entre signes qui sont soulignées ci-dessous par des ovales. En effet 10 signes sur 12 sont en proximité (la seule exception nette étant le signe du Taureau, celui du Verseau étant moins net).



Cette situation s'explique si on examine les écarts à l'indépendance : ici on n'a retenu que les écarts positifs supérieurs à 9. On constate que tous les écarts de la diagonale sont positifs, ce qui explique les rapprochements précédents.

	Femmes											
	Ver	Poi	Bel	Tau	Gém	Can	Lio	Vie	Bal	Sco	Sag	Cap
H-Verseau	**33**				20			29				24
H-Poisson		**61**								30		
H-Bélier	15		**26**				29					
H-Taureau			27	**44**	17		27					
H-Gémaux			30		**69**			15				
H-Cancer		12		11		**29**						
H-Lion		12	17				**11**			29		17
H-Vierge			10	14		19		**52**				
H-Balance	13						28		**37**		33	
H-Scorpion					10					**9**	43	
H-Sagittaire	11								18		**31**	13
HCapricorne	23	10		45					15			**43**

Cependant, si on examine l'ensemble des PEM individuels, on s'aperçoit que ces écarts diagonaux sont du même ordre de grandeur que les autres et d'ailleurs moins nombreux qu'eux.



Si le premier facteur (horizontal) de l'analyse des correspondances propose bien un ordre, cet ordre correspond à une quasi-absence de liaison puisque le PEM global est égal à 2,0%

Un autre indice que ce tableau est très proche de l'indépendance nous est fourni par la première valeur propre de l'analyse des correspondances qui est très faible et égale à 0,0006

Quant à l'effet diagonal, il s'explique par le fait que même les gens qui ne croient pas à l'astrologie pensent que l'astrologie affirme que les personnes de même signe s'attirent. Il s'agit donc d'un effet auto-réalisateur faible mais perceptible.

Avant d'étudier les problèmes de significativité des résultats, comparons les différents indices étudiés pour les différents exemples étudiés.

	PEM	1ère VP	Gamma	V Cramér	% Cramér	Khi-deux	p=
Londres23	26,6%	0,049	0,382	0,220	4,85%	83,63	0,000
Londres33	20,4%	0,066	0,368	0,198	3,90%	134,69	0,000
Londres33b	21,7%	0,050	0,415	0,158	2,51%	86,62	0,000
Londres46	23,3%	0,079	0,332	0,184	3,38%	174,82	0,000
Confiance Synd	37,9%	0,238	0,525	0,295	8,73%	221,05	0,000
Signes Conjoints	2,0%	0,0006	0,017	0,014	0,02%	139,17	0,124

- l'indice gamma n'a pu être calculé sur les deux derniers tableaux qu'en faisant l'hypothèse qu'ils possédaient une structure d'ordre obtenue par le premier facteur de l'analyse des correspondances. Dans ce cas gamma donne des résultats analogues à ceux du PEM et tout autant interprétables.

- bien qu'il soit critiqué pour son emploi du Khi-deux, le V de Cramèr réagit comme les autres indices mais sur une autre plage ; comme les autres il est très faible quand on est proche de l'indépendance (signes du zodiaque).

- on appelle pourcentage de Cramér l'indice tel qu'il a été défini par Cramér lui-même et non tel qu'il a été interprété par les autres auteurs ensuite qui en ont pris la racine carrée. En effet Cramér dit (1946) que « φ^2 / q-1 [q étant la plus petite dimension du tableau] may be used as a measure, on a standardized scale, of the degree of dependance between the variables » (p. 282). Cette proportion du maximum peut être lue en pourcentage. On constate que cet indice est très pessimiste.

- le problème de la significativité du PEM peut être résolu d'une façon simple en considérant la significativité du Khi-deux pour le tableau. Quand le PEM est calculé sur un tableau significatif, il acquière de ce fait sa significativité. C'est le cas ici pour tous les tableaux sauf le dernier.

En ce qui concerne les plages d'utilisation du PEM, l'expérience montre que les PEM intéressants se situent entre 10% et 50%. Les liaisons plus fortes sont souvent l'indice d'une redondance entre indicateurs Quand la liaison est inférieure à 10%, elle peut être l'effet du hasard et le test du Khi-deux permet de s'en rendre compte.

3. Conclusion

Le PEM global peut donc être utilisé comme indicateur de l'intensité d'une liaison entre lignes et colonnes sur tous les tableaux croisés. S'il y a un ordre existant sur les lignes et les colonnes, il sera retrouvé par le premier facteur de l'AFC. Si ce n'est pas le cas, il faudra éventuellement remettre en cause l'ordre défini *a priori* ou comprendre pourquoi il y a divergence. Si l'on fait confiance à l'ordre défini préalablement, on peut alors l'utiliser pour calculer le tableau maximum. Si l'ordre déterminé par le premier facteur d'une AFC n'est

pas interprétable, on est alors dans une situation liée à une structure d'écarts aléatoires et le PEM sera vraisemblablement faible et le tableau non significatif au sens du Khi-deux.

Le PEM est disponible dans les logiciels Trideux[2] et Modalisa[3]. Sa programmation ne pose pas de problèmes particuliers : on trouvera en annexe la programmation en R faite par Nicolas Robette[4]

Annexe

```
# ================================
# Fonction R pour le calcul du PEM
# (pourcentage de l'écart maximum,
#  proposé par Philippe Cibois)
# ================================
# x doit être un objet table ou matrix
# la fonction retourne les PEM locaux ($peml) et le PEM global ($pemg)

pem <- function(x) {
  tota <- colSums(x)
  totb <- rowSums(x)
  total <- sum(x)
  theo <- matrix(nrow=nrow(x),ncol=ncol(x))
  for(i in 1:nrow(x)) { for(j in 1:ncol(x)) theo[i,j] <-
tota[j]*totb[i]/total }
  ecart <- x-theo
  max <- matrix(nrow=nrow(x),ncol=ncol(x))
  emax <- matrix(nrow=nrow(x),ncol=ncol(x))
  pem <- matrix(nrow=nrow(x),ncol=ncol(x))
  for(i in 1:nrow(x)) { for(j in 1:ncol(x)) {
    if(ecart[i,j]>=0) max[i,j] <- min(tota[j],totb[i])
    if(ecart[i,j]<0&tota[j]<=(total-totb[i])) max[i,j] <- 0
    if(ecart[i,j]<0&tota[j]>(total-totb[i])) max[i,j] <- tota[j]+totb[i]-
total
    emax[i,j] <- max[i,j] - theo[i,j]
    pem[i,j] <- ifelse(ecart[i,j]>=0,ecart[i,j]/emax[i,j]*100,0-
ecart[i,j]/emax[i,j]*100)
    }}
  dimnames(pem) <- dimnames(x)
  cor <- corresp(x,nf=1)
  z <- x[order(cor$rscore),order(cor$cscore)]
  tota <- colSums(z)
  totb <- rowSums(z)
  maxc <- matrix(0,nrow=nrow(z),ncol=ncol(z))
  i <- 1; j <- 1
  repeat {
    m <- min(tota[j],totb[i])
    maxc[i,j] <- m
    tota[j] <- tota[j] - m
    totb[i] <- totb[i] - m
    if(sum(tota)+sum(totb)==0) break
    if(tota[j]==0) j <- j+1
    if(totb[i]==0) i <- i+1
```

[2] http://cibois.pagesperso-orange.fr/Trideux.html

[3] http://www.modalisa.com/

[4] http://nicolas.robette.free.fr/outils.html Je remercie Nicolas Robette pour cette procédure ainsi que pour les commentaires faits à propos du présent texte.

```
  }
  pemg <- (sum(ecart)+sum(abs(ecart)))/(sum(maxc-
theo[order(cor$rscore),order(cor$cscore)])+sum(abs(maxc-
theo[order(cor$rscore),order(cor$cscore)])))
  rm(tota,totb,total,theo,ecart,max,emax,cor,z,m,maxc,i,j)
  PEM <- list(peml=round(pem,1),pemg=round(100*pemg,1))
  return(PEM)
}
```

Références

Adam, G. Bon,F. Capdevielle, J. Mouriaux, R.(1970). *L'ouvrier français en 1970*, Paris, Presses de la FNSP.

Benzécri, J.-P., (1976). *L'analyse des données. Tome I La taxinomie*, Paris, Dunod.

Bertin, J. (1967). *Sémiologie graphique*. Paris ; La Haye : Mouton ; Paris : Gauthier-Villars

Cibois, Ph. (1984). *L'analyse des données en sociologie*. Paris, Presses universitaires de France.

Cibois, Ph. (1993). Le PEM, pourcentage de l'écart maximum : un indice de liaison entre modalités d'un tableau de contingence, *Bulletin de méthodologie sociologique*, n°40, p.43-63.

Cibois, Ph. (1997). Les pièges de l'analyse des correspondances, *Histoire & Mesure*, 12 (3/4), pp. 299-320.

Pearson, K. (1904) On the theory of contingency and its relation to association and normal correlation. *Drapers'Co. Memoirs, Biometric Series*, N°. 1, London (repris dans *Karl Pearson's Early Statistical Papers*. Cambridge Univ. Press, London, 1948)

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton Univ. Press.

Goodman, L. A. and Kruskal, W. H. (1954). Measures of Association for Cross Classifications. *Journal of the American Statistical Association*, 49, 732-764.

Goodman, L. A. and Kruskal, W. H. (1959). Measures of Association for Cross Classifications. II: Further Discussion and References. *Journal of the American Statistical Association*, 54, 123-163.

Goodman, L. A. and Kruskal, W. H. (1963). Measures of Association for Cross Classifications. III: Approximate Sampling Theory. *Journal of the American Statistical Association*, 58, 310-364.

Goodman, L. A. and Kruskal, W. H. (1972). Measures of Association for Cross Classifications. IV: Simplification of Asymptotic Variances. *Journal of the American Statistical Association*, 67, 415-421.

Goodman, L. A. and Kruskal, W. H. (1979). *Measures of Association for Cross Classifications*. New York, Springer-Verlag.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer et al.: *Measurement and prediction*. Princeton, N. J. : Princeton Univ. Press, p. 60-90.

Kendal, M. and Stuart, A (1961). The Advanced Theory of Statistics, Volume 2, London : C. Griffin and C.

Tschuprow, A. A. (1925). *Grundbegriffe und Grundprobleme der Korrelationstheorie.* Leipzig, Berlin, Teubner.

Yule, G.U. (1900). On the Association of Attibutes in Statistics : with Illustrations from the Material of the Childhood Societey, &c. *Philosophical Transactions of the Royal Society of London*, Series A. Vol. 194.