Essai de Typologie Structurelle des Indices de Similarité Vectoriels par Unification Relationnelle

François Marcotorchino

Thales Communications : 160, boulevard de Valmy – BP 82
92704 Colombes Cedex et Laboratoire de Statistique Théorique et Appliquée (Paris VI)
jeanfrancois.marcotorchino@fr.thalesgroup.com

Résumé : Cet article a pour but de proposer un regard nouveau et unificateur à la problématique des Indices de Similarité et des Critères de structuration ou de partitionnement. Une catégorisation des indices, des propriétés non connues ainsi qu'une présentation dans différents axes de structuration seront suggérées. La recherche des significations et des filiations associées sera donnée comme résultat dérivé de ce travail.

1 Introduction

La recherche sur les indices de similarité a, depuis longtemps, donné lieu à une abondante littérature, les références aux indices ayant été faites souvent, même dans les meilleurs articles, sous forme de listes (type inventaire) sans qu'une réelle structuration n'ait été proposée. Nombre de ces indices ont été introduits et donnés dans différents articles et dans différents domaines fondamentaux ou d'application, au fur et à mesure de leur utilisation potentielle. Ainsi n'est-il pas étonnant de trouver ces indices proposés et introduits dans des domaines aussi variés que: *les Sciences Humaines* et plus particulièrement la Sociologie et l'Ethnologie et ses dérivées : Ethnopsychologie, Ethnogénétique, etc.., *la Linguistique* proprement dite, dont: Lexicologie, Lexicométrie, Ethnolinguistique, Text Mining, *les Mathématiques*: Analyse des données, Classification et « Clustering », *les Sciences du vivants* : la Biologie, la Biométrie, la Physiologie, la Phylogénie, la Zoologie, la Médecine, *les Sciences organiques* : la Chimie moléculaire, la Biochimie et enfin de nombreux domaines plus « business », comme : le « Customer Relationship Management », le « Business Intelligence », la « Géo-cartographie », le «Profiling » etc..

Presque tous les indices courants ont été introduits à des périodes et à des dates différentes, pour des buts et motifs variés sans structuration et explication claire du rôle de chacun et sans aucun regard sur une filiation ou une hérédité sous jacentes permettant de mieux les comprendre ou de les interpréter (ceci étant sans doute dû à la provenance très différenciée des inventeurs). Ce phénomène s'est traduit, de fait, soit par une impression de fouillis, soit, et on le verra plus loin dans ce document, par des successions de redécouvertes (parfois très récentes) d'indices existants depuis fort longtemps, ou de mises en évidence de propriétés connues depuis très longtemps par des chercheurs de disciplines différentes.

Preuve que le sujet est toujours d'actualité, un article vraiment très récent de Matthijs J. Warrens (2009) vient de paraître dans la Revue Journal of Classification, au moment

même où nous envoyons le notre pour publication. Il aborde quelques uns des points exposés dans cet article, en particulier l'ordonnancement de certains des indices présentés dans notre document, en cherchant une unification des bornes par moyennage (Harmonique, Géométrique et Arithmétique), sans exploiter pourtant d'autres considérations comme les structures homographiques associées.

Même des articles synthétiques et plus théoriques, extraits d'ouvrages plus mathématiques n'ont pas couvert complètement cette problématique. Ainsi l'article de F.B. Baulieu (1989) dans la Revue « Journal of Classification » (Springer Verlag), qui représente l'un des meilleurs survols du sujet, propose une liste très abondante d'indices, sans qu'elle soit, à son tour, ni complète ni exhaustive. Cette liste est d'ailleurs présentée, en vrac, sans organisation aucune (sans citer le nom des « inventeurs »), le but étant de notifier un certain nombre de propriétés possédées ou non par ces indices; aucune mention n'est faite à propos de leurs filiations, de leurs ressemblances ou de leur hérédité, en un mot de leur structuration ou unification réciproque. Parmi les auteurs qui se sont, en France, intéressés le plus en profondeur à ce domaine, qui semble simple (uniquement en apparence), mais qui est en fait difficile, on pourra consulter utilement les travaux de S. Joly et G. Le Calvé (1986) et (1994) ainsi que ceux de B. Fichet (1994) dans le livre édité par B. Van Cutsem (ouvrage d'ensemble, excellent et de référence publié chez Springer Verlag en (1994), ceux de IC Lerman (1970), (1981), (1987), de G. Bisson (2000) auxquels on rajoutera, toujours dans le livre sus-cité de B. Van Cutsem (1994), les contributions additionnelles de B. Van Cutsem et F. Critchley .

Nous avions déjà présenté, il y a quelque temps, dans un article la Revue de Statistique Appliquée titré « Agrégation des Similarités », F. Marcotorchino et P. Michaud (1981) , un essai de catégorisation des indices de similarité les plus courants. Bien que certains résultats originaux aient été présentés dans cet article, nous n'avions pas vu à cette époque, tout l'intérêt de cette restructuration et surtout, nous ne l'avions pas appliquée à un ensemble étendu d'indices, nous nous étions contentés des plus connus et des plus représentatifs.

C'est à cette tâche que nous nous attelons dans cet article, et ce sont certaines des conclusions associées qui peuvent soulever un intérêt à la fois d'un point de vue théorique et pratique. Les différents axes de structuration des Indices de Similarité que nous allons proposer, donneront une vue d'ensemble et permettront d'en déduire des propriétés, qui, bien que souvent naturelles, ont été peu proposées sous cette forme dans la littérature. Nous revendiquons le fait que nous ne sous sommes pas intéressé en profondeur ici aux propriétés très importantes de « métricité » et de « semi-definie positivité » des matrices de similarité associées, du fait qu'elles avaient déjà été amplement étudiées, en particulier on trouvera de larges développements sur ce domaine dans l'ouvrage pré-cité de B. Van Cutsem (1994), mais également dans l'article de J. Gower et P. Legendre (1986), où dans celui de G. Le Calvé (1994) également.

Cela peut paraître légèrement présomptueux de rédiger un « nième » article sur le sujet, à la date d'aujourd'hui, vu le nombre incroyable d'articles ayant de près ou de loin traité de ce problème, nous pensons, néanmoins, qu'un certain nombre des résultats qui sont donnés dans cet article sont originaux ou du moins peu connus, même de spécialistes de l'analyse des données. C'est d'ailleurs essentiellement la raison qui nous a fait reprendre un article déjà publié en interne Thales en 2002, en lui rajoutant un certain nombre de concepts et de résultats et dont ce nouvel article est une substantielle amélioration. Bien que nous abordions en fin d'article les indices calculés sur « matrices d'abondance », l'essentiel de ce travail a trait aux indices de similarités calculés sur variables ou entités binaires. Compte

tenu de la difficulté générale du sujet lié à la « similarité », il semble que des articles du même type que celui ci devraient être consacrés à l'étude des indices de similarités entre structures plus complexes que les structures de listes ou vectorielles (similarités entre graphes , similarités profondes entre textes, similarités topologiques entre objets ou entre structures diverses..etc..). Malgré le nombre très important d'articles sur ces thématiques, il semble qu'une synthèse globale reste néanmoins à faire.

1.1 Considérations Axiomatiques Générales

L'axiomatique relative aux principes structurants le domaine des indices de similarités est, somme toute, assez pauvre, non seulement au niveau mathématique mais également au niveau philosophique. Nous avions signalé ce fait dans un article écrit conjointement avec H. Benhadda, H. Benhadda et F. Marcotorchino (1996), suivant en cela les traces d'une réflexion du philosophe D. Parrochia (1992) sur le sujet. De même on trouve chez les philosophes logiciens Willard van Orman Quine, en particulier dans l'un de ses derniers articles, voir W.V. O. Quine (1998) et Rudolf. Carnap (1928) , comme l'avait également fait en son temps le philosophe épistémologiste L. Brunchwicg (1921), quelques réflexions intéressantes sur la notion de « similarité » ainsi que sur les structures ontologiques et les principes descriptifs associés. Mais ceci aboutit a une axiomatique relativement faible dont nous donnons ci-dessous quelques principes de base et auxquels nous ferons parfois référence dans la suite.

a) Axiome n°1 : **Positivité**

Un indice de similarité S, est une fonction prenant ses valeurs dans I x I, où I est un Ensemble *fini* d'objets, à valeur dans $\Re^+$ (ensemble des réels positifs ou nuls), donc S est une fonction positive[1], d'où :

$$S(x, y) \geq 0 \quad \forall \ (x, y) \in \ IxI$$

x, peut représenter un vecteur dans un espace à plusieurs dimensions (disons R^m, ou appartenir à R, tout simplement) . Sauf cas particulier, nous identifierons :

$\vec{x} = \{ x_1, x_2, x_3 .... x_m \}$ à x , et $\vec{y} = \{ y_1, y_2, y_3 .... y_m \}$ à y , pour éviter les notations vectorielles.

b) Axiome n°2 : **Symétrie**

Un indice de similarité S, est une fonction prenant ses valeurs dans I x I à valeur dans $\Re^+$ (ensemble des réels positifs ou nuls) et telle que, pour deux objets x et y appartenant à un ensemble fini I , S doit vérifier l'axiome ou le principe de symétrie :

$$S(x, y) = S(y, x) \quad \forall \ (x, y) \in \ IxI$$

---

[1] La positivité est un axiome non totalement obligatoire, car certains indices notés « ρ_{xy} » (comme on le verra) varient de -1 à +1 et ont des comportements de coefficients de corrélation. Ceci étant dit, il suffit de remplacer ρ_{xy} par $S(x,y)=1/2(\rho_{xy}+1)$, pour se trouver titulaire d'un indice « S », vérifiant alors tous les « bons » axiomes. De la même façon , on passe d'un « indice de distance » d_{xy} ou d'une « distance vraie » à un indice de similarité en posant $S(x, y) = \frac{1}{1 + d_{xy}}$, ou $S(x,y)=1-d(x,y)$, ou bien encore $S(x,y)=1-½\ d^2(x,y)$ (on abordera ce point dans le texte)

c) Axiome n°3 : **Auto-similarité maximale** (ou Axiome de « **propreté** »)
Un indice de similarité S, est dit « propre », on dit aussi qu'il vérifie l'axiome d' « auto-similarité maximale » si :
S vérifie l'inégalité :

$$S(x, x) \geq S(x, y) \quad \forall \; y \neq x, \quad y \in I$$

Puisque on a de même :

$$S(y, y) \geq S(x, y) \quad \forall \; y \neq x, \quad y \in I$$

on en déduit que l'axiome d'auto-similarité maximale induit la propriété :

$$S(x, y) \leq \text{Min}\left(S(x, x), S(y, y)\right) \quad \forall \; x, y \in IxI$$

d) Axiome n°4 : Transitivité Généralisée Indicielle et dérivés
Dans le cas où un indice de similarité vérifie $0 \leq S(x,y) \leq 1$

1. et que son auto-similarité maximale implique $S(x,x)= 1$ (l'indice est alors dit : « normé »)
   (si ce n'est pas le cas il suffit simplement d'effectuer la transformation suivante :

$$S(x, y) \rightarrow S'(x, y) = \frac{S(x, y)}{\underset{x, y}{\text{Max}}\, S(x, y)}$$

   pour se trouver titulaire d'un indice variant de 0 à 1. On dit alors que l'indice est « re-normé »

2. et qu'il vérifie la propriété suivante dite de « Transitivité Généralisée Indicielle » :

(f 1) $$S(x,y) + S(y,z) - S(x,z) \leq 1 \quad \forall\, (x,y,z)$$

Alors, la quantité $d(x,y)=1 - S(x,y)$ vérifie l'inégalité triangulaire :
En effet :
L'inégalité précédente sur $S(x,y)$ en posant $S(x,y) = 1 - d(x,y)$, implique :
$(1-d(x,y)) + (1-d(y,z)) - (1-d(x,z)) \leq 1 \;\Rightarrow\; d(x,z) \leq d(x,y) + d(y,z)$ (inégalité triangulaire)

De plus du fait que $S(x,x)=1$ on a : $d(x,x)=0$ (dans ce cas $d(x,y)$ est une « semi-distance »)
Si en plus on a : $S(x,y)=1 \Rightarrow x = y$ alors $d(x,y)=0 \Rightarrow x = y$ et $d(x,y)$ est une « distance » vraie.

Par ailleurs, J.C. Gower (1966) a généralisé un résultat de I. Schoenberg datant de (1938) I. Schoenberg (1938) à propos du fait que si la matrice $\{S(x,y)\}$ est définie non négative (NND) et que $S(x,x) =1, \forall\, x$, alors la quantité :

$$d(x, y) = \sqrt{2[1 - S(x, y)]}$$

représente une distance euclidienne.

La « **Transitivité Généralisée Indicielle** » (TGI) est, pour un indice de similarité, la propriété « duale » de l' « **Inégalité Triangulaire** » pour une distance ou une dissimilarité.

En particulier si S(x,y) est un indice de similarité vérifiant les trois premiers axiomes et qu'il est construit à partir d'un indice de distance sous la forme : $S(x, y) = 1 - d(x, y)$ ou $S(x, y) = 1 - \frac{1}{2} d^2(x, y)$, on dira que l'indice de similarité est « métrisable » (premier cas) ou «métrisable euclidien » (deuxième cas) .

e) Théorème n°1 : (Généralisation homographique de la TGI)

Si $S_d(x,y)$ est un indice de similarité normé, c'est à dire dont la valeur maximale est égale à 1 et vérifiant la propriété d'auto-similarité maximale , alors tout indice $S_u(x,y)$, vérifiant l'auto-similarité maximale de la forme :

$$S_u(x,y) = \frac{S_d(x,y)}{(\alpha - \beta S_d(x,y))} \quad \Leftrightarrow \quad S_d(x,y) = \frac{\alpha S_u(x,y)}{\beta S_u(x,y) + 1}$$

vérifiera la propriété de « Transitivité Généralisée Indiciale » et sera donc métrisable si les coefficients α et β de la fonction homographique précédente, liant l'indice $S_d(x,y)$ à $S_u(x,y)$ vérifient eux –mêmes des propriétés particulières que nous allons expliciter.

Tout d'abord la fonction homographique associée $S_u(x,y)$ doit vérifier :

$$S_u(x,x) = \frac{S_d(x,x)}{\alpha - \beta S_d(x,x)} = \frac{1}{\alpha - \beta} = 1 \Rightarrow \text{l'autosimilarité maximale sera obtenue si : } \alpha = \beta + 1$$

et elle doit s'annuler si $S_d(x,y)=0 \Rightarrow$ le numérateur du rapport est forcément de la forme : $\alpha S_u(x,y)$, car sinon on ne peut annuler la fonction homographique associée.
D'autre part si l'indice $S_u(x,y)$ est métrisable, il doit pouvoir vérifier une inégalité de la forme donnée ci dessous, (en effet il suffit de remplacer $S_d(x,y)$ par sa valeur en fonction de $S_u(x,y)$ dans l'inégalité TGI, pour obtenir :

$$S_u(x,y) + S_u(y,z) - S_u(x,z) \le 1 - \frac{1}{\mu}[1 - S_u(x,y)][1 - S_u(y,z)][\delta S_u(x,z) + \lambda] \quad \forall x,y,z$$

la quantité :

$$\frac{1}{\mu}[1 - S_u(x,y)][1 - S_u(y,z)][\delta S_u(x,z) + \lambda] \quad \forall x,y,z \text{ devant être positive et inférieure à } 1$$

ce qui est vrai car $(1-S_u(x,y)) \le 1$ idem pour $(1-S_u(y,z)) \le 1$, ce qui implique donc que les coefficients : δ, λ, μ soient positifs.
Ceci induit, après calculs d'identification des coefficients des formes monomiales associées, les relations suivantes entre les différents coefficients :

(i) $\alpha = \beta + 1$ (déjà vue)
(ii) $\mu = \alpha^2$
(iii) $\delta = \beta^2$
(iv) $\lambda = \alpha^2 - 1$

On voit que ces différentes valeurs sont paramétrables par rapport à un seul paramètre : α ou β, de ce fait, cette inégalité se réécrit en fonction d'un seul paramètre par exemple « β », selon l'expression :

(f2) $$S_u(x,y) + S_u(y,z) - S_u(x,z) \le 1 - [1 - S_u(x,y)][1 - S_u(y,z)]\left[\left(\frac{\beta}{\beta+1}\right)^2 S_u(x,z) + \frac{\beta(\beta+2)}{(\beta+1)^2}\right] \quad \forall x,y,z$$

soit également sous la forme équivalente:

$$S_u(x,y)+S_u(y,z)-S_u(x,z)\le 1-\frac{\beta}{(\beta+1)^2}[1-S_u(x,y)][1-S_u(y,z)]\,[\beta\,S_u(x,z)+(\beta+2)]\quad\forall x,y,z$$

Comme la quantité : $\frac{\beta}{(\beta+1)^2}[1-S_u(x,y)][1-S_u(y,z)]\,[\beta\,S_u(x,z)+(\beta+2)]\quad\forall x,y,z$

est positive si $\beta\ge 0$, l'indice $S_u(x,y)$ vérifie bien :

$$S_u(x,y)+S_u(y,z)-S_u(x,z)\le 1\quad\forall x,y,z$$

c'est à dire l'inégalité TGI, il est donc métrisable. (cqfd)

En remplaçant $S_u(x,y)$ par $(1-d_u(x,y))$ (son indice de distance) associé, l'inégalité (f 2) se transforme en :

$$[1-d_u(x,y)]+[1-d_u(y,z)]-[1-d_u(x,z)]\le 1-d_u(x,y)\,d_u(y,z)\left[\left(\frac{\beta}{\beta+1}\right)^2[1-d_u(x,z)]+\frac{\beta(\beta+2)}{(\beta+1)^2}\right]\quad\forall x,y,z$$

soit :

$$d_u(x,z)\le\frac{d_u(x,y)+d_u(y,z)-\frac{2\beta}{\beta+1}d_u(x,y)\,d_u(y,z)}{1-\left(\frac{\beta}{\beta+1}\right)^2 d_u(x,y)\,d_u(y,z)}\qquad\forall x,y,z$$

Théorème n°1 : Tout indice de similarité $S_u(x,y)$, fonction homographique d'un indice $S_d(x,y)$ vérifiant l'inégalité TGI, normé et exprimable sous la forme : $S_u(x,y)=\frac{S_d(x,y)}{\beta[1-S_d(x,y)]+1}$, vérifie à son tour l'inégalité TGI et est donc métrisable.

Théorème n°1 (bis) : D'autre part, si l'indice $S_d(x,y)= 1-1/2d^2(x,y)$, alors l'indice $S_u(x,y)=\frac{S_d(x,y)}{\beta[1-S_d(x,y)]+1}$ est exprimable suivant la formule $S_u(x,y)=1-1/2d'^2(x,y)$ et est alors métrisable Euclidien .

En effet si $S_u(x,y)$ est bien exprimable suivant la formule précédente alors il existe une métrique Euclidienne $d'^2(x,y)$ telle que :

$1-\frac{1}{2}d'^2(x,y)=\frac{1-\frac{1}{2}d^2(x,y)}{\frac{\beta}{2}d^2(x,y)+1}$, montrons que ceci implique que $d'^2(x,y)=2(1-S'(x,y))$, donc que $d'^2(x,y)$ vérifie le théorème de Gower-Schoenberg .

De l'égalité précédente on tire : $d'^2(x,y)=\frac{2(\beta+1)\,d^2(x,y)}{\beta\,d^2(x,y)+2}$ soit :

$$d'^2(x,y)=2\left[1-\frac{2-d^2(x,y)}{\beta\,d^2(x,y)+2}\right]$$

En remplaçant $d^2(x,y)$ par $2(1-S_d(x,y))$ dans la formule ci-dessus, il vient :

$$d'^2(x,y)=2\left[1-\frac{S_d(x,y)}{\beta[1-S_d(x,y)]+1}\right]$$

soit $d'^2(x,y) = 2(1-S'(x,y))$ puisque l'on reconnaît dans $S'(x,y)$ la quantité $S_u(x,y)$ (indice de similarité normé) que nous avions définie précédemment. $d'^2(x,y)$ vérifiant le Théorème

de Gower Schoenberg est donc une distance Euclidienne. Et l'indice $S_u(x,y)$ est bien métrisable Euclidien.(cqfd)

De la même façon, si un indice de similarité normé, est construit à partir d'une expression comme : $S(x,y) = 1 - d(x,y)$ et si $d(x,y)$ vérifie l'inégalité « ultramétrique » suivante :

$$d(x,z) \leq \text{Max}\,(d(x,y), d(y,z)) \quad \forall\ x, y,z$$

Alors l'indice de similarité associé vérifiera l'inégalité de *Transitivité Ultramétrique Indicielle* suivante :

(f 3) $$S(x,z) \geq \text{Min}\,(S(x,y), S(y,z))$$

Ce résultat s'obtient en remplaçant $d(x,y)$ par $1\text{-}S(x,y)$ dans l'inégalité ultramétrique sur $d(x,y)$ précédente. Cette inégalité implique la Transitivité Généralisée Indicielle, en effet comme $S(x,z) \geq \text{Min}\,(S(x,y), S(y,z))$ on a :

$$S(x,y) + S(y,z) - S(x,z) \leq S(x,y)+S(y,z)- \text{Min}\,(S(x,y), S(y,z))$$

Soit donc :

$$S(x,y) + S(y,z) - S(x,z) \leq S(x,y)+ S(y,z)- [\ 1/2(S(x,y)+ S(y,z)) - 1/2\ \left|S(x,y)-S(y,z)\right|\ )$$

Et donc finalement :

$$S(x,y) + S(y,z) - S(x,z) \leq \text{Max}\,(S(x,y),S(y,z)) \leq 1$$

(cqfd)

f) Pseudo Axiome n°5 : **Axiome dit du « typage de Carnap »**

On pourrait rajouter à cette liste un « pseudo » axiome « logique » associé à l'axiome 3, celui dit « du typage » dû au philosophe logicien Autrichien Rudolf Carnap (1928)qui peut se traduire par :

Un « type TY » étant défini par l'ensemble des individus vérifiant un ensemble fini de propriétés :

$$TY(x) = \{x \mid x, \text{ vérifie les propriétés } (p_1,p_2,\ldots p_u)\}$$

alors pour tout individu « w » tel que $w \notin TY$, on doit avoir :

$$\underset{(x,y)\in TY\times TY}{\text{Min}}\ S(x,y) \geq S(x,w)\cdot$$

En d'autres termes tous les objets vérifiant un type TY donné doivent être plus semblables entre eux qu'ils ne le sont à tout objet quelconque extérieur à l'ensemble caractérisé par ce Type . Attention, ce « pseudo » axiome est tout à fait caractéristique des classes ou types « monothétiques » , mais peut ne pas être systématiquement vérifié par les classes ou types « polythétiques ».

g) Règle et borne dites de « Solomon et Fortier »

Dans leur article datant de 1966, H. Solomon et J. Fortier (1966) définissent, pour tout indice de similarité $S(x,y)$ variant de 0 à 1, une règle, que l'on peut appliquer de façon générale et qui stipule que dès lors que la valeur d'un indice est supérieure à une borne

correspondant au « milieu » de son intervalle de variation, on considérera que x et y seront « plus semblables » que « dissemblables », en un mot on pourra parler de similarité (par rapport à cet indice) entre x et y. Cette règle s'écrit donc :

$$S(x, y) \geq \frac{1}{2}$$

On peut étendre cette règle au cas où l'intervalle de variation n'est plus (0-1), on écrira alors cette règle « étendue » sous la forme :

$$S(x, y) \geq \frac{\underset{x,y}{\text{Min}}\ S(x, y) + \underset{x,y}{\text{Max}}\ S(x, y)}{2}$$

La borne de Solomon-Fortier est intéressante au sens qu'elle permet d'étalonner un indice au milieu de son intervalle de variation (disons un point de passage obligé) et ainsi d'autoriser des comparaisons structurelles entre indices. Nous étalonnerons surtout, pour un indice donné, la valeur des « matchings » entre les profils de « x » et « y » en comparant cette valeur au nombre de variables dans le contexte spécifique et particulier de la « disjonction complète » (voir § 2.2.2.1) . C'est le processus complet dérivé de l'application de cette règle qui sera utilisé comme critère de valorisation d'un indice dans les paragraphes suivants.

1.2 Quelques Définitions de Base

Dans le suite du texte nous parlerons de « Descripteurs » ou d' « Attributs » ou de « propriétés » quand nous aurons affaire à des variables de « Présence - Absence », nous noterons : **J** cet ensemble, et nous poserons :

$P = |\mathbf{J}|$ le cardinal de cet ensemble J (indice d'un attribut = j).

D'autre part, nous parlerons de « Variables » ou de variables « vraies » au sens statistique du terme lorsque nous aurons affaire à des colonnes de tableaux de données, que ces variables soient nominales (catégorielles), hiérarchisées (notes, rang, fréquences etc..), ou continues, nous appellerons : **M** cet ensemble et :

$m = |\mathbf{M}|$ le cardinal de cet ensemble (indice d'une variable = k).

Enfin, et très classiquement cette fois, nous appellerons : **I**, l'ensemble des individus ou sujets étudiés, entre lesquels, on calculera des indices de similarité. Nous poserons, ce qui est un classique de l'analyse des données : **I** = {ensemble des individus} et :

$N = |\mathbf{I}|$ = le cardinal de cet ensemble (indice d'un individu = i).

2 Les Représentations possibles et les différences entre l'Espace IxJ, l'Espace IxM et l'Espace Relationnel : IxI.

2.1 Les Tableaux K=IxJ et T=IxM

Considérons maintenant les tableaux suivants : Tableau 1, à valeurs {0-1} **,** de dimensions (N,P) (croisant des individus notés $\{O_i\}$ avec des variables binaires ou des modalités $\{m_j\}$, et le Tableau 2, de dimensions (N,m), (croisant les mêmes individus $\{O_i\}$ avec des variables catégorielles ou hiérarchisées, ou numériques $\{V^k\}$), on a :

Tableau 1

	m_1	m	m_j	$m_{j'}$		m_P
O_1	1	0	1	0	0	1
O_2	0	1	0	1	0	1
O_3	1	0	1	0	1	0
O_i						
$O_{i'}$						
O_N	1	0	0	1	1	1

Tableau 2

	V_1	V_2	V_k	V_m
O_1	1	3	2	5
O_2	4	3	1	4
O_3	5	4	3	5
O_i	0	1	2	5
$O_{i'}$	5	3	1	1
O_N	4	4	1	2

Dans le cas du Tableau IxJ, simple, de dimensions (N,P), cas des données de « Présence – Absence », il apparaît que toutes les données sont 0 ou 1, (l'individu possède = 1, ou ne possède pas = 0, une propriété j (Attribut j)).
Dans le cas du Tableau T = IxM, de dimensions (N,m) - cas de variables vraies - les quantités données à l'intersection d'une ligne i et d'une colonne k peuvent prendre des valeurs quelconques. Dans le cas d'une variable catégorielle ou hiérarchisée, le nombre de valeurs (*finies et assez peu nombreuses*) en général prises par la variable V^k**,** donnera le nombre total de modalités (*les différentes valeurs possibles*) de cette variable.

Soit p_k ce nombre, p_k = (Ensemble des valeurs différentes de V^k) et l'on a :

$P = \sum_{k=1}^{m} p_k$, $k \in \{1,2.....m\}$, est le nombre total de <u>modalités</u> du Tableau, (à ne pas confondre avec « m », qui lui est le nombre total de variables).

En effet, si l'on prend par exemple l'individu N° 2 du tableau (N× m) précédent , on voit que le profil de i_2, est donné par:

$i_2 =$ | 4 | 3 | 1 | 4 |

la décomposition en profil disjonctif complet de l'individu i_2 du fait que V^1 a 5 modalités, V^2 a 4 modalités, V^3 a 3 modalités, enfin V^4 a 5 modalités s'écrit :

V^1					V^2				V^3			V^4				
0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	1	0

Sur ce tableau disjonctif on a bien m (Nombre de variables) = 4, $N = |I|$

P, la dimension la plus large = Nombre total de modalités = 5+ 4 +3 + 5 = 17, dans le cas présent, donc : $P = \sum_{k=1}^{m} p_k$, $k \in \{1,2.....4\}$; vaut ici: P = 17.

Donc, il y a équivalence au niveau de l'information apportée entre le tableau Nxm et le Tableau NxP, qu'on appellera **Tableau Disjonctif Complet** dans ce cas, et on le notera K, son terme général est donné par :

$$k_{ij} = \begin{cases} 1 \text{ si l'objet i possède la modalité j} \\ 0 \text{ sinon} \end{cases}$$

seuls le nombre de colonnes (P au lieu de m) et les valeurs dans les cases ont changé : {0,1} dans le tableau NxP, valeurs quelconques dans le tableau Nxm.

Posons maintenant: $\overline{p} = \frac{1}{m}\sum_{k=1}^{m} p_k$ alors $P = \overline{p}\, m$, si en particulier, toutes les variables ont le même nombre de modalités soit « μ » ce nombre, alors $P = \mu\, m$.

Remarque n°1: *on peut considérer que le tableau de «présence - absence» est un tableau disjonctif particulier lorsque l'on découpe une variable de «présence – absence » en deux variables disjonctives, l'une de « présence » et l'autre d' « absence » d'une propriété. On respecte dès lors les contraintes d'un tableau disjonctif complet avec comme changement majeur, le fait que l'on double le nombre de colonnes de la matrice de « présence absence » avec un nombre de variables initial égal à* m *d'où* P=2m *est le nombre des modalités . Nous verrons ultérieurement que, de ce fait, le modèle particulier de « présence – absence » disparaît au profit d'un modèle disjonctif particulier où toutes les variables ont deux modalités.*

2.2 Cas des Tableaux Relationnels C=IxI et B=JxJ

Une dernière notation très utile, et particulière, est à retenir et à rajouter à la liste précédente. Lorsqu'on a affaire à des données représentables sous forme vectorielle (cas des « m » variables) ou dans un cas plus complexe encore: le cas de données non représentables sous forme vectorielle, mais sous forme de graphe, il existe une autre représentation appelée, représentation relationnelle qui permet de déployer une Théorie qui lui est consacrée, à savoir : « l'Analyse Relationnelle ou l'Analyse des Représentations Relationnelles » . Une abondante littérature existe sur ce domaine que nous donnons en bibliographie (voir en particulier J. Ah-Pine (2007), H.Benhadda et F.Marcotorchino (1998) et (1996), F. Marcotorchino (1984), (1989), (1991), F.Marcotorchino et P.Michaud (1981), F. Marcotorchino et N. El Ayoubi (1991) en tant qu'échantillon de cette approche). Mais

pour la suite de l'exposé, il nous paraît important d'en donner quelques définitions basiques, (voir les explications ci-après) :

2.2.1 Tableau Relationnel de « Condorcet » C, croisant IxI de Taille (NxN)

Le modèle mathématique sur lequel est basée l'Analyse Relationnelle des Données est lié à une façon particulière de prendre en compte des variables, quel qu'en soit le type. En fait, il s'agit de prendre en compte les données individuelles sous forme de Relation Binaire (ceci permet de croiser des clients par rapport à une variable en Marketing , des textes en linguistique, des Patients en Médecine etc..). A titre d'illustration, considérons le cas suivant où est représentée une variable catégorielle (nominale) à savoir la « **nationalité** » appliquée à un petit ensemble de 5 personnes notées { A, B, C, D, E} ayant 3 Nationalités. La variable V "nationalité", est codée 1 pour les citoyens Français: {A et B}, 2 pour l'Espagnol {C}, 3 pour les Anglais : {D et E}. Dans ce cas, deux individus sont en relation s'ils ont la « même Nationalité ». C'est bien entendu une relation binaire, dont le graphe de représentation sous forme matricielle C est donné ci – après:

		A	**B**	**C**	**D**	**E**
1	**A**	1	1	0	0	0
1	**B**	1	1	0	0	0
2	**C**	0	0	1	0	0
3	**D**	0	0	0	1	1
3	**E**	0	0	0	1	1
V		**C**				

Remarque n°2: *Si nous avions changé le codage* 1 *en* 2 , 2 *en* 7 *et* 3 *en* 13 *(par exemple), il est évident que la matrice Relationnelle* C *n'aurait pas été affectée par ce changement, (en fait en cas de variables catégorielles, les codages n'ont pas de signification particulière, seule l'appartenance à des classes de valeurs de modalités différentes, compte). En fait la représentation de* **C** *sous forme matricielle s'obtient en posant la valeur* « 1 » *à l'intersection de la Ligne* A *et de la Colonne* B *si: l'« Objet* A *» a « la même modalité » que l' « Objet* B *» et* « 0 » *sinon. Ceci est fait sans tenir compte explicitement de la valeur de la catégorie mais seulement de façon implicite. En fait dans ce cas, nous sommes dans une situation où la Relation est une pure « **Relation d'Equivalence** ». Il existe une correspondance claire entre la représentation codée de* **V,** *et sa représentation associée* **C** *en termes de graphe ou de matrice* (NxN) *binaire, mais comme nous l'avons vu précédemment cette correspondance n'est pas biunivoque. Néanmoins, la matrice* **C** *représentant la variable* **V** *est une Matrice* {0-1}, *contenant la même information que celle apportée par la variable catégorielle* **V**. *En général, une Matrice Binaire contient au minimum, la même information que celle contenue dans sa*

variable nominale associée, codée sous forme « vecteur linéaire ». Mais bien plus, si nous considérons le cas, où, dans notre premier exemple, on suppose à titre illustratif que l'Individu C a la triple "Nationalité" : (multi-appartenance : Passeports Français, Espagnol et Anglais). Il est dès lors impossible de coder cette nouvelle information dans la variable **V**, *telle qu'elle est. Mais ceci ne pose aucun problème en environnement relationnel, voir la nouvelle représentation* **C'** *de* **C** *ci-après :*

		A	B	C	D	E
1	A	1	1	**1**	0	0
1	B	1	1	**1**	0	0
2	C	**1**	**1**	1	**1**	**1**
3	D	0	0	**1**	1	1
3	E	0	0	**1**	1	1
V		**C'=Nouveau C**				

On voit sur le tableau précédent, qu'il est impossible de tenir compte de la triple appartenance dans le codage de **V**, alors que dans le codage relationnel, il suffit de rajouter des « 1 » (*en gras sur la figure*) à l'intersection des lignes et des colonnes de l'individu C à l'intersection des lignes et des colonnes des autres individus de départ.
Dans le cas de cette configuration relationnelle, on peut créer des critères de classification sur l'ensemble des individus ou sur l'ensemble des variables (voir l'article sur Analyse Factorielle Relationnelle dans F.Marcotorchino (1989) et (1991)) ou des indices de comparaisons entre deux tableaux (matrices) relationnels, c'est-à-dire entre deux variables V et V', ou plus généralement entre deux matrices C et C'. Des résultats fondamentaux résultent de cette notation relationnelle .

2.2.2 Additivité des Matrices Relationnelles de type « Condorcéen »

Quelle que soit une matrice relationnelle C^k, représentant une relation binaire associée à une variable V^k quelconque, elle s'écrit de la façon suivante :
a) Si la variable, V^k, est une variable purement qualitative, on a le codage suivant :

$$C^k_{ii'} = \begin{cases} 1 & \text{si} \quad V^k(i) = V^k(i') \\ 0 & \text{sinon} \end{cases}$$

b) Dans le cas d'une variable quantitative le codage associé peut prendre plusieurs formes possibles dont une assez simple,(mais non optimale), décrite ci-dessous :

$$C_{ii'}^{k} = \begin{cases} 1 & \text{si} \quad \left| V^{k}(i) - V^{k}(i') \right| \leq s \\ 0 & \text{sin on} \end{cases}$$

Où « s » est un seuil donné.

Alors, posons :

(f 4) $$C_{ii'} = \sum_{k=1}^{m} C_{ii'}^{k}$$

où « m » est le nombre de variables de départ, et où le nombre total d'objets $\{O_i\}$, on l'a vu, est posé égal à « N » par la suite. (N pouvant atteindre des valeurs égales à plusieurs millions dans les problématiques réelles, CRM et Marketing bancaires par exemple)).
Cette matrice est la matrice dite de **« Condorcet »** de l'Analyse Relationnelle. La méthodologie relationnelle s'appuie sur la définition de cette matrice globale de similarités qui dans les cas les plus usuels de similarité s'exprime comme un produit scalaire (équivalent à un « noyau » (« Kernel de la théorie de l'apprentissage).:

(f 5) $$C_{ii'} = \sum_{j=1}^{P} k_{ij} k_{i'j}$$ [2]

(où k_{ij} représente le codage disjonctif (vu au § précédent) pour l'ensemble des modalités de l'individu O_i), c'est à dire dans le cas ou les valeurs {0 ou 1} sont obtenues par disjonction du Tableau 2, du précédent paragraphe.
Cette matrice traduit le degré de ressemblance entre deux objets i et i'.
A partir de cette matrice de ressemblance, on peut construire une matrice « duale » de celle de Condorcet dite matrice des « dissemblances » ou dissimilarités soit $\overline{C}_{ii'}$, le-terme général de cette matrice alors on a :

$$\overline{C}_{ii'} = m - C_{ii'} \quad \text{(si pas de données manquantes)}$$

$$\overline{C}_{ii'} + C_{ii'} \leq m \quad \text{(si données manquantes)}$$

Tout tableau de Condorcet de terme général $C_{ii'}$, vérifie (voir F.Marcotorchino et P.Michaud (1981)), la condition de « transitivité générale » :

(f 6) $$C_{ii'} + C_{i'i''} - C_{ii''} \leq m \quad \forall (i, i', i'')$$

Cette condition de transitivité générale provient du fait que le tableau $C_{ii'} = \sum_{k=1}^{m} C_{ii'}^{k}$ (voir formule (f 4)), est la somme de « m » relations d'Equivalence, qui vérifient chacune (par définition d'une relation d'équivalence) la condition de transitivité, c'est à dire :

[2] L'analyse relationnelle a été le cadre de développement de nombreuses mesures de similarités dites régularisées qui permettent de tenir compte de la structure interne des unités lexicales dans le calcul de similarité entre deux documents. Cette approche consiste à donner un poids, calculé de manière empirique, à chaque unité lexicale qui permet de mettre en avant son caractère discriminant. Nous renvoyons le lecteur intéressé aux travaux de H.Benhadda et F.Marcotorchino (1998) et H.Benhadda (1996).

(f 7) $$C^{k}_{ii'} + C^{k}_{i'i"} - C^{k}_{ii"} \leq 1 \quad \forall (i, i', i"), \forall k$$

qui s'interprète en disant que si « i » est en relation avec « i' » (soit $C^{k}_{ii'} = 1 \quad \forall (i, i'), \forall k$) et si « i' » est en relation avec « i″ » soit $C^{k}_{i'i"} = 1 \quad \forall (i', i"), \forall k$ alors on doit avoir « i » en relation avec « i″ » c'est à dire que $C^{k}_{ii"} = 1 \quad \forall (i, i"), \forall k$, c'est justement ce que traduit la formule (f 7)

En sommant pour toutes les « m » valeurs de « k » , l'inégalité (f 7), on trouve l'inégalité proposée dans la formule (f 6), la condition « duale » sur la matrice de Condorcet $\overline{C}_{ii'}$, est une condition d'**« inégalité triangulaire » :**

(f 8) $$\overline{C}_{ii"} \leq \overline{C}_{ii'} + \overline{C}_{i'i"}$$

Comme nous l'avons vu ces deux relations sont souvent « duales » l'une de l'autre. La relation d' « inégalité triangulaire », relation liée à des approches « métriques » de distances, correspond à la relation de « transitivité généralisée» pour les structures de « similarité » vue sous l'angle relationnel.

Dans le cas général , dire qu'il y a plus de variables pour lesquelles les objets « i et i' » sont en relation, que de variables pour lesquelles ils ne le sont pas, se traduit par la règle suivante :

$$\overline{C}_{ii'} \leq C_{ii'}$$

Quand il n'y a pas de données manquantes , du fait que $\overline{C}_{ii'} + C_{ii'} = m$ dans ce cas , il vient la condition dite de majorité par paires :

(f 9) $$C_{ii'} \geq \frac{m}{2} \Rightarrow \frac{C_{ii'}}{m} \geq \frac{1}{2}$$

(ceci est une représentation de la règle majoritaire des comparaisons par paires de Condorcet (1785), qui, dans le cas où l'on considère le tableau de Condorcet comme un indice de similarité est équivalente à une borne de Solomon Fortier, il suffit de diviser le tableau C par « m », pour voir ce fait immédiatement).

Cette définition est très générale car elle permet d'additionner tous types de relations binaires quelconques en allant bien au delà des structures vectorielles dont nous avons parlé ici. Ainsi la définition de la matrice de Condorcet (que nous avons donnée en (f 5)) est moins générale, elle ne s'applique seulement que dans le cas où l'on considère des variables linéaires, telle que la variable V définie précédemment (exemple des nationalités cas 1) ou en cas de variable représentable sous forme de vecteur colonne (linéaire).

2.2.3 Additivité longitudinale modalitaire

Cette additivité est valable dans le cas de variables linéaires transformables en variables disjonctives (*formule moins générale mais utilisée souvent en Analyse des Données*)

Dans ce cas, comme nous l'avons vu, si K représente la matrice de terme $\{k_{ij}\}$ représentative d'une variable V quelconque (appelée également matrice disjonctive de V), on a, si p_k représente le nombre de modalités de la variable V_k :

$k_{ij} = 1$ si i possède la modalité j de V

$k_{ij} = 0$ si i ne possède pas la modalité j de V

dans ce cas particulier , la matrice relationnelle représentative de V (numéro k) s'écrit :

$$C^k_{ii'} = \sum_{j=1}^{p_k} k_{ij} k_{i'j} \quad \text{et} \quad C_{ii'} = \sum_{k=1}^{m} C^k_{ii'} = \sum_{k=1}^{m} \sum_{j=1}^{p_k} k_{ij} k_{i'j} = \sum_{j=1}^{\sum_{k=1}^{m} p_k} k_{ij} k_{i'j} = \sum_{j=1}^{P} k_{ij} k_{i'j}$$

On retrouve ici la justification de la formule (f 5)

Les résultats obtenus dans le cas général restent valables, avec néanmoins pour la forme duale $\overline{C}^k_{ii'}$ de $C^k_{ii'} = \sum_{j=1}^{p_k} k_{ij} k_{i'j}$, une nouvelle forme donnée par :

$$\overline{C}^k_{ii'} = \frac{1}{2} \sum_{j=1}^{p_k} \left(k_{ij} - k_{i'j}\right)^2 ,$$

soit pour la sommation longitudinale :

(f 10)
$$\overline{C}_{ii'} = \sum_{k=1}^{m} \overline{C}^k_{ii'} = \frac{1}{2} \sum_{j=1}^{P} \left(k_{ij} - k_{i'j}\right)^2$$

On vérifie également les formules précédentes dans ce cas disjonctif :

$$\overline{C}^k_{ii'} + C^k_{ii'} = 1 \quad \forall \, (i, i') \qquad \text{et} \qquad \overline{C}_{ii'} + C_{ii'} = m \quad \forall \, (i, i')$$

En effet, comme dans le cas d'un tableau K (disjonctif complet) on a :

$$\sum_{j=1}^{P} k_{ij} = m \ , \ \forall \ i \ , \text{ il vient :}$$

$$C_{ii'} + \overline{C}_{ii'} = \sum_{j=1}^{P} k_{ij} k_{i'j} + \frac{1}{2} \sum_{j=1}^{P} \left(k_{ij} - k_{i'j}\right)^2 = \frac{1}{2} \left(\sum_{j=1}^{P} k^2_{ij} + \sum_{j=1}^{P} k^2_{i'j} \right) = \frac{1}{2} \left(\sum_{j=1}^{P} k_{ij} + \sum_{j=1}^{P} k_{i'j} \right) = m$$

Ce qui se traduit au niveau des tableaux de Condorcet associés par la définition littérale suivante:

« *Dans le cas où il n'y a pas de données manquantes, dire que le nombre de modalités (des variables) que i et i' partagent est plus grand que le nombre de modalités qu'ils ne partagent pas, se traduit par* [3] » :

$C_{ii'} \geq \overline{C}_{ii'}$ ce qui implique là encore que : $C_{ii'} \geq m/2$

[3] il est à noter ici que la phrase est différente de celle présentée au cas précédent, on parle ici de « modalités » et non de « variables »

2.2.4 Représentation sous forme Vectorielle de la Matrice C^k

Si l'on représente en **extension vectorielle** (vecteur de longueur N²), la matrice représentative de la variable **C**, et celle de la matrice **C'** (nouveau C), données dans l'exemple précédent (*nationalité*) , on obtient :

C	1	1	0	0	0	1	1	0	0	0	0	0	1	0	0	0	0	0	1	1	0	0	0	1	1
C'	1	1	**1**	0	0	1	1	**1**	0	0	**1**	**1**	1	**1**	**1**	0	0	**1**	1	1	0	0	**1**	1	1

On voit bien, que nous avons rajouté 8 valeurs « 1 » dans le vecteur représentatif de **C'**, les huit cases en « 1 » gras de la matrice (Nouveau **C**) présentée auparavant. Chacun des vecteurs associés à **C** et **C'**, se compose donc de 25 (N^2) éléments, en 5 blocs de 5 juxtaposés, chaque bloc représentant une ligne de la matrice C ou de la matrice C'. Ainsi sur les mêmes données on peut avoir 4 représentations différentes : à savoir: des tableaux N×P, des tableaux N×m , des tableaux N×N et des vecteurs N^2.

a) Bref aperçu sur la réécriture du Tableau Relationnel C^k sous forme vectorielle.

Pour toute matrice relationnelle C^k, on définira son **Extension Vectorielle** γ^k comme un vecteur de longueur $\mathbf{N^2}$ tel que :

$\vec{\gamma}^k = (\gamma_1^k, \gamma_2^k, \gamma_3^k, ...\gamma_s^k, ...\gamma_{N^2}^k)$, où si $C_{ii'}^k$ est le terme général de la matrice C^k, alors :

(f 11) $C_{ii'}^k = \gamma_s^k$, si et seulement si l'indice courant « s » vérifie:

$$s = (i-1)N + i' \qquad \forall i \text{ et } i'$$

(donc à tout couple (i,i') correspond une valeur de l'indice « s » et une seule).
Inversement connaissant « s » et bien sûr la dimension N, pour avoir la valeur de i et i', il suffit de procéder trivialement de la façon suivante :

a) On divise s par N $\Rightarrow$ s=a N+b,
b) si a=0 alors : i=1 et i'=b
c) si a $\neq$0 alors i= a+1 et i'=b

Exemple pour N=5 , que valent les indices i et i' de la matrice relationnelle C^k , pour s=18 ?
On divise 18 par 5 soit 18=3x5+3, on a donc i=4 et i'=3

b) Propriétés héritées au niveau vectoriel

Un bon nombre de propriétés relatives aux notations relationnelles se retrouvent, de façon quasi héréditaire, vérifiées au niveau des notations vectorielles, mais avec néanmoins quelques précautions à prendre, comme le montre l'induction de la propriété de transitivité.
En effet, pour toute variable C^k, relation d'équivalence, on a vu en (f 7) que l'on avait l'inégalité :

$$C_{ii'}^k + C_{i'i''}^k - C_{ii''}^k \leq 1 \quad \forall (i, i', i''), \forall k$$

En utilisant maintenant la convention d'extension vectorielle (f 11), cette inégalité précédente induit au niveau du vecteur $\vec{\gamma}^k = (\gamma_1^k, \gamma_2^k, \gamma_3^k, ...\gamma_s^k, ...\gamma_{N^2}^k)$ les contraintes suivantes :

(f 12) $$\gamma_s^k + \gamma_{s'}^k - \gamma_{s''}^k \leq 1 \quad \forall\ k$$

Cependant les indices s, s' et s'' , ne sont pas quelconques, car ils sont liés par les trois égalités suivantes, impliquées par les formules (f 11) et (f 7) :

$$\begin{aligned} s &= (i-1)\,N + i' \\ s' &= (i'-1)\,N + i'' \qquad \text{avec } i < i' < i'' \\ s'' &= (i-1)\,N + i'' \end{aligned}$$

<u>Propriété n°1</u> : Inégalité sur les indices représentant une relation d'équivalence

En éliminant deux à deux les indices ou quantités présentes dans 2 des égalités ci-dessus, et en tenant compte que si $N = |I|$ est donné, alors chacun des indices : i,,i', i'' tels que $i < i' < i''$ vérifie : $i \leq N$, $i' \leq N$, $i'' \leq N$, on peut montrer que les indices s, s' et s'' vérifient alors l'inégalité non triviale suivante :

(f 13) $$s'' \leq s + N - \frac{s'}{N}$$

2.2.5 Tableau Relationnel croisant JxJ de taille (PxP) ou « Matrice de Burt »

De la même façon que nous avons défini le tableau de Condorcet C du paragraphe précédent, nous pouvons définir , dans le cas où les modalités j de l'espace J sont les modalités de variables discrètes ou catégorielles une matrice dite « Matrice de Burt », notée B, qui se calcule également à partir des valeurs du tableau Disjonctif Complet K :

$$k_{ij} = \begin{cases} 1 \text{ si l'objet i possède la modalité } j \\ 0 \text{ sinon} \end{cases}$$

Ce tableau B a pour terme général la valeur $B_{jj'}$, donnée par :

(f 14) $$B_{jj'} = \sum_{i=1}^{N} k_{ij}\, k_{ij'}$$

Ce tableau très utilisé en Analyse Factorielle des Correspondances Multiples (voir G. Saporta (1990) , F.Caillez et JP.Pages (1976) ou F.Marcotorchino (1991)) a de nombreuses propriétés structurelles dont on donne ci dessous les principales :

$$B_{jj'} \leq \text{Min}\ (B_{jj}, B_{j'j'})$$

$$\sum_{j=1}^{P}\sum_{j'=1}^{P} B_{jj'} = \sum_{j=1}^{P}\sum_{j'=1}^{P}\sum_{i=1}^{N} k_{ij}k_{ij'} = N.m^2$$

Dans le cas ou les variables V^k de départ sont des variables catégorielles, il existe une « dualité » évidente entre les Tableaux de « Burt » et de « Condorcet » (voir F.Marcotorchino (1989), (1991)). En effet on a:

$C_{ii'} = \sum_{j=1}^{P} k_{ij} k_{i'j}$, $B_{jj'} = \sum_{i=1}^{N} k_{ij} k_{ij'}$, soit en notations matricielles : $C = K\,{}^tK$ et $B = {}^tK\,K$

Propriété n°2 : **Egalité des Normes de Frobenius des matrices de Condorcet et de Burt**

De ces relations précédentes on tire la relation « unificatrice » suivante :

$$\sum_{i=1}^{N}\sum_{i'=1}^{N} C_{ii'}^2 = \sum_{i=1}^{N}\sum_{i'=1}^{N}\left[\sum_{j=1}^{P} k_{ij}k_{i'j}\right]\left[\sum_{j'=1}^{P} k_{ij'}k_{i'j'}\right] = \sum_{j=1}^{P}\sum_{j'=1}^{P}\left[\sum_{i=1}^{N} k_{ij}k_{ij'}\right]\left[\sum_{i'=1}^{N} k_{i'j}k_{i'j'}\right] = \sum_{j=1}^{P}\sum_{j'=1}^{P} B_{jj'}^2$$

soit :

(f 15)
$$\sum_{i=1}^{N}\sum_{i'=1}^{N} C_{ii'}^2 = \sum_{j=1}^{P}\sum_{j'=1}^{P} B_{jj'}^2 \ ,$$

En d'autres termes la somme des carrés des termes généraux des deux tableaux sont égales. Ce qui, en termes matriciels, est équivalent à l'égalité du carré des normes de Frobenius associées :

$$\|\mathbf{B}\|_F^2 = \|\mathbf{C}\|_F^2$$

3 Les Quantités de Base du Calcul des Similarités

Quelle que soit la façon dont les données sont prises en compte ou mises en matrices adéquates : (NxP, Nxm, NxN), l'information de base, quand on travaille sur les similarités entre deux « objets » x et y, revient à calculer les 4 quantités suivantes : représentables dans un tableau de contingence (2x2) :

Si l'on note les deux vecteurs de représentation d'un objet **x** et d'un objet **y** :

$\vec{x} = (x_1, x_2, ..x_j, ...x_P.)$, j variant de 1 à P,

$\vec{y} = (y_1, y_2, ..y_j, ...y_P.)$, j variant de 1 à P

En posant « x_j » le fait que pour la modalité « j », **x** vaut 1, et $(1\text{-}x_j)$ le fait que **x** vaut 0 alors les quantités suivantes caractérisent l'ensemble de tous les cas possibles de combinaisons vectorielles associées:

1. Nombre de «matchings » de **x** sur **y** =>
$$11(x,y)=\sum_{j=1}^{P} x_j y_j$$

où 11(x,y)[4] représente les configurations où **x** et **y** valent 1 simultanément

2. Nombre de «non matchings » de **x** sur **y** =>
$$00(x,y)=\sum_{j=1}^{P}(1-x_j)(1-y_j)$$

où 00(x,y) représente les configurations ou **x** et **y** valent simultanément 0

3. Nombre d'erreurs en **y** de **x** sur **y** =>
$$10(x,y)=\sum_{j=1}^{P} x_j(1-y_j)$$

où 10(x,y) représente le nombre de configurations où **x** vaut 1 et **y** vaut 0

4. Nombre d'erreurs en **x** de **x** sur **y** =>
$$01(x,y)=\sum_{j=1}^{P}(1-x_j)y_j$$

où 01(x,y) représente le nombre de configurations où **x** vaut 1 et **y** vaut 0

5. De la même façon , on voit que :
$$11(x,x)=11(x,y)+10(x,y)=\sum_{j=1}^{P} x_j y_j+\sum_{j=1}^{P} x_j(1-y_j)=\sum_{j=1}^{P} x_j$$

cette somme représente le nombre de valeurs pour lesquelles **x** =1, on note par convention 11(x,x) cette quantité.

6. A l'identique, on a :
$$11(y,y)=11(x,y)+01(x,y)=\sum_{j=1}^{P} x_j y_j+\sum_{j=1}^{P}(1-x_j)y_j=\sum_{j=1}^{P} y_j$$

qui représente le nombre de valeurs pour lesquelles **y** =1, on note par convention 11(y,y) cette quantité.

Ceci étant représentable sous la forme du tableau de contingence, dit :« Tetrachorique » suivant :

	y = 1	**y = 0**
x = 1	11(x,y)	10(x,y)
x = 0	01(x,y)	00(x,y)

[4] Par la suite pour alléger les notations, on identifiera le vecteur $\vec{x}=(x_1,x_2,..x_j...x_p)$ à **x**, puis, au lieu de noter, $11(\vec{x},\vec{y})$, le nombre de matchings on se contentera de la forme 11(x,y), par extension on identifiera **x** et son profil vectoriel, sous la lettre générique « x ». Enfin pour lier les deux notations vues précédemment, on aurait pu écrire en identifiant x à « i » et « x_j » à « k_{ij} » :
$$11(x,\ y)=\sum_{j=1}^{P} k_{ij}k_{i'j}=\sum_{j=1}^{P} k_{xj}k_{yj}$$

4 Processus de Structuration des Indices

Dans la forme calculable des indices de similarité, le fait avéré est que toutes les quantités du tableau de contingence (Tetrachorique) précédent, ne jouent pas un rôle équivalent, et que suivant le poids accordé à l'une ou l'autre de ces quantités, on tombe sur des formes différentes et des constructions différentes d'indices. Nous insisterons de façon répétée sur la valeur et les propriétés de ces indices dans le cas de recodage sous forme disjonctive complète entre les profils d' individus $\{O_i\}$ mesurés sur des variables catégorielles. Les Groupe I et Groupe II dont il va être question dans les pages qui suivent se distinguent par le fait qu'ils font ou ne font pas jouer un rôle à la quantité 00(x,y).

4.1 Indices du Groupe I, donnant priorité aux structures 11(x,y) et ne faisant jouer aucun rôle à 00(x,y)

4.1.1 Indices du Groupe I, Type I (indices obtenus par ratios directs)

4.1.1.1 Indice de Dice –Czekanowski (1945-1913)

Cet indice, bien qu'introduit initialement par Czekanowski en 1913 dans la théorie des matrices de confusions, puis redécouvert en 1920 par H.A. Gleason (dans le domaine de la botanique), a surtout été étudié par L.R. Dice (1945) (utilisé par lui en botanique et en phylogénie, en tout cas c'est sous son nom qu'il est présenté aujourd'hui). Quoique ces auteurs l'aient introduit séparément, et pour des besoins justifiés dans leur discipline de spécialité respective, aucun d'eux n'a explicité et compris le rôle central et globalisant de cet indice dont on verra par la suite le caractère fondamental dans la structuration des indices de similarité, d'autre part nous montrerons que sous certaines conditions il est équivalent à la notion de mesure majoritaire de comparaisons (due à A. de Condorcet en 1785). Il s'exprime au travers de la formule :

(f 16)
$$S_d(x,y)=\frac{11(x,y)}{11(x,y)+\frac{1}{2}\left[10(x,y)+01(x,y)\right]}=\frac{2.11(x,y)}{2.11(x,y)+10(x,y)+01(x,y)}$$

En fonction des quantités introduites au § 3, il s'écrit également du fait que:
$11(x,x)=11(x,y)+10(x,y)$ et $11(y,y)=11(x,y)+01(x,y)$,
d'où $11(x,x)+11(y,y) = 2.11(x,y)+ 10(x,y)+ 01(x,y)$:

(f 16')
$$S_d(x,y)=\frac{2.11(x,y)}{11(x,x)+11(y,y)}=\frac{2\sum_{j=1}^{P} x_j y_j}{\sum_{j=1}^{P} x_j+\sum_{j=1}^{P} y_j}$$

De ce fait , la quantité $d_d(x,y)=1-S_d(x,y)$ s'écrit:
$$d_d(x,y) = 1 - S_d(x,y) = \frac{\sum_{j=1}^{P} (x_j - y_j)^2}{\sum_{j=1}^{P} x_j+\sum_{j=1}^{P} y_j}$$

on voit que dans le cas où : $\forall$ x, 11(x,x)= 11(y,y)= Constante =μ, l'indice de Dice s'écrit:

(f 16") $$S_d(x, y) = 1 - \frac{\sum_{j=1}^{P} (x_j - y_j)^2}{2.\mu} = 1 - \frac{1}{2}\sum_{j=1}^{P}\left[\frac{x_j}{\sqrt{\mu}} - \frac{y_j}{\sqrt{\mu}}\right]^2 = 1 - \frac{1}{2}\delta^2(x, y)$$

Il apparaît clairement que dans cette configuration où la somme en ligne des "1" de tous les vecteurs x,y,z... est une constante, l'indice de Dice s'exprime bien sous la forme $S=1-(½)\ \delta^2$, où δ^2 est le carré d'une distance euclidienne, d'après le Théorème de Schoenberg-Gower (voir axiome n°4, § 1.1), $S_d(x,y)$ est donc un indice métrisable euclidien.

4.1.1.2 Indice de Jaccard (1908)

Cet indice, l'un des tous premiers indices à avoir été décrit dans la littérature scientifique par Paul Jaccard (1908) (ressortissant suisse du canton de Vaud, spécialiste de la phylogénie des plantes), est une variante du précédent, bien que découvert auparavant, il s'écrit :

(f 17) $$S_j(x,y) = \frac{11(x,y)}{11(x,y) + [10(x,y) + 01(x,y)]}$$

Il s'écrit aussi sous une forme très connue des spécialistes du traitement du langage naturel:

(f 17') $$S_j(x,y) = \frac{11(x,y)}{11(x,x) + 11(y,y) - 11(x,y)}$$

En effet, comme nous l'avons vu au § 3: 11(x,x)=11(x,y)+10(x,y) et 11(y,y)=11(x,y)+01(x,y), d'où 11(x,x)+11(y,y)=2.11(x,y)+10(x,y)+01(x,y), d'où l'obligation de soustraire 11(x,y) au dénominateur.

Si nous explicitons cet indice grâce aux expressions développées du §.3. nous obtenons:

$$S_j(x, y) = \frac{11(x, y)}{11(x, x) + 11(y, y) - 11(x, y)} = \frac{\sum_{j=1}^{P} x_j y_j}{\sum_{j=1}^{P} x_j + \sum_{j=1}^{P} y_j - \sum_{j=1}^{P} x_j y_j}$$

et l'indice de distance associé s'écrit:

$$d_j(x, y) = 1 - S_j(x, y) = \frac{\sum_{j=1}^{P} (x_j - y_j)^2}{\sum_{j=1}^{P} x_j + \sum_{j=1}^{P} y_j - \sum_{j=1}^{P} x_j y_j}$$

4.1.1.3 Indice d' Anderberg (1961)

Cet indice est l'un des plus récents de la famille des indices de Type I, puisqu'il a été introduit d'abord en 1958 par R.R. Sokal et P.H.A. Sneath, (à qui nous attribuerons un indice du Groupe II (voir § ultérieurs), puis redécouvert ensuite par M. R. Anderberg en 1961, voir Anderberg (1961). Il a été introduit beaucoup plus tard que les deux précédents (lesquels étaient plus intuitifs et correspondaient à des règles de similarité plus logiques et plus simples), ce n'est, lui aussi, qu'une variante de celui de Dice, lorsque l'on donne plus de poids aux situations de non concordance.

Il s'écrit :

(f 18) $$S_{an}(x,y) = \frac{11(x,y)}{11(x,y) + 2\left[10(x,y) + 01(x,y)\right]}$$

4.1.1.4 Indice de Sørensen (1960)

Cet indice a été introduit par Thorvald Sørensen, ce dernier, souvent cité comme co-inventeur en (1948) avec Dice de l'indice qui porte son nom est également l'auteur en 1960 d'une variante, donnée ci dessous, variante des indices de Type I, avec une définition très voisine de celle de l'indice de Dice, mais en donnant moins de poids aux situations de « non concordance ». Il s'écrit :

(f19) $$S_{so}(x,y) = \frac{11(x,y)}{11(x,y) + \frac{1}{4}\left[10(x,y) + 01(x,y)\right]}$$

cet indice semble un peu " tiré par les cheveux" à première vue, mais réécrit comme celui de Jaccard en fonction des quantités 11(x,x) et 11(y,y) il gagne en signification en effet on a :

(f 19') $$S_{so}(x,y) = \frac{4.11(x,y)}{\left[11(x,x) + 11(y,y)\right] + 2.\left[11(x,y)\right]}$$

Tous ces indices sont des variantes par rapport au poids donnés aux structures (10(x,y)+01(x,y)), en effet : Dice = ½, Jaccard = 1, Anderberg = 2 ; Sørensen = ¼ , il manque par symétrie à celui de Dice, un indice dont les le poids serait de 1/8 pour la configuration globale de non concordances : Nous appellerons cet indice, l'indice d' « Anderberg Complémentaire » S_{ac}, il sera défini par :

4.1.1.5 Indice d' « Anderberg Complémentaire »

Cet indice est introduit « artificiellement » ici, comme indice complémentaire et symétrique de celui qu' avait proposé en 1961, M. Anderberg (1961), , d'où le nom que nous lui donnons, pour des raisons d'équilibrage.

(f20) $$S_{ac}(x,y) = \frac{11(x,y)}{11(x,y) + \frac{1}{8}\left[10(x,y) + 01(x,y)\right]}$$

4.1.2 Unification Homographique des Indices de Type I

L'ensemble des différents indices connus, appartenant au Groupe I :Type I, peut se définir à partir de fonctions homographiques de l'indice de Dice (c'est cette propriété que nous avions déjà décrite dans F. Marcotorchino , P. Michaud (1981), étudiée et reprise dans l'article fort structuré de S Joly et G. Le Calvé (1994), publié chez Springer-Verlag dans le livre complet sur la « Dissimilarity Analysis », édité par B. Van Cutsem (1994).
En effet en exprimant la quantité (10(x,y)+01(x,y)) par rapport à 11(x,y) et à l'indice de Dice on trouve :

$$[10(x,y)+01(x,y)] = \frac{2.11(x,y)[1-S_d(x,y))]}{S_d(x,y)}$$

En remplaçant la quantité de « non concordance », précédente par son expression en fonction de $S_d(x,y)$ dans les formules relatives à chaque indice, la quantité 11(x,y) disparaît et l'on obtient des expressions ne dépendant que de $S_d(x,y)$:

(f 21) $S_j(x,y) = \frac{S_d(x,y)}{2-S_d(x,y)}$ et réciproquement $S_d(x,y) = \frac{2.S_j(x,y)}{1+S_j(x,y)}$

De la même façon on a pour Anderberg:

(f 22) $S_{an}(x,y) = \frac{S_d(x,y)}{4-3.S_d(x,y)}$ et réciproquement $S_d(x,y) = \frac{4.S_{an}(x,y)}{1+3.S_{an}(x,y)}$

De même, il vient pour l'indice de Sørensen :

(f 23) $S_{so}(x,y) = \frac{2.S_d(x,y)}{S_d(x,y)+1}$ et réciproquement $S_d(x,y) = \frac{S_{so}(x,y)}{2-S_{so}(x,y)}$

Enfin pour l'indice d'Anderberg « Complémentaire », on obtient:

(f 24) $S_{ac}(x,y) = \frac{4.S_d(x,y)}{3.S_d(x,y)+1}$ et réciproquement $S_d(x,y) = \frac{S_{ac}(x,y)}{4-3.S_{ac}(x,y)}$

Du fait que ces indices du Groupe I, Type I, soient tous fonction homographique de l'indice de Dice, nous permet de les représenter sur un diagramme unique, où l'on peut voir de façon simple comment ils se comportent et comment ils varient les uns par rapport aux autres. Par ailleurs bien évidemment tous ces indices varient de 0 à 1, ce que nous constatons sur le graphique :

$$0 \le S_\alpha(x,y) \le 1$$

- Ces indices valent 0 si le nombre de concordances (« matchings ») entre **x** et **y** est égal à 0 ⇔ 11(x,y)=0
- Ils valent 1 si la quantité 10(x,y)+01(x,y) = 0 ⇔ **x** et **y** ont le même profil de 1 et de 0
- Ils **ne sont pas définis** (division par zéro) si 11(x,y)=10(x,y)=01(x,y) =0, c'est à dire si 00(x,y) = P . On peut par extension proposer des valeurs d'extension aux différents indices cités, dans ce cas, pour éviter ces situations d'indétermination ou d'ambiguïté. Les situations où le profil de deux individus « i » et « i' » n'ont aucune valeur « 1 » présentes dans leur profil sont exceptionnelles. C'est à cette

problématique rare mais néanmoins possible que Mattijs J. Warrens (2008) s'attaque dans un article très récent Mars 2008, de la Revue « Journal of Classification ».

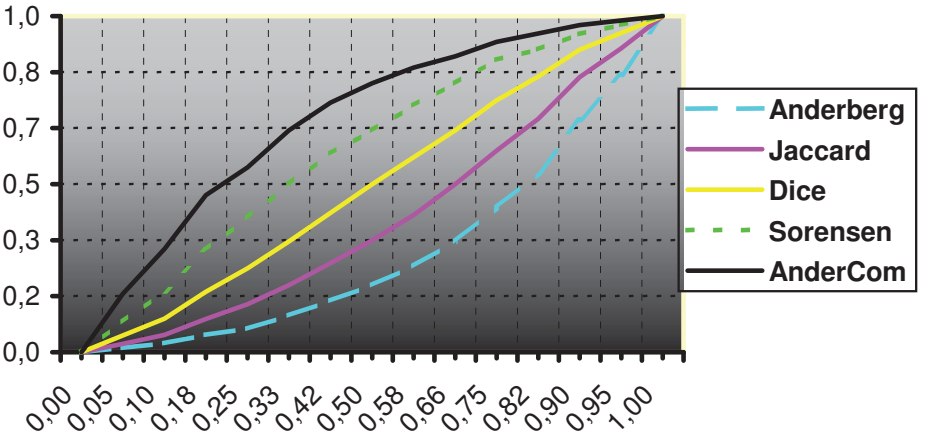


Figure 1: ***Graphique de variation des indices du Groupe I, Type I***

On constate sur ce graphique que la courbe de l'indice de Dice est la bissectrice du graphique, l'indice d'Anderberg est le symétrique, par rapport à la bissectrice (indice de Dice), de l'indice d' « Anderberg Complémentaire », l'indice de Sørensen est le symétrique de l'indice de Jaccard par rapport à cette même bissectrice.
D'autre part, dans le cas où les données de similarités sont calculées sur des tableaux disjonctifs complets (c'est ce qui se produit en AFCM (Analyse factorielle des correspondances multiples), ou en AFR (Analyse Factorielle Relationnelle)), on a de plus, les propriétés supplémentaires suivantes:

Propriété 3 : Relation avec la Valeur Disjonctive de la Distribution 11(x,y)

Dans le cas où les données (0-1) sont issues de tableaux disjonctifs complets ou de tableaux de « présence-absence » dédoublés, voir remarque du § 2-2, on a la propriété suivante :

(i) $10(x,y) = 01(x,y)$
(ii) $(10(x,y) + 01(x,y)) = 2m - 2.11(x,y)$ (cas des variables disjonctives)
(ii') d'où compte tenu de (i) $10(x,y)=01(x,y)=m-11(x,y)$
(iii) $(10(x,y) + 01(x,y)) = P - 2.11(x,y)$ (cas des variables « présence - absence » dédoublées) (avec $2m= P$)

La démonstration est très simple, en effet, on sait que sur un tableau disjonctif complet le nombre de 1 par ligne est une constante, il est égal à « m », nombre de variables initiales. Dès lors à toute présence de 1 dans le vecteur de **x**, correspond un 1 ou un 0 dans le profil de **y** .

- Si à 1 de **x** correspond un 1 de **y**, cette concordance est comptée dans 11 (x,y)

- Si à 1 de **x** correspond un 0 de **y**, cette concordance est comptée dans 10 (x,y)
- Si à 0 de **x** correspond un 1 de **y**, cette concordance est comptée dans 01 (x,y)

A toute « non concordance » de **x** sur **y**, correspond par symétrie une « non concordance de **y** sur **x**. Du fait que la somme des « 1 » pour chaque individu est une constante **m**, le nombre de non concordances, soit (10(x,y)+01(x,y)) est compté deux fois dans la somme totale m, d'où le résultat :

(iv) $$11(x,y)+\frac{1}{2}\left[\,10\,(x,y)+01\,(x,y)\,\right]=m$$

d'où évidemment les formules (i) et (ii)
De ce fait l'indice de Dice se simplifie en :

(f 25) $$S_d(x,y)=\frac{11(x,y)}{m}$$

En utilisant les formules (i) et (iii), il vient pour les autres indices après simplification et division du numérateur et dénominateur par m et application de (iii) :

(f 26) $$S_j(x,y)=\frac{11(x,y)}{2.m-11(x,y)}$$

de la même façon on a:

(f 27) $$S_{an}(x,y)=\frac{11(x,y)}{4.m-3.11(x,y)}$$

De même, il vient pour les indices de Sørensen et d' « Anderberg Complémentaire »:

(f 28) $$S_{so}(x,y)=\frac{2.11(x,y)}{m+11(x,y)}$$

(f 29) $$S_{ac}(x,y)=\frac{4.11(x,y)}{m+3.11(x,y)}$$

<u>Propriété 4</u> : Comparaison des Indices précédents par Rapport à la Règle Majoritaire de Condorcet

Si nous utilisons la formule donnant la définition d'un tableau de Condorcet, dans le cas de la somme de la représentation de variables « linéaires » (approche disjonctive), on constate que la règle majoritaire, donnée par : $C_{ii'} \geq \frac{m}{2}$ est équivalente à : $S_d(i,i') \geq \frac{1}{2}$

En effet puisque, d'après la formule (f 25) $S_d(i,i') = \frac{11(i,i')}{m}$ en remplaçant x par i , et y par i' car 11(i,i') n'est rien d'autre que :

$$C_{ii'} = \sum_{j=1}^{P} k_{ij} k_{i'j} \quad \text{(vue précédemment (f 5))}$$

si l'on remplace dans les notations du §2.2, x_j par k_{ij} et y_j par $k_{i'j}$, dès lors : $S_d(i,i') = \frac{C_{ii'}}{m}$ (cqfd).

Remarque n°3 : **Application directe de la Règle de « Solomon-Fortier » (définie au §1.1-5)** : *Au niveau indiciel si l'on applique la règle de Solomon et Fortier ,(voir* Solomon et Fortier (1966)), *de façon formelle, cette règle stipule simplement que : quelque soit l'indice de similarité* S(x,y) *ou* S(i,i') *considéré, on dira que deux objets* « x *et* y » *ou* « i *et* i' » *sont en relation de « similarité au sens de Solomon - Fortier » si* : $S(i,i') \geq \frac{1}{2}$, *ou plus généralement si un indice* S *varie de* « a » *à* « b » $a \leq S(i,i') \leq b$, *la règle dit que* i *et* i' *seront en relation de « similarité » si* : $S(i,i') \geq \frac{a+b}{2}$, *où* $\frac{a+b}{2}$ *est le « milieu » de l'intervalle de variation. La borne à ½ précédente est due au fait que* a = 0 *et* b=1, *puisque le milieu de l'intervalle de variation correspondant est d'évidence égal à ½.*
Ainsi si :

$$S_d(i,i') \geq \frac{1}{2} \text{ alors } C_{ii'} \geq \frac{m}{2} \quad \text{(règle de la majorité usuelle[5], ou Condorcéenne)}$$

Ceci s'interprète en disant que pour que deux objets « i et i' » soient considérés comme semblables au sens de la règle de Solomon et Fortier, il faut qu'ils soient semblables pour une majorité de variables.

[5] Il est d'ailleurs intéressant à ce propos de voir que l'Indice de Similarité de Gower (que ce dernier a introduit en 1971 ,J. C. Gower (1971)) qui s'écrit : $S_G(x,y) = \frac{1}{\sum_k w_k} \sum_k w_k S_{xy}^k$ est une variante pondérée de l'approche Condorcéenne.

De même on aura pour l'indice de Jaccard:

$S_j(i,i') \geq \frac{1}{2}$ implique => $C_{ii'} \geq \frac{2.m}{3}$ (règle de la majorité aux « deux tiers »)

Pour l'indice d' Anderberg:

$S_{an}(i,i') \geq \frac{1}{2}$ implique => $C_{ii'} \geq \frac{4.m}{5}$ (règle de la majorité aux « quatre cinquièmes »)

Pour l'indice de Sørensen:

$S_{so}(i,i') \geq \frac{1}{2}$ implique => $C_{ii'} \geq \frac{m}{3}$ (règle de la majorité au « tiers »)

Pour l'indice de Andersen Complémentaire

$S_{ac}(i,i') \geq \frac{1}{2}$ implique => $C_{ii'} \geq \frac{m}{5}$ (règle à la majorité au « cinquième »)

Si nous interprétons ce résultat sur une échelle « peu sévère » <==> « sévère », par rapport à la règle de Solomon et Fortier, il vient :

Peu Sévère ← ==========**Medium**==========➔**Sévère**
Ander Comp….. ***Sørensen*** …… …...***Dice***..... .. ***Jaccard***…… *Anderberg*

Pour le lecteur intéressé, on trouvera dans S. Joly et G. Le Calvé (1994), dans l'article de B.G. Baulieu (1989), ainsi que dans le livre de I. C. Lerman (1981), une présentation des propriétés de métricité, de monotonie ou d'autres encore que vérifient ces indices. Notre propos étant ici plus de structurer et classifier l'ensemble des indices que de lister leurs propriétés, nous renvoyons le lecteur à ces auteurs et à leurs articles pour de plus amples informations .

4.1.3 Un cadre unificateur pour les indices du Groupe I-Type I : le Modèle de Tversky

Une approche unificatrice a été proposée par Amos Tversky dans son article de fond A. Tversky (1977), ayant pour conséquence une formulation un peu plus complexe que celles des indices présentés précédemment, en particulier il permet dans sa formulation une influence non systématiquement équilibrée des quantités 10(x,y) et 01(x,y).

L'indice généralisé de Tversky s'écrit :

(f 30) $$S_{Ty}(x,y) = \frac{11(x,\ y)}{11(x,\ y) + \alpha\, 10(x,\ y) + \beta\, 01(x,\ y)}$$

avec : α et $\beta \geq 0$

On a bien : $0 \leq S_{Ty}(x,y) \leq 1$

Cet indice varie de 0 à 1 est est donc un indice « normé », en effet il vaut 1 si 10(x,y)=01(x,y)=0 et il vaut 0 si 11(x,y)=0. Il n'est pas défini si 11(x,y)=10(x,y)=01(x,y)=0 (voir la remarque citée page 23 sur les travaux de Mattijs Warrens(W2008))

ll semble que la non symétrie de l'influence des quantités 01(x,y) et 10(x,y) trouve un domaine particulièrement intéressant d'applications en recherche de « similarité moléculaire » en pharmacodynamique, et peut-être dans les essais cliniques, voir à ce propos le rapport de John Bradshaw de la Société Pharmaceutique 'Glaxo Wellcome', J. Bradshaw (1997). En effet la question à poser peut prendre deux formes différentes, nous citons ici son rapport :

"*There are two forms of question which we can ask,*

1. *Assess the degree to which object* **x** *and object* **y** *are similar to each other.*
2. *Assess the degree to which object* **y** *is similar to object* **x***".*

À la question de type 1., correspond un équilibre de l'influence , de ce fait : $\alpha = \beta$, à la question 2. peut correspondre un déséquilibre de **x** vers **y** d'où : $\alpha \neq \beta$, voici d'ailleurs le texte complet de J. Bradshaw : "*In this case the query is directional and we are more interested in the features in molecule* **x** *than we are in the features unique to molecule* **y,** α *and* β *, do not need to be equal. Molecule* **x** *may be regarded as the prototype and molecule* **y** *as the variant".* Cependant comme il est d'ailleurs remarqué dans le texte de J. Bradshaw, plus expérimental que théorique, on se retrouve facilement, même dans ce contexte élargi, à utiliser l'indice de Dice ou celui de Jaccard, sans oublier bien entendu les indices extrêmement connus que l'on découvrira, présentés et étudiés en détail au § 4.1.3. et qui sont dissymétriques par essence, à savoir : les indices de « Précision P » et de « Rappel :R », donnés par :

$$P = \frac{11(x,y)}{11(x,y)+10(x,y)} \quad \text{et} \quad R = \frac{11(x,y)}{11(x,y)+01(x,y)}$$

<u>Propriété n°5</u> : Equivalence de l' « Indice de Tversky » et de l' « Indice de Van Rijsbergen »
Ace propos, il est intéressant de noter que dans le cas particulier où $\alpha + \beta = 1$ dans la formule (f 30), un indice cité dans l'article de C. Michel (2000) et connu sous le nom d' « Indice de Van Rijsbergen », introduit par ce dernier dans un livre faisant référence chez les spécia-

listes de l'analyse documentaire Van Rijsbergen (1979) et qui s'exprime par rapport aux indices de Rappels et de Précision sous la forme :

(f 30') $$S_{vr}(x, y) = \frac{P.R}{\alpha P + \beta R} \quad \text{avec} \quad \alpha + \beta = 1$$

est <u>totalement équivalent</u> en remplaçant P et R par leur valeurs données précédemment, à l'indice de Tversky, puisque après simplifications on a :

$$S_{vr}(x, y) = \frac{P.R}{\alpha P + \beta R} = \frac{11(x, y)}{(\alpha + \beta)\,11(x, y) + \alpha\,10(x, y) + \beta\,01(x, y)} \quad \text{avec} \quad \alpha + \beta = 1$$

et comme $\alpha+\beta=1$, on retrouve[6] bien l'expression de l'Indice de Tversky. Conclusion si l'initiative de Tversky est intéressante pour avoir permis une justification de la dissymétrie d'influence, on connaît peu d'indices dissymétriques ayant eu une utilisation généraliste suffisante pour figurer comme des indices standard. En effet peu (voire pas) d'indices de type I (modèle de Tversky) avec $\alpha \neq \beta$, autres que ceux qui ont été créés pour des applications particulières et spécifiques ont obtenu un statut d'indices « connus » généralisables dans la littérature scientifique.

Le cas des indices $S_{Max}(x,y)$ (Indice de **Simpson** (1943)) et $S_{Min}(x,y)$ (Indice de **Braun-Blanquet** (1932)) rentrent également comme cas particuliers du modèle de Tversky, au sens que, par exemple, l'indice de Simpson qui s'écrit :

$$S_{Simpson}(x, y) = Max(P, R) = \frac{11(x, y)}{Min(11(x, y) + 10(x, y), 11(x, y) + 01(x, y))} = \frac{11(x, y)}{11(x, y) + Min(10(x, y), 01(x, y))}$$

est équivalent à l'Indice de **Précision** si $10(x,y) \leq 01(x,y)$
est équivalent à l'indice de **Rappel** si $01(x,y) \leq 10(x,y)$

On peut tabuler par rapport aux différentes valeurs de α et β , les indices qui relèvent de l'approche « Tverskienne », on obtient :

α \ β	**0**	**1/2**	**1**	**2**
0	~~**Pas de sens**~~	Comparaison à **y**	Indice de **« Rappel »**	?
1/2	Comparaison à **x**	**Dice**	On prédit* plus **y** que **x**	?
1	Indice[7] de **« Précision »**	On prédit plus **x** que **y**	**Jaccard**	

[6] On remarque une fois encore ici, que des indices ont été redécouverts par des auteurs différents au fil du temps, sans qu'aucune tentative de rapprochements avec des formalismes pré existants n'ait été proposée.
[7] Les indices de Rappel et de Précision sont définis explicitement et en extension au § 4-2.

(*) exemple d'un tel indice : le ratio suivant :

$$S(x, y) = \frac{11(x, y)}{11(x, y) + \frac{1}{2}[10(x, y) + 01(x, y)] + \frac{1}{2}01(x, y)}$$

dont le dénominateur est voisin de celui de Dice, mais pondère davantage que dans Dice les situations d'erreurs 01(x,y) . En cas de disjonction complète, cet indice particulier de la famille Tversky vaut:

$$S(x,y) = \frac{2.11(x,y)}{3.m - 11(x,y)}$$

Ce qui montre qu'il est bien intermédiaire entre l'indice de Jaccard et celui de Rappel, par ailleurs sa borne de Solomon-Fortier donne une majorité au 3/5 ème:

$$S(x,y) = \frac{2.11(x,y)}{3.m - 11(x,y)} \geq \frac{1}{2} \quad \Rightarrow \quad 11(x,y) \geq \frac{3}{5}m$$

4.1.4 Propriétés des Indices du Type I dans les cas de Disjonction des Variables

Il est important de comprendre à ce moment précis du discours que la disjonction complète permet des simplifications considérables des propriétés des indices. En effet dans un souci de généralité, rappelons que le seul cas qui n'entre pas dans la catégorie de situations à « disjonction complète », est le cas de données de « présence-absence » pures. Cependant, nous avons vu que, dans ce cas, par un artifice logique qui consiste à dédoubler chaque paramètre de description par son inverse (où à 1 pour le dit paramètre correspond 0 de l'inverse, et vice versa), on se ramène à une situation de disjonction complète, avec P nombre total de descripteurs, donné par P=2m (où m est le nombre de descripteurs initiaux et P le nombre nouveau de descripteurs après l'action de dédoublage. De ce fait le recours aux propriétés engendrées par la structure de disjonction complète devient un passage obligé qui prend toute son importance.
Ainsi les propriétés que nous allons proposer sont-elles quasi générales .

4.1.4.1 Liste de propriétés des Indices du Type I dans les Cas de Disjonction des Variables

Dans le cas particulier de tableaux disjonctifs complets, l'indice de Dice (du fait de son entière similitude à la constante « m » près, avec le tableau de Condorcet) vérifie la propriété de « transitivité généralisée» suivante (voir § 1.1 sur l'axiomatique) :

Propriété n°6 : **L'Indice de Dice vérifie la propriété de « Transitivité Généralisée Indicielle »**

(f 31) $$S_d(x, y) + S_d(y, z) - S_d(x, z) \leq 1 \quad \forall\ x, y, z$$

Il suffit pour cela de diviser la formule (f 6) relative au tableau de Condorcet par «m» .
Par ailleurs, nous avions vu (voir formule (f 16″) que lorsque le nombre de « 1 » du profil de **x,** $\forall$ x , était une constante, alors:

$S_d(x,y) = 1 - 1/2d^2(x,y)$ (où $d^2(x,y)$ est une métrique euclidienne)

Comme dans le cas de la disjonction complète, nous avons 11(x,x)=m, $\forall$ x, nous sommes tout à fait dans un cas d'application de cette formule (f 16″). Nous venons de montrer en

utilisant le recours à l'écriture Condorcéenne une autre façon de voir que l'indice de Dice était métrisable car il vérifie la TGI (Transitivité Généralisée Indicielle)

Propriété n°7 : **L'Indice de Jaccard vérifie la propriété de « Transitivité Généralisée Indicielle »**

De la même façon en remplaçant la valeur de l'indice de Dice par son expression homographique en fonction des autres indices, il vient par exemple la relation suivante pour l'indice de Jaccard :

(f 32) $$S_j(x,y) + S_j(y,z) - S_j(x,z) \le 1 - \frac{1}{4}\left[1 - S_j(x,y)\right]\left[1 - S_j(y,z)\right]\left[S_j(x,z) + 3\right] \quad \forall x, y, z$$

En fait l'indice S_j vérifie bien la « transitivité généralisée indicielle » (donc est métrisable) car « 1 » est bien une borne du premier membre de (f 32) , en effet la quantité

¼ρ (1- S_j (x,y))(1- S_j(y,z))(S_j(x, z)+3) est toujours positive du fait que $S_j(x,y) \le 1$ ∀ « x et y » .

En fait pour démontrer ce résultat de façon directe et très facilement, il suffisait d'utiliser le Théorème n°1 qui stipule que si un indice $S_u(x,y)$ s'exprime sous la forme :

$$S_u(x,y) = \frac{S_d(x,y)}{\beta[1 - S_d(x,y)] + 1}$$

par rapport à un indice métrisable $S_d(x,y)$, alors il est lui même « métrisable » .

En fait l'indice de Dice étant métrisable, comme l'indice de Jaccard s'écrit (voir formule (f 21)) $S_j(x,y) = \frac{S_d(x,y)}{2 - S_d(x,y)}$, il vérifie les conditions d'applications du Théorème n°1, avec **β=1** dans ce cas, (cqfd) .

Donc $S_j(x,y)$ est métrisable.

Quoiqu'on puisse se passer d'utiliser l'inégalité (f 32) pour obtenir ce résultat, l'inégalité (f 32) nous permet néanmoins de déduire une relation très surprenante sur l'indice de distance $d_j(x,y)=1-S_j(x,y)$ associé, en effet en remplaçant : $S_j(x,y)$ par $1-d_j(x,y)$, il vient :

(f32') $$d_j(x,z) \le [d_j(x,y) + d_j(y,z)] - d_j(x,y).d_j(y,z) + \frac{1}{4} d_j(x,y).d_j(y,z).d_j(x,z) \qquad \forall x, y, z$$

soit encore:

(f 33) $$d_j(x,z) \le \frac{[d_j(x,y) + d_j(y,z)] - d_j(x,y).d_j(y,z)}{1 - \frac{1}{4} d_j(x,y) d_j(y,z)} \qquad \forall x, y, z$$

Sous cette forme, on voit que le deuxième membre de cette inégalité est indépendant de $d_j(x,z)$. En effet comme chacune des valeurs $d(x,y)$ et $d(x,z) \le 1$, de fait la division par $(1-1/4dj(x,y)dj(y,z))$ qui est une valeur > 0, donc différente de 0, est possible. Ceci permet d'avoir une borne meilleure que celle de l'inégalité triangulaire usuelle.

La soustraction par une quantité positive permet dans les formules (f 32), (f32') et (f33) d'avoir une borne plus fine permettant souvent l'obtention d'une égalité comme le montre le petit exemple suivant :

x	1	0	0	0	1	0	1	0	0	0	1
y	1	0	0	0	0	1	1	0	0	0	1
z	0	1	0	0	0	1	0	1	0	0	1

Ici on a affaire à 4 variables disjonctivées (ayant respectivement : 3; 3; 3; et 2 modalités), d'où m=4
L'indice de Jaccard entre les différents individus vaut :
$S_j(x,y) = 3/5$; $S_j(y,z) = 2/6$; $S_j(x,z) = 1/7$
Le premier membre de l'inégalité vaut : 3/5 +2/6 -1/7 = 166/210
Le deuxième membre de l'inégalité vaut 1 – ¼(2/5.4/6. 22/7)= 1-44/210= 166/210
De même au niveau de la formule (f 29″) il vient :

$$\frac{6}{7} \le \frac{\frac{2}{5}+\frac{4}{6}-\frac{2}{5}\times\frac{4}{6}}{1-\frac{1}{4}\times\frac{2}{5}\times\frac{4}{6}} = \frac{24}{28} = \frac{6}{7}$$

Ce qui montre que dans ce cas, les inégalités précédentes deviennent des égalités.
Nous avions d'ailleurs une égalité dans le cas de la formule valable pour l'indice de Dice puisque pour cet exemple :
$S_d(x,y) = ¾$; $S_d(y,z) = ½$; $S_d(x,z) = ¼$, d'où puisque ¾+½ -¼ =1,
on a bien : $S_d(x,y) + S_d(y,z) - S_d(x,z) = 1$
En fait on peut démontrer que si l'inégalité (f 31) est une égalité alors l'inégalité (f 33) sera également une égalité .

Comme précédemment pour l'indice de Jaccard , en jetant un oeil à la formule (f 21) on voit que l'Indice d'Anderberg s'écrit comme une fonction homographique particulière de l'indice de Dice, à savoir : $S_{an}(x,y) = \frac{S_d(x,y)}{4-3.S_d(x,y)}$ **,** dans ce cas, en appliquant de nouveau le Théorème n°1 (page 3) on voit que cet indice vérifie les conditions d'applications du Théorème n°1, avec une valeur de **β=3**. . De ce fait l'Indice d'Anderberg vérifie la propriété 8 :

Propriété n°8 : L'indice d'Anderberg est un indice « métrisable ».

De plus si le terme 11(x,x)=11(y,y) est constant (exemple s'il est issu de tableaux disjonctifs complets), les indices de Jaccard et d'Anderberg sont métrisables euclidiens .

4.2 Indices du Groupe I, Type II (indices obtenus comme fonction des ratios de « Rappel » et de « Précision »)

Par analogie avec la « recherche documentaire », et en droite ligne avec les remarques liées à la présentation de l'indice général de Tversky (dissymétrie des influences), on définit les ratios dits de « Précision » « P »et de « Rappel » « R » , respectivement, comme les quantités:

(f 34) $$P = \frac{11(x,y)}{11(x,y)+10(x,y)} \quad \text{et} \quad R = \frac{11(x,y)}{11(x,y)+01(x,y)}$$

En effet supposons que nous obtenions : **x** comme résultat d'une recherche sur un domaine **y**, la quantité : 10(x,y) peut être considérée comme un facteur de *« bruit »*, on génère dans **x** des « items » (des valeurs 1 de **x** qui ne correspondent pas à ce qu'on aurait du s'attendre par rapport à **y** (0 de **y** correspondant à des 1 de **x**)), ces items parasites de **x** qui n'ont pas de correspondants dans **y** peuvent être considérés comme du *« bruit »* que l'on a généré à tort, lors d'un processus de recherche documentaire. Inversement la quantité 01 (x,y) peut être considérée comme du *« silence »,* on « rate » par **x** de l'existant sur **y** c'est la raison du mot *« silence »*. La précision est d'autant plus forte qu'on ne crée pas de bruit en effet si 10(x,y)=0 la précision est maximale, inversement si le silence est égal à 0, si 01(x,y)=0, le rappel est alors égal à 1, on récupère donc par les « matchings » de **x** l'ensemble de ce que l'on devait trouver sans oubli sur **y.**
Bien que cette présentation des ratios P et R ne soit qu'analogique, la référence au domaine de la recherche documentaire, (on dirait de nos jours « recherche Internet »), permet de se faire une idée concrète de la signification intuitive de ces ratios. Ayant intuitivement la connaissance d'une « signification » de ces ratios, nous allons voir qu'ils jouent un rôle important dans un certain nombre d'indices du Groupe I en particulier leurs différentes moyennes. On appellera **Groupe I, Type II** , le regroupement des indices appartenant à cette famille très « structurée » d'indices.

4.2.1 Indices de Type II-A (Indices obtenus comme différentes « moyennes » des ratios de « rappel » et de « précision »)

4.2.1.1 Indice de Kulczynski (1923)

L'indice de Kulczynski a été défini par l'auteur (voir Kulczynski (1927)) comme la « moyenne arithmétique » des quantités P et R. Il s'écrit donc:

(f 35) $$S_k(x,y) = \frac{1}{2}[P+R] = \frac{1}{2}\left[\frac{11(x,y)}{11(x,y)+10(x,y)} + \frac{11(x,y)}{11(x,y)+01(x,y)}\right]$$

A cette occasion, on voit que si l'on note:
11(x,x)=Somme des valeurs "1" du profil de x
11(y,y)=la somme des valeurs "1" du profil de y,
alors :

$$S_k(x,y) = \frac{1}{2}[P+R] = \frac{1}{2}\left[\frac{11(x,y)}{11(x,x)} + \frac{11(x,y)}{11(y,y)}\right]$$

soit encore: $$S_k(x,y) = \frac{1}{2}[P+R] = \frac{11(x,y)}{2}\left[\frac{1}{11(x,x)} + \frac{1}{11(y,y)}\right]$$

En posant : $\frac{1}{2}\left[\frac{1}{11(x,x)} + \frac{1}{11(y,y)}\right] = \frac{1}{MH[11(x,x),11(y,y)]}$ (moyenne harmonique des 1 de **x** et des 1 de **y**)
L'indice de Kulczynski s'écrit également sous la forme suivante, sous laquelle il est également connu:

(f 36) $$S_k(x,y) = \frac{11(x,y)}{MH\,[11(x,x), 11(y,y)]}$$

Cet indice varie de 0 à 1, il vaut:

- 1 si 10(x,y)=01(x,y)=0
- 0 si 11(x,y)=0

<u>Propriété N°9</u>: Propriété de l'indice de Kulczynski en cas de Disjonction complète

Comme nous l'avons vu pour les indices du Groupe I -Type I, l'étalonnage d'un indice par rapport à ses valeurs en cas de disjonction complète est un moyen efficace de mesurer la "valeur intrinsèque" du dit indice par rapport à son pouvoir de discrimination . En effet en utilisant le processus suivant dans l'ordre précis indiqué:

a) Donner l'expression de l'indice en cas de disjonction complète,
b) Calculer la borne induite sur la quantité 11(x,y), (seule quantité réellement variable du tableau de contingence dans les conditions de disjonction complète) par rapport à la règle de Solomon-Fortier, que nous avons indiquée précédemment
c) Evaluer cette borne qui permet une sorte d'étalonnage pour un indice par rapport aux autres.

Appliquons le principe:a)➔b➔c sur l'indice de Kulczynski, il vient:
Tout d'abord rappelons les conditions que vérifient les cases du tableau de contingence Tetrachorique si il y a "Disjonction complète": En cas de disjonction complète on a vu en effet que le nombre de variables « m » était tel que :

$$11(x,y) + \frac{1}{2}\left[10(x,y) + 01(x,y)\right] = m$$

d'autre part on a par définition : $11(x,y) + 00(x,y) + 10(x,y) + 01(x,y) = P$

et bien sûr et de façon évidente : $10(x,y) = 01(x,y) = m - 11(x,y)$

D'où l'expression nouvelle de l'indice de Kulczynski:

$$S_k(x,y) = \frac{1}{2}\left[P + R\right] = \frac{1}{2}\left[\frac{11(x,y)}{m} + \frac{11(x,y)}{m}\right] = \frac{11(x,y)}{m} = S_d(x,y)$$

On reverra les conséquences qu'impliquent ce résultat au paragraphe §4.2.3.2.

4.2.1.2 Indice d'Ochiaï (1957)

L'indice d'Ochiai[8] se définit comme la « moyenne géométrique » des indices P et R, sa paternité est attribuée par Sokal et Sneath au zoologue Japonais Ochiai (1957), qui l'aurait introduit, comme un « cosinus » entre profils binaires de description dichotomique en classification d'espèces de poissons dès 1957, il s'écrit donc :

[8] L'indice connu sous le nom d'indice de **Sorgenfrei** , introduit un an plus tard que celui de Ochiaï, par T. Sorgenfrei (1958) n'est en fait que le carré de l'indice d'Ochiaï. $S_{Sorg} = \frac{11^2(x,y)}{11(x,x) + 11(y,y)} = [S_O(x,y)]^2$ et n'a pas d'intérêt particulier, sinon d'être inférieur à celui de Jaccard, et bien entendu à celui d'Ochiaï. Par ailleurs, même si l'indice dit « d'Ochiai » ait été atribué à Ochiai, il semble qu'il ait été introduit antérieurement par **H. Driver et A. Kroeber** , voir H. Driver et A. Kroeber (1932)

(f 37)

$$S_o(x,y) = \sqrt{P.R} = \frac{11(x,y)}{\sqrt{[11(x,y)+10(x,y)][(11(x,y)+01(x,y)]}}$$

D'autre part en utilisant la définition des "1" de x et des "1" de y, l'indice d'Ochiai s'écrit également:

(f 38)

$$S_o(x,\ y) = \frac{11(x,\ y)}{\sqrt{11(x,\ x)11(y,\ y)}} = \frac{11(x,\ y)}{MG\ [11(x,\ x),11(y,\ y)]}$$

Sous cette forme, il peut s'interprêter comme un "cosinus" entre les profils de **x** et de **y**

(sous cette forme, également, on voit qu'il ne différe formellement de l'indice de Kulczynski, que par le choix d'une moyenne géométrique[9] plutôt que d'une moyenne harmonique au dénominateur). Par ailleurs la distance définie par: $d_o(x,y) = \sqrt{2(1-S_o(x,y)}$ est appelée "distance d'Ochiai" dans la littérature[10], ce qui fait de l'Indice d'Ochiai un indice métrisable

Propriété N°10: Propriété de l'indice d'Ochiai en cas de Disjonction complète

Comme nous l'avons vu pour l'indice de Kulczynski, appliquons à l'indice d'Ochiai le principe a)➔b)➔c) défini au § précédent. On obtient la nouvelle formulation de l'indice d'Ochiai:

(f 38)

$$S_o(x,y) = \frac{11(x,y)}{\sqrt{m}^2} = \frac{11(x,y)}{m} = S_d(x,y)$$

[9] Il est intéressant de constater dans ce cas particulier (nous l'avions déjà vu pour l'indice de Van Rijsbergen), que la littérature scientifique produit un nombre considérable d'articles retrouvant des évidences pour certains, et des progrès scientifiques pour les autres, comme le montre la citation suivante extraite d'un article de G.M. Maggira, J.D. Petke et J. Mestres, paru en Avril 2002 dans la Revue « *Journal of Mathematical Chemistry », Vol 31, N°3*,(MPM 2002) … « A formalism is presented that incorporates the entirety of all field-based molecular similarity indices of general form Sij=Ωij/h(ii,jj), where the numerator is given by the **inner product or "overlap"** of field functions Fi and Fj corresponding to the ith and jth molecules, respectively, and the denominator is given by a suitable **mean function** of the self-similarities ii and jj. This family of similarity indices includes the index initially introduced **by Carbó** nearly twenty years ago, where h(ii, jj) is taken to be the **geometric mean** of ii and jj, and the well-known indices due to **Hodgkin and Richards, and Petke**, where h(ii,jj) is taken to be the **arithmetic mean** and maximum of ii and jj, respectively. Two new indices based upon the **harmonic mean** and minimum of ii and jj are also defined, and it is demonstrated that the entire set of field-based similarity indices can be generated from a one-parameter family of functions, **called generalized means**, through proper choice of the parameter value and suitable limiting procedures. Ordering and rigorous bounds for all of the indices are described as well as a number of inter-relationships among the indices. The generalization of field-based similarity indices, coupled with the relationships among indices that have been developed in the present work, place the basic theory of these indices on a more unified and mathematically rigorous footing that provides a foundation for a better understanding of the quantitative aspects of field-based molecular similarity». Nous voyons que « nos amis chimistes » retrouvent des propriétés « indicielles » connues depuis la création des indices de Kulczynski ou Ochiai (depuis 80 ans pour l'un et 63 ans pour l'autre..)

[10] Voir en particulier page 521 du remarquable livre de F. Caillez et J.P. Pages « *Introduction à l'Analyse des Données* » aux éditions SMASH voir (F. Caillez et J.P. Pages (1976))

La borne étant donnée par le milieu de l'intervalle de variation qui est: (0,1), on doit comparer l'indice d'Ochiai à la borne de Solomon Fortier qui vaut ici 1/2 , ceci implique :
$11(x,y) \geq m/2$
On retrouve ici le fait que la borne de Solomon Fortier induit la règle de la majorité simple de Condorcet.

4.2.1.3 Indice de Moyenne Harmonique de P et R

Cet indice que nous noterons $S_h(x,y)$ correspond à la moyenne harmonique du Rappel et de la Précision
Il est donc défini par:

(f 39) $$\frac{1}{S_h(x,y)} = \frac{1}{2}\left[\frac{1}{P} + \frac{1}{R}\right]$$ soit en d'autres termes :

$$S_h(x,y) = \frac{2.P.R}{P+R} = \frac{\frac{211(x,y)}{11(x,y)+10(x,y)} \times \frac{11(x,y)}{11(x,y)+01(x,y)}}{\frac{11(x,y)}{11(x,y)+10(x,y)} + \frac{11(x,y)}{11(x,y)+01(x,y)}}$$

Soit après de premières simplifications:

$$S_h(x,y) = \frac{2.P.R}{P+R} = \frac{2.11^2(x,y)}{11(x,y)[2.11(x,y)+10(x,y)+01(x,y)]}$$

Puis en simplifiant Numérateur et Dénominateur simultanément, et les divisant par 2, il vient :

(f 40) $$S_h(x,y) = \frac{2.11(x,y)}{[2.11(x,y)+10(x,y)+01(x,y)]} = \frac{11(x,y)}{11(x,y)+\frac{1}{2}[10(x,y)+01(x,y)]} = S_d(x,y)$$

En fait, l'indice de moyenne harmonique[11] entre P et R n'est autre que l'indice de Dice, que nous avions déjà défini précédemment (voir formule(f16)) .
En utilisant le passage aux « 1 » de **x** et « 1 » de **y**, on voit que l'indice de Dice peut s'écrire également :

[11]On a déjà vu que l'on peut généraliser l'indice de Moyenne Harmonique c'est ce qu'a proposé C.J. Van Rijbergen (1979) en introduisant une moyenne pondérée du type : $S_{vr}(x,y) = \frac{.P.R}{\alpha P + \beta R}$ et $\alpha + \beta = 1$, (d'ailleurs on aurait put le faire également pour les autres moyennes, mais l'intérêt est faible en particulier au niveau du concept sous-jacent et de la possibilité d' interprétation simple des influences réciproques de P et R).

(f 41)
$$S_d(x,y) = \frac{11(x,y)}{\frac{1}{2}[11(x,x)+11(y,y)]} = \frac{11(x,y)}{MA[\,11(x,x),11(y,y)]}$$

(où MA désigne la moyenne Arithmétique des " 1" de x et des "1" de y) , ceci, en plus des remarques et des propriétés introduites aux paragraphes précédents, montre bien le caractère « central » de l'indice de Dice, puisqu'il était déjà membre du Groupe I, Type I.
Comme il est connu que :

Moyenne Harmonique ≤ Moyenne Géométrique ≤ Moyenne Arithmétique,

Nous avons donc :

$$\frac{2P.R}{P+R} \le \sqrt{P.R} \le \frac{1}{2}(P+R) \quad :$$

On obtient de ce fait les inégalités suivantes entre cette série de 3 indices :

$$S_d(x,y) \le S_o(x,y) \le S_k(x,y) \qquad \forall\, x, y$$

De plus, on a également:

$$S_o(x,y) = \sqrt{S_d(x,y)\ \ S_k(x,y)}$$

L'indice d'Ochiai est la moyenne géométrique de l'indice de Dice et de l'indice de Kulczynski.

Le tableau suivant récapitule les propriétés de ces 3 indices :

	En fonction de P et de R	En fonction de 11(x,x) et 11(y,y)
DICE	**MH**(P,R)	11(x,y)/**MA**(11(x,x),11(y,y))
OCHIAI	**MG**(P,R)	11(x,y)/**MG**(11(x,x),11(y,y))
KULCZYNSKI	**MA**(P,R)	11(x,y)/**MH**(11(x,x),11(y,y)

Si nous regardons en détail et développons les expressions des indices d'Ochiai et de Kulczynski , on peut montrer que ces deux indices peuvent s'écrire en fonction des quantités

$$\alpha^2 = (11(x,y) + \tfrac{1}{2}(10(x,y) + 01(x,y)))^2 \text{ et } \beta^2 = \tfrac{1}{4}(10(x,y) - 01(x,y))^2$$

Selon les formules suivantes:

$$S_k(x,y) = \frac{\alpha\, 11(x,y)}{\alpha^2 - \beta^2} \quad \text{et} \quad S_o(x,y) = \frac{11(x,y)}{\sqrt{\alpha^2 - \beta^2}}$$

4.2.2 Propriétés des indices du Groupe I, Type II-A, en cas de Disjonction Complète

Propriété n°11: Dans le cas particulier, où l'on travaille sur des matrices disjonctives complètes, ces trois indices précédents sont égaux et en particulier leurs expressions sont équivalentes à celle de l'indice de Dice.
En effet du fait que (10(x,y) + 01 (x,y)) = 2(m- 11(x,y)) (voir formule (i) §3.1.2.1),, et que d'autre part, il y a autant de configurations 10(x,y) que de configurations 01(x,y), on a 01(x,y) = 10(x,y) ; dès lors la quantité β^2 est égale à 0. Comme β=0, il vient:

$$S_k(x,y) = \frac{11(x,y)}{\alpha} \quad \text{et} \quad S_o(x,y) = \frac{11(x,y)}{\sqrt{\alpha^2}} = \frac{11(x,y)}{\alpha}$$

En remplaçant α par sa valeur, on retrouve bien la définition de l'indice de Dice.
On vérifie bien de ce fait la Propriété sus-dite :

$$S_d(x,y)\,\rho = \rho\, S_o(x,y)\,\rho = \rho\, S_k(x,y)$$

4.3 Indices du Groupe I, Type II-B (indices obtenus comme autres fonctions des ratios de « Rappel » et de « Précision »)

4.3.1 Indice de moyenne harmonique quadratique normée de P, R

Cet indice noté $S_N(x,y)$ **,** est défini par :

(f 42)
$$S_N(x,y) = \frac{\sqrt{2}.11(x,\ y)}{\sqrt{[11(x,y)+10(x,y)]^2 + [11(x,y)+01(x,y)]^2}} = \frac{\sqrt{2}}{\sqrt{\frac{1}{P^2}+\frac{1}{Q^2}}}$$

Cet indice varie bien de 0 à 1, il vaut 1 si 01(x,y)=10(x,y) =0 et 0 si 11(x,y) = 0.
En utilisant les notations en α^2 et β^2 définies précédemment, on peut montrer que cet indice se simplifie selon la formule suivante:

$S_N(x,y) = \frac{11(x,\ y)}{\sqrt{\alpha^2+\beta^2}}$ et si l'on compare son expression à celle de l'indice d'Ochiai $S_o(x,y) = \frac{11(x,\ y)}{\sqrt{\alpha^2-\beta^2}}$

une relation évidente est dérivable de ce qui précède, il suffit de comparer les dénominateurs, puisque $\sqrt{\alpha^2+\beta^2} \geq \sqrt{\alpha^2-\beta^2}$ on voit que :

(f 43)
$$S_N(x,y) \leq S_o(x,y)$$

De plus , on a de façon complémentaire le résultat suivant :

$$\frac{1}{[S_o(x,y)]^2} + \frac{1}{[S_N(x,y)]^2} = \frac{2.\alpha^2}{[11(x,\ y)]^2}$$

(cette dernière quantité étant 2 fois l'inverse de l'indice de Dice au carré) , on a donc la relation suivante liant ces trois indices :

(f 44) $$\frac{1}{[S_d(x,y)]^2} = \frac{1}{2}\left[\frac{1}{[S_o(x,y)]^2} + \frac{1}{[S_N(x,y)]^2}\right]$$

Sous cette forme il apparaît clairement que l'indice de Dice au carré est la **moyenne harmonique** de l'indice N au carré et de l'indice d'Ochiai au carré.
D'autre part on peut montrer que l'indice $S_N(x,y)$ s'exprime en fonction de l'indice de Dice, l'indice de Kulczynski et de l'indice d'Ochiai suivant les deux formules suivantes :

(f 45) $$S_N(x,y) = \frac{.S_d(x,y)}{\sqrt{\left[2 - S_d(x,y)/S_k(x,y)\right]}} = \frac{S_d(x,y)}{\sqrt{\left[2 - \left[S_d(x,y)/S_o(x,y)\right]^2\right]}}$$

On constate, une fois encore le <u>rôle « central »</u> de l'indice de Dice, du fait de son rôle de « moyenne Harmonique carrée » de $S_N(x,y)$ et $S_o(x,y)$ **,** et compte tenu de la formule (f 43), liant ces deux quantités et du fait que $S_o(x,y) \geq 0$ et $S_N(x,y) \geq 0$, on tire la relation d'inégalités suivante :

(f 46) $$S_N(x,y) \leq S_d(x,y)\ ) \leq S_o(x,y)$$

comme par ailleurs , nous avons vu au paragraphe précédent que :

$$S_d(x,y) \leq S_o(x,y)\ ) \leq S_k(x,y)$$

ceci implique que dans le cas général:

(f 47) $$\boxed{S_N(x,y) \leq S_d(x,y) \leq S_o(x,y) \leq S_k(x,y)}$$

4.3.2 Indices du Min et du Max de P et de R

Ces indices que nous noterons $S_{Min}(x,y)$ et $S_{Max}(x,y)$ varient également de 0 à 1 ils s'écrivent :

$$S_{Min}(x,y) = \text{Min}\ (P,R)$$
$$S_{Max}(x,y) = \text{Max}\ (P,R)$$

Le premier de ces deux indices est connu dans la littérature anglo-saxonne sous le nom d'indice de « **Braun Blanquet** », (voir J. Braun-Blanquet (1932)), utilisé par l'auteur en phytosociologie également, le second d'entre eux $S_{Max}(x,y)$ est également connu sous le nom d' « **<u>indice de Simpson</u>** » dans la littérature anglo-saxonne (voir G. Simpson (1943)), introduit en 1943 par l'auteur, il est surtout utilisé dans le domaine de la similarité moléculaire.
Comme les numérateurs sont égaux, et comme on a la propriété suivante :

$$\text{Min}\ (a,b) = 1/2(a+b] - 1/2\rho\,|\,a-b\,|\,\rho \text{ et } \quad \rho\ \text{Max}\ (a,b) = 1/2(a+b) + 1/2\,|\,a-b\,|\,,$$

(f48) $S_{Min}(x,y) = \dfrac{11(x,y)}{Max\left[11(x,y)+10(x,y),11(x,y)+01(x,y)\right]}$ et

$$S_{Max}(x,y) = \frac{11(x,y)}{Min\left[11(x,y)+10(x,y),11(x,y)+01(x,y)\right]}$$

mais comme :

Max (11(x,y)+10(x,y), 11(x,y)+01(x,y)) = 1/2[2.11(x,y) + 10(x,y)+01(x,y)) +1/2| 10(x,y)-01(x,y) |

Soit :

Max (11(x,y)+10(x,y), 11(x,y)+01(x,y)) = $\alpha + \beta$ où

α et β ont été définies précédemment : α = (11(x,y) + ½ (10(x,y) + 01(x,y))) et

β = ½ |10(x,y) - 01(x,y)|

de même,

Min (11(x,y) +01(x,y),11(x,y)+01(x,y)) = $\alpha - \beta$

Dès lors les nouvelles formulations de $S_{Min}(x,y)$ et de $S_{Max}(x,y)$ sont les suivantes:

$$S_{Min}(x,y) = \frac{11(x,y)}{\alpha+\beta} \quad \text{et} \quad S_{Max}(x,y) = \frac{11(x,y)}{\alpha-\beta}$$

4.3.2.1 Quelques Propriétés des Indices du Min et du Max de P et R

Si l'on calcule les moyennes arithmétique, géométrique et harmonique de $S_{Min}(x,y)$ et de $S_{Max}(x,y)$

on obtient :

$$\frac{1}{2}\left[S_{Min}(x,y)+S_{Max}(x,y)\right] = \frac{1}{2}\left[\frac{2\,\alpha\;11(x,y)}{\alpha^2-\beta^2}\right] = S_k(x,y)\,,$$ comme indiqué au § 4.2.1.3

a) Première Propriété:

L'indice de Kulczynski est la « moyenne arithmétique» des indices $S_{Min}(x,y)$ et $S_{Max}(x,y)$ $\qquad S_k(x,y) = \frac{1}{2}\left[S_{Min}(x,y)+S_{Max}(x,y)\right]$

b) Deuxième Propriété:

De la même façon , on peut montrer que l'indice d'Ochiai $S_o(x,y)$ est la « moyenne géométrique » de ces deux quantités

$$S_o(x,y) = \sqrt{S_{Min}(x,y)\;S_{Max}(x,y)}$$

En effet comme Max (P,R).Min (P,R) = P×R , le résultat précédent s'en déduit simplement.

c) Troisième Propriété:

De même on retrouve un résultat connu pour la moyenne harmonique, puisque comme nous l'indique la formule :

(f 49) $$\frac{1}{S_h(x,y)} = \frac{1}{2}\left[\frac{1}{P}+\frac{1}{R}\right] = \frac{1}{2}\left[\frac{1}{Max(P,R)}+\frac{1}{Min(P,R)}\right] = \frac{1}{2}\left[\frac{1}{S_{Min}(x,y)}+\frac{1}{S_{Max}(x,y)}\right].$$

On en déduit que d'après la formule (f 40), l'indice de Dice est la moyenne harmonique des indices $S_{Min}(x,y)$ et $S_{Max}(x,y)$ sous la forme :

$$\frac{1}{S_d(x,y)} = \frac{1}{2}\left[\frac{1}{S_{Min}(x,y)} + \frac{1}{S_{Max}(x,y)}\right]$$

4.3.2.2 Indice de Mac Connaughey (Variante Corrélative de l'Indice de Kulczynski)

Cet indice, qui aurait pu être introduit plus tard dans cet article, au § 4.4, comme cas particulier des indices du Groupe II-Type II, car issu de valeurs calculables sur le tableau de contingence Tetrachorique, n'est en fait après simplification d'écriture qu'une variante « indice de corrélation » de l'indice de Kulczynski et , de fait, se doit d'être rattaché aux indices du Groupe I. En effet cet indice dont on trouvera mention dans l'article Boyce et Ellison (2001) , s'écrit de la façon suivante :

(f 50)
$$S_{McC}(x, y) = \left[\frac{[11(x, y)]^2 - 10(x, y).01(x, y)}{[11(x, y) + 10(x, y)].[11(x, y) + 01(x, y)]}\right]$$

Donné sous cette forme, l'indice de Mac Connaughey semble être un indice, nouveau qui varie de –1 à +1 (-1 si 11(x,y)=0, +1 si 10(x,y)=01(x,y)=0). En fait il n'en est rien , il se décompose en deux parties symétriques et duales relativement aux indices de Rappel et de Précision. En effet par un simple jeu d'écriture, on voit que cet indice s'exprime en fonction de P et de R selon la formule suivante :

(f 51)
$$S_{McC}(x, y) = [P\,R - (1-P)(1-R)]] = P + R - 1 = 2\left[\frac{P+R}{2}\right] - 1 = 2\,S_K(x, y) - 1$$

C'est donc une **simple translation linéaire de l'indice de Kulczynski** , il vaut 1 quand l'indice de Kulczinski vaut 1 et il vaut –1 quand l'indice de Kulczynski vaut 0.

4.3.2.3 Indice ou fonction F(γ) de compromis P,R

Cette fonction, qui est souvent utilisée par les spécialistes du TAL (Traitement et Analyse de la Langue) (voir J. Beney (2008)) comme "bonne" mesure de compromis entre Rappel et Précision a été définie, par l'expression:

$$F_\gamma(x, y) = \frac{(\gamma^2 + 1)P.R}{\gamma^2 P + R}$$

En divisant le numérateur et le dénominateur par la quantité : (γ^2+1), on voit que l'on retrouve exactement l'indice de Van Rijsbergen défini en (f 30') , une fois posé :

$\alpha = \frac{\gamma^2}{\gamma^2+1}$ et $\beta = \frac{1}{\gamma^2+1}$ on vérifie bien dès lors que : $\alpha+\beta=1$ et que:

$$F_\gamma(x, y) = \frac{(\gamma^2 + 1)P.R}{\gamma^2 P + R} = \frac{P.R}{\alpha P + \beta R} = S_{vr}(x, y) = S_{Ty}(x, y) \quad \text{pour} \quad \alpha + \beta = 1$$

en conclusion par rapport à la typologie que nous venons de faire, cette fonction F n'apporte rien de plus que l'Indice de Tversky écrit sous la forme :

$$S_{Ty}(x, y) = \frac{11(x, y)}{11(x, y) + \frac{\gamma^2}{\gamma^2+1}10(x, y) + \frac{1}{\gamma^2+1}01(x, y)}$$

cependant on peut noter que: F(1)=Indice de Dice , F(0)=Précision, F(∞) =Rappel, en dehors de ce petit résultat cette écriture est donc totalement équivalente à celle de Tversky et nous ne conserverons pas cette fonction dans notre discussion finale. .

4.3.3 Bilan comparatif de positionnement des Indices du Groupe I

Propriété n°12: Egalité des indices en cas de disjonction complète

Dans le cas particulier, où l'on travaille sur des matrices disjonctives complètes, les trois indices du GROUPE I, Type II-B, auxquels on peut rajouter P et R (ils jouent en fait le rôle de $S_{Min}(x,y)$ et $S_{Max}(x,y)$ suivant que l'un est supérieur à l'autre) sont égaux et en particulier leur expression est équivalente à celle de l'indice de Dice et vaut : $11(x,y)/\alpha$
Ceci s'obtient en posant $\beta=0$ dans chacune des formules en α et β relatives à ces indices .
Dans le cas disjonctif complet on a donc:

$$S_d(x,y)\rho = \rho\, S_o(x,y)\rho = \rho\, S_k(x,y) = S_N(x,y) = S_{Max}(x,y) = S_{Min}(x,y) = P = R$$

Une conséquence évidente, est que, de facto, tous ces indices du **Groupe I Type II** sont, en cas de disjonction complète, de la forme:

$$S=1-(1/2)d^2$$

et dès lors sont <u>des indices de similarités métrisables euclidiens</u>

Propriété n°13: Ordonnance des Indices du Groupe I au sens Général

Dans le cas <u>général</u> nous obtenons l'inégalité suivante entre tous les indices du Type II:

$$S_{Min}(x,y)\ \rho \le S_N(x,y)\,\rho \le S_d(x,y) \le S_o(x,y) \le S_k(x,y)\ \rho \le S_{Max}(x,y)$$

Si nous rajoutons tous les indices du GROUPE I (Type I, Type II - (A et B)), les inégalités entre eux nous permettent d'écrire , pour tout couple (x,y) :

$$S_{an}(x,y) \le S_j(x,y) \le S_{Min}(x,y)\,\rho \le S_N(x,y)\rho \le S_d(x,y) \le \rho\ S_o(x,y) \le S_k(x,y)\rho \le S_{Max}(x,y)$$

Il reste à positionner $S_{as}(x,y)$et $S_{so}(x,y)$, comme $S_{so}(x,y)\,\rho \le S_{as}(x,y)$, il reste à positionner $S_{so}(x,y)$ par rapport à l'échelle précédente.
Comme $S_{so}(x,y)$ s'exprime sous la forme suivante par rapport à α : $S_{so}(x,y) = \dfrac{11(x,y)}{\alpha-\delta}$

où $\alpha = 11(x,y) + 1/2\,(10(x,y) + 01(x,y))$ et $\delta = \frac{1}{4}\,(01(x,y) + 10(x,y))$

:

comparons donc $S_{so}(x,y) = \dfrac{11(x,y)}{\alpha-\delta}$ à $S_{Max}(x,y) = \dfrac{11(x,y)}{\alpha-\beta}$ (voir précédemment)

Pour que $S_{Max}(x,y) = \dfrac{11(x,y)}{\alpha-\beta}$ soit inférieur à $S_{so}(x,y) = \dfrac{11(x,y)}{\alpha-\delta}$

il suffit que : $\dfrac{11(x,y)}{\alpha-\beta} \le \dfrac{11(x,y)}{\alpha-\delta}$, soit donc : $\alpha-\delta \le \alpha-\beta$, c'est-à-dire : $\beta \le \delta$, ce qui implique:

$$\tfrac{1}{2}\,|\,01(x,y) - 10(x,y)\,| \le \tfrac{1}{4}\,(01(x,y) + 10(x,y))$$

Pour enlever le signe « | », il suffit de passer au Max et Min, il vient alors :

$$2(\mathrm{Max}\,(01(x,y),10(x,y)) - \mathrm{Min}\,(01(x,y),10(x,y))) \le (\mathrm{Max}\,(01(x,y),10(x,y)) + \mathrm{Min}\,(01(x,y),10(x,y)))$$

soit:

(f 52) $\quad \text{Max}\,[\,01(x,y),\,10(x,y)) \;\le\; 3\,\text{Min}\,(01(x,y),10(x,y))$

Si cette condition (f 52) est vérifiée on a donc: $S_{Max}(x,y) \le S_{so}(x,y)$ et de ce fait :

(f 52') $S_{an}(x,y) \le S_j(x,y) \le S_{Min}(x,y) \le S_N(x,y) \le S_d(x,y) \le S_o(x,y) \le S_k(x,y) \le S_{Max}(x,y) \le S_{so}(x,y) \le S_{ac}(x,y)$

Si la condition inverse se produit: $\text{Max}\,[01(x,y),10(x,y)) > 3\,\text{Min}\,(01(x,y),10(x,y))$, nous ne sommes plus assurés de la séquence précédente, il faut alors comparer $S_{ac}(x,y)$ à $S_{Max}(x,y)$, nous avons à comparer $\alpha - \delta'$ et $\alpha - \beta$ (avec $\delta' = 3/8\,(01(x,y) + 10(x,y))$).
De la même façon que précédemment, il faut comparer ces quantités à l'aide des notations en Min et Max et pour que $S_{Max}(x,y)$ soit inférieur à $S_{ac}(x,y)$: il faut et il suffit que :
$\alpha - \delta' \le \alpha - \beta$ et de ce fait, il faut que : $\beta \le \delta'$, c'est à dire :

$$\text{Max}\,(01(x,y),10(x,y)) - \text{Min}\,(01(x,y),10(x,y)) \le 3/4\,(\text{Max}\,(01(x,y),10(x,y)) + \text{Min}\,(01(x,y),10(x,y)))$$

soit:

(f 53) $\quad \text{Max}\rho\,[01(x,y),10(x,y)) \le 7\,\text{Min}(01(x,y),10(x,y))$

dès lors si: $\quad 3\,\text{Min}(01(x,y),10(x,y)) \le \rho\ \text{Max}\rho\ (01(x,y),10(x,y)) \le 7\,\text{Min}\,(01(x,y),10(x,y))$, alors:

(f 53') $S_{an}(x,y) \le S_j(x,y) \le S_{Min}(x,y) \le S_N(x,y) \le S_d(x,y)) \le S_o(x,y) \le S_k(x,y) \le S_{so}(x,y) \le S_{Max}(x,y) \le S_{ac}(x,y)$

Si cette dernière condition n'est pas vérifiée, soit maintenant si :

(f 54) $\quad 7\,\text{Min}\,(01(x,y),10(x,y))\boldsymbol{\alpha} \le \rho\,\text{Max}(01(x,y),10(x,y))$

alors on aura:

(f 54') $S_{an}(x,y) \le S_j(x,y) \le S_{Min}(x,y) \le S_N(x,y) \le S_d(x,y)) \le S_o(x,y) \le S_k(x,y) \le S_{so}(x,y) \le S_{ac}(x,y) \le S_{Max}(x,y)$

Dans le cas particulier, où l'on a <u>disjonction complète</u>, les résultats se simplifient en :

a) Puisque alors $\beta=0$, on a une première série d'égalités (pour les indices fonctions de P et de R),

$$S_{Min}(x,y) = S_N(x,y) = S_d(x,y) = S_o(x,y) = S_k(x,y) = S_{Max}(x,y) = \frac{11(x,y)}{m}$$

b) Pour les indices classiques du Groupe I-Type I (Ratios directs), on obtient les inégalités suivantes:

$$S_{an}(x,y) = \frac{11(x,y)}{4m - 3.11(x,y)} \le S_j(x,y) = \frac{11(x,y)}{2m - 11(x,y)} \le S_d(x,y) = \frac{11(x,y)}{m} \le S_{so}(x,y) = \frac{2.11(x,y)}{m + 11(x,y)} \le S_{ac}(x,y) = \frac{4.11(x,y)}{m + 3.11(x,y)}$$

<u>En conclusion</u>, dans le cas de disjonction complète le **<u>rôle central</u>** de $S_d(x,y)$ est amplifié car : dans le cas du système d'égalités (a) tous les indices sont égaux à $S_d(x,y)$, dans le système d'inégalités (b), $S_d(x,y)$ joue le rôle de « médiane » des indices, indices qui sont tous, par ailleurs, fonctions homographiques de $S_d(x,y)$.

4.3.4 En guise de Conclusion sur les Indices du Groupe I

Comme nous venons de le voir, les indices du Groupe I se divisent en deux classes distinctes, même si nous avons vu que cette classification n'est pas une partition puisque l'Indice de Dice, par exemple, se retrouve dans les deux Types. Par ailleurs, le modèle de Tversky (1977) qui tend à unifier ces deux types par une formulation unificatrice ne permet pas d'exploiter la logique réelle de formation des indices. Nous avons vu néanmoins à ce propos que les cases connues de la nomenclature récapitulative représentent des indices d'usage courant et intuitifs. Plus intéressante est la tentative faite par Julien Ah Pine dans sa thèse de l'université Paris VI (voir J. Ah-ine (2007)), sur les aspects mathématiques : algébriques et combinatoires de l'Analyse Relationnelle ; en effet ce dernier redonne (pp73 à 88) une interprétation « métrique » (scalaire généralisée) de ces indices du Groupe I permettant leur extension au delà du domaine des données « binaires » vers le domaine des données » continues ».
Pour ce faire il introduit deux paramètres géométriques présents de façon sous-jacente dans chacun de ces indices du Groupe I, à savoir : le cosinus de l'angle formé par les vecteurs binaires » d'une part, le rapport de la plus grande norme des deux sur la plus petite. Il montre ainsi que l'extension continue potentielle de certains des indices du Groupe I , vers un formalisme de « bon » indice totalement général s'applique en particulier à l'indice de **Dice** et à l'indice d'**Ochiaï** . Outre qu'il retrouve par là même et d'une façon totalement différente ce que nous avions mentionné précédemment sur les « bons » indices, il donne également une interprétation géométrique en terme de « projecteurs » aux deux indices partiels de « Rappel » et de « Précision ». Nous renvoyons le lecteur à son travail pour de plus amples renseignements.

4.4 Indices appartenant au GROUPE II, faisant jouer un rôle aux structures 11(x,y) et 00(x,y).

4.4.1 Indices du Groupe II, Type I (indices obtenus par ratios directs, faisant jouer un rôle linéaire aux structures 00(x,y))

4.4.1.1 Indice de Sokal et Michener (1958), Green-Rao (1969)

Cet indice (voir R. Sokal et C. Michener (1958)), qui porte également le nom de « simple matching index » est l'un des plus intuitifs de ceux introduits dans la littérature des indices de similarité, en effet si l'on revient à l'expression du tableau contingentiel « tetrachorique » (2×2) introduit précédemment au § 3, on voit que ce coefficient, l'un des premiers à avoir été décrit dans la littérature, est le rapport des cases de la diagonale du tableau, divisé par la somme des quatre cases du tableau précédent.
En fait si la variante « similarité » est attribuée généralement à Sokal et Michener (1958), Green et Rao[12] (1969) ayant introduit, quant à eux, la variante « dissimilarité » de

[12] I.C. Lerman (1970), attribue à Hamann (1961), l'indice qui est défini par : $S_{Ha}(x, y) = \frac{11(x, y) + 00(x, y) - [10(x, y) + 01(x, y)]}{P}$, c'est en fait la différence entre l'Indice de Sokal et Michener et celui de Green et Rao, il varie de(–1à +1)

l'indice (qui est d'ailleurs le complémentaire de celui de Sokal et Michener par rapport à la somme des cases du tableau « Tetrachorique »), il se trouve que l'idée originale de ce concept est bien plus ancienne encore. En effet on peut la faire remonter à A. de Condorcet (1785), voir F. Marcotorchino (1981), F. Marcotorchino et N. El Ayoubi (1991), ce critère de Condorcet , qui dans le cas de comparaisons de vecteurs relationnels est équivalent au coefficient de Rand (voir W. Rand (1971)) lequel coefficient semble avoir été introduit avant Rand par H.Borko et All. (1968) (voir à ce propos le livre de M. Anderberg (1973) page 206). En tout état de cause c'est ce principe de « simple matching » qui prévaut dans l'approche Condorcéenne, introduite en 1785. Ce qui d'une certaine façon met tout le monde d'accord quant à la paternité de l'idée. Le critère de « simple matching » s'écrit donc :

(f 55) $$S_{sm}(x,y) = \frac{11(x,y)+00(x,y)}{P} = \frac{11(x,y)+00(x,y)}{11(x,y)+00(x,y)+10(x,y)+01(x,y)}$$

Cet indice varie de 0 à 1,

- il vaut 0 si 11(x,y)=00(x,y)=0
- il vaut 1 si 10(x,y)=01(x,y)=0

Propriété n°14 : l'indice de Sokal et Michener vérifie la condition de « Transitivité Généralisée Indicielle »

Montrons par ailleurs que $S_{sm}(x,y)$ vérifie la condition de Transitivité Généralisée Indicielle, c'est à dire que l'on a :

(f 56) $$S_{sm}(x,y) + S_{sm}(y,z) - S_{sm}(x,z) \le 1 \quad \forall\ x,y,z$$

soit : $$11(x,y) + 00(x,y) + 11(y,z) + 00(y,z) - [11(x,z) + 00(x,z)] \le P$$

En remplaçant ces valeurs en fonctions des quantités x_j et y_j , il vient après simplifications:

$$\sum_{j=1}^{P} x_j y_j + \sum_{j=1}^{P} y_j z_j - \sum_{j=1}^{P} y_j - \sum_{j=1}^{P} x_j z_j \le 0$$

Le membre de gauche de cette inégalité peut également s'écrire sous la forme produit suivante :

(f 57) $$\sum_{j=1}^{P} x_j y_j + \sum_{j=1}^{P} y_j z_j - \sum_{j=1}^{P} y_j - \sum_{j=1}^{P} x_j z_j = \sum_{j=1}^{P} (x_j - y_j)(y_j - z_j)$$

Comme x_j,y_j et z_j sont des valeurs égales à 0 ou 1 $\forall$j, il suffit de montrer que pour toutes les valeurs de l'indice « j », le produit $(x_j-y_j)(y_j-z_j)$, ne peut en aucun cas être strictement positif. Pour ce faire explorons exhaustivement les $8=2^3$ configurations possibles des valeurs x_j, y_j, z_j., on obtient le tableau suivant :

x_j	y_j	z_j	Valeur du produit (x_j-y_j) (y_j-z_j)
0	0	0	0
0	0	1	0
0	1	0	-1
1	0	0	0
0	1	1	0
1	0	1	-1
1	1	0	0
1	1	1	0

On constate que le produit $(x_j-y_j)(y_j-z_j)$ est toujours négatif ou nul pour toutes les configurations des valeurs x_j, y_j,,z_j, donc que la quantité somme de ces produits sera négative, soit :

$$\sum_{j=1}^{P}(x_j - y_j)(y_j - z_j) \le 0$$

La formule (f56) est donc vraie (cqfd)

Conclusion : L'indice de Sokal et Michener est donc un <u>indice de similarité « métrisable</u> » l'indice de distance associé s'écrit :

$$d_{sm}(x,y) = 1 - S_{sm}(x,y) = \frac{\sum_{j=1}^{P}(x_j - y_j)^2}{P}$$

Si P est un nombre pair on peut écrire P=2p', il vient alors:

(f 58) $$S_{sm}(x,y) = 1 - \frac{1}{2}\frac{\sum_{j=1}^{P}(x_j - y_j)^2}{p'} = 1 - \frac{1}{2}\sum_{j=1}^{P}\left(\frac{x_j}{\sqrt{p'}} - \frac{y_j}{\sqrt{p'}}\right)^2 = 1 - \frac{1}{2}\delta^2(x,y)$$

L'indice de Sokal et Michener est donc un <u>indice métrisable euclidien</u> , d'après le théorème de Gower –Schoenberg.

4.4.1.2 Indice de Rogers et Tanimoto (1960)

Cet indice, introduit par les auteurs dans un article sur l'écologie botanique (voir (D. Rogers et T. Tanimoto (1960)), joue, pour les indices du Groupe II -Type I, un rôle analogue à celui joué par l'indice d'Anderberg pour les indices du Groupe I - Type I, en effet il pondère légèrement plus (poids =2) les cas d'« erreurs » de matching. Il est défini par :

(f 59) $$S_{rt}(x,y) = \frac{11(x,y) + 00(x,y)}{11(x,y) + 00(x,y) + 2[10(x,y) + 01(x,y)]} = \frac{11(x,y) + 00(x,y)}{2.P - [11(x,y) + 00(x,y)]}$$

En éliminant les quantités communes : 11(x,y)+00(x,y) et (10(x,y)+01(x,y)), entre les deux indices précédents on obtient la relation suivante entre l'indice de Rogers et Tanimoto et celui de « Simple Matching » :

(f 60) $$S_{rt}(x,y) = \frac{S_{sm}(x,y)}{[2 - S_{sm}(x,y)]}$$ et réciproquement : $$S_{sm}(x,y) = \frac{2.S_{rt}(x,y)}{[S_{rt}(x,y) + 1]}$$

<u>Propriété n°15</u> : l'indice de Rogers et Tanimoto vérifie « la condition de Transitivité Généralisée Indicielle et est métrisable »

En effet en se reportant à la formule (f57) ci-dessus , on voit que les conditions d'application des Théorème n°1 et Théorème n°1 bis sont vérifiées par l'indice de Rogers et Tanimoto , il suffit de voir que comme

$S_{rt}(x,y) = \frac{S_{sm}(x,y)}{[2 - S_{sm}(x,y)]}$ s'écrit $S_{rt}(x,y) = \frac{S_{sm}(x,y)}{\beta[1 - S_{sm}(x,y)] + 1}$ avec β=1, il est donc <u>métrisable</u>, d'après le Théorème n°1 et <u>métrisable euclidien</u> d'après le Théorème n°1 bis, si P est pair.

4.4.1.3 Indice de Sokal et Sneath (1963)

Cet indice, introduit par les auteurs en 1963 (voir (Sokal et Sneath (1963)) et rejustifié dans leur livre de 1973 (voir Sokal et Sneath (1973)) est une variante des indices précédents, avec une définition très voisine de celle qu'a l'indice de Dice au regard des indices du Groupe I-Type I, en effet il donne moins de poids que les deux configurations précédentes aux situations de non concordance. Il s'écrit :

(f 61) $$S_{ss}(x,y) = \frac{11(x,y)+00(x,y)}{11(x,y)+00(x,y)+\frac{1}{2}[10(x,y)+01(x,y)]} = \frac{2.[11(x,y)+00(x,y)]}{P+[11(x,y)+00(x,y)]}$$

En éliminant les quantités communes : (11(x,y)+00(x,y)) et (10(x,y)+01(x,y)), entre l'indice de Sokal et Sneath et celui de « Simple Matching », on obtient la relation homographique ci dessous :

(f 62) $$S_{ss}(x,y) = \frac{2.S_{sm}(x,y)}{[S_{sm}(x,y)+1]} \text{ et réciproquement : } S_{sm}(x,y) = \frac{S_{ss}(x,y)}{[2-S_{ss}(x,y)]}$$

Enfin en éliminant $S_{sm}(x,y)$ dans les deux formules (f 61) et (f 62), on obtient la relation suivante entre les indices de Sokal- Sneath et Rogers - Tanimoto :

(f63) $$S_{ss}(x,y) = \frac{4.S_{rt}(x,y)}{[3.S_{rt}(x,y)+1]} \text{ et la formule réciproque : } S_{rt}(x,y) = \frac{S_{ss}(x,y)}{[4-3.S_{ss}(x,y)]}$$

Du fait que ces 3 indices du Groupe II, Type I, soient tous fonction homographique de l'indice de « Simple Matching », nous permet de les représenter sur un diagramme unique, où l'on peut voir de façon simple comment ils se comportent et comment ils varient les uns par rapport aux autres. Par ailleurs bien évidemment tous ces indices varient de 0 à 1, ce que nous constatons d'ailleurs sur le graphique de la Figure N°2 :

$$0 \le S_{\alpha}(x,y) \le 1$$

- Ces indices valent 0 si le nombre de concordances (« matchings ») positives et concordances négatives sont nulles entre x et y $\Leftrightarrow$ 11(x,y)=0 et 00(x,y)=0
- Ils valent 1 si la quantité (10(x,y)+01(x,y) = 0 $\Leftrightarrow$ x et y ont le même profil de 1 et de 0
- Ces trois indices vérifient les inégalités suivantes $\forall$ (x,y) :

(f64) $$0 \le S_{rt}(x,y) \le S_{sm}(x,y) \le S_{ss}(x,y) \le 1$$

.

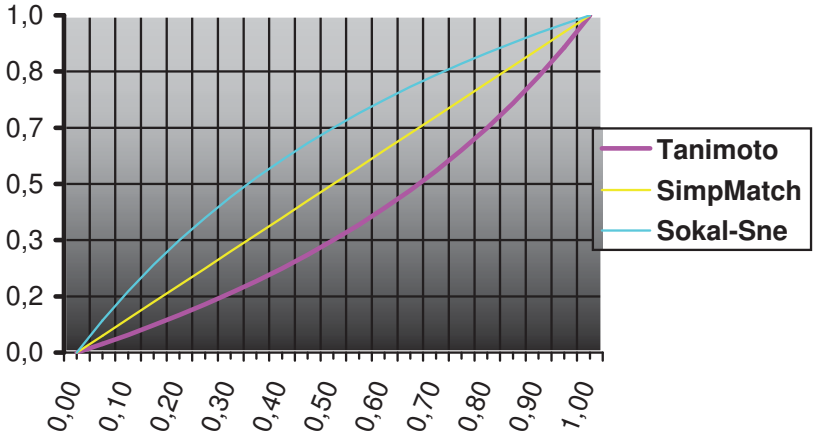


Figure 2: Graphique de variation des indices du GROUPE II, Type I

4.4.2 Quelques propriétés des indices du Groupe II-Type I en cas de disjonction complète

Dans le cas ou les indices précédents sont calculés sur les tableaux obtenus par disjonction complète, des simplifications vont se produire qui permettent des convergences.

En cas de disjonction complète on a vu en effet que le nombre de variables « m » était tel que :

$$11(x,y)+\frac{1}{2}[10(x,y)+01(x,y)]=m$$

d'autre part on a : $11(x,y)+00(x,y)+10(x,y)+01(x,y)=P$

De ce fait la quantité $[10(x,y)+01(x,y)]=2m-2.11(x,y)$ remplacée dans l'égalité sur P nous donne :

(f 65) $$00(x,y)=P+11(x,y)-2m$$

Alors en remplaçant cette valeur dans les trois indices précédents nous obtenons pour chacun d'entre eux des indices qui ne dépendent que de 11(x,y) et des constantes P et m:

a) Expression du coefficient de « Simple Matching »

Le coefficient S_{sm} vaut alors :

$$S_{sm}(x,y)=\frac{11(x,y)+00(x,y)}{P}=\frac{11(x,y)+P+11(x,y)-2.m}{P}=1+\frac{2.11(x,y)}{P}-\frac{2.m}{P}$$

b) Expression du coefficient de Sokal et Sneath

Le coefficient S_{ss} vaut dans ce cas :

$$S_{ss}(x,y)=\frac{2.[11(x,y)+00(x,y)]}{P+[11(x,y)+00(x,y)]}=\frac{2P+2[2.11(x,y)-2.m]}{2P+2.11(x,y)-2.m}=1+\frac{11(x,y)-m}{P+11(x,y)-m}$$

c) Expression du coefficient de Rogers-Tanimoto

Le coefficient S_{rt} vaut dans ce cas :

$$S_{rt}(x,y) = \frac{11(x,y)+00(x,y)}{2.P-[11(x,y)+00(x,y)]} = \frac{P-[2.m-2.11(x,y)]}{P+[2.m-2.11(x,y)]} = 1-\frac{2[2.m-2.11(x,y)]}{P+[2m-2.11(x,y)]}$$

Comme dans le cas de disjonction complète, on a toujours le résultat suivant : $S_d(x,y)=\frac{11(x,y)}{m}$ (voir formule (f 25)) montrons que ces trois indices sont dans le cas de disjonction complète des fonctions de l'indice de Dice :

Il suffit pour cela de diviser Numérateur et Dénominateur des indices précédents par « m », il vient :

- $S_{sm}(x,y) = 1+\frac{2.11(x,y)}{P}-\frac{2.m}{P} = 1-\frac{2.m}{P}[1-S_d(x,y)]$ ou réciproquement :

  $S_d(x,y) = 1-\frac{P}{2.m}[1-S_{sm}(x,y)]$

- $S_{ss}(x,y) = 1+\frac{11(x,y)-m}{P+11(x,y)-m} = 1-\frac{m[1-S_d(x,y)]}{P-m[1-S_d(x,y)]}$ ou réciproquement :

  $S_d(x,y) = 1-\frac{P}{2.m}\frac{[1-S_{ss}(x,y)]}{[1-\frac{1}{2}S_{ss}(x,y)]}$

- $S_{rt}(x,y) = 1-\frac{2[2.m-2.11(x,y)]}{P+[2m-2.11(x,y)]} = 1-\frac{4m[1-S_d(x,y)]}{P+2m[1-S_d(x,y)]}$ ou réciproquement :

  $S_d(x,y) = 1-\frac{P}{2m}\frac{[1-S_{rt}(x,y)]}{[1+S_{rt}(x,y)]}$

Dès lors la comparaison de ces indices au travers de la « borne de Salomon -Fortier » que nous avions utilisée précédemment pour l'indice de Dice (qui se traduisait par la règle majoritaire de Condorcet) induit les bornes suivantes pour ces trois indices du Groupe II-Type I :

$$S_d(x,y)\geq\frac{1}{2} \Leftrightarrow 1-\frac{P}{2.m}[1-S_{sm}(x,y)]\geq\frac{1}{2} \Leftrightarrow S_{sm}(x,y)\geq 1-\frac{m}{P} \Leftrightarrow 11(x,y)\geq\frac{m}{2}$$

$$S_d(x,y)\geq\frac{1}{2} \Leftrightarrow 1-\frac{P}{2.m}\frac{[1-S_{ss}(x,y)]}{[1-\frac{1}{2}S_{ss}(x,y)]}\geq\frac{1}{2} \Leftrightarrow S_{ss}(x,y)\geq 1-\frac{m}{2P-m} \Leftrightarrow 11(x,y)\geq\frac{m}{2}$$

$$S_d(x,y)\geq\frac{1}{2} \Leftrightarrow 1-\frac{P}{2m}\frac{[1-S_{rt}(x,y)]}{[1+S_{rt}(x,y)]}\geq\frac{1}{2} \Leftrightarrow S_{rt}(x,y)\geq 1-\frac{2.m}{P+m} \Leftrightarrow 11(x,y)\geq\frac{m}{2}$$

Remarque n°5 : *Dans le cas particulier où l'on se trouve dans les conditions de la Remarque n°1, tableau de « présence-absence » dédoublé , on a* P=2m*, de ce fait les inégalités précédentes se simplifient et l'on obtient :*

$S_{ss}(x,y)\geq 1-\frac{m}{2P-m} \Leftrightarrow S_{ss}(x,y)\geq 1-\frac{m}{4m-m}=\frac{2}{3} \Leftrightarrow$ Equivalence de l'indice de "Sokal et Sneath" avec l'"indice de Sorensen" pour la règle de Salomon Fortier appliquée sur l'"Indice de Dice"

$S_{rt}(x,y)\geq 1-\frac{2.m}{P+m} \Leftrightarrow S_{ss}(x,y)\geq 1-\frac{2.m}{2.m+m}=\frac{1}{3} \Leftrightarrow$ Equivalence de l'indice de "Rogers Tanimoto" avec l'"indice de Jaccard" pour la règle de Salomon Fortier appliquée sur l'"Indice de Dice"

Remarque n°6 : *Dans la configuration précédente, nous avions privilégié une comparaison de ces trois indices en référence au fait que l'indice de Dice vérifiait lui même la condition de Salomon et Fortier et nous avons voulu mesurer l'impact induit au niveau de chacun d'entre eux sur les bornes ainsi générées.*
Une autre approche intéressante consiste en la démarche inverse, (comme nous l'avions fait dans le cas des indices du Groupe I –Type I), c'est à dire à exhiber les contraintes que doit vérifier la quantité : 11(x,y) relativement à chacun de ces indices pour que la condition de Salomon –Fortier soit vérifiée.

- $S_{sm}(x, y) \geq \frac{1}{2} \quad \Leftrightarrow \quad 1 + \frac{2.11(x, y)}{P} - \frac{2.m}{P} \geq \frac{1}{2} \quad \Leftrightarrow \quad 11(x, y) \geq m - \frac{P}{4}$
- $S_{ss}(x, y) \geq \frac{1}{2} \quad \Leftrightarrow \quad 1 + \frac{11(x, y) - m}{P + 11(x, y) - m} \geq \frac{1}{2} \quad \Leftrightarrow \quad 11(x, y) \geq m - \frac{P}{3}$
- $S_{rt}(x, y) \geq \frac{1}{2} \quad \Leftrightarrow \quad 1 - \frac{2[2.m - 2.11(x, y)]}{P + [2m - 2.11(x, y)]} \geq \frac{1}{2} \quad \Leftrightarrow \quad 11(x, y) \geq m - \frac{P}{6}$

On constate immédiatement à l'aune de ces valeurs, dépendantes de P et de m, que pour les tableaux disjonctifs complets à grand nombre de modalités, ces bornes ont peu de signification, en effet supposons que $\overline{p}.$ soit le nombre de modalités moyen de l'ensemble des variables alors $P = \overline{p}.m$, de ce fait les bornes précédentes s'écrivent respectivement : $m(1-\frac{\overline{p}}{4})$, $m(1-\frac{\overline{p}}{3})$, $m(1-\frac{\overline{p}}{6})$, supposons, par exemple, que $\overline{p} = 5$ (ce qui est un chiffre peu élevé, très souvent rencontré dans applications concrètes), alors les bornes sont négatives pour l'indice de Simple Matching et Sokal et Sneath, donc la condition de Salomon-Fortier est toujours vérifiée, et il suffit que 11(x,y) soit supérieur à m/6 pour que la condition de Salomon -Fortier soit vérifiée par l'indice de Rogers et Tanimoto . Ces indices ne seront donc pas très utilisables dans le contexte de tableaux disjonctifs associés à des variables qualitatives dès lors que le nombre moyen de modalités est supérieur à 3, on privilégiera de ce fait les indices du Groupe I-Type I dans ce contexte.

4.4.3 Indices « non Classiques » du Groupe II - Type I (indices obtenus par ratios directs, faisant jouer un rôle aux structures 00(x,y), au Numérateur et au Dénominateur)

En plus des indices dits « classiques », en un mot les plus utilisés de ce Groupe II-Type I, il existe des variantes jouant sur des pondérations à la fois au numérateur pour la configuration 00(x,y) et au dénominateur pour les configurations 10(x,y) et 01(x,y). En fait la forme[13] générale $S_G(x,y)$ des indices du Groupe II-Type I est donnée par :

(f 66) $$S_G(x, y) = \frac{11(x, y) + \alpha\, 00(x, y)}{11(x, y) + 00(x, y) + \beta[10(x, y) + 01(x, y)]}$$

[13] En jouant sur l'intersection du modèle S_G précédent et du Modèle de Tversky , on peut fabriquer un indice encore plus général avec 4 paramètres variables, de la forme : $S_{GG}(x, y) = \frac{11(x, y) + \lambda 00(x, y)}{11(x, y) + \mu 00(x, y) + \alpha 10(x, y) + \beta 01(x, y)}$

On retrouve dans le tableau ci-dessous en fonction des valeurs de α et β le nom des indices déjà rencontrés et nous expliciterons dans ce paragraphe ceux qui n'ont pas encore été définis.

α \ β	1/2	1	2
0	Rao2	Rao (1945)	
1/2	Marco-Michaud2 (1980)	Marco-Michaud (1980)	
1	Sokal- Sneath (1968)	Sokal-Michener (1958)	Rogers-Tanimoto(1960)

4.4.3.1 Indice de T.R. Rao (1945)

Cet indice, bien qu'introduit initialement par PF. Russel et T..R. Rao en 1940, redétaillé ensuite par R. Rao seul en 1945 est assez peu utilisé dans la pratique car il est souvent trop sévère et s'éloigne de structures d'équilibre interprétables. Il consiste dans le simple rapport entre le nombre de Matching 11(x,y) divisé par le nombre total d'attributs descriptifs, qu'on notera, comme défini au début de l'article , par P. Ce nombre définissant soit le nombre de variables descriptives de « présence-absence », soit le nombre de modalités de l'ensemble des variables, quand on travaille sur un tableau disjonctif complet. Du fait de la présence des configurations 00(x,y) au dénominateur (quantités qui sont souvent très grandes), les valeurs de l'indice de Rao sont souvent très faibles. Par ailleurs contrairement aux autres indices du groupe II, il ne fait pas jouer de rôle aux configurations 00(x,y) au numérateur (ici α=0 ce qui ne permet pas le rééquilibrage). Suite à ce que nous venons de mentionner au paragraphe précédent, il apparaît que cet indice est encore moins bien adapté que les indices précédents aux calculs de similarités sur tableaux disjonctifs complets. Sa forme est donc :

(f 67) $$S_r(x,y) = \frac{11(x,y)}{P} = \frac{11(x,y)}{11(x,y)+00(x,y)+10(x,y)+01(x,y)}$$

Il varie bien de 0 à 1, il vaut:

- 0 si 11(x,y)=0
- 1 si simultanément : 10(x,y)=01(x,y)=00(x,y)=0 (ce qui est une contrainte extrêmement dure à obtenir)

Cet indice n'a pas de propriétés particulières en cas de disjonction complète , en effet en remplaçant 00(x,y), 10(x,y) et 01(x,y) par leurs valeurs en fonction de 11(x,y) et de P, du fait que 00(x,y)=11(x,y)+P-2m, 01(x,y)=10(x,y)=m-11(x,y), on retrouve : $S_r(x,y) = \frac{11(x,y)}{P}$

L'application de la règle de "Solomon Fortier" en cas de disjonction complète donne le résultat "paradoxal" suivant:

(f 68) $$S_r(x,y) \geq \frac{1}{2} \Rightarrow \frac{11(x,y)}{P} \geq \frac{1}{2} \Rightarrow 11(x,y) \geq \frac{P}{2}$$

Ce résultat est impossible à atteindre dans le contexte de "disjonction complète", sauf dans un cas . En effet, le minimum pour P par rapport à m est obtenu sur les tableaux de "présence-absence dédoublés" pour lesquels P=2m en remplaçant P par 2m dans l'inégalité précédente, on obtient:

$$11(x,y) \geq m$$

Comme, par définition $11(x,y) \leq m$, le seul cas possible est $11(x,y) = m$, pour toutes les autres valeurs de P: 3m, 4m, 5m...x.m, l'inégalité (f 68) est impossible à vérifier.
Ceci montre, encore une fois, que l'indice de Rao ne doit être utilisé que dans des configurations spéciales et en particulier, jamais sur des problèmes disjonctivés, hormis le cas mentionné.

4.4.3.2 Indice de Marcotorchino-Michaud (1980)

En 1980 (voir F. Marcotorchino et P. Michaud (1980)), nous avions introduit l'indice suivant :

(f 69) $$S_{mm}(x,y) = \frac{11(x,y) + \frac{1}{2}\,00(x,y)}{P} = \frac{11(x,y) + \frac{1}{2}\,00(x,y)}{11(x,y) + 00(x,y) + 10(x,y) + 01(x,y)}$$

Cet indice (pour lequel α=1/2) "neutralise" la trop forte influence négative des configurations 00(x,y) au dénominateur, vue dans le cas de l'indice de Rao, tout en leur donnant un poids moins positif toutefois que dans le "simple matching " index de Sokal et Michener. C'est un indice de compromis linéaire entre l'indice de Rao et l'indice de "Simple Matching", d'ailleurs, il s'écrit:

$$S_{mm}(x,y) = \frac{1}{2}\left[S_r(x,y) + S_{sm}(x,y)\right]$$

Comme l'indice de Rao, il varie de 0 à 1, il vaut:

- 0 si 11(x,y)=00(x,y)=0. Cependant on peut se poser la question, dans ce cas, de l'extrême rareté d'une telle situation qui correspond au fait que les profils de x et y sont totalement inverses l'un de l'autre, cas des situations de vecteurs de présence-absence à produit scalaire nul. Ceci dit, dans de grands tableaux cette situation peut néanmoins se produire, et de ce fait, la valeur 0 est rare, mais potentiellement atteinte.
- 1 si 10(x,y)=01(x,y)=0 et 00(x,y)=0, cependant dans le cas où 10(x,y)=01(x,y)=0, puisque dans ce cas P=00(x,y)+11(x,y), il vaut:

$$S_{mm}(x,y) = \frac{11(x,y) + \frac{1}{2}[P - 11(x,y)]}{P} = \frac{1}{2}\left[\frac{11(x,y) + P}{P}\right] = \frac{1}{2} + \frac{11(x,y)}{2P}$$

- le fait marquant ici est que si 10(x,y)= 01(x,y)=0, l'indice est toujours supérieur à ½ et il vaut 1 si en plus 00(x,y)=0, soit 11(x,y)=P . La remarque, faite pour le cas de l'obtention de la valeur 0, s'applique ici également, pour une situation inverse. Cette situation , ne peut se produire, en effet, que pour des tableaux de données de présence-absence, lorsque deux individus x et y possèdent l'ensemble complet de toutes les propriétés, en d'autres termes que leurs profils sont deux vecteurs composés de valeurs "1".

Il est à noter cependant, qu'en cas de disjonction complète, du fait des relations entre les différentes cases du tableau, déjà présentées auparavant, la valeur de l'indice se simplifie en :

$$S_{mm}(x, y) = \frac{\frac{3}{2}11(x, y) + \frac{1}{2}[P - 2m]}{P} = \frac{3}{2}\frac{11(x, y)}{P} + \frac{1}{2} - \frac{m}{P}$$

Comme dans ce cas, (disjonction complète), on ne peut avoir $11(x,y)$ et $00(x,y)$ simultanément nuls tous les deux, l'indice prend sa valeur minimale pour $11(x,y) = 0$ et vaut alors :

$$S_{mm}(x, y) = \frac{1}{2} - \frac{m}{P}$$

On voit qu'il ne vaudra 0 que si $P=2m$ (cas de la disjonction tableaux de présence-absence, où chaque variable ne peut avoir que 2 modalités), on se retrouve alors dans le cas , déjà discuté plus haut, de la valeur 0 de l'indice (vecteurs de présence-absence à produit scalaire nul).

Il prend sa valeur maximum dans le cas où $11(x,y)=m$,valeur pour laquelle il vaut :

$$S_{mm}(x, y) = \frac{1}{2} + \frac{m}{2P}$$

On voit qu'il ne prendra la valeur maximale 1 que lorsque $P=m$, c'est à dire dans le cas où l'on identifie un tableau de présence absence avec un tableau dédoublonné de présence-absence). Cependant cette situation n'est qu'un artefact, on parlera en fait de tableau disjonctif "vrai", dès lors qu'une variable possède au minimum 2 modalités. Si toutes les variables ont deux modalités , on obtient la valeur maximale qui vaut :

$$S_{mm}(x, y) = \frac{1}{2} + \frac{m}{4m} = \frac{3}{4}$$

Propriété de l' indice de Marcotorchino-Michaud en cas de disjonction complète

En cas de disjonction complète, la règle de Solomon-Fortier est applicable ici à une borne ($\frac{MinS + MaxS}{2}$), (voir §1.1), puisque l'indice varie « potentiellement » de 0 à 1, mais plus exactement selon des bornes dépendant du rapport m/P. Ici, compte tenu des bornes précédentes, la borne de Solomon-Fortier associée s'écrit :

$$\frac{MinS + MaxS}{2} = \frac{1}{2}\left(\left(\frac{1}{2} - \frac{m}{P}\right) + \left(\frac{1}{2} + \frac{m}{2P}\right)\right) = \frac{1}{2} - \frac{m}{4P}$$

La vérification de la borne de Solomon-Fortier au sens particulier du contexte disjonctif, induit donc ici :

$$S_{mm}(x, y) \geq \frac{1}{2} - \frac{m}{4P},$$

soit:

$$\frac{3}{2}\frac{11(x, y)}{P} + \frac{1}{2} - \frac{m}{P} \geq \frac{1}{2} - \frac{m}{4P} \quad \Rightarrow \quad \frac{3}{2}11(x, y) \geq \frac{3}{4}m \quad \Rightarrow \quad 11(x, y) \geq \frac{m}{2}$$

Ce résultat est intéressant au titre qu'il prouve que $S_{mm}(x,y)$ est l'un des deux seuls, parmi les indices du Groupe II–Type I, pour lequel, en cas de disjonction complète, l'application de la règle de Solomon-Fortier fait disparaître la référence à la valeur de P et où l'application du principe fait réapparaître la règle majoritaire de Condorcet.

Ce résultat montre également que, quand l'on parle de borne de Salomon-Fortier, c'est bien à la quantité $\frac{\text{Min S+Max S}}{2}$ que l'on doit faire référence et non simplement à la valeur ½, qui suppose, elle, une variation réelle de 0 à 1 (bornes atteintes) .

Une autre approche, nous montre bien la différence entre la borne vraie de Salomon Fortier et la borne théorique Si l'on fait apparaître le cas disjonctif en fin de calculs, on illustre clairement ce point.
En effet si l'on reprend la formule (f 69), et, puisque dans le cas général, sans référence aux limitations du cas disjonctif , on a vu que l'indice S_{mm} varie de 0 à 1, on peut donc lui appliquer la borne simple à ½. Ceci nous donne :

$$\frac{11(x,y)+\frac{1}{2}00(x,y)}{11(x,y)+00(x,y)+10(x,y)+01(x,y)} \geq \frac{1}{2} \quad \Rightarrow \quad \frac{11(x,y)}{2} \geq \frac{10(x,y)+01(x,y)}{2}$$

On voit déjà à ce stade que la référence à 00(x,y) disparaît, cette formule est générale. Si maintenant on revient dans un processus où il y a eu disjonction, la quantité 10(x,y)+01(x,y) est égale à 2(m-11(x,y)), soit en remplaçant cette valeur dans l'inégalité précédente :

$$\frac{11(x,\ y)}{2} \geq \frac{2[m\ -11(x,\ y)]}{2} \quad \Rightarrow \quad 11(x,\ y) \geq \frac{2m}{3}$$

On retrouve bien ici une majorité Condorcéenne des 2/3, que nous avions obtenue précédemment dans le cas de l'Indice de Jaccard, l'un des indices représentant le Groupe I – Type I .

On constate donc que, contrairement aux indices de Sokal et Michener, Sokal et Sneath et Rogers et Tanimoto, la valeur P **disparaît** de la borne sur 11(x,y) d'une part et que l'indice impose une « majorité simple » des variables pour la valeur des « matchings » 11(x,y) (cas disjonctif particulier), d'autre part une règle des 2/3 dans le cas général. Cette dernière propriété fait de cet indice, une mesure tout à fait intéressante dans le cas de structures disjonctivées, ou de présence –absence avec beaucoup de valeurs nulles, qui ,quoique étant différente, se comporte comme l'indice de Jaccard (qui, lui, appartient au Groupe I-TypeI). Cette légère dépondération au numérateur des configurations 00(x,y), redresse quelque peu le défaut inhérent aux indices du Groupe II - Type I, à savoir : ne pas être adaptés aux traitements des données disjonctives ou à valeurs nulles très nombreuses.

4.4.3.3 Indice de Marcotorchino-Michaud-2 (1980)

S'appuyant sur le même principe que précédemment, un autre indice du même genre que le précédent donne des résultats également intéressants, il s'agit de l'indice suivant :

(f 70)
$$S_{mm2}(x,y) = \frac{11(x,y)+\frac{1}{2}00(x,y)}{11(x,y)+00(x,y)+\frac{1}{2}[10(x,y)+01(x,y)]}$$

Comme le précédent cet indice varie de 0 à 1, il vaut:

- 0 si 11(x,y)=00(x,y)=0. Cependant on peut lui appliquer la même remarque que dans le cas précédent , a savoir l'extrême rareté d'une telle situation qui correspond au fait que les profils de x et y sont totalement inverses l'un de l'autre.

- 1 si 10(x,y)=01(x,y)=0 et 00(x,y)=0, La remarque, faite pour le cas de l'obtention de la valeur 0, s'applique ici également, pour une situation inverse. Cette situation , ne peut se produire, en effet, que pour des tableaux de données de présence-absence, lorsque deux individus x et y possèdent l'ensemble complet de toutes les propriétés.

Il est à noter cependant, comme dans le cas de l'indice précédent, qu'en cas de disjonction complète, du fait des relations entre les différentes cases du tableau, déjà présentées auparavant, la valeur de l'indice se simplifie en :

$$S_{mm2}(x,y) = \frac{3\,11(x,y) + P - 2m}{2\,11(x,y) + 2P - 2m}$$

Comme dans ce cas, (disjonction complète), on ne peut avoir 11(x,y) et 00(x,y) simultanément nuls, puisque 11(x,y)+00(x,y)=P, l'indice prend sa valeur minimale pour 11(x,y) = 0 et vaut alors :

$$S_{mm2}(x,y) = \frac{P - 2m}{2P - 2m}$$

On voit qu'il ne vaudra 0 que si P=2m (cas de la disjonction tableaux de présence-absence, où chaque variable ne peut avoir que 2 modalités), on se retrouve alors dans le cas , déjà discuté pour l'indice $S_{mm}(x,y)$.

Il prend sa valeur maximum dans le cas où 11(x,y)=m ,valeur pour laquelle il vaut :

$$S_{mm2}(x,y) = \frac{P+m}{2P} = \frac{1}{2} + \frac{m}{2P}$$

De façon surprenante, cette valeur est identique à celle trouvée pour l'indice précédent. On voit qu'il ne prendra la valeur maximale 1 que lorsque P=m, c'est à dire dans la même situation que $S_{mm}(x,y)$. La remarque faite pour $S_{mm}(x,y)$ s'applique ici aussi . Le maximum "vrai" (par opposition au "Maximum théorique") sera obtenu si les variables ont toutes deux modalités, ce qui permet d'obtenir une la valeur maximale qui vaut :

$$S_{mm2}(x,y) = \frac{1}{2} + \frac{m}{4m} = \frac{3}{4}$$

La borne « générale de Salomon Fortier » s'écrit dans ce cas :

$$\frac{MinS + MaxS}{2} = \frac{1}{2}\left(\left(\frac{P - 2m}{2(P - m)}\right) + \left(\frac{P+m}{2P}\right)\right) = \frac{1}{2}\frac{(P-2m)P + (P-m)(P+m)}{2(P-m)P}$$

Si l'on suppose que $P=\bar{p}m$, il vient en remplaçant P par cette valeur:

$$\frac{MinS + MaxS}{2} = \frac{1}{2}\left[1 - \frac{1}{2\bar{p}(\bar{p}-1)}\right]$$

L'application de la règle de Salomon Fortier associée au cas de cet indice en environnement disjonctif, nous donne, après remplacement de P par $m\bar{p}$:

$$\frac{3\,11(x,y) + m\bar{p} - 2m}{2\,11(x,y) + 2m\bar{p} - 2m} \geq \frac{1}{2}\left[1 - \frac{1}{2\bar{p}(\bar{p}-1)}\right] \Rightarrow 11(x,y) \geq \frac{m}{2}\left[\frac{2\bar{p}-2}{2\bar{p}-1}\right]$$

Contrairement au cas précédent, qui donnait une règle majoritaire indépendante de $\bar{p}$. La règle associée ici est dépendante de $\bar{p}$, mais tend assez vite vers une règle de la majorité Condorcéenne $\frac{m}{2}$, dès lors que $\bar{p}$ est suffisamment grand.

Propriété de l' indice de Marcotorchino-Michaud-2 en cas de disjonction complète théorique

En cas de disjonction complète, la règle de Solomon - Fortier , si l'on revient aux bornes max et min **« théoriques »** de l'indice, induit l'inégalité suivante, où l'on a remplacé ½(10(x,y)+01(x,y)) par sa valeur en fonction de 11(x,y) à savoir m-11(x,y):

$$S_{mm2}(x,y) \geq \frac{1}{2} \;\Rightarrow\; \frac{11(x,y)+\frac{1}{2}00(x,y)}{11(x,y)+00(x,y)+[m-11(x,y)]} \geq \frac{1}{2} \;\Rightarrow\; 11(x,y)+\frac{1}{2}00(x,y) \geq \frac{m}{2}+\frac{1}{2}00(x,y) \;\Rightarrow\; 11(x,y) \geq \frac{m}{2}$$

On constate, dans ce cas, que l'on fait encore disparaître les configurations 00(x,y) de chaque côté de l'inégalité de la règle de Solomon Fortier. Dans ce cas particulier l'indice qui a été doublement rééquilibré , aboutit cette fois ci sur une borne pour 11(x,y), qui n'est rien d'autre que la règle de la « majorité simple ». Cette dernière propriété en fait un indice encore plus intéressant que le précédent dans le cas de structures disjonctivées ou de tableaux de présence-absence à grands nombre de valeurs nulles, il se comporte comme l'indice de Dice , quoique étant d'un Groupe différent.

4.4.4 Indices du GROUPE II, Type II (indices obtenus comme ratios de fonctions non linéaires des quantités du tableau « Tetrachorique »)

4.4.4.1 Indice lié au Coefficient de corrélation Tetrachorique

Cet indice calcule le coefficient de corrélation à partir du tableau Tetrachorique (il est également connu sous le nom de coefficient de Bravais-Pearson) , tel qu'introduit au paragraphe §.3 (page 19), il est défini par :

(f71)
$$S_T(x,y) = \frac{[11(x,y).00(x,y) - 10(x,y).01(x,y)]}{\sqrt{[11(x,y)+10(x,y)][11(x,y)+01(x,y)].[00(x,y)+10(x,y)][00(x,y)+01(x,y)]}}$$

Contrairement aux autres indices, rencontrés précédemment, cet indice varie de -1 à +1, comme tout coefficient de corrélation. En particulier :

- Il vaut +1 si 10(x,y) et 01(x,y)=0
- Il vaut –1 si 11(x,y) et 00(x,y)=0

Comportement asymptotique de ce coefficient pour des situations particulières

Dans le cas où l'on se sert de ce coefficient ou indice de similarité sur des processus de « matching » de listes comme dans le cas de processus de recherche sur Internet (voir § 4.2.), la quantité 10(x,y) peut être considérée comme un facteur de « bruit », inversement la quantité 01(x,y) peut être considérée comme du « silence », on « rate » par x de l'existant sur y, c'est la raison du mot « silence ».Mais qu'en est-il de la quantité 00(x,y) ?. Dans le contexte décrit ci dessus, la quantité 00(x,y) représente ce qui n'est pas lié à l'univers d'un questionnement, c'est à dire « *tout le reste* » hors de la recherche représentée par x ou de

l'univers qu'on veut atteindre représenté par y. De ce fait, il arrive qu'on ne soit pas capable de connaître la taille de 00(x,y), qui par ailleurs est forcément très très grande par rapport à 11(x,y), 01(x,y) ou 10(x,y) puisqu'il s'agit du reste de l'univers. Les quantités 10(x,y) et 01(x,y) sont négligeables par rapport à 00(x,y) et dès lors le coefficient $S_T(x,y)$ se simplifie suivant la formule :

(f 72) $$S_T(x,y) \cong \frac{[11(x,y).00(x,y)]}{\sqrt{[11(x,y)+10(x,y)][11(x,y)+01(x,y)].[00(x,y)]^2}}$$

Cette simplification permet d'éliminer 00(x,y), à la fois au numérateur et au dénominateur, ce qui aboutit à :

(f 73) $$S_T(x,y) \cong \frac{11(x,y)}{\sqrt{[11(x,y)+10(x,y)][11(x,y)+01(x,y)]}} = \sqrt{\frac{11(x,y)}{[11(x,y)+10(x,y)]}}\sqrt{\frac{11(x,y)}{11(x,y)+01(x,y)}} = \sqrt{P.R.}$$

On retrouve ainsi la moyenne géométrique des indices de Rappel et de Précision, en d'autres termes l'Indice d'Ochiai.

<u>Propriété n°17</u>: dans le cas où 00(x,y) est grand par rapport à 10(x,y) et 01(x,y), l'indice de similarité fondé sur le coefficient de corrélation tetrachorique a un comportement très voisin de celui de l'indice d'Ochiai et de fait varie de 0 à 1 et non de –1 à +1:

$$S_T(x,y) \cong S_o(x,y)$$

<u>Comportement de l'indice $S_T(x,y)$ en cas de Disjonction Complète</u>

En cas de disjonction complète, comme nous l'avons vu au § 4.1.2 (Propriété n°3), nous avons à notre disposition les formules d'équivalence suivantes : :

$$11(x,y)+\frac{1}{2}[10(x,y)+01(x,y)]=m$$
$$11(x,y)+00(x,y)+10(x,y)+01(x,y)=P$$
$$00(x,y)=P+11(x,y)-2m$$

Nous rajoutons en plus, une relation évidente, que nous n'avions pas exploitée précédemment et qui va nous servir ici :

10(x,y)=01(x,y) ce qui implique que : $10(x,y)=01(x,y)=m-11(x,y)$

L'indice « Tetrachorique » se reformule alors selon l'expression suivante :

(f 74) $$S_T(x,y)=\frac{11(x,y).[P+11(x,y)-2m]-[m-11(x,y)]^2}{\sqrt{[m][m].[P-m][P-m]}}=\frac{11(x,y).P-m^2}{m.(P-m)}$$

Sous cette nouvelle expression, on voit que l'indice vaut 1 si 11(x,y)=m, et vaut :

$-\frac{m}{(P-m)}$ si 11(x,y)=0

Il n'a donc plus un intervalle de variation symétrique puisque : $-\frac{m}{(P-m)} \le S_T(x,y) \le 1$

Par ailleurs si l'on étalonne son comportement vis à vis de la « Borne de Solomon et Fortier », on obtient un résultat qui prouve que cet indice, outre son comportement lié asymptotiquement à l'indice d'Ochiai , est un « bon » indice de similarité.

En effet comme $-\frac{m}{(P-m)} \le S_T(x,y) \le 1$, le véritable milieu de l'intervalle de variation de l'indice $S_T(x,y)$, n'est plus ½ comme pour les indices rencontrés précédemment, mais le milieu de l'intervalle: $\left[-\frac{m}{P-m},\ 1\right]$

C'est à dire que la quantité: $1/2\left[1-\frac{m}{P-m}\right]$, soit: $\frac{P-2m}{2(P-m)}$, devient la "borne de Solomon- Fortier" associée. Dès lors on voit que la règle devient:

$$S_T(x,y) \ge \frac{.P-2m}{2.(P-m)} \Rightarrow \frac{11(x,y).P-m^2}{m(P-m)} \ge \frac{P-2.m}{2(P-m)} \Rightarrow 11(x,y) \ge \frac{m}{2},$$

On voit que dans ce contexte, le critère se comporte exactement comme l'indice de Dice au voisinage du milieu du segment de variation et l'on retrouve la "règle de la Majorité de Condorcet" .

4.4.4.2 Indices Dérivés, variantes du Coefficient Tétrachorique

Un nombre important d'indices ont été définis par de nombreux auteurs, en tenant compte du même numérateur que l'indice Tetrachorique, mais en faisant varier son dénominateur. Dans cette famille citons en quelques uns :

a) Indice de Maxwell et Pilliner (1968)

Ce coefficient qui varie de –1 à +1 est défini par :

$$S_{MP}(x,y) = \frac{2[11(x,y).00(x,y) - 10(x,y).01(x,y)]}{[11(x,x).00(x,x)] + [11(y,y).00(y,y)]} = \frac{2[11(x,y).00(x,y) - 10(x,y).01(x,y)]}{[11(x,y)+10(x,y)][00(x,y)+01(x,y)] + [11(x,y)+01(x,y)][00(x,y)+10(x,y)]}$$

On constate que si 10(x,y)=01(x,y) , alors:

$$S_{MP}(x,y) = \frac{2[11(x,y).00(x,y) - (10(x,y))^2.]}{2[11(x,y) + .10(x,y)][[00(x,y)+10(x,y)]} = S_T(x,y)$$

c'est exactement ce que vaut le coefficient Tétrachorique dans ce cas. En particulier en cas de disjonction complète , on a:

$$11(x,y) + \frac{1}{2}[10(x,y) + 01(x,y)] = m$$

$$10(x,y)=01(x,y)$$

$$11(x,y) + 00(x,y) + 10(x,y) + 01(x,y) = P$$

$$00(x,y) = P + 11(x,y) - 2m$$

En remplaçant ces quantités dans le coefficient de Maxwell et Pilliner, on obtient sa valeur en cas de disjonction complète, soit:

$S_{MP}(x,y) = \frac{11(x,y).P-m^2}{m.(P-m)}$, c'est exactement la valeur qu'on avait obtenue pour le coefficient de similarité Tetrachorique, voir formule (f74), ce qui veut dire que le coefficient de Maxwell et Pilliner a le même comportement que le coefficient Tetrachorique vis à vis de la borne de Solomon et Fortier.

Par ailleurs comme l'a montré M.J. Warrens (2008), en revenant sur le positionnement des Moyennes: Harmonique, Géométrique et Arithmétique (comme nous l'avons fait au § 4.2.1.3), en utilisant le recodage "astucieux" suivant :

$$u = \frac{11(x,y)00(x,y) - 10(x,y)01(x,y)}{[11(x,y)+10(x,y)][00(x,y)+01(x,y)]} \text{ et } v = \frac{11(x,y).00(x,y) - 10(x,y)01(x,y)}{[11(x,y)+01(x,y)][00(x,y)+10(x,y)]}$$

on peut exprimer les indices de Maxwell et Pilliner et Tétrachorique comme respectivement les moyennes : Harmonique des quantités u et v et Géométrique des quantités u et v sous la forme:

$$S_{MP}(x,y) = \frac{2u.v}{(u+v)} \text{ et } S_T(x,y) = \sqrt{u.v}$$

Comme ces indices varient de –1 à +1 ceci implique la propriété suivante au niveau des valeurs absolues des coefficients:

$$0 \le |S_{MP}(x,y)| \le |S_T(x,y)| \le 1$$

b) Indice de Fleiss (1975)

Dans le même ordre d'idée, la moyenne arithmétique de "u" et "v" existe, il s'agit d'un coefficient fort peu connu et signalé également par M.J. Warrens dans M.J. Warrens (2008), sous le nom de coefficient de J.L. Fleiss (1975), qui s'écrit:

$$S_F(x,y) = \frac{1}{2}(u+v)$$

$$S_F(x,y) = \frac{1}{2}[11(x,y)00(x,y) - 10(x,y)01(x,y)][\frac{1}{[11(x,y)+10(x,y)][00(x,y)+01(x,y)]} + \frac{1}{[11(x,y)+01(x,y)][00(x,y)+10(x,y)]}]$$

Là encore, si 10(x,y)=01(x,y) , le coefficient de Fleiss est égal au Coefficient Tetrachorique (c'est ce qui se produit en cas de disjonction complète).
D'autre part comme : Moyenne géométrique (u,v) ≤ Moyenne Arithmétique de (u,v), on obtient:

$$0 \le |S_{MP}(x,y)| \le |S_T(x,y)| \le |S_F(x,y)| \le 1$$

Dans cette famille d'indices dérivés de l'indice Tetrachorique, il apparaît clairement que ce dernier du fait de sa position de « médiane » des deux autres aura la préférence des utilisateurs, d'autre part nous l'avons vu, il joue un rôle de « vrai » coefficient de corrélation.

c) Indices déduits de l'Inégalité « au Determinant » de J. Hadamard

Ces indices de similarité ne sont pas, à proprement parler , des indices découverts par Jacques Hadamard (1865-1963), mais parmi les multiples travaux du grand mathématicien, une inégalité célèbre qui porte son nom va nous servir, ici, pour donner l'expression d'un indice qui peut être considéré comme un dérivé de l'Indice Tétrachorique.
En effet l'inégalité d'Hadamard s'établit comme suit :

Théorème n°2 :Soit M une matrice carrée dont les vecteurs colonnes sont { $V_1, V_2, \ldots V_n$ }, on note $\|V_k\|_2$, la norme euclidienne du vecteur colonne V_k : $\|V_k\|_2 = \sqrt{\sum_{s=1}^{n} v_{ks}^2}$, alors le théorème dit de « l'inégalité d'Hadamard » stipule que l'inégalité suivante est vraie :

$$|\det[M]| \le \|V_1\|_2 \|V_2\|_2 \ldots \|V_n\|_2$$

Dans le cas où aucun des Vk n'est nul, on a l'égalité si et seulement si les vecteurs Vk sont orthogonaux deux à deux.

Considérons alors le tableau tétrachorique suivant :

	y = 1	y = 0
x = 1	11	10
x = 0	01	00

Posons :

M =

11(x,y)	10(x,y)
01(x,y)	00(x,y)

De ce fait les vecteurs colonnes associés sont V_1={11(x,y), 01(x,y)} et V2={10(x,y), 00(x,y)} et l'application du Théorème N°2 se fait trivialement en se rappelant que pour cette matrice 2x2 , le determinant vaut : $\det[M] = [11(x,y).00(x,y) - 10(x,y)01(x,y)]$

L'inégalité de Hadamard implique donc :

$$\left|11(x,y).00(x,y) - 10(x,y)01(x,y)\right| \le \sqrt{[11(x,y)]^2 + [01(x,y)]^2}\sqrt{[10(x,y)]^2 + [00(x,y)]^2}$$

Dès lors la quantité suivante dénommée : $S_{Had}(x,y)$, c'est à dire:

$$S_{Had}(x,y) = \frac{11(x,y).00(x,y) - 10(x,y)01(x,y)}{\sqrt{[11(x,y)]^2 + [01(x,y)]^2}\sqrt{[00(x,y)]^2 + [10(x,y)]^2}}$$

varie de –1 à +1 , c'est donc bien un indice de similarité dérivé de l'Indice Tétrachorique. Cet indice vaut 1 si 01(x,y)=10(x,y)=0 et vaut –1 si 11(x,y)=00(x,y)=0 (c'est à dire lorsque les vecteurs colonnes V_1 et V_2 sont orthogonaux, comme il en est fait mention dans l'intitulé du Théorème N°2. D'autre part, comme on pouvait intervertir les roles de x et y dans la définition de la Matrice M, on aurait tout aussi bien pu obtenir l'indice suivant à partir de l'inégalité de Hadamard:

$$S_{Had'}(x,y) = \frac{11(x,y).00(x,y) - 10(x,y)01(x,y)}{\sqrt{[11(x,y)]^2 + [10(x,y)]^2}\sqrt{[00(x,y)]^2 + [01(x,y)]^2}}$$

Les bornes 1 et –1 étant atteintes dans les mêmes configurations que précédemment.

Ainsi lors de la même façon que dans le cas des indices de Maxwell et Pilliner ou Fleiss , en posant: $u = S_{Had}(x,y)$ et $v = S_{Had'}(x,y)$

On peut définir la famille des indices (symétrisés) dérivés de Hadamard suivants:

$$S_{HAD1}(x,y) = \frac{2u.v}{(u+v)} \quad S_{HAD2}(x,y) = \sqrt{u.v} \qquad \text{et} \quad S_{HAD3}(x,y) = \tfrac{1}{2}(u+v)$$

Ainsi par exemple on a:

$$S_{HAD\ 2}(x,y) = \frac{11(x,y).00(x,y) - 10(x,y)01(x,y)}{\sqrt{\sqrt{[11(x,y)]^2 + [10(x,y)]^2}\sqrt{[00(x,y)]^2 + [01(x,y)]^2}\sqrt{[11(x,y)]^2 + [01(x,y)]^2}\sqrt{[00(x,y)]^2 + [10(x,y)]^2}}}$$

Comme pour tout couple de nombres positifs {a,b}: on l'inégalité $(a+b) \geq \sqrt{a^2+b^2}$, il apparaît de façon évidente que le dénominateur de $S_T(x,y)$ est supérieur à celui de $S_{HAD2}(x,y)$ et donc que l'on a la les inégalités suivante:

$$0 \leq \left|S_T(x,y)\right| \leq \left|S_{HAD\ 2}(x,y)\right| \leq 1$$

En jouant ensuite sur l'ordonnance des moyennes pour les indices de Hadamard , couplée aux résultats obtenus précédemment en **b)** il vient :

$$0 \leq \left|S_{MP}(x,y)\right| \leq \left|S_T(x,y)\right| \leq \left|S_F(x,y)\right| \leq \left|S_{HAD2}(x,y)\right| \leq \left|S_{HAD3}(x,y)\right| \leq 1$$

4.4.4.3 Les Indices Y et Q de Yule

Les indices de similarité Y et Q de Yule dont une première version partielle est donnée dans le livre de Yule et Kendall (voir G. Yule et M.G. Kendall (1950)), et réétudiée par la suite par G. Yule, seul, se calculent également à partir du tableau de contingence Tetrachorique, puisqu'ils font jouer un rôle aux quatre quantités principales en jeu. Sur le principe, l'indice Y est le plus connu des deux, on parle à son sujet de coefficient de « colligation », en effet il a été construit à partir des notions d' « odd ratios », souvent utilisés en statistique des contingences par les anglo-saxons. Il est défini par :

(f 75) $$S_Y(x,y) = \frac{\sqrt{11(x,y).00(x,y)} - \sqrt{10(x,y).01(x,y)}}{\sqrt{11(x,y).00(x,y)} + \sqrt{10(x,y).01(x,y)}}$$

Comme on le voit immédiatement cet indice varie de -1 à +1, il vaut :

- 1 si 10(x,y) ou 01(x,y)=0
- -1 si 11(x,y) ou 00 (x,y) =0

Contrairement au cas du Coefficient $S_T(x,y)$, il est a noter la présence du « ou » et non du « et » pour l'obtention des cas +1 et –1 de l'indice, ceci, entre autre, permet de différencier les deux indices, qui par ailleurs ont des comportements assez voisins.
Contrairement au cas de l'indice $S_T(x,y)$, il est à noter, également ici, que cet indice est peu (voire pas du tout) adapté aux situations où 00(x,y) est très grand, en effet il est pratiquement voisin de 1 dans ces cas là et donc peu discriminant.

L'indice Q de Yule est une variante de l'indice précédent, en effet il s'écrit :

(f 76) $$S_Q(x,y) = \frac{11(x,y).00(x,y) - 01(x,y).10(x,y)}{11(x,y).00(x,y) + 10(x,y).01(x,y)}$$

Il varie également de –1 à +1 et vaut :

- 1 si 10(x,y) ou 01(x,y)=0
- -1 si 11(x,y) ou 00 (x,y) =0

Si l'on compare l'indice Q de Yule et l'indice Tetrachorique, il est évident que l'indice Q de Yule a une valeur qui sera toujours supérieure à l'indice de l'indice Tetrachorique . En effet on a à comparer :

$$S_T(x,y) = \frac{[11(x,y).00(x,y) - 10(x,y).01(x,y)]}{\sqrt{[11(x,y)+10(x,y)][11(x,y)+01(x,y)].[00(x,y)+10(x,y)][00(x,y)+01(x,y)]}} \text{ et}$$

$$S_Q(x,y) = \frac{11(x,y).00(x,y) - 01(x,y).10(x,y)}{11(x,y).00(x,y) + 10(x,y).01(x,y)}$$

Ces deux indices ont la même valeur du numérateur et en revanche ils ont un dénominateur différent .

Or si l'on pose $A = 11(x,y).00(x,y) + 10(x,y).01(x,y)$ et si l'on élève le dénominateur de l'indice $S_T(x,y)$ au carré , on voit que le dénominateur concerné s'écrit:

(A+11(x,y)10(x,y)+00(x,y)01(x,y)) (A+11(x,y)01(x,y)+00(x,y)10(x,y))=A^2+ Δ (où $\Delta \geq 0$)

Or le carré du dénominateur de l'indice Q de Yule est justement égal à A^2 , de ce fait on a:

Dénominateur de l'indice Q de Yule ≤ Dénominateur de l'indice Tétrachorique

Ceci implique donc la relation:

$$|S_T(x,y)| \leq |S_Q(x,y)| \quad \forall \ x,y$$

(car les deux indices peuvent être positifs ou négatifs)

Par ailleurs, il serait intéressant de voir comment varient $S_Y(x,y)$ et $S_Q(x,y)$, l'un en fonction de l'autre, puisqu'ils utilisent tous les deux les mêmes quantités. Pour plus de simplicité, on va travailler sur leurs variantes transformant leur formalisme d'indices de « corrélation » variant de –1 à +1 en indices de similarité variant de 0 à 1 , pour ce faire , il suffit de reprendre la note de bas de page n°1 et calculer :

$$S'_Y(x,y) = \frac{1}{2}[S_Y(x,y)+1] \quad \text{et} \quad S'_Q(x,y) = \frac{1}{2}[S_Q(x,y)+1] \quad \forall \ x,y$$

Les deux nouveaux indices s'écrivent alors :

$$S'_Q(x,y) = \frac{11(x,y).00(x,y))}{11(x,y).00(x,y)+10(x,y).01(x,y)} \quad \text{et} \quad S'_Y(x,y) = \frac{\sqrt{11(x,y).00(x,y)}}{\sqrt{11(x,y).00(x,y)}+\sqrt{10(x,y).01(x,y)}}$$

Dire que :

$$S'_Q(x,y) \geq S'_Y(x,y) \Rightarrow \sqrt{11(x,y).00(x,y)} \geq \sqrt{10(x,y).01(x,y)} \Rightarrow 11(x,y).00(x,y) \geq 10(x,y).01(x,y)$$

ceci implique que :

$$S'_Q(x,y) \geq S'_Y(x,y) \geq \frac{1}{2},$$

dans le cas inverse l'inégalité se produira dans le sens contraire et l'on aura :

$$S'_Q(x,y) \leq S'_Y(x,y) \leq \frac{1}{2}$$

Comportement des indices $S_Y(x,y)$ et $S_Q(x,y)$ en cas de Disjonction Complète

En utilisant la série de formules déjà mentionnée au § précédent , on peut simplifier l'expression de ces deux indices dans le cas de disjonction complète et l'on obtient pour l'indice $S_Q(x,y)$, (le seul à pouvoir donner des simplifications interprétables puisque ne faisant pas appel à des racines carrées) :

(f 77) $$S_Q(x,y) = \frac{1\mathbb{1}(x,y).[P+1\mathbb{1}(x,y)-2m]-[m-1\mathbb{1}(x,y)]^2}{1\mathbb{1}(x,y).[P+1\mathbb{1}(x,y)-2m]+[m-1\mathbb{1}(x,y)]^2} = \frac{P.1\mathbb{1}(x,y)-m^2}{(P.1\mathbb{1}(x,y)-m^2)+2[m-1\mathbb{1}(x,y)]^2}$$

Sous cette dernière forme on voit que l'indice vaut : 1 si $1\mathbb{1}(x,y) = m$ et –1 si $1\mathbb{1}(x,y)=0$

Puisque l'intervalle de variation de $S_Q(x,y)$ est ici (-1,+1) le milieu de l'intervalle de variation est égal à 0,. et de ce fait la borne de Solomon- Fortier pour $S_Q(x,y)$ est donnée par : $S_Q(x,y)\geq 0$, ce qui implique :

(f78) $$S_Q(x,y) = \frac{P.1\mathbb{1}(x,y)-m^2}{(P.1\mathbb{1}(x,y)-m^2)+2[m-1\mathbb{1}(x,y)]^2} \geq 0 \;\Rightarrow\; P.1\mathbb{1}(x,y)-m^2 \geq 0 \;\Rightarrow\; 1\mathbb{1}(x,y) \geq \frac{m^2}{P}$$

Montrons que cette borne pour $1\mathbb{1}(x,y)$, bien que non aberrante lorsque le nombre moyen de modalités par variable n'est pas trop fort, le devient si ce nombre augmente . En effet nous savons que $P = m\bar{p}$, où $\bar{p}$ est le nombre moyen de modalités par variable, de ce fait la borne donnée dans l'équation (f 78) devient:

$$1\mathbb{1}(x,y) \geq \frac{m^2}{P} = \frac{m^2}{m\bar{p}} = \frac{m}{\bar{p}}$$

Pour $\bar{p}=2$ on retrouve la règle majoritaire simple ou de Condorcet , pour $\bar{p}=3$, on obtient une règle au tiers etc., en fait on obtient une règle inversement proportionnelle au nombre de modalités moyen.

En d'autre termes cet indice n'aura aucun pouvoir discriminant dès lors que le nombre moyen de modalités est fort (si par exemple on travaille sur m=10 variables, chacune ayant 10 modalités, on voit qu'il suffit que $1\mathbb{1}(x,y)=1$ pour que l'indice atteigne la borne de similarité de Solomon-Fortier).
Le lecteur pourra vérifier facilement par lui-même que la borne sur $1\mathbb{1}(x,y)$ associée à la règle de Solomon-Fortier est exactement la même pour l'indice $S_Y(x,y)$ que celle associée à l'indice $S_Q(x,y)$ et donnée par la formule (f 78).

4.4.4.4 L'indice de Moyenne Arithmétique des « quatre ratios » d'Anderberg (1973)

Cet indice qui est explicitement défini dans le livre de Michael Anderberg (1973), page 91, a été introduit par ce dernier comme une généralisation de la moyenne arithmétique des indices de « Rappel » et « Précision » de Kulczynski à l'ensemble des ratios probabilistes possibles, issus du tableau Tetrachorique. Certain l'attribue d'ailleurs à Sokal et Sneath (1963). Il se définit de la façon suivante :

(f79) $$S_{R4}(x,y) = \frac{1}{4}\left[\frac{11(x,y)}{11(x,y)+10(x,y)} + \frac{11(x,y)}{11(x,y)+01(x,y)} + \frac{00(x,y)}{00(x,y+01(x,y)} + \frac{00(x,y)}{00(x,y)+10(x,y)}\right]$$

Cet indice varie bien de 0 à 1, il vaut:

- 1 si 10(x,y)=01(x,y)=0
- 0 si 11(x,y)=00(x,y)=0

En s'inspirant de la propriété du passage aux moyennes harmoniques des quantités 11(x,x), 11(y,y), 00(x,x), 00(y,y) , déjà vue au sujet de l'indice de Kulczynski (voir (f 35)) on a une autre expression posssible et plus compacte de cet indice :

(f 80) $$S_{R4}(x,y) = \frac{1}{2}\left[\frac{11(x,y)}{MH\ [11(x,x),11(y,y)]} + \frac{00(x,y)}{MH\ [00(x,x),00(y,y)]}\right]$$

Le développement de la formule (f 79) par réduction au même dénominateur est possible mais entrainerait des calculs fastidieux, le dénominateur serait identique par ailleurs au carré de celui de la formule (f 71) concernant le coefficient de corrélation Tetrachorique.

Comportement de l'indice $S_{R4}(x,y)$ en cas de Disjonction Complète

Comme nous l'avons fait précédemment, un bon moyen de comprendre un peu plus avant comment se comporte cet indice revient à le calculer dans le contexte particulier de la situation de disjonction complète. En effet dans ce cas la formule globale se simplifie et un certain nombre de propriétés caractéristiques peuvent être mises en évidence. En effet en cas de disjonction complète l'indice $S_{R4}(x,y)$ se simplifie selon la formulation suivante :

$$S_{R4}(x,y) = \frac{1}{4}\left[\frac{11(x,y)}{m} + \frac{11(x,y)}{m} + \frac{P+11(x,y)-2m}{P-m} + \frac{P+11(x,y)-2m}{P-m}\right] = \frac{1}{2}\left[\frac{11(x,y)}{m} + \frac{P+11(x,y)-2m}{P-m}\right]$$

ce qui, après réduction au même dénominateur et simplifications au numérateur, nous donne:

(f 81) $$S_{R4}(x,y) = \frac{1}{2}\left[\frac{11(x,y)}{m} + \frac{P+11(x,y)-2m}{P-m}\right] = \frac{11(x,y).P-m^2+m(P-m)}{2.m(P-m)} = \frac{1}{2} + \frac{11(x,y).P-m^2}{2.m(P-m)}$$

Sous cette dernière forme on voit que l'indice vaut 1 si 11(x,y)=m (sa valeur maximum) et vaut:

$\frac{1}{2} - \frac{m}{2.(P-m)}$, soit $\frac{P-2.m}{2.(P-m)}$ si 11(x,y)=0, son intervalle de variation est donc:

$\frac{P-2.m}{2.(P-m)} \le S_{R4}(x,y) \le 1$, et le milieu du segment associé est donné par:

$$\frac{1}{2}\left[1+\frac{P-2.m}{2.(P-m)}\right] = \frac{3P-4m}{4(P-m)} = 1-\frac{P}{4(P-m)}$$

Si l'on applique la règle de Solomon-Fortier sur cette valeur précédente comme étalonnage de l' indice, on obtient:

(f 82) $$S_{R4}(x,y) \ge 1-\frac{P}{4(P-m)} \quad \Rightarrow \quad \frac{1}{2}+\frac{11(x,y)P-m^2}{2m(P-m)} \ge 1-\frac{P}{4(P-m)} \quad \Rightarrow 11(x,y) \ge \frac{m}{2}$$

Là encore, on voit que la règle de Solomon Fortier appliquée à cet indice, avec un seuil "milieu d'intervalle de variation", redonne la règle de la majorité simple de Condorcet au niveau d'une borne sur 11(x,y). Cet indice bien que complexe à calculer est donc plus cohérent que les deux indices de Yule. En particulier il semble avoir un comportement voisin de celui de l'indice Tetrachorique à une homothétie près.

4.4.4.5 L'indice de Moyenne Géométrique des « quatre ratios » d'Ochiai (1973)

Cet indice comme le précédent est une généralisation sur les 4 cases du tableau Tetrachorique de l'indice d'Ochiai , on peut d'ailleurs l'attribuer également à Anderberg, même si ce dernier en donne la paternité au zoologiste Japonais, par filiation. C'est de façon simple la moyenne géométrique d'ordre 4 des quatre ratios précédemment définis pour l'indice $S_{R4}(x,y)$. Cet indice est défini par :

(f83) $$S_{o4}(x,y) = \sqrt[4]{\frac{11(x,y)}{11(x,y)+10(x,y)} \times \frac{11(x,y)}{11(x,y)+01(x,y)} \times \frac{00(x,y)}{00(x,y)+10(x,y)} \times \frac{00(x,y)}{00(x,y)+01(x,y)}}$$

Cet indice varie de 0 à 1, il vaut :

- 1 si 10(x,y)=01(x,y)=0
- 0 si 11(x,y)=00(x,y)=0

Comportement de l'indice $S_{o4}(x,y)$ en cas de Disjonction Complète

Comme nous l'avons fait pour l'indice d'Anderberg précédent, pour avoir une bonne idée du comportement de cet indice, calculons sa valeur dans le contexte particulier de la situation de disjonction complète. En effet dans ce cas la formule globale se simplifie, et comme dans le cas de $S_{R4}(x,y)$ un certain nombre de propriétés caractéristiques peuvent être mises en évidence. En effet en cas de disjonction complète l'indice $S_{o4}(x,y)$ se simplifie selon la formulation suivante :

(f 84) $$S_{o4}(x,y) = \sqrt[4]{\frac{11(x,y)}{m} \times \frac{11(x,y)}{m} \times \frac{P+11(x,y)-2m}{P-m} \times \frac{P+11(x,y)-2m}{P-m}}$$

Soit après simplifications :

(f 85) $$S_{o4}(x,y) = \sqrt{\frac{11(x,y)}{m} \times \frac{P+11(x,y)-2m}{P-m}}$$

Sous cette forme simplifiée, on voit que cet indice vaut 1 si 11(x,y)=m (valeur maximum du « matching » en cas de disjonction complète) et il vaut 0 si 11(x,y)=0.
A titre de comparaison et pour l'étalonner, calculons la borne induite sur 11(x,y) dès que l'on applique la règle de Solomon et Fortier. Ici l'intervalle de variation, même dans ce cadre disjonctif est (0, 1), la borne de la règle de Solomon- Fortier est donc fixée à ½ , d'où :

(f 86) $$S_{o4}(x,y) \geq \frac{1}{2} \quad \Rightarrow \quad \sqrt{\frac{11(x,y)}{m} \times \frac{P+11(x,y)-2m}{P-m}} \geq \frac{1}{2}$$

Le développement de la formule (f 85) aboutit à résoudre l'inégalité du second degré en 11(x,y) suivante[14] :

$11^2(x,y) + (P - 2m).11(x,y) - \frac{1}{4}m.(P-m) \geq 0$, avec $\Delta = (P-2m)^2 + m(P-m) = P^2 - 3mP + 3m^2$

le calcul du déterminant de cette équation et le fait que 11(x,y) varie de 0 à m, donc est toujours positif, nous permet de définir la racine positive de l'équation qui est donnée par :

$$z_1 = \frac{1}{2}\left[-m(\overline{p}-2) + \sqrt{m^2[(\overline{p}-1)(\overline{p}-2)+1]}\right] \quad \text{où} \quad P = m\overline{p}$$

Comme :

$$\frac{1}{2}\left[-m(\overline{p}-2) + \sqrt{m^2[(\overline{p}-1)(\overline{p}-2)+1]}\right] = \frac{1}{2}\left[-m(\overline{p}-2) + m\sqrt{\frac{1}{\overline{p}}((\overline{p}-1)^3+1)}\right],$$

on obtient :

$z_1 = \frac{1}{2}\left[-m(\overline{p}-2) + m(\overline{p}-1)\sqrt{\frac{1}{\overline{p}}[(\overline{p}-1) + \frac{1}{(\overline{p}-1)^2}]}\right]$ en simplifiant l'expression sous le signe « racine carrée » et en utilisant un développement limité de la racine carrée au voisinage de 0 (ce qui a d'autant plus de sens que $\overline{p}$ est non négligeable), il vient :

$$\left[\sqrt{1 - \frac{1}{\overline{p}} + \frac{1}{(\overline{p}-1)^2}}\right] \cong 1 - \frac{1}{2}\left[\frac{1}{.\overline{p}} - \frac{1}{(\overline{p}-1)^2}\right] + \frac{1}{8}\left[\frac{1}{\overline{p}} - \frac{1}{(\overline{p}-1)^2}\right]^2$$

en remplaçant cette valeur dans l'expression de z_1, on obtient la une valeur très légèrement supérieure pour z_1, à savoir :

$$z_1 = \frac{1}{2}\left[-m(\overline{p}-2) + m(\overline{p}-1)\sqrt{1 - \left[\frac{1}{\overline{p}} - \frac{1}{(\overline{p}-1)^2}\right]}\right] \cong \frac{1}{2}\left[-m\overline{p} + 2m + m(\overline{p}-1)\left(1 - \frac{1}{2\overline{p}} + \frac{1}{8\,\overline{p}^2} + \frac{1}{2\,(\overline{p}-1)^2} + O(\frac{1}{p^3})\right)\right],$$

Pour satisfaire l'inégalité (f 86), 11(x,y) doit être supérieur à l'approximation de $z_{1,}$ donnée ci-dessus .
La borne de la règle de Solomon-Fortier devient alors :

$$11(x,y) \geq \frac{m}{4} + \frac{m[9.\overline{p}-5]}{8\overline{p}(\overline{p}-1)}$$

Si le nombre moyen de modalités est égal à 5 par exemple, on voit que 11(x,y) doit vérifier approximativement une règle de majorité Condorcéenne simple (m/2).

On voit qu'à la limite, si $\overline{p}$ croît, la borne tend vers $\frac{m}{4}$, c'est à dire vers une « majorité au quart ».

Dans ce cas là, comme d'ailleurs dans le cas d'indices du Groupe I Type II auquel l'indice d'Ochiai , vrai, appartient, on voit que cet indice est plus « généreux » que l'indice des « 4 ratios d'Anderberg » , voire trop généreux ce qui montre qu'il n'a que peu d'intérêt comme mesure discriminante.

[14] Les valeurs de 11(x,y) étant entières, les calculs sur l'inégalité (f 86) se font en supposant une relaxation en valeurs réelles de 11(x,y) , $0 \leq 11(x,y) \leq m$, au lieu d'avoir $11(x,y) \in \{ 0,1,2,\ldots m \}$

4.4.4.6 L'indice de Moyenne Harmonique des « quatre ratios »

Continuant dans le même principe de généralisation que précédemment, nous définirons un indice de moyenne harmonique des quatre ratios tetrachoriques, en définissant ce nouvel indice $S_{h4}(x,y)$ par :

(f 84) $$\frac{1}{S_{h4}(x,y)} = \frac{1}{4}\left[\frac{11(x,y)+10(x,y)}{11(x,y)} + \frac{11(x,y)+01(x,y)}{11(x,y)} + \frac{00(x,y)+10(x,y)}{00(x,y)} + \frac{00(x,y)+01(x,y)}{00(x,y)}\right]$$

Cet indice peut être simplifié selon la formule:

$$\frac{1}{S_{h4}(x,y)} = \frac{1}{4}\left[\frac{2.11(x,y)+10(x,y)+01(x,y)}{11(x,y)} + \frac{2.00(x,y)+10(x,y)+01(x,y)}{00(x,y)}\right]$$

Comme par ailleurs : 11(x,y)+00(x,y)+10(x,y)+01(x,y)=P, il peut être exprimé uniquement en fonction des quantités 00(x,y) et 11(x,y), il vient:

$$\frac{1}{S_{h4}(x,y)} = \frac{1}{4}\left[\frac{P-00(x,y)+11(x,y)}{11(x,y)} + \frac{P-11(x,y)+.00(x,y)}{00(x,y)}\right]$$

Après réduction au même dénominateur et inversion de la formule il vient:

$$S_{h4}x,y) = \left[\frac{4\,11(x,y).00(x,y)}{P[11(x,y)+00(x,y)]-[00(x,y-11(x,y)]^2}\right]$$

Sous cette dernière formulation on voit que cet indice vaut :

- 1 si 10(x,y)=01(x,y)=0 , en effet dans ce cas P=00(x,y)+11(x,y) et le dénominateur revient à : 4.11(x,y).00(x,y)
- 0 si 11(x,y) =0

<u>Comportement de l'indice S_{h4}(x,y) en cas de Disjonction Complète</u>

Comme précédemment donnons un étalonnage de cet indice par rapport à la règle de Solomon et Fortier en cas de disjonction complète. En remplaçant 00(x,y) par P+11(x,y)-2m, et en revenant à la formule (f74) nous avons l'expression suivante de l'indice:

(f88) $$\frac{1}{S_{h4}(x,y)} = \frac{1}{2}\left[\frac{m}{11(x,y)} + \frac{(P-m)}{P+11(x,y)-2m}\right]$$

A partir de cette expression obtenue lors d'une disjonction complète, on voit bien que cet indice est maximum lorsque 11(x,y)=m, puisque qu'alors on trouve 2/2.

Calculons maintenant la valeur de la borne sur 11(x,y), lorsqu'on applique la règle de Solomon et Fortier.
Puisque l'intervalle de variation est bien: (0, 1), la borne choisie pour l'indice sera ½.

La formule (f 88) induit donc l'inégalité fonction de 11(x,y) suivante:

$$S_{h4}x,y) \geq \frac{1}{2} \quad \Rightarrow \quad \frac{2.11(x,y)[P+11(x,y)-2.m]}{11(x,y).P+m(P-2m)} \geq \frac{1}{2}$$

Ce qui implique l'inégalité quadratique suivante pour 11(x,y):

(f 89) $$4.11^2(x,y)+11(x,y)[3.P-8.m]-m(P-2m)\geq 0$$

En tenant compte de la remarque faite dans la note de bas de page n°4, la borne inférieure sur 11(x,y) sera la racine positive de l'équation du second degré précédente. Le déterminant n'est pas simple, il est égal à : $\Delta=[3P-8m]^2+16m(P-2m)=8[.P-2m]^2+P^2$

La racine positive z_1 est donnée par :

$z_1=\dfrac{8m-3P+\sqrt{8[P-2.m]^2+P^2}}{8}$ en jouant sur le fait que $P=m\overline{p}$, on obtient :

$$z_1=\frac{m(8-3\overline{p})+m\overline{p}\sqrt{8[1-\frac{2}{\overline{p}}]^2+1}}{8}=\frac{m}{8}\left[8-3\overline{p}+3\overline{p}\sqrt{1-\frac{32}{9\overline{p}}+\frac{32}{9\overline{p}^2}}\right]\cong\frac{m}{8}\left[8-3\overline{p}+3\overline{p}[1-\frac{16}{9\overline{p}}+\frac{272}{81\overline{p}^2}]\right]=\frac{m}{8}\left[8-\frac{16}{3}+\frac{272}{27\overline{p}}\right]=\frac{m}{3}\left[1+\frac{34}{9\overline{p}}\right]$$

L'approximation sur la racine carrée étant d'autant plus significative que $\overline{p}$ est fort, on voit que la borne pour 11(x,y) se traduit ici par :

$$11(x,\ y)\geq\frac{m}{3}\left[1+\frac{3.77}{\overline{p}}\right],$$

4.4.4.7 Indice de Baroni-Urbani et Buser (1976)

Cet indice, introduit en 1976 (voir Baroni, Urbani et Buser (1976)) par des spécialistes de systématique zoologique se distingue des précédents en ce sens qu'il dissymétrise l'influence des configurations de « matchings », qu'il traite non linéairement , des configurations de « non matching » , qu'il traite linéairement, le tout en pondérant légèrement plus le « matching 11 », bref un indice pas « naturel du tout » quant à sa filiation . Cependant il semble donner satisfaction aux experts des domaines de la zoologie (systématiciens, phylogénistes), de la biologie (spécialistes des mesures de bio-diversité) ou de l'écologie en général (voir à ce propos l'article de Richard Boyle et Paula Ellison (2001)).

(f 90) $$S_{BuB}(x,y)=\left[\frac{\sqrt{11(x,y).00(x,y)}+11(x,y)}{\sqrt{11(x,y).00(x,y)}+11(x,y)+01(x,y)+10(x,y)}\right]$$

Cet indice varie bien de 0 à 1, il vaut 1 quand 10(x,y)=01(x,y)=0 et il vaut 0 quand 11(x,y)=0 .
Pour ce faire une idée de plus précise de cet indice on peut le comparer à l'indice Y de Yule, donné par :

$$S_Y(x,y)=\frac{\sqrt{11(x,y).00(x,y)}-\sqrt{10(x,y).01(x,y)}}{\sqrt{11(x,y).00(x,y)}+\sqrt{10(x,y).01(x,y)}}$$

qui met en jeu des quantités voisines. Mais ce dernier, on l'a vu,, se comporte comme un indice de type indice de corrélation puisqu'il varie de -1 à 1, mais en utilisant la remarque de bas de page n°1, on sait que : $S(x,y)=1/2(\rho_{xy}+1)$ construit à partir d'un indice de corrélation redevient un indice de similarité variant de 0 à 1.
On peut donc comparer l'indice $S_{BuB}(x,y)$

à l'indice $$S'_Y(x,y)=\frac{1}{2}[S_Y(x,y)+1]=\frac{\sqrt{11(x,y).00(x,y)}}{\sqrt{11(x,y).00(x,y)}+\sqrt{10(x,y).01(x,y)}}$$

Pour ce faire montrons que :

$$\frac{\sqrt{11(x,y).00(x,y)}}{\sqrt{11(x,y).00(x,y)}+\sqrt{10(x,y).01(x,y)}} \geq \frac{\sqrt{11(x,y).00(x,y)}+11(x,y)}{\sqrt{11(x,y).00(x,y)}+11(x,y)+10(x,y)+00(x,y)}$$

En effet après développement on obtient :

(i) $[10(x,y)+01(x,y)]\sqrt{11(x,y).00(x,y)} \geq [\sqrt{11(x,y).00(x,y)}+11(x,y)]\sqrt{10(x,y).01(x,y)}$

Or (ii) $[\sqrt{10(x,y)}-\sqrt{01(x,y)}]^2 = 10(x,y)+01(x,y)-2\sqrt{10(x,y).01(x,y)}$,

en remplaçant l'expression de 10(x,y)+01(x,y) , telle qu'elle apparaît dans (ii) et en la reportant dans (i) , on obtient :

(iii) $[\sqrt{10(x,y)}-\sqrt{01(x,y)}]^2+2\sqrt{10(x,y).01(x,y)}\,]\sqrt{11(x,y).00(x,y)} \geq [\sqrt{11(x,y).00(x,y)}+11(x,y)]\sqrt{10(x,y).01(x,y)}$

soit après simplification:

$$[\sqrt{10(x,y)}-\sqrt{01(x,y)}]^2\,]\sqrt{11(x,y).00(x,y)} \geq [\sqrt{10(x,y).01(x,y)}[11(x,y)]-\sqrt{11(x,y).00(x,y)}]$$

Le membre de gauche de cette inégalité est toujours positif , alors que le membre de droite est toujours négatif , sauf en cas de disjonction bivalente, i.e ; P=2m. En effet en général 00(x,y)≥ 11(x,y) d'où : $\sqrt{11(x,y).00(x,y)} \geq 11(x,y)$

De ce calcul nous déduisons :

$$S'_Y(x,y) \geq S_{BuB}(x,y) \quad \forall x,y$$

<u>Comportement de l'indice $S_{BuB}(x,y)$ en cas de Disjonction Complète</u>

En cas de disjonction complète, on développe l'expression précédente en fonction de 11(x,y), de P et de m selon les formules obtenues en remplaçant 00(x,y), 10(x,y) et 01(x,y) par leurs valeurs en fonction de 11(x,y) et de P, du fait que on retrouve : 00(x,y)=11(x,y)+P-2m, 01(x,y)=10(x,y)=m-11(x,y),

$$S_{BuB}(x,y) = \left[\frac{\sqrt{11(x,y)[11(x,y)+P-2m]}+11(x,y)}{\sqrt{11(x,y)[11(x,y)+P-2m]}+2m-11(x,y)}\right]$$

Comme la borne de Solomon Fortier vaut **½** , puisque l'indice varie bien de 0 à 1, la borne de référence sera obtenue pour la valeur 11(x,y) , telle que:

$$S_{BuB}(x,y) \geq \frac{1}{2} \;\Rightarrow\; \left[\frac{\sqrt{11(x,y)[11(x,y)+P-2m]}+11(x,y)}{\sqrt{11(x,y)[11(x,y)+P-2m]}+2m-11(x,y)}\right] \geq \frac{1}{2}$$

Cette inégalité implique l'équation du second degré suivante:

$$-8.11^2(x,y)+11(x,y)[\overline{p}+10]m-4m^2 \geq 0 \text{ (où l'on a remplacé P par } m\overline{p}\text{)}$$

On prend la racine positive z_1, comme nous l'avons fait précédemment en simplifiant les expressions sous les racines, nous obtenons:

$$z_1 = \frac{m[\overline{p}+10]-m\overline{p}\sqrt{[1+\frac{20}{\overline{p}}-\frac{28}{\overline{p}^2}]}}{16} = \frac{m}{16}\left[10+\overline{p}-\overline{p}\sqrt{1+\frac{20}{\overline{p}}-\frac{28}{\overline{p}^2}}\right] \cong \frac{m}{16}\left[10+\overline{p}-\overline{p}\left[1+\frac{10}{\overline{p}}-\frac{14}{\overline{p}^2}-\frac{50}{\overline{p}^2}+\frac{140}{\overline{p}^3}\right]\right] = \frac{m}{16}\left[\frac{64}{\overline{p}}-\frac{140}{\overline{p}^2}\right] = \frac{m}{\overline{p}}\left[4-\frac{35}{4\overline{p}}\right]$$

Ceci implique donc que : $11(x,y) \geq \frac{m}{\overline{p}}\left[4-\frac{35}{4.\overline{p}}\right]$

On voit par exemple que si $\overline{p}=4$, alors 11(x,y) doit être supérieur à 0,45 m , c'est à dire que si m=9 (par exemple) on doit avoir 11(x,y)≥4, (puisque les valeurs de 11(x,y) sont toujours des valeurs entières), on vérifie bien d'après la formule(f87) que :

$$S_{BuB}(x,y) = \left[\frac{\sqrt{11(x,y)[11(x,y) + P - 2m]} + 11(x,y)}{\sqrt{11(x,y)[11(x,y) + P - 2m]} + 2m - 11(x,y)} \right] = \frac{\sqrt{4.(4+36-18)} + 4}{\sqrt{4.(4+36-18)} + (18-4)} = \frac{\sqrt{88} + 4}{\sqrt{88} + 14} = \frac{13,38}{23,38} = 0,57$$

On voit d'ailleurs sur cet exemple qu'il faut que 11(x,y) soit strictement supérieur à 4 , si 11(x,y) n'était égal qu' à 3 la valeur de l'indice serait égale à 0,4768, ce qui ne permettrait pas de vérifier la Borne de Solomon Fortier.

Le comportement limite de l'indice n'est pas évident si $\overline{p}$ prend des valeurs grandes, on voit que la borne implique une majorité au $\frac{4}{\overline{p}}$ èmes, ce qui n'en fait pas un indice de caractère général utilisable dans des conditions standard.

4.4.5 Indices du GROUPE II, Type III (indices obtenus comme ratios de fonctions complexes non standard des quantités du tableau « Tetrachorique »)

Dans ce paragraphe nous introduirons certains indices utilisés en statistique pour mesurer les interactions et les associations sur tableaux de contingences, puisque le tableau Tetrachorique n'est, après tout, qu'un tableau de contingence (2x2) particulier.

Dans cette configuration l'utilisation des coefficients d'association sur tableau de contingence peut ou non (cela dépend du contexte) servir de base à des indicateurs de similarité, à condition d'une part de distinguer les configurations dites de « matchings » positifs des configurations de « matchings » négatifs, ce que ne font pas d'amblée ces coefficients d'association .
Rappelons ici qu'un tableau de contingence croisant deux variables catégorielles **x** et **y**, x ayant p modalités et y ayant q modalités se présente sous la forme suivante :

x \ y	1	2	v	..	q	
1						$n_{1.}$
2						$n_{2.}$
u			$n_{u\,v}$			$n_{u.}$
..						
p						
	$n_{.1}$	$n_{.2}$	$n_{.v}$			$n_{..}$

Où
- $n_{u\,v}$ = Nombre d' objets ayant à la fois la modalité u de x et v de y ,
- $n_{u.}$ = Nombre d' objets ayant la modalité u de x
- $n_{.v}$ = Nombre d' objets ayant la modalité v de y
- $n_{..}$ = Nombre total d' objets

Nous donnerons dans les paragraphes suivants des coefficients ou indices d'association que nous évaluerons en tant qu'indices de similarité. Ceci se fera au travers d'un processus con-

sistant à remplacer les valeurs n_{uv} du tableau de contingence général précédent par les valeurs correspondantes du tableau particulier de contingence que nous avons nommé « Tetrachorique » et qui se présente sous la forme :

	y = 1	**y = 0**
x= 1	11(x,y)	10(x,y)
x= 0	01(x,y)	00(x,y)

Dans le cas du tableau Tetrachorique les valeurs p et q du tableau général sont égales à 2 , la valeur $n_{..} = P$, et les différentes valeurs n_{uv} sont au nombre de 4, à savoir : 11(x,y), 10(x,y , 01(x,y) et 00(x,y).

4.4.5.1 Indice de similarité déduit du Coefficient d'Ecart à l'Indétermination Contingentielle, du coefficient de Janson-Vegelius et du Critère de Rand

Rappelons que cet indice d' « Ecart à l'Indétermination » , dont on pourra trouver les propriétés dans F. Marcotorchino (1984) et F. Marcotorchino et N. El Ayoubi (1991), se présente sous la forme :

$$\text{Ind}(x,y) = \sum_{u=1}^{p}\sum_{v=1}^{q}\left[n_{uv} - \left(\frac{n_{u.}}{q} + \frac{n_{.v}}{p} - \frac{n_{..}}{pq}\right)\right]^2$$

Cet indice est en fait le numérateur du coefficient J(x,y) de Janson et Vegelius (voir S. Janson et J. Vegelius (1982)) , coefficient spécial de corrélation, donné lui-même sous la forme d'un ratio composite :

(f 91)
$$J(x,y) = \frac{N(x,y)}{D(x,y)}$$

Le numérateur est donné par :

$$N(x,y) = \sum_{u=1}^{p}\sum_{v=1}^{q} n_{uv}^2 - \frac{\sum_{u=1}^{p} n_{u.}^2}{q} - \frac{\sum_{v=1}^{q} n_{.v}^2}{p} + \frac{n_{..}^2}{pq}$$

le dénominateur étant donné par :

$$D(x,y) = \sqrt{\sum_{u=1}^{p} n_{u.}^2\left[1 - \frac{2}{p}\right] + \frac{n_{..}^2}{p^2}}\sqrt{\sum_{v=1}^{q} n_{.v}^2\left[1 - \frac{2}{q}\right] + \frac{n_{..}^2}{q^2}}$$

Par ailleurs, on peut montrer (voir les articles précités de (1984) et (1991)) que :

$$\text{Ind}\ (x, y) = \sum_{u=1}^{p}\sum_{v=1}^{q}\left[n_{uv} - \left(\frac{n_{u.}}{q} + \frac{n_{.v}}{p} - \frac{n_{..}}{pq}\right)\right]^2 = \sum_{u=1}^{p}\sum_{v=1}^{q} n_{uv}^2 - \frac{\sum_{u=1}^{p} n_{u.}^2}{q} - \frac{\sum_{v=1}^{q} n_{.v}^2}{p} + \frac{n_{..}^2}{pq} = N(x, y)$$

Cet indice J(x,y) varie bien de 0 à 1 .

- Il vaut 0 en cas d'Indétermination totale sur toutes les cases du tableau de contingence voir F. Marcotorchino (1984) pour des détails sur cette structure particulièrement intéressantes des données), c'est à dire si :

$$\sum_{u=1}^{p}\sum_{v=1}^{q}\left[n_{uv} - \left(\frac{n_{u.}}{q} + \frac{n_{.v}}{p} - \frac{n_{..}}{pq}\right)\right]^2 = 0 \Rightarrow n_{uv} = \left(\frac{n_{u.}}{q} + \frac{n_{.v}}{p} - \frac{n_{..}}{pq}\right) \quad \forall\ u \text{ et } v$$

- Il vaut 1 en cas d'association complète (voir (M1984)), c'est à dire d'une part si : p=q, d'autre part si pour toute cellule (u,v) du tableau de contingence on a :

$$n_{uv} = n_{u.} = n_{.v} \quad \forall\ u = v$$

$$n_{uv} = 0 \quad \forall\ u \neq v$$

dans cette configuration on voit, de façon simple, que :

$$D\ (x, y) = N(x, y) = \sum_{u=1}^{p} n_{u.}^2\left[1 - \frac{2}{p}\right] + \frac{n_{..}^2}{p^2}$$

Considérons alors le Coefficient d' « Association de Janson –Vegelius » donné sous forme développée par :

(f 92)
$$J(x, y) = \frac{N(x, y)}{D(x, y)} = \frac{\sum_{u=1}^{p}\sum_{v=1}^{q}\left[n_{uv} - \left(\frac{n_{u.}}{q} + \frac{n_{.v}}{p} - \frac{n_{..}}{pq}\right)\right]^2}{\sqrt{\left[\sum_{u=1}^{p} n_{u.}^2\left(1 - \frac{2}{p}\right) + \frac{n_{..}^2}{P^2}\right]\left[\sum_{v=1}^{q} n_{.v}^2\left(1 - \frac{2}{q}\right) + \frac{n_{..}^2}{q^2}\right]}}$$

En remplaçant dans l'expression du coefficient complexe de Janson -Vegelius : $n_{..}$ par P, p et q par 2 et n_{uv} par l'une des valeurs : 11(x,y), 10(x,y), 01(x,y), 00(x,y), on obtient , l'expression très simplifiée suivante:

$$J(x, y) = \frac{\frac{1}{4}[11(x, y) + 00(x, y) - (10(x, y) + 01(x, y))]^2}{\sqrt{\frac{P^2}{4}}\sqrt{\frac{P^2}{4}}} = \frac{[11(x, y) + 00(x, y) - (10(x, y) + 01(x, y))]^2}{P^2}$$

Dès lors, on peut définir à partir du Coefficient de similarité de Janson-Vegelius, l' indice de similarité $S_{JV}(x,y)$ associé défini par :

(f 93)
$$S_{JV}(x, y) = \sqrt{J(x, y)} = \frac{[11(x, y) + 00(x, y) - (10(x, y) + 01(x, y))]}{P}$$

On constate sous cette forme et après simplification des calculs, que l'indice obtenu n'est ni plus ni moins que la version « corrélative » de l'indice de Sokal et Michener (ou indice de « Simple Matching ») (voir (f 55)), c'est à dire qu'il varie de –1 à +1.

+1 si 10(x,y)=01(x,y)=0 et –1 si 11(x,y)=00(x,y)=0

On aurait pu également obtenir ce résultat de façon encore plus simple, en se rappelant (voir F. Marcotorchino (1984)) que le coefficient de Janson-Vegelius dans le cas de tableaux de contingence (2x2), est égal à :

2x Coefficient de contingence de Rand -1 ,

où le Coefficient de Rand , (voir W. Rand (1971)) est donné par :

$$R(x,y) = \frac{2\sum_{u=1}^{p}\sum_{v=1}^{q} n_{uv}^2 - \sum_{u=1}^{p} n_{u.}^2 - \sum_{v=1}^{q} n_{.v}^2 + n_{..}^2}{n_{..}^2}$$

On retrouve ainsi l'illustration des remarques indiquées au paragraphe 4.4..1.1:

En effet :

$$2R(x,y) - 1 = \frac{4\sum_{u=1}^{p}\sum_{v=1}^{q} n_{uv}^2 - 2\sum_{u=1}^{p} n_{u.}^2 - 2\sum_{v=1}^{q} n_{.v}^2 + n_{..}^2}{n_{..}^2}$$

En remplaçant les valeurs du tableaux Tetrachorique dans l'expression précédente , on obtient trivialement :

$$2R(x,y) - 1 = \frac{4\sum_{u=1}^{p}\sum_{v=1}^{q} n_{uv}^2 - 2\sum_{u=1}^{p} n_{u.}^2 - 2\sum_{v=1}^{q} n_{.v}^2 + n_{..}^2}{n_{..}^2} = \frac{[11(x,y) + 00(x,y) - (10(x,y) + 01(x,y))]^2}{P^2}$$

d'où :

$$S_{JV}(x,y) = \sqrt{J(x,y)} = \sqrt{2.R(x,y) - 1} = \frac{[11(x,y) + 00(x,y) - (10(x,y) + 01(x,y))]}{P}$$

(cqfd)

Sous cette dernière forme on voit que cet indice est strictement équivalent à l'Indice de Hamann (voir note de bas de page n°11), c'est à dire à la différence entrel'Indice de Sokal et Michener et l' Indice de Green –Rao.

4.4.5.2 Indice de similarité déduit du Coefficient de Light et Margolin , Haldane et du Coefficient de Goodman-Kruskal

De la même manière que dans le cas du coefficient de Janson et Vegelius , on peut s'inspirer de la littérature sur les indices d'association sur tableaux de contingence, pour déduire et définir d'autres indices de similarité par un processus équivalent à celui utilisé dans le cas précédent. Parmi les choix possibles, une excellente famille d'indices d'association (non standard quant à leur construction) est celle créée autour de la filiation des indices de Goodman-Kruskal, Light et Margolin , Haldane. Ainsi historiquement l'indice le plus complet est celui qui a été introduit en 1954 par Goodman Kruskal et redéfini dans leur livre (voir L. A.Goodman et W.H. Kruskal (1979)), sous le nom de « Tau » de Goodman- Kruskal. Il est donné ci dessous dans sa formulation pour tableau de contingence à caractère général :

(f 94)

$$\tau(x,y) = \frac{\sum_{u=1}^{p}\sum_{v=1}^{q} \frac{n_{uv}^2}{n_{u.}} - \frac{1}{n_{..}}\sum_{v=1}^{q} n_{.v}^2}{n_{..} - \frac{1}{n_{..}}\sum_{v=1}^{q} n_{.v}^2}$$

On peut montrer que le Coefficient de Light et Margolin (1971) est équivalent au Numérateur du Coefficient de Goodman Kruskal, et qu' historiquement en (1940), Haldane avait

déjà proposé un coefficient quasi équivalent au Coefficient de Light et Margolin (pour plus de détails sur propriétés de ces indices et sur leur filiation on consultera l'article F . Marcotorchino (1984)).
Quoique dissymétrique, ce coefficient de Goodman-Kruskal se comporte pour ses valeurs maximum et minimum comme le coefficient de Tchuprow, ou comme le Coefficient du χ^2.

- Il vaut 0 en cas d'Indépendance contingentielle sur toutes les cases du tableau de contingence (voir Goodman , Kruskal (1979)), c'est à dire si :

$$n_{uv} = \frac{n_{u.} n_{.v}}{n_{..}} \quad \forall \text{ u et v}$$

- Il vaut 1 en cas d'association complète , c'est à dire d'une part si : p=q, d'autre part si pour toute cellule (u,v) du tableau de contingence on a :

$$n_{uv} = n_{u.} = n_{.v} \quad \forall \ u = v$$
$$n_{uv} = 0 \quad \forall \ u \neq v$$

Pour évaluer la valeur du Coefficient d'association de Goodman Kruskal sur le tableau de contingence (2x2) Tetrachorique, il suffit de remplacer : $n_{..}$ par P, p et q par 2 et n_{uv} par l'une des valeurs : 11(x,y), 10(x,y), 01(x,y), 00(x,y), on obtient après les remplacements proposés, l' expression suivante:

$$\tau(x,y) = \frac{\dfrac{11^2(x,y)+10^2(x,y)}{11(x,y)+10(x,y)} + \dfrac{00^2(x,y)+01^2(x,y)}{00(x,y)+01(x,y)} - \dfrac{1}{P}\left[(11(x,y)+01(x,y))^2 + (00(x,y)+10(x,y))^2\right]}{P - \dfrac{1}{P}\left[(11(x,y)+01(x,y))^2 + (00(x,y)+10(x,y))^2\right]}$$

qui après développements et nouvelles simplifications revient à :

$$\tau(x,y) = \frac{[11(x,y)00(x,y) - 01(x,y)10(x,y)]^2}{[11(x,y)+01(x,y)][00(x,y)+10(x,y)][11(x,y)+10(x,y)][00(x,y)+01(x,y)]}$$

Dès lors, on peut définir à partir du Coefficient d'association de Goodman Kruskal, un indice de similarité dit de Goodman Kruskal, $S_{GK}(x,y)$ défini, comme pour Janson-Vegelius, par la racine carré du Coefficient d'association, calculé sur le tableau de contingence Tetrachorique, qui peut comme on l'a vu être positive ou négative ; il vient :

(f95)
$$S_{GK}(x,y) = \sqrt{\tau(x,y)} = \frac{[11(x,y)00(x,y) - 10(x,y)01(x,y)]}{\sqrt{[11(x,y)+10(x,y)][11(x,y)+01(x,y)][00(x,y)+10(x,y)][00(x,y)+01(x,y]}}$$

En se référant au § 4.4.4.1, formule (f 71) , on voit que malgré les calculs complexes occasionnés, cet indice de similarité n'est rien d'autre que l'indice de similarité Tetrachorique $S_T(x,y)$ (ou de Bravais-Pearson) , défini directement à partir du tableau 2x2 associé. En conclusion on a donc :

$$S_{GK}(x,y) = \sqrt{\tau(x,y)} = S_T(x,y)$$

4.4.5.3 Indice de similarité déduit du Coefficient d' « Ecart carré à l'Indépendance » et du Coefficient Contingentiel de I.C. Lerman

Une autre famille d'indices d'association est issue d'une Représentation Relationnelle générale des indices d'Association sur tableau de contingence. Ces indices font partie de la version la plus complète des indices d'association de types relationnels dont la formule générale (voir F. Marcotorchino (1984), F. Marcotorchino et N. El Ayoubi (1991), A. Idrissi (2000) s'écrit sous la forme :

(f 96)
$$A(x,y) = \frac{\sum_{i=1}^{N}\sum_{i'=1}^{N}(x_{ii'} - \mu_x)(y_{ii'} - \mu_y)}{\sqrt{\sum_{i=1}^{N}\sum_{i'=1}^{N}(x_{ii'} - \mu_x)^2}\sqrt{\sum_{i=1}^{N}\sum_{i'=1}^{N}(y_{ii'} - \mu_y)^2}}$$

où la variable x (catégorielle) est représentée, comme nous l'avions vu au § 2.2.1, par son expression relationnelle sous forme d'une matrice binaire de terme général : $\{x_{ii'}\}$, idem pour y, représentable sous forme d'une matrice relationnelle binaire : $\{y_{ii'}\}$. Et où μ_x et μ_y sont, respectivement, une moyenne dérivée de la distribution des $\{x_{ii'}\}$ et une moyenne dérivée de la distribution des $\{y_{i\,i'}\}$. Comme ceci a été montré (en particulier dans la thèse de N. Amal Idrissi (2000)), puis repris par A. Hicham et G.Saporta (2003) et revu et raffiné s dans le travail de J. Ah-Pine (2007)), suivant les valeurs choisies pour μ_x et μ_y, on retrouve des coefficients d'association connus. Ainsi , on a :

- si $\mu_x = 1/2$ et $\mu_y = 1/2$ (Moyennes logiques) ➔ $A(x,y) = 2.R(x,y) - 1$, on retrouve le critère de Rand
- si $\mu_x = 1/p$ et $\mu_y = 1/q$ (Moyennes Probabilistes) ➔ $A(x,y) = J(x,y)$, on retrouve le critère Janson Vegelius
- si $\mu_x = \frac{x_{..}}{n^2}$ et $\mu_y = \frac{y_{..}}{n^2}$ (Moyennes empiriques) ➔ $A(x,y) = L(x,y)$, on retrouve un critère voisin du critère de « Vraisemblance du Lien » de I.C. Lerman (1987), dont la formulation contingentielle (adaptée à notre problématique) est donnée ci dessous :

(f 97)
$$L(x,\ y) = \frac{\sum_{u=1}^{p}\sum_{v=1}^{q} n_{uv}^2 - \frac{\sum_{u=1}^{p} n_{u.}^2 \sum_{v=1}^{q} n_{.v}^2}{n_{..}^2}}{\sqrt{\sum_{u=1}^{p} n_{u.}^2 \left(1 - \frac{\sum_{u=1}^{p} n_{u.}^2}{n_{..}^2}\right)}\sqrt{\sum_{v=1}^{q} n_{.v}^2 \left(1 - \frac{\sum_{v=1}^{q} n_{.v}^2}{n_{..}^2}\right)}}$$

Le numérateur de ce coefficient à savoir la quantité :

$$\sum_{u=1}^{p}\sum_{v=1}^{q} n_{uv}^2 - \frac{\sum_{u=1}^{p} n_{u.}^2 \sum_{v=1}^{q} n_{.v}^2}{n_{..}^2}$$

est connue sous le nom de Critère de l' « Ecart Carré à l'Indépendance[15] », critère qui possède de nombreuses propriétés structurelles intéressantes (on trouvera dans Marcotorchino (1984,) , une étude complète de ce critère, d'un point de vue formel, structurel et applicatif).

Ce coefficient L(x,y), dérivé du Coefficient de I. C. Lerman, possède les propriétés suivantes à propos de ses valeurs maximales et minimales.

- Il vaut 0 en cas d'Indépendance contingentielle sur toutes les cases du tableau de contingence, c'est à dire si :

$$n_{uv} = \frac{n_{u.} n_{.v}}{n_{..}} \quad \forall \text{ u et v}$$

  En effet dans ce cas le numérateur de L(x,y) s'annule de façon évidente.
- Il vaut 0 également dans le cas d'une solution « parasite » (qui se présente uniquement sur des tableaux 2x2) (voir (F. Marcotorchino (1984 partie III)), et pour laquelle si le tableau 2x2 s' écrit :

	y	Non y
x	a	b
Non x	c	d

la configuration suivante se produit:

(i) $a^2 + d^2 = b^2 + c^2$ avec $n = a+b+c+d$ (on identifie ici n et $n_{..}$)

Du fait que l'équation précédente est équivalente à :

(ii) $(a+d)^2 - 2ad = (b+c)^2 - 2bc$

on a également :

(iii) $(a+d)^2 - (b+c)^2 = 2(ad-bc) \Rightarrow n([a+d]-[b+c]) = 2(ad-bc)$

Parmi les solutions de l'équation Diophantienne (i), (il y en a une infinité) nous en donnons ci dessous une famille générale et paramétrée (voir (F. Marcotorchino (1984)) :

$\{a = 2s+7, b = 2s+5, c = s+5, d = s+1\}$ (s prenant des valeurs entières : 0,1,2,..m, ..)

Il est facile de vérifier que si s=1 par exemple, les valeurs : a=9, b=7, c=6 , d=2 , sont telles qu'elles annulent le coefficient L(x,y), calculé sur le tableau :

	y	Non y
x	9	7
Non x	6	2

bien que l'on ne soit pas (loin de là) dans une configuration d'indépendance contingentielle.
On peut voir , de façon évidente, que pour cette série de solutions $\forall$ s on a : $([a+d]-[b+c]) = -2$, de ce fait ceci implique que (ad – bc) = - n puisque l'on travaille sur des valeurs entières.

En effet :

n= (2s+7)+(s+1)+(2s+5)+(s+5)=6s+18, ad = 2 s^2+ 9s + 7, bc = 2 s^2+15s+ 25, a+d = 3s +8 , b+c = 3 s +10
d'où : (ad – bc) = - (6s+18) =-n et (a+d) –(b+c) = (3 s +8) –(3s+10) = -2

[15] A ne pas confondre avec le carré de l'écart à l'indépendance (Numérateur du Coefficient de Tchuprow) ou la forme de base du Coefficient du χ^2 qui s'écrit : $\sum_{u=1}^{p}\sum_{v=1}^{q}(n_{uv} - \frac{n_{u.} n_{.v}}{n_{..}})^2$

- Il vaut 1 en cas d'association complète , c'est à dire d'une part si : p=q, d'autre part si pour toute cellule (u,v) du tableau de contingence on a :

$$n_{u\,v} = n_{u.} = n_{.v} \quad \forall \; u = v$$

$$n_{u\,v} = 0 \quad \forall \; u \neq v$$

En effet dans ce cas Numérateur et Dénominateur sont égaux.

- **Enfin le calcul de la valeur minimale de cet indice n'est pas du tout évident**. Nous allons donner la solution dans le cas d'un tableau 2x2 et montrer la difficulté associée pour trouver la solution optimale de ce problème.

Calculons, dans un premier temps, la valeur minimale du Numérateur dans le cas du tableau de contingence (2x2) général suivant:

	y	Non y
x	a	b
Non x	c	d

On peut montrer, que dans le cas d'un tableau 2x2, la quantité correspondant au numérateur de L(x,y) s'écrit sous la forme suivante:

$$\sum_{u=1}^{p}\sum_{v=1}^{q} n_{u\,v}^2 - \frac{\sum_{u=1}^{p} n_{u.}^2 \sum_{v=1}^{q} n_{.v}^2}{n_{..}^2} = \frac{(a+b+c+d)^2[a^2+b^2+c^2+d^2]-[(a+b)^2+(c+d)^2][(a+c)^2+(b+d)^2]}{n_{..}^2}$$

Après développement de la formule précédente et au bout de calculs assez fastidieux , on montre que cette quantité se factorise selon la formule suivante :

(f 98) $$\frac{(a+b+c+d)^2[a^2+b^2+c^2+d^2]-[(a+b)^2+(c+d)^2][(a+c)^2+(b+d)^2]}{n_{..}^2} = \frac{2}{n^2}(ad-bc)[a^2+d^2-(b^2+c^2)]$$

Le « dénominateur du numérateur » en question étant la constante n^2, on peut se contenter de travailler au niveau du « *numérateur du numérateur* ». On voit que cette expression est positive si , à la fois « a et d » sont positifs et « b et c » nuls (ce qui correspond à un cas favorable) , mais il se trouve qu'elle est également positive si « a et d » sont nuls et « b et c » positifs (cas d'une configuration non favorable).

Il apparaît donc évident que si l'on veut obtenir une configuration où la valeur de (f 98) soit négative , il faut Minimiser la fonction des quatre inconnues (a,b,c,d) suivante :

$$\min_{a,b,c,d} \; 2(ad-bc)\left(a^2+d^2-(b^2+c^2)\right)$$

vérifiant les contraintes :

(i) $a + d + b + c = n$ (où n est constante)

(ii) $0 \le a \le n, \quad 0 \le b \le n, \quad 0 \le c \le n, \; 0 \le d \le n$

Soit en utilisant une approche Lagrangienne :

$$\underset{abcd}{\text{Min}} \quad 2\,(ad - bc)\,[\,(a^2 + d^2) - (b^2 + c^2)\,] - \lambda(a + b + c + d - n)$$

vérifiant :

(ii) $0 \le a \le n, \quad 0 \le b \le n, \quad 0 \le c \le n, \; 0 \le d \le n$

Un raisonnement de bon sens va nous permettre de simplifier la résolution de ce Problème d'optimisation non linéaire à 4 variables[16]. En effet on constate sur la formule (f98) que pour que cette fonction F(a,b,c,d) soit négative, il faut que l'on ait simultanément:

$(ad - bc) < 0$ et $[a^2 + d^2 - (b^2 + c^2)] > 0$ (cas n°1)

ou $(ad - bc) > 0$ et $[a^2 + d^2 - (b^2 + c^2)] < 0$ (cas n°2)

On voit que l'on peut donc séparer le comportement de « a » et « d » de celui de « b » et « c ». En effet pour Minimiser F(a,b,c,d) :
* Il faut pour le couple {a, d} que l'on ait : la quantité produit « a.d » minimale et la quantité «a^2+d^2» maximale (ou l'inverse cas n°2)
* Il faut à l'inverse pour le couple {b,c} : que la quantité produit « b.c » soit maximale et la quantité «b^2+c^2» minimale (ou l'inverse cas n°2)

Nous choisissons ici le cas n°1 qui va nous garantir une solution où « d » sera maximum . En supposant que l'on ait arbitrairement découpé n en deux parties quelconques n_1 et n_2 telles que :

(i) $a + d = n_1$

(ii) $b + c = n_2$

(iii) avec $n_1 + n_2 = n$ (où n est constante)

(ii) $0 \le a \le n, \quad 0 \le b \le n, \quad 0 \le c \le n, \; 0 \le d \le n$

Du fait que $(a+d)^2 = a^2+d^2-2ad$
on voit que : si $(a+d) = n_1$=constante, alors il y a équivalence entre : Maximiser (a^2+d^2) et Minimiser (2ad) ou de façon duale : Maximiser (2ad) et Minimiser (a^2+d^2)
On déduit des remarques faites précédemment que : rendre « ad » minimum rendra automatiquement «a^2+d^2» maximum et de même pour le couple {b,c} rendre le couple « cd » maximum rendra la quantité «b^2+c^2» minimale

➨Dans le cas particulier du couple {b,c} puisque l'on veut rendre la quantité produit «bc » maximale, on va se servir d'un résultat qui découle des remarques calculatoires faites précédemment à savoir : **«** le produit de 2 nombres dont la **somme S est constante** est **maximum** si ces deux nombres sont égaux **»,** la conséquence pour notre problème est que : quelle que soit la valeur n_2 , on doit avoir $b = c = n_2/2$ à l'optimum.

➨De la même façon le produit de 2 nombres dont la **somme S est constante** est **minimum** si l'un des deux nombres est égal à S et l'autre est nul. **»,** la conséquence pour notre problème est que : quelle que soit la valeur n_1 , on doit avoir a ou d=0 à l'optimum . On vient donc de voir qu'au moins l'une ou l'autre des valeurs « a » et « d » est nulle (nous choisissons a=0) et que de plus « b = c » .

Le problème que l'on veut résoudre s'écrit maintenant sous une forme certes non linéaire mais très simplifiée :

$$\underset{a,b}{\text{Min}} \quad 2(-b^2)\,[\,d^2 - 2b^2\,]$$

vérifiant les contraintes :

[16] On ne pose pas encore à ce niveau une contrainte d'intégrité des quantités {a,b,c,d}, qu'on suppose continues dans un premier temps.

(i) $d + 2b = n$ (où n est constante)

(ii) $0 \le d \le n, \quad 0 \le b \le n,$

(iii) $d^2 \ge 2b^2$ (cette contrainte garantit la négativité du produit)

La contrainte (iii) alliée à la contrainte (i) impliquent immédiatement que :

$$b \le \frac{n}{2+\sqrt{2}} = 0{,}2929\ n$$ (nous nous en resservirons ultérieurement)

Du fait que l'on n'a pas encore introduit la contrainte d'intégrité, la solution du problème précédent revient à résoudre un programme non linéaire en (d, b) à variables continues positives. En utilisant la contrainte (i) et en remplaçant « b par x » (pour se ramener à des notations familières et classiques de variables continues inconnues) , on peut ramener ce problème à un programme d'optimisation non linéaire (du 4ème degré) à 1 seule variable x qui s'écrit :

$$\underset{x}{\text{Max}} \quad F(x) = 2\,x^2[(n-2x)^2 - 2x^2]$$

vérifiant

$$\text{(iii)}\ 0 \le x \le \frac{n}{2+\sqrt{2}}$$

(On passe du Min au Max, en enlevant le signe «- » devant la fonction)

Etudions les conditions d'optimalité du premier et du deuxième ordre par rapport à x, on obtient :

((i)) $\dfrac{dF(x)}{dx} = 2\left[2n^2 x - 12\,n x^2 + 8x^3\right] = 0$ soit $2x[n^2 - 6nx + 4x^2] = 0$

((ii)) $\dfrac{\partial^2 F(x)}{\partial x^2} = \left[2n^2 - 24nx + 24x^2\right] \le 0$ cette relation est vraie du fait que x vérifie la condition (iii)

La première condition ((i)) (condition du premier ordre) signifie que le x* cherché doit annuler la première dérivée de F(x) par rapport à x .

La deuxième condition ((ii)) (condition du second ordre) signifie, du fait que $\dfrac{\partial^2 F(x)}{\partial x^2} \le 0$, que la fonctionnelle F(x) est concave par rapport à x, donc que si x* est trouvé, ce sera un maximum unique pour F(x) (c'est justement ce que l'on cherche) et non un minimum. Puisqu'à l'optimum on doit avoir x* vérifiant la condition du 1er ordre ((i)), on est amené à résoudre l'équation du second degré en x suivante :

$$4x^2 - 6nx + n^2 = 0$$

dont la solution qui nous intéresse est donnée par :

$$x^* = \frac{3n - \sqrt{5n^2}}{4} = (3-\sqrt{5})\frac{n}{4} = 0{,}19098\ n$$

en effet la deuxième racine :

$$x'^* = \frac{3n + \sqrt{5n^2}}{4} = (3+\sqrt{5})\frac{n}{4} = 1{,}3090\ n$$

ne vérifie pas la condition (iii) et est donc éliminée, puisque :

$$0 \le x \le \frac{n}{2+\sqrt{2}}$$

n'est pas compatible avec le fait que x'*est supérieur à n, x* peut également s'écrire :

$$x^* = \frac{4-(1+\sqrt{5})}{4}n = \left[1 - \frac{\Phi}{2}\right]n \quad \text{où} \quad \Phi \quad \text{est le "Nombre d'Or"} = \frac{1+\sqrt{5}}{2} = 1{,}61803$$

Rappelons que nous voulons calculer la valeur Minimum par rapport à d, b , c de la quantité :

$$F(d,b,c) = 2\,(-bc)\,[\,(d^2) - (b^2 + c^2)]$$

sous la contrainte:

(i) $d + b + c = n$

comme nous avions vu que : b =c et puisque : b* = x* à l'optimum, on a donc :

$$b^* = c^* = \left[1 - \frac{\Phi}{2}\right] n \quad \text{où} \quad \Phi \quad \text{est le "Nombre d'Or"}$$

de ce fait comme d= n - 2b , il vient : $d^* = (\Phi - 1)n$

On vérifie bien que :

$(d^*)^2 \geq 2(b^*)^2$ car $(\Phi - 1)^2 n^2 \geq 2n^2 (1 - \frac{\Phi}{2})^2$ puisque $\Phi(1 + \frac{1}{\sqrt{2}}) \geq 1 + \sqrt{2}$

En remplaçant d*, b*,c* par leurs valeurs dans F(d,b,c) il vient :

$$F(d^*, b^*, c^*) = -2\left(1 - \frac{\Phi}{2}\right)^2 \left[(\Phi - 1)^2 - 2(1 - \frac{\Phi}{2})^2\right] n^4 = -2\left[\frac{5\Phi}{8} - 1\right] n^4 = -0{,}0225424859 \quad n^4$$

Pour obtenir cette expression simplifiée, on a utilisé les propriétés suivantes du Nombre d'Or :

$$\begin{cases} \Phi^2 = \Phi + 1 \\ \Phi^3 = \Phi^2 + \Phi = 2\Phi + 1 \\ \Phi^4 = \Phi^3 + \Phi^2 = 3\Phi + 2 \end{cases}$$

Nous rappelons que la valeur trouvée ici est la valeur minimale de F(d,b,c) pour d, b, c, à valeurs continues, nous n'avons pas tenu compte des conditions d'intégrité de d, b, c .
Ceci nous aurait obligé à utiliser une solution par Programmation non linéaire en nombres entiers ; en utilisant par exemple un algorithme dérivé des approches « cutting planes » de M. Grötschel, M. Jünger, G. Reinelt (1982)».

Outre que cette approche ne permet pas d'avoir des solutions explicites et que les calculs sont très complexes, même pour des valeurs faibles du nombre de variables, elle sortirait du cadre de cet article. Une solution approchée, et tout à fait réaliste, pour trouver la meilleure solution en nombres entiers du problème précédent, vu que le nombre de situations possibles à explorer est très faible puisqu'il n'y a que deux degrés de liberté ici, consiste alors à prendre les valeurs entières :

$\tilde{d}, \tilde{b}, \tilde{c}$ telles que : $|\tilde{d} - d^*|$ et $|\tilde{b} - b^*| = |\tilde{c} - c^*|$ soient minimales

et telles que : $|\tilde{d} - d^*| + 2|\tilde{b} - b^*|$ ait une valeur minimale sous la contrainte : $\tilde{d} + 2\tilde{b} = n$

d' où $\tilde{b} = \tilde{c}$ = partie entière de $\left[1 - \frac{\Phi}{2}\right] n = 0{,}19098$ n et idem pour $\tilde{d}$ = partie entière de $(\Phi - 1)N = 0{,}618033$ n

choisira la solution pour laquelle :

$\tilde{d}$ et $\tilde{b}$ soient telles que l'on ait Min $|\tilde{d} - d^*| + 2|\tilde{b} - b^*|$ sous la contrainte $\tilde{d} + 2\tilde{b} = n$

A titre d'exemple , prenons n=24 , alors on trouve la solution entière optimale représentable par le tableau suivant :

	y	Non y
x	0	5
Non x	5	14

Et F(d,b,c) « optimale[17], » (valeur minimale) , calculée sur ce tableau est donnée par : **-7300**
Rappelons que la valeur théorique optimale continue aurait été donnée par : - 0,0225424859 n^4= **- 7479, 055** , l'erreur d'approximation est ici de l'ordre de 2,3%.
On vérifie facilement que toute autre solution est moins bonne, en effet prenons par exemple :

	y	Non y

[17] Rappelons nous qu' il faudrait pour calculer à sa juste valeur le numérateur du coefficient de Lerman, diviser cette quantité par n². mais ceci a peu d'importance car cette quantité n², disparaîtra en fait du numérateur de L(x,y) dès que nous le calculerons en simultanéité avec le dénominateur.

x	0	4
Non x	4	16

La valeur associée de F(a,b,c) = - **7168** dans ce cas.
Considérons maintenant le tableau suivant, légèrement modifié par rapport au précédent:

	y	Non y
x	0	5
Non x	3	16

La valeur de F(d,b,c) associée à ce deuxième tableau est égale à : **-6660** (on constate bien sur ce petit exemple que dès que « $b \neq c$ » , la valeur « d » restant inchangée , la valeur de la fonction est inférieure en valeur absolue au cas où « $b = c$ »).

Revenons maintenant à l'expression du critère L(X,Y) complet et non plus au simple Numérateur qui concerne **le seul « Critère d'Ecart carré à l'Indépendance ».** Pour évaluer la valeur Minimale du Coefficient d'association de Lerman sur un tableau de contingence (2x2), il suffit de remplacer dans la formule (f 94) : $n_{..}$ par P, p et q par 2 et $n_{u\,v}$ par l'une des valeurs : { a, b, c, d }.

Le dénominateur de la formule (f94) s'écrit alors :

$$\sqrt{[(a+b)^2+(c+d)^2][2(a+b)(c+d)][(a+c)^2+(b+d)^2][2(a+c)(b+d]}$$

Soit encore sous une forme plus symétrique :

$$2\sqrt{[(a+b)^2+(c+d)^2][(a+c)^2+(b+d)^2][(a+c)(b+d)(a+c)(b+d)]}$$

En vérité, cette formule est à diviser par n^2, tout comme pour le numérateur de la formule (f95), de ce fait la quantité n^2 disparaît au numérateur et au dénominateur, de même pour la constante multiplicative 2. On obtient au final l'expression suivante (après les remplacements proposés, des simplifications des formules complexes obtenues et enfin une factorisation ,) :

$$\boxed{L(x,y) = \frac{N_1(x,y)}{D_1(x,y)} \times \frac{N_2(x,y)}{D_2(x,y)}}$$

où :

$$\frac{N_1(x,y)}{D_1(x,y)} = \frac{[ad - bc]}{\sqrt{[a+b]\,[a+c]\,[b+d][c+d]}}$$

et où

$$\frac{N_2(x,y)}{D_2(x,y)} = \frac{[a^2+d^2-(b^2+c^2)]}{\sqrt{([a+b]^2+[c+d]^2)}\sqrt{([a+c]^2+[b+d]^2)}}$$

L'indice de Lerman Modifié est donc le Produit de deux indices dont l'un est l'indice Tetrachorique que nous avons déjà défini , l'autre étant un indice que nous allons expliciter dans les paragraphes suivants. Mais revenons au calcul du Minimum de L(x,y)

Comme nous avons vu que dans le cas où le Numérateur atteint sa valeur minimale soit « a » soit « d » devait être égal à 0 , posons « a=0 », les valeurs précédentes se simplifient en :

$$\frac{N_1(x,y)}{D_1(x,y)} = \frac{[-bc]}{\sqrt{bc\,[b+d][c+d]}} \qquad \frac{N_2(x,y)}{D_2(x,y)} = \frac{[d^2-(b^2+c^2)]}{\sqrt{[b^2+(c+d)^2]\left(c^2+(b+d)^2\right]}}$$

En utilisant la contrainte : d+2b=n et en remplaçant de nouveau b=c par x , il vient le problème d'optimisation suivant :

(f 99) $$\underset{x}{\text{Max}} \quad F(x) = \frac{\left[n^2\ x^2\ -4\ n\ x^3 + 2\,x^4\right]}{\sqrt{(n-x)^2\,x^2}\ \sqrt{\left[(n-x)^2 + x^2\right]\left[(n-x)^2 + x^2\right]}} = \frac{x\,[n^2 - 4\,n\,x + 2\,x^2]}{(n-x)\ \left((n-x)^2 + x^2\right)}$$

vérifiant :

(ii) $$0 \le x \le \frac{n}{2+\sqrt{2}}$$

Comme nous avons affaire maintenant à une fonction d'une seule variable , il suffit d'appliquer la condition du premier ordre à la fonctionnelle F(x) précédente.

Comme F(x) se présente comme le rapport $\frac{u(x)}{v(x)}$ de deux fonctions de x .

La condition de dérivation du premier ordre s'écrit alors après simplifications:

$$\frac{dF(x)}{dx} = \frac{8x^3 - 14nx^2 + 8n^2x - n^3}{(n-x)^2[(n-x)^2 + x^2]^2} = 0 .$$

Il faut donc résoudre l'équation du 3ème degré[18] suivante pour obtenir x* :

(f100) $$8x^3 - 14n\,x^2 + 8\,n^2x - n^3 = 0$$

sous réserve que : $x \neq n$ (ce qui est d'office vrai voir contrainte (ii) précédente)

On obtient la solution optimale exacte: x*=0,171350939.n et de fait a*=n-x*=0,6572982.n Dans la formule (f 98) on constate néanmoins que le coefficient numérique 2,210929 est assez voisin de $\sqrt{5}$ et plus exactement :

$$0{,}210929 \cong \sqrt{5} - 0{,}025 = \sqrt{5}\left(1 - \left[\frac{\sqrt{5}}{10}\right]^3\right)$$

D'où après développements une deuxième approximation de x* fournie par :

(f 102) $$x^* \cong \frac{n}{12}[7 - (\sqrt{5} - 0{,}025)\sqrt{5}] = \frac{n}{12}[2 + \frac{\Phi}{20} - \frac{1}{40}] = \frac{n}{480}[79 + 2\,\Phi] = 0{,}17135141\ n$$

On constate ici que l'erreur n'intervient qu'au bout du 6ème chiffre après la virgule, ce qui nous permettra d'utiliser la formule (f 100) comme formule explicite pour la suite.

En particulier, on aura en fin de compte :

$$b^* = c^* = [79 + 2\Phi]\frac{n}{480} \quad \text{et} \quad d^* = n - 2b^* = \frac{n}{240}[161 - 2\Phi]$$

Comme précédemment, ces valeurs optimales ont été calculées pour un environnement de valeurs continues et sans tenir compte des contraintes d'intégrité.

A titre d'exemple , prenons n=24 , on trouve alors :

$\tilde{b} = \tilde{c} =$ partie entière de $[79 + 2\Phi]\frac{24}{480} = 4{,}1124$ et $\tilde{d}$ = partie entière de $[161 - 2\Phi]\frac{24}{240} = 15{,}7752$

d'où $\tilde{b} = \tilde{c} = 4$ et $\tilde{d} = 16$

La solution entière optimale est donc donnée par le tableau suivant :

---

[18] La solution de ce problème s'obtient en utilisant le procédé de Cardan- Tartaglia, pour une équation de la forme : $\alpha\,x^3 + \beta x^2 + \gamma x + \delta = 0$ avec $\alpha \neq 0$,. On sait alors que la solution Z* (réelle) de Z^3+r Z+ t est donnée par : $x^* = Z^* + \frac{14}{24}n$, on pose ici encore $\sqrt{5} = 2\Phi - 1$

il vient : $x^* = \frac{n}{12}[7 - 2{,}21092949\ \sqrt{5}] = \frac{n}{12}[7 - 2{,}210929\ 49\ (2\,\Phi - 1)] = \frac{n}{12}[9{,}2109294\ - 4{,}42185898\ \Phi]$

	y	Non y
x	0	4
Non x	4	16

Et L(x,y) minimum absolu calculé pour n=24 est donné par :

$$L(\tilde{x},\tilde{y}) = \frac{-4.4[16^2 - 2.16]}{20.4.[20^2 + 4^2]} = -0{,}10769$$

rappelons que la valeur théorique optimale continue aurait été donnée en calculant L(x,y) sur un tableau théorique où au lieu de '4' on aurait : '4,1124' , et au lieu de '16' , on aurait '15,775,' soit :

$$L(x^*, y^*) = -\frac{3636{,}50}{33730{,}78} = -0{,}107811 \quad \text{cette valeur}$$

on voit qu'elle est très légèrement supérieure à la valeur précédente l'écart $L(x^*, y^*) - L(\tilde{x},\tilde{y}) = 0{,}00011$,

n'étant que de 11/10000 .

En fait cette valeur minimale théorique L(x*,y*) est indépendante de n, en effet : posons b*=ω.n (avec ω=0,1713509 et d*=η n (avec η= 0,6572982) (valeurs données en (f102) plus haut). Il vient :

$$L(x^*, y^*) = \frac{-[b^*]^2\left[[d^*]^2 - 2[b^*]^2\right]}{(n-b^*)b^*\left[[n-b^*]^2 + [b^*]^2\right]} = -\frac{\omega^2 n^2[\eta^2 n^2 - 2\omega^2 n^2]}{(n-\omega n)\omega n\left[(n-\omega n)^2 + \omega^2 n^2\right]} = -\frac{\omega^2[\eta^2 - 2\omega^2]}{(1-\omega)\omega[(1-\omega)^2 + \omega^2]}$$

On voit que les valeurs n disparaissent et que d'autre part comme η=(1- 2ω) on peut tout exprimer en fonction de ω.

C'est d'ailleurs ce que nous avions fait lors de l'optimisation en jouant sur l'inconnue x (voir formule (f 96)).

Après simplification du Numérateur et du Dénominateur par n et élimination de η , il vient :

$$L(x^*, y^*) = -\frac{\omega[1 - 4\omega + \omega^2]}{(1-\omega)[1 - 2\omega + 2\omega^2]}$$

Il suffit de remplacer « ω » par sa valeur 0,1713509 dans l'expression précédente, il vient :

$$L_{Min}(x^*, y^*) = -\frac{0{,}17135\ [1 - 4\ .0{,}17135\ + 2\ .0{,}17135^{\ 2}]}{0{,}828649[1\ -2\ \ 0{,}171350\ + 2\ .0{,}17135^{\ 2}]} = -\frac{0{,}063968}{0{,}5933296} = -0{,}1078119$$

Cette valeur indépendante de n est le **minimum théorique** de **l'indice de Lerman modifié** sur un tableau de contingence (2x2) . On voit ici la complexité de la situation, car les valeurs minimisant le Numérateur de L(x,y) c'est à dire celles pour lesquelles le critère d' « Ecart Carré à l'Indépendance » est minimum, ne sont pas celles qui minimisent L(x,y) .En effet reprenons ces valeurs , on avait :

$$b^* = c^* = \left[1 - \frac{\Phi}{2}\right] n \quad \text{où } \Phi \text{ est le "Nombred'Or" et } d^* = (\Phi - 1)n$$

Remplaçons ces valeurs dans l'expression de L(x*,y*) donnée précédemment en fonction de a*,b*,c*, il vient :

$$L(x^*, y^*) = \frac{-[b^*]^2\left[[d^*]^2 - 2[b^*]^2\right]}{(n-b^*)b^*\left[[n-b^*]^2 + [b^*]^2\right]} = -\frac{\left[1 - \frac{\Phi}{2}\right]\left((\Phi-1)^2 - 2\left[1 - \frac{\Phi}{2}\right]^2\right)}{\frac{\Phi}{2}\left(\left[\frac{\Phi}{2}\right]^2 + \left[1 - \frac{\Phi}{2}\right]^2\right)} = -\frac{5\Phi - 8}{3\Phi - 4} = -0{,}1055728$$

La formule précédente peut d'ailleurs s'exprimer selon la suite de Fibonacci (liée elle même au Nombre d'Or): en effet si l'on pose F_1=1, F_2=1, F_3=2, F_4=3, F_5=5, F_6=8, F_7= 13 ….. on voit que la formule précédente s'écrit :

$$L(x^*, y^*) = -\frac{5\Phi - 8}{3\Phi - 4} = \frac{F_5\Phi - F_6}{F_4\Phi - F_5 + 1}$$

On voit que cette valeur, même si elle est très proche de la précédente (0,1078118 versus 0,1055728) , n'est pas optimale, même si elle optimisait par ailleurs le numérateur seul. Si on exprimait la solution Minimale totale par rapport au nombre d'Or on voit que seul le dénominateur serait légèrement modifié en effet :

$$L_{Min}(x^*, y^*) = -0{,}1078119 = \frac{F_5\Phi - F_6}{F_4\Phi - F_5 + 0{,}9829}$$

Revenons maintenant au calcul de l'Indice de Lerman modifié sur le vrai tableau Tetrachorique et non sur un tableau (2x2) quelconque, dès lors, comme précédemment, pour évaluer la valeur du Coefficient d'association de Lerman sur le tableau de contingence (2x2) Tetrachorique, il suffit de remplacer : $n_{..}$ par P, p et q par 2 et $n_{u\,v}$ par l'une des valeurs : { 11(x,y), 10(x,y), 01(x,y), 00(x,y)}, on obtient après les remplacements proposés, des simplifications des formules complexes obtenues et une factorisation finale, qui se traduit par l'expression suivante:

$$L(x,y)=\frac{N_1(x,y)}{D_1(x,y)}\frac{N_2(x,y)}{D_2(x,y)}$$

où :

$$\frac{N_1(x,y)}{D_1(x,y)}=\frac{[11(x,y)00(x,y)-01(x,y)10(x,y)]}{\sqrt{[11(x,y)+01(x,y)][00(x,y)+10(x,y)][11(x,y)+10(x,y)][00(x,y)+01(x,y)]}}$$

et

$$\frac{N_2(x,y)}{D_2(x,y)}=\frac{[11^2(x,y)+00^2(x,y)-(10^2(x,y)+01^2(x,y))]}{\sqrt{[[11(x,y)+01(x,y)]^2+[00(x,y)+10(x,y)]^2][[11(x,y)+10(x,y)]^2+[00(x,y)+01(x,y)]^2]}}$$

On reconnaît dans le rapport :

$$\frac{N_1(x,y)}{D_1(x,y)}=\frac{[11(x,y)00(x,y)-01(x,y)10(x,y)]}{\sqrt{[11(x,y)+01(x,y)][00(x,y)+10(x,y)][11(x,y)+10(x,y)][00(x,y)+01(x,y)]}}$$

l'indice de similarité Tetrachorique $S_T(x,y)$, défini par la formule (f 71).

Mais qu'en est-il de l'indice suivant ? :

$$\frac{N_2(x,y)}{D_2(x,y)}=S_L(x,y)=\frac{[11^2(x,y)+00^2(x,y)-(10^2(x,y)+01^2(x,y))]}{\sqrt{[[11(x,y)+01(x,y)]^2+[00(x,y)+10(x,y)]^2][[11(x,y)+10(x,y)]^2+[00(x,y)+01(x,y)]^2]}}$$

C'est un indice (de type corrélatif) non rencontré jusqu'ici , qui est donc **nouveau** et qui vaut :

- 1 si 10(x,y)=01(x,y)=0 (configuration où les deux profils sont identiques)
- -1 si 11(x,y)=00(x,y)=0 (configuration où les deux profils sont systématiquement en opposition)

Pour avoir un indice $S'_L(x,y)$ variant de 0 à 1, nous devons faire la translation suivante :

$$S'_L(x,y)=\frac{1}{2}\left[\frac{N_2(x,y)}{D_2(x,y)}+1\right]$$

Ce nouvel indice vaut donc :

- 1 si 10(x,y)=01(x,y)=0 (configuration où les deux profils sont identiques)

- ½ si $11^2(x,y)+00^2(x,y)= 01^2(x,y) + 10^2(x,y)$ (configuration d'équilibre quadratique entre les configurations de « matching » et les configuration de « non matching ») . Nous avons déjà vu page 72 que cette situation était équivalente à :

  $$P\,[11(x,\,y) + 00(x,\,y) - (10(x,\,y) + 01(x,\,y)] = 2.[11(x,\,y)00(x,\,y) - 01(x,\,y)10(x,\,y)]$$

  c'est à dire si :
  - soit les 4 valeurs sont égales : $11(x,y)=00(x,y)=10(x,y)=01(x,y)$
  - soit également pour toutes les décompositions entières paramétrées, solutions de l'équation Diophantienne associée (voir page 72 où une telle solution paramétrée est donnée ; et voir F. Marcotorchino , P. Michaud (1981) pour d'autres solutions)
- Il vaut : 0 si $11(x,y)=00(x,y)=0$.

Comportement de l'indice $S_L(x,y)$ en cas de Disjonction Complète

En cas de disjonction complète on a $10(x,y)=01(x,y)=(m-11(x,y))$ et $00(x,y)=P+11(x,y)-2m$
L'indice (forme corrélative) s'écrit alors :

$$S_L(x,y) = \frac{\left[11^2(x,y) + (P+11(x,y)-2m)^2 - 2(m-11(x,y))^2\right]}{\left[[m]^2 + [P-m]^2\right]} = \frac{[P-m]^2 + m^2 - 2P[m-11(x,y)]}{[m]^2 + [P-m]^2} = 1 - \frac{2P[m-11(x,y)]}{[m]^2 + [P-m]^2}$$

et l'indice de similarité vrai, s'écrit alors :

$$S'_L(x,y) = \frac{1}{2}\left[1 - \frac{2P[m-11(x,y)]}{[m]^2 + [P-m]^2} + 1\right] = 1 - \frac{P[m-11(x,y)]}{m^2 + [P-m]^2}$$

si l'on pose $P=m\,\bar{p}$ où $\bar{p}$ est le nombre moyen de modalités de l'ensemble des variables
l'indice précédent se simplifie en :

$$S'_L(x,y) = 1 - \frac{P[m-11(x,y)]}{m^2 + [P-m]^2} = 1 - \frac{m\,\bar{p}\,[m-11(x,y)]}{m^2 + [m\bar{p}-m]^2} = 1 - \frac{\bar{p}[m-11(x,y)]}{m[1+(\bar{p}-1)^2]} = 1 - \frac{\bar{p}}{1+(\bar{p}-1)^2}\left[1 - \frac{11(x,y)}{m}\right]$$

- La valeur 1 pour $S'_L(x,y)$ est obtenue quand $m=11(x,y)$ soit en cas de « matching » total entre les profils de x et y
- La valeur minimale n'est pas 0 dans le cas de disjonction totale car on ne peut pas avoir $00(x,y)=0$, en fait dans ce cas $00(x,y)=P-2m$ et $11(x,y)=0$, ce qui donne une valeur minimale égale à :

  $$S'_L(x,y) = 1 - \frac{Pm}{m^2 + [P-m]^2} = 1 - \frac{\bar{p}}{1+(\bar{p}-1)^2} = \frac{(\bar{p}-1)(\bar{p}-2)}{(\bar{p}-1)^2+1}$$

  - Si $\bar{p}=1$ ou 2,

    on constate que l'indice vaut 0 : il s'annule réellement dans le cas où la valeur moyenne vaut 1 du fait que nous sommes alors dans la situation de tableau de « présence –absence » vrai, il vaut 0 également dans le cas où la valeur moyenne du nombre de modalités vaut 2 (ce qui se

produit dans le cas de dédoublement disjonctif de variables binaires pures.,

- Si $\bar{p}$ croît fortement, $S'_L(x,y)$ tend vers 1

De ce fait la borne de Solomon Fortier n'est donc pas égale à ½ mais au milieu de l'intervalle allant de la valeur minimale ci dessus à la valeur 1, soit :

$$\text{Borne}_{SF} = \frac{1}{2}\left[1 - \frac{Pm}{m^2 + [P-m]^2} + 1\right] = 1 - \frac{Pm}{2\left(m^2 + [P-m]^2\right)}$$

la valeur de 11(x,y) pour satisfaire cette borne doit donc vérifier l'inégalité :

$$S'_L(x,y) = 1 - \frac{P[m - 11(x,y)]}{m^2 + [P-m]^2} \geq 1 - \frac{Pm}{2\left(m^2 + [P-m]^2\right)} \quad \Rightarrow \quad 11(x,y) \geq \frac{m}{2}$$

11(x,y) doit donc être **supérieur à la majorité de Condorcet** .

Conclusion sur l'Indice de Lerman modifié

Nous venons de voir qu'à partir de l'indice d'association de Lerman (version modifiée) , nous avons pu générer deux indices de similarités différents et distincts définis à partir de la formule :

(f103) $$L(x,y) = S_T(x,y).S_L(x,y)$$

Le premier est l'indice Tetrachorique (ou de Bravais –Pearson), l'autre un peu plus complexe vient d'être étudié ci dessus sous le nom d'Indice de similarité de Lerman modifié.
Chacun des deux indices vérifie une condition de Solomon-Fortier impliquant une majorité à m/2. Chacun des deux sous leurs formes corrélatives varient de –1à +1.
Chacun séparément a du sens, en revanche, leur produit n'en a pas en tant qu'indice de similarité.

4.4.5.4 Indice de similarité déduit du Coefficient « λ » de Goodman-Kruskal

Goodman et Kruskal on proposé en 1954 (voir leur livre déjà cité (1954)) un indice d'association qu'ils ont appelé « indice de prédiction optimale symétrique ». Cet indice qui possède d'excellentes propriétés statistiques en structure est malheureusement formé à partir d'expressions mathématiques faisant intervenir des Maxima , donc difficiles à calculer dans un processus automatique à caractère systématique, comme c'est le cas dans les problèmes d ' « Association Maximale ». Nous donnons ci-dessous son expression, calculable sur tableau de contingence général :

(f 104) $$\lambda(x,y) = \frac{\sum_{u=1}^{p} \underset{v}{\text{Max}}\, n_{uv} + \sum_{v=1}^{q} \underset{u}{\text{Max}}\, n_{uv} - \underset{u}{\text{Max}}\, n_u - \underset{v}{\text{Max}}\, n_v}{2n_{..} - \underset{u}{\text{Max}}\, n_u - \underset{v}{\text{Max}}\, n_v}$$

- L'indice λ(x,y) varie entre 0 et 1 :

- Il vaut 1 en cas d'Association complète
- Il vaut 0 en cas d'Indépendance statistique

Comme précédemment, pour évaluer la valeur de ce Coefficient d'association $\lambda(x,y)$ de Goodman Kruskal sur le tableau de contingence (2x2) Tetrachorique, il suffit de remplacer : $n_{..}$ par P, p et q par 2 et n_{uv} par l'une des valeurs : { 11(x,y), 10(x,y), 01(x,y), 00(x,y)}, on obtient après les remplacements proposés, les résultats suivants :

$$\sum_{u=1}^{p} \underset{v}{Max}\ n_{uv} + \sum_{v=1}^{q} \underset{u}{Max}\ n_{uv} = P + \frac{1}{2}\Big[\big|11(x,y-10(x,y)\big| + \big|01(x,y)-00(x,y)\big| + \big|11(x,y)-01(x,y)\big| + \big|10(x,y)-00(x,y)\big|\Big]$$

$$\underset{u}{Max}\ n_{u.} + \underset{v}{Max}\ n_{.v} = P + \frac{1}{2}\Big[\big|\,11(x,y)+10(x,y)-[(01(x,y)+00(x,y)]\,\big| + \big|\,11(x,y)+01(x,y)-[10(x,y)+00(x,y)]\,\big|\Big]$$

Mais pour nous éviter des calculs fastidieux, comme le tableau Tetrachorique est un tableau (2x2) particulier revenons au cas d'un tableau (2x2) général sous la forme :

	y	Non y	
x	a	b	a+b
Non x	c	d	c+d
	a+c	b+d	N

Reformulons les expressions précédentes au moyen des quantités {a, b , c, d }, il vient :

(i) $$A = \sum_{u=1}^{p} \underset{v}{Max}\ n_{uv} + \sum_{v=1}^{q} \underset{u}{Max}\ n_{uv} = N + \frac{1}{2}\Big(|a-b|+|a-c|+|b-d|+|c-d|\Big)$$

(ii) $$B = \underset{u}{Max}\ n_{u.} + \underset{v}{Max}\ n_{.v} = N + \frac{1}{2}\Big(|(a+b)-(c+d)|+|(a+c)-(b+d)|\Big)$$

Comme il existe a priori 4 ! =24 ordres totaux possibles (il y a isomorphisme entre l'ensemble des ordres totaux et l'ensemble Ω_n={Groupe symétrique des permutations de n objets}), ceci sans compter les possibilités de préordres associées. Nous illustrerons la valeur du coefficient $\lambda(x,y)$ de Goodman Kruskal en caractérisant les familles d'ordres induits et en le calculant par rapport aux séquences des lettres {a,b,c,d}, ou ce qui est équivalent à la structure des permutations associées.

Par convention on notera par exemple <abcd> la permutation ou l'ordre total généré par la séquence : a>b>c>d c'est à dire : 11(x,y)>10(x,y)>01(x,y)>00(x,y)

Etudions alors les différentes configurations possibles :

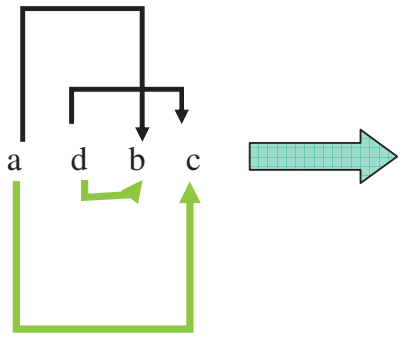
- Cas de la **FamilleI** des ordres respectant la séparation des couples blocs {a,d} et {b,c}, elle regroupe deux sous-familles :
  - **La sous famille N°1**, elle est constituée des ordres ou permutations :
    <ad bc>, <ad cb>,<da bc>, <da cb>
    Pour cette famille la valeur de A est identique pour tous ces ordres, car toutes les valeurs absolues de la quantité A sont désambiguïsées, elle est donnée par:

$$A = N + [a + d - (b + c)]$$

la valeur de B ne peut être obtenue directement , elle nécessite des comparaisons par rapport aux structures induites par les valeurs absolues non désambiguïsées . Par exemple si l'on prend l'ordre <ad bc> on voit que

B=N+ (a-d) si (a+c)> (b+d) et B=N+(b-c) dans le cas inverse .

En fait la construction se fait autour du principe représenté ci-dessous caractéristique de la permutation <adbc> et facilement généralisable.



a+b>d+c ⇒ calcul sûr

a+c ? d+b ⇒ comparaison à faire

Ainsi pour la permutation <ad cb> B vaudra soit N+ (a-d) soit N+ (c-b)

Pour la permutation <dabc>, B vaudra : soit N+ (d -a) soit N + (b-c)

Pour la permutation <dacb>, B vaudra : soit N+ (d -a) soit N + (c-b), pour cette famille de permutations on voit que le coefficient $\lambda(x,y)$ vaudra :

$$\lambda(x, y) = \frac{A - B}{2N - B} = \frac{(a + d) - (b + c) - \Psi}{N - \Psi}$$

ψ valant soit (a-d), soit (b-c),soit (d-a), soit (c-b) suivant les cas avec de ce fait respectivement:

$$\lambda(x, y) = \frac{2d - (b + c)}{2d + (b + c)} \text{ ou } \lambda(x, y) = \frac{(a + d) - 2b}{(a + d) + 2c} \text{ ou } \lambda(x, y) = \frac{2a - (b + c)}{2a + (b + c)} \text{ ou } \lambda(x, y) = \frac{a + d - 2c}{a + d + 2b}$$

- Sous **Famille N°2** elle regroupe les ordres : <bcad>, <cbad>, <bcda>, <cb da>

Pour les 4 ordres de cette famille on a :

$A = N + [(b + c) - (a + d)]$ = N+((Item rang n°1+item rang n°2)- (item rang n°3 + item rang n°4))

Ainsi pour la permutation <bcad > : B vaudra soit N+ (a -d) soit N+(b -c)

Pour cette famille de permutations on voit que le coefficient $\lambda(x,y)$ vaudra :

$$\lambda(x, y) = \frac{A - B}{2N - B} = \frac{(b + c) - (a + d) - \Psi}{N - \Psi}$$

ψ valant soit (a-d), soit (b-c),soit (d-a), soit (c-b) suivant les cas avec de ce fait respectivement :

$$\lambda(x, y) = \frac{b + c - 2a}{b + c + 2d} \text{ ou } \lambda(x, y) = \frac{2c - (a + d)}{2c + (a + d)} \text{ ou } \lambda(x, y) = \frac{b + c - 2d}{(b + c) + 2a} \text{ ou } \lambda(x, y) = \frac{2b - (a + d)}{2b + (a + d)}$$

D'une façon générale pour toutes les permutations $\{ < \sigma_1\ \sigma_2\ \sigma_3 \sigma_4 > \in \text{Famille I} \}$

A= N+((Item rang n°1+item rang n°2)- (item rang n°3 + item rang n°4))

Soit $A=N+((\sigma_1+\sigma_2)-(\sigma_3+\sigma_4))$

et $B=N+(\sigma_1-\sigma_2)$ ou $B=N+(\sigma_3-\sigma_4)$

Dès lors d'une façon générale l'indice $\lambda(x,y)$ pour toute permutation appartenant à cette Famille I

vaudra : $\lambda(x,y)=\dfrac{2\sigma_2-(\sigma_3+\sigma_4)}{2\sigma_2+(\sigma_3+\sigma_4)}$ ou $\lambda(x,y)=\dfrac{\sigma_1+\sigma_2-2\sigma_3}{\sigma_1+\sigma_2+2\sigma_4}$

- Cas de la **Famille II** des ordres <u>imbriquant</u> les couples blocs {a,d} et {cd}, elle se compose de la famille des 8 ordres suivants :
  - <abdc>,<acdb>,<dbac>,<dcab>,<bacd>, <bdca>, <cabd>, <cdba> :

    pour cette famille, on peut montrer que :

    A= N + (item en 1ère position - item en 4ème position)

    ainsi on aura A= N + (a – c) pour la 1ere permutation <**a**bd**c**>

    A= N+ (a-b) pour la deuxième <**a**cd**b**>, A=N + (d-c) pour la troisième etc..

    Pour B on aura B=N+(item en 1ème position – item en 3ème) ou B=N+(item en 2ème – item en 4ème), ces deux valeurs possibles de B dépendant de la désambiguïsation de la valeur absolue restante.

    Ainsi B=N+(a-d) pour <**a**b**d**c> ou B=N+(b-c) pour cette même permutation <a**b**d**c**>

    De fait pour cette permutation on a les valeurs suivantes de $\lambda(x,y)$:

    $$\lambda(x,y)=\frac{A-B}{2N-B}=\frac{(a-c)-(a-d)}{N-(a-d)}=\frac{d-c}{2d+(b+c)} \text{ ou } \lambda(x,y)=\frac{(a-c)-(b-c)}{N-(b-c)}=\frac{a-b}{2c+(a+d)}$$

    D'une façon générale l'indice $\lambda(x,y)$ associé à une permutation :

    $<\sigma_1\,\sigma_2\,\sigma_3\sigma_4>\in$ Famille II

    Aura deux valeurs possibles soit:

    $$\lambda(x,y)=\frac{\sigma_3-\sigma_4}{2\sigma_3+(\sigma_2+\sigma_4)} \quad \text{ou} \quad \lambda(x,y)=\frac{\sigma_1-\sigma_2}{2\sigma_4+(\sigma_1+\sigma_3)}$$

- Cas de la **Famille III** des ordres tels que les structures {a,d}et {b,c} s'<u>englobent</u> ou s'<u>emboîtent</u> mutuellement , elle est composée des 8 ordres ou permutations suivantes :
  - <abcd>,<acbd>,<dbca>, <dcba>, <badc>,<bdac>,<cadb>,<cdab>

    pour cette famille <u>toutes les valeurs absolues sont désambiguïsées</u> et on aura uniquement une seule valeur pour B , en effet on peut vérifier que :

    A=N+(item en 1ère position – item en 4ème position)

    et B=N+(item en 1ère position – item en 4ème position)

    Comme le Numérateur du coefficient $\lambda(x,y)$ est égal à A-B et qu'ici A=B

    l'indice $\lambda(x,y)$ associé à une permutation : $<\sigma_1\sigma_2\sigma_3\sigma_4>\in$ Famille III

    aura une seule et unique <u>valeur nulle</u> , soit :

    $$\lambda(x,y)=0$$

. .

Revenons au tableau de contingence Tetrachorique proprement dit et appliquons les résultats précédents à deux configurations qui peuvent se produire fréquemment (en particulier dans le cas général d'un tableau de données de type « présence-absence » pour lequel il n'y a pas a priori de structure sur les éléments qui permettent d'éliminer des configurations). Choisissons par exemples les 3 configurations d'ordres totaux suivants :

N°1= 00(x,y)>11(x,y)>01(x,y)> 10(x,y)

N°2= 11(x,y)>10(x,y)>00(x,y)> 01(x,y)

N°3= 11(x,y)>10(x,y)>01(x,y)> 00(x,y)

- L'ordre N°1 appartient de façon évidente à la **Famille I**, il est équivalent à <dacb> avec les notations précédentes. Dès lors les valeurs possibles de l'indices sont données par :

$$\lambda(x,y)=\frac{2\sigma_2-(\sigma_3+\sigma_4)}{2\sigma_2+(\sigma_3+\sigma_4)} \quad \text{ou} \quad \lambda(x,y)=\frac{\sigma_1+\sigma_2-2\sigma_3}{\sigma_1+\sigma_2+2\sigma_4}$$

après identification de σ_1, σ_2, σ_3, σ_4 respectivement à 00(x,y), 11(x,y), 01(x,y), 10(x,y) il vient :

  - $\lambda(x,y)=\frac{2.11(x,y)-[01(x,y)+10(x,y]}{2.11(x,y)+01(x,y)+10(x,y)}$ si $00(x,y)+10(x,y)>11(x,y)+01(x,y)$

    (On retrouve ici un indice connu puisque l'indice de Dice est égal à : $S_d(x,y)=\frac{2.11(x,y)]}{2.11(x,y)+01(x,y)+10(x,y)}$)

    d'où $\lambda(x,y)=2S_d(x,y)-1$
  - $\lambda(x,y)=\frac{11(x,y)+00(x,y)-2.01(x,y)}{11(x,y)+00(x,y)+2.10(x,y)}$ si $00(x,y)+10(x,y)<11(x,y)+01(x,y)$

  Compte tenu de l'ordre induit , on voit que dans les deux cas de figure : $\lambda(x,y)\geq 0$

- L'ordre N°2 appartient, lui, à la **Famille II** des ordres à <u>imbrication</u>, il est équivalent à <abdc> . Les valeurs possibles de l'indice (si l'on identifie σ_1, σ_2, σ_3, σ_4 respectivement à 11(x,y), 10(x,y), 00(x,y), 01(x,y)) sont données de façon générale par :

$$\lambda(x,y)=\frac{\sigma_3-\sigma_4}{2\sigma_3+(\sigma_2+\sigma_4)} \quad \text{ou} \quad \lambda(x,y)=\frac{\sigma_1-\sigma_2}{2\sigma_4+(\sigma_1+\sigma_3)}$$

  on aura donc :

  - $\lambda(x,y)=\frac{00(x,y)-01(x,y)}{2.00(x,y)+10(x,y)+01(x,y)}$ si $00(x,y)+10(x,y)>11(x,y)+01(x,y)$
  - $\lambda(x,y)=\frac{11(x,y)-10(x,y)}{2.01(x,y)+11(x,y)+00(x,y)}$ si $00(x,y)+10(x,y)<11(x,y)+01(x,y)$

  On voit ici encore que dans les deux cas de figure précédents , compte tenu de l'ordre induit :
  $\sigma_3>\sigma_4$ et $\sigma_1>\sigma_2$, on a : $\lambda(x,y)\geq 0$

- L'ordre N°3 appartient à la **Famille III** des ordres à <u>emboîtement</u> , il est équivalent à <abcd>, de facto
  on aura donc : $\lambda(x,y)=0$ dans cette configuration

Comportement de l'indice λ(x,y) de Goodman-Kruskal en cas de Disjonction Complète

En cas de disjonction complète , le tableau Tetrachorique se ramène à :

	y	Non y	
x	11(x,y)	m-11(x,y)	m
Non x	m-11(x,y)	P+11(x,y)-2m	P-m
	m	P-m	N

Les formules suivantes :

(i) $$A = \sum_{u=1}^{p} \underset{v}{Max}\ n_{uv} + \sum_{v=1}^{q} \underset{u}{Max}\ n_{uv} = N + \frac{1}{2}\left(|a-b|+|a-c|+|b-c|+|b-d|\right)$$

(ii) $$B = N + \frac{1}{2}\left[\,|\,a+b-(c+d)\,|+|\,a+c-(b+d)\,|\,\right]$$

se simplifient, puisqu'ici N=P, en:

(i) A= P+|m-2.11(x,y)|+ |P+2.11(x,y)-3m|

(ii) B= P+ | P-2m|

(iii) En cas de disjonction complète :00(x,y)>11(x,y) (sauf pour P=2m), 10(x,y)=01(x,y)=m-11(x,y) seule reste posée la question du positionnement de 11(x,y) par rapport à 10(x,y) .
Si 11(x,y)>10(x,y)=01(x,y) alors 11(x,y)>m-11(x,y) => 11(x,y)>m/2 (majorité de Condorcet) et l'ordre associé est : 00(x,y)>11(x,y)>10(x,y)=01(x,y) (cas des ordres à séparation vus précédemment)
Si 11(x,y)<10(x,y)=01(x,y) alors 11(x,y) <m-11(x,y) => 11(x,y)≤m/2 l'ordre associé est l' ordre à emboîtement : 00(x,y)>10(x,y)=01(x,y)>11(x,y)

- Si P=2m alors B=P=2m et :
  si 11(x,y) > m/2 alors $\lambda(x,y) = \frac{2.11(x,y) - m}{m} = 2S_d(x,y) - 1$
  si 11(x,y) < m/2 alors $\lambda(x,y) = 0$

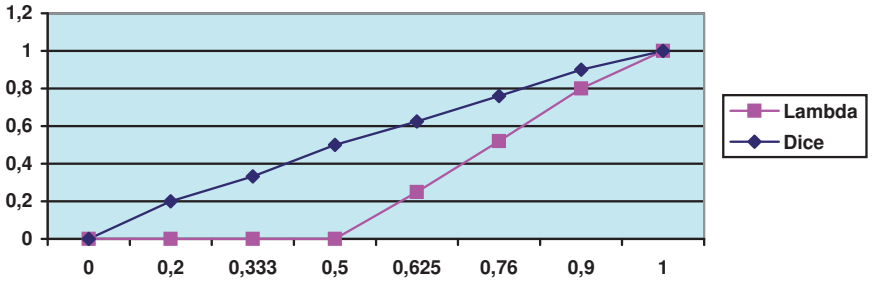
- Si P ≥3m alors B=2P-2m et A= 2P +211(x,y) –3m + |211(x,y)-m|b
  si 11(x,y)>m/2 alors $\lambda(x,y) = \frac{4.11(x,y) - 2m}{2m} = 2S_d(x,y) - 1$
  si 11(x,y) ≤ m/2 alors λ(x,y) =0
  (en effet dans ce cas on se retrouve comme précédemment dans une situation d'emboîtement avec 00(x,y)>01(x,y)=10(x,y)>11(x,y))

En conclusion l'indice de Goodman Kruskal calculé sur Tableau Tetrachorique en cas de disjonction complète se présente comme un indice de similarité à « **cliquet** » ou à seuil, qui, suivant la valeur de 11(x,y), est nul ou équivalent à 2 S_d(x,y)-1 (Homothétie sur l'indice de Dice) ce qui veut dire par ailleurs que dans le cas de disjonction complète l'indice:

$\mathbf{S_\lambda(x,y)}$ =1/2 (λ(x,y)+1) est équivalent à S_d (x,y) si 11(x,y) > m/2

La condition de Solomon Fortier appliquée à cet indice comme il varie de 0 à 1 si 11(x,y)>m/2 se traduira par l'inégalité :
λ(x,y) ≥ 1/2 ➔ 2 S_d(x,y)-1≥ 1/2 ➔ 2 Sd(x,y) ≥ 3/2➔2.11(x,y) ≥ 3/2.m➔ 11(x,y) ≥ 3/4 m (majorité au ¾)

On a tracé ci-dessous le graphe le l'indice Lambda de Goodman Kruskal λ(x,y) en fonction de l'indice de Dice ou de façon équivalente du rapport 11(x,y)/m dans le contexte de disjonction complète.



4.4.5.5. Indice de similarité déduit du Coefficient « κ» de Cohen

Comme le souligne G. Saporta dans son livre de (2006) (réédition de son premier ouvrage de 1990), ce coefficient de contingence est destiné à mesurer l'accord entre deux variables qualitatives (partitions) ayant le même nombre de modalités p. ; $n_{..}$ unités statistiques sont réparties suivant ces p catagories pour les deux partitions. Il y aura accord total entre ces deux partitions si les termes diagonaux sont les seuls termes non nuls du tableau de contingence suivant. (La première référence à ce coefficient est donnée dans un article du biostatistien et psychométricien Jacob Cohen datant de 1960 (1960)).

x \ y	1	2	v	..	p	
1						$n_{1.}$
2						$n_{2.}$
u			$n_{u\,v}$			$n_{u.}$
..						
p						
	$n_{.1}$	$n_{.2}$	$n_{.v}$			$n_{..}$

L'indice s'écrit alors :

(f 105)

$$\kappa(x,y) = \frac{\frac{1}{n_{..}}\sum_{u=1}^{p} n_{u\,u} - \frac{1}{n_{..}^2}\sum_{u=1}^{p} n_{u.}\, n_{.u}}{1 - \frac{1}{n_{..}^2}\sum_{u=1}^{p} n_{u.}\, n_{.u}}$$

Cet indice vaut :
1 en cas d'association complète et totale
-1 en cas où tous les termes de la diagonales sont nuls et les termes hors diagonales sont égaux

Dans le cas qui nous concerne nous travaillons sur un tableau (2x2), vérifiant donc p= q =2 (les deux variables ont même nombre de classes). Ce tableau est du type suivant :

	y	Non y	
x	a	b	a+b
Non x	c	d	c+d
	a+c	b+d	N

De ce fait nous pouvons calculer le coefficient κ(x,y) sur le tableau précédent, il vient

(f 106)
$$\kappa(x,y) = \frac{\frac{1}{N}(a+d) - \frac{1}{N^2}[(a+b)(a+c)+(b+d)(c+d)]}{1 - \frac{1}{N^2}[(a+b)(a+c)+(b+d)(c+d)]}$$

Après simplification on obtient:

(f 107):
$$\kappa(x,y) = \frac{2[ad-bc]}{N(b+c)+2[ad-bc]} = \frac{N[(a+d)-(b+c)]+(b^2+c^2)-(a^2+d^2)}{(a+d)N+(b^2+c^2)-(a^2+d^2)}$$

Cet indice possède une autre expression équivalente en effet il s'écrit :

(f 107')
$$\kappa(x,y) = \frac{2[ad-bc]}{(a+b)(b+d)+(a+c)(c+d)}$$

En effet les dénominateurs de (f 107) et (f 107') sont égaux après développement, néanmoins grâce à l'expression (f 107') , on voit que, en posant :

$$u = \frac{(ad-bc)}{(a+b)(b+d)} \quad \text{et} \quad v = \frac{(ad-bc)}{(a+c)(c+d)}$$

$$\kappa(x,y) = \text{Moyenne Harmonique } (u,v) = \frac{2\,u.v}{(u+v)}$$

et

$$S_T(x,y) = \text{Moyenne Géométrique}(u,v) = \sqrt{u\,v}$$

On en déduit que :

$$0 \le \left|\kappa(x,y)\right| \le \left|S_T(x,y)\right| \le 1$$

Le coefficient de Cohen vaut :

- +1, en cas d'association complète et totale $\Rightarrow$ b=c=0 et a et d $\neq$0

- -1, au cas où tous les termes de la diagonales sont nuls et les termes hors diagonale sont égaux

En effet, on voit grâce à la formule (f 106) que quand a+d=N (c'est à dire b=c=0), alors :

$$\kappa(x,y) = \frac{1 - \frac{1}{N^2}(a^2 + d^2)}{1 - \frac{1}{N^2}(a^2 + d^2)} = 1$$

grâce à la deuxième formulation on voit que quand a=d=0, alors b+c=N et $\kappa(x,y)$ devient :

$$\kappa(x,y) = \frac{-2bc}{N^2 - 2bc} = \frac{-2bc}{(b+c)^2 - 2bc} = \frac{-2bc}{b^2 + c^2}$$

ce coefficient est minimal lorque « bc » est maximum et « b^2+c^2 » minimum , or nous avons vu au §4.4.5.3 que ceci se produisait quand les deux nombres sont égaux en effet : **« le produit** de 2 nombres dont la **somme S est constante** (ici égale à N) est **maximum** si ces deux nombres sont **égaux »,** la conséquence pour notre problème est que : quelle que soit la valeur N , on doit avoir b = c = N/2 **à l'optimum**. D'où le fait signalé que l'indice de Cohen Kappa est égal à –1 dans ce cas .

Comme l'indice Kappa de Cohen varie de –1 à +1, il évoque un indice corrélatif, pour en faire un indice de similarité vrai, variant de 0 à 1, il suffit de construire l'indice de similarité « affine » associé au Kappa de Cohen en remplaçant dans les formulation précédentes « a par 11(x,y) », « b par 10(x,y) », « c par 01(x,y) », « d par 11(x,y) », et N par P et en calculant la quantité ½(κ(x,y)+1), il vient:

$$S_\kappa(x,y) = \frac{1}{2}[\kappa(x,y)+1] = \frac{2(ad-bc) + \frac{N}{2}(b+c)}{2(ad-bc) + N(b+c)} = \frac{2[11(x,y)\,00(x,y) - 10(x,y)\,01(x,y)] + \frac{P}{2}[10(x,y) + 01(x,y)]}{2[11(x,y)\,00(x,y) - 10(x,y)\,01(x,y)] + P[10(x,y) + 01(x,y)]}$$

sous cette forme on voit que l'indice de similarité vaut :

- $S_\kappa(x,y) = 0$ si $11(x,y)=00(x,y)=0$ avec la condition supplémentaire : $10x,y)=01(x,y)=N/2$
- $S_\kappa(x,y) = 1/2$ si $11(x,y).00(x,y)= 10(x,y).01(x,y)$ (condition d'indépendance)
- $S_\kappa(x,y) = 1$ si $10(x,y)=01(x,y)=0$

Si l'on compare le coefficient Kappa de Cohen et le coefficient Q de Yule on voit qu'ils ont des points communs puisqu'ils qu'ils utilisent les mêmes quantités, mais pas de la même manière.

$$S_Q(x,y) = \frac{(ad-bc)}{(ad+bc)} = \frac{11(x,y).00(x,y) - 01(x,y).10(x,y)}{11(x,y).00(x,y) + 10(x,y).01(x,y)}$$

Comportement de l'indice $S_\kappa(x,y)$ en cas de Disjonction Complète

En cas de disjonction complète l'indice Kappa précédent du fait des symétries entre valeurs se simplifie selon la formule :

$$S_\kappa(x,y)=\frac{2[11(x,y)\,(P+11(x,y)-2m)-[m-11(x,y)]^2]+P.[m-11(x,y)]}{2[11(x,y)(P+11(x,y)-2m)-[m-11(x,y)]^2]+2P[m-11(x,y)]}$$

Dire que $S_\kappa(x,y) \geq 1/2$ implique que :

$$S_\kappa(x,y)=\frac{2[11(x,y)\,(P+11(x,y)-2m)-[m-11(x,y)]^2]+P.[m-11(x,y)]}{2[11(x,y)(P+11(x,y)-2m)-[m-11(x,y)]^2]+2P[m-11(x,y)]} \geq \frac{1}{2}$$

c'est à dire :

$$11(x,y)(P+11(x,y)-2m) \geq (m-11(x,y))^2$$

soit :

$$11(x,y)P+11^2(x,y)-2m11(x,y) \geq m^2-2m11(x,y)+11^2(x,y)$$

ce qui implique : $11(x,y) \geq \frac{m^2}{P} \quad \Rightarrow$ si l' on pose $P=\overline{p}\,m$, $11(x,y) \geq \frac{m}{\overline{p}}$

On retrouve ici une borne équivalente à celle obtenue pour l'indice de similarité dérivé du Coefficient de contingence S_Q de Yule (voir § 4.4.4.2).

4.5 Indices à « valuations », ou « pseudo-indices de similarités »

Nous appelerons indices de similarité à « valuations » ou « indices d'abondance », les indices que nous allons présenter maintenant. En fait ces indices n'ont pas été définis a priori pour des tableaux de données binaires de présence–absence mais pour des tableaux de données du type du tableau n°1, introduit au § 2.1, mais où les valeurs ne sont plus {0 ou 1} mais des « comptages » d'occurrences (ce que les anglo-saxons appelent « abondance matrices » de terme général $\{t_{ij}\}$.

	$\mathbf{m_1}$	$\mathbf{m_2}$	$\mathbf{m_j}$	$\mathbf{m_{j'}}$		$\mathbf{m_P}$
$\mathbf{O_1}$	t_{11}	t_{12}	t_{1j}			
$\mathbf{O_2}$						
$\mathbf{O_3}$	t_{31}	t_{32}				
$\mathbf{O_i}$			t_{ij}			
$\mathbf{O_{i'}}$						
$\mathbf{O_N}$	t_{N1}		t_{Nj}			t_{NP}

En fait les matrices de « présence-absence » sont des cas particuliers de « matrices d'abondance », mais bien entendu les indices associés, sont a priori différents de ceux que nous avons présentés jusqu'ici. Il est intéressant de noter par ailleurs que la plupart des indices introduits dans ce contexte l'ont été par des spécialistes de l'étude des plantes et

principalement des « phytosociologues » ou des spécialistes de ce que les anglo-saxons appellent « plant ecology ».

Notre propos, dans ce paragraphe, est de voir si certains de ces indices, dédiés aux matrices d'abondance, vont générer dans leur restriction aux cas de tableaux de « présence-absence », des indices différents de ceux qui auraient déjà été présentés. Même dans le cas où nous retrouverons des indices de similarité déjà étudiés précédemment (c'est ce qui va se produire), la structure intrinséque de ces indices est différente de celle que nous avons rencontrée jusqu'à maintenant. Ceci est riche d'enseignements sur la façon dont les phytosociologues les ont imaginés et pensés.

Par convention, ici, les individus x et y seront décrits par leurs profils vectoriels:

$x = \{t_{x1}, t_{x2}, t_{x3}, \ldots\ldots t_{xj}..t_{xP}\}$ et $y = \{t_{y1}, t_{y2}, t_{y3}, \ldots\ldots t_{yj}..t_{yP}\}$

4.5.1 Indices d'abondance (ou à valuations) fonction de la différence (t_{ij} -$t_{i'j}$)

4.5.1.1 Indice de Lance et Williams (1966)

Cet « indice d'abondance » est défini par :

(f108)
$$I_{LW}(x,y) = \frac{1}{P}\sum_{j=1}^{P}\frac{\left|t_{xj} - t_{yj}\right|}{(t_{xj} + t_{yj})}$$

On voit de façon évidente qu'il varie de 0 à 1 ,

Il vaut en effet 0 si $t_{xj} = t_{yj}$ $\forall j$, et il vaut 1 si $\forall j$, $t_{xj} \neq 0 \Rightarrow t_{yj} = 0$ et réciproquement si $t_{yj} \neq 0 \Rightarrow$. $t_{xj} = 0$, en effet dans ce dernier cas on obtient un ratio concernant l'indice « j » qui vaut toujours 1. Si cette situation se produit P fois, la division par P ramène la valeur de l'Indice $I_{LW}(x,y)$ à 1. La façon de transformer un indice d'abondance en un indice de similarité normé paraît donc ici évidente, il suffit de poser :

$$S_{LW}(x,y) = 1 - I_{LW}(x,y)$$

L'indice $S_{LW}(x,y)$ varie bien maintenant de 0 à 1 , il vaut 1 si x et y ont un profil identique, et 0 s'ils sont totalement en opposition.

Intéressons nous maintenant à la restriction de cet indice dans le cas où les valeurs $\{t_{ij}\}$ en l'occurrence t_{xj} et t_{yj} sont des valeurs {0 ou 1},
(Attention : on suppose par convention ici que les configurations où $t_{xj}=t_{yj}=0$ qui a priori impliquent une division par 0) se traduisent par un ratio égal à 1 :

Dans ce cas, $\forall$ j , le ratio : $\dfrac{\left|t_{xj} - t_{yj}\right|}{t_{xj} + t_{yj}}$ vaut 0 ou 1, il vaut 0 si la valeur t_{ij} est identique pour x et y et il vaut 1 en cas contraire.

De ce fait : $$I_{LW}(x,y)=\frac{1}{P}\sum_{j=1}^{P}\frac{|t_{xj}-t_{yj}|}{(t_{xj}+t_{yj})}=\frac{1}{P}[10(x,y)+01(x,y)]$$

On a alors la propriété suivante:

<u>Propriété n°18</u> : L'Indice de Similarité de Lance Willliams dans sa restriction 0-1 avec <u>la convention précédente</u> (pour les cas 0-0) n'est rien d'autre que le critère de <u>Sokal et Michener</u> (Simple Matching)

En effet on a:

$$S_{LW}(x,y)=1-\frac{1}{P}\sum_{j=1}^{P}\frac{|t_{xj}-t_{yj}|}{(t_{xj}+t_{yj})}=1-\frac{1}{P}[10(x,y)+01(x,y)]=\frac{11(x,y)+00(x,y)}{P}=S_{SM}(x,y)$$

4.5.1.2 Indice de Odum (1950), plus connu sous le nom d'Indice de « Bray-Curtis » (1957)

Cet indice est défini par :

(f109) $$I_{BC}(x,\ y)=\frac{\sum_{j=1}^{P}|t_{xj}-t_{yj}|}{\sum_{j=1}^{P}(t_{xj}+t_{yj})}$$

Ici on n'a pas besoin (et c'est un avantage), comme dans le cas précédent, de convention particulière pour les cas (0-0). Cet indice vaut 0 si $t_{xj}=t_{yj}\ \forall\ j$, et il vaut 1 si $\forall j$, $t_{xj}\neq 0 \Rightarrow t_{yj}=0$ et réciproquement si $t_{yj}\neq 0 \Rightarrow . t_{xj}=0$,

Dans le cas où l'on se restreint à des valeurs $t_{ij}\in\{0,1\}$, le numérateur de l'indice n'est rien d'autre que la quantité

$$\sum_{j=1}^{P}|t_{xj}-t_{yj}|=10(x,y)+01(x,y)$$

Le dénominateur est égal à :

$$\sum_{j=1}^{P}(t_{xj}+t_{yj})=11(x,x)+11(y,y)=2.11(x,y)+10(x,y)+01(x,y)$$

L'indice de similarité associé défini par $S_{BC}(x,y)= 1- I_{BC}(x,y)$ est donc égal à :

$$S_{BC}(x,y)=\frac{2.11(x,y)}{211(x,y)+10(x,y)+01(x,y)}=\frac{11(x,y)}{11(x,y)+\frac{1}{2}[10(x,y)+01(x,y)]}=S_{d}(x,y)$$

<u>Propriété n°19</u> : L'Indice de Similarité de Bray-Curtis (Odum), dans sa restriction 0-1, est strictement identique à <u>l'indice de Dice</u>

4.5.2 Indices d'abondance (ou à valuations) fonction de Max(t_{ij},$t_{i'j}$) ou Min (t_{ij}, $t_{i'j}$)

4.5.2.1 Indice de Kulczinski formel (1930)

Cet indice est défini par :

(f110)
$$I_K(x,\ y) = \frac{\sum_{j=1}^{P} \text{Min}\ (t_{xj}, t_{yj})}{\sum_{j=1}^{P} (t_{xj} + t_{yj})}$$

Du fait que Min (a,b)=½ (a+b) - ½ |a - b|, il apparaît de facon évidente que cet indice est égal à :

$$I_K(x,\ y) = \frac{\sum_{j=1}^{P} \text{Min}\ (t_{xj}, t_{yj})}{\sum_{j=1}^{P} (t_{xj} + t_{yj})} = \frac{1}{2} - \frac{1}{2} I_{BC}(x,\ y)$$

Cet indice vaut ½ si $t_{xj}=t_{yj}\ \forall\ j$, et il vaut 0 si $\forall j$, $t_{xj}\neq 0 \Rightarrow t_{yj}=0$ et réciproquement si $t_{yj}\neq 0 \Rightarrow . t_{xj} = 0$,

Dans le cas où l'on se restreint à des valeurs $t_{ij} \in \{0,1\}$, on a vu que : $S_{BC}(x,y)= 1- I_{BC}(x,y)$
D'où comme $I_{BC}(x,y)= 1-2\ I_K(x,y)$, il vient alors :

$$S_{BC}(x,y)=1-(1-2I_K(x,y))= 2\ I_K(x,y)$$

en posant $S_K(x,y)=2I_K(x,y)$, il apparaît que cet indiced'abondance induit au coefficient 2 près , un indice de similarité égal à l'indice de « Dice ».

4.5.2.2 Indice de Marczewski et Steinhaus (1937)

Cet indice est défini par :

(f111)
$$I_{MS}(x,\ y) = \frac{\sum_{j=1}^{P} \left| t_{xj} - t_{yj} \right|}{\sum_{j=1}^{P} \text{Max}\ (t_{xj}, t_{yj})}$$

Dans ce cas encore, en jouant sur le fait que Max(a,b) =½ (a+b) + ½ |a - b| , on voit que la quantité $1/I_{MS}(x,y)$ peut se simplifier, en effet on obtient :

$$\frac{1}{I_{MS}(x,\ y)} = \frac{\frac{1}{2}\sum_{j=1}^{P} \left| t_{xj} - t_{yj} \right| + \frac{1}{2}\sum_{j=1}^{P} (t_{xj} + t_{yj})}{\sum_{j=1}^{P} \left| t_{x\,j} - t_{y\,j} \right|} = \frac{1}{2} + \frac{1}{2} \text{ x } \frac{1}{I_{BC}(x,\ y)} \Rightarrow I_{MS}(x,\ y) = \frac{2I_{BC}(x,\ y)}{I_{BC}(x,\ y) + 1}$$

Sous cette forme, il apparaît de façon évidente que $I_{MS}(x,y)$ varie de 0 à 1 , puisque $I_{BC}(x,y)$ varie de 0 à 1 lui-même.
Maintenant définissons l'« indice de similarité de Marczewski et Steinhaus » comme la quantité :

$$S_{MS}(X,Y)=1-I_{MS}(X,Y)$$

Dès lors, dans le cas où l'on se restreint à des valeurs t_{ij} binaires on peut montrer que l'indice précédent s'écrit :

$$S_{MS(}\ x,\ y) = \frac{1 - I_{BC}(x,\ y)}{1 + I_{BC}(x,\ y)} = \frac{S_d(x,\ y)}{2 - S_d(x,\ y)}$$

où $S_d(x,y)$ est l'indice de Dice (voir 4.1.1.1) . La fonction homographique précédente n'est autre que la formule (f17) caractéristique de l'indice de Jaccard.
Donc en conclusion , on a la propriété n°20 suivante :

Propriétén°20 : L'indice de similarité déduit de l'indice d'abondance de Marczewski –Steinhaus n'est rien d'autre dans sa restriction booléenne que l'indice de Jaccard

En guise de conclusion sur les indice du type « indices à valuations ».
Dans la liste précédente il n'a pas été fait mention (et d'aucuns pourraient s'en étonner) des indices qui se calculent à partir de données de fréquences ou d'occurrences de phénomènes (comme par exemple le comptage de termes apparaissant dans des textes).
Outre les indices de contingences dont nous avons vu quelques exemplaires au § 4.4., qui peuvent être utilisés dans ce contexte, comme également pourrait l'être le classique Critère du χ^2; il se trouve qu'un certain nombre d'auteurs provenant de la Linguistique computationnelle ou de la recherche documentaire ont élaboré des coefficients beaucoup plus adaptés au domaine de la « comparaison documentaire » ou totalement en adaquation avec la recherche sur Internet. C'est le cas de l'américain Gerard Salton (voir G. Salton (1968) et (1983)) qui a proposé avec son équipe, tout d'abord son « Cosinus simple de profils fréquentiels pondérés», puis son fameux « TF-IDF Cosine », (« Term Frequency –Inverse Document Frequency », coefficient qu'il a d'ailleurs appellé le « best fully weighted system »). Lui et l'anglaise Karen Sparck Jones (1971), ont été les pionniers de ces recherches sur les mesures de matching entre documents. On trouvera dans l'article simple et bien structuré de A. Lelu (2002) un essai de comparaison des mesures de Salton avec d'autres approches issues de méthodologies « neuronales » ou « factorialistes ».

Essai de Typologie Structurelle des Indices de Similarités

Indices	Groupe	Type	Formule	Borne S.F.	Variation
Dice	Group I	Typ1	$S_d(x,y)=\frac{2.11(x,y)}{211(x,y)+10(x,y)+01(x,y)}$	m/2	$0\le S\le 1$
Jaccard	Group I	Typ1	$S_j(x,y)=\frac{11(x,y)}{11(x,y)+[10(x,y)+01(x,y)]}$	2m/3	$0\le S\le 1$
Anderberg	Group I	Typ1	$S_{an}(x,y)=\frac{11(x,y)}{11(x,y)+2[10(x,y)+01(x,y)]}$	4m/5	$0\le S\le 1$
Sorensen	Group I	Typ1	$S_{so}(x,y)=\frac{11(x,y)}{11(x,y)+\frac{1}{4}[10(x,y)+01(x,y)]}$	m/3	$0\le S\le 1$
Anderberg 2	Group I	Typ1	$S_{ac}(x,y)=\frac{11(x,y)}{11(x,y)+\frac{1}{8}[10(x,y)+01(x,y)]}$	m/5	$0\le S\le 1$
Kulczynski	Group I	Typ 2A	$S_k(x,y)\frac{1}{2}\left[\frac{11(x,y)}{11(x,y)+10(x,y)}+\frac{11(x,y)}{11(x,y)+01(x,y)}\right]$	m/2	$0\le S\le 1$
Ochiaï	Group I	Typ 2A	$S_O(x,y)=\frac{11(x,y)}{\sqrt{[11(x,y)+10(x,y)][(11(x,y)+01(x,y)]}}$	m/2	$0\le S\le 1$
Rappel	Group I	Typ 2A	$S_R(x,y)=\frac{11(x,y)}{11(x,y)+10(x,y)}$	m/2	$0\le S\le 1$
Précision	Group I	Typ 2A	$S_p(x,y)=\frac{11(x,y)}{11(x,y)+01(x,y)}$	m/2	$0\le S\le 1$
Quadra/ Norm	Group I	Typ 2B	$S_N(x,y)=\frac{\sqrt{2}.11(x,y)}{\sqrt{[11(x,y)+10(x,y)]^2+[11(x,y)+01(x,y)]^2}}$	m/2	$0\le S\le 1$
Braun Blanquet	Group I	Typ 2B	$S_{Min}(x,y)=\frac{11(x,y)}{Max[11(x,y)+10(x,y),11(x,y)+01(x,y)]}$	m/2	$0\le S\le 1$
Simpson	Group I	Typ 2B	$S_{Max}(x,y)=\frac{11(x,y)}{Min[11(x,y)+10(x,y),11(x,y)+01(x,y)]}$	m/2	$0\le S\le 1$
McConaughey	Group I	Typ 2B	$S_{McC}(x,y)=\left[\frac{[11(x,y)]^2-10(x,y).01(x,y)}{[11(x,y)+10(x,y)].[11(x,y)+01(x,y)]}\right]$	m/2	$-1\le S\le+1$
Sokal- Michen	Group II	Typ 2B	$S_{sm}(x,y)=\frac{11(x,y)+00(x,y)}{11(x,y)+00(x,y)+10(x,y)+01(x,y)}$	m-P/4	$0\le S\le 1$
Sokal-Sneath	Group II	Typ I	$S_{ss}(x,y)=\frac{2.[11(x,y)+00(x,y)]}{P+[11(x,y)+00(x,y)]}$	m-P/3	$0\le S\le 1$
Rogers-Tanim	Group II	Typ I	$S_{rt}(x,y)=\frac{11(x,y)+00(x,y)}{2.P-[11(x,y)+00(x,y)]}$	m-P/6	$0\le S)\le 1$
Rao	Group II	Typ I	$S_r(x,y)\frac{11(x,y)}{11(x,y)+00(x,y)+10(x,y)+01(x,y)}$	P/2	$0\le S\le 1$
Marco/ Mich1	Group II	Typ I	$S_{mm}(x,y)=\frac{11(x,y)+\frac{1}{2}00(x,y)}{11(x,y)+00(x,y)+10(x,y)+01(x,y)}$	2m/3	$0\le S\le 1$
Marco/ Mich2	Group II	Typ I	$S_{mm2}(x,y)=\frac{11(x,y)+\frac{1}{2}00(x,y)}{11(x,y)+00(x,y)+\frac{1}{2}[10(x,y)+01(x,y)]}$	m/2	$0\le S\le 1$
Tetrachorique	Group II	Typ II	$S_T(x,y)=\frac{[11(x,y).00(x,y)-10(x,y).01(x,y)]}{\sqrt{[11(x,y)+10(x,y)][11(x,y)+01(x,y)].[00(x,y)+10(x,y)][00(x,y)+01(x,y)]}}$	m/2	$-1\le S\le+1$
Yule Y	Group II	Typ II	$S_Y(x,y)=\frac{\sqrt{11(x,y).00(x,y)}-\sqrt{10(x,y).01(x,y)}}{\sqrt{11(x,y).00(x,y)}+\sqrt{10(x,y).01(x,y)}}$	m^2/P	$-1\le S\le+1$
Yule Q	Group II	Typ II	$S_Q(x,y)=\frac{11(x,y).00(x,y)-01(x,y).10(x,y)}{11(x,y).00(x,y)+10(x,y).01(x,y)}$	m^2/P	$-1\le S\le+1$
MA4 Ratios	Group II	Typ II	$S_{R4}(x,y)=\frac{1}{4}\left[\frac{11(x,y)}{11(x,y)+10(x,y)}+\frac{11(x,y)}{11(x,y)+01(x,y)}+\frac{00(x,y)}{00(x,y+01(x,y)}+\frac{00(x,y)}{00(x,y)+10(x,y)}\right]$	m/2	$0\le S\le 1$
MG4 Ratios	Group II	Typ II	$S_{G4}(x,y)=\sqrt[4]{\frac{11(x,y)}{11(x,y)+10(x,y)}x\frac{11(x,y)}{11(x,y)+01(x,y)}x\frac{00(x,y)}{00(x,y)+10(x,y)}x\frac{00(x,y)}{00(x,y)+01(x,y)}}$	$\equiv$ m / 4	$0\le S\le 1$
MH4 Ratios	Group II	Typ II	$\frac{1}{S_{h4}(x,y)}=\frac{1}{4}\left[\frac{11(x,y)+10(x,y)}{11(x,y)}+\frac{11(x,y)+01(x,y)}{11(x,y)}+\frac{00(x,y)+10(x,y)}{00(x,y)}+\frac{00(x,y)+01(x,y)}{00(x,y)}\right]$	$\frac{m}{3}\left[1+\frac{34}{9\bar{P}}\right]$	$0\le S\le 1$
Urbani-Buser	Group II	Typ II	$S_{Bub}(x,y)=\left[\frac{\sqrt{11(x,y).00(x,y)}+11(x,y)}{\sqrt{11(x,y).00(x,y)+11(x,y)+01(x,y)+10(x,y)}}\right]$	$\frac{m}{\bar{P}}\left[4-\frac{35}{4.\bar{P}}\right]$	$0\le S\le 1$
Lerman mod	Group II	Typ III	$S_L(x,y)=\frac{[11^2(x,y)+00^2(x,y)-(10^2(x,y)+0^2(x,y))]}{\sqrt{[[11(x,y)+01(x,y)]^2+[00(x,y)+10(x,y)]^2][[11(x,y)+10(x,y)]^2+[00(x,y)+01(x,y)]^2]}}$	m/2	$-1\le S\le+1$
Kappa-Cohen	Group II	Typ III	$S_\kappa(x,y)=\frac{2[11(x,y)\,00(x,y)-10(x,y)\,01(x,y)]+(P/2)[10(x)+01(x,y)]}{2[11(x,y)\,00(x,y)-10(x,y)\,01(x,y)]+P[10(x,y)+01(x,y)]}$	m/2	$0\le S\le 1$

5. Liaison « Indices de Similarité » - « Critères de Contingence », vers un bouclage du processus typologique

De même que nous avions utilisé au §4.4.5, les expressions de certains critères ou coefficients de contingence pour retrouver ou définir de nouveaux indices de similarité (par exemple l'indice de similarité de « Lerman modifié »), nous allons tenter l'inverse dans ce paragraphe en partant de la notion de codage vectoriel de matrices relationnelles tel que nous l'avons introduit au § 2.4.4 .

En d'autres termes, et pour boucler la boucle, nous allons nous servir des définitions d'indices de similarité sur vecteurs binaires, pour retrouver ou éventuellement redécouvrir des coefficients d'association sur tableaux de contingence au sens statistique usuel. Rappelons que le codage vectoriel de matrices relationnelles se définit selon le principe :
Pour toute matrice relationnelle C^k*, on définit son* ***Extension Vectorielle*** $\vec{\gamma}^k$ *comme un vecteur de longueur* $\mathbf{N^2}$ *tel que :*

$\vec{\gamma}^k = (\gamma_1^k, \gamma_2^k, \gamma_3^k, ...\gamma_s^k, ...\gamma_{N^2}^k)$, où si $C_{ii'}^k$ est le terme général de la matrice C^k, alors :

$C_{ii'}^k = \gamma_s^k$, si et seulement si l'indice courant « s » vérifie:

$$s = (i-1)N + i' \qquad \forall i \text{ et } i'$$

Partant de ce principe en posant :

$$\vec{\gamma}^k = (\gamma_1^k, \gamma_2^k, \gamma_3^k, ...\gamma_s^k, ...\gamma_{N^2}^k) \text{ et } \vec{\gamma}^{k'} = (\gamma_1^{k'}, \gamma_2^{k'}, \gamma_3^{k'}, ...\gamma_s^{k'}, ...\gamma_{N^2}^{k'})$$

On peut utiliser, comme pour n'importe quel tableau binaire, la présentation des données sous forme d'un tableau de contingence Tetrachorique , en remplaçant x et y par $\vec{\gamma}^k$ et $\vec{\gamma}^{k'}$, il vient :

	$\vec{\gamma}^{k'} = 1$	$\vec{\gamma}^{k'} = 0$
$\vec{\gamma}^k = 1$	$11(\vec{\gamma}^k, \vec{\gamma}^{k'})$	$10(\vec{\gamma}^k, \vec{\gamma}^{k'})$
$\vec{\gamma}^k = 0$	$01(\vec{\gamma}^k, \vec{\gamma}^{k'})$	$00(\vec{\gamma}^k, \vec{\gamma}^{k'})$

De ce fait, certains des indices de similarité définis précédemment peuvent être utilisés comme indices sur les linéarisations vectorielles relationnelles, une fois retraduites en termes des égalités : $C^{k}_{ii'} = \gamma^{k}_{s}$. Donnons quelques exemples, choisis parmi les indices les plus connus:

5.1 Indices du Groupe I

5.1.1 Indice de Dice ➔(Morey-Agresti)

L'indice de Dice défini par :

$$S_d(x,y)=\frac{11(x,y)}{11(x,y)+\frac{1}{2}\left[10(x,y)+01(x,y)\right]}=\frac{2.11(x,y)}{2.11(x,y)+10(x,y)+01(x,y)}$$ se transforme dans cette configuration en **:**

$$S_d(\gamma^k,\gamma^{k'})=\frac{11(\gamma^k,\gamma^{k'})}{11(\gamma^k,\gamma^{k'})+\frac{1}{2}\left[10(\gamma^k,\gamma^{k'})+01(\gamma^k,\gamma^{k'})\right]}=\frac{2.\sum_{i=1}^{N}\sum_{i'=1}^{N}C^{k}_{ii'}\,C^{k'}_{ii'}}{\sum_{i=1}^{N}\sum_{i'=1}^{N}C^{k}_{ii'}+\sum_{i=1}^{N}\sum_{i'=1}^{N}C^{k'}_{ii'}}$$

et du fait des liaisons entre "expressions relationnelles" et "notations contingentielles" (voir F. Marcotorchino (1984)), on peut dans ce cas réécrire l'indice de Dice, (sous réserve que C^k et $C^{k'}$ représentent des variables nominales ayant respectivement "p" et "q" modalités) sous forme d'un coefficient de contingence:

(f112)
$$S_d(\gamma^k,\gamma^{k'})=\frac{2.\sum_{i=1}^{N}\sum_{i'=1}^{N}C^{k}_{ii'}\,C^{k'}_{ii'}}{\sum_{i=1}^{N}\sum_{i'=1}^{N}C^{k}_{ii'}+\sum_{i=1}^{N}\sum_{i'=1}^{N}C^{k'}_{ii'}}=\frac{2\sum_{u=1}^{p}\sum_{v=1}^{q}n^2_{uv}}{\sum_{u=1}^{p}n^2_{u.}+\sum_{v=1}^{q}n^2_{.v}}$$

On retrouve dans ce cas un indice de contingence connu sous le nom de Morey-Agresti , d'où:
Indice de Dice (sur matrice relationnelle vectorielle) ➔ Coefficient d'association de Morey-Agresti (sur tableau de contingence)

5.1.2 Indice d'Ochiaï ➔(Fowlkes-Mallows)

De la même façon que dans le cas de l'Indice de Dice, l'indice d'Ochiaï appliqué au tableau Tetrachorique précédent nous redonne un Coefficient de contingence connu. En effet nous avons vu que l'indice d'Ochiaï était défini par :

$$S_o(x,y)=\frac{11(x,y)}{\sqrt{[11(x,y)+10(x,y)][(11(x,y)+01(x,y)]}}$$

Soit après remplacement de x et y par γ^k et $\gamma^{k'}$:

Il vient alors:

$$S_o(\gamma^k,\gamma^{k'}) = \frac{11(\gamma^k,\gamma^{k'})}{\sqrt{11(\gamma^k,\gamma^{k'})+10(\gamma^k,\gamma^{k'})}\sqrt{11(\gamma^k,\gamma^{k'})+01(\gamma^k,\gamma^{k'})}} = \frac{\sum_{i=1}^{N}\sum_{i'=1}^{N} C^k_{ii'} C^{k'}_{ii'}}{\sqrt{\sum_{i=1}^{N}\sum_{i'=1}^{N} C^k_{ii'}}\sqrt{\sum_{i=1}^{N}\sum_{i'=1}^{N} C^{k'}_{ii'}}}$$

et de façon identique au cas de l'indice de Dice, en utilisant également les liaisons entre "expressions relationnelles" et "notations contingentielles" (voir F. Marcotorchino (1984)), on peut dans ce cas réécrire l'indice d'Ochiaï, (sous réserve que C^k et $C^{k'}$ représentent des variables nominales ayant respectivement "p" et "q" modalités) sous forme d'un coefficient de contingence:

(f113)
$$S_o(\gamma^k,\gamma^{k'}) = \frac{\sum_{i=1}^{N}\sum_{i'=1}^{N} C^k_{ii'} C^{k'}_{ii'}}{\sqrt{\sum_{i=1}^{N}\sum_{i'=1}^{N} C^k_{ii'}\sum_{i=1}^{N}\sum_{i'=1}^{N} C^{k'}_{ii'}}} = \frac{\sum_{u=1}^{p}\sum_{v=1}^{q} n^2_{uv}}{\sqrt{\sum_{u=1}^{p} n^2_{u.}}\sqrt{\sum_{u=1}^{q} n^2_{.v}}}$$

Quiconque , familier des statistiques contingentielle peut reconnaître dans la formule (f113) l'expression du Coefficient de Contingence de **Fowlkes et Mallows** (voir Fowlkes et Mallows(1983)). On trouvera également une étude détaillée de ce critère et les bornes qui lui sont associées dans la thèse de H. Messatfa (1989) .

Indice d'Ochiaï (sur matrice relationnelle vectorielle) ➔ *Coefficient d'association de Fowlkes et Mallows (sur tableau de contingence)*

5.1.3 Indice de Kulczynski ➔(Hubert, Arabie)

Nous avons vu au § 4.2.1.1, que l'indice de Kulczynski était donné par :

$$S_k(x,y) = \frac{1}{2}\left[\frac{11(x,y)}{11(x,y)+10(x,y)} + \frac{11(x,y)}{11(x,y)+01(x,y)}\right]$$

En appliquant le même raisonnement que pour les indices de Dice et Ochiaï , on obtient le résultat suivant:

$$S_k(\gamma^k,\gamma^{k'}) = \frac{1}{2}\left[\frac{11(\gamma^k,\gamma^{k'})}{[11(\gamma^k,\gamma^{k'})+10(\gamma^k,\gamma^{k'})]} + \frac{11(\gamma^k,\gamma^{k'})}{[11(\gamma^k,\gamma^{k'})+01(\gamma^k,\gamma^{k'})]}\right] = \frac{\sum_{i=1}^{N}\sum_{i'=1}^{N} C^k_{ii'} C^{k'}_{ii'}}{2}\left[\frac{1}{\sum_{i=1}^{N}\sum_{i'=1}^{N} C^k_{ii'}} + \frac{1}{\sum_{i=1}^{N}\sum_{i'=1}^{N} C^{k'}_{ii'}}\right]$$

soit en notations contingentielles et en appliquant toujours les remarques et restrictions définies pour Dice et Ochiaï:

(f114)
$$S_k(\gamma^k,\gamma^{k'}) = \frac{\sum_{i=1}^{N}\sum_{i'=1}^{N} C^k_{ii'} C^{k'}_{ii'}}{2}\left[\frac{1}{\sum_{i=1}^{N}\sum_{i'=1}^{N} C^k_{ii'}} + \frac{1}{\sum_{i=1}^{N}\sum_{i'=1}^{N} C^{k'}_{ii'}}\right] = \sum_{i=1}^{N}\sum_{i'=1}^{N} C^k_{ii'} C^{k'}_{ii'} \frac{\sum_{i=1}^{N}\sum_{i'}^{N} C^k_{ii'} + \sum_{i=1}^{N}\sum_{i'=1}^{N} C^{k'}_{ii'}}{2\sum_{i=1}^{N}\sum_{i'=1}^{N} C^k_{ii'}\sum_{i=1}^{N}\sum_{i'=1}^{N} C^{k'}_{ii'}} = \sum_{u=1}^{p}\sum_{v=1}^{q} n^2_{uv}\left[\frac{\sum_{u=1}^{p} n^2_{u.} + \sum_{v=1}^{q} n^2_{.v}}{2\sum_{u=1}^{p} n^2_{u.}\sum_{v=1}^{q} n^2_{.v}}\right]$$

En fait les indices $S_d(\gamma^k,\gamma^{k'}), S_o(\gamma^k,\gamma^{k'}), S_k(\gamma^k,\gamma^{k'})$ possèdent le même numérateur $\sum_{i=1}^{N}\sum_{i'=1}^{N} C_{ii'}^{k} C_{ii'}^{k'} = \sum_{u=1}^{p}\sum_{v=1}^{q} n_{uv}^2$ et ne différent que par leurs dénominateurs qui jouent le rôle de borne supérieure pour le numérateur. Les dénominateurs respectifs sont: la moyenne Arithmétique, la moyenne Géométrique, et la moyenne Harmonique des quantités $\sum_{u=1}^{p} n_{u.}^2$ et $\sum_{v=1}^{q} n_{.v}^2$, qui jouent effectivement le rôle de bornes de variation pour le numérateur. Ceci a été étudié pour l'ensemble de ces trois critères par L. Hubert et P.Arabie (1985), par A. Agresti et L. Morey (1984) ainsi que par H. Messatfa (1989).

Ces trois critères sont donc connus, le premier a été étudié par Agresti et Morey , le second a été étudié par **Fowlkes et Mallows** qui lui ont donné leurs noms, le dernier a été étudié par L.Hubert et P. Arabie ainsi par H. Messatfa qui a fait une étude comparative d'autres bornes possibles au dénominateur, en plus des trois précédentes.

5.2 Indices du Groupe II

5.2.1 Indice de Sokal et Michener et Green Rao ➔ (Condorcet, Rand)

L'indice de Sokal et Michener est défini au § 4.4.1.1 par :

$$S_{sm}(x,y) = \frac{11(x,y)+00(x,y)}{P} = \frac{11(x,y)+00(x,y)}{11(x,y)+00(x,y)+10(x,y)+01(x,y)}$$

il se transforme dans la configuration en notations relationnelle vectorielles selon l'expression suivante**:**

$$S_{sm}(\gamma^k,\gamma^{k'}) = \frac{11(\gamma^k,\gamma^{k'})+00(\gamma^k,\gamma^{k'})}{11(\gamma^k,\gamma^{k'})+00(\gamma^k,\gamma^{k'})+\left[10(\gamma^k,\gamma^{k'})+01(\gamma^k,\gamma^{k'})\right]} = \frac{\left[\sum_{i=1}^{N}\sum_{i'=1}^{N} C_{ii'}^{k} C_{ii'}^{k'} + \sum_{i=1}^{N}\sum_{i'=1}^{N} \overline{C}_{ii'}^{k} \overline{C}_{ii'}^{k'}\right]}{N^2}$$

ceci du fait des liaisons entre "expressions relationnelles" et "profils vectoriels". On reconnaît dans le numérateur de l'expression relationnelle présentée à droite de l'indice de Sokal et Michener, la forme symétrique du **critère de Condorcet** valable pour le croisement entre deux variables qualitatives.

De plus sous réserve que C^k et $C^{k'}$ représentent des variables nominales ayant respectivement "p" et "q" modalités et en utilisant les formules de passage relations➔ contingences (voir encore F. Marcotorchino (1984), on obtient la formule contingentielle suivante (en identifiant les notations N et $n_{..}$):

(f115) $$S_{sm}(\gamma^k,\gamma^{k'}) = \frac{\left[\sum_{i=1}^{N}\sum_{i'=1}^{N} C_{ii'}^{k} C_{ii'}^{k'} + \sum_{i=1}^{N}\sum_{i'=1}^{N} \overline{C}_{ii'}^{k} \overline{C}_{ii'}^{k'}\right]}{N^2} = \frac{2\sum_{u=1}^{p}\sum_{v=1}^{q} n_{uv}^2 - \sum_{u=1}^{p} n_{u.}^2 - \sum_{v=1}^{q} n_{.v}^2 + n_{..}^2}{n_{..}^2}$$

On reconnaît dans l'expression à droite ci dessus, l'une des formes possibles du **critère de Rand** , W. Rand (1971) , en tout cas celle étudiée par F. Marcotorchino (1984) .

En conclusion l'indice de similarité de Sokal et Michener nous permet de retrouver en consécution et dans la foulée le Coefficient de Condorcet (approche relationnelle) et celui de Rand (approche contingentielle).

5.2.2 Indice de Similarité Tetrachorique (ou de Bravais –Pearson) ➔(Lerman)

De même que dans le cas de l'Indice de Sokal et Michener, intéressons nous maintenant à un autre indice de similarité du Groupe II (en l'occurrence du Type II) , à savoir l'indice « Tetrachorique » donné (page 55) par :

$$S_T(x,y)=\frac{[11(x,y).00(x,y)-10(x,y).01(x,y)]}{\sqrt{[11(x,y)+10(x,y)][11(x,y)+01(x,y)].[00(x,y)+10(x,y)][00(x,y)+01(x,y)]}}$$

Soit en remplaçant x et y par γ^k et $\gamma^{k'}$:

$$S_T(\gamma^k,\gamma^{k'})=\frac{11(\gamma^k,\gamma^{k'}).00(\gamma^k,\gamma^{k'})-10(\gamma^k,\gamma^{k'})01(\gamma^k,\gamma^{k'})}{\sqrt{11(\gamma^k,\gamma^{k'})+10(\gamma^k,\gamma^{k'})}\sqrt{11(\gamma^k,\gamma^{k'})+01(\gamma^k,\gamma^{k'})}\sqrt{00(\gamma^k,\gamma^{k'})+10(\gamma^k,\gamma^{k'})}\sqrt{00(\gamma^k,\gamma^{k'})+01(\gamma^k,\gamma^{k'})}}$$

en utilisant également les liaisons entre "expressions vectorielles relationnelles" et "notations relationnelles (voir §2.4.4), on peut dans ce cas réécrire l'indice Tétrachorique ou de Bravais Pearson sous forme de l'expression suivante:

$$S_T(\gamma^k,\gamma^{k'})=\frac{\sum_{i=1}^{N}\sum_{i'=1}^{N}C_{ii'}^{k}\,C_{ii'}^{k'}\sum_{i=1}^{N}\sum_{i'=1}^{N}\overline{C}_{ii'}^{k}\overline{C}_{ii'}^{k'}-\sum_{i=1}^{N}\sum_{i'=1}^{N}C_{ii'}^{k}\overline{C}_{ii'}^{k'}\sum_{i=1}^{N}\sum_{i'=1}^{N}\overline{C}_{ii'}^{k}C_{ii'}^{k'}}{\sqrt{\sum_{i=1}^{N}\sum_{i'=1}^{N}C_{ii'}^{k}}\sqrt{\sum_{i=1}^{N}\sum_{i'=1}^{N}C_{ii'}^{k'}}\sqrt{\sum_{i=1}^{N}\sum_{i'=1}^{N}\overline{C}_{ii'}^{k}}\sqrt{\sum_{i=1}^{N}\sum_{i'=1}^{N}\overline{C}_{ii'}^{k'}}}$$

En utilisant maintenant les formules de passages contingence<=> relations (voir F. Marcotorchino (1984)) nous allons transformer l'expression de l'indice ci dessus en une forme contingentielle, on obtient après développements et simplifications (après identification des notations N et $n_{..}$):

(f116)
$$S_T(\gamma^k,\gamma^{k'})=\frac{n_{..}^2\sum_{u=1}^{p}\sum_{v=1}^{q}n_{uv}^2-\sum_{u=1}^{p}n_{u.}^2\sum_{v=1}^{q}n_{.v}^2}{\sqrt{\sum_{u=1}^{p}n_{u.}^2}\sqrt{\sum_{v=1}^{q}n_{.v}^2}\sqrt{n_{..}^2-\sum_{u=1}^{p}n_{u.}^2}\sqrt{n_{..}^2-\sum_{v=1}^{q}n_{.v}^2}}$$

Si nous divisons le numérateur et le dénominateur de la formule précédente par $n_{..}$, il est intéressant de constater que l'expression donnée par la formule (f116) est strictement équivalente à la formule (f97) caractérisant le **coefficient de contingence de IC Lerman**. Ici il apparaît bien que la boucle est bien bouclée, l'utilisation des tableaux de contingence et des coefficients nous avait permis de déboucher sur un indice de similarité nouveau : l'indice de similarité de « Lerman modifié ». Le passage par les notations « relationnelles vectorielles en extension » nous permet réciproquement de retrouver le coefficient de **contingence de Lerman** comme un cas d'illustration contingentielle du coefficient de similarité Tetrachorique ou de Bravais Pearson.

Attention, c'est bien par un « jeu d'écritures » que nous avons obtenu cette équivalence, car nous avions vu que l'indice L(x,y) était égal au produit $S_T(x,y)S_L(x,y)$, nous ne démontrons pas ici que l'égalité $S_T(x,y)=L(x,y)$ est vraie, mais que dans l'« espace relationnel » l'équivalence des formules existe.

Ce que nous venons d'obtenir pour quelques indices de similarité nous permettant de redécouvrir des coefficients de contingence connus ou moins connus, pourrait être généralisé à l'ensemble des indices de similarité présentés dans ce travail, nous laissons le soin au lecteur de voir (à titre d'exercice) si certains indices ainsi listés auraient des formes contingentielles intéressantes ou originales ?

6. Conclusion sur les Indices de Similarités et leur Usage

Comme nous venons de le voir , nous avons passé en revue plus d'une trentaine d'indices de similarités différents (auxquels on peut ajouter ceux issus de la définition des restrictions binaires des « indices d'abondance »). Nous en avons présenté plus de 35 et listé 28 des principaux, certains de ces indices sont d'usage courant, d'autres sont plus ésotériques, mais ils existent en tant que tels et peuvent donc être utilisés dans des conditions particulières.

Comme nous l'avons vu également un certain nombre d'entre eux ont été définis soit explicitement , soit historiquement, soit par l'usage.

Ainsi les indices de Dice, de Jaccard, de Sørensen, d'Anderberg, de Rogers-Tanimoto, de Sokal et Sneath, de Sokal et Michener, de Simpson, de Braun-Blanquet, d'Ochiai, de Yule, d' Urbani-Baroni-Buser, de Kulczinski, de Mac Connaughey , de Russel-Rao, ont été introduits directement par ces auteurs, pour des usages qui se rapportaient à leurs besoins. A ce propos un nombre important des indices cités ont été reinventés au fil du temps, après la découverte initiale, par des spécialistes de disciplines différentes (voir à ce propos la note de bas de page n°8). Nous avons tenté dans cet article d'attribuer aux indices le nom de l'« inventeur » le plus « ancien » ou le plus sûr . Par ailleurs nous avons vu qu'il y a une deuxième classe d'indices, ceux qui n'ont pas été introduits de façon directe, mais qui l'ont été à la suite de raisonnements ou de déductions.

Ainsi en construisant des indices d'association à partir de tableaux de contingence (2x2) avons nous retrouvé en fin de calculs, soit des indices, déjà introduits par ailleurs, soit des indices originaux comme le sont par exemple les indices dérivés du Coefficient de Lerman ou du coefficient de Goodman Kruskal, soit enfin nous en avons obtenus par déduction par rapport à des propriétrés mathématiques générales (exemple : Indices d'Hadamard) . Ils n'étaient, à ma connaissance pas connus des utilisateurs d'indices classiques (ce qui prouve qu'il y a de la place pour en trouver de nouveaux). Les indices bâtis autour des « 4 ratios » (à partir de l'idée originale d'Anderberg) ne sont pas, eux aussi, très connus, et très pratiqués. Pour s'éloigner des pistes usuelles des approches comme celle de Tversky sont intéressantes, au sens qu'elle a permis d'introduire de nouvelles familles d'indices à la suite de raisonnements d'ordre logique, même si les indices les plus connus de cette famille d'indices « Tverskiens » sont les indices de Rappel et de Précision (qui ne sont rien d'autre au fond , suivant le cas, que l'indice de Simpson ou l'indice de Braun –Blanquet). Cette approche fondée sur le raisonnement logique nous avait également permis d'introduire à la suite d'un travail sur les bornes « à la majorité » de ces indices, les indices auxquels nous avons donné notre nom. Enfin une place particulière doit être faite à l'indice que nous avons appellé indice Tetrachorique, au sens qu'il est au centre des convergences des calculs effectués sur les 4 quantités du tableau de contingence du même nom.

D'un point de vue pratique le lecteur de cet article, serait sans doute soucieux, tout au moins nous le pensons, d'avoir une clé de lecture lui permettant de savoir quels sont les « bons indices de similarités », en d'autres termes de faire une selection des meilleurs d'entre eux, pour un usage généraliste.

C'est exactement ce que nous proposons ci après.

Liste d' « Indices Candidats » comme Indices validés d'usage courant

Avant tout il convient de se donner des critères permettant de définir ce qu'est un « bon » indice. En voici une liste possible, hiérarchisée, correspondant aux cas d'usage les plus classiques, d'autres pourraient également convenir dans certaines configurations particulières d'utilisation ou de domaines d'application :

1. Un « bon » indice se devra de varier de 0 à 1 (ou du moins de valoir 1 pour S(x,x)) (autosimilarité maximale « normée »)
2. Un « bon » indice devra vérifier une borne de Solomon Fortier égale à « m/2 » pour 11(x,y), en cas de disjonction complète
3. Un « bon » indice devra vérifier si possible la Transitivité Indicielle Généralisée, ce qui signifie que la dissimilarité associée pourra être une « distance » et dans le cas « favorable » une distance euclidienne
4. Un « bon » indice se doit d'avoir une structure de calcul linéaire des valeurs du numérateur et du dénominateur de l'indice faisant jouer un rôle aux quantités : 11(x,y), 01(x,y), 10(x,y) et 00 (x,y). ceci du fait que sinon les calculs générés sont génants et coûteux pour les applications induites.

Par rapport à cette liste hierarchisée, nous voyons que pour les **Indices du Groupe I** : ceux de « **Dice** », et d'« **Ochiai** », qui satisfont aux quatre exigeances : 1, 2, 3 , 4 , sont donc les indices ayant le spectre le plus général, qu'il faut choisir en priorité. On choisira ensuite les indices de **Jaccard** (1,3,4) et de **Kulcsinsky** (1,2,4) qui ne vérifient que 3 des propriétés précédentes Puis suivant l'usage qui peut en être fait, les indices de « **Rappel** » et de « **Précision** » (ou indices de Simpson & Braun Blanquet) qui sont incontournables dans le domaine du Text Mining , ou du moins dans le domaine de la recherche documentaire.

On peut rajouter à cette liste pour les I**ndices du Groupe II** : les indices suivants : l'indice « **Tetrachorique affine** » (c'est à dire la forme ramenée à varier de 0 à 1 de cet indice) qui vérifie les points 1 et 2 , l'indice de « **Marcotorchino-Michaud2 »** (qui vérifie les conditions 1, 2, 4), de même pour l'indice de **Moyenne Arithmétique des 4 ratios d'Anderberg** (qui vérifie les conditions 1, 2 et 4 également) . On rajoutera dans la mesure où l'usage des configurations 00(x,y) s'impose, le coefficient de **Sokal et Michener** (puisqu'il est une variante du critère de Rand ou de Condorcet) .

D'une façon plus précise, l'utilisation et le choix d'un indice peut dépendre d'habitudes ou de normes d'usage, ainsi nous avons vu qu'en « chimie moléculaire » un certain nombre d'indices sont souvent proposés et utilisés, qui ne le sont par aucune autre communauté scientifique (par exemple l'indice de Mac Caunnaughey, cité dans N. Jeliazkova (JN2007) qui est quasiment inconnu des autres domaines d'usage).
On voit également que même au niveau du vocabulaire, l'usage d'une communauté fait souvent référence à un nom d'indice, pour un usage approprié, sans se poser la question de sa signification en dehors de ses propres normes. A ce propos les remarques intéressantes sur l'Indice de Simpson ou (de celui de Braun Blanquet) dans Bradshaw (1997) et Boyce et Ellison (2001) caractérisant leur usage auprès de la communauté des chimistes moléculaires est à comparer aux rôles joués par les indices de « Rappel » et de « Précision » auprès de la communauté des spécialistes du Traitement des langues (TAL) Parfois, on aboutit à des

redécouvertes ou à des impressions de complexité apparente qui n'existent pas, en fait, si l'on revient à l'expression originelle d'une mesure de similarité et bien entendu si l'on sait à quelle filiation elle appartient. Ainsi dans un ouvrage très récent (2008), et fort intéressant par ailleurs, intitulé « Classification Supervisée de Documents » (2008)(éditions Hermes-Lavoisier), l'auteur, J. Beney fait une digression de trois pages sur les propriétés de la Fonction F(γ) (« dite Fonction F », fonction non triviale du Rappel et de la Précision, chère aux tenants des « métriques » et des évaluations dans le domaine du TAL), alors que cette fonction, même utile, n'est que partielle et reprend des indices connus et très anciens (Dice ou Ochiai .. Kulczinski etc..) amplement étudiés et validés par d'autres[19]. Ce qui est dit ici à propos des communautés du TAL ou de la Chimie Moléculaire aurait pu être étendu tout aussi bien aux communautés de la zoologie, de la biométrie, de la phylogénie, de la phytosociologie etc…

Il existe encore beaucoup de choses à dire sur les Indices de Similarité, nous arrêtons ici ce long descriptif , mais d'autres analyses complémentaires et exhaustives seraient encore utiles. A titre d'exemple, les réflexions autour d'approches logiques simples ou logiques floues, telles que celles proposées par l'équipe de B. Bouchon-Meunier du LIP6, qu'on pourra consulter dans Bouchon Meunier-Marsala (2003) et (2000) ainsi que dans Lesot, Rifqi, Benhadda (2009), d'une part, ou des approches jouant sur la pondération de l'ordre d'apparition des « matchings », comme celles exposées dans les articles de C. Michel (2001) ou L. Egghe (2006), d'autre part, permettraient d'organiser les familles d'indices suivant des axes de nature, certes différente, mais de fait complémentaire à celle dont la structure a été proposée dans ce texte.

7. Références

Anderberg M.(1973) : « *Cluster Analysis for Applications* », Book n°19, Probability and Statistics Series, Academic Press, New-York,

Agresti A., Morey L (1984).: « *An adjustment of the Rand's Statistic for chance agreement* », Journal of Educational and Psychological Measurement, Vol44, pp.33-37,

Ah-Pine J. (2007): « *Sur les Aspects Algébriques et Combinatoires de l'Analyse Relationnelle* », Thèse de l'Université Paris VI, (2007)

Arabie P. , Hubert L.(1985) :« *Comparing Partitions* », Proceedings of the Fourth European Meeting of the Classifications Societies, Cambridge,

[19] Ainsi F(1)= Indice de Dice , F(0)= Indice de Précision,, F(∞)= Indice de Rappel . Ok tout çà existe déjà , mais quid des autres fonctions évidentes de R et P appartenant au Groupe I Type II (comme l'indice d'Ochiai (qui est limite du Chi²de contingence), où l'indice de Tversky), pourquoi les gens du TAL ne les considèrent-ils pas et ont-ils privilégié la fonction F ?.

Benhadda H. (1998) : « *La Similarité Régularisée et ses applications en Classification automatique* », Thèse de l'Université Paris VI,

Baulieu F. B. (1989) : «*A classification of presence/absence based dissimilarity coefficients*», Journal of Classification, Vol. 6,, pp. 233-246,Springer-Verlag, New York, Berlin, Braun –Blanquet J. (1932) : « *Plant Sociology, the Study of Hart Communities* » Livre, Mac-Graw Hill, New York,

Benavent C. (2001): « *Analyse des Proximités* », Rapport de l'IAE des Pays de l'Adour, 17 pages, disponible sur le site :christophe.benavent @free.fr

Beney J. (2008): « *Classification Supervisée de Documents* », livre publié aux éditions Hermes-Lavoisier),

Bisson G. (2000): « *La similarité: une notion symbolique/numérique* ». Livre : Apprentissage symbolique-numérique (tome 2). Eds Moulet Brito. Editions CAPADUES,

Bradshaw J. (1997) : « *Introduction to Tversky similarity measure*», Proceedings of MUG '97 - 11th Annual Daylight User Group Meeting Février 1997,

Brunchwicg L. (1921) : «*Les Etapes de la Philosophie Mathématique* » , Livre, A. Blanchard Editeur, (nouvelle Edition 1981), ancienne édition 1921,

Benhadda H., Marcotorchino F.(1998) : « *Introduction à la Similarité Régularisée* », Revue de Statistique Appliquée, n°56, pp. 45-69,, Dunod, Paris,

Bouchon-Meunier B., Marsala C. (2003) : « *Logique floue, principes, aide à la décision* », livre Collection Information-Commande-Communication, Hermes-Lavoisier Editeurs , Paris,

Boyce R. L., Ellison P. C.(2001): « *Choosing the Best Similarity Index When Performing Fuzzy Set Ordination on Binary Data*», Journal of Vegetation Science, Vol. 12, No. 5 pp. 711-720,

Baroni-Urbani C., Buser M.W. (1976) : « *Similarity of binary data* », Journal of Syst. Zool. n° 24 , pp**:** 165-178.,

Burnaby T.P. (1970): «*On a method for character weighting similarly coefficient, employing the concept of information* », Journal of Math. Geology., Vol. 2, n° 1, pp. 25-38, Carnap R. (1928) : « *Der Logische Aufbau der Welt* », Weltkreis Verlag , Berlin ,

Caillez F., Pages J.P. (1976): : «*Introduction à l'Analyse des Données* », Publications de l'ASU et du BURO, éditeur la SMASH,

Cohen J. (1960) : «*A coefficient of agreement for nominal scales* », Educational and Psychological Measurement Journal, Vol 20, pp37-46,

Chandon J.L**.,** Pinson S. (1981) : « *Analyse Typologique : Théorie et Applications* », Masson, Paris,

Decaestecker C. (1992): « *Apprentissage en Classification Conceptuelle Incrémentale* » , Thèse de l'Université Libre de Bruxelles (Faculté des Sciences),

Dice L.R. (1945): « *Measures of the amount of ecological association between species* », Ecology Journal, Vol 26, pp.297-302,

Deheuvels P., Marcotorchino F. (2000) : « *Statistique et Informatique, la Nouvelle Convergence* », Revue RST de l'Académie des Sciences n°8, Juillet TECD ,

Driver, H. E., et Kroeber, A. L.(1932) : «*Quantitative expression of cultural relationship*». The University of California Publications in American Archaeologyand Ethnology, 31, 211-256,

Egghe L. , Rousseau R (2006).: «*Classical retrieval and overlap measures satisfy the requirements for rankings based on a Lorenz curve*» , in Information Processing & Management, Vol 42, Issue 1, pp.106-120,

Fleiss J.L.(1975): « *Measuring Agreement between two judges in the presence or absence of a Trait*», Biometrics, N°31, pp 651-659,

Fowlkes E.B., Mallows J. (1983) : « *A method for comparing two hierarchical clusterings*», JASA (Journal of the American Statistical Association), Vol 78, pp.553-584,

Goodall D.W.: (1966) « *A new Similarity Index based on Probability*» , Biometrics, n° 22, pp. 882- 907,

Goodman L.A., Kruskal W.H. (1979) : « *Measures of Association for Cross Classification* », Book by Springer Verlag , Berlin, New-York,

Gower J.C. (1966): « *Some distance properties of latent root and vector methods used in multivariate analysis*», Biometrika, n°53, pp:325-338,

Gower J.C. (1971): « *A General Coefficient of Similarity and some of its Properties*», Journal of Biometrics, n°27, pp:857-874,

Gower J., Legendre P. (1986) : « *Metric and Euclidean properties of dissimilarity coefficients*», Journal of Classification, N°3, pp.5-48, North Holland,

Grötschel M., Jünger M., Reinelt G. (1982) : « *A Cutting Plane Algorithm for the Linear Ordering Problem*», Research Report N°82219 Operations Research, Universität zu Bonn, Germany,

Green P., Rao V.R.(1969) : «*Note on proximity Measures and Cluster Analysis» , Journal of Marketing Research*» , Vol6, pp.359-364,

Guttmann L. (1941): « *The Quantification of a Class of Attributes , A theory and Method of Scales Construction*», Horst P. Editor, Social Sciences Research Council, New York ,

Hicham A.,Saporta G. (2003) : «*Mesures de distance entre modalités de variables qualitatives; application à la classification* », Revue de Statistique Appliquée, Vol 51, n°2, pp. 75-90,

Idrissi Amal N. (2000): «*Contribution à l'Unification de Critères d'Association pour Variables Qualitatives* » , Thèse de l'Université Paris VI,

Jaccard P. (1908): *« Nouvelles Recherches sur la distribution florale »* , Bulletin de la Société Vaudoise des Sciences Naturelles, Vol n°44, pp.223-270,

Joly S., Le Calvé G.(1986) : « *Metric and Euclidean Properties of Dissimilarity Coefficients* », Revue de Statistiques et Analyse des Données n°11, pp :30-50)

Joly S., Le Calvé G. (1994) : « *Similarity functions* », Lecture Notes in Statistics, (In Van Cutsem, B. editor), Springer-Verlag, vol. 93, pp. 67-86,

Jeliazkova N. (2005): « *Chemical Similarity* », European Chemicals Bureau (ECB) Workshop on Chemical Similarity and Threshold of Toxicological Concern (TTC) Approaches , Ispra, Italy,

Jackson, A.A., Somers, K.M. et Harvey, H.H. (1989): « *Similarity coefficients: measures for co-occurrence and association or simply measures of occurrence?* », Am. Nat. Journal Vol:*133*: pp.436-453,.

Janson S. , Vegelius J. (1982): «*Correlation Coefficient for more than one Scale Type*», Multivariate Behavioral Research Journal, Vol 17, Issue 2, pp.271-284,

Kulczinski S. (1927) : «*Classes des Sciences Mathématiques et Naturelles*», Bulletin International de l'Académie Polonaise des Sciences et des Lettres, pp57-203,

Lerman I.C. (1970): « *Les Bases de la Classification Automatique*», Livre chez Gauthier-Villars, Paris,

Lerman I.C. (1981): « *Classification et Analyse Ordinale des Données* », Livre, Dunod, Paris,

Lerman I.C. (1987) : « *Construction d'un indice de similarité entre objets décrits par des variables d'un type quelconque, application au problème du consensus en classification* » Revue de Statistique Appliquée, vol. 35, n° 2, pp :39-60,

Lelu A. (2002): «*Comparaison de trois mesures de similarités utilisées en documentation automatique et analyse textuelle* », Proceedings des 6ème JADT (6 èmes Journées d'Analyse des Données Textuelles),

Lesot M.J., Rifqi M. et Benhadda H. (2009): *« Similarity Measures for binary and numerical Data* » , pp 63-84 in Journal of Knowledge Engineering and Software Data Paradigms, Interscience Editor,

Marcotorchino F. (1989): « *Liaison Analyse Factorielle-Analyse Relationnelle (I) : "Dualité Burt-Condorcet*», Etude du Centre Scientifique IBM France, No F142,

Marcotorchino F. (1984) : « *Présentation des Critères d'Association en Analyse des Données Qualitatives* », Publication AD0185, Université Libre de Bruxelles, pp :1-57, (1984).

Marcotorchino F. (1987): « *Block seriation problems: A unified approach*», Applied stochastic models and Data Analysis Journal N°3, pp.73–91,

Marcotorchino F.(1991) : « *L'analyse Factorielle-Relationnelle (parties 1 et 2*) ». Etude du Centre Scientifique IBM France, M06 pp :1à 122,

Marcotorchino F., Benhadda H.(1996) : « *Introduction à la Similarité Régularisée* », Etude MAP011, pp1-78 du Centre Européen de Mathématiques Appliqués, ECAM/IBM,

Marcotorchino F., El Ayoubi N. (1991): «*Paradigme logique des écritures relationnelles de quelques critères fondamentaux d'association* » Revue de Statistique Appliquée, Vol 39, n°2, pp :25-46 ,

Messatfa H. (1989) : « *Unification Relationnelle des Critères et des Structures de Contingences*», Thèse de l'Université Paris VI, LSTA,

Marcotorchino F., Michaud P. (1978) : « *Optimisation en analyse ordinale des données* ». Livre Masson, Paris,

Marcotorchino F., Michaud P. (1981) .: « *Agrégation des Similarités en Classification Automatique* », Revue de Statistique Appliquée, Vol 30, n°2 , Paris,

Michel C. (2000) : «*Ordered similarity measures taking into account the rank of documents* », Journal d' Information Processing and Management, n°37, pp. 603-622 ,

Maggira G.M., Petke J.D., Mestres J. (2002): « *Similarity Indexes for Chemistry*» in Journal of Mathematical Chemistry », Vol 31, N°3,

Ochiai A. (1957): «*Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring* », Bulletin of the Japanese Society for Scientific Fisheries, N°22, pp.526-530,

Parrochia D. (1992): «*Mathématiques & existence :Ordres Fragments, Empiètements* », Publication :Seyssel, Champ Vallon , Collection Milieux,

Quine W. van O. (1998): *The Pre-Established Harmony of Subjective Perceptual Similarity"* in Proceedings of the Twentieth World Congress of Philosophy, Volume 6**,** (Boston, August, 1998) (reprinted in W. V. Quine.: *Confessions of a Confirmed Extensionalist*) in Floyd, Juliet and Shieh, Sanford (editors) (2008).

Rand W.H. (1971): «*Objective Criteria for the Evaluation of Clustering Methods* », Journal of the American Statistical Association, Vol. 66,

Rifqi M., Berger V. , Bouchon- Meunier B. (2000): « *Discrimination Power of Measures of Comparison* » , Fuzzy Sets and Systems, vol.110, pp. 189-196,

Rogers D.J., Tanimoto T.T. (1960): « *A computer program for classifying plants* », Vol 132, pp 1115-1118, Science Journal ,

Saporta G. (1990): «*Probabilités Analyse des Données et Statistique* », livre (2ème édition augmentée) , Editions Technip (2006), (1ère édition (1990))

Salton G. (1968) : « *Automatic Information Organization and Retrieval* », book, Mac Graw Hill Editor,New York,

Salton G, Mac Gill M.J. (1983) : «*Introduction to Modern Information Retrieval* », Mac Graw Hill Editor, New York)

Schoenberg I.J. (1938): «*Metric Space and Positive Definite Functions* », Transactions of the American Mathematical Society, n°44, PP:522-536, New York,

Solomon H., Fortier J. (1966): «*Clustering Procedures* », Multivariate Analysis, P. Krishnaiah (Editor), Academic Press, New york,

Simpson G.G.(1943) : «*American Journal of Science* », N°241, pp.1-31,

Sneath, P.H.A. and Sokal, R.R. (1973). «*Numerical Taxonomy: the Principles*», Book Mac Graw Hill,

Sokal R.R., Michener C.D.(1958) : «*A Statistical Method for Evaluating Systematic Relationships* », University of Kansas Science Bull., n°38, pp:1409-1438,

Sokal R.R,. Sneath P.H.A., (1963): «*Principles of Numeric Taxonomy* », livre, Freeman Editeur, San Francisco,

Sørensen T. (1948) « *A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons* »,. K. Dan. Vidensk. Selsk. Biol. Skr. Vol n°5: pp.1-34,

Sparck Jones K. (1971) : «*Automatic keyword classification for retrieval* », book by Butterworth, London,

Snijders T.A. et al. (1990): «*Distribution of some similarity coefficients for dyadic binary data in the case of associated attributes* », Journal of Classification n° 7, , pp. 5-31, Springer-Verlag,

Tversky A. (1977): « *Features of Similarity* » Psychological Reviews Vol 84 n°4 pp:327-352,

Van Cutsem B. (editor) (1994) : «*Classification and Dissimilarity Analysis* », book by Springer-Verlag , Collection: Lecture Notes in Statistics n°93, New-York, Berlin,

Van Rijsbergen C.J.(1979): «*Information Retrieval*», Livre Publié par Butterworth-Heinemann , Newton, Massachussets, USA,

Warrens M.J. (2008): «*On the Indeterminacy of Resemblance Measures for Binary (Presence/Absence) Data*», Journal of Classification, Vol 25, Issue n°1, pp.125-136, Springer Verlag, Berlin,

Warrens M.J. (2009): «*Bounds of Resemblance Measures for Binary (Presence-Absence) Variables*», Journal of Classification, Vol 25, Issue n°2, pp.195-208, Springer Verlag, Berlin,

Yule G.U. , Kendall M.G (1950).: «*An Introduction to the Theory of Statistics* », 14th edition, book by Hafner , New York, (1950). New published edition by Arnold (1976).

	Nom de l'Indice	Emplacement de définition ou Renvois
1	Anderberg Simple	Défini en 4.1.1.3
2	Anderberg Complémentaire	Défini en 4.1.1.4
3	Anderberg (4 Ratios)	Défini en 4.4.4.4
4	Baroni-Urbani -Buser	Défini en 4.4.4.7
5	Borko	(voir ➔ Sokal et Michener)
6	Braun -Blanquet	Défini en 4.3.2 (voir ➔ Min, voir ➔ Simpson➔ voir Max)
7	Bravais - Pearson	Défini en 4.4.4.1 (voir ➔Tetrachorique)
8	Bray et Curtis	Défini en 4.5.1.2 (voir ➔Dice, voir➔ Odum)
9	Carbo	Cité en note de bas de page n°9 (Voir ➔ Ochiai)
10	Cohen (Kappa)	Défini en 4.4.5.5
11	Condorcet	(voir ➔ Rand➔ voir Sokal et Michener)
12	Czekanowski	(voir ➔ Dice)
13	Dice	Défini en 4.1.1.1 (voir ➔Czekanowski)
14	Fonction F	Défini en 4.3.2.3 (voir ➔Van Rijbergen➔ voir ➔Tversky)
15	Fleiss	Défini en 4.4.4.2.b
16	Fowlkes-Mallows	Défini en 5.1.2
17	Goodman - Kruskal (Lambda)	Défini en 4.4.5.4 (voir ➔Dice)
18	Goodman - Kruskal (Tau)	Défini en 4.4.5.2 (voir ➔ Tetrachorique)
19	Gower	Cité en note de bas de page n°5
20	Green et Rao	(voir ➔Sokal et Michener➔ voir Hamann)
21	Hadamard	Défini en 4.4.4.2.c
22	Hamann	Défini en note de bas de page n°12
23	Hodgkin -Richards -Petke	Défini en note de bas de page n°9 (voir ➔ Dice)
24	Hubert-Arabie	Défini en 5.1.3 (voir Kulczynski)
25	Indice du Max (P,R)	Défini en 4.3.2 (voir ➔ Simpson)
26	Indice du Min (P,R)	Défini en 4.3.2 (voir ➔Braun-Blanquet)
27	Jaccard	Défini en 4.1.1.2
28	Janson -Vegelius	Défini en 4.4.5.1 (voir ➔ Hamann, voir➔ Sokal-Michener)
29	Kulczynski (abondance)	Défini en 4.5.2.1 (voir ➔ Dice)
30	Kulczynski (0-1)	Défini en 4.2.1.1
31	Kulczynski (4 Ratios)	Défini en 4.4.4.4
32	Lance- Williams (abondance)	Défini en 4.5.1.1 (voir ➔ Sokal Michener)
33	Lerman « Modifié »	Défini en 4.4.5.3
34	Mac Connaughey	Défini en 4.3.2.2 (voir ➔ Kulczynski)
35	Marcotorchino-Michaud (1)	Défini en 4.4.3.2
36	Marcotorchino-Michaud (2)	Défini en 4.4.3.3
37	Marczewski- Steinhaus (abondance)	Défini en 4.5.2.2 (voir ➔ Jaccard)
38	Maxwell-Pilliner	Défini en 4.4.4.2.a
39	Morey-Agresti	Défini en 5.1.1
40	Ochiai	Défini en 4.2.1.2
41	Ochiai (4 Ratios)	Défini en 4.4.4.5
42	Odum	Défini en 4.5.1.2 (voir ➔ Bray-Curtis)
43	Précision (Indice de)	(voir ➔ Rappel, voir ➔ Tversky)
44	Rand	Défini en 4.4.5.1 ➔ (voir Janson et Vegelius)
45	Rao	Défini en 4.4.3.1
46	Rappel (Indice de)	(voir ➔ Précision, voir➔ Tversky)
47	Rogers et Tanimoto	Défini en 4.4.1.2
48	Sokal et Michener	Défini en 4.4.1.1 (voir ➔Green et Rao)
49	Sokal et Sneath	Défini en 4.4.1.3
50	Sørensen	Défini en 4.1.1.4
51	Sorgenfrei	Cité en note de bas de page n°8
52	Tetrachorique (Coefficient)	Défini en 4.4.4.1 (voir ➔ Bravais Pearson)
53	Tversky	Défini en 4.1.3
54	Van Rijsbergen	Cité en note de bas de page n°11
55	Yule Q et Y	Définis en 4.4.4.3