Présentation et Structuration de quelques indices de similarité et Analyse Relationnelle

François Marcotorchino(1,2), Hamid Benhadda(1)
JeanFrancois.Marcotorchino, hamid.benhadda@fr.thalesgroup.com

(1) : Thales Division Land and Joint / CeNTAI (Centre des Nouvelles Technologies de l'Analyse de l'Information) 160, boulevard de Valmy – BP 82 – 92704 Colombes Cedex – France

(2) : Directeur de Recherches au LSTA (Laboratoire de Statistiques Théoriques et Appliquées) Paris VI

Résumé : Dans cette communication, nous présenterons une restructuration de quelques indices de similarité parmi les plus étudiés dans la littérature. Nous montrerons le rôle central occupé par l'indice de Dice et nous évoquerons, succinctement, la théorie de la « similarité régularisée » qui apporte une réponse à la problématique de pondération des attributs décrivant les objets étudiés. L'idée fondamentale à la base de cette similarité est le ré-équilibrage des poids intrinsèques des attributs en diminuant l'influence des attributs à fort poids structurel et en augmentant celle des attributs à faible poids structurel.

Mots-clés : Analyse relationnelle, Indices de similarité, Similarités régularisées

Abstract : In this communication, we shall present a reorganization of some similarity coefficients among the most studied in the literature. We shall show the key role played by the Dice's similarity index and we shall evoke, briefly, the theory of the "regularized similarity" which brings an answer to the problem of attribute's weights. The basic idea behind this similarity is to balance of the intrinsic weights of the attributes by decreasing the influence of the attributes with strong structural weight and by increasing those with weak structural weight.

Keywords : Relational Analysis, Similarity Indexes, Regularized Similarity

1. Introduction

Dans la nature, tous les êtres sont par essence différents les uns des autres, et toute assertion selon laquelle deux objets se ressemblent ne peut donc être que relative. En effet, deux objets peuvent être si ressemblants qu'ils en arrivent à être considérés comme identiques par rapport à quelques attributs mesurées sur eux, mais ils ne peuvent l'être dans l'absolu puisqu'il existera toujours des attributs non mesurés mais spécifiques à l'un ou l'autre des deux objets sur lesquels ils se différencieront. La recherche pour quantifier ces ressemblances sous la forme d'indices de similarité, dans un but d'aide à la décision ou de classification automatique par exemple, a donné lieu à une abondante littérature. Bon nombre de ces indices, proposés dans différents domaines tels que les sciences humaines, la sociologie, l'ethnologie, la linguistique ou la classification automatique, pour ne citer que ceux là, ont fait l'objet de nombreuses publications au fur et à mesure de leur utilisations potentielles.

La notion de similarité a plus un caractère ontique qu'ontologique et lorsque l'on cherche à définir une « similarité » entre objets, on doit s'appuyer largement sur l'empirisme et l'intuition plutôt que sur la logique pure. Nous reprenons ici trois propositions, relatives à la similarité, de W.V.O. Quine telles qu'elles sont présentées par D. Parrochia[1] (page 125), en les transformant légèrement en une vue plus mathématique que philosophique. Nous noterons $S_{ii'}$ la similarité entre deux objets i et i'.

[1] Parrochia Daniel (1991). « *Mathématiques & existence. Ordres. Fragments. Empiétements* », Champ Vallon, Paris.

Proposition 1 : *On ne peut simplement et uniquement fonder la similarité entre deux objets* i et i', *sur le nombre de propriétés (ou d'attributs) partagés par ces deux objets.*

En effet, le problème « extra logique » qui subsiste ici est tout simplement celui de savoir ce qui devra être considéré comme une propriété.

Proposition 2 : *Si les objets* i et i' *partagent plus de « propriétés » que les objets* j et j', *alors on doit avoir :* $S_{ii'} \geq S_{jj'}$, *dès lors que notre intuition désigne les propriétés de descriptions comme équivalentes en poids et en signification.*

Proposition 3 : *On ne pourra pas dire que* $S_{ii'}$ *a une valeur de similarité maximale si* i et i' *partagent toutes les propriétés de l'univers de description choisi, si cet univers étant extensible, ils différent sur les nouvelles propriétés rajoutées.*

Dans notre communication, nous présenterons une restructuration de quelques indices de similarité les plus étudiés dans la littérature. Ces similarités seront calculées sur des variables catégorielles ou de type présence-absence. Nous parlerons également, brièvement, de la théorie de la « Similarité Régularisée », théorie co-développée par F. Marcotorchino et H. Benhadda[2][3], qui apporte une réponse à la problématique de pondération des variables.

Cette similarité consistera donc à ré-équilibrer les poids intrinsèques des variables (descripteurs) pour diminuer l'influence des variables à fort poids structurel et pour augmenter celle des variables à faible poids structurel.

Grâce à cette approche qui trouve toute son acception dans sa liaison avec la « Théorie de l'Analyse Relationnelle », nous pouvons dériver des standard généraux et des formes invariantes qui trouveront ensuite une application évidente dans le domaine des mesures de similarités (on dit aussi dans ce cas Mesures d'Association) entre variables[4] (mesures de : Rand, Janson-Vegelius, Khi 2, Michalsky, de Jordan etc..), l'apport de la « Similarité Régularisée » permet de retrouver des paradigmes d'écritures dérivées entre tous ces critères permettant une interprétation très fructueuse par rapport aux notions classiques d' « indépendance statistique » et moins classique d' « indétermination contingentielle », qui leurs sont propres.

2. Les mesures de similarité (définition générale)

Les mesures de similarité ont pour objet de quantifier le degré de ressemblance ou de dissemblance que peuvent entretenir entre eux les individus (ou objets) d'une population I donnée. Ces similarités sont quantifiées sur la base d'attributs (ou de modalités) dérivés de variables (ou descripteurs) mesurées sur cette population.

D'un point de vue mathématique une mesure de similarité S est définie sur l'ensemble $I \times I$ et à valeurs dans l'ensemble des réels positifs R^+ :

[2] Benhadda Hamid.(1998). « *La similarité régularisée et ses applications en classification automatique* », Thèse de l'université de Paris VI, Paris

[3] Benhadda Hamid. et Marcotorchino François. (1998). « *Introduction à la similarité régularisée en analyse relationnelle* », pages 45-69, Revue de Statistique Appliquée, vol. 46, N° 1,

[4] Idrissi Amel.(2000). « *Contribution à l'Unification de critères d'Association pour Variables Qualitatives* », Thèse de l'université de Paris VI.

$$S : \quad I \times I \to R^+$$
$$(i,i') \to S_{ii'}$$

et qui vérifie les propriétés suivantes :

La symétrie : *La similarité entre i et i' est égale à la similarité entre i' et i :*

$$S_{ii'} = S_{i'i} \qquad \forall i,i' \in I$$

L'auto similarité maximale : *La similarité d'un objet i avec lui même est supérieure ou égale à sa similarité avec n'importe quel autre objet de la population :*

$$S_{ii'} \leq S_{ii} \qquad \forall i' \in I$$

ou de façon plus symétrique :

$$S_{ii'} \leq Min(S_{ii}, S_{i'i'}) \qquad \forall i,i' \in I$$

3. L'analyse relationnelle : principes de base

L'analyse relationnelle est une théorie à vaste champ applicatif, développée à IBM dans les années 1970, par F. Marcotorchino et P. Michaud. Son principe consiste à construire, à partir d'une matrice rectangulaire croisant des objets et des variables, une matrice carrée C, appelée matrice globale de Condorcet, représentant les relations qui existent entre tous les couples d'objets de la population étudiée sur l'ensemble des variables.

3.1. Matrice relationnelle unitaire

L'étape préliminaire pour la construction de la matrice de Condorcet[5] C consiste à transformer chaque variable d'origine, V^k, mesurée sur la population, en une matrice relationnelle unitaire de Condorcet C^k de terme général[6] $c^k_{ii'}$ tel que :

$$c^k_{ii'} = \begin{cases} 1 & \text{si } i \text{ et } i' \text{ possèdent la même modalité (valeur) de la variable } V^k \\ 0 & \text{dans le cas contraire} \end{cases}.$$

$c^k_{ii'}$ peut être considéré comme une mesure de similarité entre les deux objets i et i' par rapport à la variable V^k.

3.2. Matrice relationnelle globale de Condorcet

Dans le cas où l'on est en présence de M variables (catégorielles ou de présence-absence) mesurées sur N objets, la matrice C s'obtient en sommant les M matrices relationnelles unitaires représentant les M variables d'origine :

$$c_{ii'} = \sum_{k=1}^{M} c^k_{ii'}$$

[5] Du nom du célèbre mathématicien et homme politique français : Antoine Caritat Marquis de Condorcet qui a introduit la notion de « comparaisons par paires » en 1785.

[6] Le terme général de la matrice relationnelle unitaire peut avoir différentes signification selon le type d'attribut. Il peut prendre les valeurs 0 ou 1 dans le cas d'un attribut catégoriel ou de présence-absence, il peut aussi prendre une valeur réelle quelconque si l'attribut est de type numérique.

$c_{ii'}$ représente l'accord (ou le nombre d'attributs communs) partagés par les deux objets i et i' sur l'ensemble des variables.

Comme le nombre d'attributs communs a deux objets donnés i et i' ne peut jamais excéder le nombre d'attributs possédés par chacun d'eux, on en déduit que leur accord $c_{ii'}$ possède les bornes suivantes :

$$0 \le c_{ii'} \le \frac{c_{ii} + c_{i'i'}}{2}$$

la borne supérieur de $c_{ii'}$ représente l'accord maximum potentiel $am_{ii'}$ que deux objets peuvent avoir. Autrement dit :

$$am_{ii'} = \frac{c_{ii} + c_{i'i'}}{2}$$

à partir de l'accord et de l'accord maximum entre deux objets i et i', on définit de façon générique un indice de similarité relationnel $\rho_{ii'}$ comme le rapport de leur accord sur leur accord maximum potentiel :

$$\rho_{ii'} = \frac{c_{ii'}}{am_{ii'}} = \frac{2c_{ii'}}{c_{ii} + c_{i'i'}}$$

cet indice est bien entendu compris entre les valeurs 0 lorsque les deux objets n'ont aucun attribut en commun et 1 lorsqu'ils sont de profils identiques.

Par analogie avec l'accord $c_{ii'}$, qui représente une mesure de similarité, on définit le désaccord représentant une dissimilarité entre i et i' par la relation : $\bar{c}_{ii'} = am_{ii'} - c_{ii'}$

3.3. Règle de la majorité de Condorcet

Que les variables représentent des relations d'ordre du type « i est préférable à i' » ou des relations d'équivalence du type « i est dans le même classe que i' » ou enfin de présence-absence « i et i' possèdent l'attribut u de V^k », la règle de la majorité de Condorcet stipule que les assertions précédentes sont vraies dès lors que l'accord des deux objets est supérieur ou égal à leur désaccord, autrement dit, dès que la condition $c_{ii'} \ge \bar{c}_{ii'}$ est vraie. On peut montrer sans difficulté que cette condition est équivalente à : $\rho_{ii'} \ge \frac{1}{2}$

où la borne $1/2$ représente la borne de Solomon-Fortier[7].

3.4. Exemple illustratif

Considérons une population formée de cinq individus notés {a, b, c, d} sur lesquels on a mesuré les trois variables qualitatives V^1 = « Sexe » , V^2 = « Catégorie socioprofessionnelle » ayant les deux modalités {Cadre , Non Cadre} et V^3 = « nationalité » ayant les trois modalités {Française, Espagnole, Anglaise}. Supposons que, relativement à la variable V^1

[7] Solomon H. et Fortier J.J. (1966). « *Clustering procedures* », In Multivariate Analysis, éd. Par Krishnaiah, Academic Press, New York

les trois premiers individus soient de sexe féminin et les deux derniers de sexe masculin, que relativement à la variable V^2 les deux premiers individus soient des cadres et les trois derniers des non cadres et que relativement à la variable V^3 les deux premiers individus soient de nationalité Française, les deux suivants de nationalité Espagnole et le dernier de nationalité Anglaise.

Représentations vectorielles				**Représentations relationnelles**															
				C^1					C^2					C^3					
V^1	V^2	V^3	Objets	a	b	c	d	e	a	b	c	d	e	a	b	c	d	e
1	1	1	a	1	1	1	0	0	1	1	0	0	0	1	1	0	0	0
1	1	1	b	1	1	1	0	0	1	1	0	0	0	1	1	0	0	0
1	2	2	c	1	1	1	0	0	0	0	1	1	1	0	0	1	1	0
2	2	2	d	0	0	0	1	1	0	0	1	1	1	0	0	1	1	0
2	2	3	e	0	0	0	1	1	0	0	1	1	1	0	0	0	0	1

Remarque : la variable V^1 est codée 1 pour le sexe féminin : individus **a**, **b** et **c** et 2 pour le sexe masculin : individus **d** et **e**, tandis que la variable V^2 est codée 1 pour les cadres : individus **a** et **b** et 2 pour les non cadres : individus **c, d** et **e** et enfin la variable V^3 est codée 1 pour les français : individus **a** et **b**, 2 pour les espagnols : individus **c** et **d** et 3 pour l'anglais : individu **e**. Si nous avions changé dans V^1 le codage 1 en X et le codage 2 en Y ou dans V^3 le codage 1 en 2, 2 en 45 et 3 en 26, il est clair que les relations correspondantes C^1 et C^3 n'auraient pas été changées[8]. Le codage relationnel consiste à représenter la relation entre deux objets en mettant 1 dès lors que deux individus ont la même modalité d'une même variable. Par sommation, terme à terme, des trois matrices relationnelles unitaires C^1, C^2 et C^3, nous obtenons la matrice relationnelle globale de Condorcet C :

	C				
	a	b	c	d	e
a	3	3	1	0	0
b	3	3	1	0	0
c	1	1	3	2	1
d	0	0	2	3	2
e	0	0	1	2	3

[8] Le codage n'a pas de signification spécifique dès lors qu'on a affaire à des variables qualitatives, i.e. sans hiérarchie.

Dans cet exemple, l'accord maximum possible entre deux individus quelconques est égal à 3. La règle de la majorité de Condorcet est vérifiée pour toutes les paires d'individus i et i' tels que : $c_{ii'} \geq \frac{3}{2}$, autrement dit pour les paires : {a , b}, {c , d} et {d , e}.

4. Structuration de quelques indices de similarités de la littérature

Lorsque les données disponibles sont de type présence-absence, on les représente souvent sous la forme d'un tableau binaire K de dimensions $n \times p$ où n est le nombre d'objets présents dans la population étudiée et p le nombre total d'attributs décrivant la population. Le terme général k_{ij} du tableau K est tel que :

$$k_{ij} = \begin{cases} 1 & si\ l'individu\ i\ possède\ le\ descripteur\ j \\ 0 & dans\ le\ cas\ contraire \end{cases}$$

chaque objet i sera représenté par un vecteur $\vec{O}_i = \{k_{i1}, k_{i2}, \ldots, k_{ip}\}$. Grâce au tableau K on définit, pour deux objets i et i' les quantités suivantes :

- Le nombre d'attributs communs aux deux objets $11_{ii'} = \sum_{j=1}^{P} k_{ij} k_{i'j}$
- Le nombre d'attributs possédés par i et non par i' : $10_{ii'} = \sum_{j=1}^{P} k_{ij} \left(1 - k_{i'j}\right)$
- Le nombre d'attributs possédés par i' et non par i : $01_{ii'} = \sum_{j=1}^{P} \left(1 - k_{ij}\right) k_{i'j}$
- Le nombre d'attributs absents chez les deux objets : $00_{ii'} = \sum_{j=1}^{P} \left(1 - k_{ij}\right)\left(1 - k_{i'j}\right)$

Si l'on note de façon générique toute composante d'un objet i par la quantité X_i, alors pour deux objets i et i' quelconques, on peut représenter les quatre quantité précédentes sous forme du tableau de contingence suivant :

	$X_{i'} = 1$	$X_{i'} = 0$
$X_i = 1$	$11_{ii'}$	$10_{ii'}$
$X_i = 0$	$01_{ii'}$	$00_{ii'}$

Les indices de similarité existants dans la littérature peuvent être divisés en deux groupes, ceux qui prennent en compte la configuration 00 dans le calcul des similarités et ceux qui ne le font pas. Les indices dont nous traiterons ici font partie du second groupe.

4.1. Indices obtenus par ratios directs

Parmi le indices qui ont été le plus étudiés et qui sont fonctions de ratios directs des quantités 11, 10 et 01 on peut citer les indices de : Dice-Czekanowski, Jaccard, Sokal-Sneath et Sorensen. Tous ces indices peuvent se présenter de façon générique sous la forme :

$$S_{ii'}(n) = \frac{11_{ii'}}{11_{ii'} + 2^{-n}\left(10_{ii'} + 01_{ii'}\right)} \qquad (F1)$$

où :

$$n = -1 \Rightarrow \quad Sokal - Sneath - Anderberg(S_{ii'}^{SSA})$$
$$n = 0 \Rightarrow \quad Jaccard(S_{ii'}^{J})$$
$$n = 1 \Rightarrow \quad Dice(S_{ii'}^{D})$$
$$n = 2 \Rightarrow \quad Sorensen(S_{ii'}^{S})$$

on pourrait ajouter à cette liste, pour des raisons de symétrie par rapport à l'indice de Dice, un indice dans lequel $n = 3$ et que nous appellerons indice « symétrique » d'Anderberg $S_{ii'}^{SA}$.

Remarque : si l'on définit 1_i comme le nombre d'attributs possédés par un objet i, i.e. $1_i = \sum_{j=1}^{P} k_{ij}$ il est facile de montrer que :

$$10_{ii'} + 01_{ii'} = 1_i + 1_{i'} - 2 \times 11_{ii'}$$

ce qui permet d'obtenir une autre formulation générique, fonction des nombres d'attributs possédés par chacun des objets et de n :

$$S_{ii'}(n) = \frac{11_{ii'}}{\left(1 - 2^{-n+1}\right)11_{ii'} + 2^{-n}\left(1_i + 1_{i'}\right)} \qquad (F2)$$

l'indice de Dice-Czekanowski, a été introduit initialement par Czekanowski dans la théorie des matrices de confusions en 1913. Il a ensuite été étudié par Dice en 1945 dans le domaine de la botanique et de la phylogénie. Cet indice a un rôle central et globalisant qui n'a pas été explicité jusqu'à présent, et dont nous montrerons le caractère fondamental dans la structuration des indices de similarité que nous venons de citer. Nous montrerons d'autre part qu'il est équivalent, sous certaines conditions, à la notion de mesure majoritaire de comparaisons par paires (due au Marquis A. de Condorcet 1785).

L'ensemble des différents indices que nous avons cités peuvent se définir à partir de fonctions homographiques de l'indice de Dice. C'est cette propriété que nous avions décrite[9] en 1981. elle a été reprise et étudiée dans l'article fort complet de S. Joly et G. Le Calvé[10].

En effet, en exprimant la quantité de non concordance $10_{ii'} + 01_{ii'}$ en fonction de la quantité de concordance $11_{ii'}$ et de l'indice de Dice (en supposant qu'il soit non nul) on a :

---

[9] Marcotorchino François (1981). «*Agrégation des similarités en classification automatique*», Thèse de Doctorat d'Etat de l'université de Paris VI

[10] Joly S. and Le Calvé G. (1994). «*Similarity Functions*», Lecture Notes in Statistics, (In Van Cutsem, B. ed), Springer-Verlag.

$$10_{ii'} + 01_{ii'} = \frac{2\left(1 - S_{ii'}^{D}\right)11_{ii'}}{S_{ii'}^{D}}$$

en remplaçant la quantité de non concordance par son expression en fonction de $S_{ii'}^{D}$ dans les formules relatives à chaque indice, on obtient :

pour l'indice de Jaccard : $S_{ii'}^{J} = \frac{S_{ii'}^{D}}{2 - S_{ii'}^{D}}$

de même pour l'indice de Sokal-Sneath : $S_{ii'}^{SSA} = \frac{S_{ii'}^{D}}{4 - 3S_{ii'}^{D}}$

pour l'indice de Sorensen on a : $S_{ii'}^{S} = \frac{2S_{ii'}^{D}}{S_{ii'}^{D} + 1}$

et enfin, pour l'indice « symétrique » d'Anderberg on a : $S_{ii'}^{SA} = \frac{4S_{ii'}^{D}}{3S_{ii'}^{D} + 1}$

en utilisant la formulation (F2), on peut montrer que la représentation homographique générique permettant d'exprimer les indices cités ci-dessus en fonction de l'indice de Dice est donnée par la relation :

$$S_{ii'}(n) = \frac{2^{n-1} S_{ii'}^{D}}{\left(2^{n-1} - 1\right)S_{ii'}^{D} + 1} \qquad (F3)$$

4.2. Relation Indice de Dice majorité de Condorcet : cas où les variables sont qualitatives

Dans le cas où les variables mesurées sur la population sont de type qualitatif (ou catégoriel) en nombre M, il est de coutume de transformer le tableau original des données en un tableau K, appelé tableau disjonctif complet[11] représentant les modalités des variables en présence. Ces modalités deviennent à leur tour les nouveaux attributs de présence-absence pour chacune des variables. Dans le cas où il n'y a pas de données manquantes, on a les propriétés, facilement démontrables, suivantes :

Propriétés :

$$1_i = M \qquad \forall i \in I \qquad (P1)$$

$$10_{ii'} + 01_{ii'} = 2\left(M - 11_{ii'}\right) \qquad \forall i, i' \in I \qquad (P2)$$

grâce à ces deux propriétés on peut montrer la relation suivante entre l'indice de Dice et l'indice relationnel $\rho_{ii'}$:

$$\rho_{ii'} = S_{ii'}^{D}$$

en effet, comme pour tout objet i, on a $c_{ii} = 1_i = M$ on en déduit :

[11] C'est ce qui se produit en AFCM (Analyse Factorielle des Correspondances Multiples) ou an AFR (Analyse Factorielle Relationnelle).

$$\rho_{ii'} = \frac{211_{ii'}}{2M} = \frac{11_{ii'}}{M}$$

Maintenant si nous remplaçons n par 1 dans la formulation (F2) on obtient :

$$S_{ii'}^{D} = \frac{11_{ii'}}{2^{-1}(M+M)} = \frac{11_{ii'}}{M}$$

C.Q.F.D.

L'indice de Dice est donc étroitement lié à la règle de la majorité de Condorcet (c.f.3.3.).

5. Apport de la similarité régularisée

Les formulations mathématiques des indices cités ci-dessus présupposent une « équi-pondération » de toutes les variables mesurées sur les objets de la population étudiée. Or ces variables ont des poids intrinsèques induits par leur distribution de présence dans la population. Ce qui confère, de façon sous-jacente, des poids plus importants à certaines de ces variables. En effet, une modalité d'une variable présente dans 99% des objets de la population aura une contribution à la similarité entre objets beaucoup plus importante que celle d'une modalité présente dans 50% de cette population. Or selon la théorie de l'information, cette dernière modalité apporte plus d'information que la première et par conséquent devrait avoir plus de poids dans le calcul de la similarité entre les objets.

De même, lorsque les variables sont catégorielles, plus le nombre de modalités d'une variable est grand, moins elle contribuera à la similarité globale (il est intuitivement plus difficile, pour deux objets donnés, de partager la même modalité d'une variable à vingt modalités, comme la région d'habitation que de partager la même modalité d'une variable à deux modalités, comme le sexe, par exemple).

Les variables ont donc des poids intrinsèques dont il faut tenir compte si l'on ne veut pas que certaines d'entre elles soient, involontairement, privilégiées par rapport à d'autres dans le calcul des similarités[12]. La similarité régularisée consiste à introduire des pondérations dans le calcul des similarités afin de compenser les biais introduits par ces poids.

A titre d'exemple, nous proposons trois type de pondérations. Ces pondérations tiennent compte, pour chaque attribut (modalité) A_j, de l'effectif des objets possédant cet attribut.

l'accord global $c_{ii'}$ entre deux objets i et i' sera donné par la relation :

$$c_{ii'} = \sum_{j=1}^{P} \pi_j k_{ij} k_{i'j}$$

où π_j est le poids attribué à l'attribut A_j et P est le nombre total de modalités de l'ensemble des variables. En général les pondérations naturelles pour chaque variable qualitative ou de présence-absence sont fonction des effectifs du nombre d'objets dans la population possédant un attribut, on peut citer comme poids possibles :

[12] Ceci est particulièrement important lorsqu'il s'agit de faire de la classification automatique. En effet, l'un des points fondamentaux dans ce cas étant la découverte de thèmes rares (ou signaux faibles), sans le ré-équilibrage des poids, on risque de se trouver avec des classes dans les seuls attributs ayant participé à leur constitutions sont les attributs plus courants dans la population. Ce qui est contraire au but recherché.

$$\pi_j = \frac{1}{k_{.j}}, \quad \pi_j = 1 - \frac{k_{.j}}{N}, \quad \pi_j = 1 - \frac{k_{.j}^2}{N^2}$$

Les deux dernières formulations du poids d'une modalité montrent que la similarité de deux objets quelconques par rapport à une modalité présente dans toute la population[13] sera nulle.

Dans le cas particulier, des variables qualitatives, s'ajoute une autre pondération qui tient compte non plus des effectifs, mais uniquement du nombre de modalités de chaque variable. Autrement dit, si une variable possède p_k modalités, afin de renforcer les variables à grand nombre de modalités on peut utiliser le poids $\pi_j = 1 - \frac{1}{p_k}$ pour chaque modalité du tableau disjonctif complet correspondant. Plus le nombre de modalités de la variable est grand plus ce poids sera fort[14]. Une autre pondération tient compte du nombre de modalités des variables ainsi que de la densité des « 1 » présents dans les matrices relationnelles correspondantes, le poids que l'on utilise dans ce cas est $\pi_j = 1 - \frac{\sum_{j=1}^{p_k} k_{.j}^2}{n^2}$, dans ce cas ce sont les variables à modalités équi-distribuées qui seront privilégiés. Pour plus de détails, on peut se reporter aux références citées plus hauts (2,3).

Un point important à signaler est que les matrices relationnelles obtenues par l'utilisation du poids $\pi_j = \frac{1}{k_{.j}}$ s'appellent matrices de Condorcet pondéré. F. Marcotorchino[15] a montré la liaison entre Condorcet pondéré et l'Analyse Factorielle des Correspondances Multiples (AFCM).

[13] Ce qui est logique, car une variable ayant une seule modalité distribuée sur l'ensemble de la population n'apporte aucune information.

[14] Contrairement à ce qui se produirait si on n'utilisait pas cette régularisation.

[15] Marcotorchino François. (1991). « *L'analyse factorielle relationnelle : parties I et II* », Etude du CEMAP, IBM France, vol. MAP-03