Analyse cohésitive et interprétations des données dans le champ de l'éducation

Nadja Acioly-Regnier, Jean-Claude Regnier

Analyse cohésitive et interprétations des données dans le champ de l'éducation

Jean-Claude Régnier*, Nadja Maria Acioly-Régnier**,

* Université de Lyon
86 Rue Pasteur 69007 Lyon France
jean-claude.regnier@univ-lyon2.fr

**EA 3729 Laboratoire « Santé, Individu, Société »
5 Av. Mendes France Université Lyon 2 69500 Bron France
*IUFM de Lyon
5 rue Anselme 69317 Lyon cedex 04
acioly.regnier@wanadoo.fr

Résumé. Cet article s'inscrit dans le prolongement de l'approche développée dans « *Repérage d'obstacles didactiques et socioculturels au travers de l'ASI des données issues d'un questionnaire* » (Acioly-Régnier, Régnier, 2005) et se centre tant sur les apports de l'analyse cohésitive en ASI que sur les difficultés rencontrées par le chercheur non spécialiste pour construire ses interprétations. La cohésion de classe est un indice construit à partir de l'indice d'implication (Gras, 1979 ; Gras & al, 1996) avec des variables binaires ou de propension (Lagrange, 1998) avec des variables modales, pour construire une classification hiérarchique de classes de variables organisées par la relation d'implication statistique, i.e. une hiérarchie orientée. Cette organisation est traduite par la notion de R-règle. La transcription graphique donne l'arbre cohésitif élagué par un seuil d'arrêt fixé sur la cohésion. L'affinement de l'aide à l'interprétation par une mesure de la cohérence a été étudié dans (Gras & al 2004).

1 Introduction

Dans notre précédente contribution (Acioly-Régnier, Régnier, 2005), nous avons abordé une problématique qui relevait du champ des recherches sur les rapports entre *culture* et *cognition*. Nous rappelions que traditionnellement celles-ci sont centrées sur les connaissances développées hors du système scolaire. Les rôles de la *culture écrite* et des stratégies d'enseignement-apprentissage constituaient notre objet en ce sens que ceux-ci peuvent engendrer des obstacles spécifiques au développement de la conceptualisation. Pour tenter d'identifier ces obstacles, leur nature et leur origine, nous avons construit un corpus de données par l'intermédiaire d'une enquête par questionnaire organisé autour d'une situation-problème, auprès d'un échantillon de 198 individus. Cette situation-problème est centrée sur un objet de l'astronomie : la lune et ses phases. Cet objet de connaissance est soumis à une double influence : celle de l'expérience quotidienne et celle de la formation scolaire. Quelle que soit sa position géographique terrestre et qu'il soit scolarisé ou non, un sujet connaît la lune et ses phases. Toutefois cette connaissance peut être située à différents niveaux de conceptualisation, plus ou moins éloignée de la connaissance scientifique apportée par les modèles des sciences physiques et de l'astronomie. N'oublions pas non plus que ce même objet de connaissance est aussi objet d'un autre champ, celui de l'astrologie, qui n'a plus de nos jours le même statut épistémologique, mais qui n'en constitue pas moins une référence culturelle. Nous avons tenté d'expliciter les propriétés de ces données pour mieux comprendre comment l'expérience scolaire pouvait générer des obstacles aux processus de construction de nouveaux concepts. Nous nous intéressons, en particulier, aux régularités associées à des enjeux conceptuels. Nous avons alors eu recours à une approche ASI pour expliciter des relations entre les variables, en particulier, des relations non symétriques fournissant des règles de quasi-implication et, de là, une structure de préordre sur les associations entre réponses fournies par les individus.

Ici nous souhaiterions revenir sur l'analyse instrumentée par la classification hiérarchique orientée développée dans le contexte de l'ASI. Dans un premier temps, nous aborderons succinctement les aspects théoriques sur lesquels se fonde la construction de cette classification ainsi que quelques questions qui nous préoccupent, soulevées par la modélisation. Dans un second temps, nous aborderons les apports instrumentaux espérés par le chercheur non spécialiste mais nous tenterons aussi d'expliciter les difficultés auxquelles il est confronté pour conduire les interprétations.

La question de l'interprétation est nodale dans toute approche statistique dans l'étude d'un phénomène. Nous avons déjà abordé cette question récurrente (Régnier 2002). Pour reprendre l'idée exprimée par Brigitte Escofier

et Jérôme Pagès (Escofier & Pagès 1990 p.217-218), nous caractérisons *l'interprétation statistique* selon trois directions :

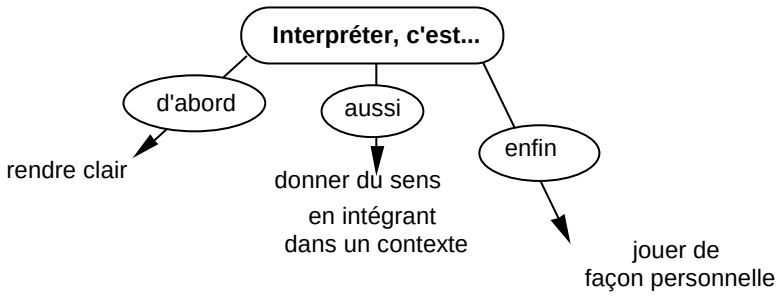


Figure 1: Interpréter, c'est...

Par ailleurs comme nous l'avons déjà abordé à propos du cadre de la didactique de la statistique (Régnier 2006) réexaminons les tâches auxquelles un chercheur est confronté seul ou avec l'aide d'un statisticien, dans une approche statistique dans une étude d'un phénomène.

[TS01]	Problématiser dans un cadre théorique T pertinent pour l'étude du phénomène qui requiert un modèle Mod(T).
[TS02]	Construire un modèle Mod(S) dans le cadre de la statistique qui intègre un modèle Mod(M) du cadre des mathématiques.
[TS03]	Formuler des énoncés hypothétiques dont la mise à l'épreuve par une approche statistique est pertinente.
[TS04]	Construire les données à partir d'un protocole explicite et congruent aux modèles Mod(S) et Mod(T).
[TS05]	Traiter les données dans le cadre du modèle Mod(S) et du modèle sous-jacent Mod(M)
[TS06]	Interpréter dans le cadre du modèle Mod(S) et du modèle sous-jacent Mod(M)
[TS07]	Interpréter dans le cadre du modèle Mod(T)
[TS08]	Décider dans le cadre de modèle Mod(S) et du modèle sous-jacent Mod(M)
[TS09]	Décider dans le cadre du modèle M(T)

Tab. 1 Tâches impliquées dans une approche statistique

Devenu usager de la statistique, le chercheur non statisticien est confronté d'abord à l'interprétation [TS06] [TS08] dans le modèle statistique Mod(S), des produits du traitement des données [TS04] dans ce modèle sous contrainte des propriétés du modèle mathématique Mod(M) pour dégager [TS07] [TS09] des significations dans le cadre théorique rapporté au modèle Mod(T).

Dans le cas auquel nous nous intéressons, nous pouvons identifier le modèle Mod(T) comme celui qui se réfère à la psychologie cognitive développementale et culturelle ainsi qu'à la didactique des sciences dans lequel le phénomène étudié : la lune et ses phases est un objet de connaissance à acquérir et faire acquérir. La construction [TS01] du problème est organisée autour de la compréhension du comment s'affrontent les règles culturelles et la rationalité pour mieux saisir le statut psychologique des procédures et des concepts mis œuvre par des acteurs sociaux dans diverses circonstances de travail ou d'étude

La construction [TS04] des données a été fondée en partie sur les méthodes de l'enquête par questionnaire et de l'enquête par entretien. Le modèle statistique Mod(S) choisi [TS02] s'est inscrit dans une référence à l'approche descriptive classique univariée ou bivariée et l'approche A.S.I. en modélisant les variables-questions par des vecteurs-variables binaires. Le modèle mathématique Mod(M) sous-jacent se réfère à un champ conceptuel complexe organisé, en particulier, par les concepts probabilistes de variables aléatoires usuelles :

hypergéométrique, binomiale, de Poisson et de Laplace-Gauss, celui d'indépendance statistique/stochastique ainsi que ceux de mesure d'écart.

Nous voyons une ébauche des références autour desquelles la compétence du chercheur non statisticien doit se développer pour parvenir à produire des significations pertinentes dans l'interprétation des données construites pour résoudre le problème qu'il se pose dans un cadre théorique autre que celui des mathématiques et de la statistique. La formation nécessaire, sans viser celle du statisticien professionnel, se doit cependant de conduire le chercheur à parvenir à un niveau de conceptualisation qui lui donne un niveau d'autonomie suffisant pour lui permettre de construire des interprétations pertinentes, valides et fiables. Du point de vue de la construction des instruments statistiques par le statisticien, il se pose des questions relatives aux interfaces à mettre en œuvre qui facilitent l'interaction entre ce que produit le logiciel ad hoc, ici C.H.I.C., et ce que cherche à produire le chercheur, à savoir des interprétations pour expliquer et comprendre le phénomène étudié.

2 La classification hiérarchique orientée

Nous allons revenir à la classification hiérarchique orientée pour la situer dans la problématique qui organise ici notre propos. Cette approche s'inscrit dans le projet de produire un modèle statistique et mathématique pour définir et traiter le concept d'implication statistique entre classes de variables (Gras & al, 1996 p. 59) de l'opérationnaliser, de l'instrumentaliser pour que des chercheurs non spécialistes du domaine puissent analyser leurs données et construire des interprétations riches et pertinentes. Notons qu'une des sources de la problématique de l'implication statistique entre classes de variables a surgi dans le champ de la psychologie comme cela est écrit dans (Gras & al, 1996 p. 58) « L'intérêt de G. Vergnaud relatif à la recherche de relation implicative entre groupes de variables afin d'observer des mouvements non symétriques plus amples, plus globaux, plus systémiques qu'ils ne le sont dans le cas de variables isolées nous a conduits à élargir la notion d'implication statistique à des classes de variables. ». Ainsi des études conduites dans des modèles Mod(T) requièrent des instruments produits aux travers de modèles Mod(S) et Mod(M). Cela illustre aussi notre conception épistémologique du développement de la science statistique dans une tension dialectique dont les deux pôles sont la statistique mathématique et la statistique appliquée (Régnier 2002). Il s'est alors agi de construire une méthode de classification sous contrainte de la fonder sur des relations non symétriques (du type relation d'implication) entre variables. Ce fut le concept de *cohésion implicative* qui fut construit pour résoudre ce problème, comme indicateur d'ordre implicatif au sein d'une classe. Cette cohésion « généralement nourrie de cohérence sémantique ou, dans le cas de la didactique, de conditions psychologiques, cognitives, situationnelles, etc., doit se traduire par une mesure (quantitative) » (Gras & al, 1996 p. 59).

Plus récemment ce modèle est abordé dans (Gras, Kuntz, Régnier 2004) où il y est resitué d'une manière synthétique. Les auteurs rappellent que deux points de vue complémentaires sont à considérer pour aborder la problématique de l'implication entre classes de variables : « un point de vue global qui cherche à quantifier la qualité de chacune des partitions associées à chaque niveau de la hiérarchie, et un point de vue local qui se focalise sur la qualité des *R*-règles –assimilables dans une première approche à des classes- construites à chaque niveau. ».

2.1 Le point de vue global

Le point de vue global a été traité dans (Gras et Ratsimba-Rajohn, 1996) en prenant appui étroitement sur une démarche exposée par I.C. Lerman, (Lerman, 1981). Ils rappellent que A étant un ensemble de variables binaires décrivant les attributs d'un ensemble d'individus, « le critère de significativité d'un niveau de la hiérarchie orientée est défini à partir d'une préordonnance Ω induite par un indice sur $A \times A$, appelé indice de cohésion, défini pour valider la qualité implicative des *R*-règles. Il s'agit alors de comparer l'ensemble des couples de couples de $A \times A$ qui respectent la préordonnance initiale Ω avec celui des couples de couples qui respecteraient une préordonnance aléatoire Ω^* dans l'ensemble de toutes les préordonnances de même cardinal que Ω, muni d'une probabilité uniforme. » La notion de R-règles est conçue comme extension aux règles de règles, des règles binaires (a) $\Rightarrow$ (b). Le nombre de variables binaires qui constituent une classe implicative est rendu par un degré, noté d°R, établi selon la définition suivante : une R-règle composée par une variable binaire (une classe élémentaire) est de degré d°R = 0. Une R-règle constituée par une règle binaire (classe contenant deux éléments) a un degré d°R = 1. Si nous considérons les deux R-règles ((a) $\Rightarrow$ (b))$\Rightarrow$ (c) et (a) $\Rightarrow$ ((b)$\Rightarrow$ (c)), elles sont de degré d°R = 2. Par récurrence, une R-règle (R') $\Rightarrow$ (R") admet un degré d°R = d°R' + d°R" + 1.

Ajoutons aussi que le terme « significativité » renvoie à sens plus général que celui en usage dans le domaine de la statistique en exprimant « … ce qui est révélateur d'un phénomène d'intérêt sémantique majeur ». Cette

précision a son importance dans la mesure où les termes de *cohésion* et de *significativité* vont entrer en résonance avec le sens que leur donne le chercheur non spécialiste et vont en partie déterminer la construction des significations. Comme nous le verrons plus bas, le terme *cohésion* désigne à la fois *un état structurel de la classe* et la mesure qui rend compte du niveau de désordre implicatif qui affecte cette *cohésion de classe*.

Définition 2.1. Une *hiérarchie orientée* H_A sur A de cardA=m est un ensemble d'éléments de l'ensemble Ω_A des k-permutations de A (k=1 à m), appelés *classes*, vérifiant les trois conditions suivantes :

1. H_A contient tous les attributs de *A*, appelés classes élémentaires ;
2. Pour chaque couple *C'*, *C''* de classes de H_A, on a $C' \tilde{\cap} C'' = \{ \emptyset, C', C'' \}$, où l' « intersection » $\tilde{\cap}$ entre deux séquences de Ω_A est définie comme étant la plus grande sous-séquence d'attributs contigus communs à *C'* et C'' ; en cas d'égalité on retient la première, selon l'ordre de lecture de gauche à droite sur les attributs d'une *k*-permutation, noté $<_I$, sous-séquence de *C'* ;
3. Pour toute classe non élémentaire *C* de H_A, il existe un unique couple (*C', C''*) tel que $C = C' \tilde{\cup} C''$, où l' « union » $\tilde{\cup}$ de deux séquences disjointes de Ω_A est définie par la concaténation de *C'* et C'' selon l'ordre $<_I$.

Cette approche conduit à produire une classification hiérarchique orientée qui sera traduite graphiquement par un arbre élagué. C'est cette représentation graphique qui est le premier instrument d'aide à l'interprétation du chercheur.

Considérons a et b deux variables binaires de A pour examiner les R-règles telles que d°R = 1 déterminées par les classes constituées de deux variables.

La première étape consiste à établir si une relation non symétrique de quasi-implication entre a et b est admissible à un niveau de confiance 1-α fixé à une valeur de seuil supérieure ou égale à 0.5. La démarche repose sur le croisement présenté dans le tableau ci-après :

		Variable b		
		1	0	
Variable a	1	N(a∧b)	N(a∧¬b)	N(a)
	0	N(¬a∧b)	N(¬a∧¬b)	N(¬a)
		N(b)	N(¬b)	N

Tab. T 1

N(…) est la variable aléatoire CARDINAL dont les réalisations empiriques sur la population-cible ou sur un échantillon de taille N, sont respectivement N(a), etc. Nous supposons que N(a)<N(b) et nous nous intéressons aux individus qui contredisent l'implication mathématique dont l'effectif est N(a∧¬b). Ce choix s'appuie en particulier sur le fait qu'il est plus aisé de considérer ce cas que les trois autres cases pour évaluer l'implication. Sous l'hypothèse Ho d'absence de lien *a priori* entre les deux caractères a et b, l'effectif théorique espéré serait de $\frac{N(a)N(\neg b)}{N}$. Nous savons que $0 \le N(a \wedge \neg b) \le \min \{ N(a); N(\neg b) \}$, mais nous pouvons aussi interpréter :

$0 \le N(a \wedge \neg b) < \frac{N(a)N(\neg b)}{N}$	L'effectif observé étant inférieur à l'effectif théorique signifie que la dépendance entre la présence du caractère (a) et l'absence du caractère (b) est répulsive. En terme d'implication statistique, nous interprétons une observation qui va dans le sens (a)⇒(b)
$0 < \frac{N(a)N(\neg b)}{N} < N(a \wedge \neg b)$	L'effectif observé étant supérieur à l'effectif théorique signifie que la dépendance entre la présence du caractère (a) et l'absence du caractère (b) est attractive. En terme d'implication statistique, nous interprétons une observation qui va dans le sens de la négation de (a)⇒(b)

À la base de la théorie, R. Gras (Gras 1996 p.32) définit alors une mesure d'écart appelée l'*indice d'implication statistique* par :

$$q(a, \neg b) = \frac{N(a \wedge \neg b) - \frac{N(a)N(\neg b)}{N}}{\sqrt{\text{var}[N(...)]}}$$

en tant que réalisation de la statistique Q(a, ¬b) dont la loi de probabilité va dépendre du modèle aléatoire choisi pour la variable N(…). Trois modèles ont été éprouvés pour cette variable discrète : hypergéométrique, binomial et de Poisson. Le modèle hypergéométrique n'est pas retenu car ne permet pas de générer la non-symétrie recherchée pour distinguer les deux couples (a, b) et (b, a). Pour caractériser la relation d'implication statistique entre les deux variables binaires a et b, un second indicateur est introduit, fondé sur une approche stochastique,

Définition 2.2.. On appelle $\varphi(a,\neg b)$ *intensité d'implication statistique du couple (a, b)*, l'indicateur calculé à partir de la probabilité que le nombre de contre-exemples obtenus par le hasard soit inférieur à celui observé empiriquement, à savoir : la *p-value* $\Pr ob\{N(...) \le n_{a\wedge\neg b}\}$, de la manière suivante :

$$\varphi(a,\neg b) = 1 - \Pr ob[Q(a,\neg b) \le q(a,\neg b)]$$

Définition 2.3.. Si $N(a) \le N(b)$, on dit que l'implication statistique a⇒b est *admissible à un niveau de confiance 1-*α, si et seulement si $\varphi(a,\neg b) \ge 1-\alpha \ge 0.5$

Pour revenir à l'aide qui peut être apportée au chercheur non spécialiste, nous notons que dans les résultats fournis par le logiciel CHIC, un flottement demeure dans les termes employés entre *indice* (écart) et *intensité* (probabilité) :

Indices d'implications : (selon la théorie classique)
Calcul avec la loi de poisson

	WA3	WA4	WA5	WA6	WA7	WA8	WA9	WA10	WA11	WA1
WA3	0	86	97	92	90	79	81	73	40	84
WA4	99	0	95	100	100	96	99	98	74	69
WA5	93	77	0	100	91	75	82	90	63	77
WA6	92	91	100	0	99	82	94	95	73	92

FIG 2 : extrait des sorties de logiciels CHIC

Cela est à considérer dans l'étude des origines des difficultés que peut rencontrer l'usager non spécialiste.

La seconde étape consiste à déterminer l'*indice de cohésion* de la classe (a, b) correspondant à une R-règle (admissible au niveau 1-α) de degré d°R=1, c'est à dire a⇒b.

Définition 2.4.. L'*indice de cohésion* est le nombre c(a,b) tel que :

1. si $0{,}5 \prec 1-\alpha \le p = \varphi(a,\neg b) \prec 1$ et l'entropie $E = -\left[p\log_2 p + (1-p)\log_2(1-p)\right]$ alors $c(a,b) = \sqrt{1-E^2}$
2. si p=1 alors c(a,b)=1
3. Par extension, pour une classe (a,b) qui ne correspond pas à une R-règle de d°R=1, c'est à dire si p≤0,5 alors on pose c(a,b)=0

Si nous nous intéressons aux R-règles de d°R=2 qui émergent du sous-ensemble {a, b, c,} de A. Nous dénombrons a priori 6 couples possibles à chacun desquels va correspondre un indice et une intensité d'implication et un indice de cohésion. Supposons que nous ayons déjà constitué la classe (a, b), il ne reste plus alors que les deux couples ((a, b), c) et (c, (a, b)) en lice. Si nous considérons la 3-classe *C*=((a, b), c) correspondant à la R-règle (a⇒b)⇒c, nous définissons l'indice de cohésion par la moyenne géométrique des indices de cohésion des trois classes (a, b), (a, c) et (b, c)

Définition 2.5.. L'*indice de cohésion de la 3-classe C*=((a, b), c) est le nombre c(*C*) tel que :

$$c(C) = \left(c(a,b)c(a,c)c(b,c)\right)^{\frac{1}{3}}$$

La généralisation à des k-classes pour k>3 est réalisée sous cette contrainte précédemment opérationnalisée. Ainsi nous posons la définition suivante :

Définition 2.6.. L'*indice de cohésion* de la k-classe *C* décrite selon l'ordre implicatif par $(a_1, a_2,\ldots,a_k)$ est le nombre c(*C*) qui est la moyenne géométrique des $\frac{k(k-1)}{2}$ indices de cohésion c(a_i, a_j) i=1 à k-1, j>i et j=2 à k,

c'est à dire : $c(C) = \left(\prod_{\substack{i=1 \\ j>i \\ j=2}}^{\substack{i=k-1 \\ j=k}} c(a_i, a_j) \right)^{\frac{2}{k(k-1)}}$

Pour rester dans la perspective épistémologique fixée, l'indice d'implication statistique entre variables binaires a été généralisé pour mesurer l'implication statistique entre classes de variables binaires. Il s'est agi de déterminer un indice qui puisse croître avec les indices de cohésion de chaque classe, s'annuler dès qu'une des cohésions est nulle, croître avec les liaisons extrêmes et décroître avec les cardinaux des classes.

Définition 2.7.. Soit *C'* et *C''*, une k-classe et une h-classe quelconques dont les attributs respectifs sont décrits dans l'ordre implicatif par ($c'_1,\ldots,c'_k$) et ($c''_1,\ldots,c''_h$). On définit l'*indice d'implication généralisé* de la classe *C'* vers la classe *C''* de la façon suivante : $\psi(C',C'') = \left(\underset{\substack{i=1 à k \\ j=1 à h}}{Sup\varphi(c'_i, c''_j)} \right)^{hk} \left(c(C')c(C'')\right)^{\frac{1}{2}}$

Nous considérons les méta-règles R' et R'' associées aux classes implicatives *C'* et *C''* et la R-règle, de degré d°R=k+h+1=r+1, R'⇒R'', sa cohésion est mesurée par l'indice c(*R'*, *R''*) suivant :

$$c(R',R'') = \left(\prod_{\substack{i=1 \\ j=2 \\ i<j}}^{\substack{j=k \\ i=k-1}} c(c'_i, c'_j) \prod_{\substack{i=1 \\ j=2 \\ i<j}}^{\substack{j=h \\ i=h-1}} c(c''_i, c''_j) \prod_{\substack{i=1 \\ j=1}}^{\substack{j=h \\ i=k}} c(c'_i, c''_j) \right)^{\frac{2}{r(r-1)}}$$

$$c(R',R'') = C(R')^{\frac{k(k-1)}{r(r-1)}} C(R'')^{\frac{h(h-1)}{r(r-1)}} \left(\prod_{\substack{i=1 \\ j=1}}^{\substack{j=h \\ i=k}} c(c'_i, c''_j) \right)^{\frac{2}{r(r-1)}}$$

2.2 Le point de vue local

Le point de vue local a été abordé dans (Gras, Kuntz, Régnier, 2004) en portant une attention sur le « préordre défini sur les couples d'attributs « agrégés » à un même niveau de la hiérarchie orientée pour former une *R*-règle, en . comparant le nombre d'inversions entre l'ordre observé dans la classe et celui induit d'un modèle statistique, l'intensité d'implication, au nombre d'inversions attendu avec un ordre aléatoire sur un ensemble de même cardinal. »

Pour rendre compte de cette qualité des classes d'une classification hiérarchique orientée, la notion de *cohérence* a été introduite : *cohérence* entre le rangement issu de la classification et celui déterminé par la relation de quasi-implication sur les variables binaires, ce dernier s'appuyant sur l'ordre des effectifs logiquement congruent à une idée de quasi-emboîtement des groupes d'individus déterminés par chaque attribut. En d'autres termes, « une classe *C* de la hiérarchie orientée H_A formée au niveau *k* est considérée comme *cohérente* pour un seuil α, s'il y a conformité ou quasi-conformité au seuil α entre l'ordre –ou le préordre- ω_0 dans lequel s'organisent les attributs de *C* selon la cohésion et l'ordre –ou le préordre- théorique ω_t défini par leurs intensités d'implication mutuelles. » (Gras, Kuntz, Régnier, 2004, p. 43).

Cette *cohérence* est évaluée par un indice aléatoire qui rend compte de l'écart entre les deux rangements d'attributs en dénombrant les inversions qui apparaissent. Dit autrement « La conformité est mesurée par le nombre d'inversions entre les différents ordres : *i* est le nombre d'inversions observées entre ω_0 et ω_t et *I* est le nombre d'inversions entre ω^* et ω_t. Le nombre d'inversions entre deux ordres est simplement défini ici par le nombre de paires d'attributs (a_i, a_j) telles que a_i est avant a_j dans le premier ordre et après dans le second. » (Gras, Kuntz, Régnier, 2004, p. 43).

Définition 2.8.. Une classe C d'une hiérarchie orientée étant donnée, son indice de *cohérence* est : o(*C*)=Prob{I>i}

Cette approche permet de respecter deux propriétés utiles à l'interprétation par le chercheur non spécialiste :

- plus le nombre d'inversions est faible, eu égard à la cardinalité de la classe, plus grande est la *cohérence* de la classe.
- pour un même nombre d'inversions observées dans deux classes *C'* et *C''*, si la classe *C'* contient plus d'attributs que la classe *C''*, la *cohérence* de *C'* est meilleure que celle de *C''*.

La détermination de la loi de probabilité de la variable aléatoire I, nombre d'inversions dans une permutation, est abordée dans (Gras, Kuntz, Régnier, 2004, p. 44-46).

Cela a conduit à définir un nouvel indice de significativité pour l'interprétation d'une classification hiérarchique orientée prenant en compte la double information apportée respectivement par la *cohésion* et la *cohérence* de chaque classe. C'est ce que vise à rendre compte *l'indice de cohésion-cohérence*.

Définition 2.9.. L'indice *co* de *cohésion-cohérence* qui mesure la significativité de la classe C_k formée au niveau *k* est défini par $co(C_k) = \frac{c(C_k)}{c(C_{k-1})}.o(C_k)$. Par convention, $co(C_0) = 1$. Un niveau *k* de la hiérarchie H_A est *significatif* s'il correspond à un maximum local de *l'indice de cohésion-cohérence* de la classe formée à ce niveau.

Pour aider le chercheur non spécialiste dans sa tâche d'interprétation, nous reprenons les arguments qui ont présidé au choix de la forme de cette expression algébrique. Ce concept d'*indice de cohésion-cohérence* « satisfait à quatre contraintes liées à la sémantique de la significativité :

1. être fonction des mesures de la *cohérence* et de la *cohésion* en majorant les valeurs de la *cohérence* ;
2. conserver l'aspect probabiliste que possède la mesure de la *cohérence* ;
3. pondérer la *cohérence*, en tant qu'indice de "bon ordre" des attributs dans la classe selon la relation de l'implication statistique par un facteur qui pourrait être qualifié d'affaiblissement de la *cohésion* et visant selon les cas : *(i)* à prendre en compte favorablement le fait que la classe formée au niveau *k* ait une *cohésion* peu différente de la classe formée à niveau k-1 précédent, *(ii)* à prendre en compte défavorablement le fait qu'une différence élevée affecte la crédibilité de la classe formée en *k*, même si elle a une bonne *cohésion*.
4. diminuer la significativité d'une classe au niveau *k* qui, bien qu'ayant une bonne *cohérence*, a une *cohésion* qui décroît entre *k-1* et *k*. »

Là encore nous pouvons constater combien le langage familier au contexte de l'A.S.I. introduit pour des raisons d'économie cognitive, un double sens pour *cohérence* tout comme pour *cohésion*, qui renvoie à la fois à une propriété utile au chercheur non spécialiste mais aussi à sa mesure, nommée *indice*.

Une mesure de la qualité de l'ensemble des niveaux constituant une classification hiérarchique orientée établie à un niveau k a été construite de la façon suivante :

Définition 2.10. La *qualité* de l'ensemble des niveaux *i*, $0 \le i \le k$, est définie par $q_k(H_A) = \left(\prod_{i=1}^{k} co(C_i) \right)$ où C_i désigne la classe formée au niveau *i*. La hiérarchie orientée H_A est *significative* au niveau *k* si sa qualité $q_k(H_A)$ admet un minimum local.

Nous n'irons pas plus loin ici dans l'explicitation des concepts dont la maîtrise concourt à la compétence de l'interprétation par le chercheur dans son cadre théorique de référence. Il en est ainsi des concepts de *significativité* des niveaux de la hiérarchie orientée selon *l'indice de similarité* $S(\Omega, h)$, de *contribution* et de *typicalité* des individus ou des groupes d'individus déterminés par des variables supplémentaires.

3 Interprétations de l'arbre cohésitif élagué : attentes, apports et limites pour le chercheur non-spécialiste

Nous aborderons maintenant le second volet de cette communication en partant du point de vue du chercheur non spécialiste qui a choisi de conduire son étude en prenant un appui méthodologique sur les outils de l'ASI, et plus particulièrement sur la classification hiérarchique orientée.

L'exemple que nous prenons, est comme nous l'avons annoncé, celui que nous avons traité dans (Acioly-Régnier, Régnier, 2005) mais avec l'intégration de nouvelles données issues d'une extension de l'échantillon.

3.1 Description statistique de l'échantillon d'observation

Les données construites par une enquête par questionnaire (Annexe 1) sont relatives aux représentations que les individus se font des phases de la lune. Ici, nous nous en tenons aux données construites à partir des 27 variables binaires principales et des 7 variables secondaires (Annexe 2) à partir d'un échantillon de 320 individus

dont 172 habitent l'hémisphère nord et 148, l'hémisphère sud. Les 7 variables secondaires ne sont autres que celles du groupe d'appartenance situé dans les hémisphères ainsi que la variable « sexe »

Dans les trois tableaux ci-après, nous décrivons l'échantillon du point de vue de ces variables secondaires ainsi du point de la variable « âge ».

Hémisphère SUD		Hémisphère NORD (France métropolitaine)		
Nouvelle Calédonie	Brésil Nordeste	IUFM	FAD_Master1	Cadre de Santé
119	29	95	35	42

TAB 2 : Constitution de l'échantillon par zone géographique.

	Homme	Femme
Effectif	97	222

TAB 3 : Sexe

	Total	Effectif	Non Rép.
	320	283	37
Age min.	Age median	Age moyen	Age max.
17	34	33,1	56
		Ecart-type	
		8,7	

TAB 4 : Caractéristique de l'âge des individus de l'échantillon

Figure 3 : Arbre cohésitif

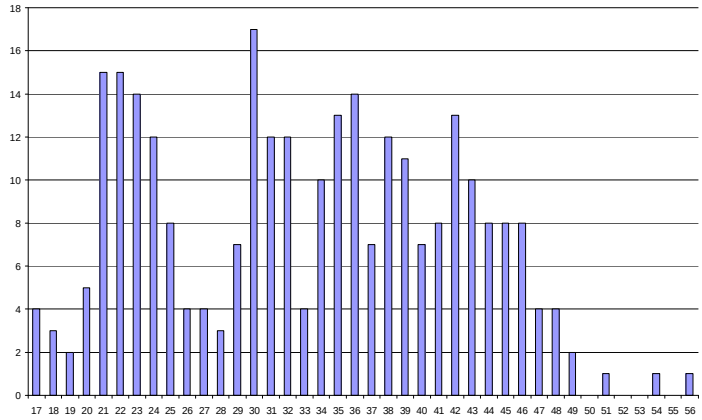
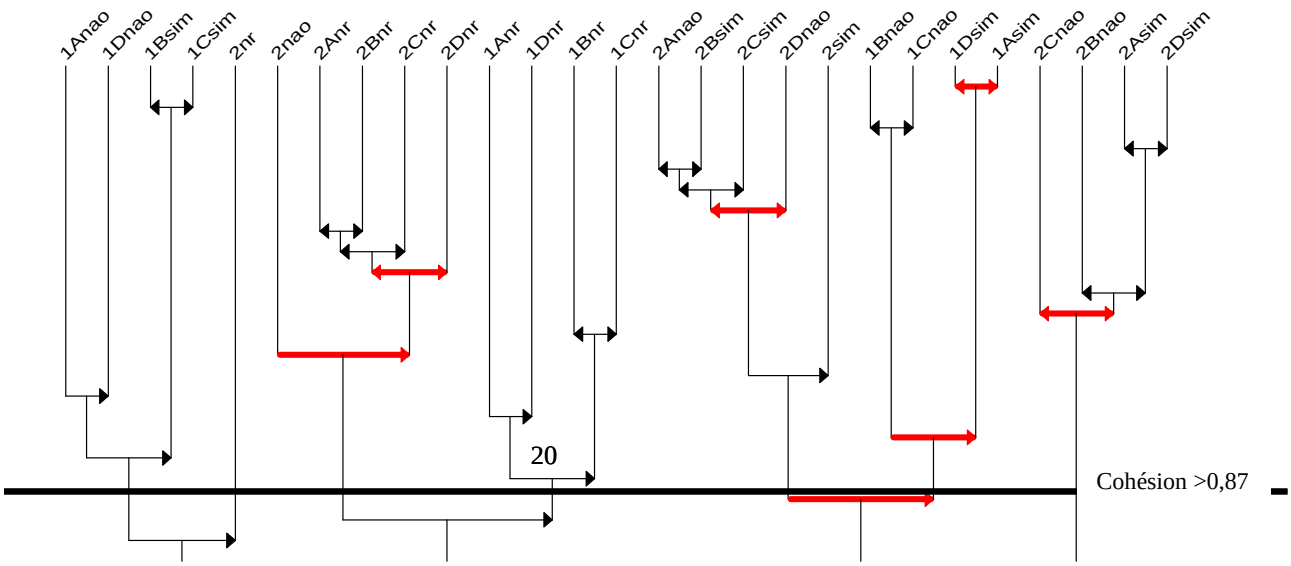


FIG 4 : diagramme de la distribution des fréquences de l'âge

3.2 Classification hiérarchique orientée

La construction de la classification hiérarchique orientée, obtenue avec le logiciel CHIC (Couturier, 2007), conduit à l'arbre cohésitif présenté par la figure ci-dessous. Cet arbre comporte 23 niveaux. Dans un premier temps, nous retenons la partition établie à un niveau de cohésion supérieur à 0,87, ce qui correspond à une coupure au niveau 20. Nous saisissons alors la structure émergente en 7 classes déterminant les 7 R-règles allant du degré d°R=0 à d°R=4



Arbre cohésitif : F:\Pesquisa\ASI4Castellon\ASI4REGNIER_ACIOLYREGNIER\BASE_CHIC\EchanQ1Q2Q3_320CHIC.csv

FIG 5 : Arbre cohésitif

Nous pouvons noter que les indices de cohésion fournis dans les fiches de résultats associées au graphique « arbre cohésitif » sont très élevés comme le montre le tableau suivant :

niveau	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
cohésion	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
niveau	17	18	19	20	21	22	23									
cohésion	0,997	0,989	0,896	0,871	0,640	0,464	0,226									

TAB 5 : Indices de Cohésion

Nous notons ici que les effets de l'arrondi des résultats ne permet pas de séparer les indices de cohésion dont la distinction apparaît graphiquement au travers des 16 niveaux de regroupement.

Nous n'avons alors retenu que les niveaux de cohésion déterminés par un indice de cohésion inférieur à 0,87. Cela a pour corollaire de couper l'arbre entre le niveau 20 et le niveau 21. Les 7 classes qui en résultent avec leurs R-Règles respectivement associées, sont par une lecture de gauche à droite du graphique « arbre cohésitif » :

C1niv19(4 ; 0,896) = (1Anao⇒1Dnao)⇒(1Bsim⇔1Csim)
C2 = (2nr)
C3niv14(5 ; 1) = 2nao⇒(((2Anr⇔2Bnr)⇔2Cnr)⇔2Dnr)
C4niv20(4 ; 0,871) = (1Anr⇒1Dnr)⇔(1Bnr⇒1Cnr)
C5niv15(5 ; 1) = (((2Anao⇔2Bsim)⇔2Csim)⇔2Dnao)⇒2sim
C6niv18(4 ; 0,989) = (1Bnao⇔1Cnao)⇒(1Dsim⇔1Asim)
C7niv12(4 ; 1) = (2Cnao⇔(2Bnao⇔(2Asim⇔2Dsim)))

TAB 6 : Classes et R-Règles retenues (cohésion>0,87)

Nous examinons maintenant les niveaux de cohérence à partir de l'indice de cohérence. Ce dernier est calculé à partir d'une prise en considération des inversions dans la comparaison entre le rangement établi sur la base de l'ordre des effectifs congruent à l'ordre des emboîtements des sous-ensembles, et le rangement déterminé par l'organisation d'une classe à un niveau de cohésion donné.

Classe R-régle	1	2	3	4	5	R-règle d°R	Nombre d'inversions Cohérence	Indice de cohésion
C1niv19	1Anao	1Dnao	1Bsim	1Csim		3	0	0,896
Eff.(rang)	20 (1)	21 (2)	68 (3)	75 (4)			$\frac{23}{24} \approx 0{,}9583$	
C2	2nr					0	0	
Eff.(rang)	5(1)						1	
C3niv14	2nao	2Anr	2Bnr	2Cnr	2Dnr	4	3	1
Eff.(rang)	38 (1)	80 (3)	81 (4)	81 (5)	79 (2)		$\frac{91}{120} \approx 0{,}758333$	
C4niv20	1Anr	1Dnr	1Bnr	1Cnr		3	0	0,871
Eff.(rang)	7(1)	12(2)	37(3)	37(4)			$\frac{23}{24} \approx 0{,}9583$	
C5niv15	2Anao	2Bsim	2Csim	2Dnao	2sim	4	3	1
Eff.(rang)	166 (2)	167 (4)	166 (3)	154 (1)	277 (5)		$\frac{91}{120} \approx 0{,}758333$	
C6niv18	1Bnao	1Cnao	1Dsim	1Asim		3	1	0,989
Eff.(rang)	215 (2)	208 (1)	287 (3)	293 (4)			$\frac{20}{24} \approx 0{,}8333$	
C7niv12	2Cnao	2Bnao	2Asim	2Dsim		3	1	1
Eff.(rang)	73 (2)	72 (1)	74 (3)	87 (4)			$\frac{20}{24} \approx 0{,}8333$	

TAB 7 : Indices de Cohérence

À partir des indices de cohésion et de cohérence, nous calculons l'indice de cohésion-cohérence.

Classes	Indice **co** de cohésion-cohérence
C1niv19	0,8682
C3niv14	0,7583
C4niv20	0,9315
C5niv15	0,7583
C6niv18	0,8241
C7niv12	0,8333

TAB 8 : Indices de Cohésion-Cohérence

3.3 La compréhension d'un outil statistique au travers des interprétations de l'arbre cohésitif par un non-spécialiste.

Il est important de considérer que les outils statistiques que nous développons dans le cadre de l'ASI, visent à être mis au service d'usagers non spécialistes de ce courant d'analyse statistique, ni parfois même de la statistique en général. Ici nous nous intéressons à un usager spécialiste du champ de la psychologie. Pour lui, travailler avec des outils statistiques pour essayer de répondre à des questions de recherche de ce champ n'est pas sans se placer face à des difficultés qui méritent d'être prises en considération. Pour ce faire, il nous a fallu nous approcher au plus près de notre propre compréhension du concept de classification hiérarchique orientée mis en œuvre au tant du point de vue de l'outil statistique avec lequel nous travaillons que de celui de la théorie psychologique de référence. Nous abordons ici de façon plus spécifique les obstacles liés à la question de l'ensemble des systèmes symboliques pour représenter les résultats, question importante dans cet article.
Reprenons la question à partir du point de départ, à savoir celui de la lecture d'une représentation graphique ou d'un tableau. Cette lecture-compréhension conduite par une psychologue centrée sur sa question de recherche et non spécialiste de la statistique n'est pas identique à celle d'un statisticien. Pour plus de précision, nous tentons de décrire de manière chronologique, les difficultés rencontrées pour l'interprétation des données, les obstacles repérés ainsi que les étayages utilisés au travers de cadre théorique de la recherche comme aide au dépassement des obstacles pour parvenir à une lecture des produits de l'outil statistique à un niveau de compréhension

acceptable. En d'autres termes et en référence au cadre théorique de la psychologie mobilisé, à savoir en particulier la théorie des champs conceptuels, la description indique au lecteur, d'une certaine façon, différents niveaux de conceptualisation

Même si les difficultés d'interprétation peuvent être diverses, nous abordons celles liées à la représentation graphique « arbre cohésitif » de la classification hiérarchique orientée.

3.3.1 Les difficultés face à une lecture verticale de l'arbre cohésitif et la question des niveaux de cohésion

En procédant à la lecture selon l'axe vertical, de haut en bas, c'est à dire, pour le spécialiste de l'ASI, selon les valeurs décroissantes des indices de cohésion, nous saisissons une première information pertinente dans le cadre de notre étude, située au niveau 1 de la hiérarchie au travers de la classe (1Dsim⇔1Asim). A première vue et en l'état de notre niveau de conceptualisation d'usager non spécialiste, ceci semble alors contrarier nos précédentes analyses (Acioly-Régnier, Régnier, 2005) conduites à partir de l'arbre de similarités et des représentations des graphes implicatifs. Traduit dans le langage du cadre théorique de la psychologie, nous serions amenés à dire que l'*adhésion* à des figures prototypiques apparaît au niveau le plus élevé de la hiérarchie. Par ailleurs au niveau 2, nous saisissons une seconde information avec la classe (1Bsim⇔1Csim) concernant alors l'*adhésion* à des figures non prototypiques.

En effet dans l'arbre des similarités, la classe {1Dsim ; 1Asim} apparaît au niveau 19 sur 23 tandis que la classe {1Bsim ; 1Csim} apparaît au niveau 3. Cela donne l'impression que la classe qui se constitue autour des figures prototypiques est moins homogène au regard de la relation de similitude que celle constituée autour des figures non-prototypiques. Dans l'arbre cohésitif, l'ordre déterminé par l'indice de cohésion est inversé comme nous l'avons ci-dessus.

Pour tenter d'expliciter la difficulté qui est apparue à ce stade et qui s'est répercutée sur l'interprétation du phénomène étudié, il nous semble que celle-ci s'origine dans la compréhension même de l'outil d'analyse statistique et de la représentation graphique qui vise à jouer un rôle de médiation. Il appert que nous nous trouvons ici par rapport à des représentations qui, pour les êtres humains, sont régies par les mêmes lois de perception que celles que nous étudions à propos des phases de la lune. Il ne faut pas oublier que les images peuvent elles-mêmes induire des obstacles au passage à des niveaux supérieurs de la conceptualisation d'un phénomène. Tout comme ce que nous évoquons dans le cadre de notre étude de représentations des phases de la lune, les représentations graphiques produites par le logiciel CHIC amènent l'usager à se confronter à des interprétations dans le cadre des modèles statistique et mathématique pour accéder à des interprétations dans le cadre du modèle psychologique. Ainsi comment lire cette représentation graphique nommée « arbre cohésitif » dans le jargon du spécialiste de l'ASI ? comment comprendre la nature même du concept de cohésion d'un point de vue statistique d'une manière différente de celle du concept quotidien ? Devons-nous alors privilégier la cohésion des figures A et D prototypiques et celles des figures B et C non prototypiques indépendamment de la nature des réponses négatives ou positives, ou bien encore la force qui unit les réponses oui et non associées à ces figures ?

Dans les deux cas, nous nous sommes trouvés face à une sorte d'*incohérence interne ressentie*. En effet, dire oui à des figures non prototypiques et à des figures prototypiques contrarient autant notre modèle théorique psychologique que les informations précédemment évoquées issue du propre modèle statistique au travers l'arbre de similarités et les graphes implicatifs.

Mais nous pouvons introduire une question encore plus simple, ce type de difficulté n'est-il pas dû au sens même selon lequel est réalisée la lecture de la représentation graphique ? Ici nous avions procédé à une lecture selon l'axe vertical. Nous avons alors tenté de réaliser une lecture selon l'axe horizontal, de gauche à droite.

3.3.2 Les difficultés face à une lecture horizontale de l'arbre cohésitif

La décision prise étant alors de ne retenir que les classes formées à un niveau de cohésion considéré comme fort ici, à savoir supérieur à 0.87, nous sommes confrontés à la question de la signification des 7 classes ainsi formées. Il ne s'agit plus de la similitude ni de la simple relation d'implication reliant deux variables binaires. Pourtant que signifient ces flèches dans une seule direction de gauche vers la droite et même d'autres dans les deux sens ? Que signifie le marquage à la couleur rouge de certaines de ces flèches ? Certes, la définition de ces signes est déjà explicitée dans l'aide qui accompagne le logiciel. Il y est indiqué qu'il s'agit de « niveaux significatifs ». Pour le non spécialiste, s'agit-il de prendre cette qualification « significatif » comme guidant le choix en allant jusqu'à penser que les flèches en noir n'indiqueraient que des liens « non significatifs » !

Comment donc rendre plus claire, pour le non spécialiste, le « sens » de ces outils visant l'aide à l'interprétation de données d'une recherche et à la communication des résultats qui s'en suivent ? Comment traduire cette représentation graphique pour la rendre accessible à un niveau de compréhension acceptable lors d'une présentation à un public de non spécialistes et même non familiarisés avec la terminologie ? Le langage sert à représenter, à traiter mais aussi à communiquer, il nous faut ainsi accéder à un niveau de compréhension rendant plus efficient l'usage de l'outil pour qu'il devienne un amplificateur culturel pour les chercheurs et au-delà pour la diffusion des résultats des recherches scientifiques.

Prenons alors l'exemple d'une classe et par conséquent d'une R-Règle associée de degré d°R=3.

((1Bnao⇔1Cnao)⇒(1Dsim⇔1Asim)).

D'un point de vue de la théorie psychologique, c'est à dire pour reprendre ce que nous avons dit en introduction, dans le modèle Mod(T), nous observons que cette classe est issue de l'agrégation au niveau 18 de deux classes binaires formées aux niveaux 3 et 1. Elle rassemble la classe référant à l'attribut de « rejet par négation des deux figures non prototypiques B et C » et celle référant à l'attribut de « acceptation par affirmation des deux figures prototypiques D et A ». Que signifie donc cette R-Règle, ce « théorème » dans le cadre du modèle théorique de la psychologie dans lequel nous étudions notre phénomène d'un rapport entre la culture et la cognition ? Comment même formuler ce « théorème » ?

Risquons-nous à une formulation dans un registre textuel : *Sachant que les sujets qui auraient une tendance à ce qu'en niant percevoir la forme 1B, nient aussi voir la forme 1C, ont une tendance à ce qu'affirmant percevoir la forme 1D, affirment alors aussi percevoir la forme 1A.*

D'un point de vue du psychologue, il y a une nécessité de se placer à un niveau d'abstraction encore plus élevé pour concevoir cette propriété qui unit par un lien d'implication, deux classes de variables. Dans la formation habituelle, le chercheur en psychologie est plutôt confronté à des situations dans lesquelles il va chercher à établir des liens statistiques entre deux variables. Dans le jargon du domaine psychologique, ces deux variables sont la plupart du temps désignées : variable indépendante et variable dépendante. Ce quasi-théorème dont la validité n'est admise que parce que, en un certain sens, les contre-exemples sont jugés négligeables, oblige le chercheur en psychologie à travailler selon un raisonnement hypothético-déductif complexifié.

Dans notre recherche, ce quasi-théorème traduit : le rejet des figures non-prototypiques entraîne une acceptation des figures prototypiques. Cela traduit un comportement cohérent de réponses au questionnaire et montre la force de l'image dans la lecture du monde.

3.3.3 La compréhension des réponses du point de vue du modèle théorique de la psychologie comme étayage de la compréhension de l'outil statistique

Le chercheur non spécialiste apprend que l'arbre cohésitif nous montre des niveaux de cohésion entre classes de variables binaires formées à partir des réponses. Dans le cas de notre recherche, l'indice de cohésion qui détermine la classe (1Dsim⇔1Asim) qui apparaît au niveau 1 et qui traduit une *adhésion* aux figures prototypiques, est plus fort que celui qui détermine l'apparition de la classe (1Bsim⇔1Csim), au niveau 2 et qui traduit l'*adhésion* aux figures non-prototypiques. Comme nous l'avons déjà dit plus haut, pour le chercheur en psychologie, ceci peut paraître comme une sorte d'incohérence vis-à-vis de son modèle théorique. Et même lui sembler incohérent dans le cadre du modèle statistique.

Ce constat positionne le chercheur dans une situation de conflit cognitif concernant son ressenti vis-à-vis de la cohérence interne de l'outil statistique CHIC et le pousse à chercher à comprendre plus avant le sens de ces résultats du point de vue du modèle statistique.

Cette situation nous a conduit à nous reporter au tableau des fréquences (Annexe 2) qui constitue une autre représentation symbolique, afin de tenter de résoudre ce conflit cognitif. Nous constatons alors que les fréquences pour les composantes binaires de la classe (1Asim ⇔1Dsim) sont respectivement 293 et 287 réponses affirmatives aux figures prototypiques A et D, alors que pour la classe (1Bsim⇔1Csim), l'effectif est de 68 pour la figure B et de 75 pour la figure C , considérées ici comme non-prototypiques. C'est à dire que les proportions d'apparition de 1Asim et 1Dsim sont de l'ordre de 90% tandis que pour 1Bsim et 1Csim, elles sont de l'ordre de 20%. Il y a une différence importante. Ce recours aux fréquences révèle aussi un obstacle qui conduit à chercher à interpréter la cohésion en référence directe aux fréquences, dans le sens d'une sorte de corrélation positive que nous pourrions traduire par « plus il y en a, plus la cohésion sera forte ».

Il nous semble alors avoir compris que nous devons nous décentrer de la notion de fréquence pour nous centrer sur la notion même de cohésion de classe. Jusqu'à l'étape précédente, nous interprétions ces données sous l'influence directe des fréquences. A partir de maintenant nous passons à un autre niveau de

conceptualisation. Alors que les différences de fréquences sont importantes, les indices de cohésion sont presque identiques à un niveau élevé (1 et 2).

Concernant les sujets de l'étude, cela peut conduire à l'interprétation suivante dans le cadre théorique psychologique. Le fait que les deux classes constituées aux deux premiers niveaux de cohésion soient respectivement (1Dsim⇔1Asim) et (1Bsim⇔1Csim) organisées autour des figures prototypiques et non prototypiques, prend un sens dans le domaine de culture et cognition. La première classe peut renvoyer au poids de la culture. La seconde peut renvoyer à des niveaux de conceptualisation plus élevés dans le sens d'une adhésion à des représentations de phases de la lune pourtant non présentes habituellement dans la culture écrite de l'actualité.

Si ce niveau d'interprétation dans le modèle théorique de la psychologie nous semble encore insuffisant pour une meilleure connaissance du phénomène étudié, il nous aide à mieux comprendre le sens de l'outil statistique.

4 Conclusion

Pour conclure, nous ne pouvons nous détacher de la question centrale qui nous préoccupe, associant production d'outils statistiques et utilisation pertinente de ces outils par des non-spécialistes de la statistique dans leur domaine de spécialité : de quels outils complémentaires a-t-on besoin pour fournir une aide à l'interprétation pertinente ?

Nous avons tenté d'expliciter quelques références autour desquelles la compétence du chercheur non statisticien devrait se développer pour parvenir à produire des significations pertinentes dans l'interprétation des données construites pour résoudre le problème qu'il se pose dans un cadre théorique autre que celui des mathématiques et de la statistique. En nous plaçant dans la lignée de travaux déjà conduits (Ratsimba-Rajohn , 1992), nous avons vu que la lecture de l'arbre des cohésions se confrontait à des pré-requis de formation pour conduire le chercheur à un niveau de conceptualisation le plaçant à un niveau d'autonomie suffisant pour produire des interprétations pertinentes, valides et fiables en adéquation avec l'outil statistique. Cela nous renvoie à la question des interfaces facilitatrices de l'interaction entre ce que produit le logiciel ad hoc, ici C.H.I.C., et ce que cherche à produire le chercheur, à savoir des interprétations pour expliquer et comprendre le phénomène étudié : ici, les rapports entre culture et cognition à partir de l'observation des perceptions et des représentations des phases de la lune. Il appert que le choix de signifiants et l'organisation des représentations graphiques doivent être l'objet d'une attention particulière dans le langage de l'ASI. Un des obstacles à franchir par le chercheur non-spécialiste est celui de la décentration de la notion de fréquence pour se centrer sur la notion même de cohésion de classe.

Références

Acioly-Régnier, N M, Régnier, J-C, (2005) Repérage d'obstacles didactiques et socioculturels au travers de l'A.S.I. des données issues d'un questionnaire. R. Gras, F. Spagnolo, J. David (coord). *Proceedings Third International Conference A.S.I. Implicative Statistic Analysis* Palerme 6-8 octobre 2005 ISSN 1592-5137 p.63-87

Couturier, R. (2007) *CHIC: Classification Hiérarchique, Implicative, Cohésitive*. [Logiciel version 4.1 diffusé par ARDM]

Escofier, B., Pagès, J., (1990) *Analyses factorielles simples et multiples : objectifs, méthodes et interprétation*, Paris : Dunod 2ème éd. 267 p.

Gras, R., (1979) *Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques,* Thèse d'État Université Rennes I

Gras, R, & al. (1996) *L'implication statistique, nouvelle méthodes exploratoire des données.* Grenoble, La Pensée Sauvage.

Gras, R., Kuntz, P., Régnier, J.-C., (2004) Significativité des niveaux d'une hiérarchie orientée en analyse statistique implicative. *Revue des Nouvelles Technologies de l'Information* RNTI-C-1 pp. 39-50

Ratsimba-Rajohn H. (1992) *Contribution à l'étude de hiérarchie implicative. Application à l'analyse de la gestion didactique des phénomènes d'ostension et de contradiction*. Université Rennes I, 1992, Thèse d'Université Mathématiques et applications.

Lagrange, J. B., (1998) Analyse implicative d'un ensemble de variables numériques ; application au traitement d'un questionnaire aux réponses modales ordonnées. *Revue de Statistique Appliquée* XLVI-1, Paris, pp. 71-93

Régnier, J.-C. (2002) A propos de la formation en statistique. Approches praxéologiques et épistémologiques de questions du champ de la didactique de la statistique. In *Questions éducatives. L'école et ses marges.* Revue du Centre de Recherche en éducation de l'Université Jean Monnet de Saint-Étienne, n°22-23 décembre 2002 *Didactique des mathématiques.* pp. 157-201

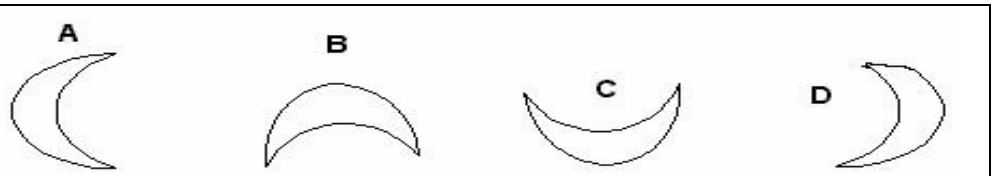
Régnier, J.-C. (2006) *Formation de l'esprit statistique et raisonnement statistique. Que peut-on attendre de la didactique de la statistique ?* in C. Castela et C. Houdement (Dir.) Actes du séminaire national de Didactique des Mathématiques. Année 2005. Editeurs: ARDM & IREM de Paris 7 (pp.13-37)

Annexe 1

Extrait du Questionnaire.

Sexe : ( ) M [MASC] ( ) F [FEMI] **Âge** ( ) **Profession :**
Si enseignant, discipline et niveau enseigné ? _____For**mation antérieure :** ( ) école normale ( ) autres. Laquelle ? _____A**nnées d'expérience professionnelle :_____Lieu du domicile :**

Q1 Avez-vous déjà vu la lune comme ça dans la réalité ?

Représentation de la lune								
Réponses	A-Oui [1Asim]	A-Non [1Anao]	B-Oui [1Bsim]	B-Non [1Bnao]	C-Oui [1Csim]	C-Non [1Cnao]	C-Oui [1Dsim]	C-Non [1Dnao]

Q2 Si un (*individu de l'autre hémisphère*) vous dit qu'il n'avait jamais vu certaines de ces lunes dans la réalité, le croyez-vous ? ( ) OUI [2sim] ( ) NON [2nao] Si oui : Laquelle ou lesquelles ?

( ) A [2Asim]/[2Anao] ( ) B[2Bsim]/[2Bnao] ( ) C[2Csim]/[2Cnao] ( ) D[2Dsim]/[2Dnao]
Pourquoi ? Si non : Pourquoi ?

Annexe 2

	Echan. global		NC		IUFM		BR		CADRE		FADM	
Effectif	320	%	119	%	95	%	29	%	42	%	35	%
Homme	97	30,31%	50	42,02%	20	21,05%	4	13,79%	14	33,33%	9	25,71%
Femme	222	69,38%	68	57,14%	75	78,95%	25	86,21%	28	66,67%	26	74,29%
[1Asim]	293	91,56%	100	84,03%	90	94,74%	26	89,66%	42	100,00%	35	100,00%
[1Anao]	20	6,25%	16	13,45%	4	4,21%	0	0,00%	0	0,00%	0	0,00%
[1Anr]	7	2,19%	3	2,52%	1	1,05%	3	10,34%	0	0,00%	0	0,00%
[1Bsim]	68	21,25%	43	36,13%	10	10,53%	4	13,79%	5	11,90%	6	17,14%
[1Bnao]	215	67,19%	57	47,90%	78	82,11%	20	68,97%	37	88,10%	23	65,71%
[1Bnr]	37	11,56%	19	15,97%	7	7,37%	5	17,24%	0	0,00%	6	17,14%
[1Csim]	75	23,44%	47	39,50%	12	12,63%	3	10,34%	6	14,29%	7	20,00%
[1Cnao]	208	65,00%	54	45,38%	76	80,00%	21	72,41%	35	83,33%	22	62,86%
[1Cnr]	37	11,56%	18	15,13%	7	7,37%	5	17,24%	1	2,38%	6	17,14%
[1Dsim]	287	89,69%	96	80,67%	90	94,74%	25	86,21%	42	100,00%	34	97,14%
[1Dnao]	21	6,56%	17	14,29%	4	4,21%	0	0,00%	0	0,00%	0	0,00%
[1Dnr]	12	3,75%	6	5,04%	1	1,05%	4	13,79%	0	0,00%	1	2,86%
[2sim]	277	86,56%	98	82,35%	85	89,47%	26	89,66%	39	92,86%	29	82,86%
[2nao]	38	11,88%	20	16,81%	6	6,32%	3	10,34%	3	7,14%	6	17,14%
[2nr]	5	1,56%	1	0,84%	4	4,21%	0	0,00%	0	0,00%	0	0,00%
[2Asim]	74	23,13%	25	21,01%	24	25,26%	3	10,34%	10	23,81%	12	34,29%
[2Anao]	166	51,88%	72	60,50%	44	46,32%	21	72,41%	22	52,38%	7	20,00%
[2Anr]	80	25,00%	22	18,49%	27	28,42%	5	17,24%	10	23,81%	16	45,71%
[2Bsim]	167	52,19%	69	57,98%	50	52,63%	18	62,07%	23	54,76%	7	20,00%
[2Bnao]	72	22,50%	27	22,69%	18	18,95%	6	20,69%	9	21,43%	12	34,29%
[2Bnr]	81	25,31%	23	19,33%	27	28,42%	5	17,24%	10	23,81%	16	45,71%
[2Csim]	166	51,88%	72	60,50%	49	51,58%	15	51,72%	23	54,76%	7	20,00%
[2Cnao]	73	22,81%	24	20,17%	19	20,00%	9	31,03%	9	21,43%	12	34,29%
[2Cnr]	81	25,31%	23	19,33%	27	28,42%	5	17,24%	10	23,81%	16	45,71%
[2Dsim]	87	27,19%	33	27,73%	24	25,26%	6	20,69%	11	26,19%	13	37,14%
[2Dnao]	154	48,13%	65	54,62%	44	46,32%	18	62,07%	21	50,00%	6	17,14%
[2Dnr]	79	24,69%	21	17,65%	27	28,42%	5	17,24%	10	23,81%	16	45,71%

Summary

This article falls under the prolongation of the approach developed in "*Identifying didactic and sociocultural obstacles to conceptualization through Statistical Implicative Analysis*" (Acioly-Régnier, Régnier, 2005) and focuses as much on the contributions of cohesitive analysis in ASI as well as on the difficulties encountered by the non-specialist researcher to build his interpretations. The cohesion of class is an index based on the index of implication (Gras, 1979; Gras & al. 1996) with binary variables or those of propensity (Lagrange, 1998) with modal variables, to build a hierarchical classification of classes of variables organized by the relation of statistical implication, i.e. a directed hierarchy. This organization is translated by the concept of the R-rule. The graphic transcription gives the cohesitif tree *pruned* by a threshold fixed on cohesion. The refinement of the aid to interpretation by the measurement of coherence was studied in (Gras & al. 2004).