Revue des Nouvelles Technologies de l'Information
Sous la direction de Djamel A. Zighed et Gilles Venturini

RNTI-E-16

Sous la direction de Régis Gras

Analyse Statistique Implicative

Une méthode d'analyse de données pour la recherche de causalités

Rédacteurs invités :
Régis Gras (École Polytechnique de l'Université de Nantes),
Jean-Claude Régnier (Université de Lyon II),
Fabrice Guillet (École Polytechnique de l'Université de Nantes)

CÉPADUÈS-ÉDITIONS
111, rue Vauquelin
31100 TOULOUSE – France
Tél. : 05 61 40 57 36 – Fax : 05 61 41 79 89
(de l'étranger) + 33 5 61 40 57 36 – Fax : + 33 5 61 41 79 89
www.cepadues.com
courriel : cepadues@cepadues.com

RNTI-Revue des Nouvelles Technologies de l'Information
Sous la direction de Djamel A. Zighed et Gilles Venturini

n°1 : Entreposage fouille de données
E1 : Mesures de qualité pour la fouille de données
E2 : Extraction et gestion des connaissances EGC 2004
C1 : Classification et fouille de données
E3 : Extraction et gestion des connaissances EGC 2005
B1 : 1re Journée Francophone sur les Entrepôts de Données EDA 2005
E4 : Fouille de données complexes
E5 : Extraction des connaissances : Etat et perspectives
E6 : Extraction et gestion des connaissances EGC 2006
E7 : Visualisation en extraction des connaissances
E8 : Systèmes d'Information pour l'Aide à la Décision en Ingénierie Système
B2 : 2e Journée Francophone sur les Entrepôts de Données EDA 2006
E9 : Extraction et gestion des connaissances EGC 2007
E10 : Défi fouille de textes
B3 : 3e Journée Francophone sur les Entrepôts de Données EDA 2007
W1 : Fouille du Web
A1 : Data Mining et Apprentissage Statistique :
applications en assurance, banque et marketing
A2 : Apprentissage artificiel et fouille de données
SM1 : ISoLA 2007 Workshop On Leveraging Applications
of Formal Methods, Verification and Validation
E11 : Extraction et gestion des connaissances EGC 2008
L1 : Langages et Modèles à Objets LMO 2008
L2 : Architectures Logicielles CAL 2008
C2 : Classification : points de vue croisés
B4 : 4e Journée Francophone sur les Entrepôts de Données EDA 2008
E12 : Modélisation des connaissances
E13 : Extraction et gestion de connaissances dans un contexte spatio-temporel
E14 : Relation spatiales : de la modélisation à la mise en œuvre
E15 : Extraction et gestion des connaissances EGC 2009
L3 : Langages et Modèles à Objets LMO 2009
L4 : Conférence sur les Architectures Logicielles CAL 2009
B5 : 5e Journée Francophone sur les Entrepôts de Données EDA 2009

ISBN : 978.2.85428.897.1

Dépôt légal : juin 2009
N° éditeur : 897

PREFACE

Djamel A. Zighed

Professeur des Universités
Directeur du laboratoire ERIC Lyon 2
Président de l'association internationale francophone d'Extraction et de Gestion des Connaissances

Quand les éditeurs m'ont sollicité pour leur rédiger une préface, j'ai été à la fois touché par leur estime et leur confiance mais également intéressé à contribuer, même modestement, à une réflexion à la fois scientifique et épistémologique sur le sujet.

L'Analyse Statistique Implicative (ASI) se définit comme un cadre mathématique et statistique pour étudier la relation entre un « **phénomène** » que l'on pourrait appeler « **antécédent** » et un autre que l'on pourrait appeler « **conséquent** ». Par phénomène, on peut comprendre un fait naturel, physique, social, etc., comme par exemple « il fait soleil », « l'objet est cassé » ou « l'enfant est mécontent ». D'une manière plus abstraite, il s'agit de l'observation de l'état d'une variable qui peut être une grandeur ou une simple modalité particulière. Le terme **Relation** que nous utilisons peut refléter différentes réalités. Il peut signifier la causalité entre les deux phénomènes, l'antécédent est alors la cause et le conséquent, son effet. Par exemple, la crise économique produit du chômage. Dans ce cas, il y a une relation de cause à effet entre la crise et le chômage. Le terme **Relation** peut également traduire la concomitance entre l'apparition des deux phénomènes. Par exemple, l'achat d'un pack d'eau et du pain au supermarché sont deux faits concomitants. On ne peut pas dire que c'est à cause de l'un que l'autre produit a été acheté.

Pour paraphraser la célèbre phrase de Platon « Sans l'intervention d'une cause, rien ne peut être engendré », on pourrait dire, de manière plus large, que les phénomènes, naturels ou artificiels, n'existent qu'à travers les relations qu'ils entretiennent entre eux. Dans cette optique, si l'accès aux phénomènes est relativement simple et même quasi naturel et immédiat pour les animaux dotés de capacités perceptives, l'accès aux relations qui les assemblent est moins évident. C'est d'ailleurs dans la compréhension de ces relations que s'exerce et s'exprime pleinement l'intelligence humaine. L'accès à la connaissance, à la compréhension et de manière plus vaste au sens s'opère à travers l'identification des relations qui relient les phénomènes.

Je ne pense pas trahir la vision des éditeurs de cet ouvrage en disant qu'ils proposent un cadre rigoureux pour étudier les relations, qu'elles soient causales ou non, entre phénomènes naturels ou artificiels. Certes, une hiérarchie qualitative existe sur la nature des relations. En effet, il me semble naturel que les relations causales soient considérées, qualitativement, plus intéressantes que les relations de concomitance qui, même si elles ne sont pas complètement dépourvues d'intérêt, la redondance qu'elles expriment est porteuse de moins d'attention pour l'homme. Pourquoi disons-nous cela ? Une des raisons, ce n'est certainement pas la seule, est l'usage que l'on peut faire de la connaissance des relations. En effet, une relation causale permet, d'une certaine manière, d'acquérir une capacité prédictive sur le conséquent. Cette capacité prédictive confère à celui qui la possède la possibilité de mieux maîtriser

l'enchaînement des phénomènes et, par conséquent, l'opportunité d'anticiper sur des enchainements qui peuvent lui être nocifs ou défavorables ou simplement de mieux servir ses besoins. Comprendre les causalités entre les phénomènes météorologiques par exemple, permet d'éviter certaines catastrophes ou de mieux tirer parti de la nature, par anticipation de ce qui va se produire.

Il est par conséquent naturel que le sens dominant accordé au concept de Relation dans les différentes contributions qui composent cet ouvrage soit celui de la causalité.

Je n'entrerai pas dans le débat sur le déterminisme et la causalité que nous laisserons aux épistémologues. Je me contenterai de dire que la causalité est la réponse au pourquoi ce fait s'est-il produit ?

Rechercher la causalité entre certains phénomènes naturels ou artificiels peut se faire selon deux approches :

- La première consisterait à recréer, en laboratoire par exemple, comme en biologie ou en physique-chimie, les expériences à travers lesquelles on essaierait d'observer si les mêmes causes produisent les mêmes effets. De là, on pourrait dégager une connaissance ou une théorie explicative. Malheureusement, une large partie des phénomènes que nous observons ne se laissent pas emprisonner dans un laboratoire. Par exemple, un phénomène social comme le comportement d'un individu dans une foule ne peut être extrait et reproduit en laboratoire. Pour ces phénomènes, c'est une autre démarche qui s'est imposée. Dans ce cadre, l'information utilisée pour inférer sur la relation est construite par l'expérimentateur.
- La seconde consisterait à identifier et à comprendre des relations entre phénomènes impossibles à isoler en laboratoire. Dans ce contexte, la démarche requise reposerait essentiellement sur la comparaison empirique où il s'agirait de faire émerger des connaissances à partir des données, on parle alors d'extraction de connaissances (ECD). L'information de base qui est la donnée ne peut être produite par l'expérimentateur. Elle émane de la survenue ou non, selon le cycle naturel, du phénomène. Le scientifique n'est plus un expérimentateur mais un modélisateur qui tente d'identifier la Relation entre les phénomènes grâce aux méthodes d'inférence statistique ou d'induction dont la ASI fait partie.

L'Analyse Statistique Implicative tente, à mon sens, de fournir un cadre méthodologique pour étudier ces Relations. On appelle également Relation un Modèle, dans le sens où l'identification de la relation peut conduire à une prédiction.

L'ASI se propose alors de répondre à la question « en quoi le phénomène **a** est-il la cause du phénomène **b**». La réponse à cette simple question requiert, d'une part, un cadre théorique et, d'autre part, une ingénierie de calcul adaptée à la nature des données qui matérialisent les observations.

Le cadre théorique général est celui de la logique car la question précédente peut se réduire à : si **a** est vrai, est-ce que **b** l'est aussi ?

Pour répondre à la question, le modélisateur observe des réalisations de **a** et de **b** et évalue ensuite la proposition : **a** (vrai) implique **b**(vrai). Si la proposition est vraie, alors on parlera, dans ce cas précis, d'un modèle implicatif sous-jacent entre **a** et **b**. Dans ce contexte, on peut bénéficier du cadre formel de la logique classique binaire.

Mais de façon plus générale, cette logique binaire connait vite des limites. En effet, la mise en évidence d'une relation, et plus spécifiquement d'un modèle, exige de disposer d'un cadre théorique et opérationnel capable de :

- Prendre en compte toutes les formes que peut revêtir un phénomène qui peut être une grandeur quantitative comme la température du corps d'un patient ou un état qualitatif booléen ou multimodal comme l'apparition ou non d'une fièvre chez un patient. Bien également prendre garde au fait que l'observation du phénomène peut être entachée d'erreur de mesure, d'imprécision, d'incertitude etc..
- Évaluer si la proposition (relation ou modèle) est vrai ou fausse ou les deux à la fois ce qui peut être paradoxal de prime abord mais qui n'en est pas. Par exemple, une proposition logique peut être vraie selon un degré qui peut s'apparenter à une probabilité d'erreur quant à l'affirmation de la véracité, à une incertitude, à une possibilité, etc..

L'ASI offre à la fois le cadre théorique et les outils pour répondre à toutes ces exigences. Pour convaincre le lecteur de la justesse de leur thèse, les éditeurs ont réuni les différentes contributions selon trois niveaux :

- Le premier rappelle les fondements de l'ASI, sa genèse et ses fondements épistémologiques, mathématiques et statistiques. Le lecteur est ainsi préparé au débat.
- Le second réunit des travaux au cœur du processus concret d'induction. Il s'agit de la mise en œuvre effective de l'ASI dans le contexte de la fouille de données. Il convient de préciser que ce domaine scientifique émergent va certainement constituer le cadre de travail de toute la recherche scientifique basée sur la systémique, c'est-à-dire la connaissance des objets et de leurs interactions.
- Le troisième propose enfin des applications réelles notamment dans les domaines réputés difficiles ou les phénomènes ne sont pas reproductibles en laboratoire. Cela concerne la didactique, la sociologie, la psychologie etc. L'ASI a su montrer qu'elle offre moyen original et contrôlable pour percer le sens caché par les lourdes couches de carapaces que sont les données.

Il est clair que les travaux réunis dans cet ouvrage ne sont que la partie visible d'un immense chantier entamé par Régis Gras et les différents contributeurs depuis de nombreuses années. Ces travaux font maintenant autorité dans le domaine. Outre son caractère pédagogique et pragmatique, l'un des grands mérites de cet ouvrage est d'avoir regroupé, dans un même corpus, les résultats de plusieurs centaines de publications depuis plus de 20 ans. Ce fait est assez rare dans les publications scientifiques françaises pour qu'il mérite d'être salué. Je suis convaincu que cet ouvrage ne manquera pas d'intéresser aussi bien les praticiens que les théoriciens notamment dans le domaine de la fouille de données et de l'informatique décisionnelle.

Je félicite les auteurs et les éditeurs pour l'excellent travail qu'ils ont réalisé et je leur souhaite un grand succès.

Table des Matières

INTRODUCTION

1 Guide de lecture de l'ouvrage (Fabrice Guillet)

L'Analyse Statistique Implicative (ASI) puise son origine en didactique des mathématiques dans un questionnement sur la modélisation des relations implicatives entre deux observations a et b, du type « si j'observe a alors j'ai aussi tendance à observer b, et ceci, de manière statistiquement significative ».

Cet ouvrage a pour objectif d'en rappeler la genèse, de dresser un panorama récent des concepts, des modèles, des méthodes et des applications de l'ASI. Afin de faciliter l'accès à son contenu, les 31 chapitres qui le composent, sont regroupés en 3 parties principales dont la dynamique est la suivante :

- La Partie 1 constitue une sorte de cours, à la manière d'un manuel, qui balaie l'ensemble des concepts qui forment le cœur de l'Analyse Statistique Implicative ;
- La Partie 2 aborde des compléments et des extensions de ces concepts qui en montrent le caractère stimulant et fécond ;
- La Partie 3 présente une sélection d'applications qui illustrent les apports de la démarche développée par l'Analyse Statistique Implicative dans des domaines variés, à la recherche de pistes causales.

En complément, nous proposons dans cette section d'offrir une cartographie résumée des chapitres. Nous espérons que cette « carte de lecture » aidera le lecteur à s'orienter et à effectuer une lecture sélective adaptée à ses préférences.

PARTIE 1 : FONDEMENTS THÉORIQUES DE L'A.S.I.

L'ensemble des textes de cette Partie a été rassemblé, coordonné, mis en forme ou reformulé par Régis Gras et Jean-Claude Régnier.

Les chapitres de cette première partie s'enchaînent comme des réponses successives aux questions naturelles nées de la problématique de l'implication statistique ou quasi-implication rappelée ci-dessus.

- **Chapitre 1 : Analyse implicative des variables binaires. Intensité implicative. Intensité entropique**. Ce premier chapitre pose les bases épistémologiques et les fondements mathématiques de l'A.S.I.. On y trouvera l'ensemble des définitions des modèles statistiques utilisés : les quasi-règles, l'indice et l'intensité d'implication, leurs propriétés et leurs extensions.
- **Chapitre 2 : Représentation des règles d'implication. Graphe implicatif**. Ici les auteurs développent la notion de graphe implicatif engendré par le réseau de quasi-implications. Ils définissent l'algorithme de construction, et illustrent la démarche par un jeu de données didactiques.
- **Chapitre 3 : Extension de l'A.S.I. aux variables non binaires**, où sont proposées deux principales extensions de l'A.S.I. à des situations non binaires. Une pour les variables

modales, variables fréquentielles et variables numériques à partir d'un indice de propension. Et une seconde pour les variables sur intervalles et les variables-intervalles

- **Chapitre 4 : Extension de l'Analyse Statistique Implicative à des hiérarchies de règles**. Complétant le chapitre 2 sur la nécessité de représentation, ce chapitre propose une extension de l'A.S.I. à des hiérarchies orientées de règles sur la base d'un indice de cohésion. Y sont définies les notions de hiérarchie de règles ou R-règles, de classe de variables, d'indice cohésitif et sa justification théorique : une distance ultramétrique. L'algorithme de construction des hiérarchies y est détaillé, puis illustré sur un exemple pédagogique.
- **Chapitre 5 : Dualité entre variables actives et variables supplémentaires : typicalité et contribution.** Ce chapitre offre des outils complémentaires pour analyser la structure des hiérarchies et graphes de règles en introduisant les notions de variables supplémentaires et deux nouvelles mesures : la typicalité et la contribution aux classes cohésitives et aux chemins implicatifs. Une topologie duale est ébauchée entre les deux espaces sujets-variables. Cette approche est illustrée de manière détaillée sur un exemple, puis appliquée à l'analyse d'un questionnaire présenté à des professeurs de terminale.
- **Chapitre 6 : Règle et R-règle d'exception en A.S.I. ou encore l'exception confirme-t-elle la règle ?** Les auteurs présentent une extension de l'A.S.I. aux règles d'exception, véritables paradoxes par rapport au bon sens « implicatif ». Ils posent le problème, puis proposent deux approches de résolution possibles pour détecter les règles d'exception sur la base de l'intensité d'implication.
- **Chapitre 7 : Extraction de Règles en Incertain par l'A.S.I..** Les auteurs s'attachent à étendre l'A.S.I. aux variables incertaines, dites « floues ». Pour établir un pont entre logique floue et logique A.S.I., ils posent les liens entre variables-intervalles et variables floues et illustrent l'approche sur un exemple.
- **Chapitre 8 : Réduction du nombre de variables**. Ce chapitre propose une solution pour la réduction de variables en définissant un indice entropique d'équivalence fondé sur l'intensité d'implication entropique. En l'illustrant sur un exemple, il présente un algorithme de construction des classes d'équivalence basé sur l'optimisation d'un critère global d'inertie implicative.
- **Chapitre 9 : Règles superflues ou redondantes en A.S.I.**. Enfin, et en complément du chapitre précédent, les auteurs présentent, sur la base de critères convenablement choisis, un ensemble de mesures adaptées à la détection des redondances ou des règles superflues : la superfluité des règles, et le gain d'information entre 2 règles.

PARTIE 2 : COMPLEMENTS ET EXTENSIONS

- **Chapitre 1 : Statistique de rangs et Analyse Statistique Implicative,** par J.-C. Régnier et R. Gras. Ce chapitre traite de l'analyse des rangs accordés par des juges à des objets, au sens de Friedman ou de Kendall. Pour cela, ils affectent une mesure de qualité des énoncés de la forme : «l'objet b est généralement rangé à un rang meilleur que a», et représentent les relations de préférences dans un graphe.
- **Chapitre 2 : Qualité d'un graphe implicatif : variance implicative,** par R. Gras et J.-C. Régnier. Ce chapitre définit et utilise la notion de variance implicative comme instrument d'évaluation de la qualité d'un graphe implicatif. L'approche est illustrée sur deux exemples.

- **Chapitre 3 : Mesurer l'écart entre une analyse a priori et la contingence en didactique**, par F. Spagnolo, R. Gras et J.-C. Régnier. Ce chapitre propose des instruments statistiques permettant d'évaluer l'écart entre des hypothèses formées a priori et les observations relevées a posteriori, la contingence. L'approche est appliquée à des données en didactique des mathématiques
- **Chapitre 4 : Problème de données manquantes dans un tableau numérique. Une application de l'A.S.I.,** par R. Gras. Ce chapitre s'intéresse au traitement des données manquantes ; et plus particulièrement à l'attribution de la valeur la plus vraisemblable à une variable non renseignée au moyen d'une distance pondérant les liens certains. Un exemple numérique illustre l'usage de la méthode proposée sur un tableau incomplet.
- **Chapitre 5 : Analyse Implicative Séquentielle**, par J. Blanchard, F. Guillet et R. Gras. Dans ce chapitre, les auteurs proposent de définir une mesure inédite nommée SII pour évaluer la qualité des règles séquentielles. SII évalue la significativité des règles au regard d'un modèle probabiliste. Les simulations montrent que SII a des caractéristiques uniques en comparaison aux autres mesures de qualité de règles séquentielles.
- **Chapitre 6 : Analyse statistique implicative entre variables vectorielles**, par R. Gras et R Couturier. Ce chapitre étend les instruments de l'A.S.I. à des variables complexes : des variables vectorielles. Il définit un indice permettant de mesurer la qualité d'une règle entre variables vectorielles. L'approche est illustrée sur deux exemples.
- **Chapitre 7 : Arbre de décision pour données déséquilibrées : sur la complémentarité de l'intensité d'implication et de l'entropie décentrée**, par G. Ritschard, S. Marcellin et D. A. Zighed. Ce chapitre porte sur l'induction d'arbres de classification pour des données déséquilibrées. Il adapte d'une part les critères de construction de l'arbre à la nature déséquilibrée des données, et d'autre part la pertinence de la conclusion à associer aux feuilles de l'arbre. Deux approches distinctes sont comparées. Une première expérimentation sommaire est présentée.
- **Chapitre 8 : Graphe de règles d'implication statistique pour le raisonnement courant. Comparaison avec les réseaux bayésiens et les treillis de Galois**, par M. Cadot. On montrera dans ce chapitre comment fonctionne l'enchaînement des règles, notamment à travers la construction du graphe implicatif et on comparera ce modèle statistique à deux autres modèles proches : un modèle algébrique, les treillis de Galois, et un modèle probabiliste, les réseaux bayésiens. Le fonctionnement des trois modèles sera illustré à l'aide d'un même jeu de données médicales librement disponible sur Internet.
- **Chapitre 9 : Une étude comparative pour la détection de dépendances multiples** par E. Salehi, J. Nyayachavadi et R. Gras. De nombreuses méthodes d'optimisation ont été proposées pour la recherche de dépendances entre variables. Ce chapitre propose d'évaluer une nouvelle méthode issue de l'A.S.I. pour résoudre ce problème dans le cadre d'une heuristique maxmin. Cette nouvelle méthode est comparée aux méthodes classiques, puis testée sur grand nombre d'expériences.
- **Chapitre 10 : Test de Mac Nemar et Analyse Statistique Implicative**, par J.-C. Régnier. Ce chapitre compare trois approches pour étudier des liens vraisemblables entre deux variables binaires, que sont : l'A.S.I., le test de Mac Nemar et le test d'indépendance fondé sur la mesure du $\chi 2$. Cette comparaison est illustrée sur un jeu de données issu de la didactique des mathématiques.
- **Chapitre 11 : Historique et fonctionnalités de CHIC**, par R. Couturier et S. Ag Almouloud. Le logiciel CHIC, acronyme de Classification Hiérarchique, Implicative, Coesitive, permet d'utiliser la plupart des méthodes de l'A.S.I. : de la découverte des impli-

cations entre variables, aux hiérarchies cohésitives et graphes d'implication, en passant par les hiérarchies de similarités entre variables. Outre un rappel de l'historique de ce logiciel CHIC, ce chapitre décrit en particulier les caractéristiques et les condition d'usage de CHIC.

- **Chapitre 12 : Guide d'utilisation des principales fonctionnalités du logiciel CHIC**, par H. Ratsimba-Rajohn. Complétant et illustrant le chapitre précédent, celui-ci présente un cas d'utilisation complet d'un exemple avec le logiciel CHIC. Sous une forme « tutorielle », il permet de guider pas à pas l'utilisateur dans l'ensemble des nombreuses fonctionnalités de CHIC.

PARTIE 3 : APPLICATIONS de l'A.S.I.

Thème 1 : Applications à la didactique des mathématiques

- **Chapitre 1 : Analyse statistique implicative et didactique des mathématiques**, par D. Lahanier-Reuter. Ce chapitre montre l'intérêt d'interpréter les liens implicatifs découvert par l'A.S.I. comme des règles de régulation d'actions. Nous étayons ceci par des exemples précis en didactique des mathématiques, portant sur la reconstruction et la compréhension d'actions tant matérielles que discursives, actions régulées des élèves, actions régulées et régulatrices des enseignants.
- **Chapitre 2 : Approche bayésienne "cachée" et approche fréquentiste "ambiguë" dans les manuels français de Première S et ES**, par P. Carranza et A. Kuzniak. Ce chapitre propose une analyse de manuels présentant un premier enseignement des probabilités en relation avec la statistique en classe de Première en France. L'approche statistique implicative permet de dégager certaines règles de fonctionnement et de décrire une tendance soit très fortement formelle et calculatoire soit assez confuse où joue alors ce que nous appelons l'intrication des signifiés.

Thème 2 : Applications à la psychologie

- **Chapitre 3 : Interprétation de graphes implicatifs: étude clinique auprès d'une chercheure en iconographie médiévale**, par N. M. Acioly-Régnier et J.-C. Régnier. Le logiciel CHIC offre une interface accessible au non-spécialiste, mais ce dernier est confronté à l'interprétation des représentations graphiques produites par ce logiciel. Ce chapitre met ainsi l'accent sur les interprétations erronées des graphes d'implication qui ont été effectuées dans le cadre d'une recherche en iconographie médiévale.
- **Chapitre 4 : Étude de représentations d'élèves en éducation physique et sportive**, par I. Verscheure et C.-M. Chiocca. Ce chapitre concerne le traitement par le logiciel CHIC d'un questionnaire proposé à des élèves de première de l'enseignement agricole français. Les questions abordent les représentations des élèves vis-à-vis des activités physiques et sportives et plus particulièrement du volley-ball. Les résultats obtenus permettent d'améliorer la sélection des élèves « représentants des réseaux »pour des futurs travaux basés sur des études de cas.
- **Chapitre 5 : Un exemple d'analyse implicative en neuro-psychologie : la comparaison de groupes contrastés,** par T. Bellaj et D. Pasquier. Alors que la psychométrie classique utilise des indices symétriques de liaisons entre variables, l'A.S.I. permet d'effectuer des analyses asymétriques plus précises sur les séquences implicatives entre variables. Ce chapitre, présente un exemple de ce type d'analyse pour comparer des groupes contrastés.

- **Chapitre 6 : Utilisation de la statistique implicative pour élaborer des référentiels de compétences comportementales**, par D. Follut, L. Diot, M. Lasserre Azema et S. Baquedano. Cinq Centres de Bilan de Compétences de la Région Midi-Pyrénées se sont groupés pour créer une méthode d'appui individualisé à la Validation des Acquis par l'Expérience. La Société PerformanSe a conçu pour cet usage un outil d'évaluation permettant aux conseillers des CIBC d'être informés sur les comportements - favorables ou défavorables - susceptibles d'être adoptés par chacune des personnes qu'ils accompagnent.

Thème 3 : Applications à la sociologie

- **Chapitre 7 : Analyse statistique implicative des transitions professionnelles dans la Genève du 19e siècle**, par G. Ritschard, M. Studer et M. Oris. Ce chapitre reprend une étude de la dynamique socioprofessionnelle du 19e à Genève et la compare à une analyse supervisée des dissimilarités entre transitions. L'étude se concentre sur le changement de groupe socioprofessionnel (GSP) entre deux recensements. L'A.S.I. donne une vision synthétique des liens entre ces dynamiques et les GSP. L'analyse des dissimilarités permet quant à elle de segmenter la population en groupes homogènes en fonction des caractéristiques démographiques et culturelles.
- **Chapitre 8 : Derrière les réseaux de variables, il y a des individus… à écouter !** L'exemple d'une étude chez des enseignants de lycée professionnel, par M. Bailleul et S. Godard. Afin de comprendre comment des enseignants en lycée professionnel se représentent les tensions résultant du croisement de plusieurs logiques (scolaire, économique, politique et administrative), un large questionnaire comportant quelques réponses ouvertes a été proposé aux enseignants des lycées professionnels de l'académie de Caen. Ce chapitre présente l'analyse, menée avec le logiciel CHIC, des 257 questionnaires qui ont été complétés.

Thème 4 : Application à la bio-informatique

- **Chapitre 9 : Une méthode implicative pour l'analyse de données d'expression de gènes**, par G. Ramstein. Ce chapitre présente une méthode d'extraction d'associations basée sur l'A.S.I. et la notion de rang. L'intensité d'implication a été adaptée à des classements pour découvrir des relations partielles robustes vis à vis du bruit et des variations d'amplitude. Appliquée aux données de puces à ADN, cette méthode met en évidence des relations entre des formes d'expressions particulières de gènes. Elle apporte également une connaissance plus fine des relations entre les gènes que les méthodes usuelles de corrélation.

Thème 5 : Application à l'histoire de l'art

- **Chapitre 10 : Iconographie médiévale en histoire de l'art et Analyse Statistique Implicative,** par M. Guénot et J.-C. Régnier. Dans ce contexte singulier de l'histoire de l'art, et en particulier en iconographie médiévale, l'A.S.I. apporte un regard nouveau sur la lecture du sujet : elle permet de connecter des éléments sélectionnés au préalable par le chercheur, faisant apparaître d'autres liens internes à l'image et offre une nouvelle manière d'appréhender l'étude thématique en histoire de l'art.

2 Origine et développement de l'Analyse Statistique Implicative (Régis Gras et Jean-Claude Régnier)

A ce jour, l'analyse statistique implicative désigne un champ théorique centré sur le concept d'implication statistique ou plus précisément sur le concept de quasi-implication pour le distinguer de celui d'implication logique des domaines de la logique et des mathématiques. L'étude de ce concept de quasi-implication en tant qu'objet mathématique, dans les champs des probabilités et de la statistique, a permis de construire des outils théoriques qui instrumentent une méthode d'analyse de données. Force est de constater que les racines épistémologiques de ce concept se sont nourries de questions qui ont surgi principalement d'un autre champ : celui de la didactique des mathématiques. Historiquement, une des questions abordées concernait la mise en évidence des niveaux de complexité des exercices de mathématiques posés à de jeunes élèves s'énonçant ainsi :

R « *si un exercice est plus complexe qu'un autre, alors tout élève qui résout le premier devrait réussir à résoudre le second* ».

Plus précisément, Régis Gras, (Gras, 1979) avait conçu a priori en 1976 une taxonomie d'objectifs cognitifs, c'est-à-dire un préordre partiel entre des compétences attendues de l'élève au cours de l'apprentissage et du fonctionnement opératoire des concepts mathématiques. Par exemple, « Choix et ordonnancement d'arguments » y précèderait « Critique d'argumentation et construction de contre-exemples » et y suivrait « Effectuation d'algorithmes simples ». Il attendait de tests divers, constitués de variantes d'exercices présentés à des élèves de collège (13 à 15 ans), la validation de cette taxonomie a priori. Sous forme de graphe orienté sans cycle, l'organisation des performances observées devrait permettre d'étudier l'adéquation de la taxonomie au préordre restitué par le graphe et, accessoirement d'étudier des distorsions liées à deux méthodes d'enseignement différentes. Tout enseignant, comme tout chercheur en didactique des mathématiques, sait par la pratique pédagogique ou par l'observation que des contre-exemples surgissent dans les situations observées par rapport aux hypothèses émises sur la performance. Un outil statistique s'avérait alors nécessaire pour évaluer et représenter les quasi-règles dégagées de la contingence sur la base des résultats obtenus[1].

Revenant à l'énoncé **R** ci-dessus, il exprime une règle qui n'est que rarement strictement observée. Il ne peut donc pas avoir le statut de théorème au sens défini dans le champ des mathématiques. Cependant, celui-ci s'inscrit pleinement dans le cadre paradigmatique de la relation d'implication statistique qui fait l'objet central de cet ouvrage et où les règles s'expriment sous la forme : « Si on observe *a*, alors on observe généralement *b* ».

Tout au long des vingt dernières années, le développement théorique de l'analyse statistique implicative a été principalement stimulé par une dialectique entre pratique et théorie, dans une tension entre deux cadres : statistique appliquée à… et statistique mathématique. Dans divers champs scientifiques tels que la didactique des mathématiques, la psychologie, la sociologie, la bio-informatique, etc., des données construites ont été soumises à cette méthode d'analyse. Cette mise en œuvre a montré l'efficience de la méthode dans sa capacité à

[1] A notre connaissance aucune autre taxonomie, par exemple la plus célèbre celle de Bloom (Taxonomy of Educational Objectives, 1956), dont s'est inspiré R.Gras, n'a été éprouvée **et** validée au moyen de méthodes statistiques comparables.

faire émerger des propriétés que d'autres éclairages ne permettaient pas, mais elle a aussi permis de montrer ses limites qui ont alors suscité de nouvelles problématiques autour du concept-objet de la quasi-implication. Le raisonnement qui fonde l'interprétation des résultats de l'analyse statistique implicative est essentiellement de nature statistique et probabiliste. Ce mode de raisonnement s'inscrit dans une perspective dégagée par le développement de la pensée statistique, de l'esprit statistique.

Une part importante du développement de l'analyse statistique implicative revient donc aux travaux conduits, dirigés ou impulsés par Régis Gras depuis les années 70. Mais également aux rencontres internationales portant sur l'Analyse Implicative [2](IUFM de Caen en 2000, Université PUC à Sao Paulo en 2003, Université de Palerme en 2005, Université de Castellon en 2007) où discussions et débats ont permis de développer la théorie et d'ouvrir la variété des applications.

L'amplification du développement de l'A.S.I. a aussi trouvé son étayage dans l'assistance informatique qu'apporte un logiciel dédié désigné par l'acronyme CHIC (Classification Hierarchique, Implicative, Cohésitive) dont Régis Gras initia la programmation, reprise ensuite dans les thèses de Saddo Ag Almouloud (1992) et de Harrison Ratsimba-Rajohn (1992) et dont le développement actuel est assuré par Raphaël Couturier (Couturier et Gras 2005)

Le présent ouvrage aborde, d'une part, les fondements théoriques, les avancées conceptuelles et les extensions de l'A.S.I. au niveau actuel de leur développement avec un chapitre consacré au logiciel CHIC de traitement automatique, d'autre part, rapporte un ensemble d'études choisies conduites dans divers domaines qui exposent les résultats obtenus par la méthode d'analyse de données A.S.I..

2.1 Un entretien entre Régis Gras et Jean-Claude Régnier

Q1 : [JCR] Quand on consulte des ressources, on découvre Régis Gras dans le domaine de la didactique des mathématiques et dans celui d'une branche de la statistique : Analyse Statistique Implicative. Une étape de ton parcours fut ta thèse de doctorat d'état soutenue en 1979. Mais quel fut le chemin qui t'a conduit vers ce centre d'intérêt qu'explicite ta thèse ?

[RG] *J'ai toujours eu le souci de comprendre quelles difficultés se présentaient au cours de l'apprentissage des mathématiques pour les élèves petits ou grands, quelles en étaient l'origine (psychologique, didactique, cognitive, sociale,…), l'expression (erreurs, échecs, lapsus, blocage, rejet, répulsion,), quelles re-médiations (répétition, changement de cadre, de registre, tutorat, exemples, contre-exemples, …) pouvaient être apportées. Peut-être parce que je l'avais appliqué à moi-même : mes propres cours suivis en licence et en maîtrise étaient notés au brouillon et je m'efforçais de me les réexpliquer par une rédaction personnelle, en m'appuyant quelquefois sur un exemple numérique ou géométrique illustratif.*

Ainsi, c'est toujours la question du sens qui s'est placée en amont de la compréhension et de l'assimilation. Je suivais en cela la philosophie de mes mentors que sont Guy Brousseau (Brousseau, 1986) et Gérard Vergnaud (Vergnaud, 1994). Que j'aie alors cru bon d'accompagner pas à pas, par des apprentissages « déformalisés », des élèves de 13-14 ans en difficulté dans une expérience nationale novatrice, ne peut surprendre. Que j'aie travaillé pendant plus de dix ans au sein de l'Association des Professeurs de Mathématiques

[2] Dans la suite, par facilité, nous désignerons le champ de l'analyse statistique implicative par l'acronyme ASI.

(A.P.M.E.P.) à tenter de présenter les notions à enseigner en réponse à des problématiques et non pas à des injonctions de programme, s'explique de la même façon. En fait, vis-à-vis des autres ou de moi-même, le « pourquoi » a toujours précédé le « quoi ».

Q2 : Quelles expériences as-tu de l'enseignement, de la formation et de la recherche en mathématiques ?

J'ai enseigné plus ou moins longtemps à tous les niveaux d'enseignement, donc rencontré tous les âges d'apprenant : classes de l'école primaire, du collège, du lycée classique et professionnel, de l'université (DEUG, licence, maîtrise, DEA, CAPES, Agrégation) et formation continue d'adultes, enseignants ou non. Ceci m'a permis de percevoir les différences de représentation et de conception des notions mathématiques suivant la place que leur enseignement prenait dans leur cursus : formation générale, fonction diplômante, fonction professionnalisante, recyclage, etc., et corrélativement, les types de blocages rencontrés par ces différents apprenants.

J'ai pratiqué la recherche fondamentale en mathématiques après l'obtention de mes Diplômes d'Enseignement Supérieur (DEA) de Calcul Automatique (première rencontre avec des concepts d'algorithmique et d'informatique) et de Probabilités. Vivement intéressé par les concepts de fonction de risque, par la prédictabilité des probabilités, par l'organisation rationnelle de tâches, j'ai travaillé, en thèse de 3ème cycle, sur un sujet de processus de Markov avec politique de contrôle sous la conduite du génial et éclectique Professeur Michel Métivier. Après l'obtention de cette thèse, j'ai été happé à l'ouverture des I.R.E.M. par les problèmes de recyclage et de formation continue des enseignants, et ceux des changements de programmes d'enseignement, problèmes que je trouvais bigrement plus passionnants encore, en raison des contacts humains qui en découlaient. A cette occasion, j'ai donc rencontré les difficultés d'enseignement tant au niveau de l'école primaire que du collège et du lycée.

Q3 : Il me semble que tu te situes plutôt comme un ***probabiliste*** : que cela signifie-t-il pour toi ?

Il est vrai que la conceptualisation de l'incertain, en contact immédiat avec les applications aux paris, aux défis, à la relativité de la notion de mesure, m'a plus séduit que l'analyse ou la géométrie que je n'ai que superficiellement fréquentées. Et comme des probabilités à la recherche opérationnelle, il n'y a qu'un pas, j'ai pu réinvestir et alimenter certaines connaissances d'une discipline dans l'autre car les notions d'optimisation et de risque leur sont communes et les traversent.

Q4 : Qu'est alors pour toi la statistique ?

Je n'ai vraiment fréquenté la statistique qu'à travers l'analyse de données (AFC, ACP, Clustering,..). C'est donc à travers son filtre, certainement réducteur, que nous avons sympathisé ! C'est donc une vision tronquée que j'en ai et je m'en excuse auprès des authentiques statisticiens. Pour faire court, je dirai donc que je perçois la statistique comme un domaine disciplinaire où les probabilités puisent du sens, offrent des modèles tout en y exerçant un certain contrôle ; en effet, c'est aussi le lieu où la démarche inductive y retrouve

avec bonheur une certaine fonction prédictive, où l'incertain se mesure sous vigilance déductive.

Q5 : D'où te vient l'intérêt pour la relation d'implication, et l'idée de son approche par un modèle statistique ?

La nécessité faisant loi, afin d'évaluer les résultats d'une innovation pédagogique et de les comparer avec une taxonomie de complexité a priori (voir plus haut), j'ai dû me tourner vers des méthodes statistiques susceptibles de hiérarchiser des comportements de réponse, c'est-à-dire permettant d'énoncer : « quand l'élève réussit ceci alors, en général, il réussit cela » ou l'inverse. Accorder une mesure à cette quasi-implication, structurer hiérarchiquement l'ensemble des réussites en catégories, tels étaient les défis de ces objectifs. On en perçoit l'intérêt pour l'enseignant qui pourrait ainsi, par un transfert dans sa propre pratique, hiérarchiser a priori ses exercices et conjecturer les difficultés que rencontreront ses élèves. Aucune méthode statistique connue ne permettait véritablement d'y répondre de façon globale et systémique. D'où les premiers pas de l'A.S.I. (1979).

Si je voulais m'aventurer sur le terrain psychologique, voire psychanalytique, je dirais que dans mes recherches, j'ai assemblé inconsciemment des ingrédients épistémologiques quelquefois en opposition et que je pense avoir conciliés dans l'A.S.I. : la structure, la rigueur, la logique, la déduction, la mesure, l'invariant, la règle, mais également, le « quasi », le « presque », la valeur approchée, estimée, l'induction, le relatif, l'appliqué, le restructuré, la représentation, etc..

Q6 : Quels sont pour toi les apports marquants de l'approche A.S.I. ?

L'A.S.I. se présente comme une méthode classificatoire non supervisée, non symétrique d'analyse de dépendances orientées sur une base probabiliste à partir de données statistiques. Elle ne fait appel qu'à des outils mathématiques relativement élémentaires. Tout comme la plupart des méthodes d'analyses classificatoires courantes[3] *qui étudient et représentent des structures de données à partir de critères de ressemblances ou de similarités. Une distance euclidienne ou non euclidienne traduit généralement la proximité entre deux variables. Ainsi, ces méthodes sont essentiellement symétriques. Peu de méthodes sont dissymétriques. Les plus courantes admettent comme critère la confiance (probabilité conditionnelle) et son extension bayésienne (réseaux bayésiens). Or la probabilité conditionnelle ne permet pas de faire apparaître des pépites de connaissances car les événements rares, pourtant surprenants, inattendus, en sont écartés et y sont généralement écrasés par les événements les plus fréquents*[4]*. De plus, elle est invariante par dilatation des données, ce qui s'harmonise peu avec la philosophie de la statistique. Deux écueils qu'évite ou contourne l'A.S.I..*

Mais notre approche A.S.I. apporte bien plus que le simple traitement des dépendances entre variables par le seul indice d'implication et que la représentation, pourtant originale,

[3] Cet aspect élémentaire ne se retrouve toutefois pas dans les méthodes factorielles qui exigent de sérieuses connaissances en algèbre linéaire.

[4] Par exemple, si l'événement B est certain ou presque certain, Prob(B/A) **(**notation signifiant : probabilité de B sachant A) est peu différente de 1. Ainsi l'information apportée sur B est voisine de 0, donc pauvre, voire superflue. Nous aborderons plus en détail ce point dans la partie théorique.

de leur structure sous forme de graphe implicatif orienté, pondéré, sans cycle, constitué d'un ensemble de réseaux. Chaque arc du graphe représente une règle de type a→b. Il est pondéré par l'intensité d'implication. Un ensemble connexe de chemins du graphe est un réseau.

Tout d'abord, l'A.S.I. présente une dimension paradigmatique ne se limitant pas à la définition de cet indice de dépendance, définition à laquelle bien souvent des recherches en statistique se limitent et au sujet duquel elles se chamaillent. En effet, l'A.S.I. structure en système dynamique[5]*, une hiérarchie, l'ensemble des variables non seulement en règles de niveau 1 (a implique b) mais également en méta-règles ou règles de règles et les représente sous forme de hiérarchie orientée, la seule « sur le marché international », pourrait-on dire, à ce jour. La forme systémique et dynamique obtenue se différencie ainsi des structures établies à partir d'indices de distance ou de dissimilarité qui elles-mêmes conduisent à des partitions ou des nuages d'un espace géométrique.*

Ainsi, l'information apportée par un « tout » organisé en système - graphe et hiérarchie - est plus riche que la somme des informations apportées par chacune de ses parties –arc ou classe- Rappelons à ce sujet, ce qui en conforte le crédit inductif, qu'un système complexe résiste au changement en raison de mécanismes homéostatiques internes, ce qui en assure une sorte d'autoconservation. Ainsi, sur ces bases, l'expert pourra émettre des conjectures assez stables quant à la signification en termes de concepts qu'il pourra accorder à un chemin ou un réseau du graphe, une classe ou un niveau de la hiérarchie.

De plus, tout aussi original, à l'instar de certaines méthodes factorielles, l'A.S.I. établit une véritable dualité entre les deux espaces en jeu dans l'analyse : celui des variables et celui des individus (sujets, objets). Cette dualité permet de mettre en évidence la responsabilité relative des individus et de leurs descripteurs dans les structures obtenues. En outre, elle conduit à une structuration de l'espace des individus (sujets, objets) au moyen d'une topologie induite par leurs relations avec les variables.

Sur nombre de ces points, nous nous sommes souvent rejoints. Rappelle-toi, combien nos discussions, dans le début des années 80, sur les problèmes de dépendance non symétrique, au moment de ton DEA puis de ta thèse, ont été fructueuses pour l'un comme pour l'autre.

---

[5] Aux mots « structuralisme" et « structure », on trouve sur internet la référence suivante :

« L'idée de structure, de laquelle sont d'ailleurs souvent synonymes les concepts de figure, de forme et de système, ne repose pas sur une opposition à l'idée de fonction mais place au premier rang la notion de totalité : la totalité prime les parties et n'apparaît pas comme l'assemblage de plusieurs éléments (Mengal, dans Doron & Parot, p. 653). La définition très générale d'une structure est d'être un "ensemble d'éléments organisés selon certaines relations" (Richelle, dans Doron & Parot, 1991, p. 653). En biologie, on distingue habituellement la structure d'un organisme, c'est-à-dire le "groupe d'éléments dont les relations ont été repérées et décrites" (Richelle, dans Doron & Parot, 1991, p. 653), et sa fonction, c'est-à-dire l'"activité physiologique assurée par la structure" (idem) .

En psychologie, et par analogie avec la biologie, les structures désignent des "systèmes organisés qui sous-tendraient les comportements" (Richelle, dans Doron & Parot, 1991, p. 563). Il faut souligner que les "structures psychologiques" ne sont pas observables en tant que telles (on dit aussi que ce sont des "compétences"). Elles sont inférées (donc "construites") à partir des comportements qui eux, sont observables (on dit aussi que ce sont des "performances"). Selon Marc Richelle, en psychologie, la notion de structure tend à rendre compte de la "relative stabilité dans le temps des conduites", c'est-à-dire leur permanence ou aussi, leur "conservation" ».

Q7 : Quelles sont pour toi les grandes problématiques que nous devons affronter pour le développement de l'A.S.I. ?

Compte tenu de l'usage répandu des techniques des réseaux bayésiens, je souhaiterais que l'A.S.I. se positionne plus clairement dans l'examen des techniques et apports respectifs de ces deux grandes approches. Que l'on puisse faire ressortir les avantages et les limites de chacune d'elles, les spécificités de leurs champs d'application.

Quant à l'A.S.I., je souhaiterais que l'on approfondisse ses réponses aux problèmes des données floues, des données manquantes, de la qualité des structures obtenues ; aux problèmes de la sensibilité des paramètres aux variations des occurrences, de la robustesse des divers indices construits, et des arbres de décision, que l'on exploite moins chichement la dualité des deux espaces croisés, etc.. Et puis la multiplication des domaines d'application me paraît la façon la plus efficace pour crédibiliser une théorie, fût elle déjà relativement validée par de beaux résultats.

A travers tous ces développements et ces applications, l'objectif majeur vise à approcher au plus près la mise en évidence de la ***causalité*** *entre des phénomènes : quelle(s) variable(s), quelle(s) conjonction(s) de variables-prémisses est (sont) déterminante(s) à l'apparition d'une autre variable-conclusion ?avec quelle force, quelle intensité ?*

2.2 Corrélation, consécutivité, causalité dans la perspective de l'A.S.I..

Si nous regardons l'A.S.I. du point de vue de l'Exploration des Connaissances à partir des Données (ECD ou Fouille de données ou Data Mining), nous retrouvons l'objectif principal de l'établissement inductif de règles et quasi-règles entre des variables observées à travers des instances x d'un ensemble E d'objets ou de sujets. Une règle stricte (ou théorème dans ce cas) s'exprimera sous une forme symbolique : $\forall x,(a(x) \Rightarrow b(x))$. Une quasi-règle présentera des contre-exemples, c'est-à-dire que l'énoncé suivant sera observé : $\exists x,(a(x) \wedge \overline{b(x)})$. Le but de l'A.S.I. est de fournir une mesure à de telles règles afin d'en estimer la qualité. Dans un premier temps, dans le cadre de l'A.S.I. un indice de qualité a été construit, que nous avons nommé **intensité d'implication**, dans le but, à l'instar d'autres indices, d'apporter une réponse probabiliste à ce problème. Mais en recherchant parmi les règles[6] celles qui exprimeraient une causalité, une relation de cause à effet, ou tout au moins une relation de type causal, il nous a paru absolument nécessaire d'appuyer la satisfaction de la règle directe par une mesure de sa contraposée : $\forall x,(\overline{b(x)} \Rightarrow \overline{a(x)})$. En effet, si statistiquement, que ce soit avec la confiance mesurée par la probabilité conditionnelle ou avec l'intensité d'implication, la vérité d'une règle stricte est aussi obtenue avec sa contraposée, ce n'est plus nécessairement le cas avec une quasi-règle. Aussi, nous cherchons à construire de façon nouvelle et originale une mesure qui permette de dépasser le paradoxe de Hempel (1943) afin d'obtenir une mesure tendant à confirmer la satisfaction de l'induction en termes de causalité. Rappelons que, selon Carl G. Hempel, ce paradoxe est lié à la non-pertinence de la contraposition par rapport à l'induction alors que serait observée la non-satisfaction empirique (de fait) de la prémisse a. Il est la conséquence de l'application du 3ème principe de Hempel : « Si un objet observé x ne satisfait pas l'antécédent (soit a(x) = faux), il ne compte

[6] Dorénavant, « règles » signifiera aussi bien règles strictes que quasi-règles. Nous ne spécifierons que lorsqu'il pourrait y avoir équivoque ou pour souligner la spécificité de la qualité.

pas ou il est dénué de pertinence par rapport au conditionnel (= la proposition directe) ». C'est-à-dire que la confirmation de la contraposée n'apporte rien quant à la version directe de la proposition bien qu'elle lui soit logiquement équivalente. Par exemple, ce n'est pas de l'observation confirmatoire de la contraposée de « Tous les corbeaux sont noirs » par celle d'un chat rouge (donc non noir) que l'on confirme la validité de « Tous les corbeaux sont noirs ». Ni en poursuivant l'observation d'autres objets non-noirs. Car pour confirmer cet énoncé et donc valider l'induction, il nous faudrait passer en revue tous les objets non-noirs qui sont en nombre infini.

1	2	3	4	5	6
a(x)	b(x)	a(x)∧b(x)	a(x)∨b(x)	a(x)⇒b(x)	a(x)⇔b(x)
v	v	v	v	v	v
v	f	f	v	f	f
f	v	f	v	v	f
f	f	f	f	v	v

TAB.*1 :Tables de vérité*

Autrement dit encore, selon Hempel, dans la table de vérité de l'implication (Tab. 1 col. 1, 2 et 5), les cas où a(x) est faux sont inintéressants pour l'induction : seules les lignes [a(x)=vrai et b(x)=vrai] qui confirme et [a(x)=vrai et b(x)=faux] qui infirme, sont retenues. Or en A.S.I., ce paradoxe ne tient pas pour deux raisons :

1) les objets x relèvent d'un même ensemble de référence fini E dans lequel tous les x sont susceptibles, avec pertinence, de satisfaire ou non les variables en jeu ; c'est-à-dire que leur attribuant une valeur (de vérité ou numérique), la proposition directe et/ou sa contraposée sont elles-aussi valuables (par ex. la proposition est vraie même si a(x) est faux alors que b(x) est vraie) ;

2) Comme nous avons le plus souvent affaire à des quasi-règles, l'équivalence entre une proposition et sa contraposée ne tient plus ; mais c'est à partir de la conjonction des qualités respectives et évaluées de ces énoncés que nous induisons ou non un caractère causal. Par ailleurs si la règle est stricte, l'équivalence logique avec sa contraposée l'est et la règle contraposée est satisfaite en même temps.

Revenons maintenant sur les termes que la littérature utilise pour évoquer, sinon traiter, les questions relevant de l'implication et de la causalité.

2.2.1 **Corrélation totale, corrélation partielle**

Il y a corrélation totale (resp. partielle) entre deux variables a et b lorsque ces événements apparaissent (resp. presque) en même temps ainsi que leurs contraires. Or nous savons par des contre-exemples numériques (Gras et al 1996 c) que corrélation et implication ne se ramènent pas l'une à l'autre, qu'il peut y avoir corrélation sans implication et réciproquement. Avec la corrélation, le sens de la relation n'y est pas transparent puisqu'elle est symétrique, ce qui n'est pas le parti pris dans l'A.S.I.. D'une relation statistique donnée par la corrélation on peut déduire deux propositions empiriques contraires.

Citons à ce sujet, Alain Ehrenberg, sociologue, directeur de Recherches au C.N.R.S. qui écrit dans la revue *Sciences Humaines*, n° 198, nov. 2008 :

« *Le constat d'une corrélation ne lève pas l'ambiguïté entre "quand je fais X, mon cerveau est dans l'état Y" et "si je fais X, c'est parce que mon cerveau est dans l'état Y", c'est-*

à-dire entre quelque chose qui se passe dans mon cerveau quand je fais une action et quelque chose que je fais quand j'agis parce que mon cerveau en est la cause. Deux choses différentes sont ainsi mélangées ».

2.2.2 **Consécutivité**

Il y a consécutivité lorsque b se produit après a mais sans que soit nécessairement vérifié que b puisse se produire indépendamment de a. En effet, le paramètre « temps » bruite la relation et peut conduire à un artefact non contrôlé. De plus, l'axe du temps qui semble ici fondamental pourrait exclure la considération de l'observation de non-b lorsque non-a se produit.

2.2.3 **Causalité**

Nous dirons qu'il y a relation de type causal[7] de a sur b si nous observons les conditions suivantes : d'une part, b se produit quand ou dès que a se produit, et d'autre part, b ne se produit pas lorsque ou alors que a ne se produit pas. Autrement dit (TAB. 2) $(a(x) \wedge \overline{b}(x)$ est faux dans toute circonstance x.

1	2	3	4
a(x)	b(x)	non-b(x)	a(x)∧non-b(x)
v	v	f	f
f	f	v	f

TAB. 2 *:Tables de vérité*

C'est le cas de l'implication formelle ou stricte de a sur b car ces conditions permettent la vérification simultanée de l'implication directe et de sa contraposée. Du point de vue de la quasi-implication, nous considérerons la causalité avec un niveau de confiance et nous parlerons de (1-α)–causalité lorsque statistiquement ces deux événements sont satisfaits au seuil de risque α. Aussi, nous désignerons cette relation causale par causalité statistique. Cette modération de la causalité renvoie à un degré de plausibilité de son existence qui est une fonction décroissante du niveau de risque α. Notons que dans ces conditions formelles ou non, plusieurs prémisses peuvent être causales par rapport à la même conclusion.

Dans *International Statistical Review* [8], D. R. Cox et Nanny Wermuth (2004) dressent la liste des auteurs et de leurs articles publiés sur le thème de la causalité. Le premier, semble-t-il, est Yule (1900) qui mit l'accent sur la différence que nous venons de rappeler entre causalité et corrélation. Cochran (1965) se vit répondre par Fischer, attaché à une technique de randomisation, qu'il se devait d'établir une théorie de la causalité s'il persistait à vouloir en faire l'étude. Hill (1965) en donna une ligne directrice à la faveur d'un contexte épidémiologique. Box (1966) insista sur la soin à apporter à l'interprétation causale aux équations de régression issues d'observations. Ce que firent de façon plus systématique Cox et Snell (1981) en contrôlant les effets de changements d'une ou plusieurs variables. Rubin (1974) développa et appliqua les idées de Neyman (1923) dans des contextes de sciences sociales.

[7] Dans divers sens en fonction de la contingence : psychologique, physique, occurrentiel, etc..

[8] Causality : a Statistical View, *International Statistical Review* (2004), 72, 3, 285-305, International Statistical Institute

Robins (1997) explora les notions de causalité dans des essais cliniques et un cadre épidémiologique. Rosenbaum (2002) a posé les questions conceptuelles et méthodologiques ouvrant ainsi un axe de recherche important. A sa suite, D. R. Cox et Nanny Wermuth font le choix de retenir les questions suivantes :

- Choisir une forme générale adaptée de la relation de régression,
- Déterminer quelles sont les variables à explorer qui peuvent légitimement être incluses dans celles ayant une interprétation potentiellement causale,
- Examiner les effets non linéaires et interactifs qui peuvent induire une interprétation correcte,
- Combiner les tendances à preuve à partir de nombreuses études.

Toutefois l'étude théorique qu'ils mènent dans cet article se fonde sur la notion de probabilités conditionnelles simples ou composées, de propriétés bayésiennes et explore plus en détail le cas où les variables sont des combinaisons linéaires de certaines d'entre elles. Cette étude sérieuse n'est pas reprise dans le cadre de l'A.S.I. pour les trois raisons suivantes :

1° l'approche bayésienne, conduisant à un réseau bayésien, n'est pas celle adoptée dans la cadre de l'A.S.I. au regard des restrictions que nous avons développées (Gras et al 2004). Celles-ci portent essentiellement sur la sensibilité réduite dans le cas de croissance des observations et sur la trivialité de la mesure de qualité associée à la probabilité conditionnelle lorsque la variable « conditionnée » est évidente ;

2° aucune étude restrictive au cas linéaire n'est faite dans le cadre de l'A.S.I.. Elle ne constitue pas un préalable pour mettre en évidence les conditions d'observation de la causalité ;

3° l'examen de la confirmation de l'hypothèse causale par la contraposition n'est pas faite en analyse bayésienne alors qu'elle semble fondamentale et effectuée en conséquence dans le cadre de l'A.S.I..

Et pourtant, les objectifs de ces deux méthodes d'analyse de la causalité instrumentées, pour l'une, par les réseaux bayésiens ou les treillis de Galois et, pour l'autre, par les graphes implicatifs, sont les mêmes : tenter de déterminer parmi les prémisses issues d'une variable observée celle ou celles, si elles existent, qui impliqueraient le plus intensément la variable-conclusion. En fait, revenant à la couverture conceptuelle recouvrant une variable a qui, en général, est une conjonction de variables, on cherche à établir une relation hiérarchique de subsomption afin de déterminer le concept qui subsume le plus le concept b. Autrement dit, utilisant la représentation donnée par le logiciel CHIC, il s'agit de trouver le concept optimal en amont du sommet b d'un cône implicatif. Ce qui n'interdit pas qu'il puisse exister plusieurs variables ayant une relation apparemment causale vis-à-vis de la même conclusion. L'émission d'une hypothèse forte de causalité appelle nécessairement l'utilisation conjointe d'une distance sémantique permettant de faire le choix parmi les antécédents d'un nœud « conclusion », non seulement celui qui optimise une intensité d'implication mais aussi celui qui admet une faible distance sémantique avec ce nœud.

La littérature nous offre également des tests de relation causale. C'est ainsi que ce que les auteurs établissent en tant que tests (Granger, Haugh-Pierce et Sims cités par Jean-Marie Dufour, rapport de recherche à l'Université de Montréal, avril 2002) consistent à accepter la

non-causalité en testant l'indépendance par le rejet d'une relation linéaire. Ce qui, à notre avis, semble insuffisant en raison des hypothèses limitatives et par trop symétriques.

D'une certaine manière, nous pouvons dire que l'analyse statistique implicative dans sa forme que nous nommons classique exprime la surprise, l'étonnement statistique provoqué par l'observation d'une quasi-implication de la variable a sur la variable b ; c'est-à-dire encore, de la quasi-inclusion de l'ensemble A des observations de a dans l'ensemble B de celles de la variable b. L'intensité d'implication qui attribue une valeur de 0 à 1, est une mesure de la qualité de cette surprise. Elle peut être considérée comme apportant une présomption de causalité. Mais l'indice d'implication, dit classique, qui est à la base de l'intensité d'implication ne distingue pas les qualités respectives de l'implication directe et de sa contraposée. Il leur accorde la même valeur en s'opposant à la sémantique de nombreux cas observés.

Pour aborder cette problématique limitative, nous avons construit une mesure de la double qualité de l'inclusion de A dans B et du complémentaire de B dans celui de A, satisfaisant la sémantique classique de la causalité : « à chaque fois que j'observe a, j'observe b et à chaque fois que b n'apparaît pas alors je n'observe pas a. je présume alors que a est cause de l'effet b ». Nous l'avons nommée : intensité d'inclusion, nombre qui attribue une valeur à la causalité de a sur b ou tout au moins dont elle en renforce la présomption. Précisons.

Le concept sur lequel nous basons **l'intensité d'inclusion** est celui d'entropie de Shannon (1949). Pour ce faire, nous avons choisi comme critère, l'expression de la faiblesse de l'entropie de la réalisation de b lorsque a se réalise, c'est-à-dire la qualité de l'information sur b quand a se réalise et, de la même façon, la qualité de l'information sur non-a lorsque non-b se réalise. Notons que, dans le contexte ECD que nous avons évoqué, cette approche nous semble unique parmi les méthodes répertoriées dans la littérature consacrée. Elle pallie aussi une des limites des réseaux bayésiens. En effet, une relation causale entre deux événements sera plus volontiers ressentie dans la vie courante si prémisse et conclusion sont rares. En tout cas elle le sera davantage que lorsque la prémisse est soit rare soit fréquente avec une conclusion fréquente, autrement dit triviale. Par suite, les autres méthodes de recherche de causalité qui filtrent les règles sur la base du support et de la confiance laissent de côté ces règles où, tout au moins psychologiquement, se nichent les relations causales. C'est en ce sens que l'A.S.I. offre un apport fondamental

En ce sens, en effet, nous pouvons dire l'analyse entropique développée dans le cadre de l'A.S.I. associe la surprise et la causalité. D'une part, nous avons construit un indice, l'indice d'inclusion qui intègre l'information délivrée par la réalisation d'un nombre faible de contre-exemples à la règle et à son contraposée. D'autre part, à partir de cet indice d'inclusion, nous avons construit une mesure que nous nommons **intensité d'implication-inclusion** ou **intensité entropique** qui qualifie donc la surprise statistique d'une certaine causalité de a sur b en intégrant la qualité de l'inclusion. Nous y reviendrons en détail dans le chapitre suivant. Lors d'une analyse statistique implicative selon l'approche que nous dénommons entropique, si nous observons une valeur forte de l'intensité entropique entre deux variables, nous pouvons alors présumer une relation de causalité de la prémisse sur la conclusion. Il est même possible, d'une part, de faire le choix de la relation la plus forte associant UNE des prémisses et LA conclusion, d'autre part, d'utiliser la conjonction entre prémisses pour extraire la relation la plus intense et ainsi assurer une présomption maximum.

2.2.4 A propos de causalité en sciences sociales

La notion de cause partage les sociologues tout comme les philosophes. A. Comte affirme que la recherche de cause doit être remplacée par celle de lois qui présupposent un certain déterminisme. Ce qui ne manque pas de conduire certains à rejeter la causalité. E. Durkheim pense que cette recherche passe par l'utilisation d'une méthode basée sur un modèle mathématique comme nous le rapportons ci-dessous. Les sociologues contemporains (par ex. P. Lazarsfeld ou R. Boudon**)** ont élaboré des outils d'analyse de la causalité avec la même rigueur que ceux de l'économétrie (analyse causale). Du fait de l'interdépendance entre les variables, ces outils prennent en compte non seulement en un schéma simple, les relations duales entre variables mais surtout la structure causale des variables, c'est-à-dire un schéma complexe où des variables explicatives sont liées les une aux autres.[9]

Dans son livre *Le suicide, Etude de Sociologie*, Durkheim (1897) mathématise par un modèle linéaire, mais opérationnel, les relations entre trois phénomènes (a = âge, c = état civil, t = taux de suicides) selon une matrice 3x3 :

$$\begin{pmatrix} \mathbf{1} & \mathbf{0} & \mathbf{0} \\ \boldsymbol{x_{12}} & \mathbf{1} & \mathbf{0} \\ \boldsymbol{x_{13}} & \boldsymbol{x_{23}} & \mathbf{1} \end{pmatrix}$$

Il transite par la relation matricielle $\begin{pmatrix} a \\ c \\ t \end{pmatrix} = M \begin{pmatrix} e_1 \\ e_2 \\ e_3 \end{pmatrix}$

Cette matrice indique :

1. l'indépendance de a par rapport à c et t : $a = e_1$ (effet de facteurs)
2. la dépendance linéaire de c par rapport à a : $c = x_{12}.a + e_2$
3. la dépendance linéaire de t par rapport à c et à a : $t = x_{13}.a + x_{23}.c + e_3$

Ces relations peuvent également être représentées par un graphe implicatif suivant :

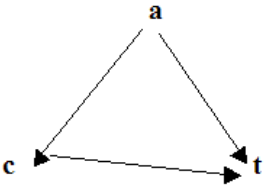


FIG. 1 – Graphe implicatif.

[9] C'est ce que l'on se propose en ASI avec le graphe implicatif et, de façon plus complexe encore, avec la hiérarchie orientée en tant que représentant une méta-causalité.

PARTIE 1

Textes inventoriés, coordonnés, mis en forme ou reformulés par

Régis Gras* et Jean-Claude Régnier**

*LINA– Ecole Polytechnique de l'Université de Nantes, UMR 6241
La Chantrerie BP 60601 44306 Nantes cedex
regisgra@club-internet.fr
** Université de Lyon - UMR 5191 ICAR
ENS-LSH 15, Parvis René Descartes BP 7000 69342 LYON cedex 07
jean-claude.regnier@univ-lyon2.fr

FONDEMENTS THÉORIQUES DE L'ANALYSE STATISTIQUE IMPLICATIVE

Résumé. La partie 1 vise à exposer en 9 chapitres, la théorie de l'Analyse Statistique implicative (ASI). Nous cherchons à y définir le plus précisément possible les concepts et les théorèmes de cette théorie ainsi que leurs fondements épistémologiques et méthodologiques. Parmi ceux-ci, citons : relation de quasi-implication, coefficient et indice d'implication, de propension, graphe implicatif, classification hiérarchique orientée, etc. De plus, chaque concept est illustré par un exemple.

Chapitre 1 : Analyse implicative des variables binaires. Intensité implicative. Intensité entropique.

1 Approche épistémologique de l'ASI

Les connaissances opératoires que les êtres humains construisent sur le monde se constituent principalement selon deux ordres : celui des faits et celui des règles entre les faits ou entre règles elles-mêmes. Ce sont leurs apprentissages qui, à travers leur culture et au travers d'expériences sociales ou singulières, leur permettent une élaboration progressive de ces formes de connaissances, en dépit des régressions, des remises en cause, des ruptures qui surgissent au détour d'infirmations décisives. Cependant, on sait que celles-ci contribuent dialectiquement à lui assurer un équilibre opératoire. Or les règles se forment inductivement de façon relativement stable dès lors que le nombre de succès, quant à leur qualité explicative ou anticipatrice, atteint un certain niveau de confiance à partir duquel elles seront susceptibles d´être mises en oeuvre. En revanche, quand ce niveau subjectif demeure non atteint, l´économie de l´individu le fera résister, dans un premier temps, à son abandon ou à sa critique. En effet, il est coûteux de substituer à la règle initiale une autre règle lors de

l´apparition d´un faible nombre d'infirmations, dans la mesure où elle aurait été confortée par un nombre important de confirmations. Un accroissement de ce nombre d'instances négatives, fonction de la qualité de robustesse du niveau de confiance en la règle, conduira peut-être à un réajustement de celle-ci, voire à son abandon. Laurent Fleury (Fleury 1996), dans sa thèse, cite avec pertinence l´exemple de la règle tout à fait admissible : « toutes les Ferrari sont rouges ». Cette règle, très robuste, ne sera pas pour autant abandonnée lors de l´observation d´un seul ou de deux contre-exemples. D´autant plus qu´elle ne manquerait pas d´être rapidement re-confortée.

Ainsi, à l´opposé de ce qui est légitime en mathématiques, où aucune règle (théorème) ne souffre d´exception et où le déterminisme est total, les règles dans les sciences humaines et sociales, plus généralement dans les sciences dites "molles", sont acceptables et donc opératoires tant que le nombre de contre-exemples restera "supportable" en rapport à la fréquence de situations où elles seront positives et efficaces. Le problème dans le champ de l'analyse des données, est alors d´établir un critère numérique, relativement consensuel, pour définir la notion de niveau de confiance ajustable au niveau d´exigence de l´usager de la règle. Qu´il soit alors établi sur des bases statistiques peut alors ne pas surprendre. Qu´il possède une propriété de résistance non linéaire au bruit (faiblesse du ou des premiers contre-exemples) peut également paraître naturel, conforme au sens "économique" évoqué plus haut. Qu´il s´effondre si les contre-exemples se répètent semble aussi devoir guider notre choix dans la modélisation des critères recherchés.

Différentes approches théoriques ont été adoptées pour modéliser l'extraction et la représentation de règles d'inférence imprécises (ou partielles) entre variables binaires (ou attributs ou caractères) décrivant une population d'individus (ou sujets ou objets). Mais les situations de départ et la nature des données ne modifient pas la problématique initiale. Il s'agit de découvrir des règles inductives non symétriques pour modéliser des relations du type *"si a alors presque b"*. C'est, par exemple, l'option des réseaux bayésiens (Amarger et al 1991, Pearl 1988) ou des treillis de Galois (Simon 2000). Mais le plus souvent, l'indice de corrélation linéaire et le test du χ^2, s'avérant inadaptés du fait de leur caractère symétrique, la probabilité conditionnelle (Loevinger 1947, Agrawal et al. 1993, Gras et al. 2004) reste le moteur de la définition de l'association, même quand l'indice de cette association retenu est de type multivarié (Bernard et Poitrenaud 1999).

De plus et à notre connaissance à ce jour, d'une part, les différents développements intéressants se centrent le plus souvent sur des propositions d'un indice d'implication partielle pour des données binaires (Lerman et al. 2004) ou (Lallich et al 2005) et, d'autre part, cette notion n'est pas étendue à d'autres types de variables que les variables binaires, à l'extraction et à la représentation selon un graphe de règles ou selon une hiérarchie de méta-règles en tant que structures visant l'accès à la signification d'un tout non réduit à la somme de ses parties[1], c'est-à-dire fonctionnant comme un système complexe non linéaire. Par exemple, on sait fort bien, par l'usage même, que la signification d'une phrase ne passe pas complètement par le sens de chacun des mots qui la compose. Par ailleurs il semblerait que, dans la littérature consacrée à ces questions, la notion d'indice d'implication ne soit pas non plus étendue à la recherche de sujets et de catégories de sujets responsables des associations. Ni même que

[1] C'est ce que souligne le philosophe L. Sève : « …dans le passage non additif, non linéaire des parties au tout, il y a *apparition de propriétés* qui ne sont d'aucune manière *précontenues* dans les parties et qui ne peuvent donc s'expliquer par elles » (*Émergence, complexité et dialectique* , Odile Jacob, mai 2005).

cette responsabilité soit quantifiée et conduise, de ce fait, à une structuration réciproque de l'ensemble des sujets, conditionnée par leurs relations aux variables.

2 Situation fondamentale et fondatrice de l'approche classique.

Une population E d'individus, objets ou sujets, est croisée avec des variables (caractères, critères, réussites, etc.) que l'on explore de la façon suivante : "*dans quelle mesure peut-on considérer qu'instancier la variable*[2] *a implique instancier la variable b ?* Autrement dit, *les individus ont-ils tendance à être b si l'on sait qu'ils sont a ?"*. Dans les situations habituelles de la vie humaine ou dans les domaines des sciences de la vie ou des sciences humaines et sociales, où les théorèmes (si a alors b) au sens déductif du terme ne peuvent être établis du fait des exceptions qui les entachent, il est important pour le chercheur comme pour le praticien de "*fouiller dans ses données*" afin de dégager, malgré tout, des règles suffisamment fiables (des sortes de "théorèmes partiels", des inductions) pour pouvoir conjecturer[3] une possible relation causale ou pour le moins quasi-causale, par exemple, pour décrire, structurer une population et faire l'hypothèse d'une certaine stabilité à des fins descriptives et, si possible, prédictives. Ainsi Hervé Londeix[4] recourant à l'ASI dans un cadre de psychologie différentielle met en évidence un ordre partiel des stades piagétiens à partir d'épreuves proposées à de jeunes enfants Mais cette fouille exige la mise au point de méthodes pour la guider et pour la dégager du tâtonnement et de l'empirisme.

3 Modélisation mathématique de l'approche classique.

Pour ce faire, à l'instar de la méthode de mesure de la similarité de I.C. Lerman (1981), à l'instar de la démarche classique dans les tests non paramétriques (ex. Fischer, Wilcoxon, etc.), nous définissons (Gras 1979, Gras et al. 1996 c) la mesure de qualité confirmatoire de la relation implicative $a \Rightarrow b$ à partir de l'invraisemblance de l'apparition, dans les données, du nombre de cas qui l'infirment, c'est-à-dire pour lesquels a est vérifié sans que b ne le soit. Ceci revient à comparer l'écart entre le contingent et le théorique si seul le hasard intervenait[5]. Mais, dans le cadre de l'analyse de données, c'est cet écart qui est pris en compte et non pas l'énoncé d'un rejet ou de l'admissibilité d'hypothèse nulle. Cette mesure est relativisée par le nombre de données vérifiant respectivement a et non b, circonstance dans laquelle l'implication est précisément mise en défaut. Elle quantifie "*l'étonnement"* de

[2] Ici, le mot « variable » désigne aussi bien une variable isolée en prémisse (ex. : « être blonde ») qu'une conjonction de variables isolées (ex. : « être blonde **et** avoir moins de 30 ans **et** habiter Paris »)

[3] « L'exception confirme la règle » nous dit l'adage populaire qui devrait être pris au sens où il n'y aurait pas d'exceptions s'il n'y avait pas de règle. Dans le contexte du raisonnement déductif, il faudrait bien sûr dire « L'exception infirme la règle. »

[4] Londeix H. (1983) *Approche génétique et différentielle du développement intellectuel*. Thèse de doctorat. Université de Bordeaux II.

[5] « … [en accord avec Jung] si la fréquence des coïncidences n'excède pas de façon significative la probabilité qu'on peut leur calculer en les attribuant au seul hasard à l'exclusion de relations causales cachées, nous n'avons certes aucune raison de supposer l'existence de telles relations. », Atlan H., (1986) *A tort et à raison. Intercritique de la science et du mythe*, Paris : Seuil.

l'expert devant le nombre invraisemblablement petit de contre-exemples sous l'hypothèse d'une indépendance entre les variables eu égard aux effectifs en jeu.

Précisons plus avant la modélisation. Un ensemble fini V de v variables, désignées par des lettres a, b, c, etc., est donné. Dans la situation paradigmatique classique, il s'agit des performances (réussite-échec) à des items d'un questionnaire de connaissances. A un ensemble fini E de n sujets désignés x, on associe, par abus d'écriture, les fonctions du type : $x \rightarrow a(x)$ où a(x)=1 (ou a(x)=vrai) si x satisfait ou possède le caractère a et 0 (ou a(x)=faux) sinon. En intelligence artificielle, on dira que x est un exemple ou une instance pour a si a(x)=1 et un contre-exemple dans le cas contraire.

La règle $a \Rightarrow b$ est logiquement vraie si pour tout x de l'ensemble E, b(x) n'est nul que dans le cas où a(x) l'est aussi, autrement dit si l'ensemble A des x pour lesquels a(x)=1 est contenu dans l'ensemble B des x pour lesquels b(x)=1. Cependant, cette inclusion stricte n'est qu'exceptionnellement observée dans les expériences réelles. Dans le cas d'un questionnaire de connaissances, on pourrait en effet observer quelques rares élèves réussissant un item a et ne réussissant pas l'item b, sans que ne soit contestée la *tendance* à réussir b quand on a réussi a. Relativement aux cardinaux de E (soit n), de A (soit n_a) et de B (soit n_b), c'est le poids des contre-exemples (soit $n_{a \wedge \bar{b}}$) qu'il faut donc prendre en compte pour accepter statistiquement de conserver ou non la **quasi-implication** ou **quasi-règle** $a \Rightarrow b$. Ainsi, c'est à partir de la dialectique entre les exemples et les contre-exemples que la règle apparaît comme le dépassement de la contradiction.

4 Formalisation de la quasi-règle implicative dans l'approche classique.

Pour formaliser cette quasi-règle, nous considérons, comme le fait I.C. Lerman pour la similarité, deux parties quelconques X et Y de E, choisies aléatoirement et indépendamment (absence de lien a priori entre ces deux parties) et de mêmes cardinaux respectifs que A et B. Soit $\overline{Y}$ et $\overline{B}$ les ensembles complémentaires respectifs de Y et de B dans E de même cardinal $n_{\bar{b}} = n - n_b$.

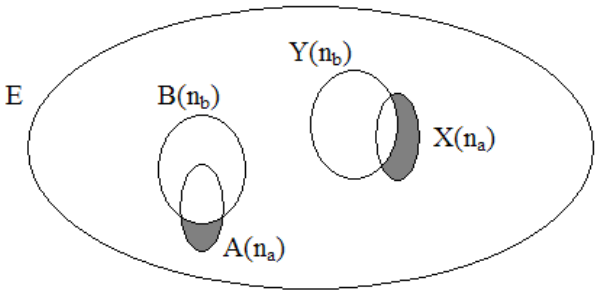


FIG. 1- *représentation par les diagrammes d'Euler*

Nous dirons alors :

Définition 1: la quasi-règle a $\Rightarrow$ b est *admissible au niveau de confiance* 1-α si et seulement si $\Pr\left[Card(X \cap \overline{Y}) \leq card(A \cap \overline{B})\right] \leq \alpha$]

Intuitivement et qualitativement, ceci signifie que la quasi-implication a $\Rightarrow$ b sera admissible à l'issue d'une expérience si le nombre d'individus de E la contredisant est invraisemblablement petit par rapport au nombre d'individus attendu sous une hypothèse d'absence de lien. Par exemple, si Card E = 100, Card A = 36, Card B = 50, alors $card(A \cap \overline{B}) = 3$ est "invraisemblablement petit" sous l'hypothèse d'une absence de lien entre a et b. On constate, en effet, que A est "presque" contenu dans B, alors que, sans liaison de A et B, on pourrait s'attendre à ce qu'environ la moitié des éléments de A soient aussi dans B.

Définition 2: On appelle intensité d'implication de la quasi-règle $a \Rightarrow b$, le nombre $\varphi(a,b) = 1 - \Pr\left[Card(X \cap \overline{Y}) \leq Card(A \cap \overline{B})\right]$ si $n_b \neq n$ et $\varphi(a,b) = 0$ si $n_b = n$

L'intensité d'implication est une valeur probabiliste, et non une fréquence, qui fonde la décision de retenir ou non une relation de quasi-implication entre les variables binaires a et b.

Cette modélisation de la quasi-implication est pertinente pour mesurer l'étonnement face au constat de la petitesse du nombre des contre-exemples en regard du nombre surprenant des instances de l'implication. Il s'agit d'une mesure de la qualité inductive et informative de l'implication. Par conséquent, si la règle est triviale, comme dans le cas où B est très grand ou coïncide avec E, cet étonnement devient petit. Nous démontrons (Gras R., 1996 c) d'ailleurs que cette trivialité se traduit par une intensité d'implication très faible, voire nulle : *Si, n_a étant fixé et A étant inclus dans B, n_b tend vers n (B "croît" vers E), alors $\varphi(a,b)$ tend vers 0*. C'est pourquoi nous définissons par « continuité »: $\varphi(a,b) = 0$ si $n_b = n$. De même, si $A \subset B$, $\varphi(a,b)$ peut être inférieure à 1 dans le cas où la confiance inductive, mesurée par l'étonnement statistique, est insuffisante.

5 Différents modèles pour évaluer l'intensité d'implication.

La détermination de l'intensité d'implication dépend du modèle retenu pour définir la loi de probabilité de la variable aléatoire « nombre de contre-exemples ».

5.1 Modèle de Poisson

Reprenant une idée abordée dans (Bodin et al., 1997), au lieu de considérer la situation statistique figée par les données du croisement sujets-variables, envisageons un processus temporel transactionnel discret, durant lequel, à certains instants, apparaît une transaction. Celle-ci représente, par exemple, tout aussi bien un sujet, une feuille d'enquête qu'une opération bancaire. Par exemple, si la transaction *s* satisfait (resp. ne satisfait pas) la variable *a* (par exemple un attribut) de *V*, nous l'instancions par 1 (resp. 0) à l'intersection de la ligne s et la colonne *a*. Les transactions peuvent être considérées comme tirées d'une population-mère supposée infinie. Le choix de cette approche dynamique, en harmonie avec la philosophie de l'ASI, puise sa justification dans l'éventail de situations illustratives où se

produisent des arrivées successives d'informations, de données, d'observations, capitalisées à un moment T fixé, dans un tableau de croisement E x V.

Dans la contingence, au bout d'un nombre *n* d'instants (donc *n*=T par convention), on a observé *n* transactions dont, parmi elles, n_a et n_b transactions de types respectifs *a* et *b*. Le nombre de contre-exemples à l'événement « si *a* alors *b* » est $n_{a\wedge\bar{b}}$ où $\bar{b}$ désigne l'événement « non *b* ».

Afin de mesurer la qualité de la quasi-règle $a \Rightarrow b$, comme nous l'avons déjà dit, nous construisons alors un modèle aléatoire en comparant ce nombre de contre-exemples à celui qu'aurait donné le seul hasard si les types *a* et *b* apparaissaient, de façon indépendante, au cours d'un processus respectant des hypothèses « raisonnables » que nous préciserons plus loin. Pour évaluer cette qualité, les événements aléatoires qui nous intéressent sont ceux où seraient réalisées les apparitions de *a* et de non *b* parmi les n transactions.

Notons A l'événement réalisant la variable *a*, soit [*a*=1], au cours du processus. De même notons B l'événement [*b*=1]. Du fait que dans la contingence, nous observons n_a fois *a* et n_b fois *b*, nous attribuons dans le modèle aléatoire à A (resp. B) la probabilité estimée par $\frac{n_a}{n}$ (resp. $\frac{n_b}{n}$). De plus, A et B devant être indépendants par hypothèse, la réalisation simultanée de A ([*a*=1]) et non B ([*b*=0]), lors d'une transaction aléatoire, y aura pour probabilité estimée $\frac{n_a}{n}.\frac{n_{\bar{b}}}{n}$. On pourrait alors ici s'arrêter sur le modèle binomial permettant de préciser la loi de la variable aléatoire « nombre de contre-exemples » qui pourrait s'imposer d'évidence. Nous y reviendrons plus loin.

Pour préciser et légitimer le modèle de processus d'extraction des transactions spécifiées, nous énonçons les hypothèses sémantiquement admissibles suivantes, relativement à la réalisation de l'événement : A et non B = [*a*=1 et *b*=0].:

- h1 : les temps d'attente successifs d'un événement [A et non B] sont des variables aléatoires indépendantes. Cette hypothèse est légitimée par l'indépendance a priori de A et B ;
- h2 : la loi du nombre d'événements survenant dans un intervalle de temps de durée T ne dépend que de T indépendamment de l'origine de temps ; ceci nécessite que le dépouillement des données soit régulier, ce qui est une moindre exigence ;
- h3 : deux tels événements ne peuvent arriver simultanément, ce qui est le cas du dépouillement des transactions qui est séquentiel

On démontre alors (Saporta, 2006) que le nombre d'événements, se produisant pendant une période de durée *n* fixée, suit une loi de Poisson de paramètre *c.n* où *c* est appelé cadence du processus d'apparitions de [*a*=1 et *b*=0] pendant l'unité de temps. Par suite, dans notre modèle, la cadence choisie est estimée par $c = \frac{n_a}{n}.\frac{n_{\bar{b}}}{n}$. Ainsi pour une durée de temps n, les apparitions de l'événement [A et non B] suivent une loi de Poisson de paramètre λ dont nous connaissons une estimation : $\lambda_{estimé} = \frac{n_a.n_{\bar{b}}}{n}$

Par suite $\forall s \in \{0,1,2,...,n\}$ Pr[Card($X \cap \overline{Y}$) = s]= $\frac{\lambda^s}{s!} e^{-\lambda}$ dans laquelle λ est remplacé par $\hat{\lambda} = \lambda_{\text{estimé}}$.

Remarquons que la nullité du nombre de contre-exemples à $a \Rightarrow b$ est équivalente à l'inclusion de X dans Y, en tant que parties aléatoires extraites de la population-mère supposée infinie.

En conséquence, la probabilité pour que le hasard conduise, sous l'hypothèse d'absence de lien a priori entre *a* et *b*, à au plus le nombre de contre-exemples observé, est estimée par :

$$\Pr\left[Card(X \cap \overline{Y}) \leq n_{a \wedge \overline{b}}\right] = \sum_{s=0}^{n_{a \wedge \overline{b}}} \frac{\hat{\lambda}^s}{s!} . e^{-\hat{\lambda}}$$

L'intensité de l'implication estimée, notée $\varphi(a,b)$ est alors :

$$\varphi(a,b) = 1 - \Pr\left[Card(X \cap \overline{Y}) \leq n_{a \wedge \overline{b}}\right] = 1 - \sum_{s=0}^{n_{a \wedge \overline{b}}} \frac{\hat{\lambda}^s}{s!} . e^{-\hat{\lambda}}$$

Définition 3: On appelle indice d'implication de la quasi-règle $a \Rightarrow b$, la variable aléatoire, notée $Q(a,\overline{b})$, déduite de la variable aléatoire $Card(X \cap \overline{Y})$ par centrage réduction.

Pour λ≥5 (λ paramètre estimé par $\frac{n_a n_{\overline{b}}}{n}$, rappelons-le), la variable « indice d'implication », notée : $Q(a,\overline{b}) = \dfrac{Card(X \cap \overline{Y}) - \frac{n_a n_{\overline{b}}}{n}}{\sqrt{\frac{n_a n_{\overline{b}}}{n}}}$ qui résulte du centrage-réduction de la variable de Poisson, Card(X ∩ $\overline{Y}$), peut être approchée par la variable gaussienne centrée réduite N(0 ;1).

Si nous considérons la **valeur empirique de l'indice** $q(a,\overline{b}) = \dfrac{n_{a \wedge \overline{b}} - \frac{n_a n_{\overline{b}}}{n}}{\sqrt{\frac{n_a n_{\overline{b}}}{n}}}$, alors l'intensité d'implication estimée de la quasi-règle $a \Rightarrow b$, est approximativement :

$$\varphi(a,b) = 1 - \Pr\left[Q(a,\overline{b}) \leq q(a,\overline{b})\right] = \frac{1}{\sqrt{2\pi}} \int_{q(a,\overline{b})}^{\infty} e^{-\frac{t^2}{2}} dt$$

Insistons sur le sens de cette intégrale. Elle représente la probabilité gaussienne pour que le nombre de transactions observées satisfaisant la quasi-règle $a \Rightarrow b$, soit supérieur à celui qui serait observable sous l'hypothèse d'indépendance de *a* et b. Autrement dit, $\Pr[Q(a,\overline{b}) \leq q(a,\overline{b})]$ est la **p-value** du test visant à réfuter l'hypothèse de l'indépendance de *a* et *b* au profit d'une relation de type quasi-implication

5.2 Modèle binomial

Examiner la qualité de la quasi-règle a⇒b, dans le cas où les variables sont binaires, revient à mesurer de façon équivalente celle de l'inclusion du sous-ensemble des transactions satisfaisant *a* dans le sous-ensemble des transactions satisfaisant *b*. Les contre-exemples relatifs à l'inclusion sont en effet les mêmes que ceux qui sont relatifs à l'implication exprimée par : « toute transaction satisfaisant a satisfait aussi b ». Dans cette optique ensembliste, dès que $n_a \leq n_b$, la qualité de la quasi-règle a⇒b, ne peut qu'être sémantiquement meilleure que celle de $b \Rightarrow a$. Nous supposerons donc, par la suite, que $n_a \leq n_b$ lors de l'étude de a⇒b. Dans ce cas, la population-mère est finie et *Card E = n.*

La modélisation binomiale fut chronologiquement la première adoptée (Gras 1979 chap. 2). Elle fut comparée à d'autres modélisations dans (Lerman et al. 1981). Rappelons brièvement en quoi consiste **le modèle binomial**. Avec les notations adoptées, *X* et *Y* sont deux sous-ensembles aléatoires, indépendamment choisis dans l'ensemble des parties de E, respectivement de mêmes cardinaux n_a et n_b que les sous-ensembles de réalisations de *a* et de *b*. La valeur observée $n_{a\wedge\bar{b}}$ peut être considérée comme la réalisation d'une variable aléatoire Card ($X \cap \bar{Y}$) qui représente le nombre aléatoire de contre-exemples à l'inclusion de *X* dans *Y*, contre-exemples observés au cours de *n* tirages successifs indépendants. De là, Card ($X \cap \bar{Y}$) peut être considérée comme une variable binomiale de paramètres *n* et π où π est elle-même estimée par $p = \frac{n_a}{n}\frac{n_{\bar{b}}}{n}$. Ainsi :

$$\Pr\left[Card(X \cap \bar{Y}) = k\right] = C_n^k \left(\frac{n_a n_{\bar{b}}}{n^2}\right)^k \left(1 - \frac{n_a n_{\bar{b}}}{n^2}\right)^{n-k}$$

La variable centrée réduite estimée $Q(a,\bar{b})$ admet alors comme réalisation :

$$q(a,\bar{b}) = \frac{n_{a\wedge\bar{b}} - \frac{n_a n_{\bar{b}}}{n}}{\sqrt{\frac{n_a n_{\bar{b}}}{n}\left(1 - \frac{n_a n_{\bar{b}}}{n^2}\right)}}$$

Comme précédemment, nous obtenons l'intensité d'implication empirique estimée :

$$\varphi(a,b) = 1 - \Pr\left[Q(a,\bar{b}) \leq q(a,\bar{b})\right] = 1 - \sum_0^{n_{a\wedge\bar{b}}} C_n^k \left(\frac{n_a n_{\bar{b}}}{n^2}\right)^k \left(1 - \frac{n_a n_{\bar{b}}}{n^2}\right)^{n-k}$$

La loi de probabilité de $Q(a,\bar{b})$ peut être approchée par celle de la loi de Laplace-Gauss centrée réduite N(0,1). Généralement, l'intensité calculée dans le modèle de Poisson est plus « sévère » que l'intensité découlant du modèle binomial au sens où $\varphi(a,b)_{Poisson} \leq \varphi(a,b)_{binomiale}$

Remarque : Nous pouvons noter que l'indice d'implication est nul si et seulement les deux variables a et b sont indépendantes. En effet

$$q(a,\bar{b}) = \frac{n_{a\wedge\bar{b}} - \frac{n_a n_{\bar{b}}}{n}}{\sqrt{\frac{n_a n_{\bar{b}}}{n}(1-\frac{n_a n_{\bar{b}}}{n^2})}} = 0 \Leftrightarrow n_{a\wedge\bar{b}} - \frac{n_a n_{\bar{b}}}{n} = 0$$

$$q(a,\bar{b}) = 0 \Leftrightarrow n_{a\wedge\bar{b}} = \frac{n_a n_{\bar{b}}}{n} \quad \text{ou encore} \quad q(a,\bar{b}) = 0 \Leftrightarrow \frac{n_{a\wedge\bar{b}}}{n} = \frac{n_a}{n}\frac{n_{\bar{b}}}{n}$$

Cette dernière relation traduit la propriété d'indépendance statistique.

5.3 Modèle hypergéométrique.

Rappelons brièvement la 3ème modélisation proposée dans (Lerman et al. 1981) et (Gras et al 1996 c). Nous reprenons la même démarche : A et B sont les parties de E représentant les individus satisfaisant respectivement a et b et dont les cardinaux sont card (A)=n_a et card (B)=n_b. Puis considérons, deux parties aléatoires indépendantes X et Y telles que card (X)=n_a et card (Y)=n_b. La variable aléatoire Card ($A\cap\overline{Y}$) représente le nombre aléatoire d'éléments de *E* qui, étant dans A ne sont pas dans Y. Cette variable suit une loi hypergéométrique et l'on a pour tout $k \leq n_a$:

$$\Pr\left[Card(A\cap\overline{Y}) = k\right] = \frac{C_{n_a}^{k} C_{n-n_a}^{n-n_b-k}}{C_n^{n-n_b}} = \frac{n_a! n_{\bar{a}}! n_b! n_{\bar{b}}!}{k! n! (n_a-k)! (n_{\bar{b}}-k)! (n_b-n_a+k)!} =$$

$$\frac{C_{n-n_b}^{k} C_{n_b}^{n_a-k}}{C_n^{n_a}} = \Pr\left[Card(X\cap\overline{B}) = k\right]$$

Ceci montre, en échangeant le rôle de *a* et *b* que l'indice d'implication empirique $Q(a,\bar{b})$ correspondant à la quasi-règle a⇒b, est le même que celui correspondant à la réciproque soit $Q(b,\bar{a})$**.** Nous obtenons ainsi la même intensité pour la quasi-règle a⇒b et pour la quasi-règle réciproque b⇒a.

5.4 Choix des modèles pour évaluer l'intensité d'implication.

Si la modélisation binomiale reste compatible avec la sémantique de l'implication, relation binaire non symétrique, il n'en est plus de même pour la modélisation hypergéométrique puisqu'elle ne distingue pas la qualité d'une quasi-règle de celle de sa réciproque et présente un faible caractère pragmatique. En conséquence, nous ne retiendrons que le modèle de Poisson et le modèle binomial comme modèles adaptés à la sémantique de l'implication entre variables binaires.

La coexistence légitimée de trois modélisations différentes de notre problématique de mesure de qualité d'une quasi-règle n'est pas incohérente : elle tient au mode de prise en compte du tirage une à une de transactions (loi de Poisson) ou d'ensembles de transactions groupées (loi binomiale ou loi hypergéométrique). Par ailleurs, nous savons que, lorsque le nombre total de transactions devient très grand, les trois modèles convergent vers le même modèle gaussien. Dans (Lallich S. et al. 2005), on trouve, à titre de généralisation, une paramétrisation des trois indices obtenus par ces modélisations, qui permet d'évaluer l'intérêt des règles obtenues en les comparant à un seuil donné.

6 Quelques propriétés de l'indice d'implication et de l'intensité d'implication.

6.1 Stabilité de l'indice d'implication et de l'intensité d'implication

Le problème de la sensibilité aux faibles perturbations des paramètres en jeu, donc de la stabilité des indices de mesure de qualité des règles d'association se pose dès lors que les données sont susceptibles d'être bruitées. Trois méthodes nous semblent appropriées dans le but d'examiner cette sensibilité des règles, en particulier celles de la forme $a \Rightarrow b$ où a et b sont des variables observées sur un ensemble de sujets :

1- la simulation consistant à partir de fichiers plus ou moins artificiels à travers lesquels sont modifiés les paramètres intervenant dans la définition des indices (Gras et al. 2004) ;

2- la méthode du bootstrap consistant à effectuer des changements de certaines valeurs des paramètres, tout en conservant constantes certaines d'entre elles dont en particulier l'effectif de la population des individus ;

3- une méthode mathématique consistant à étudier, par l'analyse, les variations des paramètres en examinant leurs dérivées partielles et donc le gradient de l'indice global. (Gras 2005), (Lenca et al. 2006) et (Vaillant et al. 2006).

C'est cette dernière méthode que nous retenons ici. Nous porterons notre attention de façon privilégiée sur l'indice d'implication à la base de l'ASI et comparerons les résultats obtenus à ceux dérivant d'autres indices retenus pour des mesures de qualité de règles.

6.2 Analyse des variations de l'indice d'implication en fonction des cardinaux

Étudier la stabilité de l'indice d'implication q, revient à examiner ses petites variations au voisinage des 4 valeurs entières observées (n, n_a, n_b, $n_{a\wedge\bar{b}}$). Pour ce faire, il est possible d'effectuer différentes simulations en croisant ces 4 variables entières dont q dépend (Fleury, 1996, Gras et al., 2004). Mais, considérons ces variables comme nombres réels et q comme une fonction continûment différentiable par rapport à ces variables contraintes à respecter les inégalités : $0 \le n_a \le n_b$ et $n_{a\wedge\bar{b}} \le \inf\{n_a, n_b\}$ et $\sup\{n_a, n_b\} \le n$. Il suffit alors d'examiner la différentielle de q par rapport à ces variables et d'en conserver la restriction aux valeurs entières des paramètres de la relation $a \Rightarrow b$. La différentielle de q s'exprime de la façon suivante :

$$dq = \frac{\partial q}{\partial n} dn + \frac{\partial q}{\partial n_a} dn_a + \frac{\partial q}{\partial n_b} dn_b + \frac{\partial q}{\partial n_{a\wedge\bar{b}}} dn_{a\wedge\bar{b}}$$

Si l'on veut étudier comment varie q en fonction de $n_{\bar{b}}$, il suffit de remplacer n_b par $n-n_b$ et donc changer le signe de la dérivée de n_b dans la dérivée partielle. En fait, l'intérêt de cette différentielle réside dans l'estimation de l'accroissement (positif ou négatif) de q que nous notons Δq par rapport aux variations respectives Δn, Δn_a, Δn_b, $\Delta n_{\bar{b}}$ et $\Delta n_{a\wedge\bar{b}}$.

Si nous examinons le cas où seuls varient n_b et $n_{a\wedge\bar{b}}$, c'est à dire où les dérivées partielles de n et n_a sont nulles, on obtient alors :

$$\frac{\partial q}{\partial n_b} = \frac{1}{2} n_{a\wedge\bar{b}} (\frac{n_a}{n})^{-\frac{1}{2}} (n-n_b)^{-\frac{3}{2}} + \frac{1}{2} (\frac{n_a}{n})^{\frac{1}{2}} (n-n_b)^{-\frac{1}{2}} > 0$$

$$\frac{\partial q}{\partial n_{a\wedge\bar{b}}} = \frac{1}{\sqrt{\frac{n_a n_{\bar{b}}}{n}}} = \frac{1}{\sqrt{\frac{n_a (n-n_b)}{n}}} > 0$$. Ainsi, si les accroissements Δn_b et $\Delta n_{a\wedge\bar{b}}$ sont positifs, l'accroissement de $q(a,\bar{b})$ est également positif. Ceci s'interprète ainsi : si le nombre d'exemples de b et celui des contre-exemples de l'implication augmentent alors l'intensité d'implication diminue pour n et n_a constants. Autrement dit, cette intensité d'implication est maximum aux valeurs observées n_b et $n_{a\wedge\bar{b}}$ et minimum aux valeurs $n_b+\Delta n_b$. et $n_{a\wedge\bar{b}}+\Delta n_{a\wedge\bar{b}}$.

Si nous examinons le cas où seul n_a varie, nous obtenons la dérivée partielle de q par rapport à n_a est :

$$\frac{\partial q}{\partial n_a} = -\frac{1}{2\sqrt{n_{\bar{b}}/n}} \cdot \frac{n_{a\wedge\bar{b}}}{n} \cdot (\frac{n}{n_a})^{\frac{3}{2}} - \frac{1}{2}\sqrt{\frac{n_{\bar{b}}}{n_a}} < 0$$

Ainsi, sur [0, n_b], la fonction indice d'implication $q(a,\bar{b})$ est toujours décroissante par rapport à n_a et est donc minimum pour $n_a = n_b$. Par suite, l'intensité d'implication y est croissante et maximum pour $n_a = n_b$

6.3 Analyse des variations de l'indice d'implication en fonction des fréquences

On examine maintenant les variations de q en fonction des fréquences relatives des variables précédentes où le référentiel a pour cardinal n. Ainsi, on note $f_i = \frac{n_i}{n}$ chacune des fréquences des variables respectives n, n_a, $n_{\bar{b}}$ (que nous privilégions par rapport à n_b pour des raisons de calculs) et $n_{a\wedge\bar{b}}$. Dans ces conditions $q(a,\bar{b})$ s'écrit alors :

$$q(a,\bar{b}) = \sqrt{n}\frac{(f_{a\wedge\bar{b}} - f_a f_{\bar{b}})}{\sqrt{f_a f_{\bar{b}}}} = \sqrt{n}\frac{f_{a\wedge\bar{b}}}{\sqrt{f_a f_{\bar{b}}}} - \sqrt{n}\sqrt{f_a f_{\bar{b}}}$$

On étudie alors la stabilité à partir des dérivées partielles de q par rapport aux 4 variables fréquentielles :

$$dq = \frac{\partial q}{\partial n} dn + \frac{\partial q}{\partial f_{a\wedge\bar{b}}} df_{a\wedge\bar{b}} + \frac{\partial q}{\partial f_a} df_a + \frac{\partial q}{\partial f_{\bar{b}}} df_{\bar{b}} = \vec{grad}\, q. \begin{bmatrix} dn \\ df_{a\wedge\bar{b}} \\ df_a \\ df_{\bar{b}} \end{bmatrix}$$

Or les dérivées partielles respectives sont :

$$\frac{\partial q}{\partial n} = \frac{1}{2\sqrt{n}} \frac{f_{a\wedge\bar{b}} - f_a f_{\bar{b}}}{\sqrt{f_a f_{\bar{b}}}} = \frac{1}{2} \frac{f_{a\wedge\bar{b}} - f_a f_{\bar{b}}}{\sqrt{n f_a f_{\bar{b}}}} \; ;$$

$$\frac{\partial q}{\partial f_{a\wedge\bar{b}}} = \sqrt{n} \frac{1}{\sqrt{f_a f_{\bar{b}}}} > 0 \; ;$$

$$\frac{\partial q}{\partial f_a} = -\frac{\sqrt{n}}{2} \left[\frac{f_{a\wedge\bar{b}}}{\sqrt{f_{\bar{b}}}} f_a^{-\frac{3}{2}} + \sqrt{f_{\bar{b}}} f_a^{-\frac{1}{2}} \right] < 0$$

$$\frac{\partial q}{\partial f_{\bar{b}}} = -\frac{\sqrt{n}}{2} \left[\frac{f_{a\wedge\bar{b}}}{\sqrt{f_{\bar{a}}}} f_{\bar{b}}^{-\frac{3}{2}} + \sqrt{f_a} f_{\bar{b}}^{-\frac{1}{2}} \right] < 0$$

Nous pouvons remarquer qu'en calculant $\frac{\partial \mathrm{q}}{\partial \mathrm{f_b}}$ au lieu de $\frac{\partial \mathrm{q}}{\partial \mathrm{f_{\bar{b}}}}$, on constate que cette dérivée partielle est alors positive. En effet : $f_b = 1 - f_{\bar{b}}$ et $\frac{\partial q}{\partial f_b} = \frac{\partial q}{\partial f_{\bar{b}}} \frac{\partial f_{\bar{b}}}{\partial f_b} = -\frac{\partial q}{\partial f_{\bar{b}}}$

Par ailleurs, à n et n_a constants, la vitesse d'accroissement de q (en valeur absolue) quand s'accroît le nombre de contre-exemples $a \wedge \bar{b}$, est inversement proportionnelle à celle de la racine carrée de $n_{\bar{b}}$. Autrement dit : si $n_{\bar{b}}$ décroît, par exemple deux fois plus petit, cette vitesse est accélérée et multipliée par 2. Conséquence, l'intensité d'implication diminue et la qualité de l'implication devient moins bonne.

6.4 Examen d'autres indices d'implication

Contrairement à l'indice d'implication q de base et à l'intensité d'implication qui mesure la qualité à travers une probabilité, d'autres indices parmi les plus courants se veulent eux-mêmes directement des mesures de qualité . Nous examinons leurs sensibilités respectives aux variations des paramètres retenus dans la définition de ces indices. Nous conservons les notations adoptées et choisissons des indices qui sont rappelés dans (Lenca et al., 2004).

L'indice Lift

Il s'exprime par : $l = \frac{n.n_{a\wedge b}}{n_a n_b}$. Cette expression peut encore s'écrire pour mettre en évidence le nombre de contre-exemples : $l = \frac{n(n_a - n_{a\wedge\bar{b}})}{n_a.n_b}$.

Pour étudier la sensibilité de l aux variations des paramètres, nous formons :

$$\frac{\partial l}{\partial n_{a\wedge\bar{b}}} = -\frac{n}{n_a n_b}$$

Ainsi, la variation de l'indice Lift est indépendante de celle du nombre de contre-exemples. C'est une constante qui ne dépend que des variations des occurrences de a et de b. L'indice Lift l décroît donc lorsque le nombre de contre-exemples croît, ce qui

sémantiquement, est acceptable mais la vitesse de décroissance ne dépend pas de la vitesse de croissance de $n_{a \wedge \bar{b}}$.

L'indice m multiplicateur de cote

Cet indice d'implication s'exprime ainsi : $m = \frac{n_a - n_{a \wedge \bar{b}}}{n_b n_{a \wedge \bar{b}}} n_{\bar{b}}$ (Lallich et al, 2004). Remarquons qu'en étant indépendant de n, il n'a pas un sens statistique aussi intéressant. Sa dérivée partielle par rapport au nombre de contre-exemples est :

$$\frac{\partial m}{\partial n_{a \wedge \bar{b}}} = -\frac{n_a n_{\bar{b}}}{n_b} (\frac{1}{n_{a \wedge \bar{b}}})^2 .$$

L'indice m multiplicateur de cote décroît donc lorsque $n_{a \wedge \bar{b}}$ croît et la vitesse de décroissance est même plus rapide qu'avec l'indice Lift et qu'avec l'indice d'implication q de base dans l'intensité d'implication. Il ne résiste pas à l'instabilité du nombre de contre-exemples.

L'indice c, confiance

Cet indice c est le plus connu et, historiquement, après celui de J. Lovinger, le plus utilisé grâce à la caisse de résonance dont dispose une publication anglo-saxonne (Agrawal et al. 1993). Il est à l'origine de plusieurs autres indices communément employés qui n'en sont que des variantes satisfaisant telle ou telle exigence sémantique. De plus, il est simple et s'interprète aisément et immédiatement.

$$c = \frac{n_{a \wedge b}}{n_a} = 1 - \frac{n_{a \wedge \bar{b}}}{n_a}$$

Cet indice c s'interprète comme une fréquence conditionnelle des exemples de b quand a est connu. La sensibilité de cet indice aux variations des occurrences des contre-exemples se lit avec la dérivée partielle :

$$\frac{\partial c}{\partial n_{a \wedge \bar{b}}} = -\frac{1}{n_a} < 0$$

Par conséquent, la confiance c croît quand $n_{a \wedge \bar{b}}$ décroît ce qui est sémantiquement acceptable, mais la vitesse de variation est constante, indépendante de la vitesse de décroissance de cette quantité ainsi que des variations de n et de n_b. Le gradient de c ne s'exprime que par rapport à $n_{a \wedge \bar{b}}$ et à n_a. Ceci peut apparaître comme une restriction du rôle des paramètres dans l'expression de la sensibilité de l'indice.

6.5 Coefficient de corrélation linéaire et indice d'implication

La quasi-implication définie par l'indice d'implication $q(a, \bar{b})$ non symétrique ne coïncide pas avec le coefficient de corrélation $\rho(a,b)$ qui est symétrique et qui rend compte d'une liaison linéaire entre les variables a et b. En effet, nous démontrons la proposition suivante :

Proposition 1 si $\rho(a,b) \neq 0$ alors $\frac{q(a,\bar{b})}{\rho(a,b)} = -\sqrt{\frac{n_b n_{\bar{a}}}{n}}$

En effet, d'une part, $q(a,\bar{b}) = \frac{n_{a\wedge\bar{b}} - \frac{n_a n_{\bar{b}}}{n}}{\sqrt{\frac{n_a n_{\bar{b}}}{n}}} = \frac{n_a n_b - n n_{a\wedge b}}{\sqrt{n n_a n_{\bar{b}}}}$ et, d'autre part,

$\rho(a,b) = \frac{n n_{a\wedge b} - n_a n_b}{\sqrt{n_a n_b n_{\bar{a}} n_{\bar{b}}}} = \frac{n_{a\wedge b} n_{\bar{a}\wedge\bar{b}} - n_{a\wedge\bar{b}} n_{\bar{a}\wedge b}}{\sqrt{n_a n_b n_{\bar{a}} n_{\bar{b}}}}$ d'où la relation annoncée

Proposition 2 $q(a,\bar{b}) = 0 \Leftrightarrow \rho(a,b) = 0$ et $\rho(a,b) \geq 0 \Leftrightarrow \varphi(a,b) \geq 0{,}5$.

Ceci signifie que quasi-implication et corrélation linéaire vont plutôt "*dans le même sens*". Cependant, on peut observer une croissance de l'implication en même temps qu'une décroissance de la corrélation. Ce qui montre bien, qu'outre la dépendance aux effectifs n, n_a et n_b, l'expression du rapport $\frac{\rho}{q}$ indique la non-coïncidence des deux concepts et par conséquent une différence dans le sens de l'information apportée. L'ASI n'est pas une étude de la dépendance au sens statistique habituel où elle est essentiellement symétrique.

	b	$\bar{b}$	marge
a	82	18	100
$\bar{a}$	45	55	100
marge	127	73	4000

TAB. 1 – *exemple n°1*

L'indice d'implication $q(a,\bar{b}) = -3{,}06$

L'intensité d'implication est $\varphi(a,b) = 0{,}9994$

Le coefficient de corrélation linéaire $\rho(a,b)=0{,}3842$

	b	$\bar{b}$	marge
a	78	22	100
$\bar{a}$	49	51	100
marge	127	73	4000

TAB. 2 – *exemple n°2*

L'indice d'implication $q(a,\bar{b}) = -2{,}4000$

L'intensité d'implication est $\varphi(a,b) = 0{,}9931$

Le coefficient de corrélation linéaire $\rho(a,b)=0{,}3011$

On constate que, d'une part, a et b sont moins corrélées dans l'exemple 1 que dans le second, mais que, d'autre part l'intensité d'implication est plus forte dans le premier que dans le second cas..

6.6 Mesure du χ^2 d'indépendance et indice d'implication

La pratique fréquente du test d'indépendance à partir de tableaux 2x2 par la méthode du χ^2 nous conduit à montrer les apports respectifs des concepts de χ^2 d'indépendance et d'indice d'implication.

Dans la théorie des tests inférentiels, la mesure du χ^2 sert d'appui au raisonnement pour rejeter l'hypothèse d'indépendance entre deux variables a et b, à un seuil déterminé à l'avance. Dans la théorie de l'ASI, l'indice d'implication permet, non seulement d'assurer ce rejet, mais aussi, d'expliciter le **sens de la relation de dépendance**. Pragmatiquement, questionnés sur l'implication de la variable a sur b, certains chercheurs utilisent la mesure du χ^2 pour étudier et rejeter l'indépendance, puis constatant la faiblesse numérique de $n_{a \wedge \bar{b}}$, à partir du tableau de croisement de a et b, décident de retenir que a implique b. Or, par cette stratégie décisionnelle qualitative, prise à partir de nombres bruts, même à bon escient, ils négligent le rôle des valeurs relatives des nombres du tableau et ne peuvent, en outre, énoncer le niveau de qualité de l'implication conjecturée.

La proposition suivante met en évidence la relation fonctionnelle entre les deux concepts en jeu

Proposition 3 Considérons la mesure χ^2 et l'indice d'implication $q(a,\bar{b})$ entre les variables binaires a et b alors $\dfrac{\chi^2}{q(a,\bar{b})^2} = \dfrac{n^2}{n_a n_{\bar{b}}}$

En effet, $q(a,\bar{b}) = \dfrac{n_{a\wedge\bar{b}} - \dfrac{n_a n_{\bar{b}}}{n}}{\sqrt{\dfrac{n_a n_{\bar{b}}}{n}}} = \dfrac{n_a n_b - n n_{a\wedge b}}{\sqrt{n n_a n_{\bar{b}}}}$ et $\chi^2 = \dfrac{n(n_{a\wedge b} n_{\bar{a}\wedge\bar{b}} - n_{\bar{a}\wedge b} n_{a\wedge\bar{b}})^2}{n_a n_b n_{\bar{a}} n_{\bar{b}}}$

La démonstration est directe dès que l'on remarque que $\chi^2 = n\rho^2$. On constate ainsi que les deux concepts χ^2 et indice d'implication ne se superposent pas, ce qui, compte tenu de leur définition analytique n'a pas lieu de surprendre. Nous pouvons interpréter la valeur de l'indice d'implication $q(a,\bar{b})$ comme la contribution absolue de la case $(a,\bar{b})$ du tableau 2x2 croisant les deux variables binaires a et b au χ^2.

A. Totohasina (1992) a étudié de plus près les relations entre le χ^2 d'indépendance (ici à 1 degré de liberté) et l'indice d'implication. En particulier, les valeurs limites qui permettent de réfuter l'indépendance mutuelle entre deux variables, dans le cas où q est négatif, c'est-à-dire lorsque $q(a,\bar{b}) = -\sqrt{\dfrac{n_a n_{\bar{b}}}{n^2}\chi^2}$ peuvent éventuellement servir à définir les valeurs limites d'acceptabilité de la règle de quasi-implication de a sur b.

6.7 Mesure du χ^2 de Mac Nemar et indice d'implication

Si nous voulons comparer deux séries successives de données binaires de type présence-absence ou échec-réussite relevées sur le même échantillon d'individus comme cela est le cas

en ASI, nous pouvons aussi utiliser le test du χ^2 de **Mac Nemar**. Comme nous l'avons déjà présenté, l'information est alors résumée dans un tableau 2x2 dont nous donnons un exemple ci-dessous qui sera repris plus tard (Partie 2 Chap. 10). Pour rester congruent au mode de présentation des recherches de liens génériquement notés a⇒b dans le contexte ASI qui présuppose que N(a)≤N(b), nous représentons systématiquement la variable binaire a en ligne et la variable binaire b en colonne.

		Variable b = Épreuve Finale		
		1 = Réussite	0 = Échec	**Total**
Variable a = Épreuve Initiale	1 = Réussite	56	6	62
	0 = Échec	14	26	40
	Total	70	32	102

TAB. 3- *Tableau de contingence du type AVANT/APRES*

La question que nous nous posons alors est de savoir si les fréquences de réussite aux deux épreuves sont significativement différentes ou non.

L'idée de Mac Nemar pour étudier ce type de lien entre les deux épreuves est qu'il est plus pertinent de ne prendre en compte que les discordances entre les deux épreuves. Dans le tableau ci-dessus, ce sont les deux effectifs 14 et 6 correspondant aux couples (A_Echec, B_Réussite) et (A_Réussite, B_Echec) qui sont considérés comme des informations majeures. Cette idée n'est pas rendue par le test du χ^2 d'indépendance que nous avons déjà évoqué précédemment (Partie 1 Chap. 1-6.6) en établissant la relation algébrique entre l'indice d'implication et la mesure du χ^2.

Si nous nous remettons dans le contexte de l'ASI, le tableau de référence est donc celui-ci :

		Variable b		
		1	0	**Total**
Variable a	1	$n(a \wedge b)$	$n(a \wedge \bar{b})$	$n(a)$
	0	$n(\bar{a} \wedge b)$	$n(\bar{a} \wedge \bar{b})$	$n(\bar{a})$
	Total	$n(b)$	$n(\bar{b})$	n

TAB 4 - *Tableau de contingence avec les notations ASI*

Dans l'hypothèse d'une équivalence entre les deux épreuves, la fréquence de ceux qui sont passés de l'état 1 à l'état 0 parmi ceux qui ont changé d'état est égale à la fréquence de ceux qui sont passés de l'état 0 à l'état 1 parmi ceux qui ont changé d'état, c'est à dire égale à 0,5. D'une certaine manière, cela revient à comparer une fréquence observée à une fréquence théorique de 0,5.

Mac Nemar a montré qu'il suffisait de prendre comme indice, la mesure suivante que nous nommerons χ^2 de Mac Nemar, $\chi^2_{MacNemar} = \frac{(n(\bar{a} \wedge b) - n(a \wedge \bar{b}))^2}{n(\bar{a} \wedge b) + n(a \wedge \bar{b})}$ dont la loi de probabilité est approximativement celle de la variable de Pearson χ^2 de degré de liberté ddl=1.

Nous ne chercherons pas ici à expliciter une relation algébrique en $Q(a,\bar{b})$ et $\chi^2_{MacNemar}$.

Dans le cas présenté, nous calculons la valeur empirique comme suit $\chi^2_{MacNemar} = \frac{(14-6)^2}{14+6} = \frac{8^2}{20} = 3{,}2$ et nous la confrontons à la valeur critique au niveau de risque α. Si nous choisissons un niveau de risque de 0,05, la valeur critique est alors de 3,84. Comme 3,2<3,84, nous ne rejetons pas l'hypothèse d'équivalence des deux épreuves que nous considérons comme telle avec un risque de 2ème espèce β inconnu.

En résumé les 4 étapes de la démarche de ce test sont les suivantes :

- Étape 1 : formulation des hypothèses :

H_0 : symétrie des changements d'état entre les deux épreuves
H_1 : non-symétrie des changements d'état entre les deux épreuves

- Étape 2 : calcul de la valeur empirique du χ^2 (Mac Nemar)
- Étape 3 : lecture de la valeur critique dans la table du χ^2 de Pearson de ddl=1 pour un risque α donné
- Étape 4 : décision statistique rejet ou non rejet de H_0

Si nous revenons à la perspective de recherche de lien par rejet de l'indépendance en appliquant le test du χ^2 d'indépendance, nous trouvons une valeur empirique de 34,56 qui est très largement supérieure à la valeur critique de 3,84 pour un niveau de risque α=0,05 et même à la valeur critique 6,63 pour un niveau de risque α=0,01. Au sens du test du χ^2 d'indépendance, il existe donc un lien fort entre les deux variables.

Si nous nous plaçons dans la perspective de recherche de lien au sens de l'ASI, le calcul de l'intensité d'implication $\varphi_P(a,b)$ avec le modèle de Poisson et le calcul de l'intensité d'implication $\varphi_{BIN}(a,b)$ avec le modèle binomial

	χ^2 =34,56	χ^2_{MC} =3,2		Intensités d'implication
a \ b	b=1	b=0		$\varphi_P(a,b)$
a=1	56	6	62	0,9996
a=0	14	26	40	$\varphi_{BIN}(a,b)$
	70	32	102	0,9998

TAB. 4- *analyse selon les trois perspectives:* χ2 d'indépendance, χ2 Mac Nemar, ASI,

Les valeurs qui figurent dans le tableau ci-dessus, indiquent un niveau de confiance en l'implication statistique (a)⇒(b) supérieur à 0,99.

Face à ce qui semble paradoxal dans la mesure où le même tableau de contingence est susceptible d'être interprété de manière contradictoire, il y a tout lieu de considérer les logiques à l'œuvre dans ces trois approches : ASI, χ^2 de Mac Nemar, χ^2 d'indépendance.

Comme nous avons pu le voir au travers des propos tenus tout au long de ce qui précède, le raisonnement s'appuie sur un point de vue soutenu par I.-C. Lerman (1992) appliqué à l'étude d'une certaine relation de dépendance orientée entre des variables descriptives. Ce point de vue oppose la logique des tests statistiques, comme celui dit du χ^2 d'indépendance

ou encore celui du χ^2 de Mac Nemar, à celle des méthodes classificatoires de la manière suivante : pour les premiers, dit I.-C. Lerman, « relativement à l'existence d'un lien, on a FAUX, jusqu'à preuve du contraire » par le rejet de l'hypothèse nulle ; pour les secondes, « pour l'optique des données, on a VRAI, jusqu'à preuve du contraire », c'est-à-dire vrai selon une certaine échelle de probabilité du lien.

6.8 Indice de similarité et indice d'implication

Étudions maintenant la **relation** qui ne peut manquer d'exister entre **l'indice de similarité de I.-C. Lerman et l'indice d'implication** tel que nous le définissons dans le modèle de Poisson. Nous rappelons que l'indice de similarité poissonnien, défini sous la condition de la vraisemblance du lien tout comme l'indice d'implication, est donné par la formule :

$$s(a,b) = \frac{n_{a \wedge b} - \frac{n_a n_b}{n}}{\sqrt{\frac{n_a n_b}{n}}}$$

Proposition 4 : L'indice de similarité et l'indice d'implication sont liés par la relation suivante :

$$\frac{q(a,\bar{b})}{s(a,b)} = -\sqrt{\frac{n_b}{n_{\bar{b}}}} = -\sqrt{\frac{n_b}{n - n_b}}$$

On peut note que ce rapport ne dépend que de la réalisation de la variable b, de la réalisation concomitante de son contraire et non pas de celle de a. Par ailleurs nous pouvons aussi interpréter $s(a,b) = \frac{n_{a \wedge b} - \frac{n_a n_b}{n}}{\sqrt{\frac{n_a n_b}{n}}} = q(a,b)$ comme étant l'indice d'implication de $a \Rightarrow \bar{b}$. Ainsi, la similarité et l'implication vont, comme la corrélation et la dépendance, plutôt dans le même sens mais, bien entendu, ne coïncident pas. En effet, voici un exemple où la similarité ne change pas alors que l'implication varie très sensiblement:

	b	$\bar{b}$	marge
a	10	0	10
$\bar{a}$	70	20	90
marge	80	20	100

TAB. 5 – *exemple n°1*

L'indice d'implication $q(a,\bar{b})$ = -1,414

L'intensité d'implication φ(a,b)=0,864

L'indice de similarité est s(a,b) = 0,707

	b	$\bar{b}$	marge
a	3	7	10
$\bar{a}$	17	73	90
marge	20	80	100

TAB. 6– *exemple n°1*

L'indice d'implication $q(a,\bar{b})$ = -3,54

L'intensité d'implication φ(a,b)=0,547

L'indice de similarité est s(a,b) = 0,707

6.9 Comparaison avec d'autres approches de l'implication statistique.

Approche de J. Loevinger (1947)

L'approche de J. Loevinger fonde l'analyse de la quasi-implication de a sur b sur l'indice H qui prend ses valeurs sur tout]-∞,1] :

$$H(a,b) = 1 - \frac{n_{a\wedge\bar{b}}}{\frac{n_a n_{\bar{b}}}{n}} = 1 - \frac{n n_{a\wedge\bar{b}}}{n_a n_{\bar{b}}} = \frac{\frac{n_a n_{\bar{b}}}{n} - n_{a\wedge\bar{b}}}{\frac{n_a n_{\bar{b}}}{n}}$$

Si H(a,b) est *assez proche* de 1, la quasi-implication est *presque satisfaite*. Rappelons que cet indice, assez naturel il est vrai, avait été dans un premier temps "redécouvert" par R. Gras en 1978 à la suite de tâtonnements numériques, comme il l'a été également par A. Bodin en 1985 sous la forme suivante : notons x = P[B/A], probabilité conditionnelle de B sachant A et p = P[B]. Alors, H(a,b)= $\frac{x-p}{1-p}$. Mais cet indice présente l'inconvénient, en ne se référant pas à une échelle de probabilité, de ne pas fournir de seuil de vraisemblance et d'être invariant dans toute dilatation de E, A, B et A∩nonB.

Proposition 5 : L'indice H de Loevinger et l'indice d'implication sont liés par la relation suivante :

$$\frac{q(a,\overline{b})}{H(a,b)} = -\sqrt{\frac{n_a n_{\bar{b}}}{n}}$$

Dans l'approche de J. Pearl (1988), de S. Acid (Acid et *al*., 1991), de A. Gammerman et Z. Luo (Gammerman et Luo, 1991), c'est l'écart entre la distribution conjointe de a et b, et non pas celle de a et de non-b, et la distribution produit qui tient lieu de critère de comparaison. Cet écart est évalué par une expression de la forme :

$$\left|\Pr ob[A \cap \overline{B}] - \Pr ob[A]\Pr ob[\overline{B}]\right|$$

En ce qui concerne le système GENRED (Ralambrodrainy 1991) conçu dans une perspective de génération de règles d'inférence en intelligence artificielle, il est considéré tout simplement qu'une règle est pertinente dès lors que pour deux seuils α et β donnés par

l'utilisateur, le nombre de contre-exemples $card[A \cap \overline{B}]$ et celui des exemples $card[A \cap B]$ vérifient les conditions $card[A \cap \overline{B}] \leq \alpha$ et $card[A \cap B] \geq \beta$

Une autre façon comparable d'aborder cette question est (Sebag et Schoenauer, 1991) est de ne retenir qu'une règle au seuil α par une condition sur le rapport entre le nombre d'exemples et celui des contre-exemples : $\frac{card[A \cap B]}{card[A \cap \overline{B}]} \geq \alpha$ ou encore au travers d'une relation équivalente dans son principe :

$$\frac{card[A \cap B] - card[A \cap \overline{B}]}{card[A \cap B]} \geq \alpha$$

Comme nous l'avons déjà évoqué, la probabilité conditionnelle P[B/A] est aussi fréquemment utilisée comme indice de référence pour juger la plausibilité de la règle a=>b (Schektman, Trejos et Troupe, 1992) (Diday et Menessier, 1991). Par exemple, dans le domaine de l'apprentissage dans les bases de connaissances (Ganascia 1991), l'incertitude sur l'implication a=>b est évaluée par l'indice : $2\Pr ob[B|A] - 1$ et s'applique même aux classes de variables. Parmi les inconvénients, notons que ce dernier indice ne sépare pas, numériquement, deux implications dont l'une serait triviale et l'autre hautement informative.

Examinons deux situations-limites qui nous semblent probantes.

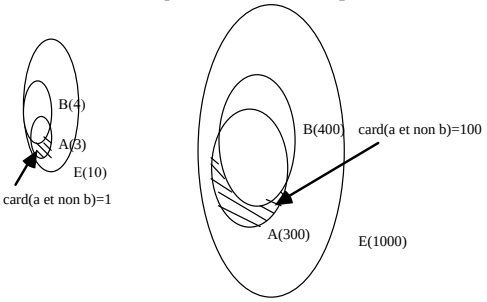


FIG. 2– *Deux situations cardinales*

Dans les deux cas, P[B/A]= 0,667, alors que, dans le premier cas, l'indice d'implication de a vers b est de l'ordre de $\varphi(a,b) \approx 0{,}7$, dans le second cas $\varphi(a,b)$ est très proche de 1. Le crédit que nous pouvons effectivement accorder, dans ce dernier cas, à une relation étroite et *invraisemblable* entre a et b y est nettement plus grand du fait de la taille des exemples la confirmant. Notre indice en rend compte alors que l'indice conditionnel ne change pas et ne souligne pas, de ce fait, la densité du phénomène.

7 Situation fondamentale et fondatrice de l'approche entropique.

Deux raisons nous ont conduits à améliorer le modèle formalisé par l'intensité d'implication dans l'approche classique:

- lorsque les tailles des ensembles d'individus traités augmentent, atteignant des effectifs de l'ordre du millier ou plus, l'intensité d'implication φ(a,b) a tendance à ne plus être suffisamment discriminante car ses valeurs peuvent être très voisines de 1, alors que l'inclusion dont elle cherche à modéliser la qualité, est loin d'être satisfaite. Ce phénomène a été déjà signalé par A. Bodin (1997) dont les travaux traitent avec des ensembles de grande taille d'élèves impliqués dans des enquêtes internationales;

- le modèle classique de la quasi-implication retient essentiellement la mesure de l'intensité de la quasi-règle $a \Rightarrow b$. Or comme nous l'avons abordé en introduction à propos de la causalité, la prise en compte concomitante de la contraposée de l'implication de $\bar{b} \Rightarrow \bar{a}$ est indispensable pour renforcer l'évaluation de la qualité suffisamment bonne de la relation de quasi-implication, voire quasi-causale, de a sur b[6]. En même temps, elle pourrait permettre de corriger la difficulté évoquée en relation à la taille des ensembles en jeu. En effet si A et B sont des ensembles de petite taille par rapport à E, leurs complémentaires seront importants et réciproquement.

8 Formalisation de la quasi-règle implicative dans l'approche entropique.

La solution[7] que nous apportons utilise à la fois l'intensité d'implication et un autre indice qui rend compte de la dissymétrie entre les situations S_1=(a et b) et S'_1=(a et non b) qui concerne la quasi-règle a⇒b ainsi que celle entre les situations S_2=(non a et non b) et S'_2 = (a et non b) qui concerne la quasi-règle contraposée. Notons que ce sont les mêmes instances qui contredisent la quasi-implication et sa contraposée. Les valeurs relatives de ces instances sont fondamentales dans notre approche.

8.1 Construction d'indice d'inclusion

Pour rendre compte de l'incertitude liée à un éventuel pari de l'appartenance à une des deux situations S_1 ou S'_1, (resp. S_2 ou S'_2), nous avons choisi le concept **d'entropie de Shannon** (1949).

Ainsi nous déterminons **l'entropie conditionnelle** relative à S_1 et S'_1 quand a est réalisée

$$H(b/a) = -\frac{n_{a\wedge b}}{n_a}\log_2\frac{n_{a\wedge b}}{n_a} - \frac{n_{a\wedge \bar{b}}}{n_a}\log_2\frac{n_{a\wedge \bar{b}}}{n_a}$$

puis l'entropie conditionnelle relative à S_2 et S'_2 lorsque non b est réalisée ou encore b n'est pas réalisée

[6] Ce phénomène est signalé par Y. Kodratoff dans son article publié dans les Actes du Colloque « Fouille dans les données par la méthode implicative », IUFM de Caen, juin 2000. Nous avons aussi abordé cette question du paradoxe de Hempel en Introduction de l'ouvrage

[7] J. Blanchard apporte dans (Blanchard J. et al. 2005) une réponse à ce problème par une mesure de « l'écart à l'équilibre ».

$$H(\bar{a}/\bar{b}) = -\frac{n_{a\wedge\bar{b}}}{n_{\bar{b}}}\log_2\frac{n_{a\wedge\bar{b}}}{n_{\bar{b}}} - \frac{n_{\bar{a}\wedge\bar{b}}}{n_{\bar{b}}}\log_2\frac{n_{\bar{a}\wedge\bar{b}}}{n_{\bar{b}}}$$

Ces entropies conditionnelles sont à valeurs dans [0,1] et devraient être simultanément faibles. En conséquence, les dissymétries entre les situations S_1 et S'_1 (resp. S_2 et S'_2) devraient être simultanément fortes si l'on souhaite disposer d'un bon critère d'inclusion de A dans B. En effet les entropies conditionnelles représentent l'**incertitude** moyenne des expériences qui consistent à observer si b (resp. non a) est réalisé lorsque l'on a observé a (resp. non b). Le complément à 1 de cette incertitude représente donc **l'information** moyenne recueillie par la réalisation de ces expériences. Plus cette information est importante, plus forte est la garantie de la qualité simultanée de l'implication et de sa contraposée. Nous devons maintenant adapter ce critère numérique entropique au modèle attendu dans les différentes situations cardinales.

Pour que le modèle ait la signification attendue, il doit satisfaire, selon nous, les contraintes épistémologiques suivantes :

1° il devra intégrer les valeurs de l'entropie et même, pour amplifier les contrastes, prendre le carré de ces valeurs ;

2° ce carré varie aussi de 0 à 1, pour rendre compte du déséquilibre, c'est-à-dire de l'inclusion en s'opposant à l'entropie, c'est-à-dire à l'incertitude, la valeur retenue sera le complément à 1 de son carré tant que le nombre de contre-exemples restera inférieur à la moitié des observations de a (resp. de non b). Si ce nombre dépasse la moitié, , nous affecterons la valeur 0 au critère compte tenu du fait que sémantiquement les implications perdent leur sens inclusif ;

3° afin de prendre en compte les deux informations propres à la quasi-implication et à sa contraposée, c'est le produit des valeurs que nous retiendrons. Le produit a la propriété de s'annuler dès que l'un de ses termes s'annule, i.e. dès que cette qualité s'efface ;

4° enfin, le produit ayant une dimension 4 par rapport à l'entropie, nous prendrons sa racine quatrième pour revenir à la même dimension.

Posons $\alpha = \frac{n_a}{n}$ la fréquence de a, $\overline{\beta} = \frac{n_{\bar{b}}}{n}$ la fréquence de non b et $t = \frac{n_{a\wedge\bar{b}}}{n}$ la fréquence des contre-exemples. Nous construisons alors deux fonctions h_1 et h_2 définies dans [0 ;1] comme suit :

$$h_1(t) = H(b/a) = -(1-\frac{t}{\alpha})\log_2(1-\frac{t}{\alpha}) - \frac{t}{\alpha}\log_2\frac{t}{\alpha}$$ si $t\in[0,\frac{\alpha}{2}[$ et $h_1(t) = 1$ si $t\in[\frac{\alpha}{2}, \alpha]$

$$h_2(t) = H(\bar{a}/\bar{b}) = -(1-\frac{t}{\overline{\beta}})\log_2(1-\frac{t}{\overline{\beta}}) - \frac{t}{\overline{\beta}}\log_2\frac{t}{\overline{\beta}}$$ si $t\in[0,\frac{\overline{\beta}}{2}[$ et $h_2(t) = 1$ si $t\in[\frac{\overline{\beta}}{2}, \overline{\beta}]$

De là, nous proposons la définition suivante permettant de déterminer le critère entropique.

Définition 4: **L' indice d'inclusion** de A, support de a, dans B, support de b, est le nombre :

$$i(a,b) = \left(\left[1-h_1^2(t)\right]\left[1-h_2^2(t)\right]\right)^{\frac{1}{4}}$$

qui intègre l'information délivrée par la réalisation du nombre de contre-exemples, d'une part à la quasi-règle $a \Rightarrow b$ et, d'autre part, à la quasi-règle $\bar{b} \Rightarrow \bar{a}$

8.2 Construction d'un indice d'implication-inclusion

Pour prendre en compte à la fois, ce que nous avons appelé l'étonnement statistique et la qualité de l'inclusion de l'ensemble A des instances de a dans celui B des instances de b, nous proposons une nouvelle mesure de la qualité inductive suivante que nous nommons intensité d'implication-inclusion ou intensité entropique.

Définition 5: **L'intensité d'implication-inclusion ou intensité entropique**, est le nombre suivant:

$$\psi(a,b) = [i(a,b)\ \varphi(a,b)]^{1/2}$$

La fonction ψ de la variable t admet une représentation qui a la forme indiquée par la figure ci-dessous, pour n_a et n_b fixés. On remarquera sur cette figure, la différence de comportement de la fonction ψ par rapport à la probabilité conditionnelle P(B/A), indice fondamental des autres modélisations de la mesure des règles, comme par exemple chez Agrawal et son école. Outre son caractère linéaire, donc peu nuancé, cette dernière probabilité conduit à une mesure qui décroît trop vite dès les premiers contre-exemples et résiste ensuite trop longuement lorsque ceux-ci apparaissent en nombre important.

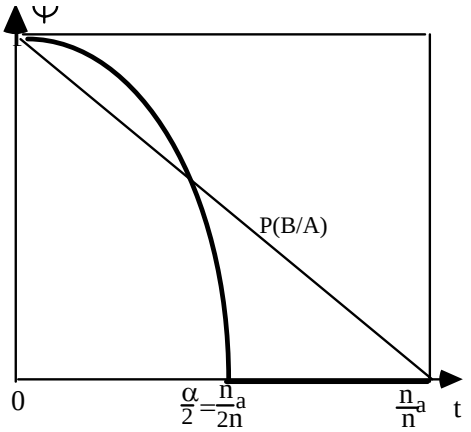


FIG. 3- *représentation graphique de la fonction ψ*

On constate que cette représentation de la fonction ψ continue de t traduit les propriétés attendues du critère d'inclusion :

1- réaction lente aux premiers contre-exemples (résistance au bruit),

2- accélération du rejet de l'inclusion au voisinage de l'équilibre soit $\frac{n_a}{2n}$,

3- rejet au-delà de $\frac{n_a}{2n}$ ce que n'assurait pas à elle seule, l'intensité d'implication φ(a,b).

8.3 Exploration succincte des propriétés de ψ

	b	$\bar{b}$	marge
a	200	400	600
$\bar{a}$	600	2800	3400
marge	800	3200	4000

TAB. 7 – *exemple n°1*

L'indice d'implication $q(a,\bar{b}) = -3,65$ (modèle de Poisson)

L'intensité d'implication est φ(a,b) = 0,999869

L'indice d'implication $q(a,\bar{b}) = -3,89$ (modèle binomial)

L'intensité d'implication est φ(a,b) = 0,999950

Les valeurs entropiques de l'expérience sont $h_1 = 1$ et $h_2 = 1$

La valeur du coefficient modérateur est donc : i(a,b) = 0

Par suite ψ(a,b) = 0 alors que P(B/A) = 0,33333

Ainsi, les fonctions entropiques modèrent l'intensité d'implication dans ce cas où justement l'inclusion est médiocre.

	b	$\bar{b}$	marge
a	400	200	600
$\bar{a}$	1000	2400	3400
marge	1400	2600	4000

TAB. 8 – *exemple n°2*

L'intensité d'implication est φ(a,b) = 1 pour un indice d'implication $q(a,\bar{b}) = -9,621$ (modèle de Poisson)

L'intensité d'implication est φ(a,b) = 1 pour un indice d'implication $q(a,\bar{b}) = -10,127$ (modèle binomial)

Les valeurs entropiques de l'expérience sont $h_1 = 0,918$ et $h_2 = 0,391$

La valeur du coefficient modérateur est donc : i(a,b) = 0,6036

Par suite ψ(a,b) = 0,777 alors que P(B/A) = 0,66666

	b	$\bar{b}$	marge
a	40	20	60
$\bar{a}$	100	240	340
marge	140	260	400

TAB. 9 – *exemple n°3*

L'intensité d'implication est φ(a,b) = 0,9988 pour un indice d'implication $q(a,\bar{b}) = -3,04$ (modèle de Poisson)

L'intensité d'implication est φ(a,b) = 0,9993 pour un indice d'implication $q(a,\bar{b})$ = -3,20 (modèle binomial)

Les valeurs entropiques de l'expérience sont h_1 = 0,918 et h_2 = 0,391

La valeur du coefficient modérateur est donc : i(a,b) = 0,603

Par suite ψ(a,b) = 0,776 alors que P(B/A) = 0,66666

Ainsi, φ(a,b) a diminué du 2ème au 3ème exemples, puisque le cardinal de l'ensemble de référence E a crû dans l'homothétie cardinale de rapport 1/10. Mais i(a,b) a augmenté de même que ψ(a,b).

Notons que, dans les deux cas, la probabilité conditionnelle ne change pas.

Pour plus de précisions, nous renvoyons à (Lenca et al, 2004) pour une étude comparative, très fouillée, des indices d'association pour des variables binaires. En particulier, les intensités d'implication classique et entropique (inclusion) présentées y sont confrontées à d'autres indices selon une entrée « utilisateur ».

Chapitre 2 : Représentation des règles d'implication et graphe implicatif

1 Problématique

A l'issue des calculs des intensités d'implication, que ce soit dans le modèle classique ou celui entropique, nous disposons d'un tableau p x p qui croise les p variables entre elles, quelle que soit leur nature, et dont les éléments sont les valeurs de ces intensités d'implication, nombres de l'intervalle [0 ;1]. Force est de constater que la structure sous-jacente de l'ensemble de ces variables est loin d'être explicite et demeure largement inapparente. L'utilisateur reste aveugle face à un tel tableau carré de taille p^2. Il ne peut embrasser simultanément les multiples enchaînements éventuels des règles qui sous-tendent la structure globale de l'ensemble des p variables. Afin de faciliter une extraction plus claire des règles et d'en examiner leur structure, nous avons associé à ce tableau, et pour un seuil d'intensité donné, un **graphe implicatif,** orienté, pondéré par les intensités d'implication, sans cycle dont l'utilisateur peut contrôler la complexité de la représentation en fixant lui-même le seuil de prise en compte de la qualité implicative des règles. Chaque arc de ce graphe représente une règle : si $n_a < n_b$, l'arc $a \rightarrow b$ représente la règle $a \Rightarrow b$; si $n_a = n_b$, alors l'arc $a \leftrightarrow b$ représentera la double règle $a \Leftrightarrow b$, en d'autres mots, l'équivalence entre ces deux variables. En faisant varier le seuil d'intensité d'implication, il est évident que le nombre d'arcs varie dans le sens opposé : pour un seuil fixé à 0,95, le nombre d'arcs est inférieur ou égal à ceux qui constitueraient le graphe au seuil 0, 90. Nous en reparlerons plus loin.

2 Algorithme

La relation définie par l'implication statistique, si elle est réflexive et non symétrique, donc sans cycle, n'est pas **transitive** bien évidemment, comme l'induction et au contraire de la déduction. Or nous voulons qu'elle modélise la relation d'ordre partiel entre deux variables. Ainsi en est-il de la question des réussites évoquées dans notre exemple fondamental.

Proposition 6 : Par convention, si $a \Rightarrow b$ et si $b \Rightarrow c$, il y a **fermeture transitive** $a \Rightarrow c$ si et seulement si $\varphi(a,c) \geq 0,5$, c'est-à-dire si la relation implicative de a sur c, qui traduit une certaine dépendance entre a et c, est meilleure que sa réfutation. Notons que, pour tout couple de variables (x ; y), l'arc $x \rightarrow y$ est pondéré par l'intensité d'implication $\varphi(x,y)$.

Prenons un exemple formel en supposant qu'entre les 7 variables *a, b, c, d, e f et g* existent, au seuil supérieur à 0,5, les règles suivantes :

$e \Rightarrow c \quad e \Rightarrow a \quad e \Rightarrow f \quad e \Rightarrow b$
$c \Rightarrow a \quad c \Rightarrow f$
$b \Rightarrow a \quad b \Rightarrow f$
$g \Rightarrow d \quad g \Rightarrow f$
$a \Rightarrow f$

On pourra alors traduire cet ensemble de relations par le graphe suivant[8] :

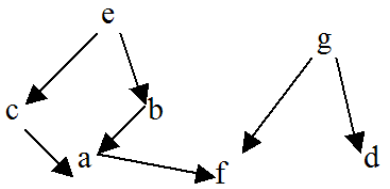


FIG. 4- *Un exemple de graphe implicatif*

Une des difficultés liées à la représentation graphique tient au fait que le graphe n'est pas planaire. L'algorithme qui en permet la construction, doit le prendre en compte et, en particulier, se doit de « décroiser » les chemins du graphe afin d'en permettre une lisibilité acceptable pour l'expert qui va l'analyser.

Le nombre d'arcs du graphe peut être réduit (*resp*. augmenté) si nous élevons (*resp.* abaissons) le seuil d'acceptation des règles, le niveau de confiance en les règles retenues. Corrélativement, des arcs peuvent apparaître ou disparaître selon les variations du seuil. Rappelons que ce graphe est nécessairement sans cycle, qu'il n'est pas un treillis puisque, par exemple la variable a n'implique pas la variable (a ou non a) dont le support est E. A fortiori, il ne peut être un treillis de Galois. Des options du logiciel C.H.I.C. de traitement automatique des données sous l'A.S.I., comme nous le verrons (Partie 2, chap. 11), permettent de supprimer à volonté des variables, de déplacer leur image dans le graphe afin de décroiser les arcs ou de se centrer sur certaines variables dites sommets d'une sorte de « **cône » dont les deux « nappes** » sont constituées respectivement des variables « **parents** » et des variables « **enfants** » de cette variable sommet (Partie 2, chap. 11). Nous désignons les extrémités des arcs par le terme « **nœuds** ». A un nœud d'un graphe donné, correspond une et une seule variable ou une conjonction de variables. Le passage d'un nœud S1 à un nœud S2 est appelé aussi « transition » qui est représentée par un arc du graphe.

3 Un exemple numérique à des fins didactiques

Nous partirons de la situation d'un tableau simulé joint ci-dessous croisant 30 individus (de i_1 à i_30) et 5 variables binaires (V1, V2, V3, V4 et V5). Deux autres variables (F s et H s) seront prises en compte comme variables supplémentaires que nous aborderons plus loin (Partie 1 Chap. 5). Ce tableau constitue un fichier dénommé RAF.

[8] Les traitements automatiques des calculs et des graphiques sont exécutés à l'aide du logiciel C.H.I.C. (acronyme de **C**lassification **H**iérarchique **I**mplicative et **C**ohésitive) disponible sous Windows 95, 98, NT, XP et Vista. Ce logiciel, à partir d'une première version établie par R. Gras et H. Rostam, révisée sous Pascal par S. Ag Almouloud, (Ag Almouloud S. 1992), est maintenant développé par R. Couturier (Couturier et Gras. 2005), constamment étendu par lui aux nouveaux concepts et nouveaux algorithmes et entretenu pour sa convivialité

	V1	V2	V3	V4	V5	F s	G s
i1	1	1	1	0	0	1	0
i2	1	1	1	0	0	1	0
i3	1	1	1	0	0	1	0
i4	1	1	1	0	0	1	0
i5	1	1	1	0	0	1	0
i6	1	1	1	1	0	1	0
i7	1	1	0	1	1	1	0
i8	1	1	0	1	1	1	0
i9	1	1	1	1	1	1	0
i10	1	1	1	1	1	1	0
i11	1	0	1	1	0	1	0
i12	1	0	1	1	0	1	0
i13	1	0	1	1	0	1	0
i14	1	0	1	1	1	0	1
i15	1	0	1	1	0	1	0

	V1	V2	V3	V4	V5	F s	G s
i16	1	0	1	1	1	0	1
i17	1	0	1	1	1	0	1
i18	1	0	1	1	0	0	1
i19	1	0	1	1	0	0	1
i20	1	0	1	0	1	1	0
i21	1	0	1	1	0	0	1
i22	1	0	1	0	0	0	1
i23	1	0	0	1	1	0	1
i24	1	0	0	0	0	0	1
i25	0	0	0	0	0	1	0
i26	0	0	0	0	0	0	1
i27	0	0	0	0	0	0	1
i28	0	0	0	1	0	0	1
i29	0	0	1	1	1	0	1
i30	0	0	0	1	0	0	1

TAB. 10- *tableau de données du fichier Raf*

Afin de prendre pleine conscience des concepts de l'implication statistique définis précédemment (Partie 1 Chap. 1) nous proposons la situation suivante dans laquelle les calculs seront réalisés une calculatrice ou un tableur :

- Étape n°1 : calculer les indices d'implication des couples de variables puis les intensités d'implication en se plaçant selon le modèle de Poisson puis dans le modèle recourant à l'approximation gaussienne de la loi de la variable $Q(a,\bar{b})$
- Étape n°2 : réaliser la représentation graphique du graphe des règles.

Pour ne pas avoir à retourner aux formules déjà présentées (Partie 1 Chap. 1), pour évaluer la quasi-implication a⇒b, nous les redonnons ici. En utilisant le modèle de Poisson de paramètre estimé $\hat{\lambda} = \frac{n_{a\wedge\bar{b}}}{n}$, l'intensité d'implication est calculée ainsi :

$$\varphi(a,b) = 1 - \sum_{s=0}^{n_{a\wedge\bar{b}}} \frac{\hat{\lambda}^s}{s!}.e^{-\hat{\lambda}} = 1 - e^{-\hat{\lambda}}(\hat{\lambda} + \frac{\hat{\lambda}^2}{2} + ... + \frac{\hat{\lambda}^{n_{a\wedge\bar{b}}}}{(^{n}{}_{a\wedge\bar{b}})!})$$

En utilisant l'algorithme de calcul de l'intensité d'implication par l'approximation gaussienne, nous obtenons :

En posant : $q(a,\bar{b}) = \dfrac{n_{a\wedge\bar{b}} - \dfrac{n_{a\wedge\bar{b}}}{n}}{\sqrt{\dfrac{n_{a\wedge\bar{b}}}{n}}}$, on a $\varphi(a,b) = \dfrac{1}{\sqrt{2\pi}} \int_{q(a,\bar{b})}^{\infty} e^{-\frac{t^2}{2}} dt$

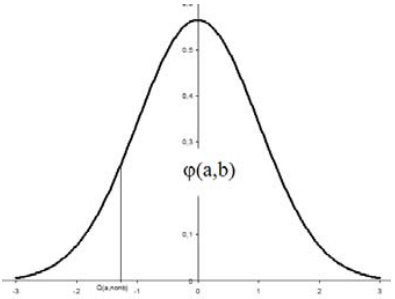


FIG. 5- *Image de l'intensité d'implication dans le cas du calcul fondé sur l'approximation gaussienne*

Il nous faut établir les 25 croisements entre les 5 variables binaires dont nous pouvons réduire les résultats en 10 tableaux suivants. Nous y avons indiqué les valeurs respectives du χ^2 dont seules celles correspondant aux tableaux V3-V1, V2-V3 indiquent une liaison significative au seuil de risque α=0,05 pour lequel la valeur critique est de 3,84 . La situation du tableau V2-V1 est proche de la liaison significative avec une valeur empirique de 3,75. Abordé d'une autre façon, nous pourrions préciser que dans ce cas la p-value serait de p=0,052. Comme nous avons pu le voir progressivement au travers des propos tenus tout au long de ce qui précède, notre perspective s'appuie sur un point de vue soutenu par I.-C. Lerman (1992) appliqué à l'étude d'une certaine relation de dépendance orientée entre des variables descriptives. Ce point de vue oppose les tests d'indépendance, comme celui bien connu dit du χ^2, aux méthodes classificatoires de la manière suivante : rappelons-le, pour les premiers, dit I.-C. Lerman, « on a, relativement à l'existence d'un lien, FAUX, jusqu'à preuve du contraire » par le rejet de l'hypothèse nulle ; pour les secondes, on a VRAI, jusqu'à preuve du contraire », c'est-à-dire vrai selon une certaine échelle de probabilité. Cette posture permet de lever une observation qui aurait pu paraître paradoxale.

Nous avons présenté les tableaux de façon à ce que la variable-ligne corresponde à la variable générique a et la variable-colonne, à la variable b, avec le respect de la contrainte $n_a \leq n_b$. Nous portons dans chaque tableau, la valeur de l'indice d'implication Q= $Q(a,\bar{b})$, les valeurs des intensités d'implication $\varphi_P(a,b)$ et $\varphi_{LG}(a,b)$ correspondant respectivement au modèle de Poisson et au calcul par l'approximation gaussienne. Pour donner plus de précision sur la mise en œuvre des algorithmes de calcul, nous développons le cas du tableau V2-V1.

D'une part $n_{V2} = 10 < n_{V1} = 24$ et $n_{V2\wedge\overline{V1}} = 0$ ce qui conduit à :

$$Q(V2,\overline{V1}) = \frac{n_{V2\wedge\overline{V1}} - \frac{n_{V2}n_{\overline{V1}}}{n}}{\sqrt{\frac{n_{V2}n_{\overline{V1}}}{n}}} = \frac{0-\frac{10x6}{30}}{\sqrt{\frac{10x6}{30}}} = -\sqrt{\frac{10x6}{30}} = -\sqrt{2} \approx -1{,}414$$

D'autre part $\hat{\lambda} = \frac{n_{V2}n_{\overline{V1}}}{n} = 2$ et $\varphi_P(V2,V1) = 1 - \sum_{s=0}^{0} \frac{\hat{\lambda}^s}{s!}.e^{-\hat{\lambda}} = 1 - e^{-2} \approx 0{,}86466472$

Si nous utilisons l'approximation gaussienne, nous obtenons

$$\varphi_{LG}(V2,V1) = \frac{1}{\sqrt{2\pi}} \int_{-\sqrt{2}}^{\infty} e^{-\frac{t^2}{2}} dt \approx 0{,}921319$$

χ^2 = 3,75		Q= -1,41			χ^2 = 1,14		Q= 0,696		
V2-V1	V1oui	V1non		φ_P(V2,V1)	V2-V4	V4oui	V4non		φ_P(V2,V4)
V2oui	10	0	10	**0,864**	V2oui	5	5	10	0,165
V2non	14	6	20	φ_{LG}(V2,V1)	V2non	14	6	20	φ_{LG}(V2,V4)
	24	6	30	0,921		19	11	30	0,243
χ^2 = 10,15		Q= -1,56			χ^2 = 0,3		Q= -0,258		
V3-V1	V1oui	V1non		φ_P(V3,V1)	V2-V5	V5oui	V5non		φ_P(V2,V5)
V3oui	20	1	21	**0,922**	V2oui	4	6	10	0,499
V3non	4	5	9	φ_{LG}(V3,V1)	V2non	6	14	20	φ_{LG}(V2,V5)
	24	6	30	0 ,940		10	20	30	0,60
χ^2 = 0,574		Q= -0,41			χ^2 = 0,334		Q= -0,293		
V4-V1	V1oui	V1non		φ_P(V4,V1)	V4-V3	V3oui	V3non		φ_P(V4,V3)
V4oui	16	3	19	0,526	V4oui	14	5	19	0,505
V4non	8	3	11	φ_{LG}(V4,V1)	V4non	7	4	11	φ_{LG}(V4,V3)
	24	6	30	0,659		21	9	30	0,615
χ^2 = 0,937		Q= -0,70			χ^2 = 0		Q= 0		
V5-V1	V1oui	V1non		φ_P(V5,V1)	V3-V5	V5oui	V5non		φ_P(V3,V5)
V5oui	9	1	10	0,593	V3oui	7	3	10	0 , 352
V5non	15	5	20	φ_{LG}(V5,V1)	V3non	14	6	20	φ_{LG}(V3,V5)
	24	6	30	0,760		21	9	30	0,5
χ^2 = 0,714		Q= - 0,57			χ^2 = 4,59		Q= -1,39		
V2-V3	V3oui	V3non		φ_P(V2,V3)	V5-V4	V4oui	V4non		φ_P(V5,V4)
V2oui	8	2	10	**0,576**	V5oui	9	1	10	0,880
V2non	13	7	20	φ_{LG}(V2,V3)	V5non	10	10	10	φ_{LG}(V4,V5)
	21	9	30	0,718		19	11	30	0,918

TAB. 11

On remarquera que les valeurs de l'intensité données par l'approximation gaussienne sont plus élevées que celles données par la valeur « vraie » calculée dans le modèle de Poisson.

Mais comme l'ordre entre les valeurs est inchangé, ce qui est essentiel, le graphe correspondant reste le même quel que soit le modèle choisi.

En fonction du seuil choisi avec le modèle de Poisson, on obtient deux graphes différents.

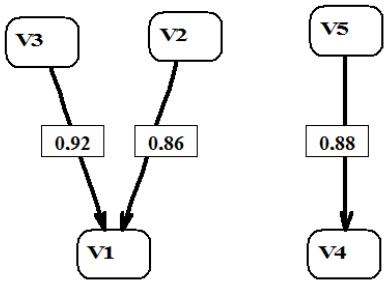


FIG. 6- *Graphe implicatif au seuil de confiance de 0,85 (Intensité calculée directement dans le modèle de Poisson)*

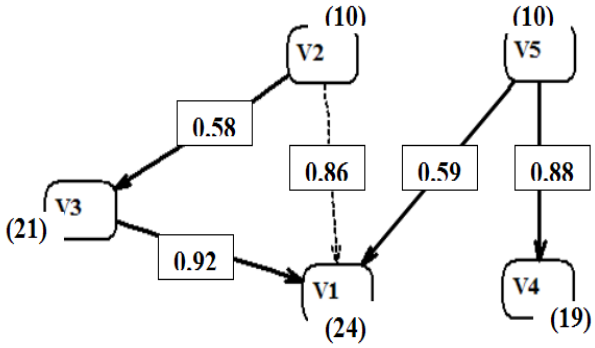


FIG. 7- *Graphe implicatif au seuil de confiance de 0,57 (Intensité calculée directement dans le modèle de Poisson)*

Modifiant le seuil à 0,57, la liaison entre V2 et V3 peut s'établir et permettre de créer un chemin transitif entre V2 et V1. L'arc V2 → V1 pourrait même être supprimé car il devient redondant en raison de l'information V2 → V3. C'est pourquoi nous l'avons mis en ligne pointillée.

En fonction de ce seuil choisi 0,57 et en procédant au calcul de l'intensité d'implication en recourant à l'approximation gaussienne, on obtient le graphe suivant.

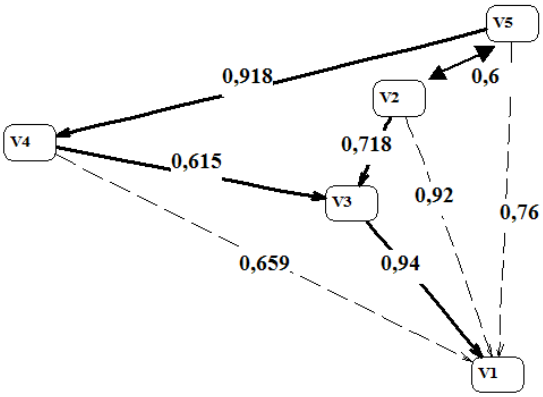


FIG. 8- *Graphe implicatif au seuil de confiance de 0,57 (Intensité calculée avec l'approximation gaussienne)*

Nous observons l'apparition d'une équivalence statistique entre V2 et V5 repérée par l'arc double V2↔V5. Par ailleurs nous avons marqué les fermetures transitives par les flèches en pointillé lesquelles pourraient être retirées pour alléger la représentation du graphe comme nous le présentons ci-dessous :

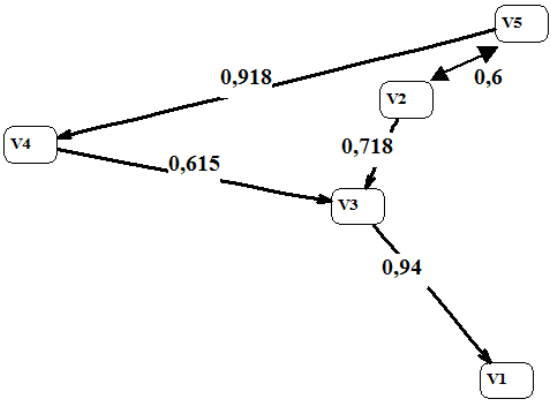


FIG. 9- *Graphe implicatif au seuil de confiance de 0,57 sans les fermetures transitives)*

4 Question sur la transition d'un nœud d'un graphe implicatif à un suivant

Nous avons déjà abordé précédemment (Partie 1, Chap. 1 – 6.2, 6.3) quelques propriétés de l'indice d'implication et de l'intensité d'implication. Nous reprenons ici l'étude de leurs

variations en fonction des variables effectifs (n, n_a, n_b, $n_{\bar{b}}$, $n_{a\wedge\bar{b}}$) ou plutôt des variables fréquences (f_a, f_b, $f_{\bar{b}}$, $f_{a\wedge\bar{b}}$). Ce problème de contrôle des variations de l'indice et de l'intensité d'implication se présente lors de fouille de règles d'association généralisées (cf. Chap 4) ou encore pour l'alignement de deux ontologies[9] (David J. et al., 2006) et (Gras et al. 2007 a).

Pour illustrer l'étude visée et lui justifier son sens, partons de l'exemple suivant où 3 variables attributs binaires (d'un repas) sont en jeu : a = {manger des huîtres} ; b_1 = {boire du vin blanc} ; b_2 = {boire du muscadet}.

En général, au cours d'un repas, on observe la règle au sens de l'ASI : $r_1 = [a \Rightarrow b_1]$, mais aussi la règle $r_2 = [a \Rightarrow b_2]$. Dans une taxonomie (ou **ontologie**) des vins, on a $b_2 \Rightarrow b_1$ parce que « muscadet » est plus spécifique que « vin blanc ». Cette dernière règle est stricte au sens de la logique mathématique. On remarque aisément que la règle r_1 est plus générale (en terme de support la réalisant) que la règle r_2. On peut ainsi se demander si la règle r_2 apporte plus d'information prédictive que r_1, autrement dit si r_2 apporte un réel gain de force ou d'intensité implicative que celle attendue par rapport à r_1. Pour cela nous proposons d'étudier la variation d'intensité d'implication entre r_1 et r_2.

On voit alors que le problème peut se généraliser dans la recherche d'élimination de redondances ou de superfluité de la façon suivante : sachant que $a \Rightarrow b_i$ pour un même attribut a et un certain nombre d'attributs spécifiques b_i, n'existe-t-il pas une règle $a \Rightarrow b_0$ qui conduise à une intensité d'implication optimale parmi les règles découlant de taxonomies T_i et qui permette l'élimination de règles moins riches en information inductive ?

4.1 Variations de l'indice d'implication q

Dans les hypothèses ci-dessus, les occurrences n de la population et les occurrences n_a de l'attribut a étant fixées, étudions les conditions de gain quand on élimine la règle r_1 au profit de la règle r_2 en fonction des paramètres associées. Désignons par b_1 (resp. b_2) les extrémités (ou nœuds) des arcs représentant respectivement ces deux règles. Supposons, comme dans l'exemple, que n_b varie en décroissant selon que l'on échange b_1 en b_2, Autrement dit, $n_{b1} > n_{b2}$. En conséquence, dans le même temps et nécessairement, le nombre de contre-exemples $n_{a\wedge\bar{b}}$ croît en passant de b_1 à b_2.

Nous voulons alors comparer les deux états contingents b_1 et b_2 associés respectivement aux règles r1 et r_2. Pour ce faire, il suffit de calculer les dérivées partielles par rapport aux fréquences variables $f_{\bar{b}}$ et $f_{a\wedge\bar{b}}$ au nœud b_1.

$$\Delta q(b_1; b_2) \approx \frac{\partial q}{\partial f_{a\wedge\bar{b}}}(b_1)\Delta f_{a\wedge\bar{b}} + \frac{\partial q}{\partial f_{\bar{b}}}(b_1)\Delta f_{\bar{b}} > 0$$

avec $\Delta f_{a\wedge\bar{b}} = f_{a\wedge\bar{b}_2} - f_{a\wedge\bar{b}_1} > 0$ et $\Delta f_{\bar{b}} = f_{\bar{b}_2} - f_{\bar{b}_1} > 0$

[9] De façon raccourcie, une **ontologie** est une théorie logique qui décrit des termes (concepts) d'un domaine et leurs propriétés (relations entre concepts). Une ontologie peut être plus ou moins formelle (et exprimée de manière plus ou moins expressive) Souvent ces concepts sont organisés en taxonomie (du concept le plus général au plus spécifique).

et pour $k \in \{1;2\}, q(b_k) = q(a, \overline{b}_k) = \sqrt{n}\,\dfrac{(f_{a\wedge\overline{b_k}} - f_a f_{\overline{b_k}})}{\sqrt{f_a f_{\overline{b_k}}}}$

D'où $\dfrac{\partial q}{\partial f_{a\wedge\overline{b_k}}} = \sqrt{n}\,\dfrac{1}{\sqrt{f_a f_{\overline{b_k}}}}$ et $\dfrac{\partial q}{\partial f_{\overline{b_k}}} = -\dfrac{\sqrt{n}}{2}\left[\dfrac{f_{a\wedge\overline{b_k}}}{\sqrt{f_{\overline{a}}}} f_{\overline{b_k}}^{-\frac{3}{2}} + \sqrt{f_a}\, f_{\overline{b_k}}^{-\frac{1}{2}}\right]$

Comme les valeurs de l'indice d'implication $q(b_k)$ sont négatives dans le cas où n_{a_k} est inférieur à n_{b_k}, il suffit alors de comparer la variation observée $q(b_2) - q(b_1)$ que l'on souhaite négative, c'est-à-dire qu'il y ait une meilleure intensité d'implication en b_2 qu'en b_1, à la variation attendue par le gradient calculé. Ainsi donc c'est le signe de la différence $\Delta q(b_1; b_2) - [q(b_2) - q(b_1)]$ qui va nous informer sur l'amélioration ou non de l'intensité d'implication au cours de la transition T de b_1 vers b_2.

Si cette différence est positive, l'amélioration observée est plus intéressante pour l'intensité d'implication qu'elle n'aurait été si l'évolution d'un nœud b_1 au suivant b_2 avait suivi le gradient de q en b_1.

4.2 Variations de l'intensité d'implication

Nous cherchons maintenant à déterminer les variations de $\varphi(a,b)$ lors de la transition T du nœud S_1 vers le nœud S_2. Pour cela, nous calculons l'intensité d'implication $\varphi(T)$, en recourant à l'approximation gaussienne de la loi de la v.a. centrée réduite $Q(a,\overline{b})$ établie à partir des contre-exemples à l'implication, afférente à la variation théorique attendue du gradient de q, à savoir

$$q(T) = q(a,\overline{b_1}) + \Delta(S_1, S_2) \text{ et } \varphi(T) = \frac{1}{\sqrt{2\pi}} \int_{q(T)}^{\infty} e^{\frac{-t^2}{2}} dt$$

Pour estimer le gain dû à la transition T, il suffit alors de comparer cette intensité à celle qui a été observée $\varphi(a,b_2) = \dfrac{1}{\sqrt{2\pi}} \displaystyle\int_{q(a,\overline{b_2})}^{\infty} e^{\frac{-t^2}{2}} dt$.

Le gain sera positif ou négatif selon que l'intensité observée sera supérieure ou égale à l'intensité attendue du gradient. Il sera exprimé en pourcentage de $\varphi(T)$ par le rapport :

$$\frac{\varphi(a,b_2) - \varphi(T)}{\varphi(T)}$$

.

Nous utiliserons alors la formulation suivante que le gain lié à la transition T est de g% de l'intensité attendue de l'observation en S_1 et du gradient de q en ce nœud.

Notons que cette méthode peut être généralisée quelles que soient les variables en jeu, c'est-à-dire dans les cas où d'autres variables peuvent être modifiées. Il suffit d'avoir recours à la différentielle de q selon les 4 variables actives (f_a, f_b, $f_{\overline{b}}$, $f_{a\wedge\overline{b}}$).

Remarque. Considérons l'intensité d'implication φ comme fonction de $q(a,\bar{b})$:

$$\varphi(q) = \frac{1}{\sqrt{2\pi}} \int_q^{\infty} e^{-t^2/2} dt$$

On peut alors examiner comment φ(q) varie lorsque q varie au voisinage d'une valeur donnée (a,b), sachant comment q varie lui-même en fonction des 4 paramètres qui le déterminent. Par dérivation de la borne d'intégration, on obtient :

$$\frac{d\varphi}{dq} = -\frac{1}{\sqrt{2\pi}} e^{\frac{-q^2}{2}} < 0$$

Ce qui confirme bien que l'intensité croît lorsque q décroît, mais la vitesse de croissance est précisée par la formule, ce qui permet d'étudier avec plus de précision les variations de φ(q).

.

Chapitre 3 : Extension de l'Analyse Statistique Implicative aux variables non binaires

1 L'A.S.I. des variables modales, variables fréquentielles et variables numériques

Nous abordons dans ce chapitre la question du traitement de variables autres que les variables binaires.

1.1 Situation fondatrice de la relation de propension entre deux variables modales

La thèse de Marc Bailleul (1994) porte en particulier, sur la représentation que se font les enseignants de mathématiques de leur propre enseignement. Afin de la mettre en évidence, des mots significatifs leur sont proposés qu'ils doivent hiérarchiser. Leurs choix ne sont donc plus binaires, les mots retenus par un enseignant quelconque sont ordonnés du moins au plus représentatif. L'interrogation de M. Bailleul se centre alors sur des questions du type : « si je choisis tel mot avec telle importance alors je choisis tel autre mot avec une importance au moins égale ». Pour analyser les données, il a fallu étendre la notion d'implication statistique à des variables autres que binaires. C'est le cas des **variables modales** qui sont associées à des phénomènes où les valeurs a(x) sont des nombres de l'intervalle [0,1] et qui décrivent des degrés d'appartenance ou de satisfaction comme le sont en logique floue, par exemple, les modificateurs linguistiques "peut-être", "un peu", "quelquefois", etc.. Cette problématique se retrouve également dans des situations où la fréquence d'une variable traduit un préordre sur les valeurs attribuées par les sujets aux variables qui leur sont présentées. Il s'agit de variables fréquentielles qui sont associées à des phénomènes où les valeurs de a(x) sont des réels positifs quelconques. On trouve une telle situation lorsque l'on considère le pourcentage de réussite d'un élève à une batterie de tests portant sur des domaines distincts.

1.2 Formalisation de la relation de propension entre variables modales

J.B. Lagrange (1998) a construit dans le cas des variables modales, un indice de propension (auparavant M. Bailleul (1994) fit autrement) entre variables modales qui généralise l'indice d'implication entre variables binaires. En posant la définition suivante :

Définition 6:

- si a(x) et $\overline{b}(x)$ sont les valeurs prises en x par les variables modales a et $\overline{b}$, où $\overline{b}(x) = 1 - b(x)$

- si s_a^2 et $s_{\overline{b}}^2$ sont les variances empiriques des variables a et $\overline{b}$

alors

$$\tilde{q}(a,\bar{b}) = \frac{\sum_{x\in E} a(x)\bar{b}(x) - \frac{n_a n_{\bar{b}}}{n}}{\sqrt{\frac{(n^2 s_a^2 + n_a^2)(n^2 s_{\bar{b}}^2 + n_{\bar{b}}^2)}{n^3}}}$$ est l'indice de propension de variables modales.

Cette solution apportée au cas modal est aussi applicable au cas des **variables fréquentielles**, voire des **variables numériques** positives, à condition d'avoir normalisé les valeurs observées sur les variables, telles que a et b, la normalisation dans [0,1] étant faite à partir du maximum de la valeur prise respectivement par a et b sur l'ensemble E.

1.3 Modélisation et propriétés de l'indice de propension et de l'intensité de propension entre variables non binaires

Nous nous intéressons désormais aux variables qui ne sont plus nécessairement binaires mais à valeurs réelles normalisées sur l'intervalle [0, 1]. Elles sont observées sur un ensemble de transactions E de cardinal n. Parmi ces variables *numériques*, les variables dites *modales* admettent un nombre fini de modalités respectivement *ordonnées* sur l'intervalle [0,1].

A l'instar de J.B. Lagrange (1998) et de S. Guillaume (2000), nous utilisons l'expression : propension (ou tendance) de a vers b, si l'on rencontre généralement dans E peu de transactions $i \in E$ pour lesquels $a_i > b_i$, pour la relation d'ordre sur [0, 1] où a_i et b_i sont respectivement les valeurs de *a* et *b* observées en *i*.

Notons $\bar{b}_i$ le complément à 1 de b_i : $\bar{b}_i = 1\text{-}b_i$. On choisit donc, comme pour l'implication entre variables binaires, l'indice $\sum_{i\in E} a_i \bar{b}_i$ - qui prend la valeur $n_{a\wedge\bar{b}}$ dans le cas binaire - comme indice de non-propension (ou de non-tendance) de *a* vers *b*. Ainsi, intuitivement et grossièrement, plus cet indice sera petit, plus on pourra s'attendre à une propension de *a* vers *b* (Régnier et Gras 2004)). Nous préciserons ce point plus loin. `

Cet indice présente un caractère voisin de l'indice de corrélation linéaire. Mais, d'une part, il n'est pas centré et, d'autre part, notre intérêt portera surtout sur ses valeurs fortes obtenues dans les cas où la variable *b* domine la variable *a*

Nous reprenons la démarche de J.B. Lagrange (1998), mais à travers une modélisation qui sera différente. En effet, il visait une *modélisation de Poisson* dans la restriction au cas binaire. Nous en avons cité le résultat principal dans l'article (Gras et al., 2001). En revanche, nous visons ici une modélisation de l'indice de propension en tant qu'extension stricte de la *modélisation binomiale* dans le cas des variables binaires.

Considérons pour cela, *n* couples de variables aléatoires X_i et Y_i indépendantes et identiquement distribuées, pour tout sujet *i*, dont la moyenne et la variance estimées sont celles des observations empiriques de E. Les réalisations de ces variables sont respectivement les n couples (a_i, b_i) de valeurs de $[0, 1]^2$.

Ainsi, par exemple, pour tout *i*, l'espérance E(X_i) peut être estimée par $m_a = \frac{1}{n}\sum_{k\in E} a_k$ et la variance Var(X_i) peut l'être par la réalisation de la variance empirique de *a* :

$$\upsilon_a = \frac{1}{n}\sum_{k\in E}(a_k - m_a)^2$$

On note également $\overline{Y_i} = 1 - Y_i$.

On pose : $M_1 = E[X_i \overline{Y_i}]$, espérance de la variable aléatoire produit $X_i \overline{Y_i}$ et $M_2 = E[(X_i \overline{Y_i})^2]$, moment d'ordre 2 de la variable aléatoire $X_i \overline{Y_i}$.

Par suite, M_1 est égale à $E[X_i]E[\overline{Y_i}]$ en raison de l'indépendance a priori entre X_i et $\overline{Y_i}$ et M_2 est égale à $E[(X_i^2)]E[(\overline{Y_i})^2]$. On a donc : $M_1 = m_a m_{\bar{b}}$ et $M_2 = (v_a + m_a^2)(v_{\bar{b}} + m_{\bar{b}}^2)$ où $m_{\bar{b}} = 1 - m_b$. Il est aisé de démontrer que $v_{\bar{b}} = v_b$.

La propension (ou tendance) sera alors mesurée par l'écart entre ce qui est attendu sous l'hypothèse d'indépendance a priori entre les variables numériques X_i et Y_i et ce qui a été réellement observé à travers les réalisations a_i et b_i. Plus la moyenne des n observations $a_i(1-b_i)$ est petite par rapport à la moyenne attendue, plus la propension sera grande.

En utilisant le théorème de la limite centrale dite de Lindeberg-Lévy puisque l'indice est une moyenne de variables aléatoires indépendantes identiquement distribuées, on démontre que Z suit approximativement, pour n grand, la loi normale d'*espérance* $E(Z) = M_1$ et de *variance* $\frac{1}{n}(M_2 - M_1^2)$.

Autrement dit, la loi de la variable « indice de propension empirique » $\tilde{Q}(a,\bar{b}) = \dfrac{Z - M_1}{\sqrt{\frac{1}{n}(M_2 - M_1^2)}}$ est approximativement la loi gaussienne centrée réduite N(0 ; 1).

Plus explicitement, pour obtenir la valeur estimée de la réalisation de la variable « indice de propension empirique », notons m_a la moyenne des observations a_i et m_b la moyenne des observations b_i ; la moyenne M_1 est alors égale à $m_a(1-m_b)$, ou $m_a . m_{\bar{b}}$ où $m_{\bar{b}}$ est la moyenne des $\overline{b_i}$.

Notons également v_a la variance des a_i et v_b celles des b_i, le moment M_2 d'ordre 2 des observations a_i et b_i est égal à $(v_a + m_a^2)(v_b + m_{\bar{b}}^2)$ puisque les observations b_i et $\overline{b_i}$ ont la même variance.

L'indice empirique d'implication devient :

$$\tilde{q}(a,\bar{b}) = \frac{\frac{1}{n}\sum_{i\in E} a_i \bar{b}i - m_a m_{\bar{b}}}{\sqrt{\frac{v_a(v_b + m_{\bar{b}}^2) + m_a^2 v_b}{n}}}$$

Quant à l'estimation de l'intensité de propension, elle est encore obtenue par :

$$\varphi(a,b) = 1 - \Pr\left[\tilde{Q}(a,\bar{b}) \le \tilde{q}(a,\bar{b})\right] = \frac{1}{\sqrt{2\pi}} \int_{\tilde{q}(a,\bar{b})}^{\infty} e^{-\frac{t^2}{2}} dt$$

1.4 Restriction de l'indice de propension aux variables binaires et indice d'implication.

Si *a* et *b* sont deux variables binaires, les lois des variables aléatoires associées X_i et Y_i indépendantes sont celles de Bernoulli.

Par ailleurs, $n_a = \sum_{i \in E} a_i$ et $n_b = \sum_{i \in E} b_i$ et $n_{a \wedge \bar{b}} = \sum_{i \in E} a_i \bar{b}_i$.

Par suite, la réalisation de $M_1 = E(X_i)E(\bar{Y}_i)$ peut être estimée par $\frac{n_a}{n} \frac{n_{\bar{b}}}{n}$. D'autre part pour les variables de Bernoulli, X_i et Y_i, nous avons $X_i^2 = X_i$ et $\bar{Y}_i^{\,2} = \bar{Y}_i = 1 - Y_i$.

Ainsi $M_2 = E[X_i^2]E[\bar{Y}_i^{\,2}] = E[X_i]E[\bar{Y}_i]$ peut aussi être estimée par $\frac{n_a}{n} \frac{n_{\bar{b}}}{n}$. Donc les valeurs estimées de M_2 et M_1 sont égales à $\frac{n_a}{n} \frac{n_{\bar{b}}}{n}$. Par suite, il en résulte que l'estimation de $\frac{1}{n}(M_2 - M_1^2) = \frac{1}{n}(M_1 - M_1^{\,2})$ vaut :

$$\frac{1}{n}\left[\frac{n_a n_{\bar{b}}}{n^2} - \left(\frac{n_a n_{\bar{b}}}{n^2}\right)^2\right] = \frac{1}{n^2}\frac{n_a n_{\bar{b}}}{n^2}\left[1 - \frac{n_a n_{\bar{b}}}{n^2}\right]$$

Ainsi l'estimation de la réalisation de Z est : $\frac{1}{n}\sum_{i \in E} a_i \bar{b}_i = \frac{1}{n} n_{a \wedge \bar{b}}$ ce qui conduit à ce que les estimations respectives des réalisations des variables aléatoires $\tilde{Q}(a, \bar{b})$ et $Q(a, \bar{b})$ coïncident. Dit autrement,

$$\tilde{q}(a,\bar{b}) = \frac{\frac{n_{a \wedge \bar{b}}}{n} - \frac{n_a n_{\bar{b}}}{n^2}}{\frac{1}{n}\sqrt{\frac{n_a n_{\bar{b}}}{n}\left(1 - \frac{n_a n_{\bar{b}}}{n^{2)}}\right.}} = \frac{n_{a \wedge \bar{b}} - \frac{n_a n_{\bar{b}}}{n}}{\sqrt{\frac{n_a n_{\bar{b}}}{n}\left(1 - \frac{n_a n_{\bar{b}}}{n^{2)}}\right.}} = q(a,\bar{b})$$

Ce modèle adopté correspond strictement à une extension de la notion d'implication statistique fondée sur des variables binaires, à celle de propension statistique fondée sur les variables modales à valeurs dans [0,1]. En effet l'indice de propension statistique coïncide exactement avec l'indice d'implication statistique en considérant une variable binaire comme une variable modale dégénérée ne prenant que les valeurs extrêmes 0 et 1. Ceci permet d'utiliser dans C.H.I.C. les mêmes algorithmes que les variables soient binaires ou numériques.

2 L'ASI des variables sur intervalles et variables-intervalles

2.1 Situation fondatrice de l'ASI des variables sur intervalles

On recherche, par exemple, à extraire d'un ensemble de données biométriques, la règle suivante, en estimant sa qualité : « si un individu pèse entre 65 et 70 kg alors en général il mesure entre 1.70 et 1.76 m ». Une situation comparable se présente dans la recherche de relation entre des intervalles de performances d'élèves dans deux disciplines différentes. La situation plus générale s'exprime alors ainsi : deux variables réelles a et b prennent un certain nombre de valeurs sur 2 intervalles finis $[a_1 ; a_2]$ et $[b_1 ; b_2]$. Soit A (resp. B) l'ensemble des valeurs de a (resp. b) observées sur $[a_1 ; a_2]$ (resp. $[b_1 ; b_2]$. Par exemple ici, a représente les poids d'un ensemble de n sujets et b les tailles de ces mêmes sujets.

Deux problèmes se posent :

1° peut-on définir, et comment le faire, des sous-intervalles adjacents de $[a_1 ; a_2]$ (resp. $[b_1 ; b_2]$.) afin que la partition la plus fine obtenue respecte au mieux la distribution des valeurs observées dans $[a_1 ; a_2]$ (resp. $[b_1 ; b_2]$.) ?

2° peut-on trouver, et comment le faire, les partitions respectives de $[a_1 ; a_2]$ et $[b_1 ; b_2]$ constituées de réunions des sous-intervalles adjacents précédents, partitions qui maximisent l'intensité d'implication moyenne des sous-intervalles de l'un sur des sous-intervalles de l'autre appartenant à ces partitions ?

Nous répondons à ces deux questions dans le cadre de notre problématique en faisant choix des critères à optimiser pour satisfaire l'optimalité attendue dans chaque cas. A la première question, de nombreuses solutions ont été apportées dans d'autres cadres (par exemple : Lahanier-Reuter, 1998).

2.2 Caractérisation et propriétés des variables sur intervalles.

Approche du premier problème

On va s'intéresser à l'intervalle $[a_1 ; a_2]$ en le supposant muni d'une partition initiale triviale de sous-intervalles de même longueur, mais pas nécessairement de même distribution des fréquences observées sur ces sous-intervalles.

Notons $P_0 = \{A_{01}, A_{02},..., A_{0p}\}$, cette partition en p sous-intervalles. On cherche à obtenir une partition de $[a_1 ; a_2]$ en p sous-intervalles $A_{q1}, A_{q2},..., A_{qp}$ de telle façon qu'au sein de chaque sous-intervalle, on ait une bonne homogénéité statistique (faible **inertie intra-classe**) et que ces sous-intervalles présentent une bonne hétérogénéité mutuelle (forte **inertie inter-classe**). On sait que si l'un des critères est vérifié, l'autre l'est nécessairement en vertu de l'application du théorème de Koenig-Huyghens. Pour ce faire, nous adoptons une méthode directement inspirée de la méthode des **nuées dynamiques** conçue par Edwin Diday (1972), exposée par Lebart (Lebart et al. 2006) et que nous avons adaptée à la situation présente. Nous obtenons ainsi la partition optimale visée.

Approche du second problème

On suppose maintenant que les intervalles $[a_1 ; a_2]$ et $[b_1 ; b_2]$ sont munis de partitions optimales P et Q, respectivement, au sens des nuées dynamiques. Soit p et q les nombres respectifs de sous-intervalles composant P et Q. A partir de ces deux partitions, il est possible d'engendrer 2^{p-1} et 2^{q-1} partitions obtenues par réunions itérées de sous-intervalles adjacents respectivement de P et de Q [10].

On calcule les intensités d'implication respectives de chaque sous-intervalle réuni ou non à un autre de la première partition sur chaque sous-intervalle réuni ou non à un autre de la seconde, puis les valeurs des intensités des implications réciproques.

Il y a donc au total $2.2^{p-1}.2^{q-1}$ familles d'intensités d'implication, chacune d'entre elles nécessitant le calcul de tous les éléments d'une partition de $[a_1 ; a_2]$ sur tous les éléments d'une des partitions de $[b_1 ; b_2]$ et réciproquement.

On choisit comme *critère d'optimalité* la moyenne géométrique des intensités d'implication, moyenne associée à chaque couple de partitions d'éléments, réunis ou non, définies inductivement. On note les deux maxima obtenus (implication directe et sa réciproque) et on retient les deux partitions associées en déclarant que l'implication de la variable-sur-intervalle a sur la variable-sur-intervalle b est optimale lorsque l'intervalle $[a_1 ; a_2]$ admet la partition correspondant au premier maximum et que l'implication réciproque optimale est satisfaite pour la partition de $[b_1 ; b_2]$ correspondant au deuxième maximum.

Comme nous allons le voir dans le sous-chapitre suivant, cette approche est applicable à la notion de variable-intervalle étudiée par ailleurs par E. Diday et ses collaborateurs. En effet, les modalités sont nominales ou numériques. Dans le premier cas, il est possible de calculer les implications des modalités ou des réunions de modalités de l'une sur les modalités ou les réunions de modalités de l'autre, comme ci-dessus. Dans le second cas, il peut être intéressant et utile de redéfinir les partitions des variables sous-jacentes en optimisant ces partitions comme nous l'avons fait dans le premier problème posé dans le cadre des variables sur intervalles.

2.3 Situation fondatrice de l'ASI des variables-intervalles

On dispose des données fournies par une population de n individus (qui peuvent être chacun ou certains des ensembles d'individus, par ex. une classe d'élèves) selon p variables (par ex. notes sur une année en français, math, physique, …, mais aussi bien : poids, tailles, tour de poitrine, …). Les valeurs prises par ces variables selon chaque individu sont des intervalles de réels positifs. Par exemple, l'individu x donne la valeur [12 ; 15,50] à la variable note de math. E. Diday parlerait à ce sujet de p **variables symboliques** à valeurs intervalles définies sur la population.

On cherche à définir une implication d'intervalles, relatifs à une variable a, constitués eux-mêmes des intervalles observés, vers d'autres intervalles pareillement définis et relatifs à une autre variable b. Ceci permettra de mesurer l'association implicative, donc non

[10] Il suffit de considérer l'arborescence dont A_1 est la racine, puis de le réunir ou non à A_2 qui lui-même sera ou non réuni à A_3, etc. Il y a donc 2^{p-1} branches dans cette arborescence.

symétrique, de certain(s) intervalle(s) de la variable a avec certain(s) intervalle(s) de la variable b, ainsi que l'association réciproque à partir de laquelle on retiendra la meilleure pour chaque couple de sous-intervalles en jeu, comme nous l'avons fait avec les variables-sur-intervalles. Par exemple, on dira que le sous-intervalle [2 ; 5,5] de notes de mathématiques implique généralement le sous-intervalle [4,25 ; 7,5] de notes de physique, ces deux sous-intervalles appartenant à une partition optimale au sens de la variance expliquée des intervalles respectifs de valeurs [1 ; 18] et [3 ; 20] prises dans la population. De même, on dira que [14,25 ; 17,80] en physique implique le plus souvent [16,40 ; 18] en mathématiques.

2.4 Caractérisation et propriétés des variables-intervalles.

En suivant la démarche de E. Diday et ses collaborateurs, si les valeurs prises selon les sujets par les variables a et b sont de nature symbolique, en l'occurrence des intervalles de R^+, il est possible d'étendre les algorithmes ci-dessus (Gras, 2001a). Par exemple, à la variable a sont associés des intervalles de poids et à la variable b des intervalles de tailles, intervalles dus à une imprécision des mesures. Effectuant la réunion des intervalles I_X et J_X décrits par les sujets x de E selon respectivement chacune des variables a et b, on obtient deux intervalles I et J recouvrant toutes les valeurs possibles de a et de b. Sur chacun d'eux on peut définir une partition en un certain nombre d'intervalles respectant comme plus haut un certain critère d'optimalité. Pour cela, les intersections des intervalles tels que I_X et J_X avec ces partitions seront munies d'une distribution prenant en compte les étendues des parties communes. Cette distribution peut être uniforme ou d'un autre type discret ou continu. Mais ainsi, nous sommes ramenés à la recherche de règles entre deux ensembles de variables-sur-intervalles qui prennent leurs valeurs sur [0 ; 1] et à partir desquelles nous pouvons chercher les implications optimales.

Chapitre 4 : Extension de l'Analyse Statistique Implicative à des hiérarchies de règles[11].

1 Introduction

Les développements théoriques de l'analyse de données offrent des retombées enrichissantes pour l'E.C.D. (Extraction des Connaissances dans des Données) et sa vitalité n'est pas étrangère aux échanges induits. Par exemple, la construction d'indices permettant d'affecter une mesure non symétrique à des règles d'inférence partielle fournit des points d'application à l'extraction et à la représentation de règles d'association imprécises entre attributs binaires décrivant une population. Les démarches fondamentales convergent vers une problématique commune aux deux domaines ; il s'agit, avons-nous vu, de découvrir et de quantifier des règles non symétriques pour modéliser des relations du type *"si a alors presque b"*. C'est, par exemple, un objectif majeur des réseaux bayésiens (Pearl 1988) ou de certains travaux utilisant les treillis de Galois (Simon, 2000 ; Bernard 1999). Le plus souvent, notamment en Fouille de Données (Agrawal 1993), la probabilité conditionnelle est l'indice fondamental de l'association, même dans l'approche multivariée. Cela peut être aussi la corrélation comme dans (Brin 1997). Cependant à notre connaissance, d'une part, les développements s'arrêtent à la proposition d'un indice d'implication partielle pour des données binaires, et d'autre part, cette notion n'est pas étendue à l'extraction de règles de règles où les prémisses et les conclusions peuvent être elles-mêmes des règles. Nous proposons ici ces prolongements en formalisant la notion de hiérarchie orientée, en charge de la représentation graphique de la structuration de l'ensemble des variables selon ces règles de règles, donc d'un niveau conceptuel supérieur, dites règles généralisées. Cette formalisation peut nous rappeler la notion d'abstraction réfléchissante selon J. Piaget qui fait passer de la strate « objet » à celle des opérations sur les objets, à celle des opérations sur ces opérations, etc.. « Elle est réfléchissante aux deux sens suivants : elle transpose sur un plan supérieur de conceptualisation ce qu´elle emprunte à un palier précédent » écrit Sylvie Lucas dans Le Tome 52 du Bulletin de Psychologie, juillet-août 1999.

Rappelons qu'une représentation structurée des relations implicatives dans l'ensemble des variables instanciées a été obtenue, dans le cadre de l'A.S.I., par un graphe implicatif sans cycle pondéré par les intensités, fermé transitivement à un seuil donné. Il est utile dans des situations d'évolution, par exemple, en psychologie cognitive, pour interpréter des chemins de ce graphe, constitués de suites d'arcs qui lient certaines variables, en termes de genèses différentielles. Mais la hiérarchie que nous présentons maintenant va doubler les informations fournies par les règles d'association en organisant leur ensemble selon une structure ordonnée en méta-règles, en méta-méta-règles, etc..

[11] Une partie de ce chapitre a été publiée sous une forme voisine dans une sélection RNTI-C-1 des Actes des 11èmes Rencontres de la Société Francophone de Classification sous le titre : « Significativité des niveaux d'une hiérarchie orientée en analyse statistique implicative » (septembre 2004) par Régis Gras, Pascale Kuntz et Jean-Claude Régnier

2 Hiérarchie de classes de variables

Afin de soutenir l'intuition, nous baserons le modèle de la **hiérarchie orientée cohésitive** sur la métaphore linguistique suivante :

1. *les variables* (ou attributs) de l'ensemble V (card V=p) constitueront l'ensemble les *lettres de l'alphabet V,*
2. *les classes de k variables,* éléments de V^k, par exemple $(a_1, a_2,\ldots, a_k)$, constitueront les *syllabes du vocabulaire,*
3. *les classes maximales,* i.e. telles qu'aucune variable ne la complète, constitueront *les mots* du vocabulaire,
4. *l'organisation hiérarchique* de l'ensemble des classes constituera une *phrase,* structurée par des propositions incises.

D'autres métaphores peuvent illustrer le modèle que nous allons construire comme, par exemple, l'ensemble des séquences constituant le génome ou encore une théorie mathématique organisée en théorèmes et corollaires. Mais nous verrons que ces métaphores ne satisfont pas totalement le modèle.

On va également constater que ce modèle hiérarchique, où l'ordre intervient, ne s'apparente pas au modèle classique d'une hiérarchie ascendante, par exemple celle basée sur un indice de similarité entre attributs, car les classes d'une telle hiérarchie sont des sous-ensembles de variables et non pas des k-uplets.

2.1 Hiérarchie orientée. Définitions. Propriétés

Définition 7: On appelle hiérarchie orientée **H** sur l'ensemble des variables V, une suite d'arrangements (au sens de la combinatoire) des éléments de V, vérifiant les axiomes 1. 2 et 3 énoncés ci-dessous. Ces arrangements sont appelés classes de H.

Par exemple, {(j), (f,g), (b,c), (e,f,g), (b,c,d), (h,i), (a,b,c,d), (e,f,g,h,i)} est une hiérarchie orientée sur V={a,b,c,d,e,f,g,h,i} et (a,b,c,d) est une classe de cette hiérarchie.

Définition 8: On appelle classe C de degré k de la hiérarchie **H** un arrangement de k éléments de V appartenant à **H**. On notera $\prec$ la relation d'ordre induite sur **C** par le tirage d'un arrangement.

Par exemple, $(a_1, a_2,\ldots, a_k)$, pour $k \leq p$, est une classe de degré k et $a_1 \prec a_2 \prec \ldots \prec a_k$. Mais également, par convention, (a) est une classe de degré 1. Elle est dite élémentaire

Définition 9: On appelle **troncation de C**, tout sous-arrangement des éléments de **C** respectant la structure d'ordre $\prec$ et la consécutivité.

Par exemple, si **C** = $(a_1, a_2,\ldots, a_k)$, la classe **C'** = $(a_i, a_{i+1},\ldots, a_j)$ où $1 \leq i$ et $j \leq k$, est une troncation de **C**.

Définition 10: On note **C'** $\widehat{\subset}$ **C** si et seulement si **C'** est une troncation de **C**. Une classe est dite maximale s'il n'existe pas de classe qui la contienne dans **H**. Elle est dite minimale si elle ne contient aucune classe de la hiérarchie **H**. En particulier, une classe élémentaire est donc minimale (mais elle peut être aussi maximale).

On peut comparer, sans la confondre cependant, cette relation à l'inclusion ensembliste. Dans l'exemple initial, les classes (a,b,c,d), (e,f,g,h,i) et (j) sont maximales. Cette dernière est aussi minimale.

Définition 11: La trace de **C** sur **C'** est constituée d'éléments communs à **C** et **C'** et elle respecte la structure d'ordre $\prec$ et la consécutivité. La trace est une opération commutative notée $\widehat{\cap}$.

Ainsi on peut comparer, sans la confondre, cette opération à l'intersection ensembliste.

Par exemple, (d,f,g,a,e) $\widehat{\cap}$ (f,g,a,e,b,h)= (f,g,a,e)

Définition 12: Si les deux classes quelconques **C'** et **C''** ont une trace vide (**C'** $\widehat{\cap}$ **C''**= $\varnothing$), la concaténation de **C'** et **C''** notée **C'** $\widehat{\cup}$ **C''** est la classe **C** dont les éléments appartiennent à **C'** et **C''** et à elles exclusivement. Elle respecte les ordres au sein de **C'** et **C''**et le plus grand élément de **C'** précède le plus petit de **C''**. On dira que **C'** et **C''** sont des classes génératrices de **C'** $\widehat{\cup}$ **C''.**

Cette opération, comparable à la concaténation ordinaire, ainsi qu'à la réunion ensembliste sans se confondre avec elle, est non commutative.

Par exemple, si **C'** = (d,f,g,a) et **C''** = (b,u,r,p,y), **C'** $\widehat{\cup}$ **C''** = (d,f,g,a,b,u,r,p,y), alors que **C''** $\widehat{\cup}$ **C'** = (b,u,r,p,y,d,f,g,a)

2.2 Axiomes d'une hiérarchie orientée

Axiome 1 : $\forall$ **C** et $\forall$ **C'** classes de **H** , **C'** $\widehat{\cap}$ **C''** $\in$ **{** $\varnothing$, **C, C'}**

Axiome 2 : $\forall$ **C** $\in$ **H**, si **C** n'est pas élémentaire ou minimale, elle est la concaténation de classes de **H**

Axiome 3 : Il existe une permutation des éléments de V qui coïncide avec la concaténation de toutes les classes maximales de **H**

2.3 Algorithme de construction de l'ensemble des classes

Nous définissons ci-dessous un critère algébrique en vue de nous permettre de construire de façon ascendante, la hiérarchie organisatrice de l'ensemble V des variables et qui respecte les trois axiomes d'une hiérarchie orientée.

2.3.1 Critères algébriques

Définition 13: La **cohésion d'une classe** de degré 2, correspondant au couple (a,b) est définie, à partir de l'entropie au sens de Shannon, par la formule,

$coh(a,b) = \left(1 - \left[-p(\log_2(p)) - (1-p)(\log_2(1-p))\right]^2\right)^{\frac{1}{2}}$ où p=φ(a,b)≥0,50 et $coh(a,b) = 0$ si p=φ(a,b)<0,50.

Définition 14: La cohésion d'une classe **C** = (a_1, a_2,…, a_r) de degré r, est définie par la formule : $coh(C) = \left[\prod_{i=1,...,r-1}^{j=2,...,r;\, j>i} coh(a_i, a_j)\right]^{\frac{2}{r(r-1)}}$

Définition 15: La cohésion d'une classe **C** = (a) de degré 1, est définie par $coh(a) = 1$

2.3.2 Algorithme de construction de la hiérarchie

Niveau 0
Les classes sont élémentaires et toutes les cohésions sont égales à 1

Niveau 1
On compare toutes les cohésions des arrangements 2 à 2 de V.
On conserve celle, notée C_1, qui correspond au maximum, soit par ex. C_1=(a,b).

Définition 16: On appelle **nœud** 1, la règle $a \Rightarrow b$

Niveau 2
On compare toutes les cohésions des classes à 2 éléments, sauf C1, à celles des classes à 3 éléments du type (x, a, b) et (a, b, x).
On conserve celle, notée C_2, correspondant au maximum obtenu.
Le nœud 2 sera :

1. soit une classe à 2 éléments, et dans ce cas le nœud sera du type : c=>d
2. soit une classe à 3 éléments, et dans ce cas le nœud sera noté (a=>b)=>c ou c=>(a=>b).

Ces dernières règles sont dites **composées ou généralisées ou R-règles**. Pour restituer l'ordre dans lequel est constituée la classe, on notera, par exemple ici, ((a,b),c) ou, dans l'autre cas, (c,(a,b)).

Niveau k
On compare toutes les cohésions des concaténations de 2 des classes déjà formées aux niveaux inférieurs, du type C_i et C_j avec i<k et j<k. On conserve celle C_k=$C_i \cup C_j$ qui satisfait le maximum et est la concaténation de C_i et C_j.
Le nœud k correspondant sera noté par extension C_i=>C_j. Mais on peut expliciter des nœuds correspondant à la formation des troncations respectives et génératrices de C_i et C_j aux niveaux inférieurs.
Par exemple, la règle composée $((f \Rightarrow (e \Rightarrow u)) \Rightarrow ((a \Rightarrow b) \Rightarrow (c \Rightarrow d))$ est l'explicitation d'un nœud particulier. Afin de faire apparaître les classes formées à des niveaux successifs, on notera aussi la règle sous la forme : ((f(eu))((ab)(cd)))

L'algorithme s'arrête, au plus tard, au niveau p-1, lorsque toute concaténation conduirait à une classe de cohésion nulle ou à une permutation de V. Les classes formées au niveau ultime et qui n'admettent pas de classes qui les contiennent sont donc maximales. Certaines classes maximales peuvent aussi être élémentaires. La hiérarchie est composée de l'ensemble des classes maximales et de toutes leurs parties.

2.4 Conformité de la construction aux axiomes d'une hiérarchie orientée

La hiérarchie ainsi construite vérifie les trois axiomes d'une hiérarchie orientée. En effet :

Axiome 1 :
Deux classes de **H**, **C'** et **C''** étant données,

1. ou bien elles sont associées dans une même concaténation et la constituent entièrement, alors $\mathbf{C'} \hat{\cap} \mathbf{C''} = \varnothing$
2. ou bien l'une est la concaténation de l'autre et d'une troisième et alors $\mathbf{C'} \hat{\subset} \mathbf{C''}$ ou $\mathbf{C''} \hat{\subset} \mathbf{C'}$
3. ou bien elles ne sont pas associées dans une concaténation et alors elles sont des arrangements sans élément commun, donc $\mathbf{C'} \hat{\cap} \mathbf{C''} = \varnothing$

Axiome 2 :
$\forall C \in H$

1. ou bien elle est constituée d'un élément et c'est une classe élémentaire
2. ou bien elle est constituée de plus d'un élément et elle est alors la concaténation de deux ou plusieurs classes par construction.

Axiome 3 :
On range toutes les classes maximales par ordre croissant de la cohésion ; les classes élémentaires seront les éléments maximaux de cet ordre. Toutes les classes sont 2 à 2 disjointes et tous les éléments de V appartiennent à l'une et l'une seulement des classes. La concaténation de leur ensemble constitue alors une permutation particulière de tous les variables de V.
Notons qu'à une permutation de V peuvent correspondre plusieurs hiérarchies.

Exemple 1: Si l'on range les classes maximales de la hiérarchie donnée au début du texte par ordre croissant de la cohésion, on obtient par exemple :
$coh(e,f,g,h,i) \leq coh(a,b,c,d) \leq coh(j)$ et (e,f,g,h,i,a,b,c,d,j) est une permutation de V.
Mais à cette permutation, peut aussi correspondre la hiérarchie :
{(g,h), (b,c), (e,f), (a,b,c), (g,h,i), (d,j), (a,b,c,d,j)} dont les classes maximales sont (e,f), (g,h,i) et (a,b,c,d,j).

Exemple 2 : Reprenant encore l'exemple initial, l'autre hiérarchie {(j), (f,g), (b,c), (e,f,g), (b,c,d), (h,i), (a,b,c,d), (h,i,e,f,g)}ne coïncide pas avec la première. La permutation correspondante de V est (h,i,e,f,g,a,b,c,d,j).

La figure ci-dessous montre la hiérarchie obtenue, artificiellement, à partir de 7 variables. Des interprétations de telles règles généralisées sont quelquefois complexes, comme par exemple, la règle $(x \Rightarrow (y \Rightarrow z)) \Rightarrow (t \Rightarrow v)$. Mais quelques règles sont réductibles à des assemblages plus aisément interprétables. Par exemple, la règle $x \Rightarrow (y \Rightarrow z)$ se ramènerait, dans le cas formel, à $x \wedge y \Rightarrow z$. La règle $(d \Rightarrow b) \Rightarrow (a \Rightarrow f)$ ou ((db)(af)), illustrée par la figure (FIG 10), peut s'interpréter comme : le « théorème » $d \Rightarrow b$ a généralement pour conséquence le « théorème » $a \Rightarrow f$. Cette figure montre aussi que la variable *e* n'a ni prémisse ni conclusion.

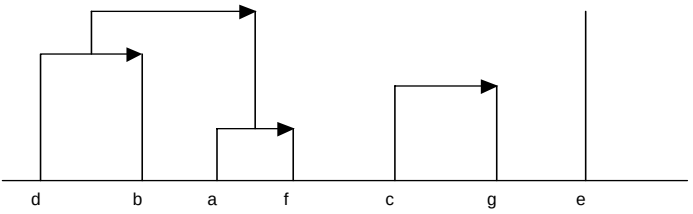


FIG. 10 *Un exemple de hiérarchie orientée*

3 La hiérarchie cohésitive basée sur une distance ultramétrique

On part de l'algorithme de construction de **H** déjà défini et par lequel la cohésion de la classe en voie de formation, à un niveau donné, est inférieure à la valeur de la cohésion au niveau immédiatement inférieur et est supérieure à celle du niveau immédiatement supérieur. C'est donc une fonction décroissante des niveaux et, a fortiori, avec l'inclusion des parties de **H**. C'est la propriété de la cohésion qui va permettre de définir la **distance ultramétrique** d, pour laquelle tous les triangles sont isocèles. Cette propriété d'ultramétricité de la cohésion va justifier, a posteriori, le bien-fondé de l'expression « hiérarchie » employée pour la construction.

Il suffit de choisir pour un couple quelconque de variables (x, y) :

$$d(x,y) = 1 - coh(C_{(x,y)})$$

où $coh(C_{(x,y)})$ est la cohésion de la plus petite classe $C_{(x,y)}$ contenant x et y. Rappelons les propriétés suivantes de la hiérarchie :

1. quelles que soient les classes C et C' de H, ou bien $C \hat{\subset} C'$ ou bien $C \hat{\supset} C'$ ou bien $C \hat{\cap} C' = \varnothing$
2. si $C \hat{\subset} C'$, alors $coh(C) \geq coh(C')$ par construction.

Vérifions alors que les trois axiomes d'ultramétricité sont bien valides sur l'ensemble V des variables

Axiome 1 :

Pour tout $x \in V$, d(x,x) = 0 par construction de d car $coh(x,x) = 1$

Axiome 2 :

Pour tout couple (x,y)∈AxA, d(x,y)=d(y,x) par construction

Axiome 3 :

(x,y,z)∈AxAxA, d(x,y)≤sup[d(x,z), d(y,z)] (1)

La définition adoptée respecte aussi l'axiome 3. En effet, soit $C_{x,y}$, $C_{x,z}$, $C_{y,z}$ les plus petites classes contenant respectivement x et y, x et z, y et z.

Alors $z \in C_{(x,z)} \hat{\cap} C_{(y,z)} \neq \varnothing$ d'où $C_{(x,z)} \hat{\subset} C_{(y,z)}$ ou bien $C_{(x,z)} \hat{\supset} C_{(y,z)}$ d'après les propriétés des classes de H. Supposons alors: $C_{(x,z)} \hat{\supset} C_{(y,z)}$. On en déduit d(x,z)≥d(y,z) et par suite d(x,z)=sup[d(x,z) ; d(y,z)] (2). De plus, on a à la fois : $x \in C_{(x,z)}$ et , par conséquent $C_{(x,y)} \hat{\subset} C_{(x,z)}$

Comme l'indice d croît avec l'inclusion en raison de la décroissance de la cohésion, alors : d(x,z)≥d(x,y) (3)

Par (2) et (3) on obtient donc (1) :d(x,y)≤sup[d(x,z), d(y,z)]

H est donc bien une **hiérarchie indicée** par la distance d au sens strict de hiérarchie mathématique.

4 Présentation d'une application de l'approche par hiérarchie cohésitive

Une enquête de l'Association française des Professeurs de Mathématiques de l'Enseignement Public (APMEP) a été proposée récemment aux professeurs de mathématiques de classes terminales de l'enseignement secondaire dans différentes filières : S (à dominante sciences dures) et ES (à dominante scientifique et sociale), littéraires L et technologique T. Ces variables constitueront des variables supplémentaires de l'analyse, donc n'entrent pas directement dans la constitution des règles. Nous avons recueilli et analysé (Bodin et Gras 1999) les réponses de 311 professeurs, à des questions portant sur les objectifs (15 ont été retenus) qu'ils assignent à leur enseignement et sur leurs opinions relatives à 11 phrases susceptibles d'être communément énoncées. Des objectifs jugés non pertinents sont ajoutés aux variables supplémentaires (cf. chap. 5) « filières ». Nous présentons ci-après le questionnaire mis dans un format conforme au présent ouvrage :

QUESTIONNAIRE -Professeurs de Terminale

Q1- Au nom de quelle série répondez-vous ? :…….

(*Si plusieurs séries, utiliser un questionnaire par série*)

Q2-Objectifs de la formation mathématique

A votre avis, quels sont les objectifs essentiels de la mission d'un professeur de mathématiques dans la série pour laquelle vous répondez ?

(*Choisir 6 objectifs et ranger par ordre préférentiel décroissant de 1 à 6*)

Code	*Objectifs*	Code	*Objectifs*
A-	acquisition de connaissances	B-	préparation à la vie professionnelle
C-	préparation à la vie civique et sociale	D-	préparation aux examens, concours, au passage dans l'enseignement supérieur
E-	développement de l'imagination et la créativité	F-	développement de la capacité à prouver et valider sa preuve
G-	développement de la capacité d'accepter des points de vue différents	H-	développement de la volonté et la persévérance
I-	développement de l'esprit critique	J-	développement de la capacité à communiquer avec objectivité, clarté et précision par des modes de représentation divers
K-	développement de compétences utiles dans les autres disciplines	L-	développement de la pratique de calculs formels, donc sans nécessité de signification
M-	développement de la capacité à mathématiser et à formaliser	N-	acquisition de savoir-faire
O-	participation au développement d'une culture générale		

Rang	1	2	3	4	5	6
Objectif						

Les objectifs ci-dessus vous paraissent-ils pertinents (PER): OUI NON (entourez votre choix)
Si non, précisez lesquels en utilisant le codage proposé

Q3- Votre opinion sur des opinions

Voici quelques opinions entendues dans la salle des professeurs. Entourez votre choix :

Code	**Opinion** 1= D'accord ; 2= Peu d'accord ; 3= Pas d'accord	**Choix**
OP1	C'est vrai que les math constituent un instrument de sélection excessif.	1 2 3
OP2	Au bac, je préfère qu'il y ait un grand problème avec plusieurs parties plutôt qu'un ensemble de petits problèmes indépendants.	1 2 3
OP3	Dans ma notation, j'attache plus d'importance à la démarche qu'au résultat.	1 2 3
OP4	Quand je corrige, j'aime bien un barème très détaillé sur les résultats à obtenir.	1 2 3
OP5	La démonstration est la seule façon rigoureuse de faire des mathématiques.	1 2 3
OP6	Je préfère des programmes bien définis indiquant ce que je dois et ce que je ne dois pas faire.	1 2 3
A la sortie de la terminale de la série sur laquelle vous répondez, un élève devrait pouvoir ou avoir...		
OP7	reconnaître si un nombre entier écrit dans la base 10 est divisible par 4	1 2 3
OP8	donner un exemple ou un contre-exemple personnels à l'affirmation : "si deux applications *f* et g sont strictement croissantes sur un intervalle, l'application produit *fxg* y est également croissante".	1 2 3
OP9	avoir appris à faire un test statistique pour pouvoir réfuter ou accepter l'hypothèse d'adéquation d'une loi théorique à une distribution empirique	1 2 3
OPX	estimer à vue, à 30% près, le périmètre et l'aire du plancher ainsi que le volume de la salle de classe	1 2 3

Les 26 variables correspondantes ne sont pas binaires, mais ordinales, elles prennent des valeurs décimales sur [0 ;1]. Ainsi l'analyse intègre l'intensité des attitudes, le choix prioritaire d'un objectif pondérant de façon différente un choix plus secondaire, voire non retenu. Pour ce faire, les enseignants font choix de 6 objectifs parmi 15 qu'ils assignent à leurs enseignements (ex : A : « Acquisition de connaissances », B : « Préparation à la vie professionnelle ») et d'opinions relatives à dix phrases communément énoncées (par exemple : OP 1 : « les maths constituent un instrument de sélection excessif ») (Bodin et Gras, 1999). Les poids des 26 variables figurent dans le tableau ci-dessous, compte tenu des pondérations décimales accordées aux variables ordinales (rangs de 6 choix pondérés par 1, 0.8, 0.6, etc. et accords modulés 1, 0.5, 0 suivant l'accord avec les opinions).

A	B	C	D	E	F	G	H		
105.7	8.8	9.7	140.0	21.8	138.7	19.5	44.8		
I	J	K	L	M	N	O	PER		
83.1	108.4	77.6	4.6	90.2	66.6	33.2	254		
OP1	OP2	OP3	OP4	OP5	OP6	OP7	OP8	OP9	OPX
81.5	147.5	242.5	229.0	190.0	240.0	200.0	165.0	98.0	207.0

TAB. 12 – *Occurrences des variables de l'enquête sur les professeurs de mathématiques*

La hiérarchie orientée obtenue structure les 26 variables en plusieurs classes qui définissent des R-règles de longueur, d'interprétation et d'intérêt variés. Une aide à l'interprétation peut être apportée si l'on se souvient de la tautologie en logique formelle :

$$(a \Rightarrow (b \Rightarrow c)) \Leftrightarrow ((a \wedge b) \Rightarrow c)$$

De plus, relativement à chaque classe maximale, le logiciel C.H.I.C. indique quelle variable supplémentaire contribue le plus à la formation de la classe. Cette information permet d'améliorer la compréhension et la signification de la classe.

Voici une partie de la hiérarchie où nous nous limitons dans un souci de clarté à trois classes maximales. Le logiciel CHIC fournit cette hiérarchie par une symétrie orthogonale par rapport à sa base.

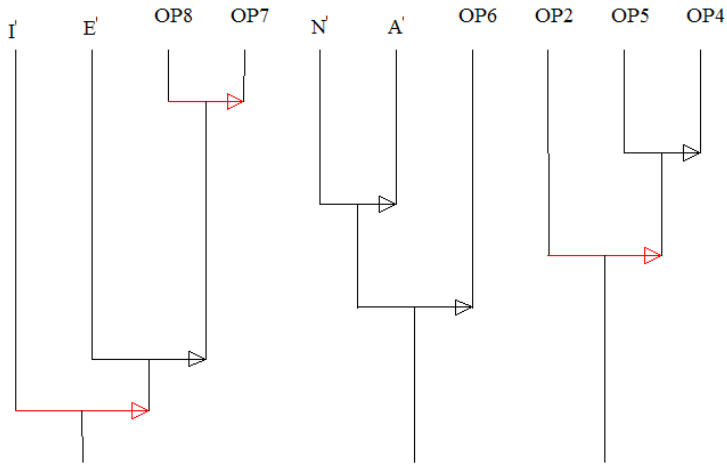


FIG. 11 *Hiérarchie orientée pour le questionnaire*

Des règles généralisées apparaissent sur cette hiérarchie et sont significatives :

- **(si N alors A) alors OP6**. Cette R-règle peut être lue ainsi : « si l'acquisition de savoir-faire (N) doit s'accompagner de celle des connaissances (A), alors le professeur demande que les programmes soient bien définis (OP6). On peut faire l'hypothèse que cette relation soit de type causal : « je suis très attaché aux contenus d'enseignement, mais inscrits dans l'institution, je demande que celle-ci définisse et précise ses choix ». Autrement dit, toute centration sur les savoirs exige un encadrement préalable de l'institution à travers des programmes. On observe alors que la règle généralisée permet de donner une signification plus synthétique aux règles qui la constituent : on passe des comportements, au sens behavioriste, à une conduite supérieure qui piloterait le comportement. Les enseignants qui considèrent que n'est pas pertinent l'objectif C ("contribuer à la préparation à la vie civique et sociale") sont les principaux responsables de cette règle ; ces enseignants possèdent donc une représentation de la formation mathématique très fermée sur la matière, dont l'enseignement, conformément aux programmes, n'est pas discutable.

- **si OP2 alors (si OP5 alors OP4).** Explicitement, cette R-règle se lit de la façon suivante : si l'on considère un grand problème indispensable à l'examen (OP2), alors, considérant que la démonstration est le seule façon rigoureuse de faire des mathématiques (OP5), le barème de correction doit être précis. On a là une évidente relation de type causal induite par une conception d'une catégorie d'enseignants très soumise à l'institution et conservatrice dans ses choix pédagogiques. Ne soyons pas surpris, la démonstration en France est le fondement de l'activité mathématique (pays de Descartes), tout en étant difficile à évaluer ; le grand problème en est le critère d'évaluation. On retrouve ici une image plus synthétique des règles d'association qui sous-tendent la R-règle, à savoir une conception de l'enseignement très classique qui exige un soutien institutionnel explicite et libérateur. Les enseignants de ES contribuent à cette association plus que les autres enseignants. Nous

reparlerons dans le chapitre suivant du critère dit de contribution qui nous permet d'établir cette affirmation.

-si I alors (si E alors (si OP8 alors OP7)) qui peut s'interpréter ainsi : si un enseignant choisit I (développement de l'esprit critique) et E (imagination et créativité) alors en général il considère que pour que l'élève découvre un caractère de divisibilité par 4 (OP7), il suffit qu'il ait été entraîné à trouver lui-même exemple et contre-exemple (OP8). D'une philosophie éducative ouverte vont découler des savoirs spécifiques laissant l'élève en situation de découverte, de construction personnelle. Cette règle est d'ailleurs constituée à un niveau significatif de la hiérarchie. Ce sont les enseignants des classes S qui contribuent le plus à son instauration. Elle met l'accent sur la relation entre des comportements non dogmatiques de l'enseignant et, en conséquence, la volonté de placer l'élève en situation de recherche personnelle. Ainsi, nous pouvons interpréter cette R-règle comme l'indice d'une conception d'ouverture didactique.

Insistons sur l'accroissement, dans chacun de ces cas, de la richesse de l'analyse obtenue par l'association des règles d'association en des R-règles. Ce ne sont plus seulement des faits ou des comportements isolés qui sont extraits, mais plutôt des conduites générales, révélatrices elles-mêmes de phénomènes plus globaux, moins singuliers ou de représentations psychologiques profondes. Une typologie comme en fournissent les classifications traditionnelles (donc symétriques), ne pourrait pas rendre compte de la dynamique des faits ou comportements sous-jacents. C'est pourtant cette dynamique restituée par les règles généralisées qui, appuyée sur des nécessités (les prémisses des règles), conduit à des élucidations vives d'un fragment de théorie, éventuellement en voie de construction.

En conclusion, quelles que soient les difficultés d'interprétation des règles généralisées, on constate le changement significatif d'information fournie par les classes orientées formées par rapport aux seules règles binaires de type $a \Rightarrow b$. Il s'agit ici de plonger l'ensemble des attributs dans une sorte de théorie globale, organisatrice des attributs où à chaque nœud on définit un théorème complexe dont la signification permet d'entrer plus profondément dans l'extraction des cohérences locales.

La construction d'une hiérarchie de règles généralisées par l'analyse implicative est implémentée dans le logiciel C.H.I.C.. Outre sa construction, un indice statistique, inspiré par celui que I.C. Lerman (1981b) a défini pour la hiérarchie des similarités, permet de mesurer la qualité de la hiérarchie globalement (ensemble des classes) et localement (nœuds). Nous avons construit un nouvel indice sur d'autres bases plus combinatoires (voir § 5 suivant).

Afin de permettre à l'expert qui doit faire l'analyse de la hiérarchie en intégrant la sémantique des variables, il semble absolument nécessaire de lui indiquer quelles sont le classes les plus pertinentes de la hiérarchie. C'est-à-dire celles où il doit porter son attention maximale, où l'interprétation serait plus cohérente avec le phénomène statistique qui a conduit, par les concepts jugés pertinents dans la théorie, à la formation de la classe. C'est dans cette intention que nous avons ensuite porté notre attention modélisante sur la construction successive des niveaux de la hiérarchie et, pour ce faire, défini un critère de significativité.

5 Significativité des niveaux d'une hiérarchie orientée

5.1 Critère de cohérence des niveaux

Une classe C de la hiérarchie orientée H 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. Pour évaluer précisément cette conformité, nous nous basons sur une propriété de l'intensité d'implication (Gras et Larher, 1992): si le nombre d'occurrences de a_i est inférieur au nombre d'occurrences de a_j, alors la qualité de $a_i \Rightarrow a_j$ au sens de φ est meilleure que celle de sa réciproque $a_j \Rightarrow a_i$. Ainsi, l'ordre théorique ω_t défini par les intensités d'implication mutuelles coïncide avec celui défini par les occurrences des attributs. Nous comparons la conformité entre ω_0 et ω_t avec celle entre un ordre aléatoire ω^* et ω_t. Nous mesurons la conformité 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.

Intuitivement cela signifie que, si α est petit, la conformité entre ω_0 et ω_t est invraisemblablement très grande puisqu'il paraît exceptionnel que le hasard « fasse mieux » que ce qui est observé.

Définition 17: La ***cohérence*** *o(C)* d'une classe C d'une hiérarchie orientée est définie par la probabilité $Pr(I > i)$.

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

Exemple 1 : Considérons une classe C d'une hiérarchie orientée H constituée de cinq variables, a_i, i =1 à 5 structurée selon l'ordre $\omega_0 = \{a_1, a_4, a_3, a_2, a_5\}$. On suppose d'autre part que leurs occurrences sont telles que $n_{a_1} < n_{a_2} < \ldots < n_{a_5}$, ce qui induit selon la propriété rappelée ci-dessus, un ordre théorique $\omega_t = \{a_1, a_2, a_3, a_4, a_5\}$ pour les intensités d'implication. On vérifie aisément que le nombre d'inversions i entre ω_0 et ω_t est *3* (échanges de a_3 et a_2, a_4 et a_3, a_4 et a_2). Afin d'évaluer la cohérence de la classe, il faut déterminer la loi de la variable I. Pour 5 variables on peut obtenir pas à pas la distribution en énumérant les cas où chacune des variables est minimale dans l'ordre ω^* (TAB.13). Ici, on a :

$$\Pr(I > i) = \Pr(I > 3) = \frac{91}{120} \approx \frac{3}{4}$$

Nombre d'inversions	0	1	2	3	4	5	6	7	8	9	10
a_1 minimal	1	3	5	6	5	3	1	0	0	0	0
a_2 minimal	0	1	3	5	6	5	3	1	0	0	0
a_3 minimal	0	0	1	3	5	6	5	3	1	0	0
a_4 minimal	0	0	0	1	3	5	6	5	3	1	0
a_5 minimal	0	0	0	0	1	3	5	6	5	3	1
Total des permutations	*1*	*4*	*9*	*15*	*20*	*22*	*20*	*15*	*9*	*4*	*1*

TAB. 13 – *Détermination de la distribution de I dans l'exemple*

D'une façon générale, la mise en œuvre de la cohérence définie en 3.1. nécessite de déterminer la loi de *I*, que nous noterons I_m dans la suite puisqu'elle dépend du nombre *m* de variables. Notons que le recours à la variable aléatoire I_m donnant le nombre d'inversions entre deux permutations est présent dans le calcul du coefficient de corrélation des rangs τ de Kendall qui peut effectivement s'écrire

$$1-\tau=\frac{4I_m}{m(m-1)} \qquad \textbf{(1)}$$

Cependant, à notre connaissance, la loi de I_m n'est ni explicitement donnée ni formalisée (Kendall et Stuart, 1991). Nous proposons et établissons, dans la suite, une formule de récurrence permettant de calculer ses valeurs dans l'indice de cohérence.

5.2 Loi de la variable I_m, nombre d'inversions dans une permutation

Sous l'hypothèse d'équiprobabilité des permutations, nous considérons la variable aléatoire $N(I_m(k))$ donnant le nombre total de permutations aléatoires correspondant à un nombre d'inversions avec ω_t égal à *k* pour un nombre de variables égal à *m*. Notons que l'on a trivialement $Pr(I_m = 0) = 1/m\ !$ puisque le nombre d'inversions est nul si et seulement si ω_t coïncide avec ω^*.

Proposition 7 Pour tout $k < m$, on a

$$N(I_m(k)) = \sum_{j=0}^{k} N\left(I_{m-1}(j)\right) \qquad \textbf{(2)}$$

et, pour tout $k \geq m$, on a

$$N(I_m(k)) = \sum_{j=k-m+1}^{k} N\left(I_{m-1}(j)\right) \quad \textbf{(3)}$$

Preuve.

Remarquons tout d'abord, que pour tout *k*, on a

$$N\left(I_m(k)\right) = \sum_{i=1}^{m} N\left(I_m(k); a_i\right) \qquad \textbf{(4)}$$

où $N(I_m(k) ; a_i)$ est le nombre total de permutations lorsque la variable a_i est minimale dans l'ordre aléatoire ω^* et placé au $i^{ème}$ rang dans l'ordre théorique ω_t.

Supposons maintenant que k < m. Pour tout i de 1 à m, la place minimale de a_i dans entraîne (i-1) inversions ; les k-(i-1) autres inversions sont donc provoquées par les m–1 autres variables. Ainsi, pour tout i de 1 à (k +1) on a

$$N(I_m(k)) = \sum_{i=1}^{k+1} N(I_{m-1}(k-(i-1)) = \sum_{i=0}^{k} N(I_{m-1}(k-i)) = \sum_{j=0}^{k} N(I_{m-1}(j)) \quad \textbf{(5)}$$

et, pour tout i de k+2 à m, $N(I_m(k) ; a_i) = 0$ puisque dans ce cas la variable a_i étant minimale il y a au moins i-1 inversions auxquelles aucune permutation ne peut conduire.

La preuve de la formule de récurrence pour le cas k≥l est basée sur un raisonnement similaire.

Les relations (2) et (3) de la proposition précédente permettent de calculer les lois des variables I_m selon une formule de récurrence. En effet, connaissant la distribution de I_{m-1} on peut déterminer celle de I_m, et les valeurs initiales sont directement calculables. On vérifie aisément que $N(I_2(0) ; a_1) = 1$ (c'est la permutation de a_1 et a_2), $N(I_2(1) ; a_1) = 0$; $N(I_2(0) ; a_2) = 0$, $N(I_2(1) ; a_2) = 1$, d'où $N(I_2(0)) = 1$ et $N(I_2(1)) = 1$. On en déduit ainsi la loi de I_2 : $Pr(I_2 = 1) = Pr(I_2 = 0) = 0.5$, puis celle de I_3, etc..

Proposition 8 : Pour tout k<m, on a $N(I_m(k))=N(I_{m-1}(k))+N(I_m(k-1))$ et, pour tout k ≥m, on a $N(I_m(k))=N(I_m(k-1))+N(I_{m-1}(k))-N(I_{m-1}(k-m))$.

Cette proposition se déduit d'arguments similaires à ceux employés dans la proposition précédente (Gras, 1997 a) et de la relation (4).

On peut ainsi déduire de façon récurrente les différentes valeurs de la loi de I_m utiles pour le calcul de la cohérence. En effet, pour k=1 et m>1, on déduit l'équation linéaire aux différences d'ordre 1 en m suivante : $N(I_m(1))= N(I_m(0))+N(I_{m-1}(1))=1+N(I_{m-1}(1))$.

D'où, $N(I_m(1))-N(I_{m-1}(1))=1$ dont la solution avec second membre est $N(I_m(1))=m-1$.

Par conséquent, rappelant que l'on a pour tout m>1, $Pr(I_m=0)=1/m$! on obtient :

$$Pr(I_m = 1) = \frac{m-1}{m!} \quad \textbf{(6)}$$

Ainsi, par exemple si m = 2, la cohérence associée à une situation sans inversion est :

$$Pr(I_2>0)=Pr(I_2=1)=0{,}5.$$

De la même façon, pour m > 2, on obtient en utilisant d'une part, le résultat donnant $N(I_m(1))$ et d'autre part, le fait que $N(I_m(1 ; a_m))=0$, la relation $N(I_m(2))=N(I_{m-1}(2))+m-1$. D'où, $N(I_m(2)) = m^2/2 – m/2 –1$, et pour m > 2, on a donc

$$Pr(I_m = 2) = \frac{m^2 - m - 2}{2m!} \quad \textbf{(7)}$$

Puis, comme $N(I_m(3))=N(I_{m-1}(3))+\frac{m^2}{2}-\frac{m}{2}-1$, on obtient pour m> 3 :

$$Pr(I_m(3)) = \frac{m^2 - 7}{6(m-1)!} \quad \textbf{(8)}$$

Par exemple, on a $Pr(I_6 \leq 2) = 0.028$. Par suite, $Pr(I_6 > 2)=0.972$ est la valeur de la cohérence d'une classe de *6* variables dans laquelle on observerait *2* inversions entre l'ordre associé à la classe ω_0 et l'ordre théorique ω_t .

Remarque 1. Pour une classe *C* réduite à un singleton, sa cohérence ne peut se déduire des relations précédentes. Nous posons dans ce cas *o(C)*=0,5 du fait que l'absence d'inversion n'apporte aucune information puisqu'elle est nécessaire.

Remarque 2. Lorsque deux variables d'une classe *C* ont le même nombre d'occurrences et ont donc ainsi le même rang dans le préordre ω_t , le nombre d'inversions qui leur sont relatives est calculé comme si les variables avaient un rang distinct.

Proposition 9 . L'espérance de la variable aléatoire I_m vaut $E(I_m) = \frac{m(m-1)}{4}$ et sa variance vaut $V(I_m) = \frac{m(m-1)(2m+5)}{72}$. Et la loi de probabilité de I_m converge vers une loi normale quand m tend vers l'infini. L'espérance et la variance peuvent se déduire, par la relation (1), des résultats connus de la statistique τ dont l'espérance est $E(\tau)=0$ et la variance $V(\tau) = \frac{2(2m+5)}{9m(m-1)}$. De plus, dans son article séminal (Kendall, 1938), Kendall expose les grandes lignes d'une démonstration permettant de déduire que, quand *m* tend vers l'infini, la variable

$$Z = \frac{\tau}{\sqrt{\frac{2(2m+5)}{9m(m-1)}}} \qquad \textbf{(9)}$$

est asymptotiquement distribuée comme une variable de Laplace-Gauss centrée réduite, dont il donne la table, à partir de ses moments centrés d'ordre pair, sous l'hypothèse d'équiprobabilité des permutations.

$$\mu_{2k} \approx \frac{(2k)!}{2^k k!}(\mu_2)^k \qquad \textbf{(10)}$$

Compte tenu de la relation entre I_m et τ, il est clair que la distribution de I_m est également asymptotiquement distribuée comme une variable gaussienne, dont l'approximation est acceptable à partir de m=10 (Siegel, 1956). A titre d'exemple, voici une comparaison :

$Pr(I_{20} = 50) \approx 0.0003197$ et $Pr(I_{20} > 50) \approx 0.9985028$.

Avec la loi de Laplace Gauss on obtient :

Pr[49.5<LG(95 ;15.41)<50.5]≈0.0003647 et Pr(LG(95 ; 15,41> 50,5) ≈ 0.998059

5.3 Vers un nouvel indice de significativité

A un niveau *k*>0 donné, une hiérarchie orientée H_A présente plusieurs classes déjà formées et associées à des *R*-règles de degré supérieur strictement à 0, et éventuellement, quelques variables non encore associés. Nous cherchons maintenant à quantifier la significativité d'une classe, ainsi que la qualité de la hiérarchie à ce niveau.

Afin de restituer l'information maximale relative à l'ensemble des classes constituées, cette significativité doit intégrer deux paramètres majeurs :

- les cohésions des classes dont, par construction de H_A, les valeurs décroissent avec la croissance des niveaux de la hiérarchie,

- les cohérences des classes qui peuvent croître ou décroître selon les niveaux en fonction de la probabilité associée à la variable aléatoire I_m, eu égard aux inversions observées et à la taille de la classe.

Le concept que nous proposons pour associer ces deux paramètres satisfait aux quatre contraintes suivantes liées à la « sémantique » de la significativité :

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

Définition 18 : L'indice *co* de **cohésion-cohérence** qui mesure la significativité de la classe C_{k+1} formée au niveau $k + 1$ est défini par

$$co(C_{k+1}) = \frac{c(C_{k+1})}{c(C_k)}.o(C_{k+1}) \quad (11)$$

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.

En effet, l'indice *co* n'étant pas une fonction monotone, il apparaît des maxima locaux correspondant d'une part à une meilleure adéquation entre les restrictions, à la classe formée à ce niveau, des préordres théorique ω_t et contingent ω_0, et d'autre part à une bonne cohésion.

Définition 19: La *qualité* de l'ensemble des niveaux h, $0 \le h \le k$, est définie par

$$q_k(H_A) = \left(\prod_{i=1}^{k} co(C_i) \right) \quad (12)$$

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.

5.4 Retour sur l'application

Conservant cette fois toutes les variables en jeu, la hiérarchie orientée obtenue avec le logiciel CHIC (Couturier et Gras 2005) comporte *16* niveaux (figure ci-dessous). Le tableau ci-dessous donne les cohésions des *R*-règles correspondantes. Par exemple, au niveau 13, la *R*-règle met l'accent sur la relation dérivée des comportements d'ouverture des élèves (I : esprit critique, E : imagination et créativité) vers des situations mathématiques les réalisant : OP8 : exemple et contre-exemple personnels, OP7 : test de réfutation). Cette interprétation globale est difficile par le seul emploi du graphe implicatif qui opère de façon binaire. Ainsi, il y a complémentarité et non redondance entre les deux approches.

Niveaux	R-règles	cohésion	Maxima locaux de l'indice co (cohésion-cohérence)
1	OP8 → OP9	0.998	0.499
2	OP5 → OP4	0.981	
3	N → A	0.955	
4	OP2 → (OP5 → OP4)	0.941	0.821
5	OP9 → OPX	0.92	
6	H → PER	0.92	0.5
7	F → OP3	0.903	
8	(N → A) → OP6	0.865	0.8
9	B → K	0.858	
10	E → (OP8 → OP7)	0.856	0.831
11	G → OP1	0.783	
12	J → (OP9 → OPX)	0.752	
13	I → (E → (OP8 → OP7))	0.707	0.752
14	C → O	0.669	
15	M → (F → OP3)	0.661	0.823
16	L → (J → (OP9 → OPX))	0.404	

TAB. 14 – *Cohésion des R-règles associées aux niveaux de la hiérarchie et maxima locaux de l'indice* co *de cohésion-cohérence*

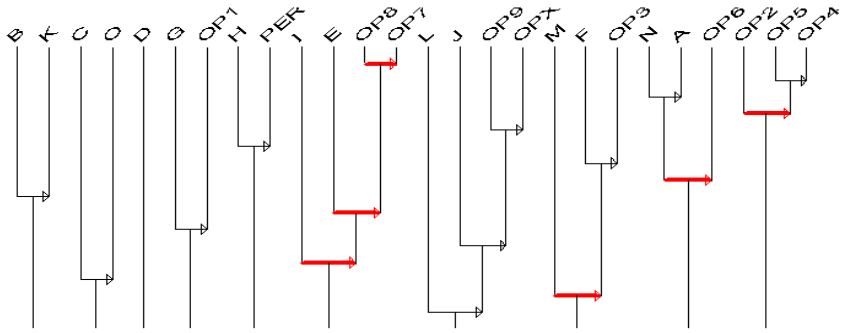


FIG. 12– *Hiérarchie orientée pour l'enquête auprès des professeurs de mathématiques*

Le calcul des cohérences à l'aide de l'algorithme implémenté dans CHIC conduit aux probabilités suivantes : $Pr(I_2 > 0) = Pr(I_2 \geq 1) = Pr(I_3 > 1) = 0.5$, $Pr(I_3 > 0) = 0.833$ et $Pr(I_4 > 2) = 0.625$. Une inversion seulement, par rapport aux occurrences, est observée pour les classes des niveaux *12*, *13* et *16*. Les maxima locaux de la cohésion-cohérence sont indiqués dans TAB 15. On observe également des maxima de l'indice de qualité *q* aux niveaux 1, *4*, *8*, *10*, *13* et *16*. Les niveaux significatifs, indiqués en gras sur la figure ci-dessus, à l'exclusion du niveau intéressant 6 (déclarer la non-pertinence des objectifs, c'est choisir de secondariser le développement de la volonté et la persévérance), avaient déjà été obtenus précédemment avec la méthode globale inspirée des travaux de Lerman (1981). Cependant, ce résultat n'a

pas valeur de généralité. Dans la situation expérimentale, la sémantique semble bien respectée dans les deux cas.

En conclusion, nous avons ici développé une approche complémentaire pour évaluer la significativité des niveaux d'une hiérarchie orientée et la qualité d'une hiérarchie orientée partielle qui tient compte du préordre défini sur les attributs de chaque *R*-règle constituée à chaque niveau de la hiérarchie. Cette approche ne nécessite pas, contrairement à une approche globale précédemment employée, la détermination d'une préordonnance sur l'ensemble des couples selon le critère de cohésion. De plus, lorsque le nombre *m* d'attributs « classés » devient grand, les calculs du nouveau critère peuvent être simplifiés par le recours à l'approximation à une loi de Laplace Gauss. Cette approche pourrait être généralisée à la recherche d'une mesure de distorsion entre deux permutations, prolongeant ainsi des travaux de Kendall. Mais, de nouvelles mises à l'épreuve sur des données réelles et, en particulier, des corpus de grande taille tels qu'on les trouve en fouille de données, permettront une comparaison plus robuste sur le plan de l'information restituée au cours des analyses que pourraient en faire des experts des domaines étudiés.

Chapitre 5 : Dualité entre variables actives et variables supplémentaires : typicalité et contribution[12]

1 Introduction

Après avoir étudié la fonction structurante (graphe, hiérarchie) du comportement des individus de E sur l'ensemble V des variables, nous nous interrogeons maintenant au sujet du rôle d'identificateurs de ces individus sur les éléments des différentes structures obtenues dans l'ensemble des variables. Afin d'y parvenir, dans ce chapitre, nous chercherons à établir une correspondance entre des éléments de certaines partitions de E et une structure de V du corpus de données, que ce soit un graphe ou une hiérarchie. Comme nous le verrons, cette correspondance devrait permettre d'opérer dans les deux sens individus↔variables au moyen de critères quantitatifs.

Cette analyse implicative de la correspondance entre deux ensembles E et V, une fois munis de métriques liées entre elles, est prometteuse et mériterait des développements que nous ne ferons qu'aborder ici. Nous mettrons cependant en évidence une dualité entre deux structures comme il est fait en Analyse Factorielle des Correspondances (A.F.C.), mais ici de façon plus modeste. Nous construirons pour cela, et dans un premier temps, une mesure visant à quantifier la « responsabilité » d'individus ou de groupes d'individus à l'élaboration des structures diversement et graduellement construites[13] sur l'ensemble des variables. Ces groupes homogènes rassemblent en une partition des individus sur la base d'un lien qui les identifie et les discrimine des autres comme, par exemple, le feraient une classe de collège ou une tranche d'âge. La « responsabilité » s'appliquera sur une structure, peut-être causale, obtenue dans l'ensemble des variables par l'A.S.I.. Inversement, sur la base de cette première correspondance, il sera possible d'identifier quelle(s) variable(s) peut ou peuvent caractériser tel individu ou tel groupe d'individus par rapport à cette structure dissymétrique donnée par l'A.S.I.. On cherchera enfin à détecter le couple individu(s)-variable(s) le plus spécifique du croisement des structures retenues sur E et sur V.

Pour ce faire, nous introduisons, en élargissant la problématique, la notion de **variable supplémentaire** en A.S.I. à l'instar de la même notion définie en A.F.C. (Benzecri, 1973). Il s'agit d'une variable exogène, un descripteur par exemple, n'intervenant pas directement dans l'établissement et la représentation (graphe ou hiérarchie) des liaisons exprimées par la classification et l'arborescence entre les variables dites principales de V. Il lui correspond, le plus souvent, une partition de l'ensemble des individus. Par exemple, une variable supplémentaire pourra représenter une catégorie de individus (âge, sexe, attitude, catégorie socio-professionnelle, etc.), mais aussi bien, si nécessaire, une des variables jugée principale, dans un premier temps..

---

[12] Ce chapitre a été présenté sous une forme voisine en atelier lors du Congrès EGC 6 sous le titre : « Typicalité et contribution des sujets et des variables supplémentaires en Analyse Statistique Implicative » avec pour auteurs : Régis Gras , Jérôme David, Jean-Claude Régnier, Fabrice Guillet. La publication de l'atelier figure dans *Volume 2, Cépaduès Editions(2006)*

[13] …graduellement construites car liées au seuil retenu par exemple dans l'élaboration des graphes implicatifs

Au cours de l'analyse, à un niveau quelconque de la hiérarchie se forme une classe C de cohésion non nulle ou un chemin C au seuil choisi. Notre objectif, particulièrement dans le cas d'un nœud significatif de la hiérarchie, est de définir un critère permettant d'identifier un ou des individus, puis la catégorie ou le groupe de individus, ou tout autre variable supplémentaire, qui seraient associés à l'apparition de cette classe ou de ce chemin, à savoir :

- ou bien des individus ou des variables supplémentaires **typiques** du comportement global de la population ; en d'autres termes, le comportement de ces individus ou de ces variables sera ainsi en harmonie avec le comportement statistique de la population à l'origine de la classe C,
- ou bien des individus ou des variables supplémentaires les plus **contributives**[14] c'est-à-dire *contribuant* formellement le plus à l'agrégation conduisant à C, c'est-à-dire en se référant à une population respectant formellement les règles constitutives de C, en d'autres termes, plus ou moins responsables de l'agrégation conduisant à C.

D'autres auteurs ont cherché à quantifier la relation qu'un individu et/ou un groupe d'individus entretiennent avec une ou des variables, par exemple, (Lerman,1981a) pour l'analyse des similarités, mais au moyen d'une modélisation et de concepts différents.

Notre approche se ramène à trouver des réponses originales aux questions suivantes :

1. quoi retenir de la forêt des règles qui sous-tendent un certain chemin d'un graphe ou une certaine classe d'une hiérarchie afin de conserver la « charpente » constituée des règles les plus consistantes ?
2. comment quantifier les positions respectives des éléments de E et de V eu égard à cette dominance ?
3. quel critère statistique permettrait de retenir avec un risque d'erreur minimal la variable supplémentaire la plus « responsable » de la « charpente » ?

Nous nous appuierons pour illustrer notre propos, sur l'exemple développé dans le chapitre 4 de la présente Partie 1. Nous rapportons les deux représentations graphiques :

[14] Les concepts et leurs propriétés définis ici diffèrent de ceux donnés dans (Gras et al., 1996 c) où l'on n'y distingue pas les deux notions « typicalité » et « contribution ».

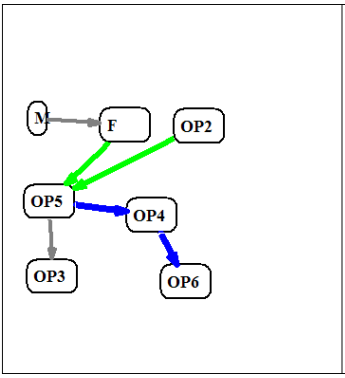


FIG. 13- *Graphe implicatif à 7 variables*

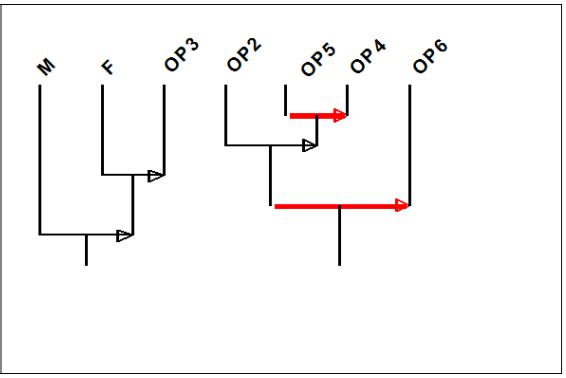


FIG. 14 *Hiérarchie cohésitive à 7 variables. Ces représentations portent sur l'exemple du questionnaire « enseignants » du chapitre 4*

2 Puissance implicative de classe et de chemin

2.1 Couples génériques

L'idée directrice suivie consiste à porter notre attention sur les « lignes de force », (ou, selon une autre métaphore : les « lignes de crête » ou la « charpente ») des règles d'association, plutôt que de les retenir toutes avec le risque afférent d'être submergé par leur nombre et de brouiller l'essentiel par les bruits confus qui les accompagnent. Plaçons-nous à un niveau k de la hiérarchie où viennent de se réunir, pour former C, deux classes $\underline{A}$ et $\underline{B}$ telles que $\underline{A} \Rightarrow \underline{B}$. Ainsi dans la figure ci-dessus rapportant la hiérarchie cohésitive, nous pouvons dire que :

- au niveau 2, on lirait $\underline{A} = OP2$ et $\underline{B} = (OP5, OP4)$;
- au niveau 4, on lirait : $\underline{A} = (\mathrm{OP2}, (\mathrm{OP5}, \mathrm{OP4}))$ et $\underline{B} = OP6$

Définition 20: Étant donné les intensités d'implication $\varphi(i, j)$, le couple (a,b) tel que:

$\forall i \in \underline{A}, \forall j \in \underline{B}, \varphi(a,b) \geq \varphi(i,j)$ est appelé **couple générique** de C[15], qui présente un caractère dominant sur le plan implicatif. En cas d'ex-æquos, on choisit le couple selon le critère des occurrences maximales.

Définition 21: Le nombre $\varphi(a,b)$ est appelé **intensité générique de la classe** C. C'est sa force qui domine les autres intensités des couples non retenus.

[15] C'est ce couple, généralement unique, qui intervient par le sup. dans le calcul de l'implication de $\underline{A} \Rightarrow \underline{B}$ (Gras et al, 1996 c).

Mais, dans chaque sous-classe de C, il existe également un couple générique. Précisément, si C est constituée de g (g≤k) sous-classes (C comprise), il y a g couples génériques à l'origine de C et g intensités maximales d'implication notées $\varphi_1, \varphi_2, \ldots, \varphi_g$, qui leur correspondent.

Dans le cas d'un chemin ch(i) du graphe implicatif, chemin fermé transitivement (chaque arc de la fermeture admet une intensité d'implication au moins égale à 0,50), composé de g nœuds, ch(i) présente $\frac{g(g-1)}{2}$ arcs transitifs. A chacun de ces arcs, par ex. (a,b), on associe, comme pour une classe, l'intensité d'implication de la règle correspondante.

Définition 22: Parmi les $\frac{g(g-1)}{2}$ intensités, l'une d'entre elles, au moins, est maximale et est appelée encore **intensité générique du chemin**.

Définition 23: Le vecteur $(\varphi_1, \varphi_2, ..., \varphi_g)$ de $[0,1]^g$, est appelé **vecteur puissance implicative** de la classe ou du chemin. Il traduit une force implicative interne à la classe ou au chemin. Ce vecteur a la propriété, en ne retenant, métaphoriquement, que les lignes de force (ou de crête) de représenter une sorte de « flux » implicatif au sein de la classe ou du chemin. Ce vecteur occupe une place stratégique par rapport à la classe ou au chemin. Toujours métaphoriquement, c'est un indicateur de la visibilité maximale d'un certain « courant » qui transporterait le flux implicatif au sein de la classe ou du chemin.

2.2 Puissance implicative d'un individu sur une classe ou sur un chemin du graphe implicatif et distance à cette classe ou à ce chemin

Dans le cas où les variables V sont binaires, un individu x quelconque respecte ou non l'implication du couple générique d'une classe ou d'un arc de chemin avec un ordre de qualité comparable. Associant logique formelle et considération sémantique, nous noterons $\varphi_x(a,b)$ cette qualité de respect en x de l'implication a=>b, par exemple et en fonction des valeurs prises en a et b par l'individu x :

$$\varphi_x(a,b) = 1 \text{ si a=1 ou a=0 et b=1}$$

$$\varphi_x(a,b) = 0 \text{ si a=1 et b=0}$$

$$\varphi_x(a,b) = p \text{ si a= b=0 avec } p \in]0;1]$$

Dans nos premières expériences, nous choisissions p=0,50, valeur neutre[16]. Ainsi, à l'individu x, nous pouvons associer g nombres notées $\varphi_{x,1}, \varphi_{x,2}, \ldots, \varphi_{x,g}$ correspondant aux g valeurs respectivement prises par x selon les g règles génériques de la classe ou du chemin.

[16] Dans le logiciel CHIC, pour des variables **modales** ou **numériques**, le calcul des typicalités (et des contributions) se fait cependant en modulant ces valeurs, à l'aide d'une fonction ad hoc, munie de propriétés adaptées afin de mieux prendre en compte la sémantique des valeurs attribuées par x à a et à b. Par exemple, pour a=0 et b=1, la fonction prend, dans CHIC, la valeur 0.682.

Définition 24 : Le vecteur $(\varphi_{x,1},\varphi_{x,2},...,\varphi_{x,g})$ est appelé **vecteur contingent générique** de l'individu x ou puissance implicative de x sur la classe ou le chemin.

Définition 25 : L'individu fictif, théorique x_t qui admettrait $(\varphi_1,\varphi_2,...,\varphi_g)$ comme vecteur contingent générique est appelé **individu typique optimal**.

En effet, on peut interpréter ce vecteur comme étant celui d'un individu « typique » des règles génériques puisque les valeurs prises par cet individu selon ces règles sont exactement celles de l'ensemble de la population. Cet individu, image conforme d'un individu fictif typique de E, n'existe pas réellement en général mais, s'il existe, il peut ne pas être unique. Dans ces conditions, on peut munir l'espace du produit $[0,1]^g$ d'une métrique afin d'obtenir un contraste accentuant les effets de fortes intensités génériques ou, réciproquement, minorant les effets d'une faible intensité générique.

Définition 26 : On appelle **distance de typicalité** d'un individu quelconque x à la classe ou au chemin C, le nombre $d(x,C)=\left[\frac{1}{g}\sum_{i=1}^{i=g}\frac{[\varphi_i-\varphi_{x,i}]^2}{1-\varphi_i}\right]^{\frac{1}{2}}$

Ce nombre, qui vérifie formellement les trois axiomes d'une distance, n'est autre également que la distance du type χ^2 entre les deux distributions $\{1-\varphi_i\}$ et $\{1-\varphi_{x,i}\}$ pour i=1 à g, qui expriment les écarts entre les implications génériques contingentes et l'implication stricte. Elle exprime, aussi et en particulier, l'écart observé sur les règles génériques entre l'individu x considéré et l'individu théorique typique optimal, écart nuancé par ces intensités. Elle exprime aussi la place que prend l'individu x par rapport à la classe ou au chemin C. C'est pour cette raison que nous avons choisi le mot **typicalité** pour quantifier le comportement de l'individu x selon les règles génériques. Nous allons le préciser plus loin.

Remarque 1 : Lorsque φ_i=1, une légère correction sur cette valeur permet d'éviter la division par zéro (par exemple, prendre φ_i = 0,99999999) ce qui ne change pas fondamentalement la distance.

Remarque 2 : Une classe C étant donnée, on peut définir une structure d'espace métrique sur E par la donnée de la distance indicée par C entre deux individus quelconques de E, distance qui mesure la différence de comportement des individus x et y à l'égard de C :

$$d_C(x,y)=\left[\frac{1}{g}\sum_{i=1}^{i=g}\frac{[\varphi_{x,i}-\varphi_{y,i}]^2}{1-\varphi_i}\right]^{\frac{1}{2}}$$

Cette distance établit une première correspondance entre l'ensemble des individus et l'ensemble des variables structuré par un graphe implicatif ou une hiérarchie cohésitive. On voit alors que la distance de typicalité donnée plus haut n'est que la spécification de d_C aux individus respectivement x et x_t. La distance d_C permet de conférer à E une d_C-structure topologique discrète. Cette topologie est équivalente à celle qui serait définie sur l'ensemble des vecteurs contingents $\vec{X}=(\varphi_{x,1},\varphi_{x,2},...,\varphi_{x,g})$, sous-ensemble d'un espace vectoriel normé de dimension g et de norme : $\left\|\vec{X}-\vec{Y}\right\|=d_C(x,y)$. L'opérateur symétrique associé à la forme

quadratique qui conduit à cette distance, a pour matrice, la matrice diagonale d'éléments $(g(1-\varphi_i))^{-1}$ pour i=1,…, g. Toutefois, l'interprétation de la somme de deux tels vecteurs n'a de sens que dans cadre théorique mathématique, c'est-à-dire hors du contexte dans lequel nous travaillons en A.S.I..

Une application intéressante peut consister à déterminer le ou les individus appartenant à une boule de diamètre donné et de centre l'un des individus pré-désignés, comme par exemple, l'individu optimal. En prolongement de cette approche métrique, le problème de complétion des données manquantes pourrait y puiser une solution originale. Nous aborderons (Partie 2, Chapitre 4) cette question, toutefois sous un angle un peu différent. Nous restons par ailleurs convaincus que les perspectives de recherche sur ce sujet sont nombreuses et stimulantes.

2.3 Typicalité, spécificité et contribution d'un individu ou d'une variable supplémentaire à une classe d'une hiérarchie cohésitive ou à un chemin d'un graphe implicatif

2.3.1 La notion de typicalité

Nous définissons la mesure de **typicalité** à partir du rapport entre la distance de typicalité relative à l'individu considéré et la distance à C, classe ou chemin, la plus grande dans l'ensemble des individus. Cette distance maximale est celle des individus y dont les $\varphi_{y,i}$ sont tous nuls ou très faibles. Ces individus sont donc ceux les plus opposés aux règles génériques. La typicalité d'un individu est alors d'autant plus grande qu'il s'écarte de ces mêmes individus, donc qu'il manifeste un comportement comparable à celui de l'individu théorique optimal. La typicalité d'une catégorie d'individus ou d'une variable supplémentaire G[17] s'en déduit alors :

Définition 27 : La typicalité de l'individu x à la classe ou au chemin C est mesurée par :

$$\gamma(x,C) = 1 - \frac{d(x,C)}{\max_{y \in E}\{d(y,C)\}}$$

Définition 28 : La typicalité de la variable supplémentaire G à la classe ou au chemin C est mesurée par :

$$\gamma(G,C) = \frac{1}{cardG}\sum_{x \in G}\gamma(x,G)$$

Afin de donner au chercheur le moyen de savoir ou de vérifier rapidement si telle catégorie d'individus qui l'intéresse est statistiquement déterminante dans la constitution d'une classe ou d'un chemin transitif, un algorithme a été élaboré en s'appuyant sur les deux notions que l'on définit ci-dessous : groupe optimal et catégorie déterminante.

Définition 29 : du **groupe optimal** d'une classe ou d'un chemin.

Soit E l'ensemble des individus étudiés. Un groupe optimal, noté GO(C), d'une classe de la hiérarchie cohésitive ou d'un chemin du graphe implicatif, noté C, est le sous-ensemble de

[17] Les deux mots « catégorie » et « variable supplémentaire » seront utilisés indifféremment, le premier ayant une charge sémantique plus forte que le second.

E qui accorde à C une typicalité plus grande que le complémentaire de GO(C) et qui forme avec celui-ci une partition en deux groupes maximisant la variance inter-classe de la série statistique des typicalités individuelles des individus les constituant (Ratsimba-Rajohn, 1992). Une telle partition est dite *significative*.

L'existence de ce groupe optimal est également démontrée dans (Gras R. et al., 1996 b, 1996 c). Les propriétés utilisées sont aussi celles qui le sont pour établir l'algorithme sur lequel se basent les modules des programmes informatiques qui construisent, automatiquement dans C.H.I.C., chaque sous-groupe optimal.

Considérons une partition $\{G_i\}_i$ de E. Cette partition peut être définie par une variable supplémentaire correspondant à une variable catégorielle, à un descripteur de E, admettant deux modalités binaires ou plus, par exemple, la variable « catégories socio-professionnelles ». Soit X_i une partie aléatoire de E de même cardinal que G_i et Z_i la variable aléatoire $Z_i = card(X_i \cap GO(C))$. Selon un modèle équiprobable admissible, Z_i suit une loi binomiale de paramètres : $card(G_i)$ et $\frac{card(GO(C))}{cardE}$ qui est la fréquence du groupe optimal de la classe ou du chemin C dans l'ensemble E.

Définition 30 : On appelle **variable supplémentaire**, ou catégorie, **la plus typique** de la classe ou du chemin C, celle qui minimise l'ensemble $\{p_i\}_i$ des probabilités p_i telles que:

$$\forall i, p_i = \Pr ob\{card(G_i \cap GO(C)) < Z_i\}$$

Ainsi, établir que G_j est la catégorie la plus typique revient à déceler, parmi les catégories, celle dont le nombre d'individus appartenant en même temps à celle-ci et au groupe optimal, est le *plus étonnamment grand eu égard à son cardinal*. Nous retrouvons ici la philosophie sous-jacente à la construction de l'indice d'implication.

Définition 31 : Une catégorie G_0 est dite déterminante au niveau de risque ou au seuil α si la probabilité associée p_0 est inférieure à α. Autrement dit, le risque de se tromper en affirmant cette propriété est donc au plus égale à α.

Par suite, la signification d'une classe ou d'un chemin ayant été donnée par l'expert, il lui associera le sous-ensemble le plus porteur de ce sens, celui correspondant au niveau de risque qu'il juge acceptable.

Groupe optimal et indice de similarité. Par ailleurs, nous pouvons associer au groupe optimal, une variable binaire définie par la fonction indicatrice de ce sous-ensemble de E. De la même façon, nous pouvons également associer à chaque catégorie G_i ou bien à la variable supplémentaire correspondante, une variable binaire dont l'indice de similarité $s = \frac{n_{a\wedge b} - \frac{n_a n_b}{n}}{\sqrt{\frac{n_a n_b}{n}}}$, au sens de I.C. Lerman, vérifie la condition $p_i = \Pr ob\{S \geq s\}$ où S est la variable aléatoire dont s est une réalisation. Ainsi, minimiser l'ensemble des probabilités $\{p_i\}_i$ revient à maximiser l'indice de similarité entre les variables binaires, indicatrices de

sous-ensembles, associées l'une au groupe optimal GO(C) et les autres aux différentes catégories $\{G_i\}_i$.

Cette remarque permet d'étendre efficacement la notion de variable supplémentaire la plus typique à des variables numériques, prenant leurs valeurs sur l'intervalle [0 ;1]. Il suffit alors d'extraire la plus forte des valeurs de similarité entre la variable binaire indicatrice définie par le groupe optimal et les différentes variables numériques placées en variable supplémentaire, l'indice de similarité étant calculé selon le principe retenu en analyse statistique implicative pour les variables numériques. Nous savons que sa restriction au cas binaire coïncide avec sa valeur s dans le cas où les 2 variables sont binaires.

Ainsi en résumé, il est possible de dégager à la fois les individus et les groupes d'individus typiques d'une règle ou d'un ensemble (classe ou chemin) de règles généralisées. Ce sont donc ceux qui sont le plus en accord avec la qualité de ces liaisons au sein de l'ensemble E considéré. Si, par exemple, la liaison implicative entre les variables a et b est quantifiée par $\varphi(a,b) = 0{,}92$, les individus x qui lui attribuent la valeur $\varphi_x(a,b) = 0{,}90$ sont plus typiques que ceux qui lui attribuent la valeur 0,98. Ceux-ci sont à une distance plus grande que les premiers pour le comportement statistique de l'ensemble E. La nuance entre cette notion et celle de contribution définie plus loin prend tout son sens dans l'étude des variables modales ou numériques.

2.3.2 La notion de spécificité

Si à chaque classe ou chemin C_j, on peut associer au moins un groupe typique, il est pertinent de mettre en évidence le couple (variable supplémentaire G_i , classe ou chemin C_j) remarquable quant à l'optimalité de sa conjugaison. D'où le recours à la notion de spécificité que nous introduisons par la définition suivante :

Définition 32 : La variable supplémentaire G_i étant donnée, le couple (G_i, C_j) est dit **mutuellement spécifique** lorsque G_i est la variable la plus spécifique de la règle associée à C_j et que la probabilité (le risque)

$$p_i^k = \Pr ob\left\{card(G_i \cap GO(C_k)) < Z_{i,k}\right\}$$

de G_i par rapport aux C_k, autres classes de la hiérarchie cohésitive ou autres chemins du graphe implicatif est supérieure à un seuil β (arbitrairement fixé par l'utilisateur).

Une analyse étant donnée, il peut n'exister aucun couple mais il peut aussi apparaître un ou plusieurs couples mutuellement spécifiques. Ce ou ces couples offrent l'intérêt de faire porter l'attention de l'expert sur les plus fortes associations prenant origine dans une variable supplémentaire.

Définition 33 : De façon analogue, un individu x étant donné, le couple (x, C_j) est mutuellement spécifique lorsque cet individu appartient au groupe optimal relatif à la règle associée à C_j et que la mesure de typicalité à C_j est maximale par rapport à toutes les autres valeurs de typicalité aux classes de la hiérarchie cohésitive ou aux chemins du graphe implicatif.

2.3.3 La notion de contribution

A ce jour, nous distinguons cette notion de celle de typicalité, ce que nous ne faisions pas en 1996. Cette distinction se manifeste dans la manière dont nous examinons la responsabilité des individus, puis des variables supplémentaires qui peuvent en être des descripteurs, *dans l'existence* d'une règle ou d'une règle généralisée entre variables principales.

Supposons que deux variables *a* et *b* (*resp.* plusieurs variables sur un chemin du graphe implicatif ou bien deux classes de la hiérarchie) soient réunies par un arc sur un graphe à un certain seuil (*resp.* en un chemin transitif C du graphe ou bien en une classe C dans une hiérarchie à un certain niveau). Connaissant la valeur $\varphi_{x,i}$ attribuée par l'individu x à la règle i : $a \Rightarrow b$ (resp. règle i du chemin C ou bien de la classe C constituée de g règles génériques) supposée admissible, nous posons alors la définition suivante.

Définition 34 : On appelle **distance de contribution** de x à l'arc (a,b) ou à C :

$$d(x,C) = \left[\frac{1}{g} \sum_{i=1}^{i=g} \left[1 - \varphi_{x,i}\right]^2 \right]^{\frac{1}{2}} \text{ où g=1 dans le cas de l'arc (a,b)}$$

Cette distance, de type euclidien, mesure l'écart entre le vecteur contingent générique de x : $(\varphi_{x,1}, \varphi_{x,2}, ..., \varphi_{x,g})$ et le vecteur à g composantes (1,1,1,…,1). Ce dernier est le vecteur d'un *individu théorique optimal qui satisferait strictement toutes les règles génériques.*

C'est donc en ce sens que les notions de typicalité et de contribution sont distinctes.

Toutefois à l'instar de ce que nous avons fait avec la notion de typicalité, nous pouvons définir sur E une topologie discrète d'espace normé dont la norme est associée à la distance entre deux individus quelconques suivante :

$$d_C(x,y) = \left[\frac{1}{g} \sum_{i=1}^{g} (\varphi_{x,i} - \varphi_{y,i})^2 \right]^{\frac{1}{2}}$$

Définition 35 : On appelle **contribution de x à C** le nombre : $\gamma(x,C) = 1 - d(X,C)$

Cette définition est la restriction de celle de la typicalité au cas où, cette fois, on compare l'individu x aux « pires » individus par rapport aux règles génériques : leur comportement s'oppose à l'implication de chaque règle (1 pour la prémisse et 0 pour la conclusion). Cette contribution a pour maximum 1 dans le cas où l'individu x a donné la valeur 1 à toutes les règles i. Ceci permet de concilier la sémantique avec la définition formelle. En effet, plus la différence est importante, plus l'individu observé a un comportement voisin de celui de l'individu théorique optimal et plus il s'éloigne de ceux qui réfutent les règles génériques. Nous pouvons donc dire qu'en contribuant à l'émergence de la classe, ils en sont aussi responsables.

La suite des définitions et des algorithmes de calcul (contribution d'une catégorie ou d'une variable supplémentaire G, groupe optimal d'individus, catégorie ou variable supplémentaire la plus contributive, couple mutuellement spécifique) se transpose immédiatement à partir des principes explicités pour la typicalité et la spécificité. Mais, dans les situations réelles, nous observons la nuance entre les deux concepts ce qui enrichit

l'information exploitable par l'utilisateur. Notons cependant que le concept de contribution est plus volontiers retenu pour l'interprétation dans une perspective inductive.

3 Application à l'étude du fichier Raf du chapitre 2

Reprenons l'exemple présenté dans le chapitre 2 dont le fichier RAF (30 sujets, 5 variables binaires, 2 variables supplémentaires : Fs et Gs) est rapporté dans le tableau (TAB 11). A titre didactique, nous allons conduire les calculs à la main pour que chaque définition et algorithme du présent chapitre soient clairement développés. Nous calculerons donc la mesure de typicalité des sujets à l'égard de la règle $V_3 => V_1$, c'est-à-dire tout autant à l'égard du chemin C ou arc $V_3 \rightarrow V_1$ qu'à celui de la classe $C=(V_3, V_1)$ puisque dans ce cas simplifié les deux représentations coïncident.

	V1	V3	F s	G s		V1	V3	F s	G s		V1	V3	F s	G s
i1	1	1	1	0	i11	1	1	1	0	i21	1	1	0	1
i2	1	1	1	0	i12	1	1	1	0	i22	1	1	0	1
i3	1	1	1	0	i13	1	1	1	0	i23	1	0	0	1
i4	1	1	1	0	i14	1	1	0	1	i24	1	0	0	1
i5	1	1	1	0	i15	1	1	1	0	i25	0	0	1	0
i6	1	1	1	0	i16	1	1	0	1	i26	0	0	0	1
i7	1	0	1	0	i17	1	1	0	1	i27	0	0	0	1
i8	1	0	1	0	i18	1	1	0	1	i28	0	0	0	1
i9	1	1	1	0	i19	1	1	0	1	i29	0	1	0	1
i10	1	1	1	0	i20	1	1	1	0	i30	0	0	0	1

TAB. 15- *extrait du tableau de données du fichier Raf*

Nous avons trouvé l'intensité d'implication (modèle de Poisson) de la règle $\varphi(V_3,V_1)$ =0,92.

Notons $v_x(V_k)$ la valeur que l'individu x prend en V_k, c'est-à-dire 1 ou 0 suivant que V_k est observée ou non en x.

Cas	$v_X(V_3)$=	$v_X(V_1)$=	$\varphi_x(V_3,V_1)$ =	Nombre d'individus
(a)	1	1	1	20
(b)	0	1	0,68 (voir note 16)	1
(c)	0	0	0,50	5
(d)	1	0	s0	4

TAB. 16

Ce qui est justifié par les données fournies dans le tableau ci-après :

	V3	V1	Cas		V3	V1	Cas		V3	V1	Cas
i1	1	1	(a)	i11	1	1	(a)	i21	1	1	(a)
i2	1	1	(a)	i12	1	1	(a)	i22	1	1	(a)
i3	1	1	(a)	i13	1	1	(a)	i23	0	1	(b)
i4	1	1	(a)	i14	1	1	(a)	i24	0	1	(b)
i5	1	1	(a)	i15	1	1	(a)	i25	0	0	(c)
i6	1	1	(a)	i16	1	1	(a)	i26	0	0	(c)
i7	0	1	(b)	i17	1	1	(a)	i27	0	0	(c)
i8	0	1	(b)	i18	1	1	(a)	i28	0	0	(c)
i9	1	1	(a)	i19	1	1	(a)	i29	1	0	(d)
i10	1	1	(a)	i20	1	1	(a)	i30	0	0	(c)

TAB. 17-

L'individu fictif x_t tel que $\varphi_{x_t}(V_3,V_1) = 0,92$ est un individu typique optimal

On rappelle que la distance implicative d'un sujet x à la classe C ayant g sous-classes, le nombre:

$$d(x,C) = \left[\frac{1}{g} \sum_{i=1}^{i=g} \frac{[\varphi_i - \varphi_{x,i}]^2}{1-\varphi_i} \right]^{\frac{1}{2}}$$

Dans ce cas $\varphi_i = \varphi(V_3,V_1)$ et $\varphi_{x,i} = \varphi_x(V_3,V_1)$ avec g=1 car il n' y a pas de sous-classe.

Nous obtenons donc les valeurs suivantes :

$v_x(V3)=$	$v_x(V1)=$	$\varphi_x(V_3,V_1) =$	d(x,C)
1	1	1	$\left[\frac{(1-0,92)^2}{1-0,92} \right]^{\frac{1}{2}} = 0,28$
0	1	0,68	$\left[\frac{(0,68-0,92)^2}{1-0,92} \right]^{\frac{1}{2}} = 0,85$
0	0	0,50	$\left[\frac{(0,50-0,92)^2}{1-0,92} \right]^{\frac{1}{2}} = 1,48$
1	0	0	$\left[\frac{(0,92)^2}{1-0,92} \right]^{\frac{1}{2}} = 3,25$

TAB. 18

Ainsi, $\max_{y \in E} (d(y,C)) = 3,25$

La mesure de la typicalité d'un individu x relativement à la classe (V_3,V_1) est donnée par la formule :

$$\gamma(x,C) = 1 - \frac{d(x,C)}{\max_{y \in E}\{d(y,C)\}}$$

(v_x(V1) ; v_x(V3))	(1 ; 1)	(0 ; 1)	(0 ;0)	(1 ; 0)
γ(x, C)=	$1-\frac{0,28}{3,25} \approx 0,913$	$1-\frac{0,85}{3,25} \approx 0,738$	$1-\frac{1,48}{3,25} \approx 0,544$	$1-\frac{3,25}{3,25} = 0$

La valeur moyenne de la mesure de typicalité d'un individu x relativement à la classe (V3,V1) est alors $\bar{\gamma} = \frac{1}{30}\left[(20.0,913)+(4.0,738)+(5.0,544)+1.0\right] \approx \frac{2,953}{30} \approx 0,798$

Catégorie la plus typique.

GO((V3,V1))=
{i1; i2; i3; i4; i5; i6; i9; i10; i11; i12; i13; i14; i15; i16; i17; i18; i19; i20; i21; i22}

	V1	V3	F s	G s		V1	V3	F s	G s
i1	1	1	1	0	i13	1	1	1	0
i2	1	1	1	0	i15	1	1	1	0
i3	1	1	1	0	i20	1	1	1	0
i4	1	1	1	0	i14	1	1	0	1
i5	1	1	1	0	i16	1	1	0	1
i6	1	1	1	0	i17	1	1	0	1
i9	1	1	1	0	i18	1	1	0	1
i10	1	1	1	0	i19	1	1	0	1
i11	1	1	1	0	i21	1	1	0	1
i12	1	1	1	0	i22	1	1	0	1

TAB. 19

Dans ce groupe, on trouve 13 sujets F_s ou filles parmi les 16 filles du fichier, et 7 sujets G_s ou garçons parmi les 14 garçons. La variable Z_1 qui représente le cardinal de l'intersection d'une partie aléatoire X_1 (de cardinal g_1=card(F)=16) avec le groupe optimal E_1 = GO((V3,V1)) de cardinal 20 suit la loi binomiale $B(16;\frac{20}{30})$. Le caractère typique de F relativement à E_1 est mesuré par la probabilité de dépasser dans une expérience aléatoire le nombre d'observations de filles dans le groupe optimal. Plus ce nombre est faible, plus il est "surprenant" de constater un tel effectif dans E_1.

$$Prob\{Z_1 > 13\} = \sum_{k=14}^{16} C_{16}^{k} (\frac{2}{3})^{k} (\frac{1}{3})^{16-k} \approx 0,0593$$

La variable Z_2 qui représente le cardinal de l'intersection d'une partie aléatoire X_2 (de cardinal g_2= card(G) = 14) avec le groupe optimal E_1 de cardinal 20 suit la loi binomiale $B(14;\frac{20}{30})$.

$$\Pr ob\{Z_2 > 7\} = \sum_{k=8}^{14} C_{14}^{k} (\frac{2}{3})^k (\frac{1}{3})^{14-k} \approx 0{,}8505$$

Par suite, la catégorie F est la plus typique relativement à la relation implicative $V_3 => V_1$, avec un niveau de risque de 0,0593 (niveau de risque de se tromper en affirmant qu'il ne s'agit pas de cette catégorie F = probabilité de trouver plus de 13 filles parmi les 16 dans un groupe de 20 sujets si la répartition fille-garçon se faisait au hasard dans E_1).

Remarquons que si nous prenions l'inégalité au sens large, nous obtiendrions les valeurs suivantes :

$$\Pr ob\{Z_1 \geq 13\} = \sum_{k=13}^{16} C_{16}^{k} (\frac{2}{3})^k (\frac{1}{3})^{16-k} \approx 0{,}166$$

$$\Pr ob\{Z_2 \geq 7\} = \sum_{k=7}^{14} C_{14}^{k} (\frac{2}{3})^k (\frac{1}{3})^{14-k} \approx 0{,}942$$

4 Application au questionnaire « Professeurs de Terminale »

Rappelons brièvement la situation développée dans le chapitre 4 de cette partie 1. Dans le cadre d'une enquête de l'Association des Professeurs de Mathématiques de l'Enseignement Public (APMEP) auprès de professeurs de mathématiques de classes terminales (séries scientifiques S et ES, littéraires LI et technologiques TE sont les variables supplémentaires), nous avons recueilli et analysé (Bodin et Gras., 1999) les réponses de 311 professeurs, à des classements (de 1 à 6) portant sur quinze objectifs qu'ils assignent à leur enseignement (A, B, C, …, O)[18] et sur leurs opinions relatives à dix phrases susceptibles d'être communément énoncées (OP1, OP2, ..., OPX)[19]. La variable PER donne la possibilité de désigner les objectifs jugés non pertinents. Les 26 variables correspondantes ne sont pas binaires, sauf PER, mais ordinales (valeurs {1, 0.8, 0.6, 0.4, 0.2, 0.1, 0} pour les objectifs et {1, 0.5, 0} pour les opinions). Ainsi l'analyse intègre l'intensité des attitudes, d'un choix prioritaire d'un objectif à un choix plus secondaire, voire non retenu.

Les variables supplémentaires sont : S(cientifique), ES(économique et sociale), LI(ttéraire), TE(chnologique).

Les occurrences des 30 variables sont les suivantes :

[18] Par exemple, E symbolise l'objectif : « développement de l'imagination et de la créativité »

[19] Par exemple, OP4 symbolise : « Pour corriger, j'aime bien un barème très détaillé sur les résultats à obtenir »

A	B	C	D	E	F	G	H	S s	ES s
105.7	8.8	9.7	140.0	21.8	138.7	19.5	44.8	155	68
I	J	K	L	M	N	O	PER	LI s	TE s
83.1	108.4	77.6	4.6	90.2	66.6	33.2	254	22	66
OP1	OP2	OP3	OP4	OP5	OP6	OP7	OP8	OP9	OPX
81.5	147.5	242.5	229.0	190.0	240.0	200.0	165.0	98.0	207.0

TAB. 20 – *Occurrences des variables de l'enquête sur les professeurs de mathématiques*

La hiérarchie cohésitive obtenue par CHIC à partir d'un nombre réduit des variables, afin de conserver les niveaux les plus significatifs, est donnée par la figure ci-dessous.

Considérons la classe C = [E ⇒ (OP8 ⇒ OP7)] ⇒ OPX. Son sens, analysé plus en détail dans (Bodin et Gras., 1999), est fortement marqué par l'importance accordée à l'imagination et à la recherche personnelle, par les enseignants d'accord avec ces objectifs et ces opinions, La variable la plus typique pour cette classe est S (série Scientifique) avec un niveau de risque de : 0,00393.

En effet, 116 des enseignants de S parmi les 155 de cette série qui ont répondu au sondage, figurent dans le groupe optimal (GO) de cardinal 201 relatif à C. Soit X une partie aléatoire de même cardinal (155) que S et Z la variable aléatoire égale au cardinal de l'intersection de X et du groupe optimal GO. Selon un modèle équiprobable de distribution des enseignants, Z suit la loi binomiale de paramètres 155 et 201/311, soit 0,656. La probabilité pour que Z soit plus grande que 116 est le risque annoncé, soit 0,00393. Mais pour S, c'est le couple (S, (OP8, OP7)) qui est mutuellement spécifique au seuil $\beta = 2.10^{-5}$. On retrouve une telle spécificité mutuelle pour TE avec le couple (TE, (B,K)) à un seuil $\beta = 5.10^{-7}$ nous confirmant, sans surprise, que les enseignants des sections techniques (TE) considèrent que les mathématiques doivent être utiles à la vie professionnelle (B) et, en conséquence, aux autres disciplines (K) et y sont les plus attachés.

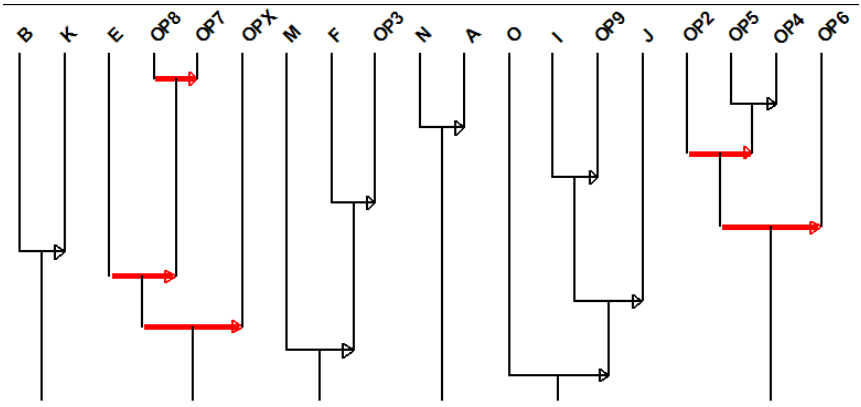


FIG. 15 - *Hiérarchie cohésitive significative*

Les calculs de contribution à la classe C montrent que, cette fois, 111 enseignants sur les 311 sondés, participent au groupe optimal. Le nombre d'enseignants de S a diminué (il passe de 116 à 67) et, surtout, sa proportion est bien moindre que précédemment dans le GO. Ceci

se ressent dans le seuil qui est 0,0251, soit un risque 6 fois plus élevé que pour la typicalité. Ce sont les enseignants sondés de S qui sont les plus typiques, c'est-à-dire « conformes » au comportement général de la population elle-même sondée. Mais ils sont moins contributeurs dans les relations strictes entre les 4 variables constituant C. Cette remarque nous montre les nuances apportées par les deux concepts : typicalité et contribution**.**

Certaines liaisons apparues et commentées ci-dessus se retrouvent dans le graphe. Les contributions calculées dans CHIC montrent encore que les enseignants de la série S contribuent le plus au chemin : $(E \Rightarrow (OP8 \Rightarrow OP7)) \Rightarrow OPX$ avec un risque d'erreur de 0,00746. La transitivité le long de ce chemin est assurée au niveau 0,75.

Ainsi on peut voir que les différents concepts mis en place ont permis d'établir une véritable dualité d'échanges d'informations d'un espace (celui des sujets) à un autre (celui des variables). Identifier le sujet ou le groupe le plus responsable de la formation de classe ou de chemin se ramène à l'identification de la variable ou du groupe de variables le ou les caractérisant.

Chapitre 6 : Règle et R-règle d'exception en Analyse Statistique Implicative ou encore l'exception confirme-t-elle la règle ?[20]

1 Introduction

On suppose une extraction de règles effectuée sur un ensemble de données binaires. Si l'on étudie localement avec attention les relations obtenues, on peut découvrir une situation parmi les associations qui défie l'intuition. C'est ainsi qu'il arrive que nous observions, entre trois variables (par exemple, des attributs) a, b et c, éventuellement conjonctions de variables binaires dans l'étude présente, les règles suivantes :a⇒c et b⇒c. Et que dans ce cas exceptionnel, on n'ait pas (a et b)⇒c, (que l'on peut aussi noter, $a \wedge b \Rightarrow c$) comme le bon sens l'attendrait, mais plutôt (a et b)⇒non(c) (que l'on peut aussi noter, $a \wedge b \Rightarrow \bar{c}$)). Cette dernière règle sera, de façon naturelle, appelée ici **règle d'exception**.

Rappelons que dans des travaux antérieurs (Suzuki et Kodratoff, 1999 ; Suzuki et Zytkow, 2005) les auteurs considèrent comme situation d'exception, la situation suivante :

a⇒c (dite règle de **sens commun**), non(b⇒c') (dite règle de **référence**) et (a et b)⇒c' (dite règle d'**exception**) où c≠c' et où *a* et *b sont respectivement des conjonctions (a = a_1 et a_2 et…et a_m)* et *(b= b_1 et b_2 et… et b_p)*. Notre définition de règle d'exception se distingue ainsi de celle-ci, mais présente, comme chez E. Suzuki et Y. Kodratoff, un caractère surprenant. Son expression plus simple lui permet d'être mieux saisie par l'intuition.

Or, il existe, comme nous le verrons en donnant des exemples, des situations naturelles où un caractère exceptionnel associe les trois variables. Pour le prendre en compte et en étudier un modèle, nous étendons ici le sens précédent en accentuant ainsi le caractère surprenant (le caractère *d'exception*) d'une règle dérivée de deux règles simples.

Pour illustrer ce type de règle, nous faisons référence tout d'abord au cas de l'incompatibilité de groupes sanguins en ce qui concerne le facteur Rhésus. Certaines femmes, non primo-parturientes, dont les globules rouges sont porteurs de deux allèles Rh- et dont l'immunisation anti-Rh+ est active, possèdent alors le phénotype Rh- (caractère *a*). Quel que soit le père en général, l'enfant qu'elles portent ne présentera pas, à la naissance, de problème sur le plan sanguin (caractère *c*). Nous sommes en présence de la règle : *a⇒c*.

Un homme, de génotype Rh+ et Rh+, possède le phénotype Rh+ (caractère *b*). Quelle que soit la mère en général, l'enfant qu'il engendrera n'aura pas de problème à sa naissance (caractère *c*). C'est la situation où la règle *b⇒c* est valide.

En revanche, un couple où la femme est Rh- et remplit les conditions *a* et l'homme est Rh+ (caractère *b*) pourra donner naissance à un enfant qui présentera un risque important du fait de l'incompatibilité Rhésus (caractère $\bar{c}$). Dans des cas exceptionnels, en effet, la mère s'immunisant contre le facteur Rh du fœtus, fabrique des anticorps, qui détruisent les globules rouges de l'enfant. Même si la conjugaison des caractères *a* et *b* est rare, on

[20] Ce chapitre a aussi été publié dans les Actes de ASI 4, Castellon, sous une forme et un contenu voisins, avec le titre : "Règle et R-règle d'exception en Analyse Statistique Implicative", Régis Gras, Einoshin Suzuki, Pascale Kuntz.

rencontre cependant la réalisation de la règle, que nous appelons « règle d'exception », (*a* et *b*) $\Rightarrow \overline{c}$. On sait d'ailleurs que des précautions sont prises pour éviter ce problème dès que sont connus les phénotypes des parents grâce à une prévention adaptée (par ex. l'exsanguino-transfusion).

On trouve une situation comparable d'apparition de règle d'exception dans l'étude des phénomènes d'interférences lumineuses, par exemple, dans l'expérience classique des franges de Young (Bruhat, 1959). La même source lumineuse franchissant deux fentes identiques (*a* et *b*) conduit à des franges d'interférences où alternent des zones d'intensité lumineuse (*c*) variable susceptible de faiblir et/ou s'annuler ($\overline{c}$).

Ces deux exemples montrent l'intérêt d'examiner le fonctionnement, la représentation et surtout les conditions d'apparition des règles d'exception afin de se prémunir en A.S.I. des inférences un peu rapides sur la stabilité implicative de la conjonction de règles simples et, comme nous allons le constater de règles généralisées ou R-règles.

2 Interprétation et illustration des règles d'exception

Soient *A*, *B*, *C* et $A \cap B$ respectivement les sous-ensembles d'individus de *E* qui satisfont respectivement les variables *a*, *b*, *c* et (*a* et *b*). Dans la situation illustrée ici, elles sont binaires, mais l'A.S.I. permet de considérer également d'autres types de variables (Gras, 2005).

2.1 Deux approches pour la caractérisation des règles d'exception

Supposons la situation prototypique des règles d'exception : $a \Rightarrow c$, $b \Rightarrow c$ et (*a* et *b*) $\Rightarrow \overline{c}$ (alors que (*a* et *b*) $\Rightarrow c$ est de piètre qualité). Elle s'exprime, en termes ensemblistes, par une quasi-inclusion des ensembles d'instances à savoir : *A* et *B* sont presque contenus dans C, mais $A \cap B$ est plutôt contenu dans le complémentaire de *C*. L'illustration ci-dessous rend compte de la situation ensembliste.

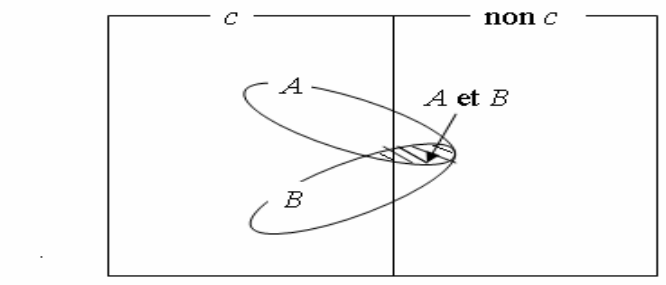


FIG. 16- – *Apparition d'une règle d'exception ensembliste*

Dans le cadre de l'A.S.I., deux approches pourraient nous permettre de mettre en évidence cette situation.

2.1.1 La première approche

Elle est basée sur l'analyse de l'intensité d'implication φ(*a*,*b*) selon la théorie présentée dans le chapitre 1 de cette partie 1. Elle nous permet de conclure au rejet de (a et b)⇒c et, a contrario, à l'apparition d'une intensité, non négligeable quelquefois, de (*a* et *b*)⇒$\bar{c}$ qui en justifie la prise en compte en tant que règle d'exception. Une représentation en graphe des relations implicatives entre règles élémentaires ci-dessus sera illustrée plus loin.

2.1.2 La deuxième approche

Elle est basée sur l'extension, que nous avons proposée et exposée au chapitre 4, des règles en *R*-règles (règles de règles) de type *R*⇒*R'* où *R* et *R'* sont elles-mêmes des règles (Gras et Kuntz, 2005). Rappelons la métaphore intuitive : ces règles sont comparables à celles qui apparaissent en mathématiques où un théorème *R* a pour conséquence un autre théorème *R'* ou est suivi d'un corollaire *R'*. Bien évidemment, il ne s'agit que d'une métaphore puisque en A.S.I. on considère des règles partielles, qui ne sont donc pas strictes et ne relèvent de la logique formelle qu'exceptionnellement. Nous avons vu qu'elles étaient construites selon un algorithme récursif utilisant un indice appelé « cohésion » et que celui-ci rendait compte de la qualité des liaisons implicatives des variables de la règle *R* avec les variables de la règle *R'*.

Rappelons, qu'en logique formelle, la règle généralisée, règle de règles, ou *R*-règle, *a*⇒(*b*⇒*c*), composée de la règle R_1 =(*b*⇒*c*) et R_2 = *a*⇒R_1 est vraie en même temps que (a et b)⇒c, donc lui est logiquement équivalente, où les variables *a, b* et *c* peuvent être elles-mêmes des règles (Gras et Kuntz., 2005). Or, nous avons vu, dans l'examen des règles élémentaires de la forme α⇒β que (*a* et *b)* ⇒ $\bar{c}$, règle d'exception, est, généralement, en contradiction sémantique avec (*a*⇒*c* et *b*⇒*c*) et que cette conjonction est plutôt formellement compatible avec (*a* et *b*) ⇒*c.*

De la même façon, la R-règle *a*⇒ *(b*⇒*c*) est en contradiction formelle avec (*a* et *b*) ⇒ $\bar{c}$. Mais comme nous sommes dans le cadre de l'A.S.I. où les règles sont partielles, cette dernière règle peut apparaître bien qu'elle soit inattendue. Nous dirons alors, comme précédemment, que (*a* et *b*)⇒$\bar{c}$ est une règle d'exception de la R-règle *a*⇒*(b*⇒*c*). Un arbre hiérarchique, présenté plus loin, permet d'illustrer cette approche par des R-règles.

2.2 Exemple numérique à partir de données fictives

Nous avons construit un fichier fictif de 200 sujets sur lesquels nous observons les variables binaires : a, b , a ∧ b, c et non(c) dont nous rapportons un extrait des 20 premiers sujets.

sujets	a	b	a ∧ b	c	non(c)
1	1	1	1	0	1
2	1	1	1	0	1
3	1	1	1	0	1
4	0	1	0	0	1
5	1	0	0	1	0
6	0	1	0	1	0
7	1	0	0	1	0
8	0	1	0	1	0
9	1	0	0	1	0
10	0	1	0	1	0

sujets	a	b	a ∧ b	c	non(c)
11	1	0	0	1	0
12	0	1	0	1	0
13	1	0	0	1	0
14	0	1	0	1	0
15	1	0	0	1	0
16	0	1	0	1	0
17	1	0	0	1	0
18	0	1	0	1	0
19	1	0	0	1	0
20	0	1	0	1	0
….	….	….	….	….	….

TAB. 21

Ce sont les 4 premiers qui vont principalement intervenir dans l'apparition de la règle d'exception. Les valeurs associées des différentes intensités sont données dans TAB 23. Elles sont obtenues par le logiciel CHIC (Couturier et Gras, 2005) qui permet les calculs et les représentations graphiques des ensembles de règles extraites des instances,

	a	*b*	*c*	$\bar{c}$	$a \wedge b$
a	0	.79	.89	.08	.89
b	.79	0	.84	.10	.88
c	.68	.67	0	0	.36
$\bar{c}$	.32	.33	0	0	.64
$a \wedge b$	1.00	1.00	.03	.97	0

TAB. 22–*Intensités d'implication associées à un jeu de données*

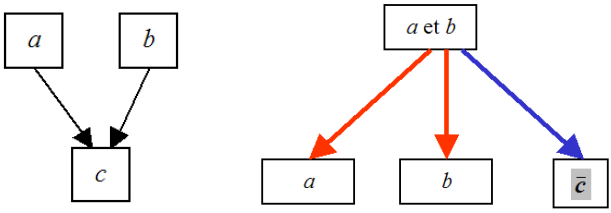
Notons les fréquences des occurrences des variables : $n_a = n_b = 12$; $n_{a \wedge b} = 7$; $n_c = 50$. Les intensités d'implication associées sont :

$$\varphi(a,c) = 0{,}89 \text{ ; } \varphi(b,c) = 0{,}84 \text{ ; } \varphi((a \text{ et } b), c) = 0{,}03$$

alors que $\varphi((a \text{ et } b), \bar{c}) = 0{,}97$; ce qui confirme la présence d'une règle d'exception.

2.2.1 Selon la première approche de règles élémentaires

Une analyse par CHIC sur le tableau complet donne le graphe implicatif et l'on constate la bonne qualité d'implication de *a* et de *b* sur *c*. On constate aussi que l'on a bien $a \Rightarrow c$ et $b \Rightarrow c$. Lorsque CHIC conjoint les variables, on obtient cette fois le phénomène lié à l'existence d'une règle d'exception.



Règles simples *Apparition de la règle d'exception*

FIG. 17– *Graphe implicatif avec les règles simples puis conjointes*

D'une façon générale en A.S.I., trois conditions nous semblent favorables à l'apparition d'une règle d'exception de $a \wedge b$ sur non c :

4. Une certaine qualité d'implication de a et de b sur c ; cette condition de bon sens conduit à ce que la règle $(a \wedge b) \Rightarrow c$ soit attendue et non pas $(a \wedge b) \Rightarrow \bar{c}$ qui en définitive va l'être;
5. Une mauvaise qualité de ressemblance entre $(a \wedge b)$ et c ($n_{a \wedge b \wedge c}$ est faible) ;
6. Une bonne qualité de ressemblance entre a et b si le référentiel devient $\bar{c}$ ($n_{a \wedge b \wedge \bar{c}}$ est grand relativement à $n_{a \wedge b}$).

2.2.2 Selon la deuxième approche

Selon la deuxième approche, relativement à cet exemple numérique, $a \Rightarrow b$, faiblement (.81), et $a \Rightarrow c$ et $b \Rightarrow c$ un peu plus fortement (.85 et .89). Par suite, la R-règle $a \Rightarrow (b \Rightarrow c)$ est validée par l'ASI et restituée au moyen du logiciel CHIC. Les arbres cohésitifs suivants illustrent respectivement d'une part la règle généralisée $a \Rightarrow (b \Rightarrow c)$ où l'on voit que a et b n'ont aucune relation implicative avec non(c), d'autre part que la règle d'exception $(a$ et $b) \Rightarrow \bar{c}$ est obtenue lorsque l'on conjoint a et b.

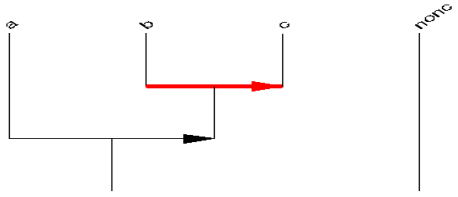


FIG. 18- *Représentation hiérarchique de la R-règle* a⇒(b⇒c)

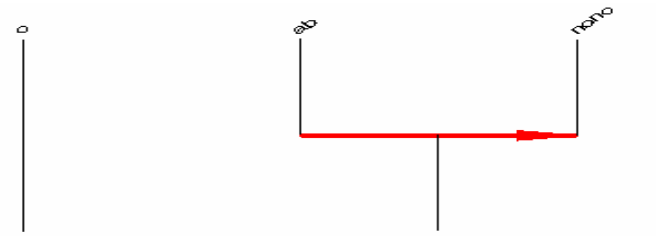


FIG. 19- *Représentation hiérarchique de la R-règle d'exception (a et b) $\Rightarrow \bar{c}$*

Autrement dit, l'absence de cohérence entre l'arbre extrait des données et l'arbre après conjonction des variables témoigne de l'apparition de la règle d'exception. Celle-ci a pu être observée de façon analogue à travers l'absence de cohérence entre les deux représentations des graphes implicatifs.

Ainsi, comme pour l'approche graphique ci-dessus, la double construction de la hiérarchie implicative : variables élémentaires puis variables conjointes, permet par examen de la non-cohérence de repérer l'existence d'une R-règle d'exception.

3 Relation entre les intensités d'implication de *a et b* sur *c* et sur non(*c*)

Rappelons ce nous avons exposé en détail dans le chapitre 1 de cette partie 1, que, en A.S.I., nous modélisons l'implication de *a* sur *b* de deux manières :

1. par une loi de Poisson de paramètre estimé $\lambda = \frac{n_a n_{\bar{b}}}{n}$;
2. par une loi binomiale de paramètres *n* et $p = \frac{n_a n_{\bar{b}}}{n^2}$

Une modélisation hypergéométrique est écartée car elle n'induit pas de différence entre une implication et sa réciproque (Gras et al., 1996 c)

Établissons pour chacun de ces deux modèles retenus les intensités d'implication de la conjonction $a \wedge b$ sur les variables *c* et non(*c*) (encore notée $\bar{c}$). Nous utiliserons la relation simple : $n_{a \wedge b \wedge \bar{c}} = n_{a \wedge b} - n_{a \wedge b \wedge c}$.

3.1 L'intensité d'implication de la conjonction dans le modèle de Poisson

Il convient d'expliciter le calcul des intensités d'implication mettant en jeu des conjonctions de variables.

3.1.1 Première approche par règles élémentaires

Dans ce modèle, pour respectivement l'implication $a \wedge b \Rightarrow \bar{c}$ et l'implication $a \wedge b \Rightarrow c$, les indices $q_1(a \wedge b,$ c) et $q_2(a \wedge b, \bar{c})$ sont :

$$q_1 = \frac{n_{a\wedge b\wedge c} - \frac{n_{a\wedge b}.n_c}{n}}{\sqrt{\frac{n_{a\wedge b}.n_c}{n}}} \qquad \text{et} \qquad q_2 = \frac{n_{a\wedge b\wedge \bar{c}} - \frac{n_{a\wedge b}.n_{\bar{c}}}{n}}{\sqrt{\frac{n_{a\wedge b}.n_{\bar{c}}}{n}}} \qquad \textbf{(1)}$$

Pour que l'implication $a \wedge b \Rightarrow \bar{c}$ soit de bonne qualité, il est nécessaire que q_1 soit négatif. En effet, le nombre de contre-exemples observé $n_{a\wedge b\wedge c}$ doit être inférieur à celui auquel seul le hasard pourrait conduire, dans l'hypothèse d'indépendance de $a \wedge b$ et de c, soit la moyenne $\frac{n_{a\wedge b}.n_c}{n}$. Des formules (1), on déduit que

$$q_1 = \frac{n_{a\wedge b} - n_{a\wedge b\wedge \bar{c}} - \frac{n_{a\wedge b}.n_c}{n}}{\sqrt{\frac{n_{a\wedge b}.n_c}{n}}} = - \frac{n_{a\wedge b\wedge \bar{c}}.n - n_{a\wedge b}(n - n_c)}{\sqrt{n.n_{a\wedge b}.n_c}}$$

$$= \frac{n_{a\wedge b\wedge \bar{c}} - \frac{n_{a\wedge b}.n_{\hat{c}}}{n}}{\sqrt{\frac{n_{a\wedge b}.n_c}{n}}} = -q_2 . \frac{\sqrt{\frac{n_{a\wedge b}.n_{\bar{c}}}{n}}}{\sqrt{\frac{n_{a\wedge b}.n_c}{n}}}$$

Finalement : $q_1 = -q_2\sqrt{\frac{n_{\bar{c}}}{n_c}}$ ou encore que le rapport $\frac{q_1}{q_2} = -\sqrt{\frac{n_{\bar{c}}}{n_c}}$.

q_1 et q_2 sont bien de signes opposés, ce qui est conforme à l'intuition. Mais de plus, l'amplitude de la positivité de q_2 induit celle de la négativité de q_1.

Au sens de l'intensité d'implication (classique), pour que la règle $a \wedge b \Rightarrow \bar{c}$ soit considérée comme une exception et apparaisse, la différence $\varphi(a \wedge b, \bar{c})$ - $\varphi(a \wedge b, c)$ suivante doit être positive et suffisamment grande :

$$\frac{1}{\sqrt{2\pi}} \int_{q_1}^{+\infty} e^{-\frac{t^2}{2}} dt - \frac{1}{\sqrt{2\pi}} \int_{-q_1\sqrt{\frac{n_c}{n_{\bar{c}}}}}^{+\infty} e^{-\frac{t^2}{2}} dt = \frac{1}{\sqrt{2\pi}} \int_{q_1}^{-q_1\sqrt{\frac{n_c}{n_{\bar{c}}}}} e^{-\frac{t^2}{2}} dt \qquad \textbf{(2)}$$

Proposition 10 Il y a donc apparition de règles d'exception, lorsque q_1 est négatif, c'est-à-dire lorsque $\frac{n_{a\wedge b\wedge c}}{n} < \frac{n_{a\wedge b}}{n}.\frac{n_c}{n}$ (sous indépendance) et de qualité d'autant meilleure que l'ensemble C des instances satisfaisant c est supérieur à celui qui satisfont sa négation non(c). La force de l'intensité d'exception sera à la mesure de la valeur de l'intégrale gaussienne sur l'intervalle $[q_1 ; -q_1\sqrt{\frac{n_c}{n_{\bar{c}}}}]$. De même, règle attendue et règle d'exception coïncident lorsque $q_1 = q_2 = 0$.

Ainsi, c'est à l'occasion de l'indépendance de $a \wedge b$ et $\bar{c}$ et donc de $a \wedge b$ et c que disparaît la règle d'exception.

3.1.2 Deuxième approche par R-règles

Dans le cadre de l'approche par R-règles, considérant la règle ($b \Rightarrow c$) comme une variable binaire, c'est-à-dire prenant respectivement la valeur 0 lorsque b=1 alors que c=0 et la valeur 1 dans les autres cas, les contre-exemples à la règle sont en nombre $n_{b \wedge \bar{c}}$. Dans ces conditions, les contre-exemples à la R-règle $a \Rightarrow (b \Rightarrow c)$ apparaissent lorsque a prend la valeur 1 alors que ($b \Rightarrow c$) prend la valeur 0, c'est-à-dire lorsque $b \wedge \bar{c}$ prend la valeur 1. Par suite, le nombre de ces contre-exemples est $n_{a \wedge b \wedge \bar{c}}$ et l'indice d'implication associé à la R-règle, dans une modélisation de Poisson de l'implication statistique, est alors :

$$q_3 = \frac{n_{a \wedge b \wedge \bar{c}} - \frac{n_a . n_{b \wedge \bar{c}}}{n}}{\sqrt{\frac{n_a . n_{b \wedge \bar{c}}}{n}}} = \frac{n . n_{a \wedge b \wedge \bar{c}} - n_a . n_{b \wedge \bar{c}}}{\sqrt{n . n_a . n_{b \wedge \bar{c}}}}$$

On constate que cet indice est différent de celui associé à la règle $a \wedge b \Rightarrow c$ qui est la règle élémentaire attendue de la conjonction $a \Rightarrow c$ et $b \Rightarrow c$. En conséquence, les indicateurs qui nous permettront de prévoir l'existence d'une règle d'exception dans cette approche hiérarchique seront différents de ceux qui nous permettent d'anticiper l'exception dans l'approche par le graphe implicatif.

Rappelons alors que l'indice d'implication de (a et b) $\Rightarrow \bar{c}$ est :

$$q_1 = \frac{n_{a \wedge b \wedge c} - \frac{n_{a \wedge b} . n_c}{n}}{\sqrt{\frac{n_{a \wedge b} . n_c}{n}}} = \frac{n . n_{a \wedge b \wedge c} - n_{a \wedge b} . n_c}{\sqrt{n . n_{a \wedge b} . n_c}}$$

On démontre, par transformation des deux indices, que q_1 est négatif (donc la règle d'exception est valide) alors que q_3 est positif (donc la règle attendue n'apparaît pas) si et seulement si : $\frac{n_{a \wedge b}}{n_a} > \frac{n_{b \wedge \bar{c}}}{n_{\bar{c}}}$, c'est-à-dire si la fréquence conditionnelle de b dans a est supérieure à sa fréquence dans $\bar{c}$.

Proposition 11 : $q_1 < 0$ et $q_3 > 0 \Leftrightarrow \frac{n_{a \wedge b}}{n_a} > \frac{n_{b \wedge \bar{c}}}{n_{\bar{c}}}$

Inversement, si cette inégalité est observée dans l'autre sens, la règle attendue apparaît alors que la règle d'exception n'existe pas.

Dans l'exemple numérique qui illustre la règle d'exception $a \wedge b \Rightarrow \bar{c}$, nous avons :

a) d'une part : $n_{a \wedge b} = 7$, $n_b = 24$, soit $\frac{n_{a \wedge b}}{n_a} = 0{,}29$,

b) d'autre part : $n_{b \wedge \bar{c}} = 8$, $n_{\bar{c}} = 100$ soit $\frac{n_{b \wedge \bar{c}}}{n_{\bar{c}}} = 0{,}08$.

L'inégalité est bien satisfaite.

Ce résultat analytique est différent de celui obtenu par l'approche graphique ce qui confirme bien la différence de signification des deux représentations de l'implication.

3.2 L'intensité d'implication de la conjonction dans le modèle binomial

Posons q'$_1$ et q'$_2$ respectivement les indices respectifs d'implication de $a \wedge b \Rightarrow \bar{c}$ et $a \wedge b \Rightarrow c$, lorsque le modèle de tirage aléatoire des parties *A*, *B* et *C* est binomial. Dans ce cas, par un calcul comparable au modèle précédent, on obtient

$$\frac{q'_1}{q'_2} = - \sqrt{\frac{n_{\bar{c}}\,(n^2 - n_{a\wedge b}.n_{\bar{c}})}{n_c\,(n^2 - n_{a\wedge b}.n_c)}} \qquad \textbf{(3)}$$

Posons

$$k(a,b,c) = \left[\frac{\left(1 - \frac{n_{a\wedge b}.n_{\bar{c}}}{n^2}\right)}{\left(1 - \frac{n_{a\wedge b}.n_c}{n^2}\right)} \right]^{\frac{1}{2}} \qquad \textbf{(4)}$$

Donc,

$$\frac{q'_1}{q'_2} = - \sqrt{\frac{n_{\bar{c}}}{n_c}}.\,k(a,b,c) \qquad \textbf{(5)}$$

et la différence $\varphi(a \wedge b, \bar{c})$ - $\varphi(a \wedge b, c)$ entre les intensités d'implication est

$$-\frac{1}{\sqrt{2\pi}} \int_{q'_1}^{+\infty} e^{-\frac{t^2}{2}} dt - \frac{1}{\sqrt{2\pi}} \int_{-q'_1\sqrt{\frac{n_c}{n_{\bar{c}}}}.\,k\,(a,b,c)}^{+\infty} e^{-\frac{t^2}{2}} dt = \frac{1}{\sqrt{2\pi}} \int_{q'_1}^{-q'_1\sqrt{\frac{n_c}{n_{\bar{c}}}}.\,k(a,b,c)} e^{-\frac{t^2}{2}} dt \qquad \textbf{(6)}$$

Proposition 12 : Pour le modèle binomial, la différence entre les intensités d'implication sera non seulement fonction du rapport $\frac{n_c}{n_{\bar{c}}}$ mais aussi de *k*(*a*,*b*,*c*) (**4**). Ce coefficient est d'autant plus grand et renforce ainsi l'effet du rapport $\frac{n_c}{n_{\bar{c}}}$ que $1 - \frac{n_{a\wedge b}.n_{\bar{c}}}{n^2} >> 1 - \frac{n_{a\wedge b}.n_c}{n^2}$, c'est-à-dire $\frac{n_{a\wedge b}}{n}.\frac{n_c}{n} >> \frac{n_{a\wedge b}}{n}.\frac{n_{\bar{c}}}{n}$. Les deux membres de cette inégalité ne sont autres que, de gauche à droite, les probabilités respectives du nombre de contre-exemples aléatoires - dans le modèle binomial où les variables $a \wedge b$ et c seraient indépendantes - des implications $a \wedge b \Rightarrow \bar{c}$ et $a \wedge b \Rightarrow c$. Ainsi, plus on s'attend à une réfutation de $a \wedge b \Rightarrow \bar{c}$, au vu de $n_{a\wedge b}$ et de $\bar{c}$, plus le caractère surprenant, *exceptionnel*, de cette règle est manifeste

par le constat de la modicité des contre-exemples observés à savoir $n_{a \wedge b \wedge c}$. Ceux-ci la valident au détriment de $a \wedge b \Rightarrow c$. L'inégalité montre la « contribution active à l'exception » au rapport $\frac{n_c}{n_{\bar{c}}}$, contribution qu'apportent les instances, de cardinal $n_{a \wedge b}$ dans celles de cardinal $n_{\bar{c}}$.

Cette conséquence, liée au modèle binomial, nous apparaît donc plus riche que celle énoncée dans le modèle de Poisson. En effet, elle nous fournit une relation de contrôle entre les paramètres plus fine du caractère d'exception que dans le modèle de Poisson. Ce phénomène, certes lié au nombre de paramètres de définition du modèle binomial, le gratifie cependant d'un intérêt que le logiciel CHIC permet d'exploiter à travers l'offre de son menu.

Remarque : A titre de comparaison, si nous nous intéressons à un autre indice de mesure de qualité de règle, à savoir la *confiance c*, qui est à la base des principaux autres indices de qualité (Lenca et al. 2004), nous obtenons les propriétés suivantes. Celle-ci s'exprime ainsi :

c($a \Rightarrow c$)= $\frac{n_{a \wedge c}}{n_a}$ (souvent notée : $\frac{\Pr[a \wedge c]}{\Pr[a]}$, autrement dite probabilité conditionnelle de *c* sachant *a*). La relation entre les règles que nous avons examinées est alors :

$$c(a \wedge b \Rightarrow \bar{c}) = \frac{n_{a \wedge b \wedge \bar{c}}}{n_{a \wedge b}} = 1 - \frac{n_{a \wedge b \wedge c}}{n_{a \wedge b}} = 1 - c(a \wedge b \Rightarrow c)$$

La règle d'exception a pour mesure le complément à 1 de la règle attendue. Ainsi, elle est indépendante des valeurs des occurrences.

4 Conclusion

Pour conclure et résumer, lorsque deux variables impliquent une 3ème, que leur conjonction implique plutôt la négation de cette 3ème, nous considérons que cette règle est d'exception, en un sens voisin mais différent de celui de E. Suzuki et Y.Kodratoff (1999). Nous avons étudié et illustré par un exemple numérique et un exemple de génétique, l'expression de ce caractère exceptionnel. Puis nous avons précisé les relations entre les paramètres des variables dans les deux modélisations selon lesquelles est construite l'Analyse Statistique Implicative : un modèle de Poisson et un modèle binomial, l'un et l'autre convergeant vers le même modèle gaussien.

Nous avons évoqué une approche complémentaire pour la détection de ces règles qui se base sur les travaux menés ces dernières années sur les R-règles (Gras et Kuntz, 2005 et chap. 4). La construction associée d'une hiérarchie implicative n'a pas été initialement développée dans ce but. Cependant, elle constitue une piste à explorer tant d'un point vue algorithmique que méthodologique concernant l'interprétation de ce que pourraient être des « R-règles d'exception ». Mais que ce soit pour des règles ou des R-règles d'exception, le signalement de ces semi-paradoxes par rapport au sens commun prouve, s'il en était nécessaire, *le saut conceptuel qu'impose le passage de la logique formelle de l'implication à la logique des quasi-implications,* objet de l'A.S.I..

Chapitre 7 : Extraction[21] de Règles en Incertain par l'Analyse Statistique Implicative[22]

Partant du cadre défini et formalisé par (Zadeh 1979, 2001), par (Dubois et Prade 1987), ce texte vise à étudier les proximités formelle et sémantique des cadres de l'incertain et de l'analyse statistique implicative (A.S.I.) entre variables à valeurs intervalles et variables-intervalles (Gras 2001a). On ne rappellera pas les formalisations classiques des notions premières et de chaque opérateur de la **logique floue**. On s'intéressera plus particulièrement à l'opérateur **«** implication » à l'aide duquel on extrait des règles d'association. Nous considérons celles qui croisent des sujets (ou des objets) et des variables, présentant des modalités nettes ou floues. Rappelons qu'une règle entre deux variables ou entre conjonctions de variables est établie sur la base de la rareté statistique du nombre de ses contre-exemples, dans l'hypothèse de l'indépendance a priori des variables en jeu (Gras 1979), (Lerman et al 1981) et chapitre 1 de la partie 1 de cet ouvrage. La qualité de la règle, avons-nous écrit, sera évidemment d'autant plus grande que ce nombre de contre-exemples sera invraisemblablement petit sous cette hypothèse, eu égard aux occurrences des variables et des instances totales.

Dans le premier sous-chapitre, nous présentons la problématique. Puis, nous construisons, de façon peu classique, une distribution floue à partir de données objectives. Ensuite, nous abordons la recherche de règles d'association dans une situation « floue » en nous appuyant auparavant sur la notion de variables modales. Enfin, nous revenons sur la construction des règles en ramenant les variables floues à des variables-intervalles.

1 Problématique. Un exemple prototypique de situation en incertain

Bien que les applications de la logique floue soient nombreuses en intelligence artificielle (par exemple en matière de diagnostic médical, de reconnaissance des formes ou de recherche de panne), plusieurs questions restent bien souvent latentes : comment obtient-on des distributions des degrés d'appartenance à un intervalle dans le cas de variables numériques ? Sur quelles connaissances sont-elles établies ? Sont-elles données a priori et mises à l'épreuve de la réalité ou bien sont-elles des construits ? S'il s'agit de ce dernier cas, quel processus d'extraction de connaissances à partir de données peut y conduire et quel type de règle peut-on alors extraire dans ce cadre ? Quelle signification peut-on donner à une règle associant deux sous-ensembles ou deux attributs flous ? On rejoint alors une des problématiques du data mining et de la qualité des règles, ce qui justifie notre préoccupation en A.S.I...

[21] Ce texte a été présenté avec une forme voisine et un contenu réduit sous le titre : « Extraction de règles en incertain par la méthode statistique implicative », dans les *Comptes rendus des 12èmes Rencontres de la Société Francophone de Classification, Montréal 30 mai-1er juin 2005, UQA.* Les auteurs étaient Régis Gras, Raphaël Couturier, Fabrice Guillet et Filippo Spagnolo.

[22] Remerciements à Maurice Bernadet pour sa lecture du texte et ses précieux conseils

Voici un exemple prototypique d'une situation où les données sont floues. U est l'univers de référence, du discours dit-on, portant sur 3 **modalités ou attributs flous** relatifs à la taille d'individus : {petit, moyen, grand}et se confondant avec l'ensemble E des observations potentielles ou réalisées des attributs. Si l'on veut accorder un **degré d'appartenance**, de la forme $\mu_T(x)$, aux sous-ensembles associés à ces 3 attributs respectivement : T_1=petit, T_2 =moyen, T_3 = grand, à des individus x de l'ensemble E des sujets, on obtient par exemple selon 3 d'entre eux :

a) $x=i_1$ est un individu dit plutôt petit, alors $\mu_{T1}(i_1)$ =0,8, $\mu_{T2}(i_1)$ =0,5, $\mu_{T3}(i_1)$ = 0,2, c'est-à-dire pour ce sujet, un fort degré d'appartenance à la classe T_1, faible degré à la classe T_3.

b) $x=i_2$ est un individu dit pas très grand, alors $\mu_{T1}(i_2)$ = 0,1, $\mu_{T2}(i_2)$ = 0,6, $\mu_{T3}(i_2)$ = 0,7.

c) $x=i_3$ est un individu dit plutôt grand, alors $\mu_{T1}(i_3)$ = 0, $\mu_{T2}(i_3)$ = 0,7, $\mu_{T3}(i_3)$ = 0,9.

Les données sont ici floues. Aux sous-ensembles de U, on associe une **pseudo-partition** définie par d'autres modalités telles que « pas très grand », « plutôt grand », etc.. « pas très », « plutôt »,… qui sont appelées des **modificateurs linguistiques**. Ils permettent de définir de nouveaux **sous-ensembles flous** à partir des sous-ensembles précédents. On peut dans ce cas considérer qu'à chaque sujet, à chaque observation, correspond un vecteur « net » dans un espace de dimension le nombre des modalités de U.

Traditionnellement, dans la théorie du flou, on représente les distributions des tailles floues d'une façon comparable à la suivante en plaçant les 3 sujets, en fonction de ces distributions supposées ici affines par morceaux :

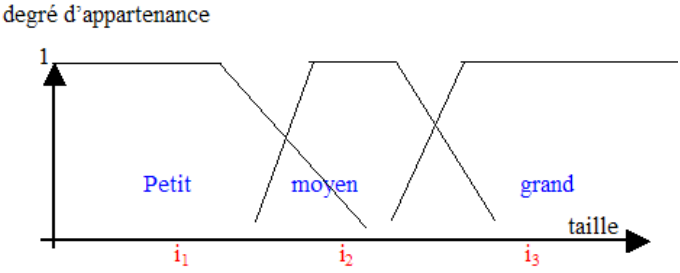


FIG. 20-– *Une représentation d'un exemple*

Mais les fonctions d'appartenance peuvent être différentes de celle choisie ici ; citons : courbe Bêta, courbe de Gauss, courbe triangulaire, etc.

On constate que les sous-ensembles flous ont des intersections généralement non vides.

2 Deux méthodes de construction de distributions floues par extraction de connaissances

Dans le cadre que nous retenons, les distributions des degrés d'appartenance seront le fruit de l'interaction entre connaissances objectives (une vraie valeur de la variable, un

attribut net ou modificateur linguistique *consensuel*) et connaissances subjectives. Dans la littérature, les degrés sont des données. D'où proviennent-elles ?

L'exemple ci-dessus illustre le cas d'un échantillon d'individus donné. On disposera **effectivement** de leur taille s (un nombre) ou des caractères ou attributs nets : « petit », « moyen » et « grand » au vu d'une décision consensuelle du type : les caractères « petit », « moyen » et « grand » seront attribués **objectivement** au regard de leur taille mesurée. Face à ces données, on pourra comparer le point de vue **subjectif** portant sur les mêmes individus qui énoncera qu'un sujet de 179 cm n'est pas petit, mais peut être considéré de taille grande ou moyenne, non contradictoirement.

Différentes méthodes pour définir la distribution des attributs visent à effectuer un processus de « **fuzzification** » (Bernadet, 2004) : définition des classes floues pour chaque attribut, puis mise en correspondance de chaque attribut avec un degré d'appartenance à un sous-ensemble flou, comme nous le voyons dans l'exemple introductif. Dans (Zadeh, 1997), une méthode de discrétisation optimale est donnée. Ici, nous procèderons autrement.

3 Relation entre intervalles nets et attributs flous

Notre objectif, dans ce paragraphe, est de « *fuzzifier* » en quantifiant le degré d'appartenance d'un sujet à un intervalle numérique donné. Pour ce faire, la méthode de type « clustering » que nous proposons consiste, à partir du choix d'un indice de similarité, ici celui de la vraisemblance du lien de I.C. Lerman (1981), d'extraire, tout d'abord, la proximité entre les attributs nets et les attributs flous.

Auparavant, selon le procédé défini dans (Gras 2001) et présenté au chapitre 3, nous choisissons de transformer l'ensemble des valeurs observées sur les sujets en sous-intervalles disjoints de variance inter-classe maximale afin de pouvoir attribuer à chaque sous-intervalle un attribut net de même désignation que celle attribuée aux attributs flous. Cette partition nette est établie par la méthode des nuées dynamiques de (Diday 1972). Enfin, pour chaque classe de similarité entre **attribut net** et **attribut flou**, nous déterminons le degré d'appartenance des sujets à une classe floue à partir de la mesure normalisée de typicalité associée à chaque individu. En effet, cette typicalité, définie dans (Gras et al. 2006) et présentée au chapitre 5, rend compte d'un degré de responsabilité dans la proximité d'attributs, soulignant l'accord entre net et flou. Ainsi, nous disposerons d'une mesure vérifiant les axiomes de Zadeh relatifs au concept de « possibilité ». Mais, son avantage par rapport à la détermination subjective classique est qu'elle est établie à l'épreuve statistique de la réalité et qu'elle varie avec la dilatation de l'ensemble des sujets.

En résumé, les données initiales sont de deux ordres :

- d'une part, des variables **objectives, consensuelles** aux valeurs numériques réparties sur des intervalles auxquels on associe respectivement un **attribut net** ,
- d'autre part, un **attribut flou** attribué **subjectivement** à chaque sujet.

Exemple : Les données portent sur 60 sujets. Leurs tailles T nettes vraies varient de 168 et 198 cm. Appliquant l'algorithme de la variance inter-classe maximale, nous obtenons, par l'algorithme, une partition constituée des intervalles nets: « Tpeti » de 168 à 174, « Tmoy » de 175 à 183, Tgran de 184 à 198. En outre, les attributs flous, issus d'un jugement subjectif, sont notés respectivement TP, TM et TG. Ainsi, par exemple, un sujet de taille vraie 180 cm

sera classé dans la classe nette Tmoy mais aussi simultanément dans les classes floues TM et TG. La hiérarchie des similarités donnée par CHIC entre ces 6 variables, est alors :

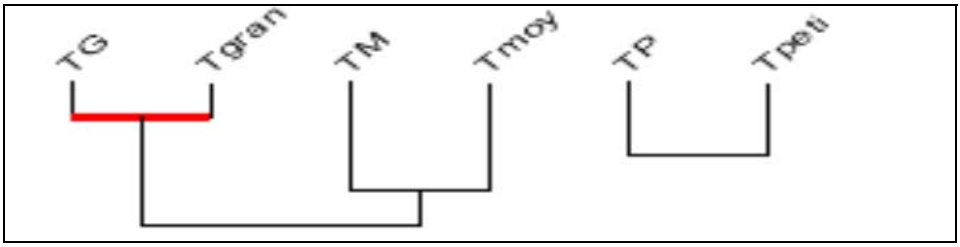


FIG. 21- – *Hiérarchie des similarités entre les intervalles nets et flous*

On note que les attributs nets s'associent aux intervalles flous correspondants, ce que raisonnablement on pouvait attendre d'une subjectivité normale. Un sujet pourra donc posséder TM et TG suivant le point de vue du juge si sa taille n'est pas manifestement grande.

Le logiciel CHIC ((Couturier et Gras 2005) et Partie 2, chap 11 et 12), restitue les mesures de typicalité des sujets selon les 3 classes de similarité. Rappelons qu'un individu est d'autant plus typique d'une classe qu'il a une attitude conforme à la constitution de la classe par la population de sujets.. On observe alors en consultant les calculs de typicalités donnés par CHIC que, par ex., le sujet i06 a une typicalité de 0 sur les tailles petites, 0,056 sur les tailles moyennes et 0,95 sur les tailles grandes. On peut faire de même pour les autres sujets de l'échantillon. Ce sont ces valeurs que nous retenons comme degrés d'appartenance respectifs par rapport aux attributs flous.

Pour itérer le procédé dans le but d'affiner les distributions, il suffit, par ex, de remplacer une des modalités d'attribut ou chacune d'entre elles par 2 modalités. Ainsi, « grand » sera subdivisé en « très grand » et « assez grand », « moyen » en « pas très grand » et « pas très petit », etc. La partition de l'intervalle des tailles se fera sur la base de 6 intervalles.

« Attribuer des degrés d'appartenance à partir des typicalités aux associations observées sur un échantillon » nous paraît atténuer l'arbitraire habituel des affectations de ces degrés.

4 Construction de l'histogramme d'une variable-intervalle à partir des données floues des sujets

Cette fois, on dispose de la distribution des valeurs floues prises par chaque sujet d'une population sur un intervalle. On cherche à en déduire une distribution des degrés d'appartenance sur cet intervalle. L'objectif final est de définir une variable symbolique, dite aussi variable-intervalle, qui soit l'histogramme d'un intervalle sur lequel on pourra déterminer des sous-intervalles optimaux selon le critère de la variance.

Soit f_1, f_2, …,f_n les fonctions d'appartenance respectives des n sujets à un intervalle A. On suppose, par analogie avec les densités, que ces fonctions sont normalisées sur A. Dans la majorité des cas, chaque sujet contribue de la même façon à la densité, sinon une pondération adaptée ramène à un problème analogue. Alors la fonction $f=(f_1+f_2+ \ldots.+f_n)/n$ intègre en un histogramme sur A la distribution des fonctions d'appartenance. Il suffit ensuite de

discrétiser A en une suite de points pondérés selon f ; enfin, d'appliquer sur A l'algorithme de la variance selon la méthode des nuées dynamiques pour obtenir une variable-intervalle a dont on pourra étudier les relations implicatives avec les autres variables du même type.

Par ex., on donne les valeurs floues de notes obtenues sur [0 ; 20] par 3 étudiants i_1, i_2, i_3 : une correction multiple affecte à i_1, i_2 et i_3 respectivement des notes : 5 à 9, 6 à 11 et 8 à 15

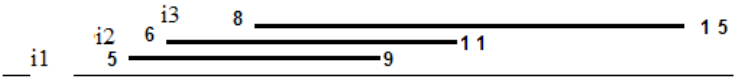


FIG. 22– *Une représentation de l'échelonnement de notes*

Supposant la distribution uniforme des valeurs floues, normalisées sur [0 ; 20], selon chacun des intervalles, on obtient le tableau (TAB. 23) des fonctions d'appartenance : par ex. sur [0,3 ; 0,55], correspondant à l'intervalle de notes [6 ;10], f_2=1/5 sur chacun des 5 intervalles d'amplitude 0.05 et f_2= 0 ailleurs.

individus	Modalités de a				
↓	[0.25; .30]	[0.30 0.40].	[0.40; .45]	[0.45; 0.55]	[0.55; 0.75]
i1	1/4	2/4	1/4	0	0
i2	0	2/5	1/5	2/5	0
i3	0	0	1/7	2/7	4/7

TAB. 23 – *Valeurs prises par les modalités sur les 3 sujets*

par ex. A est discrétisable en 420 (ppcm de 3, 4, 5, 7) points

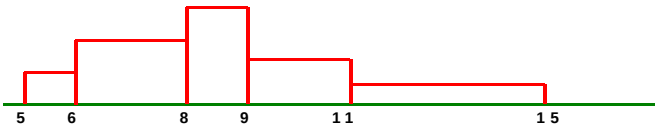


FIG. 23- *Histogramme associé*

5 Règles d'association pour des variables numériques

On suppose dorénavant, à titre d'exemple prototypique, que les distributions des variables floues sont connues selon 2 variables observées sur les mêmes sujets : taille et poids. Mais cet exemple a la vertu de généralité. On veut étudier, maintenant, comme en ASI, les règles de déduction entre l'attribut taille et l'attribut poids, présentant des modalités, l'un **Taille** = {petit, moyen, grand}, l'autre **Poids** = {léger, moyen, lourd}.

On dispose de données sous forme d'un tableau numérique des degrés d'appartenance aux modalités d'attributs flous, valeurs relatives à un échantillon de 20 sujets. Les 3 premiers de ces sujets constituent le tableau (TAB. 24). L'un d'entre eux, i_1, n'est donc pas très grand et pas très lourd, l'autre i_2 assez grand et assez lourd, le dernier i_3 plutôt grand et plutôt lourd.

	taille			**poids**		
	petit T_1	*moyen* T_2	*grand* T_3	*léger* P_1	*moyen* P_2	*lourd* P_3
i_1	8/15	5/15	2/15	7/14	4/14	3/14
i_2	1/14	6/14	7/14	2/15	5/15	8/15
i_3	0	7/16	9/16	1/16	6/16	9/16

TAB. 24 – *Valeurs prises par les modalités sur les 3 sujets*

5.1 Un premier traitement de variables numériques

On propose ici un traitement implicatif, selon l'A.S.I., en considérant les 6 variables tailles-poids comme des variables numériques sur l'ensemble des 20 sujets. On obtient le graphe implicatif en utilisant l'indice de (Lagrange 1998), réactualisé par (Régnier et Gras. 2004) et (Partie 1, chap. 3 et Partie 2, chap. 1). Ainsi, les implications T3 ⇒ P3 et P1 ⇒ T1 sont valides au seuil 0,90 et signifient que les propositions grand ⇒ lourd et léger ⇒ petit, règle qui est sémantiquement contraposée de la première, sont acceptables. Une autre implication à un seuil supérieur à 0,6 apparaît : $P_2 \Rightarrow T_1$.

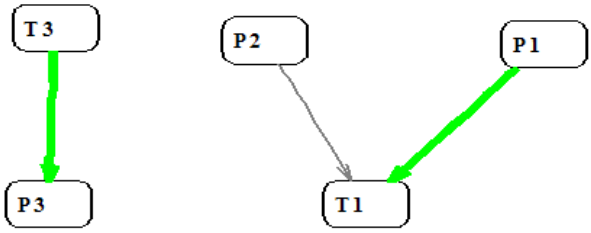


FIG. 24– *Graphe implicatif taille x poids*

Ces résultats ne s'opposent pas, bien entendu, au bon sens. Les autres règles d'association confirment une meilleure adéquation à la sémantique de l'implication qu'avec les approches de Reichenbach et Lukasiewicz (cité dans (Dubois et Prade, 1987)). On ne retrouve pas, par ex. : léger ⇒ grand.

Mais, l'approche proposée ici présente l'inconvénient de considérer que les 6 modalités des variables « taille » et « poids » sont actives dans le traitement et ne restituent pas, ainsi, les nuances de leur structure. Il semble donc intéressant, sémantiquement parlant, de revenir à la considération de modalités de variables de type intervalles où les modalités apparaissent comme sous-intervalles d'une variable-intervalle principale.

5.2 Second traitement par des variables à valeurs intervalles

Ce second traitement (Gras et al. 2001 a) et (chapitre 3) va permettre de prendre en compte de façon plus fine les nuances des observations prises selon des sous-ensembles flous et de répartir leurs valeurs de façon optimale sur un intervalle numérique [0 ;1], selon une partition dont l'utilisateur définit le nombre de classes pour chacun de 20 sujets.

Nous disposons d'un nouveau tableau donnant les distributions des 6 modalités des 2 attributs « taille » et « poids » relativement à chacun des individus et les valeurs binaires prises par 2 variables supplémentaire « Femme », « Homme ». En voici les 2 premières lignes.

	Taille petite pt	**Taille moyenne m**	**Taille grande T**	**Var. supp. Femme**	**Var. supp. Homme**	**Poids léger L**	**Poids moyen o**	**Poids grand P**
$\mathbf{i_1}$	0,7	0,4	0,3	1	0	0,8	0,3	0,1
$\mathbf{i_2}$	0,2	0,5	0,8	0	1	0,1	0,4	0,9

TAB. 25 – *Distributions des attributs flous « taille » et « poids »*

Par ex., le sujet i_1 admet un degré d'appartenance 0,7 à la classe des sujets petits, 0,4 à celle des sujets de taille moyenne et 0,3 à la classe des sujets de grande taille. De plus (variable supplémentaire), ce sujet est une femme et la distribution de ses degrés d'appartenance aux 3 classes de poids, sont respectivement 0,8, 0,3 et 0,1. Le traitement va emprunter cette fois la méthode des variables à valeurs intervalles. Chaque modalité conduira à la construction de sous-intervalles optimaux, c'est-à-dire la détermination de sous-intervalles optimisant, du moins localement sinon globalement, l'inertie inter-classe. Utilisant ensuite CHIC de traitement de ce type de variable, on établit les règles telles que : si un sujet relève de l'intervalle t_i de la modalité t de l'attribut « taille » alors, généralement, il relève de l'intervalle p_j de la modalité p de l'attribut « poids ». Ainsi, si par ex., il a tendance à être plutôt petit, alors il a généralement tendance à être plutôt léger.

Les partitions en 3 sous-intervalles calculées par CHIC sont données dans le tableau ci-dessous.

tailles petites:	tailles moyennes :	grandes tailles :
t1 de 0 à 0.1	m1 de 0.1 à 0.3	T1 de 0 à 0.1
t2 de 0.2 à 0.5	m2 de 0.4 à 0.6	T2 de 0.2 à 0.5
t3 de 0.6 à 1	m3 de 0.8 à 0.8	T3 de 0.8 à 0.9
poids légers :	poids moyens :	poids lourds :
L1 de 0 à 0.2	o1 de 0.2 à 0.3	P1 de 0 à 0.1
L2 de 0.3 à 0.6	o2 de 0.4 à 0.5	P2 de 0.2 à 0.4
L3 de 0.8 à 1	o3 de 0.6 à 0.7	P3 de 0.7 à 0.9

TAB. 26 – *Partitions optimales calculées par CHIC*

Le graphe implicatif au niveau de confiance 0,90 est également donné par CHIC :

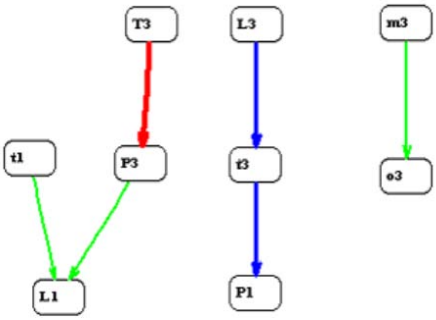


FIG. 25– *Graphe implicatif taille x poids*

On voit par exemple que :

- l'individu de grande taille (T3) admet généralement un poids important (P3) et donc n'est pas considéré comme léger(L1). Ce sont les hommes qui apportent, et de très loin (risque de se tromper = 0.07), la plus importante contribution ;

- l'individu de poids plutôt léger (L3) est généralement de petite taille (t3) ;dans ce cas, ils ne sont que très rarement considérés lourds (P1). Ce sont les femmes qui sont les plus contributives à ce chemin (risque = 0.25) ;

- les deux variables t1 et L1, liées par la règle t1⇒L1, correspondent à des fréquences rares. Si donc, on rencontre un sujet petit alors il est généralement léger. Le sexe Homme contribue à la formation de cette règle.

6 Conclusion

A l'aide de l'A.S.I., nous avons cherché à **objectiver** la notion de degré d'appartenance. Situant le modèle d'implication entre attributs par rapport à des modèles classiques, nous avons mis en évidence par un graphe, les relations implicatives entre des modalités de variables numériques. Nous avons, semble-t-il, amélioré la formalisation de la sémantique en faisant référence à des variables-intervalles. Les règles les plus consistantes ont pu être extraites selon leur qualité. Enfin, la relation entre des variables extrinsèques et ces règles ont permis d'enrichir notre connaissance sur ces règles. Des applications à des situations réelles tenteront de valider cette nouvelle approche de l'incertain. D'ores et déjà, nous savons quel intérêt, par exemple dans le détection de l'origine de pannes les industriels ont accordé au traitement du problème à l'aide de la logique du flou.

Chapitre 8 : Réduction du nombre de variables[23]

1 Introduction

L'Extraction de Connaissances dans les Données (ECD) a pour objectif la découverte de connaissances utiles cachées dans des données volumineuses (Frawley et al., 1992), (Fayyad et al., 1996). Lorsque ces données se ramènent à une table croisant sujets et variables, la notion de volume se traduit par un grand nombre de lignes (sujets), ce qui est statistiquement intéressant, mais aussi un grand nombre de colonnes (variables) ce qui peut l'être moins.

En effet, dès que le nombre de variables devient pléthorique, la plupart des techniques disponibles deviennent impraticables. En particulier, lorsque l'on procède à une analyse implicative par calcul de règles d'association (Agrawal et al., 1993), le nombre de règles découvertes, subit une explosion combinatoire avec le nombre de variables, et devient rapidement inextricable pour un décideur, pour peu que soient demandées des conjonctions de variables. Dans ce contexte, il s'avère nécessaire de procéder à une réduction préliminaire du nombre de variables.

Ainsi, (Ritschard et al., 2000) ont proposé une heuristique efficace permettant de réduire à la fois le nombre de lignes et de colonnes d'une table, à partir d'une mesure d'association servant de critère de quasi-optimalité pour piloter l'heuristique. Cependant, à notre connaissance, dans les différentes autres recherches, le type de situation à l'origine de la nécessité du regroupement de lignes ou de colonnes n'est pas pris en compte dans les critères de réduction, que la problématique et la visée de l'analyste soient la recherche de similarité, de dissimilarité, d'implication, etc., entre variables.

Aussi, dans la mesure où il existe des variables très voisines au sens de l'implication statistique, il pourrait être opportun de substituer à ces variables une seule qui serait leur leader en termes de représentation d'une classe d'équivalence de variables similaires pour la visée implicative.

Nous nous proposons donc, à l'instar de ce qui est fait pour définir la notion de quasi-implication, de définir une notion de quasi-équivalence entre variables, afin de construire des classes d'où nous extrairons un leader. Nous l'illustrerons par un exemple. Ensuite, nous envisagerons la possibilité d'utiliser un algorithme génétique afin d'optimiser le choix du représentant de chaque classe de quasi-équivalence.

2 Définition de la quasi-équivalence

Deux variables binaires a et b sont logiquement équivalentes pour l'A.S.I. lorsque sont simultanément satisfaites, à un seuil donné, les deux quasi-implications : $a \Rightarrow b$ et $b \Rightarrow a$. Nous avons conçu des critères pour évaluer la qualité d'une quasi-implication : l'un est l'étonnement statistique inspiré de la vraisemblance du lien de Lerman (1981), l'autre est la

[23] Ce chapitre s'inspire fortement de l'article intitulé, tout en l'étendant : « Réduction des colonnes d'un tableau de données par quasi-équivalence entre variables », publié dans les actes de EGC 2 (janvier 2002), Cépadués, avec pour co-auteurs Régis Gras, Fabrice Guillet, Robin Gras et Jacques Philippé.

forme entropique de la quasi-inclusion (Gras et al. 2001) qui est présentée dans le chapitre 3 de cette partie de l'ouvrage.

Selon le *premier critère*, on pourrait dire que deux variables a et b sont quasi-équivalentes lorsque l'intensité d'implication φ(a,b) de a⇒b est peu différente de celle de b⇒a. Cependant, sur des ensembles importants (plusieurs milliers), ce critère n'est plus suffisamment discriminant pour valider l'inclusion.

Selon le *deuxième critère*, on se base, sur une mesure entropique du déséquilibre entre, d'une part, les effectifs $n_{a\wedge b}$ (individus qui satisfont a et b) et $n_{a\wedge \bar{b}}$ (individus qui satisfont a et non(b), contre-exemples à l'implication a⇒b) pour signifier la qualité de l'implication a=⇒b, et d'autre part, les effectifs $n_{a\wedge b}$ et $n_{\bar{a}\wedge b}$ pour évaluer la qualité de l'implication réciproque b⇒a.

Ici nous allons utiliser ici une méthode comparable à celle utilisée dans le chapitre 3 pour définir l'indice d'implication entropique.

En posant n_a et n_b, respectivement effectifs de a et de b, le déséquilibre de la règle a⇒b est mesuré par une entropie conditionnelle K(b|a=1), et celui de b⇒a par K(a|b=1) avec :

$$K(b \mid a=1) = -(1-\frac{n_{a\wedge b}}{n_a})\log_2(1-\frac{n_{a\wedge b}}{n_a}) - (\frac{n_{a\wedge b}}{n_a})\log_2(\frac{n_{a\wedge b}}{n_a}) \text{ si } \frac{n_{a\wedge b}}{n_a} > 0{,}5 \quad (\mathbf{1})$$

$$K(b|a=1)=1 \text{ si } \frac{n_{a\wedge b}}{n_a} \le 0{,}5 \quad (\mathbf{2})$$

$$K(a \mid b=1) = -(1-\frac{n_{a\wedge b}}{n_b})\log_2(1-\frac{n_{a\wedge b}}{n_b}) - (\frac{n_{a\wedge b}}{n_b})\log_2(\frac{n_{a\wedge b}}{n_b}) \text{ si } \frac{n_{a\wedge b}}{n_b} > 0{,}5 \quad (\mathbf{3})$$

$$K(a|b=1)=1 \text{ si } \frac{n_{a\wedge b}}{n_b} \le 0{,}5 \quad (\mathbf{4})$$

Ces deux entropies doivent être suffisamment faibles pour que l'on puisse, avec une bonne certitude, parier sur b (resp. sur a) lorsque a (resp. b) est réalisé. Par conséquent leurs compléments respectifs à 1 doivent être simultanément forts.

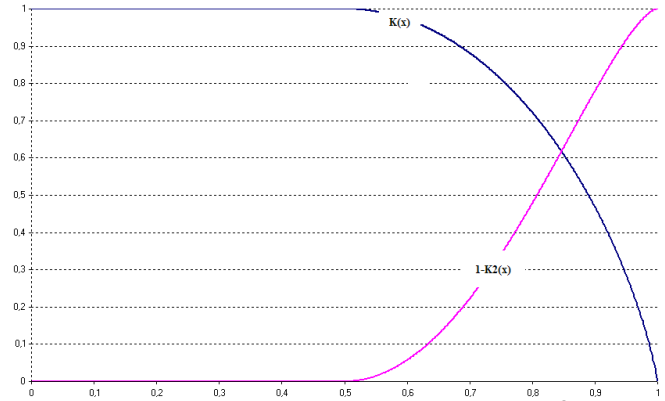


FIG. 26- *représentation graphique des fonctions* K *et* 1-K^2 *sur* [0 ;1]

Définition 36: Un premier **indice entropique d'équivalence** est donné par :

$$e(a,b) = \left(\left[1 - K^2(b \mid a = 1)\right]\left[1 - K^2(a \mid b = 1)\right]\right)^{\frac{1}{4}}$$

Quand cet indice prend des valeurs dans le voisinage de 1, cela traduit une bonne qualité d'une double implication. De plus, afin de mieux prendre en compte $a \wedge b$ (les exemples), nous intégrons ce paramètre à travers un indice de similarité s(a,b) des variables, par exemple au sens de I.C. Lerman (1981). L'indice de quasi-équivalence est alors construit par conjugaison de ces deux notions.

Définition 37: Un second **indice entropique d'équivalence** est donné par la formule

$$\sigma(a,b) = [e(a,b).s(a,b)]^{\frac{1}{2}}$$

Partant de ce point de vue, nous énonçons alors le critère de quasi-équivalence que nous retenons.

Définition 38: On dit que la **paire de variables {a,b} est *quasi-équivalente*** pour la qualité choisie β si σ(a,b)≥ β

Par exemple, une valeur β=0,95 pourra être considérée comme désignant une bonne quasi-équivalence entre a et b.

3 Algorithme de construction des classes de quasi-équivalence

Soit un ensemble V={a,b,c,…} de v variables muni d'une relation valuée R induite par la mesure de quasi-équivalence σ sur l'ensemble des paires de V. On supposera les paires de variables classées selon un ordre décroissant de quasi-équivalence. Si nous avons fixé le seuil de qualité de la quasi-équivalence à β, seules seront conservées les premières des paires {a,b} vérifiant l'inégalité σ(a,b)≥β. En général, seule une partie V', de cardinal v', des variables de V vérifiera cette inégalité. Si cet ensemble V' est vide ou trop réduit, l'utilisateur pourra ramener son exigence à une valeur seuil β plus petite. La relation étant symétrique, on disposera au plus de $\frac{v'(v'-1)}{2}$ paires à étudier. Quant à V-V', il ne contient que des variables non réductibles.

On propose d'utiliser l'algorithme glouton suivant :

1° On constitue une première classe potentielle C_1^0 = {e,f} telle que σ(e,f) représente la plus grande des valeurs de β-équivalence. Si cela est possible, on étend cette classe selon une nouvelle classe C_1 en prélevant dans V' tous les éléments x tels que toute paire de variables au sein de cette classe admette une quasi-équivalence supérieure ou égale à β.

2° On poursuit par :

a) Si o et k formant la paire (o,k) immédiatement inférieure à (e,f) selon l'indice σ, appartiennent à C_1, alors on passe à la paire immédiatement inférieure à σ(o,k) et on procède comme dans le 1°.

b) Si o et k n'appartiennent pas à C_1, on procède comme dans 1° à partir de la paire qu'ils constituent en formant la base d'une nouvelle classe C_2^0.

c) Si o ou k n'appartient pas à C_1, l'une de ces deux variables pourra soit former une classe singleton, soit appartenir une classe future. Sur celle-ci, on pratiquera bien entendu comme ci-dessus.

Après un nombre fini d'itérations, on dispose d'une partition de V en r classes de σ-équivalence : $\{C_1, C_2, \ldots, C_r\}$. La qualité de la réduction pourra être évaluée par un indice brut ou proportionnel à $\beta \frac{p}{k}$. Cependant nous lui préférons le critère défini ci-dessous qui présente l'avantage d'intégrer le choix du représentant.

De plus, on pourrait choisir k variables représentantes des k classes de σ-équivalence sur la base du critère élémentaire suivant : la qualité de liaison de cette variable avec celles de sa classe. Mais, ce critère ne permet pas d'optimiser la réduction puisque le choix du représentant est relativement arbitraire et peut-être signe de trivialité de la variable. Nous allons donc revenir sur cette question.

4 Recherche d'un critère pour déterminer un optimum de la réduction

Nous nous sommes inspirés ici des notions de variance implicative et de cohésion développées dans (Gras et al., 1996 c) et que nous retrouverons dans le chapitre 2 de la Partie 2. En particulier, on montre que si l'on considère chaque variable binaire comme un vecteur de $[0,1]^n$, alors le carré scalaire du vecteur $\vec{ab}$ traduit l'implication a⇒b et sa réciproque b⇒a. Dans le cas de variables binaires, ce carré scalaire ne s'annule que si les deux implications sont strictes (elles prennent les mêmes valeurs, 0 ou 1, selon les mêmes sujets). Toute valeur de ce carré proche de 0 traduit donc une ressemblance de a et b, tout en évitant les biais consécutifs au centrage/réduction requis par exemple pour le calcul du coefficient de corrélation.

Aussi, les k classes d'équivalence étant construites comme ci-dessus, nous pouvons expliciter une certaine inertie que nous désignons par **inertie implicative** I, à partir de l'ensemble $\{a_1, a_2, \ldots a_k\}$ de k variables qui les représentent :

$$I = \frac{1}{\sum_{i=1}^{k} k_i c_i} \left(\sum_{i=1}^{k} k_i c_i \left[\overrightarrow{a_i g}\right]^2 \right)$$

où C_i est une classe, k_i son effectif, c_i sa cohésion implicative, et g le barycentre des représentants. On vérifie que cette inertie croît avec la cohésion et l'effectif de la classe C_i. Cette grandeur I rend compte, dans un contexte implicatif, de la séparabilité des représentants des classes ainsi que de leur consistance.

5 Un exemple illustratif

Une enquête auprès d'étudiants de l'École Polytechnique de l'Université de Nantes a porté sur la recherche de règles d'association entre 41 variables, animaux sauvages ou domestiques, possédant ou non certaines qualités parmi 75 qualités proposées (par ex. « affectueux », « féroce », etc.) auxquelles on associe les 41 animaux. Ainsi «bruyant » sera accordé à « baleine », « canard », etc.. Comme la représentation du graphe exprimant les règles implicatives entre animaux risque de ne pas présenter une qualité de clarté facilitant l'analyse, nous avons utilisé le logiciel REDUCHIC, élaboré par Raphaël Couturier, afin de réduire le nombre de variables selon l'algorithme ci-dessus implémenté dans REDUCHIC. On obtient les informations suivantes :

Animaux		**Classe**
Lapin, Dauphin		*Dauphin*
Poule, Canard		*Poule*
Crocodile, Couleuvre		*Couleuvre*
Loup, Tigre, Requin		*Tigre*
Rat, Mygale		*Rat*
Ours, Lion	représentant de la classe	*Lion*
Renard, Chat		*Renard*
Lynx, Crotale		*Crotale*
Vache, Âne		*Âne*
Vautour, Corbeau		*Corbeau*
Mouche, Cigale		*Mouche*
Mouton, Chien		*Chien*

TAB. 27

Le nouveau fichier des animaux, après cette réduction par équivalence et le choix du meilleur représentant minimisant l'inertie implicative, ne comporte plus que 28 animaux, ce qui est manifestement plus aisé à représenter et à analyser. On observera la vraisemblance des équivalences extraites comme par exemple celle de la classe (Loup, Tigre, Requin)

En opérant de la même façon avec un questionnaire devant faire apparaître des traits de personnalité, nous disposions de 142 variables et 2299 sujets. Une réduction a conduit à un fichier plus compact de 34 variables, plus aisément traitable. Certaines classes d'équivalence contiennent jusqu'à 12 variables dont une seule sera retenue par la suite.

6 Détermination d'un optimum par un algorithme génétique

Notre objectif est de trouver un ensemble de k représentants maximisant l'inertie implicative I (ici de complexité en $(\frac{p}{k})^k$). Pour cela, on dispose de différentes méthodes heuristiques : par exemple, les nuées dynamiques selon la méthode développée par E. Diday et son usage dans le cadre de l'implication statistique dans (Gras et *al.* 2001), la programmation dynamique, un algorithme génétique (Goldberg 1994), etc.. C'est cette dernière méthode que nous utiliserons, car elle a l'avantage d'être efficace pour de grands espaces de recherche. Dans la population considérée par l'algorithme génétique, chaque

individu est codé par chromosome constitué de k gènes dont chaque allèle code une variable représentante d'une classe. Au début du processus chaque représentant est choisi aléatoirement.

Précisons la forme des différents opérateurs génétiques :

a) -**Sélection** : le critère de qualité utilisé pour la sélection est l'inertie implicative I.
b) -**Reproduction** : le cross-over entre deux chromosomes s'effectue par le choix aléatoire d'un rang de coupure.
c) -**Mutation** : La mutation d'un gène se fera en respectant un critère de contiguïté : la probabilité de mutation de a vers b devra être proportionnelle au carré scalaire du vecteur $\vec{ab}$.

7 Conclusion

Une implémentation spécifique a été effectuée par des étudiants de l'École Polytechnique de l'Université de Nantes à l'occasion d'un projet. Nous disposons également d'un algorithme implémenté dans le logiciel REDUCHIC qui a permis de réduire, au gré de l'utilisateur, et aussi sensiblement qu'il le souhaite, le nombre de variables à traiter. Ce logiciel a été suivi de nombreuses expérimentations et applications permettant, d'une part d'en évaluer la pertinence et l'efficacité et, d'autre part, les performances en temps de calcul. Le nouveau fichier réduit est directement intégrable au logiciel C.H.I.C et traitable par lui comme tout autre fichier .csv (Couturier et Gras, 2005). La complexité liée à la prise en considération des conjonctions de variables s'en trouve fortement allégée.

Chapitre 9 : Règles superflues ou redondantes en Analyse Statistique Implicative[24]

1 Introduction

Certaines recherches, aux objectifs différents mais avec le même souci de réduction, ont permis de diminuer sensiblement le nombre de variables en établissant des indices de similarité entre elles. C'est le cas, par exemple, des indices de Jaccard, de Russel et Rao, de Rogers et Tanimoto, de Piatetsky-Shapiro cités dans (Matheus et al, 1996). Couturier et al. (2004) définissent un indice de similarité dans le cadre de l'A.S.I. et comparent, avantageusement dans leur approche, les différents indices visant la réduction de variables. Gras et al. (2002) établissent une relation d'équivalence entre les variables en colonnes visant leur réduction comme nous venons de le voir dans le chapitre 8 précédent. Blanchard et al. (2004) ne traitent pas ce sujet mais, par contre, étudient, via la notion d'entropie, la qualité des règles de la forme a⇒b et de leur contraposée, également dans le cadre de l'A.S.I., en définissant le taux informationnel d'une règle à partir du gain d'information apporté sur b par la réalisation de a. Nous abordons ici le problème de la redondance en suivant une démarche comparable à celle de J. Blanchard. L'objectif de ce chapitre est alors : comment réduire l'ensemble des règles obtenues en ne retenant que celles qui fournissent des informations différentes, donc non surabondantes, voire non superfétatoires ?

En d'autres mots, la problématique de ce chapitre est la suivante :

On suppose que des règles ont été extraites de l'ensemble des données et, qu'en particulier, une classification orientée, dite cohésitive, organise l'ensemble des attributs en règles et méta-règles ou règles généralisées,(Gras et Kuntz., 2004) comme nous l'avons abordé dans le chapitre 4. On souhaite réduire le nombre de ces règles en conservant l'information maximale qu'elles contiennent selon deux types d'indice. Pour cela, nous envisagerons les règles généralisées d'ordre quelconque, c'est-à-dire celles dont la prémisse et la conclusion sont des règles simples, d'ordre 0 entre variables ou bien des règles généralisées, c'est-à-dire des règles de règles, d'ordres supérieurs.

Une première solution, que nous avons présentée déjà dans le chapitre 8, pourrait consister à procéder de la même façon. Il suffirait de définir une relation d'équivalence entre deux règles R et S d'ordre quelconque dès lors que nous aurions à la fois R⇒S et S⇒R à un haut niveau de qualité implicative. Ici, nous privilégions la relation non symétrique d'implication entre règles. Par suite, c'est la qualité de l'information de l'une sur l'autre qui servira plutôt de critère de réduction.

Pour ce faire, nous présentons l'approche de la réduction du nombre de règles et leur redondance au sens de l'entropie de Shannon. Dans la même perspective, nous proposons ensuite l'emploi de l'indice de Gini afin d'évaluer l'information apportée par une règle sur une autre.

[24] Une version voisine de ce texte figure en anglais sous le titre : « Reduction of Redundant Rules in Statistical Implicative Analysis. *Selected Contributions in data Analysis and Classification*, P. Brito, P. Bertrand, G. Cucumel, E. de Carvalho, (eds), Springer, p. 367-376, avec pour auteurs Régis Gras et Pascale Kuntz

2 Entropies de Shannon réduites et conditionnelles

2.1 Entropie d'une règle

Soit 2 règles. R=(a⇒b) et S=(c⇒d) où les variables a, b, c et d peuvent être elles-mêmes des conjonctions d'attributs ou, plus généralement, des règles. Soit p (resp. q) la fréquence de réalisation de R (resp. S) dans l'ensemble E des sujets.

L'entropie, au sens de Shannon, liée à R (resp. S) est :

$$H(R) = -(1-p)\log_2(1-p) - (p)\log_2(p) \text{ et } H(S) = -(1-q)\log_2(1-q) - (q)\log_2(q)$$

On rappelle que H(R) (resp. H(S)) est l'incertitude moyenne de « l'expérience » liée à la réalisation de la règle R (resp. S). En d'autres termes, c'est l'information moyenne attachée à la connaissance du résultat de l'expérience réalisant R (resp. S) ; c'est-à-dire encore, cette information est égale à la quantité d'information contenue dans la réalisation de R (resp . S). Comme ce qui nous intéresse en matière de gain informationnel est le cas où le nombre de contre-exemples à la règle est faible, eu égard aux occurrences en jeu (c'est-à-dire le cas où la fréquence p (resp. q) pour laquelle elle est vraie, est forte), nous limiterons notre étude à p (resp. q) à des valeurs supérieures ou égales à 0,5, ce qui permet une bijection de H sur l'intervalle [0,5; 1].

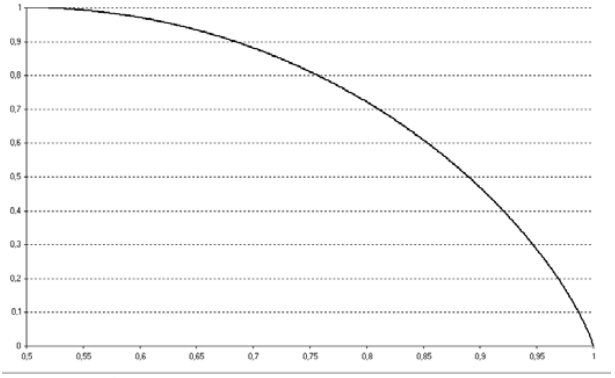


FIG. 27- *représentation graphique de la fonctions* H *sur* [0,5 ;1]

Donc, pour des raisons sémantiques, ne souhaitant pas la symétrie de H par rapport à p ou q égales à 0,5, nous posons la définition suivante.

Définition 39: On appelle ***entropie réduite*** **de la règle** R (resp. S) la quantité $H_r(R)$ (resp. $H_r(S)$) donnée par :

$H_r(R)= H(R)$ si $p \geq 0,5$ et $H_r(R)= 1$ si $p<0,5$

Si p=1 alors $H_r(R)=0$, l'incertitude au sujet de R est nulle car cette règle est certaine.

Si p=0,5 alors $H_r(R)=1$, l'incertitude au sujet de R est maximum. On ne peut faire le pari qu'elle se réalisera.

Les propriétés sont respectivement les mêmes pour S

2.2 Entropie conditionnelle d'une règle

Considérant maintenant le tableau croisant E (ensemble des individus) et les deux règles R et S considérées comme des variables prenant sur chaque individu, la valeur 1 ou 0 suivant que celui-ci vérifie ou non la règle en question.

Notons :

a) p_R (resp. p_S) la fréquence de R (resp. S) ; $p_{\overline{R}}$ (resp. $p_{\overline{S}}$) la fréquence de non(R) ou $\overline{R}$ (resp. non(S) ou $\overline{S}$)

b) p_{RS} (resp. $p_{\overline{RS}}$), la fréquence des règles R et S conjointes (resp. non(R) et S); $p_{R\overline{S}}$ (resp. $p_{\overline{RS}}$) la fréquence des règles R et non(S) (resp. non(R) et non(S))

L'entropie conditionnelle de S sachant R est alors :

$$H(S \mid R) = -p_{RS}\log_2(\frac{p_{RS}}{p_R}) - (p_{R\overline{S}})\log_2(\frac{p_{R\overline{S}}}{p_R}) - (p_{\overline{RS}})\log_2(\frac{p_{\overline{RS}}}{p_{\overline{R}}}) - (p_{\overline{RS}})\log_2(\frac{p_{\overline{RS}}}{p_{\overline{R}}})$$

Proposition 13 si R et S sont indépendantes, alors $H_r(S \mid R) = H_r(S)$

En effet $p_{RS} = p_R p_S$

$$H(S \mid R) = -p_{RS}\log_2(p_S) - (p_{R\overline{S}})\log_2(p_{\overline{S}}) - (p_{\overline{RS}})\log_2(p_S) - (p_{\overline{RS}})\log_2(p_{\overline{S}})$$

$$H(S \mid R) = -p_S\log_2(p_S) - (p_{\overline{S}})\log_2(p_{\overline{S}}) = H(S)$$

Ainsi, la connaissance de la réalisation de l'événement R ne modifie pas l'incertitude sur S.

Autre situation, si les règles R et S sont liées fonctionnellement, comme par exemple :

$$p_{RS} = p_R,\ p_{\overline{R}\,\overline{S}} = p_{\overline{R}},\ p_{R\overline{S}} = p_{\overline{RS}} = 0,$$

alors $H(S \mid R) = 0$ et il n'y a plus d'incertitude sur S.

En effet, nous obtenons en admettant que f(0)=0 est un prolongement de $f(x)=x\log_2(x)$ en x=0 : $H(S \mid R) = -p_R\log_2(\frac{p_R}{p_R}) - (0)\log_2(\frac{0}{p_R}) - (0)\log_2(\frac{0}{p_{\overline{R}}}) - (p_{\overline{R}})\log_2(\frac{p_{\overline{R}}}{p_{\overline{R}}})$

Enfin, la fréquence p_{RS} et les marges étant fixées, les autres fréquences étant déterminées, nous posons alors :

Définition 40: On appelle ***entropie conditionnelle réduite*** **de la règle S sachant R** (resp. la règle R sachant S) la quantité :

$H_r(S \mid R) = H(S \mid R)$ si $p_{RS} \geq 0{,}5$ et $H_r(S \mid R) = 1$ si $p_{RS} < 0{,}5$

$H_r(R \mid S) = H(R \mid S)$ si $p_{RS} \geq 0{,}5$ et $H_r(R \mid S) = 1$ si $p_{RS} < 0{,}5$

Ainsi, la différence $H_r(S) - H_r(S \mid R)$ est la quantité d'information ou l'incertitude contenue dans R au sujet de S, lorsque la fréquence p_{RS} est supérieure à 0,5, puisque c'est

l'accroissement de l'information sur S quand on connaît R. Autrement dit, c'est aussi la diminution de l'incertitude sur S dans l'hypothèse où la réalisation de R est connue.

2.3 Superfluité d'une règle

On pose $h(S) = \frac{\mathrm{H}(S)}{\log_2 N}$ où N est le nombre de valeurs que peut prendre S (ici N = 2 car S(x) = 1 ou 0 suivant que l'individu x satisfait ou non la règle S). Sachant que $H(S) \leq \log_2(N) = 1$ (Roubine, 1970) et que l'égalité n'a lieu que dans le cas uniforme où p=(1-p)=0,5 et, par conséquent, où l'incertitude est maximum, le rapport h(S) est une entropie *réduite relative* de S. Elle est toujours inférieure ou égale à 1. Si cette valeur est très voisine de 0, l'expérience S est quasiment *superflue* puisque l'une des deux probabilités p_S et $1\text{-}p_S$ est beaucoup plus grande que l'autre ; le « pari » sur l'une est inutile car on est presque sûr de l'issue de l'expérience.

On pose encore r(S) = 1 – h(S), pour une simple raison de compatibilité sémantique, à savoir que plus r(S) est grand, plus la superfluité est grande. r(S) est appelé ***coefficient de superfluité*** de S. On définit de la même façon h(R), entropie relative de R, et r(R), puis r(S/R) et r(R/S). Notons que $0 \leq r(S) \leq r(S/R) \leq 1$.

Définition 41: L'expérience S (donc la règle elle-même) est dite ***ε-superflue***, quand on connaît R, lorsque r(S/R), coefficient de superfluité de S sachant R, vérifie $r(S/R) \geq 1\text{-}\varepsilon$

On dit de même que R est ε-superflue, quand on connaît S, lorsque $r(R/S)) \geq 1\text{-}\varepsilon$

A travers cette notion de superfluité, nous possédons un moyen de faire décroître sensiblement le nombre de règles simples et généralisées formées au cours de l'étude hiérarchique des variables. L'utilisateur dispose d'un contrôle de la suppression éventuelle de règles en choisissant une valeur pour ε suffisamment petite. Mais, nous proposerons plus loin, un autre critère plus puissant.

Il est aisé de démontrer la symétrie :

$$H_r(S) - H_r(S \mid R) = \mathrm{I}(S,R) = \mathrm{I}(R,S) = H_r(R) - H_r(R \mid S)$$

c'est-à-dire que l'accroissement de l'information sur S quand on connaît R, autrement dit le ***gain informationnel***, est le même que l'accroissement de l'information sur R quand on connaît S. Cependant, les deux valeurs r(S) et r(R) ne sont pas nécessairement égales. Ainsi, si l'on cherche à éliminer une règle R ou S ou de type $R \Rightarrow S$, il sera bien sûr plus intéressant d'éliminer celle dont le coefficient de superfluité est le plus grand.

2.4 Redondance de règles

Afin de distinguer les gains associés aux expériences conditionnantes, nous considérons maintenant le rapport : $\frac{H_r(S) - H_r(S \mid R)}{H_r(S)}$ qui n'est pas négatif et qui, cette fois, n'est pas symétrique par rapport à R et S. Il représente un gain d'information relativisé par la grandeur de $H_r(S)$. Comme $H_r(S) \geq H_r(S \mid R)$ pour tout R, ce rapport est, par définition, égal à 0 quand $H_r(S) = 0$. Il varie entre 0 et 1.

Si R et S sont strictement liées, alors $H_r(S \mid R) = 0$, car il n'y a aucune incertitude sur S quand R est connue. La réciproque n'est pas vraie et, tout au moins, peut-on suspecter une certaine liaison de S à R.

Par ailleurs, si R et S sont indépendantes, alors $H_r(S) = H_r(S \mid R)$. La réciproque n'est pas vraie ; par contre, on peut suspecter l'indépendance.

Définition 42: Soit deux règles R et S. On appelle ***gain d'information relatif*** de S par R le rapport : $G_r(S \mid R) = \dfrac{H_r(S) - H_r(S \mid R)}{H_r(S)}$ et $G_r(S \mid R) = 1$ si $H_r(S) = 0$

Notons que :

Si $G_r(S/R) = 1$, alors nécessairement $H_r(S \mid R) = 0$. Dans ce cas, la diminution de l'incertitude sur S, soit la quantité d'information contenue dans R au sujet de S est maximum. Il est fort probable que S et R soient liées. S serait ***redondant****e* par rapport à R

Si $G_r(S/R) = 0$, la quantité d'information contenue dans R ou apportée par R au sujet de S est égale à ce qu'elle était sans connaître R. Il est fort probable que R et S soient indépendantes.

Si $H_r(S) = 0$, alors $H_r(S \mid R) = 0$. Un simple calcul montre alors que, de la même façon, $H_r(R \mid S) = 0$ et le prolongement par continuité en 0 de $G_r(S/R)$ se justifie aussi sémantiquement.

Définition 43: Soit deux règles R et S. On dit que S est ***ε-redondante*** *par rapport à R* si $G_r(S/R) \geq 1 - \varepsilon$

Étant donnée une suite de règles produite par une hiérarchie ordonnée, selon l'analyse statistique implicative, cette définition nous permet de réduire sensiblement, au gré de l'utilisateur, par action sur la valeur minimum acceptable 1- ε, le nombre de ces règles en conservant celles qui assurent un maximum d'information. L'algorithme que nous utilisons pour une automatisation de la réduction par rapport à la règle R de plus forte intensité d'implication et de plus forte fréquence dans E. R étant donnée, de proche en proche, on compare les superfluités et les redondances des règles d'intensité décroissante en éliminant celles qui sont superflues ou redondantes à un seuil 1- ε. A l'issue de ces éliminations, on itère le processus avec une règle R' d'intensité inférieure à R et on effectue les mêmes calculs parmi les règles restantes. Et ainsi de suite. Avec le même objectif de réduction du nombre de règles, nous allons maintenant examiner une autre approche dans un cadre conceptuel différent.

3 Information mutuelle au sens de l'indice de Gini

L'indice de Gini permet de mettre en évidence une distribution « inégalitaire » dans une population (comme par exemple les revenus !). Il semble donc adapté, comme l'était, dans notre choix précédent, l'entropie de Shannon, pour signifier et quantifier la dispersion au sein d'une distribution relative à la réalisation ou non de règles dans E. Étalonner d'une deuxième manière le « désordre » et donc la qualité informative d'une telle distribution nous paraît

d'un intérêt certain. Rappelons que l'indice de Gini est un cas particulier, pour α=2 de l'α-entropie de Havrda et Charvàt (1967) définie ainsi pour une variable R dont la distribution des probabilités (ou des fréquences) sur ses k valeurs est (p_1, p_2,…,p_k) ;

$$H(R) = \frac{1}{1-\alpha}\left[\sum_i p_i^{\alpha} - 1\right]$$

L'indice de Gini pour une telle variable binaire est donc :

$$Gini(R) = 1 - \sum_i p_i^2$$

On peut démontrer, par un développement au premier ordre, que quand α tend vers 1, la limite de l'ε-entropie est justement l'entropie de Shannon. D'où la proximité des sémantiques de l'une et l'autre.

Interprétons l'indice de Gini :

a) $Gini(R) = 1 - \sum_i p_i^2$ peut encore s'écrire $Gini(R) = \sum_i (1 - p_i) p_i$ puisque la somme des probabilités est égale à 1. L'indice de Gini peut donc s'interpréter en terme de variance, donc de quantité d'information : celle d'une somme de variables aléatoires indépendantes de Bernoulli de paramètres respectifs p_i ;

b) $Gini(R) = 1 - \sum_i p_i^2$ peut s'interpréter comme l'écart entre les normes de deux vecteurs de dimension k et respectivement de composantes toutes égales à $\frac{1}{\sqrt{k}}$ pour le 1er vecteur, les autres étant égales à p_i pour le 2ème vecteur.

Par exemple, dans le cas qui nous intéresse (satisfaction ou non d'une règle), k est égal à 2 et le minimum de variance ou celui de d'écart est obtenu pour p = 1-p = 0,5. Propriété que nous avions déjà observée avec l'entropie de Shannon. On démontre, par étude de la limite, que lorsque p tend vers 1 (très bonne représentativité de R), la différence entre H(R) et Gini(R) est approximativement 1,6 p(1-p)>0. De plus, quand p>0,5, H(R) reste meilleure que Gini(R). Autrement dit, pour l'intervalle de variation de p qui nous intéresse, l'entropie de Shannon est plus informative que l'indice de Gini. Ou, de façon équivalente, sauf pour l'utilisateur, que l'indice de Gini est plus sévère que l'entropie de Shannon.

Comme nous l'avons fait pour l'entropie de Shannon, on considère l'indice conditionnel de Gini dans la définition suivante (Jaroszewicz et Simovici 2001)[25] :

Définition 44: Deux règles R et S étant données, de modalités respectives avec les fréquences p_i pour R, p_{ij} pour (R, S), q_j pour S, **l'indice conditionnel de Gini** de S sachant R est :

$$Gini(S \mid R) = 1 - \sum_i \sum_j \frac{p_{ij}^2}{p_j}$$

[25] Nous reprenons ici la démarche de Jaroszewicz S. et Simovici D. mais en spécifiant leur approche, comme celle de J. Blanchard portant sur l'indice informationnel entre variables, à des règles généralisées.

On pourra remarquer que la somme a le sens d'une somme de variances conditionnelles généralisées et, donc, contient encore le sens d'une information.

Dans le cas qui nous intéresse ici, les modalités de R et de S sont « vrai » et « faux », suivant que l'individu x satisfait ou non la règle R ou la règle S. La formule donnant l'indice est alors avec les notations adoptées dans ce chapitre :

$$Gini(S \mid R) = 1 - \left(\frac{p_{RS}^2}{p_R} + \frac{p_{R\bar{S}}^2}{p_R} + \frac{p_{\bar{R}S}^2}{p_{\bar{R}}} + \frac{p_{\bar{R}\bar{S}}^2}{p_{\bar{R}}} \right)$$

On peut alors définir, comme avec l'entropie de Shannon, le gain de Gini qui sera aussi un gain informationnel sur S quand on connaît R :

Définition 45: Deux règles R et S étant données, on appelle ***gain de Gini*** *pour S sachant R*, l'accroissement d'information suivant :

$$gainGini(S \mid R) = Gini(S) - Gini(S \mid R)$$

Ce gain apparaît comme une différence de variances et, par conséquent, un indicateur de la qualité de l'information fournie sur S par la connaissance de R. Remarquons que si R et S sont indépendantes, alors $Gini(S) = Gini(S \mid R)$ et $gainGini(S \mid R) = 0$, comme pour le gain de Shannon. La réciproque n'est pas vraie. On peut, comme avec le gain informationnel relatif selon l'entropie de Shannon, relativiser le gain de Gini et nous donner ainsi un deuxième critère pour obtenir une réduction de l'ensemble des règles acceptées sous une qualité implicative ou cohésitive compatible avec l'attente de l'utilisateur. Prochainement, une comparaison de l'efficacité de réduction selon ces deux modes sera conduite par simulation.

4 Conclusion

Dans le cadre de l'ASI et après l'obtention de règles simples ou généralisées, nous avons cherché un moyen qui puisse conduire à la réduction du nombre de telles règles obtenues. Le moyen retenu dans ce texte est centré sur la prise en compte de l'information respective apportée par la réalisation d'une règle si l'on connaît la réalisation d'une autre. Pour cela, nous avons fait appel à deux méthodes qui ne se ramenant pas conceptuellement l'une à l'autre pourraient s'avérer complémentaires : la première en examinant l'entropie conditionnelle d'une règle quand on en connaît une autre ; la deuxième en utilisant l'indice de Gini conditionnel. Comparer ces deux méthodes fera l'objet de prochaines recherches.

Références

Acid S. de Campos, L.M., A. Gonzalez, R. Molina and N. Perez de la Blanca (1991). Learning with Castle, in R. Kruse, P. Siegel (Eds) *Symbolic and quantitative Approaches to incertainty*, Springer-Verlag, 99-106

Ag Amouloud, S. (1992). *L'ordinateur, outil d'aide à l'apprentissage de la démonstration et de traitement de données didactiques*, Thèse de doctorat de l'Université de Rennes 1.

Agrawal, R., T. Imielinsky and A. Swami (1993). Mining association rules between sets of items in large databases, *Proc. of the ACM SIGMOD'93*, 207-216

Amarger, S., D. Dubois and H. Prade (1991). Imprecise quantifiers and conditional probabilities, in R. Kruse, P. Siegel (Eds), *Symbolic and quantitative approaches to uncertainty* Springer-Verlag, 33-37.

Aze, J. et Y. Kodratoff (2001). Évaluation de la résistance au bruit de quelques mesures d'extraction de règles d'association, *Extraction des connaissances et apprentissage*, Hermès, Vol 1, n° 4, 143-154

Bailleul, M. (1994). *Analyse statistique implicative: variables modales et contribution des sujets. Application à la modélisation de l'enseignant dans le système didactique*, Thèse de doctorat de l'Université de Rennes 1..

Bailleul, M. et R. Gras (1995). L'implication statistique entre variables modales, *Mathématique, Informatique et Sciences Humaines*, Paris : E.H.E.S.S., n°128, 41-57

Benzecri, J.P. (1973*). L'analyse des données* (vol 1), Paris : Dunod,.

Bernard, J.-M. et S. Poitrenaud (1999). L'analyse implicative bayesienne d'un questionnaire binaire : quasi-implications et treillis de Galois simplifié", *Mathématiques, Informatique et Sciences Humaines*, n° 147, 25-46

Bernadet, M., G. Rose and H. Briand (1996). FIABLE and fuzzy FIABLE : two learning mechanisms based on a probabilistic evaluation of implications, *Conference IPMU'96*, Granada, 911-916

Bernadet, M. (2004). Qualité des règles et des opérateurs en découverte de connaissances floues. *Mesure de qualité pour la fouille de données*, RNTI-E-1, Toulouse *:* Cépaduès Éditions,169-192

Blanchard, J., P. Kuntz, F. Guillet and R. Gras (2003). Implication intensity: from the basic statistical definition to the entropic version , *Statistical Data Mining and Knowledge Discovery*, Chapman & Hall/CRC, Washington, 473-485

Blanchard, J., P. Kuntz, F. Guillet and R. Gras (2004), Mesure de la qualité des règles d'association par l'intensité d'implication entropique, *Mesures de qualité pour la fouille de données*, RNTI-E-1, Toulouse *:* Cépaduès Éditions, 33-44

Blanchard, J., F. Guillet, R. Gras et H. Briand (2004) Mesurer la qualité des règles et de leurs contraposées avec le taux informationnel TIC, *Extraction et Gestion des Connaissances*, Volume 1, RNTI, Cépaduès. 287-298.

Références

Blanchard, J., F. Guillet, H. Briand et R. Gras (2005). Ipee : Indice probabiliste d'écart à l'équilibre pour l'évaluation de la qualité des règles, *Extraction et Gestion des Connaissance : état et perspectives*, RNTI-E-5, Toulouse *:* Cépaduès Éditions, 391-395.

Bodin, A. (1997). Modèles sous-jacents à l'analyse implicative et outils complémentaires. *Prépublication* IRMAR. n°97-32, 1-24

Bodin, A. et R. Gras (1999). Analyse du préquestionnaire enseignants avant EVAPM-Terminales, *Bulletin n°425 de l'Association des Professeurs de Mathématiques de l'Enseignement Public*, Paris 772-786,

Brin, S., R. Motwani and C. Silverstein (1997) Beyond market baskets : generalizing association rules to correlations", *Proc. Of ACM SIGMOD Conf. On Management of Data SIGMOD'97*, 265-276

Brousseau, G. (1986). Fondements et méthodes de la didactique des mathématiques, *Recherche en Didactique des Mathématiques, 4/2*, Grenoble La Pensée Sauvage.

Bruhat, G. (1959). *Optique*, Paris : Masson.

Couturier, R. (2001). Traitement de l'analyse statistique implicative dans CHIC, Actes des Journées sur la *Fouille dans les données par la méthode d'analyse implicative*, IUFM Caen, 33-50.

Couturier, R, R. Gras and F. Guillet (2004). Reducing the number of variables using implicative analysis *In International Federation of Classification Societies, IFCS 2004, Springer Verlag: Classification, Clustering, and Data Mining Applications*, Chicago, 277--285.

Couturier, R. et R. Gras (2005). CHIC : Traitement de données avec l'analyse implicative, *Extraction et Gestion des Connaissances*, Volume I1, RNTI, Toulouse *:* Cépaduès Éditions, 679-684

Cox, D. R. and N. Wermuth (2004) Causality : a Statistical View, *International Statistical Review*, International Statistical Institute, 72, 3, 285-305,.

David, J., F. Guillet, R. Gras and H. Briand (2006). Conceptual hierarchies matching : an approach based on discovery of implication rules between concepts, In Proc. ECAI 2006, 17th European Conference on Artificial Intelligence, IOS Press, Riva del Garda, Italy

Diday, E. (1972). *Nouvelles méthodes et nouveaux concepts en classification automatique et reconnaissance des formes*, Thèse d'État, Université de Paris VI.

Diday, E. et M.O Menessier (1991). Analyse symbolique pour la prévision de séries chronologiques pseudo-périodiques in *Induction symbolique-numérique à partir de données"*, Cépaduès Editions.

Dubois, D. et H. Prade (1987). *Théorie des possibilités. Applications à la représentation des connaissances en informatique*, Paris : Masson.

Durkheim, E.(1897). *Le suicide*, Paris :PUF.

Fayyad, U., G. Piatetsky-Shapiro and P. Smyth (1996). From Data Mining to Knowledge Discovery. In *Advances In Knowledge Discovery and Data Mining, Fayyad U., Piatetsky-Shapiro G., Smyth P, and Uthurusamy R. eds*, AAAI/MIT Press, 1-31,.

Frawley, W., G. Piatetski-Shapiro and C. Matheus (1992). Knowledge discovery in databases : an overview. *AI Magazine*. 14(3), 57-70.

Fleury, L. (1996). *Extraction de connaissances dans une base de données pour la gestion de ressources humaines*, Thèse de doctorat de Université de Nantes.

Gammerman, A. et Z. Luo (1991). Constructing Causal Trees from a medical database, *Technical Report TR 91 002*, Dept of computer Sci. Heriot-Watt, Univ Edimburgh

Ganascia, J.G. (1987). *AGAPE et CHARADE : deux techniques d'apprentissage symbolique appliquées à la construction de bases de connaissances*, Thèse d'Etat, Université de Paris Sud

Ganascia, J. G. (1991). *CHARADE : Apprentissages de bases de connaissances dans "Induction symbolique - numérique à partir de données*, Toulouse *:* Cépaduès Éditions.

Goldberg ,D. (1994). *Algorithmes génétiques*, Addison-Wesley France

Goodman, R.M. et P. Smyth (1989). The induction of probabilistic rule set. The ITRULE algorithm, *Proceedings of sixth international conference on machine learning*, 129-132

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'Etat, Université de Rennes 1.

Gras, R. et A. Larher (1992). L'implication statistique, une nouvelle méthode d'analyse de données, *Mathématique, Informatique et Sciences Humaines*, E.H.E.S.S. Paris, 120, 5-31

Gras, R., H. Briand and P. Peter (1996a). Structuration sets with implication intensity, *Proceedings of the International Conference on Ordinal and Symbolic Data Analysis* - OSDA 95, E. Diday, Y. Chevallier, O. Opitz (Eds.), Paris : Springer, 147-156

Gras, R. et H. Ratsimaba-Rajohn (1996b). Analyse non symétrique de données par l'implication statistique. *RAIRO-Recherche Opérationnelle, 30-3*, AFCET, Paris, 217-232

Gras, R., S. Ag Almouloud, M. Bailleul, A. Larher, M. Polo, H. Ratsimba-Rajohn et A. Totohasina (1996 c). *L'implication Statistique*, Collection Associée à Recherches en Didactique des Mathématiques, Grenoble : La Pensée Sauvage.

Gras, R. (1997a). Nœuds et niveaux significatifs en Analyse Statistique Implicative, Prépublication IRMAR, 97-32, 1-11 (3)

Gras, R., H. Briand, P. Peter and J. Philippé (1997b). Implicative statistical analysis, *Proceedings of International Congress I.F.C.S., 96*, Kobé Tokyo: Springer-Verlag, 412-419

Gras, R., E. Diday, P. Kuntz and R. Couturier (2001a). Variables sur intervalles et variables-intervalles en analyse statistique implicative, *Actes du 8ème Congrès de la Société Francophone de Classification*, Université des Antilles-Guyane, 166-173

Gras, R., P. Kuntz, R. Couturier et F. Guillet (2001b). Une version entropique de l'intensité d'implication pour les corpus volumineux. *Extraction des Connaissances et Apprentissage (ECA*), vol. 1, n° 1-2, 69-80. Hermès Science Publication.

Gras, R., P. Kuntz et H. Briand (2001c). Les fondements de l'analyse statistique implicative et quelques prolongements pour la fouille de données, *Mathématiques et Sciences Humaines*, n° 154-155, 9-29

Gras, R(égis), F. Guillet, R(obin) Gras et J. Philippé (2002). Réduction des colonnes d'un tableau de données par quasi-équivalence entre variables, *Extraction des connaissances et apprentissage*, Paris : Hermès, Volume 1, n°4/2001, 197-202.

Gras, R., P. Kuntz et H. Briand (2003). Hiérarchie orientée de règles généralisées en analyse implicative, *Extraction des Connaissances et apprentissage*, Paris : Hermès, 145-157.

Gras, R., R. Couturier, J. Blanchard, H. Briand, P. Kuntz et P. Peter (2004a). Quelques critères pour une mesure de qualité de règles d'association. Un exemple : l'implication statistique, *Mesures de qualité pour la fouille de données*, RNTI-E-1 Toulouse *:* Cépaduès Éditions, 3-32.

Gras, R., P. Kuntz et J.C. Régnier (2004b). Significativité des niveaux d'une hiérarchie orientée en analyse statistique implicative, *Classification et fouille de données*, M. Chavent et M. Langlais (Eds), *RNTI-C-1*, Toulouse *:* Cépaduès Éditions, 39-50.

Gras, R. et P. Kuntz (2005). Discovering R-rules with a directed hierarchy, *Soft Computing, A Fusion of Foundations, Methodologies and Applications*, Volume 1, Springer Verlag, 46-58..

Gras, R., J. David, J.C. Régnier et F. Guillet (2006). Typicalité et contribution des sujets et des variables supplémentaires en Analyse Statistique Implicative. *Extraction des Connaissances* (EGC'06), Volume2, Toulouse *:* Cépaduès Éditions, 359-370,

Gras, R., J. David, F. Guillet et H. Briand (2007a). Stabilité en A.S.I. de l'intensité d'implication et comparaisons avec d'autres indices de qualité de règles d'association, *Proceedings atelier « Qualité des données et des connaissances »*, EGC 07, Namur

Gras, R., P. Kuntz et E. Suzuki (2007b). Une règle d'exception en Analyse Statistique Implicative, *Extraction des Connaissances* (EGC'07),Volume1, RNTI-E-9 Toulouse *:* Cépaduès Éditions 87-98.

Gras, R. et P. Kuntz P. (2007c). Reduction of Redundant Rules in Statistical Implicative Analysis. P. Brito, P. Bertrand, G. Cucumel, E. de Carvalho, (Eds) *Selected Contributions in data Analysis and Classification*, Springer, 367-376

Havrda, J.H. and F. Charvàt (1967). Quantification Methods of Classification Processes, *Concepts of Structural Entropy*, *Kybernetica*, *3*, 30-37

Hempel, C. G. (1945). Studies in the Logic of Confirmation, *Mind* 54, 1-26

Hipp, J., U. Guntzer and J. Nakhaeizadeh (2000). Mining association rules: Deriving a superior algorithm by analyzing today's approach , *Proc. of 4th Eur. Conf. on Principles of Data Mining and Knowledge Discovery*, Lect. N. in Art. Int. 1910, 160-168.

Horschka, P. et W. Klögsen (1991). A support system for interpreting statistical data. *Knowledge Discovery in Databases*, AAAI Press, 325-345,.

Jaroszewicz, S. and D. Simovici (2001). *A general Measure of Rule Interestingness* Berlin Heidelberg : Springer-Verlag, 253-265

Kendall, M. G. and A. Stuart (1991). *Kendall's advanced theory of statistics*. (Vol. 2) London : Edward Arnold.

Kuntz, P., R. Gras and J. Blanchard (2002). Discovering Extended Rules with Implicative Hierarchies, *Conference on the new frontiers of statistical data mining and knowledge discovery*, Knoxville, Tennessee

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

Lahanier-Reuter, D. (1998). *Étude de conceptions du hasard : approche épistémologique, didactique et expérimentale en milieu universitaire*, Thèse de doctorat de l'Université de Rennes 1.

Lallich S, Lenca P. et Vaillant B. (2005) Variations autour de l'intensité d'implication, *Actes ASI 03*, Université de Palerme.

Larher, A. (1991). *Implication statistique et applications à l'analyse de démarches de preuve mathématique*, Thèse de doctorat de l'Université de Rennes 1.

Lehn, R. (2000). *Un système interactif de visualisation et de fouille de règles pour l'extraction de connaissances dans une base de données*. Thèse de doctorat de l'Université de Nantes.

Lebart, L., M. Piron et A. Morineau (2006). *Statistique exploratoire multidimensionnelle*. (4ème édition), Paris : Science sup, Dunod.

Lent, B., A.N. Swami and J. Widow (1997). Clustering association rules. *Proc. of the 13th Int. Conf. on Data Engineering*, 220-231.

Lenca, P., P. Meyer, P. Vaillant, P. Picouet et S. Lallich (2004). Évaluation et analyse multi-critères de qualité des règles d'association, *Mesures de qualité pour la fouille de données*, RNTI-E-1, Toulouse *:* Cépaduès Éditions, 219-246.

Lerman, I.-C. (1981). *Classification et analyse ordinale des données*. Paris : Dunod,.

Lerman, I.-C., R. Gras. et H. Rostam (1981). Élaboration et évaluation d'un indice d'implication pour des données binaires, I et II, *Mathématiques et Sciences Humaines* ,, n° 74,, 5-35 et n° 75, 5-47

Lerman, I.C. et J. Azé (2004). Indice probabiliste discriminant (de vraisemblance du lien) d'une règle d'association en cas de « très grosses » données, *Mesures de qualité pour la fouille de données*, RNTI-E-1, Toulouse *:* Cépaduès Éditions, 69-94.

Loevinger, J., (1947). A systematic approach to the construction and evaluation of tests of abilities, *Psychological Monographs, 61*, n° 4.

Lotfi, A. et L.A**.** Zadeh (2001). From computing with numbers to computing with words from manipulation of measurements to manipulation of perception, in *Proceedings*

"Human and machine perception" (Thinking, deciding and acting), V. Cantoni, V. Di Gesù, A. Setti e D. Tegolo Eds, Kluwer Academic, New York.

Matheus, C.J., G. Piatetsky-Shapiro and D. Mc Neill (1996). Selecting and reporting what is interesting, *In Advances in Knowledge Discovery and Data Mining*, AAAI press/MIT Press, 495-515

Pearl, J. (1988). *Probabilistic Reasoning in intelligent systems*, San Mateo, CA, Morgan Kaufmann.

Pearl, J. (1995). Causal diagrams for empirical research, *Biometrika*, 82, 4, Great Britain, 669-710.

Polo-Capra, M. (1996). *Le repère cartésien dans les systèmes scolaires français et italien : étude didactique et application de méthodes d'analyse statistiques multidimensionnelles*, Thèse de doctorat de l'Université de Rennes 1.

Ralambondrainy, H. (1991). Apprentissage dans le contexte d'un schéma de base de données in *Induction symbolique-numérique à partir de données*, Toulouse *:* Cépaduès Éditions

Ratsimba-Rajohn, H. (1992). *Contribution à l'étude de la hiérarchie implicative. Application à l'analyse de la gestion didactique des phénomènes d'ostension et de contradiction*, Thèse de doctorat de l'Université de Rennes 1.

Régnier, J.C. et R. Gras (2005). Statistique de rangs et analyse statistique implicative, *Revue de Statistique Appliquée*, LIII, 5-38

Ritschard, G., D. A. Zighed et N. Nicoloyannis (2000). Maximiser l'association par agrégation dans un tableau croisé, R. Gras et Bailleul M. (Eds) *Journées sur la Fouille dans les données par la méthode d'analyse statistique implicative*, , 219-233.

Roubine, E. (1970). Introduction à la théorie de la communication, (Tome III) *Théorie de l'information*, Paris : Masson,

Saporta, G. (2006). *Probabilités, Analyse de Données et statistique*, Paris : Ed. Technip,

Schektman, Y., J. Trejos et M. Troupe (1992). Un générateur de règles floues à partir de bases de données volumineuse, *Actes des 3èmes Journées "Symboliques-Numériques"*, mai 1992, Paris.

Sebag M. et Schoenauer (1991). Un réseau de règles d'apprentissage, *Induction symbolique-numérique à partir de données*, Toulouse *:* Cépaduès Éditions.

Shannon, C.E. et W. Weaver (1949). *The mathematical theory of communication*, Univ. of Illinois Press.

Simon, A., (2000*). Outils classificatoires par objets pour l'extraction de connaissances dans des bases de données*, Thèse de doctorat de l'Université de Nancy.

Spagnolo, F. et R. Gras (2004). A new approach in Zadeh's classification : fuzzy implication through statistic implication, *NAFIPS 2004, 23nd Conference of the North American Fuzzy Information Processing Society, Banff*, AB Canada, 27-30.

Suzuki, E. and Y. Kodratoff (1999). Discovery of surprising exception rules based on intensity of implication. *Principles of data mining and knowledge discovery science.* Springer, 184-195.

Suzuki E. and J. Zytkow (2005). Unified algorithm for undirected discovery of exception rules, *Int. J. of Intelligent Systems*, vol. 20, Wiley, 673-694

Totohasina, A. (1992) *Méthode implicative en analyse de données et application à l'analyse de conceptions d'étudiants sur la notion de probabilité conditionnelle*, Thèse de doctorat de l'Université de Rennes I.

Vergnaud, G., (1994). La théorie des champs conceptuels, *Recherches en Didactique des Mathématiques*, 10/2-3, Grenoble : La Pensée Sauvage, 133-170.

Zadeh, L.A. (1979). A Theory of Approximate Reasoning, J. Hayes, D. Michie, and L.I. Mikulich (Eds)., *Machine Intelligence 9*, New York: Halstead Press, 149-194.

Zadeh, L.A. (1997). Toward a Theory of Fuzzy Information Granulation and its Centrality in Human Reasoning and Fuzzy Logic, *Fuzzy Sets and Systems 90*, 111-127.

Zighed, D. A. et R. Rakotomalala (2000). *Graphes d'induction: apprentissage et data mining*. Paris: Hermes Science Publications.

Summary

This first part, structured in 9 chapters, consists of a general presentation of the Statistical Implicative Analysis (SIA). The authors are both defining the concepts and theorems of this theoretical approach, and also the methodological and epistemological foundations. In particular, the reader will find the definitions of: the relation of quasi-implication, the implication index, the intensity of implication, the index of propensity, the tree of implication, the directed hierarchy of implications... In addition, each concept is illustrated through an example.

PARTIE 2

COMPLEMENTS ET EXTENSIONS DE L'ANALYSE STATISTIQUE IMPLICATIVE

Chapitre 1 : Statistique de rangs et Analyse Statistique Implicative

Jean-Claude Régnier [*] et Régis Gras [**]

* Université de Lyon - UMR 5191 ICAR
ENS-LSH 15, Parvis René Descartes BP 7000 69342 LYON cedex 07
jean-claude.regnier@univ-lyon2.fr

*Equipe Connaissances et Décision, Laboratoire d'Informatique de Nantes Atlantique
Ecole Polytechnique de l'Université de Nantes, UMR 6241
La Chantrerie BP 60601 44306 Nantes cedex
regisgra@club-internet.fr

Résumé. Nous discutons de l'apport de la méthode d'analyse statistique implicative au sens de R. Gras, à l'étude de la concordance/discordance des rangs accordés par des juges à des objets. Cette dernière est à comprendre au sens de Friedman ou de Kendall. Ici nous comparons une analyse de préférences exprimées par les rangs, avec l'analyse de la propension entre variables modales de J. B. Lagrange. Nous nous affranchissons de l'hypothèse d'absence de lien *a priori* entre les variables. Nous affectons d'une mesure de qualité des énoncés de la forme : « si l'objet a est rangé par les juges alors, généralement, l'objet b est rangé à un rang meilleur par les mêmes juges », et représentons par un graphe les relations de préférences de l'ensemble des objets rangés. Nous nous limitons aux deux cas des rangements complets et incomplets mais sans *ex æquo* de q objets par k juges. Le texte présenté ici reprend en partie (Régnier et Gras, 2005)

1 Introduction.

Rappelons la problématique de la concordance des rangements de q objets par k juges. Considérons q objets soumis au jugement individuel et indépendant de k (k>2) individus d'un ensemble E qui doivent fournir un rangement complet et sans *ex æquo* de ces objets. Il s'agit de savoir si on peut dégager de l'ensemble des k jugements, une concordance

relativement à un rangement de ces q objets. C'est une des formes de **l'analyse des préférences** où, ici, nous n'envisageons pas de comparaison par paires.

La réalisation d'une telle opération peut être décrite par le tableau suivant :

Objets Individus	$\mathbf{O_1}$	$\mathbf{O_2}$	…	$\mathbf{O_j}$	…	$\mathbf{O_q}$
1	$R_{11}^{(q)}$	$R_{12}^{(q)}$		$R_{1j}^{(q)}$		$R_{1q}^{(q)}$
…						
i	$R_{i1}^{(q)}$	$R_{i2}^{(q)}$		$R_{ij}^{(q)}$		$R_{iq}^{(q)}$
…						

TAB. 1 - *Tableau de rangs*

où chaque ligne i (i = 1,…, k) est une permutation de (1, 2, …,q), réalisations du vecteur aléatoire $\left(R_{ij}^{(q)}\right)_{j=1,...,q}$. Il y a donc q ! réalisations possibles par ligne. Les contraintes de rangement complet et sans *ex æquo* conduisent à ce que chaque ligne i correspondant à l'individu i vérifie la relation $\sum_{j=1}^{q} R_{ij}^{(q)} = \frac{q(q+1)}{2}$. Sous l'hypothèse Ho, généralement légitime, de l'équiprobabilité des q ! rangements des q objets par l'individu i, la probabilité pour l'objet j, $j \in \{0, 1, \ldots, q\}$, d'être placé au rang r, $r \in \{0, 1, \ldots, q\}$ est $\frac{1}{q}$. La probabilité pour que l'objet j_1, $j_1 \in \{0, 1, \ldots, q\}$, soit placé au rang r_1, $r_1 \in \{0, 1, \ldots, q\}$, et l'objet j_2, $j_2 \in \{0, 1, \ldots, q\}$ - $\{ j_1 \}$, soit placé au rang r_2, $r_2 \in \{0, 1, \ldots, q\}$ et $r_2 \neq r_1$ est de $\frac{1}{q(q-1)}$. Les caractéristiques des variables aléatoires $R_{ij}^{(q)}$ sont :

Les caractéristiques des variables aléatoires $R_{ij}^{(q)}$ sont :

Espérance	Variance	Covariance	Corrélation
$E(R_{ij}^{(q)}) = \frac{q+1}{2}$	$V(R_{ij}^{(q)}) = \frac{q^2-1}{12}$	$Cov(R_{ij_1}^{(q)}; R_{ij_2}^{(q)}) = -\frac{q+1}{12}$	$Corr(R_{ij_1}^{(q)}; R_{ij_2}^{(q)}) = -\frac{1}{q-1}$

TAB. 2 - *Tableau des caractéristiques des variables rangs aléatoires*

2 Approche par les statistiques de Friedman et de Kendall

Considérons le tableau TAB.1. Les sommes marginales des lignes sont constantes et égales à la somme des rangs attribués par un même individu. Il comporte $(q !)^k$ réalisations équiprobables sous l'hypothèse Ho d'équiprobabilité des permutations caractérisant le rangement des q objets effectué indépendamment par chaque individu i. Désignons par $S(O_j)$ la somme des rangs attribués respectivement par les k individus à l'objet O_j. Pour avoir la distribution de probabilité exacte de $S(O_j)$, il suffirait de construire les $(q !)^k$ tableaux, opération, concrètement impossible dans les cas intéressants. Sous Ho, nous pouvons cependant connaître les caractéristiques des variables $S(O_j)$, $j \in \{0, 1, \ldots, q\}$.

Espérance	Variance
$E(S(O_j))= k\frac{q+1}{2}$	$V(S(O_j))= k\frac{q^2-1}{12}$

TAB. 3 - *Tableau des caractéristiques des variables aléatoires S(Oj), somme des rangs*

2.1 Interprétation des $S(O_j)$ pour expliciter un rangement concordant

Nous pouvons utiliser l'ordre des sommes des rangs attribués à chaque objet pour déterminer un rangement que nous interprétons comme étant celui qui « s'approche au mieux » de chaque point de vue des juges. Deux cas extrêmes peuvent apparaître.

2.1.1 Concordance « parfaite »

Cette situation correspond au cas où les juges attribueraient respectivement à chaque objet le même rang.

Il y a q! tableaux de ce type sur les $(q!)^k$ possibles. Sous Ho, cet événement a donc une probabilité $\left(\frac{1}{q!}\right)^{k-1}$ de survenir.

2.1.2 Hétérogénéité « parfaite » des rangements

Cette situation correspond au cas où l'attribution des rangs aux q objets produirait par des effets de compensation des sommes de rangs égales ou presque égales à une unité près. Le nombre de tableaux vérifiant ces conditions est à l'évidence très supérieur aux q! tableaux de la concordance « parfaite ». Intuitivement, sous Ho, nous avons beaucoup plus de chance d'obtenir un tableau du type « hétérogénéité parfaite » que du type « concordance parfaite ».

2.1.3 Entre « concordance parfaite » et « hétérogénéité parfaite »

La question demeure quant à l'interprétation des cas intermédiaires. Quelle décision prendre si nous obtenons des résultats compris entre ces deux extrêmes ? La **statistique $F_{q,k}$ de Friedman ou celle W de Kendall** constituent des outils d'aide à la prise de cette décision.

$$F_{q,k} = \frac{12k}{q(q+1)}\sum_{j=1}^{q}\left(\frac{S(O_j)}{k} - \frac{q+1}{2}\right)^2$$

$$W = \frac{\sum_{j=1}^{q}\left(S(O_j) - k\frac{(q+1)}{2}\right)^2}{\frac{k^2(q^3-q)}{12}} = \frac{S_K}{S_{K\max}}$$

Ces deux statistiques sont reliées par : $F_{q,k} = k(q-1)W$. Elles constituent des mesures d'écart entre la situation observée et celle de l'hétérogénéité « parfaite ».

2.2 Test de concordance des rangements de q objets par k juges, W de Kendall

La loi de W sous Ho ne dépend que des $(q!)^k$ résultats possibles. Elle est théoriquement connaissable mais pratiquement impossible en raison de la taille des calculs en jeu. On obtient les valeurs critiques à partir d'une table qui a été établie lorsque q<7. Sinon pour q ≥ 7, on utilise la loi de la statistique k(q-1)W qui est approximativement la loi du $\chi^2_{(q-1)}$ à q-1 ddl.

3 Analyse statistique implicative, outil d'exploration des rangements complets et sans *ex aequo*.

À l'origine (Gras et al, 2001), comme nous avons pu le voir dans la première partie de cet ouvrage, l'analyse statistique implicative s'applique aux variables binaires en produisant des règles, puis aux classes de variables binaires en produisant des méta-règles (R-règles). Cette méthode d'analyse de données constitue un outil pour expliciter statistiquement des structures quasi-implicatives. Nous rappelons le tableau de contingence qui constitue la référence sur laquelle se fonde le raisonnement quasi-implicatif.

		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. 4 - *Tableau de contingence de référence du raisonnement*

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).

$0 \le n(a\wedge\neg b) < \frac{n(a)n(\neg b)}{n}$	L'effectif observé étant inférieur à l'effectif théorique signifie que la dépendance entre la présence de (a) et l'absence de (b) est répulsive. En terme d'implication statistique, l'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)

TAB. 5 - *Tableau des interprétations qui fonde le raisonnement par quasi-implication.*

Cela conduit à définir une implication statistique admissible à un niveau de confiance fixé. Par la suite le rapprochement entre variables de rangs et analyse statistique implicative apparaît avec l'extension de cette théorie aux variables et classes de variables modales à valeurs sur [0 ;1] de IR. Dans ce passage des variables binaires vers les variables modales,

J.B. Lagrange (Lagrange, 1998) avait introduit la notion de « **relation de propension** entre deux variables à valeurs dans l'intervalle [0 ;1] ». Comme nous l'avons exposé en Partie 1, il a été défini un *indice (coefficient) de propension* $\tilde{q}(a,\neg b)$ d'une variable modale a vers une variable modale b ainsi qu'une ***intensité de propension*** $\Phi(-\tilde{q}(a,\neg b))$. Le modèle adopté est une stricte extension de l'implication statistique entre variables binaires. Dans le cadre de l'approche par l'analyse implicative, le recours aux variables de rangs est apparu dans les travaux de Marc Bailleul (1994). Pour explorer les représentations de l'enseignement des mathématiques, il a été proposé un protocole qui amène les sujets interrogés à fournir un rangement. Dans le choix de l'outil mathématique d'analyse, M.Bailleul inscrit la contrainte de la prise en compte d'un indice non-symétrique permettant de répondre à la question : *dans quelle mesure peut-on affirmer, de façon relativement sûre, que le choix par un individu de telle variable va entraîner le choix de telle autre variable ?* Et concernant le cas des variables de rangs qui nous intéresse : *dans quelle mesure le choix de telle variable avec tel rang entraîne-t-il le choix de telle autre variable avec un rang meilleur ou égal à celui affecté à la première ?*

3.1 Transformation d'une variable de rang en une variable modale

En premier lieu, nous nous plaçons dans le cas d'un rangement complet et sans *ex æquo*. En général il est demandé aux k individus de ranger les q objets par ordre de préférence décroissante. Ainsi le rang 1 est le plus important tandis que le rang q correspond à l'objet le moins important. Or dans le modèle proposé par J.B Lagrange la valeur 1 représente le plus haut degré d'adhésion possible tandis que le degré minimum d'adhésion est coté par 0. Dans ces conditions, la transformation suivante est une solution à ce problème : $X_{ij}^{(q)} = \frac{q - R_{ij}^{(q)}}{q-1}$ qui respecte la contrainte sémantique.

3.2 Le modèle de la quantification de la propension de J.B. Lagrange

Étant données deux variables modales a et b, $(a_i)_{i=1,\ldots,n}$ et $(b_i)_{i=1,\ldots,n}$ deux vecteurs de $[0\ ;\ 1]^n$ représentant une réalisation de ces variables auprès des n individus d'un ensemble P. « Il y a propension de a vers b si on rencontre en moyenne peu d'individus pour lesquels (a_i) est fort et (b_i) l'est moins . La moyenne sur P des $a_i(1\text{-}b_i)$ est pris comme indice de non-propension de a vers b. Cette valeur moyenne est comparée à celle qui résulterait d'une hypothèse d'absence de lien entre a et b ». J.B. Lagrange propose alors la statistique $\tilde{Q}(a,\neg b)$ approximativement distribuée comme la variable de Laplace Gauss centrée réduite, dont la réalisation donne le *coefficient de propension de a vers b* le nombre à partir de l'hypothèse d'indépendance de A et B : $\tilde{q}(a,\neg b) = \dfrac{\dfrac{\sum_{i\in P} a_i(1-b_i)}{n} - m_a(1-m_b)}{\sqrt{\dfrac{(\nu_a + m_a^2)(\nu_b + (1-m_b)^2)}{n}}}$ à partir duquel il calcule l'*intensité de la propension*. I.C. Lerman (1992) obtenait déjà et

indépendamment une formule comparable par extension du cas booléen au cas numérique de différentes formes de l'hypothèse d'absence de liaison, dans le contexte de la similarité.

3.3 Application du modèle de la propension aux variables de rangs

L'exploration de la structure du tableau TAB.1 de rangs par l'analyse statistique de la propension requiert un réexamen des conditions d'application du modèle de Lagrange. Afin de pouvoir comparer les approches, nous supposerons que les valeurs modales sont reparties de façon régulière (équidistante) sur [0 ; 1]. Les k événements {i∈P} qui définissent les lignes du tableau, sont évidemment supposés a priori indépendants. Là où un problème nous paraît surgir, dans le cadre du traitement modal en terme de rangs, c'est dans le choix *a priori* d'une hypothèse d'indépendance entre les deux variables A_i et B_i. Dans la procédure de rangement comme nous l'avons précisé plus haut, chaque composante du vecteur i, c'est-à-dire chaque variable aléatoire rattachée au critère i et à l'objet j, dépend des q–1 autres objets. Il n'est donc pas possible ici de considérer l'indépendance de deux variables A_i et B_i — dont la réalisation donne respectivement le rang de l'objet A et celui de l'objet B — comme acceptable. Elle est en effet incompatible avec la procédure même qui conduit l'individu à se confronter à la totalité des q objets pour produire finalement une permutation de {1,…,q}. Ceci nous montre qu'il n'est pas possible de transporter le modèle modal de Lagrange dans notre modèle de rangements complets sans *ex æquo*. Comment peut-on exprimer la structure issue du rangement complet et sans *ex æquo* des q objets par k sujets en terme de relation de propension ?

Réinterprétons ce que propose Lagrange, c'est à dire la quantité $s = \frac{\sum_{i \in P} a_i (1 - b_i)}{n}$, dans le contexte des variables de rangs. Dire que a_i est fort, traduit le fait que l'objet A considéré, est placé à un haut degré de préférence. Dire que b_i est faible, traduit le fait que l'objet B considéré, est placé à un bas degré de préférence. La relation de propension de a vers b devrait traduire le fait que « la tendance à placer l'objet A à un rang r_A entraîne celle à placer l'objet B à un rang r_B correspondant à une préférence de B sur A. » Dans notre modélisation du choix préférentiel, B est préféré à A si rang(B)<rang(A) et donc si $0 \le \frac{q - rang(A)}{q-1} < \frac{q - rang(B)}{q-1} \le 1$ Dans ces conditions, en conservant la statistique $\tilde{Q}(a, \neg b) = \frac{Z - E[A_i(1 - B_i)]}{\sqrt{\frac{E([A_i(1 - B_i)]^2)}{n}}}$, il nous faut prendre en compte le fait que les variables A_i et B_i sont dépendantes pour le calcul de l'espérance et de la variance de Z. Nous conservons l'approximation gaussienne de la variable aléatoire indice Z dans la mesure où nous restons dans les conditions posées par Lagrange : d'une part, pour chaque i, l'évènement {i∈P} et les évènements définis par la variable Ai(1-Bi) sont indépendants, d'autre part, les variables Ti sont indépendantes et identiquement distribuées.

3.4 Expression du coefficient de propension pour des variables rangs

En renonçant à l'indépendance, nous avons recalculé l'espérance et la variable de Z, ce qui nous a conduits au résultat suivant par lequel le coefficient de propension s'exprime ainsi

$$\tilde{q}_{(q)}(a,\neg b) = \frac{\dfrac{\sum_{i\in P} a_i(1-b_i)}{n} - \dfrac{1}{2} + \dfrac{(q-2)(3q-1)}{12(q-1)^2}}{\sqrt{\dfrac{20q^4-46q^3+49q^2-41q+24}{180n(q-1)^4}}} = \frac{\dfrac{\sum_{i\in P} a_i(1-b_i)}{n} - \left[\dfrac{1}{4} + \dfrac{(q+1)}{12(q-1)^2}\right]}{\sqrt{\dfrac{20q^4-46q^3+49q^2-41q+24}{180n(q-1)^4}}}$$

3.5 Compatibilité de la relation de propension permutationnelle avec l'ordre des moyennes

Considérons deux variables a et b issues de la transformation des rangs attribués aux objets O_1 à O_q. Notons m_a et m_b leurs moyennes respectives et Φ la fonction de répartition de la variable de Laplace Gauss centrée réduite. Nous démontrons aisément que :

Proposition 1 : $\Phi(-\tilde{q}_{(q)}(a,\neg b)) \geq \Phi(-\tilde{q}_{(q)}(b,\neg a))$ si et seulement si $m_a \leq m_b$

Proposition 2 : $\Phi(-\tilde{q}_{(q)}(a,\neg b)) \geq \Phi(-\tilde{q}_{(q)}(b,\neg a))$ si et seulement si Rang(Objet B) ≤ Rang(Objet A) dans le rangement synthétique.

Ceci signifie que le rangement de l'objet A à un degré de préférence donné « entraîne » le rangement de l'objet B à un degré de préférence plus élevé. Ces deux propositions permettent l'accord sémantique de l'indice $\tilde{q}$ avec l'intensité implicative où la variable a serait moins bien classée que la variable b. Il advient alors que l'ordre des objets issu de l'ordre des intensités de **propension permutationnelle** est compatible avec celui issu de l'ordre des sommes des rangs attribués par les juges aux q objets dans le modèle du W de Kendall.

3.6 Apport de l'exploration des rangements des q objets à l'aide de la relation de propension

Les propriétés précédentes laissent penser que finalement la relation de propension ne fait qu'apporter une information redondante à celle fournie par l'application de la procédure W de Kendall. Toutefois elle nous paraît fournir une information plus analytique sur la structure de quasi-ordre affectée d'un niveau de confiance 1-α, qui organise l'ensemble des q(q-1) paires d'objets, que ne le permet l'approche par la procédure du W de Kendall qui fournit un rangement global synthétique. À partir des définitions posées par J.B. Lagrange (Lagrange 1998 p.82) concernant les notions de la relation de propension, de chemins de propension, de filiation et de variables jumelles, nous cherchons à étudier la structure qui organise les q objets en fonction des rangements. En particulier, nous interprétons la propriété de

covariance positive comme signifiant la tendance pour le rangement des deux objets A et B de recevoir des valeurs attribuées par les juges qui vont dans le même sens.

Dans la procédure de Kendall, le q-uplet des moyennes $(m_j)_{j=1,\ldots,q}$ est l'information à partir de laquelle nous déterminons le rangement synthétique des q objets par les juges. Il est clair que les valeurs de ces q moyennes dont la somme est constante, contiennent le niveau de préférence auquel chaque juge place chaque objet en relation avec les q-1 autres. Or dans la règle de rangement adoptée dans la procédure de Kendall, ce niveau n'est pas pris en compte. C'est alors que dans l'approche par la relation de propension (permutationnelle), nous pouvons voir qu'il est rendu compte de ce phénomène. En effet si nous évaluons

$$\Delta P_{(q)}^{(a,b)} = \frac{\dfrac{\sum_{i\in P} a_i(1-b_i)}{n} - \dfrac{\sum_{i\in P} b_i(1-a_i)}{n}}{\sqrt{\dfrac{20q^4-46q^3+49q^2-41q+24}{180n(q-1)^4}}} = \frac{\dfrac{\sum_{i\in P} a_i}{n} - \dfrac{\sum_{i\in P} b_i}{n}}{\sqrt{\dfrac{20q^4-46q^3+49q^2-41q+24}{180n(q-1)^4}}}$$

Cet écart entre les deux coefficients de propension est proportionnel à l'écart entre les deux moyennes. Plus l'écart entre les moyennes est fort, plus les coefficients de propension de a vers b et de b vers a sont différenciés. De là alors que $m_a \leq m_b$ permet de ranger l'objet B avant l'objet A, la quantité $\Delta P_{(q)}^{(a,b)} \leq 0$ complète la conclusion en fournissant un autre critère d'aide à l'interprétation basé sur une mesure, à savoir celle de l'intensité de la propension, et le niveau de confiance 1-α dont elle rend compte.

3.7 Étude des situations extrêmes

3.7.1 Concordance « parfaite »

La structure à explorer est celle d'un tableau dont les composantes de chaque colonne sont constantes. Cela signifie que tous les juges émettent le même jugement respectivement à chaque objet. Il en résulte que le vecteur $(m_j)_{j=1,\ldots,q}$ des moyennes des q colonnes est égal, à une permutation près, au vecteur $(0, \frac{1}{q-1}; \frac{2}{q-1}; \ldots; \frac{q-2}{q-1}; 1)$ de $[0;1]^q$, tandis que le vecteur $(\nu_j)_{j=1,\ldots,q}$ des variances des q colonnes est égal au q-vecteur nul. Le calcul des valeurs possibles de $s = s(A \Rightarrow B) = \dfrac{\sum_{i\in P} a_i(1-b_i)}{n}$ met en évidence que l'indice de propension dans le cadre de l'analyse implicative devient :

$$\tilde{q}(a,\neg b) = \frac{s(A \Rightarrow B) - m_a(1-m_b)}{\dfrac{m_a(1-m_b)}{\sqrt{n}}} = 0 \text{ si } m_a(1-m_b) \neq 0.$$

Dans la sémantique de cette méthode d'analyse, cette situation où l'on ne peut se prononcer ni sur la propension de a vers b, ni sur sa réciproque correspond à une structure triviale de l'ensemble des rangs et, en fait, à un vide informationnel : si tous les juges sont d'accord, les préférences vont de soi, ce sont des données *a priori*.

Si $m_a(1\text{-}m_b) = 0$, le rapport a une forme indéterminée $\frac{0}{0}$. Cette situation se produit quand l'objet A est au niveau de préférence 0 — le plus bas — pour tous les juges ou quand l'objet B est situé au niveau de préférence 1 — le plus haut — pour tous les juges. Replacé dans le contexte des rangements des q objets, cela signifie que l'objet B est l'objet préféré — donc rang(B) = 1 — alors que l'objet A est celui qui est placé en dernier rang de préférence — donc rang(A) = q —. Il ressortait de l'approche par la statistique W de Kendall que l'ensemble des objets possédait une structure d'ordre triviale significative. En analyse implicative, l'indétermination est levée par un passage à la limite, dans le cadre de la topologie de Frechet. La valeur de l'indice définie par continuité est nulle et correspond alors à une intensité d'implication statistique égale à 0,50. C'est aussi ce que vaut l'intensité d'implication dans le cas d'indépendance de a et b. En ce qui concerne le facteur de variance, il est indéterminé quand $m_a(1\text{-}m_a)=0$ et infini dans les autres cas.
Si nous revenons au coefficient de propension intégrant l'approche permutationnelle, il devient :

$$\tilde{q}_{(q)}(a,\neg b) = \frac{s(A \Rightarrow B) - \frac{1}{2} + \frac{(q-2)(3q-1)}{12(q-1)^2}}{\sqrt{\frac{20q^4 - 46q^3 + 49q^2 - 41q + 24}{180n(q-1)^4}}} = \frac{m_a(1-m_b) - \frac{1}{2} + \frac{(q-2)(3q-1)}{12(q-1)^2}}{\sqrt{\frac{20q^4 - 46q^3 + 49q^2 - 41q + 24}{180n(q-1)^4}}}$$

Or, la propension de a vers b n'est sémantiquement envisageable que si $0 \leq m_a \leq m_b \leq 1$. De là $0 \leq m_a(1\text{-}m_b) \leq m_b(1\text{-}m_b)$. Comme $m_b(1\text{-}m_b)$ est toujours au plus égal à $\frac{1}{4}$, alors le numérateur de $\tilde{q}_{(q)}(a,\neg b)$ est toujours négatif. En effet, $-\frac{1}{2} + \frac{(q-2)(3q-1)}{12(q-1)^2}$ est supérieur à $\frac{1}{4}$. Ce résultat montre que, sur la base de la propension permutationnelle, il est possible de calculer la qualité numérique de cette propension qui reste significative d'une tendance de b à dominer généralement a. Il n'y a donc pas contradiction entre les deux approches, mais l'approche permutationnelle est plus précise.

3.7.2 Hétérogénéité « parfaite »

La structure à explorer est celle d'un tableau dont les composantes de chaque colonne sont telles que leurs sommes sont constantes ou presque à une unité près. Dans cette situation, le vecteur $(m_j)_{j=1,\ldots,q}$ des moyennes des q colonnes est égal au q-vecteur constant $(\frac{1}{2} ; \ldots ; \frac{1}{2})$ ou encore à une permutation près, au q-vecteur $(\frac{1}{2}+\frac{1}{2n(q-1)} ; \frac{1}{2}-\frac{1}{2n(q-1)} ; \ldots ; \frac{1}{2}+\frac{1}{2n(q-1)} ; \frac{1}{2}-\frac{1}{2n(q-1)})$ de $[0 ;1]^q$

De là nous en déduisons que nécessairement

$$m_a(1\text{-}m_b) \in \{ \frac{1}{4} ; \frac{1}{4}-\left(\frac{1}{2n(q-1)}\right)^2 ; \left(\frac{1}{2}-\frac{1}{2n(q-1)}\right)^2 ; \left(\frac{1}{2}+\frac{1}{2n(q-1)}\right)^2 \}$$

Si nous considérons que le vecteur $(m_j)_{j=1,\dots,q}$ des moyennes des q colonnes est approximativement égal au q-vecteur constant ($\frac{1}{2}$; … ; $\frac{1}{2}$), le coefficient de propension de J.B. Lagrange s'écrit maintenant :

$$\tilde{q}(a,\neg b) = \frac{\frac{\sum_{i\in P} a_i(1-b_i)}{n} - \frac{1}{4}}{\sqrt{\frac{(\nu_a + \frac{1}{4})(\nu_b + \frac{1}{4})}{n}}}$$

3.7.3 Information apportée par la covariance : cov(a ; b)

Dans l'approche de J.B. Lagrange, le coefficient de propension peut s'exprimer en fonction de la covariance par la formule : $\tilde{q}(a,\neg b) = \frac{-\mathrm{cov}(a;b)}{\sqrt{\frac{(\nu_a + m_a^2)(\nu_b + (1-m_b)^2)}{n}}}$ d'où

Proposition 3 : $\Phi(-\tilde{q}(a,\neg b)) \geq 0{,}5 \Leftrightarrow \mathrm{cov}(a;b) \geq 0$.

La covariance de a et b contient une part importante de l'information relative à la relation de propension de a vers b. Dans notre approche permutationnelle, la relation n'est plus aussi directe. En effet nous avons :

$$\tilde{q}_{(q)}(a,\neg b) = \frac{m_a(1-m_b) - \mathrm{cov}(a;b) - \left[\frac{1}{4} + \frac{(q+1)}{12(q-1)^2}\right]}{\sqrt{\frac{20q^4 - 46q^3 + 49q^2 - 41q + 24}{180n(q-1)^4}}}$$

Si nous tenons compte du fait que les valeurs prises par a et b sont dans [0 ; 1], nous pouvons déduire quelques informations relatives aux encadrements de diverses expressions :

Proposition 4 : $0 \leq m_a(1-m_b) \leq 1$ et $-1 \leq \mathrm{cov}(a;b) \leq 1$

Par ailleurs sachant que $\frac{1}{4} < \frac{1}{4} + \frac{(q+1)}{12(q-1)^2} \leq \frac{1}{2}$ quand le nombre d'objets à ranger est supérieur à 2, alors le coefficient de propension peut être ainsi encadré :

Proposition 5 : si q>2,

$$\frac{-\frac{3}{2}}{\sqrt{\frac{20q^4 - 46q^3 + 49q^2 - 41q + 24}{180n(q-1)^4}}} \leq \tilde{q}_{(q)}(a,\neg b) < \frac{\frac{7}{4}}{\sqrt{\frac{20q^4 - 46q^3 + 49q^2 - 41q + 24}{180n(q-1)^4}}}$$

Il advient aussi que la condition $\mathrm{cov}(a;b) \geq 0$ n'est plus une condition nécessaire et suffisante pour que le coefficient de propension soit négatif. En effet elle n'est plus

nécessaire car, dans l'exemple que nous aborderons ci-dessous, nous rencontrons des cas (Annexe) où cov(a ;b) < 0 et. $\tilde{q}_{(q)}(a,\neg b)<0$ — ainsi dans notre étude cov(V45E ; V45C)<0 et $\tilde{q}_{(q)}(a,\neg b)=-3{,}3491<0$ —. Elle n'est pas non plus suffisante puisque nous avons aussi cov(a ;b) > 0 et $\tilde{q}_{(q)}(a,\neg b)>0$ —ainsi dans notre étude cov(V45A ;V45J)>0 et $\tilde{q}_{(q)}(a,\neg b)=1.4981>0$ — Nous considérons alors la condition cov(a ;b) ≥ 0 comme une condition qui réduit la relation de propension aux couples d'objets tels que les variables a et b soient corrélées positivement, prenant en compte ainsi le fait que les valeurs attribuées aux deux objets vont dans le même sens. Par là même, elle se rapporte à la même classe de variables modales mise en œuvre dans la situation que celle déterminée dans l'approche de J.B. Lagrange.

4 Analyse statistique implicative comme outil d'exploration de la structure des rangements (cas du rangement incomplet et sans *ex æquo*)

Nous allons maintenant considérer le cas où il est demandé aux individus de procéder à un choix de t objets parmi les q proposés, puis de les trier par ordre décroissant de préférence. Nous pouvons considérer alors que le résultat est une réalisation du type du tableau TAB. 1, mais avec seulement t (t<q) cases sur q comportant une des permutations (1, 2, …,t). Nous proposons alors la modélisation suivante qui nous ramène à une forme traitable selon l'approche de W de Kendall. Les individus choisissent t objets et les rangent de 1 à t. Après quoi, nous postulons que cela reviendrait à dire que les (q-t) objets non choisis sont placés en *ex æquo* aux rangs t+1 à t-(q-t) = q. Nous leur attribuons alors le rang moyen $\frac{1}{q-t}\sum_{j=1}^{j=q-t}(t+j)=\frac{t+1+q}{2}$. Nous appliquons alors la procédure du W de Kendall en prenant en compte l'existence de ces *ex æquo*. Cette prise en compte des *ex æquo* modifie la valeur S_{Kmax} obtenue dans le cas de la concordance parfaite entre les k juges qui fournissent tous le même rangement, c'est à dire à une permutation près (1, 2, 3, …, t, $\frac{t+1+q}{2}$, …, $\frac{t+1+q}{2}$). Il vient donc : $S_{K\max}=\frac{k^2}{12}\left[q^3-(q-t)^3-t\right]$ De là nous obtenons la statistique $W^*_{(q,t)}=\frac{\sum_{j=1}^{q}\left(S(O_j)-k\frac{(q+1)}{2}\right)^2}{\frac{k^2}{12}\left[q^3-(q-t)^3-t\right]}$ que nous utiliserons de la même façon que la statistique W de Kendall. Notons que si t=q-1 nous retrouvons le cas du rangement complet.

4.1 Variable modale associée au rangement de t objets parmi q réalisé par k individus

Nous adoptons la transformation qui définit la variable modale par la formule suivante :

$$X_{i,j}^{(q,t)} = \frac{(t+1) - R_{i,j}^{q}}{t} \mathrm{I}_{\{1,2,..,t\}}(R_{i,j}^{q})$$

4.2 Expression du coefficient de propension pour des variables de rangs (cas du rangement incomplet sans *ex aequo*)

Dans ce contexte particulier, le coefficient de propension s'exprime alors de la façon suivante où t est le nombre d'objets à ranger parmi les q objets soumis au jugement de n individus :

$$\tilde{q}_{(q,t)}(a,\neg b) = \frac{m_a(1-m_b) - \mathrm{cov}(a;b) - \dfrac{(t+1)(2t+1)}{6tq}}{\sqrt{\dfrac{(t+1)[(t-1)(-40t^3 - 36t^2 - 5t - 6) + 30t^2(2t+1)(q-1)]}{180nq(q-1)t^3}}}$$

Nous constatons que pour t=q-1, $\tilde{q}_{(q,q-1)}(a,\neg b)=\tilde{q}_{(q)}(a,\neg b)$

5 Application à l'exploration des représentations des étudiants à l'égard des objectifs d'enseignement de la statistique

Dans une recherche que J.C Régnier (Régnier, 2002) a conduite sur la thématique de l'enseignement-apprentissage de la statistique dans un enseignement à distance, les représentations des étudiants relativement à la statistique ont été explorées, en particulier, à partir des rangements des objectifs déclarés de son enseignement. Durant l'année 2001-2002, un échantillon de 125 étudiants concernés a été soumis à un questionnaire dont une question consistait à ranger les 10 objectifs de l'enseignement selon un ordre décroissant d'importance. De cet ensemble, nous avons obtenu un sous-échantillon d'observation de 73 sujets dont nous traitons les réponses fournies relativement au rangement des 10 objectifs suivants :

Énoncés des objectifs	*Rang*
V45A) d'expliciter les questions d'une problématique dont les réponses relèvent d'une approche statistique,	
V45B) de décrire, traiter, analyser des données de manière pertinente dans le cadre d'une étude en particulier dans le domaine éducatif,	
V45C) de faire le lien entre la réflexion analytique sur des questions relevant du champ de l'éducation, leur formalisation et leur traitement quantitatif,	
V45D) de lire avec un regard critique et distancié, les conclusions de diverses études statistiques apparaissant dans des rapports de recherche en Sciences de l'Éducation,	
V45E) de poursuivre de façon autonome et personnalisée un apprentissage en statistique afin d'enrichir ses acquis,	
V45F) de poser un regard plus positif à l'égard d'un domaine largement exploité dans les *media* , dans le sens de ne pas considérer les résultats dans l'ordre du tout ou rien mais en les replaçant judicieusement dans leur domaine de validité,	
V45G) d'exploiter des notions et des démarches mathématiques à des fins d'outils, et de ce fait de modifier dans un sens positif le rapport souvent négatif que nombre entretient avec cette science,	
V45H) de s'exercer à un raisonnement intégrant l'idée de "risque d'erreur" dans l'énoncé de ses conclusions.	
V45I) de s'exercer à l'interprétation de phénomènes éducatifs sur la base de données statistiques sur des "faits éducatifs" et sur des relations entre ces "faits"	
V45J) de s'exercer à la communication des résultats des analyses des données en distinguant clairement le modèle utilisé, de la réalité qu'il est supposé représenter, en séparant bien les traitements menés à l'intérieur du modèle, des interprétations reformulées dans le contexte du problème.	

TAB. 6 - *Tableau des q=10 items à ranger par ordre d'importance décroissante.*

Le calcul des sommes $S(O_j)_{\ j=1,\ldots,10}$ donne les résultats suivant :

Objectifs	O1 V45A	O2 V45B	O3 V45C	O4 V45D	O5 V45E	O6 V45F	O7 V45G	O8 V45H	O9 V45I	O10 V45J
Sommes	322	276	338	349	552	408	459	500	392	419
Rangs issus de l'ordre des sommes	2	1	3	4	10	6	8	9	5	7

TAB. 7 - *Tableau des q=10 items rangés par ordre d'importance décroissante selon l'approche W de Kendall.*

La mise en œuvre de la procédure du test W de Kendall fournit les informations suivantes : la réalisation empirique de la statistique W vaut $w \approx 0{,}14775$ tandis que la valeur critique au seuil de $\alpha = 0.01$ vaut $w_c \approx 0{,}03298$.

Elles conduisent à rejeter l'hypothèse Ho d'indépendance ou d'hétérogénéité des rangements fournis par les étudiants et à admettre à un niveau α=0.01 de risque de 1ère espèce qu'il existe une tendance à s'accorder sur l'ordre d'importance des objectifs :

1	2	3	4	5	6	7	8	9	10
(O2	O1	O3	O4	O9	O6	O10	O7	O8	O5)
(V45B	V45A	V45C	V45D	V45I	V45F	V45J	V45G	V45H	V45E)

5.1 Exploration des 73 rangements des 10 objectifs par l'analyse de la propension statistique

Nous avons transformé le tableau des rangs par la conversion $X_{ij}^{(q)} = \frac{q - R_{ij}^{(q)}}{q-1}$ afin d'obtenir un tableau dont les valeurs sont les réalisations de variables modales dans [0 ; 1]. Nous avons ensuite soumis ce tableau de données à la procédure de l'analyse de la propension avec le logiciel CHIC 2.3 (Couturier 2001).

Items	O1	O2	O3	O4	O5	O6	O7	O8	O9	O10
Objectifs	V45A	V45B	V45C	V45D	V45E	V45F	V45G	V45H	V45I	V45J
Occurrence	45,333	50,444	43,556	42,333	19,778	35,778	30,111	25,556	37,556	34,556
Moyenne	0,621	0,691	0,597	0,580	0,271	0,490	0,412	0,350	0,514	0,473
Ecart Type	0,264	0,256	0,254	0,308	0,298	0,333	0,342	0,271	0,269	0,334
Variance	0,070	0,065	0,064	0,095	0,089	0,111	0,117	0,073	0,073	0,111
Facteur de variance	3,3642	3,2665	3,7348	2,5700	2,2291	2,2537	2,0739	3,1046	3,4428	2,2400

TAB. 8 - *Tableau des valeurs caractéristiques des variables modales Objectifs.*

Nous obtenons le graphe « implicatif »suivant :

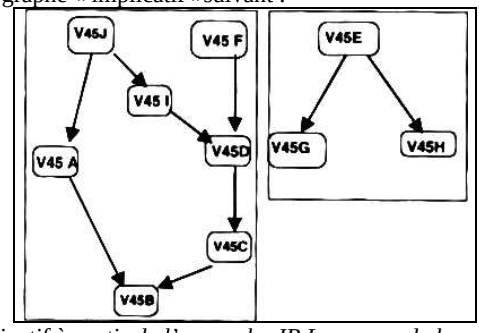


FIG. 1 – *Graphe implicatif à partir de l'approche JB Lagrange de la relation de propension.*

Les valeurs prises en compte pour construire ce graphe correspondent aux 9 cas où l'intensité de propension est supérieure à 0,50 — en d'autres termes ceux où le coefficient de propension est positif — et avec lesquelles nous procédons à l'identification des couples

(a ; b) vérifiant la **relation de propension** $a\Re b \Leftrightarrow \begin{cases} \Phi(-\tilde{q}(a,\neg b)) \geq 1-\alpha = 0.5 \\ \Phi(-\tilde{q}(a,\neg b)) \geq \Phi(-\tilde{q}(b,\neg a) \end{cases}$ et possédant la propriété de la covariance positive.

Le relation de propension est compatible avec le rangement issu de la procédure W de Kendall. Mais elle s'avère plus informative car, d'une part, l'ordre est partiel et, d'autre part, les relations de préférences sont valuées

Objectifs	O2 V45B	O1 V45A	O3 V45C	O4 V45D	O9 V45I	O6 V45F	O10 V45J	O7 V45G	O8 V45H	O5 V45E
Rangs	1	2	3	4	5	6	7	8	9	10

Voici les chemins de propension :

Chemin JAB : V45J [rang =7] ⇒V45A [rang =2] ⇒V45B [rang =1]

Chemin JIDCB : V45J [rang =7] ⇒V45I [rang =5] ⇒V45D [rang =4] ⇒V45C [rang =3] ⇒V45B [rang =1]

Chemin FDCB V45F [rang =6] ⇒ V45D [rang =4] ⇒V45C [rang =3] ⇒V45B [rang =1]

Chemin EG V45E [rang =10] ⇒ V45G [rang =8]

Chemin EH V45E [rang =10] ⇒ V45H [rang =9]

Nous constatons le respect de la propriété de la propension à ce que « l'objet O_{j1} étant rangé au rang r_1 » , « l'objet O_{j2} est rangé au rang r_2 qui traduit la préférence de cet objet O_{j2} sur l'objet O_{j1} ». Nous n'abordons pas ici l'approche sémantique que requiert l'interprétation de ces résultats.

5.2 Approche permutationnelle de la propension de X vers Y

Dans l'approche que nous avons nommée « permutationnelle » en raison du fait que la base de l'information est une permutation sur {1,…,q}, nous suivons le même raisonnement. La seule différence réside dans le fait que le calcul de l'espérance et de la variance de la statistique Z utilisée par J.B. Lagrange tient compte de la particularité de l'information fondée sur la statistique de rangs. Nous obtenons ainsi les résultats relatifs aux 90 couples de variables. 43 couples sont en relation de propension au niveau de niveau de confiance 1-α = 0,50. Pour calculer l'intensité de propension, nous utilisons encore la référence à une mesure gaussienne qui demeure monotone par rapport aux coefficients de propension permutationnelle. Nous ne retenons cependant que les couples en relation au niveau 1-$\alpha \geq$0,90. La réalisation du graphe implicatif avec le critère de la covariance positive conduit à la représentation des relations de propension suivante :

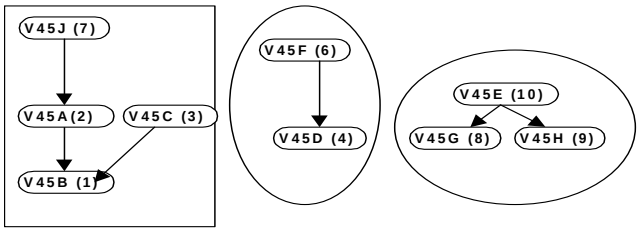


FIG. 2 –*: graphe implicatif dans l'approche permutationnelle de la relation de propension*

Si nous explicitons les chemins de propension nous obtenons alors :
Chemin JAB V45J [rang 7] ⇒V45A [rang 2] ⇒V45B [rang 1]
Chemin CB V45C [rang 3] ⇒V45B [rang 1] ;
Chemin FD V45F[rang 6] ⇒ V45D[rang 4]
Chemin EG V45E [rang 10] ⇒V45G [rang 8] ;
Chemin EH V45E [rang 10] ⇒V45H [rang 9]

Comme nous l'avons démontré, cette structure définie par la relation de propension — *versus* approche permutationnelle — respecte la propriété établissant que « l'objet O_{j1} étant rangé au rang r_1 » , « l'objet O_{j2} est rangé au rang r_2 qui traduit la préférence de cet objet O_{j2} sur l'objet O_{j1} »

6 Comparaison des résultats issus des deux approches

Nous avons calculé les écarts des intensités de propension entre les deux modèles (Annexe 4). Ce calcul met en évidence l'incidence sur la valeur de l'intensité de propension quand on ne tient pas compte des contraintes du modèle permutationnel. En moyenne, l'intensité de propension permutationnelle dépasse celle au sens de Lagrange, d'une valeur d'environ 0,13. Toutefois cette intensité de propension permutationnelle reste inférieure à celle au sens de Lagrange, pour 39 couples sur 90, soit dans plus de 43% des cas. Notons enfin que les propriétés topologiques du graphe implicatif, définies par les plus fortes intensités, sont peu affectées par ces différences, les principales liaisons étant donc respectées.

7 Conclusion

Nous avons rappelé que l'objectif du test de rang de Kendall consistait à établir un degré de concordance-discordance entre les jugements de k juges portant sur q objets, mais que sa référence à une loi de probabilité était délicate et que le résultat se limitait à rejeter l'hypothèse nulle. De même, dans le cas de rangement complet (tous les objets sont rangés) ou incomplet (certains peuvent ne pas l'être), la méthode, dite de propension, appliquée aux variables modales par J.B. Lagrange dans le cadre de la méthode implicative, ne pouvait être compatible avec la notion de rangement qui oblige à relaxer l'indépendance entre les variables. Aussi, privés de cette propriété, nous avons établi les valeurs des paramètres des

nouvelles variables en jeu dans une approche de la même propension exprimée de façon raccourcie par « l'objet b est généralement préféré à l'objet a ». Nous avons vu que la seule covariance, par son signe, n'était plus ni nécessaire ni suffisante pour conjecturer une relation de propension entre objets. Toutefois, une même référence à une mesure gaussienne pour déterminer l'intensité de propension dans les deux approches (permutationnelle et celle de Lagrange) nous a permis alors de comparer, sur une application, les structures d'ordre partiel sur l'ensemble des objets pour les deux relations de propension. On a observé que les différences étaient faibles, particulièrement lorsque les intensités de propension étaient fortes, c'est-à-dire lorsque les relations de préférences étaient nettes.

Une nouvelle question se pose alors : les différences entre les deux modèles sont-elles atténuées par l'admission des *ex æquo* où, cette fois, l'hypothèse d'indépendance entre les variables pourrait peut-être y trouver une nouvelle légitimité ?

Références

Bailleul, M., (1994) *Analyse statistique implicative : variables modales et contribution des sujets. Application à la modélisation de l'enseignant dans le système didactique*. Thèse Université de Rennes 1

Couturier, R.,(2001) Traitement de l'analyse statistique implicative dans CHIC , Actes des Journées « *Fouille dans les données par la méthode d'analyse implicative*», IUFM de Caen, 33-50.

Friedman M. (1937) The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Amer. Statist. Ass.*, 32, 675-701

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'Etat, Université de Rennes 1.

Gras, R.,. & *al*., (1996) *L'implication Statistique*. Grenoble: La Pensée Sauvage,

Gras, R., P. Kuntz et H. Briand, (2001) Les fondements de l'analyse statistique implicative et leurs prolongements pour la fouille de données, *Math. et Sc. Humaines*, n°154-155, pp. 9-29.

Kendall M.G., et B. Babington Smith (1939)The problem of *m* rankings. *Ann. Math. Statist*.,10, 275-287

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* 46 (1) : 71-93.

Lerman I.C., R. Gras et H. Rostam (1981a) Élaboration et évaluation d'un indice d'implication pour des données binaires, I et II, *Mathématiques et Sciences Humaines*, 74, 5-35 et 75, 5-47.

Lerman I.C. (1981b) *Classification et analyse ordinale des donnée.* Paris: Dunod,

Régnier J.C. (2006). Étude des difficultés d'apprentissage de la statistique dans le cadre d'un enseignement à distance. Revue *Eduquer* : *Psychologie et Sciences de l'Éducation.* 14, 15-47

Régnier J.C. et R. Gras (2005). Statistique de rangs et analyse statistique implicative. *Revue de Statistique Appliquée*. 53 (1) : 5-38

Annexe

Annexe 1 — Intensité de propension de X vers Y (JB Lagrange)

(X⇒Y)	V45A	V45B	V45C	V45D	V45E	V45F	V45G	V45H	V45I	V45J
V45A		**0,7025**								
V45B										
V45C		**0,5816**								
V45D			**0,5825**							
V45E							**0,7953**	**0,6263**		
V45F				**0,6708**						
V45G										
V45H										
V45I				**0,5587**						
V45J	**0,5190**								**0,5837**	

Annexe 2 — Corrélation (X ;Y)

(X⇒Y)	V45A	V45B	V45C	V45D	V45E	V45F	V45G	V45H	V45I	V45J
V45A	1	**0,249**	-0,125	-0,237	-0,190	-0,233	-0,152	-0,090	-0,064	**0,016**
V45B		1	**0,096**	-0,205	-0,020	-0,336	-0,138	-0,179	-0,159	-0,077
V45C			1	**0,097**	-0,111	-0,114	-0,134	-0,224	-0,133	-0,183
V45D				1	-0,389	**0,155**	-0,182	-0,142	**0,063**	-0,207
V45E					1	-0,232	**0,259**	**0,132**	-0,253	-0,218
V45F						1	-0,071	-0,303	-0,005	-0,080
V45G							1	-0,010	-0,341	-0,402
V45H								1	-0,120	-0,007
V45I									1	**0,088**
V45J										1

Annexe 3 Intensité de propension de X vers Y (approche permutationnelle)
[1-α = 0.50] et grisé [1-α = 0.90]
Les valeurs marquées avec (+) correspondent aux couples de variables dont la covariance est positive

(X⇒Y)	V45A	V45B	V45C	V45D	V45E	V45F	V45G	V45H	V45I	V45J
V45A		**0,977+**								
V45B										
V45C	**0,733**	**0,973+**								
V45D	**0,697**	**0,938**	**0,793**							
V45E	**0,999**	**0,999**	**0,999**	**0,995**		**0,990**	**0,998+**	**0,987+**	**0,994**	**0,988**
V45F	**0,900**	**0,970**	**0,895**	**0,952+**					**0,703**	
V45G	**0,983**	**0,997**	**0,973**	**0,945**					**0,755**	
V45H	**0,997**	**0,999**	**0,992**	**0,991**		**0,902**	**0,898**		**0,972**	**0,962**
V45I	**0,925**	**0,983**	**0,851**	**0,880**						
V45J	**0,974+**	**0,994**	**0,899**	**0,831**		**0,601**			**0,821**	

Annexe 4 Écart entre l'intensité de propension de X vers Y au sens de l'approche permutationnelle et celle au sens de Lagrange (moyenne = 0,13451

(X⇒Y)	V45A	V45B	V45C	V45D	V45E	V45F	V45G	V45H	V45I	V45J
V45A		0,275	0,111	0,010	-0,404	-0,296	-0,395	-0,453	-0,310	-0,445
V45B	-0,015		-0,163	-0,213	-0,491	-0,289	-0,418	-0,419	-0,385	-0,445
V45C	0,328	0,392		0,088	-0,443	-0,307	-0,400	-0,386	-0,223	-0,315
V45D	0,404	0,638	0,210		-0,277	-0,299	-0,346	-0,411	-0,185	-0,264
V45E	0,754	0,532	0,645	0,922		0,779	0,203	0,361	0,776	0,758
V45F	0,639	0,819	0,510	0,281	-0,333		-0,222	-0,263	0,208	0,028
V45G	0,666	0,682	0,626	0,667	-0,322	0,425		-0,062	0,571	0,362
V45H	0,604	0,724	0,726	0,662	0,048	0,709	0,410		0,592	0,470
V45I	0,483	0,640	0,462	0,321	-0,351	-0,009	-0,202	-0,399		-0,089
V45J	0,455	0,590	0,584	0,558	-0,335	0,187	-0,087	-0,358	0,237	

Summary

This communication explores the benefit of one data analysis method implicative analysis such as Gras defined it to study ranks concordance-discordance accorded by judges to things. This meaning is the same as in Friedman's or Kendall's works. Then we compare one analysis of judges' preferences through ranks to the propensity analysis between modal variables built by Lagrange. To achieve it, we don't consider the a priori independence hypothesis between variables. Nevertheless, we give a quality measure to "theorem" : "if an A thing is ranked by judges, then generally a B thing is ranked in a better way by the same judges" and we represent by graph relations of preferences for the whole ranked objects. We limit this study to the two cases of complete ranking and incomplete ranking, without ex æquo between q things by k judges.

Chapitre 2 : Qualité d'un graphe implicatif : variance implicative

Régis Gras*, Jean-Claude Régnier**

*LINA– Ecole Polytechnique de l'Université de Nantes, UMR 6241
La Chantrerie BP 60601 44306 Nantes cedex
regisgra@club-internet.fr
** Université de Lyon - UMR 5191 ICAR
ENS-LSH 15, Parvis René Descartes BP 7000 69342 LYON cedex 07
jean-claude.regnier@univ-lyon2.fr

Résumé. Un graphe pondéré, sans cycle, constitue une des représentations d'un ensemble de règles d'association implicative extraites d'un tableau numérique croisant variables et sujets. Le problème de son homogénéité, de sa cohérence et donc de la pertinence des interprétations de l'expert se pose dès lors qu'en Analyse Statistique Implicative (A.S.I.) il est possible de faire varier le seuil de représentation des règles partielles. Nous présentons ici le concept de variance implicative à l'instar du concept classique de variance afin de qualifier l'homogénéité de la représentation. Elle s'appuie sur une métaphore de répulsion vs consistance implicatives mutuelles entre deux variables binaires à partir de leur différence symétrique.

1 Introduction

Le texte qui suit, tente de répondre à une question posée récemment par Michel Oris et Gilbert Ritschard, (2007), question qui se ramène à celle-ci : dans quelle mesure peut-on affirmer qu'un graphe implicatif présente une « bonne » qualité de structure arborescente par rapport aux données ? Pourrait-on définir une mesure de type inertiel, par exemple, comparable à celle qui permet de qualifier une partition en classes homogènes et « convenablement séparées » ? [1]

Dans la Partie 1, chapitre 4, nous avons établi un critère probabiliste permettant de qualifier des niveaux de la hiérarchie orientée de R-règles, puis celle de la hiérarchie entière. Dans l'ouvrage *L'implication statistique. Nouvelle méthode exploratoire de données* (Gras et al, 1996), nous avons défini un critère numérique, « La variance statistique de classes cohésitives », pour quantifier la qualité d'une hiérarchie de R-règles à ses différents niveaux

[1] Je cite : « …limites de l'ASI….absence de critère permettant de juger de la pertinence statistique globale du modèle retenu. Quel pourrait être un équivalent de la déviance utilisée en modélisation statistique ou de la part d'inertie reproduite en analyse factorielle ? » (M.Oris et G. Ritschard, dans « Dynamique professionnelle dans la Genève du 19ème, enseignements d'une analyse de statistique implicative", Actes de ASI 4,, octobre 2007)

de construction ascendante. Nous procéderons de façon comparable pour quantifier l'état d'homogénéité d'un graphe implicatif[2] selon le seuil qui permet de l'établir.

2 Une remarque dans le cas où les variables sont binaires

Considérons un tableau de séries statistiques portant sur un ensemble de n sujets présentés de telle façon que soient regroupés respectivement ceux sur lesquels a et b sont vérifiées simultanément, puis ne le sont pas, puis successivement la variable a l'est sans que ne b le soit, puis la variable b sans que a ne le soit.

Dans le tableau ci -dessous, on donne un exemple d'une telle réorganisation pour n=9.

Sujets	a	b	Sujets	a	b
i1	1	1	i6	1	0
i2	1	1	i7	0	1
i3	0	0	i8	0	1
i4	0	0	i9	0	1
i5	0	0			

TAB.1

Aux variables a et b, associons les vecteurs-colonnes de leurs occurrences dans l'espace $[0 ; 1]^9$. Le vecteur $\overrightarrow{ab}$ qui se déduit des données de ce couple de variables a pour carré de longueur la valeur 4. Ce nombre caractérise d'évidence la différence symétrique (le « ou » exclusif) awb de a et b (c'est-à-dire la proposition « soit a, soit b »). Mais à ce titre, il est un indicateur de la façon dont a et b s'opposent ou ne s'impliquent pas de l'un vers l'autre, comme $a \wedge \overline{b}$ (resp. $b \wedge \overline{a}$) est un indicateur de a=>b (resp. b=>a). Plus généralement, le carré scalaire du vecteur associé à un couple de variables caractérise la non-implication de l'une sur l'autre. C'est une fonction croissante de l'opposition, de la répulsion d'une variable envers l'autre. Mais, corrélativement, elle décroît avec leur consistance[3], c'est-à-dire une certaine ressemblance. Par exemple, si le carré scalaire est nul, cela signifie que les deux variables sont identiques : les n sujets ont un comportement absolument semblable vis-à-vis de celles-ci. Nous utiliserons ces remarques dans ce qui suit pour donner du sens à la notion de « variance implicative ».

Ces remarques peuvent également s'appliquer au cas où les variables ne seraient plus binaires. Certes la référence à la différence symétrique tombe. Cependant, le carré scalaire du vecteur associé à deux variables numériques continue à exprimer qualitativement et quantitativement l'opposition ou la ressemblance entre les deux variables : plus les instances diffèrent, plus grand est le carré scalaire. Il y aura donc encore correspondance croissante entre « opposition entre deux variables » et « carré scalaire des vecteurs associés ».

[2] Rappelons qu'un graphe implicatif est orienté, sans cycle, pondéré par les intensités d'implication.

[3] Nous employons le mot "consistance" pour éviter l'usage courant, connoté, de "similarité", expressément défini en Analyse de Données, même si les sens en l'occurrence sont voisins

3 Un exemple traité par CHIC[4]

Voici un tableau récapitulant les données relatives à 5 variables binaires a, b, c, d et e, observées sur un ensemble de 20 sujets de i_1 à i_{20}.

Sujets	**a**	**b**	**c**	**d**	**e**	Sujets	**a**	**b**	**c**	**d**	**e**	Sujets	**a**	**b**	**c**	**d**	**e**	Sujets	**a**	**b**	**c**	**d**	**e**
I01	0	1	1	0	0	**I06**	0	1	0	0	0	**I11**	0	1	1	0	0	**I16**	0	1	0	0	0
I02	1	0	0	1	1	**I07**	1	0	0	0	1	**I12**	1	0	0	1	1	**I17**	1	0	0	0	1
I03	0	0	0	1	1	**I08**	0	1	1	1	0	**I13**	0	0	1	1	1	**I18**	0	1	1	1	0
I04	1	0	1	1	0	**I09**	1	0	1	1	1	**I14**	1	0	1	1	0	**I19**	1	0	1	1	1
I05	1	0	0	1	1	**I10**	0	0	0	1	1	**I15**	1	0	0	1	1	**I20**	0	0	0	1	1

	a	b	c	d	e
Occurrences	10	6	9	14	12

TAB. 2 – *Données binaires de 5 variables*

A chaque variable, point affine de V, nous pouvons associer le vecteur-colonne de ses coordonnées. Par suite, un vecteur tel que le vecteur $\overrightarrow{ab}$, est une image de la façon dont les deux variables a et b se contredisent. La valeur de son carré scalaire euclidien est 16. En effet il ressort qu'en analysant les coordonnées du vecteur $\overrightarrow{ab}$ celles-ci font apparaître une plus grande opposition entre les deux variables a et b que l'analyse des coordonnées du vecteur $\overrightarrow{ae}$ ne le montre entre les deux variables a et e.

Ainsi $\overrightarrow{ab} = (1;-1;0;-1;-1;1;-1;1;-1;0;1;-1;0;-1;-1;1;-1;1;-1;0)$ conduit à $\left\|\overrightarrow{ab}\right\|^2 = 16$ tandis que $\overrightarrow{ae} = (0;0;1;1;0;0;0;0;0;1;0;0;1;1;0;0;0;0;0;1)$ conduit à $\left\|\overrightarrow{ae}\right\|^2 = 6$. Cette propriété d'opposition est symétrique, ce que n'est pas l'implication.

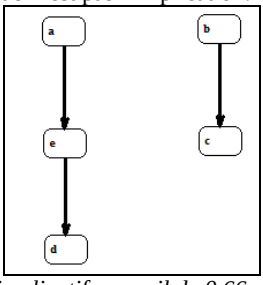


FIG. 1- *Graphe implicatif au seuil de 0,66 relatif à l'exemple*

On obtient les indices d'implication suivants :

[4] CHIC est un logiciel d'analyse de données, développé par Raphaël Couturier (Couturier, 2005) et qui sera présenté dans le chapitre 11 de la Partie 2

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

	a	b	c	d	e
a	0	5	30	60	79
b	2	0	66	3	0
c	28	64	0	52	5
d	58	11	53	0	70
e	76	3	9	73	0

TAB. 3 - *Intensités d'implication données par CHIC*

Si les variables a et e s'opposent de la même façon, par contre, a implique plus e que la réciproque.

4 Formalisation

Nous procéderons comme pour l'implication de classes de variables (Gras et al., 1996) en prenant en compte, à l'instar des méthodes de « clustering », la relation entretenue à travers deux cas entre les éléments d'une même classe (relation intra), leur barycentre et les barycentres respectifs des classes et le barycentre de l'ensemble de ces classes (relation inter). Ainsi, quand on examine des graphes implicatifs restitués par CHIC, on constate que ceux qui sont obtenus sont de deux sortes : ou bien ils se présentent d'un seul « tenant », les chemins du graphe formant un ensemble connexe ou bien, ils sont constitués de sous-ensembles connexes mais mutuellement disjoints. Par exemple, la figure 2 présente deux chemins disjoints (a→e→d) et (b→c) formant un graphe d'un seul « tenant ».

Définition 1 :
On appelle **chemin implicatif** toute suite de variables, connexe ordonnée par les occurrences croissantes des variables d'origine un nœud sans antécédent et d'extrémité un nœud sans successeur.
Définition 2 :
On appelle **grappe implicative** un ensemble connexe de chemins implicatifs connexes.

Au sein de la grappe, deux chemins peuvent partager les mêmes nœuds et les mêmes arcs. Ce partage signifie généralement à la fois une certaine ressemblance sémantique de ces chemins, mais aussi des nuances suffisantes pour que leur coexistence ait un sens dans la grappe. Par exemple, voici un graphe relatif à 6 variables constitué d'une seule grappe :

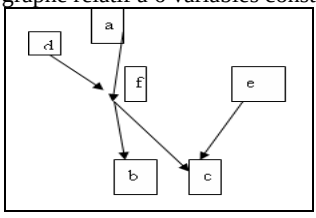


FIG. 2

(d→f→c), (d→f→b), (a→f→b), (a→f→c) et (e→c) forment, en une grappe unique, l'ensemble des 5 chemins du graphe ci-dessus. La figure 1 est, en revanche, constituée de deux grappes, réduites chacune à un chemin.

Notre objectif est de construire une expression qui, associée à un graphe présentant une ou plusieurs grappes, permette d'associer une mesure à la qualité de la structure graphique et, par là, l'homogénéité ou l'hétérogénéité de cette structure en termes de concepts implicitement sous-jacents aux relations entretenues par les variables.

Considérons maintenant un ensemble de n sujets sur lequel nous étudions des caractères représentés par des variables binaires, par l'Analyse Statistique Implicative.

4.1 1er cas : le graphe est constitué d'une grappe

A chaque arc du graphe, tel que (d→f), est associé le vecteur $\overrightarrow{df}$ de R^n des composantes des nœuds de ses extrémités d et f. Comme nous l'avons dit, la longueur de ce vecteur caractérise l'intensité de répulsion mutuelle des variables, c'est-à-dire de la non-implication de l'une sur l'autre. Plus précisément, la longueur est fonction décroissante de l'intensité implicative mesurée par le nombre de contre-exemples à l'implication.

Soit p le nombre de chemins composant une grappe.

Définition 3 :

On appelle **puissance implicative** du chemin ch(i), la quantité φ_i qui est la moyenne géométrique[5] des intensités d'implication des arcs composant le chemin ch(i), y compris les arcs obtenus par transitivité.

Cette puissance implicative a pour effet d'homogénéiser l'ensemble des barycentres des chemins du graphe puisqu'elle diminue leur variance interclasse traditionnelle.

Définition 4 :

On appelle **nodulosité** du chemin ch(i), le nombre k_i de variables ou nœuds y figurant.

On ne fait pas ici référence à la notion de longueur du chemin afin de pour pouvoir étendre la notion de chemin à celui réduit à une seule variable. Par convention, une variable isolée, non encore introduite dans le graphe implicatif, peut être considérée comme un chemin de nodulosité 1. Par exemple, le chemin (d→f→c) est de nodulosité 3.

Soit g_i le barycentre du chemin ch(i), c'est-à-dire le barycentre des k_i nœuds (ou variables) du chemin. Le barycentre G de l'ensemble des chemins est le barycentre des p chemins ch(i), pondérés par leur nodulosité k_i, constituant le graphe à un seuil choisi 1-α.

Considérons la forme bilinéaire symétrique Φ définie par la matrice carrée diagonale D d'élément générique φ_i et de dimension p. La forme quadratique associée définit à son tour une distance $\| \ \|_\Phi$ sur l'ensemble des vecteurs associés aux variables :

$$\Phi(\overrightarrow{g_iG}) = {}^t(\overrightarrow{g_iG})D(\overrightarrow{g_iG}) = \left\|\overrightarrow{g_iG}\right\|_\Phi^2 = \varphi_i^2 . \left\|\overrightarrow{g_iG}\right\|^2 \text{ où } \| \ \| \text{ est la norme euclidienne}$$

[5] Nous choisissons la moyenne géométrique de préférence à la moyenne arithmétique car la première peut alerter d'une valeur faible ou nulle et sera donc plus sensible aux variations des intensités.

Comme nous voulons, par la variance implicative, rendre compte de la façon dont les chemins de variables sont sémantiquement ou contextuellement distingués entre eux ou liés les uns aux autres, nous pondérons, à travers Φ, le carré scalaire ordinaire (euclidien) du vecteur en jeu par la puissance implicative : la différence entre un chemin et l'ensemble des autres est d'autant plus grande que la structure implicative de ce chemin est significativement et relativement importante. Une différence sémantique peut donc être mise en valeur entre deux chemins partageant en commun quelques nœuds et les arcs les reliant.

Définition 5 :

On appelle **variance intra-grappe** ou **variance implicative d'une grappe d'un graphe implicatif** admettant p chemins ch(i), la quantité $I_a = \frac{1}{\sum_{i=1}^{i=p} k_i \|\vec{u}_i\|_\Phi^2} \sum_{i=1}^{i=p} k_i \left\|\overrightarrow{g_i G}\right\|_\Phi^2$ où $\|\vec{u}_i\|_\Phi^2 = \varphi_i^2$ est interprétable comme carré de l'intensité d'implication d'un chemin dont toutes les composantes coïncideraient avec celles du barycentre commun G sauf la i^{ème} qui en diffèrerait de 1 et où $k_i \varphi_i^2$ a la signification d'une pondération implicative du chemin

En définitive, nous retrouvons la définition d'une variance statistique intra-classe ordinaire si l'on considère que les quantités $\frac{k_i \varphi_i^2}{\sum_{i=1}^{i=p} k_i \|\vec{u}_i\|_\Phi^2}$ sont des coefficients de somme 1, pondérant les chemins par leur qualité implicative. Mais, pour éviter les confusions et rappeler leur adéquation au problème, nous avons choisi la dénomination **variance intra-grappe**. Et l'expression I_a est un indicateur de dispersion des barycentres des chemins autour du barycentre commun de la grappe. Elle est minimale et nulle lorsque tous les barycentres coïncident avec le barycentre commun. Ce qui est le cas, si la grappe est constituée d'un seul chemin, c'est-à-dire lorsque toutes les variables sont alignées le long du chemin unique.

Les variables isolées, constituant un chemin de nodulosité 1 et une grappe réduite à un chemin auront manifestement une variance intra-grappe nulle. La collection des chemins est d'autant plus porteuse de sens divers et forts que ceux-ci s'opposent dans l'ensemble du graphe. Ce phénomène est accompagné d'une variance intra-grappe importante. En revanche, que deux de ces chemins, ou un réseau de chemins ne s'opposent pas globalement peut souligner l'existence d'une signification commune que le chercheur se devra d'identifier. Ce sera le cas où cette variance sera faible. C'est notamment le cas où la grappe se ramène à un chemin unique.

4.2 2ème cas : le graphe est constitué de m grappes

1. Soit G_1, G_2,…, G_m les barycentres respectifs des m grappes. Le barycentre G_j de la grappe j est affecté de la somme des pondérations des barycentres de ses chemins ;
2. Soit p_1,…, p_m les nombres respectifs de chemins implicatifs constituant chacune de ces grappes,
3. Soit Γ le barycentre commun des m grappes,

4. Soit ψ_j la moyenne géométrique des puissances implicatives des différents chemins de la grappe j. Cette moyenne ψ_j contient donc une information résumée et relative aux différentes intensités associées aux arcs de la grappe.

Rappelons qu'une variable isolée est un chemin de nodulosité 1 et donc aussi une grappe constituée d'un seul chemin. La puissance implicative d'un tel chemin se réduit à l'intensité affectée à la variable en l'occurrence ici, par définition, elle est égale à 1.

On note ψ la forme bilinéaire symétrique définie par la matrice carrée diagonale Δ d'élément générique ψ_j et de dimension m. La forme quadratique associée définit à son tour une distance $\| \ \|_\Psi$ sur l'ensemble des vecteurs associés aux vecteurs $\overrightarrow{G_j\Gamma}$.

Définition 6 :

On appelle **variance implicative inter-grappe** ou **variance expliquée d'un graphe implicatif** constitué de m grappes, la quantité $I_e = \dfrac{1}{\sum_{j=1}^{j=m} p_j \left\|\vec{v}_j\right\|_\Psi^2} \sum_{j=1}^{j=m} p_j \left\|\overrightarrow{G_j\Gamma}\right\|_\Psi^2$ où $\left\|\vec{\upsilon}_j\right\|_\Psi^2 = \psi_j^2$

est interprétable comme carré de l'intensité d'implication d'une classe dont toutes les composantes coïncideraient avec celles du barycentre commun Γ sauf la j^{ème} qui en diffèrerait de 1.

Définition 7 :

On appelle **variance implicative totale** du graphe la somme des variances intra-grappes et de la variance inter-grappe

Une variable isolée x a un apport égal à $\left\|\overrightarrow{x\Gamma}\right\|^2$, c'est-à-dire le carré de la distance euclidienne de x à Γ.

Selon une remarque déjà exprimée, I_e souligne les répulsions mutuelles entre les grappes et donc assure bien la fonction traditionnelle de la variance inter-classe dont il partage la dimension et la forme. En choisissant la dénomination **inter-grappe**, nous voulons cependant éviter les confusions. La somme de la variance intra-grappe et de la variance inter-grappe contient l'information globale implicative restituée par le graphe. Ce qui nous autorise l'appellation choisie de **variance implicative totale**. Les grappes constituées d'un grand nombre de chemins l'affectent a priori plus que les grappes à chemin unique. Si l'homogénéité autour du barycentre global est importante, cela signifie que les grappes ont des sens sans doute voisins et sont potentiellement en passe de se réunir sous la contrainte d'un seuil de confiance de niveau plus faible. Ceci justifie encore a posteriori l'appellation choisie pour la variance implicative.

Ici, comme avec certaines méthodes de classification habituelles, nous apprécions l'homogénéité de l'ensemble des classes par une recherche de minimisation de la variance implicative intra-grappe. Ce qui signifie encore, a contrario, que l'hétérogénéité sera d'autant appréciée que la variance totale sera importante, voire maximale. Ce qui peut être interprété comme signifiant l'existence de concepts différents, voire opposés, au sein de la structure exprimée en graphe. Ce qui, a posteriori encore, justifie d'autant plus la recherche d'une structure significative encore inapparente de l'ensemble des variables.

Supposons qu'à un seuil de construction d'un graphe apparaissent p grappes. Abaissons ce seuil. Apparaissent alors de nouveaux arcs d'intensités nécessairement inférieures

conformément à l'algorithme de construction. Ces arcs peuvent s'agréger aux chemins existants ou constituer de nouveaux arcs tout en supprimant éventuellement l'isolement de variables non encore reliées. La dispersion des centres diminuant, la faiblesse des intensités de ces variables isolées s'amortissant à travers les moyennes géométriques φ_i, en toute hypothèse, la variance implicative diminuera. En tant que correspondant à une variance inter-classe, elle décroît donc lorsque de nouvelles grappes se forment.

5 Retour sur l'exemple numérique

Dans l'exemple donné dans le TAB.2, et représenté par la Fig. 1, nous observons sur le graphe au seuil donné (1-a=0,66) m=2 grappes chacune constituée d'un seul chemin. Pour la grappe 1, le chemin ch(1)= (a→e→d) est de nodulosité 3 et pour la grappe 2 le chemin ch(1)= (b→c) est de nodulosité 2.

Le barycentre de l'ensemble des 5 variables vérifie la relation vectorielle :

$$\overrightarrow{O\Gamma} = \frac{1}{5}(\overrightarrow{Oa} + \overrightarrow{Ob} + \overrightarrow{Oc} + \overrightarrow{Od} + \overrightarrow{Oe})$$

et a pour coordonnées dans R^{20} $\overrightarrow{O\Gamma} = \frac{1}{5}(2;3;2;3;3;1;2;3;4;2;2;3;3;3;3;1;2;3;4;2)$

Par un raisonnement analogue, nous obtenons les barycentres respectifs des deux grappes, G_1 et G_2

$\overrightarrow{OG_1} = \frac{1}{3}(0; 3; 2; 2; 3; 0; 2; 1; 3; 2; 0; 3; 2; 2; 3; 0; 2; 1; 3; 2)$

$\overrightarrow{OG_2} = \frac{1}{2}(2; 0; 0; 1; 0; 1; 0; 2; 1; 0; 2; 0; 1; 1; 0; 1; 0; 2; 1; 0)$

La puissance implicative du chemin ch(1) de la grappe 1 est : $\varphi_1=(0,79 \text{ x } 0,73 \text{ x } 0,60)^{1/3}$ = 0,702. Comme la grappe 1 ne contient qu'un chemin, la moyenne géométrique ψ_1 des puissances des chemins de la grappe 1 coïncide avec la valeur φ_1. Le carré de la norme euclidienne du vecteur $\overrightarrow{G_1\Gamma}$ est $\left\|\overrightarrow{G_1\Gamma}\right\|^2 = 1,631$

La puissance implicative du chemin ch(1) de la grappe 2 est : $\varphi_{1=}0,66$. Et ψ_2 de la grappe 2 coïncide avec $\varphi_{1=}0,66$. Le carré de la norme euclidienne du vecteur $\overrightarrow{G_2\Gamma}$ est $\left\|\overrightarrow{G_2\Gamma}\right\|^2 = 3,67$

Les variances intra-grappes des deux grappes sont nulles puisque leurs barycentres coïncident avec les barycentres des chemins qui les composent. Ainsi, puisque $p_1=p_2=1$, la contribution à la variance inter-grappe de G_1 est

$$p_1\left\|\overrightarrow{G_1\Gamma}\right\|_\Psi^2 = p_1\Psi_1^2\left\|\overrightarrow{G_1\Gamma}\right\|^2 = (0,702)^2(1,631) = 0,8039$$

et celle de G_2 est $p_2\left\|\overrightarrow{G_2\Gamma}\right\|_\Psi^2 = p_2\Psi_2^2\left\|\overrightarrow{G_2\Gamma}\right\|^2 = (0,66)^2(3,67) = 1,598$

Ces deux nombres expriment l'intensité de répulsion respective de G_1 et de G_2 sur Γ.

$$I_e = \frac{1}{\sum_{j=1}^{j=m} p_j \left\| \vec{v}_j \right\|_\Psi^2} \sum_{j=1}^{j=m} p_j \left\| \overrightarrow{G_j \Gamma} \right\|_\Psi^2 = \frac{1}{(0{,}702)^2 + (0{,}66)^2} (0{,}8039 + 1{,}598) = 2{,}58$$

Ce nombre exprime d'une certaine façon l'opposition sémantique entre les deux grappes pour le niveau de confiance de 0,66.

La grappe 1 et la grappe 2 expliquent respectivement 33,5% et 66,5% de la variance implicative totale comme le montre le calcul ci-dessous :.

$$I_T = I_a + I_e = 0 + I_e = \frac{0{,}8039}{(0{,}702)^2 + (0{,}66)^2} + \frac{1{,}598}{(0{,}702)^2 + (0{,}66)^2} = 2{,}58$$

$$\frac{\frac{0{,}8039}{(0{,}702)^2 + (0{,}66)^2}}{2{,}58} = 0{,}3346 \quad \text{et} \quad \frac{\frac{1{,}598}{(0{,}702)^2 + (0{,}66)^2}}{2{,}58} = 0{,}6654$$

Pour poursuivre notre illustration, nous abaissons le seuil de confiance au niveau à 0,52[6], et conséquemment apparaît un nouveau graphe implication réduit à une seule grappe mais comportant deux chemins, c'est à dire que m=1 et p=2. Le chemin ch(1) de nodulosité k_1=4 et le chemin ch(2), k_2=2.

	a	b	c	d	e
a	0	0,05	0,30	**0,60**	**0,79**
b	0,02	0	**0,66**	0,03	0
c	0,28	0,64	0	0,52	0,05
d	0,58	0,11	**0,53**	0	0,70
e	0,76	0,03	0,09	**0,73**	0

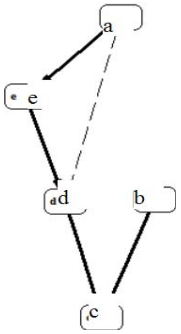


FIG. 3 - *Graphe implicatif au seuil de 0,52 relatif à l'exemple du tableau 2*

La puissance implicative du chemin ch(1)=(a, e, d, c) est de :

$\varphi_1 = (0{,}79\text{x}0{,}73\text{x}0{,}60\text{x}0{,}53)^{\frac{1}{5}} = 0{,}7123$. Celle du chemin ch(2)=(b, c) est $\varphi_2 = 0{,}66$

et les coordonnées de son barycentre g_2 sont

$$\overrightarrow{Og_2} = (1, 0, 0, 1/2, 0, 1/2, 0, 1, 1/2, 0, 1, 0, 1/2, 1/2, 0, 1/2, 0, 1, 1/2, 0)$$

Le barycentre commun reste le même que dans la situation précédente et comme il n'y a qu'une grappe, son barycentre G coïncide avec le barycentre global Γ :

$$\overrightarrow{OG} = \overrightarrow{O\Gamma} = \frac{1}{5}(2;3;2;3;3;1;2;3;4;2;2;3;3;3;3;1;2;3;4;2) .$$

[6] Le seuil est très bas en raison du faible nombre d'occurrences. Mais il s'agit d'un simple exemple d'école.

Les coordonnées du barycentre g_1 de (a, e, d, c) sont :

$$\overrightarrow{Og_1} = \frac{1}{4}(1, 3, 2, 3, 3, 0, 2, 2, 4, 2, 1, 3, 3, 3, 3, 0, 2, 2, 4, 2)$$

$$I_T = I_a + I_e = I_a + 0 = \frac{1}{k_1\varphi_1^2 + k_2\varphi_2^2}\left[k_1\varphi_1^2 \left\|\overrightarrow{g_1G}\right\|^2 + k_2\varphi_2^2 \left\|\overrightarrow{g_2G}\right\|^2 \right]$$

$$I_T = I_a = \frac{1}{4(0{,}7123)^2 + 2(0{,}66)^2}\left[4(0{,}7123)^2 \left\|\overrightarrow{g_1G}\right\|^2 + 2(0{,}66)^2 \left\|\overrightarrow{g_2G}\right\|^2 \right]$$

avec $\left\|\overrightarrow{g_1G}\right\|^2 = 0{,}4325$ et $\left\|\overrightarrow{g_2G}\right\|^2 = 3{,}67$

Par conséquent $I_T = I_a = 1{,}4048$. La contribution relative du chemin ch(1) à la variance intra-grappe est de 21,5% et celle du chemin ch(2) est de 78,5%

Comme la variance implicative inter-grappe est nulle, la variance implicative totale est donc 1,40 c'est-à-dire plus faible qu'au seuil précédent Au niveau de faible exigence relationnelle (0,52) l'hétérogénéité entre les deux chemins est atténuée grâce à la connexion établie entre les deux chemins précédemment indépendants. Mais l'information globale, restituée par la variance, s'est appauvrie car il est possible que les sens distincts de ces deux chemins se soient quelque peu dilués en un sens moins spécifique.

6 Autre exemple[7]

Sur un ensemble de 270 hommes et femmes, on dispose d'informations portant sur leur état-civil, et leur secteur d'activité. La problématique des chercheurs est la recherche de critère de décision à partir de prédicteurs à l'aide de la méthode d'arbre de décision.

Etat civil	Homme			Femme			Total
	primaire	secondaire	tertiaire	primaire	secondaire	tertiaire	
Marié	50	40	6	0	14	10	120
Célibataire	5	5	12	50	30	18	120
Divorcé/veuf	5	8	10	6	2	2	33
Total	60	53	28	56	46	30	270

TAB. 4 -

La variable à prédire est l'état civil, le sexe et le secteur d'activité étant les prédicteurs disponibles. Une analyse statistique implicative appliquée à cette question conduit aux implications suivantes :

[7] Cet exemple est extrait de l'article de G.Ritschard, D.Zighed et S.Marcellin [2007], Données déséquilibrées, entropie décentrée et indice d'implication, Nouveaux apports théoriques à l'Analyse Statistique Implicative et Applications 4èmes Rencontres Internationales d'Analyse Statistique Implicative, Université Jaume I de Castellon 18-21/10/2007, p. 315-328, ISBN 978-84-690-8241-6

règles	Intensités d'implication	
Secondaire => **M**arié	φ (**S**,**M**) = 0,94	
Homme => Marié	φ (**H, M**) = 0,99	$\varphi_{entropique}$ (**H, M**) =0,68
Tertiaire => **C**élibataire	φ (**T,C**) = 0,78	
Femme => **C**élibataire	φ (**F,C**) = 0,99	$\varphi_{entropique}$(**F,C**) =0,78

TAB. 5

Autrement dit, on peut ne pas prendre de grand risque d'affirmer que si un sujet de l'enquête est une femme (resp. un homme) elle (resp. il) est célibataire (resp. marié). Si l'échantillon interrogé respecte les règles de tirage au hasard dans une population plus large, l'induction de ces propriétés est légitime.

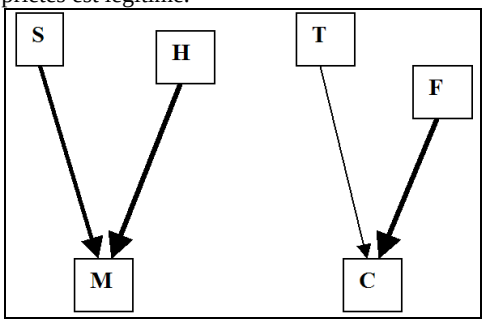


FIG. 4 - *Graphe implicatif des 4 variables au seuil 0,778*

Il s'agit ici de deux grappes composées chacune de deux chemins. Chaque chemin est de nodulosité égale à 2. Les calculs conduisent aux résultats suivants :

* la grappe 1 (**S,H,M**) a une variance intra-grappe égale à 8,954
* la grappe 2 (**T,F,C**) a une variance intra-grappe égale à 8,125
* ainsi, la somme des variances intra-grappes I_a est de 17,08
* la variance inter-grappe I_e est la somme de la contribution de la grappe 1 (**S,H,M**) soit $\frac{62}{2(0,93^2+0,77^2)} = 21,26$ et de celle de la grappe 2 (**T,F,C**) soit $\frac{51,44}{2(0,93^2+0,77^2)} = 17,64$.

D'où I_e = 38,91.

La grappe 1 **(S,H,M)** explique 54,64% de la variance inter-grappe, soit plus que la grappe (**T,F,C**) dont la structure est quelque peu affaiblie par l'intensité implicative φ (**T,C**).

On note alors que la variance *inter-grappe* est plus forte que la variance *intra-grappe* ce qui souligne la bonne cohérence interne de chacune des grappes mais leur opposition sémantique, évidemment perceptible.

Remarque

A titre indicatif, nous avons pratiqué une analyse statistique implicative à partir de certains itemsets, en l'occurrence, ceux qui associent le sexe et le secteur d'activité. On observe alors que la faible liaison entre P et C (intensité d'implication : 0,69) est très sensiblement améliorée en conjoignant P et F : (P, F) => C avec une intensité de 1. Ceci signifie qu'une Femme ayant des activités dans le secteur Primaire est presque sûrement

Célibataire. Une analyse en termes de variance implicative pourrait être conduite sur le nouveau graphe qui serait composé, cette fois, de deux grappes de trois chemins.

7 Conclusion : rôle descriptif et rôle décisionnel de la variance implicative

Nous avons vu, grâce à un exemple, que, pour un seuil 1-α du graphe implicatif donné, la variance implicative est une mesure qu'il est possible d'associer au graphe. La valeur de la fraction intra peut être comparée à la variance implicative totale et donner ainsi une information quantitative sur la structure du graphe au seuil retenu. Le rapport entre la variance implicative intra-grappe au seuil 1-α et la variance implicative totale en rend compte de façon claire mais non décisive. On peut en observer les variations en abaissant le seuil et exploiter les « vitesses » respectives de décroissance entre le seuil et la variance.

Notons aussi que la variance implicative, contrairement à l'intensité d'implication qui est une probabilité, ne fournit pas d'échelle de mesure. Par exemple, l'appréciation d'un éventuel antagonisme entre grappes ne peut s'estimer que par comparaison relative. En effet, toujours à un seuil donné, les « contributions » respectives des chemins à la variance apportent une information qui permet, par comparaison, de décider de la qualité contributive de la variance associée au chemin à la variance implicative totale. Et donc de pointer le chemin le plus structuré. L'attribution d'un sens à ce chemin peut donc être plus assurée qu'elle ne peut l'être pour les autres.

Il est possible également d'associer seuil et variance. Comme plus le seuil est abaissé, plus le nombre de chemins s'accroît de façon naturelle, il serait intéressant d'adopter un critère de significativité de la structure en formant le produit variance et seuil. Ainsi la croissance de la variance serait compensée, dans sa signification, par celle du seuil. Relativement à l'exemple donné dans le sous-chapitre 3 et traité dans le 5, nous aurions donc à comparer les variances respectives 2,58 au seuil 0.66 et 1,40 au seuil 0,52. Les deux produits respectifs 1,702 et 0,728 confirment que la deuxième structure présente une valeur inférieure et que le regroupement en deux chemins contigus est chargé de moins de sens que le graphe en deux chemins disjoints, comme si ce sens s'était affadi, avait perdu ses nuances discriminatoires.

Sur le plan décisionnel, il ne paraît pas possible, compte tenu des paramètres en jeu comme l'intensité d'implication, le nombre de chemins, le seuil de construction du graphe, de définir quelle est la structure la plus intéressante parmi celles qui peuvent être obtenues : faut-il minorer la variance implicative intra-grappe afin de disposer d'une bonne qualité de leur architecture qui renforcerait leur sens ? ou faut-il majorer la variance implicative inter-grappe afin de s'assurer d'une forte opposition des significations des grappes ? En fait, interactivement, ce sont ces deux optimisations qu'il serait bon de préserver en laissant l'expert apprécier leurs variations au fil de celle du seuil. La décision d'arrêt dépendra de la qualité de l'expression rendue par la structure graphique intégrant le doute lié à un seuil assez bas et la certitude d'une discrimination sur un sens suffisamment fort.

L'extension de cette approche quantitative pour évaluer la qualité structurelle d'un graphe au cas où les variables ne seraient pas binaires nous paraît légitimée par la référence à la différence symétrique, métaphore de la répulsion ou de la consistance.

Sur le plan formel, une autre approche semble possible et d'intérêt manifeste : Gilbert Ritschard suggère, de son côté, une approche à l'aide de l'entropie décentrée. De nature

foncièrement différente, la formalisation conduirait à des résultats qu'il serait intéressant à comparer aux précédents.

Références

Couturier R. et R. Gras (2005) : CHIC : Traitement de données avec l'analyse implicative, *Extraction et Gestion des Connaissances, Volume I1, RNTI,* Cépaduès, Paris, p.679-684, ISBN 2.85428.683.9.

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'Etat, Université de Rennes I.

Gras R., S. Ag Almouloud, M. Bailleul, A. Larher, M. Polo, H. Ratsimba-Rajohn et A. Totohasina (1996), *L'implication Statistique*, Grenoble : La Pensée Sauvage.

Oris M. et G. Ritschard (2007) , Dynamique professionnelle dans la Genève du 19ème siècle ; enseignements d'une analyse de statistique implicative, *Nouveaux apports théoriques à l'Analyse Statistique Implicative et Applications*, 4èmes Rencontres Internationales d'Analyse Statistique Implicative, Castellon, 2007, ISBN 978-84-690-8241-6, 287-300.

Ritschard G., D. Zighed et S. Marcellin (2007), Données déséquilibrées, entropie décentrée et indice d'implication, *Nouveaux apports théoriques à l'Analyse Statistique Implicative et Applications* 4èmes Rencontres Internationales d'Analyse Statistique Implicative, Université Jaume I de Castellon 2007, ISBN 978-84-690-8241-6, 315-328.

Summary

A weighted graph, without cycle, is a representation of a set of implicative association rules extracted from a digital cross variables and subjects. The problem of homogeneity, consistency and therefore the relevance of the interpretations of expert arises since in Statistical Implicative Analysis (SIA) it is possible to vary the threshold of representation of partial rules. We present here the concept of variance implicative like the classical concept of variance in order to characterize the homogeneity of the representation. It relies on a metaphor of repulsion versus mutual consistency between two binary variables from their symmetric difference.

Chapitre 3 : Mesurer l'écart entre une analyse a priori et la contingence en didactique[1]

Filippo Spagnolo *, Régis Gras **, Jean-Claude Régnier***

*G.R.I.M. (Gruppo di Ricerca sull'Insegnamento delle Matematiche),
Department of Mathematics, University of Palermo.
via Archirafi 34, 90123 Palermo (Sicily).
spagnolo@math.unipa.it

** Equipe Connaissances et Décision, Laboratoire d'Informatique de Nantes Atlantique
Ecole Polytechnique de l'Université de Nantes, UMR 6241
La Chantrerie BP 60601 44306 Nantes cedex
regisgra@club-internet.fr

***Université de Lyon – UMR 5191 ICAR
ENS-LSH 15, Parvis René Descartes BP 7000.69342 LYON cedex 07
jean-claude.regnier@univ-lyon2.fr

Résumé. En didactique des Mathématiques, mais plus généralement en sciences humaines, de nombreuses recherches utilisent des analyses qualitatives pour falsifier expérimentalement des hypothèses formulées a priori, c'est-à-dire en amont de la recherche. Une telle approche méthodologique, appliquée à une enquête, s'avère le plus souvent insuffisante pour analyser toutes les variables en jeu dans des phénomènes contingents d'enseignement/ apprentissage, même si dans certains cas (analyse ponctuelle de protocoles, de vidéos, etc.), elle permet de déceler quelques relations intéressantes. Mais si le nombre de sujets devient trop volumineux, l'analyse qualitative ne réussit plus à extraire toutes les relations existant entre les variables en jeu. Une analyse quantitative sur une base statistique s'imposera et sera complétée par une analyse qualitative, indispensable à une interprétation contextuelle. Cette communication vise à présenter une mesure permettant de confronter statistiquement, l'analyse a priori et la contingence.

1 Données

Les données d'une enquête dans un contexte d'épreuve d'évaluation d'un enseignement se composent généralement des éléments suivants :

1. un ensemble d'items ou de variables V en nombre v,

[1] Cet article a été publié dans les Actes de ASI 4 (2007), Castellon, sous une forme et un contenu voisins, avec le titre : " Une mesure comparative en didactique des mathématiques entre une analyse a priori et la contingence ", et pour auteurs Filippo Spagnolo, Régis Gras et Jean-Claude Régnier

2. un ensemble de sujets E en nombre n,

un ensemble W d'attitudes, de conceptions, de comportements généraux attendus au cours de l'épreuve dans laquelle les sujets doivent répondre aux items V.

Sur la base de ces données, les objectifs de recherche de comparaison entre ce qui est attendu et ce qui est observé sont doubles :

1. le croisement WxV qui va permettre d'extraire des relations a priori entre variables,
2. le croisement ExV qui représente le comportement observé des n sujets selon les v variables.

2 Problématique

L'analyse a priori permet de dégager des relations de similarité et des relations implicatives a priori entre les variables. Ces relations sont, bien entendu, hypothétiques et basées sur la connaissance des apprentissages connus jusqu'alors ou supposés et des conditions d'enseignement. Ici, nous nous intéressons essentiellement aux relations implicatives, à savoir des règles, selon le vocabulaire du « Data mining ». Il s'agit alors de comparer les qualités implicatives des relations du type a⇒b extraites de l'analyse a priori par exemple de conceptions d'élèves fictifs et de celles rendues par l'analyse a posteriori, que nous préférons appeler ici **analyse de la contingence**, car elle s'appuie directement sur la base des comportements observés dans une épreuve.

Précisons. Dans **l'analyse a priori**, les sujets traditionnellement en ligne sont des conceptions ou des attitudes attendues et non pas les sujets réels (élèves par exemple). Nous établissons a priori des règles dans cette « population » de « sujets » : si telle variable a est observée alors telle variable b l'est généralement, malgré d'éventuels contre-exemples en nombre acceptable pour le chercheur selon un seuil fixé. Ainsi, l'analyse a priori se fait sur la base de présupposés justifiés par la pratique, la psychologie cognitive, les observations anciennes, des résultats obtenus sur d'autres échantillons, etc. En revanche**, dans la contingence**, les sujets sont des individus, hic et nunc. Dans ce cas, on observe des quasi-règles de type a⇒b entre les mêmes variables dans la mesure où, le plus souvent, les observations sont entachées de contre-exemples.

Lors de l'épreuve a priori, épreuve fictive et pilotée par le chercheur, le nombre de contre-exemples à l'implication a⇒b peut être nul dès lors que ce chercheur a décidé que l'observation de a conduit inexorablement à celle de b. Ce qui n'est pas le cas contingent où il peut être observé un nombre faible de tels contre-exemples, sans que soit réfutée l'hypothèse que a implique généralement b. Mais, au contraire, le chercheur peut s'attendre à un nombre important de contre-exemples dans son étude a priori et, par la suite, noter un faible nombre de contre-exemples à a⇒b dans l'analyse de la contingence. C'est cette distorsion entre ce qui est prévisible et ce qui est observé qui nous intéresse. Pour ce faire, nous procèderons de façon différente de celle conduite dans une étude précédente (Spagnolo F, 1997). dans laquelle nous ne nous sommes pas référés à une échelle probabiliste.

3 Formalisation de la comparaison de l'analyse a priori et de l'analyse de la contingence

Supposons l'épreuve contingente réalisée avec n sujets. Il est possible de prendre en compte tous les couples de variables tels que (a,b). A chaque couple, le chercheur associe à la règle a⇒b une valeur a priori sur la base de ses hypothèses psycho-didactiques et des connaissances de la population de sujets en jeu. Notons $\Phi(a,b)$ cette valeur a priori établie[2], par exemple, sur la base des contre-exemples estimés dans la relation implicative de a vers b. Nous adoptons alors comme critère d'écart entre l'analyse a priori et la contingence, la différence entre les valeurs attribuées $\Phi(a,b)$ et les valeurs $\varphi(a,b)$ observées dans l'épreuve. Mais on peut aussi se contenter de ne prendre en compte qu'un nombre réduit de couples, par exemple, ceux qui figurent dans un chemin du graphe implicatif ou dans une relation implicative entre deux classes de la hiérarchie orientée des R-règles. C'est cet écart dont nous étudierons la ***significativité*** qui servira de critère pour affirmer la réfutation ou non de l'adéquation entre les relations émises a priori et celles apparues à travers la contingence. Soit m ce nombre de couples qui est au plus égal à v(v-1).

4 Construction d'une mesure comparative

Cette approche de comparaison entre les présupposés et la contingence, a été adoptée une première fois par F. Spagnolo (Spagnolo, 1997) et nous en reprenons ici les grandes lignes. Supposons connus l'ensemble de relations implicatives entretenues parmi les couples ou un sous-ensemble défini, par exemple, par un chemin du graphe implicatif ou par une classe de hiérarchie orientée. Soit $\{(a_i, b_j)\}_{i,j}$ les couples qui sont retenus pour l'analyse comparative entre les valeurs des intensités d'implication a priori $\Phi_{i,j}$ de la règle $a_i \Rightarrow b_j$ et les valeurs d'intensité réellement observées, dites contingentes, $\varphi_{i,j}$ de la même règle. Il s'agit donc de comparer les deux lignes du tableau ci-dessous.

Intensités d'implication						
a priori Φ	$\Phi_{1,1}$	$\Phi_{1,2}$	…	$\Phi_{i,j}$	…	$\Phi_{k,l}$
contingentes φ	$\varphi_{1,1}$	$\varphi_{1,2}$	…	$\varphi_{i,j}$	…	$\varphi_{k,l}$

TAB. 1

Il paraît naturel alors de calculer et sommer les écarts entre les deux séries d'intensités. En fait, nous avons choisi de définir une distance du type χ^2 entre elles sous la forme :

[Formule 1] $$\Delta = \sqrt{\sum_{i,j} \frac{1}{\Phi(a_i, b_j)} \left[\Phi(a_i, b_j) - \varphi(a_i, b_j)\right]^2} == \sqrt{\sum E_{i,j}}$$

[2] Nous ne détaillons pas ici de protocole explicitant une procédure d'attribution des valeurs d'intensité d'implication a priori

Le terme normalisateur $\frac{1}{\Phi(a_i, b_j)}$ permet de relativiser l'écart observé selon la règle $a_i \Rightarrow b_j$ à la valeur de l'intensité a priori. Sa contribution à la distance est d'autant plus importante que l'intensité a priori l'est, ce qui permet d'accentuer les grandes distorsions entre « l'avant » et « l'après », c'est-à-dire les erreurs éventuelles de jugements a priori de l'expert. Par convention, si $\Phi(a_i, b_j) = 0$, on attribuera à l'écart correspondant E_{ij}, la valeur du maximum des écarts observés dans les cas où cette valeur a priori n'est pas nulle.

F. Spagnolo, dans l'article précité (1997) fixait à 0,25 la limite supérieure de l'écart total admissible pour que la distance ne soit pas considérée comme susceptible de réfuter l'adéquation entre les deux séries. Cette décision sur le seul seuil numérique arbitraire, indépendamment du nombre de règles en jeu, nous semble nécessiter une approche statistique différente voire complémentaire qui va elle-même constituer notre deuxième approche.

5 Construction d'un test d'hypothèse d'adéquation a priori-a posteriori

Considérons alors l'épreuve aléatoire binomiale où l'on réaliserait plusieurs tirages en nombre égal à l'effectif m, de couples retenus pour l'analyse de type (a_i, b_j) et où la variable aléatoire serait le nombre de tirages conduisant à une différence fixée entre les deux valeurs d'intensités d'implication : a priori et contingente. L'écart $|\Phi(a_i, b_j) - \varphi(a_i, b_j)|$ est la valeur observée de cette différence aléatoire. Nous pouvons alors préciser que l'espace fondamental Ω est l'espace à partir duquel est définie une variable aléatoire à m composantes $\Theta(a_i, b_j)$: les intensités $\varphi(a_i, b_j)$ en sont les réalisations respectives dans l'épreuve. A partir de l'espace Ω, supposé probabilisé[3], on définit une variable aléatoire $X_{i,j}$ égale au nombre de cas où les observations telles que $\varphi(a_i, b_j)$ diffèrent au plus des intensités a priori correspondantes $\Phi(a_i, b_j)$ d'un certain nombre réel ε, $\varepsilon \in [0;1]$. Ce nombre ε est choisi librement par le chercheur qui se charge d'apprécier l'écart qu'il considère admissible entre les deux mesures associées à la règle $a_i \Rightarrow b_j$. Il peut être modifié à loisir en fonction de son niveau d'exigence de la conformité qu'il veut estimer.

Utilisons ponctuellement par commodité les notations suivantes, en considérant la généricité de cette règle :

1. $\Phi(a_i, b_j)$, valeur a priori, par Φ
2. $\varphi(a_i, b_j)$, valeur observée, par φ

[3] Nous avons alors démontré dans (Gras R. et al., 1996) que si $\Theta(a,b)$ est la variable aléatoire dont $\varphi(a,b)$ est la réalisation dans l'épreuve pour des paramètres donnés et lorsque le modèle asymptotique gaussien est choisi pour modéliser la variable $Q(a,\bar{b})$ construite à partir de $\mathrm{Card}(X \cap \overline{Y})$ alors : $\forall \partial \in [0,1]$: $\mathrm{Prob}\left[\Theta(a;b) \geq 1-\hat{\partial}\right] \approx \hat{\partial}$. Cette propriété pourrait permettre une probabilisation explicite de Ω.

Ainsi, notre intérêt se porte sur la réalisation de l'évènement A(ε)={ $|\Phi - \varphi| \leq \varepsilon$ }. Le critère de décision d'adéquation *a priori-contingence* s'exprime à partir de $n_{A(\varepsilon)}$, nombre de fois où l'écart observé satisfait l'inégalité ci-dessus à un seuil de significativité égal à α, en tant que niveau de risque de 1ère espèce.

Notre démarche s'organise autour du raisonnement suivant. Considérons l'événement A(ε), concordance ponctuelle entre ***a priori*** **et** ***a posteriori***. Notons sa probabilité Prob{A(ε)}=π. Cette valeur peut être interprétée comme un indice de confiance en la concordance ponctuelle entre la valeur de confiance (intensité d'implication) attribuée à l'implication a⇒b dans une analyse *a priori* par le chercheur et celui issu de l'observation (intensité d'implication observée)

La probabilité de A(ε) peut être estimée par $\hat{\pi} = \frac{n_{A(\varepsilon)}}{m}$

Supposons alors que le chercheur émette une hypothèse relative à la valeur de π qu'il juge acceptable, soit π_0. Cela le conduit alors à réaliser le test unilatéral suivant :

1. H_0 : $\pi \geq \pi_0$
2. H_1 : $\pi < \pi_0$

Le rejet de Ho traduit le rejet de la tendance à la concordance parce que la probabilité « vraie » de l'événement « jugement concordant au niveau ε » serait vraisemblablement plus faible que ce que le chercheur s'est donnée comme référence.

Nous rappelons un résultat classique. La variable de décision D donnant le nombre de couples favorables au critère et qui est associée à ce test, n'est autre, sous Ho, que la variable binomiale de paramètres m et π_0. Le nombre de couples vérifiant le critère $|\Phi - \varphi| \leq \varepsilon$ espéré est alors $m\pi_0$. Pour un niveau de risque α de 1ère espèce fixé, il s'agit de trouver le nombre k, tel que :

[Formule 2]
$$\begin{cases} \text{Prob}\{D \leq k\} = \sum_{i=0}^{i=k} C_m^i \pi_0^i (1-\pi_0)^{m-i} \leq \alpha \\ \text{Prob}\{D \leq k+1\} = \sum_{i=0}^{i=k+1} C_m^i \pi_0^i (1-\pi_0)^{m-i} > \alpha \end{cases}$$

On rejette Ho si le nombre de couples vérifiant le critère $|\Phi - \varphi| \leq \varepsilon$ est inférieur ou égal à k. En d'autres termes, la région critique est constituée par les nombres entiers de 0 à k.

6 Exemple

Cet exemple est tiré de travaux de thèse de Filippo Spagnolo (Spagnolo, 1995). L'essentiel de cette thèse porte sur le Postulat de Eudoxe-Archimède. Nous donnons en annexe une analyse a priori des comportements attendus des 107 élèves répondant à des questionnaires. Ce sont ces comportements qui constitueront l'ensemble des variables à partir desquelles, on estimera *a priori* puis on observera l'intensité d'implication des règles associées.

Dans une analyse a priori sur l'obstacle épistémologique (Spagnolo 2006), nous pouvons repérer des implications entre groupes des variables.

Les éléments significatifs et décisifs dans le modèle de l'obstacle épistémologique sont :

1. la résistance ;
2. la persistance ;
3. le changement du point de vue ;
4. la généralisation.

Les variables de l'analyse a priori sont les mêmes que celles qui sont retenues dans l'analyse a posteriori ou contingence (Annexe). Autrement dit W=V. Ici nous examinerons les règles généralisées suivantes :

I_1 : Si l'obstacle résiste et persiste alors on aura un changement du point de vue, c'est à dire : [Si (Q_{12},Q_{13}, Q_{14}, Q_{15}) alors (Q_8 ,Q_9 ,Q_{10})].

I_2 : Si l'obstacle résiste et persiste alors il y aura une généralisation,

c'est à dire : [Si (Q_{12} , Q_{13} , Q_{14} , Q_{15}) alors (Q_{11})].

Les élèves relevant d'un modèle d'obstacle qui résiste et persiste peuvent alors changer le point de vue.

Le tableau des intensités d'implication a priori pour 12 couples constituant une R-règle généralisée $\mathbf{I_1}$ de la de la hiérarchie orientée est donné ici [4] :

$\Phi_{12, 8}$=0.99	$\Phi_{12, 9}$=0.99	$\Phi_{12, 10}$=1
$\Phi_{13, 8}$=0.75	$\Phi_{13, 9}$=0.79	$\Phi_{13, 10}$=0.83
$\Phi_{14, 8}$=0.99	$\Phi_{14, 9}$=0.91	$\Phi_{14, 10}$=0.99
$\Phi_{15, 8}$=1	$\Phi_{15, 9}$=1	$\Phi_{15, 10}$=1

TAB. 2- *Intensités d'implication a priori*

Le tableau des intensités d'implication contingentes entre les mêmes couples est alors :

$\varphi_{12, 8}$=0.98	$\varphi_{12, 9}$=0.88	$\varphi_{12, 10}$=0.65
$\varphi_{13, 8}$=0.75	$\varphi_{13, 9}$=0.79	$\varphi_{13, 10}$=0.00
$\varphi_{14, 8}$=0.66	$\varphi_{14, 9}$=0.99	$\varphi_{14, 10}$=0.65
$\varphi_{15, 8}$=0.79	$\varphi_{15, 9}$=0.82	$\varphi_{15, 10}$=0.67

TAB. 3- *Intensités d'implication contingentes*

Par exemple, le couple (12 ; 8) formé des variables 12 et 8, a pour intensité a priori : 0.99 et pour intensité observée : 0.98.

Ici, m = 12, couples sont retenus parce qu'ils constituent une classe de la hiérarchie orientée.

Si nous appliquons le critère [Formule1], nous obtenons $\Delta \approx 1{,}17$. Si nous appliquons le seuil posé par Spagnolo, nous avons $\Delta > 0{,}25$ dont on tire la conclusion que l'adéquation entre les deux séries d'intensités a priori/a posteriori n'est pas acceptable.

Appliquons maintenant le test exprimé dans le § 5. Choisissons pour cela, $\varepsilon = 0{,}10$ et π_0=0,75, autrement dit, nous acceptons l'adéquation d'autant mieux, si, au cours de notre

[4] Nous notons par exemple : $\Phi_{12, 8}$ l'implication a priori de la variable Q_{12} sur la variable Q_8. Les occurrences respectives de ces variables sont 3 et 31.

expérience, nous trouvons au moins 12x0,75 = 9 règles pour lesquelles l'écart entre l'intensité d'implication a priori et l'intensité observée est inférieure ou égale à 0,10.

Ici l'hypothèse H_0 s'exprime alors ainsi : la probabilité « vraie »π est au moins égale à 0,75.

Par exemple, la règle $Q_{14} \Rightarrow Q_9$ où $|\Phi_{14,\,9} - \varphi_{14,\,9}| = 0{,}08$, présente un écart entre l'intensité théorique a priori et celle contingente inférieur à 0,10. En fait, nous dénombrons 4 cas seulement vérifiant l'événement

$$A(0{,}10)=\{|\Phi-\varphi| \le 0{,}10\}.$$

La résolution du système d'inéquations [Formule2] associé à la variable binomiale de paramètres (12 ; 0,75), donne k=5 et la région critique est {0, 1, 2, 3, 4, 5} pour un niveau de α=0,05. On constate donc le nombre de couples satisfaisant le critère A(0,10) est dans la zone critique. Nous rejetons Ho ce qui équivaut à ne pas conserver l'hypothèse d'adéquation entre l'a priori et la contingence.

Revenant au problème de l'obstacle épistémologique identifié par F Spagnolo, on peut dire que se trouve également réfutée l'hypothèse selon laquelle il y a effectivement changement de point de vue face à la résistance et la persistance de l'obstacle.

Supposons que nous nous placions dans le cas où le chercheur accepte un écart de ε=0,20. Il y a alors 6 règles satisfaisant le critère A(0.20). Dans ce cas, nous ne rejetons pas Ho. Ceci équivaut à conserver l'hypothèse d'adéquation entre l'a priori et la contingence avec un risque 2ème espèce.

Pour évaluer ce risque β, nous supposons que la valeur « vraie » de π est 0,5, c'est à dire 1 chance sur 2 de satisfaire au critère de concordance A(0,20). Nous obtenons :

$$\Pr ob_{H1}\{D \ge 6\} = \beta = (\frac{1}{2})^{12} \sum_{k=6}^{12} \frac{12!}{k!(12-k)!} \approx 0{,}6127$$

ou encore la puissance du test est de 1-$\beta \approx$0,3873

7 Conclusion

Dans le cadre de la didactique des disciplines, mais également dans toute situation expérimentale de recherche qui se veut scientifique, se pose constamment le problème de la question légitime, voire incontournable, de la validation ou de la réfutation d'hypothèses émises par le chercheur avant l'expérience. Se donner un moyen d'évaluer l'écart en termes quantitatifs entre les hypothèses a priori et les observations, la contingence, le *a posteriori* est l'objet de cet article. Dans l'autre publication (Spagnolo F., 1997), nous l'avions exprimé au moyen d'une distance de type χ^2 comme il est dit dans le sous-chapitre 4, mais sans référence à une échelle de mesure probabiliste. Dans ce chapitre, nous l'exprimons à l'aide d'un test dans le sous-chapitre 5 et nous en indiquons la pratique en l'illustrant par un exemple réel.

Références

Agrawal R. et al. (1993). *Mining association rules between sets of items in large databases*, Proc. of the ACM SIGMOD'93.

Bodin, A. (1996). *Improving the Diagnostic and Didactic Meaningfulness of Mathematics Assessment in France,* Annual Meeting of the American Educational Research Association AERA - New-York

Couturier R. (2001). *Traitement de l'analyse statistique implicative dans CHIC,* Actes des Journées sur la « Fouille dans les données par la méthode d'analyse implicative »

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'Etat, Université de Rennes 1.

Gras R. (2000). Les fondements de l'analyse implicative statistique, *Quaderni di Ricerca in Didattica*, Palermo, http://dipmat.math.unipa.it/~grim/quaderno9.htm

Lerman I.C. (1981) *Classification et analyse ordinale des données,* , Paris : Dunod.

Lerman I.C., Gras R., Rostam H. (1981). Elaboration et évaluation d'un indice d'implication pour des données binaires, I et II, *Mathématiques et Sciences Humaines* n° 74,. 5-35 et n° 75,. 5-47.

Spagnolo F. (1995), *Obstacles épistémologiques: Le Postulat d'Eudoxe-Archimede*, Tesi di Dottorato, Quaderni di Ricerca in Didattica, Supplemento al n.5, Palermo

Spagnolo F. (1997), L'analisi a priori e l'indice di implicazione statistica di Gras, *Quaderni di Ricerca in Didattica, n° 7, Palermo,* p. 110-117

Spagnolo F.(2005), L'Analisi Statistica Implicativa : uno dei metodi di analisi dei dati nella ricerca in didattica delle Matematiche, Troisième Rencontre Internazionale A.S.I. (Analyse Statistique Implicative), Octobre 2005, Palermo. Supplemento 2 al n.15 *Quaderni di Ricerca in Didattica* http://math.unipa.it/~grim/asi/suppl_quad_15_2.htm .

Annexe

Voici le tableau des comportements attendus. Ils doivent se révéler à travers des items de questionnaires : par exemple, la réponse positive à la question 9a instancie le comportement Q1.

Q1 (9a)	**Connaissance du Postulat d'Euclide-Archimède (P.E-A.) en termes opérationnels. Il faut déterminer un n tel que le multiple du segment na > b. (a<b). Formulation directe du Postulat. La représentation avec des petits tirets donne une liaison avec la mesure. Réponses attendue : n>4.**
Q2 (10a)	Question semblable à la précédente, mais le segment **b** est beaucoup plus grand et le segment **a** a été dessiné plus petit. Réponse attendue : n>19. Formulation directe du Postulat P.
Q3 (11a)	Répondre affirmativement à l'existence du n tel que n**a**>**b**. Formulation directe du Postulat P
Q4 (11b)	Justifier la réponse à la question précédente. Formulation directe du Postulat P
Q5 (12a)	Connaissance du P.E.-A en termes opérationnels. Il faut déterminer un n tel que (1/n)**a**<**b**). Réponse attendue : n>3. Formulation inverse du Postulat.
Q6 (13a)	Réponse affirmative à l'existence du n / (1/n)**a**<**b**. Formulation inverse de P.
Q7 (14a)	Formulation linguistique différente de la question précédente: "...il est toujours possible...".
Q8 (15a)	Changement du point de vue: Modèle du "Véronèse", non-Archimédien, en Géométrie. Réponse attendue: affirmer la non-validité du Postulat.
Q9 (18a)	Changement du point de vue, L'élève doit suivre la construction évoquée dans la proposition X.1.[5] d'Euclide et conclure à sa validité (angles rectilignes).
Q10 (19a)	Changement du point de vue, Validité de la proposition X,1 (angles curvilignes). (la question 17 indique comment confronter des angles curvilignes entre eux).L'élève doit effectuer une construction et conclure à la validité du PEA.
Q11 (20a)	Contexte généralisé (comparaison d'angles curviligne et rectiligne). L'élève doit rejeter la validité de la proposition X.1 dans ce cas.
Q12 (20b)	Résister aux contradictions avec la contingence: l'élève qui donne une justification de la X,1 (un contexte non Archimédien).
Q13 (20c)	Pour réussir, l'élève donne un argument pour le rejet de du procédé de la X,1 dans un contexte non Archimédien.
Q14 (21a)	Confirmation: l'obstacle persiste: Affirmer la validité de la proposition X,1, pour un élève, c'est montrer qu'il ne parvient pas à éprouver son modèle interprétatif dans un contexte plus général.
Q15 (22a)	Confirmation de la position Q13: 1=Affirmer la non-validité de la proposition X,1.
Q16 (3a)	Les élèves doivent chercher une relation d'ordre entre 3 triangles.
Q17 (4a)	Relation d'ordre entre 3 triangles (autres contextes).

[5] **Proposition X.1***:* Deux grandeurs inégales étant proposées, si l'on retranche de la plus grande une partie plus grande que sa moitié, et si l'on fait toujours la même chose, il restera une certaine grandeur qui sera plus petite que la plus petite des grandeurs proposées.

Q18 (5a)	Relation d’ordre entre 3 triangles (autres contextes).
Q19 (6a)	R.O. entre angles rectilignes
Q20 (7a)	R.O. entre angles curviligne (paraboles).
Q21 (8a)	R.O. entre angles curvilignes (contingence).
Q22(16a)	R.O. inclusion entre angles rectilignes.
Q23(17a)	R.O. inclusion entre angles contingents.
Q24(17b)	R.O. inclusion entre angles curviligne et contingence.

Summary

In Mathematics Education, but generally in human sciences, numerous research use qualitative analyses to falsify the formulated a priori hypotheses experimentally. Such a methodological approach, applied to an investigation, proves to be the most often insufficient to analyze all variables in game in contingent phenomena of teaching / training, even though in some cases (punctual analysis of protocols, of videos, etc.), it permits to discover some interesting relations. But if the number of pupils becomes too voluminous, the qualitative analysis doesn't succeed anymore in extracting all relations existing between the variables in game. A quantitative analysis on a statistical basis will impose itself and will be completed by a qualitative analysis, indispensable to a contextual interpretation. This communication aims to present a measure permitting to confront statistically, the analysis a priori and contingency.

Chapitre 4 : Problème de données manquantes dans un tableau numérique. Une application de l'A.S.I.

Régis Gras

Equipe COnnaissances & Décision (COD)
Laboratoire d'Informatique de Nantes Atlantique – FRE CNRS 2729
Site Ecole Polytechnique de l'Université de Nantes
La Chantrerie BP 50609 44306 Nantes cedex 3
regisgra@club-internet.fr

Résumé. Une base de données croisant variables et sujets issues d'observations présente souvent des vides dus, par exemple, à des absences de réponse ou à l'impossibilité matérielle de la recueillir. Or, pour en effectuer un traitement, il est essentiel de disposer d'un tableau complet. L'analyse statistique implicative, entre autres méthodes d'analyse de données, à l'œuvre au moyen du logiciel de traitement C.H.I.C. (Couturier et Gras., 2005), impose que les vides soient comblés. Se pose alors le problème de déterminer quelle valeur la plus vraisemblable attribuer à la variable non observée sur tel sujet ou, de façon symétrique, quelle valeur attribuer à un sujet sur une variable donnée et muette sur lui. Nous présentons ici une méthode qui, au vu du comportement de réponse observé par le sujet sur d'autres variables, intimement liées à la variable muette, permet de pallier la carence locale. Un exemple numérique illustre l'usage de cette méthode sur un tableau incomplet.

1 Problématique et contraintes sémantiques

L'analyse statistique implicative traite un tableau de données qui croise un ensemble de variables V, colonnes du tableau, de cardinal v et un ensemble E de sujets (ou d'objets), lignes du tableau, de cardinal n. L'intersection d'une ligne et d'une colonne représente donc la valeur x(i) prise par l'individu x selon la variable i. Dans le cas où les variables V sont binaires x(i) = 0 ou 1. Dans les autres cas classiques (variables modales, fréquentielles, numériques normées), les valeurs sont des nombres réels de l'intervalle [0,1].

Or il est possible que les observations ou les mesures faites conduisant au tableau présentent des "trous", c'est-à-dire des absences de réponse. Il existe donc des individus x et des variables i telles que la **valeur x(i) soit manquante**. Si l'on ne souhaite pas supprimer l'individu x, donc toutes les autres observations faites en x selon les autres variables, et par suite perdre un certain nombre d'informations permettant d'étudier les relations implicatives entre variables, le problème va consister à choisir la valeur la plus pertinente que l'on pourra affecter à x(i) pour la structure des variables sous-jacente.

L'idée intuitive est de chercher quels sont les individus qui ont des comportements semblables à x selon les autres variables sur lesquelles nous disposons d'informations pour x et d'attribuer à x(i) la valeur correspondant à celle que prend en i l'individu y, le plus "ressemblant" à x. La modélisation va donc porter sur le choix d'un critère de ressemblance entre individus.

De nombreuses méthodes[1] visant à remplacer des données manquantes par des valeurs estimées existent. Le logiciel libre R permet d'y parvenir en utilisant des procédures de régression ou de recherche du maximum de vraisemblance. Nous présentons ici deux méthodes originales : l'une exploite la proximité relative des individus eu égard à leur instanciation des variables ; l'autre exploite les relations implicatives entre les variables à données complètes.

2 Contraintes analytiques sur le modèle

Nous supposons que x n'est pas défini en i alors que y l'est par la valeur y(i). Nous supposons également que le nombre de valeurs non définies en x n'est pas trop grand et que x et y prennent des valeurs sur de nombreuses variables communes, faute de quoi l'estimation de x(i) serait entachée d'imprécision. Cette hypothèse est assez générale pour ce qui suit, le risque d'erreur croissant avec le nombre de « trous » dans le tableau de données.

Le critère de ressemblance entre deux individus x et y doit être nécessairement symétrique comme le sont tous les critères de similarité.

Il doit prendre en compte le maximum d'informations sur les valeurs prises par x et y selon les variables qu'ils affectent en commun. Ce sont donc les profils de x et y sur ces variables qu'il nous faudra envisager. Ces profils n'intégreront donc que les variables où, pour x **et** y, il n'y a pas de valeurs manquantes.

Ces profils doivent prendre en compte toutes les valeurs en jeu en intégrant la qualité de représentation relative des individus selon l'ensemble des variables sur lequel ils sont observés. Si, par exemple, la valeur x(j) est élevée et si la somme totale marginale $x. = \sum_{j \in V} x(j)$ des valeurs prises par x sur V (ou une partie de V en cas de valeurs manquantes pour x) est relativement faible, on accentuera la position de j dans le profil en calculant $\frac{x(j)}{x.}$

3 Méthodologie de substitution

Le profil de x sera donc l'ensemble $\left\{ \frac{x(j)}{x.} \right\}_{j \in V}$, distribution conditionnelle de j sachant x. L'écart entre les termes $\frac{x(j)}{x.}$ et $\frac{y(j)}{y.}$ du profil observé en x et y selon une variable j quelconque doit également être relativisé à la représentation de cette variable sur l'ensemble des individus. Si, par exemple, une variable ne prend que de faibles valeurs sur E, un écart important entre $\frac{x(j)}{x.}$ et $\frac{y(j)}{y.}$ doit être majoré par rapport à ce qu'il serait si la variable

[1] Voir par exemple : Schafer J.L. (2000), *Analysis of incomplete multivariate data*, Chapman and Hill.

n'avait pris que de fortes valeurs. On satisfera cette contrainte en divisant chaque distorsion entre les termes des profils par la somme marginale $n_j = \sum_{z \in E} z(j)$. La division de n_j par N ne change nullement le rapport entre les distorsions mais permet d'homogénéiser ce rapport en des termes relatifs. Un tableau initial se présente donc comme ceci :

V→ E↓	a	…	i	…	j	…	Marge
1	1(a)	…	1(i)	…	1(j)	…	1.
…	…	…	…	…	…	…	…
x	x(a)	…	XXXXX	…	x(j)	…	x.
…	…	…	…	…	…	…	…
y	y(a)	…	y(i)	…	y(j)	…	y.
…	…	…	…	…	…	…	…
Marge	n_a	…	n_i	…	n_j	…	N

TAB. 1

La dissemblance totale prend alors la forme d'une distance du χ^2 entre les deux profils relatifs de x et y, comparable à celle utilisée en analyse factorielle des correspondances, soit :

$$d(x,y) = \left[\sum_{j \in V', j \neq i} \frac{\left[\frac{x(j)}{x.} - \frac{y(j)}{y.} \right]^2}{\frac{n_j}{N}} \right]^{\frac{1}{2}}$$

où V' est le sous-ensemble de V des variables selon lesquelles x et y prennent en commun des valeurs. V-V' est donc l'ensemble des variables où x et/ou y présentent l'un et/ou l'autre des valeurs manquantes. Autrement dit, le calcul se fait sur le tableau extrait du tableau initial duquel ont été supprimées les colonnes relatives aux variables de V-V', y compris le calcul des valeurs marginales (individus et variables).

Cependant, un autre facteur est "aggravant" par rapport à cette dissemblance. En effet pour que l'ensemble des dissemblances entre x et les autres individus permette des comparaisons, il faut prendre en compte le cardinal k de V'. Si ce cardinal est faible, la dissemblance définie ci-dessus risque d'être faible et peu informative. On l'affecte donc d'un coefficient correcteur qui décroît si le nombre k de variables partagées en commun par x et y est important, soit le rapport : $\frac{v-k}{v}$. Par conséquent , on pose :

$$\delta(x,y) = \frac{v-k}{v} d(x,y)$$

En définitive, on affectera à x(i) la valeur y_0(i) si $\delta(x,y_0) = \inf_{y \in E / \{x\}} \delta(x,y)$. Si deux individus satisfont ce minimum, on choisira, prioritairement, celui qui est relatif à un sous-

ensemble V' le plus important. En cas de nouvelle égalité, on choisira celui dont l'effectif marginal y. est le plus fort. En cas de nouvelle égalité, on retiendra arbitrairement l'individu le plus haut placé dans le tableau ExV d'entrée.

A la suite de chaque affectation de valeurs estimées, les marges sont remises à jour, de même que la valeur de N.

Remarque 1

Il est donc évident que l'ordre dans lequel sont effectuées les opérations de substitution aura une influence sur les estimations successives. Leur rareté supposée nous permet de faire l'hypothèse que cette influence sera négligeable.

Remarque 2

On rencontre dans des situations d'évaluation de travaux d'élèves des épreuves à plusieurs modalités. Par exemple, pour des raisons de sécurité vis-à-vis du copiage, mais également pour des mises à l'épreuve d'hypothèses sur la proximité de complexité d'items, on partitionne une population en 3 groupes X, Y et Z. Chacun de ces groupes est soumis à deux cahiers d'items parmi trois cahiers A, B et C comme l'indique le tableau suivant :

X	A	B	XXXX
Y	XXXX	B	C
Z		XXXX	C

TAB.2

L'absence de résultats des élèves de X en C sera compensée par une estimation à partir de la comparaison de ses comportements de réponse selon B avec les élèves de Y et selon A avec les élèves de Z. On prendra, par exemple, la moyenne des estimations faites à partir de Y et de Z. S'il s'agissait de variable binaire (réussite-échec), on choisirait l'entier 0 ou 1 le plus proche de cette moyenne.

Remarque 3

On aurait pu utiliser le coefficient de corrélation linéaire pour signifier la ressemblance entre les individus x et y. Cependant, nous lui préférons l'indice défini plus haut qui intègre, mieux que la corrélation, les nuances distributionnelles prises en compte dans sa définition.

De même, on aurait pu faire appel au coefficient de similarité de I.C. Lerman (1981). Mais celui-ci ne paraissait pas permettre la relativisation que nous avons introduite.

4 Une autre approche dans le cadre de l'A.S.I.

Un autre point de vue mettant l'accent sur les variables plutôt que sur les individus est également envisageable dans le cadre de l'Analyse Statistique Implicative (Gras et al. 2001 et Gras, 2005) et Partie 1, chap 1. Ainsi, l'absence d'une donnée a(x) selon la variable a peut être compensée par l'attribution d'une valeur statistique en considérant la "proximité implicative" de a avec des variables dont on connaît la valeur prise en x. Par exemple, si C_a et C'_a sont respectivement les classes de variables impliquées par a et impliquant a au seuil $1-\alpha$, il suffit de considérer leur intersection avec les classes associées respectivement à celles des variables b pour lesquelles b(x) est connu. Un indice de distance, comparable à celui que nous avons défini plus haut, permettrait d'affecter un a(x) une valeur la plus vraisemblable eu

égard aux valeurs de type b(x). Le logiciel d'analyse de données CHIC (Classification Hiérarchique Implicative et Cohésitive) (Couturier et Gras. 2005) et Partie 2, chap.11 permet de mettre en évidence aisément les variables associées à a (resp. b) en choisissant l'option « Sélectionne les variables en mode cône ». La variable a sélectionnée présente, seule, ses pères et ses fils. La métaphore « cône » est justifiée par la structure qui apparaît comme le ferait un cône à deux nappes. Si, à son tour, la variable b est sélectionnée, on peut, avec un seuil convenable, examiner l'intersection des deux cônes.

Il est possible d'opérer autrement et d'utiliser les liaisons implicatives entre les variables d'une autre façon qui permette d'enrichir l'estimation de la donnée manquante x(i), faite précédemment.

Pour cela, considérons le schéma implicatif établi autour de la variable i et n'intégrant pas l'action de la donnée manquante en x mais celles des mêmes sujets. Dans le cas général, pour un seuil de l'intensité d'implication fixé (par exemple 0.95), une certaine classe de variables C, classe père disposant de valeurs en x, non totalement ordonnées, établit des liaisons implicatives vers i ; de même, une autre classe C', classe fils disposant également de valeurs en x, non totalement ordonnées, est constituée de variables impliquées par i. Si l'une des classes C et C' est vide, ce qui suit n'intéressera que celle qui est non vide. Si les deux sont vides et si un abaissement du seuil de l'intensité fait perdre toute signification , on conserve les résultats de l'estimation de x(i) précédente.

Revenons au cas général et retenons alors les variables j et k qui présentent des intensités d'implication maximales, respectivement de j de la classe C vers i et de i vers k de la classe C'. Ainsi, la variable j implique de façon la plus consistante au sein de C la variable i qui elle-même implique de façon la plus consistante au sein de C' la variable k. C'est à ces variables que l'on accordera le meilleur crédit pour affiner l'estimation de x(i).

L'une d'entre ces deux classes présente une cohésion plus forte ou égale à celle de l'autre. S'il s'agit de la classe C, nous choisirons pour x(i) le minimum de la valeur précédemment établie et de la valeur x(j) observée en x selon j, minimum qui permet de ne pas détruire, le plus souvent, la liaison de j vers i. S'il s'agit de la classe C', nous choisirons pour x(i) le maximum de la valeur précédemment établie et de la valeur x(k) observée en x selon k, maximum qui permet également de ne pas détruire, le plus souvent, la liaison de k vers i.

Ainsi, nous disposons d'une information plus complète pour déterminer la valeur manquante puisque nous intégrons simultanément et les comportements de x comparés à ceux des autres sujets relativement aux variables et ceux des variables dans l'ensemble de leurs liaisons au sein de la population où elles sont comparables.

5 Exemple didactique

Supposons donné le tableau suivant correspondant à l'observation sur 4 individus de 4 variables modales a, b, c et d , prenant les valeurs 0 ou 0,25 ou 0,5 ou 0,75 ou 1.

Il s'agit donc d'estimer les valeurs e3(c) et e4(d) sur la base des autres observations. Les marges de ce tableau n'intègrent que les variables où, dans un premier temps, e3 est estimé à l'aide de e1, e2 et e4. Donc, pour cette estimation, la variable c n'intervient pas pour e1, e2 et e4.

V → / E ↓	a	b	c	d
e1	0	0,50	0,25	0
e2	0,50	0,75	1	0,25
e3	0,50	0,50	**?**	0,25
e4	0,75	0,50	1	?

TAB.3 *: Tableau d'entrée des données*

V→ / E ↓	a	b	c	d	Marge pour e3
e1	0	0,50	0,25	0	0,50
e2	0,50	0,75	1	0,25	1,50
e3	0,50	0,50	**?**	0,25	1,25
Marges	1	1,75	1	0,50	3,25

TAB. 4 : *Tableau pour le calcul des distances de e3 à e1 et e2*

V→ / E ↓	a	b
e1	0	1
e2	1/3	1/2
e3	0,40	0,40

TAB. 5 *Fréquences conditionnelles des individus*

Les calculs successifs donnent les résultats suivants pour v=4 :

d(e1, e3) = 1,203 et δ (e1, e3) = 0,301 (pour k=3)

d(e2, e3) = 0,237 et δ (e2, e3) = 0,0592 (pour k=3)

TAB.6

Par exemple :

$$d(e1,e3) = \sqrt{\frac{(0-0,4)^2}{\frac{1}{3,25}} + \frac{(1-0,4)^2}{\frac{1,75}{3,25}} + \frac{(0-0,2)^2}{\frac{0,5}{3,25}}} = 1,203 \quad \text{d'où} \quad \delta(e1,e3) = \frac{1,203}{4} = 0,301$$

car k=3. Dans la comparaison de e3 avec e4, les variables c et d n'interviennent pas (variables non communes).

V→ / E↓	a	b	Marges
e3	0,50	0,50	1
e4	0,75	0,50	1,25
Marges	1,25	1	2,25

TAB. 7 *Tableau pour le calcul de distance entre e3 et e4*

V→ / E↓	a	b
e3	0,50	0,50
e4	0,60	0,40

TAB. 8 : *Fréquences conditionnelles de e3 et e4*

d'où d(e3, e4) = 0,201 et δ (e3, e4) = 0,101 (pour k= 2) alors que δ (e1, e3)= 0,301 et que δ (e2, e3) = 0,0592.

Par suite, e2 est l'individu le plus "ressemblant" à e3. On choisira donc l'estimation :

e3(c) = e2(c) = 1.

On obtient alors une nouvelle distribution donnée par les tableaux suivants à partir desquels nous estimerons e4(d) :

V→ E↓	a	b	c	d	Marges pour e4
e1	0	0,50	0,25	(0)	0,75
e2	0,50	0,75	1	(0,25)	2,25
e3	0,50	0,50	**1**	(0,25)	2
e4	0,75	0,50	1	?	2,25
Marge	1,75	2,25	3,25	?	

TAB. 9 *Tableau de données après estimation de e3(c)*

V→ E↓	a	b	c
e1	0	2/3	1/3
e2	2/9	3/9	4/9
e3	0,25	0,25	0,5
e4	3/9	2/9	4/9

TAB.10 : *Fréquences conditionnelles des individus*

Cette fois, toutes les informations distributionnelles relatives à d disparaissent puisque e4 n'est pas défini en d. Reprenons les calculs en tenant compte de la première estimation de e3(c). Nous obtenons :

d(e1, e4) = 1,060 et δ (e1, e4) = 0,265 (pour k = 3)

d(e2, e4) = 0,301 et δ (e2, e4) = 0,075 (pour k = 3)

d(e3, e4) = 0,397 et δ (e3, e4) = 0,099 (pour k= 3)

Par suite, e2 est l'individu le plus "ressemblant" à e4. On choisira donc l'estimation : e4(d) = e2(d) = 0,25. On obtient alors un tableau complet sur lequel on pourra pratiquer l'analyse statistique implicative habituelle.

V→ E↓	a	b	c	d
e1	0	0,50	0,25	0
e2	0,50	0,75	1	0,25
e3	0,50	0,50	**1**	0,25
e4	0,75	0,50	1	**0,25**

TAB. 11

6 Conclusion

Nous avons présenté une méthode numérique afin de pallier les carences d'un tableau de données incomplet. A l'aide d'une distance de type χ^2 intégrant le maximum d'informations observées sur les variables sur lesquelles des sujets sont mesurés, nous avons affecté à un sujet, incomplètement défini selon une variable, une valeur qu'il aurait pu prendre sur celle-ci, eu égard à sa ressemblance avec les autres sujets selon les variables partagées en commun. L'exemple numérique, traité à la main, illustre cette méthode. Nous avons aussi évoqué une autre méthode qui s'appuierait sur les données fournies par les calculs des implications respectives entre les variables partout où elles sont complètement définies. Cette stratégie qui prend plus encore en compte la structure implicative de l'ensemble des variables sera développée par ailleurs.

Références

Agrawal R., Imielinski T. and Swami A. (1993) : Mining association rules between sets of items in large databases. In Peter Buneman and Sushil Jalodia, Editors, *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, p.207-216, Washington, D.C. 26-28*

Couturier R, Gras R. (2005) : CHIC : Traitement de données avec l'analyse implicative, *Extraction et Gestion des Connaissances, Volume I1, RNTI, Cepadues, Paris, p.679-684, , ISBN 2.85428.683.9*

Gras R., Kuntz P. et Briand H.(2001): Les fondements de l'analyse statistique implicative et quelques prolongements pour la fouille de données, *Mathématiques et Sciences Humaines*, n° 154-155, p 9-29, ISSN 0987 6936

Gras R. (2005): Panorama du développement de l'A.S.I. à travers des situations fondatrices, *Actes de la 3ème Rencontre Internationale A.S.I., Supplément n° 15 de la Revue « Quaderni di Ricerca in Didattica* », *p. 9-33, ISSN 1592-5137*, Université de Palerme

Summary

The problem of missing data is a well known problem in data analysis. The Statistical Implicative Analysis (implemented in the software C.H.I.C. (Couturier and Gras, 2005)), like others data mining methods, needs that all the variables of the dataset are instantiated. Therefore, the problem is to set the missing variables with the most likely values. We present in this chapter a method to handle this problem, using information about the values of others variables, correlated with the missing one, on the same example. A numerical example is given to illustrate our method.

Chapitre 5 : Analyse Implicative Séquentielle

Julien Blanchard, Fabrice Guillet, Régis Gras

Equipe Connaissances & Décision (COD)
LINA – FRE CNRS 2729 – Ecole Polytechnique de l'Université de Nantes
julien.blanchard@polytech.univ-nantes.fr

Résumé. La découverte de motifs fréquents dans des séquences (généralement des séquences temporelles d'évènements) est l'une des tâches majeures de la fouille de données. Dans cet article, nous nous intéressons à l'évaluation de la qualité des règles séquentielles. Nous proposons une mesure inédite nommée *SII* qui évalue la significativité des règles au regard d'un modèle probabiliste. Les simulations numériques montrent que *SII* que a des caractéristiques uniques en comparaison aux autres mesures de qualité de règles séquentielles.

1 Introduction

La découverte de motifs fréquents dans des séquences symboliques (généralement des séquences temporelles d'évènements) est l'une des tâches majeures de la fouille de données. Les travaux de recherche dans ce domaine se divisent en deux catégories :

- la découverte d'*épisodes* fréquents dans une longue séquence d'évènements (approche initiée par Mannila, Toivonen, et Verkamo Mannila et al. (1995) Mannila et Toivonen (1996)),
- la découverte de *motifs séquentiels* fréquents dans un ensemble de séquences d'évènements (approche initiée par Agrawal et Srikant Agrawal et Srikant (1995) Srikant et Agrawal (1996)).

Episodes et *motifs séquentiels* sont des structures séquentielles, c'est-à-dire définies avec un ordre (partiel ou total). Une telle structure peut être par exemple :

petit-déjeuner then *déjeuner* then *dîner*

La structure est qualifiée par sa fréquence (ou support) et généralement par des contraintes sur les positions des évènements, comme une fenêtre maximale de temps "moins de 12 heures séparent *petit-déjeuner* et *dîner*" Srikant et Agrawal (1996) Mannila et al. (1997) Das et al. (1998) Höppner (2002) Sun et al. (2003).

La différence entre *épisodes* et *motifs séquentiels* réside principalement dans la mesure de leur support : la fréquence des *épisodes* est intra-séquence Mannila et al. (1997) Das et al. (1998) Weiss (2002) Höppner (2002) Sun et al. (2003) Yang et al. (2003), alors que la fréquence des *motifs séquentiels* est inter-séquences Agrawal et Srikant (1995) Srikant et Agrawal (1996) Spiliopoulou (1999) Zaki (2001) Han et al. (2005) (voir Joshi et al. (1999) pour une synthèse sur les différentes manières d'évaluer la fréquence). Ainsi, les algorithmes d'extraction d'*épisode* fréquents recherchent des structures qui se répètent souvent à l'intérieur d'une

même séquence. Au contraire, les algorithmes d'extraction de *motifs séquentiels* fréquents recherchent des structures qui se répètent dans de nombreuses séquences (indépendamment des répétitions dans chaque séquence).

L'extraction des *épisodes*/*motifs séquentiels* est souvent suivie d'une étape de génération de règles séquentielles, permettant d'effectuer des prédictions dans la limite d'une fenêtre de tempsSrikant et Agrawal (1996) Mannila et al. (1997) Das et al. (1998) Spiliopoulou (1999) Zaki (2001) Weiss (2002) Höppner (2002) Sun et al. (2003). De telles règles ont été utilisées par exemple pour faire de la prédiction de cours de bourse Das et al. (1998), ou d'évènements dans un réseau de télécommunication Mannila et al. (1997) Sun et al. (2003). Une règle séquentielle peut être par exemple :

$$\textit{déjeuner} \xrightarrow{6\text{h}} \textit{dîner}$$

Cette règle signifie : "si on observe *déjeuner* alors on observera sûrement aussi *dîner* moins de 6 heures après".

Dans cet article, nous nous intéressons à l'évaluation de la qualité des règles séquentielles. Il s'agit d'une question cruciale pour l'analyse de séquences puisque, du fait de la nature non supervisée des algorithmes d'extraction, les quantités de règles générées peuvent être énormes. Alors que la qualité des règles d'association a été largement étudiée dans la littérature (voir Blanchard et al. (2009) pour une synthèse), il existe peu de mesures dédiées à l'évaluation des règles séquentielles. En plus de la fréquence, on trouve un indice de confiance (ou précision) qui peut être interprêtée comme une estimation de la probabilité conditionnelle de la conclusion étant donnée la prémisse Srikant et Agrawal (1996) Mannila et al. (1997) Das et al. (1998) Spiliopoulou (1999) Zaki (2001) Weiss (2002) Höppner (2002) Sun et al. (2003). Une mesure de rappel est également parfois utilisée ; elle peut être interprêtée comme une estimation de la probabilité conditionnelle de la prémisse étant donnée la conclusion Weiss (2002) Sun et al. (2003). Dans Das et al. (1998) et Höppner (2002), les auteurs ont proposé une adaptation aux règles séquentielles de la J-mesure de Smyth et Goodman, un indice issu de l'information mutuelle[1]. Enfin, une mesure entropique est présentée dans Yang et al. (2003) pour quantifier l'information apportée par un épisode dans une séquence, mais cette approche n'envisage que des épisodes et non des règles de prédiction.

Poursuivant nos travaux débutés dans Blanchard et al. (2002) sur l'adaptation aux règles séquentielles de l'intensité d'implication Gras (1996), nous proposons dans cet article une mesure statistique originale pour la qualité des règles séquentielles[2]. Plus précisément, cette mesure évalue la significativité statistique des règles au regard d'un modèle probabiliste. La section suivante est dédiée à la formalisation des notions de *règle séquentielle*, d'*exemple d'une règle*, et de *contre-exemple d'une règle*, et à la présentation de la nouvelle mesure, nommée *Sequential Implication Intensity (SII)*. Dans la partie 4, nous étudions SII sur plusieurs simulations numériques et la comparons à d'autres mesures.

[1] La J-mesure est la part de l'information mutuelle moyenne relative à la vérité de la prémisse.

[2] Ces travaux ont été également présentés dans Blanchard et al. (2007).

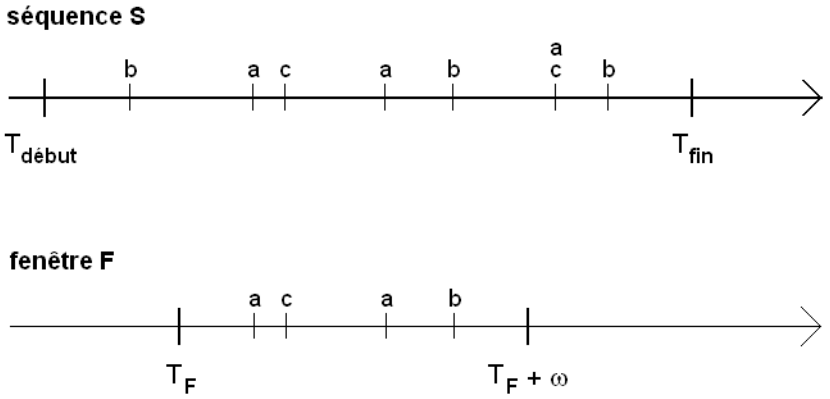


FIG. 1 – *Une séquence S d'évènements de $E = \{a, b, c\}$ et sa fenêtre F de longueur ω débutant en T_F.*

2 Mesurer la significativité des règles séquentielles

2.1 Contexte

Notre mesure SII porte sur des règles séquentielles extraites dans **une unique séquence**. Cette approche a l'avantage d'être facilement généralisable à un ensemble de séquences, par exemple en calculant une SII moyenne ou minimale sur l'ensemble. Les règles sont de la forme $a \xrightarrow{\omega} b$ où a et b sont des motifs séquentiels (ceux-ci peuvent eux-mêmes être régis par des contraintes de temps internes). Toutefois, dans cet article, nous nous limitons à des règles séquentielles où les motifs a et b sont chacun constitués d'un unique évènement.

La séquence étudiée est une séquence continue d'évènements instantanés (l'adaptation aux séquences discrètes est directe). De plus il est possible que deux évènements différents aient lieu au même instant. Ceci revient à se placer dans le cas des séquences étudiées par Mannila, Toivonen, et Verkamo Mannila et al. (1997). Pour extraire de la séquence les cardinaux nécessaires au calcul de SII, il suffit d'utiliser l'un de leurs algorithmes d'extraction de motifs, nommé Winepi Mannila et al. (1995) Mannila et al. (1997) (ou bien l'une de ses déclinaisons). Dans la suite, nous nous positionnons à l'étape des post-traitements, en considérant que Winepi a déjà été appliqué sur la séquence, et travaillons directement sur les cardinaux des motifs qui ont été extraits.

2.2 Notations

Soit E un ensemble fini de *types d'évènements* $E = \{a, b, c...\}$. Un *évènement* est un couple (e, t) où $e \in E$ est le type de l'évènement et $t \in \mathbb{R}_+$ est le temps d'apparition de l'évènement. Il est à noter que le terme d'évènement est communément employé pour désigner un type d'évènement, sans que ceci nuise à la compréhension.

Une *séquence d'évènements* S observée entre les instants $T_{début}$ et T_{fin} est une suite d'évènements

$$S = \Big((e_1, t_1), (e_2, t_2), (e_3, t_3), ...(e_n, t_n)\Big)$$

telle que :

$$\forall i \in \{1..n\}, (e_i \in E \wedge t_i \in [T_{début}, T_{fin}])$$
$$\forall i \in \{1..n-1\}, t_i \leq t_{i+1}$$
$$\forall (i,j) \in \{1..n\}^2, t_i = t_j \Rightarrow e_i \neq e_j$$

La longueur de la séquence est $L = T_{fin} - T_{début}$.

Une *fenêtre* sur une séquence S est une sous-séquence de S. Par exemple, une fenêtre F de longueur $\omega \leq L$ débutant à l'instant $t_F \in [T_{début}, T_{fin} - \omega]$ contient tous les évènements (e_i, t_i) de S tels que $t_F \leq t_i \leq t_F + \omega$.

Dans la suite, nous considérons une séquence S d'évènements de E.

2.3 Règles séquentielles

Nous établissons un cadre formel pour l'analyse des séquences en définissant les notions de *règle séquentielle*, d'*exemple d'une règle*, et de *contre-exemple d'une règle*.

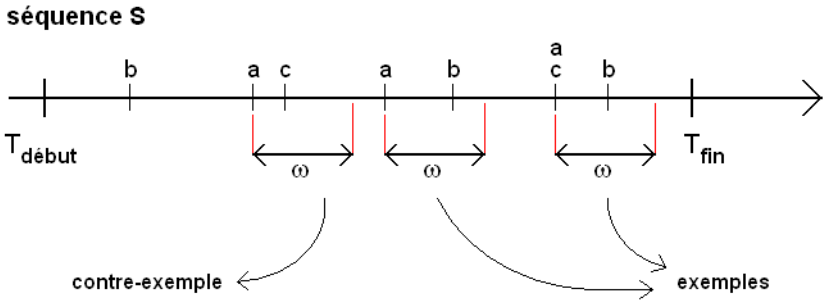


FIG. 2 – *Parmi les 3 fenêtres de longueur ω débutant sur des évènements a, on compte 2 exemples et 1 contre-exemple de la règle $a \xrightarrow{\omega} b$.*

Définition 1 Une **règle séquentielle** est un triplet (a, b, ω) noté $a \xrightarrow{\omega} b$ où a et b sont des évènements de types différents et ω est un réel strictement positif. Elle peut se lire de la manière suivante : "si la séquence comporte un évènement a alors un évènement b apparaît sûrement dans les ω unités de temps qui suivent".

Définition 2 Les **exemples** d'une règle séquentielle $a \xrightarrow{\omega} b$ sont les évènements a qui sont suivis d'au moins un évènement b à moins de ω unités de temps. Le nombre d'exemples de la règle est donc le cardinal noté $n_{ab}(\omega)$:

$$n_{ab}(\omega) = \left| (a,t) \in S \mid \exists (b,t') \in S, 0 \leq t' - t \leq \omega \right|$$

Définition 3 Les **contre-exemples** d'une règle séquentielle $a \xrightarrow{\omega} b$ sont les évènements a qui ne sont suivis d'aucun évènement b à moins de ω unités de temps. Le nombre de contre-exemples de la règle est donc le cardinal noté $n_{a\bar{b}}(\omega)$:

$$n_{a\bar{b}}(\omega) = \left| (a,t) \in S \mid \forall (b,t') \in S, (t' < t \vee t' > t + \omega) \right|$$

Contrairement aux règles d'association, n_{ab} et $n_{a\bar{b}}$ ne sont pas des constantes des données mais dépendent du paramètre ω.

La particularité de notre approche est qu'elle traite la prémisse et la conclusion de manières très différentes : les évènements a servent de référence pour la recherche des évènements b, c'est-à-dire que seules les fenêtres qui débutent par un évènement a sont prises en compte. Au contraire, dans la littérature sur les séquences, les algorithmes de type Winepi décalent (avec un pas constant) une fenêtre sur toute la longueur de la séquence Mannila et al. (1997). Cette démarche revient à considérer comme exemple de la règle séquentielle toute fenêtre qui présente a suivi de b, même si celle-ci ne débute pas par un évènement a. En comparaison, notre approche est moins complexe algorithmiquement.

Notons n_a le nombre d'évènements a dans la séquence. Nous retrouvons l'égalité bien connue $n_a = n_{ab} + n_{a\bar{b}}$. Une règle séquentielle $a \xrightarrow{\omega} b$ est entièrement caractérisée par le quintuplet $(n_{ab}(\omega), n_a, n_b, \omega, L)$. Les exemples d'une règle séquentielle étant maintenant définis, nous pouvons spécifier notre mesure pour la fréquence des règles :

Définition 4 La **fréquence** d'une règle séquentielle $a \xrightarrow{\omega} b$ est la proportion des exemples eu égard à la longueur de la séquence :

$$frquence(a \xrightarrow{\omega} b) = \frac{n_{ab}(\omega)}{L}$$

Avec ces notations, la confiance, le rappel, et la J-mesure sont donnés par les formules suivantes :

$$confiance(a \xrightarrow{\omega} b) = \frac{n_{ab}(\omega)}{n_a}$$

$$rappel(a \xrightarrow{\omega} b) = \frac{n_{ab}(\omega)}{n_b}$$

$$J\text{-}mesure(a \xrightarrow{\omega} b) = \frac{n_{ab}(\omega)}{L} \log_2 \frac{n_{ab}(\omega) L}{n_a n_b} + \frac{n_{a\bar{b}}(\omega)}{L} \log_2 \frac{n_{a\bar{b}}(\omega) L}{n_a (L - n_b)}$$

2.4 Modèle aléatoire

A l'instar de l'intensité d'implication pour les règles d'association Gras (1996), l'intensité d'implication séquentielle SII mesure la significativité des règles $a \xrightarrow{\omega} b$. Pour cela, elle quantifie l'invraisemblance de la petitesse du nombre de contre-exemples $n_{a\bar{b}}(\omega)$ eu égard à l'hypothèse d'indépendance entre les types évènements a et b. Dans une recherche de modèle aléatoire, nous supposons donc que les types d'évènements a et b sont indépendants. L'objectif est de déterminer la distribution de la variable aléatoire $\mathcal{N}_{a\bar{b}}$ (nombre de contre-exemples de la règle) étant donnés la longueur L de la séquence, les nombres n_a et n_b d'évènements de types a et b, ainsi que la taille ω de la fenêtre de temps utilisée.

Nous supposons que le processus d'arrivée des évènements de type b vérifie les hypothèses suivantes :

- les temps séparant les apparitions successives de b sont des variables aléatoires indépendantes,
- la probabilité qu'un b apparaisse dans un intervalle $[t, t+\omega]$ ne dépend que de ω.

De plus, deux évènements de même type ne peuvent arriver simultanément dans la séquence S (voir section 2.2). Dans ces conditions, le processus d'arrivée des évènements de type b est un processus de Poisson d'intensité $\lambda = \frac{n_b}{L}$. Le nombre de b apparaissant dans une fenêtre de longueur ω suit donc une loi de Poisson de paramètre $\frac{\omega.n_b}{L}$. En particulier, la probabilité pour qu'aucun évènement b ne se produise durant ω unités de temps est :

$$p = \mathrm{P}(Poisson(\frac{\omega.n_b}{L}) = 0) = e^{-\frac{\omega}{L}n_b}$$

Où qu'il apparaisse dans la séquence, un évènement a possède donc la probabilité fixée p d'être un contre-exemple, et $1-p$ d'être un exemple. Répétons n_a fois cette expérience aléatoire pour déterminer la loi du nombre de contre-exemples $\mathcal{N}_{a\bar{b}}$. Si ω est négligeable devant L, alors il est improbable que deux fenêtres de taille ω choisies aléatoirement se chevauchent, et nous pouvons considérer que les n_a répétitions de l'expérience sont indépendantes. Dans ces conditions, la variable aléatoire $\mathcal{N}_{a\bar{b}}$ est binomiale de paramètres n_a et p :

$$\mathcal{N}_{a\bar{b}} = Binomiale(n_a, e^{-\frac{\omega}{L}n_b})$$

Dans les conditions qui conviennent, la distribution binomiale peut être approximée par une seconde distribution de Poisson (même dans le cas de répétitions "faiblement dépendantes" –see Ross (2006)).

Définition 5 L'**intensité d'implication séquentielle (*SII*)** d'une règle $a \xrightarrow{\omega} b$ est définie par :

$$SII(a \xrightarrow{\omega} b) = \mathrm{P}(\mathcal{N}_{a\bar{b}} > n_{a\bar{b}}(\omega))$$

Numériquement, on a :

$$SII(a \xrightarrow{\omega} b) = 1 - \mathrm{P}(\mathcal{N}_{a\bar{b}} \leq n_{a\bar{b}}(\omega)) = 1 - \sum_{k=0}^{n_{a\bar{b}}(\omega)} \mathrm{C}_{n_a}^{k} (e^{-\frac{\omega}{L}n_b})^k (1 - e^{-\frac{\omega}{L}n_b})^{n_a - k}$$

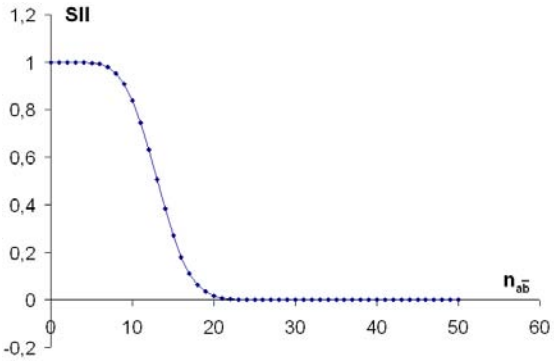


FIG. 3 – *Représentation de SII en fonction du nombre de contre-exemples.*

3 Propriétés et comparaisons

SII quantifie l'invraisemblance de la petitesse du nombre de contre-exemples $n_{a\bar{b}}(\omega)$ eu égard à l'hypothèse d'indépendance entre les types évènements a et b. En particulier, si $SII(a \xrightarrow{\omega} b)$ vaut 1 ou 0, alors il est invraisemblable que les types d'évènements a et b soient indépendants (l'écart à l'indépendance est significatif et orienté en faveur des exemples ou des contre-exemples). Ce nouvel indice peut être interprété comme le complément à 1 de la probabilité critique (*p-value*) d'un test d'hypothèse. Toutefois, à l'instar de l'intensité d'implication, il ne s'agit pas ici de tester une hypothèse mais bien de l'utiliser comme référence pour évaluer et ordonner les règles.

Dans la suite, nous testons SII dans plusieurs simulations numériques et le comparons à la confiance, au rappel, et à la J-mesure. Ces simulations soulignent les propriétés intuitives d'une bonne mesure d'intérêt pour des règles séquentielles.

3.1 Augmentation des contre-exemples

Dans cette section, nous étudions les mesures quand le nombre $n_{a\bar{b}}$ de contre-exemples augmente (avec les autres paramètres constants). Pour une règle $a \xrightarrow{\omega} b$, cela revient à espacer davantage les évènements a et b dans la séquence tout en conservant les mêmes quantités de a et de b. Cette opération fait passer les évènements a d'exemples à contre-exemples.

La figure 4 montre que SII fait clairement la distinction entre un nombre de contre-exemples acceptable (associé à des valeurs d'SII proches de 1) et un nombre de contre-exemples non-acceptable (associé à des valeurs proches de 0) au regard des autres paramètres n_a, n_b, ω, et L. Au contraire, la confiance et le rappel varient linéairement, tandis que la J-mesure produit des valeurs très peu discriminantes. A cause de sa nature entropique, la J-mesure peut même augmenter quand le nombre de contre-exemples augmente, ce qui est gênant pour une mesure de qualité de règles.

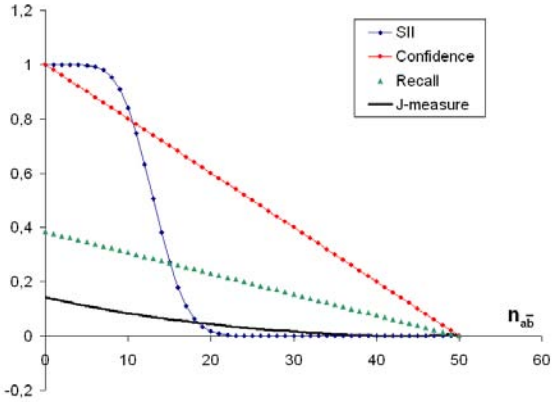


FIG. 4 – *SII, confiance, rappel, et J-mesure en fonction du nombre de contre-exemples.* $n_a = 50$, $n_b = 130$, $\omega = 10$, $L = 1000$

3.2 Allongement de la séquence

Nous désignons par allongement de la séquence l'opération qui consiste à rendre la séquence plus longue en y ajoutant de nouveaux évènements (évènements de nouveaux types) au début ou la fin. Pour une règle $a \xrightarrow{\omega} b$, une telle opération ne modifie pas les effectifs $n_{ab}(\omega)$ et $n_{a\bar{b}}(\omega)$ puisque la répartition des évènements de types a et b reste inchangée. Seule la longueur L de la séquence augmente.

La figure 5 montre que SII augmente avec l'allongement de la séquence. En effet, pour un nombre donné de contre-exemples, une règle est plus surprenante dans une séquence longue plutôt que dans une séquence courte, puisqu'il est moins probable que les a et b soient proches dans une séquence longue. Au contraire, des mesures comme la confiance et le rappel demeurent inchangées car elles ne tiennent pas compte de L (voir figure 6). La J-mesure varie avec L mais faiblement. Elle peut même décroître avec L, ce qui est contre-intuitif.

3.3 Réplication de la séquence

Nous appelons réplication l'opération qui allonge une séquence en la répétant γ fois successivement (nous faisons abstraction des éventuels effets de bord qui pourraient faire apparaître de nouvelles occurrences de motifs à cheval sur la fin d'une séquence et le début de la séquence répétée qui suit). Avec cette opération, les fréquences des évènements a et b et les fréquences des exemples et contre-exemples restent les mêmes.

La figure 7 montre que les valeurs de SII deviennent plus extrêmes (proches de 0 ou 1) avec la réplication. Ce phénomène s'explique par la nature statistique de la mesure. Une règle est en effet d'autant plus significative qu'elle est évaluée sur une séquence longue avec de nombreux évènements : plus la séquence est longue, plus on peut se fier aux déséquilibres entre exemples et contre-exemples observés dans la séquence, et plus on peut confirmer la bonne ou

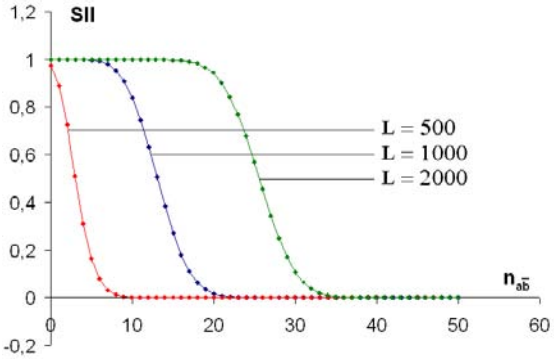


FIG. 5 – *Evolution de SII avec l'allongement de la séquence.* $n_a = 50$, $n_b = 130$, $\omega = 10$

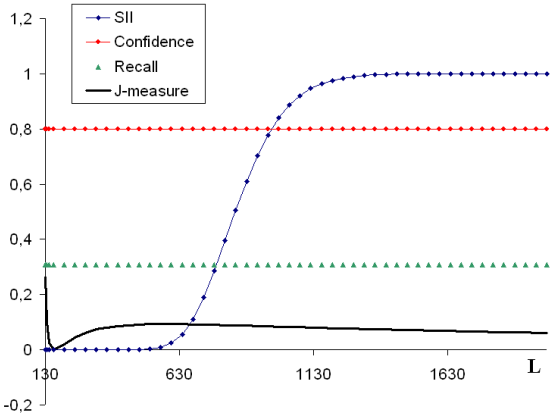


FIG. 6 – *Evolutions de SII, confiance, rappel, et J-mesure avec l'allongement de la séquence.* $n_a = 50$, $n_b = 130$, $n_{a\bar{b}} = 10$, $\omega = 10$

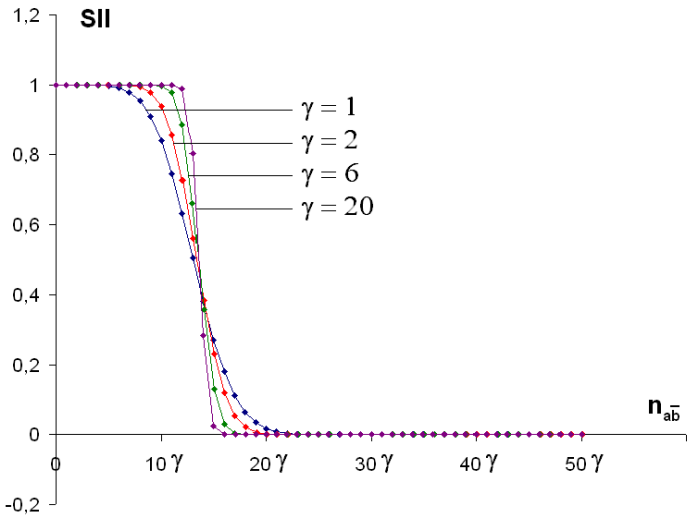
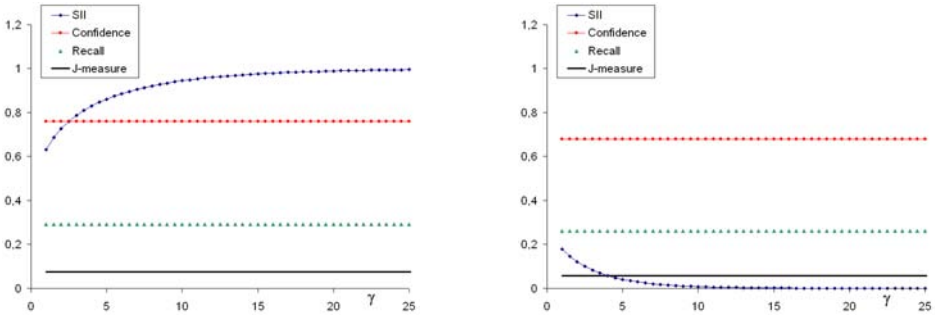


FIG. 7 – *Evolution de SII avec la réplication de la séquence.*
$n_a = 50 \times \gamma$, $n_b = 130 \times \gamma$, $\omega = 10$, $L = 1000 \times \gamma$



(a) $n_{a\bar{b}} = 12 \times \gamma$ (b) $n_{a\bar{b}} = 16 \times \gamma$

FIG. 8 – *Evolutions de SII, confiance, rappel, et J-mesure avec la réplication de la séquence.*
$n_a = 50 \times \gamma$, $n_b = 130 \times \gamma$, $\omega = 10$, $L = 1000 \times \gamma$

mauvaise qualité de la règle. Au contraire, les mesures fréquentielles comme la confiance, le rappel, et la J-mesure ne varient pas avec la réplication (voir la figure 8).

4 Conclusion

Dans cet article, nous avons étudié l'évaluation de la qualité des règles séquentielles. Tout d'abord, nous avons formalisé les notions de *règle séquentielle*, *exemple d'une règle*, et *contre-exemple d'une règle*. Nous avons ensuite présenté l'*Intensité d'Implication Séquentielle (SII)*, une mesure statistique originale qui évalue la significativité des règles séquentielles au regard d'un modèle probabiliste. Les simulations numériques montrent que *SII* que a des caractéristiques uniques en comparaison aux autres mesures de qualité de règles séquentielles. En particulier, SII est la seule mesure qui prenne en compte l'allongement de la séquence et la réplication de la séquence de manière appropriée.

Références

Agrawal, R. et R. Srikant (1995). Mining sequential patterns. In *Proceedings of the international conference on data engineering (ICDE)*, pp. 3–14. IEEE Computer Society.

Blanchard, J., F. Guillet, et H. Briand (2002). L'intensité d'implication entropique pour la recherche de règles de prédiction intéressantes dans les séquences de pannes d'ascenseurs. *Extraction des Connaissances et Apprentissage 1*(4), 77–88. Actes des journées Extraction et Gestion des Connaissances (EGC) 2002.

Blanchard, J., F. Guillet, et R. Gras (2007). On the discovery of significant temporal rules. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics SMC'2007*, pp. 443–450. IEEE Computer Society.

Blanchard, J., F. Guillet, et P. Kuntz (2009). Semantics-based classification of rule interestingness measures. In Y. Zhao, C. Zhang, et L. Cao (Eds.), *Post-Mining of Association Rules : Techniques for Effective Knowledge Extraction*, pp. 56–79. IGI Global.

Das, G., K.-I. Lin, H. Mannila, G. Renganathan, et P. Smyth (1998). Rule discovery from time series. In R. Agrawal, P. E. Stolorz, et G. Piatetsky-Shapiro (Eds.), *Proceedings of the fourth ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 16–22. AAAI Press.

Gras, R. (1996). *L'implication statistique : nouvelle méthode exploratoire de données*. La Pensée Sauvage Editions. in French.

Han, J., J. Pei, et X. Yan (2005). Sequential pattern mining by pattern-growth : Principles and extensions. In W. W. Chu et T. Y. Lin (Eds.), *Recent Advances in Data Mining and Granular Computing (Mathematical Aspects of Knowledge Discovery)*, pp. 183–220. Springer-Verlag.

Höppner, F. (2002). Learning dependencies in multivariate time series. In *Proceedings of the ECAI'02 workshop on knowledge discovery in spatio-temporal data*, pp. 25–31.

Joshi, M., G. Karypis, et V. Kumar (1999). A universal formulation of sequential patterns. Technical report, University of Minnesota. TR 99-021.

Mannila, H. et H. Toivonen (1996). Discovering generalized episodes using minimal occurrences. In *Proceedings of the second ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 146–151. AAAI Press.

Mannila, H., H. Toivonen, et A. I. Verkamo (1995). Discovering frequent episodes in sequences. In *Proceedings of the first ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 210–215. AAAI Press.

Mannila, H., H. Toivonen, et A. I. Verkamo (1997). Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery 1*(3), 259–289.

Ross, S. M. (2006). *Introduction to Probability Models*. 9th edition.

Spiliopoulou, M. (1999). Managing interesting rules in sequence mining. In *PKDD'99 : Proceedings of the third European conference on principles of data mining and knowledge discovery*, pp. 554–560. Springer-Verlag.

Srikant, R. et R. Agrawal (1996). Mining sequential patterns : generalizations and performance improvements. In *EDBT'96 : Proceedings of the fifth International Conference on Extending Database Technology*, pp. 3–17. Springer-Verlag.

Sun, X., M. E. Orlowska, et X. Zhou (2003). Finding event-oriented patterns in long temporal sequences. In K.-Y. Whang, J. Jeon, K. Shim, et J. Srivastava (Eds.), *Proceedings of the seventh Pacific-Asia conference on knowledge discovery and data mining (PAKDD2003)*, Volume 2637 of *Lecture Notes in Computer Science*, pp. 15–26. Springer-Verlag.

Weiss, G. M. (2002). Predicting telecommunication equipment failures from sequences of network alarms. In *Handbook of knowledge discovery and data mining*, pp. 891–896. Oxford University Press, Inc.

Yang, J., W. Wang, et P. S. Yu (2003). Stamp : On discovery of statistically important pattern repeats in long sequential data. In D. Barbará et C. Kamath (Eds.), *Proceedings of the third SIAM international conference on data mining*. SIAM.

Zaki, M. J. (2001). SPADE : an efficient algorithm for mining frequent sequences. *Machine Learning 42*(1-2), 31–60.

Summary

In this article, we study the assessment of the interestingness of sequential rules (generally temporal rules). This is a crucial problem in sequence analysis since the frequent pattern mining algorithms are unsupervised and can produce huge amounts of rules. While association rule interestingness has been widely studied in the literature, there are few measures dedicated to sequential rules. Continuing with our work on the adaptation of implication intensity to sequential rules, we propose an original statistical measure for assessing sequential rule interestingness. More precisely, this measure named Sequential Implication Intensity (SII) evaluates the statistical significance of the rules in comparison with a probabilistic model. Numerical simulations show that SII has unique features for a sequential rule interestingness measure.

Chapitre 6 : Analyse statistique implicative entre variables vectorielles

Régis Gras*, Raphaël Couturier**

*Equipe Connaissances et Décision, Laboratoire d'Informatique de Nantes Atlantique
Ecole Polytechnique de l'Université de Nantes, UMR 6241
La Chantrerie BP 60601 44306 Nantes cedex
regisgra@club-internet.fr

** Institut Universitaire de Technologie de Belfort,
BP 527, rue E. Gros, 90016 Belfort *cedex*
raphael.couturier@iut-bm.univ.fcomte.fr

Résumé. Nous nous plaçons ici dans le cadre de la méthode d'analyse de données, l'analyse statistique implicative (A.S.I.). A l'instar de ce que nous avons fait pour passer des variables binaires aux variables numériques ou aux variables-intervalles, nous étendons le champ des traitements aux variables à valeurs vectorielles. Nous établissons un indice permettant de mesurer la qualité d'une règle entre variables vectorielles. Nous traitons des exemples portant l'un sur le baccalauréat, l'autre sur l'examen des critères de convergence des économies de l'Union Européenne.

1 Introduction

Les développements théoriques de l'analyse de données offrent des retombées enrichissantes pour l'Extraction de Connaissances. Ses développements et sa vitalité ne sont pas étrangères aux échanges induits. Par exemple, la construction d'indices permettant d'affecter une mesure non symétrique à des règles d'inférence partielle fournit des points d'application à l'extraction et à la représentation de règles d'association imprécises entre attributs binaires décrivant une population. Les démarches fondamentales se ramènent à la prise en compte d'une problématique commune aux deux domaines ; il s'agit de découvrir et de quantifier des règles non symétriques pour modéliser des relations du type *"si a, alors presque b"*. Qu'il s'agisse de réseaux bayésiens (Pearl, 1988), de treillis de Galois (Bernard et Poitrenaud, 1999) ou de fouille de règles (Agrawal et al, 1993) de très nombreuses mesures ont été proposées pour quantifier la pertinence de ces quasi-implications et optimiser leur extraction (par ex. (Hilderman et Hamilton, 1999) ou (Tan et Kumar, 2000)), Des travaux sur la qualité des règles d'association ont permis comparer leurs mesures selon des points de vue subjectifs et objectifs (Lenca et al, 2004). Cependant, à notre connaissance, ces travaux se limitent généralement à l'étude de mesures pour des règles entre attributs binaires ou conjonction de tels attributs.

Or les situations réelles, y compris celles pour lesquelles l'analyse implicative a créé son modèle statistique (la didactique des mathématiques), conduisent au traitement d'autres types

de variables. C'est ainsi que des extensions successives, pour répondre à des applications comme dans (Gras et al. 1996) nous ont conduits à intégrer dans la même théorie, des variables modales, numériques, floues, des variables sur intervalles, des variables-intervalles, des variables-rangs et à élaborer des outils de représentation graphique. Dans cet article, nous étudions, toujours dans le cadre de l'analyse implicative, le cas de variables à valeurs vectorielles.

2 Problématique

On veut comparer l'évolution temporelle d'un système de variables à d dimensions. Par exemple, sans pratiquer de test d'hypothèse mais sans l'exclure, on envisage :

une suite de *d* notes obtenues dans des disciplines, indépendantes a priori, examinée à des périodes régulières (trimestre) sur un ensemble E d'élèves soumis à un traitement expérimental. On cherche à savoir si, globalement, les performances à l'instant *t* impliquent celles obtenues à l'instant *t+1*, *t+2*, etc. et donc si le traitement expérimental est influent ;

des mesures (analyses médicales) de *d* paramètres, indépendants a priori, obtenues à des périodes régulières afin d'étudier l'effet d'un traitement sur une population *E* de patients. On cherche s'il existe des règles d'association entre les moments d'observation et donc s'il existe un effet du traitement ;

on fait l'observation à *d* instants (année, séquence, phase,…) de valeurs de paramètres dont on veut extraire d'éventuelles règles d'évolution d'un moment à un autre ; par exemple, l'évolution comparée de la fonction en entreprise occupée par des individus à *d* instants par rapport aux salaires respectivement octroyés à ces instants.

La représentation de ces suites est vectorielle. A une observation sur un sujet correspond le vecteur des observations de chacune de ses *d* composantes. Ces observations sont quantitatives, de nature binaire, modale ou intervalle. Une règle d'association entre deux vecteurs aura d'autant plus de sens que les composantes vectorielles présenteront une sémantique commune (par exemple : «progrès », « régression », etc.). Elle se substituera d'autant plus avantageusement à l'analyse implicative classique entre les composantes, modalités de chaque vecteur, que celles-ci présenteront des caractéristiques indépendantes dans leur essence même.

Mais les deux approches peuvent être complémentaires et informatives : l'analyse vectorielle est globale, l'analyse classique est ponctuelle. La première vise, en quelque sorte, à partir de profils individuels (la suite des composantes vectorielles) à dégager des règles d'association entre profils synthétiques. La seconde permet de dégager des règles d'association entre attributs binaires ou non, élémentaires ou composites (obtenu par la conjonction d'attributs élémentaires). Mais un attribut composite n'est pas un profil, il n'en est que la contraction. Ainsi, l'analyse vectorielle ne se ramène pas à l'analyse classique.

3 Cas de vecteurs à composantes binaires

3.1 Détermination d'un indice « vectoriel »

Les situations précédentes relèvent de variables numériques qu'il est possible de garder comme telles mais qu'il est aussi possible de binariser en ramenant chaque valeur à sa position par rapport à une norme : au-dessus, au-dessous. Mais d'autres situations peuvent se présenter sous une forme immédiatement binaire : par ex., une épreuve composée de d modalités où les observations seraient réussite-échec, présence-absence, etc. [1].

On cherche donc à comparer des vecteurs de type $\vec{a} = (a_1, a_2, \ldots, a_d)$, vecteur représentatif d'une **variable vectorielle** $\vec{a}$, et $\vec{b} = (b_1, b_2, \ldots, b_d)$, vecteur représentatif d'une variable vectorielle $\vec{b}$. On veut en extraire par exemple la règle $\vec{a} \Rightarrow \vec{b}$. Pour ce faire, on associe à chaque observation selon n sujets,

- d'une part , un vecteur à *d* dimensions, associé à la variable $\vec{a}$, de la somme sur les *n* sujets de chacune des *d* composantes, c'est-à-dire le vecteur-ligne : $\vec{A} = (\sum_{i=1}^{i=n} a_1(i), \sum_{i=1}^{i=n} a_2(i), \ldots, \sum_{i=1}^{i=n} a_d(i))$. $\vec{A}$ admet ainsi pour j^{ème} composante scalaire le nombre $\sum_{i=1}^{i=n} a_j(i)$. Ce nombre représente donc le nombre de fois (*cas binaire*) où apparaît la variable composante a_j (ou la somme de ses pondérations dans le *cas numérique*) dans l'ensemble de la population. Soit aussi $\vec{A}(i)$ le vecteur-ligne des *d* composantes du sujet *i* de la forme ($a_1(i)$, $a_2(i)$,… $a_j(i)$,… $a_d(i)$). $a_j(i)$ est la j^{ème} composante du vecteur $\vec{A}(i)$, c'est-à-dire la valeur prise par le sujet *i* selon la variable composante a_j. On fait de même avec la variable $\vec{b}$ afin d'obtenir le vecteur-ligne $\vec{\mathbf{B}}_i$;

- d'autre part, deux ensembles aléatoires de vecteurs-lignes, choisis au hasard et indépendamment l'un de l'autre, dont les sommes vectorielles-lignes $\vec{X}$ et $\vec{Y}$ respectives coïncident exactement et respectivement, composante par composante, avec celles de $\vec{A}$ et de $\vec{B}$. Par analogie avec le cas des variables *a* et *b* uniques (une seule composante), ces deux ensembles sont à comparer aux ensembles *X* et *Y* de sujets, de même cardinaux que ceux de *A* et *B* supports des variables respectives *a* et *b*. Mais, bien entendu, ce ne sont pas généralement, les mêmes sujets qui les ont constitués.

A ces vecteurs $\vec{A}$ et $\vec{B}$, on associe le vecteur $\vec{w}$ des contre-exemples observés aux implications successives selon les vecteurs-sujets $\vec{\mathbf{A}}(\mathbf{i})$ et $\vec{\mathbf{B}}(\mathbf{i})$, donc d'indice identique. Par exemple, un contre-exemple en *i* à $a_j \Rightarrow b_j$ apparaît lorsque $a_j(i) = 1$ alors que $b_j(i) = 0$, c'est-à-dire lorsque $a_j(i) \wedge \bar{b}_j(i) = 1$. Par suite, le vecteur des contre-exemples observés aura pour

[1] S'il y a nécessité d'étendre la méthodologie employée dans le cas d'une variable binaire, la généralisation au cas de variable numérique se fera comme dans le cas de la composante unique en utilisant la valeur corrigée des indices définis dans (Lagrange, 1998) ou (Régnier et Gras,,2005).

composantes scalaires les *j* scalaires de la forme $\sum_{i=1}^{i=n} a_j(i) \wedge \bar{b}_j(i)$ que nous notons, comme dans le cas de la variable unique $n_{a_j \wedge \bar{b}_j}$. Aux vecteurs aléatoires $\vec{X}$ et $\vec{Y}$, on associe de la même façon, le vecteur $\vec{W}$ des contre-exemples aléatoires de composantes $\sum_{i=1}^{i=n} x_j(i) \wedge \bar{y}_j(i)$ notés également $N_{a_j \wedge \bar{b}_j}$.

Or, $N_{a_j \wedge \bar{b}_j}$ suit, en tant que variable aléatoire relative à la règle présupposée $a_j \Rightarrow b_j$, et comme nous l'avons prouvé par ailleurs (Lerman et al1981) et Partie 1, la loi de Poisson de paramètre $\frac{n_{a_j}}{n}. n_{\bar{b}_j}$ ou bien la loi binomiale de paramètres n et $\frac{n_{a_j}}{n}. \frac{n_{\bar{b}_j}}{n}$. Alors, du fait de l'indépendance des variables composantes, la probabilité pour que les nombres de tous les contre-exemples aléatoires ne soient pas inférieurs aux nombres de contre-exemples respectivement observés est le produit des probabilités pour que cette propriété soit vérifiée sur chaque composante, à savoir :

$$\prod_{j=1}^{j=d} \left[1 - \Pr[N_{a_j \wedge \bar{b}_j} \leq n_{a_j \wedge \bar{b}_j}] \right]$$

On définit alors l'intensité d'implication de la règle $\vec{a} \Rightarrow \vec{b}$ par [2]

$$\varphi(\vec{a}, \vec{b}) = \left(\prod_{j=1}^{j=d} \left[1 - \Pr[N_{a_j \wedge \bar{b}_j} \leq n_{a_j \wedge \bar{b}_j}] \right] \right)^{\frac{1}{d}}$$

Par suite, compte tenu de la définition de l'indice classique de l'intensité d'implication d'une variable vers une autre : $\varphi(\vec{a}, \vec{b}) = \left[\prod_{j=1}^{j=d} \varphi(a_j, b_j) \right]^{\frac{1}{d}}$

Le logiciel C.H.I.C. (Couturier et Gras. 2005) et chapitre 11 permet le calcul des intensités d'implication dans le cas vectoriel. Nous en aurons deux applications plus loin.

3.2 Remarques

On peut comparer cette extension à la situation originelle où l'on a, pour une variable unique, un vecteur à une dimension. Dans le cas vectoriel présenté ici, une quasi-inclusion du vecteur représentatif de *a* dans un parallélépipède de dimension *d* définie par sa diagonale $\vec{b}$ se substitue à la quasi-inclusion de l'ensemble A dans l'ensemble B, supports respectifs de *a* et *b*.

[2] Une formule acceptable et alternative de celle-ci est . $1 - \prod_{j=1}^{j=d} \left[\Pr[N_{a_j \wedge \bar{b}_j} \leq n_{a_j \wedge \bar{b}_j}] \right]^{1/d}$

Toutes les formules où figure l'intensité d'implication $\varphi(\vec{a},\vec{b})$ peuvent être réécrites en remplaçant φ par ψ qui symbolise l'implication-inclusion sur des bases entropiques (Gras et al, 2001).

De même, ces formules peuvent être étendues au cas des variables numériques selon la transformation établie par (Lagrange, 1998), qui modifient l'écart-type de la loi de Poisson. On pourrait s'intéresser à l'événement où *k* composantes seulement parmi les *d* satisferaient l'inégalité entre la valeur aléatoires des contre-exemples et la valeur observée. Ce qui correspond à une exigence affaiblie quant à la règle $\vec{a} \Rightarrow \vec{b}$. Pour cela, il suffit d'envisager toutes les parties à *k* éléments parmi $\{(a_1,b_1),(a_2,b_2),....,(a_d,b_d)\}$, de faire la moyenne arithmétique des sommes des produits correspondants dans le crochet donnant $\varphi(\vec{a},\vec{b})$ et d'en prendre la puissance correspondante.

Par exemple, si *d=3* et *k=2*, on aurait :

$\varphi(\vec{a},\vec{b})=\{\frac{1}{3}[\varphi(a_1,b_1).\ \varphi(a_2,b_2).+\ \varphi(a_2,b_2).\ \varphi(a_3,b_3)+\ \varphi(a_3,b_3).\ \varphi(a_1,b_1)]\}^{1/2}$

Le respect de la sémantique quant au jugement associé à la mise en évidence d'une implication peut être est délicat. Il est loisible, pour certaines des variables composantes de la prémisse ou de la conclusion, de changer l'ordre défini a priori sur les valeurs (binaires ou non) que prennent ces composantes. Ce changement viserait à optimiser l'intensité d'implication $\varphi(\vec{a},\vec{b})$. Par exemple, on pourrait permuter les valeurs *0* et *1* dans le cas binaire ou, de façon générale, passer à la complémentation à *1* de chaque valeur.

3.3 Exemple 1

Une population est constituée de *85* Sujets, dont certains possèdent les mêmes caractéristiques que leur prototype. Ainsi on observe, dans le tableau 1, *20* Sujets répondant *0* en a_1 et de même selon b_1, *1* en a_2 et b_2

	$\vec{a}$		$\vec{b}$	
	a_1	a_2	b_1	b_2
20Sujets1	0	1	0	1
30Sujets2	0	0	1	0
30Sujets3	1	0	1	1
5Sujets4	1	1	0	0
Totaux	**35**	**25**	**60**	**50**

TAB. 1

On observe : $n_{a_1\wedge\bar{b}_1} = n_{a_2\wedge\bar{b}_2} = 5$. En utilisant le modèle de Poisson, on obtient en se limitant au calcul par l'intensité d'implication classique (et non pas « entropique ») :

$$\Pr[N_{a1\wedge\bar{b}1}\leq 5] = 0.05 = \Pr[N_{a2\wedge\bar{b}2}\leq 5] \text{ d'où } \varphi(\vec{a},\vec{b})=(0.95\text{x}0.95)^{1/2}= 0.95$$

4 Cas de vecteurs à composantes numériques

4.1 Indices vectoriels

Nous avons établi (Lagrange, 1998) un indice (Partie 1 Chap. 3) respectant l'approche statistique du cas binaire aux cas où :

1. les variables sont modales (degré de possession d'un attribut ou d'adéuation à celui-ci),
2. les variables sont numériques (nombre d'occurrences, valeurs de contingence)

L'indice adopté permet d'accorder une mesure à l'implication entre de telles variables. De plus, cet indice restreint au cas binaire coïncide avec l'indice mesurant traditionnellement la règle où prémisses et conclusions sont binaires.

Si a et b sont deux variables numériques, dont les valeurs observées respectivement $\{a_i\}_{i \in E}$ et $\{b_i\}_{i \in E}$ dans une population *E* de taille *n* et dont les moyennes et variances respectives sont m_a, m_b, s_a, s_b, alors l'indice de base est :

$$\tilde{q}(a,\bar{b}) = \frac{\sum_{i \in E} a_i \bar{b}_i - \frac{m_a m_{\bar{b}}}{n}}{\sqrt{\frac{(n^2 s_a^2 + m_a^2)((n^2 s_{\bar{b}}^2 + m_{\bar{b}}^2)}{n^3}}}$$

que nous avons aussi nommé, indice de propension.

et l'intensité de la règle $\vec{a} \Rightarrow \vec{b}$ est dans une approximation gaussienne :

$$\varphi(a,b) = 1 - \Pr\left[Q(a,\bar{b}) \le q(a,\bar{b})\right] = \frac{1}{\sqrt{2\pi}} \int_{q(a,\bar{b})}^{\infty} e^{-\frac{t^2}{2}} dt$$

expression dans laquelle $Q(a,\bar{b})$ est la valeur centrée réduite de $N_{a \wedge \bar{b}}$, pour l'une des lois retenues dans le modèle : binomiale ou de Poisson ou gaussienne.

Considérant maintenant des vecteurs de type $\vec{a} = (a_1, a_2, \ldots, a_d)$, vecteur représentatif d'une variable vectorielle $\vec{a}$, et $\vec{b} = (b_1, b_2, \ldots, b_d)$, vecteur représentatif d'une variable vectorielle $\vec{b}$ où $a_1, a_2, \ldots, a_d$ et $b_1, b_2, \ldots, b_d$ sont *2d* variables numériques. L'intensité d'implication de la règle $\vec{a} \Rightarrow \vec{b}$ est encore : $\varphi(\vec{a}, \vec{b}) = \left[\prod_{j=1}^{j=d} \varphi(a_j, b_j)\right]^{1/d}$

4.2 Exemple 2

Nous disposons des résultats en 1997[3] aux *3* baccalauréats (Scientifique S, Littéraire L et Economique et Social ES) dans les *26* académies de France, résultats exprimés en pourcentages d'admis définitifs. Ces résultats figurent dans le TAB 2 exprimés en ‰.

[3] Références extraites du QUID 99

Académie	S	L	ES	Académie	S	L	ES
Aix	760	762	714	Montpellier	777	742	734
Amiens	753	725	715	Nancy	752	765	795
Besançon	758	767	785	Nantes	781	815	807
Bordeaux	726	779	761	Nice	750	737	735
Caen	718	720	746	Orléans	787	779	817
Clermont	730	805	799	Paris	787	754	743
Corse	794	779	702	Poitiers	765	798	795
Créteil	712	729	680	Reims	700	762	765
Dijon	757	757	783	Rennes	779	810	803
Grenoble	789	795	821	Rouen	785	765	743
Lille	761	716	742	Strasbourg	783	799	822
Limoges	719	709	742	Toulouse	815	830	775
Lyon	788	770	798	Versailles	813	823	798

TAB. 2

Une première analyse à l'aide de C.H.I.C. des règles entre les trois variables numériques S, L et ES nous conduit à constater l'absence de relation à un seuil acceptable à partir duquel une certaine dépendance existerait (>0 ,50). Une nouvelle voie s'ouvre alors : peut-**on dire que l'ordre des succès définitifs dans l'une des séries « implique » l'ordre dans une** autre et dans laquelle ? Pour tenter de répondre à cette question, il nous suffit de subdiviser chaque intervalle de la valeur minimale à la valeur maximale donnée aux 26 académies par le tableau en sous-intervalles. Nous adoptons la technique des nuées dynamiques de E. Diday qui permet, en particulier, de maximiser la variance inter-classe des sous-intervalles créés.

On obtient, à l'aide du logiciel C.H.I.C., les partitions optimales suivantes pour des partitions en 3 sous-intervalles :

S1 de 700 à 730	L1 de 709 à 742	ES1 de 680 à 746
S2 de 750 à 765	L2 de 754 à 779	ES2 de 761 à 785
S3 de 777 à 815	L3 de 795 à 830	ES3 de 795 à 822

TAB. 3

Puis utilisant la version vectorielle de CHIC, on extrait les implications suivantes :

$\vec{S}$ -> $\vec{L}$ avec l'intensité 0.63339 $\vec{L}$ -> $\vec{S}$ avec l'intensité 0.65041

$\vec{S}$ -> E$\vec{S}$ avec l'intensité 0.55874 E$\vec{S}$ -> $\vec{S}$ avec l'intensité 0.56572

$\vec{L}$ -> E$\vec{S}$ avec l'intensité 0.88015 E$\vec{S}$ -> $\vec{L}$ avec l'intensité 0.91037

que l'on peut synthétiser ainsi : $E\vec{S} \overset{0.91}{\Rightarrow} \vec{L}$, $\vec{L} \overset{0.65}{\Rightarrow} S$ et $E\vec{S} \overset{0..57}{\Rightarrow} \vec{S}$ ou par FIG 1 :

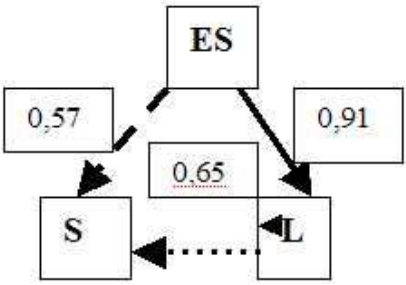


FIG. 1 – *Graphe implicatif des types de baccalauréats*

Ainsi, si les résultats au bac ES sont bons, alors ceux en L ont une propension à être meilleurs et, de façon moins décisive, ceux de S également. De plus, les résultats dans chaque académie en S ont tendance à être meilleurs, relativement à la tendance générale, à ceux de L et à en être prédicteurs dans l'ordre où ils apparaissent. Autrement dit, les académises où les résultats sont relativement faibles (resp. forts) en S, le sont aussi généralement en L et en ES.

On voit alors l'information complémentaire qu'apporte la méthode implicative par rapport au seul examen de la corrélation, voire de l'implication entre les colonnes. Elle dit dans quel sens et avec quelle intensité mesurable joue l'association entre les séries de baccalauréat.

4.3 Exemple 3

Nous disposons de « critères de convergence » entre États de l'Union européenne au 25-3-98, critères qui permettaient de comparer les économies des pays et d'en définir leur capacité à respecter la politique monétaire commune conditionnant l'ouverture à l'euro. Ces critères s'expriment selon plusieurs composantes : l'inflation, le déficit public (% du PIB), la dette publique (% du PIB) et le taux d'intérêt à long terme. Le tableau ci-dessous [4]donne les différentes valeurs observées sur 3 années : 1996, 1997 et 1998 sur les 15 pays de l'Union.

Nous avons encore affaire à des variables numériques : l'une pour 1996, une autre pour 1997, une troisième pour 1998. Les valeurs indiquées TAB. 4 ont été toutes ramenées à des nombres de l'intervalle [0,1] par division de chaque colonne par le maximum de la colonne. Seul le déficit a nécessité de translater les valeurs de telle façon que nous n'ayons que des nombres positifs, alors même que dans la plupart des pays, le déficit était toujours négatif.

[4] Tableau extrait du QUID 99, avec estimation des taux d'intérêt de la Grèce pour les trois années en question.

	96Infl	96défic	96dette	96Taux	97Infl	97défic	97dette	97Taux	98Infl	98défic	98dette	98Taux
Allemagne	0,22	0,41	0,48	0,62	0,35	0,23	0,5	0,7	0,38	0,1	0,52	0,75
Autriche	0,29	0,35	0,55	0,63	0,33	0,26	0,54	0,71	0,33	0,15	0,55	0,75
Belgique	0,27	0,43	1	0,65	0,29	0,33	1	0,73	0,29	0,31	1	0,76
Danemark	0,25	0,68	0,56	0,72	0,42	0,83	0,53	0,79	0,47	1	0,5	0,75
Espagne	0,4	0,29	0,55	0,87	0,46	0,25	0,56	0,8	0,49	0,18	0,57	0,84
Finlande	0,19	0,42	0,45	0,71	0,26	0,54	0,46	0,75	0,44	0,82	0,45	0,73
France	0,22	0,34	0,44	0,63	0,2	0,18	0,48	0,7	0,22	0	0,49	0,73
Grèce	1	0	0,88	1	1	0	0,89	1	1	0,18	0,91	1
Irlande	0,13	0,71	0,57	0,73	0,26	0,86	0,54	0,79	0,73	1	0,5	0,83
Italie	0,51	0,08	0,98	0,94	0,44	0,23	1	0,86	0,47	0,1	1	0,89
Luxembourg	0,19	1	0,05	0,63	0,26	1	0,06	0,7	0,36	1	0,06	0,75
Pays-Bas	0,15	0,52	0,61	0,62	0,4	0,46	0,59	0,7	0,51	0,33	0,59	0,73
Portugal	0,31	0,43	0,51	0,86	0,38	0,26	0,51	0,8	0,49	0,18	0,51	0,83
Royaume Uni	0,31	0,27	0,43	0,79	0,42	0,37	0,44	0,89	0,51	0,59	0,44	0,93
Suède	0,14	0,4	0,6	0,8	0,4	0,56	0,63	0,95	0,33	0,87	0,63	0,87

TAB. 4

Un premier traitement par C.H.I.C. des 12 variables numériques (4 pour chacune des 3 années) conduit au graphe implicatif suivant construit aux seuils .85 à .70:

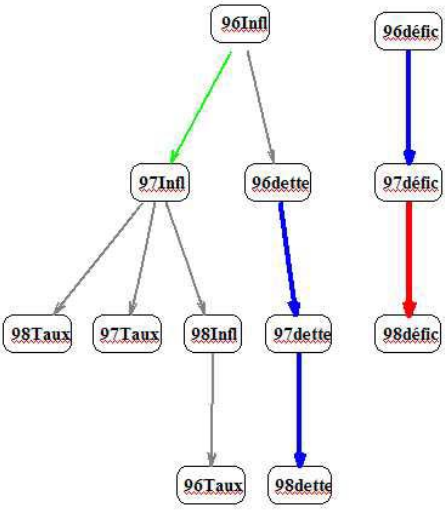


FIG. 2 – *Graphe implicatif des critères de convergence*

On observe une indépendance des 3 critères : inflation, dette et taux d'intérêt par rapport au critère déficit. Sur le plan économique, et cela paraît surprenant, le déficit n'aurait pas d'influence sur la dette et l'inflation. On constate, en revanche, la liaison attendue entre l'inflation et les taux d'intérêt, l'ensemble se présentant relativement peu lié à la dette.

On observe également qu'alors que déficit, inflation et dette croissent de 96 à 98, comme on pouvait l'attendre, les taux d'intérêt diminuent pendant cette période. Nous y reviendrons.

Arrêtant ici l'exploitation de ce graphe sur lequel bien d'autres considérations peuvent être énoncées, nous constatons que les natures des critères, bien qu'en partie liées, ne rendent pas bien compte du comportement global des économies de 96 à 98. D'où l'intérêt de considérer 3 variables vectorielles numériques : A code l'année 1996, présente les composantes A1, A2, A3 et A4 respectivement correspondant aux 4 critères, puis B code 1997, avec les composantes B1 à B4, et C code 1998 avec les composantes C1 à C4.

Le logiciel C.H.I.C. dégage les intensités d'implication suivantes :

$\varphi(A \Rightarrow B) = 0,71$; $\varphi(A \Rightarrow C)=0,70$; $\varphi(B \Rightarrow C)= 0,74$,

les intensités des relations réciproques étant plus faibles. Les trois variables s'enchaînent donc transitivement, au seuil 0,70, confirmant la croissance globale des valeurs des critères avec le temps, croissance freinée nécessairement par la décroissance des taux d'intérêt dont cette approche ne rend plus compte. Cette restriction illustre le propos tenu dans le dernier paragraphe du **§ 2**. La relation d'ordre dérivée de la sémantique des critères-composantes devrait être la même. Or pour les trois critères : inflation, déficit et dette, la sémantique de ces variables oriente l'évolution du critère vers la croissance, alors que pour le taux, dans les données, l'évolution est décroissante. Si nous complémentons, alors, à 1 chaque valeur du taux énoncé, nous obtenons des intensités d'implication supérieures aux précédentes, ce qui confirme le bien-fondé de la remarque au sujet des sémantiques d'ordre :

$\varphi(A \Rightarrow B) = 0,78$; $\varphi(A \Rightarrow C)=0,77$; $\varphi(B \Rightarrow C)= 0,79$,

5 Conclusion

Nous avons étendu la variété des variables prises en compte par la méthode d'analyse statistique implicative en conceptualisant la relation implicative entre deux variables vectorielles. De telles variables peuvent présenter des modalités binaires ou numériques représentées par les composantes vectorielles. Cette souplesse permet de dégager et quantifier une règle d'association dissymétrique entre deux vecteurs en exprimant par une mesure la dynamique inhérente aussi bien à l'ensemble des variables vectorielles qu'à celle de leurs modalités. Des exemples simples, traités à la main et des exemples réels traités par CHIC illustrent la construction théorique de l'extension de l'A.S.I. aux vecteurs.

Références

Agrawal, R., T. Imielinski. Et A. Swami (1993). Mining association rules between sets of items in large databases. In the1993 ACM SIGMOD international conference on management of data, ACM Press

Bernard J.-M., Poitrenaud S., (1999)L'analyse implicative bayesienne d'un questionnaire binaire : quasi- implications et treillis de Galois simplifié, *Mathématiques, Informatique et Sciences Humaines*, 147, 1999, 25-46

Couturier, R. et Gras R.,(2005) : CHIC : Traitement de données avec l'analyse implicative, *Extraction et Gestion des Connaissances, Volume I1, RNTI, Cepadues, Paris, p.679-684, , ISBN 2.85428.683.9*

Gras R. et al.(1996) *L'implication Statistique*, Grenoble, La Pensée Sauvage

Gras R., Kuntz P., Briand H., (2001) Les fondements de l'analyse statistique implicative et leurs prolongements pour la fouille de données , *Mathématiques et Sciences Humaines*, 154-155, 2001, 9-29.

Hilderman R.J., Hamilton H.J., (1999) Heuristics measures of efferestingness , *Proc. of the 3rd Eur. Conf. on Principles of Data Mining and Knowledge Discovery*, Lect. N. in Art. Int., 1999, p. 232-241.

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, I.H.P., 71-93

Lenca P., Meyer P., Vaillant B., Picouet P., Lallich S., (2004) Evaluation et analyse multi-critères des mesures de qualité des règles d'association, *Mesures de qualité pour la fouille de données, Collection RNTI-E-1, Cepaduès,219-246.*

Lerman, I.C., Gras R. et Rostam H., (1981) Elaboration et évaluation d'un indice d'implication pour données binaires , Mathématiques et Sc. Humaines, n°74, p. 5-35.

Pearl J., *Probabilistic Reasoning in intelligent systems*, San Mateo, CA, Morgan Kaufmann.

Régnier J.C. et R. Gras (2005). Statistique de rangs et analyse statistique implicative. *Revue de Statistique Appliquée*. 53 (1) : 5-38

Tan P., Kumar V., (2000) Interestingness measures for association patterns: a perspective, *Technical Report TR00-036, University of Minnesota, 2000.*

Summary

This text deals with statistical implicative analysis. As we have done for extending binary variables to numerical variables or interval-variables, we can extend the possible use of vectorial variables. We establish an index allowing us to measure the quality of a rule between vectorial variables. An exemple relative to first college degree and another concerning convergence criteria of European Union's economy are given.

.

Chapitre 7 : Arbre de décision pour données déséquilibrées : sur la complémentarité de l'intensité d'implication et de l'entropie décentrée

Gilbert Ritschard*, Simon Marcellin**, Djamel A. Zighed**

*Département d'économétrie, Université de Genève
**Laboratoire ERIC, Université de Lyon 2
gilbert.ritschard@unige.ch, {abdelkader.zighed,simon.marcellin}@univ-lyon2.fr
http://mephisto.unige.ch, http://eric.univ-lyon2.fr

Résumé. Cet article porte sur l'induction d'arbres de classification pour des données déséquilibrées, c'est-à-dire lorsque certaines catégories de la variable à prédire sont beaucoup plus rares que d'autres. Plus particulièrement nous nous intéressons à deux aspects: d'une part, à définir des critères de construction de l'arbre qui exploitent efficacement la nature déséquilibrée des données, et d'autre part la pertinence de la conclusion à associer aux feuilles de l'arbre. Nous avons récemment abordé cette problématique sous deux angles indépendants: l'un était axé sur le recours à des entropies décentrées, l'autre s'appuyant sur des mesures d'intensités d'implication issues de l'ASI. Nous nous proposons ici de comparer et d'établir les similarités entre ces deux approches. Une première expérimentation sommaire est présentée.

1 Introduction

Qu'il s'agisse d'induire un arbre, ou d'associer une conclusion à chacune de ses feuilles, les critères utilisés supposent en général implicitement une importance égale des modalités de la variable à prédire. Ainsi, des algorithmes comme CART (Breiman et al., 1984) ou C4.5 (Quinlan, 1993) utilisent comme critère l'amélioration d'une entropie classique, c'est-à-dire centrée sur la distribution uniforme correspondant à l'équiprobabilité des modalités. Le résultat est qu'on obtient ainsi des segmentations en classes dont les distributions tendent à s'écarter le plus possible de la distribution uniforme. De même pour le choix de la conclusion, le critère communément utilisé est simplement la règle majoritaire qui n'a évidemment de sens que si chaque modalité a la même importance. On le voit donc, cette distribution égalitaire des modalités joue le rôle de situation la moins désirable. Mais est-ce vraiment le cas ? Et sinon, de quelles solutions dispose-t-on pour d'une part favoriser les écarts à une distribution non centrée — représentative de la situation la moins désirable — et d'autre part choisir la conclusion la plus pertinente par rapport à cette référence la moins désirable ?

Une première solution nous est fournie par l'indice d'implication dont nous avons montré dans Ritschard (2005) et Pisetta et al. (2007) comment il pouvait s'utiliser avec les arbres de décision. En effet, cet indice est en fait un résidu, soit un écart par rapport à l'indépendance qui

est caractérisée dans les arbres par la distribution au nœud initial. Ainsi au lieu de mesurer des écarts par rapport à la distribution uniforme, on mesure des écarts par rapport à cette distribution initiale. Rien n'empêche cependant de considérer des résidus par rapport à d'autres distributions. Voir à ce sujet l'indice d'écart à l'équilibre de Blanchard et al. (2005) et sa généralisation dans Lallich et al. (2005). Une seconde solution consiste à utiliser des entropies décentrées (Marcellin et al., 2006; Zighed et al., 2007; Lenca et al., 2008) qui généralisent les entropies classiques en les paramétrant par le point où elles prennent leur maximum, laissant ainsi à l'utilisateur la possibilité de déterminer le point d'incertitude maximale.

Nous nous proposons dans ce papier de comparer ces deux approches en discutant leurs avantages respectifs comme critère de construction de l'arbre ainsi que comme critère de choix de la conclusion des règles. Notre discussion nous amènera à proposer une solution hybride où l'on utilise l'entropie décentrée pour induire l'arbre, et l'indice d'implication pour assigner une décision à chaque feuille.

L'article est organisé comme suit. La section 2 pose le cadre formel. Dans La section 3 nous introduisons un jeu de données qui nous servira d'illustration et rappelons le principe des arbres de décision. A la section 4 nous rappelons les définitions introduites dans Ritschard (2005) sur la notion d'indice d'implication dans le contexte des arbres de décision et examinons la possibilité de l'utiliser comme critère d'optimalité pour les éclatements successifs lors de la construction de l'arbre. Nous rappelons aussi son intérêt pour l'attribution de la conclusion aux feuilles de l'arbre. La section 5 quant à elle rappelle la forme de l'entropie décentrée introduite dans Marcellin et al. (2006) et Zighed et al. (2007) et commente son usage, en particulier comme critère de développement de l'arbre. La discussion comparative fait l'objet de la section 6 tandis que la section 7 éclaire le propos avec des résultats d'expérimentations. Enfin nous concluons à la section 8.

2 Cadre formel et notations

On se place dans un cadre supervisé où disposant d'une variable dépendante y, dite aussi variable réponse ou à prédire, on cherche à caractériser une fonction $f(x_1, x_2, \ldots)$ — un arbre de décision dans notre cas — qui permette de prédire y à partir d'un ensemble $x_1, x_2, \ldots$ de variables explicatives (prédicteurs) catégorielles, ordinales ou quantitatives. On s'intéresse ici au cas où la variable réponse est catégorielle avec ℓ modalités $y_1, \ldots, y_\ell$. Par exemple, s'agissant de diagnostiquer un cancer on aura $y_1 =$ 'a le cancer' et $y_2 =$ 'pas de cancer'. Notre propos concerne cependant plus particulièrement les situations où la réponse prend plus de 2 modalités, ce qui est par exemple le cas si l'on retient une catégorie $y_3 =$ 'requiert une analyse supplémentaire' en plus des deux classes précédentes.

Comme y est catégorielle, la prédiction de sa modalité est une *classification*. On assigne un cas j à la classe (modalité) de y que l'on prédit à partir des valeurs $x_{j1}, x_{j2}, \ldots$ que prennent les prédicteurs pour ce cas j.

3 Données illustratives et principe des arbres de décision

Pour illustrer notre propos, nous reprenons les données fictives utilisées dans Ritschard (2005) et récapitulées au tableau 1. La variable à prédire est l'état civil, le sexe et le secteur

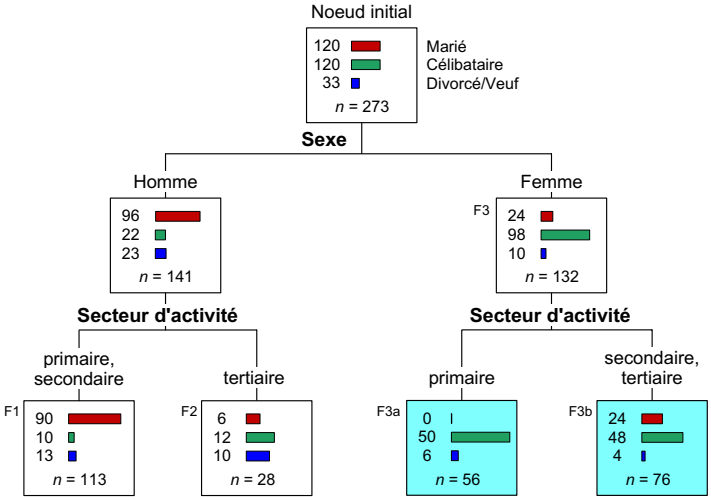


FIG. 1 – *Arbre induit. Feuilles F1, F2, F3 avec indice implication, F1, F2, F3a, F3b avec entropie décentrée.*

d'activité étant les prédicteurs disponibles.

	homme			femme			
état civil	primaire	secondaire	tertiaire	primaire	secondaire	tertiaire	total
marié	50	40	6	0	14	10	120
célibataire	5	5	12	50	30	18	120
divorcé/veuf	5	8	10	6	2	2	33
total	60	53	28	56	46	30	273

TAB. 1 – *Données illustratives.*

Les arbres de classification sont des outils supervisés. Ils déterminent des règles de classification en deux temps. Dans une première étape, une partition de l'espace des prédicteurs (x) est déterminée telle que la distribution de la variable (discrète) à prédire (y, l'état civil dans notre exemple) diffère le plus possible d'une classe à l'autre de la partition. La partition se fait successivement selon les valeurs des prédicteurs. On commence par partitionner les données selon les modalités de l'attribut le plus discriminant, puis on répète l'opération localement sur chaque nœud ainsi obtenu jusqu'à la réalisation d'un critère d'arrêt. Dans un second temps, après que l'arbre ait été généré, on dérive les règles de classification en choisissant la valeur de la variable à prédire la plus pertinente dans chaque feuille (nœud terminal) de l'arbre. On retient classiquement pour cela la valeur la plus fréquente, mais nous reviendrons précisément sur ce point.

Pratiquement, on relève dans chaque feuille j, $j = 1, \ldots, q$, le nombre n_{ij} de cas qui sont

dans l'état y_i. Ainsi, on peut récapituler les distributions au sein des feuilles sous forme d'une table de contingence croisant les états de la variable y avec les feuilles (Tableau 2). On peut noter que la marge de droite de ce tableau qui donne le total $n_{i\cdot}$ des lignes correspond en fait à la distribution des cas dans le nœud initial de l'arbre. Les $n_{\cdot j}$ désignent les totaux des colonnes.

	feuille 1	$\cdots$	feuille j	$\cdots$	feuille q	Total
y_1						$n_{1\cdot}$
$\vdots$						$\vdots$
y_i			n_{ij}			$n_{i\cdot}$
$\vdots$						$\vdots$
y_ℓ						$n_{\ell\cdot}$
Total	$n_{\cdot 1}$	$\cdots$	$n_{\cdot j}$	$\cdots$	$n_{\cdot q}$	n

TAB. 2 – *Table de contingence croisant les états de la réponse y avec les feuilles de l'arbre.*

4 Indice d'implication

L'indice d'implication (voir par exemple Gras et al., 2004, p. 19) d'une règle se définit à partir des contre-exemples. Dans le cas des arbres de classification il s'agit dans chaque feuille (colonne du tableau 2) du nombre de cas qui ne sont pas dans la catégorie qui lui a été attribuée. Ces cas vérifient en effet la prémisse de la règle, mais pas sa conclusion. En notant b la conclusion (ligne du tableau) [1] de la règle j et n_{bj} le nombre de cas qui vérifient cette conclusion dans la $j$ème colonne, le nombre de contre-exemples est $n_{\bar{b}j} = n_{\cdot j} - n_{bj}$. L'indice d'implication est une forme standardisée de l'écart entre ce nombre et le nombre espéré de contre-exemples qui seraient générés en cas de répartition entre valeurs de la réponse indépendante de la condition de la règle.

Formellement, l'hypothèse de répartition indépendante de la condition, que nous notons H_0, postule que le nombre $N_{\bar{b}j}$ de contre-exemples de la règle j résulte du tirage aléatoire et indépendant d'un groupe de $n_{\cdot j}$ cas vérifiant la prémisse de la règle j et d'un autre de $n_{\bar{b}\cdot} = n - n_{b\cdot}$ cas qui ne vérifient pas la conclusion de la règle. Sous H_0 et conditionnellement à $n_{b\cdot}$ et $n_{\cdot j}$, le nombre aléatoire $N_{\bar{b}j}$ de contre-exemples est réputé (Lerman et al., 1981) suivre une loi de Poisson de paramètre $n^e_{\bar{b}j} = n_{\bar{b}\cdot} n_{\cdot j}$. Ce paramètre $n^e_{\bar{b}j}$ est donc à la fois l'espérance mathématique et la variance du nombre de contre-exemples sous H_0. Il correspond au nombre de cas de la feuille j qui seraient des contre-exemples si l'on répartissait les $n_{\cdot j}$ cas de j selon la distribution marginale, celle du nœud initial de l'arbre (ou marge de droite du tableau 2).

L'indice d'implication de Gras est l'écart $n_{\bar{b}j} - n^e_{\bar{b}j}$ entre les nombres de contre-exemples observés et attendus sous l'hypothèse H_0, standardisé par l'écart type, soit, en ajoutant la correction pour la continuité en vue de la comparaison avec la loi normale

$$\text{Imp}(j) \quad = \quad \frac{n_{\bar{b}j} - n^e_{\bar{b}j} + .5}{\sqrt{n^e_{\bar{b}j}}} \tag{1}$$

1. Notons que b peut évidemment varier d'une colonne à l'autre.

En termes de cas vérifiant la condition, cet indice s'écrit encore

$$\text{Imp}(j) \quad = \quad \frac{-(n_{bj} - n^e_{bj}) + .5}{\sqrt{n_{\cdot j} - n^e_{bj}}} \tag{2}$$

Une valeur positive de l'indice indique que la règle fait moins bien que le hasard et n'apporte donc aucune information implicative. Seules les valeurs négatives ont donc un intérêt. Plus l'indice — l'écart par rapport au hasard — est grand (en valeur absolue), plus la force implicative de la règle est forte.

Dans Ritschard (2005), nous avons proposé des variantes inspirées des résidus utilisés en modélisation de tables de contingence multidimensionnelles. Il s'agit du résidu déviance, du résidu ajusté d'Haberman et du résidu de Freeman-Tukey qui ont une variance plus proche de 1 que le résidu standardisé utilisé par Gras de variance plus petite. Le premier a cependant un comportement tendant vers 0 quand le nombre de contre-exemples s'approche de 0 qui le disqualifie (Pisetta et al., 2007). Les deux autres évoluent de façon similaire à l'indice de Gras tout au moins du point de vue qui nous intéresse ici de l'ordre de préférence des conclusions que suggèrent les valeurs de l'indice. Nous nous contentons donc ci-après de discuter l'usage de l'indice de Gras.

4.1 Gain d'implication comme critère d'optimalité des éclatements

L'indice permet de mesurer la force implicative de la règle. On peut alors songer à l'exploiter comme critère de développement de l'arbre. L'idée est de rechercher à chaque nœud l'éclatement qui produirait le meilleur gain en termes de force implicative des règles, en admettant évidemment qu'on retienne à chaque nœud la conclusion qui maximise l'intensité d'implication. On se heurte cependant ici à une difficulté d'agrégation. En effet, s'il est aisé de calculer l'indice d'implication avant l'éclatement, on se retrouve après l'éclatement avec plusieurs nœuds et donc un ensemble de valeurs d'indices d'implication qu'il nous faut synthétiser en une seule valeur qui puisse être comparée avec l'indice d'implication avant l'éclatement. Une possibilité est de prendre simplement une moyenne pondérée par les effectifs des nœuds concernés. Une autre solution, qui ferait sens si l'on est intéressé en priorité à obtenir quelques règles très fortes tout en s'accommodant de règles peu implicatives, est de retenir le maximum des intensités obtenues. Pour rester dans la logique de l'indice d'implication, une troisième solution d'indice d'implication pour l'ensemble S de sommets résultant de l'éclatement est (en incluant la correction pour la continuité)

$$\text{ImpT}(S) \quad = \quad \frac{\sum_{j\in S} n_{\bar{b}j} - \sum_{j\in S} n^e_{\bar{b}j} + .5}{\sqrt{\sum_{j\in S} n^e_{\bar{b}j}}} \quad = \quad \frac{\sum_{j\in S}(n_{\bar{b}j} - n^e_{\bar{b}j}) + .5}{\sqrt{\sum_{j\in S} n^e_{\bar{b}j}}} \tag{3}$$

soit l'écart standardisé entre le nombre total de contre-exemples observés des règles et le total attendu.

Pour notre exemple, nous donnons au tableau 3 le gain de force implicative apporté par les différents éclatements possibles au premier niveau. Le gain est la différence entre la valeur de l'indice au nœud que l'on veut éclater (soit 0 au nœud initial) et l'indice synthétique pour les nœuds résultant de l'éclatement. Le sexe s'impose clairement comme meilleur attribut prédictif. Il est intéressant de relever que le gain mesuré avec l'indice total est en règle générale plus fort que l'écart par rapport au maximum.

Attribut utilisé	nbre sommets	moyenne pondérée	maximum	ImpT
sexe	2	4.17	4.59	5.94
secteur	3	0.82	1.34	1.50
primaire	2	0.31	0.44	0.46
tertiaire	2	0.79	0.82	1.09

TAB. 3 – *Gains de force implicative pour les éclatements possibles au premier niveau.*

Attribut utilisé	nbre sommets	moyenne pondérée	maximum	ImpT
Sommet : Homme				
secteur	3	-0.71	0.22	1.18
primaire	2	-1.30	-0.10	0
tertiaire	2	0.48	1.23	1.18
Sommet : Femme				
secteur	3	-1.83	-0.15	0
primaire	2	-1.46	-0.15	0
tertiaire	2	-0.81	-0.01	0

TAB. 4 – *Gains de force implicative pour les éclatements possibles au deuxième niveau.*

On procède donc à l'éclatement selon le sexe, et l'on donne au tableau 4 les gains possibles au niveau 2 pour chacun des sommets "Homme" et "Femme". Pour les femmes, aucun gain de force implicative n'est possible avec la seule variable qui nous reste à savoir le secteur d'activité. La raison en est simplement que quelque soit l'éclatement, la catégorie pour laquelle on a l'implication la plus forte reste la même (célibataire) dans tous les nœuds qu'on obtient. Pour les hommes, il en est de même si l'on segmente entre le secteur primaire et le reste. Par contre, une segmentation en deux, tertiaire contre le reste ou en trois, permet un gain égal en termes d'implication totale. Le partage en deux paraît cependant plus intéressant puisqu'il se traduit, contrairement à l'éclatement en 3, par un gain positif également en termes d'implication moyenne.

4.2 Choix de la conclusion des règles

Chaque feuille (nœud terminal) de l'arbre caractérise une règle dont la prémisse est définie par les conditions d'embranchement le long du chemin menant du nœud initial à la feuille, la conclusion de la règle correspondant à la modalité assignée à la feuille. Comme déjà mentionné, le choix se porte de façon classique sur la modalité la plus fréquente. Dans certaines circonstances, il est plus pertinent de retenir la modalité assurant la plus forte implication. Il en est en particulier ainsi dans le contexte du ciblage où il s'agit de déterminer les profils types de chaque modalité de la variable cible y, et non pas, comme en classification, de prévoir la modalité que prendra un individu avec un profil donné.

Notons que l'usage de l'indice d'implication pour le développement de l'arbre suppose implicitement que la conclusion attribuée est dans chaque feuille la modalité qui assure la plus

forte valeur négative de l'indice d'implication. La conclusion est ainsi dans ce cas automatiquement déterminée. A titre d'exemple, en induisant l'arbre avec l'indice d'implication on obtient les trois règles du tableau 5 qui correspondent aux feuilles F1, F2 et F3 dans la figure 1. On notera en particulier que la conclusion attribuée à la 2ème règle n'est pas la modalité majoritaire.

Le recours à l'indice d'implication pour le choix des conclusions reste cependant également possible pour des arbres induits selon d'autres critères.

Règle	Condition		Conclusion
R1	Homme et secteur primaire ou secondaire	→	marié
R2	Homme et secteur tertiaire	→	divorcé
R3	Femme	→	célibataire

TAB. 5 – *Meilleures règles en termes de force implicative.*

5 Entropie décentrée

Les mesures d'entropie ont été définies mathématiquement par un ensemble d'axiomes en dehors du contexte de l'apprentissage machine. On peut trouver des travaux détaillés dans Rényi (1960) et Aczél et Daróczy (1975). Leur transfert vers l'apprentissage s'est fait de manière hâtive et sans prêter trop attention à la pertinence de leurs axiomes fondateurs. Ainsi, nous avons souligné dans Zighed et al. (2007) l'intérêt de relâcher l'axiome exigeant que l'entropie soit maximale à la distribution uniforme, et par suite évidemment l'axiome de symétrie stipulant que l'entropie doit être insensible à l'ordre des probabilités constituant la distribution. En nous fondant sur une axiomatique plus générale, nous avons proposé une entropie décentrée d'une distribution $(p_1, \ldots, p_\ell)$ généralisant l'entropie quadratique dans le cas où $\ell = 2$. Sa forme théorique, standardisée pour que sa valeur maximale soit égale à 1, est :

$$h_w(p_1, p_2, \ldots, p_\ell) \quad = \quad \frac{1}{\ell} \sum_{i=1}^{\ell} \frac{p_i(1-p_i)}{(-2w_i+1)p_i + w_i^2} \tag{4}$$

où $\mathbf{w} = (w_1, \ldots, w_\ell)$ est un vecteur de paramètres caractérisant la distribution d'incertitude maximale. On obtient une version empirique en remplaçant les p_i par leurs estimations de Laplace $\hat{p} = (n_i + 1)/(n + \ell)$. D'autres formes d'entropies décentrées ont également été proposées par Lallich et al. (2007) et Lenca et al. (2008).

5.1 Utilisation de l'entropie décentrée

Par rapport à l'indice d'implication qui compare le nœud obtenu au nœud initial en termes de distribution entre exemples et contre-exemples, l'entropie décentrée ne privilégie pas de catégorie particulière et compare l'ensemble de la distribution. Elle ne préjuge donc pas de la catégorie qui sera assignée au nœud.

Nous donnons au tableau 6 les gains d'entropie pour les divers éclatements possibles au premier niveau. A titre de comparaison nous donnons les gains obtenus en termes de l'indice de

			entropie décentrée	
Attribut utilisé	nbre sommets	Gini	théorique	empirique
sexe	2	0.150	0.210	0.201
secteur	3	0.016	0.024	0.023
primaire	2	0.001	0.003	0.003
tertiaire	2	0.011	0.017	0.016

TAB. 6 – *Gains d'entropie pour les éclatements possibles au premier niveau.*

Gini, qui est l'entropie quadratique classique, et de l'entropie décentrée théorique et empirique. La version théorique est obtenue en remplaçant dans la formule (4) les p_i par les fréquences observées, et la version empirique en les remplaçant par les estimations de Laplace.

Pour ce premier éclatement, les entropies classiques et décentrées conduisent au même résultat. Le sexe est la variable à retenir, tout comme il l'était avec le gain d'implication. On peut relever par ailleurs pour les entropies décentrées que le gain tend à être moins fort en termes de mesure empirique que théorique.

			entropie décentrée	
Attribut utilisé	nbre sommets	Gini	théorique	empirique
Sommet : Homme				
secteur	3	0.084	0.111	0.089
primaire	2	0.020	0.025	0.012
tertiaire	2	0.082	0.106	0.098
Sommet : Femme				
secteur	3	0.042	0.075	0.048
primaire	2	0.042	0.073	0.052
tertiaire	2	0.013	0.019	0.012

TAB. 7 – *Gains d'entropie pour les éclatements possibles au deuxième niveau.*

Le tableau 7 propose la même comparaison pour les éclatements possibles au second niveau. Les résultats divergent ici selon le type d'entropie utilisé. Pour ce qui est du sommet "Homme", Gini et l'entropie décentrée théorique sélectionnerait l'éclatement en trois, tandis que la version empirique de l'entropie décentrée privilégie l'éclatement qui oppose le secteur tertiaire aux deux autres secteurs. Il n'y a donc que ce dernier indice qui donne un résultat concordant avec l'optique implication discutée précédemment.

Pour le sommet "Femme", il y a également divergence, l'entropie empirique favorisant à nouveau un éclatement en deux plutôt qu'en trois, soit le secteur primaire contre les deux autres.

Notons qu'à nouveau les gains sont plus faibles avec la version empirique, les écarts étant ici plus importants en raison des effectifs plus faibles des nœuds. C'est la sensibilité aux effectifs que nous souhaitions.

5.2 Choix de la conclusion selon la contribution à l'entropie

L'entropie qui mesure l'écart entre deux distributions ne se prête pas en tant que telle à la mesure de l'intérêt de chaque modalité dans la feuille. La contribution de chaque modalité à cette entropie nous donne par contre une information utile de ce point de vue. La seule valeur de cette contribution n'est cependant pas suffisante. Il nous faut tenir compte également du signe de l'écart. En effet, une faible contribution à l'entropie indique une classe qui se démarque fortement de sa proportion marginale, mais cet écart est pertinent seulement si l'effectif observé dépasse l'effectif attendu en cas d'indépendance. On propose alors de sélectionner dans chaque feuille j la modalité qui maximise le critère

$$\max_i \ \eta_{w,ij} = signe(n_{ij} - n^e_{ij})\,(1 - h_{w,ij}), \quad j = 1, \ldots, q \tag{5}$$

où $h_{w,ij}$ est la contribution effective de la modalité i à l'entropie de la feuille j, n_{ij} le nombre observé de cas de modalité i dans la feuille j, et n^e_{ij} le nombre attendu sous l'hypothèse d'indépendance. On retient ainsi la modalité dont la contribution $h_{w,ij}$ est la plus faible parmi celles dont on observe plus de cas qu'attendus par hasard. Notons que, l'écart $n_{ij} - n^e_{ij}$ étant nécessairement non négatif pour au moins un i, la fonction 'signe' pourrait tout aussi bien être remplacée par la fonction logique $(n_{ij} - n^e_{ij} > 0)$ qui prend la valeur 1 lorsque l'écart est positif et 0 sinon.

	F1	F2	F3a	F3b
marié	0.59	-0.79	-0.09	-0.93
célibataire	-0.42	-1.00	0.39	0.88
dicorcé/veuf	-1.00	0.81	-1.00	-0.95

TAB. 8 – *Contributions à l'entropie décentrée des feuilles.*

Le tableau 8 donne les valeurs de $\eta_{w,ij}$ pour les 4 feuilles de l'arbre de la figure 1. On note que les conclusions sélectionnées concordent avec celles d'implication maximale.

6 Discussion

Nous avons vu que tant l'indice d'implication que l'entropie décentrée pouvaient servir de critère de développement de l'arbre. Les deux approches fournissent également des éléments permettant d'attribuer aux feuilles une conclusion appropriée dans un contexte de données déséquilibrées où le rappel de catégories faiblement représentées est plus important que le taux total d'erreurs de classification. Les deux approches ne sont pas pour autant équivalentes. L'indice d'implication oppose la classe pour laquelle on a l'implication maximale aux autres, tandis que l'entropie asymétrique prend en compte tout le détail de la distribution. Quels sont alors les avantages et inconvénients respectifs ?

6.1 Avantages et limites

Considérons tout d'abord l'optique du développement de l'arbre. De ce point de vue, l'indice d'implication a quelque analogie avec le critère 'Twoing' de CART (Breiman et al., 1984) qui pour chaque éclatement possible cherche la partition en deux des valeurs de la variable cible qui maximise l'indice de Gini. L'avantage est qu'on a ainsi un critère qui devient plus robuste en se fondant sur des effectifs moins dispersés qui le rendent notamment moins sensible aux variations à l'intérieur de chacune des deux classes. Le même argument vaut pour l'indice d'implication bien que dans ce cas la première classe n'ait toujours qu'une seule catégorie.

Utiliser l'indice d'implication comme critère d'éclatement présuppose que la catégorie maximisant l'implication sera assignée au nœud. Ceci assure évidemment une cohérence à la procédure, mais limite évidemment aussi l'usage de l'arbre obtenu au contexte où ce choix de la conclusion selon la force implicative s'avère pertinent.

Pour ce qui est de l'entropie décentrée, elle mesure la proximité à la distribution de référence, proximité que l'on cherche à minimiser de sorte à obtenir des distributions aussi différentes que possible de la référence. On peut ici faire l'analogie avec le critère du khi-deux utilisé par l'algorithme CHAID (Kass, 1980) qui conduit également à choisir la segmentation pour laquelle les distributions s'écartent le plus possible de celle d'indépendance. La différence est que dans CHAID le référentiel change à chaque nœud puisque le critère consiste à s'éloigner le plus possible de la distribution du nœud qu'on éclate, tandis qu'avec le gain d'entropie décentré on cherche à se démarquer de la distribution du nœud initial qui reste la même à toutes les étapes du calcul.

Comme l'illustre en particulier notre exemple, les deux approches conduisent à des arbres relativement semblables même s'il l'on peut imaginer des situations peu claires où les éclatements proposés peuvent différer. Remarquons tout de même que l'indice d'implication ne propose pas d'éclatement lorsque les conclusions prévues pour les nœuds qui en résulteraient sont les mêmes. Ainsi, dans la figure 1, le développement s'arrête à la feuille F3 avec l'indice d'implication, alors même qu'on réalise un gain d'entropie décentrée en éclatant le nœud en F3a et F3b. On peut donc s'attendre à obtenir des arbres moins complexes avec l'indice d'implication qu'avec l'entropie décentrée.

Sur le plan de la complexité de calcul, la mise en œuvre de l'entropie décentrée semble un peu plus immédiate, l'indice d'implication nécessitant de tester à chaque fois les différentes possibilités d'opposer une catégorie aux autres. Ceci n'affecte la complexité de l'algorithme que par un facteur multiplicatif c correspondant au nombre de catégories de la variable cible.

Enfin, dans une optique de généralisation, il est important pour assurer la robustesse des résultats de disposer de critères qui soient sensibles à la taille des effectifs. L'entropie décentrée l'est dans sa forme empirique, la sensibilité à l'effectif découlant de l'utilisation des estimations de Laplace des probabilités. Quant à l'indice d'implication, qui peut être vu comme un résidu standardisé, il est calculé à partir des effectifs et non des proportions et est donc sensible aux effectifs par construction.

Si l'on considère à présent l'attribution de la conclusion aux feuilles de l'arbre, l'indice d'implication présente l'avantage d'avoir une interprétation claire abondamment discutée dans la littérature : l'implication statistique est d'autant plus forte que la règle admet étonnamment peu de contre-exemples.

Le critère $\eta_{w,ij}$, complémentaire à un de la contribution à l'entropie décentrée, est moins intuitif. Il mesure en quelque sorte l'importance de l'écart entre la fréquence de la catégorie

dans le feuille et la proportion avec laquelle cette même catégorie est observée dans l'ensemble de la population. Contrairement à l'indice d'implication, il se fonde sur la fréquence même de la catégorie, et non sur ses contre-exemples. Le critère $\eta_{w,ij}$ conduit ainsi à privilégier la catégorie dont la fréquence domine relativement le plus fortement sa proportion marginale.

Les deux critères trouvent leur justification dans une perspective de ciblage où l'on s'intéresse à savoir pour quelle valeur de la variable cible le profil décrit par la condition de la règle est la plus typique. Par exemple, un médecin sera intéressé en priorité à savoir quelle est la population la plus exposée au risque de développer un cancer. De même, il est naturel d'axer en priorité des actions de marketing, de prévention ou de contrôle sur les groupes de population qui seront les plus réceptifs même lorsque ceux-ci ne sont pas majoritairement concernés par les actions envisagées.

On peut noter que les deux indicateurs sélectionnent la même catégorie lorsqu'une seule fréquence de la feuille dépasse la proportion marginale. Les choix peuvent cependant diverger dans le cas contraire. A titre d'exemple, nous donnons au tableau 9 la distribution d'une feuille j pour laquelle on obtient des conclusions non concordantes. On observe que l'écart entre effectifs observés (les n_{ij}) et attendus selon la distribution marginale est le même, -6, pour les catégories A et B. Relativement, l'écart est plus important pour la catégorie A que privilégie la contribution à l'entropie. L'indice d'implication privilégie par contre B, pour laquelle on a moins de contre-exemples.

	en tout	distribution dans feuille j			effectifs	contre-exemples		Indice	Contrib. $\eta_{w,ij}$
catégorie	w_i	n_{ij}	f_{ij}	$\hat{p}_{ij}$	attendus	observés	attendus	implication	à l'entropie
A	10%	12	.2	0.206	6	48	54	-0.75	0.064
B	20%	18	.3	0.302	12	42	48	-0.79	0.047
C	70%	30	.5	0.492	42	30	18	2.95	-0.147
Total	100%	60	1	60	1	-	-	-	-

TAB. 9 – *Illustration de la différence entre indice d'implication et contribution à l'entropie décentrée.*

Un avantage de l'indice d'implication est qu'il peut être comparé avec une distribution normale, ce qui justifie d'ailleurs la correction pour la continuité que nous lui avons apporté. Ceci permet de dire par exemple dans le cas du tableau 9 que le choix de la conclusion n'est pas solidement établi statistiquement puisque la valeur de l'indice et en deçà du seuil critique de -1.645 pour un risque de 5%. Notons que pour la comparaison avec la loi normale il serait préférable d'utiliser l'une des variantes proposées dans Ritschard (2005), l'indice de Gras tendant à avoir une variance inférieure à 1.

6.2 Vers une approche hybride

Au vu des remarques précédentes nous proposons d'exploiter de préférence l'entropie décentrée pour le développement de l'arbre et l'indice d'implication pour l'attribution des conclusions aux feuilles.

Pour le développement de l'arbre, l'entropie décentrée est un peu plus simple à mettre en œuvre, mais nous semble surtout être un instrument plus général, l'indice d'implication étant trop étroitement lié à la procédure de choix de la conclusion et donc à la seule élaboration de

règles. Par exemple, le non éclatement lorsque les règles obtenues prennent la même conclusion peut être un handicap dans la mesure où cela empêche de repérer des sous-groupes pour lesquels la règle serait plus fiable que pour d'autres. L'entropie décentrée nous semble de ce point de vue permettre plus de nuances.

Une fois l'arbre construit par contre, l'indice d'implication nous semble mieux indiqué pour le choix de la conclusion, de par l'importance accordée aux contre-exemples et la possibilité qu'il offre de juger de la signification statistique du lien entre prémisse et conclusion de la règle.

7 Expérimentations

Le propos de cette section est de donner un éclairage empirique sur l'utilisation de l'indice d'implication et de l'entropie décentrée dans la construction d'arbres de décision. Il s'agit de vérifier empiriquement que ces critères ont bien le comportement escompté en présence de données déséquilibrées, en particulier que le recours à ces critères permet d'améliorer les résultats, notamment en termes de rappel de la ou des classes sous-représentées. Nous distinguons dans ces expérimentations l'utilisation de chacun de ces indices comme critère de développement de l'arbre ainsi que comme critère pour assigner la classe ou conclusion aux feuilles.

L'objectif étant également d'illustrer les différences entre indice d'implication et entropie décentrée, nous retenons pour cette étude empirique des données où la variable réponse a plus de deux classes. En effet, dans le cas de deux classes, en opposant la catégorie sélectionnée aux autres comme le fait le premier indice on exploite la même information que l'entropie décentrée qui prend en compte toute la distribution, et les résultats devraient donc être très similaires. Nous retenons ainsi comme point de départ pour nos expérimentation des données réelles sur la situation (réussi, redouble, éliminé) après leur première année d'études des étudiants qui ont commencé leur cursus à la Faculté des sciences économiques et sociales de l'Université de Genève en 1998 (Petroff et al., 2001). Les données étant cependant peu déséquilibrées avec une classe minoritaire (redouble) de 17%, nous avons forcé le déséquilibre en gonflant la classe majoritaire (réussi) par sur-échantillonnage, c'est-à-dire en dupliquant aléatoirement des cas de ce groupe. Le tableau 7 indique comment les données finalement retenues se répartissent selon les trois classes de la variable réponse.

Classe	Effectif	Proportion
1. éliminé	209	0.06
2. redouble	130	0.03
3. réussi	3384	0.91

TAB. 10 – *Distribution de la variable réponse du jeu de données utilisées*

L'expérimentation menée consiste à générer des arbres en utilisant successivement le gain d'entropie classique de Shannon, le gain de force implicative et le gain de l'entropie décentrée comme critère d'éclatement. Le développement des arbres est arrêté lorsqu'on ne peut plus obtenir de gain strictement positif. Aucun autre critère d'arrêt n'est utilisé et aucun élagage

	Critère de développement de l'arbre											
	Shannon				Implication				Entropie décentrée			
Règle	1	2	3	1et2	1	2	3	1et2	1	2	3	1et2
Rappel												
Majoritaire	25.4	15.6	100.0	46.6	18.6	16.7	99.9	33.3	15.3	10.8	100.0	25.1
Implication	35.4	26.1	82.6	57.8	39.2	20.0	86.9	51.3	40.2	26.2	73.1	61.4
Entropie décentrée	34.9	23.1	84.4	58.7	37.6	21.3	87.4	51.6	34.4	25.4	73.9	57.5
Précision												
Majoritaire	61.0	38.8	94.9	100.0	61.7	50.0	93.7	97.4	62.7	41.2	93.0	100.0
Implication	14.3	11.1	95.1	24.9	18.1	14.1	94.7	28.2	12.6	7.6	95.0	18.6
Entropie décentrée	15.9	9.9	95.3	27.3	18.4	15.1	94.7	29.1	12.7	6.4	94.6	18.1

TAB. 11 – *Rappel et précision, 10-validation croisée*

n'est effectué. Il s'agit là évidemment d'une première expérimentation qui permettra de voir ce qu'il se passe dans le cas extrême où l'on laisse l'arbre se développer au maximum. Pour cette expérimentation préliminaire, nous avons également simplement utilisé les simple fréquences observées sans correction de Laplace.

Le tableau 11 donne les taux de rappel et de précision pour chacune des trois classes, ainsi que pour le regroupement des deux classes sous-représentées. Il ressort très clairement de ces résultats que fonder le choix des conclusions sur l'indice d'implication ou la contribution à l'entropie, permet d'améliorer sensiblement le rappel des classes sous-représentées. Les différences entre les deux critères restent cependant non significatifs. Pour ce qui est du critère de croissance de l'arbre, les résultats sont moins clairs. Il est surprenant que ni l'indice d'implication, ni l'entropie décentrée ne domine l'entropie centrée de Shannon. L'explication tient sans doute à l'absence de critère d'arrêt et d'élagage. En effet on tend ainsi à épuiser les prédicteurs et donc à générer une partition fine qui est sans doute assez semblable quelque soit le critère utilisé.

8 Conclusion et perspectives

Nous nous sommes dans cet article intéressés à la problématique des données déséquilibrées dans le contexte des arbres de décision. Nous avons présenté et discuté avantages et inconvénients de deux approches, l'une fondée sur l'indice d'implication et l'autre sur une entropie décentrée. Il apparaît que ces deux approches conduisent à des solutions semblables bien qu'obéissant à des logiques totalement différentes. L'entropie décentrée semble être un critère plus naturel pour le développement de l'arbre tandis que l'indice d'implication présente des avantages certains pour sélectionner la catégorie à attribuer aux feuilles. Il s'agit là cependant d'une conjecture que notre expérimentation préliminaire n'a pas clairement confirmé. Nous travaillons actuellement à la mise au point d'un protocole d'expérimentation plus élaboré qui permettra de tester empiriquement notre hypothèse et d'évaluer les incidences des paramètres de contrôle du développement de l'arbre. L'expérimentation devrait aussi porter sur un ensemble de jeux de données benchmark. Enfin, il nous faudra encore populariser ces procédures en les implémentant dans des plateformes aisément accessibles.

Références

Aczél, J. et Z. Daróczy (1975). *On measures of information and their characterizations*. New York: Academic Press.

Blanchard, J., F. Guillet, H. Briand, et R. Gras (2005). Une version discriminante de l'indice probabiliste d'écart à l'équilibre pour mesurer la qualité des règles. In Gras et al. (2005), pp. 131–138.

Breiman, L., J. H. Friedman, R. A. Olshen, et C. J. Stone (1984). *Classification And Regression Trees*. New York: Chapman and Hall.

Gras, R., R. Couturier, J. Blanchard, H. Briand, P. Kuntz, et P. Peter (2004). Quelques critères pour une mesure de qualité de règles d'association. *Revue des nouvelles technologies de l'information RNTI E-1*, 3–30.

Gras, R., F. Spagnolo, et J. David (Eds.) (2005). *Actes des Troisièmes Rencontres Internationale ASI Analyse Statistique Implicative*, Volume Secondo supplemento al N.15 of *Quaderni di Ricerca in Didattica*, Palermo. Università degli Studi di Palermo.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics 29*(2), 119–127.

Lallich, S., P. Lenca, et B. Vaillant (2005). Variation autour de l'intensité d'implication. In Gras et al. (2005), pp. 237–246.

Lallich, S., P. Lenca, et B. Vaillant (2007). Construction d'une entropie décentrée pour l'apprentissage supervisé. In *QDC 2007, Actes du 3ème atelier Qualités des données et connaissances, EGC janvier 2007, Namur*, pp. 45–54.

Lenca, P., S. Lallich, T.-N. Do, et N.-K. Pham (2008). A comparison of different off-centered entropies to deal with class imbalance for decision trees. In *Advances in Knowledge Discovery and Data Mining, 12th Pacific-Asia Conference, PAKDD 2008, Osaka, Japan, May 20-23*, pp. 634–643.

Lerman, I. C., R. Gras, et H. Rostam (1981). Elaboration d'un indice d'implication pour données binaires I. *Mathématiques et sciences humaines* (74), 5–35.

Marcellin, S., D. A. Zighed, et G. Ritschard (2006). Detection of breast cancer using an asymmetric entropy measure. In A. Rizzi et M. Vichi (Eds.), *COMPSTAT 2006 - Proceedings in Computational Statistics*, pp. 975–982. Berlin: Springer. (on CD).

Petroff, C., A.-M. Bettex, et A. Korffy (2001). Itinéraires d'étudiants à la Faculté des sciences économiques et sociales: le premier cycle. Technical report, Université de Genève, Faculté SES.

Pisetta, V., G. Ritschard, et D. A. Zighed (2007). Choix des conclusions et validation des règles issues d'arbres de classification. In M. Noirhomme et G. Venturini (Eds.), *Extraction et Gestion des Connaissances (EGC 2007)*, Volume E-9 of *Revue des nouvelles technologies de l'information RNTI*, pp. 485–496. Cépaduès.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. San Mateo: Morgan Kaufmann.

Rényi, A. (1960). On measures of entropy and information. In *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability*, Volume 1, Berkeley, pp. 547–561. University of California Press.

Ritschard, G. (2005). De l'usage de la statistique implicative dans les arbres de classification. In Gras et al. (2005), pp. 305–314.

Zighed, D. A., S. Marcellin, et G. Ritschard (2007). Mesure d'entropie asymétrique et consistante. In M. Noirhomme et G. Venturini (Eds.), *Extraction et Gestion des Connaissances (EGC 2007)*, Volume E-9 of *Revue des nouvelles technologies de l'information RNTI*, pp. 81–86. Cépaduès.

Summary

This paper is concerned with the induction of classification trees for imbalanced data, i.e. for the case where some categories of the target variable are much less frequent than other ones. More specifically, we address two aspects. On the one hand, we look for growing criteria that efficiently take into account the specific imbalanced nature of the data. On the other hand, we deal with the relevance of the conclusion that should be assigned to the leaves of a grown tree. We have recently considered two independent ways for dealing with these issues. The first one consisted in defining and using out centered entropies, and the second one on relying on measures of implication strength derived from implicative statistics. The aim of this paper is to compare and establish the relationship between these two approaches. It presents a first rouph experimentation.

Chapitre 8 : Graphe de règles d'implication statistique pour le raisonnement courant. Comparaison avec les réseaux bayésiens et les treillis de Galois.

Martine Cadot

Université Henri Poincaré / LORIA, Nancy, France
Martine.Cadot@loria.fr
http://www.loria.fr/~cadot

Résumé. Les règles d'implication statistique ressemblent aux règles du raisonnement mathématique. Ce qui permet de les utiliser facilement pour raisonner sur les données. Toutefois, le modèle sous-jacent aux règles d'implication statistique n'est pas le modèle de la logique formelle utilisé en mathématique, mais un modèle statistique aboutissant à des relations approximatives. Contrairement au raisonnement mathématique, le raisonnement courant se satisfait de règles approximatives. Mais il a besoin d'un graphe pour savoir quels enchaînements de règles sont possibles car en faisant se succéder des approximations, on finit par arriver à des incohérences.
On montrera dans ce chapitre comment fonctionne l'enchaînement de ces règles, notamment à travers la construction du graphe des règles d'implication tel que proposé dans les différentes versions de CHIC et on comparera ce modèle statistique des données à deux autres modèles proches : un modèle algébrique, les treillis de Galois, et un modèle probabiliste, les réseaux bayésiens. Pour permettre des comparaisons aisées, le fonctionnement des trois modèles sera illustré à l'aide d'un même jeu de données médicales librement disponible sur Internet.

1 Introduction

Les règles d'implication statistique ont été conçues à l'origine par Régis Gras dans sa thèse de mathématique (1979). Une règle d'implication statistique entre deux « notions » A et B s'exprime sous la forme « si A alors B » (ou A→B) et indique que si l'élève a acquis la notion A alors il a acquis la notion B. La connaissance de telles règles aide l'enseignant de mathématique à organiser son cours pour faire assimiler un certain nombre de notions à ses élèves. En suivant le modèle de Régis Gras, l'enseignant peut établir un réseau de règles entre des notions (ou plutôt leur assimilation) sans avoir besoin de définir théoriquement ce qu'est chaque notion, du moment qu'il peut déterminer pratiquement si un élève donné l'a acquise ou non. En effet, les règles sont obtenues de façon statistique, par calcul à partir des résultats d'observations d'un ensemble d'élèves, à travers les résultats d'une interrogation écrite par exemple. La méthodologie de construction de règles d'implication statistique n'est pas restée cantonnée à la didactique car elle permet d'établir des liens de « cause à effet »

entre un ensemble quelconque de variables (ou propriétés, items, etc.) à partir de leurs valeurs (présence/absence) pour un ensemble de sujets (ou d'objets, enregistrements, etc.).

Les règles sont obtenues automatiquement à partir de données d'observations, et non postulées une à une avant d'être établies en utilisant un protocole expérimental adapté. Cette façon de procéder s'inscrit dans la fouille de données (Data Mining), qui permet d'explorer après coup des données pas nécessairement recueillies pour cet usage. On parle alors d'extraction de connaissances à partir de données (Knowledge Discovery from Databases). Depuis les travaux fondateurs de Régis Gras, une autre méthode de construction de relations de type causal sur des variables observables a vu le jour, il s'agit des réseaux bayésiens (Naïm 2007), issus des modèles graphiques (Whittaker 1990). Les deux méthodologies de représentation de connaissances que sont l'implication statistique et les réseaux bayésiens ont en commun leur volonté de fournir à l'utilisateur un réseau de règles dans lequel il peut naviguer pour faire des raisonnements. L'utilisateur suit une logique « de bon sens » de type mathématique, c'est pourquoi à ces deux modèles nous en avons ajouté un troisième, plus près, par construction, de la logique mathématique. Il s'agit des treillis de Galois (Godin et al. 1995). La comparaison des trois méthodes, proches par leurs objectifs, mais différentes par leur caractéristiques, va nous permettre de montrer finement les particularités des liens de type causal appelés « règles d'implication statistique ».

Afin de rendre plus concret l'exposé de ce chapitre, nous illustrerons les diverses notions abordées sur un jeu de données médicales, Asia, que nous exposons dans la deuxième partie de ce chapitre. Dans la troisième partie nous donnons les définitions des règles d'implication statistique, nous examinons de façon formelle la façon dont elles vérifient les propriétés attendues par l'utilisateur pour construire un raisonnement, puis plus concrètement le jeu de règles obtenu sur les données Asia. Dans les parties 4 et 5, nous exposons les deux autres modèles de la même façon, et nous comparons les propriétés des jeux de règles de façon formelle ainsi que sur les données Asia. Nous terminons par un bilan et des perspectives. La formulation mathématique dans ce chapitre a été réduite à l'extrême, les définitions nécessitant une écriture mathématique détaillée ont été reportées en annexe, et les preuves formelles ont été mises la plupart du temps en notes de bas de page.

2 Le jeu de données « Asia »

Le jeu de données « Asia » est souvent utilisé pour illustrer le fonctionnement des réseaux bayésiens. Il en existe de nombreuses versions. Nous avons choisi celle fournie par BayesaLab[1] (version d'évaluation, 2001-2008), logiciel de réseau bayésien, qui détaille son utilisation de façon très pédagogique. Le jeu de données initial comportait 10 variables avec des valeurs pour 10 000 sujets. Nous avons retiré les sujets pour lesquels des valeurs manquaient, il en reste 7033, et supprimé la dernière variable (localisation) qui n'était pas utilisée dans le logiciel. Puis nous avons recodé l'âge en deux classes (avant 50 ans et après) au lieu de trois. Ces transformations ont permis un traitement comparable des données par des règles d'implication statistique, des réseaux bayésiens et des treillis de Galois. Dans le tableau 1 figurent les sigles des neuf variables et leur signification, ainsi que le nombre de sujets ayant la caractéristique (valeur 1 pour la variable).

[1] http://www.bayesia.com/fr/produits/bayesialab/release/bayesialab-3-3.php.

Variable	Nombre de valeurs "1"
AG (0 : âge<50 ; 1 : âge>=50)	3391
BR (Bronchite)	3051
CA (Cancer)	271
DY (Dyspnée)	2964
SM (Fumeur)	3271
TC (Tuberculose ou Cancer)	352
TU (Tuberculose)	83
VA (Visite de l'Asie)	74
XR (Rayons X)	685
N (Nombre de sujets)	7033

TAB. 1 – *Liste des variables des données Asia et nombre de leurs valeurs 1.*

3 Les règles d'implication statistique

L'implication statistique, que l'on notera A→B tout au long de ce chapitre, a été définie pour ressembler le plus possible à l'implication logique A⇒B qui est à la base des raisonnements courants. Nous examinons donc les caractéristiques de l'implication logique les plus utilisées pour les raisonnements, et nous regardons si elles se retrouvent à l'identique dans l'implication statistique ou bien si elles sont modifiées et comment tenir compte de ces modifications dans leur utilisation.

Définition de la règle d'implication logique A⇒B. Si A et B sont deux faits qui prennent leurs valeurs dans l'ensemble {V : Vrai, F : Faux}, la règle A⇒B est un fait dont la valeur dépend de celles de A et de B comme indiqué dans le tableau 2. Elle est fausse dans un cas, et vraie dans les trois autres. Pour prouver que la règle logique A⇒B est fausse, il suffit de trouver un « contre-exemple », c'est-à-dire une situation dans laquelle A est vrai et B faux.

	B : Vrai	B : Faux
A : Vrai	A⇒B Vrai	A⇒B Faux
A : Faux	A⇒B Vrai	A⇒B Vrai

TAB. 2 – *Table de vérité de la règle d'implication logique A⇒ B*

Définition de la règle d'implication statistique A→B. Si on connaît les valeurs de N individus pour les faits A et B, la valeur de la règle d'implication A→B se calcule d'après la répartition des N individus en quatre effectifs a, b, c et d selon les valeurs de A et de B, figurant dans le tableau 3. Cette valeur est un nombre compris entre 0 et 1 dépendant des nombres a, b, c et d'autant plus élevé que le nombre b de contre-exemples est petit relativement aux nombres a, c et d.

Graphe de règles d'implication statistique pour le raisonnement courant.

	B : Vrai	B : Faux
A : Vrai	a	b
A : Faux	c	d

TAB. 3 – *Tableau de contingence de A et B pour le calcul de la valeur de la règle d'implication statistique A→B*

Algèbre booléenne ou statistique. Les deux définitions précédentes sont relatives aux mêmes faits A et B ayant pour valeurs Vrai/Faux qui peuvent être codées par 0/1. Mais la première se situe dans une logique booléenne, associée à un univers mathématique abstrait et intemporel, alors que la seconde résulte d'un décompte d'observations plus ou moins objectives faites à un moment donné dans une situation donnée. Il aurait été possible de définir une valeur booléenne pour l'implication statistique en la prenant égale à 1 (Vrai) si l'effectif b est nul (aucun contre-exemple) et à 0 (Faux) sinon. C'est le choix qui a été fait par Guigues et Duquenne (1986) pour définir leurs règles. Nous verrons dans la partie consacrée au treillis de Galois le cadre dans lequel ce choix se situe. Pour la règle d'implication statistique, le choix d'une valeur pouvant être comprise entre 0 et 1 s'appuie sur des considérations statistiques : on compare la valeur de b trouvée pour ces données aux valeurs extrêmes qu'il pourrait atteindre pour des données similaires en cas d'indépendance entre A et B. Le résultat de la comparaison est une probabilité, qui est d'autant plus proche de 1 que b est petit. Pour la calculer, on s'appuie sur des hypothèses concernant les fluctuations possibles des données (loi de Poisson, loi normale, …) comme indiqué dans (Gras et al. 2000). Plusieurs versions de ce calcul existent selon les hypothèses prises sur les lois de probabilité, ainsi qu'une correction utilisant l'entropie mutuelle. Nous précisons maintenant sur un exemple le principe de ce calcul, et nous renvoyons le lecteur intéressé par sa justification théorique à la thèse de Régis Gras ainsi qu'aux articles des différents chercheurs qui ont proposé des corrections (Gras et al. 2001).

Un exemple de calcul de l'indice d'implication statistique. Dans le tableau 4 sont donnés les 4 effectifs des variables Cancer et Bronchite de la base de données Asia afin de les utiliser pour le calcul de la valeur de la règle Cancer → Bronchite.

	Bronchite : Vrai	Bronchite : Faux
Cancer : Vrai	a =139	b = 132
Cancer : Faux	c = 2912	d = 3850

TAB. 4 – *Tableau de contingence de Cancer et Bronchite*

La valeur de b est loin d'être nulle, ce qui pourrait faire penser que la règle est loin d'être vraie. Toutefois la situation de référence n'est plus la règle vraie, avec une valeur théorique de b égale à 0, ni la règle fausse, mais une absence de règle correspondant à l'indépendance statistique entre les deux variables Cancer et Tuberculose. La valeur théorique de b en cas d'indépendance[2] est (b+a)(b+d)/N (pour Asia, N=7033), soit ici 153,4. Quand les données correspondant à une absence de règle fluctuent, leur valeur correspondante de b est supposée suivre une loi normale ayant comme paramètres une espérance E(b) de 153,4 et un écart-type

[2] On suppose ici que le lecteur a des connaissances statistiques concernant l'indépendance de deux variables aléatoires (ou test du Chi2 d'indépendance) et l'utilisation de la loi normale. Dans le cas contraire, il peut retrouver celles-ci dans les ouvrages de statistique de base traitant de l'inférence statistique, comme Morineau, A. (éd.) (1995)..

σ(b) égal à racine(153,4)=12,4. La connaissance de la loi normale nous permet de dire par exemple qu'en théorie 50% des valeurs sont inférieures à l'espérance, 15,9% sont inférieures à l'espérance diminuée de un écart-type (soit 141), 2,3% sont inférieures à l'espérance diminuée de deux écarts-types. La valeur que nous avons trouvée pour b est de 132 soit inférieure à l'espérance de 1,7 écarts-types et, en cas d'absence de règle, on ne pouvait avoir une valeur plus petite que dans 4,2% des cas, c'est-à-dire avec une probabilité[3] de 0,042. Selon la version normale de l'indice d'implication de Régis Gras, la règle Cancer → Bronchite a une valeur de 1-0,042, soit 0,958. Et sa valeur selon la version entropique figurant dans (Gras et al. 2001), dont nous ne donnons pas le détail de calcul ici, est de 0,943. C'est donc une règle qui a de grandes chances d'être vraie.

La règle et sa contraposée ont la même valeur d'indice. Nous venons de voir dans l'exemple que l'indice de A→B se calcule en comparant la valeur de b à la valeur de son espérance E(b)=(b+a)(b+d)/N, les effectifs a, b, c et d correspondant aux valeurs de A et B selon le tableau 3, puis en divisant le tout par la racine carrée de E(b), et que cette valeur de b « centrée réduite » suit la loi normale standard. Pour calculer l'indice de la contraposée de A→B, qui est nonB→nonA, on a écrit les effectifs croisant A et B dans le tableau 5. Ils ont été obtenus à partir de ceux du tableau 3 en échangeant les colonnes entre elles (B devient nonB), les lignes entre elles (A devient nonA) et les lignes avec les colonnes (A devient B et B devient A). Du tableau 3 au tableau 5, les effectifs a et d ont été échangés, alors que b et c sont restés à la même place.

	non A : Vrai	non A : Faux
non B : Vrai	d	b
non B : Faux	c	a

TAB. 5 – *Tableau de contingence de nonB et nonA déduit du tableau 3*

Pour calculer la valeur d'implication statistique de la règle nonB→nonA, on reprend donc la valeur de b que l'on compare comme précédemment à sa valeur théorique en cas d'indépendance E(b)=(b+a)(b+d)/N, et qui ne change pas de valeur quand on échange a et d. Et on divise par la racine carrée de E(b) avant de prendre la probabilité correspondante selon la loi normale standard. Ainsi la règle et sa contraposée ont la même valeur[4]. L'existence de cette propriété facilite l'utilisation des règles d'implication statistique dans les raisonnements.

Les règles A→B et A→nonB ont des valeurs complémentaires. Nous reprenons le tableau 3 en échangeant les colonnes pour obtenir nonB.

[3] Comme b suit la loi normale d'espérance E(b) et d'écart-type s(b), sa valeur centrée réduite $\frac{b - E(b)}{s(b)} = -1{,}73$ suit la loi normale d'espérance 0 et d'écart-type 1, appelée loi normale standard. Selon cette loi, la probabilité d'avoir une valeur inférieure à 61,7 est P(x<-1,73)=0,042.

[4] Nous considérons ici les indices d'implication non corrigés, quelle que soit la loi de probabilité sur laquelle ils s'appuient. Pour les propriétés des indices corrigés, nous invitons le lecteur à consulter (Gras et al. 2001).

Graphe de règles d'implication statistique pour le raisonnement courant.

	non B : Vrai	non B : Faux
A : Vrai	b	a
A : Faux	d	c

TAB. 6 – *Tableau de contingence de A et nonB déduit du tableau 3*

Le calcul de l'indice de A→B nécessite que l'on calcule d'abord la différence entre b et E(b)=(b+a)(b+d)/N. Cette différence est égale[5] à (bc-ad)/N. Le calcul de l'indice de A→nonB s'obtient en remplaçant a par b et c par d et inversement dans cette dernière formule, qui devient (ad-bc)/N, donc les deux différences sont opposées. Puis on divise par la racine carrée de l'espérance, qui est E(b) pour la première règle et E(a) pour la seconde. Cela donne deux valeurs de signe contraire, qui produisent selon la loi normale standard deux probabilités dont l'une est inférieure à 0,5 et l'autre supérieure, la valeur 0,5 étant atteinte en cas d'indépendance entre A et B. Par exemple l'indice d'implication de la règle Cancer→Bronchite vaut 1-P(x<-21,4/racine(153,4))=0,958, et celui de la règle Cancer→non Bronchite vaut 1-P(x<+21,4/racine(117,6))= 0,024. Et l'habitude étant de ne considérer que les règles ayant une valeur bien supérieure à la valeur de 0,5, on n'a jamais simultanément une règle A→B et la règle « contraire » A→nonB. Cette propriété, comme celle du paragraphe précédent, facilite les raisonnements utilisant les règles d'implication statistique.

La règle A→B et sa réciproque B→A. Les effectifs du tableau de contingence lié à la règle B→A se déduisent des effectifs de celui de la règle A→B (cf. tableau 3) par échange des lignes et des colonnes, ce qui produit l'échange de b et c alors que a et d restent inchangés. Les valeurs centrées b-E(b) et c-E(c) sont toutes deux égales à (bc-ad)/N, et les valeurs centrées réduites sont de même signe et égales si et seulement si E(b)=E(c). Ce dernier cas se produit par exemple dès que le « support » de A (le nombre de sujets vérifiant A) est le même que celui de B. Ainsi dès que l'implication statistique de la règle A→B est élevé, celui de sa réciproque l'est aussi quand les supports de A et de B sont proches. L'existence possible de deux règles réciproques l'une de l'autre et de même valeur de vérité ne pose pas de problème en logique formelle, les deux implications étant remplacées par une équivalence. Dans le cas des règles d'implication, si une règle et sa réciproque ont des valeurs élevées, le choix de Régis Gras[6] a été de n'en garder qu'une des deux, celle de plus grande valeur, quand leurs valeurs sont inégales, et de n'autoriser l'équivalence que quand les valeurs sont égales, ce qui arrive assez rarement pour des données réelles avec suffisamment de sujets. Dans le tableau 7 figurent toutes les règles tirées des données Asia ayant un indice d'implication supérieur à 0,95, la valeur de l'indice étant reportée avec cinq chiffres après la virgule. Les treize règles de R1 à R26 ont également leur réciproque d'indice supérieur à 0,95, alors que les 9 règles de R27 à R35 ne sont pas dans ce cas. On remarquera que l'extraction s'est faite parmi les 9 variables de la base de données et parmi leurs négations, soit 18 variables, mais on n'a pas fait figurer les contraposées, sinon il y aurait 70 règles d'indice supérieur à 0,95, chaque règle ayant la même valeur que sa contraposée. Si on choisissait de supprimer, parmi les 2 règles réciproques l'une de l'autre, la règle qui a l'indice le plus petit, on supprimerait les 6 règles R15, R18, R19, R21, R24, R25 et certainement une partie des autres de R1 à R14, si on avait plus de cinq décimales.

[5] Pour établir cette égalité, il suffit de réduire au même dénominateur b et (b+a)(b+d)/N, puis de développer le numérateur après avoir remplacé N par a+b+c+d.

[6] Ce choix a pour but de privilégier un graphe simple, sans boucle ni cycle, permettant des raisonnements de type causal, où un fait ne peut pas être à la fois la cause et la conséquence d'un autre.

Si on fait abstraction des valeurs des coefficients, on peut regrouper les équivalences[7] en 4 classes d'variables :

- Cancer <--> TbOrCa <--> XRay <--> Dyspnea, (R7, R8, R9, R10, R13, R14, R19, R20, R21, R22, R23, R24)
- TbOrCa <--> Tuberculosis <--> XRay (R13, R14, R15, R16, R17, R18)
- Bronchitis <--> Dyspnea <--> Smoker (R3, R4, R5, R6, R11, R12)
- Bronchitis <--> Non Age2 <--> Smoker (R1, R2, R5, R6, R25, R26)

Règles entre deux variables ayant leur réciproque au seuil 0,95			Règles sans leur réciproque au seuil 0,95		
Numéros	Composition en variables	Indices	N°	Composition en variables	Indice
R1;R2	Non Age2 <--> Smoker	1 ; 1	R27	Cancer --> Age2	1
R3;R4	Bronchitis <--> Dyspnea	1 ; 1	R28	TbOrCa --> Age2	1
R5;R6	Bronchitis <--> Smoker	1 ; 1	R29	Cancer --> Smoker	0,99999
R7;R8	Cancer <--> TbOrCa	1 ; 1	R30	Tuberculosis --> Dyspnea	0,99997
R9;R10	Cancer <--> XRay	1 ; 1	R31	TbOrCa --> Smoker	0,99993
R11;R12	Dyspnea <--> Smoker	1 ; 1	R32	XRay --> Age2	0,99971
R13;R14	TbOrCa <--> XRay	1 ; 1	R33	XRay --> Smoker	0,98983
R15;R16	TbOrCa <--> Tuberculosis	0,99999 ;1	R34	Tuberculosis-->nonBronchitis	0,96669
R17;R18	Tuberculosis <--> XRay	1 ; 0,99715	R35	Cancer --> Bronchitis	0,95884
R19;R20	Dyspnea <--> TbOrCa	0,99011 ; 1			
R21;R22	Dyspnea <--> XRay	0.98099 ; 1			
R23;R24	Cancer <--> Dyspnea	1 ; 0,96362			
R25;R26	Non Age2 <--> Bronchitis	0,991 ;0,997			

TAB. 7 – *Liste des règles d'indice d'implication supérieur à 0,95*

Dans la figure 1, on a représenté au sein du même disque les variables qui sont équivalentes entre elles deux à deux.

[7] Deux variables A et B sont liées par une règle d'équivalence, si les règles A→B et B→A ont une valeur à l'indice d'implication supérieure à 0,95. Cette équivalence étant approximative et non exacte, elle n'est pas transitive, comme on le voit dans le paragraphe suivant, ce qui interdit de parler de classes d'équivalence.

Graphe de règles d'implication statistique pour le raisonnement courant.

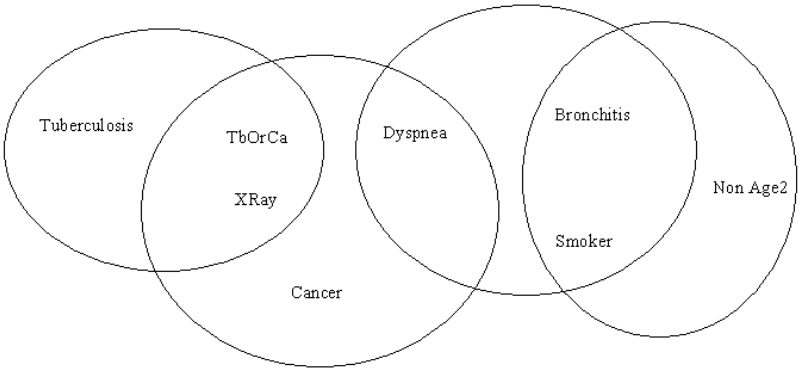


FIG. 1 – *Liste des règles d'indice d'implication supérieur à 0,95*

Les règles A→B et B→C et leur raccourci A→C. En logique formelle, quand les règles A⇒B et B⇒C sont vraies, la règle A⇒C est vraie. Pour le montrer, il suffit de faire la table de vérité croisant A, B et C. L'implication logique n'est fausse que si sa partie gauche est vraie et sa partie droite fausse, et la conjonction (« et ») est fausse dès que l'un des deux est faux. On n'a indiqué dans le tableau 8 que les valeurs fausses pour les 4 dernières colonnes, les valeurs manquantes étant vraies.

On constate que les deux implications A⇒B et B⇒C sont simultanément vraies seulement dans les cas n° 1, 5, 7 et 8, et que la règle A⇒C est vraie dans ces 4 cas. Cette propriété qui associe à deux règles qui se suivent leur raccourci, s'appelle la *transitivité* et fait partie des règles d'un niveau supérieur, les règles d'inférence, qui au lieu d'enchaîner des faits, enchaînent des règles. Elle est un des fondements du raisonnement déductif en logique formelle, mais également en logique courante. Malheureusement, la transitivité des règles (d'équivalence comme d'implication) n'est pas assurée. La figure 1 permet de l'illustrer : si on part de Tuberculosis, situé à gauche, on a la règle Tuberculosis→TbOrCa, TbOrCa→Dyspnea, puis Dyspnea→Bronchitis, mais au lieu d'avoir la règle Tuberculosis→Bronchitis qui en est le raccourci, on a la règle contraire Tuberculosis→non Bronchitis. Pour éviter ce problème qui peut gêner le raisonnement, Régis Gras a choisi de ne garder deux règles qui se suivent que si leur raccourci est de valeur supérieure à 0,5, et de supprimer le raccourci du graphe, dans la mesure où le lecteur le rétablira de lui-même automatiquement s'il a besoin de cette règle.

Cas n°	A	B	C	A⇒B	B⇒C	A⇒B et B⇒C	A⇒C
1	Vrai	Vrai	Vrai				
2	Vrai	Vrai	Faux		Faux	Faux	Faux
3	Vrai	Faux	Vrai	Faux		Faux	
4	Vrai	Faux	Faux	Faux		Faux	Faux
5	Faux	Vrai	Vrai				
6	Faux	Vrai	Faux		Faux	Faux	
7	Faux	Faux	Vrai				
8	Faux	Faux	Faux				

TAB. 8 – *Table de vérité pour 3 variables booléennes*

Pour représenter plus facilement l'ensemble des règles, on utilise l'indice d'implication corrigé[8], on retire les réciproques de moindre valeur et on obtient les 17 règles du tableau 9 ayant la valeur de cet indice supérieure à 0,95, dont il faut retirer les raccourcis pour obtenir la représentation de la figure 2.

R1	Cancer --> TbOrCa	1	R10	Dyspnea --> Bronchitis	0,97592
R2	Tuberculosis --> TbOrCa	1	R11	TbOrCa --> Age2	0,97194
R3	Cancer --> XRay	0,99996	R12	Cancer --> Smoker	0,96957
R4	TbOrCa --> XRay	0,99975	R13	TbOrCa --> Smoker	0,96605
R5	Tuberculosis --> XRay	0,99842	R14	XRay --> Dyspnea	0,96189
R6	Cancer --> Dyspnea	0,97979	R15	XRay --> Age2	0,96095
R7	Cancer --> Age2	0,97972	R16	Tuberculosis --> non Bronchitis	0,95679
R8	TbOrCa --> Dyspnea	0,97937	R17	XRay --> Smoker	0,95453
R9	Tuberculosis --> Dyspnea	0,97856			

TAB. 9– *Les règles d'indice d'implication corrigé supérieur à 0,95, sans réciproques.*

On supprime R3 qui est le raccourci de R1 et R4, ainsi que R5 celui de R2 et R4. On supprime R6 qui est le raccourci de R1 et R8, ainsi que R9 celui de R2 et R8. Bien que l'indice de 0,54 du raccourci de R2 et R11 soit trop petit pour qu'il figure dans le tableau 8 des règles d'indice supérieur à 0.95, sa valeur supérieure à 0.5 lui permet de légitimer la co-existence de ces deux règles R2 et R11. De ce fait, R7, le raccourci de R1 et R11, est supprimé. La règle R10 (avec les règles R2, R8, R16) pose un problème déjà vu précédemment en illustration de la transitivité. On la supprime. Le raccourci des règles R2 et R13 étant d'indice 0.66, la règle R13 est conservée, et la règle R12 est supprimée en tant que raccourci des règles R1 et R11. A ce stade, on a le graphe à gauche de la figure 2. La prise en compte des trois règles R14, R15 et R17 produit la suppression des règles R8, R11 et R13, qui se trouvent être respectivement les raccourcis des règles R14, R15 et R17 avec la règle R4. Il n'y a plus qu'à rajouter la règle R16, et on obtient le graphe à droite de la figure 2.

[8] Cette correction permet de redonner à l'indice d'implication le pouvoir discriminant qu'il a tendance à perdre en cas de données nombreuses, comme c'est le cas des données Asia (voir tableau 7). Toutefois pour l'étude des propriétés de l'indice, nous nous référons à sa version non corrigée.

Graphe de règles d'implication statistique pour le raisonnement courant.

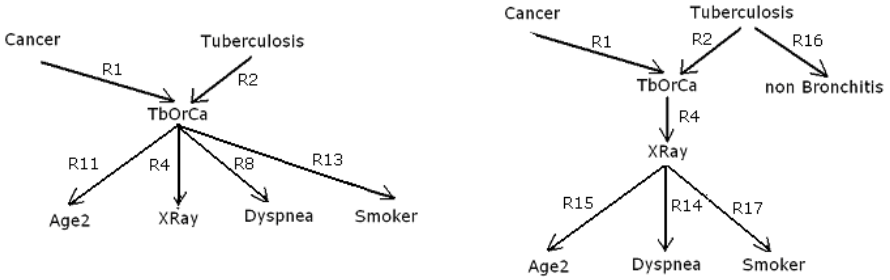


FIG. 2 – *Graphe des implications statistiques (indice corrigé). A gauche une étape intermédiaire de la construction du graphe et à droite le graphe terminé.*

Interprétation et causalité La lecture du schéma de la figure 2 fait apparaître que les implications statistiques, ne peuvent pas s'interpréter de façon directement causale. En effet, ce n'est bien sûr pas le cancer qui cause une augmentation de l'âge des sujets (règle obtenue en suivant les règles R1, R4 et R15), mais éventuellement l'inverse, l'âge avancé qui est une des causes du cancer. Dans le graphe, une bonne moitié des arcs ne peuvent s'interpréter en terme de causalité alors que leurs réciproques pourraient l'être, mais elles ont été éliminées au profit de la règle directe, qui avait une plus grande valeur d'implication. Ce problème d'interprétation n'est pas caractéristique de l'implication statistique : réussir à retrouver les effets et les causes à partir des résultats d'observations faites à un moment donné sans prise en compte d'information supplémentaire est un défi qu'aucune méthode automatique de traitement de données[9] n'a, à notre connaissance, été capable de relever à ce jour.

Pour ce qui est de l'implication statistique, dans sa forme classique, nous avons vu qu'elle est issue du raisonnement mathématique, et qu'elle est équivalente à sa contraposée, ce qui n'est plus le cas dans sa forme entropique (cf. Partie 1). La contraposée de « Cancer → Age2 » est « non Age2 → non Cancer », qui peut se traduire par « un jeune âge a pour effet l'absence de cancer ». Cette traduction en termes de causalité n'est plus du tout choquante car la règle obtenue a une variable explicative, qui est l'âge, en partie gauche et non en partie droite comme précédemment. Ainsi dès qu'une règle d'implication statistique A→B est difficile à interpréter en terme de causalité, alors que sa réciproque B→A serait interprétable, plutôt que de la remplacer par sa réciproque de valeur d'implication inférieure, il est mieux de prendre sa contraposée nonB→nonA qui a la même valeur.

Cas limites : Valeurs implicatives des règles A→A, Faux→A, A→Vrai. Ces règles sont vraies en logique formelle car elles n'admettent aucun contre-exemple. En effet elles ne peuvent avoir leur partie gauche à Vrai et leur partie droite à Faux. Pour calculer leur valeur d'implication statistique, on reprend les valeurs a, b, c et d du tableau 3 qui vérifient b=c=0 pour la première, a=b=0 pour la deuxième, et b=d=0 pour le troisième. L'espérance de b en cas d'indépendance entre partie gauche et droite de la règle est ad/(a+d) dans le premier cas, et 0 dans les deux autres cas. Le calcul de l'indice d'implication statistique est impossible dans ces deux derniers cas, et il donne un résultat très proche de 1 dans le premier cas. Pour les deux derniers cas, on peut choisir 1 si on désire que la règle d'implication statistique

[9] Les sciences expérimentales ont établi des protocoles rigoureux pour essayer d'établir quelles sont les causes parmi un ensemble de causes possibles (cf. exemples dans Cadot 2006).

ressemble le plus possible à la règle logique, 0,5 ou une valeur inférieure si on désire éliminer ces règles sans intérêt statistique. Ce type de calcul peut se produire quand la règle A→B concerne deux variables dont les valeurs sont les mêmes pour tous les sujets (calcul identique à celui de A→A), quand la variable A est fausse pour tous (A→B devient Faux→B) ou quand la valeur de B est vraie pour tous (A→B devient A→Vrai). Ces cas « limites », sous cette forme ou sous une autre équivalente[10] ont fait l'objet de développements dans les articles et livres consacrés à l'implication statistique.

4 Les treillis de Galois

Les *treillis de Galois* sont une représentation algébrique des données d'un tableau booléen de type SujetsXVariables (voir la table à gauche de la figure 5).

1. Définition du contexte
V={a, b, c, d, e}
S={1, 2, …, 7}

	a	b	c	d	e
1	x	x	x	x	x
2		x		x	x
3		x			x
4		x		x	
5	x	x	x		
6	x				
7					

Table de la relation R

2. Calcul de g({d, e})

g({d})={1, 2, 4}
g({e})={1, 2, 3}
donc
g({d, e})={1, 2}

En effet,
g({d, e})=
g({d}∪{e})=
g({d})∩g({e})=
{1, 2, 4}∩{1, 2, 3}=
{1, 2}

3. Calcul de f({1, 2})

f({1}) = {a, b, c, d, e}
f({2}) = {b, d, e}
donc
f({1, 2}) = {b, d, e}

4. Calcul de la fermeture de {d,e}
f•g({d, e})=f{1, 2}={b, d, e}

{d,e} n'est pas fermé, car il est différent de sa fermeture {b, d, e}, rectangle « maximal » grisé dans la table. {b,d,e} est un fermé.

5. Les fermés de 2^V sont ∅, {a}, {b}, {b,e}, {b,d}, {b,d,e}, {a,b,c}, {a,b,c,d,e}
Les fermés de 2^S sont {1,2,…,7}, {1,5,6}, {1,2,…,5}, {1,2,3}, {1,2,4}, {1,2}, {1,5},{1}

FIG. 3 – *Petit exemple de treillis de Galois, détail de calcul d'une fermeture.*

Ces notions, dénommées « treillis de Galois » par Barbut et Monjardet (1970) ou « treillis de concepts » par Wille (1982) sont un des objets d'études de la communauté de recherche « FCA » (Formal Concept Analysis : http://www.upriss.org.uk/fca/fca.html). Nous allons nous contenter ici de décrire la partie du formalisme des treillis de Galois qui concerne la construction des règles logiques entre variables. Puis nous comparerons les règles extraites selon cette vision algébrique des données à l'implication statistique qui provient d'une vision statistique des mêmes données. Cette comparaison se fera dans deux directions : nous examinerons d'abord les propriétés qui découlent de leurs définitions formelles, puis le type d'information qu'elles produisent sur les données Asia.

Définitions. Etant donnés un ensemble de variables V, un ensemble de sujets S, et une relation booléenne R qui les lie, les « treillis de Galois » sont les images des deux *treillis*

[10] Notamment on peut lire dans Gras et al. (2005) que le prolongement par continuité de l'indice d'implication statistique justifie l'attribution de la valeur 0 à la règle A→Vrai.

d'ensemble (voir définition en annexe et dans Davey 1990) associés à V et S par la *correspondance de Galois* associée à R (pour plus de détails voir Mephu Nguifo 1994). Cette « correspondance de Galois » est formée de deux parties duales f et g : f fait correspondre à chaque sous-ensemble de sujets le sous-ensemble de variables qui leur sont liées par R, et g fait correspondre à chaque sous-ensemble de variables le sous-ensemble de sujets qui leur sont liés par R. La figure 3 contient une illustration de ces définitions sur un petit exemple avec 5 variables et 7 sujets avec à gauche le graphe de la relation R (fig 3.1) et à droite un exemple de calcul des correspondances f et g (fig 3.2 et fig 3.3).

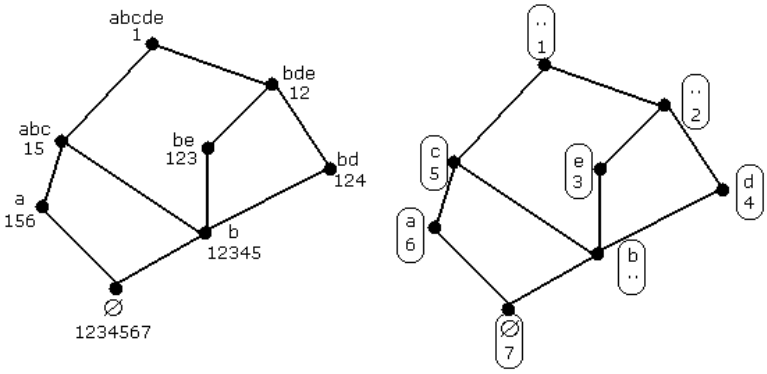


FIG. 4 – *Diagramme de Hasse du treillis de Galois du petit exemple de la figure 3, complet à gauche, simplifié à droite*

La composée des deux est une application qui associe à tout sous-ensemble un sous ensemble le contenant de même nature : de variables si on applique g puis f (fig. 3.4), de sujets si on applique f puis g. Ces deux composées sont appelées des *fermetures*, et les sous-ensembles qui leur sont stables sont appelés des *fermés* (cf. fig. 3.4 et 3.5). La fermeture étant compatible avec l'inclusion, l'intersection et la réunion, les treillis de Galois se trouvent être les deux ensembles de fermés (pour les variables et pour les sujets), munis des opérations ensemblistes et sont isomorphes. Pour retrouver les démonstrations de ces propriétés, on peut se reporter à (Wille 1982). La structure de treillis est représentée par un diagramme de Hasse, qui joint les éléments qui sont ordonnés selon la relation d'ordre, en allant des plus petits (au sens de l'inclusion) aux plus grands, en omettant les raccourcis liés à la transitivité. Comme nous avons deux treillis qui se correspondent, un seul diagramme peut exprimer les deux, chaque élément (ou concept) étant formé d'un ensemble de variables et d'un ensemble de sujets. Afin de faciliter les comparaisons avec l'implication statistique, nous choisissons l'ensemble de variables comme référence et les nœuds sont écrits sur deux lignes, la première relative aux variables, la seconde aux sujets (cf Fig. 4). En bas du diagramme figure le fermé correspondant à l'ensemble de tous les sujets, et en haut celui correspondant à l'ensemble de toutes les variables

Les notions de fermeture et de *règle logique*, c'est-à-dire qui n'admet pas de contre-exemple, sont très liées. En voici une justification rapide (pour plus de détails, voir Guigues et al. 1986) : si un ensemble B de variables, par exemple B={x,y,z,t},contient un ensemble A, par exemple A={y,z}, l'ensemble g(B) est formé des sujets vérifiant toutes les variables

de B, ceux-ci vérifiant nécessairement toutes celles de A, ce qui fait que si A⊂B, g(B)⊂g(A). Cette inclusion est stricte s'il existe des sujets qui vérifient toutes les variables de A sans vérifier toutes celles de B, par exemple qui vérifient y et z mais pas x. Quand ce cas ne se produit pas, A et B ont même fermeture et si on appelle C leur fermeture, par exemple C={x,y,z,t,u,v}, l'inclusion g(C)⊂g(B)⊂g(A) est en fait une égalité. Tous les sujets vérifiant les variables de A vérifient donc aussi celles de B et de B-A (variables de B ne figurant pas dans A), c'est-à-dire qu'il n'y a pas de contre-exemple à la règle A→B-A, qui s'écrit ici {y,z}→{x,t}, ou plus simplement yz→xt. Nous l'appelons règle logique pour la différencier de la règle d'implication statistique.

Nous venons de définir une règle logique à l'aide des deux ensembles emboîtés A et B de même fermeture C. Il peut y en avoir d'autres définies sur le même ensemble de sujets, ce sont par exemple les règles yz→x, yz→t, yz→u, yz→v, yz→uv, xuz→u, yzt→v, etc, s'écrivant de façon plus générale A∪D→E, où D et E sont deux parties disjointes[11] de l'ensemble C-A contenant les quatre variables x, t, u et v. La partie gauche de ces règles est un ensemble non fermé, appelé *lacune* par Guigues et Duquenne (1986). Ces auteurs ont proposé un parcours du treillis de Galois permettant d'extraire seulement une partie génératrice des règles logiques : pour chaque ensemble fermé C de variables, ils cherchent les lacunes A dont C est la fermeture et créent, si certaines conditions sont vérifiées, les règles correspondantes sous la forme A→C-A, ce qui donne par exemple la règle yz→xtuv. Pour chacune de ces règles, on peut déduire les autres règles construites sur les mêmes sujets en retirant des variables du membre droit de la règle, et en en mettant éventuellement une partie dans le membre gauche. Nous avons déjà évoqué dans la partie précédente au sujet de la transitivité ce procédé, appelé *règle d'inférence* qui, appliqué à des règles d'un ensemble, permet d'en créer de nouvelles. Parmi les règles d'inférence proposées par Guigues et Duquenne, il y a, outre la transitivité, deux règles qui permettent de modifier le nombre de variables des membres gauches et droits des règles, une comme celle que nous venons de voir à partir d'une seule règle, et l'autre à partir de deux. Si on applique l'algorithme proposé par Guigues et Duquenne au cas du petit exemple de la figure 3, on obtient les 6 règles e→b, d→b, c→ab, ab→c, ae→d, cd→e, qui suffisent[12] pour engendrer au moyen des règles d'inférence les 72 règles logiques associées à la relation R.

On ne peut pas lire directement les règles sur le diagramme de Hasse mais il est possible de les obtenir à partir d'une simplification de ce diagramme obtenue en retirant pour chaque nœud les variables figurant dans les nœuds en dessous et les sujets figurant dans les nœuds au dessus, comme dessiné à droite dans la figure 4. Si une arête relie deux nœuds contenant chacun une variable, la règle correspondante s'écrit en parcourant l'arête du haut vers le bas : pour la figure 4, on obtient c→a, c→b, e→b et d→b. On peut retrouver les règles plus complexes en remontant dans le diagramme à partir de deux nœuds (partie gauche de la

[11] On choisit, en général, une partie E non vide, bien que la règle {y,z}→∅ soit cohérente avec le formalisme algébrique.

[12] Les règles selon le formalisme de Guigues et Duquenne s'écrivent en mettant en partie droite la réunion des parties gauche et droite habituelles, ce qui donne e→be, d→bd, c→abc, ab→abc, ae→ade, cd→cde. Les trois règles d'inférence sont mr1, qui est la transitivité, mr2, qui transforme une règle par adjonction de variables quelconques, les mêmes en partie gauche et droite, et mr3, qui associe à deux règles la règle formée de la réunion des deux parties gauches et droites. Il est à noter qu'il peut arriver que certaines règles générées ne soient vérifiées par aucun sujet. Ce n'est pas le cas des 72 règles de ce petit exemple qui sont toutes vérifiées par au moins un sujet : le sujet 1.

règle) jusqu'à leur maximum en suivant les arêtes, puis en redescendant dans un certain nombre de nœuds en suivant les arêtes à partir de ce maximum (partie droite de la règle). Ainsi dans la figure 4, si on remonte à partir de a et b, on arrive à c ce qui donne la règle ab→c, et si on part de c et d, on doit remonter jusqu'en haut, ce qui fait qu'on peut alors redescendre dans chaque nœud et on obtient par exemple la règle cd→abe.

Réciproques et équivalences. Nous avons vu dans le paragraphe précédent de définitions du treillis de Galois que le diagramme de Hasse simplifié (cf. à droite de la figure 4) se lit plus souvent du haut vers le bas qu'inversement pour trouver les règles logiques. Du fait de ces deux sens possibles de lecture, il découle qu'un jeu de règles peut contenir une règle et sa réciproque. Par exemple les deux règles c→ab et ab→c ont été trouvées à partir du petit exemple de la figure 3. Plus généralement, les règles A→B et B→A coexistent si et seulement si les deux ensembles disjoints de variables A et B sont vérifiés par le même ensemble de sujets. Il peut arriver que ces deux ensembles de variables forment un seul nœud, ce qui rajoute aux deux façons que nous avons exposées de lire des règles à partir du diagramme de Hasse simplifié une troisième façon : boucler (c'est-à-dire joindre un nœud à lui-même) sur les nœuds qui contiennent plus d'une variable.

Règles non informatives. Dans le cas où l'ensemble de variables B est inclus dans l'ensemble A, les sujets qui vérifient A vérifient également B, ce qui permet d'écrire la règle A→B, par exemple xy→x. Cette règle est appelée règle « non informative » par Guigues et Duquenne (1986). La règle A→A fait également partie de ces règles non informatives. D'autres règles n'apportent pas non plus d'information, comme la règle A→B, où A est un ensemble quelconque de variables, et B un ensemble de variables vérifiées par tous les sujets (notamment, B peut se réduire à l'ensemble vide), ainsi que la règle A→B, où A est un ensemble de variables vérifié par aucun sujet et B un ensemble quelconque de variables.

Variables et négation. La structure de treillis de Galois n'est pas compatible avec la négation des variables. Notamment le treillis des négations peut comporter un nombre de nœuds différent (cf. figure 5). Tant que les règles du treillis des variables ont une variable en partie gauche et une en partie droite, on retrouve leurs contraposées dans le treillis des négations. Ce n'est pas toujours le cas des règles plus complexes, comme ab→c qui se réécrit « a nonc→non b » et « b non c →non a », aucune de ses règles, composées en partie seulement de négations de variables, ne figure dans le treillis des négations.

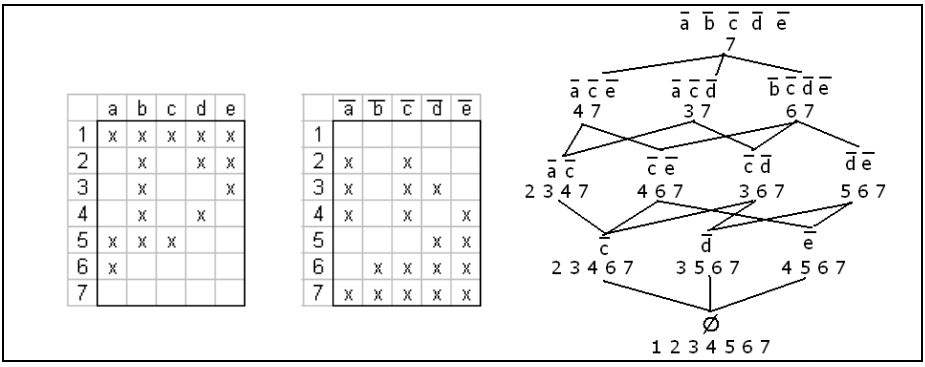


FIG. 5 – *A gauche, variables du petit exemple de la figure 2, au milieu leurs négations, à droite le diagramme de Hasse des négations.*

Application aux données Asia. Nous avons construit le treillis de Galois des données qui comporte 140 fermés et l'algorithme de Guigues et Duquenne de recherche d'une partie génératrice de règles nous a fourni 17 règles. Les voici, avec les noms de variables sous forme de deux lettres (la correspondance entre ces deux lettres et l'intitulé est faite dans le tableau 1), suivies de leur support, c'est-à-dire du nombre de sujets qui vérifient leurs parties gauche et droite, entre parenthèses :
TU→TC (83), CA→TC (271), TC VA→DY XR (7), DY XR VA→ TC (7), DY SM VA→ BR (13), BR CA→XR (139), AG SM TU→XR (14), CA TU→AG XR (2), AG BR SM TC→XR (87), AG BR SM VA→DY (6), CA VA→AG (1), BR VA XR→DY SM TU (1) et AG BR SM TU→DY XR (7), BR CA TU→DY SM VA (0), CA TU VA→BR SM (0), CA DY SM TU→BR VA (1) et AG BR TC VA→CA (0)

Notons que parmi ces règles, quatre ont un support nul, qui proviennent du fermé correspondant à toutes les variables et à aucun sujet, élément maximal du treillis de Galois, nœud au sommet du diagramme de Hasse. Bien qu'elles paraissent avoir peu d'intérêt, n'étant vérifiées par aucun sujet, elles sont correctes du point de vue algébrique, dans la mesure où elles n'admettent aucun contre-exemple. Avec les règles d'inférence, elles produisent des règles de support nul également. En combinant les 17 règles au moyen des règles d'inférence, on retrouve les 2992 règles exactes présentes dans les données Asia dont 514 de support non nul.

Comparaison des deux jeux de règles (implication statistique et règles logiques) extraits des données Asia. Les deux jeux de règles ont en commun seulement deux règles, qui sont TU (Tuberculosis) → TC (TbOrCa) ; CA (Cancer) → TC (TbOrCa). Parmi toutes les règles d'implication statistique, ce sont les seules qui ont un indice d'implication statistique corrigé égal à 1 (voir tableau 9). Et parmi toutes les règles logiques, ce sont les seules possédant une seule variable en partie gauche, et elles font partie des 17 règles génératrices. Pour faciliter la comparaison des règles qui ne sont pas communes aux deux jeux, on se limite à une partie représentative de ces règles pour chacun des deux modèles. Dans le dernier, on ne considèrera que les 13 règles logiques de support non nul de la partie génératrice, et dans le premier que les 7 règles de la figure 2.

Graphe de règles d'implication statistique pour le raisonnement courant.

On peut remarquer dans la figure 2 que la variable VA (Visit in Asia) a disparu du graphe d'implication statistique alors qu'elle est présente dans la partie gauche de presque la moitié des règles logiques de la partie génératrice. Parmi les 7033 sujets de la base de données, seuls 74 ont une valeur de 1 à cette variable, donc une partie des règles logiques a été construite sur environ 1% des données. Cette variable est présente dans la partie gauche des 6 règles TC VA→DY XR (7), DY XR VA→ TC (7), DY SM VA→ BR (13), AG BR SM VA→DY (6), CA VA→AG (1), BR VA XR→DY SM TU (1) de supports compris entre 1 et 13. Cela montre la différence de point de vue qui existe entre la représentation des données par un treillis de Galois et celle par un graphe d'implication statistique : dans le premier cas on privilégie l'extraction des règles qui caractérisent de façon exacte des groupes de sujets, quelle que soit la taille de ces groupes, même réduits à une personne[13], dans le second on cherche des relations suffisamment « fortes » (avec un fort indice d'implication statistique) entre variables pour que quelques sujets particuliers les contredisant ne puissent pas les remettre en cause.

Les règles des deux types, implication statistique et règle logique, sont créées à partir des mêmes comptages de sujets, et il en résulte une certaine ressemblance dans leur orientation (membre gauche et droit). Dans la figure 2, on peut remarquer qu'une variable comme Age (resp. Smoker, Dyspnea, non bronchitis) vérifiée par un nombre de sujets égal à 3391 (resp. 3271, 2964, 3622), est en nœud terminal, donc dans le membre droit des règles, alors qu'une variable comme Cancer (resp. TbOrCa, Tuberculosis, Xray) qui a un nombre de sujets égal à 271 (resp. 352, 83, 685) est dans un nœud non terminal, et apparaît dans le membre gauche d'une plus grande proportion de règles. On retrouve ce phénomène dans les règles logiques issues du treillis de Galois : nous avons décrit précédemment le parcours du diagramme de Hasse permettant de retrouver les règles logiques (cf. figure 4), dont le départ est généralement dans des nœuds situés plus haut que ceux de l'arrivée. Comme les nœuds les plus bas correspondent à plus de sujets, on obtient ainsi une plus grande proportion de règles logiques avec des variables fréquentes en membre droit qu'en membre gauche.

5 Les réseaux bayésiens

La création de ce type de réseau répond, d'après Olfa Ben Naceur-Mourali et Christophe Gonzales (2004) à un besoin d'assouplissement des systèmes experts à base de règles, ces derniers ne fonctionnant qu'avec des faits certains. D'après les auteurs, MYCIN, créé en 1976 par Shortliffe, fut une première tentative d'introduction de l'incertitude au moyen de "facteurs de certitude", mais la première formalisation véritablement opérationnelle de cette incertitude par des probabilités date de Pearl (1988) : ce sont les réseaux bayésiens, qui sont une représentation par un graphe de liens probabilistes entre faits.

Reprenons l'exemple très pédagogique que donne Jensen dans son introduction aux réseaux bayésiens (Jensen 1996) en nous autorisant toutefois une certaine liberté d'adaptation de l'histoire qu'il raconte :

« Sherlock Holmes sort de chez lui le matin pour aller travailler. En arrivant à sa voiture, il constate que sa pelouse est mouillée alors qu'il ne pleut pas. Il se dit qu'il a dû oublier de

[13] Dans le formalisme de treillis de Galois, il est même licite d'envisager un groupe particulier de sujets, le groupe « vide », qui vérifie toutes les règles de support nul.

couper son système d'arrosage la veille au soir. Il se dirige vers la cave pour aller l'arrêter quand il jette un coup d'oeil sur la pelouse de son voisin Watson : elle est mouillée. Il rebrousse alors chemin et monte dans sa voiture pour se rendre à son travail. »

Nous allons procéder en cinq étapes pour montrer la modélisation de ce raisonnement par un réseau bayésien : 1) définir un réseau de faits et de règles sur lequel le raisonnement peut s'appuyer, 2) lui ajouter les probabilités conditionnelles pour en faire un réseau bayésien, et montrer le raisonnement « direct » de Holmes aboutissant à sa première décision 3) Présenter le moteur du réseau bayésien, formé d'hypothèses et de formules qui ne seront pas remises en cause 4) montrer le raisonnement « bayésien » de Holmes aboutissant à la deuxième décision et 5) à la dernière décision.

Raisonnement causal avec quatre faits et trois règles. Son raisonnement peut être représenté comme un parcours dans un réseau causal comportant quatre faits et trois relations, voici les quatre faits qui peuvent prendre la valeur "Vrai" ou "Faux" :

- H : La pelouse de Holmes est mouillée
- W : La pelouse de Watson est mouillée
- A : Le système d'arrosage de Holmes n'a pas été coupé
- P : Il a plu

Et voici les trois relations causales qui les lient :

- P→H : s'il a plu, la pelouse de Holmes est mouillée
- P→W : s'il a plu, la pelouse de Watson est mouillée
- A→H : si le système d'arrosage de Holmes n'a pas été coupé alors la pelouse de Holmes est mouillée

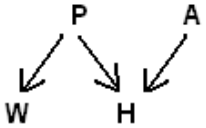


FIG. 6 – *Petit exemple de réseau bayésien emprunté à Jensen (1995)*

Holmes utilise l'information au fur et à mesure qu'elle arrive pour évaluer la probabilité qu'il ait oublié de couper l'arrosage. Ainsi, il prend trois décisions successives contradictoires:

- Le risque que A soit vrai est a priori faible. Il va directement à sa voiture
- Il constate que H est vrai. Cela augmente suffisamment à ses yeux le risque que A soit vrai pour justifier un détour par la cave
- Il constate que W est vrai. Cela diminue suffisamment à ses yeux le risque que A soit vrai pour ne plus justifier un détour par la cave. Il rebrousse chemin et monte dans sa voiture.

Les probabilités fixées au départ. Pour modéliser son comportement, il faut ajouter des probabilités. On associe à chaque fait ayant plusieurs causes une table de probabilités donnant toutes les probabilités de ce fait conditionnellement aux autres. Les probabilités proposées par l'auteur pour ces faits sont dans le tableau 10.

Graphe de règles d'implication statistique pour le raisonnement courant.

P=1	0.2
P=0	0.8

P(P)

A=1	0.1
A=0	0.9

P(A)

	P=1	P=0
W=1	1	0.2
W=0	0	0.8

P(W|P)

	P=1		P=0	
	A=1	A=0	A=1	A=0
H=1	1	1	0.9	0
H=0	0	0	0.1	1

P(H|P,A)

TAB. 10 – *Les probabilités conditionnelles des 4 faits de la figure 6*

Explicitons ces tableaux. Celui de gauche n'a que deux lignes car aucun arc n'arrive à P (cf. figure 6}, ce qui signifie qu'il n'est l'effet d'aucune cause dans ce modèle. On a en première ligne la probabilité 0.2 que P soit vrai (P(P=1)=0.2 en notant Vrai par 1 et Faux par 0), donc qu'il ait plu cette nuit. Et dans la ligne suivante 0.8, qui est la probabilité qu'il n'ait pas plu. Pour toutes les colonnes de ces tableaux la somme sera 1. A côté on a le même type de tableau pour A. Notons que Holmes ne fait pas de détour pas la cave car il estime la valeur de P(A=1) de 0.1 trop faible pour prendre la peine de ce détour. Le tableau de P(W|P) contient 4 valeurs. En haut à gauche, c'est la probabilité que W soit vraie, sachant que P est vraie. Elle est de 1, ce qui signifie que s'il a plu, on a la certitude que la pelouse de Watson est mouillée. Par contre s'il n'a pas plu, il y a quand même une probabilité de 0.2 pour que la pelouse de Watson soit mouillée, certainement parce qu'il l'arrose parfois lui aussi. Le dernier tableau s'interprète de la même façon en fonction des deux causes que sont A et P. Que Holmes ait oublié ou non le système d'arrosage ouvert (A=1 ou 0), s'il a plu, on est certain que sa pelouse est mouillée. Par contre s'il n'a pas plu, la probabilité pour que la pelouse soit mouillée est nulle s'il n'a pas arrosé et de 0.9 dans le cas contraire.

Le « moteur » probabiliste du raisonnement. Pour pouvoir faire fonctionner le modèle, il faut calculer les probabilités conjointes[14], marginales.

P=1	0.2
P=0	0.8

P(P)

A=1	**0.1**
A=0	0.9

P(A)

	P=1	P=0	
W=1	0.2	0.16	0.36
W=0	0	0.64	0.64

P(W,P) P(W)

	P=1		P=0		
	A=1	A=0	A=1	A=0	
H=1	0.02	0.18	0.07	0	0.27
H=0	0	0.1	0.01	0.72	0.73

P(H,P,A) P(H)

TAB. 11 – *Les 4 probabilités marginales et les 2 conjointes déduites du tableau10*

Les valeurs des probabilités des lois conjointes sont calculées d'après les formules de la note de bas de page 10, et celles des deux lois marginales P(W) et P(H) sont obtenues en sommant les cellules des lignes des lois conjointes. Ces formules sont le « moteur » du réseau bayésien. Elles ne changeront pas pendant son fonctionnement. Par contre les probabilités de P (il a plu) et de A (le système d'arrosage était en marche) sont des probabilités a priori, susceptibles de changement. C'est d'ailleurs un des buts du réseau bayésien que d'aider à les mettre à jour en fonction de nouvelles informations, et d'en déduire leurs probabilités a posteriori.

Prise en compte de l'information « H vrai ». On remonte l'*arc* AH (ce qu'on appelle arc est un lien orienté de A vers B correspondant à la règle A→H) de la conséquence H à la

[14]Loi conjointe de P et W : P(W,P)=P(W|P)P(P), Loi conjointe de A, P et H : P(H,P,A)=P(H|P,A)P(P,A), et avec A et P indépendants, P(P,A)=P(P)P(A).

cause A en utilisant la formule de Bayes (aussi appelée de *probabilités des causes*)[15]. Pratiquement cela consiste à multiplier les probabilités conjointes de P(H,P,A) par un coefficient pour que la première ligne ait pour somme 1 (c'est le quotient de la nouvelle valeur sur l'ancienne de P(H=1)), et pareillement pour que la deuxième soit de somme nulle. Une fois les valeurs de P(H,P,A) réactualisées, on réactualise aussi celles de P(A) et P(P) par sommation et on reprend le sens "normal" de l'arc PW pour réactualiser P(W,P), ce qui donne les valeurs du tableau 12.

					P=1	P=0				P=1		P=0			
										A=1	A=0	A=1	A=0		
P=1	0.74	A=1	**0.34**	W=1	0.74	0.05	0.79	H=1	0.07	0.66	0.26	0	1
P=0	0.26	A=0	0.66	W=0	0	0.21	0.21	H=0	0	0	0	0	0
P(P)		P(A)		P(W,P)			P(W)	P(H,P,A)					P(H)

TAB. 12 – *Modification du tableau 11 prenant en compte la certitude que H est vrai*

On constate à la lecture de ce tableau que P(A=1) est passé à 0.34, alors qu'il était de 0.10 dans le précédent. Ainsi la prise en compte de l'information sur H a augmenté la probabilité que A soit vrai, ce qui a poussé Holmes à faire un détour par sa cave.

Prise en compte de l'information supplémentaire « W vrai ». On met à jour les valeurs de la même façon que précédemment (cf. tableau 13) en commençant par la loi conjointe de P(W,P), en réactualisant celles de P(P) et P(A) par sommation, et en terminant pas celle de P(H,P,A). La probabilité de A[16] redescend à 0.16, et Holmes juge le risque que A soit vrai suffisamment faible pour repartir sans passer par la cave.

					P=1	P=0				P=1		P=0			
										A=1	A=0	A=1	A=0		
P=1	0.93	A=1	**0.16**	W=1	0.93	0.07	1	H=1	0.09	0.84	0.07	0	1
P=0	0.07	A=0	0.84	W=0	0	0	0	H=0	0	0	0	0	0
P(P)		P(A)		P(W,P)			P(W)	P(H,P,A)					P(H)

TAB. 13 – *Modification du tableau 12 en prenant en compte la certitude que W est vrai*

Nature du réseau bayésien. Le réseau bayésien est une écriture probabiliste de la connaissance d'un expert à un moment donné. Elle est formée de deux parties : la structure du réseau, constituée de règles qui sont des relations de « causes à effets », et codée par des probabilités conditionnelles, et les valeurs de probabilités des évènements (ou faits, propriétés, ou variables). La structure n'est pas susceptible de modifications alors que les probabilités des évènements sont remises à jour dès que l'une d'elles est modifiée pour intégrer une nouvelle connaissance. Les règles du réseau bayésien représentent des relations de type « cause à effet », comme les règles d'implication statistique et les règles logiques vues précédemment, mais leur mise en oeuvre dans un raisonnement est plus complexe : la propagation de l'information d'un nœud à un autre n'est pas simple, elle se fait à l'aide de

[15] D'après cette formule, si X et Y sont deux évènements, on peut écrire P(X|Y)=P(Y|X)P(X)/P(X,Y). Pour le détail de la formule concernant les trois évènements et des calculs associés, nous renvoyons le lecteur intéressé à l'ouvrage de Jensen (1996).

[16] En fait il s'agit ici de P(A=1|W=1,H=1) et non P(A=1) comme indiqué dans le tableau, de la même façon, il faudrait écrire P(A=1|H=1) au lieu de P(A=1) dans le tableau 12.

calculs de mise à jour des probabilités des nœuds qui sont à peine réalisables « à la main », comme nous venons de le voir dans cet exemple très simple.

Nature complexe des nœuds et des règles. Le fait d'avoir un seul arc AB entre deux nœuds de ce réseau ne signifie pas que B est vrai si A est vrai, mais qu'il y a une certaine probabilité que B soit vrai quand A l'est, et une autre quand A ne l'est pas. Si bien que la présence de cet arc dans la figure peut très bien indiquer la règle A→nonB, ou nonA→B, ou nonA→nonB plutôt que A→B, mais pas la règle nonB→nonA qui serait associée à une autre structure contenant à sa place l'arc BA, et avec des probabilités conditionnelles P(A|B) au lieu de P(B|A). Certes, la formule de Bayes permet le passage d'une de ces deux probabilités à l'autre, mais l'une est immuable, fixée avec la structure du réseau, et l'autre change quand les informations permettent de la mettre à jour. Quand ce sont deux arcs qui arrivent à un nœud C, cela indique non pas deux relations correspondant aux règles A→C et B→C mais une relation complexe entre A, B et C, avec dans la majeure partie des cas, une *interaction de A et B sur C*, ce qui signifie que l'action de A sur C diffère selon que B est vrai ou non. C'est le cas dans le tableau 10 du nœud H lié à chacun des nœuds P et A par un arc : la probabilité que la pelouse soit mouillée ne change pas avec le fait que l'arrosage était ou non en marche (1 si A vrai, 1 si A faux) dans le cas ou il a plu (P vrai), alors qu'elle change (0,9 si A vrai, 0 si A faux) quand il n'a pas plu (P faux). Cette relation entre les trois faits est complexe mais elle pourrait l'être plus, car on a supposé pour simplifier les calculs que les deux faits P et A sont indépendants. Dans le cas contraire, en ajoutant un arc de P vers A par exemple, P aurait deux effets sur H, un direct et un indirect par l'intermédiaire de A. Pour limiter la complexité de ces relations, des contraintes sont imposées sous forme de quelques « règles de bonne conduite » selon Naïm et al.(2007)

- se limiter à un nombre raisonnable d'arcs aboutissant à un noeud (pas plus de 4)
- éviter de boucler (même de façon indirecte, un fait ne peut pas être à la fois cause et conséquence d'un autre fait). Il faut contrôler à chaque ajout d'un nouveau lien qu'il ne crée pas de boucle dans le système.
- pas trop de noeuds intermédiaires dans un chemin (pas plus de 4)
- éviter qu'une variable agisse a la fois directement et indirectement sur une autre (bypass, ou dérivation)

Avantages et désavantages de la précision. Nous avons vu que les réseaux bayésiens mettent en lumière des relations complexes, mais précises, alors que l'implication statistique privilégie des relations imprécises, mais simples à appréhender et manipuler. Cette précision en fait un outil privilégié pour l'aide à la décision en économie par exemple. On apprend la structure à partir de données quand on en dispose, et à partir de connaissance expertes sinon, ou même en faisant des hypothèses de travail, puis on le confronte à la réalité en comparant les valeurs obtenues à celles observées, et on corrige si nécessaire. Puis on la manipule en changeant certaines valeurs (par exemple on diminue un taux de prêt, on augmente un investissement, etc.) pour envisager de multiples scénarios. On trouve des études de cas montrant ce fonctionnement dans (Naïm et al. 2007). Il est toutefois un domaine où ce mode de fonctionnement peut être dérangeant, c'est celui de la justice. Un exercice simple proposé dans (Naïm et al. 2007) permet de poser le problème :

> « Un suspect est soupçonné de meurtre. Un témoin fiable à 70% assure l'avoir reconnu, et un test ADN fiable à 99% le désigne comme coupable. Trouver la probabilité a

posteriori qu'il soit coupable sachant que la probabilité a priori qu'il le soit est de 10%. »

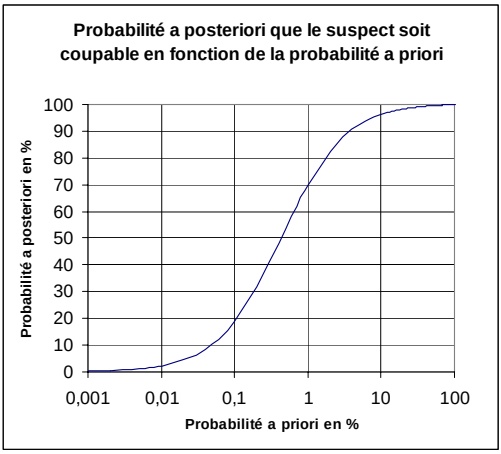


FIG. 7 – *Dépendance entre probabilités a posteriori et celles a priori dans un cas d'école*

On voit dans la figure 7 que la probabilité a posteriori dépend de la probabilité a priori. Les connaissances telles que témoignage et test ADN ne font que la renforcer, par des jeux de multiplication. Si elle est a priori de 10%, elle devient a posteriori à plus de 95%. Les auteurs signalent le problème éthique posé par la fixation de la probabilité a priori.

Au-delà de ce cas d'école, cela montre les limites d'un raisonnement précis sur des données imprécises, aléatoires ou subjectives comme ci-dessus.

Construction de la structure des données Asia. La construction de la structure peut se faire à l'aide d'experts, qui choisissent les nœuds, les arcs, et les valeurs des probabilités a priori et conditionnelles. Cette opportunité est proposée dans les logiciels de réseaux bayésiens. Le graphe de la figure 8, présent dans le logiciel BayesaLab[17] (Version d'évaluation 2001-2008), est un exemple de réseau généré de cette façon.

Cette construction peut aussi se faire par apprentissage à partir des données. Le choix des éléments du réseau se complique dès que la structure augmente de taille. Le plus délicat est de choisir des valeurs précises de probabilités là où on se contenterait d'une échelle ordinale plus ou moins vague (« un peu », « moyennement », « beaucoup ») pour qualifier l'influence de telle variable sur telle autre. La possibilité est offerte dans la plupart de ces logiciels d'apprendre la structure d'après les données. Il y a plusieurs façons de déterminer la structure, la plus « facile », du point de vue algorithmique, étant de faire préciser la structure à l'expert, puis de faire calculer les probabilités selon cette structure par le logiciel d'après les données, et la plus complexe consiste à ne fournir au logiciel que le tableau de données avec les intitulés des variables, qui deviennent toutes des nœuds et laisser le logiciel choisir les arcs et ensuite calculer les probabilités. Pour permettre une solution rapide, des choix par défaut sont faits en suivant des « règles de bonne conduite » comme celles que nous avons

[17] Il est fourni construit dans le logiciel, et un tutoriel explique de façon détaillée comment le construire soi-même (http://www.bayesia.com/fr/produits/bayesialab/).

vues précédemment, mais un paramétrage de ces choix reste possible, détaillé notamment dans BayesaLab par un tutoriel très pédagogique. Nous avons procédé selon la méthode par défaut[18] et obtenu le réseau présent en figure 9.

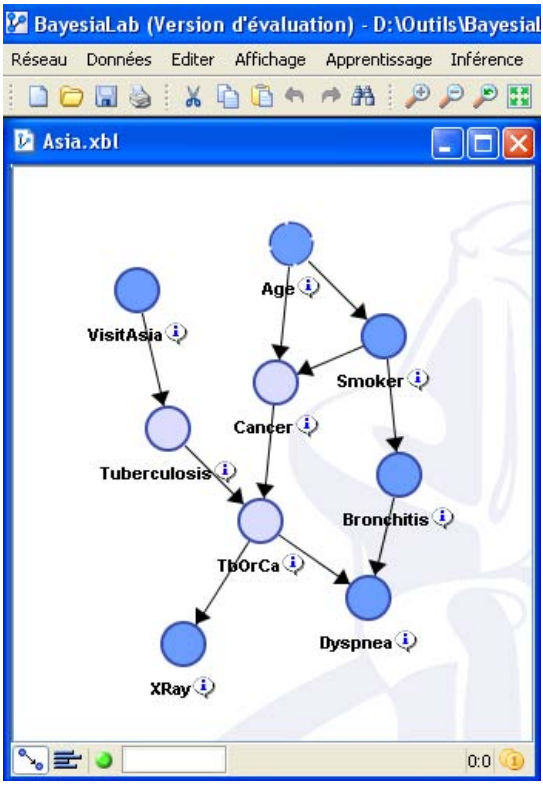


FIG. 8 – *Réseau bayésien construit sur des connaissances expertes.*

Le graphe obtenu par apprentissage total de la structure (cf. figure 9) est très proche de celui proposé par l'expert. Les nœuds sont joints par autant d'arêtes dans les deux figures, seule l'orientation de quelques-unes change : il s'agit des arcs VA→TU, AG→SM, AG→CA qui ont été retournés. Le tutoriel nous invite à intervenir dans le graphe pour le modifier éventuellement en fonction de connaissances expertes, en précisant que l'algorithme n'a pas la possibilité de déceler le sens exact de tous les arcs. Nous avons conservé sans changement le graphe de la figure 9, notre but étant de comparer les « standards » des trois méthodologies. Dans ce graphe, trois nœuds, Age, TbOrCa et

[18] Un choix devait être fait toutefois entre plusieurs méthodes de découverte d'associations (Arbre de recouvrement maximum, Taboo, EQ, SopEQ, Taboo Order). La figure 6 représente le graphe obtenu pour la plupart des options, les autres graphes ayant quelques arcs retournés ou supprimés. Aucun des graphes n'avait les arcs VA→TU, AG→SM, AG→CA dans ce sens, tous les avaient en sens contraire.

Dyspnea, sont plus complexes, étant chacun la destination de deux arcs. De façon plus ou moins importante, les variables Smoker et Cancer interagissent sur la variable Age, Bronchitis et TbOrCa sur la variable Dyspnea, et Cancer et Tuberculosis sur la variable TbOrCa (TbOrCa est l'acronyme de « Tuberculosis Or Cancer », cette variable synthétique a été construite par disjonction des deux autres pour faciliter l'écriture du réseau bayésien).

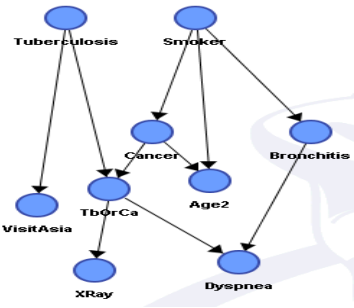


FIG. 9 – *Réseau bayésien construit sur les données Asia (N=7033).*

Comparaison du jeu de règles extrait des données Asia par réseau bayésien aux deux autres. On retrouve dans ce jeu de 10 règles trois règles d'implication statistique de la figure 2, dans le même sens (R1 : Cancer→TbOrCa, R2 Tuberculosis→TbOrCa, R4 TbOrCa→Xray). On peut retrouver les trois règles de ce jeu Smoker→Cancer, Cancer→Age, TbOrCa → Dyspnea dans le graphe d'implication statistique (figure 2, à droite) en s'aidant de la transitivité sous-jacente à ce graphe, la première de ces trois règles étant inversée. Ainsi on retrouve 6 des 10 règles. La règle Smoker→Bronchitis était encore présente dans le tableau 8 de l'indice d'implication non corrigé, et a disparu lors de la correction. L'implication statistique Tuberculosis →VisitAsia ne peut exister, la variable Asia ne donnant pas des relations statistiquement généralisables. Quant à la règle Smoker→Age, qui fait partie de la liaison complexe à trois variables entre Cancer, Smoker et Age du réseau bayésien (cf. figure 9) nous ne la trouvons dans aucun des tableaux de l'implication statistique, alors que nous trouvons à sa place dans le tableau 7 l'implication statistique Smoker→nonAge (valeur 1 de l'indice non corrigé, 0,942 de l'indice corrigé). Cela nous ramène à la remarque déjà faite qu'un arc AB du réseau bayésien indique une relation entre deux variables A et B qui ne peut se réduire à l'écriture de la règle A→B. Par exemple ici, l'arc (Smoker, Age) ne correspond pas à la règle Smoker→Age mais à la règle Smoker→nonAge, alors que les arcs (Smoker, Cancer) et (Cancer, Age) se traduisent par la bonne règle. On voit que le réseau bayésien de la figure 9, apparemment simple, ne peut s'appréhender par une simple lecture du graphique. Ce sont seulement des dépendances qui sont indiquées, et il est nécessaire de consulter les informations chiffrées, ou d'agir sur celles-ci, pour voir si les liaisons correspondantes sont positives ou négatives.

Examinons la règle Tuberculosis →VisitAsia. Sur les 83 personnes atteintes de Tuberculose, et les 74 personnes ayant visité l'Asie, seules 6 vérifient les deux variables. Ce nombre, trop petit d'un point de vue statistique pour générer une règle d'implication statistique ne l'est pas d'un point de vue probabiliste. Dans le premier cas, les sujets sont des unités statistiques, qui, pris individuellement en si petit nombre, ne peuvent être représentatifs d'une tendance, vu la variabilité naturelle des individus. Dans le second cas, 6

personnes sur 74, c'est une part non négligeable d'un sujet abstrait, pouvant être reproduit 74 fois, 7400 fois ou même une infinité de fois. Selon le premier point de vue la liaison entre Tuberculosis et VisitAsia est sans intérêt, alors qu'elle est intéressante selon le second.

Le réseau bayésien est encore plus éloigné du treillis de Galois que l'implication statistique. Bien que le nombre de sujets qui vérifient une règle logique, issue du treillis de Galois puisse être très petit, voire nul, cela ne suffit pas à assurer la ressemblance. Les règles logiques correspondent à des probabilités égales à 1, ce qui ne couvre qu'une très petite partie du réseau bayésien. On ne retrouve que les deux premières des dix règles données par le réseau, celles qui étaient déjà communes au treillis de Galois et à l'implication statistique. Les 15 autres règles logiques ne se retrouvent pas dans le réseau bayésien. La cause la plus vraisemblable de leur absence est le choix fait dans les réseaux bayésiens de ne garder que quelques dépendances très localisées parmi un grand nombre d'indépendances conditionnelles supposées.

6 Conclusion

Nous avons comparé le graphe des implications statistiques selon le modèle de Régis Gras au treillis de Galois et au réseau bayésien. Ce sont trois points de vue différents, statistique, algébrique et probabiliste, pour un même but : donner à l'utilisateur la possibilité de raisonner sur ses données. La brique de base pour chacun est le lien de type causal entre variables, mais la nature des liens et la façon de les enchaîner diffèrent. Les comparaisons ont été faites dans deux directions : aptitude formelle de la représentation à faciliter le raisonnement courant d'un utilisateur et type d'information extraite du jeu de données Asia.

La comparaison des trois ensembles de liens extraits des données Asia montre qu'aucun des modèles ne peut établir automatiquement le sens causal du lien trouvé entre deux faits. Si on sait que A est la cause de B, on peut obtenir la règle A→B comme la règle B→A. La sémantique des liens n'est donc pas directement celle attendue, mais l'utilisateur averti peut, dans le cas du graphe d'implications statistique, interpréter la contraposée nonA→nonB en lieu et place de la règle B→A, et dans le cas des réseaux bayésiens indiquer lui-même le sens attendu.

C'est le modèle à base d'implication statistique qui fournit les liens les plus approximatifs, son but étant de permettre des raisonnements prenant en compte la variabilité individuelle des sujets, comme il est de tradition en statistique. Il offre la navigation la plus simple, des choix drastiques ayant été faits pour que la transitivité manquante soit rétablie, et pour éviter les cycles aboutissant à des « cercles vicieux » du genre A→B, B→C et C→A, qui seraient difficiles à interpréter en terme de causalité. On peut le faire fonctionner sans problème dans les cas limites (variable toujours vraie ou toujours fausse) et ajouter les négations de variables sans obtenir de « fautes de bon sens » du genre A→nonA par exemple. Mais bien sûr, le choix qui a été fait de simplifier sans demander de paramétrage compliqué à l'utilisateur fait perdre certains liens complexes. Des ajouts ont été faits dans les dernières versions, par exemple la possibilité d'avoir plus d'une variable en partie gauche de la règle. Cela permet de faire apparaître des relations complexes entre variables, comme l'interaction (Cadot 2005), mais la représentation qui en est faite s'appuie sur un nombre

parfois important de nœuds contenant plus d'une variable[19], ce qui la destine à un usage expert plutôt que pour le raisonnement courant.

Les treillis de Galois offrent des liens dichotomiques entre deux groupes de variables, qui peuvent être vrais sans même qu'un sujet ne les vérifie, du moment qu'aucun ne les contredit, ce qui en fait un modèle plus adapté aux sciences exactes qu'aux sciences humaines. La rigueur de l'algèbre fait que ce modèle possède toutes les propriétés logiques attendues pour un raisonnement automatique, comme la transitivité par exemple, et le fait de pouvoir engendrer toutes les règles à partir d'une partie génératrice. Il est largement utilisé pour générer les règles d'association (Agrawal 1994), souvent accompagné d'un pré-traitement et d'un post-traitement pour limiter l'explosion combinatoire des règles en éliminant celles qui sont vérifiées par peu de sujets, ou tous « les cas limites » (si les variables x, y et z sont vérifiées par les mêmes sujets, cela fournit un grand nombre de règles sans intérêt) mais en acceptant les règles approximatives. Ces traitements variés ont pour inconvénients de changer sa structure, et de diminuer sa rigueur qui doit être alors renforcée en rajoutant des contraintes qui le rendent difficile à paramétrer pour l'adapter aux besoins de l'utilisateur final. A notre connaissance, aucune amélioration n'a été faite pour rendre ce modèle compatible avec la négation des variables.

Les réseaux bayésiens fournissent des liens très précis, soumis à un certain nombre de contraintes, appelées « règles de bonnes conduite », afin de rendre le réseau plus simple, mais la navigation ne peut se faire au simple vu du réseau, même s'il ne comporte que peu de variables. Les liens sont en effet complexes et l'information circule ou est bloquée selon que l'information dans les nœuds est ou non connue et suivant la façon dont les liens se succèdent. Cette modélisation ne fournit pas de connaissance abstraite sur les données, simple à manipuler pour un cerveau humain, mais de la connaissance concrète, bien appréciée dans les sciences économiques et de gestion pour simuler des faits comme des choix d'investissements, et prendre des décisions optimales pouvant être justifiées de façon comptable. A noter que dès qu'on s'écarte un peu de ce domaine très quantitatif et qu'on se rapproche des sciences humaines hors du domaine économique, on ne peut que constater que cette mécanique de haute précision est mal adaptée aux faits humains. L'exemple qui est fourni dans le domaine de la justice montre comment la probabilité qu'un accusé soit coupable dépend de probabilités variées qui sont établies a priori, c'est-à-dire provenant d'un expert, ou qui sont estimées d'après des données collectées. Ces deux sources sont susceptibles d'erreurs, dues à la subjectivité d'un individu ou à la variabilité d'un groupe d'individus, qui ne sont, à notre connaissance, pas prises en compte par les modèles de réseaux bayésiens.

Références

Agrawal R. and R. Srikant (1994), *Fast algorithms for mining association rules in large databases*, Research Report RJ 9839, IBM Almaden Research Center, San Jose, California.

[19] Sur les données « Asia », l'ajout de la possibilité d'avoir 2 variables en partie gauche des règles a presque quadruplé le nombre de nœuds et le nombre de règles.

Graphe de règles d'implication statistique pour le raisonnement courant.

Barbut M. et B. Monjardet (1970), *Ordre et classification*, Tome 2, Paris, Hachette.

BayesiaLab - Evaluation version (c) Copyright Bayesia S.A. 2001-2008 http://www.bayesia.com/fr/produits/bayesialab/release/bayesialab-3-3.php

Ben Naceur-Mourali, C. Gonzales (2004). Une unification des algorithmes d'inférence de Pearl et de Jensen. *Revue d'intelligence artificielle. RSTI série RIA*, Vol 18, no 2/2004. Lavoisier, Paris. 229-260

Birkhoff, G. (1948). *Lattice theory*, American Mathematical Society colloquium publications volume 25. New York.

Cadot, M. (2006). *Extraire et valider les relations complexes en sciences humaines : statistiques, motifs et règles d'association*. Thèse de doctorat en informatique. Université de Franche-Comté.

Cadot, M., Maj, J.-B. and Ziadé T. (2005), Association Rules and Statistics, in *Encyclopedia of Data Warehousing and Mining*, édité par John Wang, Montclair State University, USA, 94-98.

Davey B.A., Priestley H.A. (1990) *Introduction to Lattices and Order*, Cambridge University Press.

Formal Concept Analysis (2009): http://www.upriss.org.uk/fca/fca.html

Godin R., Mineau G., Missaoui R., Mili H. (1995). Méthodes de classification conceptuelle basées sur les treillis de Galois et applications, *Revue d'Intelligence Artificielle*, 9(2), 105-137

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

Gras R., Bailleul M. (2000). La fouille dans les données par la méthode d'analyse implicative statistique, *Journées du 23 et 24 juin 2000 organisées par l'IUFM de Caen et l'ARDM*.

Gras R., P. Kuntz, R. Couturier, F. Guillet (2001). Une version entropique de l'intensité d'implication pour les corpus volumineux, *Extraction des connaissances et apprentissage*, 2001, Volume 1, Numéro 1-2, 69-80.

Guigues J.L. et V. Duquenne (1986). Familles minimales d'implications informatives résultant d'un tableau de données binaires, *Math. Sci. Hum*. n°95, 5-18

Guillet F. (2004). Mesure de qualité des connaissances en ECD, *Tutoriel EGC 2004*, Clermont-Ferrand, 20 janvier 2004.

Jensen F.V. (1995). *An introduction to Bayesian Networks*. University College London.

Mephu Nguifo E**.** (1994). Une nouvelle approche basée sur le treillis de Galois pour l'apprentissage de concepts, *Mathématiques, Informatique, Sciences Humaines*, vol. 134, 19-38,

Morineau, A. (éd.) (1995). *Aide-mémoire statistique.*. Saint-Mandé : CISIA-CERESTA

Naïm P., P. H. Wuillemin, P. Leray, O. Pourret et A. Becker (2007). *Réseaux bayésiens*, Collection Algorithmes, Eyrolles, Paris.

Muller P ; et L. Vieu. (2009) Structures d'ordre, dans *Sémanticlopédie, dictionnaire de sémantique*, http://www.semantique-gdr.net/dico/index.php/Structures_d'ordre

Pearl J. (2000) *Causality models, reasoning, and inference*, Cambridge University Press, 267 - 279.

Reinhardt F. et H. Soeder (1974) (édition française de J. Cuenat de 1997). Atlas des mathématiques. Collection la pochothèque. Librairie Générale Française..

Whittaker J. (1990). *Graphical models in applied multivariate Statistics*, John Wiley.

Annexes

Un treillis peut être défini de deux façons, soit par une relation d'ordre (cf. définition 1), soit pas deux lois internes (cf. définition 2). Et ces deux définitions sont équivalentes (Reinhardt et al. 1974).

Définition 1 : Un treillis (E, R) est un ensemble E muni d'une relation d'ordre R, c'est-à-dire une relation binaire sur E *réflexive*, *transitive* et *antisymétrique*, telle que toute paire d'éléments admette une *borne supérieure* et une *borne inférieure*. Voici les définitions de ces termes, dans lesquelles b, x, y et z sont des éléments de E :

- réflexivité : $\forall x, (x\ R\ x)$
- antisymétrie : $\forall x, y,$ si $(x\ R\ y)$ et $(y\ R\ x)$ alors $(x = y)$
- transitivité : $\forall x, y, z,$ si $(x\ R\ y)$ et $(y\ R\ z)$ alors $(x\ R\ z)$
- b est une borne inférieure de {x, y}si :
  - $(b\ R\ x)$, $(b\ R\ y)$,
  - $\forall z,$ si $(z\ R\ x)$ et $(z\ R\ y)$ alors $(z\ R\ b)$

La relation réciproque R' de R (x R'y est définie par y R x) est également une relation d'ordre sur E. Ce qui permet de définir la borne supérieure pour R de deux éléments comme étant la borne inférieure pour R'.

Un treillis est *complet* si tout sous-ensemble de E admet un plus petit élément et un plus grand élément. C'est le cas notamment des treillis construits sur des ensembles finis.

Définition 2 : Un treillis $(E, \wedge, \vee)$ est un ensemble muni de deux lois internes possédant un certain nombre de propriétés : *commutativité*, *associativité*, *absorption*. Voici les définitions de ces termes, dans lesquelles x, y et z sont des éléments de E :

- commutativité : $\forall x, y,\ (x \vee y) = (y \vee x)$; $(x \wedge y) = (y \wedge x)$
- associativité : $\forall x, y, z, ((x \vee y) \vee z) = (x \vee (y \vee z))$; $((x \wedge y) \wedge z) = (x \wedge (y \wedge z))$
- absorption : $\forall x, y, (x \wedge (x \vee y)) = y$; $(x \vee (x \wedge y)) = y$

Un treillis *booléen* est un treillis possédant des propriétés supplémentaires : il a un *élément nul*, un *élément universel*, ses lois sont *distributives* l'une par rapport à l'autre, et ont la propriété de *complémentation*. Voici les définitions de ces termes, dans lesquelles x, y et z sont des éléments de E :

Graphe de règles d'implication statistique pour le raisonnement courant.

- distributivité de la loi $\vee$ par rapport à la loi $\wedge$: $\forall x, y, z, (x \vee (y \wedge z)) = ((x \wedge y) \vee (x \wedge z))$
- distributivité de la loi $\wedge$ par rapport à la loi $\vee$: $\forall x, y, z, (x \wedge (y \vee z)) = ((x \vee y) \wedge (x \vee z))$
- n élément nul : $\forall x, (n \wedge x) = n$; $(n \vee x) = x$
- u élément universel $\forall x, (u \wedge x) = x$; $u \vee x) = u$
- complémentation : $\forall x, \exists x', (x \wedge x') = n, (x \vee x') = u$

Un tel treillis est aussi appelé treillis distributif complémenté.

Equivalence des deux définitions : on passe du treillis (E, R) au treillis (E, $\wedge$, $\vee$) en définissant pour loi $\wedge$ (respectivement $\vee$) l'application de ExE vers E qui à toute paire d'éléments {x,y} de E associe leur borne inférieure (resp. supérieure). Et on passe dans le sens inverse en définissant la relation R pour toute paire d'éléments {x,y} de E par (x R y) si x est la borne inférieure de {x,y} (ou, ce qui revient au même, si y est la borne supérieure de {x,y}). Le lecteur intéressé par la démonstration de cette équivalence peut consulter les ouvrages relatifs aux treillis (Birkhoff 1948). On note (E, $\wedge$, $\vee$, R) le treillis pour lequel les lois et la relation d'ordre sont ainsi associées.

Treillis d'ensemble. Un ensemble E étant donné, le *treillis d'ensemble* est une structure de treillis sur l'ensemble de toutes les parties de E (on peut le noter 2^P), les lois internes étant l'intersection et la réunion ensemblistes, et la relation d'ordre étant l'inclusion ensembliste. C'est un treillis booléen dont l'élément nul est l'ensemble vide et l'élément universel l'ensemble E. On peut établir aisément que ce sont respectivement le plus petit élément et le plus grand élément du treillis. D'après P. Muller et L. Vieu. (2009), on obtient une structure d'algèbre de Boole, qui est en correspondance directe avec la structure des règles logiques, « Boole a montré que l'ensemble des parties d'un ensemble muni de l'inclusion est isomorphe au calcul des propositions en logique : l'élément universel E correspond à Vrai, $\varnothing$ à Faux, l'inclusion à l'implication, l'intersection à la conjonction, l'union à la disjonction et l'opérateur de complément à la négation ».

Summary

The statistical implication rules may look similar to the rules of mathematical reasoning. However the model underlying the former is not the formal logic model underlying the latter; it is a statistical model giving rise to approximate relations, in accordance with the common sense logic. The problem is that chaining approximate rules may yield inconsistencies, unless justifiable rule connections may be available in a graph. We will show in this chapter how to set up this chaining in order to keep clear of this drawback, particularly through the building of the implicative graph such as proposed by the successive versions of CHIC. We will compare this statistical model of the data to two alternative models: one is an algebraic model, the Galois lattice, the other one is probabilistic, i.e. bayesian networks. For the ease of comparison, we will illustrate the operation of the three models through a medical dataset freely accessible via internet.

Chapitre 9 : Une étude comparative pour la détection de dépendances multiples

Elham Salehi, Jayashree Nyayachavadi, Robin Gras

Université de Windsor
401 Sunset Avenue
N9B 3P4 Windsor, Ontario Canada
rgras@uwindsor.ca
http://cs.uwindsor.ca/faculty/rgras

Résumé. La recherche de dépendances entre variables à partir d'exemples est un problème important en optimisation. De nombreuses méthodes ont été proposées pour résoudre ce problème mais peu d'évaluations à grande échelle ont été effectuées. La plupart de ces méthodes reposent sur des mesures de probabilité conditionnelle. L'ASI proposant un autre point de vue sur les dépendances, il était important de comparer les résultats obtenus grâce à cette approche avec l'une des meilleures méthodes existantes actuellement pour cette tâche : l'heuristique max-min. L'ASI n'étant pas directement utilisable pour traiter ce problème, nous avons conçu une extension à cette mesure spécifiquement adaptée. Nous avons réalisé un grand nombre d'expériences en faisant varier des paramètres tels que le nombre de dépendances, le nombre de variables concernées ou le type de prédiction effectuée pour comparer les deux approches. Les résultats montrent une forte complémentarité des deux méthodes.

1 Introduction

Il existe de nombreuses situations pour lesquelles il est nécessaire de trouver une relation entre les variables d'un domaine. Cela est particulièrement vrai lorsque l'on s'intéresse à des problèmes d'**optimisation**. La notion d'optimisation est une notion très répandue du fait que la résolution dans grand nombre de problèmes concrets nécessite, à un moment ou à un autre du processus de résolution, d'optimiser un ensemble de paramètres (ou variables) vis-à-vis d'une **fonction objectif** F donnée. Dans beaucoup de problèmes réels, en particulier en bioinformatique, la fonction objectif n'est partiellement voire pas du tout connue. On dispose cependant de la possibilité d'évaluer la valeur de la fonction en chaque point de l'espace de recherche. On appelle ce type de problème, des problèmes "boite noire".

De nombreux travaux ont été menés pour étudier la complexité des problèmes d'optimisation combinatoire. Dans un premier temps, ce sont principalement des preuves de NP-complétude qui ont été réalisées pour certains problèmes classiques (sac à dos, le coloriage de graphe, la recherche d'une clique de taille k…) (Garey M.R. and Johnson D.S., 1979; Papadimitriou C.H. and Steiglitz K., 1998). Ce n'est cependant pas un indice de complexité très informatif. On sait que pour une classe de problèmes démontrés NP-complet

il existe dans la plupart des cas des instances de ces problèmes solubles en temps polynomial. Malheureusement, les propriétés des problèmes qui les rendent plus ou moins faciles à résoudre sont souvent mal connues. Cette information est d'autant plus importante à découvrir qu'elle est en général la clé permettant la conception d'un algorithme de résolution efficace. Des notions comme la décomposabilité d'un problème ou le nombre de solutions du problème ont une implication forte sur sa complexité. La principale propriété influant sur la complexité est le degré d'**épistasie** d'un problème, c'est-à-dire le nombre maximum de variables dont dépend, dans le calcul de F, chacune des n variables. Par exemple, un problème d'épistasie zéro est un problème où toutes les variables sont indépendantes. L'épistasie est un concept plus général que celui de décomposabilité puisqu'un problème de degré d'épistasie k n'implique pas que le problème soit décomposable en n/k sous-problèmes complètement indépendants. En effet, les sous-problèmes peuvent contenir des variables communes et on peut, par exemple, tout à fait avoir n sous-problèmes de taille k. k variables dépendantes sont appelées un bloc. Les blocs peuvent être ou non chevauchants s'ils partagent des variables. Dès que k est égal à deux et que le problème n'est pas décomposable en problèmes indépendants (une même variable est contenue dans plusieurs contextes différents), il a été démontré que le problème est NP-complet[1] (Thompson R.K. and Wright A.H., 1996).

Disposer d'un modèle décrivant les dépendances entre les variables de la fonction à optimiser apporte des informations essentielles sur la structure de cette fonction et permet donc de simplifier considérablement le problème d'optimisation. Par exemple, savoir quelle(s) variable(s) influe(nt) sur quelle(s) autre(s) peut être très utile pour le problème de sélection de variables (feature selection), Zeng et Hernandez (2008), Goldebberg et Moore (2004) ainsi que pour la décomposition du problème en sous-problèmes indépendants ; pouvoir prédire la valeur d'une variables en fonction de celles d'autres variables permet de résoudre le problème de la classification ; savoir quelles combinaisons d'instanciation d'un ensemble de variables conduisent à la valeur optimale d'une certaine fonction…

Le modèle utilisé classiquement pour cette tâche de détection de dépendances est le **réseau bayésien**. Ce réseau correspond en fait à la factorisation de la distribution de probabilités représentant un ensemble d'exemples composé de l'instanciation de l'ensemble des variables de F. L'avantage majeur d'un tel modèle, en plus de donner une représentation visuelle des dépendances du problème, est qu'il peut servir de générateur de nouveaux exemples compatibles avec le modèle. On peut ainsi l'utiliser pour explorer l'espace de recherche de F en tenant compte de propriétés précédemment découvertes. Les approches de types **Algorithmes Génétiques** par Construction de Modèles Probabilistes (AGCMP) reposent sur ce principe. A chaque génération, un modèle (réseau bayésien) de l'ensemble d'exemples courants est construit. Ce modèle est ensuite utilisé pour générer un nouvel ensemble d'exemples. Puis F est évaluée sur chacun de ces nouveaux exemples et un nouvel ensemble est constitué par tirage aléatoire biaisé avec remise dans l'ensemble précèdent. Le biais de sélection donne la préférence aux exemples pour lesquels la valeur de F est la plus grande, ce qui conduit à un nouvel ensemble d'exemples dont la valeur moyenne est plus haute que celle du précèdent. En itérant ce procédé, à chaque étape un nouveau modèle, plus représentatif de la fonction F, est construit, permettant lui-même de générer de nouveaux exemples pour lesquels la valeur de F sera plus élevée. Il a été démontré que ce processus converge vers l'optimum global de la fonction.

[1] Le problème peut en effet être transformé en MAX-2-SAT qui est NP-complet.

Cependant, avec cette approche, chaque étape demande la construction d'un réseau bayésien à partir d'un nouvel ensemble d'exemples. Or il est bien connu que la construction d'un réseau bayésien à partir d'exemples est un problème qui est en lui-même NP-difficile. Ainsi, pour résoudre un problème NP-difficile, il faudrait résoudre une succession de problèmes eux aussi NP-difficiles. Il faut toutefois remarquer qu'à chaque étape de cette méthode il n'est pas nécessaire de construire un modèle parfait, une simple approximation peut suffire. Chaque étape permet de découvrir de nouvelles propriétés du problème qui seront transmises à l'étape suivante par l'intermédiaire d'un nouvel ensemble d'exemples représentatif de ces propriétés. Il est donc important de disposer d'algorithmes heuristiques efficaces permettant de construire un réseau bayésien de bonne qualité. La difficulté de cette tâche réside bien évidemment dans la détection des dépendances du problème pour éviter d'avoir à considérer un nombre exponentiellement grand (en fonction du nombre de variables) de réseaux. Toute méthode permettant de découvrir où se situent les dépendances les plus fortes au sein d'un ensemble de variables est donc particulièrement importante. Une tâche connexe à celle là, et toute aussi importante, est de détecter les sous-ensembles de variables de telle sorte que toute variable au sein d'un même sous-ensemble a une relation de dépendance avec toutes les autres variables du même sous-ensemble et d'indépendance avec celles des autres sous-ensembles. Un cas particulier de cette situation est de détecter les variables qui ne dépendent d'aucune autre. Ce genre d'information est essentiel dans l'optique d'une décomposition du problème. En effet, chaque sous-ensemble peut être optimisé indépendamment des autres ce qui réduit considérablement la complexité du problème.

2 L'heuristique max-min

La détection de **dépendances multiples** à partir d'un ensemble d'exemples est un problème difficile. Il est clair que ce problème ne peut être résolu de manière exacte dès que le nombre de variables considérées dépasse cinq ou six et que le nombre maximum de variables dont peut dépendre une variable approche la dizaine. Or, pour certains problèmes, le nombre de variables peut atteindre plusieurs centaines ou plusieurs milliers. Il est donc particulièrement important de disposer de méthodes permettant d'obtenir une solution approchée de bonne qualité. L'approche utilisée généralement est une approche par recherche locale. Dans ce cas, un modèle des dépendances est recherché incrémentalement, chaque étape consistant à ajouter ou retirer une ou plusieurs dépendances dans le modèle. Le choix des dépendances à ajouter ou retirer se fait en utilisant un score qui évalue la qualité du nouveau modèle en fonction de l'ensemble d'exemples. L'espace de recherche de cette approche reste exponentiel en fonction du nombre maximum de variables dont peut dépendre une variable. Il est donc nécessaire de disposer de méthodes permettant d'augmenter les chances de construire un modèle de bonne qualité sans explorer exhaustivement tout l'espace de recherche. Une approche possible est d'utiliser une méthode moins coûteuse en temps de calcul pour déterminer un sous-ensemble de l'espace de recherche prometteur sur lequel on peut appliquer ultérieurement une méthode plus systématique et plus coûteuse.

Un intérêt de cette approche est que l'on peut utiliser une mesure de dépendance pour la première phase permettant la réduction de l'espace de recherche puis une autre mesure pour la construction du modèle final. Le modèle final utilisé le plus fréquemment est le réseau bayésien qui permet de représenter des dépendances correspondant à des probabilités

conditionnelles. Cependant, pour construire ce modèle, il est tout à fait possible d'utiliser des informations provenant d'autres mesures que les probabilités conditionnelles. Dans cette optique, les mesures effectuées lors de la première phase sont utilisées comme filtre pour éliminer les variables indépendantes ou grouper les variables en sous-groupes partageant des dépendances. La deuxième phase utilise ces informations filtrées pour construire un réseau bayésien. Le but de notre étude est de comparer la capacité de deux approches pour la détection des dépendances pour la première phase. Dans cette section nous considérons directement la mesure de probabilité conditionnelle et elle sera comparée dans la section suivante à une mesure basée sur ASI.

2.1 Définitions et notations

Nous considérons un problème composé de n variables $\{v_1, v_{2, ..., } v_n\}$. Chaque variable v_i peut prendre différentes valeurs parmi l'ensemble de modalités $M_i = \{m_{i,1}, m_{i,2}, \ldots m_{i,k}\}$. Pour la détection des dépendances un ensemble de N exemples est disponible. Chaque exemple correspond à l'instanciation de chacune des n variables dans l'une des k modalités possibles.

Soit Par_i, l'ensemble **parent** de v_i, l'ensemble des variables dont dépend la variable v_i. On dira que $v_j \in Par_i$ est un parent de v_i et que v_i est un **enfant** de v_j.

2.2 L'approche grimpeur max-min

Bien que la construction de réseaux bayesiens soit un domaine de recherche suscitant de nombreuses publications et que des algorithmes exacts aient été donnés pour des problèmes pour lesquels le nombre de variables est inférieur à 30, Koivisto et Sood (2004), l'application de ces algorithmes de construction à des problèmes réels, tels que l'étude des réseaux biologiques ou sociaux, doit toujours faire face au problème du grand nombre de dimensions. Ces dernières années, plusieurs algorithmes ont été développés spécifiquement pour résoudre ses problèmes par des approches basées sur la restriction de l'espace de recherche des structures de réseaux possibles en utilisant différentes heuristiques, Friedman et al. (1999), Tsamardinos et al. (2006). L'un de ces algorithmes, le « Spare Candidate », Friedman et al. (1999), a une complexité polynomiale. Le principe de cette méthode est de restreindre les ensembles parents de chaque variable en supposant que, si deux variables sont quasiment indépendantes dans l'ensemble d'exemples, il est très improbable qu'elles soient connectées dans le réseau bayésien. En utilisant ce principe, l'algorithme « Spare Candidate » construit des ensembles parents de taille identique et contenant peu d'éléments pour toutes les variables. Le problème majeur avec cette approche est de choisir quelle taille d'ensemble parent utiliser ce qui nécessite une très bonne connaissance préalable du problème à résoudre. Du plus, cette taille étant la même pour toutes les variables, cela induit une présupposition très contraignante d'uniformité de densité du réseau.

Plus récemment, un autre algorithme, permettant d'obtenir de meilleurs résultats sur un plus large panel de structure de réseau, a été proposé, Tsamardinos et al. (2006). Cet algorithme, appelé **grimpeur max-min** (MMHC pour max-min Hill Climber), permet de découvrir les possibles relations parents-enfants en utilisant une méthode par contrainte et les utilise pour construire un réseau bayésien. La première étape de cet algorithme, celle qui nous intéresse car c'est elle qui détecte les dépendances, est appelée max-min parents-enfants (MMPC pour max-min Parent Children). L'algorithme MMPC utilise une structure de

données appelée ensemble parents-enfants associée à chaque variable v_i qui contient l'ensemble des variables qui sont soit parent soit enfant de v_i en respectant les distributions mesurées dans l'ensemble d'exemples. Cette notion de respect des distributions des exemples est définie dans Tsamardinos et al. (2006), Neapolitan (2003). MMPC utilise pour cela le test statistique G^2, Spirtes et al. (2000), appliqué sur l'ensemble d'exemples pour déterminer les indépendances conditionnelles entre des paires de variables sachant un ensemble d'autres variables. L'algorithme MMPC est composé de deux phases. Dans la première, un ensemble vide de candidat parents-enfants (CPC) est associe à v_i. Il essaie ensuite d'ajouter des variables une par une dans cet ensemble en utilisant l'heuristique max-min. Cette heuristique sélectionne une variable v_j qui maximise l'association minimale de v_i relativement au CPC courant de v_i. L'association minimale de v_j et v_j relativement à l'ensemble de variable CPC est définie par :

$$\text{MinAssoc}(v_i ; v_j | \text{CPC}) = \text{argmin Assoc}(v_i ; v_j | S) \text{ pour tout sous-ensemble S de CPC}$$

$\text{Assoc}(v_i ; v_j | S)$ est une estimation de l'importance de la corrélation entre v_i et v_j sachant le CPC et est égale à zéro si v_i et v_j sont indépendantes conditionnellement sachant le CPC. La fonction Assoc utilise la p-value renvoyée par le test d'indépendance G^2 comme une mesure de corrélation : plus la valeur de la p-value est petite plus la corrélation est forte. La première phase de MMPC s'arrête lorsque toutes les variables restantes sont considérées comme indépendantes de v_i sachant le CPC courant. Cette approche étant gloutonne, une variable ajoutée à une étape de cette première phase peut très bien être en fait inutile après que d'autres variables aient été ajoutées au CPC. La deuxième phase de MMPC essaie de corriger ce problème en effectuant un deuxième passage pour toutes les variables du CPC de v_i et en enlevant toutes celles qui sont indépendantes de v_i sachant un sous-ensemble de CPC.

Ce qui n'apparaît pas clairement dans les publications associées à ces méthodes c'est leur capacité à découvrir tout type de structure et comment différentes distributions de probabilités conditionnelles et structures du réseau du modèle réel (celui que l'on cherche à découvrir) influent sur la qualité des résultats obtenus. Nous présentons dans la section suivante les résultats que nous avons obtenus en utilisant l'algorithme MMPC sur des exemples générés à partir de différents réseaux bayésiens.

2.3 Effet de la structure du model utilisé pour générer les exemples sur l'efficacité de l'heuristique max-min

Dans cette section, nous étudions la capacité de l'algorithme max-min à découvrir les bons ensembles parents-enfants des variables d'un réseau bayésien à partir de données générées par ce réseau.

Un réseau bayésien est un outil permettant de présenter la distribution conjointe d'un ensemble de variables aléatoires. Les propriétés de dépendances de cette distribution sont codées sont la forme d'un graphe acyclique direct (GAD). Les noeuds de ce graphe sont les variables aléatoires et les arcs correspondent aux influences directes entre les variables. Une table de distribution de probabilités conditionnelles (DPC), appelée aussi paramètres locaux d'une variable, est associée à chaque nœud du graphe. Elle représente la distribution de probabilités $P(v_i | \text{Par}_i)$.

Dans les sections 2.3.1-2.3.3, nous utilisons des données artificielles générées à partir de réseaux bayésiens construits aléatoirement. Chacun de ces réseaux comporte A arcs et n =

100 variables divisées en deux ensembles : un ensemble D de variables pour lesquelles il existe des relations de dépendances directes avec au moins une des n-D-1 autres variables ; un ensemble de I de variables ne possédant aucune relation de dépendance avec aucune des n-1 autres variables. La DPC de chaque variable est générée aléatoirement en tenant compte des éventuelles relations de dépendance. Le nombre de modalités de chaque variable est k = 3. Dans notre étude, nous changeons les caractéristiques des réseaux pour analyser les conséquences de ces changements sur l'efficacité de l'algorithme MMPC. Ces changements incluent la distribution des variables indépendantes I, le nombre de variables dépendantes D et le nombre de dépendances parmi les D variables dépendantes (c'est-à-dire le nombre d'arcs A dans le réseau). Les résultats sont présentés dans les tables 1 à 3. Chaque ligne de ces tables représente une moyenne des résultats obtenus pour 10 ensembles d'exemples différents générés à partir de 10 réseaux différents mais possédant les même caractéristiques. Dans chaque expérimentation, nous calculons la moyenne et l'écart type du nombre de Vrais Positifs (VP), Faux Positifs (FP), Faux Négatifs (FN) et du temps de calcul. VP correspond au nombre de relations parent-enfant correctement prédîtes par l'algorithme sans tenir compte du sens de la dépendance. Donc, le nombre de VP peut être au maximum égal au double du nombre d'arcs du réseau. De la même façon, le nombre de FN, c'est-à-dire les arcs existants dans le réseau mais qui n'ont pas été prédits par l'algorithme, peut au maximum être égal au double du nombre d'arcs. La somme VP + FN est donc égale au double du nombre d'arcs du réseau. Le nombre de FP est le nombre de dépendances prédîtes par l'algorithme et qui n'existent pas dans le réseau.

2.3.1 Distribution des variables indépendantes

Dans cette section, nous étudions l'effet de la distribution des variables indépendantes sur l'efficacité de l'algorithme max-min. Les réseaux bayésiens utilisés dans cette section comportent I = 75 variables indépendantes et D = 25 variables dépendantes. La distribution utilisée pour générer les variables indépendantes varie de pratiquement constante à complètement uniforme. Nous représentons la distribution des variables indépendantes comme un triplet tel que (p1, p2 ,p3). Par exemple, (80, 10, 10) signifie que chaque variable aléatoire a une probabilité de 0.8 de prendre comme valeur l'une des trois modalités possibles, et une probabilité de 0.1 pour les deux autres. La modalité ayant une probabilité de 0.8 est choisie aléatoirement parmi les trois pour chaque variable aléatoire. Les résultats sont présentés dans la table 1. Le nombre d'arcs pour tous ces réseaux est A = 40. Chaque ligne du tableau correspond à un résultat moyen obtenu pour 10 réseaux différents mais de mêmes caractéristiques. On peut constater, à partir de ces résultats, que la distribution des variables indépendantes n'a pratiquement aucun effet sur l'efficacité de l'algorithme max-min. L'algorithme max-min, dans ces conditions, permet de découvrir environ 37% des dépendances. On peut remarquer également que le nombre de FP est élevé, ce qui signifie que l'algorithme a tendance à prédire beaucoup plus de dépendances qu'il n'en existe réellement.

Distribution des variables independantes	Moyenne des VP	DS des VP	Moyenne des FP	DS des FP	Moyenne des FN	DS des FN	Temps de calcul (s)	Précision= VP/(VP+FN)
(80, 10, 10)	30	7.29	116	11.33	49	7.28	6.5	37.5%
(50, 25, 25)	30	8.12	113	11.9	49	8.16	6.4	37.5%
(40, 30, 30)	29	7.28	117	11.63	51	7.28	6.7	36.25%
(33, 33, 33)	29	6.76	118	10.54	50	6.78	6.2	36.25%

TAB. 1 – *Efficacité de l'algorithme max-min en fonction de la distribution des variables indépendantes. Chaque ligne comporte la moyenne et la déviation standard (DS) du nombre de VP, de FP, de FN et du temps de calcul pour dix exécutions de l'algorithme sur des données générées à partir de dix réseaux différents mais présentant les mêmes propriétés. Chaque réseau possède 25 variables dépendantes, 75 variables indépendantes et 40 arcs.*

2.3.2 Proportion de variables dépendantes

Dans cette seconde expérimentation, nous avons fait varier les nombres D et I (n reste égal à 100). Nous avons également fait varier A de façon à conserver pratiquement constant le ratio A/D. Comme l'on peut le constater dans les résultats présentés dans la table 2, lorsque le réseau ne comporte que des variables dépendantes (D = 100), l'algorithme max-min a de bien meilleurs performances : plus de 80% des dépendances sont découvertes. En revanche, comme on peut le voir dans les deux premières lignes de la table, lorsque le nombre de variables dépendantes est égal à 25 ou 50, seules environ 35% des dépendances sont découvertes. Le nombre de FP est aussi très faible lorsque toutes les variables sont dépendantes. Cela semble donc indiquer que cette méthode a des difficultés à déterminer que certaines des variables sont indépendantes. On peut cependant remarquer que le temps de calcul augmente considérablement dans le cas où toutes les variables sont dépendantes. Cela peut s'avérer problématique lorsque le nombre de variables dans le problème est très supérieur a 100.

D	A	Moyenne des VP	DS de VP	Moyenne des FP	DS des FP	Moyenne des FN	DS des FN	Temps de calcul (s)	Précision = VP/(VP + FN)
25	40	29	6.76	118	10.54	50	6.78	0.31	36.2%
50	80	53.4	8.81	99.4	11.35	106.6	8.81	0.34	33.4%
100	150	243.8	8.17	16.6	6.81	56.2	8.17	21.1	81.3%

TAB. 2 – *Efficacité moyenne de l'algorithme max-min en fonction du nombre de variables dépendantes. Chaque ligne comporte la moyenne et la déviation standard (DS) du nombre de VP, de FP, de FN et du temps de calcul pour dix exécutions de l'algorithme sur des données générées à partir de dix réseaux différents mais présentant les mêmes propriétés. Chaque réseau possède 100 arcs.*

2.3.3 Complexité des réseaux

Dans cette section, nous étudions l'effet de la complexité du réseau sur l'efficacité de l'algorithme max-min. Nous faisons varier le nombre de variables n et le nombre d'arcs A. Toutes les variables sont dépendantes, donc D = n. Comme dans l'étude précédente, puisqu'il n'y a pas de variables indépendantes, le pourcentage de dépendances découvertes est élevé. En revanche, ce pourcentage diminue légèrement avec le nombre d'arcs et le nombre de variables. Il semble que la complexité du réseau ait moins d'importance que la mixité des variables (dépendantes et indépendantes). La complexité du réseau influe tout de même fortement sur le temps de calcul.

D	A	Moyenne des VP	DS des VP	Moyenne des FP	DS des FP	Moyenne des FN	DS des FN	Temps de calcul (s)	Précision = VP/(VP + FN)
25	30	52.4	3.55	2.4	1.96	7.6	3.55	0.31	87.3%
25	40	65.6	4.17	2.2	1.89	14.4	4.17	1.83	82%
25	60	91.8	4.51	3.2	3.37	28.2	4.51	6.63	76.5%
100	120	200.6	5.51	25.2	7.28	39.4	5.52	9.98	83.6%
100	150	243.8	8.17	16.6	6.81	56.2	8.17	21.1	81.3%
100	200	312.4	9.67	10.8	3.37	87.6	9.67	28.3	78.1%

TAB. 3 - *Efficacité moyenne de L'algorithme max-min en fonction de la complexité du réseau bayésien. Chaque ligne comporte la moyenne et la déviation standard (DS) du nombre de VP, de FP, de FN et du temps de calcul pour dix exécutions de l'algorithme sur des données générées à partir de dix réseaux différents mais présentant les mêmes propriétés.*

2.3.4 Problèmes classiques

Les réseaux que nous avons utilisés dans les expériences précédentes ont tous été générés aléatoirement. Pour évaluer l'efficacité de l'algorithme max-min sur des données provenant de problèmes réels, nous l'avons appliqué sur un ensemble de problèmes classiquement utilisés pour valider les algorithmes de construction de réseaux bayésiens, Binder et al. (1997), Jensen et Jensen (1996), Kristensen et Rasmussen (2002), Beinlich et al. (1989). Les résultats obtenus semblent meilleurs que ceux présentés dans les tables 1 et 2 et de qualité similaire à ceux présentés dans la table 3. Cela peut s'expliquer vraisemblablement par le fait que les problèmes réels comportent plus de régularités que ceux générés aléatoirement et que les dépendances sont de ce fait plus faciles à localiser.

Problèmes	VP	FP	FN	Temps de calcul (s)
insurance	70	2	34	137
halfinder	124	136	8	1200
barley	136	160	32	205
alarm	84	2	8	68

TAB.. 4 – *Résultats obtenus avec l'algorithme max-min pour quatre problèmes classiques. Chaque ligne contient le nombre de VP, de FP, de FN et le temps de calcul.*

2.4 Utilisation de l'algorithme max-min pour la « sélection de variables »

Nous avons évoqué dans l'introduction la possibilité d'utiliser les méthodes de détection des dépendances pour la sélection de variables. L'idée est de décomposer le problème initial en localisant les variables indépendantes (celles pour lesquelles Par_i est vide après application de l'algorithme max-min) pour lesquelles l'optimisation peut être réalisée indépendamment des autres variables. On sélectionne alors uniquement le sous-ensemble des variables pour lesquelles on soupçonne qu'il existe des liens de dépendances avec d'autres variables pour n'appliquer la méthode d'optimisation combinatoire que sur le sous-problème correspondant. L'espace de recherche ayant ainsi été réduit par cette étape préliminaire, les chances de découvrir une solution de bonne qualité s'en trouvent augmentées. Le problème est donc ici légèrement plus facile que celui étudié dans la section 2.3 car on cherche uniquement à déterminer la listes des variables impliquées dans des relations de dépendances sans vouloir découvrir précisément ces dépendances. Nous avons donc réalisé une série d'expériences pour mesurer les capacités de l'algorithme max-min sur ce problème réduit.

Nous avons utilisé des réseaux en utilisant la méthode présentée dans la section 2.3.1 pour la distribution des variables indépendantes. Pour la distribution (33, 33, 33) nous avons également fait varier la complexité du réseau en changeant le nombre d'arcs. Nous donnons les résultats de ces expérimentations dans la table 5. Chaque ligne représente la moyenne obtenue pour 10 réseaux différents mais possédant les mêmes propriétés. Il apparaît que bien que la méthode max-min puisse découvrir plus de 90% des variables dépendantes (les Vrai Positif de la table 5), elle ne découvre qu'environ 17% des variables indépendantes (les Vrai Négatifs de la table 5). Une méthode basique qui prédirait toutes les variables comme indépendantes aurait des résultats à peine moins bons. Cela signifie que cette méthode a tendance à surestimer fortement le nombre de dépendances. Les résultats sont peu affectés par les différentes distributions de variables indépendantes et par la complexité du réseau. Il semble donc que cette méthode ne puisse pas être utilisée pour le problème de la sélection de variables car pratiquement toutes les variables sont sélectionnées.

Distribution	D	A	VP	VN
(33, 33, 33)	25	60	24.2	11.4
(33, 33, 33)	25	40	23.2	12.4
(33, 33, 33)	25	30	24	10.6
(80, 10, 10)	25	40	23.8	12.6
(50, 25, 25)	25	40	23.8	13.6
(40, 30, 30)	25	40	23.8	13.8

TAB.. 5 – *Résultats obtenus par l'algorithme max-min pour le problème de la sélection de variables. Chaque ligne comporte la moyenne du nombre de VP et de VN pour dix exécutions de l'algorithme sur des données générées à partir de dix réseaux différents mais présentant les mêmes propriétés.*

3 Approche basée sur l'analyse statistique implicative

Nous nous intéressons aux capacités de la méthode d'Analyse Statistique Implicative pour la détection de dépendances multiples. Nous voulons étudier plus particulièrement si les spécificités de l'ASI, Gras et Kuntz (2008), Gras et al. (2004), permettant de prendre en considération les contre-exemples à une hypothèse d'implication, peuvent être utiles pour découvrir des dépendances dans des situations qui sont difficiles pour les méthodes classiques se basant sur des mesures de probabilités conditionnelles. Par exemple, une situation dans laquelle deux variables sont indépendantes mais prennent toutes les deux très souvent la même modalité dans un grand nombre d'exemples. Dans cette situation, une mesure basée sur la probabilité conditionnelle détectera une dépendance. Or si cette modalité est bien présente fréquemment simultanément pour les deux variables, il n'en reste pas moins que les occurrences des ces modalités sont décorrélées, puisque les variables sont indépendantes, donc la prise en compte du nombre de contre-exemples, même peu nombreux, pourrait permettre de réfuter l'hypothèse de dépendance.

Cependant, pour pouvoir utiliser l'ASI dans des situations très générales, certaines modifications sont nécessaires. En effet, nous ne voulons a priori pas nous limiter aux modalités binaires ni présupposer une relation d'ordre entre les différentes modalités d'une variable. Nous voulons également pouvoir détecter une situation où une conjonction de variables implique une autre variable et, dans ce cas, en tenir compte dans une mesure globale. C'est-à-dire que nous voulons pouvoir mesurer qu'une ou plusieurs combinaisons de modalités des variables parents impliquent une ou plusieurs modalités de la variable enfant. Par exemple, soit les variables A, B et C $\in \{0, 1, 2\}$. Nous voulons définir une mesure capable de détecter une dépendance de A vis-à-vis de B et C parce que : quand $B=0 \wedge C=2$ alors 'souvent' $A = 1$ et quand $B = 0 \wedge C = 0$ alors 'souvent' $A = 0$. La version actuelle de l'ASI ne permet pas de prendre en compte simultanément toutes les combinaisons des modalités des variables considérées pour produire un score unique déterminant le niveau de dépendance global de A vis-à-vis de B et C. Nous avons donc défini une extension à l'ASI pour prendre en compte ces différentes situations.

3.1 Définitions et notations

Ces définitions et notations s'ajoutent à celles présentées dans la section 2.1. Soit Card($m_{i,j}$) le nombre de fois où la variable v_i prend la valeur $m_{i,j}$ dans les N exemples. On note Card($m^-_{i,j}$) le nombre de fois où la variable v_i prend une valeur différente de $m_{i,j}$ et Card($m_{i1,j1} \wedge m_{i2,j2}$) le nombre de fois où la variable v_{i1} prend la valeur $m_{i1,j1}$ et la variable v_{i2} prend la valeur $m_{i2,j2}$ dans les N exemples.

On note π_i une instanciation de chacune des variables parents de v_i choisie dans la liste de modalités que chacune d'elles peut prendre et Π_i l'ensemble de toutes les combinaisons d'instanciations des variables parents de v_i. Par exemple, en reprenant l'exemple précédent avec les variables A, B et C, Π_A = {(0, 0), (0, 1), (0, 2), (1, 0), (1, 1), (1, 2), (2, 0), (2, 1), (2, 2)}. Si on note k = $|M_i|$ alors pour tout $v_j \in Par_i$:

$$\left|\Pi_i\right| = k^{\left|Par_i\right|}.$$

Soit Card(π_i) le nombre de fois où l'ensemble des variables parents de v_i prend la valeur π_i dans les N exemples. La mesure q de l'ASI est alors donnée par :

$$\boldsymbol{q(\pi_i, m_{i,j})} = \frac{Card(\pi_i \wedge m^-_{i,j}) - \frac{Card(\pi_i) \times Card(m^-_{i,j})}{N}}{\sqrt{\frac{Card(\pi_i) \times Card(m^-_{i,j})}{N}}}$$

De la même façon, nous calculons la mesure de l'ASI $\tau(\pi_i, m_{i,j})$. Puis le score que nous cherchons à maximiser est donne par :

$$s(\pi_i, m_{i,j}) = -i(\pi_i, m_{i,j}) \times q(\pi_i, m_{i,j})$$

où i(.,.) est l'indice d'inclusion défini dans la partie 1, chap. 1, § 8.2.

3.2 Extension de l'ASI

Nous avons expérimenté plusieurs mesures permettant de prendre en considération les différents critères que nous avons énoncés plus haut. Pour une variable v_i nous voulons une mesure unique qui représente globalement son degré de dépendance vis-à-vis de son ensemble parent. Nous devons donc tenir compte de l'ensemble des combinaisons possibles $\prod_i$ et déterminer, en utilisant les mesures $s(\pi_i, m_{i,j})$, comment elles impliquent l'ensemble M_i des modalités de v_i. Nous construisons donc une table T_i, contenant l'ensemble $\prod_{si}$ des mesures de s pour toutes les combinaisons de π_i et de M_i et de taille :

$$k \times \left|\pi_i\right| = k^{\left|Par(v_i)\right|+1}$$

Nous avons essayé différentes méthodes pour combiner toute l'information provenant de cette table en une seule mesure. La méthode la plus simple pour cela est de considérer uniquement le maximum des $\prod_{si}$. D'autres possibilités sont de considérer la moyenne des $\prod_{qi}$, ou la moyenne des x% plus hauts scores de $\prod_{si}$. Nous avons réalisé de nombreux tests avec ces différentes approches et aucun n'a donné de résultats satisfaisants. Dans cette première série de mesures nous avons considéré indépendamment les scores obtenus pour

une même valeur de π_i mais différentes valeurs de M_i. Or, ce que l'on souhaite détecter c'est qu'une valeur de π_i, c'est-à-dire une instanciation particulière des variables parents de v_i, implique une instanciation particulière de v_i, et que ça soit le cas pour plusieurs instanciations différentes de π_i. Il faut donc une mesure qui permette de détecter que q est fort pour un couple (π_i, $m_{i,j}$) avec $m_{i,j} \in M_i$ et faible pour tous les autre $m_{i,j'} \in M_i$ et que cela soit vrai pour plusieurs π_i. Nous avons donc défini un score qui tient en compte, pour un π_i donné, de la valeur de s maximum $Sup_{\pi i}$ pour tout $m_{i,j} \in M_i$ et de l'entropie $E_{\pi i}$ de s pour toutes les valeurs $m_{i,j} \in M_i$.

Soit $$Sup_{\pi i} = \max_{1 \le j \le k}(s(\pi_i, m_{i,j}))$$

Et $$E_{\pi i} = -\sum_{j=1}^{k} \frac{p(s(\pi_i, m_{i,j}))\log(p(s(\pi_i, m_{i,j})))}{\log(k)}$$

Avec $$p(s(\pi_i, m_{i,j})) = \frac{q(\pi_i, m_{i,j})}{\sum_{j'=1}^{k} q(\pi_i, m_{i,j'})}$$

Pour calculer la mesure associée à une table T_i, on considère l'ensemble H, de taille h, des π_i correspondant aux x% des valeurs de $Sup_{\pi i}$ les plus élevées de la table. Le score de la table est alors :

$$S_{i,Par_i} = \frac{\sum_{\pi_i \in H} Sup_{\pi_i}}{\sum_{\pi_i \in H} E_{\pi_i}}$$

C'est cette mesure que l'on cherche finalement à maximiser.

En reprenant l'exemple avec les variables A, B et C, la table T_A correspond à :

B	C	A = 0	A = 1	A = 2	Sup	E
0	0	0	1.3	0.6	1.3	0.272
0	1	0	0	0	0	0
0	2	2.1	0	0.2	2.1	0.129
1	0	0	0	0	0	0
1	1	0.4	0.2	0.5	0.5	.45
1	2	1.1	0	0	1.1	0
2	0	0	0	0	0	0
2	1	0	0	0	0	0
2	2	0	0	0	0	0

TAB. 6 – *Exemple de calcul de la table T_i avec A, B et C $\in$ {0, 1, 2} et Π_A = {(0, 0), (0, 1), (0, 2), (1, 0), (1, 1), (1, 2), (2, 0), (2, 1), (2, 2)}.*

Si on sélectionne les 20% Sup les plus élevés, seules les lignes 1 et 3 seront sélectionnées et c'est avec les valeurs de Sup et de E correspondantes qu'est calculée S_A qui est égale à

8.48 dans cet exemple. Plus la valeur de S est élevée plus on considère que la dépendance est forte. Dans la section suivant nous donnons un algorithme qui va utiliser cette mesure pour déterminer quelles sont les dépendances principales d'un problème.

3.3 Algorithme de détection des dépendances

Nous avons défini la mesure S_i, associée à chaque variable v_i. Il faut cependant connaître l'ensemble parent de v_i pour pouvoir la calculer. Pour déterminer les dépendances d'un problème il faut examiner différentes configurations possibles d'ensembles parents pour toutes les variables et choisir la configuration qui mène à un score total maximal. Cependant le nombre de configurations à examiner est exponentiel avec le nombre de variables. Il faut donc définir une heuristique pour construire les ensembles parents. Nous avons choisi une approche incrémentale pour cette heuristique. L'algorithme associé commence avec des ensembles parents vides pour chaque variable, puis à chaque étape une nouvelle variable est ajoutée dans l'un des ensembles parents. C'est la mesure S qui est utilisée pour choisir à chaque étape quelle variable ajouter dans quel ensemble parent. Ce processus est appliqué jusqu'à ce qu'un nombre total maxVariable, fixé à l'avance, de variables ait été ajouté. Le calcul de la table T_i étant lui aussi exponentiel en fonction du nombre de variables parents de v_i, nous utilisons des contraintes supplémentaires, limitant à quatre le nombre maximal de variables que peut comporter chaque ensemble parent et le nombre total de variables ajoutées maxVariable. L'algorithme utilisé est également glouton.

Le choix de la prochaine variable à ajouter dans un ensemble parent se fait en comparant le meilleur score de quatre différentes tables. L'algorithme 1, que nous avons défini permet d'éviter de calculer le score pour toutes les combinaisons ayant 2, 3 et 4 variables dans l'ensemble parent. Seules les combinaisons qui ont été sélectionnées avec x variables parents conduiront au calcul du score avec x + 1 variables parents. Dans cet algorithme, la variable structMax contient trois informations : le score de dépendance de la variable enfant vis-à-vis de ses variables parents, le numéro de la variable enfant et le numéro de la variable parent candidate à être ajoutée à l'ensemble parent. Après initialisation, la table Max_1 contient la liste triée de façon décroissante de tous les scores de toutes les combinaisons comportant une variable enfant et une variable parent. Il y a donc n^2 scores dans la table initialement. Les tables Max_2, Max_3 et Max_4 sont initialement vides. Elles servent à mémoriser les scores des combinaisons enfant-parents dans les cas respectifs où il y a 2, 3 et 4 variables dans l'ensemble parent. Ainsi, à chaque étape de l'algorithme, la variable à ajouter dans l'ensemble parent d'une autre variable sera déterminée en sélectionnant le score le plus élevé des 4 tables. Si c'est la table Max_i qui est sélectionnée, l'ensemble parent de la variable associée au score maximal de cette table passera de i-1 à i variables. Le score est alors retiré de la table max_i et un nouveau score est calculé et inséré dans la table max_{i+1}. Les quatre tables sont maintenues triées par ordre décroissant de telle manière que la valeur maximale de chaque table se trouve toujours en position 0.

```
Pour tout v_i
        Par_i = {∅}
    structMax = {0, 0, 0}
    max_1 = ∅
    pour tout v_i {
        pour tout v_j {
            si ( S_{i,Par_i +v_j} > structMax.score) {
                sturctMax.score = S_{i,Par_i +v_j}
                structMax.enfant = i
                structMax.parent = j
            }
        }
        max_1 = max_1 + structMax
    }
    TrieDecroissant(max_1)
    max_2 = ∅, max_3 = ∅, max_4 = ∅
    nbVariable = 0
    tant que (nbVariable < maxVariable) {
        k = maximum(max_1[0].score, max_2[0].score, max_3[0].score, max_4[0].score)
        enf = max_k[0].enfant
        par_enf = par_enf + max_k[0].parent
        si (k < 4) {
            structMax = {0, 0, 0}
            pour tout v_j ∉ par_enf {
                si ( S_{i,Par_i +v_j} > structMax.score) {
                    sturctMax.score = S_{i,Par_i +v_j}
                    structMax.enfant = i
                    structMax.parent = j
                }
            }
            max_{k+1} = max_{k+1} + structMax
            TrieDecroissant(max_{k+1})
            max_k[0] = {0, ∅, ∅}
            TrieDecroissant(max_k)
            nbVariable = nbVariables + 1
    }
```

Algorithme 1 – *Détermination des maxVariable ensembles parents par heuristique gloutonne.*

3.4 Evaluation de l'algorithme de détection des dépendances basé sur l'ASI

Pour réaliser la comparaison la plus honnête possible, nous avons réutilisé les mêmes expériences que celles de la section 2 pour évaluer notre algorithme de détection de dépendances basé sur l'ASI. Il faut cependant remarquer que cette procédure défavorise l'ASI. En effet, les données ont été générées à partir de modèles, des réseaux bayésiens, basés sur la mesure de probabilité conditionnelle qui est celle qui est utilisée également par l'approche max-min. Or, l'approche ASI utilise une autre mesure qui n'a pas les mêmes propriétés. En particulier, une différence très importante est que le modèle du réseau bayésien est non transitif alors que celui de l'ASI l'est. Mais une comparaison totalement équitable n'étant pas possible et en tenant compte de ces différences dans notre analyse, cela nous a paru être la meilleure façon de procéder.

3.4.1 Distribution des variables indépendantes

Nous avons repris les mêmes données que dans la section 2.3.1. Notre algorithme utilise plusieurs paramètres : le pourcentage x% des meilleurs Sup de chaque table T_i et le nombre maximal maxVariable de variables ajoutées dans les ensembles parents. Pour chacun d'eux nous avons expérimenté de nombreuses valeurs. Celles qui nous ont semblé les plus pertinentes et que nous présentons ici sont 10% et 50% pour x% et 35, 50 et 150 variables pour max variables. Pour ce dernier paramètre nous pouvons évaluer ainsi trois configurations importantes correspondant à une situation réelle dans laquelle on ne sait pas à l'avance combien il y a de dépendances dans le problème. Avec 35 variables on recherche moins de variables qu'il n'en existe réellement, avec 50 variables on en recherche légèrement plus et avec 150 on en recherche nettement plus qu'il n'en existe réellement.

Les résultats présentés dans les tables 7 et 8 confirment que notre algorithme ne découvre que peu de dépendances. La mesure utilisée semble toutefois plus sensible à la distribution utilisée pour générer les variables indépendantes. Les résultats obtenus avec la valeur x% = 10% sont cependant légèrement meilleurs. Le temps de calcul est également plus élevé qu'avec l'algorithme max-min mais notre programme n'a pas encore été optimisé pour cela.

X% = 10%	Moyenne des VP	Moyenne des FP	Moyenne des FN	Précision =VP/(VP+FN)	Temps de calcul (s)
(80, 10, 10) 35 var	0.9	34.1	39.1	2.25%	37
(50, 25 ,25) 35 var	6.7	28.3	33.3	16.7%	61.7
(40, 30, 30) 35 var	7.8	27.2	32.2	19.5%	69.3
(33, 33, 33) 35 var	0.6	34.4	39.4	1.5%	44.6
(80, 10, 10) 50 var	1	49	39	2.25%	46.1
(50, 25 ,25) 50 var	8.4	41.6	31.6	21%	76.4
(40, 30, 30) 50 var	11	39	29	27.5%	89.8
(33, 33, 33) 50 var	1.2	48.8	38.8	3%	57.4
(80, 10, 10) 150 var	1.2	148.8	38.8	3%	63.6
(50, 25 ,25) 150 var	12.3	137.7	27.7	30.7%	179
(40, 30, 30) 150 var	15.3	134.7	24.7	38.2%	184
(33, 33, 33) 150 var	4.9	145.1	36.1	12.2%	194

TAB. 7- *Efficacité de notre algorithme en fonction de la distribution des variables indépendantes lorsque x% = 10%. Le paramètre maxVariable prend successivement les valeurs 35, 50 et 150 variables. Chaque ligne comporte la moyenne du nombre de VP, de FP, de FN et du temps de calcul pour dix exécutions de l'algorithme sur des données générées à partir de dix réseaux différents mais présentant les mêmes propriétés. Chaque réseau possède 25 variables dépendantes, 75 variables indépendantes et 40 arcs.*

X% = 50%	Moyenne des VP	Moyenne des FP	Moyenne des FN	Précision = VP/(VP + FN)	Temps de calcul (s)
(80, 10, 10) 35 var	0.2	34.8	39.8	0.5%	33.8
(50, 25 ,25) 35 var	7.7	27.3	32.3	19.25%	59.7
(40, 30, 30) 35 var	5.1	29.9	34.9	12.7%	66.4
(33, 33, 33) 35 var	0.4	34.6	39.6	1%	36.9
(80, 10, 10) 50 var	0.2	49.8	39.8	0.5%	42.3
(50, 25 ,25) 50 var	6.2	43.8	33.8	15.5%	70.2
(40, 30, 30) 50 var	7.1	42.9	32.9	17.7%	80.9
(33, 33, 33) 50 var	0.5	49.5	39.5	1.25%	41.3
(80, 10, 10) 150 var	0.3	149.7	39.7	0.75%	55.5
(50, 25 ,25) 150 var	6.6	143.3	33.4	16.5%	164.2
(40, 30, 30) 150 var	8	142	32	20%	177.9
(33, 33, 33) 150 var	4.2	145.8	36.8	10.5%	140.9

TAB. 8 - *Efficacité de notre algorithme en fonction de la distribution des variables indépendantes lorsque x% = 50%.. Le paramètre maxVariable prend successivement les valeurs 35, 50 et 150 variables. Chaque ligne indique la moyenne du nombre de VP, de FP, de FN et du temps de calcul pour dix exécutions de l'algorithme sur des données générées à partir de dix réseaux différents mais présentant les mêmes propriétés. Chaque réseau possède 25 variables dépendantes, 75 variables indépendantes et 40 arcs.*

3.4.2 Problème de la « sélection de variables »

Nous avons réutilisé les mêmes jeux de données pour tester la capacité de notre algorithme à résoudre le problème de la sélection de variables. Les résultats présentés dans les tables 9 et 10 montrent un très fort potentiel de notre algorithme pour ce problème. Les résultats obtenus sont en effet bien meilleurs que ceux obtenus avec l'algorithme max-min. Si le nombre de VP est légèrement plus faible, le nombre de VN est considérablement plus élevé. Ce qui est particulièrement important c'est le constat que le niveau de prédiction est ici bien meilleur à ce que l'on pourrait attendre du hasard. Comme le ratio du nombre de variables dépendantes sur le nombre de variables indépendantes est de 1/3 dans le modèle utilisé pour générer les données, une prédiction aléatoire donnerait le même ratio de VP/FP (c'est-à-dire donc ce cas VP/(75-VN)). Nous présentons dans la colonne VP/(0.33xFP) des tables 9 et 10, le gain par rapport à une sélection aléatoire des variables dépendantes. Dans les cas de distributions des variables indépendantes (40, 30, 30) et (50, 25, 25) le gain est très important allant jusqu'à 16.1. A titre de comparaison, les résultats de l'algorithme max-min montre plus de stabilité mais un gain qui ne dépasse jamais 1.18. Notre algorithme semble avoir plus de difficulté lorsque les variables indépendantes ont des distributions extrêmes (33, 33, 33) ou (80, 10, 10). Avec x%=10% et dans le cas où l'on recherche moins de dépendances qu'il en existe (35 variables) le gain est toujours d'au moins 1. Bien que ce soit une première version, notre algorithme semble donc avoir un très fort potentiel pour détecter les variables dépendantes et donc résoudre le problème de la sélection de variables. Nous avons également testé notre algorithme sur les jeux de données présentées en section 2.3.2 dans lesquels D = 50, I = 50 et A = 80 (résultats non présentés ici). Les résultats montrent qu'avec la configuration x% = 10%, le gain est compris entre 1.28 et 1.82.

X% = 10%	Moyenne des VP	Moyenne des VN	VP/(0.33xFP)
(80, 10, 10) 35 var	6.7	55.5	1.03
(50, 25 ,25) 35 var	15.5	71.8	14.7
(40, 30, 30) 35 var	15.4	72.1	16.1
(33, 33, 33) 35 var	3.4	64.7	1
(80, 10, 10) 50 var	6.8	46.4	0.73
(50, 25 ,25) 50 var	18.3	68.7	8.79
(40, 30, 30) 50 var	18.3	69.7	10.5
(33, 33, 33) 50 var	6.1	61.3	1.36
(80, 10, 10) 150 var	9.2	13	0.45
(50, 25 ,25) 150 var	22.8	31.8	1.6
(40, 30, 30) 150 var	21.8	46.7	2.33
(33, 33, 33) 150 var	17.3	32.6	1.24

TAB. 9 - *Résultats obtenus par notre algorithme pour le problème de la sélection de variables avec le paramètre x% = 10%. Chaque ligne comporte la moyenne du nombre de VP et de VN pour dix exécutions de l'algorithme sur des données générées à partir de dix réseaux différents mais présentant les mêmes propriétés. Les résultats sont présentés pour trois valeurs du paramètre maxVariable : 35, 50 et 150 variables.*

X% = 50%	Moyenne des VP	Moyenne des VN	VP/FP
(80, 10, 10) 35 var	7.6	60.4	1.58
(50, 25 ,25) 35 var	16.8	66.2	5.79
(40, 30, 30) 35 var	14.8	67	1.85
(33, 33, 33) 35 var	4	57	0.67
(80, 10, 10) 50 var	8	53.5	1.13
(50, 25 ,25) 50 var	18.3	59.9	3.67
(40, 30, 30) 50 var	18.1	63.7	4.85
(33, 33, 33) 50 var	5.5	50	0.67
(80, 10, 10) 150 var	11.8	6.7	0.52
(50, 25 ,25) 150 var	22.1	22.8	1.28
(40, 30, 30) 150 var	22	41.2	1.97
(33, 33, 33) 150 var	22	16.5	1.14

TAB. 10 - *Résultats obtenus par notre algorithme pour le problème de la sélection de variables avec le paramètre x% = 50%. Chaque ligne comporte la moyenne du nombre de VP et de VN pour dix exécutions de l'algorithme sur des données générées à partir de dix réseaux différents mais présentant les mêmes propriétés. Les résultats sont présentés pour trois valeurs du paramètre maxVariable : 35, 50 et 150 variables.*

4 Conclusion

Nous avons réalise une étude sur la capacité de découverte des dépendances d'un problème de deux méthodes basées sur des mesures différentes. La première, l'algorithme max-min repose sur le test de dépendance conditionnelle G^2. La deuxième est un algorithme que nous avons conçu basé sur une extension de la mesure ASI. Nous avons appliqué ces algorithmes à de nombreux jeux de données en faisant varier les paramètres du problème tels que la distribution des variables indépendantes, le nombre de variables dépendantes et le nombre de dépendances. Nous avons également considéré deux problèmes différents : déterminer quelles sont les dépendances et déterminer quelles sont les variables impliquées dans des relations de dépendance. Bien entendu être capable de résoudre le premier problème permet de résoudre également le second. Il n'est cependant généralement pas possible de résoudre directement et intégralement ce problème. Avoir la possibilité de découvrir dans un premier temps uniquement quel est le sous-ensemble des variables concernées par des relations de dépendance permet de réduire la complexité du premier problème et donc de fournir une solution de meilleure qualité.

Nos résultats ont montré une bonne efficacité pour l'algorithme max-min pour découvrir les dépendances lorsque toutes les variables du problème sont impliquées dans des relations de dépendance. L'algorithme semble peu affecté par la variation de complexité du modèle et par les différentes distributions des variables indépendantes. Il a cependant d'importantes limitations pour détecter les dépendances lorsqu'une partie des variables sont indépendantes. L'algorithme max-min ne semble pas non plus être efficace pour le deuxième problème : la sélection de variables. Notre algorithme, basé sur l'ASI, en revanche ne semble pas capable de détecter directement les dépendances et ce quelles que soient les configurations. Il paraît cependant très efficace pour déterminer quelles sont les variables indépendantes et quelles

sont les variables dépendantes. Il a toutefois plus de difficulté dans les situations où les variables indépendantes ont des distributions extrêmes (80, 10, 10) ou (33, 33, 33).

Les deux approches paraissent donc complémentaires et prometteuses. Il serait très intéressant de développer une méthode combinant ces deux approches. Dans une première phase notre algorithme, utilisant la version étendue de l'ASI, permettrait de sélectionner un sous-ensemble des variables pour lesquelles il y a une forte présomption de dépendances. Puis, dans une deuxième phase, l'approche max-min serait appliquée à ce sous-ensemble pour déterminer plus précisément où sont ces dépendances. L'ensemble de ces informations permettrait alors de construire un réseau bayésien modélisant bien l'ensemble des exemples disponibles et donc pouvant être utilisé pour résoudre le problème d'optimisation associé. Il serait intéressant également d'étudier d'autres variantes de l'extension de la mesure ASI pour voir dans quelle mesure il serait possible d'améliorer les résultats obtenus pour la détection directe de dépendances.

Références

Beinlich I.A., H.J. Suermondt, R.M. Chavez, G.F. Cooper (1989) The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks. *Second European Conference in Artificial Intelligence in Medicine,* 247-256

Binder J., D. Koller, S. Russell, K. Kanazawa (1997). Adaptive probabilistic networks with hidden variables. *Machine Learning,* 29(2-3):213-244

Friedman N., I. Nachman et D. Peèr (1999). Learning bayes network structure from massive datasets: The "sparse candidate" algorithm. *15th Conference on Uncertainty in Artificial Intelligence* 206-215.

Garey M. R. et D. S. Johnson (1979). *Computers and Intractability: a Guide to the Theory of NP-Completeness.* Freeman.

Goldebberg A. et A. Moore (2004). Tractable learning of large Bayes net structures from sparse data. *21ft International Conference on Machine Learning* 44-51

Gras R., R. Couturier, J. Blanchard, H. Briand, P. Kuntz, P. Peter (2004). Quelques critères pour une mesure de qualité de règles d'association. Un exemple : l'implication statistique, *Mesures de qualité pour la fouille de données, RNTI-E-1, Cepaduès – Editions,* 3-32

Gras R. et P. Kuntz (2008). An overview of the Statistical Implicative, *Statistical Implicative Analysis,* R.Gras, E. Suzuki, F.Guillet and F.Spagnolo, Eds, Springer-Verlag.

Jensen A.L., F.V. Jensen (1996). MIDAS: An Influence Diagram for Management of Mildew in Winter Wheat. In *Proceedings of 12th Annual Conference on Uncertainty in Artificial Intelligence,* 349-356

Koivisto M. et K. Sood (2004). Exact Bayesian structure discovery in Bayesian networks, *Journal of Machine Learning Research* 5:549-573

Kristensen K., I.A. Rasmussen (2002). The use of a Bayesian network in the design of a decision support system for growing malting barley without use of pesticides. *Computers and Electronics in Agriculture,* 33: 197–217

Larranaga P. et J.A. Lozano (2002). *Estimation of Distribution Algorithms: A new tool for evolutionary computing*: Kluwer Academic Publishers.

Muhlenbein H. et T. Mahnig (2001). Evolutionary Algorithms: From Recombination to Search Distributions. *Theoretical Aspects of Evolutionary Computing*: Springer Verlag, 135-173.

Neapolitan, R. (2003) *Learning Bayesian networks*. Prentice Hall

Papadimitriou C. H. et K. Steiglitz (1998). *Combinatorial Optimization: Algorithms and Complexity*. Dover.

Spirtes P., C. Glymour et R. Scheines (2000). *Causation, prediction, and search*. The MIT Press, second Edition

Thompson R. K. et A. H. Wright (1996). *Additively decomposable fitness functions.* University of Montana, Computer Science department.

Tsamardinos I., L. E. Brown et C. F. Aliferis (2006). The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78.

Zeng Y. et C. Hernandez (2008). A Decomposition Algorithm for Learning Bayesian Network Structures from Data. PAKDD 2008, *Lecture Notes in Artificial Intelligence* 5012:441–453.

Summary

Searching for dependencies between variables using a set of examples is an important problem in combinatorial optimization. Numerous methods have been proposed to solve this problem but few large scale evaluations have been performed. Most of these methods rely on conditional probability measures. The SIA propose another point of view on dependency measurement. It is therefore important to compare the results obtained using this approach with the ones obtained with one of the currently most efficient method for this task: the max-min heuristic. However, the SIA can not be used directly for this problem. We have therefore defined a dedicated extension to this measure. We have performed a large number of experiments varying several parameters like the number of dependencies, the number of variables or the goal of the prediction task. The results show a strong synergy of the two methods.

Chapitre 10 : Test de Mac Nemar et Analyse Statistique Implicative

Jean-Claude Régnier

Université de Lyon - UMR 5191 ICAR
ENS-LSH 15, Parvis René Descartes BP 7000 69342 LYON cedex 07
jean-claude.regnier@univ-lyon2.fr

Résumé. Nous tentons de comparer, dans ce chapitre, les trois approches pour étudier des liens vraisemblables entre deux variables binaires, que sont : l'ASI, le test de Mac Nemar et le test d'indépendance fondé sur la mesure du $\chi 2$. Pour ce faire, nous faisons un retour sur des données issues de nos travaux passés dans le champ de la didactique des mathématiques.

1 Introduction

Comme nous l'avons déjà évoqué succinctement dans la première partie de cet ouvrage, la comparaison de deux séries successives de données binaires de type présence-absence ou échec-réussite relevées sur le même échantillon d'individus comme cela est le cas en ASI, peut aussi être réalisée à l'aide du test du χ^2 de Mac Nemar.

2 Le test de Mac Nemar

2.1 Un exemple introductif issu d'une situation dans le cadre de la didactique des mathématiques

Pour comparer la difficulté de deux épreuves de mathématiques passées par le même groupe d'individus (séries dites appariées), c'est cette approche qui avait été choisie à la fin des années 70 et début 80 par Jean-Claude Régnier (Régnier 1980 p. 62-75, 1983) dans ses travaux de DEA et thèse en didactique des mathématiques. Comme nous l'avons déjà présenté, l'information est résumée dans un tableau 2x2 dont nous donnons un exemple ci-dessous. Pour rester congruent au mode de présentation des recherches de liens génériquement notés $a \Rightarrow b$ dans le contexte ASI qui présuppose que $N(a) \leq N(b)$, nous représentons systématiquement la variable binaire a en ligne et la variable binaire b en colonne.

		Variable b = Épreuve Finale (item 113)		
		1 = Réussite	0 = Échec	**Total**
Variable a = Épreuve Initiale (item 101 et/ou item 201)	1 = Réussite	56	6	62
	0 = Échec	14	26	40
	Total	70	32	102

TAB. 1 - *Tableau extrait (Régnier 1983, p. 164-166) : items 101 et 201 -- item 113*

Pour mieux illustrer notre propos, nous rappelons succinctement le contenu de ces items qui relèvent du champ de la trigonométrie.

A partir d'angles **aigus** représentés graphiquement et fournis dans l'épreuve et non rapportés ici :

Item 101	Déterminer le sinus et le cosinus de l'angle (101) en utilisant le demi-cercle trigonométrique donné par report avec un calque ou par une construction.
Item 201	Déterminer le sinus et le cosinus de l'angle (201) en construisant un demi-cercle trigonométrique (unité 10 cm). Faire figurer la construction sur la feuille. Utiliser l'équerre pour construire les perpendiculaires.
Item 113	Déterminer le sinus et le cosinus de l'angle (113) à l'aide du demi-cercle trigonométrique fourni ou. par construction, sur la figure, d'un demi-cercle trigonométrique (unité 10 cm)

La question que nous nous posons alors est de savoir si les fréquences de réussite aux deux épreuves sont significativement différentes ou non.

2.2 Formalisation succincte du test de Mac Nemar

L'idée de Mac Nemar pour étudier ce type de lien entre les deux épreuves est qu'il est plus pertinent de ne prendre en compte que les discordances entre les deux épreuves c'est-à-dire le cas de réussite à l'une et d'échec à l'autre et son complémentaire. Dans le tableau ci-dessus, ce sont les deux effectifs 14 et 6 correspondant aux couples (A_Echec, B_Réussite) et (A_Réussite, B_Echec) qui sont considérés comme des informations majeures. Cette idée n'est pas rendue par le test du χ^2 d'indépendance que nous avons déjà évoqué précédemment (Partie 1 Chap. 1-6.6) en établissant la relation algébrique entre l'indice d'implication et la mesure du χ^2.

Si nous nous remettons dans le contexte de l'ASI, le tableau de référence est donc celui-ci :

		Variable b		
		1	0	**Total**
Variable a	1	$n(a \wedge b)$	$n(a \wedge \bar{b})$	$n(a)$
	0	$n(\bar{a} \wedge b)$	$n(\bar{a} \wedge \bar{b})$	$n(\bar{a})$
	Total	$n(b)$	$n(\bar{b})$	n

TAB. 2- *Tableau de contingence avec les notations ASI*

Dans l'hypothèse d'une équivalence entre les deux épreuves, la fréquence de ceux qui sont passés d'une réussite à un échec parmi ceux qui ont changé d'état est égale à la fréquence de ceux qui sont passés d'un échec à une réussite parmi ceux qui ont changé d'état, c'est à dire égale à 0,5. D'une certaine manière, cela revient à comparer une fréquence observée à une fréquence théorique de 0,5.

Mac Nemar a montré qu'il suffisait de prendre comme indice, la mesure suivante que nous nommerons χ^2 de Mac Nemar, $\chi^2_{MacNemar} = \frac{(n(\bar{a} \wedge b) - n(a \wedge \bar{b}))^2}{n(\bar{a} \wedge b) + n(a \wedge \bar{b})}$ dont la loi de probabilité est approximativement celle de la variable de Pearson χ^2 de degré de liberté 1

En résumé les 4 étapes de la démarche de ce test sont les suivantes :

- Étape 1 : formulation des hypothèses :

H_0 : symétrie des changements d'état entre les deux épreuves

H_1 : non-symétrie des changements d'état entre les deux épreuves

- Étape 2 : calcul de la valeur empirique du χ^2 (Mac Nemar)
- Étape 3 : lecture de la valeur critique dans la table du χ^2 de Pearson de ddl=1 pour un risque α donné
- Étape 4 : décision statistique rejet ou non rejet de H_0

2.3 Retour à l'exemple présenté.

Dans le cas présenté, nous calculons la valeur empirique comme suit $\chi^2_{MacNemar} = \frac{(14-6)^2}{14+6} = \frac{8^2}{20} = 3{,}2$ et nous la confrontons à la valeur critique au niveau de risque α. Si nous choisissons un niveau de risque de 0,05, la valeur critique est alors de 3,84. Comme 3,2<3,84, nous ne rejetons pas l'hypothèse d'équivalence des deux épreuves que nous considérons comme telle avec un risque de 2ème espèce β inconnu.

Si nous revenons à la perspective de recherche de lien par rejet de l'indépendance en appliquant le test du χ^2 d'indépendance, nous trouvons une valeur empirique de 34,56 qui est très largement supérieure à la valeur critique de 3,84 pour un niveau de risque α=0,05 et même à la valeur critique 6,63 pour un niveau de risque α=0,01. Au sens du test du χ^2 d'indépendance, il existe donc un lien fort entre les deux variables.

Si nous nous plaçons dans la perspective de recherche de lien au sens de l'ASI, le calcul de l'intensité d'implication $\varphi_P(a,b)$ avec le modèle de Poisson et le calcul de l'intensité d'implication $\varphi_{BIN}(a,b)$ avec le modèle binomial

$\chi^2 = 34{,}56$		$\chi^2_{MC} = 3{,}2$		Intensités d'implication
a-b	b=1	b=0		$\varphi_P(a,b)$
a=1	56	6	62	0,9996
a=0	14	26	40	$\varphi_{BIN}(a,b)$
	70	32	102	0,9998

TAB. 3 - *analyse selon les trois perspectives: χ^2 d'indépendance, χ^2 Mac Nemar, ASI,*

Les valeurs qui figurent dans le tableau ci-dessus, indiquent un niveau de confiance en l'implication statistique (Réussir Item 101 et/ou Item 201)⇒(Réussir Item 113) supérieur à 0,99. Ce que nous avions pris en compte à l'époque par la mise en œuvre de l'indice d'implication que Régis Gras (1979) avait exposé dans sa thèse deux ans plus tôt. Ce que nous découvrons à ce jour, c'est qu'il y avait eu une erreur dans les références des tableaux, ce qui conduisait à prendre comme valeur $\chi^2_{MacNemar} = \frac{(56-26)^2}{56+26} = \frac{30^2}{82} = 10{,}97$ et nous avions alors appuyé le sens de la quasi-implication sur l'approche du test de Mac Nemar. Ce retour est donc l'occasion d'une rectification dans les analyses des données d'une thèse d'il y a plus d'un quart de siècle ! Nous étions satisfait de cette concordance entre les deux approches. Dans la conséquence de la confusion que nous avions faite dans les références des tableaux, le second cas où nous avions mis en œuvre cette comparaison relevait bien d'une situation de discordance entre les deux approches. Nous rapportons le tableau de contingence :

		Variable b = Épreuve Finale (item 114)		
		1 = Réussite	0 = Échec	**Total**
Variable a = Épreuve Initiale (item 102 et/ou item 204)	1 = Réussite	40	7	47
	0 = Échec	21	34	55
	Total	61	41	102

TAB. 4 - *Tableau extrait (Régnier 1983, p. 164-166) : items 102 et 204 -- item 114*

Pour mieux illustrer notre propos, nous rappelons succinctement ces items qui relèvent du champ de la trigonométrie.
A partir d'angles **obtus** représentés graphiquement et fournis dans l'épreuve et non rapportés ici :

Item 102	Déterminer le sinus et le cosinus de l'angle (102) en utilisant le demi-cercle trigonométrique donné par report avec un calque ou par une construction.
Item 204	Déterminer le sinus et le cosinus de l'angle (204) en construisant un demi-cercle trigonométrique (unité 10 cm). Faire figurer la construction sur la feuille. Utiliser l'équerre pour construire les perpendiculaires.
Item 114	Déterminer le sinus et le cosinus de l'angle (114) à l'aide du demi-cercle trigonométrique fourni ou. par construction, sur la figure, d'un demi-cercle trigonométrique (unité 10 cm)

Si nous revenons à la perspective de recherche de lien par rejet de l'indépendance en appliquant le test du χ^2 d'indépendance, nous trouvons une valeur empirique de 23,21 qui

est encore très largement supérieure à la valeur critique de 3,84 pour un niveau de risque α=0,05 et même à la valeur critique 6,63 pour un niveau de risque α=0,01. Au sens du test du χ^2 d'indépendance, il existe donc un lien fort entre les deux variables.

Si nous nous plaçons dans la perspective de recherche de lien au sens de l'ASI, le calcul de l'intensité d'implication $\varphi_P(a,b)$ avec le modèle de Poisson et le calcul de l'intensité d'implication $\varphi_{BIN}(a,b)$ avec le modèle binomial

$\chi^2 = 23{,}21$		$\chi^2_{MC} = 7$		Intensités d'implication
a-b	b=1	b=0		$\varphi_P(a,b)$
a=1	40	7	47	0,9983
a=0	21	34	55	$\varphi_{BIN}(a,b)$
	61	41	102	0,9993

TAB. 5 - *analyse selon les trois perspectives: χ^2 d'indépendance, χ^2 Mac Nemar, ASI,*

Les valeurs qui figurent dans le tableau ci-dessus, indiquent un niveau de confiance en l'implication statistique (Réussir Item 102 et/ou Item 204)⇒(Réussir Item 114) supérieur à 0,99. Ce que nous avions pris en compte à l'époque par la mise en œuvre de l'indice d'implication malgré la valeur empirique du χ^2 Mac Nemar erronée (3,45 au lieu de 7 !). Là nous nous étions étonné de la discordance des conclusions entre les deux approches. Nous avions alors tenté de rechercher diverses configurations de tableaux de contingences pour établir une comparaison.

2.4 Comparaison des approches : test de Mac Nemar, ASI et test d'indépendance

Nous avons répété cette recherche en essayant de produire un exemple pour chacun des huit cas possibles que nous identifions dans les tableaux suivant :

Au seuil de risque α		Test du χ^2 Mac Nemar	
	Décision…	Rejet de Ho	Non rejet de Ho
test du χ^2 d'indépendance	Rejet de Ho	Cas 1	Cas 2
	Non rejet de Ho	Cas 3	Cas 4

TAB. 6 - *Cas dans la logique des tests statistiques d'hypothèses*

Au niveau de confiance 1- α		Cas 1	Cas 2	Cas 3	Cas 4
	Décision…				
Analyse Statistique Implicative	Quasi-implication retenue	Conf_1	Conf_2	Conf_3	Conf_4
	Quasi-implication non retenue	Conf_5	Conf_6	Conf_7	Conf_8

TAB. 7 - *Configurations issues du croisement de la logique des tests et de la logique ASI*

Le tableau suivant fournis les tableaux de contingences correspondant à chacune des huit configurations possibles quant à la prise de décision pour un risque de 1ère espèce d'une valeur $\alpha = 0,05$ et pour un niveau de confiance 1- α en la quasi-implication supérieur à la valeur minimale requise dans la théorie de l'ASI, à savoir que l'intensité d'implication soit supérieure à 0,50.

Nous pouvons constater dans les configurations 6 5, 6, 7 et 8 que les valeurs de l'intensité d'implication sont inférieures à 0,50. Dans celles-ci la présomption de quasi-implication ne peut absolument pas être retenue. L'approche Mac Nemar conduit cependant à conclure que dans la configuration 7, l'hypothèse de symétrie de changement d'états doit être rejetée à un seuil de risque bien inférieur à 0,05 et même inférieur à 0,005. Cette dissymétrie aurait pu alors être interprétée comme une tendance à une implication de réussite entre a et b , malgré le non rejet de l'indépendance entre les deux variables a et b.

$\chi^2 =$	9,65	$\chi^2_{MC} =$	7,81	
a-b	b=1	b=0		$\varphi_P(a,b)$
a=1	36	10	46	0,9628
a=0	27	29	56	$\varphi_{BIN}(a,b)$
	63	39	102	0,9742

TAB. 8 - *Configuration 1*

$\chi^2 =$	34,56	$\chi^2_{MC} =$	3,2	
a-b	b=1	b=0		$\varphi_P(a,b)$
a=1	56	6	62	0,9996
a=0	14	26	40	$\varphi_{BIN}(a,b)$
	70	32	102	0,9998

TAB.. 9 - *Configuration 2*

$\chi^2 =$	1,75	$\chi^2_{MC} =$	21,3	
a-b	b=1	b=0		$\varphi_P(a,b)$
a=1	38	8	46	**0,7519**
a=0	40	16	56	$\varphi_{BIN}(a,b)$
	78	24	102	0,7667

TAB.. 10 - *Configuration 3*

$\chi^2 =$	2,96	$\chi^2_{MC} =$	2,81	
a-b	b=1	b=0		$\varphi_P(a,b)$
a=1	30	16	46	**0,797**
a=0	27	29	56	$\varphi_{BIN}(a,b)$
	57	45	102	0,826

TAB. 11- *Configuration 4*

$\chi^2 =$	4,005	$\chi^2_{MC} =$	5,22	
a-b	b=1	b=0		$\varphi_P(a,b)$
a=1	24	22	46	0,101
a=0	40	16	56	$\varphi_{BIN}(a,b)$
	64	38	102	0,081

TAB.. 12 - *Configuration 5*

$\chi^2 =$	4,22	$\chi^2_{MC} =$	0,068	
a-b	b=1	b=0		$\varphi_P(a,b)$
a=1	34	28	46	**0,1318**
a=0	30	10	56	$\varphi_{BIN}(a,b)$
	64	38	102	0,1026

TAB. 13 - *Configuration 6*

$\chi^2 =$	0,259	$\chi^2_{MC} =$	8,01	
a-b	b=1	b=0		$\varphi_P(a,b)$
a=1	29	17	46	**0,320**
a=0	38	18	56	$\varphi_{BIN}(a,b)$
	67	35	102	0,310

TAB.. 14 - *Configuration 7*

$\chi^2 =$	0,019	$\chi^2_{MC} =$	1,23	
a-b	b=1	b=0		$\varphi_P(a,b)$
a=1	24	22	46	**0,413**
a=0	30	26	56	$\varphi_{BIN}(a,b)$
	54	48	102	0,409

TAB. 15 - *Configuration 8*

Face à ces configurations, il nous semble qu'un paradoxe surgisse puisque le même tableau de contingence est susceptible d'être interprété de manière contradictoire. Une façon de lever ce paradoxe est de considérer la logique sous-jacente à chacune des trois approches : approche ASI, approche χ^2 Mac Nemar, approche χ^2 d'indépendance.

3 Conclusion

Comme nous avons pu le voir au travers des propos tenus tout au long de ce qui précède, ceux-ci s'appuient sur un point de vue soutenu par I.-C. Lerman (1992) appliqué à l'étude d'une certaine relation de dépendance orientée entre des variables descriptives. Ce point de vue oppose la logique des tests statistiques, comme celui dit du χ^2 d'indépendance ou encore celui du χ^2 de Mac Nemar, à celle des méthodes classificatoires de la manière suivante : pour les premiers, dit I.-C. Lerman, « relativement à l'existence d'un lien, on a FAUX, jusqu'à preuve du contraire » par le rejet de l'hypothèse nulle ; pour les secondes, « pour l'optique des données, on a VRAI, jusqu'à preuve du contraire », c'est-à-dire vrai selon une certaine échelle de probabilité du lien. Pour terminer nous pouvons rappeler que le test de Mac Nemar se généralise à des variables catégorielles qui ont plus de deux modalités. Si k est ce nombre, le test s'appuie sur un tableau de contingence de dimension k x k. (Pupion et Pupion, 1998 p.94).

Références

Lerman I.-C (1992) Conception et analyse de la forme limite d'une famille de coefficients statistiques d'association entre variables relationnelles. *Mathématiques, Informatique et Sciences Humaines* (118)

Pupion G. et P.-C. Pupion (1998) *Tests non paramétriques. Avec applications à l'économie et à la gestion.* Paris : Economica

Régnier, J.C. (1980) *Élaboration d'un livret auto-correctif. Étude préliminaire : questionnaire sur l'équation du second degré. Projet de livret autocorrectif.* Mémoire de DEA de Didactique de mathématiques Université Nancy 1- ULP Strasbourg. Directeur du mémoire : Georges Glaeser. Irem de Nancy. 172 p.

Régnier J.C. (1983) *Étude didactique d'un test autocorrectif en trigonométrie*. Thèse de doctorat en mathématiques (mention : didactique des mathématiques) ULP Strasbbourg Directeur de thèse : Georges Glaeser. Irem de Strasbourg Tome 1 : 307 p. Tome 2 : 50 p. suivies des Annexes.

Summary

In this chapter we try to compare three approaches to study the likely links between two binary variables, which are: The SIA, the test of Mac Nemar and the test of independence based on the measurement of $\chi 2$. For this purpose, we make a return on data from our past work in the field of mathematics education

Chapitre 11 : Historique et fonctionnalités de CHIC

Raphaël Couturier* et Saddo Ag Almouloud**

* Laboratoire d'Informatique de l'université de Franche-Comté (LIFC),
IUT de Belfort-Montbéliard, BP 527, 90016 Belfort, France
raphael.couturier@iut-bm.univ-fcomte.fr
** Pontifícia Universidade Católica de São Paulo - PUC/SP
Rua Marquês de Paranaguá, 111, Consolação, São Paulo - SP - Brasil
saddoag@gmail.com
http://www.pucsp.br/pos/edmat

Résumé. CHIC permet d'utiliser la plupart des méthodes définies dans le cadre de l'ASI (Implication Statistique Implicative). Il a pour objectif de découvrir les implications les plus pertinentes entre les variables d'un ensemble de données. Pour cela, il propose d'organiser les implications sous forme d'une hiérarchie cohésitive (orientée) ou un graphe implicatif. De plus, il permet d'obtenir une hiérarchie des similarités (non orientée) basée sur les ressemblances des variables. Ce papier décrit l'historique, les caractéristiques et l'usage de CHIC.

1 Introduction

L'analyse statistique implicative (ASI) a été développée par Régis Gras et ses collaborateurs. Elle permet d'établir des règles d'association à partir d'un ensemble de données croisant sujets et variables. Le but initial de cette méthode a été de répondre à la question : "Si un objet possède une propriété, est ce qu'il en possède une autre ?". Bien entendu lorsque la réponse est totalement affirmative, il est facile de répondre à cette question. Cependant, il est possible que ce ne soit pas le cas, alors on peut constater que des tendances apparaissent. L'ASI a pour objectif de mettre en évidence de telles tendances dans un ensemble de propriétés. Comparée aux autres méthodes statistiques qui permettent de générer des règles d'association, l'ASI se distingue par le fait qu'elle utilise une mesure non linéaire qui satisfait des critères importants. Tout d'abord, cette mesure est basée sur l'intensité d'implication qui mesure le degré de surprise inhérent à une règle. Ainsi, les règles triviales qui sont potentiellement évidentes et connues de l'expert sont supprimées. Cette intensité d'implication peut être renforcée par le degré de validité, défini par rapport à l'entropie de Shannon, si l'utilisateur choisit ce mode de calcul. Dans ce cas, la mesure ne prend pas simplement en compte la validité de la règle, mais aussi sa contraposée. En effet, quand une règle d'association est estimée valide, c'est-à-dire que l'ensemble des items A est fortement proche de l'ensemble des items B, alors il est légitime et intuitif d'attendre que la contraposée soit valide, c'est-à-dire que l'ensemble des items non-B soit fortement

proche de l'ensemble des items non-A. Ces deux mesures originales sont complétées par une mesure classique basée sur la taille du support de la règle. Ainsi, en combinant les trois mesures, on peut définir une mesure pertinente qui possède les qualités des trois mesures (si on considère l'utilisation de la théorie entropique), c'est-à-dire la résistance au bruit comme la contraposée de la règle est prise en compte et le rejet des règles triviales. Pour plus d'information le lecteur intéressé peut consulter Gras et al. (2004). Grâce à cette mesure originale, CHIC permet de calculer les règles d'associations à partir d'un ensemble de données. CHIC et l'ASI ont été utilisés pour un large spectre de domaines de recherche (Couturier et al., 2004; Froissard, 2005; Couturier, 2005; Orus et Gregori, 2005; Ramstein, 2008).

CHIC permet de construire deux types de hiérarchie et un graphe. La hiérarchie la plus connue est la hiérarchie des similarités. L'index de similarité a été défini dans Lerman (1981) et il permet de construire une hiérarchie ascendante. De manière similaire, l'intensité d'implication peut être utilisée afin de construire une hiérarchique orientée. En plus de cela, CHIC offre la possibilité de générer un graphe original, appelé graphe implicatif qui permet à l'utilisateur de sélectionner les règles d'associations et les variables qu'il souhaite voir apparaître.

L'historique de CHIC est décrit dans la section 2. Dans la section 3 nous passons en revue les variables que peut traiter CHIC ainsi que les options qui peuvent aider l'utilisateur. La section 5 présente la hiérarchie des similarités et la hiérarchie cohésitive. La section 6 présente le graphe implicatif. Dans la section 7 nous présentons d'autres possibilités de CHIC. La section 8 donne une illustration du calcul avec variables intervalles et du calcul des typicalités et contributions. Finalement, la section 9 conclut ce chapitre.

2 Historique

2.1 Le cadre théorique

Dans ces paragraphes, nous retraçons la genèse historique et l'évolution du logiciel CHIC (Classification Hiérarchique Implicative et Cohésitive) parallèlement aux développements théoriques de l'ASI. Dans sa thèse d'état, R. Gras (Gras, 1979), à partir de l'idée de l'indice de similarité développé par I. C. Lerman a construit l'indice d'implication statistique entre variables binaires. A partir de nouveaux problèmes réels posés pour analyser d'autres types de variables, R. Gras et son équipe ont élargi l'étude de l'implication en concevant à chaque étape les notions mathématiques en réponse à ces problèmes :

- Entre variables non binaires, c'est-à-dire modales ou fréquentielles ;
- Entre variables-sur-intervalles, variables-intervalles et variables floues ;
- Entre des classes de variables de nature quelconque.

Les applications à l'analyse de problèmes de didactique ont mis en évidence la nécessité d'autres développements théoriques. C'est ainsi que le concept "d'implication-inclusion", les notions de typicalité et de contribution des variables supplémentaires, de nœuds significatifs, ont été développés toujours en écho aux questions posées par les praticiens.

Mais l'émission de ces questions de terrain conduisant aux développements théoriques de l'ASI a été facilitée, rendue possible grâce à l'implémentation de l'outil informatique CHIC sur micro-ordinateur. Ses élaborations successives ont pris en compte le développement théorique, mais aussi les progrès des outils informatiques. C'est donc l'interaction ternaire questions posées - réponses théoriques - réalisations informatiques qui est le moteur du paradigme systémique que serait l'implication statistique.

2.2 Les premières étapes de CHIC

Avant 1990, le logiciel CHIC (il ne portait pas encore ce nom) consistait en une version primitive de R. Gras, en Basic (1984), implémentant les calculs des intensités d'implication et surtout, plus délicat, l'algorithme de construction sur micro-ordinateur (Thomson 05) de la hiérarchie des similarités de I.C.Lerman.

Dès 1990, puis dans le cadre de sa thèse S. Ag Almouloud (Ag Almouloud, 1992), s'est attaqué à l'un des objectifs de la réalisation d'un outil informatique fiable et assez convivial pour permettre de traiter, outre l'analyse de similarité de I. C. Lerman, l'analyse implicative de R. Gras et ses extensions : la hiérarchie implicative de classe (Larher, 1991) ainsi que l'étude des variables numériques et modales. Pour ce faire, il fallait créer un logiciel d'analyse de données intégrant les traitements suivants dont S. Ag Almouloud s'est chargé :

- la saisie et les éditions de données, ainsi que les différentes opérations sur ces données (100 sujets et 54 variables maximum),
- la classification hiérarchique de similarité entre variables (binaires, ou fréquentielle s) de I. C. Lerman,
- l'intensité d'implication entre variables (binaires, modales ou fréquentielles),
- la hiérarchie implicative de classes de variables et le calcul de leur cohésion,
- la construction de graphe implicatif (programmé par H. Rostam (Rostam, 1981) et repris en turbo Pascal par M. Mouradi, Université de FES, Maroc),
- le repérage des nœuds significatifs des la hiérarchie des similarités
- du repérage des nœuds significatifs de la hiérarchie implicative de classes (programme réalisé par H. Ratsimba-Rajohn (Ratsimba-Rajohn, 1992) dans le cadre de sa thèse),
- le calcul des paramètres (moyenne, écart-type et corrélation).
- la possibilité de concaténer deux fichiers ayant le même nombre de variables binaires.

2.3 Une évolution dans une seconde étape

L'utilisation de ce premier module d'édition de données conduisant à des restrictions préjudiciables, (impossibilité d'ajouter ou d'éliminer des variables ou des individus par exemple,) S. Ag Almouloud l'a progressivement amélioré :

- l'extension de la dimension du tableau de données (1000 individus et 100 variables) grâce à une meilleure gestion de la mémoire ;
- la possibilité de supprimer ou d'adjoindre des individus et des modalités ;

- la concaténation de deux fichiers ayant le même nombre d'individus ;
- la disjonction d'un sous-programme permettant de modifier la valeur d'une donnée ;
- la relecture de fichier de données ; ces données pouvant être des données binaires, fréquentielles ou des intensités ;
- une meilleure gestion de la mémoire et des ports graphiques
- le calcul des intensités d'implication.

Le changement principal que S. Ag Almouloud a alors apporté consiste dans la précision des résultats. En collaboration avec A. Totohasina (Totohasina, 1992), la loi normale est simulée. Aussi, remplace-t-on dans le programme initial de R. Gras (traduit en turbo pascal par P. Gentil) l'approximation grossière de cette loi par une simulation plus fine. Pour l'analyse implicative, sont intégrées au programme :

- la possibilité d'une ou plusieurs sorties des intensités d'implication à 1'écran. L'option offre à l'utilisateur la sélection des intensités d'implication les plus fortes afin de faciliter la construction manuelle du graphe implicatif,
- l'impression à l'écran ou à l'imprimante des couples (i, j), j fixe et i variable tels que ($i \Rightarrow j$). Cette option s'inscrit également dans le cadre d'une aide à la construction du graphe implicatif,
- la sauvegarde des intensités d'implication dans un fichier explicitement nommé. Leur stockage était indispensable au calcul des indices de cohésion et à la hiérarchie cohésitive de classes,
- la possibilité également de stocker la matrice d'incidence dans un fichier dénommé par l'utilisateur. Pour la construction automatique d'un graphe implicatif, les intensités d'implication sont multipliées par 100. Sa construction est essentielle pour le didacticien qui cherche à mettre en évidence de façon visuelle, les enchaînements dans les comportements de réponses des élèves.

Signalons qu'à chacune des options "calcul des intensités d'implication" et "calcul des indices de similarité", nous obtenons les occurrences des variables. L'analyste peut alors en éliminer certaines s'il estime que la faiblesse de leurs occurrences n'est pas suffisante ou pertinente et ceci sans changer le contenu du fichier initial. Cette élimination entraîne éventuellement une nouvelle numérotation des variables non éliminées et une sortie à l'écran du nombre de variables sur lesquelles portera l'analyse. Pour, d'une part, vérifier la fiabilité du nouvel outil statistique qu'est la hiérarchie implicative de classes, et d'autre part, informer les utilisateurs des potentialités de ce nouvel outil, S. Ag Almouloud a intégré au programme "cohésion", disponible dans la partie "hiérarchie implicative de classes", les indices d'implication entre les classes de variables.

En résumé, la dernière version en 1992 de CHIC écrite en turbo pascal 6, se compose d'un programme principal et d'un ensemble de sous-programmes, selon l'organigramme figure 1 correspondant à la structure de CHIC réalisée par S. Ag Almouloud.

Parallèlement aux travaux réalisés par S. Ag Almouloud, Harrisson Ratsimba-Rajohn (Ratsimba-Rajohn, 1992) a introduit dans sa thèse la notion de typicalité de variables supplémentaires et des individus et a réalisé le programme informatique qui en permet le traitement.

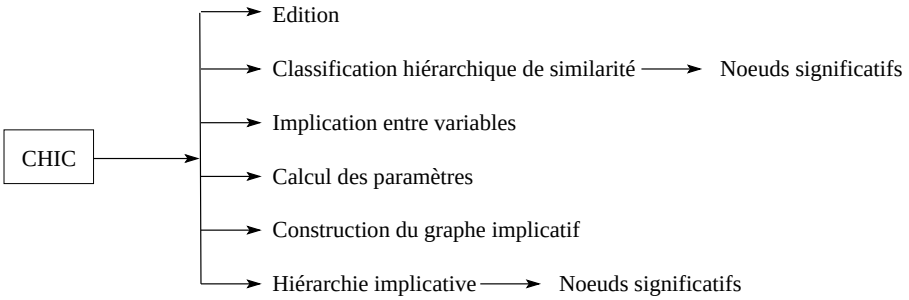


FIG. 1 – *Organigramme du logiciel CHIC*

2.4 Le CHIC contemporain

Depuis 1993, Raphaël Couturier a repris la réalisation informatique du logiciel CHIC en lui apportant des améliorations et une évolution prenant en compte les nouveaux développements théoriques de l'ASI et les vœux des utilisateurs. Cette évolution se traduit, entre autres, par l'introduction des couleurs, par l'enrichissement du menu et par son écriture en C++ qui a conduit à des dizaines de versions différentes jusqu'à l'actuelle version 4.2. Celle-ci permet comme nous allons le voir plus en détail :

- de traiter différents de variables, autres que binaires,
- de quantifier la significativité des valeurs attribuées à la qualité, la consistance de la règle associée, de classes ordonnées de règles, à la typicalité et la contribution des sujets ou de catégories de sujets à certaines règles,
- de représenter, par un graphe, pour un seuil de qualité choisi, des chemins de règles et, par une hiérarchie, des règles sur des règles que l'on appelle aussi règles généralisées,
- de supprimer, d'ajouter, de conjoindre des variables.

3 Variables

Initialement CHIC tout comme l'ASI ont été pensés pour traiter des variables binaires. Par la suite, l'ASI a été enrichie par l'ajout d'autres types de variables et CHIC en a bénéficié. Actuellement, CHIC offre la possibilité de traiter des variables binaires, des variables fréquentielles, des variables définies sur intervalles et des variables intervalles. Le cas des variables binaires est évidemment le cas le plus simple. Les variables fréquentielles, quant à elles, prennent leur valeur entre 0 et 1. Ce type de variable permet de modéliser les variables modales pour lesquelles il existe un nombre fixe de valeurs comprises entre 0 et 1 qui correspondent aux différentes modalités. La manière de définir les modalités est très importante, parce qu'elle intervient fortement dans les résultats de CHIC selon que les valeurs des modalités ordonnées sont proches de 0 ou de 1. Cette remarque est évidemment valide pour les variables fréquentielles.

L'utilisateur doit prêter une très grande attention au processus qu'il utilise pour transformer une variable réelle en une variable fréquentielle. En effet, plusieurs stratégies sont envisageables en fonction des valeurs. Si les valeurs sont positives, elles peuvent être divisées par la valeur maximale. Une autre possibilité consiste à considérer que la valeur minimale représente le 0 et que la valeur maximale représente le 1, toutes les autres variables sont, dans ce cas, proportionnellement distribuées entre la valeur minimale et la valeur maximale. Si une variable réelle possède des valeurs négatives et positives, il est possible de constituer deux variables, l'une contenant les valeurs positives et l'autre contenant les valeurs négatives. Dans ce cas, les précédentes remarques sont toujours opportunes pour les deux variables nouvellement constituées. Cependant, il est également possible de considérer que la valeur minimale (même si elle est négative) représente 0 et la valeur maximale représente 1 et, ainsi, les autres valeurs sont proportionnellement converties dans l'intervalle $[0, 1]$.

	poids p	taille p	femme s	homme s
i1	95	200	0	1
i2	43	140	1	0
i3	75	186	0	1
i4	60	174	1	0
i5	110	183	0	1
i6	140	180	0	1
i7	100	176	0	1
i8	79	172	1	0
i9	69	170	1	0
i10	45	165	1	0
i11	65	180	0	1
i12	60	175	0	1
i13	120	175	0	1
i14	80	160	1	0
i15	100	180	0	1
i16	121	175	0	1
i17	97	162	1	0
i18	72	176	0	1
i19	40	160	1	0
i20	57	175	1	0

FIG. 2 – *Un simple exemple de données avec des variables sur intervalles (avant partitionnement) et des variables supplémentaires*

Les variables sur intervalles et les variables intervalles sont utilisées pour modéliser des situations complexes. Nous les détaillons par la suite. Les variables intervalles permettent de faire face au problème rencontré par la conversion d'une variable réelle en une variable fréquentielle, comme nous l'avons expliqué précédemment (voir paragraphe suivant). En utilisant les mêmes valeurs réelles, une variable sur intervalle procède différemment. Elle découpe les valeurs de la variable en un nombre fixe d'intervalles. Le nombre d'intervalles est choisi par l'utilisateur et ensuite l'algorithme des

nuées dynamiques Diday (1971) constitue automatiquement les intervalles qui ont des limites distinctes. Cet algorithme a la particularité de construire des intervalles en minimisant l'inertie de chaque intervalle et en maximisant l'inertie interclasse de l'ensemble des intervalles. Ensuite, un intervalle est représenté par une variable binaire et un individu a la valeur 1 s'il appartient à cet intervalle et 0 sinon. En utilisant une telle décomposition, un individu appartient à un seul intervalle. Ainsi, le nombre de variables croit avec cette méthode.

Prenons un exemple. Supposons que nous disposions d'un ensemble d'individus et que pour chacun d'entre eux, nous connaissions sa taille et son poids. Supposons également que ces individus pèsent entre $40kg$ et $140kg$ et que leur taille varie entre $140cm$ et $200cm$. La figure 2 montre un exemple avec quelques individus, les valeurs ont été choisies arbitrairement et ne sont pas encore partitionnées. Supposons que nous souhaitions décomposer chaque variable en quatre intervalles, en fonction de la distribution des deux variables, nous pouvons obtenir les intervalles $[40, 60[$, $[60, 95[$, $[95, 110[$, $[110, 140]$ que nous appelons respectivement $poids1$, $poids2$, $poids3$ et $poids4$ et les intervalles $[140, 160[$, $[165, 174[$, $[174, 186[$, $[186, 200]$ pour les tailles que nous appelons respectivement $taille1$, $taille2$, $taille3$ et $taille4$. Dans la suite du calcul, toutes les unions des intervalles d'une variable sont considérées. Ainsi avec la variable $taille$, nous obtenons les intervalles $taille12$, $taille23$, $taille34$, $taille1-3$, $taille2-4$ et $taille1-4$. Les intervalles de la forme $nomAB$ correspondent à l'union de deux intervalles consécutifs, par exemple $taille23$ correspond à l'union des intervalles $taille2$ et $taille3$. Les intervalles de la forme $nomA-B$ correspondent à l'union de tous les intervalles entre $nomA$ et $nomB$, par exemple $taille1-3$ correspond à l'union des intervalles $taille1$, $taille2$ et $taille3$. Bien évidemment, l'utilisation des variables sur intervalles est d'autant plus intéressante quand il est possible de constituer des partitions les plus petites possibles, c'est-à-dire rassembler les intervalles pour lesquels on sait qu'ils sont naturellement proches les uns des autres. CHIC permet d'utiliser un tel algorithme qui est décrit mathématiquement, par exemple, dans (Gras et al., 1996; Gras, 2005). Avec l'exemple précédent, si d'autres variables renseignent sur les tendances vers telle ou telle caractéristique des individus, alors il est possible d'obtenir des informations entre ces variables et le poids et la taille de la population étudiée. Par exemple, il est possible de savoir que les personnes mesurant entre $140cm$ et $180cm$ ont plutôt une attirance pour telle ou telle chose ou que les personnes avec telles hauteurs pèsent principalement entre $90kg$ et $150kg$. Bien évidemment le nombre d'intervalles peut agir fortement sur les résultats.

Alors que pour une variable sur intervalles un individu prend la valeur 1 pour un seul intervalle, une variable intervalle offre la particularité qu'un individu ait différentes valeurs sur plusieurs intervalles. De plus, les intervalles peuvent être continus et peuvent représenter une décomposition discrète, comme c'est le cas en utilisant une méthode de décomposition automatique telle que celle des nuées dynamiques, mais ils peuvent également être définis par l'utilisateur selon les critères personnels de celui-ci. En prenant l'exemple précédent avec la taille et le poids, un utilisateur peut préférer choisir que les personnes soient élancées, normales, en surcharge pondérale, petite, moyenne ou grande. Néanmoins, une variable intervalle offre la possibilité qu'un individu puisse prendre différentes valeurs parmi les différents intervalles

mais impose que la somme de ces valeurs soit inférieure ou égale à 1. Dans la plupart des cas, la somme sera égale à 1 mais ce n'est pas une obligation. Dans la pratique, l'utilisation de variables intervalles permet de classifier plus facilement un objet ou un individu parce qu'il est fréquent que les opinions divergent sur le fait que quelque chose ou quelqu'un soit plutôt petit ou normal. Par conséquent, on peut exprimer que quelqu'un est mince en donnant à cet individu 0.75 pour élancé et 0.25 pour normal. Il faut également noter que ce mécanisme permet de prendre en compte les variables floues qui sont souvent utilisées dans certains types de problèmes Bojadziev et Bojadziev (1996). L'utilisation de variables floues vient soit d'une appréciation humaine, qui par définition est subjective, soit par une mesure imprécise qui pour une raison quelconque introduit une incertitude (cf Partie1, chap 7).

CHIC utilise le format CSV (avec un point virgule comme séparateur) comme format de données pour les fichiers, celui-ci est utilisé classiquement dans les tableurs. Les individus sont rangés dans la première colonne. Les variables sont disposées sur la première ligne. Les valeurs des individus sont représentées dans un tableau à deux dimensions tel que les valeurs pour chaque variable d'un individu sont rangées dans une ligne du tableau (le premier élément étant le nom de l'individu). Les valeurs d'une variable pour tous les individus sont disposées dans les colonnes du tableau (le premier élément étant le nom de la variable). Bien entendu, le type des valeurs diffère dans un tableau selon le type des variables (binaires, fréquentielles, ...).

Les variables supplémentaires peuvent être utilisées dans CHIC afin d'expliquer la formation de certaines règles. Ce type de variable n'intervient pas directement dans le calcul des règles mais il est utilisé dans le calcul des typicalités et des contributions. Prenons un exemple. Supposons que nous souhaitions étudier l'impact d'un nouveau tramway dans une ville et qu'un questionnaire ait été élaboré à cet effet. Ce dernier rassemble de nombreuses informations sur les besoins et les espoirs associés à ce projet. Dans ce genre de questionnaire, le sexe des personnes est renseigné. Par exemple, il est possible que CHIC, avec un tel questionnaire, génère des règles telles que les personnes travaillant et habitant loin de leur lieu de travail sont généralement très intéressées par le projet, ou les familles avec des enfants jeunes sont parties prenantes du projet. En utilisant le sexe des personnes comme variable supplémentaire, il est possible de savoir si les personnes responsables de la construction des précédentes règles sont plutôt des hommes, des femmes ou s'il n'y a pas de distinction.

Avant de lancer un calcul, l'utilisateur doit choisir quel type de calcul il désire utiliser. En effet, il est possible de choisir le calcul classique de l'intensité d'implication ou la version entropique de celle-ci, comme nous l'avons signalé en introduction. Ce choix de type de calcul influe très fortement sur les règles produites. Généralement, il faut utiliser la version entropique dès lors que l'effectif de la population devient grand, par exemple plus grand que 300 à 400. Elle est plus sévère que la version classique de l'intensité d'implication qui produit plus de règles mais qui n'est pas appropriée aux grands ensembles de données.

Pour plus d'informations sur l'intensité d'implication, le lecteur intéressé est invité à consulter la partie 1 de ce livre.

4 Calcul des conjonctions

Pour calculer efficacement les indices d'implication CHIC est basé sur l'algorithme de calcul des règles d'association défini par Agrawal Agrawal et al. (1993). Cet algorithme permet de calculer efficacement règles d'implications composées de conjonctions dans la partie prémisse. Prenons un exemple avec 5 variables A, B, C, D et E et cherchons les règles composées de 3 variables (donc de la forme $A \wedge B \Rightarrow C$). Pour cela, l'algorithme va déterminer les occurrences des triplets de variables, c'est-à-dire des 8 triplets : ABC, ABD, ABE, ACD, ACE, ADE, BCD, BCE and CDE. Pour chacun de ces triplets, il a fallu déterminer les occurrences des couples AB, AC, AD, AE, BC, BD, BE, CD, CE et DE. À partir de ces couples et triplets, il est possible de calculer de nombreuses règles d'implication. Par exemple avec les occurrences de ABC, de AB, BC, et AC il est possible de calculer l'intensité des règles $A \wedge B \Rightarrow C$, $B \wedge C \Rightarrow A$ et $A \wedge C \Rightarrow B$.

Même avec un seuil d'implication élevé, le nombre de règles produit par les conjonctions peut s'avérer très élevé si le nombre de variables initial est grand. De plus, le fait d'utiliser des conjonctions peut être source de ressemblance voire de superfluité (cf Partie 1, chap 9) entre les règles, c'est pourquoi nous avons introduit un critère d'originalité entre les règles. Celui-ci permet de sélectionner uniquement les conjonctions de règles présentant un critère d'originalité que nous définissons par le fait que les sous-règles qui la composent ne sont pas triviales. Prenons par exemple la règle $A \wedge B \Rightarrow C$, elle est originale si son intensité d'implication est forte et si les règles $A \Rightarrow C$ ont $B \Rightarrow C$ une faible intensité d'implication. Les détails des calculs sont dans Couturier (2008).

Dans la suite nous présentons les modes de représentations graphiques offerts par CHIC. L'utilisateur souhaitant plus d'informations sur la manière de calculer les règles pourra consulter par exemple Gras et al. (2004) et les références s'y trouvant.

5 Hiérarchie des similarités et hiérarchique cohésitive

Dès que CHIC a calculé l'ensemble de toutes les règles en fonction des paramètres choisis par l'utilisateur, il est possible de construire une hiérarchie à partir de ces règles. Cette hiérarchie peut s'apparenter à une méthode de classification orientée ou non en fonction du type de calcul choisi "similarité ou implication". Cependant les manières de construire chacune de ces hiérarchies comportent certaines similitudes. Dans la suite une règle est appelée classe, elle agrège deux variables dans sa forme la plus simple. À chaque niveau de la classification, CHIC choisit la classe qui possède la plus grande cohésion (de similarité ou d'implication). Ensuite, à chaque étape, CHIC calcule un ensemble de nouvelles classes à partir des classes présentes dans la hiérarchie. Pour créer une nouvelle classe, on agrège une classe existante avec soit une variable qui n'a pas été agrégée pour l'instant, soit avec une autre classe de la hiérarchie. Néanmoins, chaque couple de variables lors de l'agrégation de deux classes doit avoir une intensité valide. Par exemple, la formation de la classe $((a,b),c)$ nécessite que les classes (a,c) et (b,c) aient une bonne cohésion (avec l'implication) ou soient similaires (avec l'analyse des similarités). La classe $((a,b),c)$ représente la règle

$(a \Rightarrow b) \Rightarrow c$ avec l'analyse implicative et représente le fait que a implique b et que la classe ((a,b),c) admet une bonne cohésion et que la classe (a, b) est similaire à c avec l'analyse des similarités. Pour plus de détails sur la formation de classes, nous invitons le lecteur intéressé à se référer à (Lerman, 1981; Gras et al., 1996).

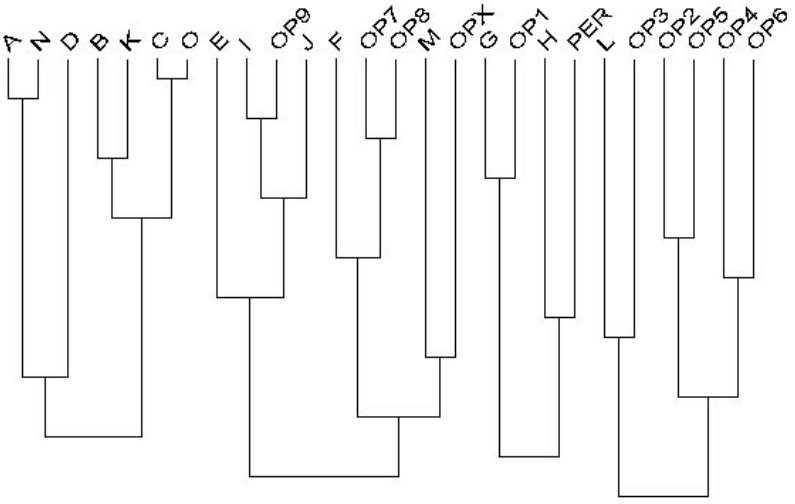


FIG. 3 – *Un exemple de hiérarchie des similarités*

Si l'utilisateur se demande quelle allure aurait la hiérarchie sans une ou plusieurs variables, il peut simplement les déselectionner grâce à l'interface prévue à cet effet. Cette possibilité est offerte pour toutes les représentations de CHIC (hiérarchie ou graphe). Malheureusement, une modification (même petite) portant sur la présence des variables implique une reconstruction totale de la hiérarchie. Cette étape dépend fortement du nombre de variables concernées dans le calcul (l'algorithme a une complexité qui dépend de la factorielle du nombre de variables dans le pire des cas). Avant de lancer une analyse, l'utilisateur peut choisir dans les options de calcul de détecter les niveaux significatifs de la hiérarchie.

La figure 3 montre une hiérarchie des similarités et la figure 4 illustre une hiérarchie cohésitive. Pour cette dernière, les niveaux significatifs sont affichés. Ils sont représentés par un trait rouge (dans CHIC) et ils signifient que le niveau signalé est plus significatif que le précédent et le suivant. Pour plus de détails sur leur construction sur les articles s'y rapportant, consulter par exemple Gras et al. (1996) et dans cet ouvrage Partie 1, chap 4.

L'indice de similarité est défini en utilisant soit la théorie classique soit la théorie entropique. Il est clair que le dernier choix est préférable avec un grand nombre d'individus. De plus, la construction de la hiérarchie des similarités avec la théorie classique conduit à la fin à une seule classe qui rassemble toutes les autres. Au contraire,

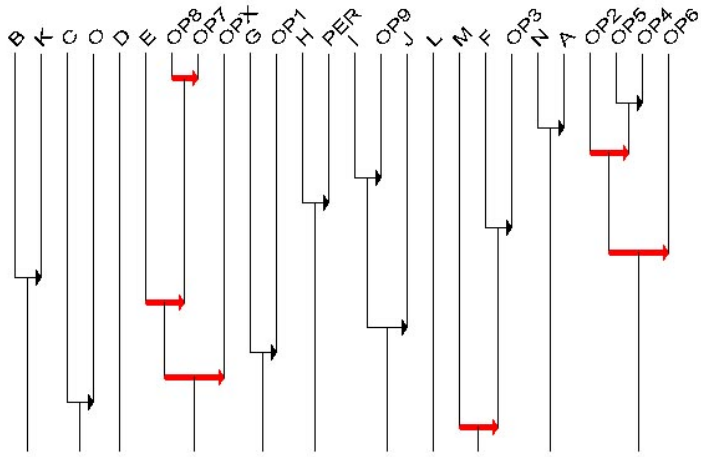


FIG. 4 – *Un exemple de hiérarchique cohésitive*

avec la version entropique de l'index de similarité, il est fréquent que l'algorithme de construction des classes conduise à plusieurs classes distinctes à la fin du processus. Le nombre de classes dépend en fait de la similarité des données.

Pour plus d'informations sur la hiérarchie cohésitive, le lecteur intéressé est inviter à consulter le chapitre 4 de la partie 1 de ce livre.

6 Graphe implicatif

Comme nous l'avons expliqué précédemment, les deux classifications de CHIC mettent en valeur certaines règles significatives en mettant de côté certaines autres règles. Si aucune variable ne doit être privilégiée, l'utilisation du graphe implicatif peut sembler judicieuse. Dans ce cas, l'utilisateur peut visualiser les règles dont l'intensité est plus grande qu'un seuil choisi. Dans la pratique, quatre seuils sont disponibles et CHIC propose des couleurs différentes les identifier plus rapidement. La figure 5 illustre un exemple de représentation d'un graphe implicatif. Une flèche est utilisée pour représenter l'implication entre deux variables (la règle $A \Rightarrow B$ est représentée par une flèche entre A et B). Comme le nombre de règles peut être important, l'utilisateur a la possibilité de sélectionner uniquement certaines variables. Ainsi seules les règles impliquant les variables présentes sont représentées. Par conséquent, ceci réduit le nombre de règles. De plus, afin de rendre le graphe plus lisible, CHIC utilise un algorithme de dessin automatique de graphes qui essaie de minimiser le nombre de croisements parmi les règles. Par défaut les fermetures transitives ne sont pas affichées sur le graphe implicatif. Un simple clique sur la souris dans la boite à outils les affiche. CHIC les calcule une fois pour toutes au début de chaque nouveau

graphe. Ensuite, même si l'utilisateur sélectionne ou désélectionne certaines variables, change le seuil d'affichage des règles, choisit d'afficher ou non les fermetures transitives, CHIC affiche le graphe sans aucun calcul supplémentaire. Cela permet à l'utilisateur de mettre en évidence les caractéristiques importantes de ses données. Néanmoins, l'utilisation de la procédure de dessin automatique de graphe est coûteuse en ressource de calculs, c'est pourquoi il n'est pas souhaitable de l'utiliser systématiquement.

Pour plus d'information sur la hiérarchie cohésitive, le lecteur intéressé pourra trouver de plus amples information dans la partie 1 de ce livre.

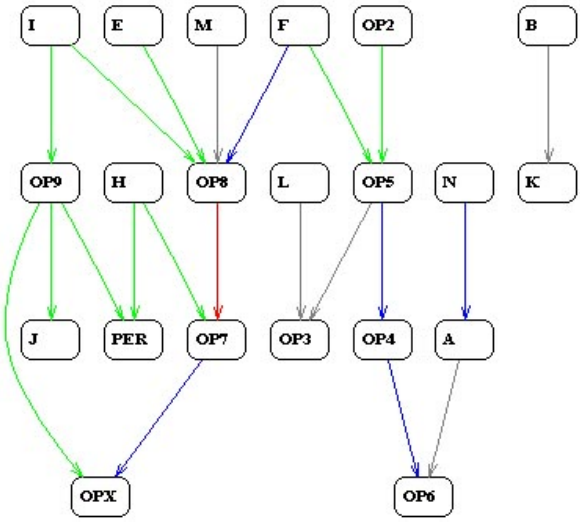


FIG. 5 – *Une exemple de graphe implicatif*

7 Autres possibilités

En plus des modes de représentation précédemment décrits, CHIC fournit quelques outils intéressants. Pour chaque mode de représentation graphique, il est possible de calculer la contribution et la typicalité d'un individu à une règle donnée. De la même manière, CHIC permet de calculer la contribution et la typicalité d'un ensemble d'individus à une règle donnée (cf Partie 1, chap 5).

La notion de contribution est définie pour déterminer les individus qui contribuent bien à la création de la règle. Ces individus sont plus responsables de la création de la règle que les autres. Par exemple si une règle $A \Rightarrow B$ possède une intensité implicative égale à $0,7$, alors les individus les plus contributifs sont ceux qui ont la valeur 1 pour les variables A et B. Par opposition, la notion de typicalité est définie par le fait que

certains individus soient "typiques" du comportement de la population, c'est-à-dire avec une intensité d'implication similaire à celle de la règle. Avec l'exemple précédent, les individus les plus "typiques" de la règle sont ceux qui possèdent respectivement des valeurs proches de $0,5$ et 1 pour A et B (ces valeurs dépendent du mode de calcul choisi pour l'analyse, classique ou entropique, et des cardinalités des ensembles A et B).

Il est facile de se rendre compte que les notions de typicalité et de contribution sont différentes. De la même manière, la notion de typicalité (respectivement de contribution) d'un ensemble d'individus (ou d'une catégorie d'individus) est définie pour savoir si un ensemble d'individus particulier est typique (respectivement contributif) d'une règle, d'une règle généralisée ou d'un chemin.

8 Illustration avec les variables intervalles et le calcul des typicalités et des contributions

Cette partie décrit un exemple simple et concret avec les deux variables sur intervalles de la section 3. La figure 6 montre un graphe implicatif issu des données de la figure 2. Les deux variables *poids* et *taille* sont automatiquement découpées en 4 intervalles par CHIC en suivant la méthode décrite dans la section 3. Dans ce graphe on peut remarquer quelques propriétés intéressantes. Par exemple, on peut remarquer les règles : $poids1 \Rightarrow taille12$ et $taille34 \Rightarrow poids2\text{-}4$. En raison du petit nombre d'individus pour cet exemple, par conséquent non significatif, et parce que les valeurs ont été générées de manière arbitraire, on ne peut rien dire de plus que : "les individus légers sont généralement petits et les personnes les plus grandes ne sont pas les plus légères". Néanmoins ces règles montrent une implication entre les partitions de deux variables. En considérant que les données puissent avoir du sens pour un expert, alors nous pourrions calculer la typicalité et la contribution d'un groupe d'individus. Par exemple, à propos de la règle $taille34 \Rightarrow poids2\text{-}4$, CHIC détermine que la variable *homme* contribue le plus à cette variable. Au contraire, la variable la plus typique à la règle $poids1 \Rightarrow taille12$ est la variable $femme$. Ces deux résultats ne sont pas surprenants compte tenu des données.

9 Conclusion

CHIC permet de mettre en pratique la plupart des méthodes et techniques liées à l'ASI. Dans ce chapitre, nous décrivons les caractéristiques principales de CHIC. Tout d'abord nous détaillons les types de variables que CHIC permet de traiter. Quelques options utiles pour comprendre CHIC sont détaillées. Ensuite, nous présentons les trois principaux modes de représentation. La hiérarchie des similarités et la hiérarchie cohésitive fournissent respectivement une classification orientée et une classification non orientée. Le graphe implicatif, qui est de loin le plus interactif, permet à l'utilisateur de "fouiller" parmi ses données et ainsi mettre en évidence les règles qui peuvent intéresser l'expert.

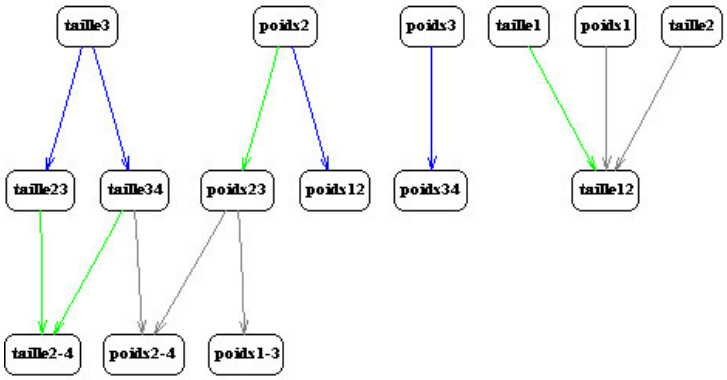


FIG. 6 – *Un exemple de graphe implicatif avec des variables intervalles*

Références

Ag Almouloud, S. (1992). *L'ordinateur, outil d'aide à l'apprentissage de la démonstration et de traitement de données didactiques*. Thèse de doctorat, Université de Rennes 1.

Agrawal, R., T. Imielinski, et A. Swami (1993). Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pp. 207–216.

Bojadziev, G. et M. Bojadziev (1996). *Fuzzy sets, fuzzy logic, applications*. World scientific.

Couturier, R. (2005). Un système de recommandation basé sur l'a.s.i. In *Troisième rencontre internationale de l'Analyse Statistique Implicative (ASI3)*, pp. 157–162.

Couturier, R. (2008). Statistical implicative analysis. In *CHIC : Cohesive Hierarchical Implicative Classification*, Volume 127 of *Studies in Computational Intelligence*, pp. 41–52. Springer Verlag.

Couturier, R., R. Gras, et F. Guillet (2004). Reducing the number of variables using implicative analysis. In *International Federation of Classification Societies, IFCS 2004*, pp. 277–285. Springer Verlag : Classification, Clustering, and Data Mining Applications.

Diday, E. (1971). La méthode des nuées dynamiques. *Revue de statistique appliquée 19*(2), 19–34.

Froissard, G. (2005). Chic et les études docimologiques. In *Troisième rencontre internationale de l'Analyse Statistique Implicative (ASI3)*, pp. 187–197.

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 de Rennes I.

Gras, R. (2005). Panorama du développement de l'A.S.I. à travers des situations fon-

datrices. In *Actes de la 3ème Rencontre Internationale A.S.I.*, pp. 9–33. Université de Palerme.

Gras, R., S. Ag Almouloud, M. Bailleul, A. Lahrer, M. Polo, H. Ratsimba-Rajohn, et A. Totohasina (1996). *L'implication Statistique*. La Pensée Sauvage.

Gras, R., R. Couturier, J. Blanchard, H. Briand, P. Kuntz, et P. Peter (2004). *Mesures de qualité pour la fouille de données*, Chapter Quelques critères pour une mesure de qualité de règles d'association. Un exemple : l'implication statistique, pp. 3–32. RNTI-E-1, Cepaduès Editions.

Larher, A. (1991). *Implication statistique et applications à l'analyse de démarches de preuve mathématique*. Thèse de doctorat, Université de Rennes I.

Lerman, I. C. (1981). *Classification et analyse ordinale des données*. Dunod.

Orus, P. et P. Gregori (2005). Des variables supplémentaires et des élèves "fictifs", dans la fouille didactique de données avec chic. In *Troisième rencontre internationale de l'Analyse Statistique Implicative (ASI3)*, pp. 279–291.

Ramstein, G. (2008). Statistical implicative analysis. In *Statistical Implicative Analysis of DNA microarrays*, Volume 127 of *Studies in Computational Intelligence*, pp. 205–225. Springer Verlag.

Ratsimba-Rajohn, H. (1992). *Contribution à l'étude de la hiérarchie implicative, application à l'analyse de la gestion didactique des phénomènes d'ostension et de contradictions*. Thèse de doctorat, Université de Rennes 1.

Rostam, H. (1981). Construction automatique et évaluation d'un graphe d'implication. Technical Report 150, IRISA, Rennes.

Totohasina, A. (1992). *Méthode implicative en analyse de données et application à l'analyse de conceptions d'étudiants sur la notion de probabilité conditionnelle*. Thèse de doctorat, Université de Rennes 1.

Summary

CHIC is a data analysis tool based on SIA. Its aim is to discover the more relevant implications between states of different variables. It proposes two different ways to organize these implications into systems: i) In the form of an oriented hierarchical tree and ii) as an implication graph. Besides, it also produces a (non oriented) similarity tree based on the likelihood of the links between states. The paper describes its history, its main features and its usage.

Chapitre 12 : Guide d'utilisation des principales fonctionnalités du logiciel CHIC[1]

Harrisson Ratsimba-Rajohn

33 Chemin des Peyrères, 33610 CANEJAN, FRANCE
ratsimba@u-bordeaux3.fr
http://mizara.ovh.org/recherche

Résumé : Dans ce chapitre**,** nous avons choisi d'utiliser les données d'une recherche en cours, pour présenter et décrire les étapes et procédures à suivre dans un traitement par le logiciel CHIC dans le but, d'une part, d'obtenir des graphes implicatifs et des représentations des hiérarchies cohésitives et, d'autre part, d'en faciliter la lecture grâce aux diverses options et fonctionnalités de ce logiciel. Nous aborderons par étape la description des données recueillies, leur transformation, le choix des variables supplémentaires, la préparation du fichier-texte au format *.CSV requis par CHIC, les traitements pour obtenir un graphe implicatif en particulier avec la notion de conjonction de prémisses et de seuil d'originalité, le traitement pour obtenir une représentation de la hiérarchie cohésitive, complété par les notions de niveaux significatifs, de contribution de groupes optimaux, de typicalité de variables supplémentaires ou d'individus. Dans un dernier nous aborderons par le traitement de variables sur intervalles.

1 Présentation

A partir d'un exemple de données, nous allons, dans cette partie, regarder les manières de manipuler CHIC afin d'obtenir des résultats d'analyses implicatives.

Les données collectées, sur lesquelles nous allons illustrer les manipulations sont issues d'une recherche[2] relative à l'utilisation des images en didactique de la biologie. Dans cette recherche on avait distingué deux types d'images : nous appellerons le premier "infographie" et le second "schéma". Brièvement, derrière "infographie" a été mise une représentation suffisamment proche de la "réalité" d'un processus cellulaire biologique et derrière "schéma" a été mise une représentation graphique suffisamment simplifiée et schématique de ce même processus. La[i] recherche qui a été menée consistait à étudier la manière dont des étudiants en

[1] Acronyme C.H.I.C pour désigner le logiciel : Classification Hiérarchique, Implicative, Cohésitive

[2] Dahmani H.R., (2008-2009) *Recherche en cours dans le cadre de sa thèse* , ENS (Ecole Normale Supérieure), Laboratoire de Didactique de la Biologie, Kouba, 16000, Alger, Algérie), sous la direction de P. Schneeberger, Equipe du DAESL (Laboratoire de Didactique, d'Anthropologie, d'Epistémologie et des Sciences du Langage), Université Bordeaux II, 33000, Bordeaux, France (http://pagesperso-orange.fr/daest/index.htm), et en partenariat avec Dr I Kramer, chercheur à l'Institut Européen de Chimie et de Biologie de Bordeaux. (http://www.cellbiol.net/home/index.php)

SVT (Bac+2) "percevait" le même processus cellulaire selon qu'ils traitaient une infographie ou un schéma le représentant durant une évaluation écrite.

Nos propos n'entreront pas sur le fond de cette recherche en didactique de la biologie ni dans sa forme mais se concentreront sur la manière d'utiliser le logiciel CHIC. C'est pourquoi, d'une part, nous ne procéderons pas comme habituellement dans une phase normale d'une recherche où, avant de se lancer dans une analyse des données, on pose d'abord les différentes questions et hypothèses qui nous aident à traiter une problématique. D'autre part nous ne procéderons pas aussi aux interprétations qui en découlent. .

Les données recueillies sont les résultats de l'évaluation écrite soumise à 82 étudiants le 28 mars 2008. Les 82 étudiants reçoivent les mêmes questions mais 42 ont comme support les images sous forme d'infographie et 40 les ont sous forme de schéma. La répartition des 2 types d'image a été faite aléatoirement.

Sur chaque étudiant ont été observés 7 caractères :

1. L'étudiant a passé l'évaluation sous la forme schéma: OUI/NON (variable valorisée 1 ou 0)
2. L'étudiant a passé l'évaluation sous la forme infographie: OUI/NON (variable valorisée 1 ou 0)
3. La note moyenne obtenue par l'étudiant lors de 3 devoirs précédents. (note /20)
4. La note obtenue par l'étudiant lors de cette évaluation écrite (note /20)
5. La note obtenue par l'étudiant sur une question demandant l'intitulé d'une image (note /20)
6. La note obtenue par l'étudiant sur une question demandant de légender l' image (note /20)
7. La note obtenue par l'étudiant sur une question demandant de décrire le processus représenté par l'image (note /20)

Nous avons désigné respectivement ces sept caractères par les codes de variables suivants :

Sch	inf	moy3	not3	titr	LEG	DES

Les données que nous allons traiter se présentent sous forme de tableau de 83 lignes et 9 colonnes.

Ainsi nous procéderons selon les étapes suivantes :

- Description des données recueillies et leur préparation;
- Mise en forme de ces données sur le tableur EXCEL, afin d'obtenir des données de type .CSV acceptées par le logiciel CHIC;
- Deux traitements avec 5 variables principales réelles et 2 variables supplémentaires binaires;
- Deux traitements avec des "variables sur intervalles" sans variables supplémentaires.

2 Description des données recueillies

Les données que nous allons traiter se présentent sous forme de tableau de 83 lignes et 9 colonnes. :

1. La première ligne contient les codes des variables et les autres lignes représentant les 82 individus qui ici sont les étudiants
2. La première colonne contient les codes des individus (ici ce sont des numéros de 1 à 86 car quelques numéros sont absents) puis les autres 7 colonnes contiennent les valeurs prises par les variables.

Ainsi dans le tableau ci-dessous l'individu 11 est absent, l'individu 12 a traité les questions avec les images sous forme de schéma, a obtenu 13/20 comme moyenne de 3 devoirs précédents, 16 comme note à l'évaluation; 0 comme note sur l'intitulé de l'image, 19 sur la légende et 18 sur la description ; par contre l'individu 8 a traité les questions ayant comme support l'infographie, il a obtenu 15/20 à la moyenne des 3 devoirs précédents, 17 à l'évaluation; 20 sur l'intitulé des images, 13 sur les légendes et 20 sur la description.

Notons que la première ligne de la première colonne est laissée vide.

	Sch	**inf**	**moy3**	**Not3**	**titr**	**LEG**	**DES**
1	1	0	19	19	10	20	18
2	1	0	12	19	10	20	20
5	1	0	13	17	10	20	16
6	1	0	14	17	20	19	20
7	0	1	16	17	20	16	18
8	0	1	15	17	20	13	20
9	1	0	10	17	20	19	18
10	1	0	18	17	10	19	20
12	1	0	13	16	0	19	18
				……			
84	0	1	5	2	0	2	3
86	1	0	9	2	10	3	2

TAB 1 – *Données brutes : source : http://mizara.ovh.org/recherche/chic/donnees_brutes.xls*

3 Transformation de données

Les données que nous allons traiter sont mises sur une feuille d'un classeur EXCEL. Mais CHIC ne peut pas traiter des données dont des valeurs de variables sont plus grandes que 1 (ni plus petites que 0), si ce ne sont pas des variables sur intervalles. Avec l'aide du tableur nous transformons ces valeurs comprises entre 0 et 20 en des valeurs proportionnelles comprises entre 0 et 1 pour obtenir des variables fréquentielles (Ici, il suffit de diviser par 20 chaque valeur) et nous obtenons le tableau :

	Sch	inf	moy3	not3	Titr	LEG	DES
1	1	0	0,95	0,95	0,50	1,00	0,90
2	1	0	0,60	0,95	0,50	1,00	1,00
5	1	0	0,65	0,85	0,50	1,00	0,80
6	1	0	0,70	0,85	1,00	0,95	1,00
7	0	1	0,80	0,85	1,00	0,80	0,90
8	0	1	0,75	0,85	1,00	0,65	1,00
9	1	0	0,50	0,85	1,00	0,95	0,90
10	1	0	0,90	0,85	0,50	0,95	1,00
12	1	0	0,65	0,80	0,00	0,95	0,90
			…				
84	0	1	0,25	0,10	0,00	0,10	0,15
86	1	0	0,45	0,10	0,50	0,15	0,10

TAB. 2 –*: Données en variables fréquentielles*

Selon la problématique étudiée on aurait pu choisir de diviser les valeurs d'une colonne par le maximum de la colonne. Ici nous avons fait le choix de diviser par la valeur maximale théorique. On aurait pu aussi obtenir pour une variable des valeurs binaires en demandant à Excel de mettre 0 si inférieur à tel seuil, sinon 1.

4 Choix des variables qui seront supplémentaires

En principe comme les variables à étudier sont les variables principales pour être prises en compte normalement dans CHIC pour les calculs des indices d'implication ou de similarité, nous pouvons alors considérer les deux premières variables Sch et inf comme pouvant être des variables supplémentaires. En effet les variables supplémentaires sont en général des descripteurs et sont ici des variables binaires (ou modales) et ils n'interviennent que dans le calcul des contributions des catégories. Par exemple, si on souhaite savoir si une implication est plutôt formée par des étudiants qui ont traité les schémas ou par ceux qui ont eu à traiter les infographies, on définit pour chaque individu le fait qu'il avait traité le schéma ou l'infographie. Ensuite, les variables Sch et inf ne sont prises en compte que lors de la recherche de la contribution ou la typicalité des catégories.. Pour indiquer à CHIC une variable supplémentaire, on ajoute au nom de la variable un espace et un "s". Ainsi les codes respectifs de ces deux variables supplémentaires sont "Sch s" et ."inf s". Nous obtenons alors le tableau suivant :

	Sch s	Inf s	moy3	not3	titr	LEG	DES
1	1	0	0,95	0,95	0,50	1,00	0,90
2	1	0	0,60	0,95	0,50	1,00	1,00
3	1	0	0,80	0,90	1,00	0,70	1,00
4	1	0	0,80	0,85	0,50	0,95	0,90
5	1	0	0,65	0,85	0,50	1,00	0,80
6	1	0	0,70	0,85	1,00	0,95	1,00
7	0	1	0,80	0,85	1,00	0,80	0,90
8	0	1	0,75	0,85	1,00	0,65	1,00
...	**...**	**...**	**...**	...	...	...	...

TAB. 3 – *Données avec variables supplémentaires*

5 Préparation du fichier-texte au format *.CSV requis par le logiciel CHIC

CHIC n'acceptant pas les données sous forme de tableau comme ci-dessus, les données doivent être sous la forme suivante (Deux valeurs adjacentes sont séparées par un "point-virgule") :

```
;Sch s;inf s;moy3;not3;titr;LEG;DES
1;1;0;0,95;0,95;0,5;1;0,9
2;1;0;0,6;0,95;0,5;1;1
3;1;0;0,8;0,9;1;0,7;1
4;1;0;0,8;0,85;0,5;0,95;0,9
5;1;0;0,65;0,85;0,5;1;0,8
6;1;0;0,7;0,85;1;0,95;1
7;0;1;0,8;0,85;1;0,8;0,9
8;0;1;0,75;0,85;1;0,65;1
…
```

TAB. 4 – *Données sous format CSV*

Ils seront contenus dans un fichier-texte de type (ou format) .CSV. On peut écrire par exemple ces données avec un traitement de texte simple (comme le bloc-note de Windows) en le sauvegardant sur un fichier-texte mais avec l'extension .CSV.

Toutefois, Il est bon de savoir que le tableur EXCEL peut enregistrer une feuille de calcul sous ce format .CSV. Il suffit de choisir dans le menu *Fichier* la rubrique *Enregistrer sous…*, de donner un nom sans tiret-bas, ni espace et enfin de choisir comme type de fichier : CSV (séparateur point-virgule) (*.csv). Signalons qu'avec le tableur Calc du logiciel libre Open Office la production de ce type de fichier .CSV est aussi faisable avec quelques précautions supplémentaires. A partir du traitement de texte WORD ou de celui WRITER d'Open Office, on peut également obtenir des valeurs séparées par des virgules en utilisant judicieusement la rubrique "convertir un tableau en texte" du Menu Tableau. Quelle que soit la manière d'obtenir le fichier-texte de type .CSV, nous l'appellerons "donnees01s.csv" en vue du

traitement par le logiciel CHIC. Il est essentiel, pour raison technique, que ce nom ne comporte ni espace, ni tiret-bas.

6 Traitement pour obtenir un graphe implicatif

Pour simplifier, mettons d'abord le fichier "donnees01s.csv" dans le même dossier que l'application Chic.exe.

6.1 Exécutons Chic.exe

A ce stade, il convient de lancer l'exécution du programme CHIC.exe.

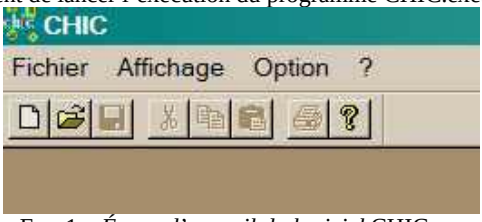


FIG. 1 – *Écran d'accueil du logiciel* CHIC

6.2 Choix des options

Choisissons les Options. Préparons les paramétrages dans le menu Options. Dans un premier temps, pour une première approche avec un graphe implicatif, nous choisissons le Type d'implication selon la théorie classique avec le type de loi qui sera la loi binomiale. Historiquement ce sont les bases initiales des calculs. Par ailleurs, pour simplifier, nous garderons les couleurs des seuils proposés par défaut, puis dans les Options d'Apriori nous choisissons, pour commencer, 1 prémisse. Ceci signifie que, dans une autre analyse ultérieure, on peut paramétrer CHIC, à travers cette Option d'Apriori, pour représenter une implication dont la prémisse est formée par un ensemble de deux ou trois variables qui impliquerait d'autres variables. Dans un second temps, avec des éclaircissements théoriques et en fonction de la problématique traitée, on pourra choisir d'autres combinaisons d'options et de paramètres.

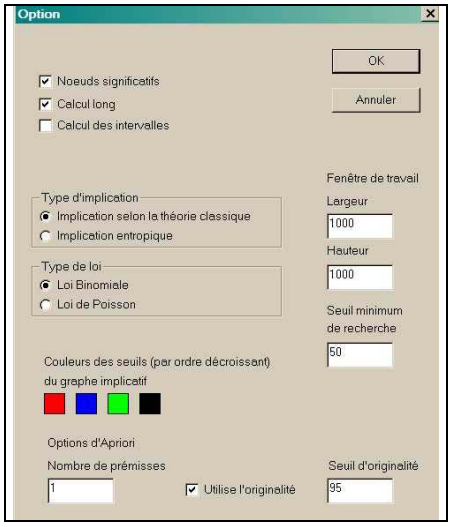


FIG. 2 – *Menu options du logiciel* CHIC

6.3 Traitement

Procédons au traitement en choisissant dans le Menu Fichier la rubrique : Nouveau traitement … et en cherchant d'ouvrir le fichier "donnees01s.csv" que nous avons préparé auparavant.

6.4 Choix de l'analyse

CHIC nous propose de choisir l'analyse qui nous intéresse parmi les trois proposés : Arbre des similarités, Graphe implicatif, Arbre cohésitif.

6.5 Choix de Graphe implicatif

Ayant choisi Graphe implicatif, trois fenêtres s'ouvrent :

- Fenêtre 1 : Elle contient 5 tableaux d'information, de résultats de calculs intermédiaires (dont la matrice des corrélations) et de la matrice des indices d'implications (compris entre 0 et 100).
- Fenêtre 2 : Elle contient le tableau des valeurs traitées en plus des mêmes tableaux que la fenêtre 1 mais présentées d'une manière plus conviviale
- Fenêtre 3 : Elle présente le graphe implicatif calculé .

Le Menu Fenêtre permet de choisir la fenêtre à visualiser.

6.6 Paramétrage : choix des seuils d'indices d'implication

Si la fenêtre 3 n'affiche rien (comme c'est le cas avec nos données), il est nécessaire de vérifier à droite de la fenêtre la Barre des paramètres du graphe implicatif en y choisissant les seuils des indices d'implication qui doivent apparaître avec les couleurs correspondantes.

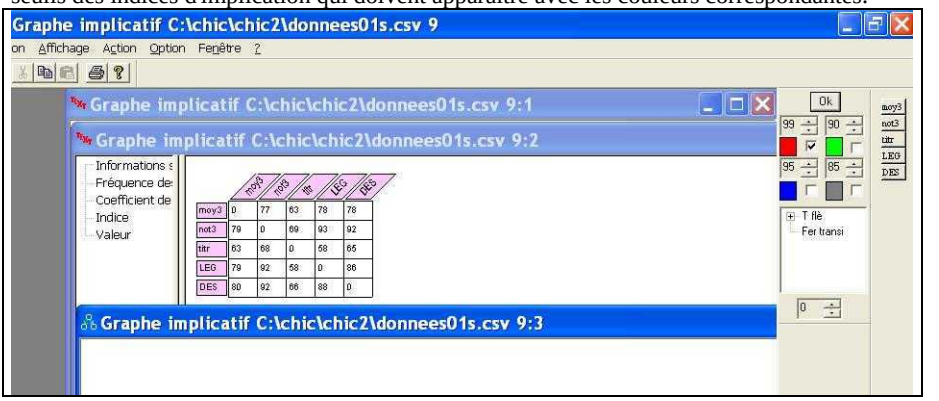


FIG. 3 – *Matrice d'implication*

Effectivement, sur l'image ci-dessus (FIG 3) seuls les indices au moins égaux à 99 ont été choisis alors que si nous consultons la matrice d'implication (Fenêtre 2) l'indice maximal obtenu sur nos données est : 93.

Ainsi nous pouvons choisir les indices à faire apparaître avec le paramétrage de ces seuils et des couleurs correspondantes

6.7 Paramétrage : différenciation des seuils

A partir de l'observation de la matrice d'implication, on peut décider de choisir les implications à faire apparaître en manipulant judicieusement les 4 seuils possibles. Dans la zone T flè, après avoir développé le (+), on peut choisir l'épaisseur d'une flèche correspondant à un seuil. En cliquant sur Fer Transi et en choisissant 1 on peut obtenir le tracé de tous les arcs associés aux fermetures transitives aux seuils choisis, et avec respect des couleurs. Ces nouveaux arcs apparaissent en pointillés. Mais le choix de faire apparaître les fermetures transitives risque d'augmenter la complexité du graphique implicatif, même si on la possibilité de choisir les variables qui apparaissent ou non soit en cliquant, soit en décliquant sur un au moins des codes des variables vus sur la Barre des items placée à l'extrême droite de la fenêtre 3.

Après les différents choix effectués sur la Barre des paramètres du graphe implicatif et sur la Barre des items, on recalcule le graphique soit en mettant Ok, soit en faisant menu contextuel sur l'espace de travail des graphes (généralement sur PC : un clic droit), puis en choisissant la rubrique "Dessine le graphe de manière automatique". Cette action est aussi accessible dans le Menu Action parmi d'autres possibilités d'analyse. Si l'espace de travail pour dessiner le graphique est insuffisant, il suffit de modifier les dimensions en pixel dans le menu Option.

Remarquons que les deux variables supplémentaires n'apparaissent pas.

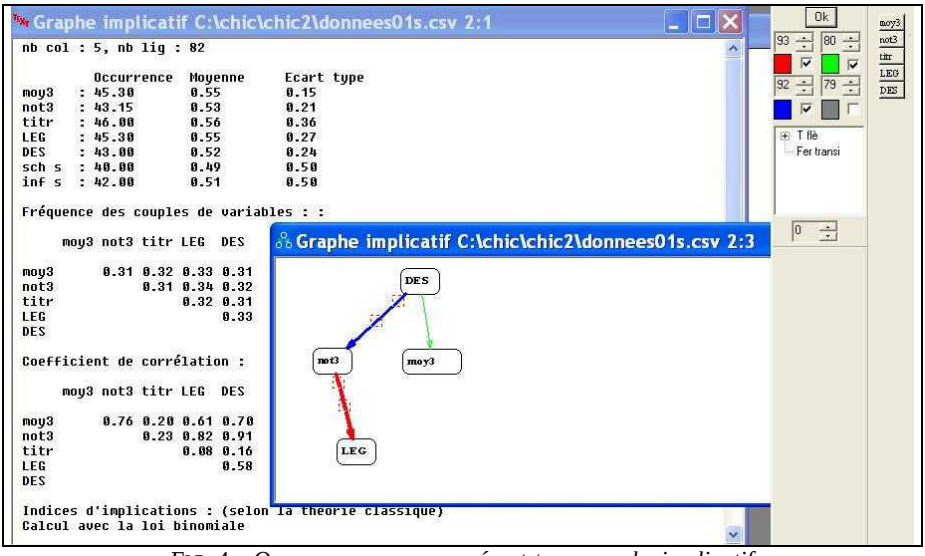


FIG. 4 – *Occurrence, moyenne, écart-type, graphe implicatif*

6.8 Utilisation de l'option : conjonction de variables

Si nous regardons de plus près, le graphe implicatif ci-dessus, il est assez dépouillé. La variable « titr » qui repère les étudiants ayant pu donner correctement ou non un titre aux images proposées n'entre même pas en jeu : elle n'apparaît pas dans la chaine des implications au niveau que nous avons choisi (seuil d'intensité implicative 80). Cette variable semble ne pas avoir d'effet significatif. Mais le fait de la « coupler » avec d'autres variables nous apportera probablement plus d'information.

En effet, CHIC permet de calculer des conjonctions entre les variables prémisses de l'implication et ceci de manière automatique.

Ainsi, si on choisit par exemple 3 prémisses, on obtient des règles de la forme $a \wedge b \wedge c \Rightarrow d$. Le choix du nombre de prémisses qui interviendront dans les règles se fait dans le menu « Option » au niveau des paramètres « Option d'Apriori ». Si vous choisissez 1, alors vous aurez seulement des règles de la forme $a \Rightarrow b$. Si vous choisissez 2, vous aurez des règles de la forme $a \wedge b \Rightarrow c$ et ainsi de suite. Le calcul des conjonctions peut rapidement engendrer un nombre considérable de règles. La figure ci-dessous donne une idée de la complexité de la représentation obtenue à partir de seulement 5 variables avec le choix de 2 prémisses et si on ne contrôle pas ce que nous appelons «le seuil d'originalité ».

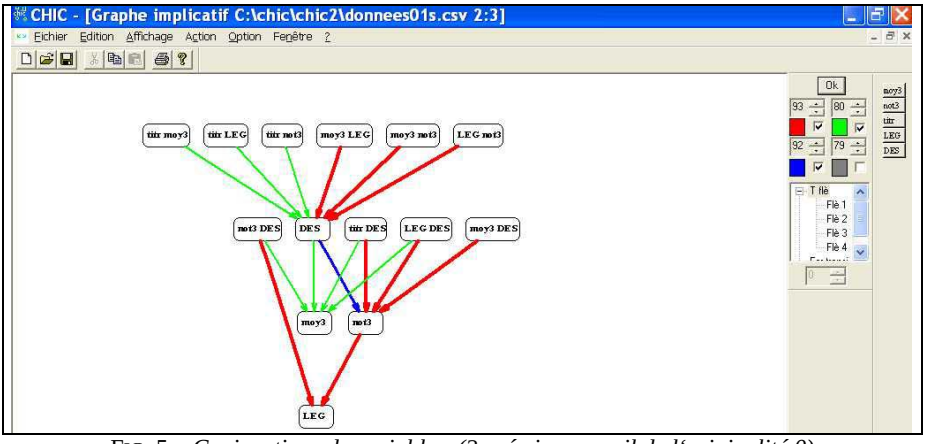


FIG. 5 – *Conjonctions de variables, (2 prémisses, seuil de l'originalité 0)*

Le graphe ci-dessus nous montre donc les 10 prémisses obtenues par les combinaisons de 5 éléments pris 2 à 2. Or, il est possible de retenir uniquement les règles ayant un caractère original, optimisant le nombre de conjonctions significatives pour le nombre de prémisses choisi. D'où l'option prévue à cet effet avec un seuil qui permet de retenir uniquement les règles supérieures à ce seuil. Le seuil d'originalité est un nombre compris entre 0 et 100. Pour approfondir cette notion, on peut consulter (Couturier R., 2004, 2008).

La figure ci-dessous, nous présente la conjonction qui serait la plus originale après qu'on ait choisi le seuil d'originalité 50 avec le seuil d'intensité implicative de 80 déjà préalablement sélectionné.. Dans notre exemple, c'est la conjonction de la variable « titr » et de la variable « moy3 » qui impliquerait la variable « DES ». Rappelons que la variable « titr » n'était même pas dans le graphe implicatif initial sans conjonction alors que sa moyenne est « la plus réussie ». Pourquoi apparaît-elle ici en conjonction avec la variable « moy3 » ?

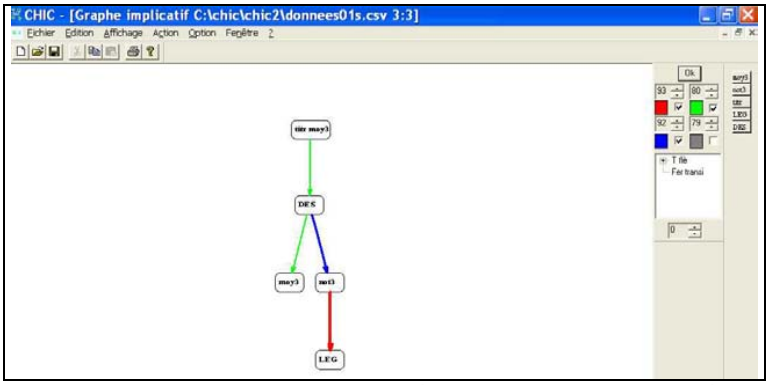


FIG. 6 – *Conjonctions de variables, (2 prémisses et seuil de l'originalité 50)*

Indéniablement, avec les conjonctions de variables le chercheur peut obtenir des informations complémentaires utiles au traitement de sa problématique.

7 Traitement pour obtenir un arbre cohésitif

7.1 Exécutons Chic.exe ou si CHIC est déjà ouvert, fermons le dernier travail

FIG. 7 – *Choix du traitement : arbre cohésitif*

7.2 Choix des options

Choisissons les Options. Préparons les paramétrages dans le menu Options. Dans un premier temps, pour une première approche avec un arbre implicatif, nous choisissons de faire apparaître les niveaux significatifs avec le Type d'implication selon la théorie classique et avec le type de loi qui sera la loi binomiale. Dans un second temps, avec des éclaircissements théoriques (cf. Partie 1) et en fonction de la problématique traitée on pourra choisir d'autres combinaisons d'options et de paramètres.

7.3 Début du traitement

Procédons au traitement en choisissant dans le Menu Fichier la rubrique : Nouveau traitement et en cherchant d'ouvrir le fichier "donnees01s.csv" que nous avons préparé auparavant.

7.4 Choix de l'analyse

CHIC nous propose de choisir l'analyse qui nous intéresse parmi les trois proposés : Arbre des similarités, Graphe implicatif, Arbre cohésitif.

7.5 Traitement "Arbre cohésitif"

Ayant choisi "Arbre cohésitif", trois fenêtres s'ouvrent (cf. FIG.8) :

- Fenêtre 1 : Elle contient en plus des 4 tableaux d'information déjà présents lors du traitement pour le graphe implicatif, la matrice des indices de cohésion (valeurs de 0 à 1), la classification des niveaux de cohésions et l'indication des nœuds les plus significatifs, option que nous avons demandée.
- Fenêtre 2 : Elle contient le tableau des valeurs traitées en plus des mêmes tableaux que la fenêtre 1 mais présentées d'une manière plus conviviale
- Fenêtre 3 : Elle présente l'arbre cohésitif avec les nœuds significatifs mis en couleurs.

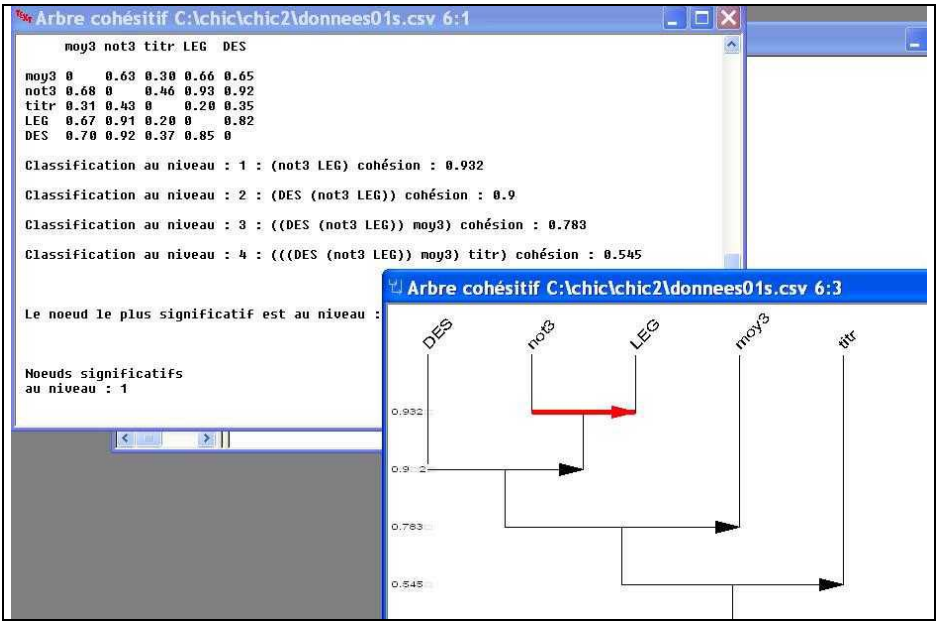


FIG. 8 – *Les cohésions à chaque niveau*

7.6 Compléments d'analyse

Dans le Menu "Action" nous pouvons choisir plus de possibilités d'analyse telle que « Calcul des contributions des variables supplémentaires ». Dès que ce complément d'analyse est demandé, dans la Fenêtre 1 apparaissent les informations correspondantes (cf. FIG 8):

7.7 Lecture des contributions des variables supplémentaires

La lecture de ces contributions nous révèle les risques d'erreur à déclarer que telle variable supplémentaire contribuerait à la formation de la classe de cohésion implicative

Exemple, avec nos données, pour la classe de cohésion implicative : DES, not3, LEG, moy3 (1,2,3) formée au niveau 3, on lit que "La variable supplémentaire sch contribue à cette classe avec un risque de : 0.00588 " et que "La variable supplémentaire inf contribue à cette classe avec un risque de : 0.996 ", autrement dit, normalement il y a 6 chances sur 1000 d'être dans l'erreur de déclarer que ce sont les étudiants ayant traité les schémas qui ont permis la construction de cette classe de cohésion implicative DES, not3, LEG, moy3 (1,2,3), alors que plus de 99 fois sur 100, on sera dans l'erreur de déclarer que ce sont les étudiants ayant traité l'infographie qui y ont contribué.

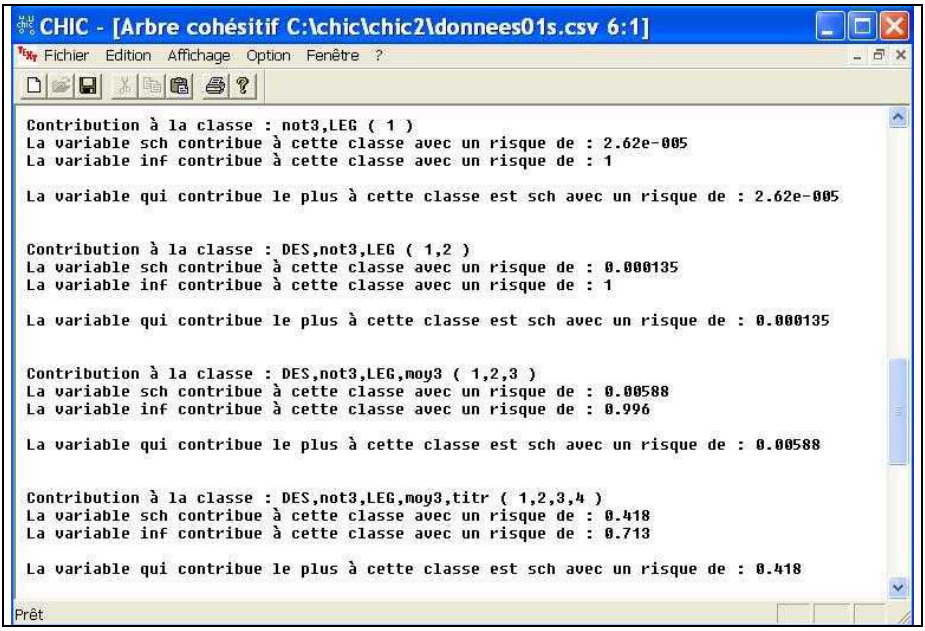


FIG. 9 – *Contributions des variables supplémentaires*

7.8 Typicalité d'une variable supplémentaire

Dans le Menu Action de la Fenêtre 3, on peut demander aussi comme complément d'analyse la typicalité des variables supplémentaires. La Typicalité permet de connaître quelle variable supplémentaire (c'est-à-dire quel groupe d'individus) serait la plus typique d'une classe de cohésion implicative. La typicalité d'une variable supplémentaire sur une classe de cohésion implicative est aussi mesurée en risque d'erreur de la déclarer comme caractéristique de la classe. Ces typicalités s'affichent dans la Fenêtre 1.

7.9 Contributions des individus

Pour des analyses plus fines, avec CHIC et une rubrique de la Menu Action de la Fenêtre 3, on a la possibilité de connaître dans quelle mesure chaque individu a contribué à l'élaboration d'une classe de cohésion implicative. CHIC calcule en plus le sous-groupe d'individus qui a permis cette classe. C'est le **groupe optimal** qui a contribué le plus à la fabrication de cette classe. Ces informations sont affichées dans la Fenêtre 1

Exemple extrait de la Fenêtre 1 avec nos données sur la classe de cohésion implicative formée au niveau 3 :

- Contribution à la classe : DES, not3, LEG, moy3 (1, 2, 3)
- Contribution des individus, évaluée de 0 à 1

1 : 0.966	2 : 0.650	3 : 0.686	4 : 0.843	5 : 0.795	6 : 0.680
7 : 0.819	8 : 0.635	9 : 0.634	10 : 0.809	12 : 0.710	13 : 0.702
14 : 0.676	15 : 0.842	16 : 0.714	17 : 0.848	18 : 0.732	19 : 0.878
20 : 0.767	21 : 0.836	22 : 0.616	23 : 0.831	24 : 0.805	25 : 0.791
26 : 0.805	27 : 0.723	28 : 0.761	29 : 0.770	30 : 0.713	32 : 0.757
33 : 0.748	34 : 0.762	35 : 0.745	36 : 0.598	37 : 0.580	38 : 0.690
39 : 0.732	40 : 0.756	41 : 0.715	42 : 0.660	43 : 0.759	44 : 0.593
45 : 0.732	46 : 0.753	47 : 0.601	48 : 0.673	49 : 0.546	50 : 0.540
51 : 0.657	52 : 0.727	53 : 0.702	54 : 0.714	55 : 0.671	56 : 0.677
57 : 0.703	58 : 0.721	59 : 0.686	60 : 0.657	61 : 0.696	62 : 0.704
63 : 0.682	64 : 0.691	65 : 0.697	66 : 0.605	67 : 0.585	68 : 0.642
69 : 0.607	70 : 0.631	71 : 0.625	72 : 0.654	73 : 0.651	74 : 0.614
75 : 0.657	76 : 0.612	78 : 0.640	79 : 0.619	80 : 0.565	81 : 0.565
82 : 0.559	83 : 0.568	84 : 0.554	86 : 0.575		

TAB. 5 – *Contribution de chaque individu à un niveau de classe de cohésions*

Groupe optimal : card 34
12 30 54 16 41 58 27 52 18 39 45
35 33 46 40 32 43 28 34 20 29 25
5 26 24 10 7 23 21 15 4 17 19
1

TAB. 6 – *Un groupe optimal d'individus contribuant à la formation d'un niveau de classe de cohésions*

7.10 Typicalité de chaque individu

D'une manière analogue à la typicalité d'une variable supplémentaire, on peut connaître d'une manière plus fine la typicalité de chacun des individus et aussi, avec la liste "groupe optimal", le sous-groupe d'individus qui serait le plus typique d'une classe de cohésion implicative.

Exemple extrait de la Fenêtre 1 avec nos données sur la classe de cohésion implicative formée au niveau 3 :

- Typicalité à la classe : DES, not3, LEG, moy3 (1,2,3)

- Typicalité des individus évaluée de 0 à 1 :

1 : 0.797	2 : 0.566	3 : 0.402	4 : 0.887	5 : 0.842	6 : 0.562
7 : 0.798	8 : 0.301	9 : 0.547	10 : 0.679	12 : 0.641	13 : 0.281
14 : 0.590	15 : 0.845	16 : 0.571	17 : 0.841	18 : 0.410	19 : 0.846
20 : 0.760	21 : 0.813	22 : 0.082	23 : 0.775	24 : 0.745	25 : 0.698
26 : 0.725	27 : 0.637	28 : 0.525	29 : 0.664	30 : 0.354	32 : 0.611
33 : 0.520	34 : 0.630	35 : 0.592	36 : 0.150	37 : 0.013	38 : 0.366
39 : 0.558	40 : 0.583	41 : 0.415	42 : 0.249	43 : 0.584	44 : 0.088
45 : 0.545	46 : 0.572	47 : 0.189	48 : 0.263	49 : 0.020	50 : 0.009
51 : 0.216	52 : 0.486	53 : 0.476	54 : 0.501	55 : 0.294	56 : 0.307
57 : 0.432	58 : 0.481	59 : 0.412	60 : 0.253	61 : 0.432	62 : 0.431
63 : 0.414	64 : 0.400	65 : 0.442	66 : 0.279	67 : 0.006	68 : 0.201
69 : 0.112	70 : 0.265	71 : 0.229	72 : 0.246	73 : 0.246	74 : 0.131
75 : 0.319	76 : 0.142	78 : 0.283	79 : 0.168	80 : 0.000	81 : 0.000
82 : 0.010	83 : 0.043	84 : 0.018	86 : 0.070		

TAB. 7–*: Typicalité de chaque individu à un niveau de classe de cohésions*

Groupe optimal : card 40
62 61 57 65 53 58 52 54 33 28 45
9 39 6 2 16 46 40 43 14 35 32
34 27 12 29 10 25 26 24 20 23 1
7 21 17 5 15 19 4

TAB. 8 – *Un groupe optimal d'individus typique d'un niveau de classe de cohésions*

7.11 Comparaison des individus avec le groupe optimal

La comparaison des individus associés à une variable supplémentaire typique d'une classe de cohésion implicative avec le groupe optimal des individus typiques de cette classe devrait apporter des informations intéressantes pour la problématique étudiée. Idem pour l'étude des contributions.

7.12 Utilité des groupes optimaux

La possibilité d'obtenir les groupes optimaux est intéressante si des variables supplémentaires ne sont pas au préalable définies.

8 Traitement de variables sur intervalles

Nous allons procéder à une analyse qui considère des **variables sur intervalles**.

Revenons sur la forme initiale de nos données mais sans les 2 variables supplémentaires Sch et Inf qui rappelons-le jouent le rôle de descripteurs. Au cours des traitements suivants nous allons regarder dans chacun des deux groupes comment se structurent les 5 variables principales du point de vue de la cohésion implicative. Nous avons choisi de procéder de

cette sorte puisque lors des analyses précédentes, nous avons observé que la variable supplémentaire Inf, contrairement à Sch, ne contribuerait nullement à la structuration des 5 variables principales sur les 82 individus.

Ainsi, à partir du tableau initial, nous créons deux tableaux. Le premier contenant les résultats du sous-groupe d'étudiants ayant 1 à la variable "Inf" (42 étudiants) et le second contenant ceux qui ont traité les images sous forme de schéma (40 étudiants). Les deux tableaux ont chacun les 5 variables principales.

1. La note moyenne obtenue par l'étudiant lors de 3 devoirs précédents. (note /20) (code : moy3)
2. La note obtenue par l'étudiant lors de cette évaluation écrite (note /20) (code : not3)
3. La note obtenue par l'étudiant sur une question demandant l'intitulé d'une image (note /20) (code : titr)
4. La note obtenue par l'étudiant sur une question demandant de légender l' image (note /20) (code : LEG)
5. La note obtenue par l'étudiant sur une question demandant de décrire le processus représenté par l'image (note /20) (code : DES)

Les 5 premières lignes du premier tableau "données-inf" contenant les notes des étudiants ayant traité les images sous forme "infographie" (la première colonne contient les codes des étudiants) se présentent ainsi :

	moy3 p	**not3 p**	**titr p**	**LEG p**	**DES p**
7	16	17	20	16	18
8	15	17	20	13	20
18	16	15	10	12	17
22	16	14	0	9	18

TAB. 9 – *1er groupe de données sous forme variables sur intervalles*

Et les 6 premières lignes du second tableau "données-sch" contenant les notes des étudiants ayant traité les images sous forme "schéma" (la première colonne contient les codes des étudiants) se présentent ainsi :

	Moy3 p	**not3 p**	**titr p**	**LEG p**	**DES p**
1	19	19	10	20	18
2	12	19	10	20	20
3	16	18	20	14	20
4	16	17	10	19	18
5	13	17	10	20	16

TAB. 10 – *2ème groupe de données sous forme variables sur intervalles*

Remarquons que la première ligne contient les codes respectifs des 5 variables principales, mais comme nous voudrions les traiter comme des variables sur intervalles, nous avons accolé à chaque code initial le caractère "espace" suivi du caractère "p". Par exemple, "**moy3**" est devenu "**moy3 p**".

Ces deux tableaux doivent donc être enregistrés dans deux fichier-textes de type .CSV nommés respectivement "donnees-inf.csv" et "donnees-sch.csv"

Les 6 premières lignes du fichier "donnees-inf.csv" sont sous cette forme :

```
;moy3 p; not3 p; titr p; LEG p; DES p
7;16;17;20;16;18
8;15;17;20;13;20
13;15;16;20;11;16
18;16;15;10;12;17
22;16;14;0;9;18
```

TAB. 11 – *données sous forme variables sur intervalles en format .CSV*

Rappelons que lorsqu'une variable est proposée au logiciel CHIC comme **variable sur intervalles**, CHIC "découpe les valeurs de cette variable en un nombre fixe d'intervalles. Le nombre d'intervalles est choisi par l'utilisateur et ensuite l'algorithme des nuées dynamiques Diday (thèse d'Etat, 1972) constitue automatiquement les intervalles qui ont des limites distinctes. Cet algorithme a la particularité de construire des intervalles en minimisant l'inertie de chaque intervalle. Ensuite, un intervalle est représenté par une variable binaire et un individu a la valeur 1 s'il appartient à cet intervalle et 0 sinon. En utilisant une telle décomposition, un individu appartient à un seul intervalle" . Durant leurs traitements, ces nouvelles présentations des données sont enregistrées par CHIC dans un fichier-texte de type .CSV dont le nom se termine par "_partition.csv"; par exemple pour nos données, nous obtenons "donnees-inf_partition.csv" ou "donnees-sch_partition.csv".

Les 6 premières lignes du fichier "donnees-inf_partition.csv" sont sous cette forme :

```
;moy3@1 i;moy3@2 i;moy3@3 i;not3@1 i;not3@2 i;not3@3 i;titr@1 i;titr@2 i;titr@3 i;LEG@1 i;LEG@2 i;LEG@3 i;DES@1 i;DES@2 i;DES@3 i
7;0;0;1;0;0;1;0;0;1;0;0;1;0;0;1
8;0;0;1;0;0;1;0;0;1;0;0;1;0;0;1
13;0;0;1;0;0;1;0;0;1;0;0;1;0;0;1
18;0;0;1;0;0;1;0;1;0;0;0;1;0;0;1
22;0;0;1;0;0;1;1;0;0;0;1;0;0;0;1
```

TAB. 12 – *fichier .CSV de données binaires générées par chic à partir de variables sur intervalles*

En effet, au moment où nous exécutons l'application CHIC, dans une première approche, nous paramétrons les options et en particulier nous ne cochons pas le paramètre "Calcul des intervalles"; En effet, si, en vue d'une analyse plus approfondie, nous le cochons, les traitements correspondants traiteront des **variables-intervalles** : c'est-à-dire "dans la suite du calcul, toutes les unions des intervalles d'une variable sont considérées", apportant par conséquent des résultats plus fins mais plus complexes, ainsi dans une première approche nous ne demandons pas cette option de "calcul des intervalles".

Après avoir paramétré les options, nous demandons un nouveau traitement sur le fichier "donnees-inf.csv" avec le type de calculs "Arbre cohésitif" et comme CHIC a repéré au moins une variable avec le code terminé par "une espace suivie de p", pour cette variable il pose à l'utilisateur une question sur le nombre de subdivisions que comportera la partition de

l'intervalle. Pour notre exemple, nous avons choisi la partition en 3 parties. Ainsi pour la variable "LEG p" nous allons obtenir les 3 nouvelles variables codées:

LEG@1 i	**LEG@2 i**	**LEG@3 i**

Dans la fenêtre 1, en plus des résultats intermédiaires habituels (5 tableaux d'information, de résultats de calculs intermédiaires (dont la matrice des corrélations) et de la matrice des indices de cohésions (compris entre 0 et 1)), nous obtenons des informations sur la partition obtenue par l'algorithme des nuées dynamiques, par exemple, le tableau suivant pour la variable sur intervalle LEG p:

Partitions optimales :
LEG1 de 2 à 4
LEG2 de 6 à 9
LEG3 de 11 à 16

TAB. 13 – *Exemple de partition en 3 parties obtenue par l'algorithme des nuées dynamiques*

Ce tableau nous informe qu'un étudiant ayant eu entre 2 et 4 inclus sera classé dans la variable LEG1, que celui qui a eu entre 6 et 9 inclus sera dans LEG2, et que dans LEG3 seront classés ceux ayant une note comprise entre 11 à 16. En gros LEG1 sont ceux qui ne savent pas légender et LEG3 ceux qui ont pu reconnaître une majorité de légendes et LEG2 ceux qui sont juste moyens. Pour information, la répartition des étudiants sur ces subdivisions est LEG1 : 8, LEG2 : 25, LEG3 : 9.

La fenêtre 3 nous donne l'arbre cohésitif suivant des variables sur intervalles : (cf FIG 9)

Brièvement, nous pouvons lire sur ce graphe qu'il y a 3 niveaux significatifs et qu' au premier niveau significatif le fait de ne pas savoir décrire entraînerait le fait d'avoir une mauvaise note à l'évaluation ; autrement dit en appliquant la contraposée "le fait de ne pas avoir une mauvaise note à l'évaluation" impliquerait que "l'étudiant ne soit pas mauvais sur la description". Et qu'au second niveau significatif de la cohésion (LEG1, (DES1,not31)) impliquerait MOY31, c'est-à-dire, en gros, "ne rien connaître sur le fait de savoir légender" ou "ayant une mauvaise note à l'évaluation même si la description n'est pas mauvaise" inpliquerait "une mauvaise moyenne au 3 devoirs précédents"

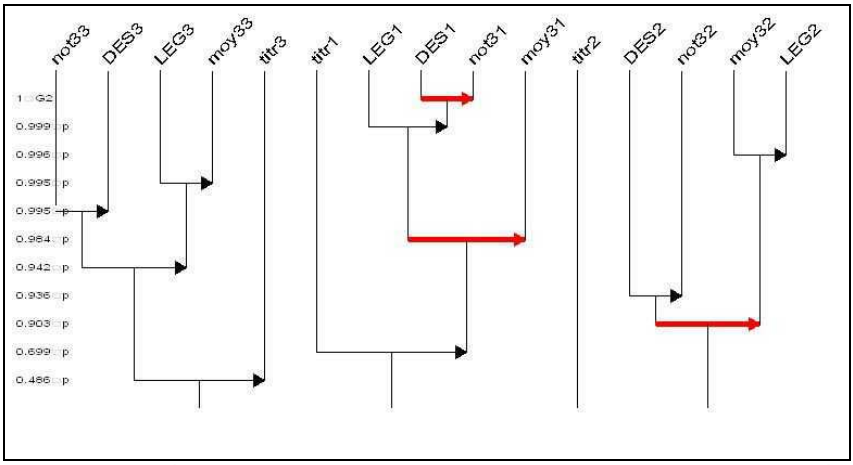


FIG. 10 – *arbre cohésitif suivant des variables sur intervalles (1er sous groupe de données)*

Or, nous ne retrouvons pas ces structures de cohésions implicatives si nous menons l'analyse sur le sous-groupe d'étudiants ayant traité les images sous forme de schéma.

En effet, le traitement des données contenues dans le fichier "données-sch.csv" où les 5 variables principales sont des variables sur intervalles (données des étudiants ayant traité l'évaluation avec les images sous forme de schémas),: nous a produit l'arbre cohésitif suivant

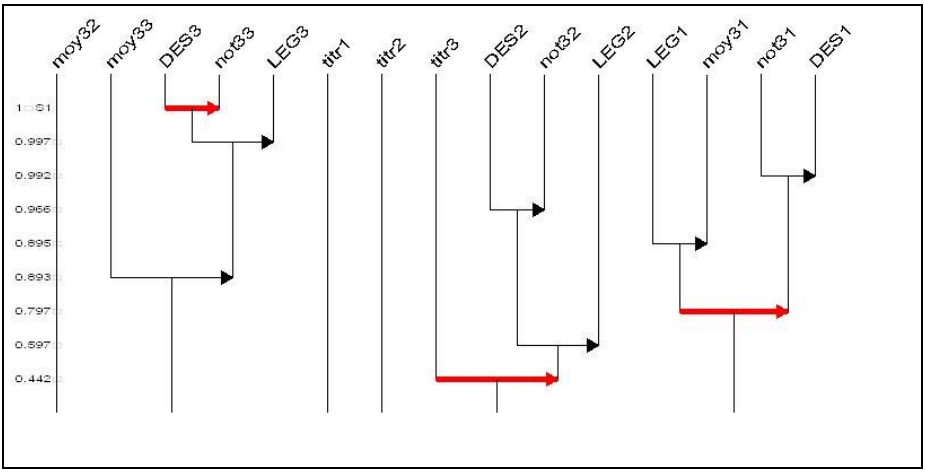


FIG. 11 – *arbre cohésitif suivant des variables sur intervalles (2ème sous groupe de données)*

Pour ce dernier traitement, nous avons aussi choisi 3 subdivisions pour chaque variable sur intervalle. Sur ce résultat on peut lire que pour ce groupe d'étudiants, au premier niveau significatif c'est le fait de savoir décrire qui impliquerait probablement le fait d'avoir une bonne note sur l'évaluation.; C'est-à-dire qu'il y aurait certains étudiants qui ont eu de bonnes

notes sur l'évaluation sans même savoir décrire. Ce qui correspond bien au tableau de contingence suivant entre les variables sur intervalles not33 et Des3 (codés respectivement dans le fichier "donnees-inf_partition.csv" **not3@3 i** et **DES@3 i)**

Sch	**not3@3 i**		
DES@3 i	**0**	**1**	**Total**
0	26	2	28
1	**0**	**12**	12
Total	26	14	40

TAB. 14 – *Tableau de contingence*

En comparant les deux structures de classification des cohésions implicatives (les deux arbres cohésitifs) on peut remarquer des différences qui doivent amener à se poser des questions sur le fond de la problématique.

En effet, d'une part, dans le groupe ayant traité l'infographie on aurait donc significativement au premier niveau DES1=>not31 alors que pour l'autre groupe on aurait not31=>DES1 (au 3ème niveau de classification). Et d'autre part, on peut remarquer entre les deux sous-groupes les différences de positions des performances des étudiants lors de l'évaluation (reflétées par MOY31, MOY32, MOY33) par rapport à l'ensemble des autres variables principales dans les classifications des cohésions implicatives. Par ailleurs la structure globale de l'arbre produite par le fichier donnees-inf.csv (étudiant ayant traité par infographie) semble plus "cohérent" que celle produite par donnees-inf.csv (étudiant ayant traité les schémas).

En outre, notons que sur notre exemple de données les informations apportées par les arbres cohésitifs ne sont pas accessibles à partir des graphes implicatifs si on les avait demandés. Et aussi, si nous avions demandé les arbres des similarités, nous n'aurions pas vu les rôles dissymétriques que jouent les variables not31 et DES1 selon que l'on soit dans tel ou tel groupe.

9 Conclusion

En conclusion, afin de montrer les principales procédures d'utilisation du logiciel CHIC nous avons procédé dans cette partie, à deux types d'analyse que permet cette application : celle qui permet d'obtenir un graphe implicatif avec des variables supplémentaires et celle qui aboutit à des arbres cohésitifs. Dans les procédures, il y a la préparation d'une part des fichier-textes au format *.CSV sur lesquels les données sont organisées dans le format exigé par CHIC, après les avoir saisies et préparées à partir d'un tableur, d'autre part, les choix des options et des paramètres spécifiques aux calculs de graphe implicatif ou à ceux de l'arbre cohésitif. Pour des analyses plus approfondies, nous avons vu comment on obtient les contributions respectives des variables supplémentaires dans les calculs et même s'il n'y a pas de variables supplémentaires, CHIC nous a donné la possibilité d'obtenir les groupes optimaux qui ont contribué aux calculs d'une classification ou ceux qui en sont typiques. Pour une analyse plus fine, nous avons demandé pour chacun des deux groupes d'individus un arbre cohésitif sur les 5 variables principales. Cela nous a permis de voir des différences

de structures et les dissymétries qui devraient aider à poser les bonnes questions concernant la problématique à approfondir.

Références

Couturier, R., (2008). *CHIC : utilisation et fonctionnalités* Laboratoire d'Informatique de l'université de Franche-Comté de Belfort-Montbéliard, BP 527, 90016 Belfort, raphael.couturier@iut-bm.univ-fcomte.fr

Couturier, R, R. Gras et F. Guillet (2004). Reducing the number of variables using implicative analysis In International Federation of Classification Societies, IFCS 2004, Springer Verlag: Classification, Clustering, and Data Mining Applications, p. 277--285, ISBN 3-540-22014-3, Chicago, USA, July 2004.

Couturier, R. (2008). *CHIC, Cohesive Hierarchical Implicative Classification,* in R. Gras, E. Suzuki, F. Guillet and F. Spagnolo (Eds) *Statistical Implicative Analysis* Berlin : Springer-Verlag, p. 41-54

Ressources bibliographiques en ligne

http://www.ardm.asso.fr/chic_web/bibliographie_chic/bibliographie_chic_base.xml

Summary

In this article, within the framework of an implicative statistic analysis, we made the choice to use the datas of a current research, in order to present and describe the steps and procedures to be follow with the CHIC software. The aim is, on the one hand, to get an implication graph and an oriented hierarchical tree, and on the other hand, to make easier the reading of the results thanks to the various possibilities and CHIC functionalities. Step by step, we will deal with the description of collected datas, the transformation of these datas, the choice of the variables which will be supplementary variables, the preparation of the text file under .CSV format which will be compatible with CHIC, the processing to get an implicative graphic, particularly with the conjunction of premises and the threshold of originality concepts, the processing to get an oriented hierarchical tree completed by the significant levels, contribution of an optimal set of individuals and computation of typicality and contribution concepts, and we will finish with processing on interval variables.

PARTIE 3

APPLICATIONS DE L'ANALYSE STATISTIQUE IMPLICATIVE

Thème 1 : Applications à la didactique des mathématiques

Chapitre 1 : Analyse statistique implicative et didactique des mathématiques

Dominique Lahanier-Reuter

Équipe CIREL-Théodile ÉA 4354
UFR Sciences de l'éducation
Université Lille 3, Villeneuve d'Ascq
dominique.reuter@numericable.fr

Résumé. L'analyse statistique implicative a toujours été considérée comme une méthode d'analyse de données particulièrement heuristique et fructueuse pour la didactique des mathématiques. Nous soutenons ici que ces choix méthodologiques tiennent au fait que les liens implicatifs que révèle l'analyse statistique implicative peuvent être interprétés en termes de règles et de régulations d'actions. Nous étayons ceci par des exemples précis de recherches menées en didactique des mathématiques tout d'abord, puis dans d'autres didactiques, portant sur la reconstruction et la compréhension d'actions tant matérielles que discursives, actions régulées des élèves, actions régulées et régulatrices des enseignants.

1 Introduction

Penser les relations entre ces deux champs de recherche en termes de producteurs de modèles et de techniques d'une part et de terrain d'application d'autre part est sans doute trop réducteur d'un point de vue historique. En effet, l'histoire, même encore très brève, de l'émergence de ces deux domaines scientifiques nous montre des connexions plus complexes : la coïncidence des temps d'émergence et de reconnaissance, celle de leurs lieux géographiques et institutionnels – certaines universités et associations de chercheurs en France– enfin et surtout la présence d'acteurs communs – Régis Gras tout particulièrement– laissent supposer une dynamique spécifique à ces relations. Ainsi l'analyse implicative de données a-t-elle pu être identifiée comme une méthode d'analyse privilégiée en **didactique**

des mathématiques et réciproquement en quelque sorte, certaines problématiques en didactique des mathématiques ont pu marquer des questionnements en analyse implicative.

Nous cherchons dans un premier temps à exposer les raisons que nous voyons à cette coopération fructueuse, en montrant en quoi la mise à jour de règles (ou quasi-règles) que permet l'analyse implicative s'avère un outil précieux et pertinent pour les didacticiens des mathématiques. Dans un second temps, nous décrivons les modes possibles de cette coopération. Nous tentons de prendre en compte à la fois la diversité des comportements étudiés (discours oraux/discours écrits), des situations de recueil (observations/enregistrements/productions d'élèves/questionnaire). Tout d'abord en montrant comment l'ASI permet d'obtenir des résultats remarquables dans le cadre de l'une des principales problématiques en didactique dans lesquelles ces règles prennent sens : l'étude des régulations de comportements observables d'élèves et d'enseignants en situation. Ensuite, nous exposons l'un des intérêts de l'ASI, non plus en tant que méthode d'obtention de résultats mais en tant qu'instrument/outil méthodologique de contrôle de résultats didactiques.

2 Règles et régulations

L'analyse implicative de données permet d'exhiber des règles (ou quasi-règles) qui structurent un ensemble de données à partir de calculs sur les cooccurrences de certaines modalités de variables. Ces règles peuvent être génériquement représentées par une expression du type « si A alors B ». Elles sont par conséquent hiérarchiques, c'est-à-dire qu'elles opèrent une dissymétrie entre les modalités réglées. En cela, l'analyse implicative de données (dorénavant A.S.I) se distingue d'autres modes d'analyses statistiques qui, si elles se fondent également sur des calculs de cooccurrences de modalités de variables, n'exhibent généralement que des règles symétriques qui ne discriminent donc pas les variables étudiées. Etudier les relations entre A.S.I. et didactique des mathématiques interroge par conséquent les statuts théoriques que les didacticiens des mathématiques peuvent accorder à ces modèles de règles ainsi que la nature ou le statut des données à partir desquelles sont construites les modalités de variables soumises à l'A.S.I.

Si l'on peut définir les **didactiques disciplinaires** par un projet de connaissance scientifique des phénomènes liés à la transmission de savoirs disciplinaires, ceci a pour conséquence que le terrain d'observation privilégié des didacticiens est la classe : la classe au sens d'espace matériel certes (puisque l'on peut y explorer les affiches, les cahiers d'élèves…) mais aussi au sens d'espace symbolique (la classe de mathématiques par exemple existe encore lorsque l'enseignant prépare ses cours, lorsque l'élève apprend ses leçons, chez lui, à l'étude…). Cette classe existe donc dès lors que des interactions entre sujets les posent l'un comme enseignant, l'autre comme élève, relativement à un objet de savoir disciplinaire. Très schématiquement, l'un des principaux objets d'étude de la didactique des mathématiques est celui des manifestations de cette relation entre ces trois éléments inter-dépendants, le maître, l'élève et le savoir disciplinaire, de son établissement et de son maintien au fil du temps. Deux conséquences peuvent être tirées de cette modélisation. Tout d'abord les observables sont construits en tant qu'interactions entre maître et élève, maître et savoir, élève et savoir. Ensuite, cette étude requiert l'analyse des *régulations* qui vont simultanément être générées par ce système d'interactions et en retour assurer son fonctionnement. Ainsi par exemple, l'une des problématiques les plus

fructueuses en didactique des mathématiques est celle des régulations qui affectent les interactions de l'élève et du savoir en jeu : si les observables sont dans ce cas les actions engagées par l'élève (langagières ou non), les régulations qui régissent ces actions (l'engagement de procédures, certains choix effectués…) ou qui sont produites par ces actions (l'abandon de manières de faire, la mise en œuvre de nouveaux contrôles…) sont essentielles à mettre en évidence.

Certaines des règles mises en évidence ou révélées au travers de l'A.S.I. sont ainsi interprétables en didactique des mathématiques en termes de **régulations d'interactions**. Deux positions peuvent alors être adoptées : soit ces règles ont le statut d'hypothèses pour le didacticien, à charge pour lui de les invalider ou de les confirmer par d'autres méthodes d'analyse (comme par exemple l'analyse de contenu d'entretiens), soit elles ont le statut de faits d'expérience (permettant ainsi de contredire ou de ne pas infirmer l'analyse *a priori*).

L'asymétrie que présentent ces règles est également à prendre en compte dans l'interprétation que peut en faire cette didactique. Elle pose le problème des asymétries explicables par la modélisation en termes de système d'interactions. Le **système didactique** que nous avons sommairement évoqué (triplet des interactions entre élève, enseignant et savoir) est un système où les dissymétries des caractéristiques liées par une règle peut être expliqué de nombreuses façons différentes.

Commençons par le cas le plus classique, où les règles établies le sont à partir de données correspondant à des observables situés (les actions ou les déclarations effectives d'élèves ou d'enseignants recueillies en situation). A la dissymétrie entre les modalités des variables liées par une règle issue de l'A.S.I. « presque tous les sujets possédant la caractéristique A possèdent la caractéristique B » doit correspondre une dissymétrie entre les observables. Cette dissymétrie conduit à interroger de façon didactique le fait que très peu d'élèves ont fait B sans avoir fait A, ont réussi B sans avoir réussi A, ont répondu B sans avoir répondu A, que très peu d'enseignants ont fait B sans avoir fait A etc. Ces régulations des faire, des dire et de leurs effets peuvent être des effets de la temporalité, des différences entre les tâches proposées, d'organisations des connaissances… Ainsi, Jeanne Guiet (Guiet, 1994) à l'aide de l'ASI, a pu hiérarchiser des caractéristiques des divisions à effectuer au CM2 en étudiant les règles qui gèrent les réussites et les échecs d'élèves à un ensemble de ces opérations.

Un autre cas est à envisager, celui où les règles établies le sont à partir de données correspondant à des observables situés, mais aussi à partir de données pérennes à ces observables. La stabilité de ces dernières rend alors compte de groupes de sujets « fixes » (les élèves d'un même milieu socio-culturel, les enseignants « novices » *vs* « expérimentés » etc.). L'A.S.I. peut alors fournir soit des règles liant des actions, des dires, des effets de ces actions de ces dires et ces groupes constitués, soit, par l'étude des contributions de sujets à des règles, établir des tendances partagées par des sujets d'un même groupe, ou au contraire des évitements également caractéristiques.

3 Régulations d'actions situées, règles établies à partir de modalités d'observables

Nous commençons donc par illustrer l'intérêt de cette méthode d'analyse de données en présentant les cas les plus remarquables d'interprétation des règles ou quasi-règles que fait surgir l'ASI. Il s'agit donc d'interpréter cette organisation « statistique » de caractéristiques

au regard d'une autre organisation « didactiquement » pertinente. Nous avons retenu pour cette dernière, le temps, la structure des représentations, et enfin celle de compétences langagières disciplinaires.

3.1 Asymétries des règles établies et chronologie des tâches

L'exemple que nous proposons de développer est celui de l'étude des réponses d'élèves de CM1-CM2 (9 à 10 ans) à un exercice qui se décompose en deux tâches successives. Les régulations entre ces réponses fournies par l'ASI peuvent de façon fructueuse être associées à ce déroulement chronologique.

En effet, comme nous le détaillons ci-dessous, ces deux tâches sont articulées. Dans un premier temps, il est demandé à ces élèves d'ordonner des écritures décimales et fractionnaires : 0,5 ; 1,5 ; 2 ; 3,5 ; 4 ; 5,15 ; 5,6 ; 6,2 ; 9,5 ; 12. Dans un second temps de les placer sur une droite graduée. Puisque les mots nombres associés à 5,15 et 5,6 s'énoncent « cinq virgule quinze » et « cinq virgule six », une erreur fréquente à ce niveau scolaire consiste à placer 5,6 avant 5,15, en ne comparant que les parties décimales de ces nombres. Cependant, les nombres ont été choisis de telle sorte que la reproduction de cette erreur de classement dans la deuxième partie de la tâche conduise à une contradiction que les élèves – toujours des niveaux scolaires considérés – peuvent appréhender. En effet, placer sur la droite graduée tout d'abord correctement un point correspondant à 5,6, puis placer celui correspondant à 5,15 en décalant le premier d'un écart de « 9 » (l'écart entre15 et 6) conduit l'élève à placer 5,15 de façon erronée sur le point qui devrait correspondre à 6,5 (5,6+0,9). Ce placement peut sembler contradictoire avec celui correspondant à 6,2 puisque dans ce cas, le point « représentant 5,15 » se trouve après celui représentant 6,2. Nous dirons dans ce cas que les informations renvoyées à l'élève par le placement erroné de 5,15 sont des éléments potentiels du *milieu* avec lequel l'élève interagit.

Si nous nous attendons par conséquent à ce que certains élèves commettent des erreurs dans le classement des écritures numériques, nous nous interrogeons en revanche sur les effets des conséquences de ces erreurs, lors de l'exécution de la seconde tâche. Deux types de considérations usuels en didactique des mathématiques permettent de les anticiper. Tout d'abord, les informations que donne le placement erroné des points sur la droite ne sont pas « naturellement » interprétées en termes de contradictions. En effet, la lecture et la compréhension de ces informations nécessitent la mise en œuvre de certaines connaissances : il s'agit de considérer le placement de 5,15 et de 6,2 comme « étrange » et de le considérer en tant que *conséquence* de l'erreur sur le classement de 5,6 et 5,15. Ensuite, les différents travaux menés sur l'erreur ou les dysfonctionnements en situation scolaire nous mènent à différencier la reconnaissance par un élève d'une erreur commise et la prise en compte de cette erreur. Pour le dire de façon rapide, la perception d'une contradiction dans ses résultats est souvent insuffisante pour amener un élève, dans une classe, à infirmer ces derniers car il ne se sent pas toujours investi de la responsabilité de la résolution du problème soulevé (Brousseau 1990, Margolinas 1993). La question de l'étude des comportements des élèves est donc une question légitime.

L'étude du corpus des productions écrites des élèves laisse apparaître des stratégies diverses pour répondre aux deux questions de l'exercice. Pour classer les écritures numériques, certains ont recours à une stratégie de classement par « genres d'écriture », en classant tout d'abord les écritures fractionnaires, puis les écritures d'entiers qui ne comprennent pas de virgule, puis les écritures décimales en séparant celles qui ne comportent

qu'un seul chiffre après la virgule de celle qui en comporte deux. D'autres, comme nous pouvons le prévoir, classent ces écritures selon leur partie entière (visible ou calculée dans le cas de ½) puis selon leur partie décimale, considérée elle aussi en tant que nombre entier (5,15 est alors placé après 5,6). Signalons enfin que certains ont « négligé » les repères de la droite graduée et ont usé de la droite comme « ligne d'écriture » sans placer de points. Cette étude permet également de décider si, finalement, l'élève produit deux ordres différents et en conséquence contradictoires ou si au contraire il produit deux ordres cohérents même s'ils sont erronés.

L'A.S.I permet de mettre en évidence les règles suivantes[1] :

(1) « Adopter, définitivement, un classement d'écritures par genres d'écritures » implique, avec l'intensité d'implication .99, d'« accepter une non-adéquation des deux ordres produits »

(2) « Travailler sur la droite graduée en tant que ligne d'écriture » implique, avec l'intensité d'implication .95, d'obtenir « deux ordres cohérents même s'ils sont erronés »

(3) « Produire, finalement, un classement exact des écritures numériques » implique, avec l'intensité d'implication .95, d'obtenir « deux ordres cohérents ».

Ces trois règles sont interprétées en tant que régulations de comportements d'élèves face à ces deux tâches. Tout d'abord, il est possible de lire dans la règle (1) le fait que certains de ces élèves perçoivent – ou décident de percevoir – les deux tâches comme distinctes. Nous considérons que ces derniers ne conçoivent pas (dans la situation explorée) l'articulation entre l'ordre que « montre » la disposition linéaire des écritures numériques et celui que « montre » la disposition des points de la droite graduée. Ensuite, la règle (2) peut être interprétée, dans le contexte de la situation, comme un « évitement » de la deuxième tâche. L'élève recopie sur la droite graduée la liste précédente des écritures et s'interdit ainsi de prendre en compte les éventuelles difficultés qu'il rencontrerait à assurer une cohérence des deux ordres. Enfin, la règle (3), en établissant la dissymétrie entre les deux modalités[2], nous laisse supposer que le contrôle de la cohérence des deux ordres permet pour certains élèves, une rectification du classement des écritures.

Ainsi, l'organisation didactique de ces deux tâches, et tout particulièrement la conception d'un milieu susceptible par l'interprétation de ses rétro-actions de faire surgir une incohérence sur les résultats est insuffisante : il est nécessaire en effet, que l'élève admette de relier ces deux tâches pour qu'il accepte d'interpréter les résultats de l'une en fonction de l'autre.

3.2 Asymétrie des règles et organisation de représentations

L'asymétrie des règles peut aussi rendre compte d'autres structures. Nous avons choisi celles des représentations, problématique qui a déjà été illustrée par les travaux de M. Bailleul (Bailleul, 1995) sur celles des enseignants de mathématiques. Nous exposons ici les résultats d'une étude centrée sur les représentations disciplinaires d'élèves de lycée. Il s'agit pour nous d'analyser les façons dont ces élèves reconstruisent la discipline scolaire questionnée, au travers de leurs souvenirs, des usages qu'ils lui reconnaissent et des délimitations qu'ils effectuent. Les différentes manières de penser ces usages, ces

[1] Nous n'avons choisi ici que quelques-uns des résultats produits.

[2] Ce que ne pourrait assurer un test du χ^2, en particulier.

délimitations, sont autant d'indicateurs de ce qu'Yves Reuter nomme la « conscience disciplinaire » (Reuter, 2007).

Les mathématiques, en tant que discipline scolaire, sont familières aux élèves de Première S[3] et de Terminale S depuis longtemps. En revanche, l'analyse est encore une sous-discipline nouvelle[4] : cela signifie pour nous que la conscience disciplinaire de l'analyse est en construction. D'après un questionnaire passé auprès de (bons) élèves de quatre classes de ces niveaux, il ressort des résultats intéressants.

Nous avons posé un ensemble de questions qui ont été systématiquement déclinées en Analyse…, en Statistique…, en Mathématiques…

Ainsi, nous avons demandé

Á quel niveau scolaire ils avaient commencé à faire…

Quelles finalités ils voyaient à l'enseignement de…

Si leur enseignant leur avait déjà mentionné les finalités de …

S'ils identifiaient des caractéristiques d'un cours de…

D'un exercice de …

Des usages de… dans d'autres cours

Des contenus enseignés/appris

S'ils utilisaient un classeur, cahier particulier pour… ;

Nous avons isolé les variables concernant les réponses portant sur l'analyse, les mathématiques et la statistique : elles sont identifiées par l'initiale de la matière concernée (A pour analyse, M pour Mathématiques, S pour Statistique. L'analyse qui suit isole d'une part les variables concernant les mathématiques, d'autre part l'analyse[5]. Commençons par ces dernières. L'ASI permet d'obtenir le graphe suivant :

[3] Le « S » est pour scientifique.

[4] Nous la qualifions de nouvelle car c'est à partir de ces niveaux scolaires que le nom apparait officiellement (parfois en 2e). Cela ne préjuge en rien du fait que certains contenus des niveaux antérieurs puissent relever de ce domaine.

[5] Pour une analyse plus détaillée, voir Lahanier-Reuter, 2008.

FIG 1 *: Les relations entre les déclarations sur l'Analyse*

Presque toutes les caractéristiques rendant compte d'une construction de l'analyse en tant que sous-discipline scolaire sont liées et *hiérarchisées*. Les règles qui déterminent cet ordre peuvent être interprétées en tant que conditions à la construction de certaines compétences caractéristiques de la conscience disciplinaire. Ainsi, la possibilité d'identifier des apprentissages spécifiques serait, sur la population étudiée, une condition à celles de désigner les débuts de cet enseignement et de discerner les exercices relevant de celui-ci. Cette capacité, à son tour, serait une condition à la perception de l'usage de l'analyse dans d'autres disciplines scolaires et à l'identification d'un cours d'analyse. Enfin, cette capacité à distinguer les enseignements/contenus d'analyse des autres serait nécessaire à la décision de matérialiser[6] la délimitation de cette matière (l'organisation des cours dans un classeur séparé) ainsi qu'à l'identification d'une finalité cognitive à l'analyse. En résumé, nous avons ici une structure de l'élaboration de cette conscience disciplinaire sur la population considérée : les éléments constitutifs peuvent être classés selon leur contribution à la « profondeur » de celle-ci. Les délimitations de cette sous-discipline s'effectueraient par conséquent d'abord sur les contenus appris/à apprendre, puis sur les exercices, ensuite sur les éléments du cours et enfin par la mise en œuvre personnelle d'une organisation matérielle.

Certains éléments contribueraient à cette structure :

[6] Nous sommes bien consciente qu'il s'agit de déclarations sur la pratique, et non de pratiques observées.

- L'identification des exercices qui lui sont propres favorise la reconnaissance des usages « hors de la classe d'analyse » de cette discipline, et par conséquent la possibilité d'identifier d'autres disciplines scolaires « liées » : (ce qui est cohérent) ;

La possibilité d'identifier des contenus disciplinaires appris ou enseignés induit la capacité à désigner le début institutionnel de l'enseignement (ce qui est également cohérent)

La conscience disciplinaire manifestée par les discours d'un élève n'est pas la somme de différentes caractéristiques perçues, mais plutôt la délicate élaboration de liens entre ces différentes caractéristiques.

Ces résultats ne peuvent être mis en évidence qu'au travers d'une analyse statistique telle que l'ASI qui préserve la dissymétrie des variables. Cependant, cette puissance de l'analyse est, *a contrario*, un frein à l'interprétation lorsque elle ne fournit aucun lien. C'est le cas justement qui se produit lorsque nous examinons l'ensemble des variables qui décrivent les réponses au questionnaire lorsque celles-ci concernent les Mathématiques. L'ASI menée ne révèle aucun lien. Il est alors nécessaire de faire appel à d'autres techniques d'analyse statistique, en particulier des tests du chi², qui font apparaitre un antagonisme entre les finalités attribuées à l'enseignement des mathématiques[7] (plutôt cognitive *vs* plutôt sociale), une co-occurence des impossibilités d'identifier un cours et un exercice de mathématiques[8] et des indépendances entre les autres variables. Une analyse des correspondances multiples résume ces résultats (voir Fig.2) ainsi que l'ASI menée cette fois, sur les items et leurs contraires (Fig.3) où n'apparaissent que l'exclusivité des finalités accordées à l'enseignement et aux apprentissages mathématiques : elles sont soit cognitives (MFCognitive) soit sociales (MFSociales).

Ces différents traitements révèlent la différence de l'élaboration de la conscience disciplinaire sur la population étudiée en ce qui concerne l'analyse, sous-discipline « nouvelle », et les mathématiques, discipline éprouvée. Celle de l'analyse est une élaboration étagée qui laisse supposer un mouvement qui irait de l'absence d'identification des contenus appris/ à apprendre jusqu'à la décision de consacrer un classeur particulier à cette sous-discipline. Les élèves interrogés se distingueraient alors, dans cette structure ordonnée par leurs positions sur quelques items. En revanche l'élaboration de la conscience disciplinaire des mathématiques est une élaboration relativement figée : il semble que les choix, les compétences interrogés aient été déjà réalisés ou opérationnalisés. Ce sont cette fois des différences entre réponses d'élèves et non plus des classifications qui permettent de les distinguer.

[7] Chi²= 17,64, s. à 0.01.
[8] Chi²= 7,04, s. à 0.01

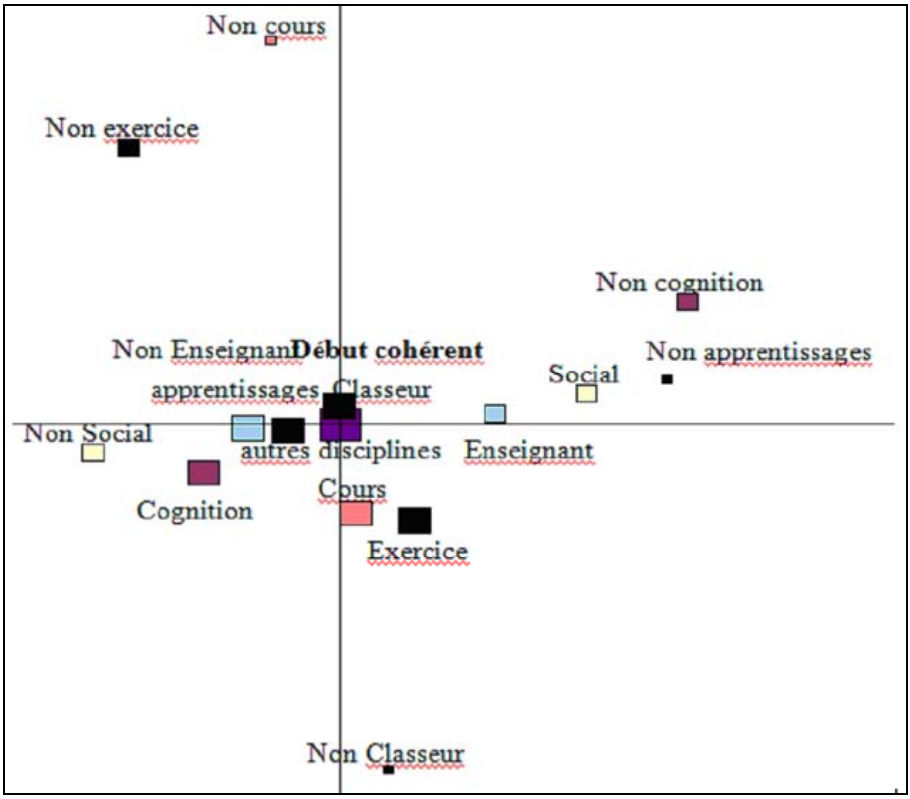


FIG. 2 *: Analyse portant sur les items et leurs contraires*

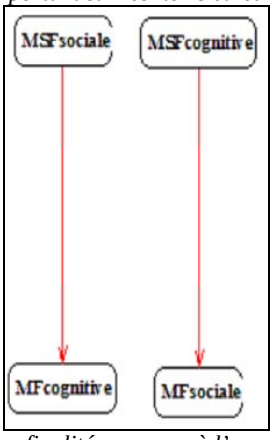


FIG. 3 *La répartition exclusive des finalités perçues à l'enseignement des mathématiques*

3.3 Asymétrie des règles et analyses de performances à l'écrit

Les régulations fournies par l'ASI peuvent enfin être utilisées pour décrire des performances d'élèves. Cette technique nous apparait comme extrêmement pertinente pour approcher des performances à l'écrit puisqu'elle nous permet, comme nous allons le voir, de retracer des chemins implicatifs reliant des caractéristiques de ces écrits : nous proposons d'interpréter un tel chemin comme la trace d'une *stratégie d'écriture* et les caractéristiques qu'il relie comme des paliers de compétences nécessaires à cette stratégie. Afin d'illustrer cette problématique, nous étudions ci-dessous des productions d'élèves de CM1 et CM2 « géométriques », qui répondent à la consigne suivante : « Comment faire pour tracer cette figure ? »

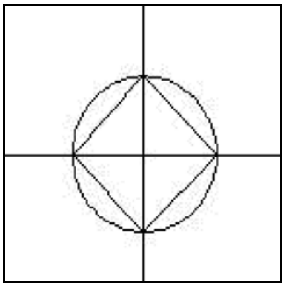


FIG. 4 *: « Comment faire pour tracer cette figure ? »*

Il s'agit par conséquent d'une tâche d'écriture d'un « programme d'instructions pour reproduire une figure complexe ». Elle requiert pour être menée à bien, d'identifier et de désigner des éléments constructibles par le lecteur, d'identifier et de désigner les relations constructibles qui existent entre ces divers éléments. Il n'est ainsi pas équivalent, lorsque l'on est élève de CM1 ou de CM2, de produire comme programme de construction : « trace deux droites perpendiculaires, trace un cercle dont le centre est le croisement des deux droites et relie les points d'intersection des deux droites et du cercle » ou « trace un carré, trace ses diagonales et son centre, trace le cercle dont le centre est celui du carré et qui passe par les sommets du carré » ou encore « trace un cercle et un carré à l'intérieur de ce cercle ». En effet le premier texte désigne des objets et des relations géométriques entre ces objets qui sont aisément constructibles par des élèves de ces niveaux scolaires. En revanche le second suppose le tracé d'un carré, ce qui l'est déjà moins. Enfin le dernier omet justement les indications des tracés nécessaires pour parvenir à inscrire un carré dans un cercle.

En conséquence, nous avons retenu comme premiers indicateurs des positions d'écriture adoptées par les élèves *les éléments choisis et leurs désignations, les relations géométriques mentionnées et leur désignation*.

Ainsi un indicateur discriminant se trouve être la mention par les élèves des points : O -le centre du carré qui est aussi intersection des droites et le centre du cercle- et des points A,B,C,D- sommets du carré ou intersections des droites et du cercle ou encore des points repérés sur les droites-. Le fait que cet indicateur soit discriminant n'est pas pour nous surprendre. En effet, l'analyse d'une figure complexe peut s'effectuer à plusieurs niveaux : soit en terme de configuration de traits et de lignes, soit en terme de configuration de points. D'autres que nous (Duval, 2003, Perrin-Glorian, 2004) ont mis en évidence le fait que les

élèves de primaire acquièrent tardivement la compétence d'analyse d'une figure en termes de configuration de points.

Différencier les productions de ces élèves peut également s'appuyer sur *l'image du scripteur* (Delcambre & Reuter) que ceux-ci tentent de construire. Certains ont une position « aidante » ou « coopérative » en fournissant des aides instrumentales, ou des conseils (« *tiens bien ta pointe*… »). D'autres au contraire, en reconstruisant un autre lecteur (sans doute l'enseignant) se montrent en train de faire (« *je prends mon compas* … ». Des indicateurs, tels ceux de la présence de conseils ou d'aides, tels celui du mode utilisé (impératif ou autre) tels celui enfin du sujet adopté (écriture en « je » ou non...) sont ainsi intéressants à mettre en œuvre.

Enfin, la conduite même du discours est à interroger. La tâche demandée requiert une *programmation des actions préconisées et en conséquence une planification du discours*. Nous avons retenu des indicateurs liés à ces opérations, tels la présence de « clôture » (« *et voilà ta figure est terminée* »), tels des marques de planification (la numérotation des actions à mener) et ceux qui sont pour nous des indices de dysfonctionnement de ces opérations, comme peuvent l'être des ajouts du type « *j'avais oublié de dire* » ou des contraintes qui surviennent tardivement, après que les actions aient été décrites.

Dans un premier temps, nous nous attacherons uniquement à l'interprétation des liens entre les items qui révèlent le mode d'analyse de la figure dans le discours des élèves.

Des chemins implicatifs séparés : l'exemple de compétences géométriques différentes

Le schéma (FIG. 5) est composé de deux graphes indépendants au seuil d'implication retenu ici. D'un côté apparait une relation entre le fait de situer topologiquement les droites « *il ya deux droites à l'intérieur de… »* et celui de ne pas relier le cercle aux autres traits et points de la figure. De l'autre côté on remarque des relations entre les critères indiquant la prise en compte des différents points de la figure et ceux indiquant de même la prise en compte des différentes lignes de cette même figure. L'analyse menée permet en conséquence de différencier nettement les productions qui restent dans le descriptif inefficace à la construction de la figure et celles qui s'appuient sur une analyse en termes de points et de lignes, nécessaire à cette construction.

Les « nœuds » des réseaux : des points cruciaux.

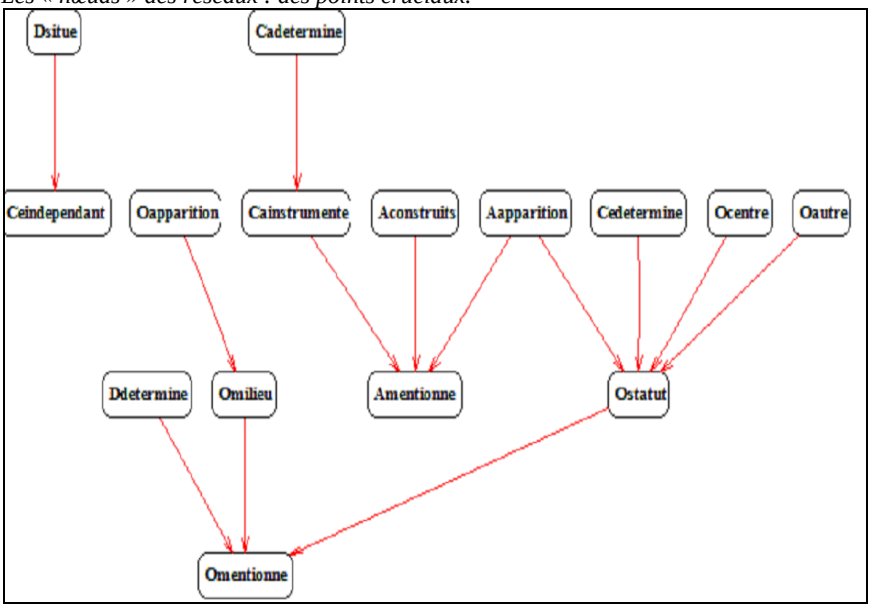


FIG 5 *: Décisions d'écriture et modes d'analyse de la figure*

Le statut des points est cependant quelque peu différent. Les points ABCD ne sont mentionnés *que si O lui même l'est tout d'abord*. Ensuite, on remarque l'importance du critère « O apparaît et change de statut » qui peut être considéré comme un nœud du réseau implicatif, à la différence du critère « O apparaît ». En effet, les élèves peuvent faire mention du centre du cercle ou du « milieu » des deux droites sans pour autant entrer dans une perspective de relier les deux tracés (ou de le dire). C'est donc bien le *changement de statut de O* (tour à tour centre du cercle et intersection des droites ou centre du cercle et intersection des diagonales) qui constitue le critère selon nous décisif pour le classement des productions des élèves et leur évaluation. Confier au lecteur ce changement et trouver les moyens discursifs pour le confier est sans doute une étape cruciale et décisive. Cela suppose en effet de trouver les moyens pour faire resurgir, rappeler un élément déjà présent dans le texte (donc un maniement ou la mise en œuvre de procédés anaphoriques), cela suppose aussi d'être capable de se détacher des évidences visuelles de la figure, bref de passer d'une appréhension visuelle immédiate à la mise en texte linéaire d'un changement invisible. Le point O ne bouge pas, mais il change de rôle.

L'A.S.I. permet donc de mettre en évidence le rôle déterminant de certains critères d'analyse et de réfléchir aux étapes décisives d'une performance d'élève.

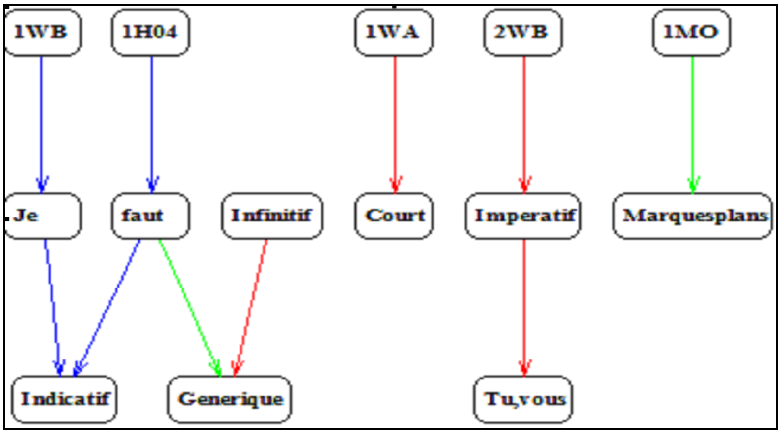


Fig 6 *: Des genres spécifiques aux classes*

Comme nous l'avons dit plus haut, quelques caractéristiques du texte écrit peuvent aussi être explorées. Les liens qui apparaissent sont pour la plupart attendus puisque les usages - même partiels- des modes impératifs et infinitifs sont bien reliés aux pronoms qui marquent le lecteur reconstruit : un « pair » pour l'impératif, marqué par la présence de pronoms « tu » et « vous », un « lecteur générique » pour l'infinitif, un lecteur « évaluatif », marqué par le recours à une écriture en « je » et un mode indicatif. Cependant, la présence de relations univoques entre les pronoms choisis et les modes utilisés permet de supposer que ces règles d'usage auxquelles obéissent les discours des élèves les définissent comme des actualisations de *genres* de discours. La rigidité de ces usages est assez extraordinaire pour qu'elle puisse être interprétée comme des contraintes fortes, issues de la situation scolaire de production écrite. Ainsi nous supposons que construire le lecteur comme un « pair » par exemple est caractéristique d'un genre d'écrits scolaires dans la classe de mathématiques qui peut être reconnu légitime par l'élève pour plusieurs raisons. Il peut se faire que les dispositifs pédagogiques et didactiques autorisent de telles positions parce que l'aide et la coopération sont des principes mis en œuvre dans les pratiques langagières disciplinaires. Il peut se faire également que les textes d'exercices des manuels de l'école primaire définissent un tel lecteur. De même, construire un lecteur « générique » est aussi une caractéristique de certains écrits disciplinaires. Cependant, cette caractéristique est moins fréquente à l'école dans les textes de construction géométrique des manuels scolaires primaires. En revanche, elle est fréquente dans ceux de « description de recette » qui occupent une place non négligeable dans les activités scolaires de ce niveau d'études, mais également dans les « programmes de construction » ou « manuel d'utilisation » rencontrés cette fois à l'extérieur de l'école. Enfin, les conduites discursives en « je » où l'élève montre au récepteur du texte ce qu'il sait faire ou réussit à faire est une caractéristique de situations d'évaluation. Ainsi, les résultats de l'A.S.I. présentés plus haut peuvent être interprétés en termes d'actualisation de genres scolaires.

Apparait alors la question récurrente en didactique des mathématiques de la raison de ces inscriptions : sont-elles les effets de l'enseignement ?

Notre projet est de montrer comment le recours à cette analyse statistique permet d'apporter des éléments de réponse à cette question.

4 Régulations relatives à des groupes de sujets

L'A.S.I. permet de distinguer les cas où les caractéristiques sont celles d'un sous ensemble du groupe d'élèves étudié de ceux où tous les élèves d'un groupe étudié partagent « presque » les mêmes caractéristiques. La problématique principale en didactique des mathématiques dans laquelle cette distinction prend sens est celle des relations entre enseignement et apprentissages, ou, pour la préciser encore davantage, celle de la mesure des effets d'un dispositif pédagogique particulier sur certains apprentissages en mathématiques. De nombreuses recherches antérieures peuvent être convoquées sur ce thème, dont deux synthèses également intéressantes ont été publiées récemment (Mercier et Buty, 2004 d'une part et Bru, Altet, Blanchard-Laville, 2004 d'autre part). Elles font apparaître comme légitime l'hypothèse selon laquelle des particularités de la gestion didactique des situations d'enseignement l'enseignant peuvent avoir une influence sur l'élaboration des savoirs mathématiques et sur l'appropriation de ,savoir-faire disciplinaires par les élèves concernés. Nous avons pu étayer cette hypothèse, et l'élargir à des attitudes, des comportements face à des tâches spécifiques (Lahanier-Reuter, 2005, 2007).

Nous reprendrons l'exemple précédent pour montrer l'intérêt de l'ASI à cet ensemble de questions. Dans le schéma précédent nous avions en effet fait apparaitre les classes en tant que « groupes d'élèves » : il s'agit dans ce cas de classes de CM1 (1WA, 1WB, 1H04, 1MO) et CM2 (2WA, 2WB, 2H03).

Nous interprétons ces relations comme autant d'inscriptions dans des genres spécifiques de « presque tous » les élèves de chacune des six classes (sur les sept) concernées. Il nous semble dans ce cas possible d'envisager le respect ces règles auxquelles se soumettent « presque » tous les élèves d'une même classe comme un effet de l'enseignement qu'ils partagent.

Après avoir évoqué cette problématique didactique et l'apport que peut représenter le recours à l'ASI dans ce cas, nous souhaitons terminer cette revue des échanges entre ASI et didactique des mathématiques par l'examen d'une dernière question, celle de l'analyse des pratiques langagières, des discours tenus par les acteurs principaux du système didactique, l'enseignant et les élèves.

5 Régulations des discours

Le champ de ces recherches étant extrêmement large, nous nous contentons, dans le cadre de cet ouvrage, de développer trois aspects complémentaires des usages de l'ASI. Le premier d'entre eux met en lumière une contribution de l'ASI aux travaux de didactique que nous avions laissés dans l'ombre jusqu'ici, celui d'apport de preuves. Dans l'exemple que nous développons, le thème de l'analyse de discours n'est tout d'abord qu'illustration de cette contribution. Mais il redevient prépondérant lorsque nous présentons des résultats qui sont interprétés dans une autre didactique que celle des mathématiques, celle de l'écrit. Enfin, il redevient le thème principal de la seconde étude que nous présentons ensuite, et qui a pour

projet de proposer une méthode d'analyse des interactions langagières dans des classes de mathématiques.

5.1 Comment l'asymétrie d'une relation devient élément de preuve : le cas du contrat de communication de réponses à un questionnaire

Nombre d'études en didactiques tendent à reconstruire des représentations d'élèves, ou encore à décrire les rapports d'élèves ou d'enseignants à certains objets enseignés, ou encore des compétences, des dysfonctionnements… bref à comprendre des pratiques ou des comportements, à inférer des réussites, des échecs et des performances. L'une des méthodes d'investigation en usage est celle du questionnaire. Cependant, cette méthode est sensiblement différente de celle qui a cours dans des sciences connexes, en sociologie par exemple : deux dimensions particularisent l'usage du questionnaire en didactique, celle quantitative tout d'abord du nombre de questionnaires étudiés, et celle plus « qualitative » de la construction de l'interviewé. En effet les questionnaires en didactique construisent leur interlocuteur en tant qu'élève, acteur d'un système didactique particulier (ou tentent de le faire), tandis que les questionnaires composés par des géographes par exemple constituent leur interviewé en tant qu'élément/acteur/sujet d'un espace géographique. Malgré cette différence voulue, des problèmes communs à cette méthode d'investigation traversent les champs. L'un d'entre eux est bien sûr la pertinence des discours recueillis et la position de scripteur de l'interviewé : les déclarations recueillies ne sont que des discours et sont soumises au contrat de communication qu'interviewé et interviewer négocient au long des questions. Le degré de crédibilité accordé à ces dires est par conséquent un problème récurrent, contre lequel certaines techniques (comme celle des questions répétées…) sont éprouvées. Dans le cas d'un traitement de ces réponses par l'analyse implicative de données, des règles intéressantes de ce contrat de communication peuvent être mises en évidence.

En voici un exemple.

Dans la recherche ANR EUIPM[9] à laquelle nous participons, c'est l'univers de l'écrit que rencontrent et élaborent des étudiants de sciences humaines[10] qui est interrogé. Un long questionnaire a été distribué à plus de 600 étudiants de ces diverses formations et années d'études. On comprend l'importance de ce contrat de communication et la vigilance toute particulière des exploitants de ce questionnaire, au vu de la longueur de ce dernier et du temps conséquent (près d'une heure) qui est nécessaire pour y répondre.

Nous avons demandé en particulier à ces étudiants de désigner parmi une liste de treize items ceux auxquels *ils faisaient attention* lors de l'écriture d'un genre représentatif de leur formation universitaire. Puis (les deux questions apparaissent sur la même page du questionnaire) nous leur avons demandé de désigner, par ordre d'importance, trois dimensions de l'écrit sur lesquels *ils estimaient être évalués* par leur enseignant.

Nous attendions de voir :

- Si les critères que les étudiants sélectionnaient parmi les treize proposés dépassaient les critères de surface (orthographe etc.) au profit des critères plus « universitaires »;
- Les critères qu'ils considéraient, à tort ou à raison, comme évaluatifs ;

[9] ANR-06-APPR-019 « Les écrits à l'université : Inventaire, pratiques, modèles », pilotée par Isabelle Delcambre.

[10] Etudiants d'histoire, de lettres modernes, de psychologie, de sciences de l'éducation et de sciences du langage, de tous les niveaux d'enseignement (L1 à M2).

- Les cohérences entre les deux réponses ;
- Enfin si nous pouvions attribuer les cohérences dans les déclarations successives à des règles du contrat de communication.

Pour répondre à ces questions, nous avons :
- Créé treize variables dichotomiques pour chacun des treize critères sélectionnés[11], variables qui se déclinent en « présence du critère *clarté du discours* comme critère d'attention » ou encore « Clarté discours ATT » etc.
- Créé treize variables dichotomiques en présence/absence pour la première dimension de l'évaluation, variables qui se déclinent en « *clarté du discours* choisie comme dimension principale de l'évaluation sur l'écrit choisi » ou encore « ClartédiscoursD1Ev »
- Fait de même pour les dimensions 2 (« CartédiscoursD2Ev ») et 3 (« ClartédiscoursD3Ev »)
- Créé trois variables dichotomiques en présence/absence d'autres critères que ceux utilisés précédemment et sur lesquels certains étudiants estiment être évalués (le respect de la longueur imposée de l'écrit par exemple).

Comme nous l'avons annoncé plus haut, notre première intention est de montrer comment l'asymétrie des relations que permet d'exhiber l'ASI peut devenir un élément de preuve. Pour cela, nous avons volontairement réduit l'ensemble des implications statistiques pour nous focaliser sur trois d'entre elles.

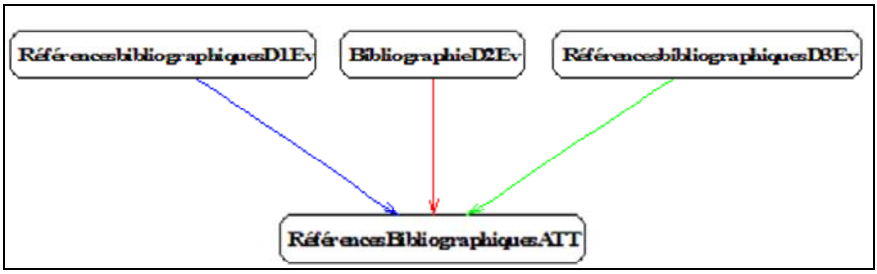


FIG 7 *: Quand l'asymétrie des relations contribue à prouver l'établissement d'un type de contrat de communication.*

Ces trois relations peuvent s'interpréter de façon similaire : « Citer les références bibliographiques comme un élément auquel on prête attention dans l'écrit universitaire est une condition pour citer ces références bibliographiques comme dimension de l'évaluation ». Cet ensemble de relations concerne également d'autres critères[12], à savoir : les articulations, les citations, la clarté du discours et la clarté formelle, les connaissances, la correction de la langue, la justesse de la réponse, l'originalité, le style, soit au total dix items sur les treize.

[11] Voici les treize items parmi lesquels les étudiants sélectionnent ceux auxquels ils déclarent porter une attention particulière, lors de leur écriture : Les articulations (entre les parties de votre texte ou entre les concepts, les idées présentées) ; L'avis personnel ; Les citations (citer des auteurs ou des extraits de textes) ; La clarté du discours (introductions ou conclusions partielles, exemples) ; La clarté formelle (graphie, mise en page, paragraphes, etc.) ; Les connaissances; La correction de la langue (syntaxe, lexique, orthographe, etc.) ; La discussion des auteurs au programme ; La justesse de la réponse ; L'originalité ; Les références bibliographiques ; La reformulation des textes lus ; Le style.

[12] Nous ne reproduisons pas les graphiques qui sont similaires à celui-ci-dessus.

Seuls les items citations, discussion des auteurs, reformulation des textes lus ne sont pas concernés par ce type de relations.

Par conséquent, il est légitime de supposer que les étudiants qui sélectionnent l'un de ces dix items comme critère auquel ils portent attention, répondent à la question sur les dimensions évaluatives en utilisant les critères que le questionnaire met à leur disposition. Ainsi, ces étudiants perçoivent comme règle locale du contrat de communication l'intention de rendre compte de la cohérence entre leurs attentions et leurs représentations de l'évaluation par les enseignants. Cette règle locale, qui gère la *chronologie* des actions et des choix, ne peut être établie que grâce à la dissymétrie des liens implicatifs.

Conjointement, nous testons si la décision d'indiquer d'autres critères pour l'évaluation que pour l'attention portée est également une règle locale à laquelle se soumettent d'autres étudiants.

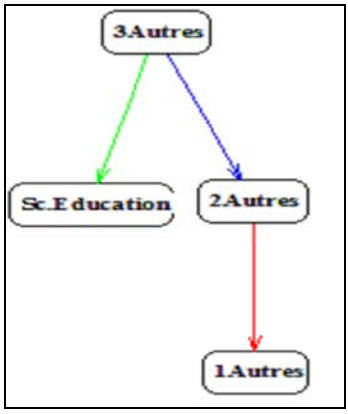


FIG. 8 *: Traces de trois représentations distinctes*

D'après le graphique ci-dessus, il est manifeste que la décision de fournir d'autres critères pour l'évaluation que ceux proposés pour rendre compte de l'attention portée à l'écrit, est immédiate : en effet, c'est lors du choix de la première dimension (1Autres) de l'évaluation que cette décision apparaît et entraîne en conséquence la même stratégie de réponse aux deux dimensions suivantes (2Autres et 3Autres). On remarque enfin que cette décision est plus ou moins caractéristique d'un groupe particulier d'étudiants, ceux qui suivent des études en Sciences de l'éducation.

5.2 Un usage de l'ASI dans une autre didactique que celle des mathématiques.

Notre projet est de montrer, sur un exemple trop partiel sans doute, que les résultats apportés par l'ASI sont encore interprétables dans une autre didactique, celle de l'écrit. Reprenons l'étude précédente et les réponses à la question des attentions portées sur les écrits universitaires (Fig. 9).

Les trois chemins implicatifs distincts apparaissent : cette séparation est interprétée comme la marque de trois systèmes de représentations différents, qui gouvernent la cohérence des réponses à cette question. Le premier est celui de l'écrit universitaire

comme monstration des connaissances, le second celui de l'écrit universitaire comme lecture critique des auteurs, le troisième enfin celui de l'écrit universitaire comme expression de soi. Ces résultats satisfont aux objectifs de la recherche ANR poursuivie puisqu'ils permettent de distinguer ces représentations, pour ensuite explorer l'influence de la discipline universitaire et du niveau d'étude sur celles-ci.

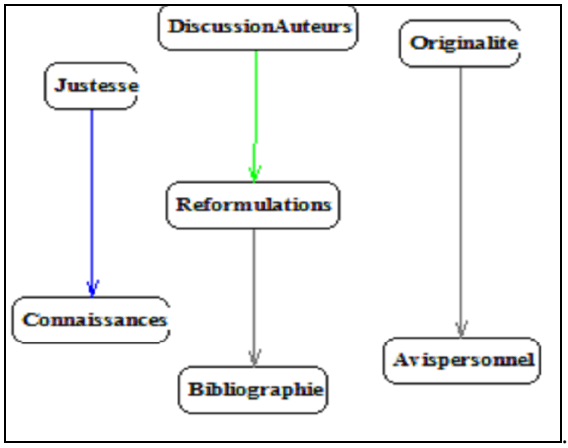


FIG. 9 *Traces de trois représentations distinctes*

5.3 Une mise en œuvre de l'analyse implicative de données sur des retranscriptions orales

En guise de dernière proposition pour rendre compte de l'apport de l'ASI aux problématiques didactiques, nous présentons dans ce paragraphe l'intérêt que revêt cette analyse statistique à l'étude des pratiques langagières dans des classes. En effet les différentes didactiques s'intéressent à ce qui se dit et les manières dont les choses sont dites dans les espaces disciplinaires. La place si importante de l'oral à l'école le constitue en un type de corpus « obligé » pour qui cherche à comprendre, décrire ou expliquer ce qui s'y passe. Certes, les façons de constituer en corpus de données des enregistrements, de les interroger et de les traiter enfin sont étroitement liées aux questions spécifiques qui les dirigent. Il est ainsi légitime, dans le champ des didactiques, de chercher à savoir si l'oral prend des formes particulières selon la discipline enseignée. L'exploration de cette question amène aussi à poser celle des raisons de ces particularités disciplinaires: peut-on expliquer leur permanence par leurs effets supposés, attendus ou effectifs sur les apprentissages ? Ou par la « nature » des contenus enseignés ? Un autre ensemble de questions est également au centre des préoccupations actuelles des didacticiens : comment décrire l'activité (ou les activités) des élèves et de l'enseignant durant ces échanges ou ces prises de parole ? Pour le dire autrement, que font-ils lorsqu'ils parlent ?

Nous n'avons listé ces questions que pour suggérer la diversité des modes de catégorisation, de description, d'isolements, bref de traitements de ces données. De façon encore une fois très générale, les analyses menées en didactiques nécessitent un cadre

théorique puissant et des systématisations dont un « étranger » ne peut guère s'emparer : en quelque sorte, les modes de traitements des données sont extrêmement lourds et ne peuvent guère s'exporter. Toujours dans cette perspective de remise en cause ou plutôt de pas de côté par rapport aux positions orthodoxes, nous dirons que ces analyses[13] reposent sur un principe de « découpages » des situations.

C'est pourquoi il nous a paru intéressant d'explorer les résultats que permet l'analyse implicative de données, en ne nous embarrassant que très peu au départ de considérations théoriques, et surtout en considérant ces traitements *comme une alternative possible à la recherche de découpages des séances observées*. C'est bien sûr une présentation plutôt non orthodoxe que la nôtre, qui ne peut être jugée qu'à l'efficacité qu'elle offre, à la pertinence des résultats et bien sûr des interprétations en retour.

Nous avons isolé les tours de paroles d'un enseignant 1, dans une situation fermée au cours d'un exercice disciplinaire, ici un exercice de calcul mental autour de la règle des zéros[14]. Chacun de ces énoncés présente des caractéristiques diverses, selon par exemple que l'enseignant1 programme le travail des élèves, le contrôle ou apporte des éléments de savoir, selon qu'il rebondit sur le discours d'un élève ou au contraire l'arrête[15] etc. Ce qui nous intéresse dans ce cas est le traitement de ces caractéristiques :

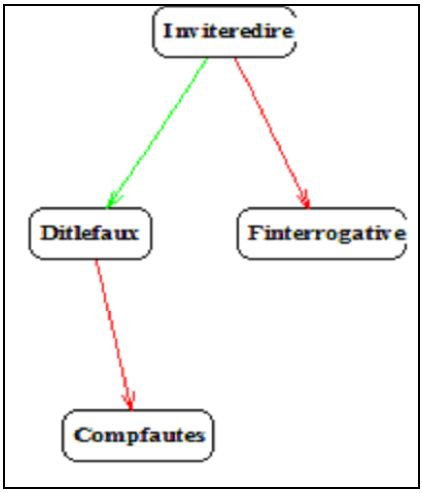


FIG. 10 *: Graphe implicatif des caractéristiques d'énoncés de l'enseignant1 de CE2 en situation de calcul mental*

Il vient que les énoncés où l'enseignant « invite à redire ou à reformuler » font partie de ceux où l'enseignant « dit ce qui est faux », « compte les fautes », et enfin s'insèrent dans ceux de « forme interrogative ».

[13] Sensévy et Mercier 2007, Marlot 2007, Chappet-Pariès, Robert et Rogalski 2008.

[14] La « règle des zéros » est celle qui permet de calculer un produit du type 45x1000 en ajoutant trois zéros à 45.

[15] Pour une description plus complète de ces caractéristiques et de leur élaboration, voir D. Lahanier-Reuter, 2007.

A contrario cela peut être dit sous la forme suivante : l'enseignant1 n'invite l'élève à reformuler ses dires uniquement si cet élève a dit quelque chose d'erroné et ne le fait qu'après avoir décompté « les fautes ».

Comparons avec une autre séance, menée dans la même classe par le même enseignant 1, qui porte cette fois sur la compréhension d'un énoncé de problème arithmétique (Fig. 11).

Cette fois, les règles auxquelles se plie son discours sont les suivantes (nous faisons abstraction des règles générales, telles que la forme interrogative des questions) :

L'enseignant 1 dans ce cas ne reformule que s'il approuve explicitement les propositions ou les résultats des élèves et toute reformulation, toute approbation s'accompagne inéluctablement d'une question factuelle. C'est un enchaînement très classique de questions, réponses, reformulations etc. Les énoncés erronés sont rejetés sans plus. Les aides apportées sont des indices de la « bonne réponse » et non pas des encouragements par exemple… L'énoncé du programme d'actions (ce que l'on va faire) est imposé.

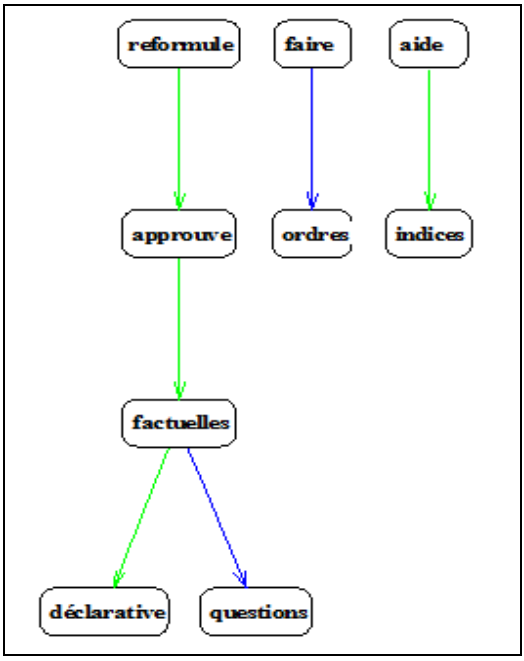


FIG. 11 *: Résolution d'un problème d'arithmétique*

Comparons cette fois avec un autre enseignant 2 et une autre séance de mathématiques consacrée elle aussi à la résolution d'un problème arithmétique.

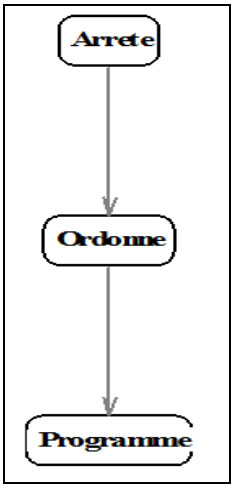


Fig. 12 *: Un autre mode d'intervention sur un contenu « identique »*

Nous voyons un autre mode d'intervention : ici l'enseignant 2 arrête (les échanges, le travail des élèves) plus tôt s'il a à programmer des actions.

Nous terminons cette très rapide exploration par l'analyse cette fois d'une séance consacrée à la règle des zéros (comme l'était la première de l'enseignant 1) mais avec une autre classe et un autre enseignant 3 :

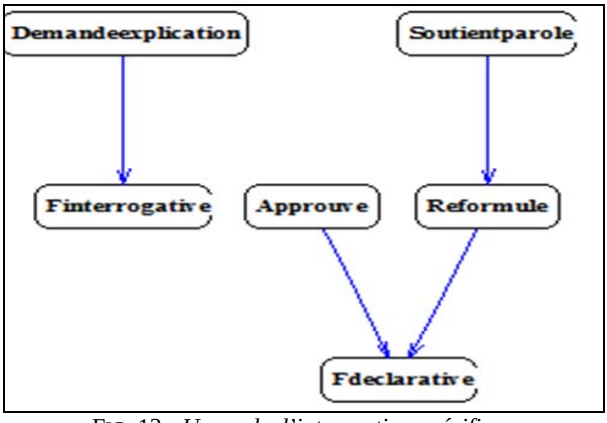


FIG. 13 *: Un mode d'intervention spécifique*

D'après ce dernier graphique, l'enseignant 3 concerné ne reformule que pour soutenir la parole de l'élève.

Ainsi cette analyse menée auprès de trois classes différentes, d'enseignants différents, dans des situations « ordinaires » de l'enseignement des mathématiques, révèle des divergences importantes quant aux pratiques langagières de l'enseignant.

Il est bien sûr possible de produire des résultats en étudiant les énoncés des élèves durant une séquence. Nous avons choisi, pour sa représentativité, la séance 2, qui tourne autour de la résolution en CE2 d'un problème d'arithmétique, menée par l'enseignant 1 (Fig.13).

Certains de ces résultats sont prévisibles : lorsque les élèves « demandent une méthode », leurs énoncés ont plutôt une forme interrogative, lorsqu'ils « fournissent des réponses » c'est plutôt sur le mode déclaratif.

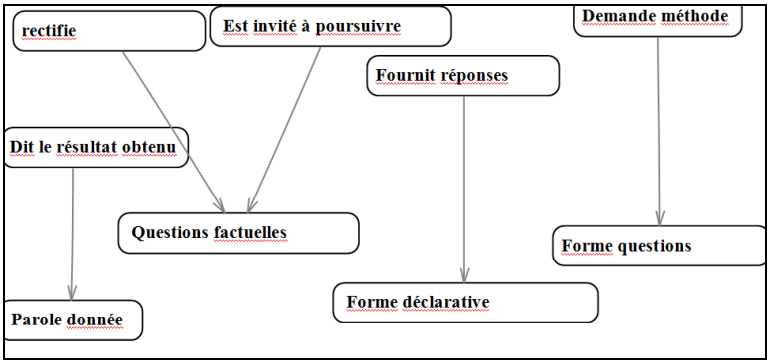


FIG. 14 *: Interventions des élèves dans une séance de résolution de problème arithmétique*

Ce qui est plus intéressant ici, est l'apparition de règles scolaires[16] : un élève ne « dit le résultat obtenu » que si on (l'enseignant) lui donne la parole explicitement. Il ne rectifie (son résultat/son énoncé) que lorsqu'il réagit aux questions factuelles posées par l'enseignant. Et il n'est invité à poursuivre que si l'enseignant lui pose une question factuelle (et non pas de méthode).

Nous voyons en conséquence apparaître ici des régulations des actions langagières dans cette séance qui semblent varier selon les enseignants et les tâches disciplinaires. Cette méthode d'analyse paraît en conséquent intéressante à poursuivre, en l'étendant de façon systématique aux séances disciplinaires menées par cet enseignant, puis aux disciplines enseignées et enfin à d'autres enseignants.

6 En guise de conclusion

Les études dont nous venons de rendre compte permettent de révéler des cas où l'analyse implicative de données nous amène à des résultats différents de ceux tirée des analyses « plus classiques ». Il va de soi que nous n'avons pu explorer toutes les interprétations possibles en didactique des mathématiques des résultats issus de l'ASI, mais nous espérons avoir pu montrer à quel point ces traitements sont fructueux. On peut retenir dès à présent l'apport de ces méthodes à l'étude des performances d'élèves, à la classification d'exercices disciplinaires, à la reconstruction de représentations. Il nous semble que les domaines d'études des interactions langagières, des régulations des discours, des stratégies d'écriture

[16] Nous ne nous hasarderons pas à les dire disciplinaires.

sont autant de domaines dans lesquels la contribution de l'ASI devrait se révéler d'un très grand intérêt.

Références

Bailleul M. (1995). Une approche statistique des représentations de l'enseignement des mathématiques chez des enseignants de mathématiques de collège et de lycée. *Recherches en didactique des mathématiques*, Vol. 15/2: 9-30.

Brousseau G. (1990). Le contrat didactique et le concept de milieu: Dévolution. *Recherches en didactique des Mathématiques*, Vol 9/3: 309-336.

Bru M., Altet M., Blanchard-Laville C. (2004). La recherche des processus caractéristiques des pratiques enseignantes dans leurs rapports aux apprentissages. Revue française de pédagogie, vol. 148: 75-87.

Chappet-Pariès M., Robert A., Rogalski J. (2008). Que font des élèves de troisième et de quatrième avec un même enseignant dans une séance de géométrie?. in Vandebrouck *coord. La classe de mathématiques: activités des élèves et pratiques des enseignants*. Toulouse: Octares Éditions : 95-138.

Delcambre I., Reuter Y. (2000). Rapports à l'écriture et images du scripteur. *Les Cahiers Théodile*, 1 : 1-18

Duval R. (2003). Décrire, visualiser ou raisonner: Quels apprentissages premiers de l'activité mathématique ?. *Annales de didactique et de sciences cognitives*, 8: 13-62.

Guiet J. (1994). *La division: une longue souffrance*. Thèse de doctorat en Sciences de l'Education, Université René Descartes Paris V.

Lahanier-Reuter D. (2005). Enseignement et apprentissages mathématiques dans une école Freinet. *Revue Française de Pédagogie*, 153: 55-65.

Lahanier-Reuter D. (2007). Enseignement et apprentissages mathématiques. in Reuter Y. (dir). *Une école Freinet, Fonctionnements et effets d'une pédagogie alternative en milieu populaire*. Paris. L'Harmattan, 185-216.

Lahanier-Reuter D. (2008). Didactics of mathematics and implicative statistical analysis. in Gras R., Suzuki E., Guillet F., Spagnolo F. (eds.). Statistical Implicative Analysis : Theory and Applications. Series: Studies in *Computational Intelligence*, Vol. 127. Berlin – Heidleberg. Springer, 277–298.

Margolinas C. (1993). *De l'importance du vrai et du faux dans la classe de mathématiques*. Grenoble: La Pensée Sauvage.

Marlot C. (2007). Analyse de l'action du professeur en classe ordinaire : formes méthodologiques de réduction du corpus et gestion de la disparité des unités de découpage de l'action. in Lahanier-Reuter D., Roditi E. (*éds.*) *Questions de temporalités, Les méthodes de recherche en didactiques (2)*. Villeneuve d'Ascq: Presses Universitaires du Septentrion:153-172.

Mercier A., Buty C. (2004). Evaluer et comprendre les effets de l'enseignement sur les apprentissages des élèves: problématiques et méthodes en didactique des mathématiques et des sciences. *Revue Française de Pédagogie*, 148: 47-59.

Perrin Glorian M-J. (2004). Eclairages et questions pour la didactique des mathématiques : cadres et registres en jeu dans la résolution de problèmes en lien avec les connaissances des élèves et recherches sur l'action des enseignants en classe. *Annales de didactique et de sciences cognitives*, 9: 67-82.

Reuter Y. (2007). La conscience disciplinaire: présentation d'un concept. *Éducation & Didactique*, 1/2 : 57-72.

Sensévy G., Mercier A. (2007) *Agir ensemble, Éléments de théorisation de l'action conjointe du professeur et des élèves.* Rennes: Presses Universitaires de Rennes.

Summary

Statistical Implicative Analysis provides a successful method for Didactics of mathematics. This is due to the possibility for the implicative links to be interpreted as rules and regulations. Describing some researches in the fields of didactics make this clear, especially when focused upon construction and interpretation of actions. These might be either regulated actions of pupils or regulated and regulating actions of teachers.

Chapitre 2 : Approche bayésienne “cachée” et approche fréquentiste “ambiguë” dans les manuels français de classes Première S et ES

Pablo Carranza*, Alain Kuzniak**

*Laboratoire André Revuz- Université Paris-Diderot
pfcarranza@gmail.com
**Laboratoire André Revuz- Université Paris-Diderot
alain.kuzniak@orleans-tours.iufm.fr
http://people.math.jussieu.fr/ kuzniak/

Résumé. A partir d'une analyse de manuels présentant un premier enseignement des probabilités en relation avec la statistique, nous esquissons l'espace de travail potentiel existant sur ce thème en classe de Première en France. L'approche statistique implicative nous permet de dégager certaines règles de fonctionnement de cet espace et de décrire une tendance soit très fortement formelle et calculatoire soit assez confuse où joue alors ce que nous appelons l'intrication des signifiés

1 Introduction

A l'issue d'une étude épistémologique, nous avons pu identifier une forme duale de la signification de la probabilité qu'expriment d'une part l'approche fréquentiste et d'autre part l'interprétation bayésienne. Cette dualité de signifié peut, en se manifestant dans un même problème, induire des erreurs d'interprétation dans la démarche inférentielle.

Les auteurs des programmes français actuels de lycée ont choisi de ne retenir que l'approche fréquentiste pour présenter la notion de probabilité sans aucune prise en compte de la notion bayésienne. Cependant, nous avons pu vérifier (Carranza et Kuzniak, 2006) que la dualité des signifiés se manifestait de manière implicite et cachée dans l'enseignement. Pour mettre en évidence ce phénomène, nous avons étudié certains exercices proposés dans les manuels et aussi dans les sujets d'examen. Nous souhaitons revenir d'une manière plus systématique sur cette question en étudiant de manière exhaustive les exercices posés dans les manuels de Première ES et S (élèves de 16–17 ans).

Après avoir rapidement présenté ce que sont pour nous les deux interprétations de la probabilité, nous décrirons les variables que nous avons retenues pour décrire le contenu des exercices de façon à analyser la présence d'un éventuel signifié de la probabilité et son mode de traitement proposé par les manuels. Plus précisément, nous cherchons à savoir si les exercices proposés dans les manuels restent dans un contexte fréquentiste ou si, comme nous le pensons, ils s'en échappent et dans ce cas quel type de signifié apparaît de manière cachée ou non.

2 La dualité de signifié

Le concept de probabilité admet plusieurs signifiés et ceci, comme l’a notamment montré Hacking (2002), depuis son émergence à la fin du 17^e siècle. Ces diverses interprétations peuvent être réunies dans deux grands groupes se rattachant à un des deux courants inférentiels, l’un fréquentiste et l’autre bayésien que nous avons retenus pour notre travail

Notion fréquentiste

Pour un certain type d’expériences susceptibles de se reproduire sous les mêmes conditions (au moins mentalement), la fréquence d’apparition d’un événement donné se stabilise progressivement lorsque le nombre de réalisations croît considérablement. Dans ce cas, la probabilité est la tendance du système à produire un événement donné. Cette première notion trouve sa justification dans la loi faible des grands nombres.

Notion Bayésienne

Dans cette autre approche, la probabilité représente une mesure de crédibilité sur une proposition donnée compte tenu des informations disponibles. Ainsi la probabilité n’est pas une caractéristique de l’objet mais la mesure de la crédibilité accordée par un sujet à une proposition donnée. Cette fois la justification de cette probabilité s’appuie sur la notion de probabilité conditionnelle et sur le théorème de Bayes : les probabilités attribuées initialement à un événement peuvent différer d’un observateur à l’autre mais elles évolueront ensuite de manière convergente en intégrant les nouvelles informations selon les critères donnés par la formule de Bayes.

L’intrication des signifiés

Les deux interprétations de la probabilité sont souvent étroitement reliées et enchevêtrées dans de nombreuses situations où elles apparaissent simultanément. Pour désigner ce phénomène, nous parlerons d’intrication des signifiés.

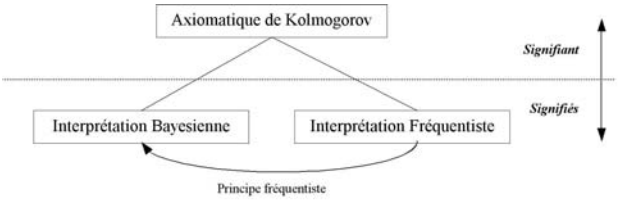


FIG. 1 – *Le principe fréquentiste*

La principale source de cette intrication s’appuie sur le principe fréquentiste (Fig. 1) où l’attente à long terme en tant que seule source d’information du système est considérée comme la raison fondant le degré de certitude (Droesbeke, Fine et al. 2002), page 25 ; (Gärdenfors, Sahlin et al. 1988), page 110 ; (Hacking and Dufour 2004), page 149). Ainsi, lorsqu’on lance

un dé non pipé, la fréquence d'apparition du trois est l/6 et on pense alors avec une certitude de 1/6 que le résultat du lancer à venir sera un trois.

3 La dimension cachée

Nous posons comme hypothèse que la dualité de signifié est une caractéristique inhérente à la probabilité et qu'elle se manifeste nécessairement dans l'enseignement même si officiellement il l'ignore comme c'est le cas dans les programmes actuels de l'enseignement français de lycée où la probabilité n'est présentée que comme fréquentiste.

Dans ce texte, nous nous interrogeons plus particulièrement sur les éventuelles manifestations de cette dualité dans les exercices proposés dans les manuels de Première ES et S. Nous avons choisi ces niveaux de programme car la probabilité y est introduite en relation avec la statistique ce qui nous permet d'observer le jeu entre les trois pôles que nous avons articulés dans la figure 1. Dans les programmes publiés en 2000, il est explicitement demandé d'éclairer le lien entre *lois de probabilité et distributions de fréquence* grâce à *un énoncé vulgarisé de la loi des grands nombres*. Pour cela, l'accent est mis sur la modélisation d'expériences aléatoires répétitives et sur la simulation d'expériences. Ces programmes marquent une rupture avec l'approche axiomatique classique et s'inscrivent dans une initiation à la statistique. Ils conseillent d'éviter les calculs systématiques et sans but d'espérance ou de variance.

La classe de ES est une section à dominante économique et sociale, les élèves ne suivront pas pour la plupart des études scientifiques par la suite à la différence de ceux qui suivent la classe de S qui est à dominante scientifique. Il est donc intéressant de voir si l'orientation différente des élèves va faire prendre en compte de manière différente la question de la dualité des signifiés et sa relation avec le modèle ensembliste de Kolmogorov.

Notre étude de manuels vise à décrire des espaces de travail potentiels (Kuzniak 2004) en dégageant certains principes de fonctionnement de ces espaces. L'analyse statistique implicative (Gras, Ag Almouloud et al. 1996 ; Gras et Bailleul 2000) nous permet de valider certaines régles d'inférence partielle du type "si a alors presque b" dans les exercices proposés dans les manuels. Nous pourrons ainsi examiner des règles telles que "si un problème réunit des conditions pour une interprétation bayésienne alors une interprétation n'est pas demandée" ou alors "si une interprétation est demandée alors il s'agit d'un problème fréquentiste".

Bien sûr, ces variables a priori du problème ne suffisent pas à déterminer une interprétation donnée. La gestion en classe est aussi essentielle et dans notre travail de recherche, nous avons aussi incorporé cette dimension qui ne figurera pas ici où nous avons choisi de nous concentrer sur l'espace de travail que proposent les manuels aux élèves et aux professeurs. Cependant les tendances que nous espérons découvrir avec l'analyse implicative doivent nous renseigner sur l'espace de travail potentiel disponible pour certaines interprétations de la probabilité à partir des exercices des manuels. Il s'agit en tout cas d'analyser la compatibilité ou la cohérence entre la notion de probabilité à institutionnaliser et l'ensemble de situations lui donnant le sens visé. L'intérêt ne se limite pas à ce seul cadre et notre étude s'intègre dans toutes celles qui portent sur les difficultés constatées sur le plan interprétatif d'une démarche statistique dont la source est la dualité de signifié (Gärdenfors, Sahlin et al. 1988 ; Régnier et Oriol 2001 ; Batanero, Godino et al. 2004)

4 Présentation des variables retenues

Notre étude s’appuie essentiellement sur deux manuels, l’un de première ES et l’autre de première S, ce qui est peu par rapport à l’ensemble des collections disponibles, mais notre propos n’est pas de comparer ces manuels mais bien de parvenir à identifier le rôle possible des différentes interprétations. En complément des deux manuels retenus, nous avons aussi étudié deux autres manuels pour nous servir de garde-fous contre les généralisations trop hâtives (Carranza 2009).

Les deux manuels observés ont été édités en 2001 : le manuel de première ES (287 pages) est édité par Bréal (Breal ES par la suite) et celui de Première S (432 pages) par Nathan (Nathan S).

Le manuel Breal ES propose pour le chapitre probabilité, trente et un exercices (page 98 à 103) répartis en trois sections. “Maîtriser le cours” : Exercices 1 à 12. “Savoir-faire fondamentaux” : Exercices 13 à 19. “S’entraîner” : Exercices 20 à 31.

Pour le manuel Nathan, le nombre d’exercices est beaucoup plus important et il y en a soixante-treize (page 224 à 234) répartis en quatre sections : “Maîtriser le cours” : Exercice 1 à 20. “Pour apprendre à chercher” : Exercices 21 à 25. “Pour progresser” : Exercice 26 à 67. “Problèmes de synthèse” : Exercice 68 à 73.

Variables étudiées

Nous avons retenu pour notre étude un certain nombre de variables, les unes d’ordre pédagogique prennent en compte des points classiques dans la gestion des exercices comme son rôle dans les apprentissages, sa longueur, sa place dans le chapitre. Mais nous avons aussi considéré un certain nombre de variables qui sont plus étroitement liées à l’articulation entre signifiant et signifié et aussi de manière encore plus spécifique à l’interprétation en termes fréquentiste ou bayésien.

Nous avons considéré 21 variables, toutes prenant des valeurs binaires (0 ou 1). Nous allons expliciter les critères à satisfaire pour que les variables prennent la valeur 1.

Variables concernant la longueur du texte de l’exercice A :

Après plusieurs regroupements nous avons retenu deux classes à effectifs supérieurs à 10 :

Codage	Etiquette	Description
A1	Lignes(1-10)	L’exercice a 10 lignes au plus
A2	Lignes(11-40)	L’exercice a plus de 10 lignes

TAB. 1 – *Longueur de l’exercice*

Variables concernant le nombre de questions B :

En relation avec la variable précédente, le nombre des questions posées à l’élève est pris en compte par cette variable.

Codage	Etiquette	Description
B1	Questions(1-2)	L'exercice a moins de trois questions
B2	Questions(3-4)	L'exercice a trois ou quatre questions
B3	Questions(5 et plus)	L'exercice a plus de quatre questions

TAB. 2 – *Nombre de questions*

Avec ces deux variables, nous souhaitons voir la relation entre la notion la probabilité en jeu et la taille des exercices. Dans l'ensemble des manuels que nous avons observés, il faut noter que plus des deux-tiers des exercices ont moins de 10 lignes et par contre plus des deux-tiers ont plus de trois questions.

Variables concernant le contexte de l'exercice E :

Cette catégorie, particulièrement importante pour notre étude, nous renseigne sur le type de contexte proposé à l'élève lors des exercices. S'agit-il notamment d'un contexte issu des supports traditionnels du calcul probabiliste, du contexte de la classe ou d'une situation de travail dans la société. Ou alors, note-t-on une absence de contexte ou une référence à d'autres domaines mathématiques. Le regroupement est dû à des raisons d'effectifs.

Codage	Etiquette	Description
E1	Contexte(Urnes ou de Jeux)	Urnes, jeux, cartes etc.
E2	Contexte(Classe-Quotidien)	Référence à l'environnement de l'élève ou d'un travail
E3	Contexte(Sans ou Maths)	Absence de contexte ou contexte purement mathématique

TAB. 3 – *Nature du contexte*

Variables portant sur les hypothèses du modèle F :

Nous cherchons, à partir du statut donné aux hypothèses, à caractériser le raisonnement mis en œuvre dans le problème, pour ainsi émettre des hypothèses sur le signifié potentiel de la probabilité. Si l'on probabilise :

- dans le sens déductif (des hypothèses vers la série infinie), on sera en présence potentiellement d'une interprétation fréquentiste ;
- dans le sens inductif (des hypothèses vers un élément de la série), alors il s'agira plutôt d'une probabilité bayésienne avec un possible apport du principe fréquentiste (Fig. 1) ;
- dans le sens d'une abduction au sens de Peirce (Everaert-Desmedt, 2006) où l'on va de quelques évidences vers des hypothèses plausibles, alors c'est la probabilité bayésienne qui est en germe.

En classe de Première où aucun outil inférentiel n'est enseigné, on ne doit pas s'attendre à probabiliser sur l'hypothèse d'un modèle.

Codage	Etiquette	Description
F1	H-Admises	Hypothèses du modèle sont admises, l’ensemble des événements possibles est spécifié
F2	H-Frequen ou P(A)	Hypothèses admises par la proportion dans un échan--tillon important ou explicitement fournies
F3	H-ADecider	Hypothèses à décider à partir de la proportion d’un petit échantillon.

TAB. 4 – *Statut des hypothèses*

Variables concernant la demande du calcul de probabilité H :

Nous nous intéressons ici au statut de la notion de probabilité en tant qu’objet mathématique : le signifiant P(A) est-il demandé (explicitement ou implicitement) ou non ? En croisant ces variables avec l’éventuelle demande d’interprétation (J) nous verrons les liens entre signifiant et signifié dans les exercices.

Codage	Etiquette	Description
H1	Cal-Oui	Calcul de P est explicitement ou implicitement demandé
H2	Cal-Non	Calcul de P n’est pas demandé

TAB. 5 – *Calcul de la probabilité*

Variables concernant l’explicitation d’un contexte fréquentiste ou non I :

Avec ces variables nous cherchons à renforcer l’identification de la notion de probabilité potentiellement en jeu en la confrontant aux variables F (nature des hypothèses) et K (nature de l’événement). Dans les exercices des manuels, nous avons repéré le phénomène suivant : en général, l’énoncé précise les éléments de l’ensemble des cas possibles puis, une action hypothétique est proposée “on tire au hasard une carte...” avec pour fonction de donner un contexte d’incertitude et aussi de fournir des indices sur la cardinalité de l’événement en question. Enfin, une demande de probabilité est explicitée (Calculer la probabilité que la carte tirée soit ...). Lorsqu’il n’y a aucune référence à la série mais à un élément de la série (la carte tirée), nous sommes dans un contexte bayésien de la probabilité.

Cependant, quelques exercices proposent de vérifier les calculs effectués grâce à des simulations ou à des répétitions de l’expérimentation. Ici, nous sommes en présence encore une fois de la manifestation du principe fréquentiste.

Variables concernant la demande d’interprétation J :

Cette variable confrontée aux variables I (répétition), F (nature des hypothèses) et K (nature des événements) nous informe sur la notion potentielle de la probabilité dans un exercice et du lien avec son signifiant (variables H).

Codage	Etiquette	Description
I1	Rep-Oui	Mention explicite ou implicite de la répétition de l'expérience aléatoire
I2	Rep-Non	Aucune mention de la répétition

TAB. 6 – *Répétition de l'expérience*

Codage	Etiquette	Description
J1	Inter-Oui	L'interprétation de P est explicitement ou implicitement demandée
J2	Inter-Non	Pas d'interprétation de P demandée

TAB. 7 – *Interprétation de la probabilité*

Variables concernant la nature de l'événement à probabiliser K :

Elles visent à observer particulièrement la nature de l'événement A dans l'application P(A).

– A est une série infinie d'événements susceptibles de se reproduire : on est en présence de la notion fréquentiste ;
– A est un événement dont la reproductibilité est ou bien impossible ou bien elle conduirait à modifier les valeurs données à la probabilité ;
– A est un cas générique élément d'une série éventuellement infinie. Dans le cas fréquentiste cela n'a pas de sens de probabiliser sur un élément, pour nous, demander par exemple la probabilité de tirer un six lors du prochain lancer avec un dé équilibré relève de l'approche bayésienne et non fréquentiste.

Codage	Etiquette	Description
K1	Nat-Série	Série infinie d'événements
K2	Nat-CasUnique	La reproductibilité est impossible
K3	Nat-CasGenerique	Cas générique d'une série
K4	Nat-Ambiguë	

TAB. 8 – *Nature de l'événement*

Variables non utilisées

Nous avions pensé que la description d'un événement déjà arrivé permettrait de donner une mesure de certitude en fonction de l'information disponible. Ceci nous a conduits à considérer la temporalité de l'événement. Cependant la très faible occurrence de la variable passé nous a conduits à la supprimer de l'étude : la plupart des exercices décrivent l'événement en utilisant un temps du présent ou un infinitif impersonnel.

D'autre part, dans les deux ouvrages que nous avons retenus, ne figurent pas deux variables importantes, l'une concerne la prise de décision et l'autre l'existence d'expériences simulées.

Si la prise de décision est également absente des autres manuels que nous avons observés, ce n’est pas le cas des expériences simulées, notamment avec tableur, qui représentent 15 % des deux autres manuels et sont naturellement mises en relation avec des expériences répétées.

5 Résultats

Nous allons commencer notre présentation des résultats avec le livre Nathan qui se prête sans doute mieux à une étude statistique compte tenu de la grande abondance des exercices. Avant de commenter nos résultats, voici l’arbre cohésitif.

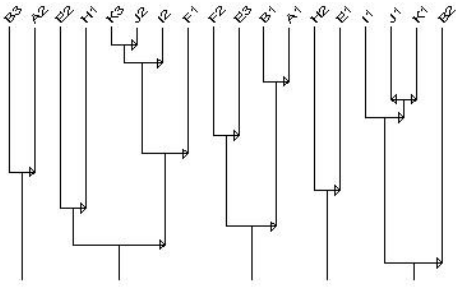


FIG. 2 – *Arbre cohésitif Nathan 1S*

Grâce à une analyse au seuil de 0.90, nous avons pu retenir quatre ensembles qui vont structurer de manière significative l’ensemble des exercices.

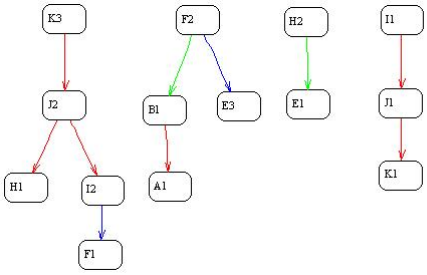


FIG. 3 – *Graphe Nathan*

Exercices génériques. Le premier ensemble, formé par [K3,J2,I2,F1,(H1)], représente près des deux tiers des exercices donnés dans ce manuel et rend compte de l’exercice type auquel est confronté l’élève de cette classe. Il s’agit de raisonner sur un cas générique (K3) sans que la répétition de l’expérience ne soit suggérée et sans qu’aucune interprétation du résultat ne

soit demandée. En complément, notons que ces exercices sont balisés par la donnée fréquente de l'ensemble des événements (E1). Les exercices prenant appui sur un contexte de classe ou de travail impliquent également la branche H1 qui demande l'explicitation d'un calcul de la probabilité. Ainsi, l'ensemble le plus typique des exercices proposés dans cette classe se situe clairement dans le modèle probabiliste. Celui-ci est donné d'entrée et aucun lien n'est demandé avec une situation à l'origine du modèle.

Exercices fréquentistes. La classe [K1,I1,J1] est très homogène dans cet ouvrage et elle présente des exercices basés sur une suite (K1) avec une répétition de l'expérience (I1) et une demande d'interprétation (J1). Ces exercices peu nombreux (en fait six) correspondent à la demande du programme qui souhaite que la probabilité soit introduite à partir de la fréquence.

Exercices d'entraînement. Notons aussi un petit ensemble [F2,E3,B1,A1] d'exercices (une dizaine) qui sont des exercices courts sans contexte et avec la donnée de la probabilité P(A). Il s'agit d'exercices d'application sur la notion de probabilité.

Exercices sur urnes et jeux. La classe [H1,E1] montre seulement que les exercices avec jeux et urnes forment une entité autonome mais on sait par ailleurs qu'elle est fortement connotée là encore par l'idée de modèle.

Ainsi, dans cet ouvrage qui respecte le programme, quelques exercices travaillent sur le sens fréquentiste de la probabilité mais l'ensemble apparaît comme un travail intensif dans un modèle qui privilégie le signifiant probabiliste. Si l'on suit l'approche de Peirce, ce travail intensif sur un signifiant laissé à la charge de l'élève (l'interprétant) débouchera sur un signifié profondément dépendant du jeu sur le signifiant. On peut s'attendre à ce que l'élève ne résume les probabilités qu'à un jeu dans le modèle laissant en jachère tout travail sur les signifiés dont nous avons parlé. Ce point sera clarifié par notre étude au niveau des BTS où nous devons nous attendre à retrouver une faible progression dans la compréhension des phénomènes liés au hasard avec des conceptions initiales assez proches de celles que Lahanier-Reuter (1999) a pu dégager chez des élèves plus jeunes.

Donnons maintenant notre analyse de l'ouvrage de première ES.

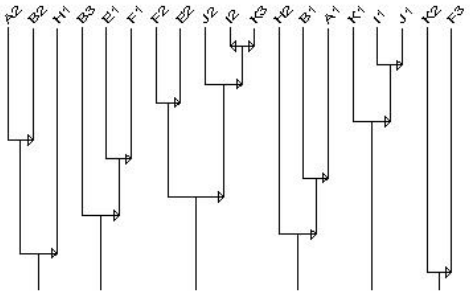


FIG. 4 – *Arbre cohésitif Bréal ES*

L'arbre cohésitif confirme bien l'impression que nous a donnée cet ouvrage assez atypique qui propose un nombre restreint d'exercices mais d'une très grande variété. La seule présentation de ces exercices ne permet pas aisément de dégager une idée dominante : l'influence du

professeur sera déterminante en supposant qu’il ait lui-même une idée nette des objectifs qu’il poursuit.

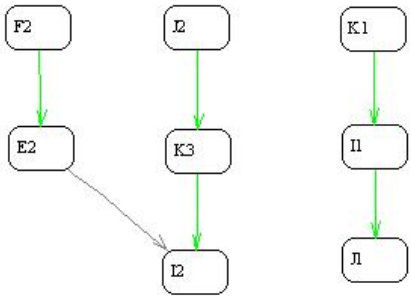


FIG. 5 – *Graphes Breal ES (seuil 0.90)*

En retenant les graphes implicatifs au même seuil que pour l’ouvrage de première S, nous retrouvons, mais de manière moins marquée, certaines des classes précédentes. C’est ainsi que nous retrouvons la classe fréquentiste [K1,I1,J1] qui s’appuie sur des répétitions d’expériences, cependant son homogénéité est moins grande et un certain nombre d’interprétations ne sont pas demandées dans le contexte de détermination d’une série, ce qui n’était pas le cas précédemment. Nous retrouvons aussi des exercices génériques [J2,K3,I2] mais de façon moins marquée qu’en Première S.

La dernière classe, reliée à la précédente par I2 (pas d’interprétation demandée) mais au seuil de 0.82 est la classe [F2,E2]. Elle est nouvelle et spécifique de la section ES, il s’agit de proposer des exercices dans un contexte quotidien (classe ou travail) : ils sont relativement très nombreux (plus de la moitié des exercices) et s’appuient prioritairement sur la donnée de la probabilité. Tout se passe comme si, du fait que cet enseignement s’adressait à des élèves qui ne deviendront pas des scientifiques, il était nécessaire d’ancrer davantage l’enseignement des probabilités dans une interaction avec la réalité source du modèle. Cependant, d’une certaine façon, les auteurs se trouvent pris dans une contradiction car ils ne peuvent interpréter correctement dans une approche fréquentiste la plupart des problèmes qu’ils donnent et qui relèvent nettement de la deuxième approche du signifié.

6 Conclusion

En examinant de manière exhaustive les exercices proposés dans des manuels de première, nous espérions dresser une carte de l’espace de travail probabiliste potentiel des élèves de Première. Nos résultats confirment la priorité qui est donnée dans l’enseignement français à l’axe qui va de l’interprétation fréquentiste à l’axiomatique de Kolmogorov mais dans sa forme molle.

Les élèves travaillent ensuite entièrement dans ce modèle en manipulant des signifiants qui finissent par acquérir un signifié assez éloigné des interprétations fondamentales qui ont donné

lieu à l'émergence de la probabilité et qui donnent aussi du sens à son usage dans le monde réel. Il faut noter cependant une différence sensible entre les deux manuels observés sans pour autant conclure à une spécificité liée à la classe, l'un s'engage moins dans la voie ensembliste et se trouve donc davantage confronté de manière plus ouverte à la question de l'intrication des signifiés. Dans les deux cas, mais de manière un peu différente, nous voyons, en germe, naître toutes les questions futures posées à l'enseignement des statistiques par la difficile adéquation que les élèves devront réaliser entre réalité et modèle dans la résolution des problèmes qui seront issus de la réalité.

REFERENCES

Batanero, C., Godino, J. D., et al. (2004). Training teachers to teach probability *Journal of statistical education 12(1)*

Carranza, P., et Kuzniak, A. (2006). Dualité de la notion de probabilité et enseignement de la statistique au Lycée en France. *Colloque EMF 2006*, Sherbrooke.

Carranza, P. (2009) *La dualité de la probabilité dans l'enseignement de la statistique* Thèse Université Paris-Diderot.

Droesbeke, J.-J., Fine, J., et al. (2002). *Méthodes bayésiennes en statistique.* Paris, Sfds.

Everaert-Desmedt, N. (2006), La sémiotique de Peirceă, dans Hébert, L. (dir.), Signo [en ligne], Rimouski.

Gärdenfors, P., Sahlin,N.-E., et al. (1988). *Decision, Probability and Utility,* Cambridge University Press.

Gras, R., S. Ag Almouloud, et al. (1996).*L'Implication Statistique* Grenoble, La Pensée Sauvage.

Gras, R. and M. Bailleul, Eds. (2000). *La Fouille dans les données par la méthode d'analyse statistique implicative.* Caen, ARDM.

Hacking, I. (2002). *L'émergence de la probabilité*, Paris,Seuil.

Hacking, I. et Dufour, M. (2004). *L'ouverture au probable.* Paris, Armand Colin.

Kuzniak, A. (2004) *Paradigmes et espaces de travail géométriques* Université Paris-Diderot

Lahanier-Reuter, D. (1999) *Conceptions du hasard et enseignement des probabilités et statistiques.* Paris, PUF

Régnier, J.-C. et Oriol, J.-C. (2001). *Fonctionnement didactique de la simulation en statistique.* Journées de Statistique Lyon 2003, Lyon, France, Sfds.

Summary

Based on a study of textbooks on the topics of probability in relation to statistics, we sketch the potential working space existing at Grade 11 in France. A strong calculative and formal approach or a more vague tendacy are higlighted by a statistical implicative analysis. In the second case, what we call the intricacy of probability's signifiers seems to have a real importance

Thème 2 : Applications à la psychologie

Chapitre 3 : Interprétation de graphes implicatifs : étude clinique auprès d'une chercheure en iconographie médiévale

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

* EA 4129 Laboratoire « Santé, Individu, Société »
5 Av. Mendès France Université Lyon 2 69500 Bron France
Pole École et Société -IUFM Université Lyon 1
5 rue Anselme 69317 Lyon cedex 04
acioly.regnier@wanadoo.fr
**Université de Lyon – UMR 5191 ICAR
ENS-LSH 15, Parvis René Descartes BP 7000.69342 LYON cedex 07
jean-claude.regnier@univ-lyon2.fr

Résumé. En Analyse Statistique Implicative, le logiciel CHIC[1] offre des interfaces efficaces pour un usage accessible au non-spécialiste. Ce dernier est confronté à la lecture et l'interprétation des représentations graphiques produites par ce logiciel. Or contrairement à une idée répandue : « un dessin, un graphique parle de lui-même », l'interprétation des graphiques requiert des connaissances minimales sur l'outil même que le chercheur doit alors acquérir. Ce travail poursuit la réflexion développée dans (Acioly-Régnier & Régnier, 2005, 2008), (Régnier & Acioly-Régnier, 2007) sur les obstacles à la conceptualisation liés aux représentations symboliques et met l'accent sur les difficultés rencontrées dans le cadre d'une recherche en iconographie médiévale. Nous cherchons à les expliciter de façon clinique lors de l'interprétation des graphes d'implication. La forme même d'une représentation graphique peut induire des lectures erronées des propriétés des données comme l'ont montré des études sur l'interprétation des signes et leur dépendance des expériences préalables des sujets.

1 Introduction

"Nous sommes appelés à n'être que des lecteurs de signes. Le monde moderne, le monde urbanisé est peuplé de signes auxquels nous répondons. Trouble quand nous ne savons pas ce qu'ils signalent (en voyage à l'étranger par exemple) ou quand ils sont

[1] Logiciel dédié aux traitements de l'Analyse Statistique Implicative, l'acronyme C.H.I.C. signifie Classification Hiérarchique, Implicative et Cohésitive. (voir Chapitres 11 et 12 de la Partie 2 de cet ouvrage)

> indécis et que nous sommes contraints de les interpréter". (J.-B. Pontalis, 2000 p.110)

Ce propos se situe dans un ensemble de réflexions autour de l'usage de l'analyse statistique implicative par de non-spécialistes. Dans notre premier travail (Acioly-Régnier & Régnier, 2005, 2008), nous avons abordé une problématique dans le domaine de culture et cognition, centrée sur un objet de l'astronomie : la lune et ses phases. Dans ce cadre, nous avons 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. La rencontre entre un spécialiste de la statistique et les premiers apprentissages d'une chercheure en sciences humaines fait alors émerger l'importance d'étudier de façon plus approfondie des obstacles liés à l'usage même de cet outil et aux difficultés posées par les interprétations. Suivant ce raisonnement nous avons ensuite (Régnier & Acioly-Régnier, 2007) abordé les aspects théoriques sur lesquels se fonde cette analyse. Ainsi avons-nous tenté de nous confronter à quelques questions soulevées par la modélisation en mettant l'accent sur les apports instrumentaux espérés par le chercheur non-spécialiste et par les difficultés auxquelles il est confronté pour conduire les interprétations des représentations graphiques. Dans ce cas précis, il s'agissait des arbres cohésitifs. Dans une perspective de même nature, nous abordons ici la question de la lecture et de l'interprétation des graphes d'implication.

Certes, cette thématique ne revêt pas un caractère original en tant que centrée sur la lecture-interprétation de représentations graphiques, mais, en revanche, elle trouve son originalité dans son inscription dans le champ de l'usage de l'ASI en tant qu'outil d'analyse des données dans divers domaines de recherche. Pour ce qui concerne les travaux réalisés, nous pouvons citer ceux sur la production du sens dans l'usage qui est fait, des représentations graphiques. Par exemple, Meira et Pinheiro (2007) se sont intéressés au processus de sémiotisation inhérent à la production de sens dans un contexte extra-scolaire. Monteiro et Ainley (2004) étudient, pour leur part, la question de l'interprétation des représentations graphiques en usage dans les journaux, les médias de la presse écrite et celle du sens critique des lecteurs. Régnier (1998) en s'intéressant à la question de la lecture d'articles de journaux de la presse ordinaire avait mis en évidence un questionnement fondamental dont nous pouvons aussi nous inspirer ici, fondé sur le point suivant : comment un article de presse régionale, que nombre de lecteurs parcourent en diagonale, s'avère requérir un traitement parallèle écrit explicitant le modèle mathématique sous-jacent pour parvenir à la compréhension et au contrôle de la validité de ce qui est dit dans celui-ci. Ce questionnement peut être décliné ainsi :

1. Quelles informations le lecteur en tire-t-il ?
2. Comment contrôle-t-il sa compréhension du texte ?
3. À quelles activités se livre-t-il pour procéder à ce contrôle ?
4. Comment contrôle-t-il la validité de ce qui est énoncé ?

Pour sa part, Monteiro (1998) avait aussi conduit une recherche sur l'interprétation des graphiques dans la presse écrite portant sur des informations économiques au Brésil. Dans ses conclusions, il fait ressortir le point suivant selon lequel il y a nécessité à adopter une perspective élargie autour du concept d'interprétation en ne le restreignant pas aux connaissances mathématiques liées aux élaborations formelles, et en mettant en évidence l'activité même d'interprétation de ces graphiques comme un processus de résolution de

problèmes. Ce dernier est marqué par l'interaction de trois éléments fondamentaux : les spécificités du sujet, celles des tâches qu'il doit accomplir et les stratégies de résolution qu'il met en œuvre dans cette activité.

Ici, dans ce chapitre, nous partons d'une étude empirique qui porte sur l'analyse d'un entretien vidéographié de 2 heures auprès d'une doctorante en iconographie médiévale qui a fait choix de l'usage de l'approche ASI dans son travail de recherche (Partie 3 Thème 5) et que nous nommerons Magali dans tout notre propos. Nous devons préciser que cet usage correspond pour elle à une situation de découverte et d'apprentissage de la méthode. Nous avons visé, en utilisant une approche clinique, l'identification de quelques difficultés et obstacles rencontrés, l'explicitation de connaissances minimales nécessaires pour les affronter efficacement et le repérage des effets éventuels de l'expérience préalable du sujet concernant l'interprétation des images dans l'expérience actuelle, *hic et nunc*. Nous basons notre propos sur des données issues du repérage d'hésitations, de dysfonctionnements et d'erreurs manifestés lors de l'interprétation des représentations graphiques. Dans le prolongement de ce travail, nous réaliserons une analyse exhaustive du discours et des gestes associés que nous publierons plus tard.

D'un point de vue théorique, nous nous situons dans une approche historico-culturelle du psychisme où les outils et les signes sont créés socialement tout au long de l'histoire humaine. Dans ce cadre théorique, les opérations avec des signes sont le résultat d'un processus prolongé et complexe, soumis à des lois de base d'évolution psychologique. Dans la perspective vygotskienne, ces signes en tant que médiateurs ne peuvent être compris que dans le contexte particulier où ils sont utilisés. Nous précisons que les réflexions que nous menons ici, même si elles ont l'ambition de participer à une meilleure compréhension du développement des compétences dans le champ de l'Analyse Statistique Implicative, se limiteront à chercher à mieux comprendre le fonctionnement du sujet à un moment donné de l'utilisation des outils ASI. En d'autres termes, ce sont davantage les aspects opératoires des connaissances construites par le sujet qui sont l'objet de notre étude que les processus d'apprentissage.

L'originalité de la **démarche clinique** proposée ici consiste dans l'étude d'une situation où la chercheure travaille elle-même, dans le cadre de ses recherches du domaine de l'iconographie médiévale, avec des images en tant qu'objets d'étude, et se trouve aussi confrontée à une autre catégorie d'images, celles produites par le logiciel CHIC. Dans cet usage, il lui faut être en mesure d'interpréter les représentations des graphes implicatifs. Habituée à analyser des images relevant d'un cadre conceptuel particulier, le modèle théorique de sa discipline – M(T) dans le sens donné par Régnier (Régnier et Acioly-Régnier 2007) il lui faut maintenant intégrer de nouveaux concepts du modèle de la statistique – M(S) pour interpréter des images qui serviront dans ce cas à donner du sens à la résolution de sa problématique. Ainsi, si la familiarité avec l'usage des images et leurs interprétations peut être facilitatrice du travail d'interprétation de graphes d'implication, elle peut aussi se constituer en source d'obstacles du fait de transferts non pertinents.

D'une certaine manière, sur la base d'une perspective relevant du domaine de la psychologie, nous proposons aussi une contribution au développement du champ de la didactique de la statistique, plus particulièrement la **didactique de l'ASI**.

2 Identification de quelques difficultés et explicitation de conditions pour les affronter

Pour expliciter les difficultés et repérer leurs natures et origines, nous avons procédé à une analyse du discours de Magali. Dans tout son propos, nous avons pu observer un fort investissement dans l'étude et l'appropriation d'un outil et repérer des connaissances prédicatives qui indiquent un niveau déjà avancé de maîtrise des concepts de base de l'ASI. Quand nous nous intéressons aux connaissances opératoires, nous repérons alors des difficultés liées à un niveau de conceptualisation qui limite un usage autonome de l'outil et l'interprétation de signifiants associés.

2.1 Difficultés précédant l'usage de l'ASI : Formation littéraire *versus* formation scientifique

Concernant les difficultés pouvant puiser leur origine dans la formation initiale, nous avons relevé dans le discours de Magali une représentation sociale forte qui oppose littéraire et scientifique.

« *Ma formation de base, dès le bac est littéraire, bac histoire des arts déjà, ensuite j'ai fait une formation (…) on n'a aucune formation dans ce qui peut être base des données, statistique, on peut bien avoir des cours d'informatique en première et deuxième années d'université, ce n'est pas le plus important car on ne nous explique pas l'importance de ces cours et il fallait mieux faire ces cours en troisième année ou en Master pour que vraiment on puisse mieux comprendre l'importance. Sinon je pars vraiment de zéro* » (Magali)

Force est de constater combien implicitement nous pouvons déjà identifier la dichotomie scientifique et littéraire qui peut facilement s'ériger en source d'obstacles. Un point de vue que Pontalis souligne remarquablement dans son propos quand il rappelle que « Vieille opposition, instituée dès le lycée, entre « littéraires » et « scientifiques » de mon temps au bénéfice des premiers, aujourd'hui des seconds. Très tôt mes professeurs m'ont fait savoir que je n'avais pas « esprit scientifique » (…) Les qualités requises – esprit logique, précision, souci de l'exactitude – ne me manquaient pas quand il s'agissait de traduire du grec ou de l'anglais. Alors ? Une incapacité à me confronter à l'abstraction, à me fier à des signes trop distants du sensible, du visible, du perçu ? Pourtant, en classe de philosophie, je sus me mouvoir avec aisance dans un monde fort éloigné du « concret » et, en maths, je finis par prendre goût à l'algèbre alors que la géométrie dans l'espace me donnait le vertige. » (Pontalis, J.-B. 2000. 113-116)

2.2 Difficultés liées à la manipulation d'un tableur et des bases mathématiques associées aux formules

Comme nous l'avons déjà abordé dans un autre cadre, (Acioly-Régnier, 2008), nous considérons que « les outils informatiques ne sont pas d'emblée des instruments psychologiques au sens de Vygotski. Ils ne le deviennent qu'à condition que des activités pertinentes et des schèmes associés rendent possible leur appropriation en tant que telle, chez

les sujets. La simple maîtrise d'un instrument technique ne conduit pas d'emblée les sujets vers l'appropriation de l'instrument psychologique ».

Ici nous avons identifié un second groupe de difficultés qui tirent son origine de la manipulation concrète d'outils informatiques tels que les tableurs comme le logiciel Excel. Ces outils prennent une importance capitale dans le traitement de base des données construites et, en particulier, pour parvenir à l'usage de CHIC. En effet ce logiciel exige une base de données mise au format csv, par exemple. Mais au préalable, il y a la construction d'un tableau de valeurs binaires, fréquemment de type disjonctif total et le recours à des fonctions intégrées.

Pour Magali « *Quand il s'agit de faire les premiers résultats, les premiers (euh) les premières sommes, les premières évaluations, on va dire, je ne connaissais pas les formules, je ne les connais d'ailleurs toujours très peu (rires) donc voilà ça a été difficile de comprendre (…) C'est vrai que c'est très répétitif, par exemple, pour les formules de base. Oui j'en reste là pour l'instant, ça a été de comprendre un petit peu les différentes formules mathématiques pour réussir ensuite à en tirer des pourcentages, ensuite passer également de ces pourcentages à l'élaboration des histogrammes, pour comprendre un petit peu, se plonger dans des connaissances de maths qui se sont arrêtées au niveau de la classe de troisième, à la fin du collège, les abscisses, les ordonnées, ce sont de vieilles réminiscences, c'est très difficile quand on est* ***purement littéraire****, c'est quelque chose de très dure, à essayer à tout comprendre, voilà après, quand on a le tableau sous les yeux, on ne regrette pas d'avoir passé autant de temps là-dessus (rires) parce que ça ouvre des champs au niveau de la recherche qui sont fantastiques, et là en restant au stade de la statistique, ça nous montre ce qu'on peut faire en se plongeant dedans*. »

Nous voyons que le rapport que Magali établit avec les outils informatiques et leurs produits ne transforme toutefois pas les difficultés en obstacles qui pourraient alors l'éloigner de l'usage de l'outil CHIC.

2.3 Difficultés liées à la nécessité d'une activation soutenue des processus cognitifs attentionnels et mnésiques

Cet aspect de l'activité du sujet ressort clairement dans le propos suivant de Magali : « *La principale difficulté depuis le début, c'est juste de faire la base des données, ça peut paraître très bête. Mais de ne pas me tromper dans les chiffres, parce que si on reprend mon tableur Excel, j'ai donc rentré différentes données, ça part de la nature du support de l'image jusqu'à son origine, l'image en elle-même et donc* ***je peux avoir jusqu'à 8 ou 9 occurrences ensemble pour une action spécifique*** *du Christ, pour prendre cet exemple là, et si je me trompe sur ces chiffres là, si je donne d'autres chiffres qui n'appartiennent pas au corpus, ça peut fausser les données. Pour la construction du tableur* ***ça demande énormément de concentration****, je n'avais pas trop l'habitude, même maintenant que j'ai plus d'habitude pour rentrer des images,* ***je suis hypervigilante pour ne pas faire d'erreurs***. »

Ce propos de Magali est interprétable dans le cadre théorique de la psychologie cognitive comme le souligne Cadet (1998 p.152) en faisant lui-même référence aux travaux de Logan (1988) qui portent sur le lien entre l'attention et la mémoire. Selon ce dernier, tout traitement qui nécessite de l'attention, laisse des traces mnésiques et la répétition des traitements s'accompagne d'une augmentation des informations disponibles et donc d'une meilleure connaissance des conduites les mieux adaptées. Ainsi l'automaticité est une restitution de la

mémoire et la rapidité du processus s'explique par le fait qu'il s'agit d'un accès direct à la restitution par la mémoire des solutions passées. Un tel parallélisme entre attention et mémoire a été confirmé par l'observation comparative de sujets novices et experts. Si les experts sont à même de réaliser des traitements de manière semi-automatique, en revanche les novices rencontrent des difficultés non par défaut d'attention mais par manque de connaissances. Nous pourrions dire que l'expert et le novice ne se situent pas au même niveau de conceptualisation et c'est qui modifie leurs pratiques.

Cette perspective confirme l'insistance dont fait preuve Magali à l'égard du niveau d'attention qu'elle ressent pour accomplir sa tâche.

Ce phénomène réapparaît dans la phase de l'interprétation des graphes implicatifs. Le sujet pour conduire cette interprétation est alors contraint de maintenir, en mémoire de travail (Ehrlich et Delafoy, 1990), une quantité importante d'informations de plusieurs types : les signifiés des codages alpha-numériques des variables binaires (exemples : JCPos1, JCPres3), les signifiés des variables du modèle construit pour l'analyse des images étudiées dans le champ de l'iconographie médiévale (exemple : position du Christ, présentation du Christ), la signification donnée par l'orientation des flèches représentant les quasi-règles et leurs positions dans le plan (CHIC : surface de travail), la couleur des flèches correspondant à des niveaux de confiance en la quasi-règle (CHIC : rouge, bleu, vert, noir), le tracé en pointillé ou en continue de ces flèches.

L'interprétation de la figure (Fig 1) est un exemple de situation qui requiert un stockage important d'informations en mémoire de travail

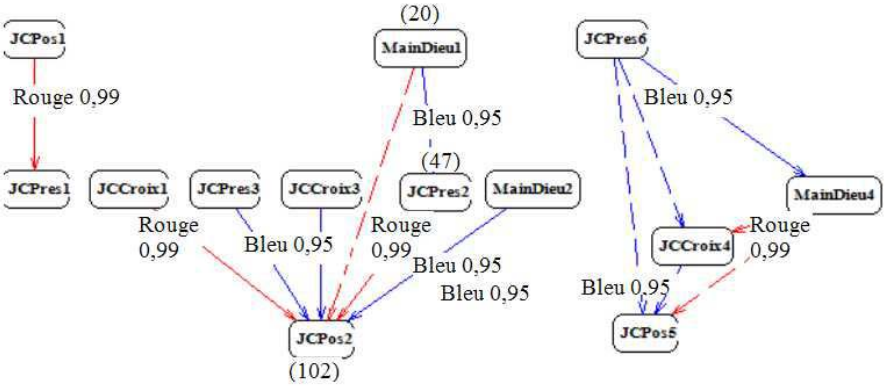


FIG. 1 - *liens entre la variable Main de Dieu et les variables Position, Présentation, Croix du Christ*

M : *voilà, alors dans ce graphique, j'ai voulu mettre en lien la présentation du Christ. Par présentation, j'entends savoir si le Christ est de face, de profil. Là j'ai pris uniquement les occurrences du Christ de face et de profil puisque j'ai choisi la position du Christ debout. Donc la position du Christ debout, sa présentation de face et de profil et, dans ces cas là* (*euh*) *comment* (*euh* – pause longue) *… non, j'ai pris toutes les présentations du Christ, en fait, parce que je veux savoir de la* (JCPos1) *position 1 (assise) et la* (JCPos6) *position 6, ça doit être...* (elle cherche dans sa base de données)*… présentation (euh), excusez-moi* (elle passe de la recherche sur sa base des données inscrite sur un papier à celle présente dans

l'ordinateur) *alors lorsque la présentation du Christ, bon ! je vais le refaire,* (JCPres1) *présentation 1, il est de profil, non, présentation 1, il est de face,* (JCPres2) *présentation 2, il est de profil,* (JCPres3) *présentation 3, il est de trois-quarts,* (JCPres4) *présentation 4, il est de dos,* (JCPres5) *présentation 5, il est de trois-quarts dos et* (JCPres6) *présentation 6, on ne connaît pas le Christ.*

N : *Alors, on a la position 2, il est de profil et la position 5 c'est laquelle ? je suis aussi perdue car je n'arrive pas à tout mémoriser*.

M : *là c'est la présentation et là c'est la position*

N : *ah ! d'accord*

M : *La position du Christ, c'est encore une autre chose,* (JCPos0), *position 0 le Christ n'est pas représenté,* (JCPos1), *position 1 on ne voit que ses pieds,* (JCPos2), *position 2 il est présenté de la tête au pied*

N : *alors ici c'est celle là ?* (en montrant sur le graphique le pôle JCPos2)

M : *Voilà,* (JCPos3) *position 3, il est assis,* (JCPos4) *position 4, il est en buste,* (JCPos5) *position 5, c'est inconnu, donc là on va trouver un schéma isolé par rapport à la position inconnue et* (JCPres6) *présentation 6, que j'ai rajoutée par la suite, les pieds du christ ne sont pas représentés,* (rires signifiant peut-être une exhaustivité des critères) *il y a vraiment des gradations* (rires). *Donc* (elle reprend le tout) *j'ai pris la position du Christ, sa présentation, ensuite je me suis basée sur* (consulte de nouveau l'ordinateur) *sur deux éléments qu'on peut mettre en lien les uns avec les autres avec la position du Christ, c'est la représentation de la croix du Christ, alors elle peut être tenue par le Christ, au-dessous du Christ, ça peut être une image spécifique d'une croix dite à bannière, (euh), ça c'est la* (JCCroix3) *croix numéro 3,* (JCCroix4) *croix numéro 4 est inconnue et* (JCCroix5) *croix numéro 5, j'ai rajouté la semaine dernière.*

Dans la suite, à ce niveau de l'entretien, elle poursuit en décrivant en détail la variable Main de Dieu. Ce besoin qu'elle ressent de procéder à cette description exhaustive paraît une nécessité à la fois pour maintenir les informations en mémoire de travail et produire l'interprétation du graphe implicatif (Fig 1)

3 Repérage des effets de l'expérience du sujet dans l'interprétation des graphes implicatifs

Le chercheur non spécialiste des domaines des mathématiques, des probabilités et de la statistique agit dans et sur l'outil CHIC en fonction de ses « représentations » explicites ou implicites, conscientes ou inconscientes. Celles-ci sont construites à partir des expériences passées qui se constituent comme des éléments d'ancrage à la compréhension de situations nouvelles. D'un point de vue piagétien, on parle d'assimilation lorsque l'information est transformée pour pouvoir être intégrée dans les structures cognitives déjà existantes. Lorsque celles-ci ne sont plus capables d'assimiler les informations, le système cognitif s'engage dans un processus d'auto-transformation désigné dans la théorie piagétienne comme accommodation. Lemmeignan et Weil-Barais (1993, p. 24-25) remarquent que l'accommodation ne peut être qu'exceptionnelle, car le changement entraîne toujours une déstabilisation dont l'issue peut être incertaine. Ils rajoutent que l'abandon d'un mode interprétatif familier impose de changer de registre de fonctionnement : d'automatique, la pensée doit fonctionner de manière contrôlée. Carraher, Schliemann et Nemirovosky (1995)

observent que la lecture et l'interprétation des représentations graphiques impliquent des processus cognitifs en lien étroit avec des connaissances mathématiques portant sur les notions de quantité, nombre, proportion, mais aussi avec les expériences antérieures des sujets.

3.1 Les signifiants de CHIC : images et mots qui représentent des concepts de l'ASI

Les **signifiants** qui représentent les concepts peuvent être des éléments facilitateurs ou encore constituer des obstacles à leur compréhension. Selon Vergnaud (1987) la représentation ne se réduit pas à un système symbolique qui renverrait directement au monde matériel, les signifiants représentant alors directement des objets matériels. En fait, les signifiants représentent des **signifiés** qui sont eux-mêmes d'ordre cognitif. Ainsi la notion de représentation recouvre celle de concept.

Le chercheur non spécialiste de l'ASI est alors confronté, d'une part, à des signifiants iconiques, à des images, des représentations graphiques telles que celles des graphes implicatifs, des hiérarchies orientées (Régnier et Acioly-Régnier, 2007), et d'autre part, à des signifiants langagiers, à des mots pour représenter des concepts, mais qui peuvent aussi évoquer des usages dans le quotidien avec un sens commun. Dans notre étude clinique, l'expérience du sujet dans le domaine de l'interprétation des images dans le champ de l'iconographie médiévale peut paradoxalement induire des difficultés dans la lecture des graphes implicatifs et des mots qui leur sont associés dans le contexte de l'ASI. Nous allons aborder deux situations qui illustrent notre propos : la fermeture transitive, le niveau de confiance.

3.1.1 La fermeture transitive

Dans la figure (FIG 2) ci-dessous représentant un graphe implicatif de nodulosité 3 (MainDieu1, JCPres2, JCPos2), il apparaît une flèche en pointillé qui indique une règle de quasi-implication entre les deux variables binaires (MainDieu1, JCPos2). Celle-ci indique localement à un niveau de confiance supérieur à 0,50, la relation vérifie la propriété de transitivité. Dans le langage de l'ASI, il est fait usage du signifiant « fermeture transitive »

Extrait du mode d'emploi :

Fer transi :
Elle permet, en choisissant 1, d'obtenir le tracé de tous les arcs associés aux fermetures transitives aux seuils choisis, et avec respect des couleurs. Ces nouveaux arcs apparaissent en pointillés

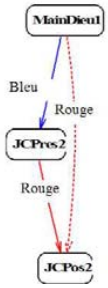


FIG 2

Dans l'entretien, il est ressorti des questions au sujet de cette fermeture transitive. Nous extrayons l'échange suivant qui permet d'expliciter un des aspects des difficultés auxquelles le sujet se confronte :

N : *Les concepts du CHIC sont désignés par des mots qui sont aussi en usage dans la vie quotidienne, comme par exemple fermeture. En ce qui concerne la fermeture transitive, tu as manifesté plusieurs hésitations quand tu as tenté d'interpréter le graphe implicatif de la figure (Fig 1). Mais qu'est-ce que cela t'évoque ?*

M : *oui, ça me brouille, effectivement parce que (…) je trouve que c'est contradictoire, au niveau purement du langage. Parce qu'au contraire, parce que ça serait plus une ouverture transitive, parce que pour moi, pour mon sujet, ça laisse le champ à d'autres interprétations, alors que fermeture transitive, on aurait tendance, par le terme fermeture à clore véritablement la donnée, alors qu'ici, au contraire, elle m'ouvre sur d'autres choses parce qu'elle permet de dire que justement elle ne passe pas par (euh, euh) par cette donnée pour arriver à la troisième, donc on peut avoir des intermédiaires.*

N. *Je comprends un peu car je suis moi-même encore d'un niveau non spécialiste. Mais peux-tu un peu plus préciser ?*

M : *ça m'évoque un lien qui peut être, euh, un lien qui peut être indirect, qui peut passer par d'autres chemins, mais dont le résultat sera le même.*

N : *Pourquoi transitive ?*

M : *Parce que justement, on laisse le champ ouvert à différents intermédiaires par lesquels on peut passer, donc transiter, pour arriver à cette donnée, et pour moi c'est le mot fermeture qui m'embarrasse.*

3.1.2 Un niveau de confiance de 95% est-il un pourcentage ?

Dans le modèle de l'ASI, nous avons recours à un critère de sélection des quasi-règles d'implication qui est une mesure de probabilité d'un certain événement comme il est exposé dans la Partie 1 de cet ouvrage. Ce critère, appelé niveau de confiance, est un nombre compris entre 0,50 et 1. Il est présenté dans le logiciel CHIC sous la forme exprimée en % comme le montre la figure (Fig 3) ci dessous. Les niveaux de confiance sont eux-mêmes représentés par des couleurs.

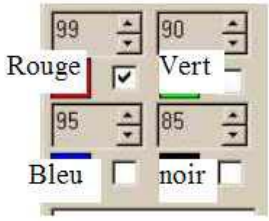


FIG 3

Du point du concepteur du logiciel comme de celui des spécialistes en ASI, cette codification et cette façon de représenter le critère de sélection sont considérées comme une présentation facilitatrice. Voilà ce qu'en dit Magali en tant novice.

M : *Le fait par exemple d'avoir de flèches de couleurs différentes, je m'en doutais que ça voulait dire quelque chose, mais après s'il n'y avait pas la légende en dessous du graphe implicatif, quel chiffre correspond à quelle couleur ? si on n'a pas ça, c'est très difficile de*

savoir, et même si on a uniquement les chiffres pour savoir s'il s'agit de pourcentage ou autre (euh) c'est difficile à comprendre pour quelqu'un qui arrive et qui est totalement extérieur à ça.

N : *Pourtant tu n'as pas semblé avoir trop de difficulté quand tu as essayé de m'expliquer dans le cadre de ta recherche sur tes images.*

M : *oui parce que j'ai vu ça auparavant mais pour une personne qui n'est pas au courant et qui voit écrit 95, 99 ça peut être n'importe quoi, ça indique au moins qu'on raisonne ici en pourcentage et même ce qui peut paraître le plus logique ça peut prêter à confusion dans l'esprit de gens. Après au niveau de tableaux, on comprend bien quand on voit le sens des flèches, telle chose, implique telle autre.*

N : *Revenons à cette question de 99% en rouge et des 95% en bleu.*

M : *alors pour moi (rires) à 99% ça veut dire qu'il y a 99%. Par exemple* (FIG 2*), « la main de Dieu implique que le Christ est dans telle position » et (euh) enfin, je prends cet exemple, (euh) avec ces deux éléments c'est plus facile, donc à 99% par exemple, « la main de Dieu qui bénit », si je reste à 99% je n'aurais aucune…* (euh hésitation*) je n'aurais aucun lien entre les deux. En revanche si je passe simplement à 95%, euh, bien là mes deux images, enfin mes deux données seront en lien. Donc, après c'est à traiter en fonction de justement toutes les images (…) c'est encore un peu difficile pour moi à expliquer correctement, mais la base de données va traiter absolument toutes les images , la main de Dieu n'apparaît pas sur toutes les images. Ce qui veut dire que la main de Dieu en lien avec la position du christ ça sera un pourcentage forcement plus faible, parce que ça sera pas uniquement les images qui ne prendront que la main de dieu mais sur absolument toutes les images du corpus, donc ça sera difficile de (*hésitation longue*) ça sera difficile (*répétition*) d'avoir un résultat sur un élément précis, une réponse qui pourra aller dans le sens de ma demande, du fait qui ça prendra absolument toutes les images. J'essaye d'expliquer ça par rapport également à mon objet* (tentative de se justifier pour l'ancrage). *Comme on a une très grande mobilité dans ces images médiévales, quelque chose qui apparaît dans 95% des images, on peut considérer que ça va être* (pause longue) *que ça va être la norme parce que la perfection n'existe pas et on ne peut pas avoir de (pause), on peut moins avoir de 99 ou de 100% en fait.*

Nous pouvons percevoir combien cette représentation du niveau de confiance par un % vient se confronter à l'usage des proportions exprimées en %. Cette interférence est génératrice de difficulté.

3.2 Un signifiant fondamental du l'ASI : le graphe implicatif

Comme nous l'avons déjà dit dans ce chapitre, nous nous situons dans un ouvrage qui présente la théorie de l'analyse statistique implicative, en particulier sur les variables binaires, en recherchant des relations non-symétriques entre celles-ci. Cette relation est définie comme une relation d'implication statistique, c'est-à-dire quasi-implication. L'opérationnalité de cette méthode d'analyse des données tient en partie à la fonctionnalité informatique qui rend possible la représentation graphique plane du graphe implicatif. Toutefois cette production requiert, pour l'usager, la capacité de lire-interpréter une telle représentation graphique, c'est-à-dire d'un signifiant de nature iconique. Ce que nous cherchons ici, est de mieux comprendre les difficultés que rencontrent les novices et non spécialistes des mathématiques et de la statistique, quand ils souhaitent faire un usage de la méthode dans leur domaine de recherche ou d'étude. Ici nous avons, par une approche

clinique, tenté de cerner les difficultés dans l'usage de l'ASI étayée par le logiciel CHIC, dans le domaine de l'iconographie médiévale qui se fonde épistémologiquement et méthodologiquement sur des objets de nature iconique. C'est-à-dire un domaine où le chercheur en iconographie médiévale est totalement familier avec l'usage et la lecture-interprétation des images.

3.2.1 Influence des approches spécifiques du domaine de l'iconographie médiévale sur l'interprétation du graphe implicatif

Voilà comment Magali explicite le lien entre la lecture et l'interprétation du graphe implicatif (Fig. 5) et son expérience acquise dans la lecture-interprétation des images en iconographie médiévale.

M. : *On commence à raisonner. La nuée 4* (Nuee4), *nuée au-dessus du Christ, implique une représentation du mont* (Sol2) *qui implique que* (JCPos2) *le Christ est debout. Alors qu'en fait c'est une lecture contraire, c'est le Christ qui est debout, qui implique la plupart du temps, une (euh), une montagne qui, elle, implique (euh), attendez* (elle regarde en haut) *c'est bien …* (éclats de rire) *c'est bien compliqué pour moi. Ça implique la nuée au-dessus du Christ* (gestes d'inversion avec les mains)

N : *D'accord, il me semble que tu es en train de faire une lecture dans le sens inverse de la façon dont est représenté sur le graphe implicatif dans la figure* (Fig. 5) *ce chemin* (Nuee4→Sol2→JCPos2)

M : (manifeste des gestes d'inversion avec les mains accompagnant son discours) *En fait, j'ai tendance à inverser les choses. Et là, en fait, on a plutôt l'impression que de la nuée 4, la nuée en-dessus du Christ, va découler tout le reste. Donc de la nuée découle une montagne de laquelle va découler la position du Christ. Alors que le plus important ici, c'est pour moi, l'élément à partir duquel part le plus de choses* (dans l'iconographie médiévale), *donc la position du Christ qui devrait être en haut de l'image* (graphe implicatif) *parce que c'est ce qu'il y a de plus important et ensuite faire une arborescence qui va partir… Du coup on part*

N : *Attend, si j'ai bien compris, tu penses que, par exemple, la représentation graphique doit être inversée ?*

M : *Pour moi, ça serait plus évident*

N : *Alors, la position du Christ debout* (JCPos2) *impliquant* le (Sol2) *qui implique la* (Nuée4) (lisant de bas en haut comme si on inversait le sens de flèches)

M : *Non* (geste de négation avec le doigt), *ce n'est pas ça en fait. Non, c'est plutôt mettre la position du Christ en haut et ensuite des flèches qui vont vers le Christ, bien sûr, mais non partir du haut vers le bas, mais bien du bas vers le haut.*

N : *Et pourquoi tu dis ça ?*

M : *Parce que pour moi, ça se lit plus facilement, c'est à dire, l'élément principal va être situé en haut (euh) de l'arborescence, ensuite (euh) on aura ce qu'il y a de moins en moins important, c'est une raison pour moi, purement hiérarchique.*

N : *et à ce moment là, en haut comme c'est représenté dans ton graphique (Fig. 5). Ici ça veut dire qu'il y a le moins d'occurrences pour les choses qui sont en haut, puis un peu plus au niveau 2 et encore plus au niveau 3* (chemin (Nuee4(22)→Sol2(35)→JCPos2(102)).

M : *Oui*

N : *Donc dans ta logique, ça serait plus lisible du point de vue graphique, si l'élément qui a plus d'occurrences se situait en haut, même si les flèches iront…*

M. : *Oui, c'est ça, iront vers lui. Oui, pour moi ça serait plus, plus lisible. En effet si je partais du fait qu'on a la position du Christ située en haut du graphique et que, vers lui, vont converger dans un mouvement ascendant, et donc du moins important au plus important, jusqu'au Christ, ça serait pour moi plus logique dans ma lecture. Certes, en fait, on peut également y voir, de mon côté, une déformation tout simplement parce que je travaille sur l'ascension du Christ et que l'image sur les supports étudiés va du bas vers le haut. En plus, même d'une façon plus générale, pour d'autres thèmes de l'iconographie médiévale que celui de l'ascension du Christ, l'élément le plus important est, la plupart du temps, placé en haut des images dans une orientation verticale.*

Il nous faut rappeler que c'est par un paramétrage par défaut du logiciel CHIC que le graphe implicatif se présente selon la direction verticale et l'orientation des flèches du haut vers le bas de la surface de travail. Nous voyons l'importance de la prise en compte des fonctionnalités qui permettent à l'utilisateur de déplacer selon son choix, chacune des variables binaires représentées. Ceci rend possible une représentation orientée verticalement de bas en haut ou encore horizontalement de gauche à droite ou inverse, ou encore tout autre disposition. L'extrait de l'entretien ci-dessus met bien évidence les premiers éléments de données qui laissent penser que ces choix d'orientation ne conduisent pas à des situations équivalentes pour l'interprétation.

3.2.2 Influence de la signification donnée par le chercheur à la notion d'implication sur l'interprétation du graphe implicatif

Voilà comment Magali explicite son recours au graphe implicatif à partir des figures (FIG 4, FIG 5) à partir du sens qu'elle donne à la relation d'implication au travers de l'expression « ceci implique cela ». Par ailleurs, il conviendrait de poursuivre des investigations pour expliciter l'influence de l'autonomie du chercheur quant au choix des niveaux de confiance (FIG 3), sur le sens donné au graphe implicatif et sur les interprétations du réseau de chemins qui en découle à chaque niveau choisi.

M : *Alors, ici justement, j'ai fait deux fois le même tableau. Avec une fois avec des occurrences uniquement à 99% et de lien entre les images. Ensuite celle à 95%, et on s'aperçoit qu'à 95%, forcement on aura beaucoup plus de liens entre les différents éléments.*

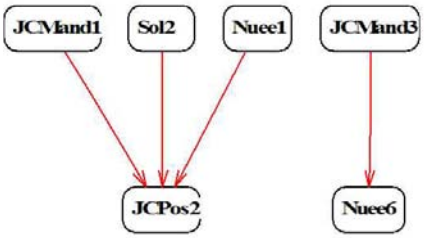


FIG 4 *: Lien Christ-Eléments naturels A*
Niveau = 99

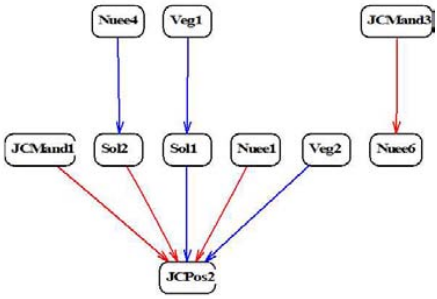


FIG 5 *: Lien Christ-Eléments naturels B*
Niveau = 95

Nous pouvons voir qu'à ce niveau de conceptualisation, il est compris l'effet du choix de la valeur du critère pour retenir les liens entre deux variables binaires.

Elle poursuit :

M : *Le graphique implicatif est le plus facile à comprendre dans l'immédiat. Le plus intéressant dans l'arbre implicatif, c'est de voir les implications d'un élément par rapport à un autre et du coup à mieux* (pause) *cerner l'image. Aussi quand on est à 99%, c'est très intéressant et en même temps, il y a beaucoup de choses qui tournent autour et qui apparaissent quand même à 95%, donc ici je me suis basée, sur ce tableau que j'ai intitulé* « Lien Christ-Eléments naturels A » (FIG 4), *c'est 99%* et « Lien Christ-Eléments naturels B » (Fig 5) *c'est 95%. J'indique une seule position du Christ, c'est* (elle recherche dans sa base de données Excel, en commentant qu'elle a un peu du mal à cause du nombre d'informations) « position du Christ numéro 2 », *qui représente le Christ quand il est débout. Donc du coup, le graphique implicatif, c'est aussi pour moi peut-être, le plus facile à comprendre dans l'immédiat, et une fois que (…) je saurai maîtriser correctement dans ses tenants et ses aboutissants, à ce moment là, après, j'irai voir les différentes sortes de graphiques* (Arbre des similarités, arbre cohésitif)

Il reste que la notion de quasi-implication dans sa confrontation à l'implication logique, l'implication mathématique et l'usage quotidien du si… alors… nécessite un travail de réflexion pour l'élévation du niveau de conceptualisation.

N : *Quel sens prend l'expression « ça implique » ?*

M : (rires) *Du coup, j'utilise le vocabulaire CHIC, ça implique, ça veut dire que* (euh) *lorsque, lorsqu'on a la mandorle* (pause longue). *Il faut que je réfléchisse pour que je pose mes idées correctement* (pause). *Dans la majeure partie des cas, lorsque la mandorle* (euh pause) *est représentée* (pause), *lorsque le Christ dans la mandorle est seul* (euh) *ça implique dans la majorité de cas qu'il est représenté debout, mais ça ne veut pas dire que lorsqu'il est représenté debout il est toujours représenté dans la mandorle*

N : *Alors est-ce pour ça que la flèche va dans un sens ?*

M : *voilà ! C'est pour ça que la flèche part de la mandorle vers le Christ. Parce que ça veut dire effectivement que lorsque la mandorle est représentée seule, dans la plupart des cas, le Christ est représenté debout et ça ne veut pas dire que lorsqu'il est debout il est dans la mandorle.*

N : *et quand tu dis, dans la plupart des cas, du point de vue statistique, ça veut dire quoi ?*

M : (rires) *Alors, c'est là, dans le coup, ça veut dire que quand la mandorle est représentée seule, dans la plupart des cas, 99% de cas, le Christ va être représenté debout, dedans.*

Nous pouvons voir dans ce dernier propos ressurgir la confusion entre le niveau de confiance $1-\alpha = 0,99$ qui est une probabilité et peut être exprimée par l'expression 99%, et la proportion des occurrences, par exemple ici une représentation du Christ que apparaîtrait dans 99% d'un catégorie de supports.

4 Conclusion

Ce chapitre se situe comme nous l'avons déjà dit dans une suite de travaux que nous menons conjointement à partir de nos deux domaines de spécialités, à savoir la psychologie et la statistique. Plus particulièrement, nous nous intéressons, non pas à l'usage de l'analyse statistique implicative dans le domaine de la psychologie, mais plutôt à l'ASI en tant qu'objet d'apprentissage à la lumière de la perspective historico-culturelle du psychisme et de la médiation instrumentale et sémiotique dans ce domaine. L'accès pour le non-spécialiste à l'ASI est médiatisé par le logiciel CHIC qui offre des interfaces efficaces pour un usage accessible à celui-ci. L'approche clinique que nous avons menée ici illustre les difficultés liées à l'usage même de ce logiciel. Nous y voyons les problèmes que rencontre l'usager novice et de formation littéraire, désireux de s'approprier la méthode ASI, face à la lecture et l'interprétation des représentations graphiques produites par ce logiciel. Nous y retrouvons aussi ce que pointe Mottet (1996) pour qui « si les images n'équivalent pas à des énoncés verbaux, elles ne reproduisent pas non plus les conditions de la perception directe des choses. Ce qu'apportent aussi les images, ce sont des modes de traitement, d'analyse et de calcul, que leurs structures de représentation rendent possibles. La pluralité des représentations imagées est un moyen de soumettre la réalité à une pluralité de traitements possibles. » Nous continuons à penser que le travail sur la question de connaissances minimales requises est à poursuivre. Par exemple, la confusion entre 99% en tant que niveau de confiance et 99% en tant que proportion mérite d'être étudiée afin d'élaborer des médiations didactiques qui puissent permettre d'atteindre un niveau de conceptualisation supérieur. Cette observation clinique confirme que la forme même d'une représentation graphique peut induire des lectures erronées des propriétés des données.

Références

Acioly-Régnier, N.M (2008). Des instruments techniques aux instruments psychologiques : béquilles intellectuelles ou aides à la conceptualisation en mathématiques ? *Carrefours de l'éducation*. N° 26 – juillet-décembre 2008 : 115-128

Acioly-Régnier, N.M. et J.-C. Régnier. (2008) Identifying didactic and sociocultural obstacles to conceptualization through Statistical Implicative Analysis In R. Gras, E. Suzuki, F. Guillet et F. Spagnolo (Eds): *Statistical Implicative Analysis : Theory and Applications* Heidelberg : Springer Verlag, 347-381

Acioly-Régnier, N.M. et J.-C. Régnier. (2005) Repérage d'obstacles didactiques et socioculturels au travers de l'A.S.I. des données issues d'un questionnaire In: *Third International Conference A.S.I. : Implicative Statistic Analysis*, Palerme: *Quaderni di Ricerca en Didattica* ISSN on line1592-4424, 2005. v.1. 63 – 87

Cadet, B. (1998) *Psychologie cognitive*. Paris : In Press éditions

Carraher, D., Schliemann, A. et R. Nemirovsky (1995). Graphing from everyday experience. *Hands on*, vol. 18, n. 2, Technical Education Research Center (TERC), Cambridge, Massachusetts.

Ehrlich, M.-F. et M. Delafoy (1990) La mémoire de travail : structure, fonctionnement, capacité. *L'Année psychologique* 90, 3 : 403-427

Lemmeignan, G. et A. Weil-Barais, (1993). *Construire des concepts en physique.* Paris : Hachette education

Meira, L. et M. A. Pinheiro (2007) Produção de sentidos no uso que se faz de *gráficos. Estudos de Psicologia* 12(2) : 99-107

Monteiro, C. et J. Ainley (2004) Interpretation of media graphs and critical sense: implications for teaching and teachers. In *10th International Congress on Mathematical Education*. Copenhagen. http://wwwicme-10.dk/

Monteiro, C. (1998) *Interpretação de gráficos sobre economia veiculados pela midia impressa*. Dissertação de mestrado em psicologia cognitiva. Universidade Federal de Pernambuco. Brasil

Mottet, G. (1996) Les situations-images : une approche fonctionnelle de l'imagerie dans les apprentissages scientifiques à l'école élémentaire. Document de travail pour un article *ASTER*, n° 22

Pontalis, J.-B. (2000). *Fenêtres*. Paris : Éditions Gallimard

Régnier, J.-C. et N. M. Acioly-Régnier, (2007) Analyse cohésitive et interprétations des données dans le champ de l'éducation In: R. Gras, P. Orús, B. Pinaud, P. Gregori (Eds) *Nouveaux apports à l'Analyse Statistique Implicative et Applications dans des Disciplines Variées* Castellon (Espanha) : Universidade JAUME I, p. 329-344.

Régnier, J.-C (1998) Lire un article de journal de la presse ordinaire In J-C Girard, D. Gros, P. Planchette, J-C. Régnier, R. Thomas (Eds) *Enseigner la Statistique du CM à la Seconde. Pourquoi ? Comment ?* Villeurbanne : Irem de Lyon. p.127-133

Vergnaud, G. (2007) Qu'est-ce qu'apprendre? « Les effets des pratiques enseignantes sur les apprentissages des élèves » Colloque organisé par les IUFM du Pôle Nord-Est les 14 et 15 mars 2007 à Besançon (version courte d'une publication à venir dans un ouvrage issu du colloque)

Vergnaud, G. (1987) Chapitre de conclusion. In C. Janvier *Problems of representation in teaching and learning of mathematics*. London: Lawrence Erlbaum associates publishers

Summary

In Statistical Implicative Analysis, the software CHIC offer efficient interfaces to non-specialist for usage. The latter is confronted with reading and interpreting graphical representations produced by this software. But contrary to popular belief, "a drawing, chart speaks for itself," the interpretation of graphs requires minimum knowledge about the tool that the researcher must acquire. This work continues the discussion developed in (Acioly-Régnier & Régnier, 2005, 2008), (Régnier & Acioly-Régnier, 2007) on barriers to the conceptualization related to symbolic representations, and focuses on the difficulties encountered in research in medieval iconography. We seek to clarify them so clinical by interpreting the involving graphs. The form of a graphic representation may lead to erroneous readings of the properties of data as shown by studies on the interpretation of signs and their dependence on prior experience of subjects.

Chapitre 4 : Étude de représentations d'élèves en éducation physique et sportive

Ingrid Verscheure*, Catherine-Marie Chiocca**

* GRIAPS-LASELDI, Besançon, France
ingrid.verscheure@univ-fcomte.fr
http://ufrstaps.univ-fcomte.fr/recherche/equipes.php
** E.N.F.A. Toulouse, France
catherine-marie.chiocca@educagri.fr
http://www.enfa.fr

Résumé. Ce chapitre concerne le traitement par le logiciel CHIC d'un questionnaire proposé à des élèves de première de l'enseignement agricole français. Les questions abordent les représentations des élèves vis-à-vis des activités physiques et sportives et plus particulièrement du volley-ball. Plusieurs réseaux de variables permettant de profiler des types d'élèves apparaissent. L'étude des contributions de deux variables, sexe et genre, ajoutées aux réseaux mis en évidence, permet d'améliorer et de faciliter la sélection des élèves « représentants des réseaux »pour des futurs travaux basés sur des études de cas.

1 Introduction

Le but de notre étude est de mettre en évidence des processus différentiels susceptibles d'éclairer les inégalités sexuées en éducation physique. Nous avons choisi pour terrain l'enseignement du volley-ball dans les lycées agricoles. De nombreux travaux en sociologie et psychosociologie se sont intéressés à l'étude de la **construction sociale** des **identités de sexe** et des **comportements sexués** au sein du système scolaire. Ils montrent l'importance de l'influence des représentations sociales et fonctionnelles dans l'engagement ou non des élèves dans le processus d'apprentissage. Une des questions fondamentales de notre travail est d'éclairer en quoi les représentations de l'activité chez les filles et les garçons engendrent-elles des types différents de comportements chez les élèves ? Nous envisageons ces questions comme le résultat de processus complexes ne pouvant être réduits à des explications déterministes en terme de curriculum « masculin », de « stéréotypes professoraux », ou d'oppositions identitaires, même si tous ces facteurs peuvent avoir une influence. Nous nous sommes placées dans le cadre de l'enseignement tel qu'il se fait et non tel qu'il pourrait se faire. Une première approche des dynamiques différentielles des interactions didactiques selon le genre en Education Physique et Sportive (EPS) et notamment dans le cas de l'attaque en volley-ball (Verscheure et Amade-Escot, 2004) a cependant montré la nécessité d'un questionnement des représentations des élèves vis-à-vis des Activités Physiques et Sportives (APS) et du volley-ball.

Par l'intermédiaire d'une enquête par questionnaires, nous souhaitions explorer les relations existant entre la **variable sexe**, la **variable genre** - telle que mise en évidence par le test du BSRI (Bem Sex-Role Inventory Bem, 1974) validé pour l'éducation physique par Fontayne (Fontayne et al., 2000) - et les représentations sociales de l'attaque en volley-ball.

Après une brève présentation des notions essentielles de notre cadre théorique et suite à notre problématique, nous présentons la méthodologie utilisée et quelques résultats. Nous montrerons dans un premier temps l'apparition de réseaux implicatifs qui semblent structurer les données. Puis nous verrons que la variable sexe pèse de façon prépondérante sur les réseaux implicatifs de représentations d'élèves concernant les activités sportives par rapport à la variable genre.

2 Cadre théorique et problématique

Les conclusions des travaux en psycho-sociologie portant sur les différences sexuées en EPS, qu'ils soient français ou anglo-saxons (Davisse, 1991, Scraton, 1992, Penney, 2002), indiquent que l'éducation physique en tant que matière d'enseignement, contribue à renforcer et légitimer les importantes inégalités de sexe et renforce également de façon implicite, les attitudes et réactions qui en découlent. Finalement, certain/e/s auteur/e/s (Penney, 2002) suggèrent que les élèves développent des compétences, des habiletés, des connaissances et des perceptions différentes puisque les enseignements dispensés sont différents. De plus, à la différence entre fille ou garçon il semble se rajouter des différences de genre qui font que les élèves ne se positionnent pas de la même façon face aux savoirs qui leurs sont proposés. Nous avons donc re-problématisé la question des inégalités de sexe en EPS dans le champ de la didactique tout en considérant la question du genre, dans la mesure où nous pensons qu'elle a une importance particulière en EPS (du fait de la connotation masculine ou féminine des pratiques sociales de références de cette discipline). En effet, la caractéristique des contenus d'enseignement en EPS renvoie à des actions, des transformations de la conduite motrice liées à des savoir-faire. Bouthier et David (1989) ont montré que la prise en compte des représentations est essentielle dans l'enseignement des activités physiques et sportives scolaires. Nous cherchons à connaître les représentations des élèves de sexe et/ou de genre différent à propos de l'EPS et des sports collectifs d'une manière générale, ainsi que les représentations de la logique interne du volley-ball.

Afin d'éclairer le lecteur, il semble nécessaire de présenter rapidement, le cadre théorique de notre recherche qui est structuré autour de concepts centraux : le concept de genre, emprunté à la sociologie et à la psychologie sociale de l'éducation et les concepts de représentation (sociales et fonctionnelles).

2.1 Le sexe et la notion controversée de « genre »

La tendance actuelle dans les travaux de sociologie différentielle en éducation ainsi que dans le champ de la psychologie sociale qui abordent les dissemblances entre les filles et garçons est de substituer la notion de genre à celle de sexe (Hurtig, Kail et Rouch, 2002 ; Marro, 2000). Or, il nous semble important de bien différencier ces termes.

Généralement, le sexe est l'expression des caractéristiques biologiques qui partagent les deux populations femmes et hommes. Chaque individu naît mâle ou femelle mais devient fille ou garçon par le biais de processus socioculturels. Dans les années 70 les scientifiques

féministes nord-américaines, cherchant à aller au-delà de la dichotomie sexuée, ont développé des théories concernant la masculinité et la féminité. Elles sont définies comme « des traits relativement permanents plus ou moins enracinés dans l'anatomie, la physiologie et l'expérience précoce, et qui servent en général à distinguer, dans l'apparence, les attitudes et le comportement, les individus de sexe masculin des individus de sexe féminin » (Hurtig et Pichevin, 1986). Ainsi, Bem (1974) se focalise sur la notion d'androgynie psychologique, qui repose sur l'hypothèse qu'« il est en principe possible pour un individu d'être à la fois masculin et féminin, en fonction de la pertinence de ces diverses modalités eu égard à une situation donnée » (p. 253). Cette auteure a construit un test : le BSRI [Bem Sex Role Inventory, Bem (1974)], dont nous avons utilisé la version française (IRSB), qui cherche à mesurer les rôles sociaux sexuellement typés qui constituent des identités de genre, servant à la fois de filtres cognitifs pour interpréter les événements et orienter les conduites. Cependant, à la suite des anthropologues, il faut considérer « la construction sociale du genre sous deux aspects : comme artefact d'ordre général fondé sur la répartition sexuelle des tâches (…) ; comme artefact d'ordre particulier résultant d'une série de manipulations symboliques et concrètes portant sur des individus, cette deuxième construction s'ajoutant à la première » (Héritier 1996, p. 21). Ainsi nous pouvons considérer à la suite de ces auteur/e/s que le genre renvoie à la dynamique singulière de chaque être allant des formes (stéréotypées) les plus traditionnellement masculines à celles les plus féminines avec toute la variation possible suivant les contextes et les situations. Nous pensons donc- avec d'autres - que « l'orientation de genre » en tant que construction sociale et culturelle de « modèles intégrés de traits, de comportements, de tâches et d'activités » pouvant être endossés ou non par les individus (Marro, 2003) détermine en partie les représentations qu'ont les élèves des activités scolaires que les institutions (didactiques ou non) leur proposent. S'il peut être décrit, ce lien serait susceptible d'expliquer certains phénomènes différentiels à l'œuvre dans les situations d'enseignement-apprentissage.

Nous avons fait l'hypothèse que vis-à-vis de toute pratique sociale (dans notre cas le volley-ball scolaire) les sujets (les élèves) construisent un ensemble singulier de représentations articulant des dimensions individuelles et collectives, jouant des fonctions de connaissance, d'orientation et de justification de l'action, susceptible de varier en contexte bien qu'ayant une certaine cohérence identitaire.

2.2 Les représentations sociales et fonctionnelles

Dans notre approche nous souhaitons approfondir les liens entre sexe, genre (mesuré par l'IRSB) et représentations des APS parce que nous pensons que ces liens sont susceptibles d'expliquer les processus différentiels à l'œuvre lors des interactions didactiques en classe. Les représentations sont considérées comme des guides pour l'action puisque finalement celles-ci, dans les situations d'interactions jouent un rôle souvent plus important que les caractéristiques objectives dans les comportements adoptés par les sujets ou les groupes (Abric, 1994). Elles sont inscrites en tant que connaissances dans des cadres sociaux et sont nourries des rapports que les sujets entretiennent avec leurs pratiques (Grize, Verges et Silem, 1987). Selon nous, la notion de représentation permet d'émettre des hypothèses explicatives des conduites humaines allant de la perception à l'action et incluant la résolution de problèmes, la mémorisation ou l'apprentissage d'habiletés motrices complexes (Fouret, 1992). Elles sont « toute forme d'activité mentale par laquelle une réalité objective est transposée, consciemment ou non, dans le corpus des connaissances du sujet » (Eloi,

2000, p.23). Cet auteur pense que c'est la qualité des représentations (celles que le sujet a du milieu dans lequel il fait mais aussi celles de l'action qu'il va devoir exécuter) qui va permettre ou non d'atteindre le but fixé.

Faisant l'hypothèse que les représentations des APS (Activités Physiques et Sportives) jouent un rôle déterminant dans la façon dont les élèves (filles et garçons, en fonction de leur orientation de genre) décodent les significations des tâches qui leur sont proposées, notre ambition était d'explorer les liens entre genre et **représentations des élèves**, étape préalable à l'étude des dynamiques différentielles de construction des savoirs en situations d'enseignement apprentissage du volley-ball (que nous ne présentons pas dans cet article).

Nous avons donc considéré que les représentations différentielles du volley-ball chez les élèves filles et garçons étaient susceptibles d'être plus finement décrites en prenant en considération la notion d'orientation de genre, que nous avons opérationnalisée (dans le cadre d'une enquête extensive) au moyen de l'IRSB. Les représentations sociales ont dans notre recherche un statut de variable que nous souhaitons mettre en relation avec les variables sujets (« sexe »et « genre » des élèves) dans le but de discuter les relations entre le sexe, le genre et les représentations du volley-ball.

3 Recueil et traitement des données

Le recueil des données s'est fait par grappe. Un questionnaire a été proposé pour connaître les représentations d'élèves de classes de première (scientifique, technologique et professionnelle) de lycées agricoles publics de la région Midi-Pyrénées à propos de l'EPS, des sports collectifs, du volley-ball...

Le questionnaire utilisé est constitué de trois grandes parties : une première question, dont la formulation est décrite ci-dessous permettait de définir « le genre » des élèves à partir d'items issus du BSRI[1].

« Nous essayons de savoir ce qui, pour toi, constitue des traits marquants de ton caractère. Réponds à chacune des phrases du questionnaire ci-dessous en utilisant l'échelle qui est fournie. Entoure le 7 si tu estimes que cette phrase correspond parfaitement à ton caractère, et le 1 si la phrase ne correspond jamais à ton caractère. Sinon, coche une des autres valeurs intermédiaires ».

[1] Pour une discussion sur la pertinence du BSRI pour la détermination des attitudes de genre, cf. Verscheure, Amade-Escot et Chiocca (2006)

	Jamais vrai			Parfois vrai			Toujours vrai
je suis toujours prêt/e à écouter les autres	1	2	3	4	5	6	7
je suis doux/ce	1	2	3	4	5	6	7
j'ai l'esprit de compétition	1	2	3	4	5	6	7
je suis sensible aux peines et aux problèmes	1	2	3	4	5	6	7
j'ai des qualités de commandement	1	2	3	4	5	6	7
je suis affectueux/se	1	2	3	4	5	6	7
je suis sûr/e de moi	1	2	3	4	5	6	7
j'aime rendre service	1	2	3	4	5	6	7
je suis énergique	1	2	3	4	5	6	7
je suis attentif/ve aux besoins des autres	1	2	3	4	5	6	7
je suis dominateur/trice	1	2	3	4	5	6	7
je suis chaleureux/se	1	2	3	4	5	6	7
j'aime les enfants	1	2	3	4	5	6	7
je suis sportif/ve	1	2	3	4	5	6	7
je suis prêt/e à consoler les gens	1	2	3	4	5	6	7
je me comporte en chef	1	2	3	4	5	6	7
je suis tendre	1	2	3	4	5	6	7
j'ai confiance en moi	1	2	3	4	5	6	7

TAB. 1 *: Inventaire des Rôles de Sexe de Bem (traduction et validation pour des adolescents français par Fontayne, Sarrazin et Famose, 2000)*

Une deuxième partie du questionnaire s'intéresse à l'EPS en général. Parmi les thèmes abordés, voici quelques exemples : utilité de l'éducation physique et sportive ? Préférence pour une APS particulière ? Adhésion ou non à la pratique de l'EPS en mixité ? Préférence pour le sexe du professeur ? etc..

La troisième partie du questionnaire se focalise plus particulièrement sur les sports collectifs et surtout le volley-ball grâce à un test d'association de mots et un différenciateur sémantique inspiré d'Osgood (Osgood et al., 1957).

Le test d'association de mots concernant le volley-ball a été adapté des tests d'association libre, qui sont des productions verbales de la psychologie sociale des représentations (Abric, 1994). Cela consiste, à partir d'un mot inducteur - ici, le volley-ball - à demander aux sujets de produire tous les mots, expressions ou adjectifs qui lui viennent à l'esprit (ici sept au maximum). Le caractère spontané et la dimension projective de cette production permettent d'accéder plus rapidement et facilement que dans un entretien aux éléments qui constituent l'univers sémantique de l'objet étudié (Rouquette et Rateau 1998).

Le **test d'association de mots** était présenté de la façon suivante aux lycéen/ne/s: *« A quoi te fais penser le mot volley-ball ? Peux-tu m'indiquer les mots (entre 5 et 7) qui te viennent à l'esprit ? »*

Au total, les élèves nous ont fourni 2386 mots (soit une moyenne de 4,7 mots par élèves). En tout, 521 mots différents ont été recensés, dont certains ont très souvent été cités. Par exemple : le mot « filet » nous a été proposé 201 fois, le mot « ballon » mentionné 198

fois ; le mot « smash » écrit 175 fois ; le mot « passe » évoqué 102 fois et le mot « équipe » cité 97 fois....

Pour faciliter le traitement de ces informations, nous avons regroupé les mots en plusieurs catégories. Nous avons d'abord associé les mots de même « racine » (balle - ballon...), puis, dans un second temps, nous avons effectué des regroupements de mots qui nous semblaient avoir un lien, une proximité, en terme de signification par rapport aux questions de recherche.

Par exemple : nous avons rattaché le mot « équipe » avec les groupes de mots qui le contenait : « bonne ambiance dans l'équipe », « coéquipiers », « équipe », « équipe soudée », « être en équipe », « jeu d'équipe », « jeu en équipe »... Puis nous avons combiné les mots de ce groupe « équipe » avec les mots relevant du thème « collectif » (« apprendre à jouer ensemble », « bon esprit de groupe », « collectif », « collègues », « complicité », « confiances en ses partenaires », « entente entre joueurs », « entraide », « groupe ») pour former la catégorie que nous appellerons finalement : « aspect collectif du volley-ball ». Nous avons procédé de la même façon pour l'ensemble des 521 mots, qui ont été regroupés en vingt catégories. Pour tester la validité de nos regroupements, nous avons eu recours à la « **méthode des juges** », c'est-à-dire que deux experts de l'activité volley-ball ont examiné les mots intégrés dans nos catégories. Leur mission était de nous dire s'ils étaient d'accord ou non avec la catégorie à laquelle nous avions rattaché les mots. Ayant obtenu plus de 80% d'accord, nous n'avons que très légèrement remanié nos catégories afin d'obtenir le meilleur consensus possible.

Catégories	**n_k=**	**Catégories**	**n_k=**
Activité ludique	96	Qualités physiques	78
Aspect collectif du jeu	287	Rapport d'opposition	103
Attaque	274	Référence jeu de plage	87
Coopération	214	Règlement /limites (notamment Filet)	306
Difficile	21	Sentiments négatifs	38
Dimension sexuée	25	Sentiments positifs	49
Equipement /matériel (notamment Ballon)	258	Tactique	99
Mouvement /énergie	44	Technique	210
Peur /douleur	55	Vers le haut	42
Qualités mentales	44	Non codés	24

TAB. 2 *: Catégorisation des « mots » associés au volley-ball et occurrences* n_k (k=1 à 20)

Le **différenciateur sémantique** d'Osgood est un modèle théorique du processus d'impression sémantique Jodelet (1972). C'est une méthode d'analyse quantitative des connotations. La connotation est la définition intensive d'un mot. Par exemple : le corbeau évoque le noir, le mauvais présage et une série de significations implicites ou explicites (Menahem, 1974). Cette méthode consiste à associer un mot à des couples d'adjectifs opposés et représentatifs (Menahem, 1968). Elle est un moyen général de recherche des significations (mots, figures, ...) et le choix des couples d'adjectifs doit être adapté à chaque cas, ici le volley-ball.

Nous nous sommes basées sur les travaux de David (1995) qui utilise un différenciateur sémantique adapté au rugby pour la mise en évidence des aspects différentiels des

représentations sociales de ce sport collectif chez un public mixte en EPS. Ici, nous posions la question suivante :

« Voici une série de termes relatifs au volley-ball. Selon que ce qu'ils évoquent pour toi, entoure le chiffre de chaque ligne le plus proche du pôle qui traduit ce que tu penses ».

Renvoyer	3	2	1	0	1	2	3	Attaquer
Faire des progrès	3	2	1	0	1	2	3	Devenir sportif/ve
Marquer son point	3	2	1	0	1	2	3	Jouer collectif
Faire durer l'échange	3	2	1	0	1	2	3	Rompre l'échange
Devenir fort/e	3	2	1	0	1	2	3	Devenir malin/e
Etre le/la meilleur/e	3	2	1	0	1	2	3	Apprendre à se maîtriser
Regarder le ballon	3	2	1	0	1	2	3	Regarder l'adversaire
Rude	3	2	1	0	1	2	3	Doux
Combatif/ve	3	2	1	0	1	2	3	Ne pas se faire mal
Jouer	3	2	1	0	1	2	3	Gagner
Etre un(e) champion(ne)	3	2	1	0	1	2	3	Se sentir bien
Statique	3	2	1	0	1	2	3	Mobile
Progresser	3	2	1	0	1	2	3	Se détendre
Entraînement	3	2	1	0	1	2	3	Match
Précision	3	2	1	0	1	2	3	Force
Assurer	3	2	1	0	1	2	3	Risquer
Rupture	3	2	1	0	1	2	3	Continuité

TAB. 3 *: Le différenciateur sémantique inspiré d'Osgood* (Osgood et al., 1957).

Les couples d'échelles bipolaires ont été construits sur la base d'adjectifs antonymes puisés dans des répertoires de mots. Le différenciateur sémantique est considéré comme une échelle d'attitude prédictive d'un comportement. Il favorise par ailleurs la construction d'une image assez lisible des représentations des groupes de sujets, notamment fonctionnelles, aspect qui nous intéresse tout particulièrement à propos du volley-ball.

Ces deux techniques empruntées aux méthodologies en usage dans le champ de l'exploration des représentations des élèves, ont été retenues pour leur aspect complémentaire. Le test d'association de mots permet d'avoir l'aspect plutôt « figé » des représentations. Le différenciateur sémantique est lui basé sur les conceptions tactiques et dynamiques du volleyball.

4 Quelques résultats

Deux traitements statistiques ont été mis en œuvre pour explorer les liens éventuels entre le sexe, le genre mesuré par l'IRSB[2] et les représentations sociales.

[2] Pour une meilleure lisibilité, nous écrirons « genre/IRSB » à chaque fois que nous parlerons du genre tel que mesuré par l'IRSB

Un premier temps du traitement des résultats a consisté à décrire la structure des données à partir d'une Classification Ascendante Hiérarchique (CAH) afin de construire une typologie des représentations (mots associés, différenciateur sémantique) et d'identifier le poids des modalités de la variable sexe et de la variable genre/IRSB dans la définition de ces classes.

Puis, dans un second temps, nous avons cherché à saisir les implications statistiques entre les différentes modalités des variables sujets (sexe et genre/IRSB) et les structures représentationnelles en utilisant le logiciel d'analyse implicative CHIC.

4.1 Caractéristiques de la population étudiée

Les questionnaires ont été envoyés fin 2002 à 9 Lycées d'Enseignement Général et Technologique Agricole publics de la région Midi-Pyrénées. Ils ont été distribués en fin de cours par les enseignant/e/s d'EPS eux-mêmes à tous leurs élèves de classes de première, quelle que soit la filière d'enseignement. 507 questionnaires exploitables nous ont été retournés.

Notre échantillon présente une fréquence importante d'élèves préparant un baccalauréat technologique (65,7%). Quelques élèves préparent un baccalauréat scientifique (18,7%) ou un baccalauréat professionnel (15,6%). La proportion de filles est plus faible (33, 5%, soit 170 élèves) que la proportion de garçons (66,5% soit 337 élèves) mais est sensiblement identique à la moyenne nationale (33,9% en 2000).

La répartition des élèves selon les différentes modalités de la variable genre a été effectuée selon la méthode de l'IRSB. Cette méthode repose sur un partage par la médiane (median split method) et permet d'établir quatre modalités de la variable genre : androgyne (A), non différencié (ND), féminin (F), masculin (M). Rappelons que les variables genre et sexe sont indépendantes, et. qu'une fille peut être de genre masculin ou de genre androgyne, et un garçon de genre féminin ou androgyne. La répartition des élèves selon le genre est assez équilibrée : 26% d'élèves de genre non-différencié, 23% d'élèves de genre féminin, 25% d'élèves de genre masculin et 26% d'élèves de genre androgyne.

4.2 Résultats de la CAH : quatre classes

Par l'intermédiaire du logiciel SPSS® nous avons réalisé des Classifications Hiérarchiques par méthode Ascendante (CAH) et une Analyse Factorielle des Correspondances (AFC) à propos du volley-ball avec nos 507 questionnaires (Verscheure, 2005). Nous ne présentons que l'essentiel des résultats ici.

L'analyse des données par la CAH nous a permis de regrouper les réponses des individus en quatre classes. Chaque classe est représentée par un certain nombre de modalités du différenciateur sémantique et des mots concernant le volley-ball.

Les profils représentationnels de notre population se répartissent en quatre classes.

La classe 1 est représentée par « Des garçons fidèles aux stéréotypes sexués « masculins » identifiés par la littérature à propos du volley-ball (Davisse et Louveau, 1991 ; Tanguy, 1992) c'est à dire qu'ils cherchent à attaquer, sans être forcément dans de bonnes positions pour le faire ».

La classe n°2 regroupe « Des garçons qui s'inscrivent également dans le rapport de force (ils parlent notamment d'être combatif) mais intègrent derrière cette idée la notion de jeu collectif ».

La classe 3 rassemble « des élèves cherchant à progresser pour jouer avec des partenaires et attaquer le camp adverse sans forcément taper fort (idée de précision, de progrès, de plaisir) dans une optique collective ».

La classe 4 regroupe « Des élèves, plutôt des filles (51%), qui cherchent à jouer et sont dans une logique de progrès technique pour ce faire ». Il y a une négation de la notion de rapport de force entre adversaires. Ce profil s'apparente plutôt aux stéréotypes sexués des filles concernant leur rapport au volley-ball (Davisse et Louveau, 1991 ; Tanguy, 1992).

Ainsi, seule la variable sexe paraît significative de certaines des classes (classe 1, 2 et 4) et il y a un écrasement de la variable genre/IRSB. Convaincues cependant qu'il doit y avoir l'influence sous-jacente du genre sur les représentations du volley-ball, nous nous sommes tournées vers l'analyse implicative.

4.3 Résultats de l'analyse implicative : trois réseaux de variables au seuil 0,70

Par l'intermédiaire du logiciel CHIC (Gras et al., 1996), nous avons traité les réponses au questionnaire qui concernaient les représentations des sports collectifs et du volley-ball. Selon Bailleul (2000) « *l'analyse statistique implicative est un outil particulièrement puissant pour travailler sur les représentations et mettre en évidence leurs structures organisatrices* ». En effet, cela semble pertinent pour notre recherche puisque le logiciel CHIC réalise des analyses implicatives. Le graphe implicatif des variables conduit à l'identification de « **réseaux** » de réponses, eux-mêmes constitués de « chemins ». Ces derniers ont un sens proche, voire identique afin d'interpréter ce que l'on observe.

Afin d'étudier les contributions des modalités des variables sexe et genre aux représentations du volley-ball, nous avons codé ces dernières en « variables supplémentaires ».

Nous menons l'analyse au seuil 0.70, où apparaissent trois réseaux relativement bien distincts (que nous avons appelé A, B et C) dans lesquelles les variables sexe et genre/IRSB jouent un rôle même si elles n'expliquent pas chacun des chemins constituant le réseau.

Pour ne pas surcharger ce chapitre en schémas, nous décrivons dans un premier temps les caractéristiques principales des réseaux obtenus au seuil 0,70. Dans un second temps, nous reprenons ces trois réseaux en présentant les graphes implicatifs, et en désignant la plus forte contribution des modalités de la variable sexe et la plus forte contribution des modalités de la variable genre/IRSB sur les chemins.

4.3.1 Réseau A

Un premier réseau regroupe différents chemins qui impliquent les variables « jouer collectif »[3] et « j'aime les sports collectifs ». Il apparaît que ce réseau se caractérise par des variable « implicantes » (qu'elles proviennent du différenciateur sémantique ou du test d'association de mots) renvoyant toutes aux caractéristiques féminines du volley-ball énoncées par la littérature (Davisse,1991 ; Tanguy ,1992), telles que « renvoyer », « ne pas se faire mal », « s'entraîner », « assurer ».

[3] Variable issue du différenciateur sémantique d'Osgood

Une certaine dimension du travail (« s'entraîner »), des gestes techniques (« renvoyer »), dans certaines limites (« ne pas se faire mal ») contribuent à développer un savoir « jouer » qui, combiné avec des préoccupations qu'on pourrait presque qualifier « d'hygiéniste » (« doux », « se détendre », « activité ludique ») me permet de « me sentir bien ». « Me sentir bien », conjointement à la condition de « faire des progrès » (deuxième référence au travail dans ce réseau), dont les éléments de « maîtrise » qui me permettent de « jouer collectif » et de me sentir bien dans le jeu collectif (« j'aime les sports collectifs »). Il apparaît que cette représentation du volley-ball, voire des sports collectifs, semble équilibrée puisqu'elle fait une part à la dimension du travail et une part à la dimension du plaisir, tout en excluant la dimension compétition.

Dans ce réseau, on retrouve principalement les idées de progrès, de jeu collectif et de plaisir dans l'activité. Ces trois idées constituent, selon nous, trois des dimensions essentielles de la logique interne du volley-ball : le jeu en équipe, la progression et le côté ludique ; mais ignore la notion de rapport de force.

4.3.2 Réseau B

Au seuil de 0.70 un 2ème réseau regroupe différents chemins qui impliquent la variable « j'aime les sports collectifs ». Les variables telles que « jouer vers le haut », « j'aime le volley-ball », « qualités mentales », « jouer collectif », tout comme celles relevant de « sentiments positifs », « activité ludique » impliquent « j'aime le volley-ball » impliquent toutes « j'aime les sports collectifs ».

Avec la présence des thèmes « sentiments positifs », « activité ludique » qui impliquent « j'aime le volley-ball » et « jeu collectif », nous pouvons faire l'hypothèse que les représentations qui structurent ce réseau de réponses sont des attitudes positives à l'égard de cette activité.

Ceci rejoint les résultats observés lors des niveaux d'analyses précédents à savoir que les élèves apprécient cette discipline car ils la considèrent comme un lieu de défoulement. Le volley-ball semble se rattacher aussi à la valorisation des qualités physiques [cette catégorie compte notamment des mots tels que « grand », « sauter haut »] ainsi qu'à la mobilité. D'autre part, le jeu semble être représenté par la « continuité » (confirmé par la présence des mots associés au jeu « vers le haut ») et la tactique. Ce réseau agrège des représentations associées à la pratique valorisée du volley-ball : jeu collectif et tactique dans lequel les qualités physiques et mentales jouent un rôle important sans oublier le côté ludique.

4.3.3 Réseau C

Dans ce troisième réseau, la représentation du jeu en tant que rapport de force collectif (« match », « gagner », « être combatif », « jouer collectif ») se caractérisant par la rupture de l'échange en force (« attaque », « rude ») ou par la ruse (« devenir malin », « risquer »), est prédominante. De plus, la représentation de la victoire de ce qu'il faut mettre en œuvre pour gagner (collectivement) est prégnante dans ce réseau. De façon sous-jacente il se dégage dans ce réseau une représentation du volley-ball en tant que sport de compétition structuré autour du gain du match.

4.3.4 Conclusion provisoire sur ces trois réseaux

Pour conclure sur ces 3 réseaux ; il s'avère que dans le réseau A, la représentation des élèves semble être que pour progresser et se faire plaisir au volley-ball il faut s'entraîner. On retrouve principalement les idées de progrès, de jeu collectif et de plaisir dans l'activité. Le réseau B, lui, met en avant la dimension collective du volley-ball et il semble que les élèves pensent que jouer collectivement au volley-ball nécessite des tactiques et des qualités (mentales et physiques). Dans le réseau C, la représentation principale qui émerge est que le volley-ball est une activité collective, où il faut chercher à gagner, principalement en attaquant (soit en force soit par une feinte).

Ces trois réseaux nous paraissent structurer l'ensemble des données recueillies et notamment les relations entre les différentes variables des questions à propos des sports collectifs, du différenciateur sémantique et du test d'association de mots à propos du volley-ball. Nous allons voir maintenant quelles sont, des différentes modalités des variables de sexe ou de genre, celles qui ont les implications les plus fortes avec ces réseaux.

4.4 Contribution des variables supplémentaires sexe et genre aux réseaux mis en évidence au seuil 0,70

Nous avons calculé pour chaque chemin, la contribution des variables supplémentaires : sexe : 2 modalités, (fille et garçon) et genre : 4 modalités (androgyne, non différencié, féminin et masculin). Ces variables supplémentaires influent différemment sur la constitution de certains chemins dans les différents réseaux.

Nous considérons que la contribution d'une variable supplémentaire est à prendre en compte dans nos commentaires, en tant qu'élément de réponse à nos hypothèses, à condition que le risque de se tromper soit inférieur à 0,10 (condition classique pour dire qu'une variable est significativement explicative de tel ou tel phénomène). Nous rappelons que plus le risque statistique est faible, plus la confiance est fiable pour énoncer la contribution la plus forte des variables supplémentaires.

Dans la suite des résultats et pour faciliter la lecture des contributions des variables sujets aux trois réseaux, nous indiquons sur les schémas la valeur de la contribution significative la plus haute. Nous indiquerons dans le texte les autres valeurs identifiées en précisant les chemins auxquels elles contribuent.

4.4.1 Contribution des variables supplémentaires au réseau A

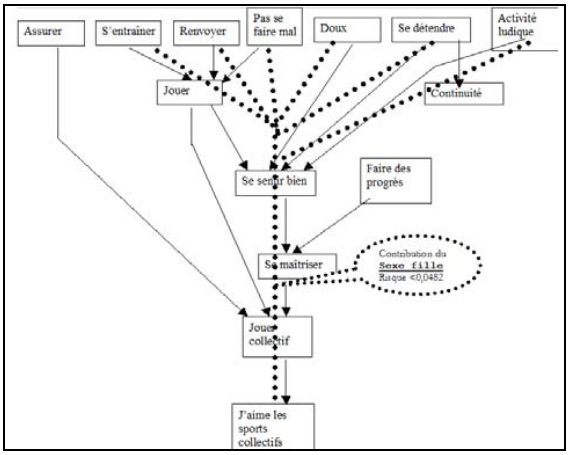


FIG. 1 – *Réseau A.*

Si l'on s'intéresse aux chemins se finissant par : « jouer », « se sentir bien », « se maîtriser », « jouer collectif » qui impliquent la variable « j'aime les sports collectifs » ; on s'aperçoit que, quelle que soit la variable « en amont », (renvoyer, s'entraîner, se détendre, ne pas se faire mal, doux), la modalité fille de la variable supplémentaire sexe contribue significativement à ces chemins (risques : 0,0438 ; 0,0137 ; 0,042). Sur le chemin « activité ludique - se sentir bien -se maîtriser-jouer collectif » c'est encore la modalité fille de la variable sexe qui contribue significativement avec un risque de 0,0482.

Le fait que la modalité fille de la variable sexe soit caractéristique de cet ensemble de chemins permet de faire l'hypothèse que ce sont les filles qui se représentent le volley-ball comme une activité physique et sportive où l'important est le jeu en équipe, la progression et le côté ludique. Ce réseau n'est pas sans évoquer certaines classes du précédent plan d'analyse notamment le profil représentationnel de la classe 3 intitulée « des élèves qui veulent progresser pour jouer et attaquer collectivement », ainsi que celui de la classe 4 à savoir « des élèves, dont une majorité de filles, qui cherchent à maîtriser le ballon pour le renvoyer ». On pourrait donc faire l'hypothèse d'un caractère **prédicteur** du « sexe fille » par rapport aux représentations du volley-ball en tant qu'activité collective et une obligation de maîtrise pour progresser. Ainsi, il semble que les représentations du réseau A se structurent autour de l'idée qu'il faut d'abord s'entraîner, progresser pour se faire plaisir en volley-ball. Cela correspondrait plutôt aux caractéristiques féminines de pratique du volley-ball décrites dans la littérature (Davisse, 1991, Tanguy, 1992).

4.4.2 Contribution des variables supplémentaires au réseau B

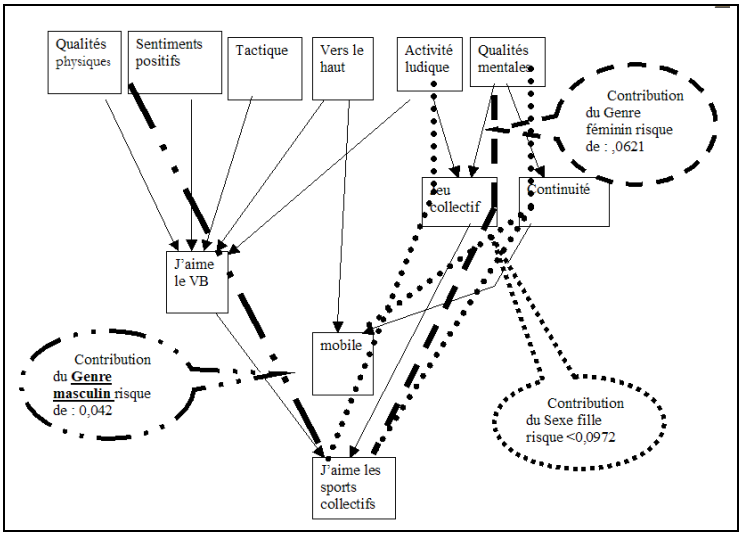


FIG. 2 – *Réseau B*

Dans ce réseau, la modalité fille de la variable sexe est caractéristique de plusieurs chemins :

- sur le chemin « activité ludique - jeu collectif - j'aime les sports collectifs » la modalité fille de la variable sexe contribue significativement avec un risque de : 0,0891.
- sur le chemin « activité ludique- j'aime le volley-ball », la modalité fille de la variable sexe contribue significativement avec un risque de : 0,0764.
- sur le chemin « qualités mentales - continuité - j'aime les sports collectifs » la modalité fille de la variable sexe contribue significativement avec un risque de : 0,0863.
- sur le chemin « qualités mentales - continuité – mobile » la modalité fille de la variable sexe contribue significativement avec un risque de : 0,0972.

Cependant deux modalités de la variable genre contribuent significativement en tant que variables supplémentaires sur deux chemins.

- sur le chemin « qualités mentales - jeu collectif - j'aime les sports collectifs » la modalité féminin de la variable genre contribue significativement avec un risque de : 0,0621.
- sur le chemin « qualités physiques -j'aime le volley-ball - j'aime les sports collectifs » la modalité masculin de la variable genre contribue significativement avec un risque de : 0,0983.

La représentation de ce réseau, qui regroupe les représentations du volley-ball en tant qu'activité ludique collective, où il est nécessaire d'avoir des qualités physiques et mentales, semble plutôt apparentée au « sexe fille ». Et l'analyse implicative semble appuyer cette représentation du volley-ball comme une activité ludique.

Cependant, il faut y apporter une nuance en fonction du genre (mesuré par l'IRSB) car le « genre féminin » contribue le plus lorsqu'on s'attache aux « qualités mentales » (rappelons qu'un garçon peut être de genre féminin). Tandis que concernant les « qualités physiques » c'est le « genre masculin » qui contribue davantage, ce qui laisserait penser que les représentations du volley-ball nécessitant des qualités physiques seraient rattachées au genre masculin. Ceci n'est pas sans relation avec les descriptions les plus classiques des identités de genre et des rôles de sexe en matière de pratique des APS (Davisse et Louveau, 1991 ; Fontayne, Sarrazin et Famose, 2001 ; Vigneron, 2004). Notre idée est qu'il s'agit d'un réseau plutôt mixte dans la constitution ou dans les représentations.

Nous pouvons émettre l'idée d'une différenciation entre ces deux modalités du genre/IRSB relative à l'activation des représentations du volley-ball comme jeu collectif sollicitant une activité tactique pour les un/e/s construites par le biais de qualités physiques pour les autres par le biais de qualités mentales au sein d'un réseau mixte.

4.4.3 Contribution des variables supplémentaires au réseau C

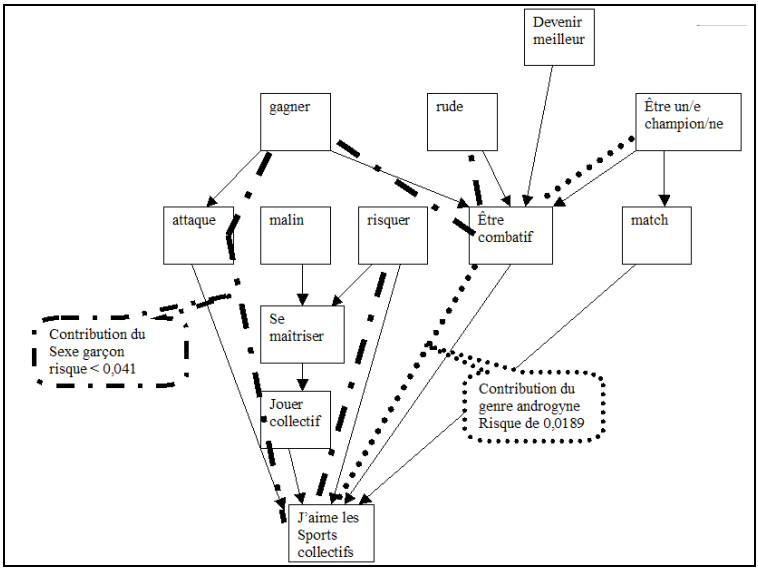


FIG. 3 – *Réseau C*

Ce réseau évoque la représentation du volley-ball en tant que rapport d'opposition collectif, où l'intérêt majeur est de faire des matchs et de se mesurer à l'adversaire. Selon ce réseau, le volley-ball c'est jouer collectif, attaquer et faire des matchs.

La modalité garçon de la variable sexe contribue significativement à plusieurs chemins de ce réseau :

- le chemin « gagner -combatif - j'aime les sports collectifs » avec un risque de 0,0184.
- le chemin « rude - combatif- j'aime les sports collectifs » avec un risque de : 0,0449.
- le chemin « gagner- attaquer- j'aime les sports collectifs » avec un risque de : 0,041.

- le chemin « risquer- j'aime les sports collectifs » avec un risque de : 0,00572.

Cependant, la modalité androgyne de la variable genre/IRSB contribue significativement sur les chemins suivants :

- « être un/e champion/ne - être combatif- j'aime les sports collectifs » avec un risque de : 0,0189.

- « être un/e champion/ne - match - j'aime les sports collectifs » avec un risque de : 0,0246.

Ce réseau s'apparente plutôt aux représentations d'un rapport d'opposition en volley-ball. La modalité garçon de la variable supplémentaire sexe contribue significativement à plusieurs chemins et la modalité androgyne de la variable genre à un chemin.

Ce serait donc plutôt les garçons qui se représenteraient le volley-ball comme un sport d'opposition permanente (d'où l'hypothèse du caractère prédicteur de l'influence du sexe garçon sur les représentations de cette activités sportive comme un rapport d'opposition). La représentation d'adversité (envers les adversaires mais aussi pour soi-même) émerge dans ce réseau sous deux formes : soit de façon agressive (attaque, rude) ; soit de façon maligne (malin, se maîtriser, risquer). On pourrait évoquer l'idée d'un réseau plus mâle que les deux réseaux précédents.

4.4.4 Conclusion sur ces trois réseaux

Contrairement aux résultats attendus, c'est la variable sexe qui apparaît le plus souvent en tant que variable supplémentaire comme contribuant significativement aux différents chemins mis en évidence. Seules les modalités de genre « féminin », « androgyne » et « masculin » contribuent chacune significativement à un seul chemin. La modalité « non différencié » de cette variable ne contribue significativement à aucun chemin. Les deux modalités filles et garçons de la variable sexe interviennent davantage. Finalement, il ressort que les modalités de la variable genre telles que construites par l'IRSB sont très marginalement liées aux représentations que les élèves de notre population ont du volley-ball scolaire.

Poussant plus loin l'analyse, nous pouvons conjecturer que les élèves appartenant à la classe 4 (des élèves, dont une majorité de filles, qui cherchent à maîtriser le ballon pour le renvoyer) doivent se trouver dans les schémas d'implication du réseau A dans lequel la représentation des élèves semble être que « pour progresser et se faire plaisir au volley-ball, il faut s'entraîner ». Mais dans le même temps nous pensons pouvoir identifier des connivences entre ce réseau A et les représentations de la classe 3 (des élèves qui veulent progresser pour jouer et attaquer collectivement). D'autant plus que nous avons vu que la variable supplémentaire qui contribue le plus au réseau A est le « sexe fille ». La représentation du volley-ball, renvoyant à la nécessité de progresser, de s'entraîner avant d'avoir la possibilité de se faire plaisir serait ainsi plutôt l'apanage des filles.

D'autre part, le réseau B met en avant la dimension collective du volley-ball, la nécessité de mettre en place des tactiques et posséder certaines qualités (mentales et physiques). Ce réseau que nous résumons par « jouer collectivement au volley-ball nécessite des tactiques et des qualités mentales » est plutôt apparenté au « sexe fille », mais il existe une implication du genre/IRSB féminin lorsqu'il s'agit de répondre qu'il faut des « qualités mentales » et du genre masculin lorsqu'il s'agit des « qualités physiques ». Cette explication d'un primat de la variable de sexe vient cependant buter sur la mise en évidence lors des analyses statistiques d'une classe représentationnelle non spécifiée par le sexe (classe 3) et

d'un réseau implicatif plutôt mixte (réseau B). Ces constats évoquent quant à eux des représentations de la classe 2 intitulée « des garçons qui recherchent un rapport d'opposition dans un jeu collectif » mais dont on a vu que le pourcentage de garçons est aussi important que dans la classe 1.

Le réseau C s'apparente plutôt aux représentations d'un volley-ball en tant que rapport d'opposition. Il y a la prégnance des matchs, l'envie de gagner, d'attaquer et de jouer collectif. La variable supplémentaire qui contribue le plus à ce réseau est le « sexe garçon » avec une contribution du genre androgyne sur le chemin de la « combativité ». Nous pensons identifier des concordances entre ce dernier réseau et certains éléments représentationnels de la classe 1 (des garçons qui cherchent à rompre l'échange à tout prix) et de la classe 2 (des garçons qui recherchent le rapport d'opposition dans un jeu collectif).

5 Discussion et conclusion générale

Au regard de ces résultats, on peut dire que la variable sexe apparaît plus souvent que la variable genre en tant que variable supplémentaire contribuant significativement à ces chemins. Grâce au logiciel CHIC, l'analyse fine permet de penser qu'il existerait des représentations sexuées des élèves concernant le volley-ball, mais il faut y apporter parfois cependant des modulations en fonction du genre. Cette analyse permet également de prédire des effets du sexe sur certains types de représentations à propos du volley-ball. Cependant, les différents constats effectués nous invitent à ne pas écarter a priori l'idée que les stéréotypes de sexe sur les sports collectifs et le volley-ball sont suffisamment partagés et vivaces pour écraser toutes les variations possibles mais marginales, susceptibles d'expliquer la diversité des représentations.

Ceci nous a conduites à discuter dans d'autres travaux (Verscheure et al., 2006) la validité du test de BSRI comme moyen d'accéder aux attitudes de genre des élèves par rapport à l'EPS. Il s'agissait après catégorisation des élèves selon leurs représentations, d'étudier les contributions des variables sexe et genre aux profils mis en évidence lors d'une classification ascendante hiérarchique (Verscheure et al., 2006) pour ensuite décrire et analyser la différentiation des interactions des enseignants avec certains représentants ou représentantes des profils d'élèves. Dans d'autres travaux non exposés ici, nous avons ainsi construit des typologies d'élèves selon le sexe, qu'il faut moduler parfois selon le genre. Ces résultats semblent permettre d'affirmer que l'orientation de genre doit être reconstruite en contextes et que suivant ces derniers, diverses dynamiques sont activées, marquant ainsi des « positions de genre » (Verscheure., 2007) susceptibles de varier au fil des interactions et des situations éducatives.

Références

Abric, J.-C. (1994). Pratiques sociales et représentations. PUF, Paris.

Bailleul, M. (2000). *Mise en évidence de réseaux orientés de représentations dans deux etudes concernant des enseignants stagiaires en IUFM*. In Actes des journées sur la fouille dans les données par la méthode d'analyse statistique implicative.

Bem, S. (1974). The measurement of psychological androgyny. Journal of Consulting and Clinical Psychology vol. 42, 2, 155–162.

Bouthier, D. et B. David (1989). Représentation et action : de la représentation initiale à la représentation fonctionnelle des APS en EPS. Méthodologie et didactique de l'éducation physique et sportive Ed. G. Bui-Xuan, 233–249.

David, B. (1995). *Rugby mixte en milieu scolaire.* Revue Française de Pédagogie. 110 : 51–61.

Davisse, A. (1991). *Sport, école et société : la part des femmes*. Paris Ed. Actio, 174–263.

Eloi, S. (2000). Représentations mentales et acquisition d'habiletés tactiques en volley-ball : effets comparés de deux cycles d'enseignement. *Sciences et motricité*, 40. 22-31

Fontayne, P., P. Sarrazin, J-P. Famose (2000). The *Bem sex-role inventory : validation of a shortversion for french teenagers*. European Review of Applied Psychology 50, 4: 405–416.

Fouret, J-L. (1992). Représentation, connaissance et action. in D., Bouthier, &P., Griffet, *Représentation et action en activités physiques et sportives*. Actes de la journée du 15 mai 1992. 37-171.

Gras, R., S. Almouloud, M. Bailleul, A. Larher, H. Ratsimba-Rajohn et M. Polo (1996). *L'implication statistique. Nouvelle méthode exploratoire de données*. La Pensée Sauvage.

Grize, J-B., Vergès, P., Silem, A. (1987). Salariés face aux nouvelles technologies : vers une nouvelle approche sociologique des représentations sociales. Paris. CNRS.

Heritier, F. (1996). Masculin / Féminin. La pensée de la différence. Paris. Odile Jacob.

Hurtig, M.C. et Pichevin, M.F. (1986). La différence des sexes : questions de psychologie. Paris : Tierce Sciences

Hurtig, M.C., Kail, M.,Rouch, H. (2002). Sexe et genre : de la hiérarchie entre les sexes (textes réunis à partir du colloque du C.N.R.S. « Sexe et genre » organisé par M. Bordeaux) Paris : CNRS.

Jodelet (1972). *L'association verbale*. in P. Fraisse et J. Piaget, Traité de psychologie espérimentale. VIII:97–153.

Marro, C. (2000). Différence des sexes et tolérance à la transgression des rôles de sexe. In Vouillot, F. Filles et garçons à l'école : une égalité à construire. CNDP : Autrement dit. p.123-129.

Marro, C. (2003). Se qualifier de "fille féminine" ou de "garçon masculin" à l'adolescence. Pratiques psychologiques, 9 (3), 5-20.

Menahem, R. (1968). *Le différenciateur sémantique, le modèle de mesure.* L'année psychologique 68: 451–465.

Osgood, C., G. Suci, et P. Tannenbaum (1957). *The mesurement of meanings*. Chicago, University of Illinois Press.

Penney, D. (2002). *Gender and Physical Education*, London, Routledge.

Rouquette, M.-L. et P. Rateau (1998). *Introduction à l'étude des représentations sociales*. Grenoble, PUG.

Scraton, S. (1992). Shaping Up to Womanhood : Gender and girl's physical education. Milton Keynes. Open University Press.

Tanguy (1992). *Le volley: un exemple de mise en œuvre didactique. Echanges et controverses.* 4 : 7–20.

Verscheure, I., C. Amade-Escot, (2004). *Dynamiques différentielles des interactions didactiques selon le genre en EPS. Le cas de l'attaque en volley-ball en seconde*. Revue STAPS, numéro spécial « filles et garçons en EPS ». 66 : 79-97.

Verscheure, I. (2005). Dynamique différentielle des interactions didactiques et co-construction de la différence des sexes en Education Physique et Sportive. Le cas de l'attaque en volley-ball en lycées agricoles. Thèse non publiée. C.L.E.S.C.O. - LEMME, Université Paul Sabatier, Toulouse III.

Verscheure, I., C. Amade-Escot, et C.-M. Chiocca (2006). Représentations du volley-ball scolaire et genre des élèves : pertinence de l'inventaire des rôles de sexe de Bem? in Revue Française de Pédagogie. 154 : 125-144

Verscheure, I., C. Amade-Escot, (2007). *Diriger l'étude d'élèves hétérogènes*. in C. Amade-Escot (Coord) Le didactique. Paris : Editions de la revue EPS, collection « Pour l'action » dirigée par M. Durand.

Summary

This chapter relates a questionnaire treatment by CHIC software. This survey was proposed to pupils of last high school year in French agricultural teaching. Questions approached representations of pupils about physical and sporting activities, and, more particularly about volley-ball. Several networks of variables appear which make it possible to profile kinds of pupils.Study of contributions of 2 additional variables, sex and gender, highlighted networks, makes it possible to improve choices of students' representatives' networks for later works based on case of studies.

Chapitre 5 : Un exemple d'analyse implicative en neuro-psychologie : la comparaison de groupes contrastés.

Tarek Bellaj*, Daniel Pasquier**

*Université de Tunis
Faculté des Sciences Humaines et Sociales de Tunis
Unité de Recherche « Psychopathologie Clinique » (URPC), Tunis
tarekbellaj@gmail.com
**15, rue du Grand Carré F-45800 Saint-Jean de Braye
Unité de Recherche « Psychopathologie Clinique » (URPC), Tunis
dpasquier@avenirentreprise.fr
http://www.avenirentreprise.fr

Résumé. La psychométrie classique utilise des indices de liaisons entre variables qui sont symétriques. Par exemple, la corrélation entre a et b sera la même que celle entre b et a, chacune des variables pouvant prendre indifféremment le statut de variable expliquée ou de variable explicative : il est impossible d'ordonner des corrélations entre variables en séquence implicative. L'analyse implicative des données offre une possibilité d'atteindre cet objectif d'ordonnancement séquentiel des variables. Dans ce chapitre, nous présenterons un exemple de recours à ce type d'analyse pour comparer des groupes contrastés. L'intérêt de ce texte ne se situe pas dans les résultats pour eux-mêmes, mais dans les perspectives ouvertes par la démarche utilisée.

1 Introduction

Un lien symétrique ne peut pas être un lien de causalité. Il est impossible d'ordonner des corrélations entre variables en séquence implicative. Pour atteindre cet objectif il convient de faire appel à des indices de liaisons qui ne soient pas symétriques comme ceux de l'A.S.I.. L'exemple portant sur la comparaison de groupes contrastés est une étude neuropsychologique comparant les performances d'un groupe de sportifs commotionnés à un groupe de sportifs de contrôle. Le choix de cet exemple revient à ce qu'offrent les paradigmes neuropsychologiques en termes d'**associations**, de **dissociations** et de **causalité**. En fait, en neuropsychologie de l'adulte, il s'agit généralement de troubles acquis qui s'inscrivent dans le temps par rapport à un ou des événements lésionnels particuliers (traumatique, vasculaire, tumoral, dégénératif, inflammatoire etc.). Parmi ces événements lésionnels, l'étiologie traumatique a été choisie du fait qu'elle s'inscrit le mieux dans le temps et du fait que l'événement traumatique permet de distinguer une période et des caractéristiques personnelles pré-morbides et une période et des changements personnels cognitifs et émotionnels post lésionnel évaluables ici et maintenant. La quasi-totalité des travaux scientifiques s'accorde sur l'existence d'une relation temporelle linéaire entre

événement lésionnel, lésion ou altération fonctionnelle et troubles neuropsychologiques. Or, une analyse implicative plus dynamique peut apporter de nouveaux éclairages sur la question. Des difficultés neuropsychologiques infra-cliniques pré-morbides peuvent précipiter l'événement lésionnel traumatique. Les performances évaluées ici et maintenant ne reflèteraient par uniquement les conséquences de la lésion mais permettraient de développer des inférences quant à leur valeur implicite et relative comme facteur participant à l'avènement traumatique.

L'intérêt de ce texte ne se situe pas dans les résultats pour eux-mêmes, mais dans les perspectives ouvertes par la démarche utilisée.

2 Méthode

Après une présentation rapide de la population, le choix des variables du réseau nomologique sera explicité. On passera ensuite à la présentation de l'analyse implicative et des possibilités du logiciel utilisé, CHIC - classification hiérarchique implicative et cohésitive - (Couturier et Gras, 2005).

2.1 Population d'étude

La population est composée de 40 sujets sportifs de haut niveau distribués dans deux groupes contrastés, 20 sportifs commotionnés au cours de leur activité sportive composant le groupe expérimental - Gexpé -, et 20 sportifs indemnes de tout traumatisme cérébral composant le groupe témoin - Gtém -. Tous ces sportifs étaient consentants et informés de toutes les modalités de notre recherche (objet de l'étude, sa finalité, le déroulement du testage, etc.).

2.1.1 Le groupe expérimental

Les 20 sportifs de l'élite tunisienne ayant subi une commotion cérébrale complexe sont répartis comme suit : 10 boxeuses, 3 boxeurs et 7 joueurs de rugby. Leur âge moyen est de 20,3 ans (σ = 2,07). 35% de ces sportifs ont un niveau de scolarité supérieur, 65% un niveau secondaire. Tous ont participé à un entretien portant sur l'historique de la commotion cérébrale (nombre, mécanisme, etc.) et ont subi l'administration de l'échelle des symptômes post-commotionnels (McCrory et *al*., 2005). Ne sont retenus dans ce groupe que les sportifs répondant aux critères d'inclusion et d'exclusion suivants :

Les critères d'inclusion dans le Gexpé sont les suivants : avoir subi une commotion cérébrale en milieu sportif définie comme l'occurrence d'un ou de plusieurs des symptômes suivants : suite à un coup à la tête, altération des fonctions mentales, présence d'une perte de connaissance, désorientation, amnésie rétrograde et/ou antérograde, ou l'un des symptômes subjectifs post-commotionnels (céphalées, intolérance à la lumière ou au bruit, nausée ou vomissement ou vertiges) ; présenter une commotion cérébrale de type complexe qui renvoie à l'existence de l'un ou de plusieurs, des symptômes comme la perte de connaissance supérieure à une minute, convulsion post-commotionnelle, persistance des symptômes post-commotionnels au-delà de 10 jours. Le sportif ne présente aucune lésion anatomique constante démontrable par la neuropathologie ou l'imagerie cérébrale.

Les critères d'exclusion sont les suivants : sportifs commotionnés dont la durée de l'amnésie post-traumatique est supérieure à une heure ou dont la durée de la perte de connaissance est supérieure à 30 minutes ; sportifs commotionnés suite à un traumatisme autre qu'en milieu sportif ; sportifs commotionnés avec une durée post-commotionnelle inférieure à 3 mois ; sportifs commotionnés présentant un trouble neurologique prémorbide (traumatisme crânien antérieur modéré ou sévère, épilepsie, etc.), ou un trouble psychiatrique avéré (dépression, anxiété généralisée, stress post-traumatique, etc.).

2.1.2 Le groupe témoin

Le groupe témoin est apparié au groupe expérimental quant à l'âge, au sexe et au niveau d'instruction. Il est composé de 20 sportifs de l'élite nationale pratiquant des sports avec un moindre risque de contact (10 handballeuses et 10 sportifs pratiquant l'athlétisme). Ce groupe comprend 11 femmes et 9 hommes. La moyenne d'âge est de 20,2 ans (σ = 1,67). Parmi les membres du groupe des témoins, 35% ont un niveau de scolarité supérieur et 65% un niveau secondaire.

Les sujets exclus du groupe sont ceux qui présentent ou qui ont pu présenter l'un des critères d'exclusion du groupe expérimental, ainsi que les sujets ayant relaté des antécédents traumatiques, neurologiques et/ ou psychiatriques.

2.2 Matériel

Il a été proposé aux 40 sujets de passer différents tests et questionnaires donnant chacun un ou plusieurs scores. L'évaluation s'est déroulée individuellement pour tous les sujets au Centre National de la Médecine et des Sciences du Sport ou dans des lieux de stage. La passation a pris environ une heure et quarante-cinq minutes.

2.2.1 Le Test tunisien d'apprentissage verbal (T.T.A.V.)

Ce **test d'apprentissage verbal** est directement inspiré des tests les plus utilisés dans l'examen des fonctions mnésiques tels que le Test des 15 mots de Rey (1964), le Test d'apprentissage verbal de Californie de Delis et *al.* (1987), ou encore le Test d'apprentissage verbal Hopkins de Benedict et *al.* (1998). Le T.T.A.V. permet de mesurer les capacités mnésiques du sujet dans des conditions standard, dans le but d'observer dans quelle mesure celui-ci est capable de mettre en place spontanément des stratégies efficaces d'encodage, de stockage et de récupération. Le matériel est composé de deux listes de mots (liste 1 et liste 2) comportant chacune 15 mots. La liste 1 est composée de 15 mots appartenant à trois catégories sémantiques différentes (fruits, vêtements, épices et herbes). La liste 2 comprend aussi 15 mots appartenant à 2 catégories non communes à la liste 1 (ustensiles de cuisine et légumes) et une catégorie commune à la liste 1 (fruits). La présentation de cette deuxième liste permet de mesurer la sensibilité du sujet à l'interférence proactive, c'est-à-dire la difficulté à apprendre un nouveau matériel en raison de l'apprentissage préalable d'un autre matériel. Chaque catégorie est représentée par 5 mots.

Tous les mots ont un rang de fréquence d'apparition dans la catégorie allant de 6 à 20, dans un ordre invariable d'un essai à l'autre. L'administration du T.T.A.V. comprend trois présentations et rappels immédiats de la liste 1 suivis de deux présentations et rappels de la liste 2. La phase qui suit comporte un rappel libre de la liste 1 puis un rappel indicé de la liste

Comparaison de groupes contrastés.

1. Après vingt minutes on procède à un rappel différé de la liste 1 et à un rappel différé indicé de la liste 1. Le test se termine par une tâche de reconnaissance où les 15 mots de la liste 1 sont entremêlés, d'une manière désordonnée, avec 30 mots distracteurs comportant 9 items de la liste 2, 9 items prototypiques des catégories utilisées dans la liste 1, 6 items phonétiquement proches des items de la liste 1 et 6 items non liés (ni phonétiquement, ni sémantiquement) aux items d'apprentissage des deux listes. Quatre indices de rappel et un indice de reconnaissance sont retenus. Il s'agit du nombre total de mots corrects rappelés, du nombre de répétitions, du nombre d'intrusions, du taux de groupements sémantiques dans les trois premiers rappels et de l'indice de discriminabilité dans la tâche de reconnaissance.

2.2.2 Le Test de Gestion des Ressources Attentionnelles (T.G.R.A)

Le Test de gestion des ressources attentionnelles de Pasquier est une épreuve inspirée du *5-Digit test* (ou *Digital Stroop*) de Sedó (1998a, 1998b) dans lequel l'auteur a remplacé les mots du test de Stroop par des chiffres afin de contourner les difficultés de lecture ou l'analphabétisme de certains sujets nord-américains. Toutes les données de la passation sont enregistrées dans un fichier .dat en arrière plan. Un développement sous Excel permet de choisir les données qu'on souhaite extraire, présenter et analyser. Sa validité a été vérifiée par Bellaj, Pasquier et Van Dam (2005) : le quotient attentionnel produit - QA - est en lien positif avec la mémoire à court terme, le facteur général d'intelligence, le temps d'inhibition au Stroop et en lien négatif avec une échelle de dépression. Un module d'entraînement en temps libre permet de familiariser le sujet avec la tâche et le temps de passation du module de test est limité à 10 minutes (contrainte de gestion de temps), le sujet devant produire le maximum de réponses justes (contrainte de rapidité du traitement de l'information et de l'efficience). Globalement, la tâche induite par le T.G.R.A évalue la gestion des ressources attentionnelles de la mémoire de travail qui sont mobilisées par la mise en oeuvre de fonctions exécutives impliquant des capacités d'inhibition et de flexibilité mentale. La tâche répond à une consigne double : lire un chiffre quand il est présenté dans un cadre délimité par un trait simple, compter les chiffres quand le cadre de présentation est entouré d'un trait double (Fig. 1).

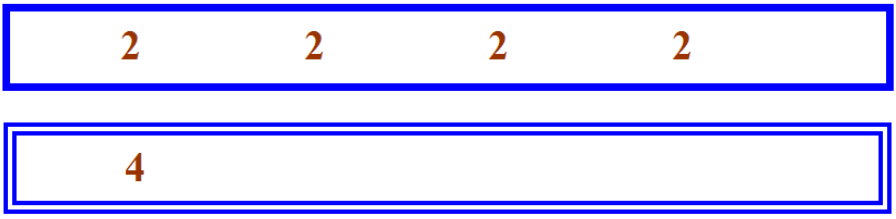


FIG. 1 – *Exemple de cadre simple et de cadre double du T.G.R.A.*

Parmi les multiples indices fournis par le test nous avons retenu l'indice QA (quotient attentionnel) qui combine vitesse et précision selon la formule

$$QA = \frac{(nombre\_de\_bonnes\_réponses)^2}{nombre\_de\_réponses\_traitées}$$

2.2.3 Le Test de classement de cartes de Wisconsin (W.C.S.T.)

Le W.C.S.T. peut être considéré comme une mesure du fonctionnement exécutif nécessitant l'élaboration et le maintien, dans des conditions changeantes, d'une stratégie appropriée de résolution de problèmes (Heaton et *al*., 2002). Le W.C.S.T. fait appel à la planification des stratégies de recherche d'informations structurées et requiert la prise en compte des rétroactions environnementales afin d'adapter le schéma de réponse qui est appliqué, et d'orienter le comportement vers la réalisation des objectifs tout en inhibant les réponses impulsives (Welsh et Pennington, 1988). Le matériel est composé de 128 cartes représentant deux fois les 64 combinaisons des 4 attributs différents des dimensions de couleur (rouge, jaune, vert et bleu), de forme (carré, cercle, étoile et croix) et de nombre (1, 2, 3 ou 4) avec pour consigne de les placer une à une en dessous de l'une ou l'autre des cartes-stimuli selon le critère d'appariement qu'il juge être pertinent et à l'endroit où il pense que la carte devrait se trouver. Les quatre cartes de référence contiennent respectivement un triangle rouge, deux étoiles vertes, trois croix jaunes et quatre cercles bleus. Suite à chaque réponse du sujet, l'expérimentateur informe le sujet sur la justesse de sa réponse mais jamais sur le critère d'appariement attendu. Le sujet doit tenter de parvenir à un maximum de réponses correctes sans toutefois pouvoir corriger ses erreurs. Il doit donc se servir des rétroactions de l'examinateur pour élaborer une nouvelle stratégie de classement. Nous avons retenu l'indice le plus utilisé dans la littérature qui est le nombre d'erreurs persévératives, soit tout classement effectué en fonction de la catégorie qui était correcte précédemment mais qui ne l'est plus actuellement.

2.2.4 La tâche du temps de réaction simple (T.R.S.)

Dans cette épreuve informatisée, on demande au sujet de réagir le plus rapidement possible à un stimulus cible apparaissant au centre de l'écran. Le stimulus est la lettre (X) ayant pour taille 1,5 cm, écrite en noir et présenté sur un fond blanc. Le sujet doit réagir à l'apparition du stimulus en appuyant sur la touche « X » du clavier. La disparition de la cible se fait immédiatement après l'appui sur le bouton réponse ou après 3000 ms en cas de non réponse. Cette tâche comporte quatre blocs d'items dont chacun est composé de 50 items avec dix intervalles différents inter-stimuli randomisés (250, 500, 750, 1000, 1250, 1500, 1750, 2000, 2250 et 2500 ms). A la fin de chaque bloc le sujet peut se reposer pendant quelques minutes. L'indice retenu est la médiane du temps de réaction aux différents essais.

2.2.5 Les Matrices Progressives de Raven (P.M. 38)

Les Matrices Progressives Standard (Raven, 1981) permettent d'apprécier le facteur g. Ce test comporte cinq séries de 12 problèmes de complètement de matrices. Le sujet a pour consigne de trouver les lois de progression, en ligne et en colonne, dans le but de découvrir laquelle des huit figures proposées correspond à la solution correcte. Ces problèmes sont de difficultés croissantes et doivent, selon la version, être réalisés en temps libre ou limité.

2.2.6 L'Echelle abrégée d'appréciation psychiatrique (BPRS 42)

L'Echelle abrégée d'appréciation psychiatrique (Pichot et al., 1973) est une échelle catégorielle de type Likert dont la cotation peut aller de 1 (absent) à 7 (extrêmement important). Durant l'entretien, l'examinateur essaye de relever l'état symptomatique et

comportemental du sujet pour juger de la présence et de l'intensité des diverses dimensions explorées : paranoïde ou délirant-hallucinatoire, mélancolique ou dépressif, hystérique, maniaque, paranoïaque ou hostilité-méfiance, organique, phobique-obsessionnel, hébéphrénique ou psychopathique. Le score retenu est le score total de la symptomatologie psychopathologique (score BPRS 42).

2.3 Traitement des données

La première étape du traitement des données a eu pour objectif de déterminer lesquelles des variables pouvaient être retenues dans le réseau nomologique dans la mesure où elles se différenciaient selon les deux groupes de sujets, Gexpé et Gtém, en termes de taille de l'effet appréciée par le d de Cohen. On s'en est tenu aux préconisations de Corroyer et Wolff (2003, p. 243) qui, en référence aux propositions de Cohen (1977), proposent les limites suivantes : 0 à 0, 35 effet négligeable ; 0,35 à 0,65 effet intermédiaire ; plus de 0,65 effet notable. La seconde étape propose une analyse en composantes des variables retenues dans le réseau nomologique afin d'en déterminer la structure. La troisième étape présentera l'analyse implicative des résultats.

3 Résultats :

3.1 Constitution du réseau nomologique :

La comparaison des résultats opposant les deux groupes de sujets, Gexpé et Gtém, a été menée pour chacun des scores produits par les différentes épreuves afin de déterminer lesquels d'entre eux différenciaient les deux groupes. Par exemple, pour la variable PM, après calcul des valeurs caractéristiques (m, σ et n) pour chacun des deux groupes (Tab. 1), on a considéré l'écart des moyennes (8 points), le d de Cohen (0,79), et la taille de l'effet (notable), ce qui a amené à retenir la variable PM comme variable différenciatrice des deux groupes, avec un avantage notable au groupe expérimental.

groupe	Gexpé	Gtém
m	34,50	42,50
σ	10,40	9,74
n	20	20

TAB. 1 - *Valeurs caractéristiques de la variable PM.*

A contrario la variable Score BPRS 42 n'a pas été introduite dans le réseau nomologique au vu d'un écart de |1,05|, d'un d de Cohen de 0,13 marquant un effet négligeable (Tab. 2).

groupe	Gexpé	Gtém
m	55,30	54,25
σ	8,43	7,46
n	20	20

TAB. 2 - *Valeurs caractéristiques de la variable Score BPRS 42.*

In fine, les variables ci-après ont été retenues pour constituer le réseau nomologique définitif. Pour quatre variables, le Gexpé obtient des scores moyens supérieurs à ceux du Gtém. Il s'agit du quotient attentionnel obtenu au Test de gestion des ressources attentionnelles (TGRA-QA ; 147,35 *vs* 150,45), du score aux Matrices Progressives de Raven (PM 38 ; 34,50 *vs* 42,50), du nombre total de mots rappelés dans les trois essais du Test tunisien d'apprentissage verbal (Cor A tot ; 29,35 *vs* 31,45) et pour ce même test, l'indice de discriminabilité (Discrim ; -250,33 *vs* -170,89). Pour les autres variables, le Gexpé obtient des scores inférieurs à ceux du Gtém, ce qui traduit un niveau moindre de difficultés. Il s'agit du nombre de répétitions d'items dans les trois rappels au T.T.A.V. (REPET ; 6,90 *vs* 4,05), du nombre d'intrusions (Intru ; 1,50 *vs* 0,85), du nombre de groupements sémantiques produits (Gp_sem ; 4,28 *vs* 3,47), ainsi que du temps de réaction simple (TRS ; 276,55 *vs* 261,70) et du nombre d'erreurs persévératives au Test d'assortiment d'images de Wisconsin (WCST-per ; 39,35 *vs* 32,15).

En d'autres termes, par rapport aux sportifs du Gtém, les sportifs commotionnés du Gexpé sont moins bons sur les indicateurs de performance et d'autre part ils éprouvent un niveau de difficulté plus marqué sur les indicateurs fonctionnels. On pourrait se contenter de ce profil différencié du Gexpé et du Gtém. On peut aller plus loin en recherchant la structure latente propre au réseau nomologique.

3.2 Structure du réseau nomologique :

Une première analyse factorielle donne le tableau des variances expliquées (Tab. 3).

Composante	Valeurs propres	Valeurs aléatoires	% de la variance	% cumulés
1	2,55	1,80	28,37	28,37
2	1,81	1,50	20,11	48,49
3	1,05	1,28	11,73	60,22
4	0,84	1,10	9,34	69,57
5	0,70	0,94	7,79	77,36
6	0,68	0,79	7,55	84,92
7	0,59	0,65	6,59	91,51
8	0,43	0,51	4,86	96,38
9	0,32	0,37	3,61	100

TAB. 3 - *Analyse factorielle du réseau nomologique ; facteurs et valeurs propres. Méthode d'extraction : Analyse des principaux composants.*

Trois facteurs présentent une valeur propre supérieure à 1, mais seulement deux de ces valeurs sont supérieures à des valeurs aléatoires[1]. Partant de là, une seconde analyse a été conduite par la méthode d'extraction du maximum de vraisemblance[2] avec rotation Oblimin et normalisation de Kaiser. La rotation a convergé en 12 itérations. Le test de qualité d'ajustement donne une valeur de khi-deux égale à 9,90 pour 19 d.d.l. et une signification de

[1] Les valeurs aléatoires ont été calculées à l'aide de *RanEigen 2.0* (Enzmann, 2003).

[2] Cette méthode d'extraction présente l'avantage de proposer un indice d'justement entre données observées et modèle théorique.

0,95. Selon la matrice des types (Tab. 4), le premier facteur regroupe, selon un effet de taille, les variables nombre de groupements sémantiques (gp_sem), nombre de répétitions (repet) et nombre d'intrusions (intru) ; le second facteur, bipolaire, oppose discrimination (discrim), total mots rappelés (cor a tot), quotient attentionnel (tgra-qa) et matrices progressives (pm) à nombre d'erreurs persévératives (wcst-per) et temps de réaction simple (trs). Ces deux facteurs sont négativement corrélés, de façon très modérée ($r_{F1.F2} = -0,21$).

Le premier facteur renvoie aux fonctions de contrôle exécutif en mémoire travail. Le facteur étant unipolaire, il apparaît que le déploiement d'efforts et de stratégies de regroupements sémantiques n'est pas incompatible avec l'échec de contrôle des répétitions et des intrusions en mémoire de travail, les efforts de mise en place de stratégies cognitives pouvant traduire des efforts de compensation des difficultés de contrôle en mémoire de travail.

variables	Facteur 1: contrôle exécutif	Facteur 2: efficience et fluidité
gp_sem	0,86	
repet	0,56	
intru	0,36	
discrim		0,76
cor a tot		0,72
tgra-qa		0,46
pm		0,45
wcst-per		-0,33
trs		-0,28

TAB. 4 - *Matrice des types.*

Gtém : groupe témoin, Gexpé : groupe expérimental ; pm : matrices progressives ; tgra-qa : quotient attentionnel ; trs : temps de réaction simple ; wcst-per : nombre d'erreurs persévératives ; scores au T.T.A.V. : discrim : discrimination, cor a tot : total mots rappelés, GP_sem : nombre de groupements sémantiques, repet : nombre de répétitions, intru : nombre d'intrusions.

Le second facteur peut référer à l'efficience et à la fluidité. Bipolaire, il oppose les variables d'efficience mnésique discrimination (discrim) et total mots rappelés (cor a tot), d'efficience attentionnelle (tgra-qa) et d'efficience intellectuelle générale (pm) à des variables de flexibilité (wcst-per) et de vitesse de traitement (trs). Cette opposition renvoie au sens des scores exprimant des niveaux de performances ou bien des niveaux de difficultés.

On compare ensuite les scores factoriels moyens des deux groupes Gexpé et Gtém. Pour le facteur 1, facteur du contrôle exécutif, le Gexpé présente un score moyen positif (m=0,35 ; σ=1,06) alors que le Gtém présente quant à lui un score moyen négatif (m=-0,35 ; σ=0,51) ; cet écart de 0,70 point renvoie un d de Cohen de 0,84, soit un effet notable. Les résultats s'inversent pour le facteur 2, facteur de l'efficience et de la fluidité, le Gexpé présentant un score moyen négatif (m=-0,38; σ=0,74) et le Gtém un score moyen positif (m=0,38; σ=0,83) ; dans ce cas, l'écart de 0,76 donne un d de 0,96, soit également un effet notable. A partir de ces valeurs on a pu situer les deux groupes sur le graphe factoriel (Fig. 2): Gtém se projetant près des indicateurs de performances (efficacité et flexibilité), alors que Gexpé se projette du côté des difficultés du contrôle exécutif.

En résumé, par une approche classique, on a mis en évidence un certain nombre de variables différenciant les sportifs commotionnés de ceux qui ne le sont pas. L'analyse factorielle du réseau nomologique a permis d'extraire la structure latente de ce réseau et de situer les deux groupes dans cette structure.

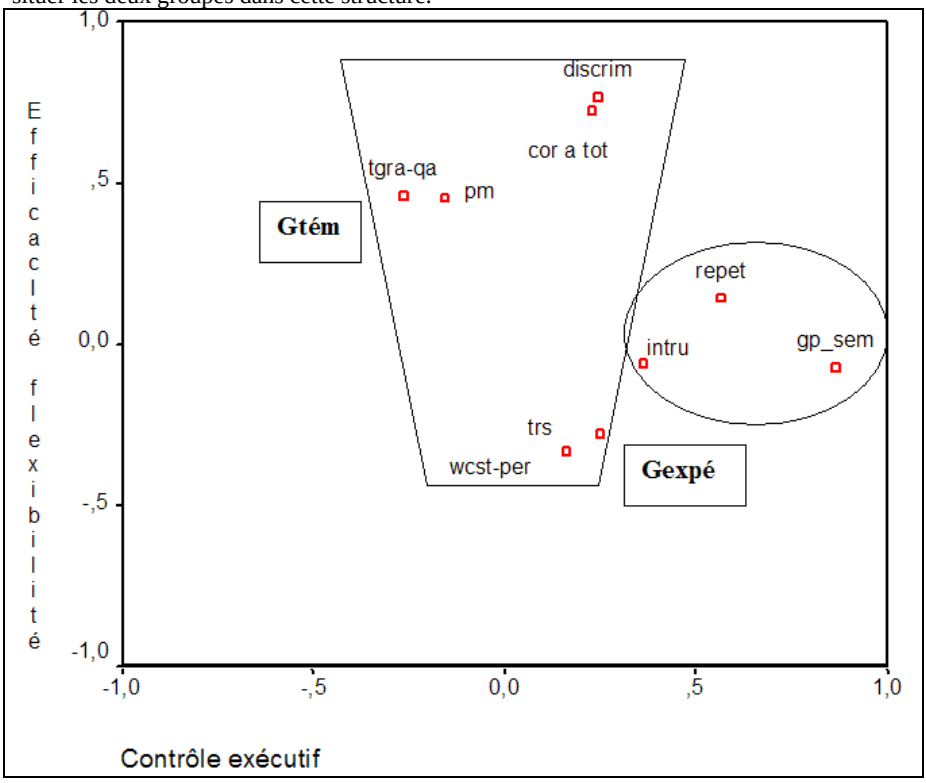


FIG. 2 – *Graphe factoriel et groupes témoin* (*Gtém*) *et expérimental* (*Gexpé*).
Gtém : groupe témoin, Gexpé : groupe expérimental ; pm : matrices progressives ; tgra-qa : quotient attentionnel ; trs : temps de réaction simple ; wcst-per : nombre d'erreurs persévératives ; scores au T.T.A.V. : discrim : discrimination, cor a tot : total mots rappelés, GP_sem : nombre de groupements sémantiques, repet : nombre de répétitions, intru : nombre d'intrusions.

Toutefois, on ne répond pas à une question centrale, théoriquement et pratiquement, relative aux déterminismes de l'accident sportif : la commotion cérébrale affecte-t-elle les fonctions psycho-neurologiques jusqu'à un mode dégradé ? Ou bien des dysfonctionnements psycho-neurologiques seraient-ils à la source de l'accident commotionnel ? L'analyse implicative offre des possibilités de réponse à cette question centrale.

Comparaison de groupes contrastés.

3.3 Analyse implicative des données :

Dans un premier temps, nous nous sommes intéressés à la structure cohésitive du réseau nomologique, en retenant comme variables supplémentaires le sexe et l'ancienneté dans le sport pratiqué. La hiérarchie cohésitive (Fig. 3) organise les variables des classes principales de manière quasi-implicative. Les nœuds significatifs sont indiqués par un trait épais rouge.

Par rapport au graphe factoriel obtenu supra, on remarquera que le nombre d'erreurs persévératives (WCST_per) est exclu de la logique quasi-implicative dans la mesure où il ne passe pas le seuil de sélectivité imposé par l'algorithme du logiciel. Nonobstant cette différence, les variables du réseau se regroupent de la même manière qu'elles s'agrègent aux deux axes factoriels. Ainsi, dans la classe figurant à droite du graphe, on retrouve les variables du pôle positif du facteur efficience et flexibilité, soit discrimination (Discrim), total mots rappelés (Cor A tot), matrices progressives (PM) et quotient attentionnel (TGRA-QA), ce pôle positif exprimant des niveaux de performances. Les individus les plus typiques du groupe optimal de cette classe appartiennent tous au groupe témoin. La variable supplémentaire la plus typique de cette classe est le sexe masculin avec un risque de 0,03.

Dans la classe figurant à gauche de la hiérarchie cohésitive, on retrouve les variables du facteur contrôle exécutif, à savoir nombre d'intrusions (Intru), nombre de répétitions (Repet), nombre de groupements sémantiques (Gp-sem) auxquelles s'est adjointe temps de réaction simple (TRS). Ce facteur montre que la mise en place de stratégies cognitives peut accompagner la mise en place de compensations des difficultés de contrôle en mémoire de travail. Les individus les plus typiques du groupe optimal de la seconde classe appartiennent aux deux groupes. La variable additionnelle la plus typique à cette classe est l'ancienneté dans le sport pratique avec un risque de 0,16 ; ce risque étant supérieur à 0,10 mais proche de ce seuil conventionnel, on peut retenir l'idée d'une tendance.

Bien que se construisant sur la cohérence de dissymétries implicatives et non sur des proximités corrélationnelles, la structure cohésitive extraite ici apparaît relativement semblable à la structure factorielle, les deux structures amenant à distinguer entre des variables d'efficience et des variables de contrôle exécutif.

Le « plus » apporté par la logique cohésitive est d'ordonner les variables à l'intérieur de chacune des classes en séquences quasi-implicatives dynamiques ouvrant la voie à des inférences relatives aux processus psychologiques sous-jacents. Dans la classe regroupant les variables d'efficience et de flexibilité, la cohésion implicative se construit à l'aide d'une règle significative (niveau 1 du graphe) et de deux méta-règles, dont l'une significative (niveau 3) : discrimination (Discrim) implique total mots rappelés (Cor_A_tot) ; l'implication Discrim => Cor_A_tot implique matrices progressives (PM), ce sous-ensemble impliquant quotient attentionnel (TGRA-QA). Dans l'autre classe, on observe une règle et deux méta-règles, dont l'une significative (niveau 2) : nombre d'intrusions (Intru) implique nombre de répétitions - Repet - => nombre de groupements sémantiques - GP_sem -, ce sous-ensemble impliquant temps de réaction simple (TRS). Il reste à interpréter ces deux structures quasi-implicatives.

Dans l'analyse factorielle, on impute la proximité corrélationnelle des variables observées à une variable latente, le facteur ou la composante, qui sature plus ou moins chacune d'elles : les variables ne sont pas reliées directement entre elles, mais par la distance à un moyen terme virtuel. Le facteur les explique toutes, mais aucune n'explique l'autre plus précisément. Par rapport à nos données, le facteur efficience et flexibilité permet de dire qu'un répondant obtenant un score élevé à discrimination (Discrim) aura tendance à obtenir

des bons scores aux autres variables du fait de sa capacité à se montrer efficient et flexible[3], mais on ne pourra rien inférer quant à l'influence de Discrim sur les autres variables.

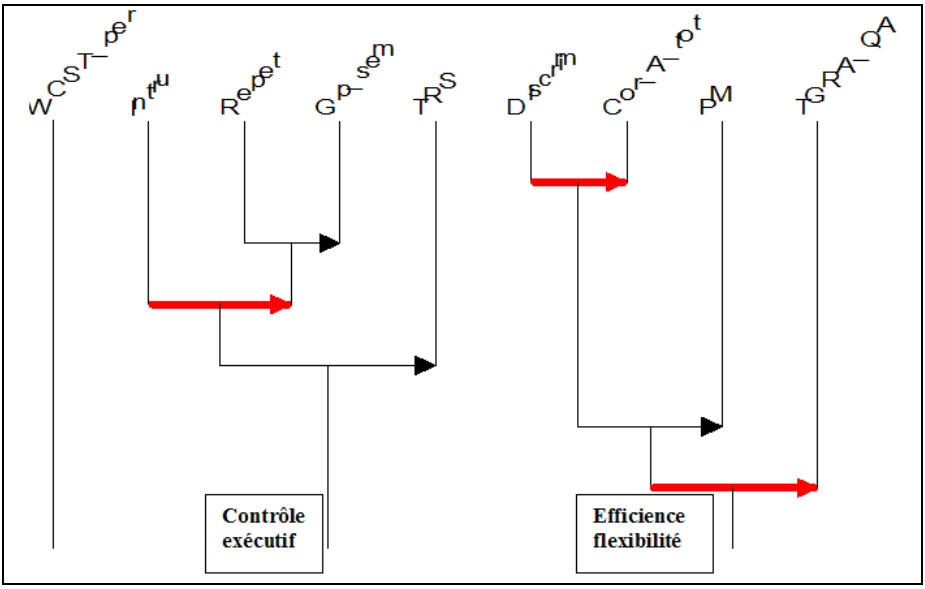


FIG. 3 - *Structure cohésitive du réseau nomologique.*
Gtém : groupe témoin, Gexpé : groupe expérimental ; pm : matrices progressives ; tgra-qa : quotient attentionnel ; trs : temps de réaction simple ; wcst-per : nombre d'erreurs persévératives ; scores au T.T.A.V. : discrim : discrimination, cor a tot : total mots rappelés, GP_sem : nombre de groupements sémantiques, repet : nombre de répétitions, intru : nombre d'intrusions.

L'analyse cohésitive complète l'analyse factorielle en ce sens qu'à l'intérieur d'une même classe, les liens qui fondent la cohésion de cette classe pourront être ordonnés de manière dynamique, de type A ⇒ B, mais pas l'inverse. Bien évidemment toute implication n'est pas nécessairement une relation causale et toute la difficulté de l'interprétation viendra de la nécessité d'ajouter des informations extérieures à la structure de la classe pour décider du caractère causal ou non des relations inter-variables, c'est-à-dire d'une règle ou d'une méta-règle.

Dans la classe des variables d'efficience et de flexibilité, au premier niveau, on observe la règle discrimination (Discrim) ⇒ total mots rappelés (Cor A tot). Cette règle unit deux scores du test d'apprentissage verbal. On dira, de manière descriptive, que c'est la capacité discriminative qui détermine le nombre de mots correctement rappelés, mais pas l'inverse, d'autres facteurs pouvant agir sur la qualité du rappel. En effet, depuis 1970, Kintsch

[3] Ce qui a pu faire dire à certains que l'analyse factorielle ressemblait à l'auberge espagnole : on y trouve ce qu'on y amène, un point c'est tout.

distingue au niveau du rappel deux processus : un processus de génération mentale de contenus mnésiques plausibles suivi d'un processus décisionnel de sélection du contenu cible parmi les contenus plausibles générés et ce en se basant sur un jugement de familiarité.

En revanche, la situation de reconnaissance n'exige que le deuxième processus décisionnel : celui de la sélection selon le jugement de familiarité. Ainsi, la discriminativité, basée sur le jugement de familiarité, est un dénominateur commun au rappel et à la reconnaissance et est donc une condition parmi d'autres, nécessaire mais pas suffisante, de la réussite de l'apprentissage verbal.

Cette conclusion peut paraître modestement limitée à première vue, mais sur un plan théorique elle est cohérente avec les données de la neuropsychologie relatives au fonctionnement en réseau des neurones. D'autre part, sur un plan pratique, le psychologue pourrait indiquer au pédagogue ou au rééducateur qu'il se donnera une chance d'améliorer l'efficience des apprentissages verbaux en exerçant les capacités discriminatives, mais qu'il serait *a priori* peu rentable d'entraîner les apprentissages verbaux dans le but d'affiner les capacités discriminatives[4].

Au second niveau se révèle une méta-règle certes non significative mais qui peut se raisonner de même manière. En psychométrie classique, les factorialistes ont produit différents modèles basés sur une structuration hiérarchique de l'intelligence (Lautrey,1990 ; Huteau et Lautrey, 2006), un facteur général se décomposant en facteurs de premier ordre, parmi lesquels un facteur verbal, un facteur spatial et un facteur numérique. Il n'est pas contre-intuitif de dire que l'apprentissage verbal relève plus largement du facteur verbal de l'intelligence, lui-même composante du facteur général. Dans notre exemple, dans le prolongement de la séquence implicative règle et méta-règle discrimination (Discrim) implique total mots rappelés (Cor_A_tot) ; l'implication Discrim => Cor_A_tot implique matrices progressives (PM), on affirmera qu'un bon niveau verbal participe à un bon niveau de facteur général évalué par le test des matrices, mais pas l'inverse, un bon niveau de facteur général pouvant être déterminé par ses autres composantes, ceci dans une certaine mesure.

Les suggestions psychopédagogiques iront dans le même sens : c'est en modifiant les caractéristiques du déterminant qu'on se donne une chance de modifier le déterminé et non l'inverse. En entraînant les capacités verbales d'un individu, on se (lui) donne une chance d'améliorer son intelligence générale. Au-delà du simple aspect opérationnel, la possibilité de lever l'incertitude du type qui de la poule et l'œuf est premier éclaire également les choix idéologiques entre les partisans du tout inné (idéologies du don au principe de la naturalisation des différences sociales) et les partisans des apprentissages culturels médiatisés au premier rang desquels Vygotski[5], pour lesquels l'intelligence s'apprend, essentiellement par la transmission intergénérationnelle et par l'entraînement aux codes et aux langages symboliques qui structurent et alimentent les fonctions mentales supérieures dans les limites d'une zone proximale de développement.

La seconde méta-règle significative, discrimination (Discrim) implique total mots rappelés (Cor_A_tot) ; l'implication Discrim $\Rightarrow$ Cor_A_tot implique matrices progressives (PM), ce sous-ensemble impliquant quotient attentionnel (TGRA-QA), suggère que le niveau

[4] On peut rapprocher ce développement de la problématique psychopédagogique des pré-requis.

[5] Sans oublier le pionnier que fut Jean Itard, ou encore les intuitions d'Alfred Binet sur l'éducabilité de l'intelligence.

de fonctionnement des ressources attentionnelles de la mémoire de travail évalué par le quotient attentionnel dépend des règles quasi-implicatives précédentes et non l'inverse.

Au niveau 1 de la classe regroupant des indices de contrôle exécutif du T.T.A.V. et du temps de réaction simple, une règle significative énonce que la fréquence de répétition des items affecte le taux de groupements sémantiques, cette règle étant elle-même affectée par les intrusions. Cet ensemble de règles et méta-règles relatives au contrôle exécutif dans l'apprentissage verbal impacte le temps de réaction simple (TRS), mais pas l'inverse. Une vue fixiste des choses pourrait laisser entrevoir un déterminisme neurologique rigide ne laissant entrevoir aucune possibilité d'intervention susceptible d'améliorer le TRS. Là encore, l'inférence dynamique abstraite des données sur le fonctionnement neuropsychologique ouvre la voie à des possibilités d'amélioration du TRS via un objectif d'apprentissage de stratégies visant l'optimisation fonctionnelle du contrôle exécutif.

Le recours au graphe implicatif des variables du réseau nomologique (Fig. 4) apporte une réponse à la question préalablement évoquée de la direction de la relation entre commotion et fonctions dégradées. L'introduction dans le graphe des deux groupes, Gexpé et Gtém, permet de préciser leurs statuts causaux et en conséquence le jeu des inférences possibles.

Sur la partie droite du graphe, on voit que le groupe témoin (Gtém) s'agrège à la classe des indicateurs d'efficience et de flexibilité, plus précisément comme point de départ du graphe, impliquant la variable discrimination (Discrim) en premier lieu, puis de manière transitive, les variables total mots rappelés (Cor_A_tot), quotient attentionnel (TGRA_QA) et matrices progressives (PM). Sur la partie gauche du graphe, le groupe expérimental (Gexpé) rejoint les indices fonctionnels de contrôle exécutif parmi lesquels il prend une place à la fois de déterminant des variables temps de réaction simple (TRS), nombre d'erreurs persévératives (WCST_per) et nombre de groupements sémantiques (Gp_sem) et de déterminé par l'implication reliant les variables nombre d'intrusions (Intru) => nombre de répétitions (Repet).

Ces constats généreront des inférences théoriques et pratiques spécifiques pour chacune des deux classes. Tout d'abord, malgré les éventuelles frustrations induites chez le chercheur, il faut souligner qu'en toute rigueur implicative on ne peut rien inférer à partir des variables temps de réaction simple (TRS), nombre d'erreurs persévératives (WCST_per) et nombre de groupements sémantiques (Gp_sem) pour une classe, ni à partir des variables total mots rappelés (Cor_A_tot), quotient attentionnel (TGRA_QA) et matrices progressives (PM) pour l'autre classe.

Du fait qu'un sportif de haut niveau n'ait subi aucun accident commotionnel, on pourra conclure, s'il présente un bon niveau de discrimination verbale, à la probabilité de bons niveaux d'intelligence générale, de gestion des ressources attentionnelles et d'apprentissage verbal. Des dysfonctionnements dans le contrôle attentionnel des contenus de la mémoire de travail (nombre d'intrusions - Intru - et nombre de répétitions - Repet -) sont à considérer comme des facteurs de risque d'appartenir au groupe expérimental, c'est-à-dire que pour un individu présentant ces caractéristiques il y a une probabilité de subir un choc commotionnel au cours de ses activités sportives. En aval du graphe, on peut lire les conséquences d'un choc commotionnel et réaliser des prévisions de dégradations comportementales pour des sujets commotionnés sur le temps de réaction simple (TRS) et aussi sur les stratégies d'apprentissage verbal (GP-sem) et de résolution de problème (WCST-per).

Comparaison de groupes contrastés.

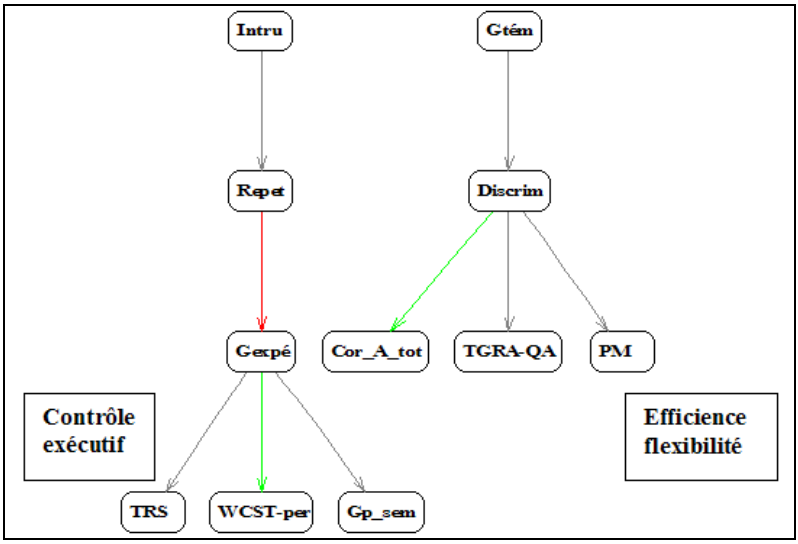


FIG. 4 : *Graphe implicatif du réseau nomologique et des deux groupes contrastés. Gtém : groupe témoin, Gexpé : groupe expérimental ; pm : matrices progressives; tgra-qa : quotient attentionnel ; trs : temps de réaction simple ; wcst-per : nombre d'erreurs persévératives ; scores au T.T.A.V. : discrim : discrimination, cor a tot : total mots rappelés, GP_sem : nombre de groupements sémantiques, repet : nombre de répétitions, intru : nombre d'intrusions.*

On peut ensuite préciser le rôle des variables supplémentaires pour chacun des chemins reliant les variables. Pour la classe du contrôle exécutif, la variable la plus typique du chemin Intru (nombre d'intrusions au T.T.A.V.) ⇒ Repet (nombre de répétions au T.T.A.V.) ⇒ Gexpé est le sexe masculin avec un risque de 0,03. En aval, la variable ancienneté dans le sport pratiqué est la plus typique du chemin Gexpé ⇒ TRS (temps de réaction simple) avec un risque de 0,11 ; la variable sexe féminin est la plus typique du chemin Gexpé ⇒ WCST_per (nombre d'erreurs persévératives) avec un risque de 0,04 et du chemin Gexpé ⇒ nombre de groupements sémantiques (Gp_sem) avec un risque de 0,00. Pour la classe de l'efficience et de la flexibilité, la variable la plus typique aux trois chemins est l'ancienneté dans le sport pratiqué avec un risque de 0,19 pour le chemin Gtém => discrimination (Discrim) ⇒ total mots rappelés (Cor_A_to) ; avec un risque de 0,16 pour le chemin Gtém ⇒ discrimination (Discrim) ⇒ quotient attentionnel (TGRA_QA) ; avec un risque de 0,11 pour le chemin Gtém ⇒ discrimination (Discrim) ⇒ matrices progressives (PM). Deux variables additionnelles interviennent donc pour caractériser la typicalité des chemins : l'ancienneté dans le sport pratiqué, principalement dans la classe de l'efficience et de la flexibilité et le sexe dans la classe du contrôle exécutif, le masculin typicalisant les déterminants d'appartenance au groupe expérimental et le féminin typicalisant les déterminés par cette d'appartenance. Au vu d'un effectif limité en taille et au vu d'un manque d'études réalisées sur le sujet, on se gardera de construire une théorie générale à partir de ces constats.

4 Conclusion

Alors que l'approche classique s'arrête au niveau descriptif statique se limitant à la caractérisation des réseaux nomologiques différentiels, l'approche implicative nous introduit ainsi dans une analyse dynamique où l'appartenance même à un groupe ne constitue pas le point de départ d'une recherche. L'analyse implicative nous fournit des arguments en faveur d'hypothèses largement émises dans la littérature qui suggèrent que le fait d'appartenir au groupe des commotionnés n'est pas dû au hasard. Des facteurs prédisposants sont à considérer. Des failles attentionnelles, de mémoire de travail ou de contrôle exécutif de l'action pourraient être impliquées dans l'occurrence de l'accident traumatique.

Le statut des fonctions exécutives dans la détermination du comportement est à revoir en fonction de ces résultats. Plusieurs travaux ont décrit des dysfonctionnements exécutifs chez les sportifs commotionnés (Collins et al., 1999 ; Echemendia et *al*., 2001 ; Koh et *al*., 2003 ; Rutherford et *al*. 2005). Aucun de ces auteurs ne s'est demandé si les déficits observés sont secondaires à la commotion ou secondaires et préludes à la commotion. En fait, les fonctions exécutives renvoient aux mécanismes cognitifs de haut niveau qui assurent la planification de l'action, régulent le comportement en fonction du contexte et contrôlent les opérations en cours (Luria, 1978). Il s'agit de mécanismes qui agissent par amplification et/ou par inhibition sur plusieurs processus dont les processus de contrôle moteur (Mahone et *al*., 2006), mnésiques (Anderson et Green, 2001), attentionnels (Tipper, 2001) et de régulation émotionnelle (Ochsner et Gross, 2005), tous processus clefs dans la pratique d'une activité sportive de haut niveau. Des difficultés pré-morbides dans la prise d'initiative, dans la planification de l'action, dans la suppression d'une réponse non appropriée, dans la modification de la réponse en fonction des changements de contexte ou dans la régulation émotionnelle peuvent entraîner des gestes risqués et précipités manquant de jugement. Une commotion peut surtout survenir lorsqu'il y a une faille, même temporaire, dans l'interface entre penser et agir.

D’une manière plus large, il ne s’agit pas d’opposer approche classique et approche implicative, mais d’établir un rapport de complémentarité entre elles, la seconde précisant la première en donnant une direction quasi-implicative, voire causale, aux liaisons abstraites entre variables observées, ou construites ou latentes. Il ne s’agit pas de nier l’intérêt des approches corrélationnelles, mais de donner un sens, voire du sens, par une interprétation dissymétrique des liaisons. C’est là une occasion d’éviter de tirer des conclusions abusives car commandées par des scripts mentaux plus ou moins bien conscientisés qui relèvent soit des idéologies de la pensée unique dérivées du *mainstream* anglo-saxon, soit des schémas de la pensée naturelle ou naïve.

On en trouvera une illustration dans cette recherche si on prend en compte l’idéologie de la performance dans un monde réduit à la catégorie des gagnants et des perdants. A partir de la différenciation des deux groupes Gtém et Gexpé sur la base des indicateurs de performances, on aurait pu spontanément expliquer le fait que les sportifs non commotionnés ne le sont pas parce qu’ils sont plus intelligents, plus… et étayer des décisions de recrutement de sportifs de haut niveau sur cette explication et sur la prise en compte des simples niveaux de performances. La directionnalité, abstraite de l’analyse implicative, montre qu’on ne peut rien conclure en termes de probabilité d’accident sportif à partir des seuls niveaux de performances mais qu’il serait plus pertinent de prendre en compte comme **prédicteurs** les indicateurs fonctionnels du contrôle exécutif d’un test d’apprentissage verbal. Ajoutons que si l’attribution d’une simple catégorie de niveau relève de différentes

formes de fatalismes (génétique, économique, socio-culturel, affectif…) par rapport auxquels l'action psychopédagogique ou médicale ne peut se montrer qu'inopérante, le centrage sur les aspects fonctionnels, donc dynamiques et interactifs, ouvre la voie à toute perspective de remédiation.

Références

Anderson, M.C., et Green, C. (2001). Suppressing unwanted memories by executive control. *Nature*, 410 (6826): 366-9.

Bellaj, T., Pasquier, D., et Van Dam, F. (2005). Une étude de validité du test de gestion des ressources attentionnelles -T.G.R.A.-. *Revue Tunisienne de Sciences Sociales*, 130, 29-59.

Benedict, R.H.B., Schretlen, D., Groninger, L., et Brandt, J. (1998). Hopkins Verbal Learning Test-Revised: Normative data and analysis of inter-form and test-retest reliability. *The Clinical Neuropsychologist*, 12, 43-55.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (revised edition). New York: Academic Press.

Collins, M. W., Grindel, S. H., Lovell, M. R., Dede, D. E., Moser, D. J., Phalin, B. R., et al. (1999). Relationship Between Concussion and Neuropsychological Performance in *College Football Players*. JAMA, 282(10), 964-970.

Corroyer, D., et Wolff, M. (2003). *L'analyse Statistique des Données en Psychologie - Concepts et méthodes de base.* Paris : Armand Colin.

Couturier, R, et Gras, R. (2005). CHIC : Traitement de données avec l'analyse implicative, Extraction et Gestion des Connaissances, Volume I1, *RNTI*, Cepadues, Paris, p.679-684, ISBN 2.85428.683.9

Delis, D.C., Kramer, J.H., Kaplan, E., et Ober, B.A. (1987). *California Verbal Learning Test*. New York: The Psychological Corporation.

Echemendia, R. J., Putukian, M., Mackin, R. S., Julian, L., et Shoss, N. (2001). Neuropsychological Test Performance Prior To and Following Sports-Related Mild Traumatic Brain Injury. *Clinical Journal of Sport Medicine*, 11, 23-31.

Heaton, R. K., Chelune, G. J., Tally, J. L., Gary, G. G., et Curtiss, G. (2002). *Test de classement de cartes de Wisconsin*. Paris: ECPA.

Huteau, M., et Lautrey, J. (2006). *Les tests d'intelligence*. Paris : Editions La Découverte

Kintsch, W. (1970). *Learning, memory and conceptual processes*. New York: Wiley.

Koh, J.O., Cassidy, J.D. et Watkinson, E.J. (2003). Incidence of concussion in contact sports: a systematic review of the evidence. *Brain Injury*, 17: 901217.

Lautrey, J. (1990). Esquisse d'un modèle pluraliste du développement cognitif. In Reuchlin, Lautrey, Marendaz, Ohlmann (eds), *Cognition : l'individuel et l'universel*; pp. 185-213, Paris : PUF.

Luria, A. R. (1978). *Les fonctions supérieures de l'homme*. Paris : Presses Universitaires de France.

Mahone, E.M., Powell, S.K., Loftis, C.W., Goldberg, M.C., Denckla, M.B., et Mostofsky, S.H.(2006). Motor persistence and inhibition in autism and ADHD. *Journal of the International Neuropsychological Society*, 12(5): 622-31.

McCrory, P., Johnston, K., Meeuwisse, W., Aubry, M., Cantu, R., Dvorak, J., Graf-Baumann, T., Kelly, J., Lovell, M., et Schamasch, P. (2005). Summary and agreement statement of the 2nd International Conference on Concussion in Sport, Prague. British *Journal of Sports Medicine*, 39 (Supplement 1): 78-86.

Ochsner, K.N., et Gross, J.J. (2005). The cognitive control of emotion. Trends in Cognitive *Sciences*, 9 (5): 242–9.

Pichot, P., Samuel, L. B., et Lebeaux, A. M. (1973). Etude d'une nouvelle forme expérimentale de la BPRS. *Annales Médico-psychologiques*, 2, 254-273.

Raven, J. C. (1981). *Matrices Progressives de Raven*. Paris : EAP.

Rey, A. (1964). *L'examen clinique en psychologie*. Paris: Presses Universitaires de France.

Rutherford, A., Stephens, R., Potter, D., et Fernie, G. (2005). Neuropsychological Impairment as a Consequence of Football (Soccer) Play and Football Heading: Preliminary Analyses and Report on University Footballers. *Journal of Clinical and Experimental Neuropsychology*, 27, 299-319.

Sedó, M. A. (1998a). *Five Digit Test*. Natick, Multigraphié.

Sedó, M. A. (1998b). *Digital Stroop*. Natick, Multigraphié (november revision).

Tipper, S.P. (2001). Does negative priming reflect inhibitory mechanisms? A review and integration of conflicting views. Quarterly *Journal of Experimental Psychology*- A, 54 (2): 321–43.

Welsh, M. C., et Pennington, B.F. (1988). Assessing frontal lobe functioning in children: views from developmental psychology. *Developmental Psychology*, 4, 199-230.

Summary

The use of classical psychometric indices of correlation between variables are symmetrical. For example, the correlation between a and b will be the same as that between b and a, each variable can take either the status variable explained or explanatory variable: it is impossible to order correlations between variables sequentially implicative. The analysis implicative data provides an opportunity to achieve this objective scheduling sequence of variables. In this chapter, we present an example of using this type of analysis to compare groups contrasted. The interest of this text lies not in the results for themselves, but in the opportunities created by the process.

Chapitre 6 : Utilisation de la statistique implicative pour la construction d'un référentiel de compétences comportementales

Dominique Follut*, Laurence Diot*, Martine Lasserre Azema**, Serge Baquedano*

*PerformanSe SAS
Atlanpole La Fleuriaye 44470 Carquefou
**CIBC de l'Ariège
18 Rue de l'Espinet 09000 Foix
dominique.follut@performanse.fr,
laurence.diot@performanse.fr
serge.baquedano@performanse.fr
mla-cibc09@wanadoo.fr
http://www.performanse.com

Résumé. Cinq Centres Interinstitutionnels de Bilan de Compétences de la Région Midi-Pyrénées se sont groupés pour créer une méthode d'appui individualisé au profit des personnes engagées dans une démarche de Validation des Acquis de l'Expérience. PerformanSe, éditeur d'outils dévaluation des compétences comportementales, a conçu pour cet usage un système informatisé permettant aux conseillers des CIBC d'être informés sur les comportements - favorables ou défavorables - susceptibles d'être adoptés par chacun des bénéficiaires qu'ils accompagnent dans le cadre d'une VAE.

1 Introduction

Depuis 20 ans, PERFORMANSE conçoit, développe et commercialise auprès des professionnels des Ressources Humaines – entreprises, cabinets-conseils, organismes institutionnels - des solutions logicielles dédiées à l'évaluation et à la gestion des **compétences comportementales.**

Ces outils permettent de structurer et d'organiser des démarches d'évaluation favorisant :

- la maîtrise des processus d'aide à la décision (recrutement, mobilité interne, formation…),
- la mise en œuvre d'une démarche d'objectivation de ces processus (traçabilité des décisions, réponse aux impératifs de non discrimination),
- l'élaboration en toute autonomie de stratégies de gestion des compétences,
- l'intégration des spécificités culturelles et de la diversité des pratiques de terrain dans les démarches d'évaluation.

PerformanSe-ECHO est un inventaire de personnalité qui permet d'expliciter et d'évaluer les caractéristiques comportementales d'une personne. Sa dernière validation psychométrique a été réalisée en 2004 en s'appuyant sur un échantillon de 4 538 dossiers individuels. Cette solution est aujourd'hui déployée dans plusieurs pays, et elle équipe plus

de 2 000 entreprises, cabinets-conseil et groupements institutionnels. Cet outil mesure 10 dimensions de personnalité bipolaires en s'appuyant sur un questionnaire de 70 items à choix forcé et une base de connaissances structurée à partir de plus de 27 000 règles.

Le modèle comportemental de PerformanSe-ECHO qui a été utilisé ici repose sur trois axes :

- La théorie psychologique des « Big Five », décrivant la personnalité à partir de 5 facteurs : *Extroversion, Consciensiousness, Openness, Agreeableness, Nevrosism.*
- L'étude des motivations, comme éléments orientant le comportement individuel.
- L'approche systémique et comportementale développée par «l'Ecole de Palo-Alto».

Il est organisé autour de dix **dimensions bipolaires** :

- Extraversion / Introversion (EXT/INT)
- Anxiété / Détente (ANX/DET)
- Affirmation / Remise en cause (AFF/RMC)
- Réceptivité / Distanciation (REC/DTN)
- Rigueur / Improvisation (RIG/IMP)
- Dynamisme intellectuel / Conformisme intellectuel (DIN/CIN)
- Combativité / Conciliation (COM/CCL)
- Motivation de Réalisation / Motivation de Facilitation (REA/FAC)
- Motivation d'Appartenance / Motivation d'Indépendance (APP/IND)
- Motivation de Pouvoir / Motivation de Protection (POU/PRO)

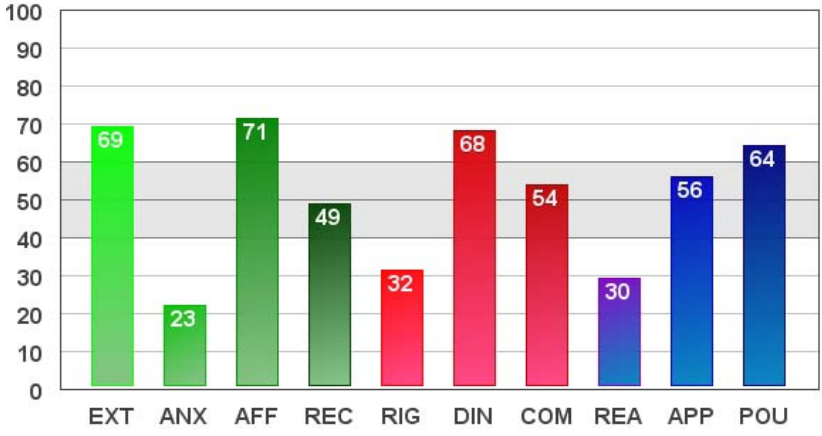


FIG. 1. - *Exemple de graphique PerformanSe-ECHO*

2 Contexte de ce travail

L'inventaire PerformanSe-ECHO est utilisé comme outil d'aide à l'accompagnement par les Centres Interinstitutionnels de Bilan de Compétences, dont une mission essentielle est l'aide à l'orientation et à l'évolution professionnelle. Dans cette perspective, les CIBC ont

été investis d'une mission de Points Relais Conseil en Validation des Acquis de l'Expérience.

En 2002 et 2003, dans le cadre de la mission "*Recherche et Développement*" visant à élaborer de meilleures pratiques d'accompagnement de la VAE, le CIBC de l'Ariège a mené un travail d'analyse approfondie sur le déroulement des parcours suivis par les bénéficiaires. Cette étude a notamment révélé un nombre excessif d'abandons en cours de VAE, et la nécessité d'améliorer ou d'enrichir l'intervention des conseillers pour atténuer ce phénomène.

Fort de ce constat, dès 2004, ce CIBC a élaboré et mis en place un dispositif original appelé "Appui renforcé", qui a ultérieurement fait l'objet d'une préconisation officielle dans le cadre d'une mission d'évaluation sous l'égide du Premier ministre. Cette prestation consiste à mettre en place une méthodologie de soutien personnalisé au profit du public s'engageant dans une VAE. Un axe important de cet accompagnement, qui prévoit une dizaine d'entretiens de suivi avec un conseiller, est fondé sur la déclinaison du concept d'image de soi – sa consistance et son évolution - au cours du processus de VAE.

Ce dispositif nécessite une approche spécifique qui tienne compte des caractéristiques individuelles de chaque bénéficiaire. C'est pourquoi les CIBC initiateurs du projet ont souhaité se doter d'un outil qui facilite le diagnostic et la prise de décision des conseillers en début d'accompagnement. Cet outil s'appuie sur la technologie des référentiels de compétences comportementales élaborée par PERFORMANSE. Il s'agit plus précisément d'un « module clignotants », méthode qui permet de croiser une évaluation comportementale à partir du questionnaire PerformanSe-ECHO avec un ensemble d'indicateurs, construits en référence à une situation donnée. Au cours d'un traitement individuel, seuls s'affichent les clignotants qui s'avèrent pertinents pour la personne concernée.

Le « Module Clignotant VAE » poursuit trois objectifs :

1. identifier des personnes ayant particulièrement besoin d'un suivi, tout en signalant leurs comportements clés.
2. sécuriser le diagnostic en orientant vers les modalités de suivi les plus pertinentes.
3. personnaliser pendant toute la durée du suivi les conseils donnés à chaque bénéficiaire par son conseiller.

3 Problématique

Toute la démarche de construction d'un « module clignotant » se fait en relation avec une réalité opérationnelle particulière. Concevoir un outillage personnalisé implique en effet de se référer à un corpus d'observations issu de la pratique quotidienne des personnes destinataires de cet outillage. C'est pourquoi la méthode de construction élaborée par PERFORMANSE s'appuie d'une part sur l'expertise des praticiens de terrain et d'autre part sur des échantillons représentatifs de la population à qui cet outil est destiné. Pour construire ce « **module clignotant** VAE », nous avons donc utilisé conjointement trois approches complémentaires :

- Les méthodes statistiques descriptives qui mettent en avant les caractéristiques fortes de l'ensemble de l'échantillon étudié.
- Les outils de statistique implicative qui permettent de mettre en évidence des relations caractéristiques entre **traits de personnalité** (pour l'ensemble de

l'échantillon, mais aussi pour des configurations comportementales individuelles) en tenant compte essentiellement de la force implicative des règles proposées.

- L'expertise, qui est recueillie auprès des conseillers en charge de l'accompagnement, afin de déterminer les comportements favorables et ceux qui ne le sont pas pour mener à son terme dans des conditions optimales un parcours de VAE.

Cette triple approche est fondamentale quand on se positionne sur le champ de la psychologie. En effet, pour atteindre l'objectif fixé, à savoir personnaliser l'accompagnement, il convient de tenir compte des comportements caractéristiques de la population mais aussi - et peut être surtout - des comportements individuels, fussent-ils rares, voire atypiques ou étonnants. La personne qui les adopte manifestera en effet des points d'appui ou de vigilance particuliers ou originaux qui doivent être immédiatement et clairement identifiés par le conseiller afin que celui-ci adapte de manière pertinente son accompagnement.

Il s'agit donc :

- d'une part, après avoir caractérisé une population de façon générale par la statistique descriptive, d'aller au-delà en qualifiant rapidement et manière fiable des comportements individuels adoptés par une proportion plus ou moins importante de cette population.
- d'autre part, de découvrir directement, grâce à la statistique implicative, des comportements atypiques (ne relevant pas de caractéristiques communes à la population ou à telle ou telle cohorte).

L'essentiel, dans une telle application, consiste à pouvoir mettre au jour et affirmer avec une sécurité suffisante des combinatoires (comportements) qu'un traitement statistique classique ne révèle pas. Cette approche permet d'identifier avec fiabilité des comportements particuliers, mais l'expert reste toutefois maître dans le choix des combinatoires qui font sens pour lui. Toute la subtilité de ce travail réside dans l'utilisation des seuils de significativité, c'est-à-dire, dans la force d'implication. En d'autres termes, si en statistique classique, une probabilité de 50% de chance de se tromper est largement à rejeter, dans le cadre de la statistique implicative, une telle probabilité permet à l'expert de retenir des combinatoires faiblement occurrentes mais dont l'importance devient cruciale lorsqu'il s'agit de créer des indicateurs centrés sur les différences interindividuelles. Ici, « se tromper » prend une valeur positive dans la mesure où l'on peut repérer des cas particuliers, en créant davantage d'indicateurs que ne le permettrait une étude classique. En l'occurrence, l'outil résultant de cette démarche permet aux conseillers des CIBC d'être alertés, pour chaque bénéficiaire, de comportements particuliers pouvant poser problème, ce qui constitue toute la logique du « module clignotant ». La problématique opérationnelle de cette étude est donc de déterminer à l'aide de la statistique implicative des indicateurs comportementaux individuels, fussent-ils peu typiques de la population d'étude.

4 Méthodologie

Notre recherche se base sur deux échantillons de bénéficiaires, les uns ayant validé leur VAE, les autres étant encore en cours de validation. Les 258 évaluations PerformanSe-ECHO se répartissent ainsi :

- 44 personnes ayant mené à bien le parcours de validation.
- 214 personnes n'ayant pas terminé le processus.

Pour chaque bilan comportemental individuel, chacune des 10 dimensions est représentée par trois variables suivant que le sujet vérifie peu (-), moyennement (0), beaucoup (+) cette dimension (par exemple : EXT- pour une faible extraversion, EXT0 pour extraversion moyenne et EXT+ pour extraversion forte). La méthodologie employée s'appuie sur un protocole en deux étapes pour chacun des échantillons. Il s'agit dans un premier de temps de décrire statistiquement les dimensions en terme d'occurrence, de fréquence et d'écart type. Dans un second temps, nous procédons à une analyse implicative entre les dimensions, à l'aide de CHIC. Ce logiciel a été choisi car il permet, d'un part de traiter les fonctions de base de la statistique classique et d'autre part d'effectuer une analyse implicative sur les dimensions. L'analyse des données consiste à mettre en perspective les résultats obtenus pour chacun des deux échantillons, c'est-à-dire VAE réussie versus VAE en cours. Le graphe implicatif permet d'identifier des groupes de combinaisons, qu'il est possible d'associer à une analyse des cardinalités.

5 Résultats

5.1 Résultats du traitement en statistique descriptive.

5.1.1 Comparaison de la population des 258 dossiers VAE et de la population d'étalonnage du questionnaire (4 538, tout venant : « français moyen »)

Par construction, le questionnaire PerformanSe-ECHO est réglé de telle sorte que, sur la population d'étalonnage (4 538), la ventilation des scores corresponde pour chacune des 10 dimensions aux normes suivantes (issue d'une répartition normale):

- dimension X : X- : 25% X0 : 50% X+ : 25%

A partir du traitement effectué sur les 258 dossiers VAE, quatre dimensions caractérisent particulièrement cette population par une fréquence d'occurrence significativement différente de l'étalonnage. Le tableau suivant montre pour ces quatre dimensions significatives l'écart observé par rapport à ce que l'on aurait pu attendre en fonction de l'étalonnage du modèle PerformanSe-ECHO. Le nuancier qui à été ajouté montre quelles sont les expressions de la dimension retenue en fonction de sa valence.

Dimension	Ecart par rapport au modèle	Valence	Score/items	Nuance d'expression de la dimension
Motivation de pouvoir (POU)	0.18	-	40	Précaution
			30	Couverture
			20	Sécurisation
			10	Immobilisme
Rigueur (RIG)	0.13	+	90	Rigidité
			80	Structuration
			70	Organisation
			60	Contrôle/ suivi
Réceptivité (REC)	0.09	+	90	Assimilation à l'autre
			80	Intérêt à ce qu'exprime l'autre
			70	Ouverture à l'autre
			60	Écoute
Anxiété (ANX)	0.09	+	90	Angoisse
			80	Anxiété
			70	Préoccupation
			60	Attention/Vigilance

TAB. 1 - *Écart et nuances des dimensions d'une caractéristique de la population VAE*

Les écarts observés ici sont significatifs et révèlent une caractéristique bien marquée de la population VAE. Cette typicité se révèle bien plus lorsque l'on regarde dans leur étendue les écarts par rapport au modèle PerformanSe-ECHO (figure 3). Il devient alors évident que des distorsions sont présentes dans les distributions, reflétant un ensemble de caractéristiques propres à cette population.

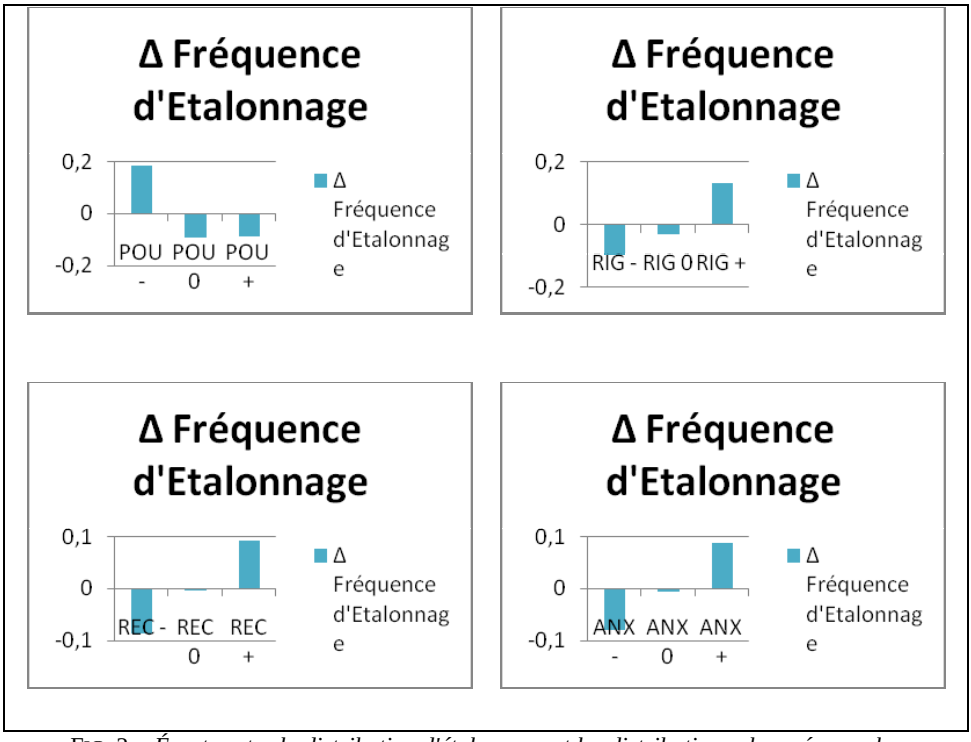


FIG. 2 - *Écarts entre la distribution d'étalonnage et les distributions observées sur les populations VAE*

5.1.2 Comparaison des deux échantillons VAE réussie et VAE en cours.

De la manière dont nous avons étudié les différences entre la population « VAE » et les résultats attendus du modèle, nous pouvons nous intéresser aux différences intra population en considérant d'une part la population qui a réussi sa VAE « VAE réussie » et celle qui n'a pas achevé sa VAE « VAE en cours ». Six dimensions sur dix distinguent ces deux populations par des différentiels d'occurrence significatifs. Ces différences sont exprimées dans le tableau ci-dessous

Dimension			Ecart entre « VAE réussie » et « VAE en cours »	Valeur	Nuance d'expression de la dimension
Extraversion	+	EXTraversion	0.25 réussie vs 0.17 en cours	90	*Envahissement*
				80	*Occupation de l'espace*
				70	*Volubilité*
				60	*Spontanéité*
	-	INTroversion	0.34 réussie vs 0.40 en cours	40	*Concentration*
				30	*Réserve*
				20	*Intériorisation*
				10	*Mutisme*
Affirmation	+	AFFirmation	0.32 réussie vs 0.19 en cours	90	*Certitudes*
				80	*Principes*
				70	*Conviction personnelle*
				60	*Confiance en soi*
	-	ReMise en Cause	0.27 réussie vs 0.42 en cours	40	*Questionnement*
				30	*Révision des positions*
				20	*Doute*
				10	*Incertitudes*
Motivation de Pouvoir	+	POUvoir	0.20 réussie vs 0.15 en cours	90	*Directivité*
				80	*Autorité*
				70	*Ascendant*
				60	*Influence*
	-	PROtection	0.36 réussie vs 0.45 en cours	40	*Précaution*
				30	*Couverture*
				20	*Sécurisation*
				10	*Immobilisme/*
Combativité	+	COMbativité	0.20 réussie vs 0.16 en cours	90	*Agressivité*
				80	*Affrontement*
				70	*Pugnacité*
				60	*Compétition*
	-	ConCiLiation	0.25 réussie vs 0.16 en cours	40	*Consensus*
				30	*Concessions*
				20	*Evitement*
				10	*Abandon*

TAB. 2 - *Différences entre "VAE Réussie" et "VAE Partielle"*

Afin de préciser ces différences une comparaison plus fine peut être réalisée en superposant les distributions de chacune des populations et la distribution d'étalonnage du modèle. Les caractéristiques de chacune des populations sont alors évidentes et montrent bien ce qui les différencie. De plus, ces différences font sens, en termes d'expertise, par

rapport à chaque contexte et sont cohérentes avec les observations des consultants de terrain. Le tableau suivant (figure 3) propose ces superpositions :

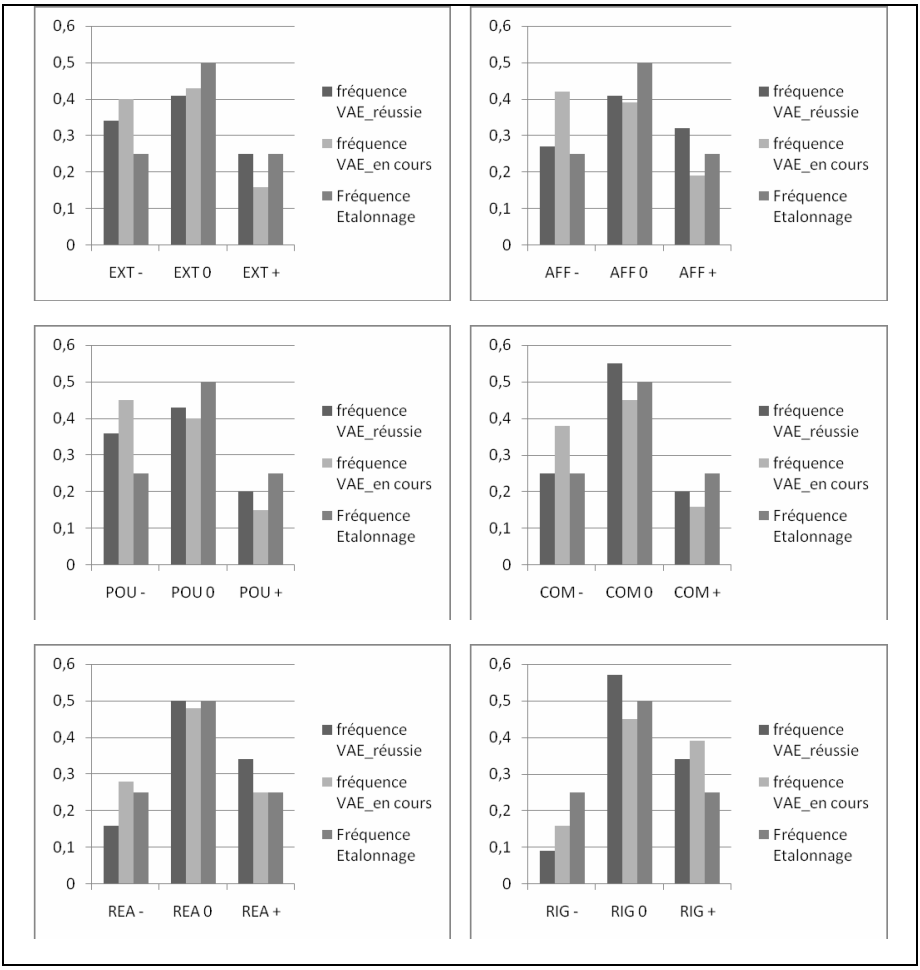


FIG. 3 - *Fréquences des différentes dimensions sur les trois populations de référence*

Ce faisceau d'indications permet sans aucun doute d'affirmer qu'il y a là des particularités bien typées qu'il est nécessaire de prendre en compte dans la réalisation d'un outil d'évaluation des point d'appuis et de vigilance pour s'engager dans une démarche de VAE et la mener à bien.

5.1.3 Synthèse.

Cette première phase de travail permet de caractériser de façon pertinente et significative les diverses populations en jeu, et d'émettre quelques hypothèses, comme l'influence de l'estime de soi, de la pugnacité, du maintien des efforts dans la durée... Ceci est nécessaire comme étape préalable, notamment pour le travail avec les experts du domaine mentionné ci-dessus, mais ne répond pas encore au problème de fond, qui consiste à identifier des comportements individuels et à valider leur pertinence.

5.2 Traitement en statistique implicative.

5.2.1 Généralités.

Le traitement séparé des deux échantillons permet d'extraire 29 règles pour la population VAE réussie et 31 règles pour la population VAE en cours. Par règle, nous entendons : combinaison sur plusieurs dimensions, significative d'un comportement observable.

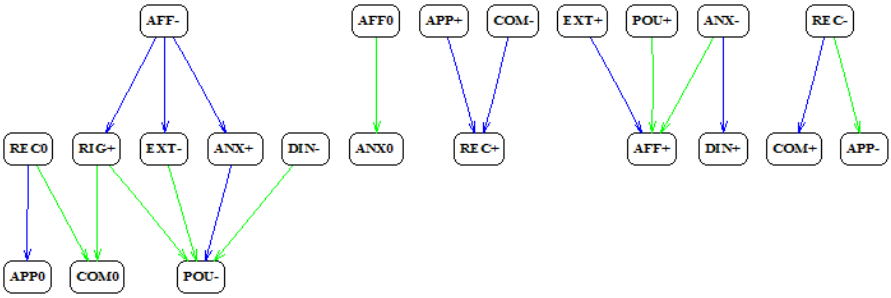


FIG. 4 - *Graphe implicatif* (Modèle entropique) – *[niveau de confiance R100 B90 V80]*

Par exemple, sur ce graphe (figure 5) :

- L'implication : AFF- / RIG+ / COM0 , nous permet de détecter un comportement rigoureux fondé sur l'incertitude, avec des réactions de « défense des normes » si celles-ci sont en péril.
- l'implication : AFF- / RIG+ / POU- , un comportement rigoureux fondé sur l'incertitude et visant la minimisation des risques.
- L'implication : AFF- / EXT- / POU- , un comportement discret et réservé dans les relations, avec la même origine et la même finalité.
- L'implication : AFF- / ANX+ / POU- , un comportement vigilant voire inquiet, avec la même origine et la même finalité.

Cette seule partie des graphes implicatifs a donc permis de mettre au jour, par observation directe dans la population, quatre comportements différents à partir de traits de personnalité qui s'avèrent liés par de très fortes implications (90 ou 80). Si un bénéficiaire vérifie l'un de ces « clignotants », le conseiller dispose immédiatement d'une information

précieuse et fiable. Celle-ci sera encore plus fine et opérationnelle si le bénéficiaire vérifie deux ou plusieurs de ces « clignotants ».

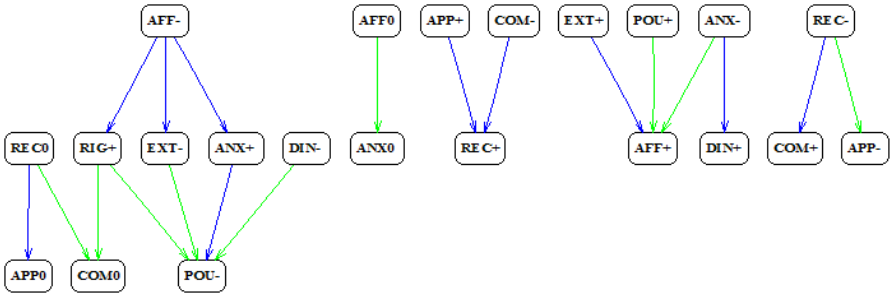
5.2.2 Élaboration de l'outil « module clignotants VAE ».

Sur les 60 règles-comportements (29 + 31, cf 5.2.1) extraites dans les deux populations (VAE réussie et VAE en cours) grâce à la mise en œuvre de CHIC à un niveau d'implication satisfaisant :

- 20 sont vérifiées par 11 à 20% de la population concernée,
- 23 par 5 à 10%,
- 17 par moins de 5%.

Une seule règle est vérifiée dans les deux populations. Le traitement est donc extrêmement différenciant et facilite grandement l'œuvre d'élaboration des experts du groupe de travail CIBC qui, après un examen détaillé, ont retenu 53 de ces 60 règles pour constituer 53 « clignotants », ventilés en 3 thèmes et 2 catégories :

Expression de soi	:	clignotants favorables	défavorables
Activité	:	clignotants favorables	défavorables
Mode relationnel	:	clignotants favorables	défavorables



Degré d'implication [100 90 80]

FIG 5 - *Graphe implicatif (Modèle entropique) – [100 90 80]*

5.2.3 Exemples de comportements retenus.

Comportements retenus
Expression de soi - favorable Equilibre – « Se caractérise par l'égalité d'humeur et des réactions modérées » EXT0 ANX0 AFF0 COM0
Expression de soi - défavorable Inquiétude – « Redoute d'échouer à cause d'une négligence ou d'un oubli » ANX+ AFF- DIN- POU0
Activité - favorable Constance – « Ne se décourage pas face aux obstacles ou à l'adversité » COM0 REA+ APP0
Activité - défavorable Dispersion – « Change de sujet sans se soucier de finir ce qui est commencé » ANX- RIG- DIN+ REA-
Mode relationnel - favorable Convivialité – « Aime entretenir un climat agréable dans son entourage » AFF0 REC+ APP+
Mode relationnel - défavorable Isolement – « Se concentre plus sur son travail que sur ses relations » EXT- RIG+ DIN- POU-

TAB. 3

L'ensemble de ces 53 indicateurs composent le « module clignotants VAE ». Cet outil permet donc un repérage opérationnel des comportements correspondants, en rapprochant un dossier individuel PerformanSe-ECHO du référentiel ainsi constitué.

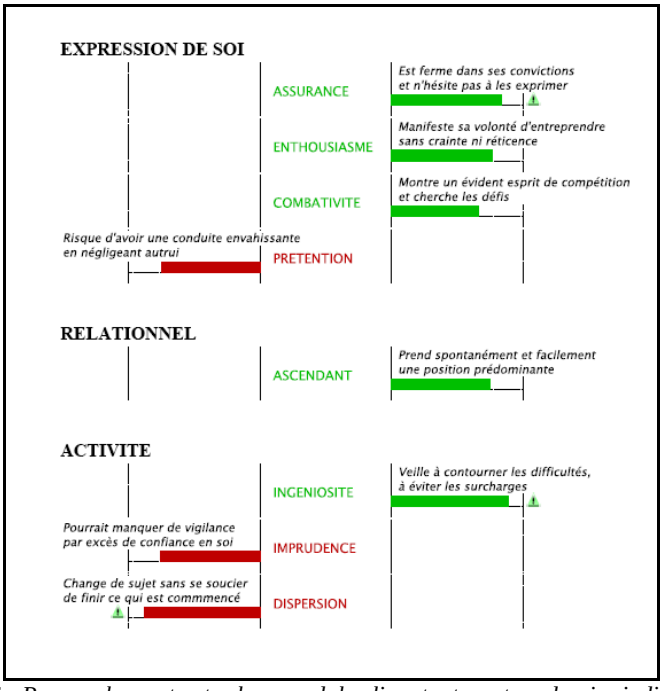


FIG. 6 - *Rapprochement entre le « module clignotants » et un dossier individuel*

6 Conclusion

Le module clignotant rend compte de 53 comportements, leviers et freins. La part de l'exploration des données dans cette élaboration est très importante dans la mesure où il faut construire cet outillage en adéquation la plus fine possible avec les caractéristiques de la population concernée. Trois angles sont à considérer de manière complémentaire dans la constitution d'un tel outil :

1. Une recherche approfondie, pour détecter le maximum de comportements favorables ou défavorables, car l'outil se doit d'être opérationnel quelque soit le cas individuel traité.
2. L'observation d'une population de référence, sans tomber dans la vision de masse car au final c'est bien un rapprochement individuel qui est opéré. Un comportement fréquent et très typé peut être trivial pour les praticiens alors qu'un comportement rare fait sens car, s'il survient, il est critique et doit absolument être accompagné.
3. Une exhaustivité mesurée et pertinente, pour ne pas noyer le praticien sous des comportements observés mais triviaux ou peut significatifs par rapport aux objectifs de l'outil.

Ce triple réseau d'objectifs s'il est contraignant n'en est pas moins indispensable pour élaborer un référentiel alliant pertinence et efficacité. La méthode CHIC s'avère particulièrement satisfaisante dans cette application. Elle permet en effet de dépasser, grâce à

la découverte directe de règles d'implication, les effets de masse tenant souvent au traitement par la statistique descriptive, et grâce au recours à une base de données, les inférences émanant d'une « expertise » *a priori.*

Un premier retour d'expérience est disponible depuis la mise en ligne de cet outil et sont utilisation systématique dans les phases de diagnostic préalable à une démarche de VAE. On observe des résultats significatifs pour les CIBC équipés de la Région Midi-Pyrénées équipés :

- 70 % de validation totale.
- 27 % de validation partielle.
- 3 % sans aucune validation.

Parmi ce public, 78 % a retrouvé un emploi en CDI avant même l'obtention du diplôme ou du titre. Alors que, au niveau régional, 13 % seulement des candidats non accompagnés obtiennent une validation totale. Sans pouvoir inférer complètement la relation qui existe entre ces méthodes d'accompagnement et ces résultats, nous pouvons supposer qu'elles y contribuent en alertant sur les comportements favorables ou défavorables.

Références

Allport, G. W. (1937). Personality *: a psychological interprétation*. New York : Henry Holt, Rinehart, and Winston.

Briand H., Baquedano S., Fleury L., Philippé J., (1993), Comment une démarche de validation sur une population importante conduit à confirmer un modèle comportemental et à l'enrichir, Colloque international, *Techniques Psychologiques de l'évaluation des personnes*, Paris, 1993.

Costa P.T., McCrae R.R. (1996), Toward a new generation of personality theories: Theoretical contexts for the five-factor model. Guilford Paunonen, New York

Couturier R. (2001), Traitement de l'analyse statistique implicative dans CHIC, *Journées sur la fouille des données par la méthode d'analyse implicative*, pp 33-55

Gras R., (1996), *L'implication statistique : une nouvelle méthode exploratoire de données*, Grenoble : La Pensée sauvage

Gras R., Briand H., Peter P., Philippé J., (1996), Implicative Statistical Analysis, 5th Conference of the International Federation of Classification Societes, Kobe Japan

Gras R., Kuntz P., Couturier R., Guillet F., (2001), Une version entropique de l'intensité d'implication pour les corpus volumineux, *Extraction des Connaissances et Apprentissage* (ECA), vol. 1, n° 1-2, pp 69-80, Hermès Science Publication

Gras R., Peter P., Baquedano S., Philippé J. ,(2003), Structuration de comportements de réponse à un questionnaire par des méthodes multidimensionnelles, *Extraction des Connaissances et apprentissage*, Hermès, , 307-312

Guingouain, G. (1999). *Psychologie sociale et évaluation*. Paris : Dunod

Hall. E.T. (1971), *La dimension cachée*. Paris : Seuil,.

Hogan, R. T., De Fruyt, F., & Rolland, J. P. (2006). Validité et intérêt des méthodes d'évaluation de la personnalité à des fins de sélection : une perspective de psychologie appliquée aux problématiques des entreprises. *Psychologie francaise*, 51(3), 245-264.

Laberon, S., Lagabrielle, C., Vonthron, A. M. (2005). Examen des pratiques d'évaluation en recrutement et en bilan de compétences. *Psychologie du travail et des organisations*, 11(1), 3-14.

Lerman I.C., (1981), *Classification et analyse ordinale des données*, Paris : Dunod

Lewin, K., Lippitt, R. (1938) An experimental approach to the study of autocracy and democracy. A preliminary note. *Sociometry*, (1), 292-300.

Lewin, K., Lippitt, R., White, R. (1939) Patterns of aggressive behaviour in experimentally created social climates. *Journal of Social Psychology*, (10), 271-299.

McCrae R.R , John O.P.(1992), An introduction to the fivre-factor model and its applications. *Journal of Personality*, 175-215

McCrae, R. R., Costa, Jr., Paul, T. (1997). Personality Trait Structure as a Human Universal. *American Psychologist*, 52(5), 509-516.

Peter P., Gras R., Philippé J. et Baquedano S., (2001), L'analyse implicative pour l'étude d'un questionnaire de personnalité, Proceedings des Journées Extraction et Gestion des Connaissances EGC.'2001 de Nantes (18-19 janvier 2001), Hermès, p 181-187

Watzlawick. P. (1980), *Le Langage du changement.* Paris : Collection Points, Editions du Seuil,

Summary

A group of French CIBCs (French inter-institutional skills assessment centers) from the Midi-Pyrenees' region has developed an individual coaching method, as part of its process concerning the validation of work experience (the VAE, or Validation des Acquis de l'Expérience, is a French procedure that allows any educational institution to grant degrees on work experience). PerformanSe, as a publisher of software tools, has been approached in order to propose a support tool that could assist this process. The method chosen consists in designing an "alert" tool that could indicate the favourable and unfavourable behaviours during a process of validation of work experience

Thème 3 : Applications à la sociologie

Chapitre 7 : Analyse statistique implicative des transitions professionnelles dans la Genève du 19e siècle

Gilbert Ritschard, Matthias Studer et Michel Oris

Laboratoire de démographie et d'études familiales, Université de Genève
gilbert.ritschard@unige.ch, matthias.studer@unige.ch, michel.oris@unige.ch,
http://www.unige.ch/ses/demog

Résumé. Cet article reprend l'analyse statistique implicative de la dynamique socioprofessionnelle dans la première moitié du 19e à Genève que nous avons présentée aux rencontres ASI4 (Oris et Ritschard, 2007) et la compare à une analyse supervisée des dissimilarités entre transitions. Les données considérées résultent de l'appariement deux à deux de 6 recensements. Plus précisément, nous considérons le groupe socioprofessionnel (GSP) des individus retenus et son changement entre deux recensements successifs. Nous nous intéressons aux types de transition (stable, devenir actif, cesser l'activité, ...) ainsi qu'aux nouveaux venus (immigrés et naissances) et disparus (émigrés et décédés). L'analyse statistique implicative donne une vision synthétique des liens entre ces dynamiques et les GSP concernés, ainsi qu'avec un certain nombre de variables démographiques et culturelles (sexe, âge, état-civil, religion). Elle met en lumière notamment des polarisations autour de variables clé. L'analyse des dissimilarités permet quant à elle de segmenter la population en groupes homogènes en fonction des caractéristiques démographiques et culturelles. Le recours à l'intensité d'implication pour identifier les transitions typiques des groupes ainsi obtenus s'avère une précieuse aide à l'interprétation et donne les éléments nécessaires à la comparaison avec les résultats du graphe implicatif.

1 Introduction

Nous reprenons et étendons ici le travail (Oris et Ritschard, 2007) présenté aux 4èmes rencontres d'Analyse statistique implicative, ASI4, où nous rapportions une expérience d'analyse statistique implicative dans un contexte de démographie historique. L'étude porte sur le recrutement et le renouvellement des groupes socioprofessionnels à Genève (Oris et al., 2006) entre 1816 et 1843. Plus précisément, nous nous intéressons aux changements individuels de statut socioprofessionnel que l'on observe entre deux recensements successifs, l'objectif étant de mieux comprendre comment ces changements ou transitions sont liés aux caractéristiques démographiques des individus ainsi que leur impact sur la démographie de la Cité de Genève.

Dans Oris et Ritschard (2007), nous avons montré comment l'analyse statistique implicative, ASI, vient enrichir les connaissances obtenues par des méthodes démographiques et d'analyse de données classiques, l'intérêt de l'ASI résidant essentiellement dans sa force synthétique. Elle donne en effet une vision globale aisément compréhensible d'un ensemble d'associations et permet ainsi, par rapport notamment aux approches qui se focalisent sur l'explication d'une caractéristique particulière, de mettre en lumière des polarisations tout à fait éclairantes. Nous complétons ici cette comparaison en proposant une analyse originale des dissimilarités entre les couples d'états d'origine et de destination qui caractérisent les transitions. Cette étude exploite les principes de l'analyse de variance pour générer un arbre de segmentation des transitions observées en fonction de caractéristiques démographiques (âge, sexe, état-civil) et culturelles (religion). Elle n'est pas totalement déconnectée de la statistique implicative, puisque nous exploitons l'intensité d'implication pour identifier les transitions caractéristiques dans chaque nœud de l'arbre.

L'article est organisé comme suit. Nous commençons à la section 2 par rappeler brièvement le contexte historique de la Cité de Genève dans cette première moitié du 19e. La section 3 décrit ensuite l'origine et l'organisation des données utilisées. L'analyse statistique implicative fait l'objet de la section 4 tandis que la section 5 présente l'analyse des dissimilarités des transitions et démontre l'intérêt de l'intensité d'implication comme aide à l'interprétation des résultats obtenus. Enfin, nous concluons à la section 6 en soulignant les apports respectifs de l'ASI et de l'analyse des dissimilarités pour notre problématique, mais aussi quelques-unes de leurs limites.

2 Le contexte historique genevois

La Réforme protestante a fait de Genève la "Rome calviniste" qui, plus de deux siècles durant, a vécu telle une forteresse menacée au milieu d'un environnement essentiellement catholique. En 1816, au terme d'une annexion à la France et suite au Traité de Vienne de 1815, la république urbaine devient un canton doublement mixte grâce à la fusion avec des municipalités rurales catholiques qui appartenaient auparavant à la France ou au royaume de Piémont-Sardaigne. A Genève intra-muros, la proportion de catholiques au sein de la population passe de 11% en 1816 à 28% en 1843, puis 46,4% en 1900. Cela a profondément marqué les esprits des "vieux Genevois" qui percevaient leur identité menacée par l'afflux d'immigrés transformant subitement en un ensemble multiculturel un bloc monolithique séculaire (Herrmann, 2003). Le choc a été d'autant plus rude que l'essentiel de la croissance démographique a été assurée par le solde migratoire durant toute la première moitié du 19e siècle. Déjà au 18e, les pratiques de contrôle des naissances s'étaient diffusées des élites aux masses populaires. Les couples genevois mariés entre 1800 et 1850 n'ont eu en moyenne (Schumacher, 2002) que 2,32 enfants ! Parallèlement, les freins traditionnels restent serrés au maximum, avec un âge moyen des femmes au 1er mariage de 28 ans et un célibat définitif de presque 20%. Genève est à cette époque probablement le seul endroit au monde où cumulent leurs effets au sein du même régime démographique et le malthusianisme traditionnel et le néo-malthusianisme moderne. Dès lors, bien que la mortalité infantile ait chuté de 200 à 100/130 ‰ entre la deuxième moitié du 18e et la première moitié du 19e siècle, la balance naturelle présente un excès des naissances sur les décès qui n'atteint que 557 unités entre 1806 et 1850. La croissance de la population, qui passe de 21327 à 31200 au cours de cette période, est due à plus de 90% à

l'excédent des immigrations sur les émigrations. Les structures de la population genevoise ont donc été bien plus transformées par les migrations que ne le suggérerait une expansion somme toute modeste, dans le contexte global d'explosion urbaine qui a caractérisé le 19e siècle.

Il est vrai que l'économie ne pouvait guère soutenir une progression plus marquée. La ville s'est dotée de la structure d'activités que lui imposait son insularité religieuse, son absence de domination sur l'arrière-pays catholique. En particulier, elle a développé la Fabrique, qui "est l'ensemble des arts et des artistes qui concourent à la création des montres et des bijoux. Le travail en est réparti en une multitude de petits ateliers" (Babel, 1942b, p 13). Après une période faste entre 1750 et 1792, la ville a beaucoup souffert durant la période française (Guichonnet, 1974, p 273). Le marasme économique a duré jusqu'en 1819, voire 1822 (Babel, 1942a, p 44). Le textile (fabrication d'indiennes) n'ayant pas su résister au retour de la concurrence anglaise après 1816, la Fabrique s'est encore plus imposée comme l'activité dominante. Elle emploie 5000 personnes en 1828 (Guichonnet, 1974, p 287) ; 35,4% des hommes qui se marient à Genève entre 1822 et 1845 travaillent dans ce secteur. Cette domination est dangereuse car industrie d'exportation par excellence, l'horlogerie est sensible aux fluctuations politico-guerrières qui affectent ses marchés extérieurs. Ainsi, si la période qui s'étend entre 1830 et 1845 fut prospère pour la Fabrique et l'ensemble de l'économie urbaine (Guichonnet, 1974, p 288), la crise qui a traversé le continent (1845-1847) sera violemment ressentie. Avec l'arrivée des radicaux au pouvoir en 1846, la ville connaît de profonds changements dont la destruction des murailles et la modernisation, en particulier le développement d'industries performantes (Perroux, 2006).

La période que nous étudions est donc à la croisée des temps. Entre 1816 et 1843, Genève reste une ville tranquille dominée par un "conservatisme éclairé" (Dufour, 1997, p 99), soucieux de préserver le modèle social ancien alors même que les flux migratoires commencent à transformer les structures sociales en profondeur.

3 Les données

Les données étudiées ont été collectées dans le cadre d'une recherche visant à reconstruire les parcours de vie dans la cité de Genève au 19e et réalisée avec le soutien financier du Fonds national suisse pour la recherche scientifique. Elles sont tirées des archives genevoises et proviennent plus particulièrement de six recensements de la population genevoise de 1816, 1822, 1828, 1831, 1837 et 1843. Plus précisément, les données de ces recensements n'étant évidemment pas informatisées, nous n'avons extrait que les informations relatives aux individus dont le patronyme commence par la lettre "B", ce qui représente environ 12,5% de la population. Au total, en cumulant les 6 recensements, ce ne sont pas loin de 30'000 notices individuelles qui ont été relevées. La technique d'historiens consistant à échantillonner les "B" est neutre socialement et ethniquement, et s'apparente à un tirage aléatoire représentatif de la population (Bardet, 1983). Elle simplifie grandement la collecte à partir d'archives dans la mesure où elle permet de dépouiller plusieurs sources distinctes et d'y retrouver les mêmes individus.

La particularité de nos données est en effet que disposant et d'un échantillon alphabétique et du nom et de caractéristiques individuelles comme l'année de naissance, le sexe et même l'adresse, nous avons dès lors pu apparier les notices individuelles des recensements successifs. C'est cet appariement que nous exploitons pour analyser les entrées et sorties de la vie active et les mobilités socioprofessionnelles entre deux états de la population. Les cas qui n'ont pu être

Statuts sociaux		Groupes socioprofessionnels	
ss_inc	Inconnu	*gsp_inac*	Inactif
ss_nqua	Manuel sans qualification	*gsp_nqua*	Sans qualification
ss_art	Manuel qualifié	*gsp_art*	Artisan
ss_colb	Col blanc	*gsp_hor*	Horloger
ss_pmb	Petite et moyenne bourgeoisie	*gsp_com*	Commerçant
ss_eli	Elites	*gsp_serv*	Services privés et publiques

TAB. 1 – *Liste des groupes socioprofessionnels et statuts sociaux*

appariés fournissent également des informations précieuses. Un individu présent à un recensement qui ne l'est plus au recensement suivant a nécessairement quitté la cité, soit parce qu'il a émigré soit parce qu'il est décédé. De même, un individu qui n'était pas présent au recensement précédent et qui n'est pas né dans l'intervalle est nécessairement un immigré. Nonobstant la perte des "feux-follets", c'est-à-dire de ceux qui n'ont fait que passer par Genève entre deux opérations censitaires, c'est un tableau dynamique assez complet qui nous est ainsi dessiné.

L'analyse est centrée sur les recensements séparés de 6 ans. On considère ainsi les états au temps t, soit 1816, 1822, 1831 et 1837, et au temps $t + 6$ (1822, 1828, 1837, 1843), ainsi que l'évolution de ces états entre t et $t + 6$.

Parmi les informations collectées, nous avons dénombré environ 1200 métiers qui ont été réorganisés d'une part en groupes socioprofessionnels, et d'autres part en statuts sociaux. Le tableau 1 liste les catégories retenues pour chacun de ces regroupements. Pour distinguer les états au début de l'intervalle de ceux à la fin nous préfixons les notations par '*t_*' pour indiquer le début. Par exemple '*t_gsp_hor*' correspond à horloger en t et '*gsp_hor*' à horloger en $t + 6$. Le tableau 2 indique comment les cas retenus se répartissent selon ces catégories au temps t. Notons la distinction entre le statut social 'non qualifié' et le groupe professionnel 'non qualifié'. Un garçon boucher ou une fille de boutique par exemple ont un statut social 'non qualifié' mais appartiennent au GSP des artisans. Les transitions qui nous intéressent sont définies à partir des groupes socioprofessionnels et sont récapitulées au tableau 3 :

L'objectif de l'étude étant de comprendre comment le recrutement et la mobilité socioprofessionnels sont liés aux facteurs démographiques nous avons retenus également comme variables pour notre analyse l'âge, le sexe et l'état civil. Par ailleurs, intéressé par l'impact

Statuts / Groupes socioprofessionnels	Inconnu	Manuel ss qual.	Manuel qualifié	Col blanc	P.M.B.	Elite	Total
Inactif	4467	23	0	79	1	344	4914
Sans qualification	274	1672	96	118	3	0	2163
Horlogerie	0	71	1330	0	213	0	1614
Artisan, manuel qualifié	0	173	1527	3	80	0	1783
Commerce	0	112	64	21	537	7	741
Services publics et privés	0	28	18	37	156	82	321
Total	4741	2079	3035	258	990	433	11536

TAB. 2 – *Croisement des groupes socioprofessionnels avec les statuts sociaux (au temps t)*

Transition	(désignation)	GSP en t	GSP en $t+6$	autre condition
reste inactif	(inactif)	inactif	inactif	
devient actif	(nv_actif)	inactif	actif	
stable	(stable)	actif	actif	$\text{GSP}(t) = \text{GSP}(t+6)$
mobile	(mobile)	actif	actif	$\text{GSP}(t) \neq \text{GSP}(t+6)$
cesse l'activité	(retraite)	actif	inactif	
nouveau venu	(nv_venu)	non présent	présent	
disparu	(disparu)	présent	non présent	

TAB. 3 – *Les types de transitions*

possible de la montée du catholicisme qui, comme mentionné plus haut fait plus que doubler entre 1816 et 1843, nous avons également inclus la religion dont nous distinguons trois modalités : protestant, catholique et autre. S'agissant de l'état civil, le nombre de divorcés étant très faible (< 10) nous ne considérons que les états célibataire, marié et veuf. Pour la religion comme pour l'état-civil nous retenons l'état au début t de l'intervalle censitaire à l'exception des nouveaux venus pour lesquels nous ne disposons que de l'état en $t + 6$. Pour l'âge, nous retenons celui du milieu de l'intervalle, soit l'âge en $t + 3$. Pour les analyses de statistique implicative, l'âge a été discrétisé en 3 classes (en minimisant la variance intra groupe) : *age*1 à moins de 16 ans, *age*2 de 16 à 41 ans et *age*3 pour les plus de 41 ans. On distingue également deux périodes, l'une couvrant les transitions de 1816 à 1822 et de 1822 à 1828 et la deuxième celles de 1831 à 1837 et de 1837 à 1843.

On trouvera dans Oris et Ritschard (2007) l'analyse de la structure de la population et des transitions avec les outils classiques d'analyse démographique. Cette analyse renseigne sur les taux d'entrée et de sortie de la vie active, ainsi que sur les arrivées et départs de la Cité. Elle fait ressortir notamment l'importance des flux migratoires et la forte croissance du secteur de l'horlogerie, mais reste essentiellement limitée à la population dans son ensemble. Le calcul de tables, avec leurs probabilités, leurs courbes de survie, leurs âges moyens, serait a priori aussi voire plus profitable s'il était fait par groupe d'âges, sexe, état-civil et religion par exemple. Mais deux obstacles se présentent. L'un est la dispersion des effectifs, surtout lorsqu'ils sont distribués par groupes d'âges quinquennaux. L'autre est la multiplicité de résultats qui en résulterait, tant à cause de la division en six groupes socioprofessionnels que de la multiplicité des dynamiques. Il y a donc nécessité d'approches plus synthétiques.

Une première possibilité est la modélisation statistique. Les régressions logistiques présentées dans Oris et Ritschard (2007) permettent ainsi de nuancer l'information des tables par la prise en compte de variables démographiques et de la religion. Elles n'en produisent pas moins une avalanche de chiffres à peu près aussi déconcertante qu'avec les tables. L'analyse des correspondances multiple est une autre alternative considérée. Elle donne une vision globale des liens entre GSP, types de mobilité et facteurs socio-démographiques. Elle reste sujette cependant aux difficultés d'interprétation des méthodes factorielles.

Nous considérons ci-après deux approches qui nous semblent pouvoir apporter un éclairage synthétique novateur des liens entre démographie et dynamique socioprofessionnelle. La première s'appuie sur la statistique implicative comme outil d'analyse de données (Gras et al.,

1996). Ceci permettra d'illustrer comment la statistique implicative vient enrichir aussi bien la vision partielle des tables démographiques, que la vue globale résultant d'une analyse factorielle. La seconde qui est une contribution originale est une analyse par arbre des dissimilarités des transitions.

4 L'analyse statistique implicative

Nous nous proposons donc d'examiner nos données avec les outils de la statistique implicative, soit plus particulièrement ceux mis à disposition par le logiciel CHIC (Couturier et al., 2006), à savoir l'arbre de similarité fondé sur la vraisemblance du lien de Lerman et al. (1981), l'arbre cohésitif orienté (Gras et Kuntz, 2006) fondé sur l'intensité implicative de Gras et al. (1996), et le graphe implicatif.

La figure 1 montre un premier graphe d'implication obtenu en n'incluant dans l'analyse que les transitions et les groupes socioprofessionnels d'origine et de destination. Le graphe a été obtenu avec des seuils relativement bas. Toutefois, cela ne remet pas en cause sa pertinence statistique. En effet, comme nous avons utilisé la mesure entropique d'implication, les seuils ne doivent pas être interprétés comme des significations statistiques. Avec la mesure classique non entropique, les relations indiquées ici sont d'ailleurs toutes significatives à des seuils supérieurs à 95%, mais se trouvent être noyées dans une quantité d'autres relations dont l'interprétation est moins intéressante.

Démonstration de l'efficience de la méthode, la figure 1 montre des relations triviales découlant des définitions même du tableau 3, à savoir que ceux qui restent inactifs sont inactifs en t et en $t+6$, que les retraités deviennent inactifs en $t+6$ alors que les nouveaux actifs étaient inactifs en t. Les autres relations sont plus instructives. Elles nous indiquent d'une part que les manuels non qualifiés en $t+6$ et dans une moindre mesure les artisans sont essentiellement constitués de nouveaux venus à Genève, d'autre part que quand on appartient à ces mêmes groupes en t, on a de forte chances de quitter la cité dans les 6 ans suivants. L'importante rotation de la population semble donc concerner principalement ces groupes de manuels qualifiés et non qualifiés.

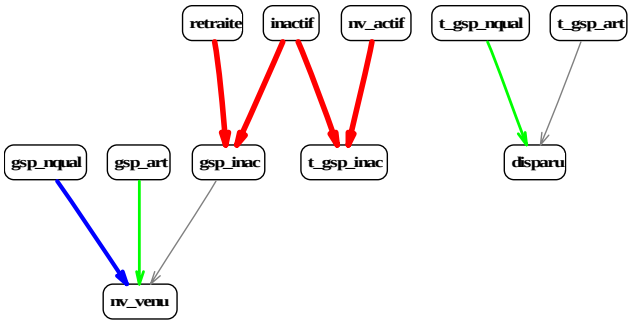


FIG. 1 – *Transitions et groupes socioprofessionnels. Mesure entropique, seuils 99%, 81%, 63%, 58%.*

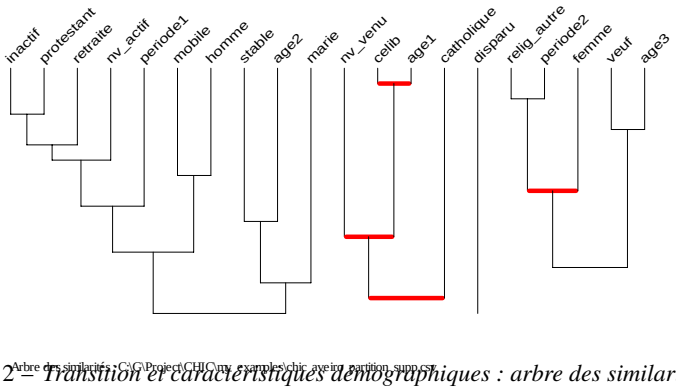


FIG. 2 – *Transition et caractéristiques démographiques : arbre des similarités.*

Les figures 2, 3 et 4 montrent respectivement l'arbre des similarités, l'arbre cohésitif et le graphe implicatif obtenus en incluant les transitions, les variables démographiques et la religion. L'arbre des similarités fait ressortir clairement quatre groupes. En partant de la gauche, le premier correspond aux enracinés qui restent dans la cité ; il englobe essentiellement les familles locales. Leurs membres présentent le paradoxe apparent d'être professionnellement et stables et mobiles, mais la prévalence de la mobilité dans ce groupe s'explique essentiellement par l'entrée des jeunes sur le marché du travail (variable "nv_actif"). Ces types de transitions se recoupent essentiellement avec les caractéristiques protestant, homme et marié, ce qui est cohérent. Le second groupe comprend les nouveaux arrivants qui sont plutôt jeunes — et donc célibataires — et catholiques, et le troisième ceux qui quittent la cité. Ces disparus ne se regroupent avec aucune caractéristique démographique. Enfin, le dernier groupe n'est guère intéressant car il ne comporte aucune variable de transition et présente des caractéristiques hétérogènes.

Le principal enseignement de l'arbre cohésitif est qu'il n'y a pas de méta-règle intéressante. La seule qui apparaît, à savoir 'nouveau venu' $\Rightarrow$ ('jeune' $\Rightarrow$ 'célibataire'), n'est rien d'autre en effet que la règle simple 'nouveau venu' $\Rightarrow$ 'jeune', puisque les jeunes sont évidemment célibataires, en particulier dans le contexte genevois de mariage tardif décrit plus haut.

Le graphe implicatif 4, obtenu comme celui de la figure 1 avec le critère entropique et des seuils relativement bas, est plus instructif. On y observe quatre modalités qui polarisent les effets des types de transition. Ce sont 'protestant', 'femme', 'homme' et 'célibataire'. Les nouveaux venus, les inactifs et ceux qui commencent une vie active sont principalement célibataires, les deux derniers groupes étant de jeunes célibataires. [1] L'inactivité, le début d'activité et la cessation d'activité sont plus le fait des femmes, tandis que la stabilité tout comme la mobilité socioprofessionnelle sont l'apanage des hommes. Quant à l'attribut 'protestant' il est associé à toutes les transitions internes, c'est-à-dire toutes sauf les nouveaux venus et les disparus. Ceci indique que si l'on trouve une majorité de protestants dans tous les groupes de transitions non migratoires, les fidèles de Calvin sont moins concernés par les mouvements

1. L'existence des fermetures transitives que nous commentons a bien entendu été vérifiée. Ainsi, les implications directes des inactifs et nouveaux venus sur célibataires ont toutes deux des intensités supérieures à 75%. Nous avons choisi de ne pas les représenter dans le graphe pour des raisons de lisibilité.

migratoires. On note, en particulier, que les catholiques sont plutôt constitués de nouveaux venus, le seuil de cette relation étant cependant inférieur à 65%. Les implications fortes de 'autre religion' sur les femmes et les célibataires sont plus difficiles à interpréter en raison de la nature résiduelle de cette modalité de la variable religion.

Pour tenter de voir comment les relations mises en évidence sont liées aux statuts sociaux, nous avons introduits ceux-ci comme variables supplémentaires ce qui permet d'obtenir avec CHIC des mesures de leur typicalité pour les implications qui nous intéressent. La typicalité fournie peut être vue comme la probabilité que l'implication soit indépendante de la variable supplémentaire. Ainsi, une typicalité inférieure à 5% indique que l'implication est caractéris-

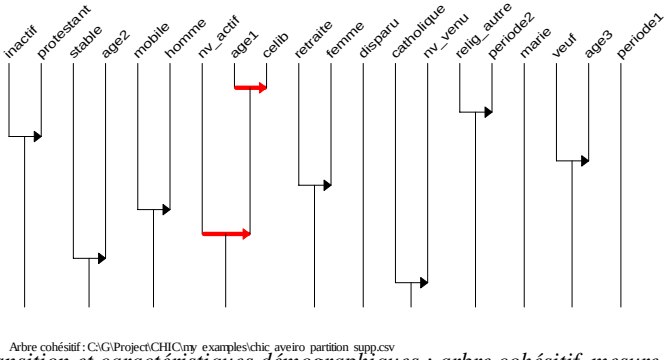


FIG. 3 – *Transition et caractéristiques démographiques : arbre cohésitif, mesure entropique.*

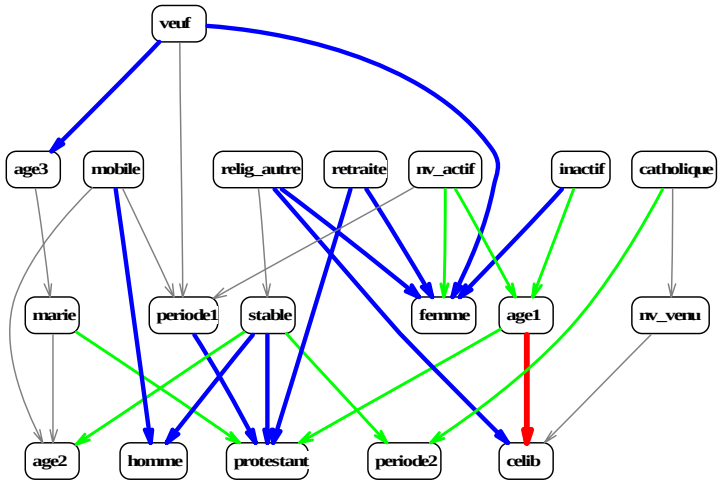


FIG. 4 – *Transitions et variables démographiques : graphe implicatif, mesure entropique, seuils 99%, 75%, 65%, 55%.*

Chemins \ Statuts	Inconnu	Manuel ss qual.	Artisan, manuel qualifié	Col blanc	P.M.B.	Elite
stable ⇒ protestant	.	x	x	x	x	.
stable ⇒ homme	.	.	x	x	.	x
mobile ⇒ homme	.	.	x	x	x	x
nv_actif ⇒ protestant	.	x	x	x	x	.
nv_actif ⇒ célibataire	.	x	x	x	x	.
nv_actif ⇒ age1 ⇒ célibataire	x	.	.	x	.	.
nv_actif ⇒ femme	.	x	.	.	.	.

TAB. 4 – *Typicalité des statuts sociaux pour quelques chemins*

tique des individus ayant le statut social en question.

Par rapport à la figure 4, il est intéressant de noter que l'implication 'stable ⇒ protestant' n'est pas — avec des typicalités (Gras et al., 2006) supérieures à 0.75 — caractéristique des membres de l'élite, ni des individus de statut social inconnu. Toutes les autres typicalités étant quasi-nulles ($< .001$), elle distingue cependant tous les autres statuts. L'implication 'stable ⇒ homme' est, quant à elle, typique des artisans, des cols blancs et de l'élite, mais ne l'est ni de la petite et moyenne bourgeoisie, ni des non qualifiés. Il en va de même du fait que la mobilité socioprofessionnelle soit l'apanage des hommes, si ce n'est que cette dernière relation est également typique parmi la petite et moyenne bourgeoisie. Si l'on considère maintenant les implications importantes de 'nouvel actif' sur 'célibataire' et 'protestant', les indices de typicalités montrent qu'ils caractérisent les non qualifiés, les artisans, les cols blancs et la petite et moyenne bourgeoisie, les statuts inconnus et les cols blancs étant typiques du chemin qui passe par la catégorie 'jeunes' (age1). Enfin, il est instructif, quoique attendu, de relever que la prédominance de femmes parmi la population qui débute (ou retourne à) la vie active est surtout propre au bas de l'échelle des statuts sociaux. Le tableau 4 récapitule les typicalités relevées.

Les figures 5 et 6 ont été obtenues en incluant comme variables régulières les groupes socioprofessionnels et les statuts sociaux respectivement du début t et de la fin $t + 6$ de l'intervalle. On retrouve évidemment les quatre mêmes modalités de polarisation, avec toutefois une polarisation plus forte sur le nœud 'homme' où convergent les quatre GSP (tant en t qu'en $t+6$) du haut de l'échelle, reflétant ainsi la dominance masculine dans les activités valorisantes. Par contraste, on voit que seul le GSP non qualifié converge sur 'femme'.

Il est intéressant de relever ici les différences entre les effets des statuts sociaux en tant que variables propres, et leurs typicalités discutées plus haut. Ici nous observons les effets sur les autres variables indicatrices, tandis que les typicalités rendent compte de l'effet sur les implications. En l'absence de méta-règles pertinentes, les typicalités et les effets comme variables régulières restent très similaires : quand une variable est typique d'une implication elle converge en général aussi sur la conclusion de la règle. Ainsi, par exemple, les statuts sociaux 'petite et moyenne bourgeoisie' et 'horlogers' qui sont typiques de l'implication 'mobile ⇒ homme', convergent sur 'homme'. Dès lors, comme à nouveau aucune méta-règle pertinente ne ressort des arbres cohésitifs (non montrés) correspondant aux graphes des figures 5 et 6, il ne nous a pas paru opportun de traiter par exemple les GSP en supplémentaires.

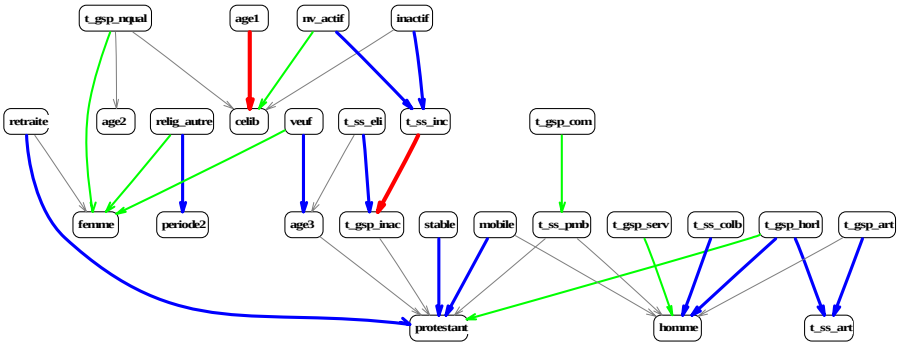


FIG. 5 – *Graphe implicatif, groupe socioprofessionnel et statut social en t. Seuils 99%, 90%, 85% et 80%.*

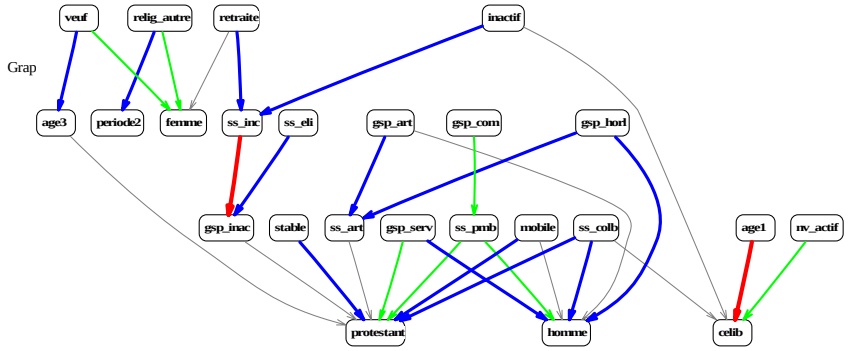


FIG. 6 – *Graphe implicatif, groupe socioprofessionnel et statut social en $t + 6$. Seuils 99%, 90%, 85% et 80%.*

5 Segmentation des transitions et intensité d'implication

L'analyse statistique implicative, tout comme l'analyse des correspondances menée dans Oris et Ritschard (2007), consiste essentiellement en une étude des liens entre modalités des variables. Elle contraste de ce point de vue avec les approches plus axées sur l'étude des liens entre unités statistiques comme les techniques visant à mettre en évidence des typologies par le biais de classification supervisées ou non des individus. Nous proposons de comparer ici les résultats obtenus avec ceux d'une analyse supervisée des dissimilarités des cas réalisée selon les techniques discutées dans Studer et al. (2009). Ce sera aussi l'occasion de démontrer l'intérêt de l'intensité d'application comme aide à l'interprétation des résultats de cette analyse.

Il s'agit dans un premier de temps de déterminer les dissimilarités entre chaque paire de cas, les cas étant dans notre contexte des transitions entre t et $t + 6$. Notons que comme la

base comprend plus de 18'000 transitions, cela représente plus de 162 millions de dissimilarités. Pour évaluer ces dissimilarités, le profil de chaque cas est considéré en termes des groupes socio-professionnels en t et $t + 6$. Chaque cas est ainsi représenté par le vecteur des 14 variables indicatrices de type *t_gsp* et *gsp*. Les dissimilarités retenues sont les distances euclidiennes entre profils selon les scores de l'analyse des correspondances multiple de ces variables. Avec cette façon de procéder, deux cas qui diffèrent sur des caractéristiques rares sont plus dissemblables que deux cas qui diffèrent sur des caractéristiques fréquentes.

Le principe de l'analyse est ensuite celui des arbres de décisions. Il consiste à construire une arborescence en cherchant à chaque pas la meilleure façon d'éclater un nœud selon les modalités d'un prédicteur. La particularité est ici le critère utilisé, à savoir retenir l'éclatement qui explique la plus grande part de la diversité ou pseudo-variance des transitions. On exploite pour cela des relations qui permettent de calculer les pseudo-variances expliquées et résiduelles à partir des dissimilarités entre paires de transitions, et l'éclatement est sélectionné en fonction de la significativité de la part expliquée que l'on évalue par un test de permutation (voir Studer et al., 2009, pour les détails).

La figure 7 montre l'arbre obtenu. Pour l'interpréter, nous nous appuyons sur le concept d'intensité d'implication et visualisons dans les nœuds les transitions d'implication statistiquement significative. Plus précisément, l'implication considérée est celle de la règle dont la prémisse est définie par la branche menant au nœud et la conclusion la transition en question. Les transitions représentées sont ainsi les plus caractéristiques du nœud au sens où ce sont celles dont la fréquence relative des contre-exemples (total d'autres transitions) au sein du groupe est la plus significativement inférieure à la proportion d'autres transitions dans le nœud initial. Dans chaque nœud les transitions d'implication significative sont ordonnées de bas en haut par ordre décroissant de significativité. Au nœud initial, où toute les implications sont nulles par définition, sont reproduits les 10 transitions les plus fréquentes.

Cet arbre met en évidence trois groupes de transitions caractéristiques, à savoir les transitions entre les GSP inconnu et non actif qui, de façon attendue, sont essentiellement le fait de jeunes (moins de 16 ans en $t + 3$), les transitions entre inconnu et sans qualification qui caractérisent les femmes adultes à l'exception notable des femmes protestantes de plus de 40 ans, et les transitions entre artisan, horloger et inconnu qui sont caractéristiques des hommes adultes. On peut relever que la stabilité dans la Fabrique (transition d'horlogerie vers horlogerie) concerne typiquement les hommes non catholiques, et que la stabilité dans le domaine du commerce est caractéristique des hommes non catholiques mariés. On peut relever également les importants mouvements de populations qu'illustrent les pavés bleus (inconnu) qui indiquent des immigrés lorsqu'ils sont à gauche (excepté pour les jeunes où il peut s'agir de naissances) et des émigrés (ou décès) lorsqu'ils sont à droite.

Si l'on compare à présent les enseignements de cette analyse des dissimilarités avec ceux de l'analyse de la section précédente, on peut relever leur similitude. Ainsi, la mobilité, c'est-à-dire la transition entre deux GSP actifs apparaît comme le fait d'hommes surtout tant dans le graphe de la figure 4 que dans l'arbre de la figure 7, où aucune des transitions typiques des femmes et des enfants ne se fait entre deux GSP actifs différents. Il en est de même de la stabilité, à savoir la transition d'un GSP actif à ce même GSP, qui au vu de la Figure 4 est l'apanage des hommes protestants, ce qui est confirmé par l'arbre puisque, si l'on excepte la stabilité dans l'état non-qualifié, les autres formes de stabilité sont typiques des hommes et essentiellement des hommes non catholiques. De façon générale, il ressort de l'arbre que

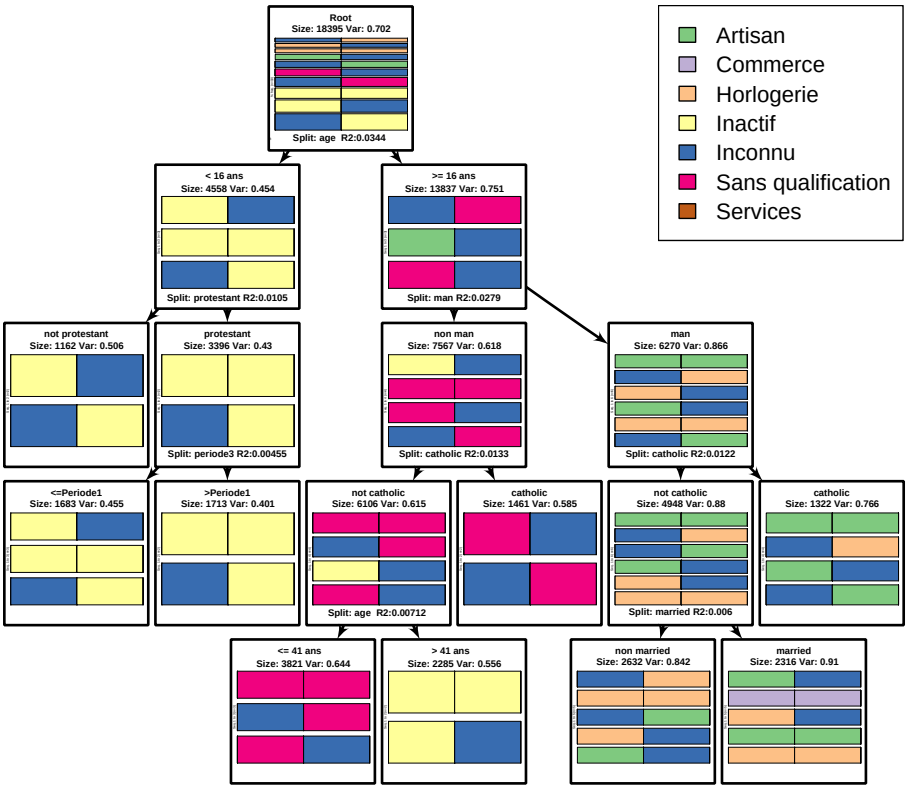


FIG. 7 – *Analyse de la diversité des transitions.* R^2 *global* $= 0.073$.

les catholiques, qu'il s'agisse d'hommes, femmes ou enfants, sont caractérisés par des trajectoires migratoires turbulentes, leurs transitions typiques comprenant presque toutes "inconnu" comme origine ou destination. Notons que si les analyses avec CHIC ont mis en évidence une relative transparence des catholiques signifiant qu'ils se fondent dans la population, il ressort de l'arbre que l'attribut catholique intervient en interaction avec les variables âges et sexe : les migrants catholiques masculins rejoignent ou quittent le GSP artisan, tandis que les femmes investissent puis abandonnent principalement le groupe des non-qualifiés (essentiellement des domestiques), et les rares enfants de cette confession se raccrochent bien sûr au groupe des inactifs. Ainsi, sans contredire les graphes d'implication, l'arbre apporte un éclairage complémentaire qui permet de nuancer la structure implicative dégagée précédemment.

6 Conclusion

Nous avons présenté la mise en œuvre d'une analyse statistique implicative dans un contexte de démographie historique où il s'agissait d'appréhender l'articulation démographique des dynamiques socioprofessionnelles dans la Genève du 19e siècle. Plus que de tirer une conclusion sur le fonctionnement de la société genevoise de cette époque — que le lecteur intéressé trouvera dans Oris et al. (2006) —, nous aimerions ici faire le bilan de ce que nous ont effectivement apporté d'une part l'ASI et d'autre part l'arbre de dissimilarité des transitions couplé à une interprétation en termes d'intensité d'implication.

S'agissant de l'analyse réalisée avec CHIC, cet apport réside dans la vision synthétique et structurée des relations d'implication entre les multiples caractéristiques considérées. Les tables, taux et quotients utilisés en démographie permettent certes d'aller dans le détail et de quantifier les relations en termes de pourcentage de population concernée. Elles nécessitent cependant de se focaliser sur des aspects spécifiques et conduisent dès que l'on cherche à segmenter la population simultanément selon plusieurs variables de contrôle — sexe, état-civil, classe d'âge quinquennale, religion — à des effectifs souvent bien trop faibles, et sinon à une avalanche indigeste de chiffres. Les régressions logistiques peuvent palier en partie à ces faiblesses en permettant de mesurer les effets propres qui subsistent quand on contrôle pour les facteurs démographiques. L'inconvénient et qu'elles nous obligent à travailler sur des sous-populations, ceux qui restent actifs, ceux qui sont inactifs en début de période, ceux qui restent présents à Genève en fin de période, etc. Dans une optique plus synthétique nous avons également dans Oris et Ritschard (2007) procédé à une analyse factorielle des correspondances multiples. Le résultat s'est avéré éclairant, mais les tendances mises en évidence restent, bien que faisant sens, relativement floues. L'arbre de similarité des modalités offre une vision plus structurée des caractéristiques individuelles qui s'associent. Il serait intéressant, ce que nous n'avons ni eu le temps ni la place de faire ici, de comparer les regroupements proposés avec ceux résultant de méthodes traditionnelles de 'clustering' de variables. Une telle comparaison réalisée dans un autre contexte (Studer et al., 2007) semble curieusement indiquer que le 'clustering' classique produit une solution plus proche de l'arbre cohésitif que de celui de l'arbre de similarités. Finalement, la structuration sous forme de graphe d'implication nous est apparue plus enrichissante encore. Nous y avons retrouvé l'essentiel de ce que nous avions appris par des analyses détaillées menées avec les outils démographiques de base, et surtout elle nous a permis de mettre en lumière la forte polarisation des caractéristiques démographiques que sont le sexe et l'état-civil, et en particulier la dichotomie entre les enracinés protestants et les turbulences migratoires qui sont plus le fait de catholiques.

En toute équité il convient évidemment aussi de mentionner les limites de l'ASI. Un premier aspect auquel nous avons été confrontés est l'important travail de recodage de nos variables multinomiales qu'a nécessité la mise en œuvre de l'ASI. Un second point, plus fondamental, tient au fait que l'ASI se fonde sur les seuls liens bruts deux à deux entre caractéristiques. Les intensités de lien ne rendent ainsi compte que d'effets bruts, alors que l'on peut dans certaines circonstances être intéressé aux effets propres contrôlés pour les effets des autres caractéristiques. Par exemple, nous avons observé que les nouveaux venus sont en général célibataires, mais on sait aussi que surtout les jeunes sont célibataires. On peut alors se demander si le fait que les nouveaux venus soient célibataires ne reflète pas indirectement un effet d'âge. La relation est-elle valable autant pour les jeunes que pour les moins jeunes ? Une réflexion sur la possibilité de pouvoir distinguer et visualiser, en ASI, les effets propres des

effets bruts pourrait être une piste de développement. Un troisième élément qui peut perturber les adeptes de la modélisation statistique, est l'absence de critère permettant de juger de la pertinence statistique globale du modèle retenu. Quel pourrait être un équivalent de la déviance utilisée en modélisation statistique, ou de la part d'inertie reproduite en analyse factorielle ?

Pour ce qui est de l'étude des dissimilarités des transitions qui s'apparente à une analyse arborescente de la variance, elle s'avère utile pour identifier les facteurs les plus discriminants et renseigne sur leurs interactions. Contrairement au graphe d'implication qui rassemble des liens bruts mesurés entre paires de modalités sans distinguer a priori entre variables à prédire et prédicteurs, l'analyse arborescente des dissimilarités proposée est conçue pour décrire la diversité d'une variable réponse, le type de transition dans notre cas, en fonction de facteurs discriminants potentiels. Il s'agit d'une méthode supervisée au sens où la segmentation se fait en fonction des valeurs des prédicteurs. La nouveauté méthodologique dans ce que nous avons présenté est l'utilisation de l'intensité d'implication pour identifier les transitions caractéristiques des nœuds, ce qui donne une lecture intelligible de l'arbre. La méthode étant destinée à expliquer la variance, une mesure naturelle de l'information apportée et la part de variance expliquée. La faible valeur de 7.3% pour l'arbre de la figure 7 nous indique qu'il reste une part importante de variabilité résiduelle dans les nœuds.

Bien sûr cette méthode arborescente a également des limites. En particulier, comme pour toute méthode arborescente, se pose la question de la stabilité de la structure obtenue. Par ailleurs, nous n'avons pas tenu compte que plusieurs transitions pouvaient concerner un même individu ce qui nécessiterait une approche multiniveau pour laquelle il n'existe pas actuellement de solution dans le contexte des arbres de décision.

Au total, la corroboration mutuelle des enseignements des deux approches est un signe clair de leur intérêt et pertinence.

Références

Babel, A. (1942a). *La crise économique de Genève à l'époque révolutionnaire et les remèdes qu'on a tenté de lui opposer*. Genève : Georg.

Babel, A. (1942b). *La Fabrique genevoise*. Neuchâtel-Paris : Attinger.

Bardet, J.-P. (1983). *Rouen aux XVIIe et XVIIIe siècles : les mutations d'un espace social*. Paris : Société d'édition d'enseignement supérieur.

Couturier, R., A. Bodin, et R. Gras (2006). CHIC v3.7 Classification Hiérarchique Implicative et Cohésitive. Guide d'utilisation, Ecole Polytechnique, Université, Nantes.

Dufour, A. (1997). *Histoire de Genève*, Volume 3210 of *Que Sais-je ?* Paris : PUF.

Gras, R., S. Ag Almouloud, M. Bailleul, A. Laher, M. Polo, H. Ratsimba-Rajohn, et A. Totohasina (1996). *L'implication statistique : Nouvelle méthode exploratoire de données*. Recherches en didactique des mathématiques. Grenoble : La pensée sauvage.

Gras, R., J. David, J.-C. Régnier, et F. Guillet (2006). Typicalité et contribution des sujets et des variables supplémentaires en analyse statistique implicative. In G. Ritschard et C. Djeraba (Eds.), *EGC'2006*, Volume RNTI-E-6 (2 volumes) of *Revue des Nouvelles Technologies de l'Information*, pp. 359–370. Cépaduès.

Gras, R. et P. Kuntz (2006). Discovering *R*-rules with a directed hierarchy. *Soft Computing 10*(5), 453–460.

Gras, R., P. Orús, B. Pinaud, et P. Gregori (Eds.) (2007). *Nouveaux apports théoriques à l'analyse statistique implicative et applications (actes des 4èmes rencontres ASI4)*, Castellò de la Plana, Espagne. Departament de Matemàtiques, Universitat Jaume I.

Guichonnet, P. (1974). *Histoire de Genève*. Toulouse : Privat.

Herrmann, I. (2003). *Genève entre république et canton. Les vicissitudes d'une intégration nationale (1814-1846)*. Québec-Genève : Presses de l'Université Laval-Editions Passé Présent.

Lerman, I. C., R. Gras, et H. Rostam (1981). Elaboration d'un indice d'implication pour données binaires I. *Mathématiques et sciences humaines* (74), 5–35.

Oris, M. et G. Ritschard (2007). Dynamique socioprofessionnelle dans la Genève du 19e, enseignements d'une analyse de statistique implicative. In Gras et al. (2007), pp. 287–300.

Oris, M., G. Ritschard, et G. Ryczkowska (2006). Recrutement et renouvellement des groupes socioprofessionnels à Genève, 1816-1843. In *14e Colloque de l'Association Internationale des Démographes de Langue Française AIDELF, Aveiro, 18-22 septembre 2006*, Aveiro, pp. 791–805. Université d'Aveiro et AIDELF.

Perroux, O. (2006). *Tradition, vocation et progrès. Les élites bourgeoises de Genève (1814-1914)*. Genève : Slatkine.

Schumacher, R. (2002). De l'analyse classique à l'analyse différentielle : nuptialité, fécondité et mortalité à Genève pendant la première moitié du 19e siècle. mémoire de DEA en Histoire économique et sociale, Université de Genève.

Studer, M., G. Ritschard, L. Baccaro, N. S. Müller, et D. A. Zighed (2007). Relations entre types de violation des libertés syndicales garanties par les conventions de l'OIT : une analyse de statistique implicative des résultats d'une fouille de texte. In Gras et al. (2007), pp. 111–122.

Studer, M., G. Ritschard, A. Gabadinho, et N. S. Müller (2009). Analyse de dissimilarités par arbre d'induction. *Revue des nouvelles technologies de l'information RNTI E-15*, 7–18.

Summary

This paper reconsiders the analysis of the socioprofessional dynamic in the first half of the 19th century Geneva that we presented at the ASI4 Conference. It complements the study with a supervised analysis of the dissimilarities between transitions. Data are two by two matches between 6 censuses. More specifically, we consider the socioprofessional group (SPG) of each considered individual and its change between two successive censuses. Our interest is in the types of transition (stable, becoming active, becoming inactive, ...) as well as in new comers (immigrants and births) and lost cases (emigrants and deaths). The statistical implicative analysis of our data gives an insightful image of the relationships between these dynamics and the concerned SPG, as well as with some demographic and cultural variables (sex, age, religion). It highlights for instance polarizations around key variables. The dissimilarity analysis permits to segment the population into homogeneous groups using demographic (age, sex, marital status) and cultural (religion) predictors. The interpretation relies on typical transitions identified for each group by means of the intensity of implication. By so doing, we get also the required elements for comparing the outcome with the results of the implicative graph.

Chapitre 8 : Derrière les réseaux de variables, il y a des individus… à écouter !

L'exemple d'une étude chez des enseignants de lycée professionnel

Marc Bailleul, Sylvain Godard

IUFM de Basse-Normandie
186 reu de la Délivrande
14053 Caen Cedex
CERSE, EA 965
marc.bailleul@caen.iufm.fr
sylvain.godard@caen.iufm.fr

Résumé. L'analyse statistique implicative permet d'organiser en réseaux orientés un ensemble de variables sur lesquelles s'est projetée une population d'individus (par questionnaire par exemple, sous forme de réponses à des questions fermées, binaires ou modales). Le recueil conjoint d'avis plus ouverts, sous forme de textes explicitant les réponses fournies, va, après identification des individus les plus représentatifs des réseaux mis en évidence par l'analyse, faciliter le travail d'interprétation du chercheur en quête du sens portés par lesdits réseaux.
Dans le cadre d'une recherche visant à comprendre comment des enseignants en lycée professionnel se représentent les tensions résultant du croisement, à l'intérieur de ce type d'établissements, de plusieurs logiques (scolaire, économique, politique et administrative), *a minima* en tension, voire contradictoires, nous avons proposé aux enseignants des lycées professionnels de l'académie de Caen un large questionnaire composé de questions essentiellement fermées, mais pour lesquelles nous avions, conformément au choix présenté plus haut, laissé des espaces ouverts sollicitant commentaires ou explicitations.
Nous présentons ci-dessous l'analyse, menée avec le logiciel CHIC, des 257 questionnaires qui nous ont été retournés.

1 Une introduction double

1.1 L'objet de la recherche

L'**enseignement professionnel** sous statut scolaire a profondément évolué. Le lycée professionnel des années 2006 n'a plus grand-chose à voir avec le CET[1] des années 1960. De

[1] CET : Collège d'Enseignement Technique

nombreuses évolutions concernant à la fois les cursus, les sections, l'organisation des enseignements, les personnels, ont amené progressivement au lycée professionnel que nous connaissons. Mais le lycée professionnel n'en a sans doute pas terminé avec les nouvelles prescriptions (Evolution des PFE[2], Lycée des métiers, participation à l'expertise et à la certification dans le cadre de la VAE) et on peut penser que les années à venir verront à leur tour de nouvelles évolutions qui modifieront le profil des établissements et le travail de acteurs de la formation professionnelle en LP[3].

Sans refaire l'historique de la **formation professionnelle** sous statut scolaire, il convient néanmoins de faire ici quelques rappels en termes de repères. Dans les années 60 le passage des centres d'apprentissage aux collèges d'enseignement technique installe une liaison organique entre les deux paliers de l'orientation du collège et les deux formations professionnelles assurées par le CET : CAP[4] en 3 ans à l'issue de la classe de 5ème, et BEP[5] en deux ans à l'issue de la classe de 3ème.

Dans les années 70, le souci est désormais de créer des formes de scolarisation permettant aux élèves en difficulté au collège de ne pas quitter le système éducatif sans un minimum de sensibilisation professionnelle. Les Lycées d'enseignement professionnel remplacent les CET. On passe d'un ordre extérieur au système éducatif (centre d'apprentissage) à une formation identifiée comme faisant partie du second cycle lycée. Dans un même temps l'augmentation du chômage et la détérioration des taux d'insertion provoquent la mise en cause de la formation professionnelle dispensée par l'école. Ces questionnements trouvent des réponses partielles dans l'introduction des séquences éducatives en entreprise pour les élèves de CAP et de BEP.

Par ailleurs et quasi simultanément les collèges ne peuvent plus orienter les élèves à l'issue de la 5ème. Les CAP en 3 ans se transforment en classes de 4ème et 3ème préparatoires. Les deux premières années de CAP, de classes de formation professionnelles, deviennent des classes d'orientation qui débouchent sur la passation du brevet des collèges.

Dès 1983 une reforme des formations de niveau V se traduit par une articulation du BEP et du CAP. Le BEP sanctionnant une maîtrise approfondie d'un champ professionnel tandis que le CAP reste plus proche de l'acquisition d'un métier. En 1985, les lycées d'enseignement professionnels deviennent des Lycées professionnels et le baccalauréat professionnel conduisant au niveau IV de qualification est créé. Désormais (années 90) le second cycle est constitué de trois voies (générale, technologique et professionnelle) quasiment identiques dans leur structure pédagogique. Le LP comprend trois cycles : technologique, de détermination (BEP) et terminal (Bac Pro[6]). En revanche le concept de formation « partagée » école et entreprise évolue en donnant d'avantage d'importance aux périodes de formation en entreprise (en quantité, et en qualité, certaines (Bac Pro seront désormais qualifiantes).

Cette brève description de l'évolution de l'enseignement professionnel montre à quel point la formation professionnelle dépend des environnements sociaux, économiques, scolaires et institutionnels.

---

[2] PFE : Période de Formation en Entreprise

[3] LP : Lycée Professionnel. Nous emploierons désormais ce sigle pour désigner le Lycée Professionnel.

[4] CAP : Certificat d'Aptitude Professionnelle

[5] BEP : Brevet d'Etudes Professionnelles

[6] Bac Pro : Baccalauréat Professionnel

Particulièrement sensible aux différentes logiques à l'œuvre dans la société, l'enseignement professionnel aura dû s'adapter en continu au cours des 50 dernières années, en but aux critiques émanant de l'entreprise et du monde politique comme échouant dans l'insertion professionnelle de ses jeunes élèves, souvent montré en exemple pour ses capacités d'adaptation et d'innovation dans le milieu pédagogique. Valorisé dans les discours « pédagogique » mais dévalorisé dans la réalité des représentations d'une société qui continue à considérer la filière professionnelle comme une voie de relégation en dépit des discours politiques et des professions de foi des ministères successifs.

En aval du collège, dont il doit scolariser une partie des élèves en rupture avec l'école dans des conditions d'orientation qui ne privilégient pas toujours la motivation et l'intérêt des jeunes pour le métier et l'insertion professionnelle, en amont de l'entreprise qui ne ménage pas ses critiques quant à la capacité du lycée professionnel à former ses futurs salariés ; les établissements de l'enseignement professionnel doivent encore, depuis la décentralisation et la dévolution de la formation professionnelles aux services décentralisés de l'état que sont les conseils régionaux, composer avec la concurrence accrue des autres formes d'enseignement professionnel privées et/ou sous contrat (CFA[7], Lycées privés)

Viennent donc s'ajouter de nouvelles logiques institutionnelles qui amènent les établissements à élaborer des stratégies diversifiées : cartes de formations, financements, en adéquation avec l'environnement économique et les possibilités d'insertion et d'emploi. Le ministère et les rectorats continuant à garder la main sur le recrutement, la gestion des personnels ainsi que sur les enseignements à travers les corps d'inspection.

La démultiplication des décideurs et prescripteurs peut poser problème à des personnels enseignants qui se voient assez régulièrement confrontés à des injonctions qui peuvent apparaître comme contradictoires et faire naître chez les équipes enseignantes des conflits de critères qui ne facilitent pas le travail au quotidien sur le terrain. Quant aux établissements s'ils veulent survivre, ils devront jouer avec les prescriptions nationales, régionales et les négocier localement au niveau des différentes équipes présentes dans les établissements sous peine d'entrer dans un jeu de concurrence qui, à terme, pourrait remettre en cause l'égalité des chances pour tous les élèves et les valeurs fondamentales de l'enseignement public[8].

1.2 Problématique : Une recherche centrée sur les enseignants de LP

L'enseignement professionnel ne représente pas un véritable objet d'intérêt pour les chercheurs et pour l'université. Les études, peu nombreuses au total, concernent quelques historiens, sociologues parfois eux-mêmes anciens enseignants de L.P. Ce peu d'intérêt peut paraître paradoxal compte tenu de l'enjeu primordial que représente la formation professionnelle, pour le pays, pour l'entreprise, pour l'éducation nationale, pour les collectivités territoriales et pour l'université. A noter que lorsque l'enseignement professionnel fait l'objet d'un intérêt pour les chercheurs c'est qu'ils y voient, et souhaitent y étudier, les prémisses de ce qui pourrait à plus ou moins long terme toucher les autres ordres d'enseignement (collèges, lycées généraux) qu'il s'agisse de l'évolution des cursus, du travail enseignant, dans un contexte de redéfinition des logiques économiques,

[7] CFA : Centre de Formation des Apprentis

[8] C. Agulhon évoque (1994, p. 66) une situation tripartite Rectorat, Conseil Régional, établissement.

institutionnelles et scolaires lui-même à l'œuvre dans un contexte européen de plus en plus prégnant dans ces domaines.

« S'il s'agit de décrire, d'expliquer et de comprendre l'action éducative aux niveaux individuel, interindividuel, institutionnel, social et politique c'est, de l'amont vers l'aval, l'ensemble des composantes et des caractéristiques de cette action dans ses contextes de réalisation que la recherche doit explorer. La recherche scientifique ne peut donc s'intéresser exclusivement aux effets et aux résultats quantifiables censés être produits par l'action éducative. » (Bru, 2006)

C'est pourquoi nous[9] avons souhaité mener un travail spécifique sur les enseignants de lycées professionnels de l'académie de Caen autour de la question suivante : comment ceux-ci prennent-ils en compte, au quotidien, les logiques parfois contradictoires à l'œuvre dans le fonctionnement du LP ?[10]

Nous avons identifié trois grandes logiques. La **logique économique** (Brucy, Troger, 2000) que l'on peut exprimer ainsi : comment trouver le meilleur équilibre entre la réponse que l'on doit apporter aux exigences immédiates du marché du travail et les exigences prévisibles de ce même marché à long terme ? La **logique sociale** : les lycées professionnels doivent à la fois insérer leurs élèves dans le monde du travail et offrir des formations susceptibles d'être assimilées par les publics scolaires qu'ils accueillent, en particulier ceux qui sont issus de catégories sociales dites défavorisées et/ou en rupture avec le collège (Agulhon, 2000). La **logique administrative**[11] : la décentralisation et la loi quinquennale sur l'emploi placent les LP dans une perspective de développement qui, si elle maintient la définition nationale des formations, permet aussi des développements différenciés selon les régions et autorise leur adaptation aux besoins locaux. Les trois logiques citées ci-dessus cohabitent dans le champ scolaire du LP avec une quatrième logique : celle de l'école et de son fonctionnement.

[9] Nous = un enseignant chercheur auteur de ce texte, un formateur d'enseignants de LP lui aussi auteur, une formateur dit « transversal », deux enseignants en LP et une proviseure adjointe de LP, réunis dans un GFR (Groupe de Formation Recherche)

[10] Pelpel, P., 2002. Une identité plurielle, *Cahiers pédagogiques*, 403, 16-17, Paris : CRAP.

[11] Entre autonomie et intégration, en reprenant le titre de la partie 2 de Pelpel, Troger, 2001.

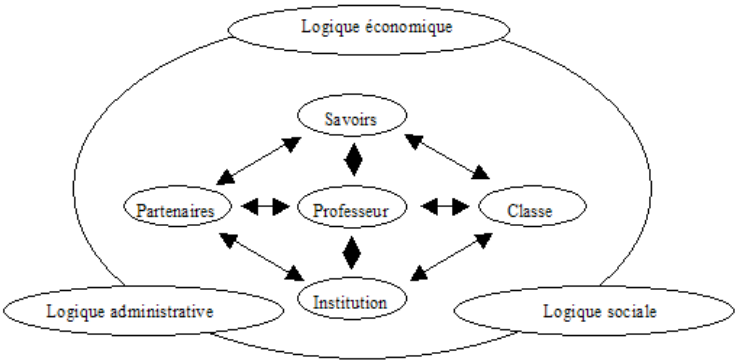


FIG. 1 – *Le LP traversé par différentes logiques*

2 Méthodologie

Se pose alors de façon aigue la question du choix de l'outil de recueil de données. Ayant pour ambition la prise en compte simultanée du sens et des faits, nous avons opté pour un questionnaire dans lequel nous avons mixé questions fermées, à valeurs modales, et questions ouvertes, comme on peut le voir ci-dessous pour une partie de la question 6[12], cherchant ainsi à concilier un certain degré d'ouverture (la place réservée aux commentaires possibles) et de « calculabilité ». Le recueil de données textuelles permet au répondant, s'il le souhaite, de s'impliquer dans ses réponses[13] et sera pour nous une aide au moment de l'interprétation des résultats de l'analyse que nous allons mener sur les données quantitatives, en prenant bien garde de ne pas donner sans précaution aux chiffres le statut d'« explicatifs »[14].

[12] Les personnes intéressées par l'exhaustivité du questionnaire peuvent s'adresser à l'auteur.

[13] Fenneteau, H., 2002. *Enquête : entretien et questionnaire*. Paris : Dunod, « Les avantages des questions ouvertes », p. 61.

[14] De Singly, F., 2001. *L'enquête et ses méthodes : le questionnaire*, Paris : Nathan Université, Collection 128, p. 16.

Q6 - POUR VOUS, EN TANT QU'ACTEUR DU SYSTÈME D'ENSEIGNEMENT EN LP, …

Pour chacune des affirmations suivantes, exprimer votre degré d'adhésion en cochant une case pour chaque item. De plus, dans chaque ligne intermédiaire du tableau, nous aimerions que vous apportiez des éléments d'élucidation de vos réponses.

	Tout à fait d'accord	D'accord	Pas vraiment d'accord	En désaccord
Le LP est un lieu de relégation				
Le LP est une filière d'excellence				

TAB.1 – *Des questions à valeurs modales avec une possibilité de réponse ouverte*

Nous avons soumis ce questionnaire, après test, à l'ensemble des professeurs de LP de l'Académie de Caen. Les principaux thèmes abordés étaient les suivants : les contenus d'enseignement (l'interaction enseignement général/enseignement professionnel et leur adaptation aux besoins actuels et à venir des élèves, dans les deux types d'enseignement), le partage de la formation entre lycée professionnel et entreprise, l'apport du passage en entreprise, les caractéristiques du lycée professionnel et la formation des enseignants.

Nous avons recueilli 257 questionnaires (taux de retour = 24 %, tout à fait acceptable dans ce genre d'enquête basée sur le volontariat), dont un bon nombre s'était emparé de la possibilité offerte de compléter ses réponses par des textes commentaires.

L'analyse statistique implicative s'intéresse aux phénomènes de quasi-implication entre variables[15] et permet de construire des graphes implicatifs, représentations graphiques desdits phénomènes. On peut ainsi mettre en évidence des **réseaux orientés** de variables. Il est ensuite possible d'identifier les individus représentatifs des différents réseaux[16], ce qui permet par simple comptage de relativiser le « poids », dans la population enquêtée, des différents réseaux, et, de plus, de retrouver les éventuels commentaires qu'ils apportent à l'appui de leurs positionnements.

[15] Voir Bailleul, 1994.
[16] Voir Bailleul, Gras, 1994 et Bailleul, 2001.

3 Quelques résultats

A l'issue d'une analyse statistique implicative appliquée au corpus des réponses aux questions fermées, au seuil d'implication de .98, nous avons dégagé six réseaux de variables. C'est maintenant à l'interprétation du graphe implicatif produit par cette analyse que nous allons nous confronter. En identifiant les individus représentatifs de chaque réseau, nous avons pu aller quérir les commentaires qu'ils ont parfois apportés à leurs réponses en termes de croix, dans les cases « ouvertes » du questionnaire.

Les chemins qui apparaissent dans la partie droite du réseau R1, dont le « puits » est le rejet de « le LP, on pourrait faire sans… », ont leurs origines dans la nécessité double d'une « culture générale au service du professionnel » et d'une « culture générale de l'honnête homme » (l'expression « honnête homme » renvoyant à son acception humaniste). A la question Q21 concernant l'adéquation des savoirs des disciplines générales aux besoins des élèves, l'individu 249 répond : « oui, même si les élèves sont sceptiques sur la nécessité de tous ces contenus » et adhère au rejet de « le LP, on pourrait faire sans… ». En écho, l'individu 37 énonce « Beaucoup d'employeurs confirment la nécessité de s'exprimer en anglais (plombiers, routiers, commerçants, comptables…) », à l'appui de la nécessité de certains éléments de culture générale pour s'insérer dans le monde du travail.

Pour l'individu 55 « le manque de culture générale est un handicap » quand l'individu 68 affirme que le LP « doit rester un lieu d'apprentissage de formation de culture protégé pour des adolescents parfois incertains de leurs choix ». L'individu 206 de renchérir : « Le LP est en premier lieu un endroit où les élèves ayant quelques difficultés peuvent s'épanouir. Cependant, nous devons aussi les former pour leur assurer une bonne intégration dans le monde de l'entreprise. L'expérience montre que, dans le bâtiment par exemple, nos jeunes vivent de belles réussites professionnelles. » En commentaire de sa réponse à la question Q62, l'individu 107 modère (à peine) sa réponse de la façon suivante : « indispensable me semble excessif, mais c'est un bon outil de formation et de suivi social » après avoir avancé « le LP était et devrait redevenir (un lieu d'émergence d'une "culture professionnelle") » quand l'individu 122 dit « préserver la culture professionnelle est un atout majeur, une culture incontestable de notre lieu de formation ».

Derrière les réseaux de variables, il y a des individus… à écouter !

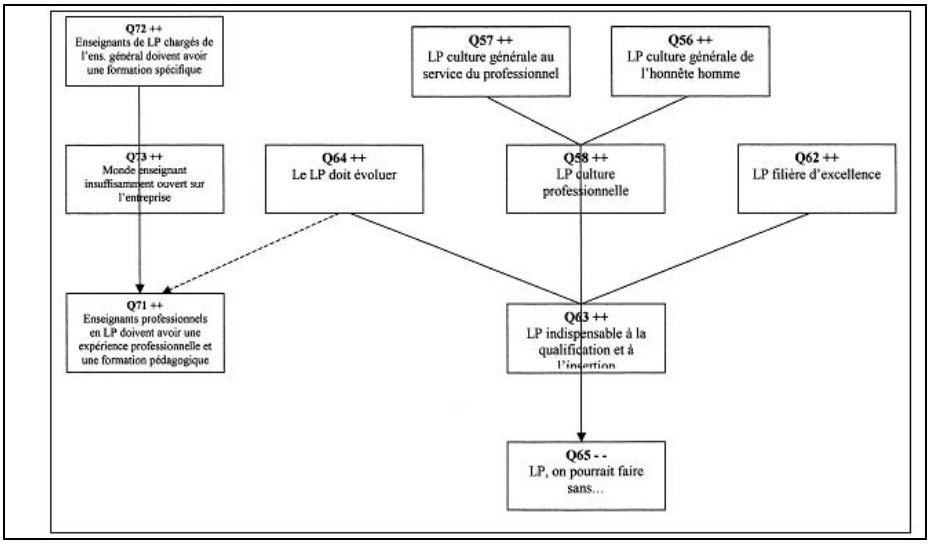


FIG. 2 – *Le réseau R1*

Il y a dans ce réseau un sous-graphe (à gauche) pointant des caractéristiques relatives aux enseignants, nécessaires aux yeux des répondants puisque leurs avis sont positifs sur ces items. On perçoit bien la logique de ce chemin. Même quand on enseigne une discipline générale, il faut une formation spécifique pour enseigner en LP (l'individu 254 répond « Ils devraient faire une formation en entreprise et en lycée professionnel. » au deuxième item de la question Q7 et l'individu 68 ajoute « une formation sur le monde du travail, certainement. Je pense que beaucoup sont en décalage avec la réalité professionnelle »), ce qui implique une ouverture vers le monde de l'entreprise, mais aussi, et réciproquement, une formation pédagogique pour les enseignants en charge des enseignements professionnels (individu 14 : « La pédagogie et la didactique doivent être enseignées, il faut au minimum les bases pour les appréhender. L'expérience professionnelle est nécessaire, cependant elle doit être variée »).

Afin de caractériser simplement ce réseau, nous le titrerons avec l'expression : **« De la nécessité sociale du LP ».**

Au cœur du réseau R2, on trouve le rejet de la phrase : « Le LP est lieu d'émergence d'une culture de l'exclusion ». L'individu 205 est particulièrement représentatif de ce réseau bien qu'il tienne des propos qui pourraient sembler contradictoires, révélant une distorsion entre les opinions de ceux qui « habitent le LP » et le ressenti « social » de ce type d'établissement. Voici ses réponses aux différents items de la question Q6 :

« Le LP est un lieu de relégation de la société, mais pas pour moi. Il suffit de valoriser le LP. Le LP est une solution qui reste très satisfaisante pour le devenir professionnel et social d'un jeune : CAP : 80 %, BEP : 80 %, BacPro : 90 % de réussite. Le LP n'est pas un lieu d'émergence de l'exclusion pour celui qui y vit. Mais il faudrait peut être changer l'esprit des média et aussi des collègues. »

Cet enseignant pointe le déficit d'image des LP, déjà signalé par de nombreux auteurs, et responsables politiques[17].

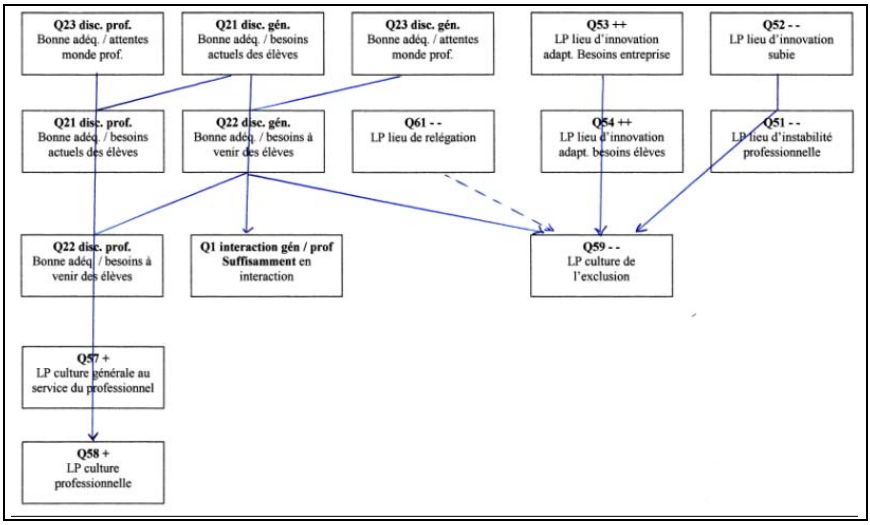


FIG. 3 – *Le réseau R2*

En conséquence, les enseignants tentent de répondre aux attentes du « monde de l'entreprise » (chemin de gauche dans le graphe) :

« La technologie évolue dans les entreprises. Il faut suivre le mouvement pour que nos élèves utilisent le même matériel que celui des entreprises. » Ind. 19

« Le monde de l'entreprise exige que nos élèves s'adaptent lors des stages. » Ind. 73

« Il nous est nécessaire de modifier nos programmations. » Ind. 160

« Le monde enseignant en LP dans le domaine professionnel est lui assez ouvert sur l'entreprise. » Ind. 205

Nous nous autoriserons à qualifier ce réseau par l'expression suivante : **« Le LP, lieu d'émergence d'une culture professionnelle et non de l'exclusion »**

L'impression générale qui se dégage du réseau R3 est celle d'une collaboration « sans difficultés » entre lycée et entreprise dans le cadre d'une alternance « tout bénéfice » : le passage en entreprise apporte sens et cohérence, aide à l'insertion professionnelle, à la construction du projet professionnel, développe l'autonomie, est une dimension indispensable des apprentissages et, finalement, apporte une plus grande connaissance de l'entreprise.

[17] Dubet, F. (2005) ; Mélenchon, J.-L., (2002, p. 25-26) ; Jellab, A. (2005, p. 297)

Derrière les réseaux de variables, il y a des individus… à écouter !

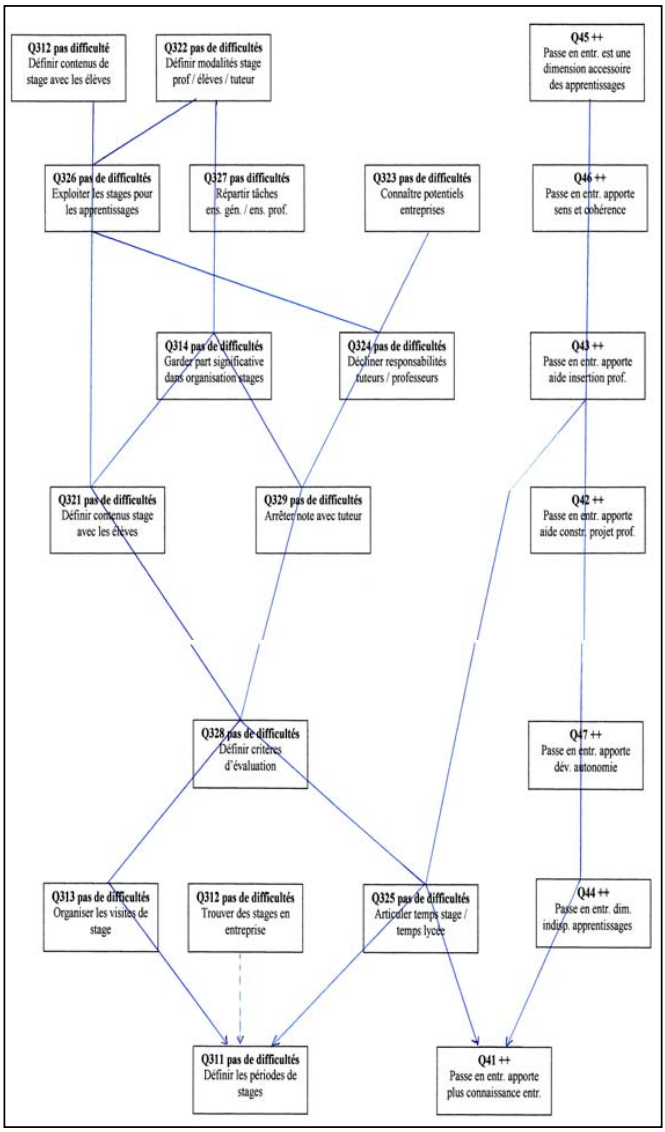


FIG. 4 – *Le réseau R3*

« L'alternance Lycée entreprise est très bénéfique pour préparer la future vie active et sociale. » Ind. 222

« car l'élève se rapproche du monde de l'industrie au fur et à mesure de sa formation » Ind. 114

« Le LP doit collaborer au monde de l'entreprise » Ind. 110

« Il semble évident qu'il doit y avoir adéquation entre les apprentissages et la réalité du monde du travail auquel nos élèves vont être confrontés » Ind. 49

« Les référentiels sont élaborés en partenariat avec le monde professionnel » Ind. 17

Il y a là, nous semble-t-il, trace d'une vision idéalisée de l'alternance lycée professionnel / entreprise : **« LP / entreprise, une collaboration sans problèmes… »**

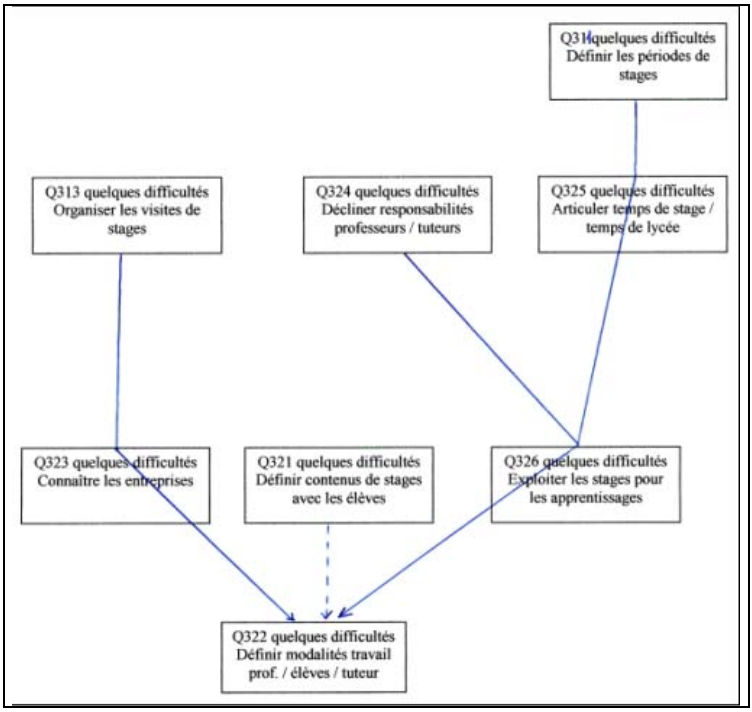


FIG. 5 – *Le réseau R4*

« Les entreprises (surtout les PME-PMI) sont trop sollicitées par des stages de toutes natures et de tous niveaux. Nos élèves ont du mal à trouver des lieux de stage. » Ind. 230

« Les semaines de stage amputent gravement les heures de cours. Il convient donc d'en limiter le nombre. » Ind. 230

La collaboration évoquée plus haut n'irait-elle pas sans générer quelques insatisfactions quant aux conditions matérielles de sa mise en place ? Ne peut-on voir derrière la deuxième phrase un problème de « leadership » de la formation ?

« Utilisation de matériel spécifique à l'entretien des locaux mais la plupart des entreprises ont encore… un simple balai et des éponges… ! » Ind. 237

« En CAP ETC[18], dans la discipline Entretien du linge, l'enseignement de la couture est décalé par rapport à la réalité : la pose de bouton ne se fait plus à partir des machines à coudre familiale. » Ind. 237

Il s'agit là d'un autre type de distorsion entre le lycée et l'entreprise : celle qui existe au niveau des matériels sur lesquels on travaille. Le décalage pouvant même être ressenti comme plus profond :

« Le référentiel appliqué en atelier est parfois trop éloigné de la réalité sur le terrain » Ind. 116

Alors, finalement, l'alternance est-elle si simple à mettre en place ? Est-elle aussi harmonieuse que le laissait pressentir le réseau R3 ? Nous croyons pouvoir dire qu'apparaît ici le réseau « miroir » du précédent. Le **réseau R4** sera donc qualifié par l'expression : **« Où il est question de quelques difficultés à gérer les PFE… »**

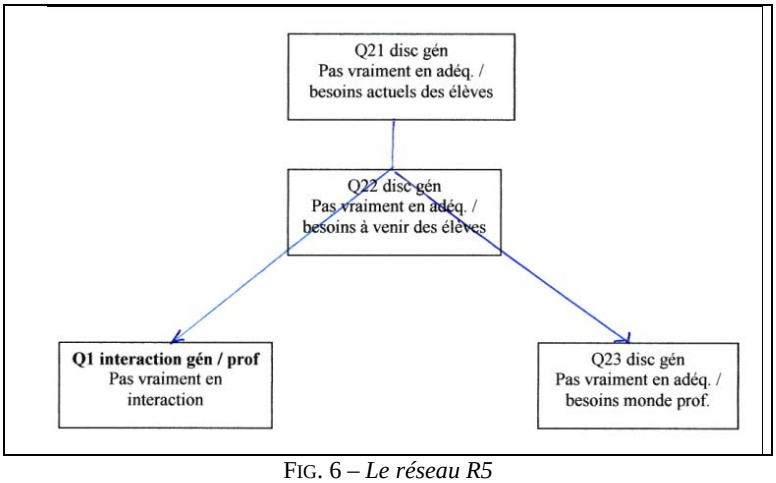


FIG. 6 – *Le réseau R5*

N'y aurait-il pas non plus quelques tensions à l'intérieur du lycée, en particulier quant à la place de l'enseignement général ?

- par rapport à l'enseignement professionnel :

« L'enseignement général semble peu intéressé par les visites en entreprises » Ind. 22

- par rapport aux attentes du monde professionnel :

« Sans doute l'enseignement du français tel que je le pratique n'est-il pas assez utilitaire pour un chef d'entreprise. » Ind. 224

« Il faudrait que le français soit maîtrisé avant de passer à autre chose. Les entreprises attendent que le français et une langue soient maîtrisés plus les maths. » Ind. 158

- par rapport aux élèves eux-mêmes :

« L'enseignement du français est éloigné des envies et des besoins des élèves. » Ind. 66

« Il faudrait que le français soit maîtrisé avant de passer à autre chose. » Ind. 158

« S'il s'agit des besoins estimés par les élèves, beaucoup n'ont aucune "utilité" à court terme. Nos logiques sont souvent contraires. » Ind. 224

[18] ETC : Employé Technique de Collectivités

R5 pose alors la question de la « **légitimité / utilité des enseignements généraux dans le cadre du LP** ».

Le réseau R6 ci-dessous est globalement un réseau traduisant une insatisfaction et des difficultés majeures, du moins beaucoup plus importantes que dans le réseau 4. Essayons d'y voir plus clair.

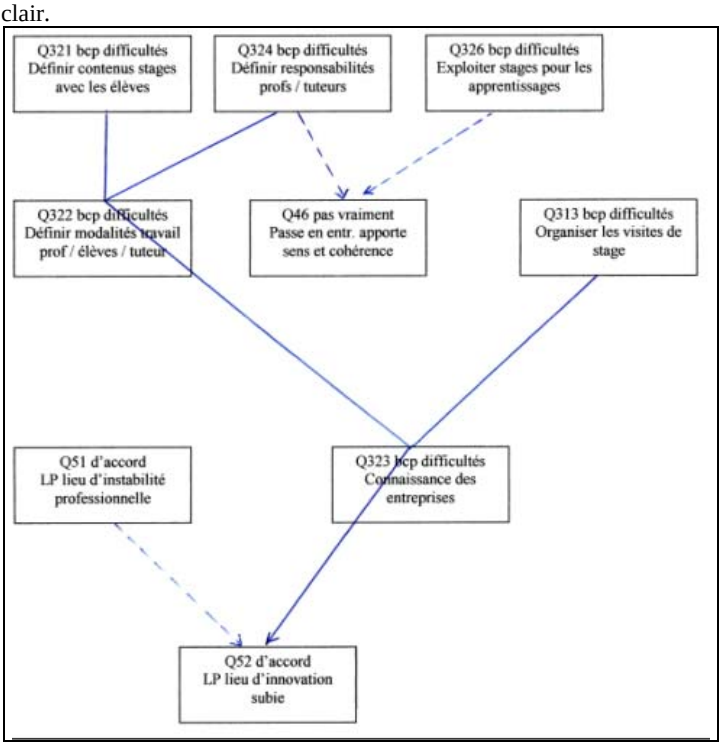


FIG. 7 – *Le réseau R6*

D'un premier point de vue, ce sont les élèves du LP qui posent problème[19]

« Formation partagée entre le LP et l'entreprise : ce qui pose problème : élèves non autonomes, ne veulent pas ou ne peuvent se déplacer, pas toujours de la bonne volonté du coté parental. » ind. 118

« A mon avis, il ne s'agit pas d'un problème de contenus d'enseignement mais ce sont nos élèves qui sont de plus en plus en inadéquation avec les attentes du milieu pro. » Ind. 245

« Le monde professionnel attend surtout de la matière grise que nous ne sommes pas toujours en mesure de lui offrir. » Ind. 224

puis les stages

« Trop de stages par rapport aux heures du lycée » Ind. 146

[19] Voir Agulhon, 1998.

« Les stages sont indispensables mais beaucoup sont trop longs. Pour le Bac Pro compta 1 stage de 4 semaines et 1 stage "emploi vacance" rémunéré conviendraient et seraient plus formateurs. » Ind. 129

« Problèmes des stages de Juin : beaucoup de demandes/ énormément de difficultés à trouver des stages, beaucoup ne sont pas adaptés. » Ind. 129

« L'exploitation des stages n'est plus possible en raison de la difficulté à les trouver; dommage, car nous gagnerions beaucoup de temps pour le programme (beaucoup de notions ne sont plus traitées en stage) » Ind. 129

« Pas assez de cours (en général et professionnel) en raison des stages. Voyez ce que sont 2 années en intégrant 16 semaines de stages + vacances + 1 semaine formation profs + 1 semaine intempérie ou épidémie + contrôles + CCF[20] + etc. » "Ind. 129

mais aussi les contenus d'enseignement[21], dans leur articulation avec les exigences du monde professionnel qui est à la tête des instances d'élaboration des référentiels

« Il existe un décalage entre les enseignements et l'examen final du diplôme et les exigences professionnelles. » Ind. 88

« Ce sont les entreprises qui rédigent le rapport des activités professionnelles (R.A.P.) duquel est décliné le référentiel de certification » Ind. 133

Finalement apparaît une divergence profonde sur les valeurs supposées des deux mondes

« Certaines activités théoriques ne seront jamais vues en entreprise pour les élèves de CAP. En effet, on ne leur demande aucune autonomie mais de mettre en application des gestes. » Ind. 119

« L'école doit-elle être un lieu d'insertion pour fabriquer des outils à peine pensants (mais productifs) ou des êtres ouverts pouvant évoluer et parfois revendicatifs ! » Ind. 215

« Qu'est-ce qu'une "culture professionnelle" ? La pratique et la connaissance d'une série de traditions et d'habitudes d'un métier ? ? Dans ce cas dans un monde où sont valorisés l'innovation, le changement, la réforme ! ! Ce n'est pas très utile ! ! » Ind. 216

d'où une impression d'injonction, de contrainte à l'innovation.

« Exigences du monde professionnel est un terme mal choisi qui tend à présenter l'entreprise comme le seul élément de progrès. Il faudrait mieux utiliser le terme général innovation technique ou industrielle » Ind. 60

En résumé, nous caractériserons ce réseau **R6** par l'expression : **« LP / entreprise : des tensions difficiles à concilier… »**

4 Une conclusion double elle aussi

La première partie de cette conclusion est d'ordre méthodologique. Il est indéniable que les propos tenus par les répondants dans les espaces qui leur étaient offerts pour une expression personnelle nous ont passablement aidés pour livrer des interprétations des graphes issus de l'analyse implicative.

Nous pouvons maintenant aborder la deuxième partie de cette phase de conclusion. On peut, comme nous l'avons indiqué à la fin du paragraphe consacré plus haut à la

[20] CCF : Contrôle en Cours de Formation

[21] Jellab, A., 2001. Scolarité et rapport aux savoirs en lycée professionnel, Paris : PUF, 232 p.

méthodologie, retrouver les individus représentatifs de chaque réseau pour essayer de « mesurer » les poids respectifs des réseaux dans la population enquêtée. Le tableau ci-dessous rend compte de la répartition de la population sur les six réseaux.

	R1	R2	R3	R4	R5	R6
Effectif	57	39	30	73	24	33
Pourcentage de la population (%)	22	15	12	28	9	13

TAB.2 – *Répartition de la population dans les six réseaux*

Remarquons que la population se distribue de façon globalement identique entre les réseaux qui révèlent une vision « positive » du LP (R1, R2, R3), 49 % de la population, et ceux qui renvoient une vision « critique » (R4, R5, R6), 50 % de la population. Peut-être même pourrions-nous accentuer cette opposition en qualifiant d'« idéalisée » la première vision et de « réaliste » la seconde. Il est aussi possible de distinguer deux niveaux de lecture du « système LP » à travers les différents réseaux : R1 et R6 renvoient aux finalités dudit système quand R2, R3, R4 et R5 renvoient à son fonctionnement.

Nous pouvons alors proposer le schéma ci-dessous comme modélisation des résultats de notre analyse.

Les logiques sociale et économique traversent les quadrants Q1 et Q2, partie supérieure du schéma, en harmonie dans Q1, à gauche, en tension forte dans Q2, à droite. Doit-on pour autant dresser un constat d'impuissance devant un trop grand écart entre les caractéristiques et les compétences des élèves d'un côté et les attentes toujours plus élevées du monde de l'entreprise de l'autre ?

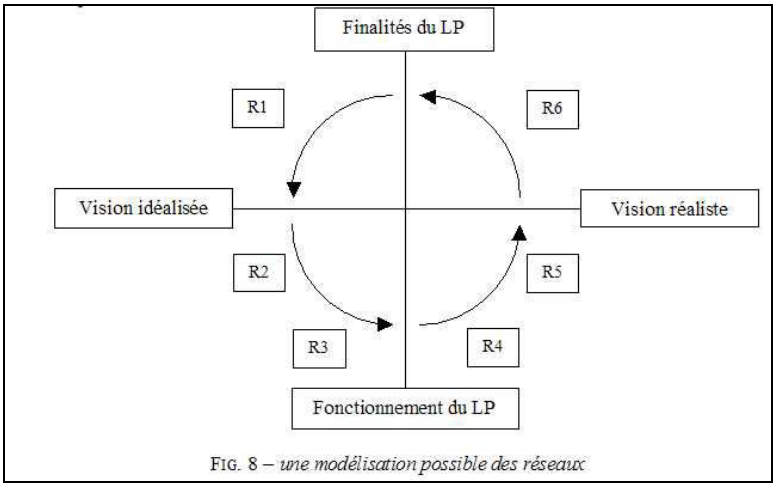


FIG. 8 – *une modélisation possible des réseaux*

Les logiques scolaire et administrative sont la traduction dans les faits des finalités assignées au LP par l'Institution Education Nationale en lien avec l'environnement local. Si, dans le quadrant Q3, partie inférieure gauche du schéma, tout donne l'impression de pouvoir fonctionner ensemble, les réseaux impliqués dans le quadrant Q4, inférieur droit, font

apparaître des difficultés. La mise en œuvre de la réforme des « lycées des métiers »[22] se veut une réponse, tant administrative que scolaire, en termes de fonctionnement, à cette tension idéal / réalité.

« "Dégager le sens" et "démêler la complexité" sont deux activités essentielles pour les chercheurs qui ont l'ambition d'appréhender les processus de changement de pratiques d'enseignement dans des contextes locaux identifiables. » (Vandenberghe, R., 2006)[23] Nous espérons, à travers ce travail, y avoir contribué.

Références

Agulhon C. (1994) *L'enseignement professionnel, Quel avenir pour les jeunes ?*, Paris : Editions de l'atelier.

Agulhon C. (1998) L'orientation scolaire, prescription normative et processus paradoxal, *L'orientation scolaire et professionnelle*, 27-3 : 353-371.

Agulhon C. (2000) L'enseignement professionnel et technologique dans la tourmente, *Cahiers pédagogiques*, 387 : 25-26.

Bailleul M. (1994) *Analyse statistique implicative : variables modales et contributions des sujets. Application à la modélisation de l'enseignant dans le système didactique*, Thèse « Nouveau Régime », Université de Rennes I.

Bailleul M., Gras R. (1994) L'implication statistique entre variables modales, *Mathématiques, Informatique et Sciences humaines*, 128 : 41-57.

Bailleul M. (2001) Des réseaux implicatifs pour mettre en évidence des représentations, *Mathématiques et Sciences humaines*, 154-155 : 31-46.

Bru M. (2006) Préface, in Paquay L., Crahay M., De Ketele J.-M., 2006. *L'analyse qualitative en éducation*, 7-11, Bruxelles : De Boeck.

Brucy G., Troger V. (2000) Un siècle de formation professionnelle en France : la parenthèse scolaire ?, *Revue Française de Pédagogie*, 131 : 9-21.

Dubet F. (2005) *Pourquoi changer l'école ?*, Paris : Editions du Seuil.

Fenneteau H. (2002) *Enquête : entretien et questionnaire*, Paris : Dunod.

[22] Le label *« lycée des métiers »* qualifie certains établissements qui offrent une palette étendue de formations et de services, grâce notamment à un partenariat actif, tant avec le milieu économique qu'avec les collectivités territoriales, et en premier lieu la région.

Le label *« lycée des métiers »* met en évidence la cohérence d'une offre de formation, la prise en compte des attentes des élèves et l'adaptation aux besoins des employeurs. Il constitue un indicateur d'excellence pour les voies technologique et professionnelle.

Les critères qui permettent à un lycée d'obtenir ce label sont des éléments clefs pour faciliter l'insertion des jeunes dans le monde du travail.

http://education.fr- Eduscol-Lycée des métiers/Le label Lycée des métiers, site consulté le 07 octobre 2008

[23] Vandenberghe, R., 2006. La recherche qualitative en éducation : dégager le sens et démêler la complexité, in Paquay, Crahay, De Ketele, 2006. *L'analyse qualitative en éducation*, Bruxelles : De Boeck, 53-64)

Jellab A. (2005). Les enseignants de lycée professionnel et leurs pratiques pédagogiques : entre lutte contre l'échec scolaire et mobilisation des élèves, *Revue française de sociologie*, 46-2 : 295-323.

Jellab A.(2001). Scolarité et rapport aux savoirs en lycée professionnel, Paris : PUF.

Mélenchon J.-L. (2002) *Pour une école globale*, Paris : Ramsay.

Paquay L., Crahay M., De Ketele J.-M. (2006) *L'analyse qualitative en éducation*, Bruxelles : De Boeck.

Pelpel P. (2002) Une identité plurielle, *Cahiers pédagogiques*, 403 : 16-17.

Pelpel P., Troger V.(2001) *Histoire de l'enseignement technique*, Paris, L'Harmattan.

Singly F. de (2001). *L'enquête et ses méthodes : le questionnaire*, Paris : Nathan université, Collection 128.

Vandenberghe R. (2006) La recherche qualitative en éducation : dégager le sens et démêler la complexité, in Paquay L., Crahay M., De Ketele J.-M., 2006. *L'analyse qualitative en éducation*, 53-64, Bruxelles : De Boeck.

Summary

Implicative statistical analysis allows to organise a set of variables (for example answers to a questionnaire) in oriented networks and to project people who answered through these networks. In order to understand how vocational secondary school teachers see the tensions between four kind of preoccupation which go through this type of secondary school (teaching and learning, economical, political and management), we have given a large questionnaire to these teachers and obtained 257 answers.

In this paper, we present the results of the implicative analyse made with CHIC software.

Thème 4 : Application à la bio-informatique

Chapitre 9 : Une méthode implicative pour l'analyse de données d'expression de gènes

Gérard Ramstein

LINA, Polytech'Nantes
Rue Christian Pauc BP 50609 44306 Nantes cedex 3, France
gerard.ramstein@univ-nantes.fr

Résumé. Nous présentons une méthode d'extraction d'associations basée sur l'analyse statistique implicative et la notion de rang. Nous avons adapté le concept d'intensité d'implication à des classements pour découvrir des relations partielles robustes vis à vis du bruit et des variations d'amplitude. Appliquée aux données de puces à ADN, cette méthode met en évidence des relations entre des formes d'expressions particulières de gènes. Ces associations peuvent être révélatrices de mécanismes de corégulation génique et donc contribuer à l'analyse biomédicale. Nous montrons que cette définition de l'intensité d'implication apporte une connaissance plus fine des relations entre les gènes que les méthodes usuelles de corrélation et qu'elle permet notamment de discriminer entre différents phénotypes avec une précision comparable aux techniques de classification les plus abouties dans ce domaine.

1 Introduction

La technologie des puces à ADN permet d'analyser l'expression simultanée de milliers de gènes. L'étude du transcriptome représente un enjeu considérable, tant du point de vue de la compréhension des mécanismes du vivant que des applications cliniques et pharmacologiques. Malheureusement, les données d'expression sont entachées de multiples bruits. D'une part, la complexité du protocole expérimental conduit à une réduction de la précision des mesures. D'autre part, la variabilité naturelle de l'activité cellulaire induit des différences notables d'amplitude d'expression entre les gènes, phénomène également perceptible en considérant plusieurs patients présentant le même phénotype. Cette étude propose une méthode d'analyse implicative des règles d'association sur les données du transcriptome. Elle utilise l'approche de Régis Gras (1996) en considérant non pas les mesures elles-mêmes, mais le rang des observations. Cette optique permet de s'affranchir des valeurs numériques en considérant des zones de classement dans les mesures d'expression. S'intéresser au classement a l'avantage d'améliorer la robustesse des algorithmes en les rendant insensibles à des transformations monotones des données. On peut émettre une analogie avec un tableau notes scolaires. Chaque enseignant possède son propre système de notation et aura une sévérité différente vis à vis des réponses

données par les élèves. Le corps professoral pourra plus facilement s'entendre sur les résultats relatifs des élèves que sur les notes. Un élève situé en haut de classement en mathématique et en physique sera ainsi considéré comme bon, même si sa note diffère entre les deux matières de quelques points. Notre approche peut aisément se généraliser à d'autres domaines. Dans le problème dit du panier de la ménagère, on s'intéressera ainsi au volume d'achat des clients. Les connaissances seront donc relatives à des niveaux de consommation des produits considérés. Nous allons dans une première partie présenter un état de l'art du domaine d'étude. Nous donnons ensuite le cadre conceptuel des règles d'association utilisé dans notre approche. Nous présentons dans une troisième partie une application à la classification de tumeurs.

2 Etat de l'art

Le traitement des données incertaines ou imprécises a déjà fait l'objet de travaux dans le cadre de l'analyse implicative. Dans Gras et al. (2001), la méthode proposée consiste à définir une partition optimale des données puis à rechercher des implications entre domaines, ces domaines étant constitués à partir de l'union d'éléments de la partition obtenue. Une approche parallèle a été développée, basée sur la logique floue (Gras et al. (2005)). Ces deux méthodes reposent sur une partition préalable des mesures avant l'analyse implicative proprement dite. Comme les distributions de nos données se sont avérées monomodales, nous avons préféré rechercher directement l'implication optimale sans passer par un prétraitement des données qui risquerait d'introduire un biais dans l'extraction des règles. A notre connaissance, l'analyse statistique implicative n'a pas encore été appliquée à l'étude du transcriptome. Cependant, plusieurs études concernent l'extraction de règles d'association à partir de données d'expression. Un travail, basé sur l'algorithme Apriori (Agrawal et Srikant (1994)) et utilisant les notions habituelles de support et de confiance, a été mené sur le génome de la levure (Creighton et Hanash (2003)). Les données ont été discrétisées à partir de seuils prédéfinis pour caractériser trois niveaux d'expression (sous-expression, expression normale, sur-expression). Tuzhilin et Adomavicius (2002) présente une analyse sur les biopuces définissant un ensemble d'opérateurs adaptés à de grands volume de règles. Ces deux méthodes présentent l'inconvénient d'être tributaires du paramétrage des seuils de discrétisation. Dans Carmona-Saez et al. (2006), l'étude est enrichie par la prise en compte de connaissances a priori sur les gènes. Une méthode originale a été développée dans Cong et al. (2004) à travers un outil dénommé FARMER. Cette méthode recherche des ensembles de règles possédant un support commun et s'appuie sur une technique de discrétisation entropique. Dans Becquet et al. (2002), les auteurs ont appliqué l'algorithme Min-Ex. L'analyse porte sur des données binaires, la valeur logique associée prenant en compte le fait que le gène est considéré comme sur-exprimé ou non. L'approche la plus proche de la nôtre est certainement celle proposée dans Jin et al. (2006). Les auteurs y définissent le concept de patron émergent comme un ensemble d'opérateurs booléens sur des valeurs d'expression. Cette méthode revient à extraire des intervalles d'expression spécifiques à chaque gène en optimisant une mesure entropique. Cette étude diffère de la nôtre par le fait qu'elle suppose la connaissance de classes d'individus et que la procédure est globale : la prise en compte de la totalité des observations induit un risque de perte d'implications statistiquement significatives.

3 Intervalles de rang

Notre étude porte sur une matrice $M(k,l)$ de données numériques, où k représente un individu, l une observation et $M(k,l)$ la mesure effectuée. Dans l'exemple cité en introduction, la matrice peut définir un ensemble de notes, k désignant une matière et l un élève. Pour les données de biopuces, k correspond à un gène et l à une expérimentation. Dans une étude clinique, l représentera par exemple un patient soumis à une condition expérimentale particulière, comme l'absorption d'une médicament. On notera m le nombre d'individus et n le nombre d'observations. Il est à noter que même si nous traitons de données réelles, notre étude peut s'appliquer à n'importe quelle valeur ordinale. De même, il est possible de transposer la matrice si on s'intéresse à des règles portant sur les observations plutôt que sur les individus. Nous appelons profil d'un individu k le vecteur $p(k) = (M(k,l), l \in [1,n])$ et supposons l'existence d'un opérateur $rang$ qui délivre les observations d'un profil dans l'ordre croissant des valeurs d'expression. Plus précisément, nous avons $rank(p(k)) = (l_1, l_2, \ldots, l_n)$, où les observations $l_i \in [1,n]$ vérifient la condition suivante : $M(k,l_1) \leq M(k,l_2) \leq \ldots \leq M(k,l_n)$. Soit par exemple un profil représentant les notes en Physique : $p(k) = (9,4,17,12)$, $rang(p(k))$ renvoie alors le vecteur $(2,1,4,3)$. Ce vecteur représente le classement des élèves par ordre croissant dans la matière, le plus mauvais élève (note 4/20) étant désigné par l'indice $l_1 = 2$ et le meilleur (17/20) par l'indice $l_4 = 3$.

Nous nous proposons d'extraire des règles d'association entre profils. Ces règles vont mettre en évidence des intervalles d'étude. Un intervalle de rang permet de déterminer des niveaux de classement (i.e. d'expression) sans avoir à définir des seuils de mesure. Pour l'analyse du transcriptome, ces intervalles vont se rapporter à des niveaux d'expression, telles que la surexpression ou la sous-expression. L'analyse de notes scolaires permet quant à elle d'extraire des règles telles que si un élève est en bas de classement en Mathématique, il sera également en bas de classement en Physique. Sur l'exemple précédent, l'intervalle $[1,3]$ désignera la zone de classement $(2,1,4)$, relative aux trois plus basses notes du profil, à savoir 4, 9 et 12, correspondant aux valeurs $M(k,l_i)$, $i \in [1,3]$. Notre approche considère tous les intervalles possibles, à savoir l'ensemble des sous-intervalles de $[1,n]$:

$$I = \{[p,q], (p,q) \in [1,n]^2, p \leq q\} \quad (1)$$

A un intervalle $i \in I$ est associé un intervalle de rang $r_k(i)$, défini comme suit : $r_k(i) = \{l_j \in rang(p(k)), j \in i\}$. Notons que par raccourci de langage nous désignons sous le terme d'intervalle de rang $[p,q]$ les observations relatives à cet intervalle, à savoir $(l_p, l_{p+1}, \ldots, l_q)$. La table 1 donne deux profils relatifs aux individus A et B. Nous remarquons par exemple que $r_A(i) = (3,5,7,9)$ pour $i = [1,4]$ et $r_B(j) = (9,7,2,5,3)$ pour $j = [5,9]$. Ces deux intervalles de rang comportent des observations communes. Ce phénomène semblerait indiquer une association entre le début de classement du profil de A (intervalle i) et la fin de classement du profil de B (intervalle j). Pour savoir si ce phénomène est statistiquement significatif, nous allons reprendre l'approche de Gras (1996) et considérer deux ensembles α et β de mêmes tailles respectives que $r_A(i)$ et $r_B(j)$. Ces deux ensembles sont évidemment inclus dans l'ensemble des observations O. Notons que dans notre approche l'hypothèse nulle consiste à affirmer que l'opérateur de classement $rang$ n'apporte aucune information utile. Dans ce cas, prendre deux intervalles de rang reviendrait à sélectionner les ensembles α et β au hasard, i.e. sans

tenir compte du classement des observations. Nous reprenons le concept d'intensité d'implication tel qu'il est défini dans Gras (1996) ainsi que la mesure de qualité $\varphi(\alpha, \beta)$, où α et β représentent les ensembles vérifiant respectivement la prémisse et la conclusion d'une règle $A \rightarrow B$. Cette mesure définit l'étonnement statistique d'observer si peu de contre-exemples dans une règle d'association. Nous étendons la définition originelle reposant sur le cardinal aux intervalles de rang :

$$\varphi_I(A, B) = max(\varphi(r_A(i), r_B(j)), (i, j) \in I^2) \quad (2)$$

L'expression (2) indique que la qualité de la règle $A \rightarrow B$ est définie comme la plus grande intensité d'implication entre deux intervalles de rang, le premier étant issu du profil $p(A)$ et le deuxième du profil $p(B)$. Cette définition revient à rechercher les intervalles i et j qui maximisent $\varphi(r_A(i), r_B(j))$. Dans l'exemple de la table 1, les intervalles $i = [1, 4]$ et $j = [5, 9]$ correspondent à la valeur maximale d'intensité d'implication : $\varphi(A \rightarrow B) = 0.86$.

profil	1	2	3	4	5	6	7	8	9
p(A)	5	6	1	8	2	9	3	7	4
p(B)	13	17	19	14	18	12	16	11	15

TAB. 1 – *Exemples de profils. Les numéros de colonne désignent les indices relatifs à neuf observations effectuées sur deux individus A et B.*

4 Intérêt de l'approche implicative pour l'étude des données d'expression

En matière d'étude du transcriptome, les analyses les plus couramment menées par les biologistes sont basées sur des mesures de corrélation entre profils d'expression. Ces mesures présentent l'inconvénient d'être globales dans le sens où elles font intervenir l'ensemble des observations, alors que la définition (2) recherche une correspondance optimale entre des sous-ensembles d'observations. Pour expliciter la différence entre ces deux approches, nous allons considérer un exemple de règle d'association entre deux gènes (figure 1). Ces gènes appartiennent à l'espèce Saccharomyces cerevisiae, communément dénommée levure du boulanger. Nous avons repris les données de puces sélectionnées dans Gasch et Eisen (2002). Nous avons retenu 89 différentes conditions expérimentales correspondant à différents stress induits tels que le choc thermique. La figure 1 représente l'implication CHA1 → SAM1. Le gène CHA1, intervenant dans le catabolisme de la threonine, est clairement sous-exprimé en réponse à un signal de déficience en acide aminé, et dans une moindre mesure, en nitrogène. SAM1, un gène interférant dans le métabolisme de la methionine, est sur-exprimé pour le même jeu d'observations. Ce jeu correspond à environ 9 % des conditions.

Comme le montre la table 2, les indices usuels de corrélation ne peuvent déceler de telles associations. Les valeurs obtenues sont trop faibles pour être retenues dans une analyse alors que la mesure implicative exprime que le risque de rencontrer une association de même nature au hasard est inférieur à un pour mille.

Méthode	Valeur
Intensité d'implication	0,9992
Indice de corrélation de Pearson	0,16
Indice de corrélation de Kendall	0,0089

TAB. 2 – *Comparaison de mesures effectuées sur l'exemple de la figure1.*

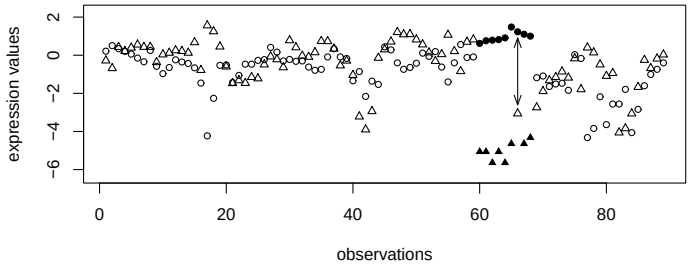


FIG. 1 – *Profils des gènes CHA1 et SAM1. L'axe des abscisses représente 89 conditions expérimentales. L'axe des ordonnées représente les mesures d'expression. Le profil du gène CHA1 (YCL064C) est représenté par des triangles et celui de SAM1(YLR180W) par des cercles. Les figures pleines correspondent aux observations appartenant aux intervalles de rang qui maximisent la valeur de l'intensité d'implication. On remarque que ces observations sont communes à une exception près, indiquée par la double flèche : l'observation de SAM1 qui est moins sous-exprimée, n'apparaît pas dans le groupe alors qu'il est bien présent dans celui de CHA1.*

5 Une mesure implicative du pouvoir discriminant des gènes

La section précédente a montré que la méthode implicative peut être utilisée dans le cadre non supervisé. Nous supposerons désormais que les observations appartiennent à un ensemble C de classes prédéfinies. Par la suite, une classe correspondra à des tissus provenant de patients présentant un phénotype commun. Nous proposons une technique de sélection des gènes les plus pertinents pour la discrimination de tumeurs basée sur la méthode implicative. La recherche de gènes informatifs est un enjeu majeur en bio-informatique. Il s'agit d'extraire dans un jeu de plusieurs milliers de gènes ceux dont l'expression est la plus significative. L'analyse la plus commune est celle de l'expression différentielle, à savoir la recherche de profils d'expression qui varient d'une classe d'observation à une autre. Nous montrons à partir d'une expérimentation sur deux types de leucémie que l'approche que nous proposons s'avère particulièrement performante. Nous proposons enfin une technique originale de représentation visuelle, appliquée à un jeu de données portant sur des tumeurs cérébrales.

5.1 Une définition implicative du concept de gène discriminant

Les données de puces concernent généralement un nombre important de gènes. La plupart d'entre eux n'aident pas à discriminer les classes, soit parce que leurs expressions ont une faible amplitude de variation, soit parce que leur expression est indépendante des classes d'observations. Une étape de sélection des gènes est donc nécessaire. La technique la plus communément partagée repose sur la puissance discriminative des tests statistiques telles que le test de Student ou ANOVA (voir Chen et al. (2005) pour une synthèse des mesures utilisées sur les données d'expression). Nous proposons d'utiliser l'intensité d'implication pour déterminer les gènes les plus discriminants. On appellera fonction d'étiquetage la fonction $L(o_k) = c_i$, où o_k est une observation de O et $c_i \in C$. Notre approche est fondée sur des règles de classification de la forme : $r_g(i) \rightarrow O_c$, où la conclusion désigne les observations de classe c et la prémisse l'intervalle de rang associé au gène g et à l'intervalle i. Cette règle peut être explicitée de la façon suivante : si pour une observation, sa mesure d'expression relative au gène g figure dans la zone de classement défini par l'intervalle i, alors cette observation appartient probablement à la classe c. Nous restreignons par la suite le domaine des intervalles défini en (1) en imposant des niveaux d'expression ayant une signification biologique précise, à savoir la sous-expression (intervalle de la forme $[1, q]$) ou la sur-expression (intervalle $[p, n]$). L'ensemble I des intervalles est donc désormais :

$$I = \{[p, q], (p, q) \in [1, n]^2, p \leq q, p = 1 \vee q = n\} \tag{3}$$

La règle $r_g(i) \rightarrow O_c$ peut dès lors s'interpréter plus simplement : elle signifie que si on observe une dérégulation du gène g sur un patient o, alors ce patient possède probablement le phénotype c. Cette implication correspond bien à la notion de gène discriminant. Le pouvoir discriminant'un gène g vis à vis d'une classe c sera exprimée par l'expression suivante :

$$\varphi_c(g) = max(\varphi(r_g(i), O_c), i \in I, O_c = \{o, L(o) = c\}) \tag{4}$$

La mesure $\varphi_c(g)$ est définie par la maximisation de l'intensité d'implication des règles de classification de type $r_g(i) \rightarrow O_c$. Il est à noter que la définition (3) diffère uniquement de l'expression (2) par une modification de l'ensemble conclusion : $r_B(j)$ est remplacé par l'ensemble O_c des observations de classe c. Une remarquable propriété de l'intensité d'implication est de prendre en compte le nombre d'observations appartenant à O_c : pour un intervalle fixe i, le fait d'accroître la population O_c entraîne une diminution de la qualité de la règle. Il est en effet normal de retrouver une proportion notable d'observations de classe c si cette dernière est sur-représentée.

5.2 Méthode de sélection

Soient G l'ensemble total des gènes définis dans l'ensemble des expérimentations O et M la matrice de données d'expression associée. Soient C les classes d'observations et L la fonction d'étiquetage de O vers C. L'algorithme présenté en figure 2 extrait les K gènes les plus discriminants pour chaque classe de C.

Notons que cet algorithme considère chaque classe de manière indépendante. Il peut arriver qu'un gène soit considéré comme discriminant pour plusieurs classes. On remarquera que la mesure $\varphi_c(g)$ sera différente selon la classe considérée. Il peut en effet par exemple arriver

Algorithme Sélection
Entrée :
$M, \mathcal{G}$: matrice et ensemble des gènes considérés
$\mathcal{C}, L$: ensemble des classes et fonction d'étiquetage
K : le nombre de gènes discriminants à retenir
Sortie :
gd : ensemble $\{(g, c, j)\}$ des gènes discriminants
où g est le gène retenu, c la classe discriminée
et j la mesure de qualité définie en (3)
début
$gd \leftarrow \emptyset$
pour chaque $c \in \mathcal{C}$ **faire**
$genelist \leftarrow \emptyset$
pour chaque gène $g \in \mathcal{G}$ **faire**
$\varphi \leftarrow \varphi_c(g)$
$listGenes \leftarrow listGenes \cup \{(g, c, \varphi)\}$
fait
trier les triplets de $ListeGenes$ par ordre décroissant de φ
Soit $selection$ l'ensemble des K premiers triplets
de la liste triée
$gd = gd \cup selection$
fait
fin.

FIG. 2 – *Algorithme pour l'extraction de gènes discriminants*

qu'un gène soit nettement sous-exprimé pour une classe donnée et qu'il soit au contraire sur-exprimé dans une autre, à un degré moindre. Il faut aussi noter que dans le cas particulier d'une expérimentation à deux classes (e.g. patients sains versus patients malades), on pourrait s'attendre à ce qu'un gène discriminant pour une classe le soit aussi pour l'autre. En ce qui concerne notre approche, ce fait n'est pas forcément vérifié. Il peut en effet arriver que les deux extrêmes du classement des observations n'aient pas la même homogénéité. Dans ce cas, les valeurs de $\varphi_{c_1}(g)$ et de $\varphi_{c_2}(g)$ différeront, et selon la qualité relative du gène vis à vis des autres gènes candidats, il se peut fort bien qu'il soit retenu pour une classe et rejeté pour une autre.

5.3 Application à une étude portant sur deux types de leucémie

Nous appliquons notre algorithme de sélection sur une étude portant sur deux types de leucémie Golub et al. (1999). La leucémie se caractérise par une prolifération maligne de cellules d'origines hématopoïétiques peu matures et rapidement diffusantes. Cette maladie se caractérise par une atteinte massive de la moelle osseuse, due au développement de lymphomes malins. On distingue les leucémies aiguës lymphoblastiques (notées ALL par la suite pour Acute Lymphoblasic Leukemia) des leucémies aiguës myéloblastiques (notées AML pour

Acute Myeloid Leukemia). La distinction entre ces deux formes est essentielle pour le succès des thérapies envisagées : le traitement diffère selon l'une ou l'autre de ces deux classes de leucémie. Le jeu de données comporte 38 patients (27 patients ALL et 11 patients AML) et concerne 3571 gènes. Nous nous proposons d'extraire dans ce jeu les gènes les plus discriminants en appliquant l'algorithme de sélection décrit précédemment. Nous avons étudié le pouvoir discriminant des gènes du jeu complet sur la base de la mesure (3), dont nous donnons ici une version logarithmique : $\lambda_c(g) = -log_{10}(1-\varphi_c(g))$. Cette transformation présente plusieurs avantages : l'indice de qualité $\lambda_c(g)$ n'est plus borné et les valeurs sont plus facilement interprétables. Une règle de qualité possède ainsi un risque de $10^{-\lambda_c(g)}$ d'être dû au hasard. L'analyse révèle que 10 % des gènes ont un pouvoir discriminant important ($\lambda_c(g) > 3{,}5$ soit $\varphi_c(g) > 0{,}9997$). Bien que 300 gènes peuvent être considérés comme discriminants, les auteurs de l'étude n'ont retenu qu'une liste de 50 gènes les plus informatifs, tout en indiquant le caractère arbitraire de ce nombre. Comme ils ont développé une technique originale de sélection, il est intéressant de comparer leur approche avec la nôtre. Nous avons donc fixé le paramètre K de l'algorithme de la figure 2 à 25, puisque nous avons deux groupes de gènes pour les classes ALL et AML. Nous avons obtenu une liste de gènes discriminants dans laquelle figurent 14 gènes appartenant à la liste publiée par les auteurs. La table 3 compare le pouvoir discriminant des deux jeux. On remarque le gène le plus discriminant est commun aux deux jeux de gènes et que la moyenne est du même ordre de grandeur. On observe cependant une plus grande dispersion dans le jeu de Golub et al. De même, la valeur médiane est significativement plus faible par rapport à notre liste.

Liste de gènes	min	médiane	max	moyenne	variance
Golub & al.	$8.3\ 10^{-10}$	$3.6\ 10^{-6}$	$2.8\ 10^{-5}$	$7.6\ 10^{-6}$	$6.9\ 10^{-11}$
Notre liste	$8.3\ 10^{-10}$	$3.4\ 10^{-7}$	$3.2\ 10^{-6}$	$1.2\ 10^{-6}$	$1.6\ 10^{-12}$

TAB. 3 – *Comparaison des valeurs de $\lambda_c(g)$ sur les 50 gènes sélectionnés par Golub et al. et selon notre méthode.*

Pour comparer la puissance discriminative de ces deux jeux de gènes, nous avons procédé à une validation croisée sur les données en utilisant le même classifieur, à savoir les 3 plus proches voisins. Cette technique a été retenue parce qu'elle prend en considération la distance entre points dans tout l'espace des gènes, contrairement à d'autres méthodes qui vont privilégier certains gènes (arbres de décision, forêt aléatoire, séparateurs à vastes marges) et qui apporteraient donc un biais pour la comparaison. Nous avons appliqué l'algorithme sur la matrice d'expression réduite comportant les mêmes patients et portant sur le jeu de gènes considéré. La table 4 montre que notre sélection donne des résultats supérieurs à celle des auteurs.

6 Application à la classification de tumeurs

La section précédente a permis de définir une mesure du pouvoir discriminant des gènes. Comme nous avons basé notre algorithme de sélection sur des règles de classification de type $r_g(i) \rightarrow O_c$, il est naturel d'envisager d'utiliser celles-ci pour prédire la classe d'un patient

% test	Golub & al.	notre méthode
50%	3.11	0.79
25%	1.67	0.11
10%	1.00	0.00
2.6%	0.00	0.00

TAB. 4 – *Comparaison des taux d'erreurs en validation croisée. La première colonne indique le pourcentage considéré en test sur le jeu de données. Les deux autres colonnes donnent les taux d'erreurs en pourcentage sur 100 jeux de validation aléatoires.*

d'après son profil d'expression. Cette catégorisation est un des enjeux majeurs de la technologie des biopuces : elle permet de diagnostiquer l'existence d'un cancer à un stage précoce, lorsque la maladie s'exprime dans les cellules sans qu'on observe encore des signes cliniques manifestes. Par ailleurs, la prédiction basée sur l'expression des gènes permet de distinguer entre différents types de tumeurs même si leur apparence morphologique tumorale est identique. Nous allons dans un premier temps présenter les algorithmes mis en oeuvre pour classer des observations, puis nous allons présenter les jeux de données ainsi que les méthodes utilisées pour comparer nos résultats avec ceux de la littérature.

6.1 Algorithmes pour la classification

Pour caractériser la capacité prédictive des règles extraites, nous proposons une approche supervisée comprenant un jeu d'apprentissage $A = \{G, M, O, L, C\}$, où G est un ensemble de gènes, M les mesures effectuées sur un ensemble O d'observations, et L la fonction qui attribue à chaque observation une classe de C. La figure 4 présente l'algorithme d'extraction des règles de classification, qui est analogue dans son principe à celui qui nous a servi à sélectionner les gènes. On notera que le cardinal de l'ensemble des règles extraites R est $|R| = K.|C|$, K étant un paramètre d'entrée de l'algorithme et $|C|$ le nombre de classes considérées.

Le principe de la prédiction d'un échantillon à partir d'une expérimentation peut être décrit comme suit : soit o une observation nouvelle de classe inconnue sur laquelle a été effectuée une mesure d'expression sur l'ensemble des gènes G et soit $p(o)$ le profil d'expression correspondant à ces mesures, il s'agit de relever les prémisses des règles de classification que respecte o et définir $L(o)$ comme étant la classe la plus souvent rencontrée en conclusion de ces règles. On ne peut en pratique vérifier directement si o respecte la prémisse P d'une règle $(r_g(i) \rightarrow O_c) \in R$. En effet, $P = r_g(i)$ est défini à partir de l'ensemble $O \in A$, ensemble dans lequel ne figure pas l'observation o. Pour pallier à cette difficulté, nous recherchons où se situerait la nouvelle mesure d'expression $M[g, o]$ relative à un ensemble $O' = O \cup \{o\}$. Autrement dit, on cherche à savoir si la nouvelle observation s'insère entre les rangs du classement effectué au moment de l'apprentissage. La pratique opératoire est donc la suivante : soit $s(P) = min(M[g, o], o \in P)$ et $S(P) = max(M[g, o], o \in P)$, on dira que o respecte P si la condition suivante est réalisée : $s(P) \leq M[g, o] \leq S(P)$.

La figure 5 décrit l'algorithme de prédiction d'une observation o. Son principe repose sur un consensus : la classe attribuée est celle qui a recueilli le maximum de suffrages, les votants

Algorithme Extraction des règles de classification
Entrée :
A : jeu d'apprentissage
K : le nombre de règles souhaitées par classe
Sortie
R : l'ensemble des règles extraites
début
pour chaque classe $c \in \mathcal{C}$ **faire**
$listeRegles \leftarrow \emptyset$
pour chaque gène $g \in \mathcal{G}$ **faire**
Rechercher l'intervalle de rang $r_g(i)$ qui maximise
$\varphi(r_g(i), O_c)$, $O_c = \{o, L(o) = c\}$
$listeRegles \leftarrow listeRegles \cup \{(r_g(i), \varphi_c(g))\}$
fait
trier les couples de $ListeRegles$ par ordre décroissant de $\varphi_c(g)$
Soit $selection$ l'ensemble des K premiers triplets
de la liste triée
$R = R \cup selection$
fait
fin.

FIG. 3 – *Algorithme pour l'extraction de règles de classification*

étant les règles et les votes leurs conclusions.

On notera que la pertinence du vote est liée à la qualité de la règle : plus celle-ci comporte de contre-exemples et plus le risque d'erreur de classification augmente. Il est envisageable de tenir compte de cette propriété en attribuant un poids de vote proportionnel à l'intensité d'implication de la règle. En pratique, nous n'avons pas observé d'améliorations notables en procédant de la sorte. La raison en est vraisemblablement que toutes les règles extraites sont suffisamment pertinentes vis à vis des jeux de données considérés. Cela est dû au nombre $K = 25$ relativement réduit qui a été appliqué pour la sélection des règles lors de l'apprentissage.

6.2 Etude comparative des performances de classification

Outre les jeux de données sur la leucémie et sur le cerveau, qui ont déjà été présentés, nous appuyons notre étude sur des données portant sur le cancer du colon. Ce jeu contient des données d'expression sur des tissus du colon. L'étude a été faite sur 62 tissus, dont 22 sains et 40 tumoraux. L'expression de 6 500 gènes a été analysée. Les données sont accessibles sur le site : Colorectal Cancer Microarray Research (http ://microarray.princeton.edu/oncology). La table 5 résume les différentes caractéristiques des jeux que nous avons étudiés.

Nous comparons les performances de notre classifieur avec les résultats obtenus, soit dans la littérature, soit obtenus en utilisant des classifieurs génériques. Pour le premier type de clas-

Algorithme Prédiction
Entrée :
M : : matrice d'expression du jeu d'apprentissage
$p(o)$: profil d'expression de l'observation o
R : ensemble des règles extraites par apprentissage
Sortie :
χ : classe prédite pour lŠobservation o
Variable intermédiaire :
count : vecteur de taille $|\mathcal{C}|$ pour compter les occurrences des prémisses satisfaites par o
début
pour chaque classe $c \in \mathcal{C}$ **faire**
$count[c] \leftarrow 0$
pour chaque règle $r \in R$ **faire**
soit g le gène relatif à la règle r,
soit m la mesure d'expression relative à g dans $p(o)$,
soit P la prémisse de r, $s(P)$ et $S(P)$ les valeurs minimales (resp. maximales) dans M relativement à g et à la prémisse de la règle r.
si $s(P) \leq m \leq S(P)$ **alors**
$count[c] \leftarrow count[c] + 1$
fait
fait
$\chi < -argmax_{c \in \mathcal{C}}(count)$
fin.

FIG. 4 – *Algorithme de prédiction d'une observation o.*

sifieur, nous avons retenu deux études majeures dans le domaine de l'analyse du transcriptome, à savoir l'algorithme Gene clustering (Dettling et Buhlmann (2002)) et l'algorithme Fuzzy c-means (Wang et al. (2003)). Outre ces méthodes spécifiquement dédiées au traitement de données d'expression, nous avons également retenu des techniques de classification supervisée habituellement utilisées dans ce domaine.

Les classifieurs utilisés dans l'étude sont les suivants :

k-plus-proches-voisins. Cette méthode requiert les faveurs des biologistes pour sa simplicité d'interprétation Yeang et al. (2001). Le classifieur recherche les k plus proches voisins d'un échantillon inconnu en fonction d'une mesure de disance $d(x, y)$. La métrique la plus courante en bioinformatique est le coefficient de Pearson absolu, la distance étant

Jeu de données	Publication	# tissus	# classes	# gènes	type
Cerveau	Pomeroy et al. (2002)	42	5	5 597	S
Colon	Alon et al. (1999)	62	2	2 000	T
Leucémie	Golub et al. (1999)	72	2	3 571	S

TAB. 5 – *Présentation des jeux de données publiques sur le cancer utilisés dans notre étude. La colonne type désigne le type d'expérimentations biomédicales (S : sous-types tumoraux, T : tissu sain/ tissu malade.)*

défini par :

$$d(x,y) = 1 - \mid \frac{\sum_{i=1}^{n}(x_i - \mu(x))((y_i - \mu(y))}{(n-1)\sigma(x)\sigma(y)} \mid \quad (5)$$

où μ et σ désignent respectivement la moyenne et l'écart-type des profils d'expression. Le classifieur attribue à l'échantillon inconnu la classe majoritaire de ses k voisins (k étant impair, souvent fixé à 3 dans la littérature). Comparée à des techniques plus élaborées, cette méthode donne des résultats satisfaisants, à condition de disposer d'un jeu de gènes pertinents. Dans le cas contraire, la grande dimension de G peut être un élément défavorable et rendre peu significatif le calcul de distance entre observations.

Forêt aléatoire. Ce classifieur est composé d'un très grand nombre d'arbres de décision (Breiman (2001)). Chacun de ces arbres reçoit les données relatives à un choix aléatoire d'observations (tirage avec remise). A chaque noeud, l'arbre sélectionne l'attribut le plus pertinent parmi un choix aléatoire de gènes. L'algorithme utilise ensuite une prédiction basée sur le consensus, à l'instar de notre méthode : la classe retenue est celle qui a été le plus souvent prédite par les arbres. Cette technique s'est montré particulièrement efficace sur les données du transcriptome. Elle possède un grand pouvoir de prédiction, même quand le nombre d'observations est réduit par rapport au nombre de gènes (Diaz-Uriarte et Alvarez de Andres (2006)).

Séparateurs à vastes marges. Cette technique est un des plus performantes en matière d'apprentissage automatique (Vapnik (1995)). Elle est particulièrement efficace dans le cas d'espace de données de haute dimension. Le principe du classifieur consiste à rechercher un hyperplan de séparation optimale entre deux classes d'échantillon dans un espace de caractéristiques. Cette méthode se prête bien à la classification tumorale à partir de biopuces (Lee et Lee (2003)).

La table 6 présente les résultats obtenus selon la technique du *leave-one-out* (validation croisée en prenant l'ensemble des observations en apprentissage moins une, cette dernière servant de jeu de test, ce principe étant répété pour chaque observation). Sur les trois jeux de données décrits précédemment, notre méthode atteint des performances comparables aux autres classifieurs. Ce résultat est d'autant plus remarquable que notre algorithme est relativement frustre, puisqu'il s'agit d'un simple comptage de règles de classification. Malgré sa simplicité, il rivalise parfaitement avec des techniques sophistiquées.

méthode	Cerveau	Colon	Leucémie
notre méthode	14.3	12.9	2.8
Gene clustering	11.9	16.1	2.8
Fuzzy c-means	14.3	11.4	4.1
forêt aléatoire	19.0	14.5	2.8
séparateurs à vastes marges	11.9	12.9	2.8
3 plus proches voisins	23.8	22.6	1.4

TAB. 6 – *Comparaison des méthodes de classification. Le tableau indique les taux d'erreurs selon la technique du leave-one-out. Bien évidemment, la méthode basée sur les règles de classification effectue à chaque test une extraction de règles sur le jeu d'apprentissage ; le candidat testé a été retiré de ce jeu.*

7 Conclusion

L'approche implicative, appliquée aux données d'expression, présente plusieurs avantages. En premier lieu, elle est plus fine que les techniques qui mesurent des relations de similarité globale. Par mesure globale, nous entendons des estimations basées sur l'ensemble des observations. Il est clair que si un certain nombre d'observations ne participent pas à une relation, ces informations apportent un bruit qui masque l'association entre gènes, comme nous l'avons montré sur un exemple d'expérimentations sur la levure. De la même manière, une association entre conditions expérimentales peut être masquée par des gènes qui ne sont pas régulés de manière coordonnée avec un autre groupe de gènes. C'est la raison pour laquelle une analyse implicative est plus performante que des techniques basées sur des corrélations.

Un deuxième intérêt de notre approche est liée à la robustesse de l'analyse de rang. On remarquera que l'analyse de classement est invariante par rapport à toute transformation monotone des données. Cette propriété est particulièrement utile dans le cadre de données de puces qui subissent un grand nombre de prétraitement (transformation logarithmique, normalisation, ...).

Enfin, on rappelle que l'implication est une information orientée, contrairement aux techniques de similarité qui sont symétriques. Cette propriété peut être exploitée dans le cadre du transcriptome. On sait en effet que les gènes sont activés par le biais de facteurs de transcription, qui sont eux-mêmes exprimés dans la cellule. Découvrir des relations de causalité entre l'expression de gènes est un enjeu majeur en bioinformatique. Les jeux de données jusqu'à présent ne permettaient pas de telles analyses, puisque la biopuce n'est que la photographie de l'activité de la cellule à un instant donné. L'accumulation des expérimentations et leur libre diffusion au sein de la communauté scientifique offrent depuis peu la possibilité d'opérer des méta-analyses : la comparaison de multiples jeux de données permet dès lors d'inférer des relations d'implications entre gènes (lorsque tel gène est exprimé, tel autre gène est exprimé, et non l'inverse). C'est une voie d'application prometteuse pour la fouille de données.

Nous avons proposé une méthode originale de sélection des gènes informatifs. La pertinence de la méthode a été vérifiée en démontrant le pouvoir prédictif du jeu sélectionné. Une forme ori-

ginale de représentation visuelle des données a été proposée pour analyser les gènes d'intérêt et la représentativité des observations. La découverte de gènes discriminants est d'une grande importance pour les applications cliniques, car elle permet de définir des méthodes de diagnostic fiables et relativement peu coûteuses. Nous avons développé un algorithme d'extraction de règles de classification. L'avantage d'une méthode de classification basée sur les règles est qu'elle délivre une information aisément interprétable par un expert biologiste, contrairement à des méthodes abstraites telles que les machines à vecteurs de support. Malgré sa simplicité de mise en oeuvre, notre algorithme s'est révélé aussi performant que les techniques les plus éprouvées dans ce domaine.

Références

Agrawal, R. et R. Srikant (1994). Fast algorithms for mining association rules. In *Proceedings of the 20th Very Large Data Bases Conference*, pp. 487–499. Morgan Kanfmann.

Alon, U., N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, et A. J. Levine (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci U S A 96*(12), 6745–6750.

Becquet, C., S. Blachon, B. Jeudy, J. F. Boulicaut, et O. Gandrillon (2002). Strong-association-rule mining for large-scale gene-expression data analysis : a case study on human sage data. *Genome Biol 3*(12).

Breiman, L. (2001). Random forests. *Machine Learning 45*, 5–32.

Carmona-Saez, P., M. Chagoyen, A. Rodríguez, O. Trelles, J. M. Carazo, et A. D. Pascual-Montano (2006). Integrated analysis of gene expression by association rules discovery. *BMC Bioinformatics 7*, 54.

Chen, D., Z. Liu, X. Ma, et D. Hua (2005). Selecting genes by test statistics. *Journal of Biomedicine and Biotechnology 2*, 132–138.

Cong, G., A. Tung, X. Xu, F. Pan, et J. Yang (2004). Farmer: Finding interesting rule groups in microarray datasets.

Creighton, C. et S. Hanash (2003). Mining gene expression databases for association rules. *Bioinformatics 19*(1), 79–86.

Dettling, M. et P. Buhlmann (2002). Supervised clustering of genes. *Genome. Biol. Res. 3*(12), research0069.1–0069.15.

Diaz-Uriarte, R. et S. Alvarez de Andres (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics 7*.

Gasch, A. et M. Eisen (2002). Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering.

Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, et E. S. Lander (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science 286*, 531–537.

Gras, R., R. Couturier, F. Guillet, et F. Spagnolo (2005). Extraction de règles en incertain par la méthode statistique implicative. In *12èmes Rencontres de la Société Francophone de*

Classification, Montreal, pp. 148–151.

Gras, R., E. Diday, P. Kuntz, et R. Couturier (2001). Variables sur intervalles et variables-intervalles en analyse statistique implicative. In *Société Francophone de Classification (SFC'01)*, Univ. Antilles-Guyane, Pointe-à-Pître, pp. 166Ű–173.

Gras, R. e. c. (1996). *L'implication Statistique*. Grenoble : La Pensée Sauvage.

Jin, X., X. Zuo, K.-Y. Lam, J. Wang, et J.-G. Sun (2006). Efficient discovery of emerging frequent patterns in arbitrary windows on data streams. In *ICDE*, pp. 113.

Lee, Y. et C.-K. Lee (2003). Classification of multiple cancer types by multicategory support vector machines using gene expression data. *Bioinformatics 19*(9), 1132–1139.

Pomeroy, S. L., P. Tamayo, M. Gaasenbeek, L. M. Sturla, M. Angelo, M. E. McLaughlin, J. Y. H. Kim, L. C. Goumnerova, P. M. Black, C. Lau, J. C. Allen, D. Zagzag, J. M. Olson, T. Curran, C. Wetmore, J. A. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, G. Stolovitzky, D. N. Louis, J. P. Mesirov, E. S. Lander, et T. R. Golub (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature 415*(6870), 436–442.

Tuzhilin, A. et G. Adomavicius (2002). Handling very large numbers of association rules in the analysis of microarray data. In *KDD*, pp. 396–404.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. New York, NY, USA : Springer-Verlag New York, Inc.

Wang, J., T. H. Bø, I. Jonassen, O. Myklebost, et E. Hovig (2003). Tumor classification and marker gene prediction by feature selection and fuzzy c-means clustering using microarray data. *BMC Bioinformatics 4*, 60.

Yeang, C. H., S. Ramaswamy, P. Tamayo, S. Mukherjee, R. M. Rifkin, M. Angelo, M. Reich, E. Lander, J. Mesirov, et T. Golub (2001). Molecular classification of multiple tumor types. *Bioinformatics 17 Suppl 1*, 316–322.

Summary

We present a rule extraction method based on the statistic implicative analysis and the concept of ranking. We adapt the definition of the intensity of implication to ranked data. This definition permits to discover partial relationships inside ordered observations. Applied to microarray DNA data, our method extracts relationships between particular forms of gene expressions. These associations may reveal underlying mechanisms of gene coregulation and help biological analysis. We show that our definition of the intensity of implication gives a finer knowledge of gene relations than correlation based techniques. Our tool can discriminate different phenotypes with a precision comparable to the most performing classifiers.

Thème 5 : Application à l'histoire de l'art

Chapitre 10 : Iconographie médiévale en histoire de l'art et Analyse Statistique Implicative

Magali Guénot* et Jean-Claude Régnier**
*Université de Lyon – UMR 5138 Archéométrie et archéologie
magali.guenot@univ-lyon2.fr
**Université de Lyon - UMR 5191 ICAR
ENS-LSH 15, Parvis René Descartes BP 7000 69342 LYON cedex 07
jean-claude.regnier@univ-lyon2.fr

Résumé. Si l'utilisation de la statistique en l'histoire de l'art, et en particulier en iconographie médiévale, est acquise depuis nombre d'années, le recours à l'analyse statistique implicative marque un pas supplémentaire dans l'approche des sujets thématiques. Les statistiques sont une aide à la recherche : elles mettent en évidence des permanences ou des originalités, mais sans pour autant participer à une réflexion sur le sujet. L'ASI apporte un regard nouveau sur la lecture du sujet : elle permet de connecter des éléments sélectionnés au préalable par le chercheur, faisant apparaître d'autres liens internes à l'image et offre une nouvelle manière d'appréhender l'étude thématique en histoire de l'art.

1 Introduction

Depuis quelques années, l'utilisation de bases de données s'est imposée dans les travaux **d'histoire de l'art** reposant sur des recherches thématiques, qu'il s'agisse d'architecture moderne ou des œuvres d'un même peintre. De telles bases facilitent le travail de l'historien de l'art puisque grâce à elles, il est possible d'élaborer des statistiques quant à, par exemple, la récurrence d'une couleur dans les œuvres d'un peintre ou l'emploi du fer dans un type donné d'architecture, et de les traduire en graphiques. De nouvelles investigations peuvent ainsi être menées grâce aux résultats obtenus.

Aussi, lorsque nous avons entamé notre thèse sur les représentations de l'Ascension du Christ en Occident entre le IXe et le XIIIe siècle, regroupant quelques 250 œuvres, le recours à cette méthode de travail nous est apparu évidente. La rencontre avec Jean-Claude Régnier a cependant bouleversé notre manière de penser et d'exploiter l'utilisation des statistiques dans des travaux d'histoire de l'art. En effet, la découverte de l'analyse statistique implicative a donné naissance à une nouvelle manière de concevoir la recherche en **iconographie médiévale** : de tels outils ouvrent de nouvelles perspectives dans l'approche et l'analyse des thèmes en montrant des liens qui resteraient inexplorés sans le recours à ces procédés de recherche.

Après une brève présentation de l'iconographie médiévale, nous présenterons les travaux antérieurs portant sur les études thématiques : des chercheurs ont déjà abordé l'image

médiévale sous la problématique du thème, mais les moyens mis à disposition à ces époques ne permettaient pas une analyse aussi poussée que celle proposée dans ce travail de thèse. Puis nous nous pencherons sur le cas précis de l'Ascension du Christ dans le cadre de l'analyse statistique implicative : de la constitution de la base de données, support de travail préliminaire, aux premiers résultats obtenus par le biais de l'ASI.

2 L'analyse statistique en iconographie médiévale

L'étude des images médiévales s'oriente autour de deux axes de recherche principaux. La monographie consiste à étudier et interpréter une œuvre ou un programme iconographique[1], offrant une interprétation spécifique à l'objet étudié. L'étude thématique se détache de l'image pensée en tant qu'individu pour la mettre en perspective selon une caractéristique précise, aussi bien le support sur lequel elle est représentée que la représentation de tel détail ou d'un thème. Ce type de recherche nécessite le recours à l'analyse statistique afin de traiter les multiples données obtenues. Comment est alors envisagée l'analyse statistique dans un tel travail ?

2.1 Définition de l'iconographie médiévale et quelques-unes de ses problématiques

Du grec *eikon* (image) et *graphein* (écriture), l'iconographie consiste à étudier et interpréter les images. On emploie aussi le terme d'iconologie (que l'on peut traduire par discours de l'image, *logos* signifiant en grec discours), dont la première utilisation remonte au XVI^{e} siècle sous la plume de Cesare Ripa : cet auteur recense tous les attributs et éléments nécessaires à la reconnaissance d'un personnage ou d'un épisode avant d'être utilisée selon une acception beaucoup plus scientifique, basée sur l'interprétation des œuvres. Le terme « iconographie » reste toutefois le plus usité. L'iconographie médiévale, s'intéressant dans une optique scientifique d'interprétation, aux images du Moyen Age est une science somme toute récente : en effet, la redécouverte du Moyen Age au XIX^{e} siècle entraîne un regain d'intérêt pour les images de cette époque. Mais celles-ci furent d'abord perçues comme une illustration des textes et même, à tort, targuées de « Bible des pauvres » (Mâle, 1986). Les recherches avançant, l'image médiévale s'est avérée beaucoup plus complexe à aborder. Il est apparu qu'elle s'interprète non pas en fonction d'un texte, mais selon son contexte de création, qu'il s'agisse de la place qu'elle occupe dans un édifice ou un manuscrit, de son lien avec les autres images du même objet sur lequel elle est représentée, ou encore de sa part active dans la vie religieuse d'une communauté au sens large[2]. Une image prise isolément ne peut être correctement analysée tant il est vrai qu'elle n'existe que très rarement par et pour elle-même : elle s'inscrit toujours dans un environnement (historique, géographique, architectural, politique, etc.) qu'il est nécessaire de prendre en compte pour la comprendre

[1] Une image médiévale appartient souvent à un ensemble d'images au sein d'un même objet (édifice, manuscrit etc.), appelé programme iconographique, et qui font sens les unes par rapport aux autres. Cela demande l'examen de toutes les images, leur situation les unes par rapport aux autres ou encore leur place dans l'objet (à l'entrée de l'église ou dans le chœur, partie sacrée de l'édifice).

[2] Il a été prouvé à maintes reprises que l'image peut servir de support lors des prêches des hommes d'Eglise, illustrer une prière liturgique ou encore prendre part à des conflits politiques.

correctement. A cette première démarche, s'en ajoute une seconde qui considérera, par-delà l'image comme objet, l'image comme thème et partira donc de ce dernier dans le but non pas de décrypter l'image, mais bien le thème en lui-même. Ainsi, le cheminement consiste à étudier une série d'images, ici des images médiévales, pour en expliciter la variabilité, car elles sont trop souvent considérées comme fixes et avec une variété de créations limitée, et pour en interpréter le thème en fonction des permanences et des particularités ainsi dégagées. Le caractère individuel, singulier, de chaque image est dans un premier temps ignoré pour favoriser une vision globale du thème, avant de lui être restitué. Le raisonnement part de la décomposition des images en fonction d'un spectre de **signifiants** pour en dégager de nouveaux **signifiés**. Celui-ci se base sur une analyse objective de l'image, dont les qualités esthétiques ou spirituelles s'effacent au profit d'un découpage formel mais dont le résultat est identique à une analyse iconographique en tant que telle[3] : élaborer des hypothèses d'interprétation selon les signifiés dégagés par l'observation et la confrontation de signifiants.

Ce type d'études ne pouvait se développer sans les moyens technologiques actuels aussi bien dans les méthodes de calcul que dans les nouvelles possibilités de recherche offertes par Internet. Quelques chercheurs entreprirent une démarche en ce sens malgré les difficultés rencontrées, en particulier, faute de moyens adaptés.

2.2 Les pionniers de l'étude thématique par une approche statistique.

Dès les années 1920, des historiens de l'art engagèrent des travaux thématiques. Si nous nous basons sur le cas d'étude proposé, en l'occurrence, l'iconographie de l'Ascension du Christ en Occident entre le IX^{e} et le XIII^{e} siècle, nous notons deux publications de référence : un article écrit en 1930 par Hubert Schrade (Schrade 1930) et un livre datant de 1935 dont l'auteur est Helena Gutberlet (Gutberlet 1935). Tous deux s'attachent à dresser le portrait du thème de l'Ascension à travers un large corpus d'images. Se basant sur leurs observations, ils parviennent à présenter des typologies et une analyse pertinente de l'Ascension. Ces premiers travaux peuvent désormais être actualisés et approfondis grâce aux outils dont les chercheurs peuvent disposer.

En 1996, l'analyse statistique fait véritablement son entrée dans la recherche de l'iconographie médiévale à travers les travaux de Jérôme Baschet. Ce dernier signe un article cette même année (Baschet 1996) montrant de quelle manière l'analyse statistique contribue à une nouvelle vision de l'image médiévale en mettant en avant son caractère mobile dans la mise en image d'une série et, par conséquent, dans son interprétation. Il emploie le terme d'analyse sérielle, partant du fait que l'objet de l'étude porte sur une série d'images. Ainsi explique-t-il que par le morcellement d'une série partageant un thème ou un signifiant identique, il s'agit de faire ressortir une synthèse collective, dont le paradoxe est d'étudier les images selon une approche globale tout en mettant en avant leur spécificité. Il se dégage plusieurs sources de renseignements :

1. des récurrences permettant la création de typologies que l'on peut qualifier de références dans leurs grandes lignes,
2. des particularités, en fonction des supports par exemple,
3. des éléments peu courants mais suffisamment réitératifs pour être retenus.

[3] Nous entendons par là l'étude spécifique d'une image dans son contexte de création

Même s'il existe une marge d'erreur due aux œuvres manquantes, les résultats restent significatifs et peuvent être exploités pour apporter une analyse pertinente de l'image, et mettre en évidence le caractère mobile de cette dernière au Moyen Age, car ainsi que le souligne Jérôme Baschet (2000, p. 180) « si certains aspects stables répétitifs doivent être décrits comme tels, la dimension créative de chaque œuvre ne peut être ignorée », mettant ainsi en valeur la « mobilité figurative » (1996, p. 110) de l'image médiévale. Nous retrouvons déjà là les conditions d'un recours pertinent à l'analyse statistique implicative – ASI, en iconographie médiévale dans la mesure où elle prend en compte la complexité de l'objet d'étude. L'ASI permet de considérer simultanément un ensemble important de variables qui caractérisent de manière extrêmement détaillée chaque image, pour pouvoir ensuite intégrer cette singularité dans un système d'analyse plus vaste qui met en comparaison un nombre qui n'est limité que par les accès aux objets, d'autres images. De cette comparaison, naît la mise en lumière des caractéristiques que l'ensemble des images a tendance à partager ou celles qui demeurent propres à la singularité de quelques-unes.

Par cette ouverture de la recherche, de nouvelles problématiques apparaissent et permettent la reprise d'études référentielles en poussant encore plus loin les investigations.

3 Les représentations de l'Ascension du Christ en Occident entre le IX^{e} et le XIII^{e} siècle : problématiques liées à l'analyse statistique implicative

Ce travail est conduit dans la perspective de l'analyse iconique de l'image, c'est-à-dire de la manière dont cette dernière est agencée dans sa structure interne à partir de l'idée que l'exploitation du contenant influe directement sur le contenu, pour aboutir à la mise en lumière de nouveaux signifiés.

3.1 Présentation de l'objet de la recherche

En iconographie médiévale, il y a lieu de tenir compte de la complexité du thème iconographique étudié, qui se situe à plusieurs degrés. Notre choix s'est ici porté sur le thème de l'Ascension du Christ. Pour ce faire, nous avons tout d'abord défini un vaste territoire géographique ainsi qu'une large période chronologique pour obtenir des résultats significatifs, probants quant à la question de la mise en image de ce thème. Ensuite, plutôt que de limiter la recherche à un type de support, nous avons jugé plus pertinent et plus riche de prendre en compte la diversité de subjectiles possibles plutôt que de nous restreindre à un seul type. En effet, il est admis que le public ayant accès à l'image se différencie en fonction du support, que la technique d'élaboration lui est soumise ou que les dimensions de l'œuvre en dépendent : ce sont autant de facteurs à intégrer dans la lecture de l'image médiévale qui en font sa richesse. Cette prise en compte de la variabilité du support permet de déterminer avec plus de précisions dans quelle mesure le thème est soumis aux contraintes physiques comme aux caractères externes liés au support ou encore s'il s'en affranchit pour servir une

pensée, une réflexion. Pour expliciter notre propos, nous prenons deux exemples de support que tout oppose a priori : les lettres enluminées et les tympans d'église[4].

FIG. 1 – *Tours, Bibliothèque municipale, ms 0013, f.130, 1225.*

FIG. 2 – *Montceaux-l'Etoile (71), église Saint-Pierre, façade occidentale.*

Les lettres enluminées comme le P enluminé conservé à la bibliothèque municipale de Tours (FIG. 1), représentent le plus souvent les pieds du Christ dans la partie haute de l'image tandis que les témoins, souvent réduits à quatre, assistent à la scène, le bas de leurs corps tronqués par le cadre de la lettre.

Les tympans disposent d'une surface d'un tout autre ordre de grandeur que les artisans ont exploité de la manière suivante comme sur le tympan occidental de Montceaux-l'Etoile, en Saône-et-Loire (FIG. 2) : le Christ est placé dans la partie haute, appelée tympan, debout, dans une mandorle qui peut être portée par deux anges s'accordant aux retombées de l'arc. Le linteau, constituant la bande verticale sous le tympan, permet de placer les témoins sous le Christ qui s'élève. Ainsi, il semble que l'image s'adapte à son support. Nous sommes facilement tentés de penser qu'une règle de représentation est ainsi établie pour chaque support. Et pourtant, loin s'en faut. S'il est somme toute logique que le support influe sur la représentation, une diversité se fait jour, que nous cherchons à déterminer et à interpréter. C'est à ce premier titre qu'intervient l'analyse statistique, afin d'expliciter les variables qui agissent au sein de l'image et de les caractériser.

Reprenons l'exemple des représentations du Christ dans les lettres enluminées qui nous semble tout à fait illustratif. Considérons alors la variable catégorielle «Position du Christ dans les Lettres enluminées » modélisée a priori en sept modalités que nous avons définie sur le sous-échantillon de 83 subjectiles « Lettres Enluminées » extrait de l'échantillon des 246 subjectiles dont nous disposons à ce jour. Nous avons nommé ces modalités : LettresEnluminéesTypk avec k=1 à 7 mais ne donnerons pas, dans le cadre de cet article, les détails des critères qui définissent chacune. Nous en rapportons le tableau :

[4] Les lettres enluminées sont les majuscules ornées des manuscrits : peintes et ne dépassant pas quelques centimètres carrés, elles accueillent souvent des saynètes historiées, et sont destinées à un public restreint, seuls les clercs ayant accès aux manuscrits. Les tympans sont pour leur part des pièces sculptées en demi-lune au-dessus des portes des églises : exposés à la vue de tous, leur taille est plus considérable mais la technique de la sculpture de même que la forme semi-circulaire ne facilitent pas la mise en image d'un thème.

LettresEnluminées						
Type1	Type2	Type3	Type4	Type5	Type6	Type7
72	2	0	1	1	3	3

TAB. 1 –*Effectifs de la variable « Position du Christ dans les lettres enluminées ».*

La pensée commune voulait que seuls les pieds du Christ soient représentés dans le cadre de la lettre, à quelques exceptions près. En réalité, tout en observant un schéma récurrent, il se fait jour une pluralité arguant d'une créativité artistique : 1/8ème des images sort de ce schéma de référence. De plus, une profonde réflexion s'en dégage puisque ces différences de représentation induisent une lecture différente de l'image.

La variabilité est aussi identifiable au niveau d'autres éléments composant iconiquement l'image, comme le nombre de témoins, la représentation ou non d'attributs, d'éléments naturels ou de décor, etc. A travers toutes ces observations, l'image prend sens en se dédoublant du thème.

3.2 Orientations de la recherche

La recherche s'organise alors autour d'un double axe, l'un basé sur la mise en image du thème et le second autour de l'image en tant que représentation d'une idée à travers le thème figuré. Dans le premier cas, il s'agit de dresser les caractéristiques et spécificités du thème iconographique de l'Ascension du Christ, par la décomposition d'un nombre considérable d'images pour obtenir des résultats probants. Cette décomposition se veut la plus objective possible, se basant sur une enquête visuelle. Elle prend en compte des éléments humains ou naturels, et à partir de ceux-ci, elle opère une sélection de nombre, de gestes ou de positions pour les personnages. Ainsi, dans une œuvre de notre échantillon, figurent quatre anges, debout, volant et parlant ou en buste, immobile, accueillant le Christ. Sont aussi pris en compte des éléments de figuration pour les objets : par exemple, la croix possède une bannière ou non, est dans la main du Christ, est située au-dessus de ce dernier, etc.. La prise en compte, sans exception, de toutes les images de l'échantillon construit au fil de notre investigation dans les ressources les plus diverses, permet d'élaborer quelques éléments d'ordre théorique quant à la mise en forme du thème. Si nous considérons la variable catégorielle « type de subjectiles » que nous avons modélisée a priori en 20 modalités en fonction de notre cadre théorique initial dans le champ de l'Histoire de l'art, nous obtenons la distribution suivante que nous rapportons dans le tableau ci-après.

Nous pouvons ainsi évaluer la fréquence de chaque type de support pour mieux percevoir la question de la représentation du thème de l'Ascension du Christ. Il ressort clairement une prédominance des subjectiles de type « Obj01 = Enluminure ». Ce résultat n'est pas surprenant en soi pour le chercheur : les manuscrits étaient nombreux au Moyen Age et à défaut d'être entièrement illustrés, ils possédaient souvent des lettres historiées.

Ces observations comme nous l'avons dit, montrent qu'au gré de nos enquêtes pour compiler des images de toutes catégories, le thème de l'Ascension du Christ apparaît sur 18 types de subjectiles parmi les 20 attendus théoriquement et selon l'ordre suivant :

Code du subjectile	effectif	%	Type	Code du subjectile	effectif	%	Type
Ob01	83	33,88%	Enluminure	Ob06	3	1,22%	Croix
Ob04	50	20,41%	Pleine page	Ob07	2	0,82%	Chapiteau
Ob02	36	14,69%	Miniature	Ob08	2	0,82%	Pilier
Ob05	26	10,61%	Plat de reliure	Ob03	1	0,41%	Médaillon
Ob11	18	7,35%	Façade	Ob10	1	0,41%	Crosse abbatiale
Ob16	6	2,45%	Peintures murales	Ob12	1	0,41%	Fonts baptismaux
Ob17	5	2,04%	Vitrail	Ob15	1	0,41%	Tissu
Ob09	4	1,63%	Autel	Ob19	1	0,41%	Porte
Ob13	4	1,63%	Reliquaire	Ob20	1	0,41%	Façade écran

TAB. 2 – *Les types de subjectile rangés par ordre décroissant de fréquences dans l'échantillon actuellement disponible (février 2009).*

Cette information ne nous renseigne toutefois pas directement sur la mise en image du thème. D'autres résultats en revanche aborderont le thème sous un angle iconographique et livreront d'autres informations.

Le cas suivant est très intéressant. En étudiant la position du Christ, nous avons établi sept positions possibles que nous présentons dans le tableau ci-dessous

Codage des modalités	Modalités de la variable « Position du Christ »	Effectifs	% (n=244)
PosJC_0	Absent	1	0,41%
PosJC_1	Pieds	108	44,26%
PosJC_2	En pied	109	44,67%
PosJC_3	Assis	9	3,69%
PosJC_4	Buste	6	2,46%
PosJC_5	Inconnu	7	2,87%
PosJC_6	Pieds non représentés	4	1,64%

TAB. 3 – *Variable « Position du Christ ».*

Il apparaît que les positions 2 et 3 sont les plus adoptées : 108 images ne représentent que les pieds du Christ et 109 le figurent en entier. Cependant, nous avions précédemment noté que la position 2 était essentiellement utilisée dans les lettres enluminées. Si l'on compare les les tableaux 1 et 3, nous constatons que sur 83 lettres enluminées, 72 d'entre elles représentent les pieds du Christ tandis que la figure 5 et le tableau 2 montrent que cette position est présente sur 108 supports. Ce type de figuration n'est donc pas l'apanage des lettres enluminées et ces résultats confirment la diversité de l'image médiévale comme le montre l'exemple des différentes positions du Christ dans les lettres enluminées, contrairement à l'idée reçue.

Les tableaux orientent alors la recherche vers l'interprétation de l'image : reprenant le cas des lettres enluminées, nous nous demandons quelles conséquences une autre position du Christ engendre sur l'organisation iconique de l'image et du même coup, sur la lecture de l'image.

Lors du morcellement des images, l'on se rend compte de l'apparition de certains détails qui avaient paru insignifiants lors de la première description de l'image, noyés dans une

masse de composants. Le travail systématique de morcellement permet de mieux évaluer leur fréquence qui, même si elle paraît moindre par rapport au nombre d'images, marque une intention qui retient l'attention. Il en fut ainsi pour la représentation du soleil et de la lune : apparaissant sur quelques images, leur présence semblait anecdotique. Le fait de l'inclure tout de même dans la base de données a mis en avant son importance. De nouvelles recherches ont été menées et il s'est avéré qu'une telle représentation reprenait un sermon du VI^{e} siècle et apportait à l'image une lecture ecclésiologique, c'est-à-dire en lien avec la naissance de l'Église.

Le deuxième axe est basé sur l'image en elle-même. Lorsqu'une donnée apparaît à une basse fréquence, elle demande une reprise des images l'incluant. Certaines se révèlent alors. Reprenons le cas du soleil et de la lune présents sur les Ascensions. La British Library de Londres possède un sacramentaire, (livre destiné à l'évêque et donc, personne d'autorité représentant l'Église), sur lequel le Christ s'élève devant quatre témoins. . Ce dernier possède une croix à bannière et est entouré du soleil et de la lune. Le regard est attiré par le personnage à la droite du Christ : ses mains sortent du col de son vêtement, qui laisse imaginer qu'il s'agit d'une aube et donc, d'un vêtement porté uniquement par les hommes d'Église. Plusieurs points de convergence sont ici pris en compte pour donner à l'image toute sa dimension ecclésiologique. En elle-même, la représentation de l'Ascension intégrant le soleil et la lune reprend un sermon de Grégoire le Grand, basée sur la naissance de l'Église et comparant le soleil s'élevant au Christ et la lune restant à l'Église, selon un verset du livre d'Habacuc dans la Septante[5]. De plus, elle distingue parmi les témoins un homme d'Église, évêque ou abbé, dépositaire de la parole du Christ sur terre depuis le départ de celui-ci. Enfin, cet ecclésiastique est le seul à avoir accès au livre, lequel sert à des fins liturgiques. Une connexion est ainsi établie entre l'image, le public et l'objet en lui-même, pour offrir un discours ecclésiologique complet, centré sur la naissance spirituelle de l'Église lors de l'Ascension, sur un livre recueillant des pratiques liturgiques, s'adressant à un ecclésiastique non seulement par l'objet, mais aussi par l'image qui l'associe à la naissance et à l'extension de l'Église.

Tous ces détails ont au préalable été saisis dans une base de données, dont la précision est la fondation et la garantie d'un travail d'analyse concluant.

4 Conception d'une base de données, travail préliminaire à l'ASI

Le logiciel utilisé pour ce travail est Excel, facilitant la création de premiers graphiques et surtout, autorisant une souplesse de travail sur la base de données en elle-même. En effet, au fur et à mesure des descriptions d'images, il se fait jour de nouvelles entrées à intégrer, comme des subtilités dans les actions des personnages ou de nouveaux éléments dont la rareté avait dans un premier temps entraîné leur exclusion, mais qui se sont révélés suffisamment réitérés pour finalement prendre place dans l'étude. Il fallait aussi pouvoir ajouter les œuvres découvertes tout au long des recherches menées.

Cependant, la difficulté majeure est de réussir à traiter la masse documentaire pour dégager les constantes et les variantes de l'image sans nier pour autant l'individualité de

[5] Hab 3, 11 : « Le soleil s'est élevé et la lune est restée ». La Septante est la traduction grecque de la Bible, qui précéda historiquement la Vulgate, version latine de la Bible traduite de la Septante

cette dernière. L'enjeu est donc double pour ce travail : il faut dresser des typologies en fonction de l'exploitation de la surface iconique afin de savoir dans quelle mesure ladite surface iconique joue un rôle dans la représentation d'un thème, et faire apparaître la fréquence de certains éléments inaccoutumés pour les mettre en parallèle avec l'image.

4.1 Élaboration des données par rapport aux images

L'image est l'individu statistique de référence. Ce que nous abordons maintenant constitue les variables ou les vecteurs-variables statistiques construites pour réaliser l'étude de l'image,. qui traduisent le morcellement dont nous avons parlé plus haut. Avant d'entrer dans sa description iconique, quelques renseignements pouvant influer sur son traitement sont livrés : la nature du subjectile, la date ainsi que l'origine géographique si elle est connue.

Vient ensuite une donnée d'ordre général, appelée typologie de rattachement. Basée sur la représentation du Christ qui constitue le schéma directeur référentiel de l'image, elle jauge l'image dans son ensemble avant d'être soumise à un découpage précis, correspondant à ses principaux groupes de composition. Le premier d'entre eux étudie les témoins : leur nombre, la présence ou non de certains, s'ils tiennent des objets spécifiques. L'exercice se complique avec l'étude des anges. Ils peuvent occuper plusieurs places dans l'image, opérer différentes actions en même temps, être situés à plusieurs niveaux sur l'image. Plutôt que de démultiplier la base de données à l'envi au risque de la rendre inexploitable, il a été décidé de tirer profit de ces polyvalences et de créer une entrée manifestant cette pluralité de sens. Le troisième ensemble concerne le Christ lui-même : constituant l'axe dynamique de l'image, il nécessite un morcellement spécifique. Celui-ci concerne aussi bien ses positions dans l'image que ses actions, ou dans son environnement immédiat la présence d'éléments spécifiques, lesquels peuvent agir directement sur la représentation. La présence de la main de Dieu illustre ce propos : la base de données permettra de mettre en lien la représentation de la main de Dieu avec l'action et la position du Christ au moment de l'analyse statistique implicative. Enfin, un dernier groupe comprend les éléments intervenant dans l'image sans être expressément rattachés à des personnages, mais dont la présence modifie la mise en place et l'interprétation de l'image.

Chaque critère est symbolisé par des codages numériques ou alphanumériques correspondant à un type spécifique. Dans un cas simple comme un personnage précis, il existe trois entrées : 0 signifie l'absence du personnage, 1 sa présence, tandis que 2 qualifie une incertitude dans l'identification. Les images sont ensuite traitées une à une en fonction des 36 critères de description, eux-mêmes subdivisés selon les items préétablis. Dit autrement, il s'agit des variables statistiques construites qui sont soit quantitatives, soit qualitatives, et de leurs résultats qui sont respectivement des nombres ou des modalités, des types, des catégories.

Ce travail minutieux est indispensable à une exploitation pertinente des données pour obtenir de premiers résultats statistiques, d'ordre quantitatif et qualitatif. Nous ne rapportons pas la base de données qui n'est autre qu'un tableau de séries statistiques.

4.2 Les méthodes d'analyse quantitatives et qualitatives conduisant à un premier bilan d'interprétation

La complémentarité du quantitatif et du qualitatif dans les méthodes d'analyse est une nécessité à prendre en compte pour nous conduire vers l'analyse statistique implicative à proprement parler. En ce qui concerne le quantitatif, il apparaît sous deux formes : dans la nature des variables construites, variables quantitatives discrètes ou continues, et dans le comptage des occurrences qu'elles soient de nature quantitative ou qualitative. Ainsi en ce qui touche aux témoins assistant à la scène, avons-nous construit une variable quantitative discrète « nombre de témoins ». Nous rendons compte de la distribution des effectifs de cette variable par le diagramme en bâtons ci-dessous :

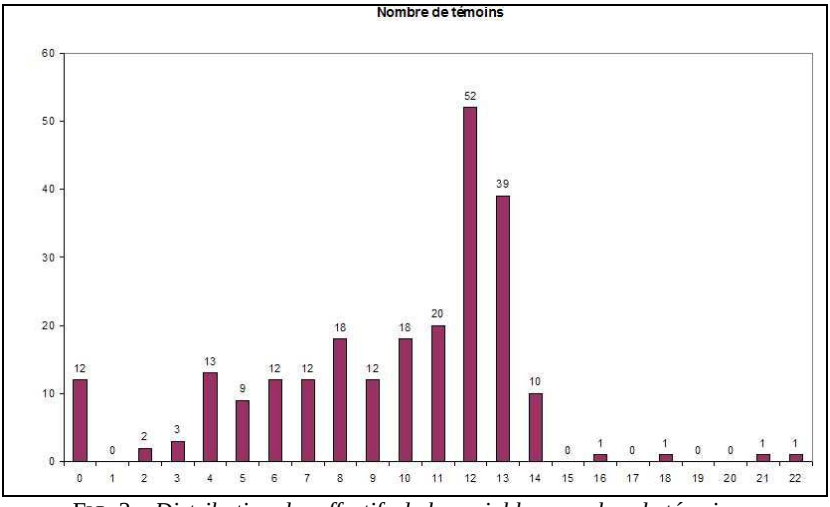


FIG. 3 – *Distribution des effectifs de la variable « nombre de témoins».*

La moyenne est de l'ordre 10 et la fluctuation moyenne, l'écart-type, de l'ordre de 4. Les deux pics (modes) orientent les recherches et mettent au jour une question amenant à une piste de réflexion. Le graphique ci-dessus (Fig.3) montre clairement que les images privilégient la présence de 12 ou 13 personnages, témoins de l'Ascension du Christ. Le texte biblique de référence mentionne la présence des apôtres qui sont onze lors de l'Ascension, Judas s'étant suicidé et Mathias n'ayant pas encore rejoint le groupe des apôtres. Ce nombre est suivi par nombre de théologiens médiévaux. Il s'agit d'expliquer ce surplus numéraire, qui s'inscrit autant dans la pensée médiévale centrée autour d'une perfection numéraire du nombre, que dans un discours réfléchi en lien avec la fondation de l'Eglise. La première question qui vient à l'esprit est l'identité de ces personnages supplémentaires. Les images livrent une première réponse : la Vierge et saint Paul. Mais la présence de l'un n'impose pas obligatoirement celle du second : la Vierge apparaît un nombre de fois bien plus important que saint Paul et pourtant, lorsque saint Paul est représenté, la Vierge n'est pas systématiquement présente. De surcroît, 105 images représentent plus de 11 témoins mais la Vierge apparaît 173 fois. Ces résultats mettent en avant le rôle prépondérant de la Vierge

dans les représentations de l'Ascension et dirigent une piste de réflexion quant à la signification de la Vierge dans un tel contexte.

En ce qui concerne le qualitatif, il apparaît aussi sous deux formes : dans la nature des variables construites, variables qualitatives nominales ou ordinales, et dans les interprétations relatives aux informations qui ressortent des quantités quant à leur sens de variation, leur position relative, etc.. L'analyse qualitative questionne pour sa part les critères déterminant un point précis de l'image. Outre les notions de présence-absence-incertitude mentionnées plus haut, les critères renseignent sur la position du Christ par exemple : savoir s'il est de face, de profil, de trois quart, ou de dos. En fonction de cette position, l'effet produit sur le spectateur ainsi que l'interprétation diffèreront, mais aussi la disposition de l'image et les éléments représentés, d'où la nécessité de prendre connaissance de cette composante afin de croiser judicieusement les données dans un travail ultérieur, comme par exemple la représentation de profil du Christ et le mouvement effectué ou encore les attributs portés tels que la croix ou le livre.

Ce traitement qualitatif met en relief le caractère dichotomique de l'image médiévale : l'étude fractionnée fait ressortir des schémas généraux tout en conservant la propriété mobile de l'image, autorisant une double étude qui s'établit soit par élément (donc par signifiant) soit par image (donc par individu).

L'analyse qualitative montre aussi qu'un seul signifiant peut aussi coordonner toute une image, iconiquement et iconographiquement. Nous basant toujours sur la position du Christ, nous voyons que celle-ci joue sur l'organisation de l'image en sa qualité d'axe vertical dynamisant la composition et bousculant le rapport du spectateur à l'image ou la vision de l'épisode et donc, son interprétation. Ainsi, la position de face du Christ le déconnecte de l'image mais en même temps, le Christ regarde le spectateur et l'implique directement dans la scène. Par ailleurs, dans 95% des images, il est immobile, ce qui renforce le caractère hiératique de l'image en opposant le Christ et les apôtres dans leurs actions et donc, dans leur essence. Les méthodes d'analyse statistique employées ont rendu possible l'explicitation de telles liaisons et même de présomptions de liens de cause à effet entre la position et l'action par exemple, influant directement sur la lecture de l'image.

4.3 Intérêt d'une base de données en iconographie médiévale

Le chercheur en iconographie médiévale est fréquemment confronté aux difficultés de la recherche inhérente à son sujet d'étude. Tout d'abord, dans une étude statistique de ce type, le corpus élaboré ne peut prétendre à l'exhaustivité : nombre d'œuvres ont disparu et dans une proportion incalculable. De plus, si l'outil Internet permet d'avoir accès à de nombreuses reproductions, il n'en reste pas moins un nombre conséquent qui demeure inconnu, soit dans des collections privées, soit dans des réserves de bibliothèques ou de musées non portées à la connaissance du public. Néanmoins, nous postulons que les œuvres restantes apportent des éléments significatifs sur la mentalité d'une époque. De même que lorsque des instituts réalisant des sondages se basent sur un panel représentatif, le chercheur tente d'adopter une attitude similaire dans l'élaboration de son corpus : il réunit le plus de représentations possibles, estimant qu'elles constituent un échantillon suffisamment représentatif. La base de données permet également de juger de la pertinence de cette méthode : les permanences des images prouvent en effet que malgré leur nombre désormais restreint au vu de la production artistique médiévale, les images traduisent une pensée commune dans la conception tant intellectuelle que formelle du thème qu'elles figurent.

La transcription des œuvres est aussi sujette à la part d’interprétation du chercheur, qui s’effectue soit de manière objective, liée à la qualité de l’œuvre, soit subjective, c'est-à-dire en fonction de la lecture que le chercheur fait de tel mouvement par exemple. Les œuvres médiévales ont pu subir les outrages de l’histoire et être altérées : il arrive souvent que les peintures de manuscrit aient été grattées pour être ingérées par le lecteur en vertu des pouvoirs miraculeux accordés à certaines. Lorsque les visages ou les personnages ont été ainsi abîmés, leur identification est malaisée mais en se basant sur des œuvres ressemblantes, il est possible d’en dresser un portrait somme toute assez fidèle, tout en conservant bien sûr une certaine réserve : la moindre réticence aboutit à une entrée neutre de type « inconnu ». Il en va de même pour les reproductions pêchant par leur qualité : certaines sont floues ou dans le cas d’images Internet, trop pixellisées, ce qui rend parfois difficile le travail de description. Une démarche identique aux images abîmées est alors utilisée par le biais de comparaison avec d’autres images du même type.

La description est aussi soumise à la perception sensible du chercheur. Des actions, des attitudes peuvent porter à confusion : nous avons mentionné plus haut les multiples actions des anges, le cas est identique pour faire la distinction entre le Christ grimpant et le Christ montant. En toute honnêteté, le chercheur tranchera en faveur d’une conjecture à un moment donné, puis hésitera et optera pour la seconde, pour parfois revenir au premier choix.

Il existe donc une part d’interprétation inéluctable et nécessaire en iconographie, mais qui doit toujours rester objectivable. La base de données, par ses critères établis, oblige à ne pas tomber dans le piège de décrire ce que l’on veut voir mais bien ce qui est représenté. Elle agit comme un garde-fou auprès du chercheur en faisant appel à une mise à distance dans la description.

La base de données ainsi que les premiers graphiques élaborés permet d’aborder l’image objectivement et de croiser les données pour mettre en avant des signifiants aboutissant à de nouveaux signifiés. Le cas du soleil et de la lune précédemment cité est en ce sens explicite : d’un détail anodin de prime abord, nous sommes parvenus à une interprétation en rapport avec la fondation de l’Église.

Ce travail révèle aussi toutes les ambivalences de l’iconographie. Outre la part d’interprétation inhérente à toute enquête descriptive, il a dévoilé la complexité de l’image qui conjugue plusieurs actions simultanément, au contraire d’un texte dans lequel les actions se succèdent au rythme de la lecture. Dans une image, le spectateur embrasse et assimile les actions d’un seul regard, d’où la difficulté ensuite de les désolidariser mais aussi l’intérêt de relever ces polyvalences.

La création de la base de données marque aussi un double mouvement qui semble contradictoire mais qui a pour résultat de compléter la lecture de l’image. En effet, en traitant ainsi un nombre considérable d’images afin de traduire les résultats dans le langage de la statistique, les images sont dans un premier temps dépossédées de leur individualité pour être noyées dans une masse commune en fonction de critères objectifs qui les décomposent en autant de pièces de puzzle qu’il s’agit ensuite de recréer non pas par image, mais par signifiant par exemple. Mais le traitement statistique, en mettant en avant les permanences et les spécificités des compositions, redonne à chacune son caractère identitaire soit par le biais des signifiants, soit par son originalité propre.

La base de données offre ainsi de premiers résultats par le biais de diagrammes, et qui peuvent être approfondis par le recourt à l’ASI étayée par l’utilisation du logiciel CHIC. Ainsi que tente de le présenter cet ouvrage même pour rendre accessible l’ASI aux non-experts, il devient alors possible de marier certaines variables dans un même graphique qui,

cette fois-ci, vise à rendre compte d'une organisation de l'ensemble des variables binaires décrivant par le détail les images du corpus par une relation de quasi-implication associée à un niveau de confiance choisi par le chercheur. Cela conduit à se doter de critères explicitables pour explorer les niveaux d'inclusion entre les catégories d'éléments sélectionnés, partant de l'occurrence la moins présente dans tout le corpus à la plus présente.

5 Utilisation de CHIC dans le contexte de l'iconographie

Dans notre étude, l'exploration des données du champ de l'iconographie médiévale par l'analyse statistique implicative apporte un nouveau point de vue pour la lecture des images. Sur une sélection de critères que traduisent des variables binaires et selon un indice de confiance que nous avons fixé au préalable, elle nous permet d'expliciter des liens entre les données et qui nous servent ensuite à l'analyse même de l'image, à condition de juger au préalable de la pertinence des connexions dans le champ de l'iconographie médiévale.

5.1 Création de la base de données de référence pour l'analyse statistique implicative

Il est nécessaire de transformer les variables catégorielles de la base de données initiale en variables binaires : c'est à dire que chaque modalité de la variable qualitative devient une variable qui prend ses valeurs dans l'ensemble {0 ; 1}. Cela revient à associer à chaque variable catégorielle à k modalités, un vecteur-variable de dimension k dont les composantes représentent la présence (1) ou l'absence (0) des propriétés traduites par les modalités. Rappelons au passage la richesse informative de la variable binaire 1-0 qui est à la fois une variable qualitative présence-absence et une variable quantitative puisque la somme des résultats donne l'effectif correspondant à la modalité.

Prenons par exemple la variable « Position du Christ » qui prend ses valeurs dans un ensemble de modalités {PosJC_0; PosJC_1; PosJC_2; PosJC_3; PosJC_4; PosJC_5; PosJC_6} se transforme en le vecteur-variable (PosJC_0 ; PosJC_1 ; PosJC_2 ; PosJC_3 ; PosJC_4 ; PosJC_5 ; PosJC_6) de dimension 7 où chaque composante PosJC_k est une variable binaire prenant ses valeurs dans {0 ; 1}. Dans ce cas particulier, le 7-uplet contient une fois la valeur 1 et six fois la valeur 0 car les modalités sont exclusives.

Sous l'effet de cette transformation, la base de données initiale qui correspond au tableau des séries statistiques est démultipliée considérablement. Cette nouvelle base n'est autre qu'un tableau disjonctif total. En ce qui nous concerne, la base initiale comporte 30 variables catégorielles et 3 variables quantitatives, c'est donc un tableau de 245 lignes (les supports de l'échantillon en date de février 2009) et de 33 colonnes dont nous rappelons le sens dans le tableau ci-dessous.

Code	Objets des variables	Code	Objets des variables
V01	Nature des subjectiles	V19	Présentation du Christ
V03	Origine	V20	Action du Christ
V04	Typologie de rattachement	V21	Situation du Christ
V06	Vierge	V22	Geste de bénédiction du Christ
V07	Pierre	V23	Croix du Christ
V08	Jean	V24	Livre du Christ
V09	Paul	V25	Mandorle du Christ
V10	Symétrie	V26	Nuées
V11	Livre	V27	Main de Dieu
V12	Phylactère	V28	Végétaux
V13	Rotulus	V29	Sol
V15	Situation des Anges	V30	Inscriptions
V16	Position des Anges	V31	Couronne du Christ
V17	Action des Anges	V32	Lune
V18	Position du Christ	V33	Cube

TAB. 4 – *Variables catégorielles de description des subjectiles.*

Dans la transformation en variables binaires, le nombre de variables passe à 118, c'est à dire que la nouvelle base est un tableau de 245 lignes sur 118 colonnes. Chaque colonne correspond à la présence ou l'absence du caractère étudié.

Nous ne reprendrons pas le détail de la démarche ASI qui est présentée dans cet ouvrage. Nous ne faisons que rappeler quelques points clés en vue d'un usage de l'ASI par des chercheurs en Histoire de l'art souvent peu familiarisés avec des méthodes statistiques. Nous avons dans un premier temps choisi la valeur $1-\alpha=0{,}99$ pour le niveau de confiance en la règle que donne la quasi-implication $a \Rightarrow b$, dite encore quasi-règle d'implication. Il s'agit d'un critère, appelé intensité d'implication, fondé sur une probabilité qu'il ne faudrait absolument pas interpréter comme indiquant que dans 99% des images prises en compte, un lien de la nature préétablie existera ! Comme cela a été défini (Partie 1 Chap. 1) :

Définition : On appelle intensité d'implication de la quasi-règle $a \Rightarrow b$, le nombre $\varphi(a,b) = 1 - \Pr ob\left[Card(X \cap \overline{Y}) \leq Card(A \cap \overline{B})\right]$ si $n_b \neq n$ et $\varphi(a,b) = 0$ si $n_b = n$

L'intensité d'implication est donc une valeur probabiliste qui fonde la décision de retenir ou non un lien du type quasi-implication entre deux variables binaires a et b. Comme il a été exposé dans la présentation théorique de l'ASI, cette modélisation est tout à fait pertinente pour mesurer l'étonnement face au constat de la petitesse du nombre des contre-exemples en regard du nombre surprenant des occurrences en faveur de l'implication. L'intensité d'implication est une mesure de la qualité inductive et informative d'un lien de type implicatif.

Rappelons qu'une variable peut prendre deux statuts dans l'ASI. Le chercheur peut lui attribuer le statut de variable principale participant directement aux calculs de l'indice d'implication. Mais il peut aussi la considérer comme une variable secondaire, supplémentaire sur laquelle il appuiera son interprétation, sans pour autant que cette variable

soit intervenue dans les calculs de l'indice d'implication. Ici 31 variables binaires ont été mises en variables supplémentaires.

Pour revenir à notre étude, dans un premier temps, nous avons pris en compte les 118=87 principales)+31(supplémentaires) variables binaires. Nous avons conduit l'analyse avec CHIC selon le paramétrage suivant :

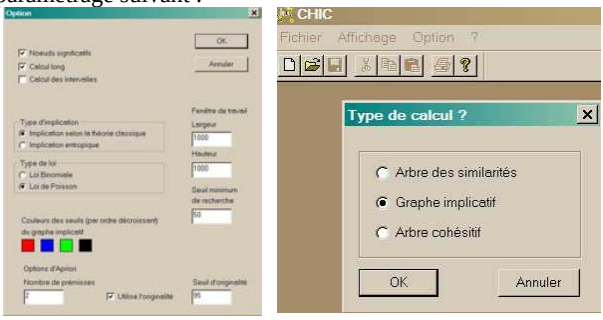


FIG. 4 – *Paramétrage du logiciel CHIC*

Le traitement conduit alors à obtenir le graphe implicatif suivant :

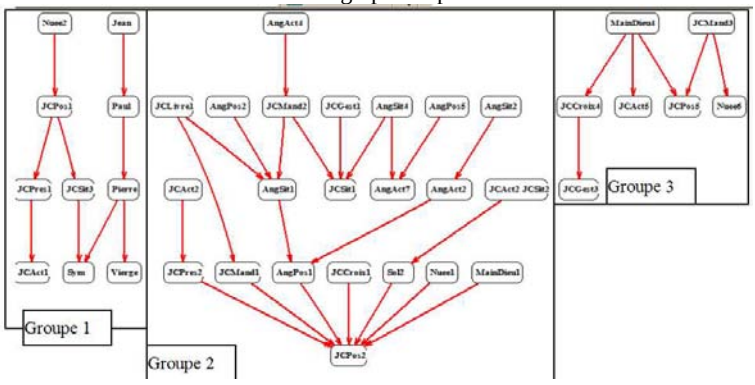


FIG. 5 – *Graphe implicatif au niveau de confiance 1- α=0,99*

La représentation graphique ainsi construite a donné un premier aperçu des liens à explorer : trois groupes se distinguent, allant de deux à cinq niveaux d'implication. Le premier groupe propose lui-même deux sous-groupes réunis par une même donnée, marquant des quasi-implications par rapport au Christ lui-même puis aux témoins. Le second groupe, qui est le plus dense, comporte aussi plusieurs sous-groupes. Les analyses se complexifient à ce degré : si l'on prend le chemin partant d'« AngPos2 » c'est à dire Ange dans la position 2 et aboutissant à « JCPos2 », c'est à dire le Christ dans la position 2,

$$\text{AngPos2} \Rightarrow \text{AngSit1} \Rightarrow \text{AngPos1} \Rightarrow \text{JCPos2}$$

on peut s'apercevoir que deux positions des anges, aux niveaux 0 et 2 du chemin sont prises en compte. La lecture du graphique se comprend à partir de la source même, en l'occurrence les images. Ainsi, si nous considérons la quasi-règle AngPos2 ⇒ AngSit1, elle indique que, lorsque les anges sont représentés dans la position 2, c'est à dire en buste, alors

très vraisemblablement ils seront aussi représentés dans la situation 1, donc près du Christ. En parcourant le chemin à partir du niveau 1, si nous considérons la quasi-règle AngSit1 ⇒ AngPos1, lorsque les anges sont représentés dans la situation 1, alors très vraisemblablement ils seront aussi représentés dans la position 1 à savoir en pieds.

Le groupe 3 pour sa part réunit toutes les variables qui traduisent un état inconnu quant à la propriété étudiée sur le subjectile. C'est cette dimension de l'inconnu du point de vue du chercheur qui fait le lien entre ces variables de type donnée manquante.

Ces trois groupes guident la réflexion quant au choix de quasi-règles à explorer en priorité. En effet, les liens représentés viennent confirmer certaines connexions qu'il semblait déjà nécessaire d'établir au sein des images selon une approche contextuelle basée sur les mœurs intellectuelles de l'époque, comme la théologie par exemple. Mais le graphe fait aussi surgir d'autres liens qu'il convient de rapprocher afin de voir s'il est possible d'en déduire des significations nouvelles vraisemblables et pertinentes. La perception de l'image et son interprétation peuvent ainsi être avantageusement guidées et affinées par ce type de graphe implicatif.

5.2 Exemple de connexions implicatives pertinentes pour l'étude du thème de l'Ascension du Christ en iconographie médiévale

Lors de la sélection d'items précis, c'est à dire des variables binaires les plus pertinentes à ce stade de l'étude, il a été jugé opportun de modifier le niveau de confiance 1- α en le réduisant à 0,95%. Cela revient à se mettre dans les conditions d'un accroissement du nombre de quasi-règles acceptables portant sur des propriétés plus ou moins rares dans les images, afin de mieux faire ressortir des quasi-implications possibles pouvant interférer sur la lecture de l'image.

Partons d'un exemple précis : la main de Dieu. La représentation de cette dernière modifie directement la typologie du Christ, aussi bien dans son action, sa présentation ou sa position. Ont donc été sélectionnés les items correspondant aux éléments mentionnés plus haut, mais en optant pour une action précise de la main de Dieu : la préemption du Christ.

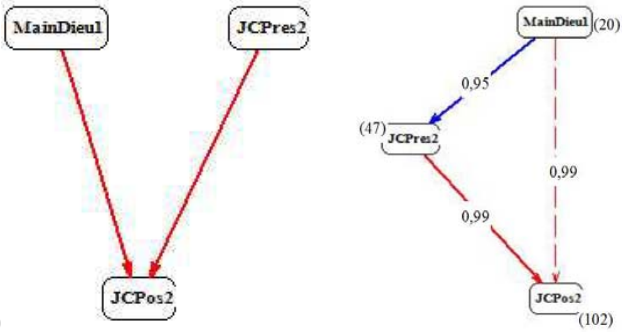
Code	Composantes binaires du vecteur-variable						
V18	= (JCPos0	JCPos1	JCPos2	JCPos3	JCPos4	JCPos5	JCPos6)
	Absent	Pieds	En pied	Assis	En buste	Inconnu	Pieds non représentés
V19	= (JCPres0	JCPres1	JCPres2	JCPres3	JCPres4	JCPres5	JCPres6)
	Absent	De face	De profil	De trois-quarts	De dos	De trois-quarts dos	Inconnu
V27	=						
	(MainDieu0	MainDieu1	MainDieu2	MainDieu3	MainDieu4	MainDieu5)	
	Absence	Tenant le Christ	Bénissant	Présentée	Inconnu	tenant une couronne végétale	

TAB. 5 – *Détails des variables V18, V19 et V27.*

Notre objectif est ici de vérifier les accointances entre la position et la présentation du Christ par rapport à la main de Dieu saisissant son fils. Au vu des images, il semblait que lorsque cette occurrence intervenait dans l'image, le Christ était systématiquement debout de

profil. Les résultats, c'est à dire les quasi-règles, les quasi-théorèmes de l'iconographie médiévale, obtenus à un niveau de confiance de 0,99 ont été avantageusement complétés par ceux qui sont apparus à un niveau de confiance de 0,95.

Reprenons nos deux étapes. En premier lieu, en fixant 1- α=0,99, les quasi-règles concernant la position du Christ (V18), la présentation du Christ (V19) et la main de Dieu (V27) sont ainsi apparues. Rappelons que ces variables déterminent les vecteurs-variables à composantes binaires respectivement de dimension 7, 7 et 5.



(1- α=0,99)

FIG. 6 – *Graphe implicatif concernant les variables V18, V19 et V27*

Le résultat montre que lorsque le Christ est en position de profil, il est la plupart du temps debout en pieds. Nous notons par ailleurs que la main de Dieu le saisit, sans qu'un lien soit fait entre la présentation de profil et la main de Dieu

Cette absence de lien entre la main de Dieu et la position de profil nous paraissant surprenant, nous avons donc décidé d'abaisser à 95% le niveau de confiance, et d'introduire le recours au marquage des fermetures transitives. Le graphique a révélé une autre possibilité de lecture.

Ainsi, en haut du graphique, la main de Dieu dans la modalité MainDieu1 apparaît. A ce niveau de confiance, un lien vraisemblable avec la présentation de profil du Christ s'établit. Nous voyons aussi apparaître une fermeture transitive liant la main de Dieu à la position debout du Christ : la main de Dieu saisissant le Christ laisserait sous-entendre qu'il y a une très forte chance pour que le Christ soit aussi debout, rejoignant le résultat du premier graphique mais en ajoutant un niveau intermédiaire d'implication entre la main et la position du Christ. Ainsi, la main de Dieu attrapant son Fils par le poignet impliquerait que le Christ est de profil, cette position même impliquant le fait que le Christ soit la plupart du temps debout, mais aussi que lorsque la main de Dieu est ainsi représentée, elle laisse entendre une figuration du Christ debout.

5.3 . Étude de cas : position, présentation et action du Christ

Nous nous intéressons maintenant aux trois variables position du Christ (V18), présentation du Christ (V19) et action du Christ (V20).

Comme nous l'avons déjà présenté dans (Tab 3), la base de données préliminaire a montré la multitude de positions, présentations et actions du Christ. En ce qui concerne la variable « action du Christ », nous la déclinons comme suit (Tab 4)

Code	Composantes binaires du vecteur-variable					
V20	=(JCAct0	JCAct1	JCAct2	JCAct3	JCAct4	JCAct5)
	Absent	Immobile	Grimpant	Volant	Montant	Image tronquée

TAB. 6 – *Détails de la variable V20.*

Désolidariser ainsi trois attitudes intimement liées les unes aux autres est le premier pas objectif vers une nouvelle approche de l'image : par cette décomposition mécanique, l'on se détache de l'expérience sensible liée à l'observation des images mais aussi à des réflexes issus du conditionnement culturel que tout un chacun effectue[6]. Le schéma implicatif intervient pour guider une analyse objective des représentations du Christ.

Si l'on choisit un niveau de confiance de 0,99 prenant aussi en compte les fermetures transitives, peu d'occurrences de lien se font jour.

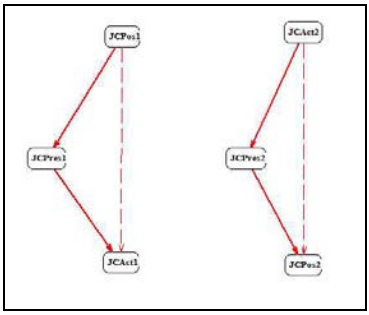


FIG. 7 – *Graphe implicatif (1- α=0,99) concernant les variables V18, V19 et V20*

Deux groupes apparaissent, à trois niveaux d'implication. A gauche, la position « 1 », ne figurant que les pieds du Christ, mène à la présentation « 1 », signifiant alors que dans cette position, il y a une forte chance que le Christ soit de face. A un deuxième niveau d'implication, il ressort que lorsque le Christ est de face, il est immobile. Notons également la fermeture transitive établie entre la position et l'action du Christ, indiquant que l'on peut être quasi certain que lorsque seuls les pieds du Christ sont représentés, ils sont immobiles.

A droite, le second groupe met en relation le Christ grimpant avec la présentation de profil du Christ et sa position debout. L'action de grimper est impliquée dans la présentation de profil du Christ, elle-même induisant à un niveau de confiance 99% le fait que le Christ soit représenté debout.

A ce stade, nous pouvons déjà faire quelques remarques. L'ensemble de gauche caractérise bien la typologie 1, correspondant surtout aux lettres enluminées. Il laisse la marge aux représentations n'entrant pas dans ce cadre mais qui apparaissent alors pour quantité négligeable. Quelques rares images ne représentant que les pieds du Christ les montrent de profil, immobiles ou en action. Le niveau de confiance permet ainsi d'évaluer la pertinence des quasi-règles que nous formulons dans le champ tout en laissant place à des exceptions.

[6] Ainsi, si l'on demande à une personne de décrire, sans image, l'Ascension du Christ, elle se basera sur son expérience sensible personnelle et décrira un personnage de profil montant au ciel, c'est-à-dire effectuant un mouvement ascendant.

En reprenant exactement les mêmes variables mais en abaissant le niveau de confiance à 95%, d'autres quasi-règles apparaissent (Fig. 8). Il est tout d'abord intéressant de constater que cela n'a aucune incidence sur les deux groupes précédents, qui restent en l'état, alors que nous aurions pensé voir de nouveaux liens s'instaurer, en vertu de la diversité de l'image médiévale.

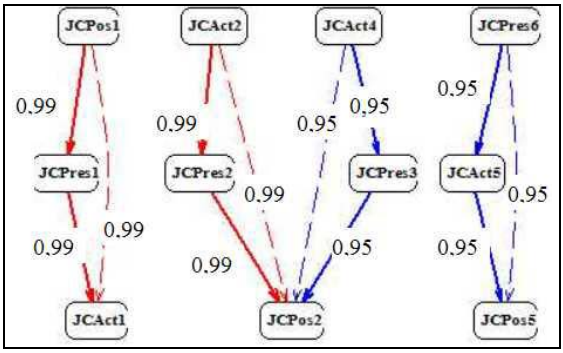


FIG. 8 – *Graphe implicatif (1- α=0,95) concernant les variables V18, V19 et V20*

A ce niveau de confiance, ce sont maintenant trois groupes qui se sont formés sur trois niveaux. Le premier est rattaché à la position debout du Christ. Partant du Christ grimpant, il connecte l'action avec la position de trois quarts du Christ puis qu'il est plutôt debout. Aucune relation n'est faite entre, par exemple, l'action de grimper du Christ et sa présentation de profil, ce qui ne laisse pas sans interrogation. En effet, le Christ apparaît plus souvent de profil que de trois quarts. Les différents graphiques laissent alors entendre que lorsque le Christ est de trois-quarts, cela induirait la plupart du temps qu'il monte et l'action est alors plus douce, alors que s'il est de profil, il grimpe et l'accent est mis sur un mouvement beaucoup plus prononcé renvoyant à des solutions iconiques qui font appel à l'expérience sensible de l'homme et qui dans le même temps, mettent en avant l'humanité du Christ. On peut supposer que l'image adopte alors un discours théologique basé sur la glorification de l'humanité à travers l'Ascension du Christ.

PrésJC1	PrésJC2	PrésJC3	PrésJC4	PrésJC5	PrésJC6
154	45	33	5	1	4

TAB. 7 – – *Distribution des effectifs de la variable « Présentation du Christ» V19*

Le dernier ensemble réunit les variables concernant les inconnues. Nous sommes surpris de voir ces liens s'établir à un taux de confiance de 95% et non 99%. En effet, concernant des œuvres dont la partie figurant le Christ est manquante, il ne peut y avoir extrapolation. Il est donc primordial de maîtriser son thème d'étude pour utiliser l'analyse statistique implicative de manière adéquate.

En tant qu'axe ordonnateur de l'image, le Christ nécessite une étude de fond. Il apparaît par le biais de ces graphiques dans toutes ses diversités qui sont autant de manières d'interpréter, pour le concepteur de l'image, l'Ascension du Christ. Le chercheur prend alors pleinement conscience de la complexité non seulement de la mise en page iconique à travers

les actions du Christ, mais aussi de la valeur iconographique du thème étudié en fonction des partis pris iconiques. Il est donc évident que la manière d'agencer iconiquement l'image a des répercussions immédiates sur la perception intellectuelle du sujet.

5.4 . Compréhension de l'utilisation de CHIC en iconographie médiévale

Une telle étude thématique demande de réunir un nombre d'images conséquent pour mener une enquête significative. La difficulté réside dans la manière de décomposer les images en signifiants et pour ainsi dire, en sous-signifiants : le signifiant principal est l'image, composée de différents signifiants, eux-mêmes subdivisés en autant de catégories. Tous sont en relation les uns avec les autres et la disposition de l'un aura des répercussions sur la représentation de l'autre, dans un ordre de réaction en chaîne.

L'analyse statistique implicative, à travers CHIC, permet de combiner différents signifiants dont le résultat permet de confirmer, affiner ou infirmer les hypothèses de départ. Elle est à considérer comme une aide estimable dans la lecture de l'image par les implications qu'elle met en évidence, tout en ne se déparant pas d'une part d'incertitude (plus ou moins élevée) que le chercheur doit prendre en compte.

Dans le cas présent, CHIC a ainsi guidé le travail : voyant que la plupart des quasi-implications étaient liées à la position 2 du Christ, c'est à dire la position debout, nous avons décidé d'axer notre recherche en partant de ce signifiant. Nous voulions par exemple définir les interactions entre la main de Dieu, la croix portée par le Christ et la position de ce dernier : le Christ porte-t-il toujours une croix lorsqu'il saisit la main de son Père ?

Pour cela nous nous sommes centrés sur les variables « position du Christ » (V18), « Main de Dieu » (V27) et « Croix du Christ » (V23).

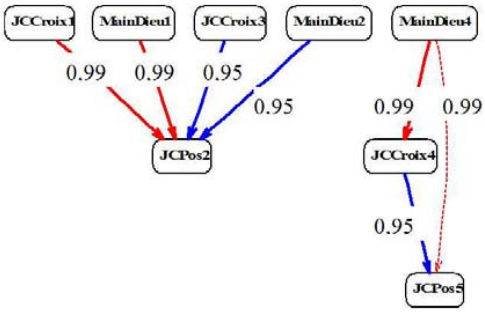


FIG. 9 – *Graphe implicatif (1- α=0,95 et 0,99) concernant les variables V18, V23 et V27*

A notre grand étonnement, aucune quasi-règle acceptable à ce niveau de confiance n'a été opérée entre la main de Dieu dans un certain type d'action et quelque sorte de croix. Il faudrait abaisser le niveau de confiance pour voir des liens s'établir.

L'analyse statistique implicative s'avère alors un outil précieux en iconographie médiévale. Tout d'abord, elle permet une approche la plus neutre possible face à l'image, support visuel soumis à la sensibilité du spectateur. Elle facilite un recul salutaire face à

l'objet iconographique traité et met ainsi une protection entre le chercheur, l'objet et l'approche sensible en obligeant dans un premier temps à décrire non pas ce que l'on veut voir mais ce qui est représenté, et dans un second temps, par les liens opérés ou non selon les requêtes demandées, à orienter de la façon la plus objective les recherches. Ensuite, elle donne à saisir la diversité de l'image médiévale. Souvent considérée comme fixe, dénuée de variété, l'image médiévale propose beaucoup plus de mobilité qu'on ne lui en accorde. Celles-ci se manifestent à travers les liens non établis entre signifiants à des taux de confiance pourtant élevés.

Cette approche apparaît donc comme primordiale dans des recherches telles que l'iconographie médiévale, car elle permet de garder un œil objectif face à l'objet.

6 Conclusion

Le recours à l'ASI en iconographie médiévale s'avère encourageant. En effet, s'il est courant, dans une étude telle que celle envisagée ici, d'utiliser des statistiques et de les traduire en tableaux et diagrammes pour guider l'exploration et la formulation de conjectures, l'utilisation de l'ASI permet d'obtenir des résultats pertinents dans la lecture interne de l'image en liant, par des règles de quasi-implication, ses signifiants, tout en laissant le champ ouvert à d'autres connexions. Ainsi, le chercheur voit se faire jour de nouveaux cheminements dans les images et peut affiner ses interprétations.

Allier l'ASI à l'iconographie médiévale était une gageure comme toute première collaboration entre deux sciences que tout semble opposer. Cette démarche a démontré une fois de plus l'intérêt de l'ouverture pluridisciplinaire d'un champ de recherche à un autre, tout en laissant à chacun sa spécificité. Par le futur, nous pouvons envisager d'appliquer de telles méthodes de travail dans d'autres champs de recherche de l'histoire de l'art et de l'archéologie, et donner ainsi naissance à une nouvelle manière d'aborder ces disciplines.

Références

Baschet, J. (1996). Inventivité et sérialité des images médiévales. Pour une approche élargie. *Annales*, 51 : 93-133.

Baschet, J. (2000). *Le Sein du Père. Abraham et la paternité dans l'Occident médiéval*, Paris : Gallimard.

Christe, Y. (1969). *Les grands portails romans. Etude sur l'iconologie des théophanies romanes*, Genève: Droz, 66-96.

Gutberlet, H. (1935). *Die Himmelfahrt Christi in der bildenden Kunst. Von den Anfängen bis ins hohe Mittelalter*, Leipzig - Straßburg- Zürich: Heitz & Co.

Mâle, E. (1986). *L'Art religieux du XIIIe siècle en France*, Paris : Armand Colin.

Schrade, H. (1930). Zur Ikonographie des Himmelfahrt Christi. *Bibliothek Warburg Vorträge (1928-1929)* : 60-190.

Summary

If the use of statistics in the history of art, especially in medieval iconography, is gained in many years, the use of statistical implicative analysis is a step in the approach to thematic issues. Statistics are an aid to research: they demonstrate the permanence or originality, but it does not participate in a debate on the subject. The SIA provides a fresh reading on the subject. It allows connecting items selected in advance by the researcher, reflecting other internal links in the image and provides a new way in understanding the thematic study in history of art.

INDEX

Biographie des Editeurs

Régis Gras est Professeur Emérite à l'Ecole Polytechnique de l'Université de Nantes, et membre de l'équipe COnnaissances et Décision (COD) du Laboratoire Informatique de Nantes Atlantique (LINA-UMR CNRS 6241) depuis 1998. Docteur ès-sciences mathématiques à l'Université de Rennes 1 en 1979, il a été président de la commission internationale d'enseignement des mathématiques (1995-1998), puis membre du comité sur l'enseignement de la société européenne de mathématiques (1997-2003). Il a effectué de nombreuses missions de formation en Afrique, Amérique du Sud et Moyen Orient. Il est le fondateur de l'ensemble des méthodes et des outils de l'Analyse Statistique Implicative, et continue d'en développer les extensions dans le cadre de la fouille de données. Il a présidé le comité de programme des 4 dernières éditions de la conférence ASI (France 2001, Sao Paulo Brésil 2003, Palerme Italie 2005 et Castellon Espagne 2007). Il est l'un des membres fondateurs de l'association « Extraction et Gestion des Connaissances ». Il est l'auteur de 5 ouvrages, 6 films pédagogiques et coéditeur de 4 livres de chapitres

Jean-Claude Régnier est Professeur des Universités à l'université Lyon 2, UMR CNRS 5191 ICAR (Interactions, Corpus, Apprentissages, Représentations) et directeur de thèse à l'ED 485 EPIC (Éducation, Psychologie, Information et Communication) de l'Université de Lyon. Il a obtenu, en 1983, un doctorat en mathématiques de l'ULP L. et, en 2000, une HDR à l'Université de Strasbourg. Il est membre de l'IASE (International Association Statistical Éducation), de l'ISI (Institut International de la Statistique), de la SFDS (Société Française de Statistique) où il préside depuis 2003 le groupe «Enseignement de la statistique ». Il a été membre des Comités de Programme des Congrès EGC, des comités scientifiques des colloques ASI3 (2005) et ASI4 (2007), SFDS (2006) et co-organisateur du 1er Colloque International Francophone sur l'Enseignement de la Statistique à l'Université de Lyon. Actuellement, dans le cadre des relations entre l'Université de Lyon et des universités brésiliennes, il est coordinateur scientifique du programme (2008-2011) ARCUS Rhône Alpes Brésil pour l'axe SHS, et responsable du Collège doctoral franco-brésilien à Lyon2.

Fabrice Guillet est Maître de Conférences, Habilité à diriger des recherches, au département informatique de l'Ecole Polytechnique de l'Université de Nantes, et membre de l'équipe COnnaissances et Décision (COD) du Laboratoire Informatique de Nantes Atlantique (LINA -UMR CNRS 6241) depuis 1997. Il a obtenu son doctorat en informatique de l'université de Rennes en 1995 à l'Ecole Nationale Supérieure des Télécommunications de Bretagne. Il est membre fondateur de l'association scientifique "Extraction et Gestion des Connaissances" (EGC, http://www.polytech.univ-nantes.fr/associationEGC) et participe au comité de pilotage de la conférence EGC depuis 2001. Ses domaines de recherche concernent la fouille de données et l'ingénierie des connaissances. Il a récemment co-édité deux livres de chapitres internationaux avec comité de lecture publiés chez Springer en 2007 puis 2008.

Vous pouvez faire part de vos remarques,
critiques, suggestions
aux auteurs à cette adresse :

auteurs@cepadues.com

Imprimé en France par Messages SAS
111, rue Nicolas-Vauquelin
31100 Toulouse