REMERCIEMENTS

Nous tenons à remercier les différents organismes qui nous ont apporté un soutien pour la tenue de ces Journées sur les plans **scientifique, financier et logistique** :

- ISET – Institut Supérieur des Études Technologiques de Radès
- ARSA – Association de Recherche en Statistique Appliquée
- Université Lumière Lyon2 (Lyon France)
- Laboratoire de recherche : Unité Mixte de Recherche UMR 5191 – ICAR (Lyon France)
- Association pour la Recherche en Didactique des Mathématiques (ARDM France),

ou leur parrainage scientifique *:*

- Association Extraction et Gestion des Connaissances (EGC) (France),
- École Polytechnique de l'Université de Nantes (France),
- Laboratoire d'Informatique de Nantes (LINA France),
- Société française de Statistique (SFdS France),
- Société Francophone de Classification (SFC France),
- G.R.I.M. (Gruppo di Ricerca sull'Insegnamento delle Matematiche), Dipartimento di Matematica, Università di Palermo (Italie)
- Revue des Sciences de l'éducation (Montréal Québec Canada)

Nous tenons à remercier tous les membres du comité d'organisation dont l'implication logistique a rendu possible la réalisation de ce 8ème colloque sur l'Analyse Statistique Implicative.

Président du comité scientifique : Jean-Claude Régnier

Vice-Président du comité scientifique : Yahya Slimani

Président d'honneur : Régis Gras

Président du comité d'organisation: Ahmed Dhouibi

Vice-Président du comité d'organisation: Makram Ben Jeddou

INTRODUCTION DE L'OUVRAGE

L'ANALYSE STATISTIQUE IMPLICATIVE : DES SCIENCES DURES AUX SCIENCES HUMAINES ET SOCIALES

Jean-Claude REGNIER[1], Régis GRAS[2]

Le huitième colloque A.S.I. 8 s'est déroulé dans un contexte d'espérance stimulée pour la démocratie tunisienne avec l'attribution du Prix Nobel de la Paix pour 2015. Comme on a pu le lire dans les journaux :

« Le comité Nobel norvégien a décidé de récompenser, vendredi 9 octobre 2015, le quartet menant le dialogue national en Tunisie, qui s'est distingué pour « *sa contribution décisive dans la construction d'une démocratie pluraliste en Tunisie après la "révolution du jasmin" de 2011* » ». (Le Monde[3])

Un tel contexte est de toute évidence plus propice à la créativité scientifique, à l'exercice de la pensée critique qui fonde les dimensions épistémologique et méthodologique des champs scientifiques. Évidemment, nous plaçons le champ de l'Analyse Statistique Implicative dans ces perspectives scientifiques.

A.S.I. – Analyse statistique implicative : une fois encore et encore, de quoi s'agit-il ?

Au risque de nous répéter, nous définissons l'Analyse Statistique Implicative comme un cadre théorique d'analyse de données fondée sur une relation non symétrique. Il s'agit d'« … 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. »[4] (Gras, Régnier, 2009 p.12).

Le présent ouvrage est constitué des articles issus d'un appel à contributions lancé dans le cadre du huitième colloque sur l'Analyse Statistique Implicative – ASI8 - organisé à Radès en novembre 2015. Ces articles ont été soumis à la lecture critique[5] des membres du comité international scientifique qui en ont assuré la qualité scientifique. Nous rapportons ici la liste en leur adressant nos remerciements les plus chaleureux pour leur travail.

- Nadja **Acioly-Régnier**, EAM 4128 SIS - Université Lyon 1 (France)

[1] Président du comité scientifique de ASI8 - UMR 5191 ICAR, Université de Lyon - Lyon 2 (France)

[2] Président d'honneur du comité scientifique de ASI8 - LINA, Université de Nantes (France)

[3] http://www.lemonde.fr/prix-nobel/article/2015/10/09/le-prix-nobel-de-la-paix-attribue-au-dialogue-national-tunisien_4786205_1772031.html

[4] Gras R., Régnier J.-C., Guillet F. (Eds) (2009) *Analyse Statistique Implicative. Une méthode d'analyse de données pour la recherche de causalités*. RNTI-E-16 Toulouse Cépaduès Editions

[5] Chaque contribution a été anonymement soumise à trois relecteurs dont les expertises ont été retournées anonymement aux auteurs.

- Saddo **Ag Almouloud**, PPG Educação Matematica - Pontifícia universidade Católica de São Paulo (Brésil)
- Marc **Bailleul**, EA 965 CERSE, Université de Caen (France)
- Makram **Ben Jeddou**, ISET de Radès, Association ARSA et Université virtuelle UVT (Tunisie)
- Sadok **Ben Yahia**, Département des Sciences de l'Informatique, Faculté des Sciences de Tunis (Tunisie)
- Younes **Boujelbene** Faculté des Sciences économiques et de gestion de Sfax (Tunisie)
- Guy **Brousseau**, DAEST - Université Bordeaux 2 (France)
- Raphaël **Couturier**, FEMTO-ST département DISC Université de Franche-Comté (France)
- Benedetto **Di Paola**, Département de mathématiques - Université di Palermo (Italie)
- Iliada **Elia**, Département de Sciences de l'Éducation - University of Cyprus (Chypre)
- Athanasios **Gagatsis**, Département de Sciences de l'Éducation - University of Cyprus (Chypre)
- Pablo **Gregori**, IMAC - Universitat Jaume I de Castellón de la Plana (Espagne)
- Fabrice **Guillet**, LINA CNRS 6241, Equipe COD - École Polytechnique de l'Université de Nantes (France)
- Pascale **Kuntz**, LINA CNRS 6241, Equipe COD - École Polytechnique de l'Université de Nantes (France)
- Dominique **Lahanier-Reuter**, Equipe Théodile CIREL - Université Lille 3 (France)
- Chiraz **Latiri**, Université de la Manouba (Tunisie)
- Philippe **Lenca**, DECIDE CID UMR 6285 Lab-STICC - Telecom Bretagne - (France)
- Dhafer **Malouche**, ESSAI - Ecole Supérieure de la Statistique et de l'Analyse de l'Information - Tunis - (Tunisie)
- Paraskevi **Michael-Chrysanthou**, Department of Education University of Cyprus (Chypre)
- Abdelmajid **Naceur** Université virtuelle de Tunis, (Tunisie)
- Maria del Pilar **Orus Baguena**, IMAC - Universitat Jaume I de Castellón de la Plana (Espagne)
- François **Pluvinage**, Depto. Matemática Educativa, CINVESTAV-IPN México and France
- Gilbert **Ritschard**, Institut d'études démographiques et des parcours de vie - Pôle de recherche national LIVES - Université de Genève (Suisse)
- Gilbert **Saporta**, Laboratoire CEDRIC - CNAM, Paris (France)
- Gérard **Vergnaud**, CNRS, Université Paris VIII (France)
- Miguel R.**Wilhelmi**, Departamento de matemáticas, Universidad Pública de Navarra (Espagne)
- Djamel Abdelkader **Zighed**, Laboratoire ERIC - Université Lyon2 (France)

A.S.I. – Analyse statistique implicative : pour mémoire, quelques-uns de ses lieux de débats et de construction.

Pour mémoire, le premier colloque ASI1 qui s'était tenu à l'Institut de Formation des Maîtres de Caen (France) les 23-24 Juin 2000 à Caen et avait été organisé par Marc Bailleul et Régis Gras, avait une thématique générale centrée sur « La fouille dans les données par la Méthode Statistique Implicative ». Ce choix correspondait à une orientation prometteuse quand on considère le développement et les activités de l'association EGC (**E**xtraction et **G**estion des **C**onnaissances). Rappelons que la « fouille dans des données » (encore appelée **K**nowledge **D**iscovery in **D**atabases ou encore "Data Mining" dans la littérature anglo-saxonne) part, en général, du croisement de sujets (ou objets) et de variables (propriétés ou attributs) binaires, ordinales ou numériques. Son objectif majeur consiste à conjecturer des modèles basés sur des relations quantitatives ou qualitatives et des structures induites à partir des données. Différentes méthodes, comme l'Analyse Factorielle des Correspondances (A.F.C.), la Classification Ascendante Hiérarchique (C.A.H.), sont communément utilisées pour de telles fouilles dans des données. Parmi elles, l'Analyse Statistique Implicative (A.S.I.) vise l'extraction de connaissances, d'invariants, de règles inductives non symétriques consistantes, et accorde une mesure à des propositions du type « quand a est choisi, on a tendance à choisir b ».

Le deuxième colloque ASI2 a été organisé à l'Université PUC de São Paulo (Brésil) les 9-11 Juillet 2003, par Saddo Ag Almouloud.sur la thématique globale « O metodo estatístico implicativo utilizado em estudos qualitativos de Régis Gras de associação. Contribuição à pesquisa em Educação »

Le troisième colloque ASI3 a été organisé par Filippo Spagnolo de l'Université de Palerme à Palerme (Italie) les 6-8 octobre 2005.

Le quatrième colloque ASI4 s'est tenu à l'Université Jaume I de Castellón de la Plana (Espagne) du 18 au 21 octobre 2007 et fut organisé par Pilar Orus et Pablo Gregori.

Le cinquième colloque ASI5 s'était tenu à Palerme (Italie) les 5-7 novembre 2010, et a été l'occasion d'entendre près de 30 communications qui avaient été retenues par le comité scientifique parmi celles qui avaient été soumises, et ainsi de poursuivre le débat. Il est possible d'accéder aux publications :

Sur la site ASI 5 - http://sites.univ-lyon2.fr/asi5/?page=15

Sur le site de la Revue QRDM - "QUADERNI DI RICERCA IN DIDATTICA" - G.R.I.M. http://math.unipa.it/~grim/QRDM_20_Suppl_1.htm

Le sixième colloque ASI6 s'était tenu à Caen (France) les 7-10 novembre 2012, et a été un fois de plus le lieu de riches échanges, l'occasion d'entendre une conférence de Guy Brousseau, et d'émettre une pensée chaleureuse à la mémoire de Filippo Spagnolo, président du comité d'organisation de ASI 5 qui nous a quittés peu de temps après, au début de 2011. Il est possible d'accéder aux publications :

Sur la site ASI 6 - http://sites.univ-lyon2.fr/asi6/?page=15

Sur le site de la Revue QRDM - "QUADERNI DI RICERCA IN DIDATTICA" - G.R.I.M. http://math.unipa.it/~grim/QRDM_22_Suppl_2.htm

Le septième colloque ASI7 s'était tenu à São Paulo (Brésil) les 27-30 novembre 2013. Il est possible d'accéder aux publications :

Sur la site ASI 7 - http://sites.univ-lyon2.fr/asi7/?page=16

Rappelons qu'en 2009 et 2013, deux ouvrages majeurs francophones ont été publiés chez Cépaduès à Toulouse : Analyse Statistique Implicative. Méthode exploratoire et confirmatoire pour la recherche de causalités[6] et Analyse Statistique Implicative. Une méthode d'analyse de données pour la recherche de causalités[7].

Ils permettent de faire un point global sur l'actualisation significative de l'ouvrage de 1996, *L'implication statistique*, édité par La Pensée Sauvage Editeur (Grenoble France). Une autre publication majeure anglophone a été réalisée avec l'ouvrage[8] de 2008, *Statistical Implicative Analysis* diffusé par Springer-Verlag, (Berlin-Heidelberg, Allemagne). Enfin une troisième en espagnol a aussi été produite avec l'ouvrage[9] de 2009 *Teoria y Aplicaciones del Analisis Estadistico Implicativo*, diffusé par Universitat Jaume-1, (Castellon Espagne).

Dans ce huitième colloque A.S.I.8 de 2015 à Radès, les dimensions internationale et interculturelle, initiée dès ASI5 par l'introduction de cinq idiomes de travail : anglais, espagnol, français, italien et portugais, a été maintenue, certes non sans difficultés et obstacles à dépasser, mais parce que cela nous semble fondamental pour enrichir nos champs de recherche.

Les thèmes privilégiés furent :

- Concepts fondamentaux en ASI : modèles statistiques, types de variables, variables principales et supplémentaires
- Avancées nouvelles en cours, stabilité des indices, intensité d'implication entropique ; extension à de nouveaux types de variables, espace des sujets continu ; règles d'exception; dualité espace des sujets - espace des règles, structure métrique et topologie de l'espace des sujets induites par leur contribution ou leur typicalité, analyse vectorielle ;
- Comparaison critique des démarches, des modèles, des représentations et des résultats de l'ASI avec ceux d`autres méthodes d'analyse de données (treillis de Galois, réseaux bayesiens, arbres d'induction, analyses factorielles, etc.)
- Pratique du logiciel CHIC, les développements actuels et attendus
- Applications traitées par l'ASI et comparativement avec d'autres méthodes, dans les domaines de la didactique, des sciences de l'éducation, de la psychologie, de la sociologie, de l'économie, de l'histoire de l'art, de la biologie, de la médecine, de l'archéologie, etc.

[6] GRAS, R., REGNIER, JC., MARINICA, C., GUILLET, F. (Eds) (2013**)** Analyse Statistique Implicative. Méthode exploratoire et confirmatoire pour la recherche de causalités. Toulouse : Cépadues

[7] GRAS R., RÉGNIER J.-C., GUILLET F. (Eds) (2009) Analyse Statistique Implicative. Une méthode d'analyse de données pour la recherche de causalités. RNTI-E-16 Toulouse : Cépaduès

[8] Gras R., Suzuki E., Guillet F.., Spagnolo F. (Eds) (2008) *Statistical Implicative Analysis*, Berlin-Heidelberg : Springer-Verlag, ISBN : 978-3-540-78982-6

[9] Orus P., .Zemora L., Gregori P. (Eds) (2009) *Teoria y Aplicaciones del Analisis Estadistico Implicativo*, Castellon : Universitat Jaume-1,, ISBN : 978-84-692-3925-4

- Présentations graphiques et numériques des résultats applicatifs, aides à l'interprétation de ces résultats, rôles respectifs et critiques des types de variables, des variables principales et supplémentaires choisies
- Spécificités de la formation à l'ASI : usage du logiciel CHIC, interprétation des représentations graphiques (graphe implicatif ; arbre de la hiérarchie cohésitive)
- Problématiques de didactique de l'ASI

Mais ce colloque était aussi le lieu de confrontation des débats autour des défis qui ont été lancés à l'issue de A.S.I. 7

Défis 2015

- Défi 1 : A chaque couple de variables de VxV, associer le vecteur q, indice centré-réduit d'implication. Étude du champ de gradient dq
- Défi 2 : Variables continues de loi uniforme ou quelconque. Exemples attendus.
- Défi 3 : Notion de variable " attracteur ", sommet du cône implicatif. Exploitation sémantique. Exemples attendus.
- Défi 4 : Relation entre l'intensité d'implication et la confiance (probabilité conditionnelle). Exemples attendus.
- Défi 5 : Exploiter la métaphore ensembliste pour définir une intensité d'inclusion prenant en compte l'implication directe et la contraposée
- Défi 6 : Dans le cadre de la conférence Genèse et développement de l'Analyse Statistique Implicative ; rétrospective intuitive parue dans les actes de ASI 7, une étude de l'organisation conceptuelle hiérarchisée de la cognition (par ex. schèmes, procédures, conception, méthode, ..) mériterait d'être faite par le psychologue ou le didacticien s'appuyant sur les classes plus ou moins emboîtées d'une hiérarchie cohésitive. Exemples attendus.

Délibérément le programme de ce colloque A.S.I. 8 s'est organisé, dans la continuité, autour d'une alternance équilibrée entre des communications portant sur des approches théoriques ou sur des mises en application, et des travaux pratiques sur le logiciel CHIC et sa nouvelle version RCHIC.

A.S.I. – Analyse statistique implicative : des sciences exactes aux sciences sociales et humaines.

Le présent ouvrage donne à voir une partie de l'activité de production scientifique suscitée par l'appel à communication du 8ème colloque A.S.I.8. Les contributions constituent les chapitres successifs de ce livre. Mais auparavant nous présentons les textes des conférences réalisées au cours de ce colloque.

La première donnée par Régis GRAS aborde un retour sur le sens de la construction du cadre théorique : ***Un survol paradigmatique de l'analyse statistique implicative***

La seconde, par Yahya SLIMANI, s'intéresse à la question des rapports entre Data Mining, Big Data et A.S.I. : ***Big Data et Datamining: entre mythes et réalités***

La troisième, par François PLUVINAGE s'intéresse à la question: Regards sur les rapports entre statistique et enseignement mathématique

Contributions retenues et publiées par le comité scientifique

A – Développement du cadre théorique A.S.I.

Notion de champ implicatif en analyse statistique implicative
Régis Gras, Pascale Kuntz, Nicolas Greffard

Un mariage arrangé entre l'implication et la confiance ?
Régis Gras, Raphaël Couturier, Pablo Gregori

Analyse d'un questionnaire « enseignants de mathématiques » par différentes méthodes
Régis Gras et Antoine Bodin

Variable nodale et cône implicatif
Dominique Lahanier-Reuter, Régis Gras, Marc Bailleul

Hiérarchie de règles en A.S.I. et conceptualisation
Régis Gras, Nadja Maria Acioly-Régnier

Ajout de la confiance au graphe implicatif
Souhila Ghanem et Raphaël Couturier

Extension de l'analyse statistique implicative au cas des variables continues quelconques
Régis Gras, Jean-Claude Régnier

Classifying objective interestingness measures based on the tendency of value variation
Nghia Quoc Phan, Hiep Xuan Huynh, Fabrice Guillet, Régis Gras

Medidas alternativas de los grados de adhesión de individuos a la implicación y similaridad para variables modales
Larisa Zamora Y Pablo Gregori

Variability of GRAS classical implication index and other rule quality measures under small size sampling from bivariate binary processes
Pablo Gregori, Raphaël Couturier

Pourquoi et comment transformer des variables quantitatives en qualitatives ? Application à la mélodie de la langue française
Martine Cadot, Anne Bonneau

Apport de la combinaison de la méthode d'analyse statistique implicative (ASI) avec la théorie de réponses aux items (IRT)
Hayette Khaled, Raphaël Couturier

B – Application aux recherches en didactiques disciplinaire et professionnelle

La coordinación de los procesos de aproximación en la comprensión del límite de una función en un punto. Una aproximación a través del análisis implicativo
Joan Pons, Julia Valls, Salvador Llinares

Les supports d'enseignement dans la représentation du métier chez des professeurs des écoles débutants.
Marc Bailleul, Laurence Leroyer

Uma análise das diferentes praxeologias no ensino das equações de segundo grau. um olhar a partir da quadro teorico da A.S.I
Marcus Bessa De Menezes, Marcelo Câmara Dos Santos

Professores que ensinam matemática nos anos iniciais do ensino fundamental: Análise das Tendências de Pesquisa no Brasil (2006-2014)
Luciana Silva Dos Santos, Marcelo Câmara Dos Santos,
Nadja Maria Acioly-Régnier, Edênia Maria Ribeiro Do Amaral

L'approche socioconstructiviste dans les situations d'enseignement-apprentissage de la biologie en Tunisie : le cas de la reproduction humaine
Fadhila Farhane Horrigue

Étude des significations données à la notion de fraction par des élèves de CM1 et de CM2 de l'École primaire en France
Abdul Aziz Alahmadati

Análise das pesquisas didáticas sobre função afim no Ensino Fundamental e Médio no quadro da análise estatística implicativa
Aveilson José de Santana, Vladimir Lira Veras Xavier De Andrade,
Jean-Claude Régnier

Représentations et valeurs implicites des enseignants gabonais sur l'éducation à la santé
Laurence Ndong

L'enseignement de la résolution de problèmes mathématiques à l'école primaire au Maroc : représentations des enseignants à l'égard de leurs pratiques

UN SURVOL PARADIGMATIQUE DE L'ANALYSE STATISTIQUE IMPLICATIVE

Régis GRAS [1]

A PARADIGMATIC SURVEY OF THE IMPLICATIVE STATISTICAL ANALYSIS

RESUME.

Cette conférence présente tout d'abord, l'origine de la situation fondamentale dans laquelle la nécessité d'organiser des comportements de réponse d'élèves à un test de didactique des mathématiques est apparue, en respectant la complexité a priori d'exercices. Cela a conduit à la création d'un indice d'implication entre items de réponses, pour évaluer des règles comme : *« si a alors généralement b »*. puis à des représentations du préordre partiel obtenu entre les réponses. La théorie, dénommée alors Analyse Statistique Implicative, s'est ensuite développée, sous la poussée des applications variées rencontrées, par extension de la nature des variables de comportement à des variables non binaires. Enfin, une relation topologique duale a été établie entre les sujets et les variables. Ce développement est de type paradigmatique car il associe un ensemble de problèmes à étudier et de techniques propres à leur étude..

Mots-clés *: taxonomie, rapport non-linéaire tout/partie, dialectique, système dynamique, stabilité structurelle, propriété émergente, règle, métarègle, intensité d'implication, graphe implicatif, hiérarchie cohésitive, variables binaire, modale, numérique, intervalle, floue, vectorielle, supplémentaire, structure topologique duale*

1 Introduction

L'A.S.I. est le dernier plat sorti de la marmite dans laquelle j'ai, au cours de mes 62 ans de service dans l'Éducation Nationale (de la maternelle à l'université), ajouté, mélangé un nombre important d'ingrédients en un ensemble jubilatoire et passionné.

Je me propose, avec une telle composition de la marmite, de vous raconter l'histoire de l'Analyse Statistique Implicative, sa raison d'être, son fondement épistémologique, ses développements, sans rentrer dans les détails techniques, les formules mathématiques, bref en ne faisant appel qu'à votre bon sens et votre intuition[2].

2 Problématique d'ordre psycho-didactique

Au cours des années 70, dans le cadre des Instituts de Recherche sur l'Enseignement des Mathématiques de France, j'ai fréquenté une nouvelle fois des classes du

[1] Ecole Polytechnique de l'Université de Nantes, Equipe DUKE, Laboratoire d'Informatique de Nantes-Atlantique (LINA), UMR 6241, E-mail : regisgra@club-internet.fr,

Site : http://math.unipa.it/~grim/homegras_03.htm

[2] « … il n'y a de connaissance vraie que par l'intuition, c'est-à-dire par un acte singulier de l'intelligence pure et attentive, et par la déduction, qui lie entre elles les évidences », M.Foucault, « Les mots et les choses », p. 103

secondaire, particulièrement du 1er cycle (11 à 15 ans) où j'y ai conduit et évalué une expérience nationale tout en participant à la formation continue et aux recherches des enseignants de ces classes. J'ai donc été le témoin des difficultés d'apprentissage des élèves et quelquefois l'acteur des tentatives de remédiation aux obstacles qui s'opposent à l'assimilation des notions enseignées. Ces obstacles n'étaient pas toujours rencontrés par les élèves mais certains étaient récurrents et relativement partagés. Leur nature relevait de la didactique mais aussi souvent de l'épistémologie en s'opposant à l'acquisition de leurs connaissances[3]. Je pouvais observer ces problèmes d'apprentissage aussi bien directement par l'attitude ou l'expression orale des élèves mais également à travers des questionnaires ou des travaux écrits. Faisant l'hypothèse que les attitudes ou les comportements de réponse étaient globalement identifiables, je disposais de données constituées de traces laissées par les élèves. A l'occasion de résolution d'exercices de mathématiques ou de problèmes, une certaine hiérarchie de difficultés segmentait l'ensemble des élèves interrogés. Plus la difficulté s'accroissait, plus le nombre de réussites diminuait, ce qui peut sembler une tautologie : on peut en effet s'attendre à ce que tout élève qui réussit une épreuve jugée difficile, dans un contexte qui serait comparable, réussirait a fortiori ce qui était facile. Ce qui pourrait **étonner et le contester**, ce sont les incohérences par rapport à cet attendu. L'idée que la difficulté a priori soit définissable objectivement par ma propre pratique ne tenait plus en tant que prédicteur, même si elle était le plus souvent respectée. Ainsi, c'est la relation stable et relativement prévisible entre réussites et échecs, entre comportements de réponse qui m'intéressait plutôt que la réussite ou l'échec à un item donné. Ce qui rejoint l'opinion de H. Poincaré qui dit « *que les mathématiciens n'étudient pas les objets mais les relations entre les objets* ».

D'où l'idée, afin d'aider les enseignants dans l'évaluation d'un niveau d'acquisition d'un concept mathématique donné et dans un projet de construction d'épreuves calibrées, de modéliser des niveaux d'acquisition, en une **taxonomie d'objectifs cognitifs** Celle-ci, à l'instar de celle de Bloom, la plus connue, visait à organiser a priori, selon un ordre de complexité croissante et à travers une analyse des tâches, la maîtrise ou l'appropriation d'un concept (et non pas les moments de son apprentissage). Par exemple, un objectif s'exprimant en termes d'utilisation d'un algorithme serait considéré de complexité inférieure à celle d'un objectif exigeant la construction d'un contre-exemple. Une relation de type causal sous-tendrait cette hypothèse : les outils cognitifs d'un objectif supérieur seraient suffisants à ceux que mobilise l'élève pour un objectif de niveau inférieur, comme une « conséquence » ou un « effet » serait le fruit d'une « cause ». Dit autrement : « résoudre un exercice complexe » impliquerait « résoudre un exercice moins complexe » et sa réussite en serait un **bon prédicteur**.

Relativement à un questionnaire constitué d'items spécifiant chacun des objectifs cognitifs de la taxonomie, en théorie on aurait pu attendre des réussites organisées linéairement selon la complexité a priori. Ce qui n'a pas été observé. A l'ordre total présumé s'est substitué un **préordre partiel**, comme sont définis les stades différentiels de développement de l'enfant chez Piaget. Nous abordons cette question avec plus

[3] « *C'est en termes d'obstacles qu'il faut poser le problème de la connaissance scientifique* » écrit G. Bachelard dans « La formation de l'esprit scientifique »

d'attention dans la communication ASI 8 : « Hiérarchie de règles et conceptualisation ». L'observation de ce préordre partiel signifie que des élèves pouvaient dans certains cas et pour quelques-uns d'entre eux, réussir à un item *a* jugé difficile tout en échouant à un item *b* jugé plus facile. Et ceci sans remettre en cause l'affirmation que «généralement la réussite à *a* s'accompagne de la réussite à *b* » et sans que sa réciproque ne soit nécessairement vraie. Mon intérêt va alors porter sur ce type de relation non symétrique, tenter de pondérer la qualité de son caractère approximatif et d'organiser si possible l'ensemble des couples de variables-items en jeu de ce préordre partiel.

Quels **outils statistiques** étaient alors à ma disposition pour qualifier et quantifier cette relation non symétrique entre deux variables ? Un **test paramétrique** non symétrique ? mais pour réfuter quelle hypothèse ? que les élèves qui ont répondu à *a* ont aussi répondu à b ? que faire de la réfutation ? la ranger sagement ? Établir une liste de cas réfutés ou acceptés ? non, pas de structure globale attendue d'une telle liste ; utiliser la mesure de liaison entre deux variables sur la base de leur **corrélation** ? Mais cette mesure quantifie la qualité des co-occurrences et est donc symétrique. Utiliser **une méthode multidimensionnelle d'analyse de donnée**s afin d'organiser les relations en un tout ? J'avais alors entretenu des collaborations avec J.P. Benzecri sur **l'A.F.C.** et I.C. Lerman sur la **classification hiérarchique** et, très souvent, enseigné et utilisé leurs méthodes d'analyse. Mais leurs fondements théoriques sont essentiellement **symétriques**. Ma réserve était la même que pour la corrélation. D'ailleurs, soulignant bien la différence fondamentale de points de vue, la métaphore ensembliste suivante éclairait la différence et facilitait la compréhension intuitive de la problématique de quasi-implication : parmi la population de sujets concernés par l'étude, le sous-ensemble A des sujets qui satisfont *a* est presque contenu dans le sous-ensemble B des sujets qui satisfont *b* . Restait la **théorie Bayesienne** qui offre le moyen de calculer ce que l'on appelle la « probabilités des causes ». Elle est d'une grande efficacité mais - je l'ai souligné dans un article et Martine Cadot en a étudié la comparaison avec l'ASI – sans nier cette efficacité, elle me semble présenter moins de sensibilité aux effectifs des échantillons (gênant en statistique) et écrase quelque peu les cas rares. Circonstances qu'évitera l'ASI et qu'Yves Kodratoff a exprimées par la recherche « *des pépites de connaissances* ».

Point de départ épistémologique, il me fallait donc établir un **indice**, compris entre 0 et 1 par exemple, capable de rendre compte de l'écart entre **prédiction** et **contingence**, c'est-à-dire entre ce qui était attendu de l'ordre a priori [4](A est inclus dans B dans la métaphore ensembliste) et ce qui était effectivement observé, à savoir une **règle de quasi-implication** « **si *a* alors généralement *b*** ». La stratégie que j'ai alors utilisée, en 1978, a consisté à prendre plutôt en considération la non-satisfaction de l'implication « **si *a* alors *b*** » qui, comme on le sait, apparaît dès lors que *a* étant vrai, *b* est faux. Ce sont donc les **contre-exemples** sur lesquels va porter mon attention. Comme je lui souhaitais une vertu inductive, je devais prendre en compte les effectifs des populations concernées : effectif total des élèves, nombre de ceux qui satisfont *a* et ne satisfont pas *b*. De plus, sans information sur la population, ni sur l'existence d'une relation entre les variables étudiées, j'ai fait l'hypothèse d'une **absence de liaison a**

[4] « Contrairement à l'opinion commune, la grande affaire de la science est moins la production de vérités absolues et universelles ou la reconnaissance d'erreurs rédhibitoires, que la délimitation des conditions de validité d'énoncés… » (J.-M. Lévy-Leblond, « Aux contraires », p.35)

priori entre elles, comme le faisait I.-C. Lerman, comme il est fréquent dans des tests non paramétriques et comme l'exprime H. Atlan[5]. L'objectif est, si possible, de la conserver et la qualifier en mettant en évidence la faible probabilité de la dérogation à la règle sur des bases statistiques. Belle illustration de l'adage : « **l'exception confirme la règle** ».

3 Une mesure statistique de la quasi-implication. Représentations associées

La stratégie a donc été la suivante : si les variables réussite-échec, **booléennes** en l'occurrence, étaient indépendantes, le nombre de contre-exemples aléatoires à la règle de quasi-implication suivrait une certaine loi de probabilité, définissable sur la base des effectifs des échantillons. Si le nombre de contre-exemples attendus avec la probabilité p est supérieur à celui des contre-exemples observés, on admet la règle assortie **d'une mesure de qualité** p. Cette probabilité a été appelée **intensité d'implication**, En tant que telle, elle s'identifie à une échelle de probabilité, propriété que ne possèdent pas d'autres indices numériques comme l'échelle de Guttman, les indices de Loevinger ou de Shapiro, par exemple. Elle représente une sorte d'étonnement ***statistique***, *donc de nature anthropologique*[6], devant le faible nombre de contre-exemples par rapport à ceux qui étaient attendus dans la théorie. Voici une illustration de son caractère subjectif :

- sur un ensemble de 10 individus, des attributs *a* et *b* sont vérifiés respectivement 6 et 8 fois, sans contre-exemple à la règle « si *a* alors *b* ». Celle-ci est logiquement acceptable ; la fréquence de *b* sachant *a* est 1,
- sur un ensemble de 1000 individus, des attributs *c* et *d* sont vérifiés respectivement 600 et 800 fois, avec un seul contre-exemple à la règle « si *c* alors *d* ». La règle logique n'est plus acceptable.

Laquelle vous étonnerait ? A laquelle accorderiez-vous la meilleure qualité prédictive ? Dans le premier cas, la règle est stricte mais la confiance en elle est fragile. Dans le second cas, c'est le contraire, l'étonnement est plus grand en dépit de la moindre valeur de la fréquence conditionnelle. Ce paradoxe relatif à l'acceptabilité de la règle est soluble dans la subjectivité. La statistique ASI va en restaurer une composante objective. Certes, l'intensité d'implication est d'autant plus proche de 1 que la qualité pressentie de l'implication est grande. La relation qu'elle établit entre variables s'exprime fréquemment en termes de causalité. Mais cette *explicabilité causale* d'une variable par une ou plusieurs autres (J.-M. Levy Leblond parle de « **champ causal** »

[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) à tort et à raison. Intercritique de la science et du mythe, Paris : Seuil, p.160

[6] C'est aussi ce qu'affirme René Thom (« Paraboles et catastrophes », 1980, p.130) : « …le problème n'est pas de décrire la réalité, le problème consiste bien plus à repérer en elle ce qui a de sens pour nous, ce qui est surprenant dans l'ensemble des faits. Si les faits ne nous surprennent pas, ils n'apportent aucun élément nouveau pour la compréhension de l'univers : autant donc les ignorer » et plus loin : « … ce qui n'est pas possible si l'on ne dispose pas déjà d'une théorie ».

dans « Aux contraires ») que l'intensité d'implication évalue, n'est en rien **déterministe.** D'ailleurs, elle n'est pas **transitive** comme nous l'avons formalisé par l'étude des **règles d'exception** où les règles a =>b et b=>c ne s'accompagnent pas de la règle a =>c. Elle ne relève pas non plus d'un **déterminisme probabiliste** comme elle est souvent interprétée à tort : si 0.95 est une intensité d'implication de *a* sur *b*, cela ne signifie pas que *b* se réalise avec la probabilité 0.95 si *a* est réalisée. Ainsi, partis de recherche de règles strictes, respectant la logique platonicienne, j'ai transigé et cherché des quasi-règles, ne respectant plus la logique formelle en raison de ses contre-exemples. Cette démarche illustre -nous y reviendrons- le mode de pensée que l'on dit **dialectique**[7] car il accepte les contradictions et les intègre pragmatiquement tout en enrichissant la connaissance. Supposons donc effectué le calcul de l'intensité d'implication pour chaque paire de variables de la situation expérimentale. On ne conserve pour chaque paire que celle relative au couple conduisant à la plus grande intensité d'implication. Que faire du tableau de toutes ces valeurs ? Comment en dégager les lignes de force qui les structureraient comme le fait un plan factoriel ? Disposant d'un ensemble de relations binaires valuées, j'ai fait le choix de le représenter par un **graphe orienté pondéré et sans cycle,** image plus facilement appréhendable par l'utilisateur expert du domaine, par exemple, la didactique, la psychologie, la sociologie, la médecine, etc.. En général, il ne se réduit pas à un chemin linéaire puisque à une même « cause » peuvent être associés plusieurs « effets » et un « effet » peut être le fruit de plusieurs « causes ». Le problème de représentation d'**ontologies** est alors compatible avec ce cadre (travaux en collaboration avec **Jérôme David**). Les graphes suivants illustrent ces deux situations.

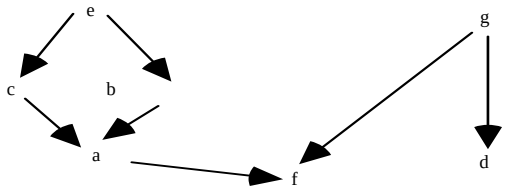


Fig.1

L'arc c →a représente la règle « si la variable c est choisie alors généralement la variable a l'est aussi » ou c=> a. e → c →a est un chemin implicatif.

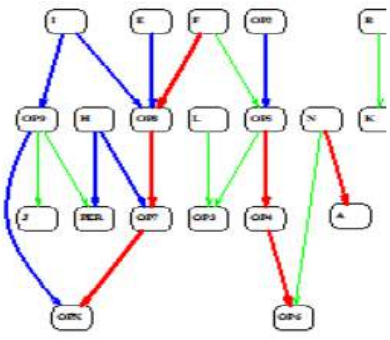


Fig. 2

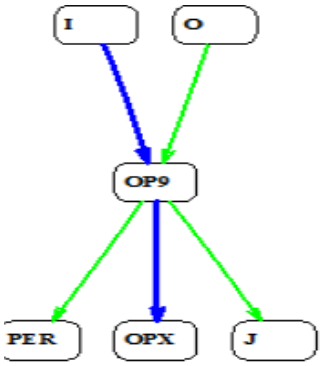


Fig. 3

[7] « …la dialectique n'entrant en scène que pour examiner et résoudre les difficultés logiques de niveau supérieur ». (« Emergence, complexité et dialectique », L. Sève et al, 2005, p. 86).

L'expert analyse et interprète alors les différents chemins en termes conceptuels en donnant du sens dans son domaine d'expertise à des **chemins** du graphe, à des **réseaux** comme dit Marc Bailleul, à un cône ascendants-descendants d'une variable au rôle d**'attracteur** qu'elle joue (Fig. 3)(voir la communication « Variable modale et cône implicatif » de D.Lahanier-Reuter et al), mais du sens aussi aux connexités ou à leur absence. Le graphe relatif au questionnaire construit selon la taxonomie cognitive l'a quasiment validée[8] en organisant généralement dans le préordre attendu les 5 classes et la vingtaine de sous-classes de cette taxonomie.

En psychologie du développement selon Piaget, la notion **d'abstraction réfléchissante** décrit le passage d'un niveau de conceptualisation à un niveau supérieur, chacun des niveaux étant constitué de règles portant sur des objets, puis d'opérations sur ces objets, puis sur des opérations sur ces opérations, etc... (par ex. schèmes, procédures, conception, méthode, ..). On retrouve, d'ailleurs, dans une théorie mathématique ces mêmes élargissements lorsque l'on passe d'un théorème à un corollaire ou, par exemple, dans l'étude des fonctions à celle, en analyse fonctionnelle, de fonction de fonctions. D'où l'idée de construire un second plan de relations implicatives, celui de **règles de règles** selon une **hiérarchie dite cohésitive** en raison de l'indice de **cohésion** qui permet d'engendrer des classes orientées de règles. A l'instar de certaines méthodes d'analyse de données, l'intérêt de passer, non linéairement, du niveau relationnel à celui de hiérarchie m'est apparu et nous avons étendu notre recherche de règles à celui de règles de règles ou méta-règles ou règles généralisées.

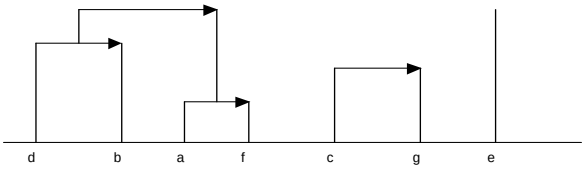


Fig. 4 - Graphe de la hiérarchie orientée

La règle $(d \Rightarrow b) \Rightarrow (a \Rightarrow f)$, illustrée par Fig.4 peut s'exprimer par la phrase : le « théorème » $d \Rightarrow b$ a généralement pour conséquence le « théorème » $a \Rightarrow f$. Par ex, de théorèmes comportementaux, on peut dégager **un trait** ou une **conception** (ex : conceptions du hasard par Dominique Lahanier-Reuter). Un indice statistique permet en outre de repérer les niveaux hiérarchiques correspondant à une **significativité des classes** formées à ces niveaux (en rouge sur la Fig.5). Nous construisons et détaillons le modèle mathématique qui fonde la hiérarchie cohésitive dans la communication « Hiérarchie de règles et conceptualisation ».

[8] Elle est d'ailleurs utilisée pour des évaluations d'acquisitions et pour des constructions de tests mathématiques en France et dans quelques pays francophones.

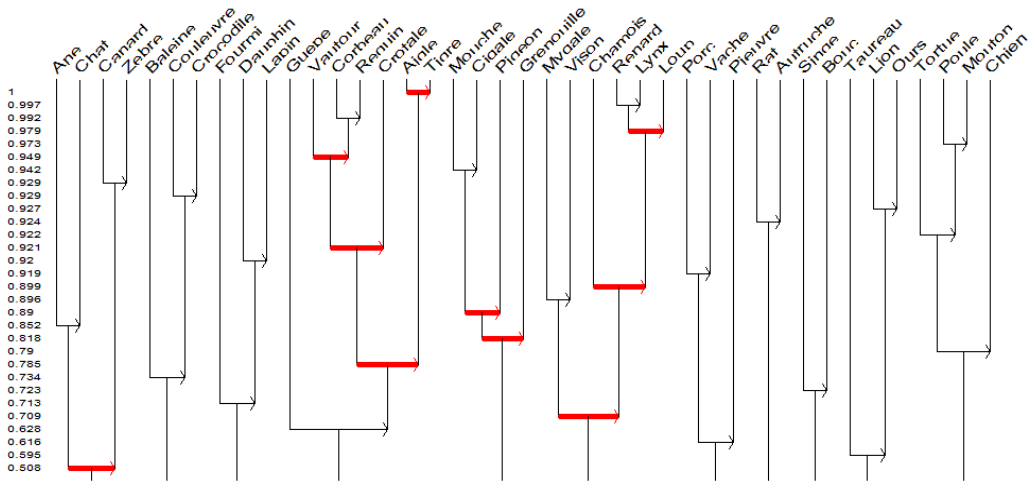


FIG.5 - Graphe de la hiérarchie orientée

4 Un détour par la philosophie des sciences.

A travers ces deux types de représentations de règles simples ou généralisées, qui mettent au jour deux types de structures dans l'ensemble des variables, nous répondons à la philosophie structuraliste qui estime que le **« tout » est plus riche que la somme de ses « parties ».** Je cite, à ce sujet, deux extraits du livre de L. Sève (ib. p. 58) « *…le **tout** ne se compose de rien d'autre que de ses **parties**, et pourtant il présente, en tant que tout, des propriétés n'appartenant à aucune de ses parties. Autrement dit, 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 ne peuvent donc s'expliquer par elles* » et de poursuivre *« Tout se passe donc comme si se produisait une génération spontanée de propriétés du tout…C'est le paradoxe de **l'émergence**».*

J'insiste sur cette propriété spécifique de l'ASI que la hiérarchie cohésitive satisfait de façon originale à travers l'extension des relations entre variables. C'est par un saut qualitatif, produit généralement par un effet de seuil dans la quantité (ex. : psychologie de groupe, vaporisation de l'eau…), que le tout, au prix d'une synthèse, prend son sens. Celui-ci s'extrait de la lecture véritablement dialectique du **rapport non-linéaire tout/partie**[9] (non-proportionnalité et non-additivité de la cause sur l'effet). Avec l'ASI, il y a alors **paradoxe** entre des contraires, lien et absence de lien, car le tout, constitué (on devrait dire *organisé*) de parties en un **système dynamique**, possède des propriétés que ne possèdent pas les parties et qui sont généralement de **niveau supérieur**. De la même façon, et métaphoriquement, en linguistique, la signification d'une phrase ne se fait pas par l'analyse de chacun de ses mots mais par l'interaction de ceux-ci[10]. La

[9].. comme le montrent les équations différentielles non-linéaires de l'indice fondamental par rapport aux paramètres cardinaux des observations, contrairement à ce qui est observable avec d'autres indices concurrents.. « La société n'est pas constituée d'individus, mais exprime la somme des relations, des rapports où ces individus se situent les uns par rapport aux autres » (K. Marx, « Manuscrits de 1857-1858). Une rue n'est pas la somme des maisons qui y figurent. De même une ville n'est pas somme de ses rues, etc.

[10] « Il n'y a ni additivité ni proportionnalité entre le sens des unités (mots) et celui de la phrase. On

logique qui sous-tend ce rapport tout/parties est **dialectique** (et non pas dichotomique) car elle concilie interactivement des contradictions : règle et non-règle, **instabilité d'un système dynamique et stabilité structurelle**. Elle se fonde en règle sur l'inexistence importante du contre-exemple et en métarègles sur la négation de l'entropie, du désordre. En cela, la logique dialectique s'oppose à la logique stricte (du mathématicien) sans, bien sûr, être un sophisme. La fécondité et l'originalité de l'ASI tiennent à ce caractère, particulièrement dans l'analyse hiérarchique manifestement non-linéaire où le tout fonde son sens, non par addition des propriétés de ses parties (sous-classes) mais par la synthèse des interactions inférentielles. C'est par la notion de **niveau significatif** que nous pouvons mettre en évidence le phénomène de **propriété émergente**. En ce sens, l'A.S.I. apparaît comme une sorte d'avatar de l'apprentissage non linéaire des connaissances, apprentissage fait de ruptures et de reconstruction dialectique (cf. l'épistémologie de G. Bachelard ou de Lev Vygotsky et notre communication déjà citée sur la conceptualisation). A l'opposé de l'A.S.I., le rapport tout-parties serait **linéaire** dans le cas d'emboîtements de classes comme en clustering fondée sur la similitude jusqu'à son nœud terminal.

Des premières applications, en didactique des mathématiques et sur des problèmes de gestion humaine, il est apparu que le premier indice, l'intensité d'implication, pêchait dans sa discrimination lorsque le nombre de sujets s'accroissait. D'où la nécessité, pour cette raison et pour approcher au plus près de l'interprétation causale des règles, d'intégrer à cette intensité l'information apportée par la **contraposée de l'implication**. Ainsi, non seulement la mesure de l'implication va prendre en compte la relation «si *a* alors généralement *b* » mais également, dialectiquement, sa contraposée « **si *non b* alors généralement *non a*** ». Cet indice nouveau est basé sur la notion d'**entropie, d'information,** des expériences évaluant les deux règles. Ce choix permet d'agir plus en profondeur pour accéder aux « pépites de connaissances », règles qui seraient rejetées ou ignorées par une méthode basée sur le support et la confiance (l'algorithme a priori d'Agrawal relève, semble-t-il, de la pensée linéaire).

Bien que nous recherchions des relations entre variables par des règles non symétriques et que ces relations puissent souvent s'exprimer en termes de causalité, nous ne prétendons pas, comme je l'ai dit plus haut, qu'elles soient déterministes, mais simplement qu'elles ont la capacité de permettre d'émettre des hypothèses quantifiables sur leur prédictibilité.

5 Un logiciel de traitement informatique de l'A.S.I.

Lorsque le nombre de sujets et de variables croît, il est difficile d'effectuer les calculs associés et surtout de fournir les deux représentations : graphe et hiérarchie. Au début, les deux tâches étaient faites à la main. La plaie ! Je l'ai fait pour ma thèse. J'ai

---

voit se dessinée une topologie du sens » (F. Gaudin dans « Emergence, complexité et dialectique »).et « …le mot isolé de la langue chinoise n'a en vérité ni signification ni existence à part, chacun ne reçoit sa signification que du parler même (de l'intonation, etc….), pris isolément il a dix, voire quarante significations,… ; si nous soustrayons ce mot à la totalité, il se perd dans une creuse infinité. « (F.-W. Schelling, « Philosophie de la mythologie », p.361).

alors écrit un premier programme en Basic (!) qui effectuait les calculs et construisait la **hiérarchie**. Un doctorant de Lerman -H.Rostam- construisit à son tour le **graphe implicatif** à l'aide d'un programme plus sophistiqué dont j'avais construit l'algorithme. Puis, deux de mes doctorants, Saddo Ag Almouloud et Harrisson Ratsimba-Rajohn, intégrèrent l'ensemble en un seul logiciel que nous avons dénommé **Classification Hiérarchique Implicative et Cohésitive** sous l'acronyme « **C.H.I.C**. ». Enfin, depuis la fin des années 90, Raphaël Couturier a unifié le traitement complet des calculs et des représentations tout en lui apportant continûment les améliorations dues au développement de la théorie sous-jacente et de la variété des applications. En particulier, une option permet de faire apparaître les différents éventuels prédicteurs et les descendants d'une variable, mais également de déterminer la conjonction optimale de ceux-ci au sens de leur originalité. La possibilité de modifier le seuil de construction du graphe implicatif met en lumière la plasticité de la structure des variables, tout en conciliant, dialectiquement l'instabilité d'un système dynamique et sa stabilité structurelle. Une version écrite en R est présentée par Raphaël Couturier lors du colloque ASI 8.

Ce logiciel est opérationnel à travers le monde puisque 26 pays, gréco-latins pour la plupart, le possèdent et l'utilisent[11]. Il y joue, pour la recherche, le double rôle de **révélateur et d'analyseur**. Si bien que des questions sur les possibilités et les limites de l'ASI se font jour à travers des questions d'ordre général mais aussi spécifiques du fait des traditions et des cultures différentes. Outre le lien épistolaire, elles se renforcent au fil de nos rencontres internationales sur l'ASI, de ASI 1 à ASI 7 en 2013 et maintenant en 2015 en Tunisie. Ce sont ces questions qui poussent la théorie et son outil vers des développements continus comme nous allons le voir. Pour illustrer ceci, je citerai Anne Lauvergeon dans « La femme qui résiste » (Plon, 2012) : « … *lorsque l'on produit, on finit également par concevoir* ».

6 Extension de la méthode à d'autres variables

C'est donc à travers les différentes situations rencontrées que la limitation aux variables binaires, ayant servi à donner un sens ensembliste à l'implication, est apparue contraignante. Celui qui a tiré le premier est Marc Bailleul qui a voulu rechercher des relations de type préférences entre des assertions d'enseignants. Des **variables modales** (« un peu », « beaucoup », « pas du tout », …), devaient être traitées. Il a proposé un premier indice satisfaisant à la mesure d'expression du type : « **si pour *a* la modalité « *un peu* » est choisie alors généralement pour *b* une modalité supérieure ou égale est choisie »**. Dans sa thèse, Marc Bailleul a obtenu d'excellents résultats, dont certains imprévisibles, relativement à 4 conceptions de l'enseignement des mathématiques sous le regard des enseignants. Nous utiliserons son étude dans plusieurs communications de ASI 8. Afin de procéder, relativement à la même problématique, par extension de l'intensité d'implication entre variables booléennes, avec J.-B. Lagrange, nous avons alors défini une nouvelle mesure portant sur des **variables numériques** et modales permettant d'attribuer une valeur à des énoncés comme : « **si $a = \alpha$ alors généralement $b \geq \beta$** ». Jean-Claude Régnier, de son côté, a ramené la problématique de la recherche de

[11] A l'heure actuelle, Pablo Gregori (Castellon) et Ruben Rodriguez (Quito) œuvrent pour traduire CHIC en une version informatique sous R

relations entre préférences à celle de **variables de rangs** qui a fourni une autre extension de l'A.S.I...

De là, suite à une question d'E. Diday se plaçant dans le cadre des **variables symboliques** et par une extension des variables numériques, nous avons affecté des valeurs à des expressions « si ***a*** **prend ses valeurs dans l'intervalle** $\mathbf{I_a}$ **alors généralement** ***b*** **prend ses valeurs dans** $\mathbf{I_b}$ ». L'idée maîtresse a consisté à partitionner l'étendue des valeurs prises par chaque variable en sous-intervalles maximisant leur variance interclasse. L'intérêt de cette nouvelle catégorie de variables, **dites intervalles**, se manifestait dans l'enseignement, pour comparer hiérarchiquement des performances dans des disciplines différentes et dans la recherche de transfert de compétences spécifiques vers d'autres compétences. On voit alors que de ces **variables-intervalles** traitées en collaboration avec E. Diday et Pascale Kuntz, nous parvenons naturellement à **des variables floues** comme nous les rencontrons dans l'étude de relations du type : « si la tension *a* d'un sujet est plutôt élevée alors généralement son rythme cardiaque *b* l'est aussi ». On pressent les applications pratiques qui peuvent en découler dans des problèmes de construction ou de détections de pannes. Grâce à la collaboration de Fabrice Guillet, Maurice Bernardet, Raphaël Couturier, Filippo Spagnolo, cela m'a permis de modéliser la notion de variable floue dans le cadre des variables-intervalles et d'en faire une présentation dans un congrès sur la logique floue. Un problème est survenu un jour dans le traitement implicatif d'un ensemble trop large de variables au point de rendre illisibles les graphes obtenus. Avec Raphaël Couturier, Fabrice Guillet et Robin Gras, nous avons défini une relation d'équivalence entre les variables, basée sur leurs comportements implicatifs voisins, qui a conduit à substituer à un paquet de variables un représentant leader de ce paquet. Cette **réduction** s'est avérée efficace dans de nombreuses autres situations, par exemple dans la thèse de Laurence Ndong.

Un petit arrêt pour parler des collaborations dans l'équipe DUKE, toujours appuyées sur des choix épistémologiques en réponse à des attendus sémantiques. L'explosion d'une multitude de questions applicatives ou théoriques, toujours non symétriques, a conduit à des travaux communs entrepris, depuis 1990, avec certains membres de l'équipe du DUKE actuel : je citerai par exemple, avec Pascale Kuntz sur **les hiérarchies de règles** (ah ! la démonstration de l'ultramétricité de la hiérarchie ! le modèle algébrique de la hiérarchie !), sur les **règles d'exception** avec Einoshin Suzuki, la **redondance de règles** avec Pascale, avec Julien Blanchard sur **l'analyse entropique**, les **variables temporelles**, **l'expression de gènes en bio-informatique** avec Gérard Ramstein, etc.. Par exemple, comment la nécessité de variables temporelles est-elle apparue ? Eh bien, pour rendre compte de l'évolution des relations entre variables économiques, sociales ou cognitives. En effet, peut-on expliquer dans l'enseignement, l'implication entre variables lorsque l'on procède à des interventions en cours d'année sur l'apprentissage ou encore en socio-psychologie par des interventions successives, par des entretiens ou par des stages (en collaboration avec D.Pasquier). Nous avons alors formalisé ces variables indexées par le temps en **variable vectorielle,** où une variable est modélisée par un vecteur paramétré par le temps. Puis, Julien Blanchard, en collaboration avec Fabrice Guillet et moi-même, a, de façon différente, défini des **variables séquentielles** modélisées par un processus de Poisson. Autant de nouveaux concepts et de nouveaux champs d'application nés de problématiques variées et

d'extensions attendues. **Des extensions généralisantes** de l'ASI ont vu le jour récemment :

- d'une part, à un ensemble continu de sujets, par exemple des couleurs, des opinions, muni d'une loi de distribution donnée, généralisation puisque jusqu'alors le modèle se limitait à des ensembles discrets et finis ;
- d'autre part, à l'ensemble des valeurs prises par les différents types de variables dans des espaces continus, par exemple des champs, munis de lois données ;
- enfin, lors du colloque ASI 8, nous présentons un prolongement à des variables continues, ce qui représente une avancée théorique reconnue majeure en analyse de données.

J'insiste sur une remarque susceptible de satisfaire tout cartésien : lors de chaque extension, nous nous sommes efforcés de prouver et nous y sommes parvenus, que **la restriction au cas fondamental de variables binaires et d'espaces discrets était toujours satisfaite**. L'emboîtement mathématique est original et rigoureux.

7 Rôle explicatif de variables supplémentaires

Une autre question m'a taraudé très tôt. Y a-t-il dans la population de sujets concernés par une étude, une structure interne qui expliquerait la structure des variables révélée par l'ASI ? Supposant une structure implicative de variables obtenue par un graphe ou une hiérarchie, est-il possible de désigner les sujets et les catégories de sujets plus ou moins responsables des éléments de ces structures ? Par exemple, si nous observons que dans l'ensemble de classes scolaires une certaine conception de notions géométriques entraîne certaines conduites de réponses, à quel type d'élèves peut-on plus spécifiquement l'attribuer ? Inversement quels sont les élèves qui y seraient réfractaires ? Avec les thèses de Harrisson Ratsimba-Rajohn et Marc Bailleul, nous avons formalisé et exploité, dans l'A.S.I., la notion de variable supplémentaire (ou **exogènes** par opposition aux variables **endogènes** analysées) et sa **contribution** (on parle aussi de **typicalité**) à des éléments de structure, des **réseaux** ou des classes cohésitives par exemple, sémantiquement expliquées. Mieux encore, nous avons pu définir de façon rétroactive une **structure topologique duale** sur l'ensemble des sujets. Ainsi, par exemple, une conception « emblématique de l'école républicaine » est renforcée dans un milieu scolaire favorisé où la distance inférée entre les élèves par un élément des structures de règles est la plus faible (travaux avec Dominique Lahanier-Reuter). Mais puisque j'ai cité au fil des pages les apports et le rôle d'aiguillons de collaborateurs pour le développement de l'ASI, en particulier de l'équipe DUKE, je dois mettre en exergue le rôle critique, poil à gratter mais constructif que joue Jean-Claude Régnier tant à l'égard de la théorie et ses applications qu'à l'égard de la diffusion, la popularisation de l'ASI puisqu'il a pris en charge la présidence des manifestations internationales ASI. Combien de lapsus, d'erreurs, de maladresses, d'excès de précipitation, Pascale Kuntz et Jean-Claude ont-ils redressés ! Je n'oublie pas non plus le dévouement fidèle et les compétences informatiques de Raphaël Couturier sans lequel l'ASI, privée de CHIC, et maintenant de RCHIC, ne serait qu'une construction théorique et peut-être spéculative à contempler. Il a comblé, avec CHIC, mon handicap en informatique comme Pascale a bien voulu le faire pour l'anglais. Je crains d'ailleurs

que cette dernière carence ait coûté une meilleure audience de l'ASI sur le plan international. Cependant, d'autres équipes internationales participent également aux travaux sur le développement et les applications de l'ASI en utilisant le logiciel CHIC. Je cite, en particulier, étant donnée leur régularité de participation :

- conduite activement par Pilar Orus (Université de Castellon en Espagne), une diversité de noyaux hispanisants en Espagne (Pablo Gregori, Eduardo Lacasta, Miguel Wilhelmi, …), à Cuba (Larisa Zamora, …), au Chili (Guzman-Retamal, …) ; en Argentine (Pablo Carranza, …), en Equateur (Ruben Rodriguez) ont créé des liens permanents ;
- en Italie, Filippo Spagnolo, récemment décédé, avait créé une équipe qui se resserre maintenant autour de Benedetto Di Paola ;
- au Brésil, autour de Saddo Ag Almouloud et grâce au renforcement apporté par des conventions avec Jean-Claude Régnier et l'université Lyon II,
- à Chypre autour de Athanasios Gagtasis,
- en Slovaquie avec Lucia Rumanova.
- de façon plus épisodique, des chercheurs de Belgique, de Suisse, d'Allemagne, de Grèce, de Roumanie, de Tchéquie, du Japon, du Vietnam, du Canada, du Mexique, du Gabon, de Tunisie, et j'en oublie, utilisent méthode et outil de l'ASI.

8 Les spécificités de l'ASI et conclusion

L'ASI, mesurée à d'autres méthodes d'analyse de données, présente des caractères originaux importants. Je les résume :

- **modèles successifs** de variables répondant à des contraintes épistémologiques explicites compatibles avec la sémantique des situations à modéliser ;
- **non-symétrie** de la méthode ;
- extension progressive de la nature des variables traitées tout en conservant les propriétés de plongement ;
- **capacités pédagogiques et ergonomiques** des représentations, en particulier pour l'examen des règles généralisées ;
- **dualité structurelle** des deux espaces en jeu : sujets et variables avec les notions de contribution et de typicalité aux structures ;
- **extension du discret fini au continu** tant pour les variables que pour les sujets ;
- **originalité du raisonnement dialectique** à la base de la définition des règles simples et généralisées ;
- la **simplicité du modèle mathématique** sous-jacent lui assurant accessibilité plasticité et fécondité utiles pour répondre à des attentes applicatives dans de larges domaines.

D'aucuns diront que l'ASI, par l'amplitude de ses champs d'application et par l'homogénéité de ses propriétés analytiques et graphiques présente une **nature paradigmatique** originale. Doit-on le reconnaître ? Djamel Zighed, Directeur de

l'Institut des Sciences de l'Homme de Lyon exprime cette idée de la façon suivante : « *l'ASI n'est pas une méthode mais un cadre théorique, large, dans lequel se traitent des problèmes modernes de l'extraction des connaissances à partir des données. C'est une théorie générale dans le domaine de la causalité parce qu'elle répond à des faiblesses d'autres théories, elle apporte un outillage formel et des méthodes pratiques de résolution de problèmes. Ses applications sont multiples... »*. Pour terminer, je voudrais remercier tous les participants à ce 8ème colloque scientifique international sur l'Analyse Statistique Implicative organisé sous la direction scientifique et organisationnelle de Jean-Claude Régnier, Yahya Slimani, Makram Ben Jeddou, Ahmed Dhouibi et Inès Ben Tarbout, pour leurs contributions au développement du cadre théorique, méthodologique et applicatif au travers de la mise en œuvre des concepts et des outils qui le constituent, ainsi que par la confrontation aux défis théoriques traduits par des problématiques nouvelles.

Références

[1] *L'implication statistique. Nouvelle méthode exploratoire de donnée,* sous la direction de R.Gras, et la collaboration de S. Ag Almouloud, M. Bailleul, A. Larher, M. Polo, H. Ratsimba-Rajohn, A.Totohasina, La Pensée Sauvage, Grenoble. ISBN : 2.85919.129.1 (1996)

[2] *Mesures de Qualité pour la Fouille de Données,* H. Briand, M. Sebag, R. Gras et F.Guillet (eds), RNTI-E-1, Cépaduès, 2.85428.646.4 (2004)

[3] *Quality Measures in Data Mining* , F. Guillet et H. Hamilton (eds), Springer, ISBN : 3.540.44911.6. (2007)

[4] *Statistical Implicative Analysis, Theory and Applications,* R. Gras, E. Suzuki, F. Guillet, F. Spagnolo, (eds), Springer, ISBN : 978.3.540.78982.6 (2008)

[5] *Analyse Statistique implicative. Une méthode d'analyse de données pour la recherche de causalités,* sous la direction de Régis Gras, réd. invités R. Gras, J.C. Régnier, F. Guillet, Cépaduès Ed. Toulouse, ISBN : 978.2.85428.8971. (2009)

[6] *Teoria y Aplicaciones del Analisis Estadistico Implicativo,* Eds : P.Orus, L.Zemora, P.Gregori, Universitat Jaume-1, Castellon (Espagne), ISBN : 978-84-692-3925-4, (2009).

[7] *L'Analyse Statistique Implicative : de l'exploratoire au confirmatoire.* J.C. Régnier, Marc Bailleul, Régis Gras, (Eds) Université de Caen, ISBN : 978-2-7466-5256-9, (2012)

[8] *L'Analyse Statistique Implicative. Méthode exploratoire et confirmatoire à la recherche de causalités* sous la direction de direction de Régis Gras, (Eds.) R. Gras, J.C. Régnier, C. Marinica, F. Guillet, Cépaduès Ed. Toulouse, 201, ISBN : 978.2.36493.056.8. (2013)

BIG DATA ET DATAMINING: ENTRE MYTHES ET REALITES

Yahya SLIMANI [12]

L'évolution de l'informatique a été et reste toujours marquée par l'apparition de concepts, de technologies, etc. dont certaines disparaissent au bout d'un certain temps (systèmes experts, téléinformatique) alors que d'autres sont plus ou moins pérennes (processus, technologie objet, etc.). Avec le développement du Web qui permet de générer de la connaissance collective et l'apparition des réseaux sociaux, nous assistons de plus en plus à l'apparition de termes qui relèvent parfois du *buzz* et ont une résonnance beaucoup plus commerciale que technique. Cette conférence a pour objectifs de :

- Démystifier certaines idées et concepts qui sont essentiellement utilisés à travers des stratégies commerciales pour attirer de plus en plus de sociétés et de clients vers des « technologies informatiques » qui s'avèrent parfois très coûteuses et inefficaces pour les entreprises, sans leur apporter une plus value dans leurs systèmes d'informations.
- Présenter le domaine du datamining, dans lequel les statistiques et l'analyse de données, de manière générale, sont très largement utilisées.
- Présenter la notion de Big Data qui est parfois utilisée à mauvais escient dans le seul but d'inciter les sociétés et les entreprises à investir dans cette nouvelle « technologie », sans se soucier des bouleversements qu'elle peut apporter au niveau de leurs systèmes d'informations et de leurs exploitations.
- Présenter les liens qui existent entre datamining et Big Data pour l'exploitation de larges volumes de données.

[12] Institut Supérieur des Arts Multimédia (ISAMM) Université de la Manouba (Tunisie) http://www.yahya-slimani.net/

REGARDS SUR LES RAPPORTS ENTRE STATISTIQUE ET ENSEIGNEMENT MATHEMATIQUE

Francois PLUVINAGE[1]

Un éclairage de nature socioépistémologique fait bien voir comment les statistiques ont pénétré les mathématiques, tardivement dans l'histoire (XIXe siècle). Jusqu'au XVIIe siècle, les pratiques de référence étaient la prise de données par écrit, et les pratiques sociales touchaient à l'anticipation, par exemple la quantité de blé à attendre de la prochaine récolte, et à des décisions, telles celles des impôts à prélever.


Die
Staatsklugheit
nach ihren erſten
Grundſätzen
entworfen
von
Gottfried Achenwall
ordentlichem Lehrer des Natur- und Völkerrechts wie auch der Politick auf der Universität zu Göttingen und auſſerord. Mitgliede der Königl. Gesellſchaft der Wiſſenſchaften daſelbſt.
Göttingen
in Verlag der Witwe Vandenhoeck 1761.

Certains textes, comme celui que donne la Brockhaus Enzyklopädie, prétendent que le terme allemand *Statistik* a été introduit par Gottfried Achenwall (1719-1762). Il semble cependant que le mot, construit sur le latin, soit apparu plus tôt. La figure ci-contre reproduit la page de couverture du livre « *Die Staatsklugheit* » (la sagesse de l'Etat, en allemand) d'Achenwall, ouvrage dans lequel le mot *Statistik* apparaît.

La pénétration des statistiques dans le champ des mathématiques s'est produite lorsque la saisie de données confiée depuis l'antiquité à quelques spécialistes, tels les scribes égyptiens, s'est combinée avec les probabilités au service de prédictions prenant en compte des risques. Dans un premier temps, c'est-à-dire avant que les mathématiques n'incluent la statistique comme l'une de ses branches, les pratiques de référence se sont enrichies de la représentation des données et les pratiques sociales de la formalisation et de la prédiction argumentée. Ensuite se sont glissées dans les pratiques de référence la modélisation et les techniques d'enquête, incluant l'échantillonnage, et dans les pratiques sociales les évaluations de risques.

[1] CINVESTAV-IPN Mexico (Mexique) email: fpluvinage@cinvestav.mx

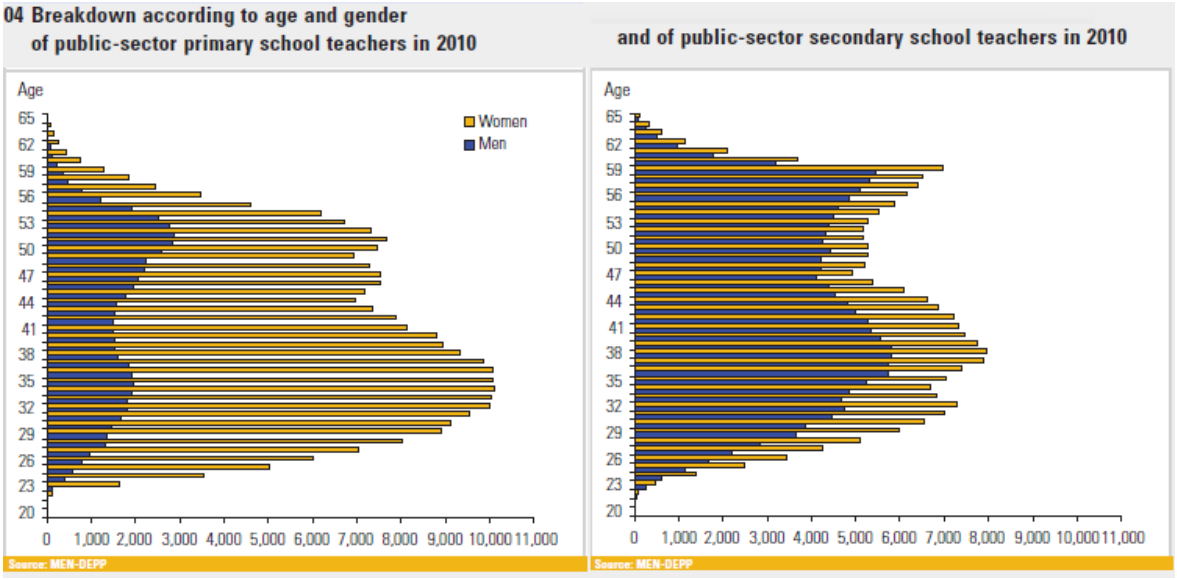


Figure 1: Une représentation qui interpelle : Pyramides des âges par genres des enseignants français des premier et second degrés.

Au service de l'institution, les statistiques d'une part alimentent le système éducatif en données factuelles et d'autre part fournissent des indicateurs de la qualité de l'enseignement. Ce dernier rôle est plus controversé au sein de la communauté éducative, allant jusqu'à être générateur de tricheries et même de fraudes importantes. C'est que les objectifs d'explications de phénomènes touchant l'enseignement et les plans d'améliorations ne coïncident pas nécessairement quand ils se situent au niveau d'un établissement ou d'une population expérimentale et quand ils se situent au niveau de tout un système éducatif.

Probabilités d'accès au lycée en fonction du niveau scolaire au CP (première année d'école primaire) et du milieu social d'origine

Niveau ⇨	Faible	Moyen	Bon
Enfants de cadres supérieurs	63,4	80,6	90,8
Enfants de prof. intermédiaires	33,2	54,3	74,0
Enfants d'ouvriers	15,8	31,0	51,9

Source : Duru-Bellat, M., Jarousse, J.-P. et Mingat, A. (1993)

Tableau 1

Il est clair qu'aujourd'hui la mise en œuvre de statistiques avec les moyens de traitement fournis par l'informatique peut donner lieu à des usages sans intérêt, par exemple en appliquant un quelconque test à une population puis en passant ses résultats à la moulinette d'un logiciel d'analyses statistiques. Mais en sciences de l'éducation, touchant en particulier l'enseignement des mathématiques, nombreuses sont les questions de recherche qui demandent de s'appuyer sur l'usage d'analyses statistiques pour leur étude.

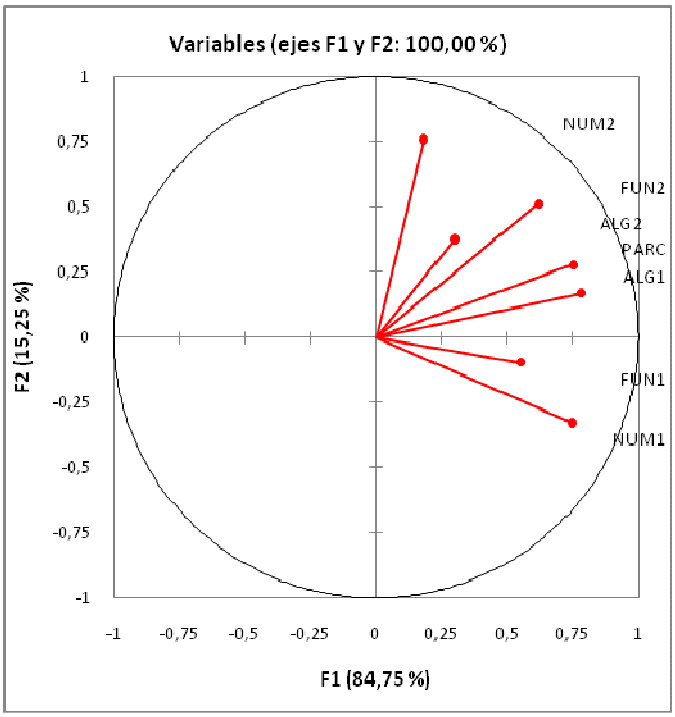


Figure 2: Sortie d'analyse discriminante extraite de Cuevas, Martinez & Pluvinage (2011)

Nous illustrerons ce propos par des exemples, utilisant pour certains des données accessibles au public et s'appuyant pour d'autres sur des résultats d'enquêtes. Au service plus spécifiquement de la didactique des mathématiques, nous verrons que les statistiques sont bienvenues dans des études comparatives et que les techniques d'analyse statistique, comme l'analyse implicative, peuvent s'imposer en contexte expérimental. En rapport avec le dernier volet de notre exposé, la participation de professeurs en exercice dans de telles études nous semble éminemment souhaitable.

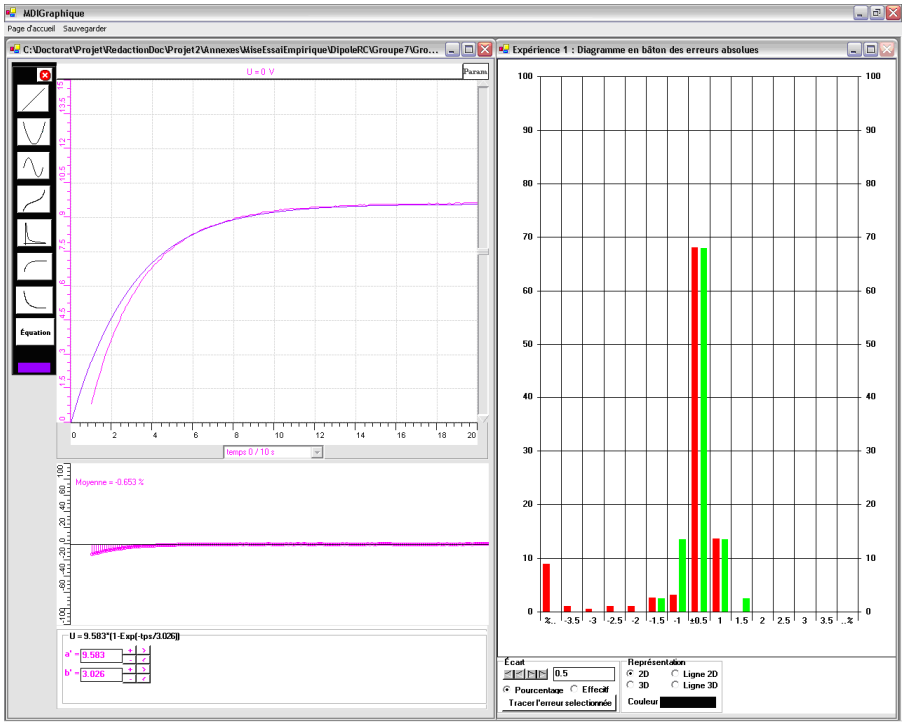


Figure 3: Modélisation avec recours à la statistique, extrait de Touma (2009)

Codage : Réussite forte ou faible, un problème ouvert pour l'ASI

Un problème que nous rencontrons dans des études expérimentales et pour lequel des recherches sont encore à conduire est tout bêtement celui du codage de réponses quand celles-ci donnent lieu, ce qui est fréquent, à des niveaux de prise en compte. Le cas le plus simple est celui d'un intermédiaire entre échec et réussite sur critère mathématique : des réponses non entachées d'erreur, mais incomplètes. On peut parler alors de réussite stricte et réussite large, et on a le choix entre une analyse qui distingue Réussite stricte et l'oppose aux autres réponses (Echec et Réussite large) et une analyse qui regroupe Réussites stricte et large, les opposant aux seuls échecs. Dans un exemple que nous avons proposé pour être analysé dans un atelier que dirigeait Maria Trigueros à Mexico, les arbres de similarités de l'une et l'autre analyse faisaient apparaître deux grands groupes, mais la composition des groupes différait entre les deux cas.

L'enseignement du domaine Probabilités et Statistique

Le dernier volet de notre présentation concerne l'enseignement mathématique de la branche *probabilités et statistique*, et la place que les enseignants lui accordent. Des situations qui nous semblent paradigmatiques et que nous commenterons sont celles de décision en présence d'incertitude et celles de modélisation avec contrôle statistique. Les unes et les autres justifient l'introduction des paramètres statistiques de position et dispersion, des lois usuelles et des outils d'analyse. Leur orchestration en classe sera d'autant plus efficace qu'elle sera conduite par un enseignant qui en a eu l'expérience personnelle dans son vécu.

Etudes citées

Cuevas Vallejo, C. A., Martínez Reyes, M. & Pluvinage, F. (2011). La promoción del pensamiento funcional en la enseñanza del cálculo: un experimento con el uso de tecnologías digitales y sus resultados. *Annales de Didactique et de Sciences Cognitives*, Vol. 17, 137-168

Duru-Bellat, M., Jarousse & J.-P., Mingat A. (1993). Les scolarités de la maternelle au lycée. Etapes et processus dans la production des inégalités sociales. *Revue française de sociologie*. Vol. 34-1. 43-60. Consulté en ligne à

http://www.persee.fr/web/revues/home/prescript/article/rfsoc_0035-2969_1993_num_34_1_4218

Touma, G. (2009). Une étude sémiotique sur l'activité cognitive d'interprétation. *Annales de Didactique et de Sciences Cognitives*, Vol. 14. 79-101

Quelques sources officielles

Online Education Database

http://www.oecd.org/document/54/0,3343,en_2649_39263238_38082166_1_1_1_37455,00.html

Panorama de l'éducation 2010 au niveau international (OCDE)

http://browse.oecdbookshop.org/oecd/pdfs/browseit/9610074E.PDF

Sites français consultés

Données internationales http://www.ciep.fr/sitographie/ries44.php
Données sur la France http://media.education.gouv.fr/file/07/0/3070.pdf
http://media.education.gouv.fr/file/etat20/41/7/The_state_of_education_168417.pdf

NOTION DE CHAMP IMPLICATIF EN ANALYSE STATISTIQUE IMPLICATIVE

Régis GRAS, Pascale KUNTZ et Nicolas GREFFARD[1]

NOTION OF IMPLICATIVE FIELD IN STATISTICAL IMPLICATIVE ANALYSIS

RÉSUMÉ

Dans le cadre de la théorie de l'Analyse Statistique Implicative, la problématique de la stabilité de l'indice qui permet de définir et évaluer la qualité de l'indice d'implication est posée par l'utilisateur qui renouvelle ses expériences dans un domaine particulier. Dans cet article, nous étudions ce problème en invoquant les concepts différentiels de l'analyse mathématique. Nous examinons un à un les paramètres intervenant dans la formule donnant l'indice d'implication. Nous comparons les variations de ces paramètres avec ceux d'autres indices classiques utilisés en fouille de données. Nous étendons cette étude par celle de la structure de l'espace vectoriel qu'ils engendrent et en centrant cette étude sur la notion de gradient implicatif. De là, nous illustrons par une représentation géométrique la problématique de l'équilibre de l'indice via une discrétisation des surfaces équipotentielles.

Mots-clés : *analyse statistique implicative**,** intensité d'implication, indice d'implication, différentielle, gradient, surface équipotentielle*

ABSTRACT

In the context of the theory of implicative statistical analysis, a user repeating some experimentation in a specific domain is faced with the issue of the robustness of the metric appraising the quality of the implicative index. In this paper, we address this problem through a differential analysis instead of bootstrapping. We study each individual parameter involved in the implicative index equation. And we compare their variations with those of other indices from the data-mining literature. Furthermore, we study the structure of the vector field they span by focusing on the notion of implicative gradient. From there, a geometrical representation is used to illustrate the index equilibrium problematic through a series of figures of equipotential surfaces.

Keywords *: statistical implicative analysis**,** implication intensity, measure of implication, differential, equipotential surface.*

1 Introduction

Le chercheur en sciences et particulièrement en sciences humaines, expérimentateur ou non, vise, à travers ses interrogations, à construire ou conforter des connaissances dans son domaine. Il est confronté à des situations, vécues par une population de sujets dans lesquelles apparaissent des phénomènes, par exemple des attributs, conduisant à de nombreuses données d'observation. De celles-ci, parmi d'autres informations, il veut extraire des relations d'association entre les variables observées. Puis, à partir de ces relations, il cherche à constituer des savoirs stables et partagés dans son domaine. Parmi

[1] École Polytechnique de l'Université de Nantes, Équipe DUKE Data User Knowledge, Laboratoire d'Informatique de Nantes-Atlantique (LINA), UMR 6241, Site de la Chantrerie, rue C. Pauc, BP 44306, Nantes cedex 3, e-mail : regisgra@club-internet.fr, pascale.kuntz@univ-nantes.fr et nicolas.greffard@univ-nantes.fr.

celles-ci, les relations causales ou règles non symétriques -« cause => effet (ou conséquence) » - figurent en bonne place pour leur intérêt de découvertes à valeur prédictive. Mais rares sont les réponses respectant les principes métaphysiques de la logique formelle c'est-à-dire s'exprimant en terme de « vrai » ou de « faux ». Ces réponses seraient plus acceptables évaluées dans une logique **dialectique,** « logique des contraires », logique qui permet de dépasser les contradictions. C'est cette composante épistémologique qui guide, de façon originale, la méthode d'analyse de données, l'Analyse Statistique Implicative -d'acronyme ASI-, que nous avons élaborée afin que prévalent les **quasi-règles**, c'est-à-dire celles qui acceptent des contre-exemples sans effacer leur plausibilité. En effet, une règle non strictement satisfaite y trouve sa place en même temps qu'elle serait logiquement réfutable ; autrement dit, simultanément, règle et sa négation ont droit de coexister. La contradiction est provisoirement levée en ASI par la donnée d'un seuil probabiliste variable dont le chercheur contrôle la flexibilité : au-delà, la quasi-règle est acceptée, en-deçà elle est mise à l'écart. Un certain déterminisme probabiliste confère à la quasi-règle une prédictibilité statistique. Autrement dit, comme l'énonce le philosophe Lucien Sève (Sève, 2005, p.104), « on perd en rigueur ce que l'on le gagne en richesse » et, ajouterons-nous, en fécondité.

La méthodologie que nous développons dans l'ASI et que nous implémentons dans le logiciel Classification Hiérarchique Implicative et Cohésitive, d'acronyme CHIC, consiste, en partant de la contingence des données, issues d'un ensemble répétitif de phénomènes, à analyser le singulier, le « surprenant ». Celui-ci qui constitue la quasi-règle, débouche dialectiquement sur une interprétation, une signification et donc une esquisse de savoir général, admis provisoirement comme une réalité indépendante, platonicienne. Car, comme le dit plus loin L. Sève : « il y a toujours de l'universel dans le singulier » (p. 200). Pour ce faire, nous construisons un modèle mathématique qui permet la comparaison statistique du contingent et du théorique aux ordres du modèle. Au sujet de son application possible en biologie, Jean-Claude Ameisen déclare dans « Sur les épaules de Darwin, tome 2» (p. 163) : « *A la recherche des régularités, des mécanismes et des relations de causalité qui rendent compte des formes et des comportements du monde vivant, la formalisation mathématique jouera un rôle de plus en plus important en biologie ».*

Plus précisément, plaçons-nous donc dans le cadre de la théorie de l'Analyse Statistique Implicative, où l'on croise des sujets ou des objets avec des variables de natures variées. Restreignons-nous, pour le moment, au cas de variables binaires telles que a et b. La problématique majeure consiste en la recherche d'une mesure qui permettrait de quantifier la règle quasi-implicative de a sur b. Cette mesure, en ASI, se fonde sur le nombre de contre-exemples à l'implication, nombre qui doit être le plus petit possible eu égard aux cardinaux des sous-ensembles de sujets vérifiant respectivement a et b (voir « *L'analyse statistique implicative, Méthode exploratoire et confirmatoire à la recherche de causalités »* [2013]) . Nous nous intéresserons à l'indice que retient l'ASI pour signifier ou non si la qualité de la règle est statistiquement acceptable et, qu'en conséquence, si la tendance à la règle implicative a de bonnes raisons d'être ou non retenue en étant *étonnamment* acceptable[2] dans le cadre d'une théorie.

[2] C'est aussi ce qu'affirme René Thom (« Paraboles et catastrophes », 1980, p.130) : « …le problème

Nous examinerons les propriétés de sensibilité de cet indice, dit d'implication, particulièrement aux variations des cardinaux en jeu dans des expériences d'extraction de règles de corpus où les variables sont instanciées sur l'ensemble des sujets. On s'intéressera en particulier au gradient de cet indice significatif de la « vitesse et de l'amplitude » de la croissance ou la décroissance de ces cardinaux. Nous étudierons le champ de vecteurs engendré par les observations de l'indice d'implication et le champ de gradient qui lui est associé. Cette approche diffère donc de celle adoptée par Eugen Barbu dans son mémoire de D.E.A.(Barbu E., 2003) intitulé « Hiérarchie cohésitive » qui, à travers un bootstrap appliqué à des mesures de qualité de règles d'association, a fait varier les paramètres de ces mesures dont celle qui fonde l'implication statistique.

Des modélisations différentes de l'ASI pour la recherche de règles quasi-implicatives se retrouvent sous la forme de treillis de Galois et de réseau Bayesien. Un algorithme dit « A priori » est le plus souvent à la base de la mesure de qualité de règles implicatives. Nous en avons comparé les propriétés face à l'ASI tant sur le plan de leur sensibilité aux variations des instances des variables que sur la représentation des structures des ensembles de règles obtenus. Nous y revenons rapidement dans le § 4. C'est donc plus sur ces phénomènes que nous porterons notre attention et non pas, *dans ce texte*, sur la représentation des règles et méta-règles. Nous rappelons ci-dessous, le processus probabiliste qui en permet l'extraction.

2 Rappel des fondements de l'A.S.I. (Gras, 1979)

Rappelons les fondements premiers de l'ASI dont l'objectif est de rechercher la plausibilité causale derrière une relation quasi-implicative là où le physicien philosophe Bernard d'Espagnat (B.d'Espagnat, 1981) parlerait d'**influence** en ces termes : *« les évènements A influencent appréciablement les évènements B si et seulement si la fréquence avec laquelle les évènements B ont lieu est (appréciablement) différente selon que l'on impose on non aux évènements A d'exister* » (p. 186). Reprenant cette perspective, nous mathématisons cette définition en quantifiant l'expression floue «appréciablement », non pas par une fréquence conditionnelle comme il est fait dans (Agrawal R. et al, 1993) et ses dérivés, par exemple dans les réseaux bayesiens, mais de la façon suivante.

Notons A et B les sous-ensembles respectifs de E d'individus qui vérifient respectivement les variables booléennes a et b (Fig. 1). Depuis 1979, nous avons étendu (cf. Ouvrages cités en référence), de façon appropriée, les propriétés qui suivent à des variables réelles continues ou non.

Pour une règle quelconque $a \rightarrow b$, observée dans E, l'ASI consiste à comparer le nombre de contre-exemples $n_{a \wedge \bar{b}}$ ($\bar{b}$ est la négation de b) à cette règle observés dans l'intersection $A \cap \overline{B}$ ($\overline{B}$ est le complémentaire de B dans E) avec le nombre de contre-exemples qui apparaîtraient lors d'un choix aléatoire et indépendant de deux parties X et Y de E de mêmes cardinaux respectifs que A et B (Fig. 1) (Gras, 1979 ; Lebart et al.,

n'est pas de décrire la réalité, le problème consiste bien plus à repérer en elle ce qui a de sens pour nous, ce qui est surprenant dans l'ensemble des faits. Si les faits ne nous surprennent pas, ils n'apportent aucun élément nouveau pour la compréhension de l'univers : autant donc les ignorer » et plus loin : « … ce qui n'est pas possible si l'on ne dispose pas déjà d'une théorie ».

2006). La variable aléatoire associée est notée $N_{a\wedge\bar{b}}$. Le principe fondamental que nous retenons consiste à comparer la contingence à la théorie par une méthodologie qui relève de la philosophie de l'expérience où sont premières les observations (cf. B. d'Espagnat, 1981, cité plus haut).

La qualité de la règle sera intuitivement d'autant meilleure que Prob[$N_{a\wedge\bar{b}} > n_{a\wedge\bar{b}}$] sera proche de 1 : autrement dit, dans ce cas, on y observe plus de contre-exemples dans des circonstances aléatoires que l'on en a observés dans la contingence, sous l'hypothèse a priori d'indépendances des variables a et b. Dans ce cas, le seul hasard conduit donc, en moyenne, à plus de contre-exemples que ce qui est observé. C'est ainsi que se définit l'ASI en falsifiant la relation implicative par la mesure relative de sa négation exprimée à travers ses contre-exemples.

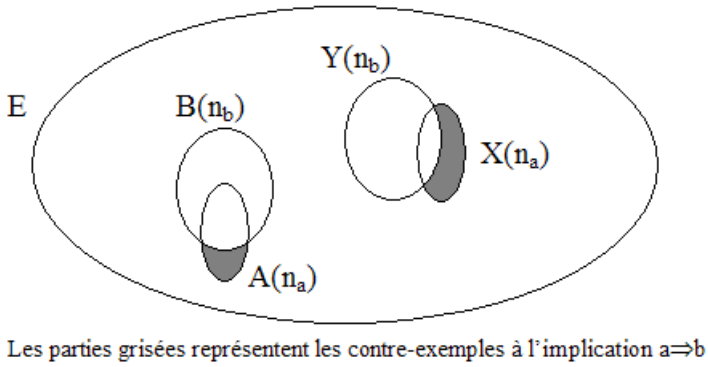


Figure 1. Représentation ensembliste

La méthode de tirage au hasard de X et Y, dans une hypothèse a priori d'indépendance de *a* et *b*, conduit à différentes options pour la loi de la variable aléatoire $N_{a\wedge\bar{b}}$. Deux modélisations de cette variable sont généralement retenues en ASI conduisant à un modèle de Poisson et à un modèle binomial (Gras et al 2009). On centre et on réduit cette variable en la variable $Q(a,\bar{b})$; l'observation contingente (empirique), sa réalisation, est $q(a,\bar{b})$. Par exemple, dans le cas du modèle de Poisson, on obtient l'indice de base pour des variables binaires :

$$Q(a,\bar{b}) = \frac{\mathrm{Card}(X\cap\bar{Y}) - \frac{n_a n_{\bar{b}}}{n}}{\sqrt{\frac{n_a n_{\bar{b}}}{n}}} \text{ alors que } q(a,\bar{b}) = \frac{n_{a\wedge\bar{b}} - \frac{n_a n_{\bar{b}}}{n}}{\sqrt{\frac{n_a n_{\bar{b}}}{n}}} \text{ est sa réalisation contingente}$$

L'intensité d'implication est alors définie par $\varphi(a,b)$ = Prob[$Q(a,\bar{b}) > q(a,\bar{b})$], dont la valeur gaussienne asymptotique, centrée et réduite est :

$$\varphi(a,b) = \frac{1}{\sqrt{2\pi}} \int_{q(a,\bar{b})}^{+\infty} e^{-\frac{t^2}{2}} dt \quad (1)$$

Cette définition (1) de l'intensité d'implication, rappelle au chercheur qu'elle ne présente un intérêt implicatif qu'à condition qu'elle soit supérieure à 0.50, c'est-à-dire

que $q(a, \bar{b})$ soit négatif. Plus $q(a, \bar{b})$ est négatif, meilleure est la qualité de la règle $a \rightarrow b$. Cependant, si $n_b = n$, nous définissons l'intensité d'implication par : $\varphi(a,b) = 0$ qui n'est pas le prolongement par continuité de l'intensité définie par (1) alors que $q(a, \bar{b})$ n'est plus défini pour cette valeur de n_b. Lever son indétermination reviendrait à lui attribuer la valeur 0, et donc $\varphi(a,b) = 0.5$, incompatible avec la sémantique de l'ASI. Cette exception tient au respect de celle-ci puisque l'implication de a sur b est alors devenue triviale, tautologique, donc non informative.

Dans cet article, nous nous intéressons aux propriétés de la fonction q(.) comme fonction des 4 occurrences observées, c'est-à-dire les variables cardinales n, n_a, n_b, $n_{a \wedge \bar{b}}$. Quel rôle jouent ces variations sur celles de l'intensité d'implication qui dépend de q(.) par intégration gaussienne ? Quelle sensibilité observe-t-on lorsque les occurrences varient quelque peu ? En quoi le gradient de q permet-il de définir un champ de gradient sur l'espace des couples de variables ? Nous sommes persuadés que la perception d'un graphe implicatif, passe certes par l'examen de sa structure qui comme l'énergie est de type « intégrale », mais aussi qu'elle se révèle à travers des procédures de dérivation. Sans mésestimer le premier que nous explorerons dans une modélisation originale mécanique de l'ASI, nous consacrerons l'étude qui vient à une approche essentiellement infinitésimale.

3 Variations de l'indice d'implication q en fonction des 4 occurrences

3.1 Stabilité de l'indice d'implication

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 (Gras et al., 2013). Mais, considérons ces variables comme variables à valeurs réelles et q comme une fonction continûment différentiable par rapport à ces variables, elles-mêmes contraintes à respecter les inégalités : $0 \leq n_a \leq n_b$ et $n_{a \wedge \bar{b}} \leq \inf\{n_a, n_b\}$ et $\sup\{n_a, n_b\} \leq n$. La fonction q définit alors un champ scalaire et vectoriel sur R^4 en tant qu'espace affine et vectoriel sur lui-même. Dans l'hypothèse vraisemblable d'une évolution d'un processus non chaotique du recueil de données, 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, au sens de la topologie de Fréchet[3], s'exprime de la façon suivante par un produit scalaire :

$$dq = \frac{\partial q}{\partial n} dn + \frac{\partial q}{\partial n_a} dn_a + \frac{\partial q}{\partial n_b} dn_b + \frac{\partial q}{\partial n_{a \wedge \bar{b}}} dn_{a \wedge \bar{b}} = \text{grad q. dM (2)}^4$$

où M est le point de coordonnées (n, n_a, n_b, $n_{a \wedge \bar{b}}$) du champ scalaire C de vecteurs,

[3] La topologie de Fréchet admet comme base de filtres des sections de N, soit des sous-ensembles de naturels de la forme {n, n+1, n+2, ….} alors que la topologie usuelle sur R admet pour filtres des intervalles de réels. Ainsi continuité et dérivabilité sont des concepts parfaitement définis et opératoires selon la topologie de Fréchet au même titre qu'ils le sont avec la topologie usuelle.

[4] Par une métaphore mécaniste, on dira que dq est le travail élémentaire de q pour un déplacement dM.

dM est le vecteur de composantes les accroissements différentiels de ces variables d'occurrences,

et grad q le vecteur de composantes les dérivées partielles de ces variables occurrences.

La différentielle de la fonction q apparaît donc comme le produit scalaire de son gradient et de l'accroissement de q sur la surface représentant les variations de la fonction $q(n, n_a, n_b, n_{a\wedge\bar{b}})$. Ainsi, le gradient de q représente ses propres variations en fonction de celles de ses composantes, les 4 cardinaux des ensembles E, A, B et $A \cap \bar{B}$. Il indique la direction et le sens de croissance ou la décroissance de q dans l'espace de dimension 4. Rappelons qu'il est porté par la normale à la surface de niveau q = cte.

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 $\Delta n_{a\wedge\bar{b}}$. On a donc :

$$\Delta q = \frac{\partial q}{\partial n}\Delta n + \frac{\partial q}{\partial n_a}\Delta n_a + \frac{\partial q}{\partial n_b}\Delta n_b + \frac{\partial q}{\partial n_{a\wedge\bar{b}}}\Delta n_{a\wedge\bar{b}} + o(\Delta q)$$

où $o(\Delta q)$ est un infiniment petit du 1er ordre.

Examinons les dérivées partielles de n_b et le nombre de contre-exemples $n_{a\wedge\bar{b}}$. On obtient :

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

$$\frac{\partial q}{\partial n_{a\wedge\bar{b}}} = \frac{1}{\sqrt{\frac{n_a n_{\bar{b}}}{n}}} = \frac{1}{\sqrt{\frac{n_a (n - n_b)}{n}}} > 0 \quad (4)$$

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 n_a varie, nous obtenons la dérivée partielle de q par rapport à n_a qui est :

$$\frac{\partial q}{\partial n_a} = -\frac{1}{2\sqrt{n_{\bar{b}}/n}} \cdot \frac{n_{a\wedge\bar{b}}}{} \cdot \left(\frac{n}{n_a}\right)^{\frac{3}{2}} - \frac{1}{2}\sqrt{\frac{n_{\bar{b}}}{n_a}} < 0 \quad (5)$$

Ainsi, pour des variations de n_a sur [0, n_b], la fonction indice d'implication $q(a,\bar{b})$ est toujours décroissante (et concave) 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 $n_a = n_b$.

Notons la dérivée partielle de q par rapport à n :

$$\frac{\partial q}{\partial n} = \frac{1}{2\sqrt{n}}\left[n_{a\wedge\bar{b}} + \frac{n_a n_{\bar{b}}}{n}\right]$$

En conséquence, si les 3 autres paramètres sont constants, l'indice d'implication décroît en $\sqrt{n}$. La qualité de l'implication n'en est donc que meilleure, propriété spécifique de l'ASI par rapport à d'autres indicateurs retenus dans la littérature (cf. Gras et Couturier, 2010). Cette propriété est en accord avec les attentes statistiques et sémantiques relatives au crédit accordé à la fréquence des observations. Les dérivées partielles de q (au moins l'une d'entre elles) étant non linéaires selon les paramètres variables en jeu, on a affaire à un système dynamique non linéaire avec toutes les conséquences épistémologiques que nous envisagerons par ailleurs.

3.2 Exemple numérique

Dans une première expérience, on observe les occurrences :

$n = 100$, $n_a = 20$, $n_b = 40$ (d'où $n_{\bar{b}} = 60$), $n_{a\wedge\bar{b}} = 4$.

L'application de la formule (1) donne $q(a,\bar{b}) = -2{,}309$

Dans une 2ème expérience, n et n_a sont inchangées mais les occurrences des b et des contre-exemples $n_{a\wedge\bar{b}}$ s'accroissent d'une unité.

Au point initial de l'espace des 4 variables, les dérivées partielles qui seules nous intéressent (selon n_b et $n_{a\wedge\bar{b}}$) ont respectivement pour valeurs en appliquant les formules (3) et (4)) : $\frac{\partial q}{\partial n_b} = 0{,}0385$ et $\frac{\partial q}{\partial n_{a\wedge\bar{b}}} = 0{,}2887$

Comme Δn_b, $\Delta n_{\bar{b}}$ et $\Delta n_{a\wedge\bar{b}}$ sont égaux à 1, -1 et 1, l, alors Δq est égal à :

$0{,}0385 + 0.2887 + o(\Delta q) = 0{,}3272 + o(\Delta q)$ et la valeur approchée de q lors de la deuxième expérience est $-2{,}309 + 0{,}2887 + o(\Delta q) = -1{,}982 + o(\Delta q)$ en utilisant le développement de q au premier ordre (formule (2)).

Or le calcul du nouvel indice d'implication q au point de la 2ème expérience est, par l'usage de (1) : -1,9795, valeur bien approchée par le développement de q.

3.3 Une première relation différentielle de φ en tant que fonction de la fonction q.

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 :

$$\frac{d\varphi}{dq} = -\frac{1}{\sqrt{2\pi}} e^{\frac{-q^2}{2}} < 0 \quad (6)$$

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. Puisque la dérivée de φ par rapport à q est toujours négative, la fonction φ est décroissante.

3.4 Exemple numérique

Reprenant les valeurs des occurrences observées dans les 2 expériences évoquées plus haut, on trouve pour q = -2,309, la valeur de l'intensité d'implication φ(q) est égale à 0,992. Appliquant la formule (6), la dérivée de φ par rapport à q est : - 0 ,02775 et l'accroissement négatif de l'intensité est alors : - 0,02775 . Δq = - 0,02775. 0,3272. L'intensité approchée au premier ordre est donc : 0, 992 - Δq soit 0,983. Or le calcul réel de cette intensité est, pour q= -1,9795 , φ(q) = 0, 976

4 Examen d'autres indices

Contrairement à l'indice de base q et l'intensité d'implication qui mesure la qualité à travers une probabilité (cf. définition 3), les autres indices les plus courants se veulent eux-mêmes directement des mesures de qualité. Nous examinerons leurs sensibilités respectives aux variations des paramètres retenus dans la définition de ces indices. Nous conservons les notations adoptées au paragraphe 2 et choisissons des indices qui sont rappelés dans (Gras et al, 2004), (Lenca et al., 2005) et (Gras et Couturier, 2010) .

4.1 L'indice de Loevinger

C'est un « ancêtre » des indices d'implication (Loevinger, 1947). Cet indice, noté H(a,b) varie de 1 à - ∞. Il est défini par :

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

Sa dérivée partielle par rapport à la variable nombre de contre-exemples est donc :

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

Ainsi l'indice d'implication est toujours décroissant avec $n_{a\wedge\bar{b}}$. S'il est "proche" de 1, l'implication est "presque" satisfaite. Mais cet indice présente l'inconvénient, 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 $A \cap \bar{B}$.

4.2 L'indice Lift

Il s'exprime par : $l = \frac{n.n_{a\wedge b}}{n_a n_b}$. Cette expression, linéaire par rapport aux exemples, 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 :

$$\frac{\partial l}{\partial n_{a\wedge\bar{b}}} = -\frac{1}{n_a.n_b}$$

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 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 $n_{a\wedge\bar{b}}$.

4.3 L'indice MC

Il s'exprime ainsi : $m = \frac{n_a - n_{a\wedge\bar{b}}}{n_b.n_{a\wedge\bar{b}}} n_{\bar{b}}$. Remarquons qu'en étant indépendant de n, il n'a pas de sens statistique aussi intéressant.

Sa dérivé 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 décroît donc lorsque $n_{a\wedge\bar{b}}$ croît et la vitesse de décroissance est même plus rapide qu'avec le Lift et qu'avec l'indice q gaussien basique pour calculer l'intensité d'implication. Il ne résiste pas à l'instabilité du nombre de contre-exemples.

4.4 La confiance

Cet indice est le plus connu et 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.

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

La première forme, linéaire par rapport aux exemples, indépendante de n_b, 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 à travers la dérivée partielle :

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

Par conséquent, la confiance 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 ce nombre, 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 :

$$\begin{pmatrix} -\frac{1}{n_a} \\ \frac{n_{a\wedge\bar{b}}}{{n_a}^2} \end{pmatrix}$$

Ceci peut apparaître comme une restriction du rôle des paramètres dans l'expression de la sensibilité de l'indice.

5 Champ de gradient, champ implicatif

5.1 Existence d'un champ de gradient

Nous revenons à l'indice implicatif $q(a,\bar{b})$. Considérons l'espace E de dimension 4 où les points M ont pour coordonnées les paramètres relatifs aux variables binaires a et b, soit $(n, n_a, n_{b,} n_{a\wedge\bar{b}})$. $q(a,\bar{b})$ définit donc un champ scalaire en tant qu'application de R^4 dans R (plongement de N^4 dans R^4).

Pour que le vecteur grad q de composantes les dérivées partielles de q par rapport aux variables n, n_a, $n_{b,}$ $n_{a\wedge\bar{b}}$ définisse un champ de gradient - champ vectoriel particulier que nous appellerons aussi champ implicatif- il doit respecter le critère de Schwartz d'une différentielle totale exacte à savoir et par exemple :

$$\frac{\delta}{\delta n_{a\wedge\bar{b}}}\left(\frac{\delta q}{\delta n_b}\right) = \frac{\delta}{\delta n_b}\left(\frac{\delta q}{\delta n_{a\wedge\bar{b}}}\right)$$

et idem pour les autres variables prises deux à deux. Or on a bien par les formules (3) et (4) :

$$\frac{\delta}{\delta n_{a\wedge\bar{b}}}\left(\frac{\delta q}{\delta n_b}\right) = \frac{1}{2}\left(\frac{n_a}{n}\right)^{-1/2}\left(\frac{n_{\bar{b}}}{n}\right)^{-3/2} = \frac{\delta}{n_b}\left(\frac{\delta q}{\delta n_{a\wedge\bar{b}}}\right)$$

Ainsi, au champ de vecteurs $C = (n, n_a, n_{b,} n_{a\wedge\bar{b}})$ de E, dont nous préciserons la nature, correspond un champ de gradient G qui est dit dérivé du **potentiel** q. Le gradient grad q est donc le vecteur qui représente la variation spatiale de l'intensité du champ. Il est dirigé des faibles valeurs du champ aux valeurs plus élevées. En suivant le gradient en chaque point, on suit l'augmentation de l'intensité d'implication du champ dans l'espace et, en quelque sorte la vitesse avec laquelle elle change sous l'effet de la variation d'un ou plusieurs paramètres.

Par exemple, si l'on fixe 3 des paramètres n, n_a, $n_{b,}$ $n_{a\wedge\bar{b}}$ donnés par la réalisation du couple (a,b), le gradient est un vecteur dont la direction indique la croissance ou la

décroissance de q, donc la décroissance ou la croissance de |q| et par suite de φ en fonction des variations du 4ème paramètre. Nous l'avions indiqué plus haut en interprétant la formule (5).

5.2 Lignes de niveau ou équipotentielles

Une ligne ou une surface équipotentielle (ou de niveau) dans le champ C est une courbe de E le long de laquelle ou sur laquelle un point variable M conserve la même valeur du potentiel q (par ex. lignes isothermiques sur le globe ou lignes de niveau d'une carte IGN). L'équation de cette surface[5] est, bien entendu :

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

Par suite, sur une telle courbe, le produit scalaire grad q. dM est nul. Ce qui s'interprète comme indiquant l'orthogonalité du gradient avec la tangente ou l'hyperplan tangent à la courbe, c'est-à-dire avec la ligne ou la surface équipotentielle. Dans une interprétation cinématique de notre problème, la vitesse de parcours de M sur la surface équipotentielle est orthogonale au gradient en M.

A titre d'illustration, relativement à un potentiel F dépendant de 2 variables seulement, la figure ci-dessous par exemple montre la direction orthogonale du gradient par rapport aux différentes surfaces équipotentielles le long desquelles le potentiel F ne varie pas mais passe de F=7 à F= 10.

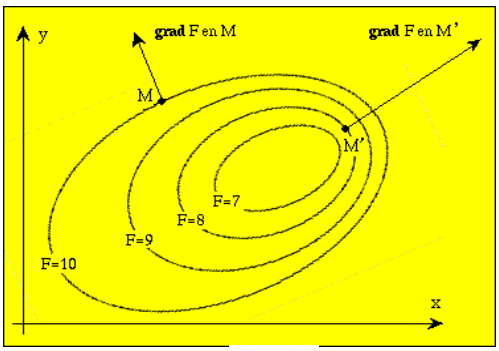


Figure 2

Il est possible dans le cas du potentiel q, de construire des surfaces équipotentielles comme ci-dessus (à deux dimensions pour la facilité de représentation). On peut comprendre que plus le champ est intense plus les surfaces sont serrées. Pour une valeur de q donnée, dans ce cas, on fixe 3 variables, par exemple n, n_a, n_b et une valeur de q compatibles avec les contraintes du champ. Soit : $n = 10^4$; $n_a = 1600 \leq n_b = 3600$ et q =

[5] En géométrie différentielle, on dirait que cette surface est une variété (quasi)différentiable à bord, compacte, homéomorphe au pavé fermé des intervalles de variations des 4 paramètres. Notons que le point dont la composante n_b est égale à n (donc $n_{a\wedge\bar{b}} = 0$) est un point singulier (« catastrophique » au sens de René Thom) de la surface et q, le potentiel, n'est pas différentiable en ce point. Partout ailleurs, la surface est différentiable, les points sont tous réguliers. Si le temps, par exemple, paramètre les observations du processus dont (n, n_a, n_b, $n_{a\wedge\bar{b}}$) est une réalisation, à chaque instant correspond une fibre morphologique du processus représentée par une telle surface dans l'espace-temps..

-2 soit |q| = 2. On trouve alors $n_{a\wedge\bar{b}}$ = 528 en utilisant la formule (1). Mais les points (10^4, 1600, 5100, 728) et (100, 25, 64, 3) appartiennent également à cette surface et la même courbe équipotentielle. Le point (10^4, 1600, 3600, 928) appartient à la courbe équipotentielle q=-3). En fait, sur toute cette surface, on obtient une sorte d'homéostasie de l'intensité d'implication. L'expression de la fonction q de la variable $n_{a\wedge\bar{b}}$ permet de montrer qu'elle est convexe. Cette propriété prouve que le segment des points $t.M_1 + (1-t).M_2$, pour $t \in [0,1]$ qui joint deux points M_1 et M_2 de la même ligne équipotentielle est entièrement contenu dans sa convexité. La figure ci-dessous représente, dans le champ implicatif deux surfaces équipotentielles voisines Σ_1 et Σ_2 correspondant à deux valeurs du potentiel q_1 et q_2. Au point M_1 le champ scalaire prend donc la valeur q_1. M_2 est l'intersection de la normale issue de M_1 avec Σ_2. Etant donné la direction du vecteur normal $\vec{n}$, la différence $\Delta = q_2 - q_1$, variation du champ quand on passe de Σ_1à Σ_2 est alors égale à l'opposé de la norme du gradient de q en M_1 soit $\frac{\partial q}{\partial n}$, si n_a, n_b et $n_{a\wedge\bar{b}}$ sont fixés .

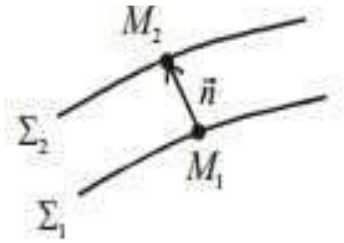


Figure 3

Ainsi, l'espace E peut être comme feuilleté par des surfaces équipotentielles correspondant à des valeurs successives de q relativement aux cardinaux (n, n_a, n_b, $n_{a\wedge\bar{b}}$) que l'on ferait varier. Cette situation correspond à celle qui est envisagée dans la modélisation de l'ASI. Fixant n, n_a et n_b, on considère les ensembles aléatoires X et Y de mêmes cardinaux que A (n_a) et B (n_b) et dont le cardinal de $X \cap \bar{Y}$ suit une loi de Poisson ou une loi binomiale, suivant le choix du modèle. Les différents champs de gradient, véritables « lignes de force », qui leur sont associés sont orthogonaux aux surfaces définies par les valeurs correspondantes de Q. Ceci nous évoque, dans le cadre théorique du potentiel, la métaphore prémonitoire de « flux implicatif » que nous avons exprimée dans (Gras et al. 1996). Derrière cette notion nous pouvons imaginer un transport d'information d'intensité variable dans un univers causal. Nous illustrons cette métaphore avec l'étude des propriétés du cône implicatif à deux nappes (cf. Lahanier-Reuter et al, en publication ASI 8). De plus et intuitivement, l'implication a=>b est d'autant de bonne qualité que la surface équipotentielle C de la contingence recouvre des surfaces équipotentielles aléatoires dépendant de la variable aléatoire $Q(a, \bar{b})$.

Rappelons la relation qui unit le potentiel q à l'intensité d'implication $\varphi(a,b)$ définie par:

$$\varphi(a,b)= \frac{1}{\sqrt{2\pi}} \int_{q(a,\bar{b})}^{\infty} e^{-\frac{t^2}{2}} dt$$

Remarque 1

On constate que l'intensité est également invariante sur toute surface équipotentielle de ses propres variations. Les portions de surfaces engendrées par q et par φ sont même en correspondance biunivoque. En termes intuitifs, on peut affirmer que lorsque l'une « enfle » l'autre se « dégonfle.

Remarque 2

Notons une fois encore une particularité de l'intensité d'implication. Alors que les surfaces engendrées par les variations des 4 paramètres des données ne sont pas invariantes par une même dilatation des paramètres, celles associées aux indices citées dans le § 4 sont invariantes et présentent une même forme géométrique indifférenciée.

6 Quelques simulations

6.1 Surfaces équipotentielles

Afin de soutenir l'intuition et l'imagination, nous avons effectué quelques simulations de surfaces équipotentielles. Pour obtenir les figures suivantes avec Matlab, nous avons fixé n = 100, utilisé 3 valeurs de $n_{a\wedge\bar{b}}$: 5, 10 et 20 et avons fait varier n_a selon le 1er axe (dit des « x ») de $n_{a\wedge\bar{b}}$ à n et n_b selon le 2ème axe (dit des « y ») de 1 à $n - n_{a\wedge\bar{b}}$. Les courbes représentent donc les variations de q, indice d'implication selon l'axe vertical (dit des « z »). [6]

Sur ce type de figure en 3D, les valeurs affichées sur les axes, pour des commodités de représentation, peuvent être trompeuses. Sur la première par exemple les valeurs de n_a vont de 0 à 100 alors que n_a n'est bien entendu défini que de 5 à 100 puisque $n_{a\wedge\bar{b}} = 5$. De même, l'examen utile de la qualité de la règle a => b exige $n_a \leq n_b$.

Les couleurs sont choisies pour signifier une certaine qualité de la règle :

- couleurs chaudes (rouge) pour les situations intéressantes où l'implication est satisfaite et d'autant plus que q est négatif,
- couleurs froides (bleu) où c'est le rejet de la règle a => b, voire la validité de la réciproque qui sont de plus en plus satisfaits et soulignés par l'intensité bleue des valeurs de q.

[6] Faisons une place à l'imagination : dans le cas présent où les variables et les paramètres associés ne prennent que des valeurs entières les points représentant les valeurs de q apparaissent comme des « petits pois » sur la « robe » équipotentielle.

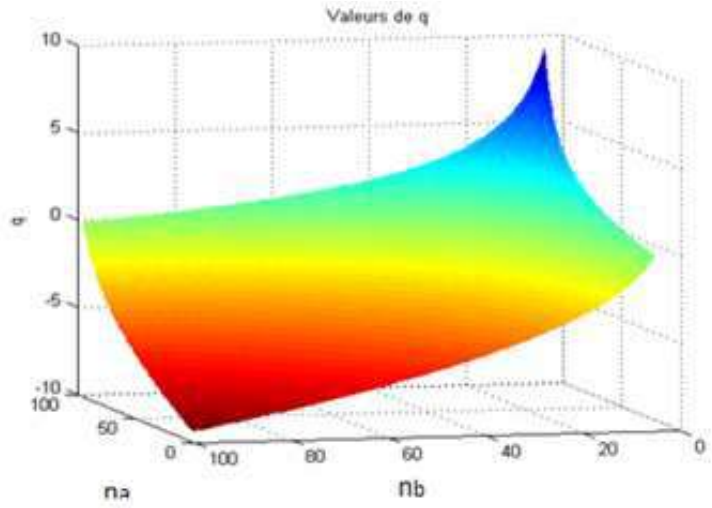


Figure 4 **Courbe 1** : $n_{a\wedge\bar{b}} = 5$

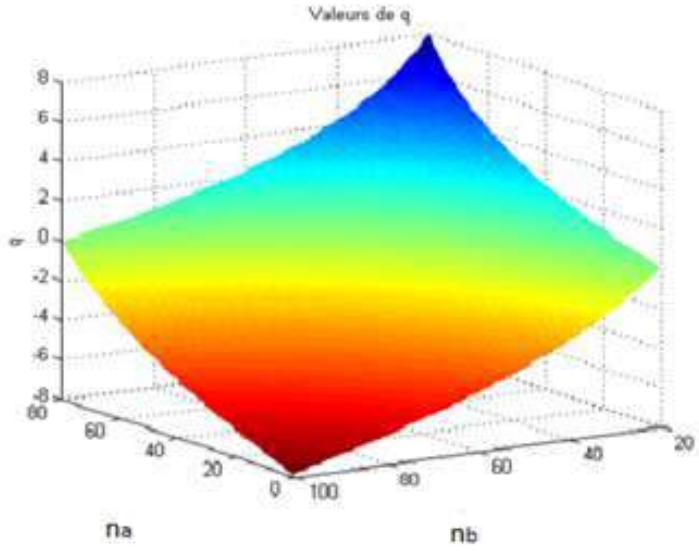


Figure 5 **Courbe 2** : $n_{a\wedge\bar{b}} = 20$

Remarquons l'étendue moins importante de la zone où l'implication est acceptable et, par contre, celle des rejets qui s'est amplifiée ainsi qu'un étalement plus marqué des valeurs indécises pour la règle.

6.2 Courbes équipotentielles de q pour différentes valeurs des contre-exemples

Courbes pour $n_{a\wedge\bar{b}} = 5$

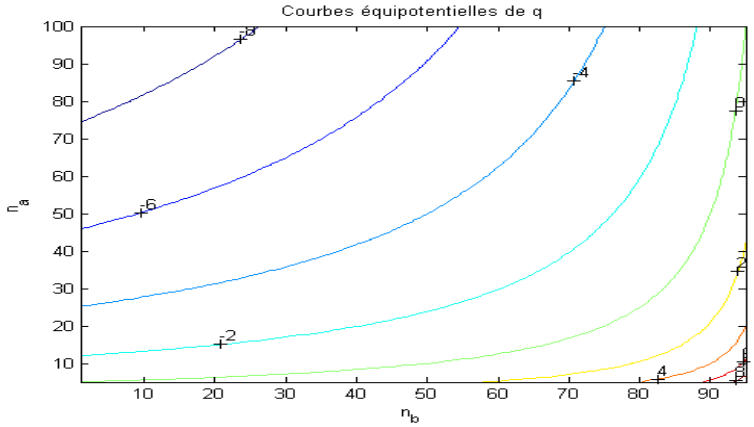


Figure 6

Ces courbes (n_b en abscisse et n_a en ordonnée) sont des projections planes des surfaces précédentes et limitées aux valeurs entières de q de -8 à 0 dans la figure 5 et de -6 à 0 pour la figure 6.

L'indice q est invariant tout le long de chaque arc de courbe. Par exemple, on trouve pour $n_a = 45$ et $n_b = 45$ (soit $n_{\bar{b}} = 55$), q = -3.9 (approximativement courbe q = -4). Mais aussi pour $n_a = 25$ et $n_b = 55$ (soit $n_{\bar{b}} = 45$), q = - 1.86.01$\cong -2$.

On peut remarquer que, pour des valeurs de n_a et de $n_{\bar{b}}$ assez élevées (donc n_b petit) les valeurs prises par q restent fortes plus « longtemps » que pour des valeurs opposées, ce que montre le resserrement des courbes pour des valeurs de q faibles conduisant à des incertitudes au sujet de la règle en raison de la décroissance de q.

Comparaison pour $n_{a\wedge\bar{b}} = 5$ (courbes rouges) et $n_{a\wedge\bar{b}} = 10$ (courbes vertes)

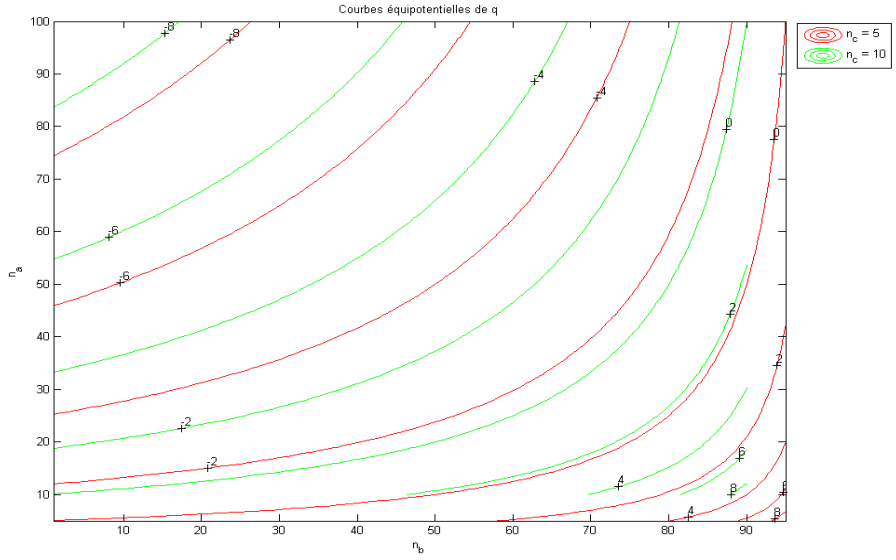


Figure 8

On notera le décalage des lignes équipotentielles rouges par rapport aux vertes où, pour une même valeur du couple (n_a, n_b), q est plus négatif sur les rouges que sur les vertes, donc accompagné d'une intensité d'implication de meilleure qualité.

Mise en évidence des vecteurs « gradient » du champ implicatif.

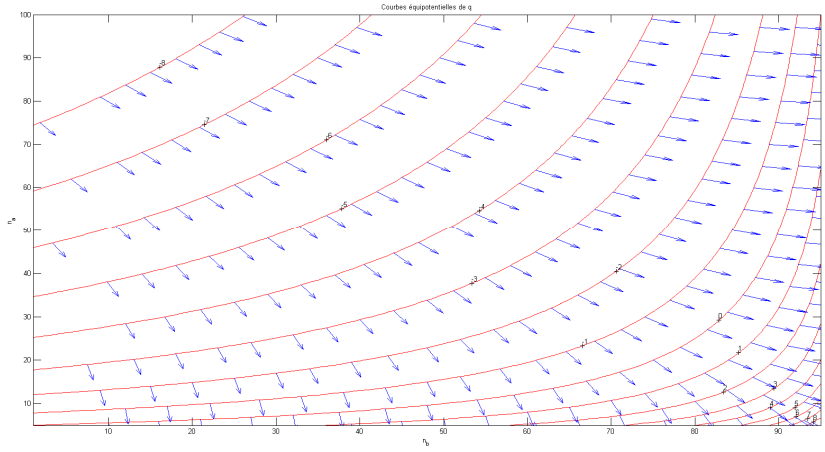


Figure 9

Si l'on fait maintenant apparaître les vecteurs « gradient » du champ implicatif normalisés et normaux aux lignes équipotentielles, on constate, métaphoriquement, une sorte de « respiration » du phénomène lié aux changements de valeurs des paramètres de la contingence. L'expansion de q, donc de l'intensité d'implication, est bien appréhendée par le sens du vecteur comme le laissaient prévoir les signes des dérivées partielles de q par rapport à n_a (négatif) et à n_b (positif). Plus les paramètres croissent, pour un même nombre de contre-exemples, meilleures sont l'intensité et par conséquent notre confiance en la règle. Notons aussi que l'instabilité de q (courbes qui se « resserrent ») va croissante de gauche à droite comme on le remarquait dans le 6.2.

7 Conclusion

Sur la base des concepts qui ont conduit à la modélisation de la notion de quasi-implication en ASI, nous avons étudié l'indice primitif gaussien qui permet de quantifier la qualité de cette notion et, de ce fait, de qualifier notre agrément ou notre rejet des règles extraites d'un corpus de données binaires (et autres par extension). Notre étude s'est bien précisément centrée sur la sensibilité de l'indice d'implication aux variations des paramètres en jeu dans l'énonciation d'une règle. C'est ainsi que nous avons focalisé l'étude sur le champ scalaire défini par cet indice, puis sur un concept d'analyse, le gradient d'une telle fonction. Nous avons alors examiné les structurations de l'espace vectoriel donné par les paramètres, puis dégagé la notion de champ de gradient (dénommé ici champ implicatif) pour illustrer géométriquement, par des surfaces et des lignes équipotentielles, la sensibilité de l'indice d'implication. Ce pas-de-côté géométrique, par rapport aux approches statistiques et analytiques originelles, restitue à l'A.S.I. une vision dynamique du concept qui en enrichit la représentation mentale. A sa charge d'entretenir nos rêves car selon une formule de G. Bachelard *: « On ne peut étudier que ce que l'on a d'abord rêvé ».*

Références

[1.] Agrawal R., Imielinsky T. et Swami A,(1993), Mining association rules between sets of items in large databases, *Proc. of the ACM SIGMOD'93*, 207-216

[2.] Barbu E. (2003), *Hiérarchie cohésitive (ou implicative*), Mémoire de D.E.A. Extraction des Connaissances à partir des Données, Ecole Doctorale Informatique et Information pour la Société, Equipe COD, LINA, Université de Nantes

[3.] Bernard J.-M. et Poitrenaud S, (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.

[4.] Cadot M., (2009), Graphe de règles d'implication statistique pour le raisonnement courant. Comparaison avec les réseaux bayesiens et les treillis de Galois, *Analyse Statistique Implicative, Une méthode d'analyse de données pour la recherche de causalités, sous la direction de Régis Gras, réd, invités R. Gras, J.C. Régnier, F. Guillet, Cépaduès Ed. Toulouse, p.223-250*

[5.] David J, Guillet F., Gras R. and Briand H. (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.*

[6.] D'Espagnat B. (1981), *A la recherche du réel*, Gauthier-Villars, Paris.

[7.] Fayyad U., Piatetsky-Shapiro G. and Smyth P. 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.

[8.] 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.

[9.] Gras R., Ag Almouloud S., Bailleul M., Larher A., Polo M., Ratsimba-Rajohn et Totohasina A (1996), *L'implication Statistique*, Collection Associée à Recherches en Didactique des Mathématiques, Grenoble : La Pensée Sauvage.

[10.] Gras R., Kuntz P. et Briand H., (2001b), Les fondements de l'analyse statistique implicative et quelques prolongements pour la fouille de données, *Mathématiques et Sciences Humaines*, n° 154-155, 9-29.

[11.] Gras R., Couturier R.., Blanchard J., Briand H., Kuntz P., Peter P., (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, Cépaduès –Editions*, 3-32.

[12.] Gras R., David J., Guillet F., Briand H. (2007). 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

[13.] Gras R., Couturier R., (2010) Spécificités de l'Analyse Statistique Implicative (A.S.I.) par rapport à d'autres mesures de qualité de règles d'association, *Quaderni di Ricerca in Didattica - GRIM (ISSN on-line 1592-4424),* Eds : J.C. Régnier, R.Gras, F.Spagnolo, B. Di Paola, Université de Palerme, *p.19-57*

[14.] Lenca P., Vaillant B., Meyer P., et Lallich S. (2007), Association Rule Interestingness Measures : Experimental and Theorical Studies, *F.Guillet and H. J.Hamilton eds, Studies in Computational Intelligence 43, Springer*, p. 51-76.

[15.] Loevinger J. (1947), "A systematical approach to the construction and evaluation of tests of ability", *Psychological Monographs*, n° 61, 1947, p. 1-49

[16.] Ritschard G., Marcellin S., Zighed D.A. (2009), Arbre de décision pour données déséquilibrées : sur la complémentarité de l'intensité d'implication et de l'entropie décentrée, *Une méthode d'analyse de données pour la recherche de causalités, sous la direction de Régis Gras, réd, invités R. Gras, J.C. Régnier, F. Guillet, Cépaduès Ed. Toulouse p.207-21.*

[17.] Sève L. (2005), *Emergence, complexité et dialectique*, Odile Jacob, Paris.

[18.] Thom, R (1980). *Paraboles et catastrophes*, Flammarion, Paris.

Ouvrages de référence

L'implication statistique. Nouvelle méthode exploratoire de donnée, sous la direction de R.Gras, et la collaboration de S. Ag Almouloud, M. Bailleul, A. Larher, M. Polo, H. Ratsimba-Rajohn, A.Totohasina, La Pensée Sauvage, Grenoble (1996)

Mesures de Qualité pour la Fouille de Données, H.Briand, M.Sebag, R.Gras er F.Guillet eds, RNTI-E-1, Cépaduès, 2004

Quality Measures in Data Mining , F.Guillet et H.Hamilton eds, Springer, 2007,

Statistical Implicative Analysis, Theory and Applications, R.Gras, E. Suzuki, F. Guillet, F. Spagnolo, eds, Springer, 2008.

Analyse Statistique implicative.Une méthode d'analyse de données pour la recherche de causalités, sous la direction de Régis Gras, réd. invités R. Gras, J.C. Régnier, F. Guillet, Cépaduès Ed. Toulouse, 2009.

Teoria y Aplicaciones del Analisis Estadistico Implicativo, Eds : P.Orus, L.Zemora, P.Gregori, Universitat Jaume-1, Castellon (Espagne), ISBN : 978-84-692-3925-4, 2009..

L'Analyse Statistique Implicative : de l'exploratoire au confirmatoire. Eds : J.C. Régnier, Marc Bailleul, Régis Gras, Université de Caen, ISBN : 978-2-7466-5256-9, 2012

L'analyse statistique implicative, Méthode exploratoire et confirmatoire à la recherche de causalités, sous la direction de Gras R., eds Gras R., Régnier J.-C., Marinica C., Guillet F., Cépaduès Editions, 522 pages, ISBN 978.2.36493.056.8, 2013.

UN MARIAGE ARRANGE ENTRE L'IMPLICATION ET LA CONFIANCE ?

Régis GRAS[1], Raphaël COUTURIER[2], Pablo GREGORI[3]

AN ARRANGED MARRIAGE BETWEEN IMPLICATION AND CONFIDENCE

RESUME

La plupart des indices d'association entre variables binaires utilisent la fréquence conditionnelle qu'ils appellent confiance ou expressions algébriques d'instanciations pour décider d'une liaison entre deux variables et en apprécier la qualité. En Analyse Statistique Implicative, une autre mesure, l'intensité d'implication, vise le même objectif en se limitant à l'implication tout en s'appuyant plutôt sur la probabilité d'apparition des contre-exemples à ce type de liaison. Dans cet article nous comparons ces deux mesures en montrant qu'elles sont étrangères mais possèdent des relations analytiques intéressantes. De ce fait, nous concevons et expérimentons une nouvelle mesure de qualité d'implication en deux approches qui associe confiance et intensité. Nous montrons l'intérêt présenté par cette combinaison pour intégrer la contraposée de l'implication, condition nécessaire pour faire jouer à ce nouvel indice une fonction d'analyse causale.

Mots-clés : *confiance, fréquence conditionnelle, intensité d'implication, intensité entropique, implifiance, règle, quasi-règle, contraposée*

ABSTRACT

Most of the indexes used with association rules are based on the conditional probability, also called confidence, to measure the quality of a rule. With Statistical Implicative Analysis, another measure, the implication intensity, focuses on the same objective using only the implication. It is built with the probability of appearance of counter-examples of a rule. In this paper, we compare those two measures, showing they are different but also have interesting analytical properties. Hence, a new measure based on those two approaches, confidence and implication is proposed. The interest of this combination is to integrate the contrapositive of the implication, in order to give this new index a causal property.

Keywords : confidence, conditional probability, implication intensity, entropic intensity, implifiance (implidence), rule, quasi-rule, contrapositive

1 Introduction

1.1 Motivation

Nous nous plaçons dans le cadre du croisement d'un ensemble de sujets et d'un ensemble de variables binaires selon lesquelles les sujets prennent leurs valeurs. De nombreux auteurs (voir Références) ont établi des indices qui permettent de mettre en

[1] École Polytechnique de l'Université de Nantes, Équipe DUKE Data User Knowledge, Laboratoire d'Informatique de Nantes-Atlantique (LINA), UMR 6241, : regisgra@club-internet.fr

[2] FEMTO-ST, Université de Franche-Comté, Belfort, raphael.couturier@univ-fcomte.fr

[3] Dept. Matemàtiques, Universitat Jaume I, Campus del Riu Sec, Castelló E-12071 (SPAIN), gregori@mat.uji.es , http://www3.uji.es/~gregori

évidence des relations implicatives entre des variables binaires telles que, par exemple, a et b. Parmi ces indices et très souvent, ils font référence à la confiance, fréquence conditionnelle de b sachant a et établissent leurs indices d'association sur celle-ci. Dans cet article, nous établirons une relation entre l'intensité d'implication et la confiance, relation qui soulignera à la fois la différence entre ces deux indicateurs et leur éventuelle covariation. *De plus, pour des raisons sémantico-statistiques, nous les combinerons pour construire un nouvel indice composite d'implication statistique qui aura la vertu d'associer la tendance inclusive donnée par la confiance et l'étonnement statistique par l'intensité d'implication.*

Nous présentons également une nouvelle approche de la modélisation en A.S.I. pour des variables binaires prenant en compte autant l'implication directe que sa contraposée. Elle vise à se substituer à la modélisation entropique jusqu'alors utilisée et qui présente un caractère jugé trop ad-hoc par les familiers de l'A.S.I. Elle va donc être construite *contre une connaissance antérieure* comme le dit G. Bachelard dans (Bachelard G. 1967). Cependant, cette précédente modélisation était loin de déplaire aux utilisateurs qui en appréciaient la capacité à accepter plus facilement la grande taille de l'échantillon des sujets considéré. D'où son intérêt pour ce que l'on appelle les « big data ». De plus, prenant en compte la contraposée, elle semble remplir le rôle d'extracteur de relations causales, ce que ne permet pas ou moins finement l'implication dite classique mesurée par l'intensité d'implication. En effet, celle-ci conduit à la même valeur d'intensité autant pour l'implication directe que pour sa contraposée. Elle ne permet pas ainsi d'intégrer l'information nuancée apportée par la contraposée. Rappelons par contre qu'en cas d'implication logique stricte les deux formes implicatives sont équivalentes. Ce qui n'est pas le cas de la quasi-implication, objet central de nos recherches. Ainsi, cette nouvelle modélisation de l'implication entre deux variables binaires, tout en intégrant à la fois l'implication directe et sa contraposée, mais également la confiance, va adopter une méthodologie comparable à celle utilisée dans le cas de l'implication mesurée par l'intensité d'implication définie par R. Gras (Gras, 1979) à la base des premiers fondements de l'A.S.I.

1.2 Remarque épistémologique importante

Il pourrait être objecté à l'égard des auteurs de cet article que ce nouveau concept soit seulement le fruit d'un mariage arrangé entre deux partenaires étrangers l'un à l'autre, regardant dans des directions différentes. Or, l'un et l'autre alimentent l'objectif que nous nous sommes donné sachant que nous traitons essentiellement des problèmes qui relèvent du réel comme en sciences humaines ou relativement à des phénomènes scientifiques non entièrement modélisés (médecine, biologie, etc.) : rendre compte de l'implication en serrant au plus près la notion logique de celle-ci sachant que leurs célibats respectifs ne le permettent pas. Nous prétendons, en plagiant G. Vergnaud (Vergnaud, 2007) s'exprimant au sujet de l'esthétisme, qu'il n'y a pas d'analyse de données sans **confiance**. Mais il n'y pas non plus d'analyse de données sans **surprise**[4],

[4] C'est aussi ce qu'affirme René Thom (« Paraboles et catastrophes », 1980, p.130) : « …le problème n'est pas de décrire la réalité, le problème consiste bien plus à repérer en elle ce qui a de sens pour nous, ce qui est surprenant dans l'ensemble des faits. Si les faits ne nous surprennent pas, ils n'apportent aucun élément nouveau pour la compréhension de l'univers : autant donc les ignorer » et plus loin : « … ce qui

ni sans **correction d'échelle** comme il est fait en analyse en composants principales, par exemple, où une division par $\sqrt{n}$ s'impose dans la détermination de la distance entre sujets. Accoupler confiance et surprise n'a rien de plus artificiel que celui auquel sont soumis les physiciens de l'acoustique, de la thermodynamique, de l'électricité, de la relativité, etc.

Par exemple, sur le problème de chute des corps, l'observateur relève expérimentalement l'influence de certaines variables qui expriment que la résistance de l'air est proportionnelle au carré de la vitesse V, proportionnelle à la surface S du maître-couple du solide, à la masse spécifique r_0 de l'air, à un coefficient de forme C. Tout ceci est résumé par la formule : $R=kCr_0SV^2$. C'est une relation construite sur la base de mesures empiriques qui conduisent à la satisfaire. Il en est de même de la formule expérimentale des gaz parfaits : pV=Cte et toutes les règles énoncées sous le terme de principe[5].

Bien entendu, d'autres phénomènes physiques seront plongés par essence dans un modèle mathématique déduit (par ex. les phénomènes vibratoires). Et d'autres seront extraits de l'expérimentation et de la mise en évidence des variables intervenantes. Nous pourrons dire que notre modèle est bon lorsque les résultats expérimentaux que nous mènerons avec satisferont à la fois l'explicabilité, le bon sens et l'observation mais aussi auront une capacité prédictive. D'autres modèles concurrents du concept implicatif de notre étude présente existent et nous avons mis en évidence leurs adéquation ou leurs limites vis-à-vis des variables requises ou de leur pondération (Gras et Couturier, 2010). De toute façon, quel qu'il soit, le modèle utilisé est bon s'il rend compte au mieux de la réalité jusqu'à sa substitution éventuelle par un autre plus performant. Or c'est cette performance que nous voulons atteindre (cf. aussi note 2).

2 Relation entre confiance et intensité d'implication.

2.1 Rappels sur l'intensité d'implication

Relativement aux cardinaux de E (soit n), espace des sujets, de A (soit n_a), sujets satisfaisant a et de B (soit n_b), sujets satisfaisant b, c'est le poids des contre-exemples (soit $n_{a\wedge\bar{b}}$), i.e. des sujets satisfaisant a et non 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.

Pour formaliser cette quasi-règle, nous formulons l'hypothèse que a et b sont indépendantes. Puis nous considérons, comme le fait I.C. Lerman dans (Lerman, 1981) 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

---

n'est pas possible si l'on ne dispose pas déjà d'une théorie ».

[5] « On nomme **principe physique** une loi physique apparente, qu'aucune expérience n'a invalidée jusque là bien qu'elle n'ait pas été démontrée, et joue un rôle voisin de celui d'un postulat en mathématiques ». (Wikipedia). « Parfois un principe peut être démontré à partir d'un ou plusieurs autres, à charge au physicien de choisir le principe de base pour ses raisonnements : par exemple, le principe de moindre action est équivalent au principe fondamental de la dynamique associé au principe de d'Alembert » (Wikipedia).

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 $n_{\overline{b}} = n\text{-}n_b$. Soit α un réel quelconque de l'intervalle [0,1].

Définition 1: la quasi-règle $a \Rightarrow b$ est *admissible avec l'intensité* 1-α si et seulement si $\Pr\left[\text{Card}(X \cap \overline{Y}) \leq \text{card}(A \cap \overline{B})\right] \leq \alpha$.

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

On démontre (Gras R, 1995, puis autre preuve en 2009) que, pour une certaine modélisation de tirage de X et de Y, $\text{Card}(X \cap \overline{Y})$ suit une loi de Poisson de paramètre $\lambda = \frac{n_a n_{\overline{b}}}{n}$. L'intensité d'implication est une valeur probabiliste, contrairement aux indices implicatifs les plus usités, qui fonde la décision de retenir ou non une relation de quasi-implication entre les variables binaires a et b. Intuitivement, elle représente une sorte d'étonnement statistique (une *surprise* ?) que le nombre de contre-exemples à $a \Rightarrow b$ soit petit alors qu'elles sont supposées indépendantes.

2.2 La confiance en tant que réalisation d'une variable aléatoire

Avec les mêmes notations, dans un corpus de données binaires, la confiance c est la fréquence conditionnelle pour qu'un sujet satisfasse à b sachant qu'il satisfait à a soit

$c = \text{Fr}[b|a] = \text{Fr}[a \wedge b|a] = [(\text{card } A \cap B) / \text{card } A] = \frac{n_{a \wedge b}}{n_a} = \frac{n_{a \wedge b}/n}{n_a/n} = 1 - \frac{n_{a \wedge \overline{b}}}{n_a}$

La variable aléatoire card $(X \cap Y) = N_{a \wedge b}$ est réalisée par card $(A \cap B)$ c'est-à-dire $n_{a \wedge b}$. Aussi, comme nous l'avons fait pour $N_{a \wedge \overline{b}}$ (Gras et Régnier, 2013) et selon une modélisation comparable, nous démontrerions que $N_{a \wedge b}$ suit approximativement, sous l'hypothèse d'indépendance, une loi de Poisson de paramètre estimé $\frac{n_a n_b}{n}$.

Considérons c comme la réalisation d'une variable aléatoire C fonction uniquement de card $(X \cap Y)$ définie par : $C = \frac{N_{a \wedge b}}{n_a}$. Cette confiance aléatoire conditionnelle prend ses valeurs réelles sur [0,1] et a pour loi la loi image, c'est-à-dire transportée de celle de $N_{a \wedge b}$. Ses valeurs sont donc des fractions des valeurs entières dont le numérateur est Poissonnien.

D'où $\Pr[C \geq \alpha] = \Pr[1 - \frac{N_{a \wedge \overline{b}}}{N_a} \geq \alpha] = \Pr[\frac{N_{a \wedge \overline{b}}}{N_a} \leq 1-\alpha]$

Or dans la contingence, la variable aléatoire N_a prend la valeur fixée et égale à n_a.

Par suite, $\Pr[C \geq \alpha] = \Pr[N_{a \wedge \overline{b}} \leq n_a(1-\alpha)]$.

2.3 Comparaison intensité d'implication et confiance

Rappelons que $Q(a, \bar{b}) = \dfrac{\text{Card}(X \cap \bar{Y}) - \dfrac{n_a n_{\bar{b}}}{n}}{\sqrt{\dfrac{n_a n_{\bar{b}}}{n}}}$ est la variable aléatoire gaussienne centrée réduite limite de la variable de Poisson $\text{Card}(X \cap \bar{Y})$ ou $N_{a\wedge\bar{b}}$, nombre aléatoire de contre-exemples à l'implication. D'où l'on tire :

$$\Pr[N_{a\wedge\bar{b}} \leq n_a(1-\alpha)] = \Pr\left[Q(a, \bar{b}) . \sqrt{\frac{n_a n_{\bar{b}}}{n}} + \frac{n_a n_{\bar{b}}}{n} \leq n_a(1-\alpha)\right].$$

Posons $r(a,\bar{b}) = \dfrac{n_a(1-\alpha) - \dfrac{n_a n_{\bar{b}}}{n}}{\dfrac{n_a n_{\bar{b}}}{n}}$. Alors $\Pr[C \geq \alpha] = \dfrac{1}{\sqrt{2\pi}} \int_{-\infty}^{r(a,\bar{b})} e^{-\frac{t^2}{2}} dt$

On vérifie bien que $\Pr[C \geq \alpha]$ est fonction décroissante de α.

A titre de comparaison, rappelons que $\varphi(a, b) = 1 - \Pr\left[Q(a, \bar{b}) \leq q(a, \bar{b})\right] = \dfrac{1}{\sqrt{2\pi}} \int_{q(a,\bar{b})}^{\infty} e^{-\frac{t^2}{2}} dt$. Posons : $\dfrac{1}{\sqrt{2\pi}} \int_{q(a,\bar{b})}^{\infty} e^{-\frac{t^2}{2}} dt = \alpha$.

Alors : si $\dfrac{1}{\sqrt{2\pi}} \int_{-\infty}^{r(a,\bar{b})} e^{-\frac{t^2}{2}} dt \geq \dfrac{1}{\sqrt{2\pi}} \int_{q(a,\bar{b})}^{\infty} e^{-\frac{t^2}{2}} dt$ la confiance est plus grande que l'intensité d'implication, sinon, c'est cette dernière qui l'emporte.

Or, avec les mêmes notations, pour α donné, $[\varphi(a,b) \geq \alpha]$ qui est aussi équivalent à $\Pr[N_{a\wedge\bar{b}} \geq n_{a\wedge\bar{b}}] \geq \alpha$ nous rappelle que l'évènement $[\varphi(a,b) \geq \alpha]$ est de moins en moins probable lorsque α croît. Nous y reviendrons plus loin.

Autrement dit, encore, partant de l'inégalité : $n_{a\wedge\bar{b}} \leq N_{a\wedge\bar{b}} \leq n_a\,(1-\alpha)$, pour un seuil comparable, si l'inégalité de gauche est satisfaite l'implication est de bonne qualité. Si celle de droite l'est, ce sera la confiance.

Les deux concepts (confiance et intensité d'implication) répondent donc à des principes relativement distincts mais non contradictoires : la confiance Fr[b|a] est fondée sur la subordination de la variable b à la variable a alors que l'intensité d'implication se fonde sur les contre-exemples à la relation de sujétion de b par a.

Ainsi, si l'on souhaite une bonne qualité d'implication (par ex. α= 0.95), il est nécessaire que le nombre de contre-exemples obtenus par hasard soit plus grand que celui de la contingence. En revanche, si l'on veut que la confiance C ait des chances d'être grande (par ex. β = 0.95), il est nécessaire que le nombre d'exemples que donnerait le seul hasard ne soit pas plus grand que 0.05 n_a. On note alors que la probabilité de la variable confiance ne dépend plus de $n_{b,}$ contrairement à l'intensité d'implication.

Rappelons, à ce sujet, que nous avons prouvé que lorsque le nombre de sujets vérifiant b tend vers n ou est égal à n, la relation implicative devient triviale, banale alors que la confiance sera maximale. Nous avons également montré que la confiance ne varie pas dans toute dilatation de E et des exemples de A et de B. Ce qui nous a conduits à refuser le critère « fréquence conditionnelle » comme seul critère de révélation d'une relation implicative. Elle est cependant, le fondement de la notion de réseau Bayesien qui permet d'organiser un ensemble de variables selon un graphe, la transition d'un nœud du graphe au suivant se faisant sur la base de la fréquence conditionnelle du premier sur le second.

Remarque prospective

Il serait alors possible, dans le cadre d'une recherche, de définir un autre indice probabiliste comme l'est φ(a,b), pour évaluer la qualité de l'implication de a sur b.

3 Lemme

Le comportement asymptotique de φ(a,b), intensité d'implication, est celui d'une variable uniforme sur l'intervalle [0 ,1].

Ce lemme a été démontré dans Gras et al (1996), mais sera rappelé et illustré par un exemple numérique dans l'annexe. Nous utiliserons ce résultat dans la suite de cet article.

4 Indicateur de proximité asymptotique entre confiance et intensité d'implication

Ainsi, $Pr[\varphi(a,b)\leq\alpha]\sim\alpha$ ou, de façon équivalente vu la continuité de la mesure, $Pr[\varphi(a,b)\geq\alpha] \sim 1-\alpha$. Or : $Pr[C\geq\alpha]= 1 - \frac{1}{\sqrt{2\pi}} \int_{r(a,\bar{b})}^{\infty} e^{-\frac{t^2}{2}} dt$

Dans ces conditions, le rapport $Pr[C\geq\alpha]/Pr[\varphi(a,b)\geq\alpha] \sim \frac{Pr[C\geq\alpha]}{1-\alpha}$ est un bon indicateur de satisfaction entre confiance et intensité d'implication : plus grand que 1, la confiance est alors meilleure que l'intensité ; inférieure à 1, c'est l'intensité qui est plus forte. Une recherche ultérieure pourrait s'appuyer sur cet indicateur.

5 Une première approche d'une intensité d'implication intégrant la contraposée

Nous présentons cette approche pour des raisons didactiques : partant d'une intuition, un développement théorique se construit mais son débouché conduit à un problème que les auteurs avaient dénoncé et déjà tenté de résoudre. Cela illustre cependant la démarche du chercheur dont les errements peuvent mener à des impasses.

L'implication a $\Rightarrow$ b, où $n_a \leq n_b$, pour des variables binaires, est d'autant plus réalisée que $n_{a\wedge b}$ est peu différent de $n_{a.}$, c'est-à-dire que l'inclusion de A dans B est vérifiée.

La contraposée de cette implication, $\bar{b} \Rightarrow \bar{a}$, est elle-aussi d'autant plus réalisée que l'inclusion de $\overline{B}$ dans $\overline{A}$ est validée, c'est-à-dire que $n_{a\wedge b} + n_{\bar{a}\wedge\bar{b}}$.est peu différent de $n_a + n_{\bar{b}}$, condition nécessaire à la satisfaction de bonnes qualités d'implication tant directe que contraposée.

	b	$\bar{\mathbf{b}}$	**Totaux**
a	$n_{a\wedge b}$	$n_{a\wedge\bar{b}}$	$\mathbf{n_a}$
$\bar{\mathbf{a}}$	$n_{\bar{a}\wedge b}$	$n_{\bar{a}\wedge\bar{b}}$	$\mathbf{n_{\bar{a}}}$
Totaux	$\mathbf{n_b}$	$\mathbf{n_{\bar{b}}}$	**n**

Tab. 1

Soit $N_{a\wedge b}$ et $N_{\bar{a}\wedge\bar{b}}$ les variables aléatoires qui représentent les cardinaux des sous-ensembles aléatoires $X \cap Y$ et $\overline{X} \cap \overline{Y}$ de E respectivement associés à $A \cap B$ et $\overline{A}\cap\overline{B}$ et de mêmes cardinaux. Posons $T = N_{a\wedge b} + N_{\bar{a}\wedge\bar{b}}$ et $t = n_{a\wedge b} + n_{\bar{a}\wedge\bar{b}}$ sa réalisation dans l'expérience de tirage uniforme.

Définition 3

On appelle **intensité d'implication inclusive** associée à a $\Rightarrow$ b, dans l'hypothèse d'indépendance a priori de a et b, la probabilité :

$K(a,b) = Pr[T \leq t]$ ou explicitement : $K(a,b) = Pr[N_{a\wedge b} + N_{\bar{a}\wedge\bar{b}} \leq n_{a\wedge b} + n_{\bar{a}\wedge\bar{b}}]$

Cette intensité inclusive est la probabilité pour que, si les variables a et b n'avaient aucun lien entre elles, le hasard des inclusions de A dans B et de leurs complémentaires inversés conduise globalement à plus d'exemples que ceux qui ont été observés. Autrement dit, plus les quasi-inclusions de A dans B et de $\overline{B}$ dans $\overline{A}$ seront proches des inclusions strictes, plus le nombre des seuls exemples satisfaisant les inclusions et dues au hasard sera réduit.

Par suite : $Pr[T \leq t] = \sum_{k=0}^{k= n_{a\wedge b} + n_{\bar{a}\wedge\bar{b}}} Pr[N_{a\wedge b} + N_{\bar{a}\wedge\bar{b}} = k]$

Choix du modèle hypergéométrique de T.

Adoptant une distribution uniforme sur l'ensemble des sujets la loi de T est hypergéométrique. Par suite :

$$Pr[T \leq t] = \sum_{k=0}^{k= n_{a\wedge b} + n_{\bar{a}\wedge\bar{b}}} . \sum_{l=0}^{l=n_{a\wedge b}} Pr[N_{a\wedge b}=l,\ N_{\bar{a}\wedge\bar{b}} = k - l]$$

$$= \sum_{k=0}^{k= n_{a\wedge b} + n_{\bar{a}\wedge\bar{b}}} \sum_{l=0}^{l=n_{a\wedge b}} \frac{C^{l}_{n_{a\wedge b}} . C^{k}_{n_{a\wedge b}-l}}{C^{l}_{n_{a\wedge b}}}$$

Les valeurs de ce nouvel indice d'implication inclusif utiliseront les moyenne et variance de la loi hypergéométrique et, de ce fait, pourront, dans la plupart des cas, être ramenés à ceux d'une loi binomiale, voire gaussienne comme il est fait dans le cas classique. Une étude ultérieure approfondie serait intéressante en partant de la loi de T.

Remarque

Une modélisation différente pourrait consister à définir séparément les lois de probabilité des variables $N_{a\wedge b}$ et $N_{\bar{a}\wedge\bar{b}}$ en comparant leurs variations aux réalisations respectives à $n_{a\wedge b}$ et $n_{\bar{a}\wedge\bar{b}}$. Cependant, tenant compte des simulations effectuées, nous abandonnerons cette démarche qui, bien qu'intégrant de façon naturelle l'information due à la contraposée, conduit aux mêmes réserves que l'implication classique, à savoir une valeur de l'intensité proche de 1 dès que les occurrences sont grandes. Aussi, nous nous tournons vers une deuxième approche.

6 Une deuxième approche d'une intensité d'implication intégrant la contraposée[6]

6.1 6.1 Motivation sémantique et épistémologique

Puisque implication et confiance admettent des racines communes (cf. § 3), nous associerons, sans agir contre nature, ces deux concepts, à la façon du physicien expérimentaliste comme il a été dit dans l'introduction. Les valeurs de la confiance relative respectivement aux règles $a \Rightarrow b$ et $\bar{b} \Rightarrow \bar{a}$ sont, en utilisant les mêmes notations que dans 2.1 :

$C_1(a,b) = Fr[Y \mid X] = ([\text{card } X \cap Y]) / (\text{card } X) = \frac{n_{a\wedge b}/n}{n_a/n} = \frac{n_{a\wedge b}}{n_a}$

$C_2(\bar{b},\bar{a}) = Fr[\bar{X} \mid \bar{Y}] = ([\text{card } \bar{X} \cap \bar{Y}]) / (\text{card } \bar{Y}) = \frac{n_{\bar{a}\wedge\bar{b}}/n}{n_{\bar{b}}/n} = \frac{n_{\bar{a}\wedge\bar{b}}}{n_{\bar{b}}}$

Or, compte tenu de leur définition, les deux intensités d'implication $\varphi(a,b)$ et $\varphi(\bar{b},\bar{a})$ sont égales. De ce fait, l'intensité ne nuance pas l'étonnement statistique de l'implication et de sa contraposée. En revanche, les deux confiances sont différentes en général. C_1 et C_2 sont d'excellents indicateurs des inclusions partielles : $A \subset B$ et $\bar{B} \subset \bar{A}$. Mais contrairement à $\varphi(a,b)$ et $\varphi(\bar{b},\bar{a})$, ils sont invariants dans toute homothétie globale des ensembles en jeu. Par suite, ils ne permettront pas de déceler des relations « surprenantes » eu égard à l'échantillon observé de taille n. C'est pour cette raison que nous leur associerons l'intensité d'implication classique,

Nous utiliserons cependant les nuances inclusives apportées par ces indicateurs d'inclusion, pour affecter l'intensité d'implication classique de coefficients permettant à la fois :

- d'intégrer les informations de l'implication directe et de sa contraposée
- d'obtenir une meilleure discrimination des valeurs critiques des contre-exemples pour des grands échantillons où la valeur 1 apparaît comme point d'accumulation de l'intensité d'implication,
- et aussi de crédibiliser le caractère inclusif de A dans B.

Ainsi, en associant les deux indices, nous devrions obtenir une mesure qui intègre à la fois la recherche des implications surprenantes (à la charge de l'intensité

[6] Ne craignons pas le changement que nous introduisons car comme l'affirme G.Bachelard : « *Accéder à la science, c'est, spirituellement rajeunir, c'est accepter une mutation brusque qui doit contredire un passé* ».(Extrait de « La Formation de l'esprit scientifique », 1938).

d'implication classique) et la crédibilité de l'inclusion conditionnelle de A dans B et de $\overline{B}$ dans $\overline{A}$ (à la charge des confiances).

Notre objectif est alors de définir une fonction de C_1 et C_2 qui permette de satisfaire les critères précédents tout en majorant l'effet implicatif dont rend compte l'intensité d'implication afin d'obtenir une relation de la forme : $\varphi(a,b).\ f(C_1, C_2)$. Une fonction du type : $f(C_1.C_2) = [C_1(a,b).C_2(\overline{b}, \overline{a})]^r$ avec r < ½ satisfait nos exigences. Nous choisirons arbitrairement r = ¼ et examinerons, à travers nos expériences si cette valeur est satisfaisante pour restituer une information riche, plausible et susceptible de prédictabilité dans un cas causal. Quoi qu'il en soit, cette valeur petite de r affecte le pouvoir de la confiance en lui réduisant son effet d'évidence intrinsèque. De ce fait, il préserve la propriété de « surprise », « d'étonnement statistique » que recèle l'intensité d'implication. Ce qui nous satisfait.

6.2 La corbeille de la mariée.

Définition 4

On appelle **implifiance**[7] la mesure de l'implication statistique qui prend en compte l'implication directe et sa contraposée ainsi que la confiance en chacune de ces deux formes inclusives. Sa valeur est :

$$\Phi(a,b) = \varphi(a,b).\ [C_1(a,b).C_2(\overline{b}, \overline{a})]^{1/4}$$

Par exemple, si l'on extrait une règle dont l'implifiance est égale à 0.95, son intensité d'implication est au minimum égale à 0.95 et chacune des confiances C_1 et C_2 est au moins égale à 0.81. Si l'implifiance est égale à 0.90, les minimas respectifs sont 0.90 et 0.66, ce qui préserve la plausibilité de la règle.

Conclusion partielle

Ainsi, même si les covariations de l'intensité d'implication et la confiance ne sont pas anarchiques, si elles n'obéissent pas ensemble à une référence en termes de probabilité, même si nous nous trouvons devant le choix arbitraire d'une définition composite d'où est partiellement exclue la part de contrôle de l'utilisateur de cette mesure, elles porteront les traces pondérées de deux indices majeurs pour évaluer la qualité implicative.

Quelques propriétés

- P1 : si $n_a = n_{\overline{b}}$, alors C1 = C2. Nous avons dans ce cas, une même confiance en l'implication directe et sa contraposée ;
- P2 : $\frac{1-C_1}{1-C_2} = \frac{n_{\overline{b}}}{n_a}$. Si ce dernier rapport reste constant, les deux confiances varient dans le même sens ;
- P3 : en exprimant $C1 = \frac{n_{a\wedge b}}{n_a} = 1 - \frac{n_{a\wedge\overline{b}}}{n_a}$ et $C2 = \frac{n_{\overline{a}\wedge\overline{b}}}{n_{\overline{b}}} = 1 - \frac{n_{a\wedge\overline{b}}}{n_{\overline{b}}}$ en fonction du nombre de contre-exemples à l'implication directe, on obtient les dérivées partielles de ces confiances par rapport à ce paramètre :

[7] Comme son nom le laisse penser, ce nouveau mot est la contraction de « **Impli**cation » et « Con**fiance** » puisque le concept associé est la combinaison des deux concepts évoqués ici.

* $\frac{\partial C_1}{\partial n_{a\wedge\bar{b}}} = -\frac{1}{n_a}$ donc C_1 décroît quand les contre-exemples augmentent et d'autant plus vite que n_a est petit ,

* $\frac{\partial C_2}{\partial n_{a\wedge\bar{b}}} = -\frac{1}{n_{\bar{b}}}$ donc C_2 décroît quand les contre-exemples augmentent et d'autant plus vite que n_b est grand, les autres paramètres étant constants.

Ces résultats sont compatibles avec la formule établie dans (Gras et Régnier, 2003) : $\frac{\partial q}{\partial n_{a\wedge\bar{b}}} = \frac{1}{\sqrt{\frac{n_a n_{\bar{b}}}{n}}} = \frac{1}{\sqrt{\frac{n_a (n-n_b)}{n}}} > 0$, qui assurent ainsi qu'intensité d'implication et confiance varient plutôt dans le même sens en accord avec le point 2.3.

Remarque

Dans notre première approche, le cadre de la modélisation imposait la nature binaire des variables en jeu. Dans cette seconde approche, les deux notions de confiance, détachée de la sémantique forte en environnement binaire, gardent des propriétés fonctionnelles ce qui autorise la définition d'implifiance dans les cas où les variables sont de nature quelconque (numériques, modales, floues, variables intervalles, etc.).

7 Quelques comparaisons

7.1 Comparaisons entre l'intensité d'implication et l'implifiance

Voici une représentation (Fig. 1) de l'intensité d'implication (II) et de l'implifiance (IC) (en ordonnée) pour des valeurs de n = 100, n_a = 40, de n_b = 50 et où l'on fait varier $n_{a\wedge\bar{b}}$ de 0 à 40 en abscisses. Sur cette même représentation, nous faisons figurer, dans un objectif de comparaison, comment varient les 3 modes de valuation de la qualité de l'implication II, IC et IE (intensité entropique utilisée précédemment). Cette dernière sera également présente dans les figures 2,3, 4, 5, 6, 7, 8 et 9 où nous faisons varier les paramètres fondamentaux : n, n_a et n_b.

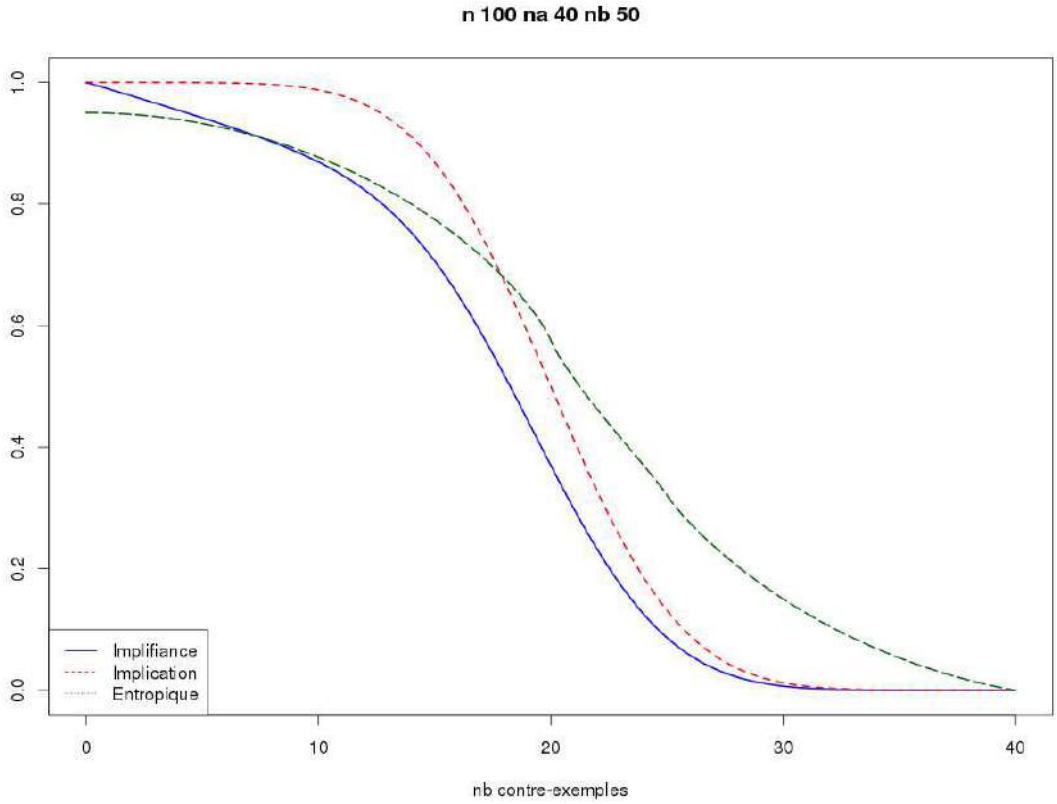


Figure 1

On remarque immédiatement, ce qui était attendu compte tenu de sa définition, que l'implifiance est toujours inférieure à l'intensité d'implication classique. Une valeur de 0.95 pour l'intensité d'implication pourra être accompagnée d'une valeur de l'implifiance compris entre 0.85 et 0.90. Que l'usager ne s'en inquiète pas. Par exemple, des confiances C_1 et C_2 supérieures à 0.50 conduisent à une implifiance supérieure à 0.67 alors que l'intensité est 0.95. De plus, l'IC n'ayant pas de dérivée nulle à l'origine décroît plus vite que l'II mais cette décroissance est lente avant de s'accélérer à n/4, voisinage du point d'inflexion de la courbe. En revanche, comme nous le savons, l'II décroît très lentement au début de la croissance des contre-exemples mais cette lenteur est préjudiciable pour des valeurs de n grandes car elle devient alors peu discriminante comme nous le voyons par une homothétie x100, dans la Fig. 2 où cette fois n = 10000, n_a = 4000, de n_b = 5000 et où l'on fait varier $n_{a \wedge \bar{b}}$ de 0 à 4000. L'II (intensité d'implication) ne varie que très peu de la valeur maximum 1 jusqu'à $n_{a \wedge \bar{b}}$ = 2000, puis plonge brusquement plus vite, ce qui la valorise, que IE, l'intensité entropique.

Remarquons que l'IC est plus fidèle à l'II que ne l'était la version précédente de la mesure de qualité entropique.

D'autres situations avec n= 10 et n_a = 1, n = 100 et n_a = 10 (Fig. 3), n = 1000 et n_a = 1000 montrent que le rejet de l'implication, en raison de la faiblesse de la valeur de IC, n'intervient pas trop tôt et seulement au voisinage du cardinal de $X \cap \overline{Y} = n_a/2$.

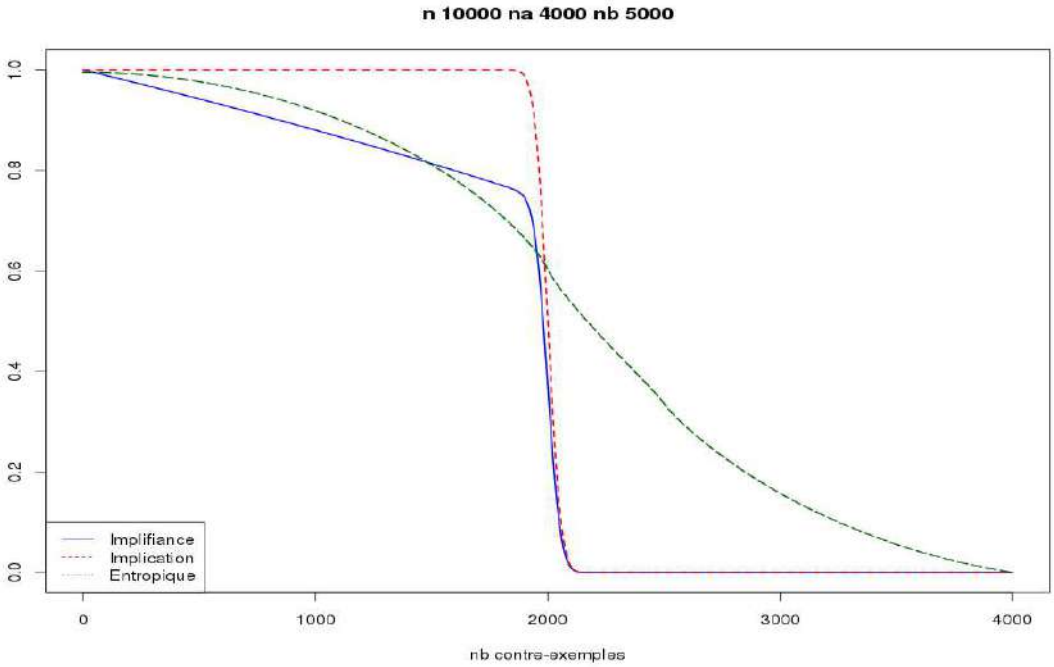


Figure 2

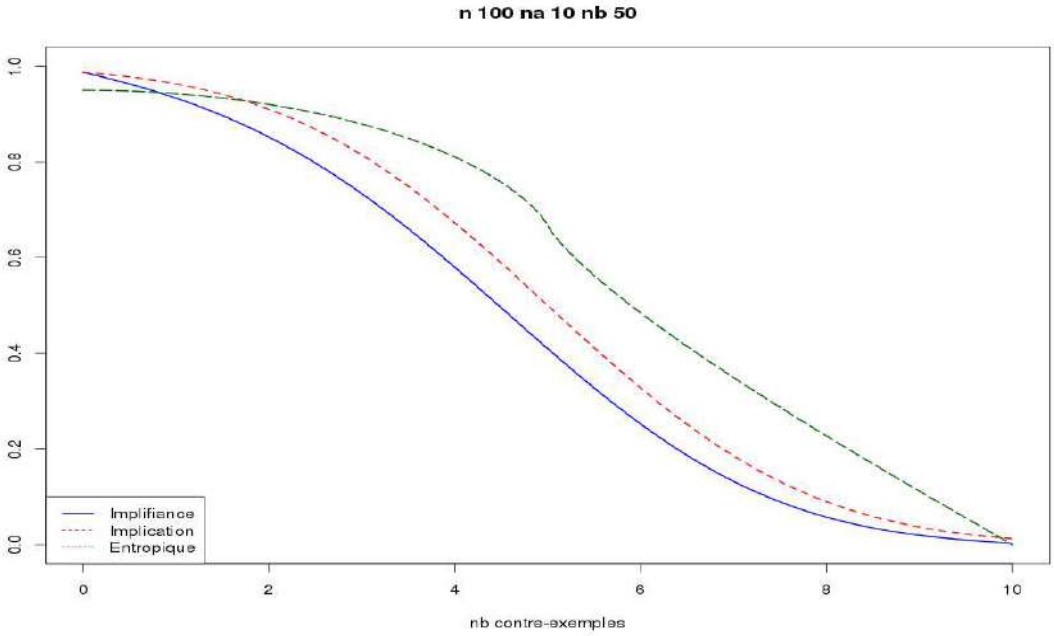


Figure 3

Lors de l'examen des spécificités de l'ASI disposant de l'intensité φ(a,b) comme outil de mesure implicative, nous avions souligné la propriété suivante : au voisinage de $N_{a\wedge\bar{b}} = 0$, l'intensité d'implication varie peu et reste très voisine de 1, comme le montrent les courbes (cf Fig.1 et Fig. 2), dont la dérivée à l'origine est presque nulle. Cette propriété a la vertu de permettre une certaine résistance de la fonction implicative, donc également prédictive, avant de céder à de trop forts écarts autour de 0. Certes, ceci

évite le rejet brutal de l'implication pour quelques contre-exemples qui pourraient n'être qu'accidentels ou dus à des erreurs de mesure. Mais le prix à payer pour garantir cette sagesse se facture en des valeurs de l'intensité difficilement discriminantes comme nous venons de le voir. Aussi, l'implifiance semble associer harmonieusement cette réserve à l'égard des écarts admissibles à l'implication et cependant la mise en garde du chercheur envers l'accumulation de contre-exemples devient insoutenable.

7.2 Comparaisons du comportement de l'implifiance par rapport à celui des confiances C1 et C2

Une autre propriété de l'intensité d'implication II a été soulignée. Ses variations en fonction de la variable $N_{a\wedge\bar{b}}$ ne sont pas linéaires contrairement à la seule confiance des multiples et autres indices d'association. Or, nous avons dénoncé l'inadéquation de la linéarité en tant que modélisante dans le fonctionnement de la pensée. La philosophie structuraliste nous rappelle que le « tout est plus riche que la somme de ses parties ». « *Autrement dit, 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 ne peuvent donc s'expliquer par elles* » (Sève L., 2005). Ce phénomène qui conduit bien souvent à l'émergence d'une propriété du tout nécessite analytiquement un caractère non dérivable ou à points d'inflexion de la fonction qui mesure le phénomène. D'où cette mise à l'écart de la linéarité. En effet, en un point d'inflexion de l'implifiance, celui-ci marque le passage de la phase d'accélération de la décroissance de l'implifiance à sa phase de décélération. Les courbes qui décrivent les variations de l'implifiance montrent bien ce caractère non linéaire qui tient à la présence multiplicative de l'intensité d'implication dans sa définition.

En revanche, comme le montrent les courbes représentant C_1 et C_2 des figures 4a, 5a et 6a, la linéarité de celles-ci ne respecte pas la philosophie dont s'arroge la nouvelle mesure implicative. Rappelons que C_1 mesure la fréquence de b sachant a et que C_2 mesure celle de non a sachant non b. Ces deux fréquences ne sont qu'exceptionnellement égales et se présentent avec une valuation indifférente : tantôt C_1 est au-dessus de C_2 et tantôt le contraire.

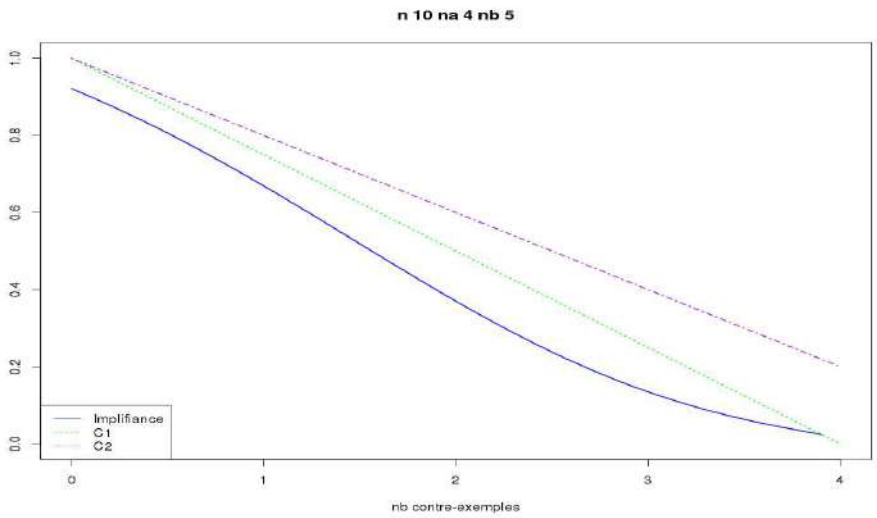


Figure 4a

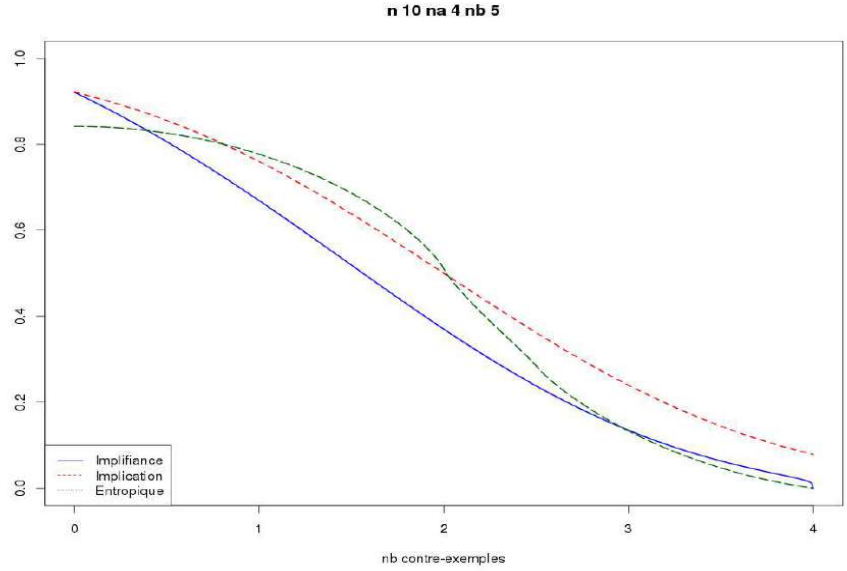


Figure 4

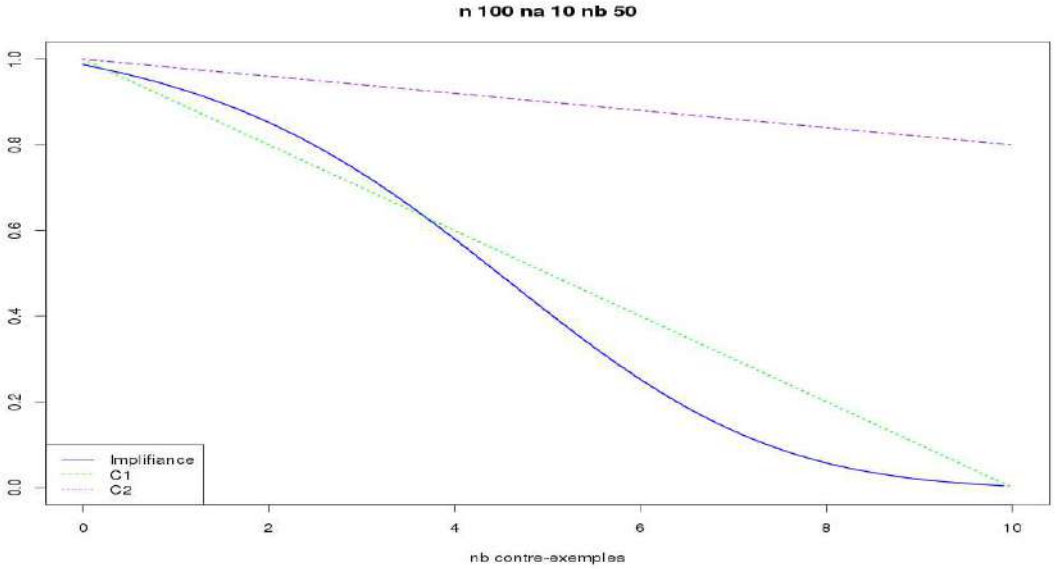


Figure 5a

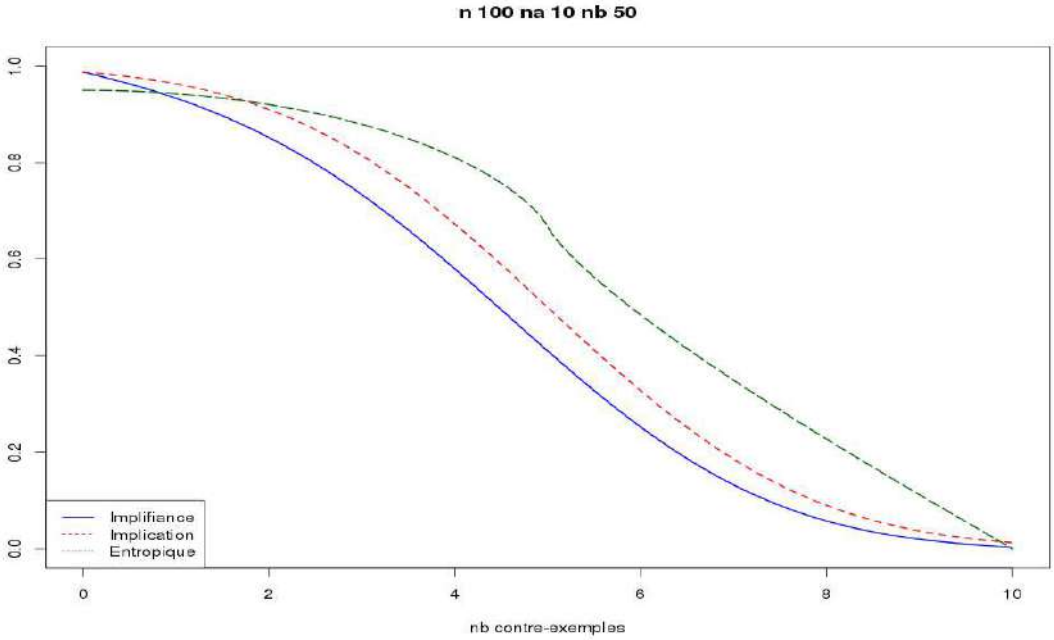


Figure 5b

En revanche, la figure 6a montre clairement l'effet « lissage » produit par les coefficients des deux facteurs de confiance mais sans en supporter le défaut majeur de linéarité sur l'ensemble des nombres de contre-exemples : décroissance de l'implifiance de 0 à $n_a/2$, puis plongeon de celle-ci lorsque la règle implicative associée devient insoutenable. Cet effet « lissage » est répercuté sur l'implifiance comme on le voit sur la figure 6.

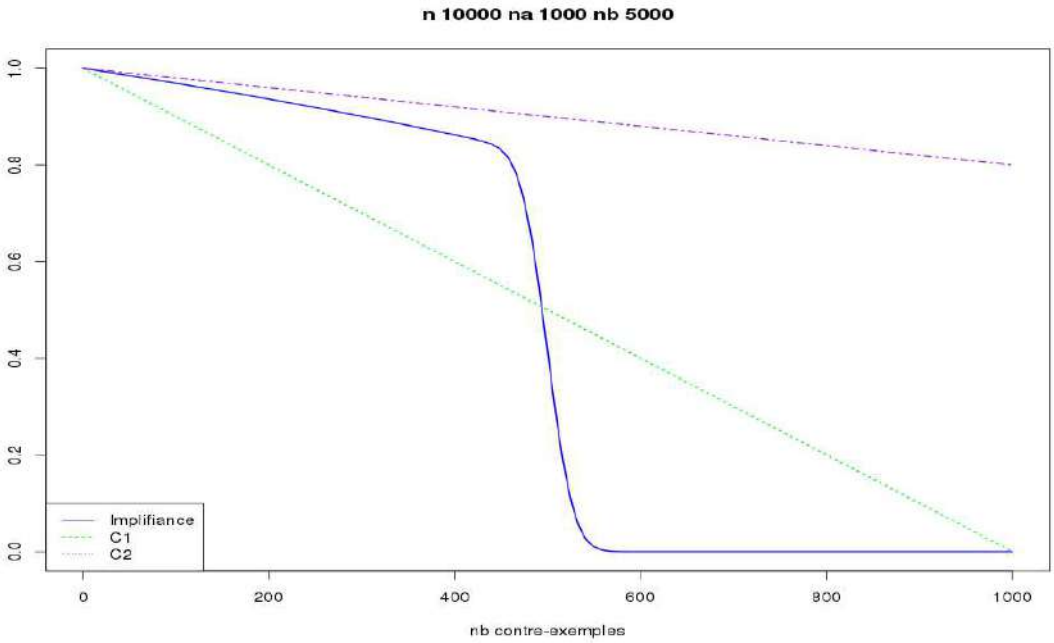


Figure 6a

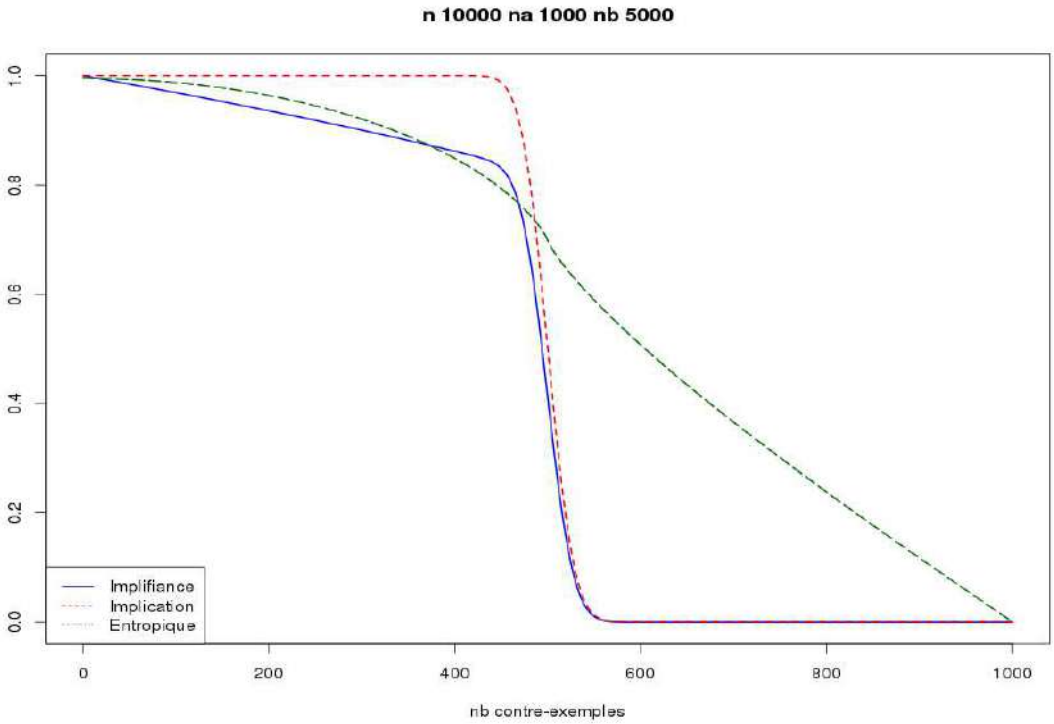


Figure 6b

7.3 Comparaisons entre l'intensité d'implication, l'implifiance et l'intensité entropique

En réaction au comportement de l'intensité d'implication II pour les corpus volumineux, insuffisamment discriminante sur une grande plage de contre-exemples, nous avions défini un nouvel indice pour mesurer la qualité des règles implicatives. La définition s'appuyait sur la notion d'entropie conditionnelle des évènements associés à l'implication directe et sa contraposée. Comme nous l'avons dit dans l'introduction, il lui fut reproché son côté quelque peu ad-hoc ainsi que la nécessité de modifier l'entropie à partir de $n_a/2$ pour lui ôter sa propriété de symétrie incompatible avec la philosophie implicative. Autre grief : comme l'II la décroissance de l'intensité entropique semblait insuffisante pour rendre compte de l'admissibilité du caractère implicatif des règles évaluées. Ceci nous a conduits à la nouvelle mesure envisagée dans cet article : l'implifiance.

Les courbes 7, 8 et 9 qui suivent, obtenues également dans la situation où le nombre de sujets est important : n =10000, illustrent la comparaison entre les 3 mesures : l'intensité d'implication classique, l'intensité entropique et l'implifiance.

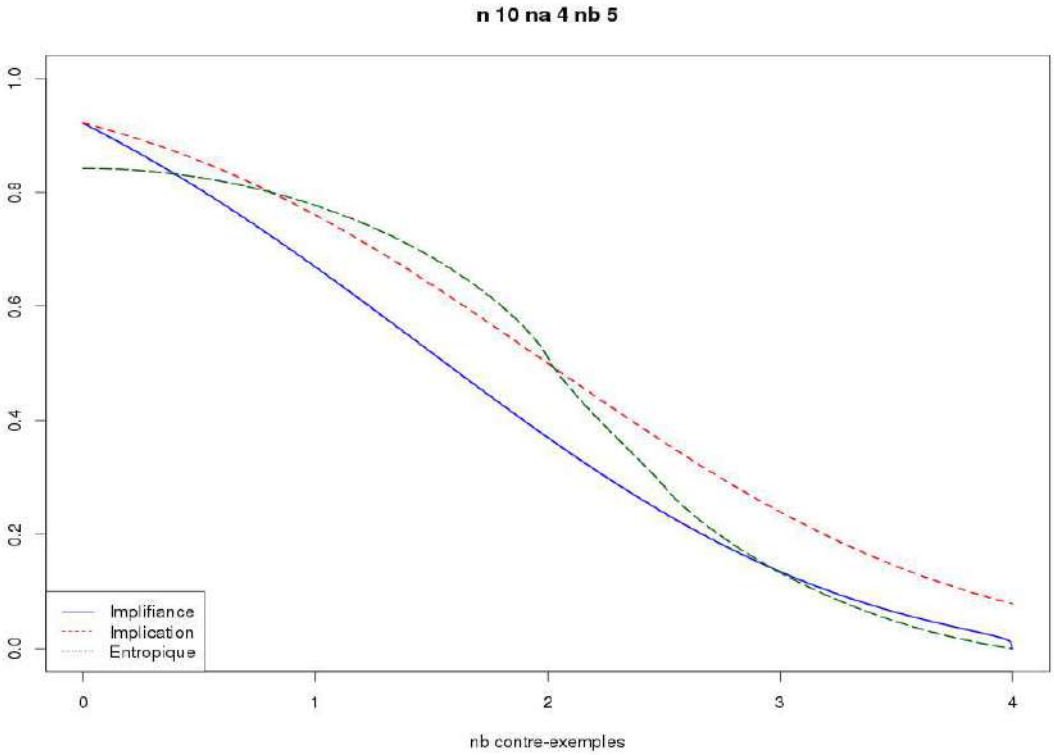


Figure 7

On remarquera, en effet, la plus grande résistance au rejet de la règle implicative dès que le nombre de contre-exemples devient trop importants eu égard aux valeurs de n_a et n_b, et particulièrement selon la courbe 9 des grandes valeurs des données.

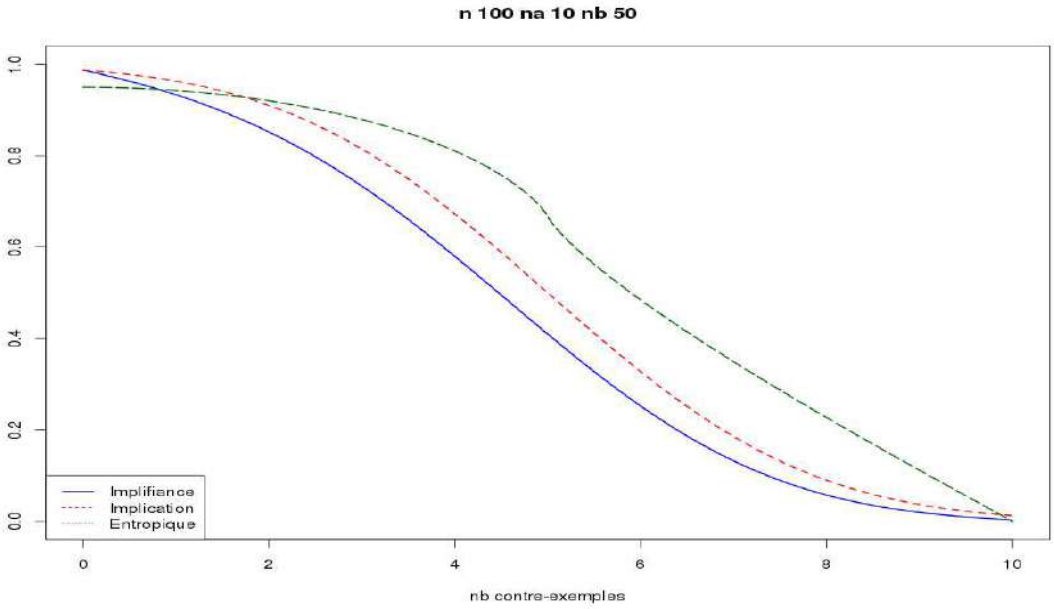


Figure 8

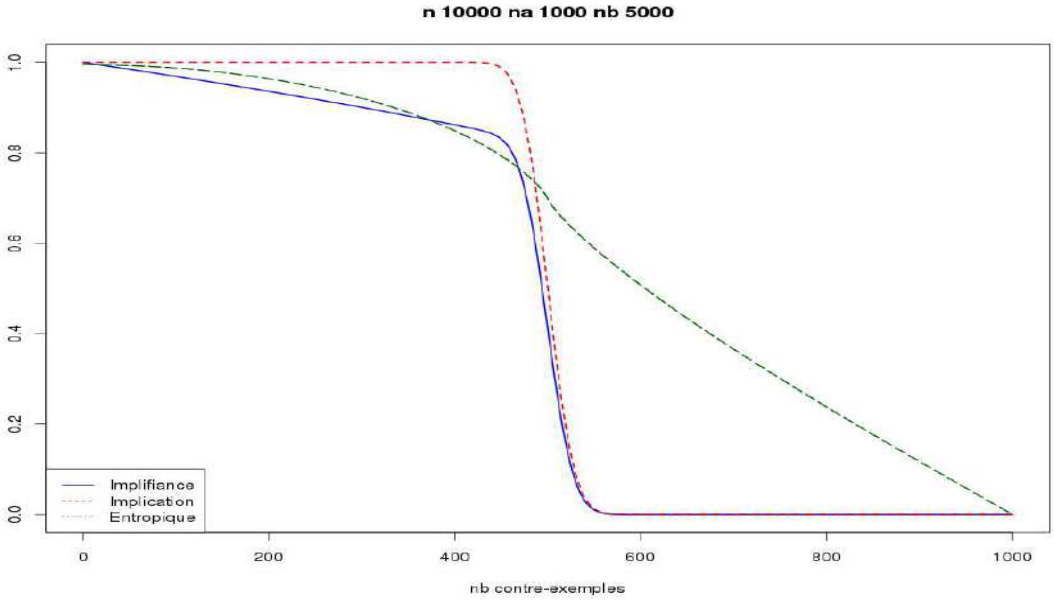


Figure 9

Comme nous l'avons déjà souligné, nous notons, avec ces 3 exemples 7, 8 et 9, une plus grande affinité entre l'intensité II et l'implifiance IC qu'entre II et l'intensité entropique IE. La fig. 9 est significative à cet égard où IE tarde à décroître avec le nombre croissant de contre-exemples. Ceci n'empêche pas l'implifiance de restituer à l'II les nuances de « confiance » que lui confèrent les deux composantes de la fréquence conditionnelle C_1 et C_2. Ceci confirme le contenu du § 3 où nous avions montré la relation entre intensité d'implication et confiance.

8 Conclusion

Nous proposons dans cet article la construction d'une nouvelle mesure qui permette d'évaluer la qualité de règles implicatives sur la base de l'importance relative des contre-exemples à l'implication. *Nous poursuivons donc la consolidation de notre édifice* ***non-réductionniste*** *de la mise en évidence et de représentation de relations présumées causales, cachées au sein d'un large ensemble de variables, édifice intégrant aussi la dualité sujets-variables, édifice qui fait de l'ASI un modèle original et fécond.* D'ailleurs, nous avons pu laisser ouvertes quelques pistes de recherche ultérieure.

Cette mesure, *l'implifiance*, comme l'intensité entropique jusqu'alors utilisée en complément de l'intensité d'implication, a la vertu de prendre en compte la contraposée de l'implication faute de quoi la fonction de recherche causale nous semblerait imparfaite. *Par rapport à l'intensité entropique, elle satisfait le principe philosophique du rasoir d'Occam*[8] *car sa définition est moins complexe.* Elle intègre cette fois et en outre, la notion de confiance ou fréquence conditionnelle afin de limiter à l'examen des règles celles qui présentent un caractère de liaison conditionnelle entre la prémisse et la

[8] Ce principe est généralement retenu par les scientifiques qui affirment qu'en présence de plusieurs théories en charge de la même réalité, il est préférable de choisir la plus simple.

conclusion. Tout comme ses mesures aînées, cette implifiance subira les assauts des données réelles *(« Expérience, source unique de vérité »* écrit Henri Poincaré dans « Science et hypothèse ») pour pouvoir prétendre être un bon prédicteur de relation causale et un bon outil d'extraction de pépites de connaissance. Et de toute façon, la qualité de la prévision restera entachée des incertitudes inhérentes au choix du modèle probabiliste. Car, comme le dit avec humour le physicien Niels Bohr : « *La prévision est un art difficile, surtout quand elle concerne l'avenir ».*

Références

[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.

[2] Amarger S., D. Dubois and H. Prade (1991), Imprecise quantifiers and conditional probabilities, In *Symbolic and quantitative approaches to* uncertainty (R. KRUSE, P. SIEGEL), Springer-Verlag, 33-37.

[3] Bachelard G. (1967), *La Formation de l'esprit scientifique*, Paris, 5e édition, Librairie philosophique J. Vrin.

[4] Briand H., M. Sebag, R. Gras et F.Guillet (2004), *Mesures de Qualité pour la Fouille de Données*, H.Briand, M.Sebag, R.Gras et F.Guillet eds, RNTI-E-1, Cépaduès, 2004.

[5] 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.

[6] Couturier, R. (2008). CHIC: cohesive Hierarchical Implicative Classification, In *Statistical implicative analysis*. Volume 127 of Studies in Computational Intelligence, Springer Verlag, p. 41–54.

[7] 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.

[8] Frawley W., G. Piatetski-Shapiro and C. Matheus (1992), Knowledge discovery in databases: an overview. *AI Magazine*. 14(3), 57-70.

[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 1.

[10] Gras R., S. Ag Almouloud, M. Bailleul, A. Larher, M. Polo, H. Ratsimba-Rajohn et A. Totohasina (1996), *L'implication Statistique*, Collection Associée à Recherches en Didactique des Mathématiques, Grenoble : La Pensée Sauvage.

[11] 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. Un exemple : l'implication statistique, *Mesures de qualité pour la fouille de données, RNTI-E-1, Cépaduès –Editions*, 3-32.

[12] Gras R. et R. Couturier (2010), Spécificités de l'Analyse Statistique Implicative (A.S.I.) par rapport à d'autres mesures de qualité de règles d'association, *Quaderni di Ricerca in Didattica - GRIM (ISSN on-line 1592-4424)*, Eds : J.C. Régnier, R.Gras, F.Spagnolo, B. Di Paola, Université de Palerme, *p.19-57.*

[13] Gras R. et J.-C. Régnier (2013), Fondements théoriques de l'Analyse Statistique Implicative, *Méthode exploratoire et confirmatoire à la recherche de causalités*, sous la direction de Gras R., eds Gras R., Régnier J.-C., Marinica C., Guillet F., Cépaduès Editions, 522 pages, ISBN 978.2.36493.056.8, p 25-186.

[14] Gras R., J.-C. Régnier et F. Guillet (2009), *Analyse Statistique implicative.Une méthode d'analyse de données pour la recherche de causalités,* sous la direction de Régis Gras, réd. invités R. Gras, J.C. Régnier, F. Guillet, Cépaduès Ed. Toulouse.

[15] Gras R., J.-C. Régnier, C. Marinica et F. Guillet (2013), *L'analyse statistique implicative, Méthode exploratoire et confirmatoire à la recherche de causalités*, Cépaduès Editions, 522 pages, ISBN 978.2.36493.056.8.

[16] Gras R., E. Suzuki, F. Guillet and F. Spagnolo (2008), *Statistical Implicative Analysis, Theory and Applications,* R.Gras, E. Suzuki, F. Guillet, F. Spagnolo, eds, Springer.

[17] Guillet F. et H. Hamilton (2007), *Quality Measures in Data Mining* , F.Guillet et H.Hamilton eds, Springer.

[18] 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.

[19] Lallich S., O. Teytaud et E. Prudhomme (2007), Association Rule Interestingness: Measure and Statistical Validation, *F.Guillet and H. J.Hamilton eds, Studies in Computational Intelligence 43, Springer*, p. 251-275.

[20] Lenca P. et S. Lallich (2011), *Le choix d'une bonne mesure de qualité, condition du succès d'un processus de fouille de données*, Atelier Data Mining, Applications, Cas d'Etudes et Success Stories, Extraction et Gestion des Connaissances, *5-8.*

[21] Lerman I.-C. (1981), *Classification et analyse ordinale des données*, Paris : Dunod.

[22] Orús P., L. Zamora et P. Gregori (2009), *Teoria y Aplicaciones del Analisis Estadistico Implicativo,* Eds: P. Orús, L. Zamora y P. Gregori, Universitat Jaume I Castellon (Espagne), ISBN : 978-84-692-3925-4.

[23] Pearl J. (1988), *Probabilistic Reasoning in intelligent systems,* San Mateo, CA, Morgan Kaufmann.

[24] Régnier J.-C., M. Bailleul et R. Gras (2012), *L'Analyse Statistique Implicative : de l'exploratoire au confirmatoire.* Eds : J.C. Régnier, Marc Bailleul, Régis Gras, Université de Caen, ISBN : 978-2-7466-5256-9, 2012.

[25] Saporta G. (2006), *Probabilités, Analyse de Données et statistique,* Paris : Ed. Technip.

[26] 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.

[27] Sève L. (2005), *Emergence, complexité et dialectique,* Odile Jacob, Paris.

[28] Vaillant B., S. Lallich and P. Lenca (2008), On the behaviour of the generalisations of the intensity of implication: a data-driven comparative study, *Statistical Implicative Analysis, Theory and Applications,* R.Gras, E. Suzuki, F. Guillet, F. Spagnolo, Studies in Computational Intelligence, 127, Springer-Verlag Berlin Heidelberg.

[29] Vergnaud G. (2007), *Activités humaines et conceptualisation,* Presses Universitaires du Mirail, p.29.

[30] Xuan-Hiep Huynh, F. Guillet, J. Blanchard, P. Kuntz, H. Briand et R. Gras (2007), A Graph-based Clustering Approach to Evaluate Interestinguess Measures: A Tool and and a Comparative Study, *F.Guillet and H. J.Hamilton eds, Studies in Computational Intelligence 43, Springer*, p. 25-50.

ANNEXE

Lemme

Le comportement asymptotique de φ(a,b), intensité d'implication, est celui d'une variable uniforme sur l'intervalle [0 ,1].

Notons $p = \varphi(a, b)$, p étant une probabilité, ne peut être considérée formellement comme une variable aléatoire qu'en changeant le modèle probabiliste jusqu'alors adopté. En effet, l'univers des possibles Ω introduit pour l'étude de l'étonnement à l'observation de la valeur $q(a, \bar{b})$, suppose que les valeurs également observées n_a et n_b soient des réalisations des cardinaux des variables aléatoires indépendantes X et Y. Si l'on veut alors mesurer un "étonnement" au sujet de la valeur de la cohésion, il faut considérer un autre univers des possibles Ω' sur lequel p suivrait une certaine loi de probabilité. Le problème revient alors à choisir la "bonne" loi de probabilité que suivrait p.

Nous allons la calculer pour n suffisamment grand, de telle façon que l'on puisse se considérer dans le cas asymptotique. Pour cela, on suppose que la valeur observée $n_{a \wedge \bar{b}}$ est la réalisation d'une variable aléatoire qui n'est plus Card $(X \cap \overline{Y})$, mais Card $(A \cap \overline{B})$, où A et B sont des sous-ensembles aléatoires indépendants dans E et fonctions d'éventualités $\omega \in \Omega' \neq \Omega$.

On obtient dans ces nouvelles conditions:

Prob [$p \leq \alpha$] = Prob Ω' [Prob Ω (Card $(X \cap \overline{Y}) \leq n_{a \wedge \bar{b}}$) $\leq \alpha$]

= Prob Ω' { ω' | Prob Ω { ω | Card $(X \cap \overline{Y})(\omega) \leq$ Card $(A \cap \overline{B})(\omega')$ } $\leq \alpha$}

= Prob Ω' { ω' | Prob Ω { ω | $N(\omega) \leq N'(\omega')$} $\leq \alpha$}}

où N et N' admettent des lois identiques, par exemple des lois normales centrées réduites et dans ce cas :

Prob [$p \leq \alpha$] = Prob Ω' { ω' | $\int_{-\infty}^{N'(\omega')} e^{\frac{-t^2}{2}} dt \leq \alpha$ }

= Prob Ω' { ω' | $N'(\omega') \leq F^{-1}(\alpha)$}, où F est la fonction de répartition de la loi normale,

$$= \int_0^{F^{-1}(\alpha)} e^{\frac{-t^2}{2}} dt = F(F^{-1}(\alpha))$$

$$= \alpha \qquad (1)$$

Ainsi, la loi de p est uniforme sur l'intervalle [0,1]. Cela signifie que, si n est très grand, un tirage au hasard de deux variables a et b étant donné, l'évènement [[$\varphi(a,b) \leq \alpha$]=α] ou [$\varphi(a,b) \geq \alpha$]=1-α] est presque sûr. De plus, la loi de l'intensité d'implication, en tant que variable aléatoire, tend vers une loi uniforme sur [0,1].

Exemple numérique fictif : 11 variables, 1081 objets (des véhicules)

On dispose de 11x10/2 soit 55 couples envisageables pour définir une règle de type $a \Rightarrow b$ respectant $n_a \leq n_b$. Le tableau des intensités d'implication donne des résultats n'invalidant pas la relation (1) :

- 5 intensités supérieures ou égales à 0.95 (1-a = 0.05) pour 5% x 55 = 2.75 intensités attendues d'après (1) ;
- 6 intensités supérieures ou égales à 0.90 (1-a = 0.10) pour 10% x 55 = 5.5 intensités attendues d'après (1) ;
- 8 intensités supérieures ou égales à 0.80 (1-a = 0.20) pour 20% x 55 = 11.5 intensités attendues d'après (1) ;

Ce qui statistiquement et pour ces trois mesures représente 19 résultats observés pour 19.75 attendus confortant notre lemme.

ANALYSE D'UN QUESTIONNAIRE « ENSEIGNANTS DE MATHEMATIQUES » PAR DIFFERENTES METHODES

Régis GRAS[1] et Antoine BODIN[2]

RÉSUMÉ

Les différents acteurs éducatifs (parents, enseignants, élèves) s'interrogent**,** particulièrement en mathématiques, sur les objectifs généraux de leur enseignement. Un groupe de travail de l'Association des Professeurs de Mathématiques a élaboré un questionnaire pour interroger les enseignants eux-mêmes sur ces objectifs tels qu'ils les perçoivent à travers ce qu'en attendent l'institution et la société. Les réponses fournies par un corpus de professeurs de filières différentes sont analysées par différentes méthodes de traitement de données dont l'A.S.I., l'A.C.P. et autres. Nous en comparons dans cet article les résultats obtenus en insistant sur l'émergence de deux variables latentes qui discriminent la population interrogée.

Mots-clés – *objectifs, analyse statistique implicative, analyse en composantes principales, analyse hiérarchique, nuées dynamiques, typicalité.*

ABSTRACT

The various agents of the Educational System try to better understand, what are, really, the general objectives of teaching mathematics. To do that, a research team from the Association of Mathematics Teachers set a questionnaire to ask the teachers themselves about what, according to them, is being expected by the educational institution and by the society at large. The answers provided by a corpus of teachers of different streams have been analyzed by different data processing methods including I.S.A., P.C.A. and some others. In this paper, we compare the results obtained with emphasizing the emergence of two latent variables that discriminate the population surveyed.

Keywords : *Objectives, Implicative Statistical Analyse, Principal Component Analysis, Hierarchical Clustering, k-means Clustering, Typicality.*

1 Introduction. Pourquoi un tel questionnaire ?

Vers la fin des années 90, un groupe de travail de l'Association des Professeurs de Mathématiques de l'Enseignement public (A.P.M.E.P.) a été chargé d'organiser une étude à grande échelle sur les acquis des élèves des classes terminales (17-18 ans) des lycées, toutes séries confondues. L'opportunité d'élaborer de telles épreuves ne s'était pas immédiatement inscrite dans le prolongement des épreuves des classes antérieures et, en particulier, de la classe de première (16-17 ans. Mais différents facteurs se sont conjugués pour que le comité de l'APMEP accepte de confier cette nouvelle (et lourde) tâche à l'Observatoire EVAPM (Évaluation des Programmes de Mathématiques. Cette étude apparaissait comme une suite logique de ce qui avait été présenté dans toutes les autres classes des premier et second cycles (de 11 à 18 ans). Pourquoi donc s'arrêter avant la terminale ? Les épreuves apparaîtront-elles comme un obstacle psychologique

[1] École Polytechnique de l'Université de Nantes, Équipe DUKE Data User Knowledge, Laboratoire d'Informatique de Nantes-Atlantique (LINA), UMR 6241 : regisgra@club-internet.fr

[2] Institut de Recherche sur l'Enseignement des Mathématiques, Université d'Aix-Marseille, 163 Avenue de Luminy - Case 901, 13288 Marseille Cedex 09, antoinebodin@mac.com

au climat pré-baccalauréat ? Au contraire agiront-elles comme un stimulant et un récapitulatif de ce qui paraît essentiel de savoir et de savoir-faire ? Ne permettraient-elles pas, comme les évaluations précédentes, de mettre en évidence les obstacles communs rencontrés par les élèves, d'éclairer ainsi sur les difficultés liées aux programmes, aux limites de leur assimilation et aux conditions institutionnelles de leur réalisation ?

Parallèlement, mais de façon indépendante, un mouvement de fond avait fait émerger des critiques sévères et convergentes au sujet du baccalauréat lui-même, de son inadéquation au moins partielle à l'esprit dans lequel veut être pratiqué l'enseignement, tout au moins jusqu'en classe de première. Dès 1997, des réunions inter-associatives (APMEP, UPS, SMAI, SMF[3]), en présence de l'Inspection Générale de Mathématiques, avaient permis de dresser un bilan très réservé au sujet de certains acquis des élèves à la sortie des classes terminales et avaient responsabilisé, en partie, le baccalauréat dans les appauvrissements constatés par effet induit. En effet, personne ne contestait plus la relation de cause à effet de l'évaluation de sortie du lycée sur l'enseignement-même qui s'y pratiquait : à un examen peu adapté à la mesure de compétences recherchées répondrait un enseignement souvent rétréci à la préparation aux conditions dans lesquelles se présente l'examen. Sans oblitérer les objectifs spécifiques d'EVAPM (cf. plus haut) et leur opérationnalisation, les épreuves de ce type s'offraient en instrument d'évaluation des curricula des classes terminales et ne pouvaient que contribuer à objectiver les jugements portés sur l'enseignement conduit dans ces classes.

En préalable à l'étude proprement dite, donc à la rédaction des épreuves, ce groupe, dans le cadre de l'observatoire EVAPM, a décidé de sonder, par un questionnaire portant sur des objectifs généraux, les enseignants ayant accepté de faire passer les épreuves dans leurs classes. Le questionnaire proposé portait sur les représentations des enseignants dans les classes terminales de lycée quant à leur rôle dans la formation en mathématique à leur charge. Il devait permettre d'obtenir une typologie et une hiérarchie plus claires des attentes des professeurs. Que privilégient-ils ? Que jugent-ils essentiel ? Qu'est-ce qui distingue les opinions des enseignants des classes littéraire ou technique ou scientifique ? Y a-t-il des invariants, des relations stables entre les opinions, des éléments prédicteurs de gestes pédagogiques ? La présente étude approfondit une partie de l'analyse publiée en 1999 (Bodin et Gras, 1999) : d'une part elle se centre sur l'analyse de la première partie du questionnaire, d'autre part elle complète l'utilisation de l'analyse implicative et s'ouvre à d'autres méthodes d'analyse des données. Les méthodes d'analyse qui vont tenter de répondre aux questions posées ci-dessus, font en effet référence aux statistiques descriptives classiques, mais également à des méthodes multidimensionnelles qui globalisent tous les comportements de réponse afin d'en dégager les dimensions majeures et les structures sous-jacentes. On distingue de manière classique les variables manifestes et les variables latentes. Les premières, qui relèvent du champ empirique, sont directement observées et / ou mesurées. Les secondes qui relèvent d'un champ inférentiel, renvoient à des caractéristiques qui ne sont ni directement observables, ni directement mesurables, ce qui est le cas des

[3] UPS : Union des Professeurs de Spécialités.
SMAI : Société de Mathématiques Appliquées et Industrielles,
SMF : Société Mathématique de France

processus mentaux. En psychologie, on rendra compte des relations entre les données observées par le jeu de variables inférées de ces données. L'inférence peut se limiter à un versant qualitatif. Par exemple, on expliquera la covariation de l'âge mental et du poids de l'enfant par un processus de maturation et de croissance. Elle peut aussi s'exprimer sur le versant qualitatif. Par exemple, l'analyse factorielle des données observées permet de dégager des dimensions latentes dont les scores factoriels pourront être calculés pour chacun des individus. Dans le cadre de l'ASI, l'analyste met en évidence des chemins implicatifs entre les variables observées. On peut également calculer les scores (contributions et / ou typicalités) de chacun des individus sur ces dimensions implicatives. Pour l'analyse factorielle et pour l'ASI, la question centrale revient à l'interprétation et à la validation des facteurs ou des chemins produits de l'analyse. A une nuance près : pour l'analyse factorielle on cherche à interpréter les regroupements corrélationnels des variables observées ; à ce premier niveau symétrique, l'ASI en ajoute un second non symétrique, l'interprétation de l'ordre implicatif des variables. En d'autres termes, sont-ce les mêmes variables latentes qui expliquent à la fois regroupement et ordre des variables observées, ou bien faut-il considérer deux strates de variables latentes explicatives ?

2 Le questionnaire : objectifs de la formation mathématique

La consigne précédant le questionnaire était la suivante : *A votre avis, quels sont les objectifs essentiels de la mission d'un professeur de mathématiques dans la série pour laquelle vous répondez. Pour répondre à cette question, classez par ordre préférentiel décroissant de 1 à 6 (1 : le plus important,...) six des objectifs majeurs de cette formation en les choisissant parmi les objectifs proposés ci-dessous :*

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

Codage utilisé :
Exemple de réponse d'un enseignant x : en 1 : I ; en 2 : G ; en 3 : M ; en 4 : D ; en 5 : A ; en 6 : J)

Pour un traitement par le logiciel CHIC (Couturier et Ag Almouloud, 2009), ces variables pourraient être considérées comme des variables modales. C'est la stratégie que nous avions utilisée dans notre ancienne présentation de l'analyse complète par l'ASI de ce questionnaire (cf. Bodin et Gras, 1999). Cela nous permettait de conserver

l'information provenant des rangs choisis par chaque enseignant. Ici, nous nous contentons de coder 1 à chaque fois qu'un objectif est choisi par l'enseignant, quel que soit le rang qu'il lui a attribué ; et nous codons 0 lorsque la variable n'est pas classée.

On s'intéresse alors à trois sujets : le classement des items selon la fréquence de choix des items ; la structure implicative des items ; la structure des items à l'aide d'autres approches et comparaison avec la structure implicative.

Nombre de sujets ayant répondu au questionnaire :

311 professeurs ont répondu à ce questionnaire :

- 50 % enseignent en série Scientifique (S),
- 21 à 22 % enseignent en Economique et Sociale (ES) ou en diverses séries technico-professionnelles (TE),
- 7% enseignent en classe littéraire (LI).

2.1 Fréquence des choix :

Nous récapitulons ci-dessous (Tab. 1) les fréquences de choix quels que soient les rangs retenus.

item	n	fréquence
D préparation aux examens, concours, au passage dans l'enseignement supérieur	230	73,95%
F développement de la capacité à prouver et valider sa preuve	230	73,95%
J développement de la capacité à communiquer avec objectivité, clarté et précision par des modes de représentation divers	205	65,92%
M développement de la capacité à mathématiser et à formaliser	178	57,23%
I développement de l'esprit critique	164	52,73%
A acquisition de connaissances	161	51,77%
K développement de compétences utiles dans les autres disciplines	161	51,77%
N acquisition de savoir-faire	137	44,05%
H développement de la volonté et la persévérance	120	38,59%
O participation au développement d'une culture générale	101	32,48%
E développement de l'imagination et la créativité	81	26,05%
G développement de la capacité d'accepter des points de vue différents	50	16,08%
B préparation à la vie professionnelle	24	7,72%
C préparation à la vie civique et sociale	19	6,11%
L développement de la pratique de calculs formels, donc sans nécessité de signification	12	3,86%

Tableau 1. Fréquences des choix.

Les fréquences des choix sont franchement inégales, de 3,85% à 73,72%. Ces fréquences observées valident les constats précédents : « ...à un examen peu adapté à la mesure de compétences recherchées répondrait un enseignement souvent rétréci à la préparation aux conditions dans lesquelles se présente l'examen. »

Des compétences plus larges, transversales ou génériques utiles aux adaptations futures apparaissent comme des objectifs secondaires. Le contexte scolaire réduit les objectifs aux seules compétences scolaires et ferme l'ouverture à la citoyenneté, à la future vie sociale et professionnelle. On peut s'attendre, compte tenu de notre expérience d'enseignants- chercheurs, à une structure double des choix :

- Facteur 1 : l'enseignement des mathématiques se présente comme une discipline institutionnelle conservatrice
- Facteur 2 : l'enseignement des mathématiques se présente comme une occasion de libre créativité, d'inventivité et de préparation à la vie citoyenne.

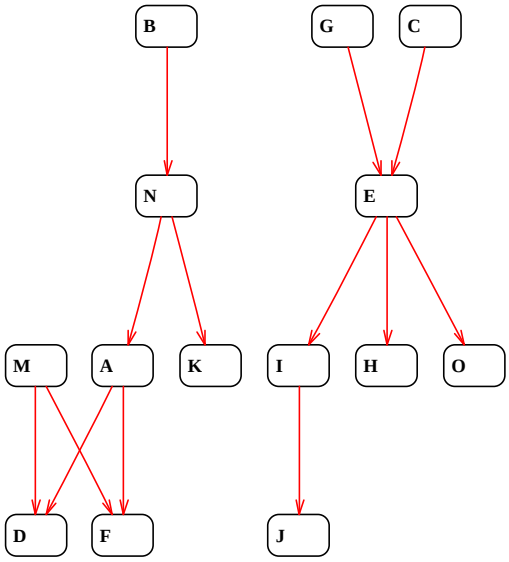
2.2 Structure implicative des choix :

On a demandé à CHIC de générer un graphe implicatif des items. Les variables supplémentaires correspondent aux différentes séries : S s pour série scientifique ; ES s pour série économique et sociale ; LI s pour classe littéraire ; TE s pour séries technico-professionnelles. On obtient la table d'implication ci-dessous (Tab. 2). Les intensités sont diversifiées et s'étendent de 0 à 100 pour le couple D - F.

item	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**	**N**	**O**
A	0	48	22	**77**	28	**77**	25	11	15	0	14	45	29	**65**	35
B	37	0	**58**	29	**51**	29	37	7	17	4	**95**	37	6	**79**	36
C	1	**59**	0	0	**99**	0	33	**62**	**54**	**78**	21	46	7	4	**63**
D	**72**	46	16	0	0	**100**	24	29	9	13	**50**	**51**	**66**	**74**	7
E	18	**54**	**86**	0	0	0	**79**	**74**	**97**	**95**	46	49	24	16	**97**
F	**72**	46	16	**100**	0	0	24	29	9	13	**50**	**51**	**66**	**74**	7
G	9	41	40	4	**83**	4	0	58	50	**87**	6	47	2	3	24
H	6	27	**58**	16	**72**	16	**58**	0	20	39	2	47	2	1	40
I	15	38	**54**	3	**91**	3	**53**	26	0	**69**	0	35	5	2	49
J	0	29	**61**	9	**84**	9	**72**	45	**68**	0	13	41	6	6	47
K	14	**74**	42	45	**50**	45	22	4	0	8	0	45	13	**92**	35
L	23	32	45	38	40	38	43	32	6	12	23	0	**58**	36	19
M	30	30	33	**67**	35	**67**	14	5	6	4	15	**55**	0	**58**	5
N	**67**	**65**	27	**82**	25	**82**	14	2	1	2	**94**	49	**56**	0	12
O	27	46	**59**	0	**95**	0	34	36	45	41	27	41	1	7	0

Tableau 2. Table d'implication.

En utilisant le mode classique pour calculer avec CHIC l'intensité d'implication, le graphe implicatif au seuil 0,66 (par la méthode entropique dont on connaît la sévérité) met en évidence une dichotomie entre deux groupes d'items, tout en excluant l'item L (Fig. 1).



Le graphe situé à droite renvoie explicitement au facteur 2 : l'enseignement des mathématiques se présente comme une occasion de libre créativité, d'inventivité et de préparation à la vie **citoyenne**.

G- développement de la capacité d'accepter des points de vue différents ou C- préparation à la vie civique et sociale impliquent E- développement de l'imagination et la créativité et de là soit O- participation au développement d'une culture générale, soit H- développement de la volonté et la persévérance soit I- développement de l'esprit critique qui implique J- développement de la capacité à communiquer avec objectivité, clarté et précision par des modes de représentation divers.

Le graphe situé à gauche renvoie au facteur 1 : l'enseignement des mathématiques se présente comme une discipline institutionnelle conservatrice :

B- préparation à la vie professionnelle implique N- acquisition de savoir-faire et de là soit K- développement de compétences utiles dans les autres disciplines, soit A- acquisition de connaissances qui implique F- développement de la capacité à prouver et valider sa preuve ou D- préparation aux examens, concours, au passage dans l'enseignement supérieur. M- développement de la capacité à mathématiser et à formaliser implique également soit F- développement de la capacité à prouver et valider sa preuve, soit D- préparation aux examens, concours, au passage dans l'enseignement supérieur.

On examine maintenant les typicalités des différents chemins (Tab. 3).

facteur	chemin	S	ES	LI	TE
2	C-E-I-J			X	
	C-E-O			X	X
	C-E-H			X	
	G-E-H		X		
	G-E-O		X	X	
	G-E-I-J				
1	B-N-A-D				X
	B-N-A-F				X
	B-N-K		X		X
	M-D		X	X	
	M-F		X	X	

Tableau 3. Typicalités des chemins**.**

S n’est typique d’aucun des chemins.

ES est typique de deux des départs G-E-x (F2) : développement de la capacité d'accepter des points de vue différents => développement de l'imagination et la créativité.

ES est typique de l’un des départs B-N-x (F1), soit B-N-K : préparation à la vie professionnelle => acquisition de savoir-faire => développement de compétences utiles dans les autres disciplines.

ES est enfin typique des départs M-x (F1) : développement de la capacité à mathématiser et à formaliser.

LI est typique des départs C-E-x (F2) : préparation à la vie civique et sociale => développement de l'imagination et la créativité.

LI est typique de l’un des départs G-E-x (F2), soit G-E-O développement de la capacité d'accepter des points de vue différents => développement de l'imagination et la créativité => participation au développement d'une culture générale.

Enfin, comme ES, LI est enfin typiques des départs M-x (F1) : développement de la capacité à mathématiser et à formaliser.

TE est typique des départs C-E-O (F2) : préparation à la vie civique et sociale => développement de l'imagination et la créativité => participation au développement d'une culture générale.

TE est typique des départs B-N-x : préparation à la vie professionnelle => acquisition de savoir-faire.

Il serait intéressant, et certainement enrichissant, d’étudier les relations contributives entre les modalités de choix des enseignants. Il suffirait pour cela de considérer, techniquement possible avec CHIC, les variables principales, comme des variables supplémentaires. Dans cet article, la place manque pour pousser l’étude plus loin.

2.3 Structures non-implicatives des choix

La question est ici de savoir si des analyses classiques font ressortir cette structure en deux grands facteurs. On utilisera successivement la classification hiérarchique, les nuées dynamiques et l'analyse en composantes principales.

2.3.1 La classification hiérarchique :

Par le moyen d'une classification hiérarchique utilisant la distance moyenne entre classes on obtient deux classes relativement proches du résultat de l'ASI (Fig. 2).

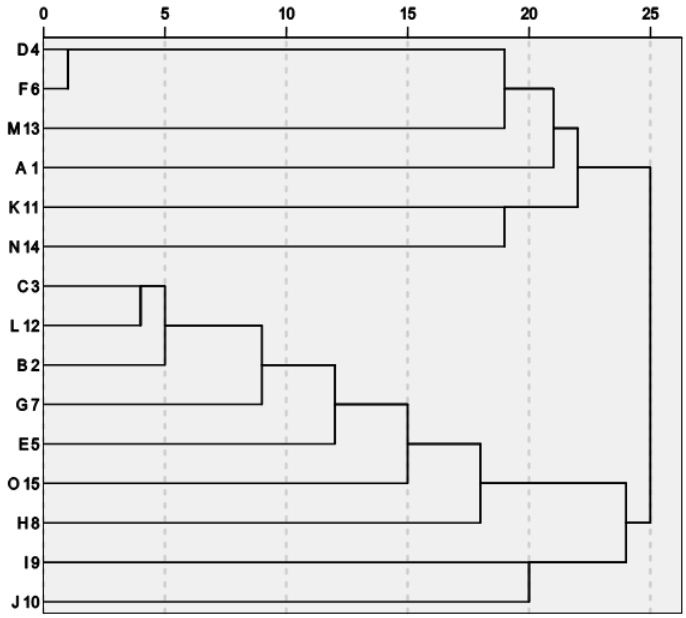


Figure 2. Arbre hiérarchique utilisant la distance moyenne entre classes.

Soit pour le facteur 2 : C- préparation à la vie civique et sociale, G- développement de la capacité d'accepter des points de vue différents, E- développement de l'imagination et la créativité, O- participation au développement d'une culture générale, H- développement de la volonté et la persévérance, I- développement de l'esprit critique et J- développement de la capacité à communiquer avec objectivité, clarté et précision par des modes de représentation divers. L- développement de la pratique de calculs formels, donc sans nécessité de signification non classé par l'ASI rejoint ce groupe et B- préparation à la vie professionnelle joue le transfuge.

Et soit pour le facteur 1 : D- préparation aux examens, concours, au passage dans l'enseignement supérieur, F- développement de la capacité à prouver et valider sa preuve, M- développement de la capacité à mathématiser et à formaliser, A- acquisition de connaissances, K- développement de compétences utiles dans les autres disciplines et N- acquisition de savoir-faire.

2.3.2 La classification en nuées dynamiques :

Là encore, pour la classification en nuées dynamiques, les classes se rapprochent de celles obtenues par l'ASI, à une variable près (Tab. 4).

items	classe 1	classe2
A	1	0
B	1	0
D	1	0
F	1	0
K	1	0
M	1	0
N	1	0
C	0	1
E	0	1
H	0	1
I	0	1
J	0	1
O	0	1
G	0	0
L	0	0

Tableau 4. Classification en nuées dynamiques.

On retrouve intégralement la classe 1. L n'est pas classée. Mais G non plus.

2.3.3 L'analyse factorielle en composantes principales :

H, L et G restent isolées. On extrait une structure en cinq facteurs, dont quatre bipolaires. Ces derniers opposent des items appartenant aux deux classes du graphe implicatif issu de l'ASI (Tab. 5).

items	Classes ASI	F1	F2	F3	F4	F5
D	1	0,99				
F	1	0,99				
E	2	-0,99				
K	1		0,73			
N	1		0,55			
I	2		-0,72			
A	1			0,80		
J	2			-0,70		
O	2				0,69	
M	1				-0,79	
B	1					0,80
C	2					0,48

Tableau 5. Analyse factorielle en composantes principales.

F1 *opposition logique, imagination* **:** F développement de la capacité à prouver et valider sa preuve et D préparation aux examens, concours, au passage dans l'enseignement supérieur *vs* E développement de l'imagination et la créativité.

F2 *opposition utilités, esprit critique* : K développement de compétences utiles dans les autres disciplines et N acquisition de savoir-faire *vs* I développement de l'esprit critique

F3 *opposition savoirs, communication* **:** A acquisition de connaissances *vs* J développement de la capacité à communiquer avec objectivité, clarté et précision par des modes de représentation divers

F4 *opposition culture générale, discipline* **:** O participation au développement d'une culture générale *vs* M développement de la capacité à mathématiser et à formaliser

F5 *préparation à la vie citoyenne* **:** B préparation à la vie professionnelle et C préparation à la vie civique et sociale

2.3.4 Introduction des variables supplémentaires (sections d'enseignement)

Pour chacune des sections d'enseignement, on a calculé la signification de la différence pour chacun des facteurs (Tab. 6).

	S	ES	LI	TE
F1				
F2	X -	X +		X +
F3				
F4	X -	X +	X +	
F5	X -			X +

Tableau 6. Résumé de l'ANOVA par section.

On ne relève aucune différenciation pour les facteurs F1 opposition logique, imagination et F3 opposition savoirs, communication.

Les enseignants des sections S, scientifiques, ont des scores différenciés pour les trois autres facteurs, ces sections présentant un score négatif sur ces facteurs, les autres sections un score positif. Pour F2 *opposition utilités, esprit critique* (-0,35 *vs* 0,35) ces enseignants valorisent le développement de l'esprit critique. Pour F4 *opposition culture générale, discipline* (-0,24 *vs* 0,24) ils valorisent le développement de la capacité à mathématiser et à formaliser. Pour F5 ils n'adhèrent pas à la *préparation à la vie citoyenne* (-0,15 *vs* 0,15).

Les enseignants des sections ES, économique et sociale, ont des scores différenciés sur les facteurs F2 et F4, y présentant des scores moyens positifs, à l'inverse des S. Pour F2 *opposition utilités, esprit critique* (0,37 *vs* -0,10) ces enseignants valorisent le développement de compétences utiles dans les autres disciplines et l'acquisition de savoir-faire. Pour F4 *opposition culture générale, discipline* (0,30 *vs* -0,08) ils valorisent la participation au développement d'une culture générale.

Les enseignants des sections LI, littéraire, ne se différencient que sur F *opposition culture générale, discipline* (0,39 vs -0,03) en privilégiant la participation au développement d'une culture générale.

Enfin, les enseignants des sections TE, technologiques, se différencient sur F2 et F5. Ils rejoignent les ES pour le facteur 2 *opposition utilités, esprit critique* (0,42 *vs* -0,11) valorisant le développement de compétences utiles dans les autres disciplines et l'acquisition de savoir-faire. Ils rejoignent également leurs collègues LI pour F5 opposition culture générale, discipline (0,46 *vs* -0,12) privilégiant la participation au développement d'une culture générale.

On note la spécificité des enseignants de la section scientifique S étroitement centrés sur les aspects formels *a contrario* de leurs collègues des autres sections plus ouvert sur la culture générale et plus largement sur la préparation à la vie professionnelle et citoyenne.

3 Conclusion

Ce papier tente de cerner et d'identifier les liens entre ASI et autres méthodes d'analyses en classes ou en facteurs. Par rapport aux données utilisées, on observe que chaque méthode de calcul pratiquée fait ressortir une opposition entre deux groupes d'items, soit sous forme de deux classes, soit sous forme de facteurs bipolaires. Cette quadruple approche méthodologique présente donc un vif intérêt alors que les outils d'analyse sont très différents. On peut inférer de cette partition en deux classes l'existence d'un processus mental qui opère ces classifications selon deux conceptions distinctes : préserver l'enseignement des mathématiques en tant que discipline institutionnelle conservatrice, conception caractérisant les professeurs de section S, gardiens du temple de l'orthodoxie scientifique ; utiliser le « prétexte » de l'enseignement des mathématiques comme occasion de libre créativité, d'inventivité et de préparation à la vie citoyenne. Mais n'y a-t-il pas variation du rapport institutionnel des enseignants en fonction de la situation (référence à la théorie anthropologique du didactique).

Toutefois, on ne sait pas si ce clivage fermeture / ouverture au principe de la relation d'appartenance à l'une des deux classes commande également la relation d'ordre implicatif structurant ces deux pôles. Il devient de plus en plus nécessaire de se donner les moyens d'avancer sur cette problématique des déterminants de l'appartenance et de l'ordre propres aux facteurs implicatifs

Références

[1] Bodin A., Gras R. (1999) : Analyse du préquestionnaire enseignants avant EVAPM-Terminales, *Bulletin n°425 de l'Association des Professeurs de Mathématiques de l'Enseignement Public, 772-786, Paris*

[2] Couturier R. et Ag Almouloud S. (2009), Historique et fonctionnalités de CHIC, *Analyse Statistique Implicative, Une méthode d'analyse de données pour la recherche de causalités, sous la direction de Régis Gras, réd, invités R. Gras, J.C. Régnier, F. Guillet, Cépaduès Ed. Toulouse p.279-293*

[3] Pasquier D. et Gras R., [2012], De l'intérêt de l'Analyse Statistique Implicative (A.S.I.) pour la recherche exploratoire en psychologie**,** *Psychologie Française, Elsevier-Masson***,** p 161-173

[4] Gras R., Régnier J.-C. [2013], Fondements théoriques de l'Analyse Statistique Implicative, *Méthode exploratoire et confirmatoire à la recherche de causalités,* sous la direction de Gras R., eds Gras R., Régnier

J.-C., Marinica C., Guillet F., Cépaduès Editions, 522 pages, ISBN 978.2.36493.056.8, p 25-186.

[5] Pasquier D. et Gras R., [2012], Analyse exploratoire a priori et analyse confirmatoire a posteriori : paradigme pour la comparaison de deux structures en analyse statistique implicative, *L'Analyse Statistique Implicative : de l'exploratoire au confirmatoire*, Eds J.C.Régnier, M.Bailleul R.Gras, Université de Caen, ISBN 978-2-7466-5256-9, p.245-261.

[6] Pasquier D., Gras R. [2013], Analyse Statistique Implicative et psychométrie, *Méthode exploratoire et confirmatoire à la recherche de causalités*, sous la direction de Gras R., eds Gras R., Régnier J.-C., Marinica C., Guillet F., Cépaduès Editions, 522 pages, ISBN 978.2.36493.056.8, p. 365-378.

Ouvrages de référence

[1] *L'implication statistique. Nouvelle méthode exploratoire de donnée*, sous la direction de R.Gras, et la collaboration de S. Ag Almouloud, M. Bailleul, A. Larher, M. Polo, H. Ratsimba-Rajohn, A.Totohasina, La Pensée Sauvage, Grenoble (1996)

[2] *Mesures de Qualité pour la Fouille de Données*, H.Briand, M.Sebag, R.Gras er F.Guillet eds, RNTI-E-1, Cépaduès, 2004

[3] *Quality Measures in Data Mining* , F.Guillet et H.Hamilton eds, Springer, 2007,

[4] *Statistical Implicative Analysis, Theory and Applications*, R.Gras, E. Suzuki, F. Guillet, F. Spagnolo, eds, Springer, 2008.

[5] *Analyse Statistique implicative.Une méthode d'analyse de données pour la recherche de causalités*, sous la direction de Régis Gras, réd. invités R. Gras, J.C. Régnier, F. Guillet, Cépaduès Ed. Toulouse, 2009.

[6] *Teoria y Aplicaciones del Analisis Estadistico Implicativo*, Eds : P.Orus, L.Zemora, P.Gregori, Universitat Jaume-1, Castellon (Espagne), ISBN : 978-84-692-3925-4, 2009..

[7] *L'Analyse Statistique Implicative : de l'exploratoire au confirmatoire*. Eds : J.C. Régnier, Marc Bailleul, Régis Gras, Université de Caen, ISBN : 978-2-7466-5256-9, 2012

[8] *L'analyse statistique implicative, Méthode exploratoire et confirmatoire à la recherche de causalités*, sous la direction de Gras R., eds Gras R., Régnier J.-C., Marinica C., Guillet F., Cépaduès Editions, 522 pages, ISBN 978.2.36493.056.8, 2013.

VARIABLE NODALE ET CONE IMPLICATIF

Dominique LAHANIER-REUTER[1], Régis GRAS[2], Marc BAILLEUL[3]

NODUL VARIABLE AND IMPLICATIVE CONE

RÉSUMÉ.

Dans le cadre de l'Analyse Statistique Implicative, il est procédé à l'extraction de règles implicatives pondérées par une mesure appelée intensité d'implication. Un graphe dit implicatif permet de représenter l'ensemble des relations non symétriques qui associent les variables. Au sein du graphe, on observe quelquefois, à partir d'une variable centrale, des relations amont et aval telles qu'il soit possible d'isoler ces deux types de part et d'autre de ce nœud, ce confluent, sous la métaphore de cône implicatif à deux grappes. Donner une signification à ce cône, sous la condition d'une certaine homogénéité et de la connexité de l'ensemble, permet de l'isoler conceptuellement du tout en des termes de causalité-conséquences. Dans cette communication, nous définissons selon deux approches un critère d'homogénéité d'un cône et des conditions d'existence de relations causales. Nous illustrons le propos théorique par l'examen de plusieurs exemples significatifs.

Mots-clés. *Analyse statistique implicative, règle, graphe, nœud, cône, sommet du cône, arête, cause et conséquence,*

ABSTRACT

Inside the Statistical Implicative Analysis frame, implicative rules may be extracted while pondered by a measure called Intensity of Implication. The set of the non-symmetrical relationships that associate the different variables may be represented by a graph so called implicative. Inside the graph, relationships upstream and downstream from a central variable parted on either sides of this node can be sometimes noticed. This situation of a confluent is metaphorically said a two clusters implicative cone. On condition that the set is rather homogeneous and linked, giving conceptual meaning to this cone allows some kind of causalities/consequences isolation of it from the whole set. In this communication, we define criteria for the homogeneity of a cone and conditions for existence of causal relationships. Some significant examples illustrate our theoretical discourse.

Keywords *Statistical Implicative Analysis, Rule, Graphic, Node, Cone, Edge, Top of the cone, Causes and consequences*

1 Introduction

Disposant d'un corpus de données croisant sujets et variables, l'Analyse Statistique Implicative offre des réponses à certains objectifs du chercheur :

– Extraire de cet ensemble de données où les sujets sont éléments d'un ensemble discret E et les variables sont de natures diverses, des règles implicatives par

[1] Équipe Théodile CIREL, Université Lille 3, Villeneuve d'Ascq, France, E-mail: dominique.lahanier@univ-lille3.fr

[2] École Polytechnique de l'Université de Nantes, Équipe DUKE Data User Knowledge, Laboratoire d'Informatique de Nantes-Atlantique (LINA), UMR 6241, : regisgra@club-internet.fr

[3] CERSE (Centre d'Etudes et de Recherches en Sciences de l'Education), Université de Caen Basse-Normandie, courriel : marc.baillleul@unicaen.fr

exemple du type : « si la variable a est instanciée alors généralement la variable b l'est aussi », dans le cas de variables booléennes,

- Affecter à chaque règle une valeur numérique de [0,1] croissante avec la qualité prédictive de la règle,
- Représenter par un graphe dit implicatif, non symétrique, pondéré et sans cycle, l'ensemble des règles de qualité au moins égale à un certain seuil d'acceptabilité,
- Interpréter en donnant du sens aux arêtes du graphe, mais aussi aux structures connexes qui le constituent.

C'est ainsi qu'il existe, pour un seuil d'intensité d'implication donné, certaines sous-structures connexes où des nœuds du graphe admettent isolément des antécédents (des « pères ») et des successeurs (des « fils »). Ces sous-structures sont extraites « à la main » par l'utilisateur lui-même au vu du graphe implicatif donné par CHIC. Antécédents et successeurs peuvent éventuellement être interprétés respectivement comme causes et conséquences du sommet du graphe dans une perspective causale.

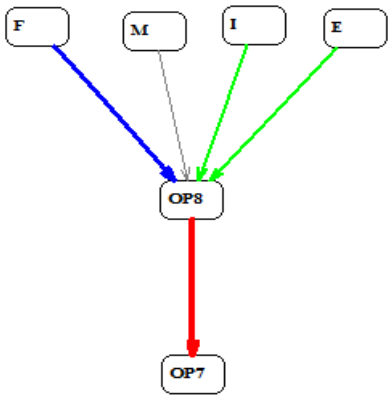


Figure 1

Par exemple, sur la figure 1, on a extrait, d'un graphe implicatif d'une quarantaine de variables et au seuil 0.69, la sous-figure ci-dessus que nous appelons *« cône implicatif »*, avatar géométrique car il admet un sommet et deux nappes dont l'amont et l'aval représentent respectivement les « pères » et les « fils ». Le sommet de ce cône, OP8, admet pour « pères » (pluriel, comme c'est bizarre !) ou « causes » les variables F, M, I et E. Il n'a qu'un « fils » ou « conséquence : OP 7. Ce sommet joue un rôle que l'on peut désigner comme « confluent » ou « variable nodale» de variables. Le mot « cône » a été retenu car il symbolise aussi, comme un entonnoir, un objet de type déversoir » où confluent divers affluents et d'où repartent de nouveaux réseaux absorbant les flux de l'amont. Jean-Marc Lévy-Leblond, dans (J.-M. Lévy-Leblond, 1996, p. 320) parlerait même d'un **champ causal** ou un écheveau enchevêtré qui « ne laisse plus individualiser les fils continus au long desquels pourrait se concevoir la propagation d'influences causales ».Au sujet des mots constitutifs de la phrase, François Gaudin dans (Sève, L., 2005) dirait que l'on voit « s'y dessiner une topologie du sens ».

Ce cône pouvant se prolonger le long de ses deux nappes, on constate que la notion de « cause » (ou de « pères ») et de « conséquence » (ou « d'effet » ou de « fils ») est toute relative. Un père n'est père qu'en tant qu'il a un fils. Nous rencontrons ici aussi le raisonnement logique de type dialectique où l'analyse a pour fonction la résolution de la contradiction. La cause détermine l'effet tout autant que l'effet détermine la cause (cf L.Sève, 2005, p.71). Son sens ne se construit que dans son rapport avec l'effet.

Notre intention est d'utiliser ce cône pour mettre en évidence un ensemble de causes non liées entre elles afin de pouvoir identifier leurs rôles, leurs responsabilités dans l'apparition de l'attracteur. Autrement dit, nous nous intéresserons de façon privilégiée aux cônes dont les « pères » du sommet sont **indépendants ou très faiblement dépendants**. Ce sommet pourrait définir une sorte de « bassin d'attraction » au sens de la théorie des catastrophes. Sous cette restriction *d'indépendance*, nous nous limiterons donc ici, un graphe étant donné comme image d'un ensemble de règles, à ce type de structure conique et à son interprétation. Mais celle-ci exige que ce cône ait des chances d'offrir une certaine consistance sémantique et, pour cela, que sa constitution présente des caractères analytiques révélateurs d'une certaine homogénéité implicative. *N'y observerait-on pas un phénomène comparable à celui que L. Sève désigne par « propriété émergente » ?* Cette problématique constituera notre premier paragraphe.

2 Indice d'homogénéité d'un cône implicatif

2.1 Rappels

Au cours de la formalisation de l'implication statistique, nous avons introduit un indicateur de qualité de la mesure d'une règle de type a=>b. Dans le cas des variables binaires, nous avons comparé les occurrences des contre-exemples à la règle à celles qui serait observables au hasard si les variables étaient indépendantes. Pour quantifier cette proposition, aux sous-ensembles A et B de E des observations de sujets effectives et respectives de a et b, nous avons associé des parties X et Y de mêmes cardinaux que A et B. En modélisant un tirage au hasard dans E des parties X et Y, nous avons comparé le cardinal aléatoire $N_{a\wedge\bar{b}}$ des contre-exemples à l'implication (soit $\text{card}(\overline{X\cap Y})$), au cardinal $n_{a\wedge\bar{b}}$ (soit $\text{card}(\overline{A\cap B})$) des contre-exemples observés. L'intensité d'implication qui mesure la qualité de l'implication est alors la probabilité que le hasard conduise à plus de contre-exemples que la contingence soit :

$$\varphi(a,b) = = \Pr\left[\text{Card}(X\cap\overline{Y}) > n_{a\wedge\bar{b}}\right] = 1-\Pr\left[\text{Card}(X\cap\overline{Y}) \le n_{a\wedge\bar{b}}\right] \quad (1)$$

Si l'on centre et l'on réduit la variable aléatoire $\text{Card}(X\cap\overline{Y})$ sous la forme :

$$Q(a,\bar{b}) = \frac{\text{Card}(X\cap\overline{Y}) - \frac{n_a\, n_{\bar{b}}}{n}}{\sqrt{\frac{n_a\, n_{\bar{b}}}{n}}} \quad (2)$$

Celle-ci dans des conditions à grands effectifs est une variable gaussienne N(0,1) en tant que valeur approchée d'une variable aléatoire de Poisson centrée et réduite (cf.

Partie 1, Chapitre 1 de Gras et Régnier, 2013). Sa réalisation contingente est $q(a,\bar{b})$. Par suite, l'intensité d'implication devient dans ce cas gaussien :

$$\varphi(a,b) = 1 - \Pr\left[Q(a,\bar{b}) \le q(a,\bar{b})\right] = \frac{1}{\sqrt{2\pi}} \int_{q(a,\bar{b})}^{\infty} e^{-\frac{t^2}{2}} dt \quad (3).$$

2.2 Rôle fonctionnel des intensités d'implication des arêtes

Une première approche pourrait consister à déterminer la *variance implicative* intra-classe du cône implicatif (cf. Partie 2, Chapitre 2 de Gras et Régnier, 2013). Cette variance du cône, basée sur une interprétation géométrique de l'implication statistique, offre une mesure qui permet d'apprécier la consistance d'une grappe connexe du graphe comme l'est un cône implicatif. L'absence de programmation des calculs associés dans CHIC nous incite, puisque nous visons des calculs conduits à la main, à retenir une autre méthode plus intuitive et aisée sur le plan calculatoire.

Une deuxième approche, plus simple et plus intuitive, consiste à examiner la consistance d'un tout à partir de celles de l'ensemble de ses arêtes, même si un tout est plus informatif que l'ensemble de ses parties. Ainsi, un cône implicatif étant donné, une mesure naturelle pour en qualifier la qualité d'homogénéité se nourrit de l'ensemble des qualités implicatives des arêtes. Il est alors possible de mesurer cette qualité à travers, par exemple, la moyenne arithmétique ou la moyenne géométrique (afin de valoriser une intensité nulle pour viser un rejet) de toutes les intensités d'implication des arêtes. Cependant, cette méthode simple a le défaut de ne pas accorder à la mesure un caractère relatif comme nous l'obtenons pour l'intensité $\varphi(a,b)$ qui est une probabilité.

Aussi, nous faisons appel à une troisième approche qui a la vertu épistémologique de s'harmoniser avec la philosophie de la modélisation employée pour évaluer la qualité d'une quasi-implication. Comme l'intensité d'implication, elle présente aussi l'avantage de fournir une échelle de mesure (une probabilité, nombre de [0 ,1]), c'est-à-dire une valeur ordinale où le « zéro » sera signe d'une hétérogénéité maximale du cône et le « un » sera signe d'une excellente homogénéité dissymétrique ou d'un flux implicatif régulier et riche dans ses liaisons.

Associons à chacune des p arêtes une variable aléatoire relative à l'implication a => b ; on la note $Q(a,\bar{b})$ (formule (2) du §2.1). Elle conduit, nous l'avons vu, à une mesure de la qualité de l'implication entre les deux extrémités de l'arête a ---> b. Notons de façon générique, Q_i , i = 1 à p, l'indice d'implication aléatoire dont la réalisation est q_i. Nous savons que les p variables Q_i , supposées indépendantes, suivent approximativement la loi normale N(0,1). Or, ces p variables indépendantes, suivant la même loi satisfont, en conséquence, les hypothèses du théorème central limite.

Posons $S_p = \sum_{1}^{p} Q_i$ et sa forme centrée-réduite $T_p = \frac{S_p - E(S_p)}{\sigma(S_p)}$ où E(Sp) est l'espérance de S_p, somme des espérances des variables Q_i qui sont toutes nulles et $\sigma(S_p)$ est l'écart-type de S_p, soit $\sqrt{p}$.

Par suite, $T_p = \frac{S_p}{\sqrt{p}}$ converge en loi vers la variable aléatoire de loi N(0,1). Pour une valeur de p convenable, par exemple $p \geq 4$ [4], T_p est donc approximable par la variable gaussienne centrée réduite N(0,1).

Comparant cette valeur attendue à la valeur observée $t_p = \frac{\sum_{1}^{p} q_i}{\sqrt{p}}$, nous obtenons alors un indice probabiliste pour mesurer, sur son versant d'entrée, la qualité du cône à savoir : $Pr[T_p \geq t_p] = \frac{1}{\sqrt{2\pi}} \int_{t_p}^{\infty} e^{-\frac{u^2}{2}} du$

3 Recherche des conditions de causalités sans liaison entre elles

Si l'on veut identifier les « pères » du sommet du cône comme étant des causes ne dérivant pas de phénomènes causaux communs ou proches, comme le seraient par exemple « il pleut » et « le sol est mouillé », il nous faut exiger que ces variables amont ne soient pas liées ou ne le soient que faiblement. Si tel est le cas, on pourra affirmer que telle variable (par ex. le sommet du cône en l'occurrence) dépend essentiellement de telles causes séparées, tout au moins dans les conditions expérimentales sources des données traitées. Ces données qu'elles soient quantitatives ou qualitatives, pour un traitement avec CHIC, sont numériques ou binaires ou décimales. Envisager une étude la corrélation linéaire de Bravais-Pearson est donc justifié puisque le rho de Spearman n'aurait été indiqué que si nous avions conservions les rangs des valeurs des variables. Or, dans l'exemple initial, la relation d'ordre des variables ordinales –des préférences- a été commuée en relation d'ordre sur R. Aussi, nous écartons le rho de Spearman.

Un test de significativité du coefficient de corrélation linéaire peut alors être joint à l'analyse interprétative pour réfuter la liaison entre deux variables amont. Par exemple, utilisant ce test, si le nombre de sujets est 150, on dira que la liaison est significative au risque de 5% si on observe une corrélation r supérieure à 0.16. Mais si r = 0.10, on admettra l'absence de liaison entre les deux variables considérées. Il est possible aussi de pratiquer un test de corrélation partielle[5] entre d variables « pères ». Pour d= 5, le risque de 5% est atteint pour une corrélation partielle de 0.164.

Il en est de même pour les « fils » en aval du sommet qui seront désignées comme conséquences non liées entre elles de ce sommet.

[4] Pour un nombre d'arêtes plus petit, l'indice envisagé risque de n'avoir qu'un rôle indicateur de tendance.

[5] Rappelons que la corrélation partielle se calcule de façon récurrente. Par exemple, si d=3 : $r_{1,2,3} = \frac{r_{1,2} - r_{1,3}\, r_{2,3}}{\sqrt{(1-r_{1,3}^2)(1-r_{2,3}^2)}}$. Pour obtenir $r_{1,2,3,4}$, on remplace dans la formule précédente les corrélations simples par les corrélations partielles.

4 Retour sur l'exemple initial

4.1 Les données

Reprenons l'exemple du § 1 Fig. 1. Ce cône est obtenu par extraction au sein d'un graphe implicatif de données issues d'un questionnaire présenté à des professeurs de mathématiques de classes terminales de lycée (17-18 ans). Nous avons recueilli et analysé (Bodin et Gras., 1999) les réponses de 311 professeurs, à des classements attendus (de 1 à 6) portant sur quinze objectifs que ces professeurs assignent à leur enseignement (A, B, C, …O)[6] et sur leurs opinions relatives à dix phrases susceptibles d'être communément énoncées (OP1, OP2,..OPX)[7]. Les 26 variables correspondantes ne sont pas binaires, 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.

La consigne était la suivante :

« A votre avis, quels sont les objectifs essentiels de la mission d'un professeur de mathématiques dans la série pour laquelle vous répondez. Pour répondre à cette question, classez par ordre préférentiel décroissant de 1 à 6 (1 : le plus important,...) six des objectifs majeurs de cette formation en les choisissant parmi les objectifs proposés ci-dessous :

A- acquisition de connaissances

……

E- développement de l'imagination et la créativité

F- développement de la capacité à prouver et valider sa preuve…….

I- développement de l'esprit critique…….

M- développement de la capacité à mathématiser et à formaliser……..

Puis, Voici quelques opinions entendues dans la salle des profs. Vous pouvez être d'accord, ou un peu d'accord ou pas d'accord avec l'une ou l'autre. Entourez votre choix :

1-(OP1) C'est vrai que les math constituent un instrument de sélection excessif.

D'ACCORD UN PEU D'ACCORD PAS D'ACCORD

……..

A la sortie de la terminale de la série sur laquelle vous répondez, un élève devrait....

7-(OP7) ....pouvoir reconnaître si un nombre entier écrit dans la base 10 est divisible par 4.

D'ACCORD UN PEU D'ACCORD PAS D'ACCORD

8-(OP8) ....pouvoir 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".

D'ACCORD UN PEU D'ACCORD PAS D'ACCORD

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

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

4.2 L'homogénéité du cône

Le logiciel CHIC nous fournit les valeurs des intensités d'implication des variables en jeu dans le graphe de la Fig. 1. De ces valeurs de l'intensité d'implication qui sont des probabilités donc des nombres de [0,1] affichées par CHIC, nous pouvons remonter à la valeur q_i prise par Q_i où i est l'un des couples qui nous intéressent du cône implicatif. Il suffit de lire la table de la loi normale N(0,1) de façon inverse.

Règle i	**Intensité d'implication φ_i**	**Indice d'implication observé q_i**
F -> OP8	0.96	- 1.75
M -> OP8	0.86	- 1.08
I -> OP8	0.93	- 1.48
E -> OP8	0.91	- 1.34
F -> OP7	0.81	- 0.88
M -> OP7	0.61	- 0.28
I -> OP7	0.75	- 0.68
E -> OP7	0.81	- 0.88
OP8 -> OP7	1	- 4.00

Tableau 1

Appliquons les relations établies dans le 2.2 pour p = 9, $t_9 = \frac{\sum_{1}^{9} q_i}{\sqrt{9}} = -4.13$

D'où $Pr[T_9 \geq t_9] = \frac{1}{\sqrt{2\pi}} \int_{t_p}^{\infty} e^{-\frac{u^2}{2}} du \cong 1$

Ceci signifie que la qualité d'homogénéité sémantique et dissymétrique du cône, eu égard à notre modèle aléatoire, est excellente puisqu'une distribution au hasard des indices d'implication conduirait à des sommes de Q_i bien supérieures à celle qui est observée à savoir – 4.13. Autrement dit, le « flux implicatif » qui traverse le cône respecte bien la dissymétrie attendue de l'implication et cela dans le même sens puisque les valeurs des q_i sont toutes négatives.

Remarquons, en passant, que la liaison OP8 => OP7 donne du sens à OP7 comme engageant la réflexion personnelle et critique de l'élève et non pas l'application d'un critère appris en cours puisque c'est précisément ce qu'exprime l'objectif OP8. La compétence de l'élève qui lui est associée apparaît comme « émergente » de la conjugaison des objectifs qui seraient réalisés en amont. Cette interprétation est renforcée par la nature de I, une des « causes » amont où l'esprit critique est explicitement visé au lycée. Cette remarque illustre le propos tenu dans l'avant-dernier paragraphe du § 1.

4.3 Les liaisons entre les « causes » et entre les « conséquences »

Dans cet exemple, seules les éventuelles « causes » sont plurielles : F, M, I et E. Mais nous n'avons qu'une conséquence : OP7. Les coefficients de corrélation linéaire - à la limite d'applicabilité en raison des faibles effectifs en jeu- entre les variables « pères » sont les suivantes (toujours données par CHIC) :

r (F ; M) = 0.13 ; r (F ; I) = - 0.04 ; r (F ; E) = 0.1 ;

r (M ; I) = - 0.1 ; r (M ; E) = - 0.07 ; r (I ; E) = 0.15

Faisant référence à la table des valeurs critiques du coefficient de corrélation d'un échantillon issu d'une population normale, nous constatons qu'un coefficient au hasard R, pour tous les couples en jeu, dans le cas d'une liaison entre ses termes, est supérieur à 0.13 avec un risque de 5%. Ceci est le cas pour (F et M) et (I et E), mais non pour les autres couples dont l'affirmation d'une liaison conduirait à un risque supérieur à 10%. On peut donc faire l'hypothèse d'ordre cognitif que deux causes fondamentales se conjuguent pour impliquer OP8 puis OP7 : l'un de type non normatif dans le cadre scolaire (imagination E, esprit critique I), l'autre de type cartésien (preuve F, formalisation M), à l'intérieur desquels des modalités seraient liées. L'examen des tâches nécessaires à la résolution de OP8 (exemple, contre-exemple, concepts mathématiques) montre que celle-ci nécessite la mise en commun des deux types causaux identifiés en tant que « pères ».

Cependant, l'examen des corrélations partielles montre que admettre l'absence de liaison globale reste légitime et donc que les « causes » présentent des caractères spécifiques non inclusives. En effet, dans les limites de seuil de 5% (0.164) et même à 10% (0.137), on obtient des corrélations partielles en permettant la validation :

r(F, M, I) = 0.127 ; r(M, I, E)) _0.092 ; r(F, I, E) = - 0.056 ; r(F , M, E) = 0.143 ;

et finalement r(F, M, I, E) = 0.131.

5 Un deuxième exemple

Relations et Co variations

D'un point de vue statistique la problématique principale que nous abordons est celle de la définition des liens statistiques entre variables lorsqu'ils sont déduits des mises en relation de leurs variations. Il s'agit de donner un sens statistique aux questions suivantes : si les valeurs d'une des variables différencient deux groupes de sujets (si ces groupes de sujets sont affectés par un changement de valeurs de cette variable ou par une modification importante de ces valeurs) en est-il de même pour l'autre variable? Et ces changements sont-ils comparables en importance ? Sont-ils significatifs ou négligeables ? etc.

Nous proposons de différencier les réponses à ces questions selon un critère particulier, celui de la symétrie exigée ou non des mises en relation des variations des variables. Ainsi nous distinguons les liens statistiques dont la définition inclut une exigence de réciprocité entre les variations des variables intéressées, et ceux dont la définition n'accorde pas le même statut aux variables considérées : les variations de la seconde (disons « impliquée » ou « seconde » ou « fils ») ne sont étudiées qu'au regard des variations de la première (disons le « père » ou l' « initiale »).

Ainsi nous comparons des traitements « classiques » qui permettent d'établir des relations symétriques entre variables tels ceux que le coefficient de corrélation linéaire r

permet (ou le chi²) aux traitements issus de l'ASI, qui permettent au contraire d'établir des dissymétries.

Ces problématiques et les décisions que nous venons d'évoquer sont inscrites dans le champ de la statistique. Notre projet est de montrer ici l'intérêt qu'elles peuvent présenter dans le champ des didactiques disciplinaires ou plus largement, dans celui des sciences de l'éducation. Pour ce faire, il est nécessaire de considérer l'interprétation ou les points de la modélisation qui posent problème d'un point de vue didactique.

Il a été dit plus haut que la problématique des liens entre variables reposait sur l'identification de modifications, de changements de valeurs d'une des variables sur un *groupe* de sujets. Par conséquent, les changements de valeurs à étudier sont plutôt à considérer du point de vue d'un groupe de sujets et non pas d'un individu isolé.

Ceci nous amène à interroger les groupes pertinents pour les études didactiques. Nous envisageons ici quelques-uns des groupes pertinents du point de vue didactique. Pour que cette pertinence soit reconnue, nous ne relevons que des groupes relativement stables afin que leurs frontières soient identifiables ainsi que leurs variations. Il nous semble que parmi ces groupes font consensus les groupes classes, les groupes de niveaux scolaires, les groupes de classes disciplinaires (nous pensons par exemple aux groupes d'élèves pour les enseignements de langues). Il nous semble également que sont moins consensuels, au regard toujours de leur pertinence pour les didactiques, les groupes identifiés par des choix pédagogiques revendiqués des enseignants ou encore les regroupements d'élèves selon leur « catégorie sociale »[8]. C'est pourquoi nous avons choisi dans ce cadre, de présenter une étude dans laquelle ces derniers regroupements seront l'objet d'une attention particulière.

5.1 La recherche « conscience disciplinaire »

Nous disposons d'un vaste corpus, construit à partir de la passation de trois questionnaires. Ces questionnaires ont été conçus et analysés dans le cadre d'une recherche en didactiques, dirigée par Cora Cohen-Azria, qui a donné depuis lieu à un ouvrage collectif. Cette recherche a pour but de questionner ce que nous avons appelé la « conscience disciplinaire » d'élèves de CM1 et CM2. Pour le dire très rapidement, nous appelons « conscience disciplinaire » les façons dont chaque élève identifie (ou ignore) les disciplines scolaires, les façons dont il se munit pour identifier ces différents espaces disciplinaires, les façons dont il se reconnait sujet de ces espaces et y agit. Ainsi par exemple, les façons de reconnaître dans quelle discipline on travaille – ce qui n'est pas toujours aussi évident que l'on peut le croire – ce qui y est valorisé et ce qui y est dénigré, ce qui y est possible, ce qui ne l'est pas, sont autant de dimensions de cette conscience disciplinaire : déclarer qu'en Français « ce que le prof attend c'est surtout du bla bla **bla** » tandis qu'en Mathématiques « c'est surtout un raisonnement correct », est une trace de la conscience disciplinaire du Français et des Mathématiques (en tant que disciplines scolaires) du lycéen interrogé. Durant cette recherche, nous avons tenté d'identifier certaines dimensions de ces consciences disciplinaires, dans quatre disciplines : le Français et les Mathématiques (disciplines scolaires reconnues comme « très importantes » et « très présentes »), les Sciences et l'Histoire-Géographie (qui le

[8] Guillemets, car ce sont plutôt celles de leur famille, avec tous les risques que cela comporte quant à l'interprétation du terme famille.

sont sans doute un peu moins). Nous avons choisi de les étudier auprès d'élèves de l'école primaire, pour plusieurs raisons : celle tout d'abord de l'organisation institutionnelle, en France, des disciplines scolaires, qui varie fortement entre l'école maternelle, l'école primaire, le collège et les divers types de lycées (nous mettons en note quelques différences). Les autres raisons tiennent à l'ancrage du travail didactique de notre équipe de recherche (Théodile CIREL) dans l'exploration des effets des modes de pédagogies alternatives (références) ainsi que dans l'intérêt que nous avons vu à explorer ces degrés de conscience disciplinaire à un moment du curriculum « charnière ».

C'est pourquoi ces questionnaires ont été soumis à quelques trois cents élèves de CM1 et CM2. Ces trois questionnaires les interrogent sur les façons dont ils dénomment, identifient, reconstruisent… les disciplines scolaires choisies: le Français, les Mathématiques, les Sciences et l'Histoire-Géographie (en note, nous expliquons que quelques questions plus globales sont importantes aussi). Ils sont pratiquement uniquement constitués de questions ouvertes.

Nous nous focalisons ici sur les variables issues de l'étude de quatre questions :

(1) Qu'est-ce que tu as appris d'important en Mathématiques ? (Questionnaire 1)
(2) Qu'est-ce que tu as appris d'important en Français ? (Q1)
(3) Qu'est-ce qui est important pour réussir en Mathématiques ? (Q3)
(4) Qu'est-ce qui est important pour réussir en Français ? (Q3)

Nous avons relevé les expressions/ locutions (terme à fixer) des réponses des élèves (présents à Q2 et à Q3). Et nous avons défini ainsi un ensemble de 51 variables binaires : elles valent 0 ou 1 selon que cette expression/locution a été employée par l'élève dans la 1e, 2e, 3e ou 4e question. Par exemple la variable FIOrtho correspond à la question (2) (F pour Français, I pour Important Appris) et à l'usage du terme « orthographe » dans la réponse de l'élève.

5.2 L'importance du flux implicatif et de la dissymétrie

Nous avons fait passer ces questionnaires dans des établissements implantés dans des quartiers « socialement défavorisés », et dans d'autres qui le sont au contraire dans des zones « socialement favorisées ». Et de fait, savoir s'il est possible, à partir des origines ou des milieux sociaux des élèves d'induire des conséquences quant à leurs représentations de l'école et des disciplines scolaires est une question très étudiée et particulièrement polémique. Les discours d'opinion abondent sur ce problème récurrent qu'ils tranchent dans un sens ou dans un autre. Du point de vue scientifique, en sociologie de l'éducation, dans les différentes didactiques (et peut-être plus particulièrement en didactique du français), les références sont nombreuses également (voir les travaux du groupe ESCOL, ceux de Lahire…). Formulons autrement notre question : est-ce que les expériences sociales possibles et effectives des *enfants* sont déterminantes pour leurs compréhensions, lectures de l'organisation de l'école où ils sont *élèves* ? Cohen-Azria, Lahanier-Reuter et Reuter, 2013).

Reprenons les termes que nous venons d'utiliser : « induire des origines sociales des conséquences sur les représentations ». Ceci recouvre quelques questions suivantes :

- est-ce que le type de quartier est considéré dans cette question comme un facteur/déclencheur de certaines dimensions des représentations? (corrélée ou « cause », « père » dans le schéma implicatif à des locutions/termes qui seraient distincts : par exemple, « être dans un milieu défavorisé » (est corrélé à) implique que/ a pour conséquence « dire que pour réussir en maths, il faut savoir compter » et « pour réussir en Français il faut bien écouter », tandis que « être dans un milieu favorisé » (est corrélé à) implique que/ a pour conséquence « dire que pour réussir en maths, il faut faire des exercices, s'entrainer » et « pour réussir en Français il faut bien écouter »). Le type d'établissement est alors « antécédent », « père » de certaines dimensions.
- est-ce qu'à un type d'établissement correspond un certain type de représentation ? (presque une fonction, donc des relations corrélées/implicatives comme ci-dessus, mais l'exigence de ne relever aucun terme/locution partagée par des élèves de milieux différents./pas de consensus possible, pas d'attitude commune entre ceux qui appartiennent à des types d'établissements différents).
- est-ce qu'un type d'établissement est particularisé/singularisé par plusieurs dimensions des représentations ? Ce qui est assez différent de ce qui précède. Dans ce cas, ce sont certaines dimensions qui seront propres à des élèves de ce type d'établissements : elles ne seront pas communes à tous les élèves de ces derniers, mais ne se relèveront que dans ceux-ci (tout cela doit être tempéré par « statistiquement »). Par exemple, le fait de reconnaître la lecture comme moyen de réussite en français est un discours que seuls tiennent des élèves des établissements d'un type particulier. Pour l'analyse que nous menons, ces dimensions apparaîtront alors comme des « pères », des antécédents des types d'établissement. La situation est l'inverse de la première.

Nous avons donc interrogé les discours des élèves qui ont été recueillis sous cet angle : les caractéristiques de ces discours, autour des contenus importants enseignés en Mathématiques et en Français et autour des contenus importants pour réussir dans ces deux disciplines, sont confrontées aux inscriptions des établissements scolaires qu'ils fréquentent, dans des quartiers « socialement favorisés » ou au contraire « socialement défavorisés ». A nos variables précédentes s'en ajoutent donc deux : « Favorisé » et «Très Défavorisé », qui sont encore des variables binaires. Les cônes implicatifs sont centrés tour à tour sur ces deux variables.

5.2.1 Les établissements situés dans des quartiers « très défavorisés »

Les facteurs « pères »

Commençons par le cône implicatif qui apparaît sous CHIC au seuil .60 centré sur la variable «Très Défavorisé ».

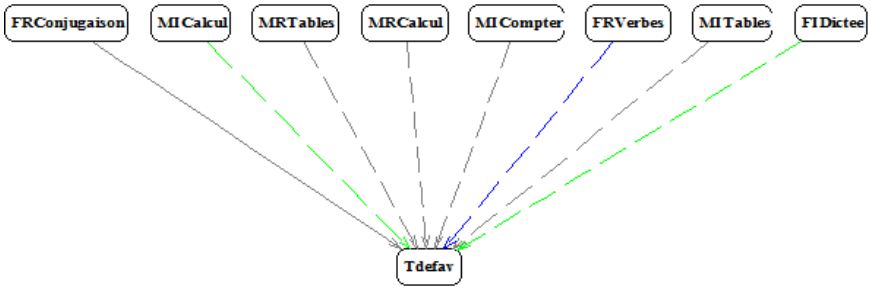


Figure 2 : Cône implicatif au seuil .60 centré sur la variable « Très défavorisé »

Nous constatons que le fait d'être inscrit dans un établissement (public) situé dans un quartier «Très défavorisé » n'a **aucune *conséquence* implicative**, mais apparaît comme une « conséquence » implicative.

Il semble donc que dans les établissements étudiés, situés dans des quartiers très défavorisés, *on puisse identifier des discours d'élèves qui sont propres à ces établissements.*

- Les contenus importants en Mathématiques sont compter/calculer/les tables de multiplication (MICompter, MICalcul, MITables);
- Il est important pour réussir en Mathématiques de [savoir] les tables de multiplication (MRTables) ;
- Les contenus importants en Français sont [faire] des dictées (FIDictées);
- Il est important pour réussir en Français de [savoir] les verbes et/ou les conjugaisons (FRVerbes, FRConjugaison).

Ces discours, spécifiques (au sens statistique du terme) des établissements étudiés situés dans des quartiers très défavorisés, sont cohérents. Ils mettent en évidence des constructions des deux disciplines essentiellement centrées sur des contenus :

- emblématiques de l'école primaire (compter, calculer, les tables de multiplication ou les dictées, les verbes et les conjugaisons) ;
- ancrés dans les représentations communes ;
- utiles ;
- gérables/contrôlables par les parents ou l'entourage ;

Pour le dire très brièvement, ces traits sont ceux d'une conscience disciplinaire qui serait celle du « certificat d'études », c'est-à-dire d'un enseignement non basé sur des sous-disciplines, destinés à des enfants dans lesquels on lit l'adulte, les futurs travailleurs et non pas de futurs collégiens par exemple.

Les corrélations partielles

Si nous considérons à présent les corrélations partielles entre ces différents facteurs (voir Annexe), seules les variables (MRCalcul, MRTables) et (FRConjugaison, FRVerbes) peuvent être considérées comme couples de variables dépendantes au seuil

de risque de 5%. Remarquons tout d'abord que ces couples de variables concernent des réponses distinctes aux deux questions :

- qu'est-ce qui est nécessaire selon toi pour réussir en Mathématiques ?
- qu'est-ce qui est nécessaire selon toi pour réussir en Français ?

Il n'y a pas trace de corrélation entre des réponses concernant des disciplines scolaires différentes, ce qui conforte notre hypothèse de conscience disciplinaire, c'est-à-dire du poids des disciplines scolaires dans les « attitudes » des élèves. Plus précisément, les deux conditions à la réussite en Mathématiques telles qu'elles sont principalement énoncées par des élèves des établissements très défavorisés (la maîtrise du calcul *vs* la connaissance des tables de [multiplication]) seraient des conditions exclusives[9] (r = -0,23). Cette tendance à l'exclusion peut être interprétée comme l'indice de deux modes d'appréhension de la réussite, de ce qui est attendu par l'enseignant, de ce qui est évalué en mathématiques par ces élèves. Y aurait-il une différence entre concevoir les Mathématiques comme l'art de calculer (au sens de « jongler avec les chiffres) et les résumer par un « slogan » (il faut savoir [ses] tables) ? Ou n'est-ce là qu'une caractéristique du discours des élèves, qui en quelque sorte « choisissent » l'un des termes (calculer ou les tables) de façon tendanciellement exclusive tant ils leur paraissent synonymes ou redondants ?

Il est plus simple nous semble-t-il d'interpréter la corrélation positive entre les deux facteurs de réussite en Français, [maîtriser] les conjugaisons et [connaître/savoir] les verbes, qui sont en effet parfois rapprochés dans le discours des élèves. Certains d'entre eux choisissent donc de considérer ces deux termes comme non redondants, non synonymes, mais plutôt sans doute comme complémentaires.

5.2.2 Les établissements situés dans des quartiers favorisés

Considérons à présent les établissements situés dans des quartiers considérés comme plutôt favorisés.

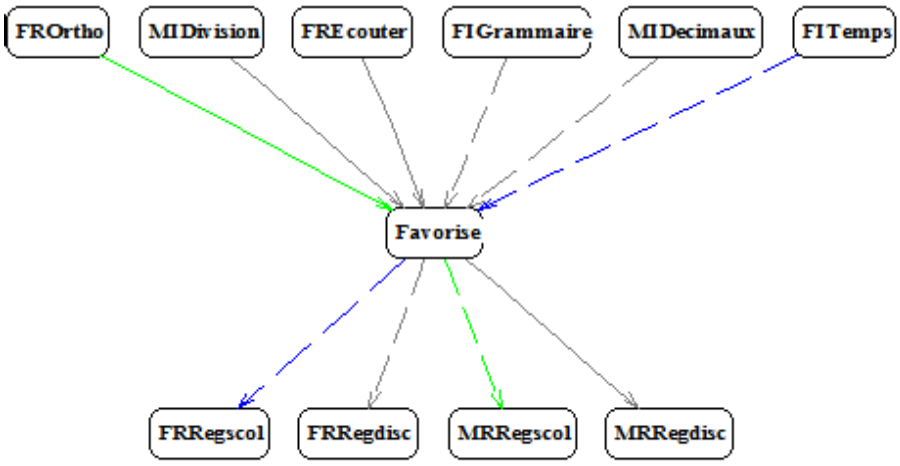


Figure 3 : Cône implicatif à .60 centré sur la variable « Favorisé »

[9] Ou des façons de dire exclusives plutôt.

Cette fois, les établissements « favorisés », centre d'un bassin d'attraction, sont à la fois des « pères » pour certaines dimensions et des « fils » pour d'autres (on retrouve la relation dialectique où la cause détermine l'effet et réciproquement, voir plus haut). Commençons par les discours qui sont ceux d'élèves de ces établissements, et d'eux (statistiquement) uniquement :

- Les contenus importants en Mathématiques sont la division et les nombres décimaux (MIDivision, MIDécimaux) ;
- Pour réussir en mathématiques… (il n'y en a pas de cités) ;
- Les contenus importants en Français, c'est la grammaire et les temps (FIGrammaire, FITemps) ;
- Pour réussir en Français, [il faut bien] écouter (FREcouter).

Par conséquent, les contenus qui sont désignés comme importants en Mathématiques sont des contenus qui sont *nouveaux* (au CM1 et au CM2), la division et les nombres décimaux. La perception de cette nouveauté présuppose son ancrage dans un apprentissage continu, ce qui particularise sans doute la conscience disciplinaire de ces élèves en ce qui concerne les Mathématiques. Ceux qui le sont en Français sont pour l'un, une sous-discipline (la grammaire) pour l'autre une désignation (scolaire?), les temps.

Les discours de ces élèves d'établissements (publics) « favorisés » privilégient parfois des contenus disciplinaires qui portent la marque de l'école: apprendre des temps (peu employés parfois hors de l'école, comme le passé antérieur), la grammaire, la division (qui est vraiment un savoir nouveau et complexe), les nombres décimaux (qui ne sont pas toujours les plus usités à l'extérieur de cette école, pensons par exemple à l'écriture 0,9764). Nous avançons, avec beaucoup de prudence, comme précédemment, l'idée que ces discours sont les traces de consciences disciplinaires qui seraient de l'ordre de l'apprentissage de contenus académiques, dont la maîtrise inscrit son possesseur dans une trajectoire/curriculum.

Cette fois, deux variables seulement sont corrélées positivement, au seuil de risque de 5% : Le fait de citer les décimaux comme contenus importants vus en Mathématiques et celui de citer les Temps comme contenus importants vus en Français : une vision globale des savoirs disciplinaires conforme à celle des programmes ?

5.2.3 Comparaison

Si nous comparons les « antécédents » mis au jour des deux types d'établissements (favorisés et très défavorisés), il nous semble qu'ils montrent des différences sensibles. Identifier des sous-disciplines et non pas des objets, identifier les savoirs nouveaux comme importants, ceux qui sont bien évalués (la maîtrise de la division) et non pas ceux qui sont intéressants à maîtriser (bien compter, savoir ses tables de multiplication) et que l'entourage ou l'opinion commune désigne comme tels seraient des positions (?) qui permettraient de parier à presque coup sûr sur l'appartenance à un établissement inscrit dans un quartier « favorisé ».

En ce qui concerne les « conséquences » de l'inscription dans un établissement « favorisé », elles présentent des aspects similaires, qui concernent les conditions de réussites :

- Que ce soit en Français ou en Mathématiques, les élèves de ces établissements s'entendent (statistiquement toujours) sur le fait que l'on réussit en obéissant à des règles scolaires : [bien] écouter/ ne pas faire de bêtises/faire ce que le maître/la maîtresse dit de faire etc.
- Que ce soit en Français ou en Mathématiques, les élèves de ces établissements s'entendent (statistiquement toujours) sur le fait que l'on réussit en obéissant à des règles cette fois disciplinaires : en Français en consultant son Bled, en Mathématiques en tenant bien droit sa règle pour tracer, etc.

Sur le corpus dont nous disposons, cette analyse confirme des résultats déjà établis : il y a des modes d'appréhension variables de l'école, ici de la structuration des enseignements scolaires en enseignements disciplinaires. Et ces variations, de ce qui est pour nous des traces des consciences disciplinaires, sont réglées (soumises ?)/ expliquées par les variations des milieux sociaux dominants des quartiers dans lesquels les établissements scolaires sont implantés. C'est le respect formel des règles qui semble dériver de l'appartenance au milieu favorisé : une institutionnalisation acceptée ?

Mais il nous est aussi possible, par le recours à ce type d'analyse, de détruire (ou de reconsidérer ?) la symétrie des relations (statistiques) entre dimensions de la conscience disciplinaire et type d'environnement social dans lequel les élèves évoluent. Ces relations (à une variation conséquente du milieu correspond une variation conséquente des dimensions reconstruites) comme nous venons de le voir, ne sont pas, loin de là, des relations simples de cause à effet. En montrant que certaines de ces dimensions sont des caractéristiques « partagées » tandis que d'autres semblent plutôt être le fait de « groupes d'élèves », l'ASI démontre aussi son intérêt.

Cependant, notre analyse n'est pas terminée. En effet, notre projet global est l'évaluation (dans un sens très large) les effets des modes de travail pédagogiques dans des quartiers populaires (voir Reuter etc.).

5.2.4 Les pédagogies dont se réclament les enseignants

Par conséquent, les variations les plus intéressantes pour nous sont celles des modes de travail pédagogiques : dans le corpus dont nous disposons, deux établissements (dans des quartiers « très défavorisés ») revendiquent plus ou moins l'inscription dans une démarche pédagogique particulière. Dans le troisième et dernier des établissements inscrits dans ce type de quartier, les enseignants en revanche, ne revendiquent pas de démarche pédagogique spécifique.

Voici un tableau descriptif des inscriptions pédagogiques déclarées par les enseignants et les codages :

Mode de travail pédagogique	**Établissements[10]**	**Codages des classes correspondantes**
Mouvement Freinet	Hélène Boucher (HB)	HB1
		HB2

[10] Les noms ont été modifiés, sauf celui d'Hélène Boucher.

Fonctionnement collectif : échanges de pratiques, de groupes de classe etc.	Pascal (Pa)	Pa1
		Pa2
Pas d'inscription particulière	Veermer (Ve)	Ve

Tableau 2

Nous nous posons donc la question de savoir si aux différentes pratiques pédagogiques des maîtres correspondent des différences entre les discours des élèves et des [reconstructions de leur] conscience disciplinaire.

L'analyse a été menée sur le même corpus, en centrant cette fois les cônes implicatifs sur les 5 classes de ces trois établissements.

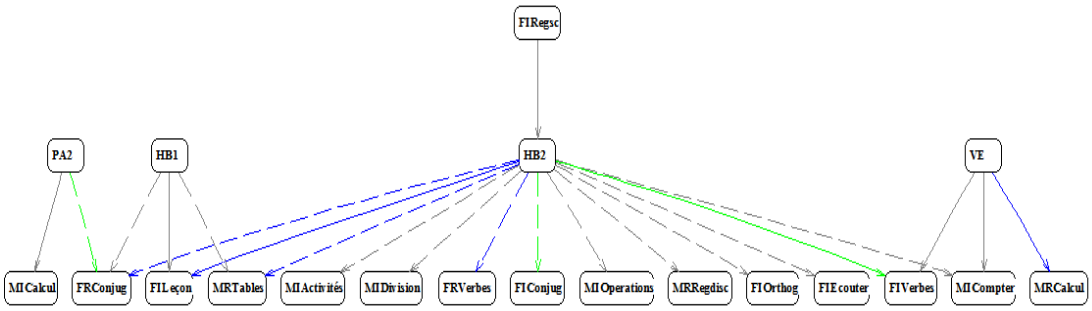


Figure 4

La particularité de la classe HB2 « saute aux yeux ». C'est tout d'abord celle où les discours sont les plus variés. C'est aussi celle où les discours des élèves sont originaux : ainsi la dimension « les contenus importants en Mathématiques sont les activités spécifiques telles les recherches mathématiques (MIActivités). Nous notons enfin que quelques-unes des « conséquences » c'est-à-dire quelques-unes des dimensions manifestées par l'appartenance à cette classe sont celles qui particularisaient des discours des élèves des établissements situés dans un autre type de quartier : MRReglesdisciplinaires, MIDivision, FIOrthographe, FIécouter.

Nous ne creuserons pas davantage ici cette singularité, pour laisser la place à une question qui illustre la relation dialectique cause-effet. Cette classe HB2 nous est familière : elle est celle d'un enseignant qui est un des piliers de ce mouvement de l'ICEM, qui a été un des premiers enseignants de cette équipe[11]. Faut-il en conclure que ce dernier peut être qualifié d'expert ? (et, parallèlement, qualifier le second enseignant de l'établissement HB, l'enseignant de la classe HB2 de « novice » ?). Nous hésitons à attribuer ces étiquettes à des sujets : ne faut-il pas plutôt parler dans la classe HB2 de pratiques pédagogiques réellement alternatives ?

6 Conclusion

[11] Nous avons suivi cette expérience (une école entière fonctionnant en pédagogie Freinet depuis 2001. L'enseignant responsable de la classe HB1 est un jeune enseignant, qui a rejoint cette école depuis un an, mais qui s'inscrit lui aussi dans ce mouvement pédagogique.

A la faveur de l'analyse non symétrique d'un corpus de données et à travers la représentation en graphe implicatif de l'ensemble des règles d'association obtenues, nous avons isolé un sous-graphe constitué d'une sorte de cône à deux nappes. Son intérêt réside dans la mise en évidence du rôle de bassin d'attraction que remplit le sommet du cône. Dans le cas où la nappe supérieure supposée causale est constituée de sommets indépendants, on dispose d'une confluence causale des pères du sommet comme effet résumé de plusieurs causes et, en un même temps dialectique, explicatif de celles-ci. De la même façon, l'examen des « fils » de ce sommet permet de donner un sens de conséquences éclatées de cet attracteur que joue le sommet. Plusieurs situations concrètes, dont l'un plus fouillée portant sur la notion de « conscience disciplinaire », ont permis d'illustrer ces démarches. Il serait intéressant de se priver de l'indépendance des « causes » pour, au contraire, examiner en quoi le sommet pourrait relever de l'agglutination de ces causes qui ne seraient alors que des modalités faisant en quelque sorte « cause commune ». Cette autre problématique rejoindrait celle de cet article dans notre ambitieuse entreprise de l'élucidation des relations causales entre phénomènes car comme l'écrit Francis Bacon : « *Vraiment connaître, c'est connaître par les causes* » et par ailleurs : « *Savoir pour prévoir, prévoir pour pouvoir* ».

Références

[1] Bailleul, M. & Godard, S. (2009), Derrière les réseaux de variables, il y a des individus… à écouter!, In Gras R. (dir.), Régnier J-C., Marinica C., Guillet F., *L'analyse statistique implicative, Méthode exploratoire et confirmatoire à la recherche de causalités,* Cépaduès Ed., Toulouse, 437-453.

[2] Bodin, A. & Gras, R. (1999), Analyse du préquestionnaire enseignants avant EVAPM-Terminales, *Bulletin de l'Association des Professeurs de Mathématiques,* n° 425, 772-786.

[3] Briand, H., Sebag, M., Gras, R.et Guillet F. (eds), (2004), *Mesures de Qualité pour la Fouille de Données,* RNTI-E-1, Cépaduès, Toulouse.

[4] Cohen-Azria, C., Lahanier-Reuter, D., Reuter, Y. (2013), *La conscience disciplinaire,* P.U.R., Rennes.

[5] Gras R. (dir.) Régnier J.-C., Marinica C., Guillet F (eds.), (2013), L'analyse statistique implicative, Méthode exploratoire et confirmatoire à la recherche de causalités, Cépaduès Editions, Toulouse.

[6] Gras, R. (dir.), Almouloud, S. Ag, Bailleul, M., Larher, A., Polo, M., Ratsimba-Rajohn, H., Totohasina, A. (coll.), (1996), *L'implication statistique. Nouvelle méthode exploratoire de donnée,* La Pensée Sauvage, Grenoble.

[7] 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, 9-29.

[8] Gras, R., Suzuki, E., Guillet, F., Spagnolo, F. (eds), (2008), *Statistical Implicative Analysis, Theory and Applications,* Springer.

[9] Guillet, F. & Hamilton, H. (eds), (2007), *Quality Measures in Data Mining,* Springer.

[10] Lahire, B. (1993), Culture écrite et inégalités scolaires, P.U.L., Lyon.

[11] Lévy-Leblond, J.-M. (1996), *Aux contraires*, Gallimard, Paris.

[12] Orus, P., Zemora, L., Gregori, P., (2009), *Universitat Teoria y Aplicaciones del Analisis Estadistico Implicativo*, Eds : Jaume-1, Castellon (Espagne).

[13] Pasquier, D. & Gras, R. (2012), De l'intérêt de l'Analyse Statistique Implicative (A.S.I.) pour la recherche exploratoire en psychologie, *Psychologie Française*, Elsevier-Masson, 161-173.

[14] Régnier, J-C., Bailleul, M., Gras, R. (eds.), (2012), *L'Analyse Statistique Implicative : de l'exploratoire au confirmatoire*, Université de Caen, Caen.

[15] Sève, L. (2005), *Émergence, complexité et dialectique*, Odile Jacob, Paris.

Annexes

Coefficients de corrélation (Figure 2)

	MR Tables	FR Conjugaison	FR Verbes	MI Tables	MI Calcul	MI Compter	FI Dictées
MR Calcul	-0,23	0,07	0,08	-0,02	0,20	0,04	0,05
MRTables		0,14	0,19	0,06	0,01	0,02	0,11
FRConjugaison			0,28	0,05	0,09	0,02	0,08
FRVerbes				0,21	0,04	0,03	0,14
MITables					-0,06	-0,03	0,07
MICalcul						0,06	-0,05
MICompter							-0,04

Corrélations partielles

r(MRC, MRT,FRC) = **-0,24** r(MRC,MRT, FRV) = **-0,25**

r(MRC, FRC, FRV)= 0,05 r(MRT,FRC, FRV) =0,09

Pour la Figure 3

	FREcouter	FROrtho	MIDivision	MIDécimaux	FITemps	FIGrammaire
FREcouter		-0,06	-0,02	0,02	0,12	0,11
FROrtho			0,05	0,2	0,09	0,18
MIDivision				0,09	0,22	0
MIDécimaux					0,32	0
FITemps						-0,08
FIGrammaire						

Cette fois, deux variables seulement sont corrélées positivement, au seuil de risque de 5% : Le fait de citer les décimaux comme contenus importants vus en Mathématiques et celui de citer les Temps comme contenus importants vus en Français : une vision globale des savoirs disciplinaires conforme à celle des programmes ?

HIERARCHIE DE REGLES EN A.S.I. ET CONCEPTUALISATION

Régis GRAS[1], Nadja ACIOLY-REGNIER[2]

HIERARCHY OF RULES IN S.I.A. AND CONCEPTUALIZATION

RÉSUMÉ

Nous présentons dans cet article une méthode alternative et complémentaire au graphe implicatif pour représenter les relations implicatives nouées au sein de variables qualifiant ou quantifiant des sujets ou des objets. Mais ici les règles extraites des données sont de degré supérieur et apparaissent comme métarègles ou règles de règles. Nous axiomatisons ces règles généralisées et les représentons selon une hiérarchie ascendante orientée. En psychologie différentielle et en didactique de sciences (champs conceptuels), la complexité cognitive nous semble apparaître comme métaphore ou avatar privilégié de ce type de hiérarchie. Nous présentons deux exemples illustratifs en montrant l'intérêt prédictif de cette représentation.

Mots-clés *: Hiérarchie, classes orientées, axiome, distance ultramétrique**,** psychologie différentielle, champ conceptuel.*

ABSTRACT

In this article we present an alternative and complementary method to the implicative graph to represent the implicative relationship to present the knotted within the variables qualifying or quantifying either subjects or objects. But here, the data extracted rules are of a superior level and appear as metarules or as rules of rules. We axiomatise those generalised rules and we represent them according to an oriented ascendant hierarchy. In differential psychology and in science didactic (conceptual fields), cognitive complexity seems to appear as a privileged metaphor or avatar for this type of hierarchy. We present two illustrative examples by showing the predictive interest of this representation.

Keywords : *hierarchy, oriented classes, axiom, ultrametric distance, differential psychology, conceptual field.*

1 Introduction

A notre connaissance, d'une part, les développements existant en matière de mesures de qualité de règles d'association s'arrêtent généralement à la proposition d'un indice d'implication partielle pour des données binaires et, au mieux, à la représentation arborescente des règles (par exemple : Agrawal, R. et al, 1993, Guillet F. et Hamilton H. 2007, Hiep J. et al 2000, Lenca P. et al, 2007 et tous les travaux portant sur les réseaux bayesiens et les treillis de Galois, voir Cadot M. 2009). D'autre part, cette notion n'est pas étendue à l'extraction de règles de règles où les prémisses et les conclusions peuvent

[1] Laboratoire d'Informatique de Nantes-Atlantique (LINA) Site Polytech Nantes - La Chantrerie, rue C. Pauc, BP 44306, Nantes cedex 3, e-mail : regisgra@club-internet.fr,

[2] ESPE-École Supérieure du Professorat /Université Claude Bernard – Lyon 1. E-mail: nadja.acioly-regnier@univ-lyon1.fr

être elles-mêmes des règles. A l'instar de certaines méthodes d'analyse de données, l'intérêt de passer, non linéairement, du niveau relationnel à celui de hiérarchie nous est apparu évident, comme une réponse à la question brûlante, dans la construction des concepts, de non-linéarité[3]. Nous pouvions nous appuyer sur notre expérience en didactique des mathématiques et en psychologie cognitive. Nous avons alors prolongé notre recherche de règles dans un corpus de données à son extension théorique, celle de règles de règles ou méta-règles (ou règles généralisées). *Ainsi, notre recherche de relations causales sera* ***non-réductionniste*** *puisqu'elle dépassera le simple dénombrement et conduira au nouveau choix d'un indice, ceci en élargissant le champ opérationnel de l'ASI à des règles de niveau supérieur, à leur représentation, à l'identification et à la mesure du rapport dual sujets-variables.*

Nous proposons *donc* ici ces prolongements dans le cadre de l'A.S.I., en formalisant la notion de hiérarchie orientée. Celle-ci vise la prise en charge de la représentation graphique de la structure de l'ensemble des associations des variables selon **des règles de règles**, susceptibles pour certaines de se situer en toute hypothèse à un niveau conceptuel supérieur. Une première version formelle a été présentée dans (Gras et, 1996). L'approche en était intuitive et pragmatique. Depuis, (Gras et Kuntz, 2005) a permis de réviser la notion de hiérarchie orientée par une présentation algébrique plus formelle mais en accord avec les propriétés rigoureuses de la notion de hiérarchie orientée.

La forme graphique obtenue peut nous faire penser, toutes réserves faites et par analogie de structure, dans son architecture logique, au modèle théorique épistémologique de J. Piaget. Ce modèle, aujourd'hui relativisé pour sa rigidité, sinon contesté, conçoit le développement cognitif de l'enfant selon une organisation élaborée passant par une succession de stades emboîtés successifs. Le passage d'un stade à un autre selon une conception établie sur un mode implicatif strict n'est pas sans poser problème (Pellois C., 2002, 2010 et 2013) car il se fait, notamment, mais pas seulement, par le bais de ce que Piaget appelle l'abstraction réfléchissante « 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. Ecartant provisoirement toute référence à cette théorie, l'objectif central de ce texte réside dans l'étude de l'adéquation métaphorique de la représentation de règles de règles avec une démarche conceptualisante.

Rappelons qu'une représentation structurée des relations quasi-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 d'implication, fermé transitivement à un seuil donné. Cette première représentation serait-elle comparable aux mécanismes de conceptualisation qui semblent se mettre en œuvre formellement pour aboutir, in fine à l'organisation cognitive du sujet épistémique, telle que la présente

[3] Relisons G. Bachelard : « Pour un esprit scientifique, toute connaissance est une réponse à une question. S'il n'y a pas eu de question, il ne peut y avoir connaissance scientifique. Rien de va de soi. Rien n'est donné. Tout est construit ». (Extrait de *La Formation de l'esprit scientifique* 1938).

Piaget (1970) ? Répondre à cette question nécessiterait très certainement bien des discussions et quelques approfondissements, mais dans des situations d'apprentissage ou d'évaluation, elle nous a permis d'interpréter des chemins de ce graphe, constitués de suites d'arcs qui lient certaines variables, en termes de genèses relevant d'une forme particulière de conceptions différentielles. Cette théorie piagétienne du développement cognitif du sujet, dans sa forme originelle exprimée en termes de stades ordonnés indépendants des contenus est cependant, comme nous l'avons dit, à relativiser dans sa rigueur formelle. Elle fait actuellement la place, non seulement à la psychologie différentielle sous différentes conceptions (Cf. par exemple, Pellois, C., 2008, en particulier p. 91.), mais aussi à la théorie des champs conceptuels[4] où le développement cognitif serait fortement lié au contenu conceptuel des domaines dans lesquels il s'opère plus qu'à des structures logico-mathématiques de la pensée. On en fera l'observation dans l'exemple 2 du § 4.2.

La méthode d'organisation hiérarchique des données, non symétrique que nous présentons maintenant, va doubler les informations fournies par les règles d'association implicative en organisant leur ensemble selon une structure ordonnée en méta-règles, en méta-méta-règles, etc.. Nous verrons, par la suite, quelle proximité de structure (quel homomorphisme ?) existe entre ce type de représentation et l'organisation de règles nécessaires à la conceptualisation. Les deux premiers paragraphes 2 et 3 présenteront une modélisation d'une hiérarchie orientée de règles de règles. Le paragraphe 4 étudiera en quoi la hiérarchie ainsi modélisée pourrait être une métaphore d'un processus de conceptualisation.

2 Hiérarchie de classes de variables[5]

Afin de soutenir l'intuition, nous baserons le modèle de la **hiérarchie orientée cohésitive** sur la correspondance 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,…, 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

[4] Vergnaud (Vergnaud 1981 et 1990) définit la notion de champ conceptuel « comme un espace de problèmes ou de situations-problèmes dont le traitement implique des concepts et des procédures de plusieurs types en étroite connexion. »

[5] Cette section §2 est une composante de l'article Gras et Kuntz (2005), reprise d'ailleurs dans Gras et al. (2009 et 2013). On la retrouve présentée dans Régnier et Acioly-Régnier (2007). Dans un souci d'homogénéisation, nous tenions à le rappeler ici.

mathématique organisée en théorèmes et corollaires. Mais nous verrons que ces métaphores ne satisfont pas totalement la structure du 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

Ce paragraphe théorique rend compte de la formalisation de la construction d'une hiérarchie dans laquelle l'ordre ascendant de classification se doit de respecter le fondement asymétrique de l'implication. C'est donc sur ce projet, cette base épistémologique que nous construirons ce nouveau modèle hiérarchique.

Définition 1: On appelle hiérarchie orientée **H** sur l'ensemble des variables V, une suite d'arrangements (au sens de la combinatoire, donc suite de k-uplets) 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 2: On appelle classe **C** de degré k de la hiérarchie **H** un arrangement (ou k-uplet) 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 3: 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 4: 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 5: 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 $\hat{\cap}$.

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

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

Définition 6: Si les deux classes quelconques **C'** et **C''** ont une trace vide (**C'** $\hat{\cap}$ **C''**=$\varnothing$), la concaténation de **C'** et **C''** notée **C'** $\hat{\cup}$ **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'** $\hat{\cup}$ **C''.**

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

Par exemple, si **C'** = (d,f,g,a) et **C''** = (b,u,r,p,y), **C'** $\hat{\cup}$ **C''** = (d,f,g,a,b,u,r,p,y), alors que **C''** $\hat{\cup}$ **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 $\hat{\cap}$ 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 7: 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, $coh(a,b) = \left(1 - [-p(\log_2(p)) - (1-p)(\log_2(1-p))]^2\right)^{\frac{1}{2}}$ où p=φ(a,b)≥0,50 et $coh(a,b) = 0$ si p=φ(a,b)<0,50.

Cette notion de cohésion nous conduit, dans le cadre de l'ASI, de parler aussi bien de **hiérarchie orientée** que de **hiérarchie cohésitive.**

Définition 8: La cohésion d'une classe **C** = (a_1, a_2,…, a_r) de degré r, est définie par la formule : $coh(C) = \left[\prod_{i=1,\ldots,r-1}^{j=2,\ldots,r;\, j>i} coh(a_i, a_j)\right]^{\frac{2}{r(r-1)}}$

Définition 9 : 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

Cet algorithme est implémenté dans le logiciel CHIC (logiciel de traitement non symétrique de données, Couturier R. et Ag Almouloud S., 2013).

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

Axiome 2 :

Pour toute classe C de la hiérarchie :

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.

2.5 Exemples

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 A.

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 A est (h,i,e,f,g,a,b,c,d,j).

Exemple 3

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 $x \wedge y \Rightarrow z$. Une cohésion de qualité (proche de 1) autoriserait cette extension sémantique. La règle $(d \Rightarrow b) \Rightarrow (a \Rightarrow f)$ ou ((db)(af)), illustrée par la figure (Fig. 1), peut s'interpréter par la phrase: 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.

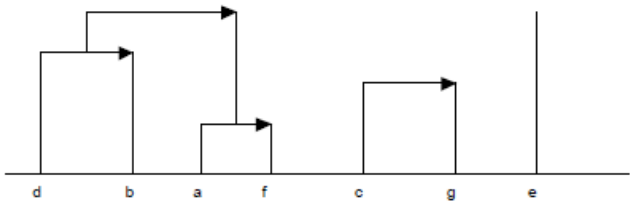


Figure 1 Un exemple de hiérarchie orientée

Exemple 4

Par exemple, de théorèmes comportementaux, on peut dégager **un trait** ou une **conception** (ex : conceptions du hasard par (Lahanier-Reuter, 1999)). Un indice statistique permet en outre de repérer les niveaux hiérarchiques correspondant à une **significativité des classes** formées à ces niveaux (en rouge sur la Fig.2). La hiérarchie de la Fig. 2 est produite, à l'aide de CHIC, à partir d'un fichier constitué par les réponses à un questionnaire figurant en annexe et présenté à des enseignants de mathématiques dans des classes terminales de lycée (18-19 ans). Ces enseignants devaient choisir et pondérer les objectifs qu'ils assignaient aux cours de mathématiques de leur programme. Nous reviendrons plus loin sur cet exemple.

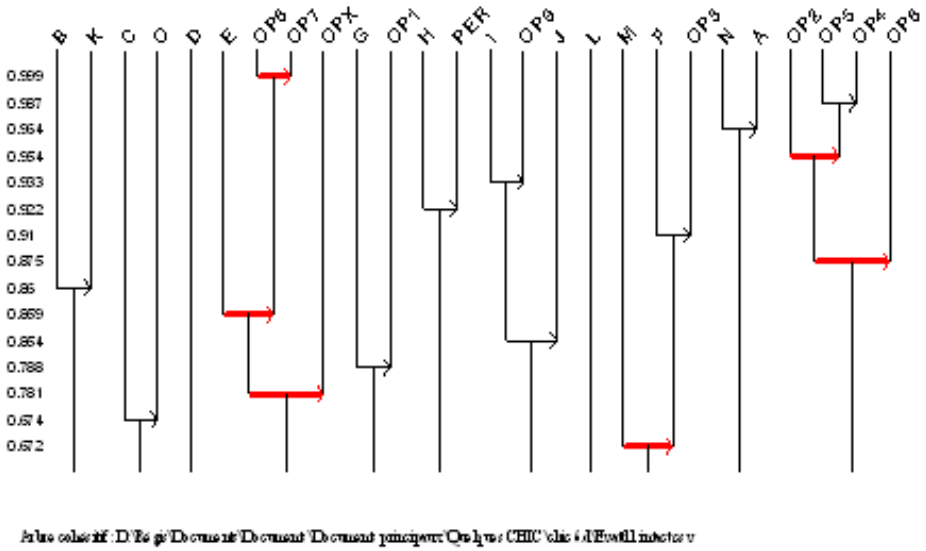


Figure 2 Hiérarchie des objectifs et attitudes propres à l'enseignement des mathématiques données par le logiciel CHIC

Il n'y a pas de raison fondamentale d'utiliser le type de représentation (« vers le haut ») de la fig.1 ou celui de la fig. 2 (« vers le bas »). La première est traditionnellement utilisée dans les hiérarchies de la théorie des graphes. La seconde, obtenue par pliage axial, organise la hiérarchie selon la cohésion croissante. Elle permet un accès privilégié aux noms des variables.

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) :

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

où $\text{coh}(C_{(x,y)})$ 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 $C \subset C'$ ou bien $C \supset C'$ ou bien $C \cap C' = \varnothing$
2. si $C \subset C'$, alors $\text{coh}(C) \geq \text{coh}(C')$ 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 $(x,y) \in VxV$, d(x,y)=d(y,x) par construction

Axiome 3 :

$(x,y,z) \in VxVxV$, $d(x,y) \leq \sup[d(x,z), d(y,z)]$ (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 $z \in C_{(x,z)} \cap C_{(y,z)} \neq \varnothing$ d'où $C_{(x,z)} \subset C_{(y,z)}$ ou bien $C_{(x,z)} \supset C_{(y,z)}$ d'après les propriétés des classes de H. Supposons alors: $C_{(x,z)} \supset C_{(y,z)}$. On en déduit $d(x,z) \geq d(y,z)$ 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 $C_{(x,y)} \subset C_{(x,z)}$

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

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

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

4 Hiérarchie cohésitive, modèle de la conceptualisation ?

En psychologie du développement, et, ceci afin de poursuivre l'analogie avec les conceptions piagétiennes, la notion d'abstraction réfléchissante associée à celle d'abstraction empirique et d'abstraction « réfléchie », constituant différentes étapes de « prise de conscience », tente de rendre compte du mécanisme d'équilibration majorante et de conceptualisation. Celle-ci fait passer d'un niveau de l'organisation conceptuelle (par exemple, un schème, une règle d'action, un stade donné du développement cognitif) à un niveau supérieur de cette organisation conceptuelle (le stade suivant du développement cognitif) ; chacun des niveaux étant constitué de règles, par exemple des théorèmes en acte dirait G. Vergnaud, portant sur des contenus à chaque fois différents, d'opérations portant sur ces contenus, enfin sur des opérations sur ces opérations. On retrouve, d'ailleurs, dans une théorie mathématique ces mêmes élargissements lorsque l'on passe d'un théorème qui établit l'implication d'une propriété sur une autre, à un corollaire qui fait découler d'un théorème un autre théorème ou une simple propriété. C'est le cas, par exemple, dans l'étude des fonctions, puis à l'étude de fonction de fonctions en analyse fonctionnelle. D'où l'idée que nous avons eue, de construire un second plan de relations implicatives, celui de **règles de règles** selon une **hiérarchie dite cohésitive** en raison de l'indice de **cohésion** qui permet d'engendrer des classes orientées de règles. Cet indice est à une règle généralisée ce que l'intensité d'implication est à une règle. C'est-à-dire que contrairement aux mathématiques où la validité est absolue, celle d'une règle de règle en ASI est relativisée à la valeur de la cohésion, nombre compris entre 0 et 1, croissant avec la qualité de l'implication d'une règle sur une autre.

4.1 Un détour par la philosophie des sciences.

A travers ces deux types de représentations de règles simples (graphe implicatif) ou généralisées (hiérarchie cohésitive), qui mettent au jour deux types de structures dans l'ensemble des variables, nous répondons à la philosophie structuraliste, donc non réductionniste, tout comme la théorie générale des systèmes de L. Von Bertalanffy (1973, p.66) qui estime que le **« tout » est plus riche que la somme de ses « parties ».** Citons, à ce sujet, deux extraits du livre de L. Sève (ib. p. 58) « *…le **tout** ne se compose de rien d'autre que de ses **parties**, et pourtant il présente, en tant que tout, des propriétés n'appartenant à aucune de ses parties. Autrement dit, 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 ne peuvent donc s'expliquer par elles* ». Et plus loin *« Tout se passe donc comme si se produisait une génération spontanée de propriétés du tout…C'est le paradoxe de **l'émergence**».*

4.2 Illustrations

Exemple 1

Appuyons-nous sur l'exemple 4 du § 2.5. La classe M => (F => OP3) est constituée de 3 objectifs et attitudes attendus par les professeurs de mathématiques au lycée :

- OP3 : « Dans ma notation, j'attache plus d'importance à la démarche qu'au résultat »,
- F : « développement de la capacité à prouver et valider sa preuve »,
- M : « développement de la capacité à mathématiser et à formaliser.

OP3 apparaît comme une opérationnalisation de l'objectif spécifique F. M exprime une position supérieure sur le plan cognitif, quasi-finalité de la discipline « Mathématiques ». Il y a donc bien élévation du niveau conceptuel du « théorème » F => OP3 à M. M apparaît comme large enveloppe émergeant d'une spécification.

On retrouve, de la même façon, le passage des comportements spécifiques exprimés par E => (OP8 => OP7) vers un objectif général E qui n'est pas la seule addition de OP8 et OP7 :

- OP7 : ....pouvoir reconnaître si un nombre entier écrit dans la base 10 est divisible par 4 impliqué par …
- OP8 : pouvoir donner un exemple ou un contre-exemple personnels à l'affirmation….
- E : « développement de l'imagination et la créativité ».

Nous constatons, ici aussi, le passage, par un saut qualitatif, d'un niveau d'activité élémentaire OP7 à un niveau supérieur E via un niveau intermédiaire OP8 dont OP7 est un exemple. La classe (E(OP8, OP7)) traduirait, en fait, l'émergence d'une qualité intellectuelle E génératrice d'une règle opératoire inférée par E.

Cette propriété disparaît ou s'estompe lorsque, sur le même fichier, en utilisant le même type d'algorithme de base[6], on construit avec CHIC la hiérarchie de similarité sur l'ensemble des variables. On observe (voir la Figure 3) des classes regroupant des variables vis-à-vis desquelles les enseignants ont adopté des attitudes identiques ou voisines. Par exemple, la classe {E, I, OP9, J} rassemble 3 objectifs généraux (E,I,J) de mise à distance non nécessairement disciplinaire et un savoir de type critique (réfuter vs accepter). Contrairement aux classes cohésitives ordonnées, celle-ci est homogène : les comportements qui l'illustrent relèvent du même modèle attendu de l'élève. Ce n'est donc pas comme avec la hiérarchie cohésitive l'image d'un modèle en construction mais en action.

[6] La similarité entre deux variables est évaluée à partir du nombre statistiquement étonnant d'exemples satisfaisant simultanément les deux variables eu égard aux nombres de sujets satisfaisant séparément chacune des variables. Des paires sont alors réunies en une classe. La liaison entre deux classes est établie de façon comparable.

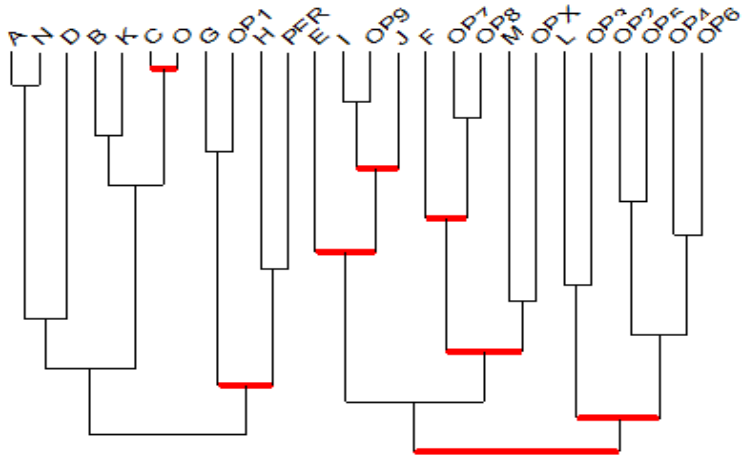


Figure 3 Hiérarchie des similarités

Exemple 2

Dans sa thèse où il a fondé l'A.S.I., Gras R. (1979) a également construit, dans le cadre de l'apprentissage des mathématiques, une taxonomie d'objectifs cognitifs selon 5 niveaux allant de la « Connaissance des outils de préhension de l'objet et du fait mathématique » jusqu'à la « Critique et évaluation ». Cette taxonomie ne décrit pas des phases d'apprentissage conceptuel cumulatives qui s'étireraient linéairement dans le temps au cours d'une scolarité, mais une hiérarchie de niveaux de complexité cognitive. Dans le but de valider ponctuellement ou simplement interroger la taxonomie[7], un questionnaire portant sur la notion de symétrie centrale (géométrique et algébrique) a été proposé à 401 élèves de 13-14 ans. Il était constitué d'une cinquantaine d'items supposés opérationnalisant a priori la taxonomie, c'est-à-dire organisant les réussites attendues selon un préordre respectant peu ou prou l'ordre taxonomique.

Les résultats obtenus par une analyse implicative montrent bien le phénomène attendu en tant qu'illustration des paragraphes plus haut. L'échelle de complexité prévue est relativement bien validée. Les items de niveaux supérieurs géométriques (resp. algébriques) (critique, évaluation,..) se placent généralement en racines des classes cohésitives en impliquant ceux de niveaux inférieurs (connaissance de faits, de définitions, de techniques,…) avec des décalages comme il s'en trouve dans des tests de psychologie différentielle. L'observation de certains décalages montre en particulier des différences de réussite très nettes entre les items algébriques et les items géométriques supposés de même niveau « piagétien ». Mais ces items, opérant dans des domaines non isomorphes, conformément à la théorie des champs, ne mobilisent pas les mêmes structures logico-mathématiques de Piaget, les mêmes schèmes, les mêmes invariants opératoires, les mêmes théorèmes et concepts en acte. On trouvera ainsi des classes

[7] C'est d'ailleurs dans cette intention que R.Gras a conçu la notion d'Analyse Statistique Implicative afin de rendre compte du préordre partiel sous-jacent aux comportements de réponse des élèves à ce questionnaire. Il ne disposait pas, parmi les méthodes d'analyse de données de moyen de mettre en évidence ce phénomène. L'échelle de Guttman ne répondait que partiellement à sa demande.

orientées spécifiques d'un champ géométrique et d'autres d'un champ algébrico-numérique.

Exemple 3

Dans Régnier et Acioly-Régnier (2007), les auteurs traitent un questionnaire sur les représentations de 320 sujets relativement aux phases de la lune. Une première analyse avait été menée déjà dans Acioly-Régnier et Régnier (2005) mais approfondie en 2007 par un usage du concept de cohésion. Ils ont été conduits, dans ces deux articles, à une interprétation dans le cadre théorique psychologique. Entre autres classes, ils s'intéressent à deux classes formées très tôt par la hiérarchie cohésitive. L'une d'entre elles, renverrait au poids de la culture, l'autre à des niveaux de conceptualisation plus élevés. On retrouve donc, dans ces textes, une démarche interprétative comparable à celle emboîtée ci-dessus.

Dans les ouvrages cités en bibliographie, on trouvera de multiples exemples tant en psychologie, sociologie, qu'en biologie, en art, en médecine, etc.

4.3 Le rapport tout-partie.

Nous insistons sur cette propriété spécifique de l'ASI que la hiérarchie cohésitive satisfait de façon originale à travers l'extension des relations entre variables. C'est par un saut qualitatif, produit généralement par un effet de seuil dans la quantité (ex. psychologie de groupe, vaporisation de l'eau…), que le tout, au prix d'une synthèse, prend son sens. Celui-ci s'extrait de la lecture véritablement dialectique du **rapport non-linéaire tout/partie**[8] (non-proportionnalité et non-additivité de la cause sur l'effet). Avec l'ASI, il y a alors **paradoxe** entre des contraires, lien et absence de lien, car le tout, constitué (on devrait dire *organisé*) de parties en un **système dynamique**, possède des propriétés que ne possèdent pas les parties et qui sont généralement de **niveau supérieur**. De la même façon, et métaphoriquement, en linguistique, la signification d'une phrase ne se fait pas uniquement par l'analyse du sens de chacun de ses mots mais par le sens de chacun de ces mots inscrit donné par l'interaction de ceux-ci[9]. La logique qui sous-tend ce rapport tout/parties est **dialectique** (et non pas dichotomique) car elle concilie interactivement des contradictions : règle et non-règle, **instabilité d'un système dynamique et stabilité structurelle**. Elle se fonde en règle sur l'inexistence importante du contre-exemple et en méta-règles sur la négation de l'entropie, du désordre comme on le constate dans les relations d'un fleuve et de ses affluents. En

[8].. comme le montrent les équations différentielles non-linéaires de l'indice fondamental par rapport aux paramètres cardinaux des observations, contrairement à ce qui est observable avec d'autres indices concurrents.. *« La société n'est pas constituée d'individus, mais exprime la somme des relations, des rapports où ces individus se situent les uns par rapport aux autres* » (K. Marx, « Manuscrits de 1857-1858). Une rue n'est pas la somme des maisons qui y figurent. De même une ville n'est pas somme de ses rues, etc.

[9] « *Il n'y a ni additivité ni proportionnalité entre le sens des unités (mots) et celui de la phrase. On voit se dessiner une topologie du sens* » (F. Gaudin dans « Emergence, complexité et dialectique »). « *…le mot isolé de la langue chinoise n'a en vérité ni signification ni existence à part, chacun ne reçoit sa signification que du parler même (de l'intonation, etc.…), pris isolément il a dix, voire quarante significations,… ; si nous soustrayons ce mot à la totalité, il se perd dans une creuse infinité.* « (F.-W. Schelling, « Philosophie de la mythologie », p.361).

cela, la logique dialectique s'oppose à la logique stricte (du mathématicien) sans, bien sûr, être un sophisme, c'est-à-dire un raisonnement faux. La fécondité et l'originalité de l'ASI tiennent à ce caractère, particulièrement dans l'analyse hiérarchique manifestement non-linéaire où le tout fonde son sens, non par addition des propriétés de ses parties (sous-classes) mais par la synthèse, voire par une reconstruction, des interactions inférentielles. C'est par la notion de **niveau significatif** que nous pouvons mettre en évidence le phénomène de **propriété émergente**. En ce sens, l'A.S.I. apparaît comme une sorte d'avatar de l'apprentissage non linéaire des connaissances, apprentissage fait de dépassements, de ruptures et de reconstruction dialectique (cf. l'épistémologie de G. Bachelard ou de Lev Vygotsky). A l'opposé de l'A.S.I., le rapport tout-parties serait **linéaire** dans le cas d'emboîtements de classes comme en classification fondée sur la similarité jusqu'à son nœud terminal (cf. § 3.2). Et ce linéaire infécond ne peut pas surprendre. Dans l'ouvrage « Dans la lumière et les ombres, Darwin et le bouleversement du monde » (2011), Jean-Claude Ameisen cite François Jacob : « On mesure l'importance d'une découverte à la surprise qu'elle cause » et il ajoute ; « A son caractère profondément inattendu. A ce qui nous manquait pour simplement nous y attendre » (p. 468).

5 Conclusion

A la suite de ces réflexions reliant psychologie de l'apprentissage et représentation hiérarchique de données, on peut se poser la question que se pose Gérard Vergnaud relativement à la correspondance entre signifiés et signifiants qu'il qualifie d'homomorphisme de structure : y aurait-il, de façon semblable, entre classes et sous-classes, au moins morphisme de structures préordonnées : celle de la pensée conceptualisante et celle de la hiérarchie des règles généralisées telle que la présente une hiérarchie cohésitive ? Celle-ci serait-elle une métaphore graphique de la pensée ? *Comme l'avance Thom R. (1980, p.142) au sujet de l'analogie : " ou bien elle est vraie et alors elle s'avère sterile, ou bien elle est audacieuse et alors elle peut être féconde"*. Les deux exemples présentés ici semblent soutenir cette conjecture. D'autres situations expérimentales l'ont confortée. La comparaison entre ces deux structures nous a conduits à aller au delà de la seule construction (spéculative ?) d'un outil classifiant d'analyse de données et, par suite, à donner un sens spécifique à ce mode de représentation par rapport aux modes classiques de représentation hiérarchique. *Mais, ce faisant, ne satisfait-on pas au voeu de Thom R. (1980, p.58) qui écrit encore : "…je localise l'effort théorique de la science dans sa capacité d'organiser les données de l'expérience selon des schémas imposés par des structures théoriques" ?*

Références

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

[2] Agrawal R., Imielinsky T. et Swami A,(1993), Mining association rules between sets of items in large databases, *Proc. of the ACM SIGMOD' 93*, p. 207-216

[3] Bachelard G., (1967), *La Formation de l'esprit scientifique*, Paris, 5e édition, Librairie philosophique J. Vrin.

[4] Cadot M., (2009), Graphe de règles d'implication statistique pour le raisonnement courant. Comparaison avec les réseaux bayesiens et les treillis de Galois, Analyse Statistique Implicative, sous la direction de Gras R., réd, invités Gras R., Régnier J-C., Guillet F. *Une méthode d'analyse de données pour la recherche de causalités*, Cépaduès Ed. Toulouse, p.223-250

[5] Couturier R. et Ag Almouloud S., (2013), Historique et fonctionnalités de CHIC, L'analyse statistique implicative, sous la direction de Gras R., eds Gras R., Régnier J.-C., Marinica C., Guillet F *Méthode exploratoire et confirmatoire à la recherche de causalités*,., Cépaduès Editions,, p.313-325.

[6] Fayyad U., Piatetsky-Shapiro G. and Smyth P. From Data Mining to Knowledge Discovery. In., Piatetsky-Shapiro G., Smyth P, and Uthurusamy R. (eds), *Advances In Knowledge Discovery and Data Mining*, Fayyad U AAAI/MIT Press, 1-31,.

[7] 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.

[8] Gras R., Kuntz P., Briand H., Couturier R. (2002), Hiérarchie de règles généralisées et notion de variable supplémentaire en analyse statistique implicative, Actes des IXèmes *Rencontres de la Société Francophone de Classification,* Université de Toulouse, 2002, p. 211-214.

[9] Gras R., Kuntz P. et Briand H. (2003), Hiérarchie orientée de règles généralisées en analyse implicative, *Extraction des Connaissances et apprentissage*, Hermès, 145-157.

[10] Gras R., Couturier R.., Blanchard J., Briand H., Kuntz P., Peter P., (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, Cépaduès –Editions, 3-32.

[11] Gras R., Kuntz P. et Régnier J.C., (2004), 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, Cépaduès, 39-50.

[12] Gras R. et Kuntz P., (2005), *Discovering R-rules with a directed hierarchy, Soft Computing, A Fusion of Foundations, Methodologies and Applications*, Volume 1, Springer Verlag, 46-58..

[13] Guillet, F. et Hamilton H.J. (2007) (Eds.). *Quality Measures in Data Mining*. Springer.

[14] Hipp J., Guntzer U., Nakhaeizadeh, J.(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.

[15] Lahanier-Reuter D. (1999), Conceptions du hasard et enseignement des probabilités et statistiques, *Éducation et Formation*, PUF, Paris.

[16] Lenca P., Vaillant B., Meyer P., et Lallich S. (2007), Association Rule Interestingness Measures : Experimental and Theorical Studies, Guillet F. and Hamilton H.J. eds, *Studies in Computational Intelligence* 43, Springer, p. 51-76.

[17] Pellois, C., (2002). Apprentissage et développement de la personne dans l'enseignement et la formation, Tome 1 : *du rationnel au complexe*. l'harmattan, Collection : « Recherches et innovations sur et pour des enseignants et des formateurs », p. 105 et suivantes

[18] Pellois, C., (2008). Contrainte et liberté du sujet : entre incertitude et prévisibilité ? In Cadet B. Chasseigne G., Foliot G., *Cognition, incertitude et prévisibilité*, Editions Publiboock, Sciences Humaines et Sociales, Coll. « Psychologie cognitive », Paris, 77-100

[19] Pellois, C., (2010), Sens et incertitude, une forme de complexité en psychologie : des contraintes aux parts de liberté, le développement et ses contextes, in Cadet, C., Chasseigne, G. *Traitement de la complexité dans les sciences humaines*, Editions Publibook Université, Coll. « Psychologie Scientifique », 177-207

[20] Pellois, C., (2013). La psychologie et l'usage du traitement mathématique des données statistiques. Nouvelles perspectives conceptuelles, *Revista Brasiliera de Ensino de Ciencia e technologia*, Vol. 6, n° 1, 230-259

[21] Jean Piaget, (1970) *L'épistémologie génétique* PUF Paris

[22] Régnier, J.-C. et Acioly-Régnier, N.M. (2007) Analyse cohésitive et interprétations des données dans le champ de l'éducation. In Régis Gras et al. *Nouveaux apports théoriques à l'analyse statistique implicative et applications.* Castellón : Innovació Digital Castelló, p.329-343

[23] Thom, R (1980). *Paraboles et catastrophes*, Flammarion, Paris.

[24] Vergnaud, G. (1981). Quelques orientations théoriques et méthodologiques des recherches françaises en didactique des mathématiques – *Recherches en Didactique des Mathématiques*, 2.2, 215-232.

[25] Vergnaud, G. (1990). La théorie des champs conceptuels. *Recherches en Didactiques des Mathématiques*, 10 (23), 133-170.

[26] Vygotsky L. (1997). *Pensée et langage* (1933) (traduction de Françoise Sève, avant-propos de Lucien Sève), suivi de « Commentaires sur les remarques critiques de Vygotski » de Jean Piaget,(Collection « Terrains», Éditions Sociales, Paris, 1985) ; Rééditions : La Dispute, Paris.

[27] Von Bertalanffy, (1973, 1980) *Théorie générale des systèmes*. Paris : Dunod

Ouvrages de référence

L'implication statistique. Nouvelle méthode exploratoire de donnée, sous la direction de R.Gras et la collaboration de S. Ag Almouloud, M. Bailleul, A. Larher, M. Polo, H. Ratsimba-Rajohn, A.Totohasina, La Pensée Sauvage, Grenoble (1996)

Mesures de Qualité pour la Fouille de Données, H.Briand, M.Sebag, R.Gras er F.Guillet eds, RNTI-E-1, Cépaduès, 2004

Quality Measures in Data Mining , F.Guillet et H.Hamilton eds, Springer, 2007,

Statistical Implicative Analysis, Theory and Applications, R.Gras, E. Suzuki, F. Guillet, F. Spagnolo, eds, Springer, 2008.

Analyse Statistique implicative.Une méthode d'analyse de données pour la recherche de causalités, sous la direction de Régis Gras, réd. invités R. Gras, J.C. Régnier, F. Guillet, Cépaduès Ed. Toulouse, 2009.

Teoria y Aplicaciones del Analisis Estadistico Implicativo, Eds : P.Orus, L.Zemora, P.Gregori, Universitat Jaume-1, Castellon (Espagne), ISBN : 978-84-692-3925-4, 2009..

L'Analyse Statistique Implicative : de l'exploratoire au confirmatoire. Eds : J.C. Régnier, Marc Bailleul, Régis Gras, Université de Caen, ISBN : 978-2-7466-5256-9, 2012

L'analyse statistique implicative, Méthode exploratoire et confirmatoire à la recherche de causalités, sous la direction de Gras R., eds Gras R., Régnier J.-C., Marinica C., Guillet F., Cépaduès Editions, 522 pages, ISBN 978.2.36493.056.8, 2013.

ANNEXE- QUESTIONNAIRE

L'APMEP conduit une réflexion sur l'enseignement des mathématiques au lycée. Elle souhaite recueillir l'opinion du plus grand nombre possible de professeurs de mathématiques. Les résultats nous aideront aussi dans la préparation de l'opération EVAPM Terminale.

Les conclusions seronty publiées dans un prochain BGV et sur le site INTERNET de l'APMEP :

http:/www.univ-lyon1.fr/apmep

An nom de quelle série répondez-vous :.....

(Vous pouvez, bien sûr, répondre pour plusieurs séries, mais utilisez un questionnaire par série)

I 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. Pour répondre à cette question, classez par ordre préférentiel décroissant de 1 à 6 (1 : le plus important,...) six des objectifs majeurs de cette formation en les choisissant parmi les objectifs proposés ci-dessous :

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
Réponse (par exemple : 1 : I, 2 : G, 3 : M, 4 : D, 5 : A, 6 : J)

1	2	3	4	5	6

II Votre opinion sur des....opinions

Voici quelques opinions entendues dans la salle des profs. Vous pouvez être d'accord, ou un peu d'accord ou pas d'accord avec l'une ou l'autre. Entourez votre choix :

1-(OP1) C'est vrai que les math constituent un instrument de sélection excessif.
D'ACCORD UN PEU D'ACCORD PAS D'ACCORD

2-(OP2) An bac, je préfère qu'il y ait un grand problème avec plusieurs parties plutôt qu'un ensemble de petits problèmes indépendants.
D'ACCORD UN PEU D'ACCORD PAS D'ACCORD

3-(OP3) Dans ma notation, j'attache plus d'importance à la démarche qu'au résultat.
D'ACCORD UN PEU D'ACCORD PAS D'ACCORD

4-(OP4) Quand je corrige, j'aime bien un barème très détaillé sur les résultats à obtenir.
D'ACCORD UN PEU D'ACCORD PAS D'ACCORD

5-(OP5) La démonstration est la seule façon rigoureuse de faire des mathématiques.
D'ACCORD UN PEU D'ACCORD PAS D'ACCORD

6-(OP6) Je préfère des programmes bien définis indiquant ce que je dois et ce que je ne dois pas faire.
D'ACCORD UN PEU D'ACCORD PAS D'ACCORD

A la sortie de la terminale de la série sur laquelle vous répondez, un élève devrait....

7-(OP7) ....pouvoir reconnaître si un nombre entier écrit dans la base 10 est divisible par 4.
D'ACCORD UN PEU D'ACCORD PAS D'ACCORD

8-(OP8) ....pouvoir 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".
D'ACCORD UN PEU D'ACCORD PAS D'ACCORD

9-(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.
D'ACCORD UN PEU D'ACCORD PAS D'ACCORD

10-(OPX) ...estimer à vue, à 30% près, le périmètre et l'aire du plancher ainsi que le volume de la salle de classe.
D'ACCORD UN PEU D'ACCORD PAS D'ACCORD

AJOUT DE LA CONFIANCE AU GRAPHE IMPLICATIF

Souhila GHANEM[1] et Raphaël COUTURIER[2]

ADDING CONFIDENCE INTO IMPLICATIVE GRAPH

RÉSUMÉ

L'analyse statistique implicative développée par Régis Gras et ses collaborateurs, porte sur les mesures statistiques de qualité des règles d'associations, plus particulièrement sur celles issues de la modélisation de la loi du nombre de contre-exemples. Cette dernière basée sur le calcul de l'intensité d'implication permet d'extraire les règles les plus étonnantes sans prendre en considération le nombre de contributions des individus à chaque règle. La probabilité conditionnelle représente la fonction la plus classique servant à la confirmation inductive d'une règle. Dans ce papier nous souhaitons ajouter cette mesure probabiliste au graphe d'implication mis en œuvre dans RCHIC, et définir un seuil laissé au choix de l'utilisateur pour sélectionner toutes les règles qui lui semble intéressantes. À travers l'étude du suivi d'une population de malades, nous montrons l'intérêt de la combinaison des deux mesures, intensité d'implication et confiance.

Mots-clés : Analyse Statistique Implicative, RCHIC, Confiance, intensité d'implication

ABSTRACT

Statistical Implicative Analysis (SIA) developed by Régis Gras and his staff is based on a statistical measure that computes the interest of association rules particularly on those that use the counter-examples number law. The latter based on the computation of the implication intensity allows to extract the most surprising rules without considering the number of contributions to each rule. The conditional probability is the most classic function for the inductive confirmation of a rule. In this paper we wish to add this probability measure to the implicative graph implemented in RCHIC, and define a threshold that the user will choose to select all the rules which seem interesting to him/her. Through the study of a population of patients we show the interest of the combination of both measures, intensity of implication and confidence.
Keywords: *Statistical Implicative Analysis (SIA), RCHIC, confidence, implication intensity*

1 Introduction

Dans le cadre de l'exploration de données, la découverte de règles d'association a comme but l'extraction de règles implicatives potentiellement utiles à partir de données initialement stimulées par des recherches. La probabilité conditionnelle est la fonction la plus classique servant à la confirmation inductive d'une règle d'association. La recherche de règles d'associations a été introduite pour une application populaire qui est l'analyse de panier des clients de supermarché. L'objectif est d'étudier les habitudes d'achats des clients en cherchant les ensembles d'articles qui sont fréquemment achetés ensemble. Par exemple pour dire qu'un client qui achète du *pain* tend à acheter du *lait*

[1] Laboratoire LIMED, Université de Bejaia, souhila.ghanem@gmail.com

[2] Institut Femto-ST, Université de Bourgogne Franche-Comté, Belfort, France, raphael.couturier@univ-fcomte.fr

nous écrivons la règle d'association suivante : *pain=>lait [support=2%. Confiance=60%]*, avec le *support* et la *confiance* qui reflètent respectivement l'utilité (2% de toutes les transactions dans lesquelles le client achète le *pain* et *lait* ensembles) et la certitude (60% des clients qui achètent le *pain* achètent aussi du *lait*) de la règle découverte. Ainsi, une règle est dite intéressante ou forte si elle satisfait les seuils minimums de support et confiance. La confiance et le calcul du support fondent la stratégie utilisée dans les principaux algorithmes d'extraction, à la suite d'Apriori, Agrawal *et al.* (1994). Ces algorithmes prennent en considération le support et le seuil de confiance pour sélectionner les règles. Ces dernières sont trop nombreuses et la majorité d'entre elles n'ont aucun intérêt Freitas, A. (2000) et Gras, R. (2000). Ces algorithmes sélectionnent l'ensemble des règles dont le support et la confiance dépassent les seuils de support et de confiance préalablement fixés. Ce type d'algorithme, exhaustif et déterministe Freitas, A. (2000), produit des règles trop nombreuses dont l'intérêt n'est pas toujours assuré Gras, R. (2000). L'Analyse Statistique Implicative est une méthode d'analyse de données dédiée à l'extraction et à la structuration des règles de quasi implication, elle se distingue des autres méthodes statistiques qui permettent de générer des règles d'association, par le fait qu'elle utilise une mesure non linéaire qui satisfait des critères importants, Couturier, R. (2007, 2008). Tout d'abord, cette mesure est basée sur l'intensité d'implication qui mesure le degré de surprise inhérent à une règle. L'Analyse Statistique Implicative est utilisable par le logiciel CHIC (Classification Hiérarchique Implicative et Cohésitive). Ce dernier a été utilisé dans plusieurs domaines, dans le cadre de la hiérarchisation des compétences tel qu'en science de l'éducation, voir par exemple Malaise, S. (2010), Montpied, P. et al (2010), Leroyer, L. (2013), Gras, R. (2008) et Demeuse, M. (2005). Certaines personnes ne sont pas satisfaites par l'ASI car la valeur de confiance de chaque règle n'est pas connu, en effet, seul le degré d'étonnement de la règle est connu.

Dans ce papier nous proposons de rassembler les deux informations : intensité d'implication et confiance dans le graphe d'implication implémenté dans RCHIC[3], pour avoir en même temps les règles les plus étonnantes et dont le pourcentage de participation est très élevé. Pour montrer l'intérêt de l'ajout de la mesure de confiance nous avons étudié les causes qui conduisent à un état de stress en utilisant des données conçues à partir des échocardiographies de stress. L'utilisation de ces deux mesures aidera les médecins à sélectionner les règles qui sont fortes en intensité d'implication et en confiance. Le choix d'un seuil de confiance élevé permet d'extraire les causes qui induisent avec une grande probabilité à avoir un état de stress. En diminuant le seuil de confiance cette probabilité d'avoir un état de stress diminue, de ce fait ce seuil permet d'avoir un graphe plus lisible

Dans la suite de ce papier nous allons calculer l'intensité d'implication et la confiance pour des données expérimentales. A travers les résultats obtenus, nous montrons comment l'expert peut utiliser ces deux informations ensemble pour extraire les règles qui l'intéressent le plus. Après nous allons appliquer notre approche avec des données issues d'une étude sur l'échocardiographie de stress des malades. Nous

[3]RCHIC est la version écrite dans R du logiciel CHIC (elle est encore en cours de développement)

présentons les expérimentations et les différents résultats obtenus avec des interprétations. Enfin nous terminons par une conclusion.

2 Motivations

Dans le but d'extraire la connaissance à partir des grandes bases de données, plusieurs algorithmes ont été conçus. Parmi, les algorithmes de type apriori, ces derniers se basent sur la recherche des règles d'association intéressantes. L'inconvénient de ces algorithmes et qu'ils produisent beaucoup de règles qui ne sont pas toutes intéressantes, et ils ignorent d'autres règles intéressantes Lallich, S et Teytaud, O. (2004). Pour pallier ces problèmes une autre méthode qui permet de structurer les données en individus et variables a été développée. Il s'agit de l'analyse statistique implicative. La méthode implicative se développe au fil de ces problèmes rencontrés et de ces questions posées. Son objectif majeur vise la structuration de données croisant individus et variables. Dans ce papier nous utilisons le graphe d'implication mis en œuvre dans RCHIC qui permet d'extraire les implications les plus étonnantes basées sur le calcul de l'intensité d'implication. Le graphe d'implication permet de donner toutes les implications étonnantes.

La question qui se pose est : est ce que toutes les règles obtenues avec l'ASI avec une intensité d'implication forte sont intéressantes ? La réponse à cette question dépend du domaine étudié et revient en principe à l'expert du domaine. Dans certains cas le nombre de règles fortes en termes d'intensité d'implication est très élevé et le graphe devient moins lisible. Ceci nous a motivé à intégrer une autre mesure qui permet de classer les règles selon leur importance. Il s'agit de calculer la valeur de confiance pour chaque règle. RCHIC calcul cette valeur et l'enregistre dans un fichier appelé **transaction.out**. En récupérant cette valeur pour chaque règle et en l'ajoutant au graphe d'implication nous pourrons déterminer les règles qui sont fortes en termes d'intensité d'implication et qui ont une confiance élevée. Ces dernières pourront être considérées comme les plus intéressantes. Pour cela nous avons effectué une expérimentation sur des données avec des valeurs représentatives, ensuite nous avons calculé l'indice d'implication et la confiance. La variable *n* correspond à la population totale, les variables *na* et *nb* correspondent respectivement aux nombres d'apparition des propriétés a et b et la variable *nabb* correspond au nombre d'apparition de *a* et de *non b*. Dans les tableaux ci-dessous nous présentons les résultats obtenus.

n	na	nb	nabb	confiance	Intensité d'implication
100	20	40	10	50%	71,81
150	20	40	10	50%	88,84
200	20	40	10	50%	93,31
250	20	40	10	50%	95,14
400	20	40	10	50%	97,03

Tableau 1 – Variation de la taille de l'échantillon n

Dans le Tableau 1, nous avons initialement pris un échantillon de *n* variables parmi lesquelles les variables *na*, *nb*, *nabb* apparaissent respectivement 20, 40 et 10 fois. La

valeur de confiance calculée est égale à 50 % et l'intensité d'implication à 71,81, par la suite nous avons fait varier la taille de l'échantillon. Nous remarquons que la confiance ne dépend pas de la taille de l'échantillon, elle est restée constante. Ceci s'explique par le fait que la confiance de la règle a=>b est égale a la probabilité conditionnelle d'avoir b sachant a, cette dernière est égale au nombre d'apparition de la variable a union b divisé par le nombre d'apparition de la variable a. Ainsi, dans le tableau 1,la confiance indépendante de la taille de l'échantillon. Par contre l'intensité d'implication augmente avec la taille de l'échantillon et montre ainsi la croissance de la surprise lorsque *n* croit.

n	na	nb	nabb	confiance	Intensité d'implication
400	20	40	2	90%	99,99
400	20	40	4	80%	99,95
400	20	40	6	70%	99,76
400	20	40	8	60%	99,07
400	20	40	10	50%	97,03
400	20	40	12	40%	92,13
400	20	40	14	30%	82,71
400	20	40	16	20%	68,13
400	20	40	18	10%	50,00

Tableau 2 – Variation de la variable nabb

Dans le Tableau 2, nous avons fixé les valeurs *n*, *na* et nb et nous avons fait varier le nombre d'apparitions de a et non b (*nabb*). Nous remarquons que les valeurs de confiance et d'intensité d'implication sont très fortes lorsque le nombre de *nabb* est faible et elles diminuent avec sa diminution. Avec l'augmentation du nombre de *nabb*les valeurs de confiance et d'intensité d'implication diminue considérablement, cela apparait logique car plus on a de contre-exemples d'une règle plus le nombre d'exemples qui vérifient la règle diminue ce qui implique la diminution de la valeur de confiance et d'intensité d'implication.

n	na	nb	nabb	confiance	Intensité d'implication
400	20	40	10	50%	97,03
400	25	40	10	60%	99,58
400	28	40	10	64,29%	99,88
400	35	40	10	71,43%	99,99

Tableau 3 – Variation de la variable na

n	na	nb	nabb	confiance	Intensité d'implication
400	20	40	10	50%	97,03
400	20	50	10	50%	96,35
400	20	55	10	50%	95,95
400	20	100	10	50%	90,16

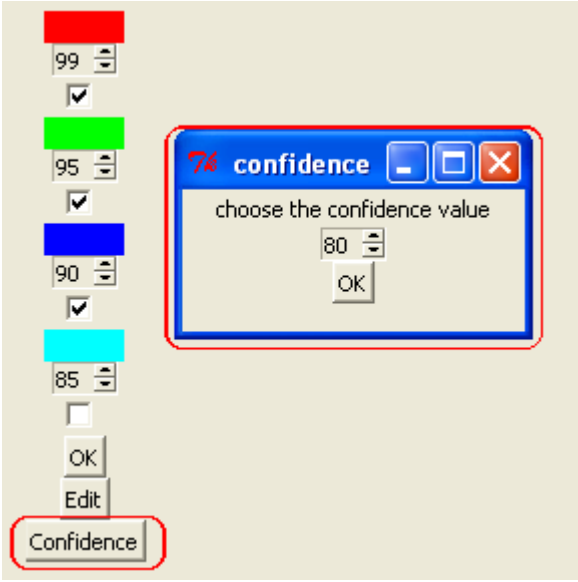
Tableau 4 – Variation de la variable nb

Nous avons fait varier la variable *na* dans le Tableaux 3 et la variable *nb* dans le Tableau 4.En augmentant la valeur de *na*, les valeurs de confiance et d'intensité d'implication augmentent. Par contre en augmentant la variable *nb* la confiance reste constante et l'intensité d'implication diminue.

À partir des résultats présentés dans ces tableaux nous constatons que la modification de la taille de l'échantillon et du nombre d'apparition de la variable en conclusion d'une règle d'implication n'ont aucune influence sur la valeur de confiance par contre en augmentant la taille de l'échantillon l'intensité d'implication augmente. En augmentant le nombre d'apparition de la variable en conclusion d'une règle l'intensité diminue. Dans les cas où la valeur de confiance est faible contrairement à la valeur de l'intensité d'implication qui est très élevée, notre approche offre la possibilité à l'expert de prendre ces implications ou de les ignorer et cela en choisissant un seuil de confiance. Ces dernières sont supposées moins importantes que les règles qui ont une valeur de confiance et d'intensité d'implication élevée qui sont supposées les plus importantes. Dans la suite nous allons voir l'impact d'élimination des règles d'implication qui ont une valeur de confiance faible et cela à travers des données issues des échocardiographies de stress.

3 L'approche proposée et son application sur des données issues des échocardiographie de stress

Notre approche consiste à combiner les deux mesures, l'intensité d'implication et la confiance. Pour cela dans RCHIC nous avons rajouté la possibilité d'afficher les valeurs de la confiance pour chaque règle en ajoutant un bouton « confidence » dans le menu du graphe d'implication. Et pour sélectionner que les règles considérées comme intéressantes nous avons rajouté une fenêtre pour choisir la valeur de confiance désirée.



La suite de cette section présente un exemple concret d'utilisation de RCHIC pour étudier les causes qui conduisent à un état de stress. Nous avons récolté des informations à partir d'une base de données sur l'analyse des données issues des échocardiographie de stress obtenue à l'Université de VANDERBILT par Frank Harrell

Harrell, F. (2015). Il a évalué les échocardiographie de stress de 558 patients jusqu'au 24 février 2015. Le jeu de données contient 31 variables.

L'échocardiographie ou bien échographie cardiaque est un test de diagnostic qui utilise des ondes ultrasonores pour créer une image du muscle cardiaque qui désigne les dimensions et le volume de cœur. Il est ainsi possible de voir la taille, la forme, et le mouvement des soupapes et cavités cardiaques ainsi que la circulation du sang à travers le cœur. L'échocardiographie peut montrer des anomalies du mauvais fonctionnement des valves cardiaques ou des dommages aux tissus du cœur à cause d'une crise cardiaque passé. Avant de présenter les résultats obtenus nous définissons les variables utilisées.

3.1 Variables utilisées

Le jeu de données utilisé comporte les variables suivantes :

bhr : Fréquence cardiaque de base
basebp : Tension de base
basepb : Le produit Bhr* basebp
pkhr: Le sommet de la fréquence cardiaque
sbp : La pression systolique[4]
dp : Le produit pkhr*sbp
dose : Dose de dobutamine[5] donnée
mbp : Pression artérielle maximale
dpmaxdo : Doubler la dose du dobutamine au maximum dans le produit
dobdose : La dose du dobutamine lorsque le produit double en maximum
gender : Le sexe, masculin ou bien féminin
baseEF : La fraction d'éjection cardiaque initiale
dobEF : Fraction d'éjection sur dobutamine
chestpain : Une douleur dans la poitrine
restwma : Repos à anomalie de mouvement de la parois sur échocardiogramme
posSE : Le stress positif échocardiogramme
newMI : Nouvelle crise cardiaque
newPTCA : Angioplastie[6] récente
hxofHT : Le malade est sujet à une hypertension
hxofDM : Le malade est sujet au diabète

[4]Correspond à la pression artérielle mesurée lors de la phase de la contraction du cœur.

[5] La dobutamine est un médicament utilisé pour l'augmentation de la contraction cardiaque, notamment en cas d'insuffisance cardiaque grave

[6] Angioplastie: une technique chirurgicale pour rétablir la circulation normale du sang dans une artère rétrécie ou bloquée par l'athérosclérose, soit par l'insertion d'un ballonnet dans la section rétrécie et le gonfler soit en utilisant un faisceau laser

hxofCig: Si le malade fume : il est soit non fumeur, fumeur modéré ou bien un grand fumeur.
hxofMI: Vérifier si le malade aune crise cardiaque aigue et grave.
hxofPTCA : Vérifier si le malade a subit la chirurgie d'angioplastie
ecg : Diagnostic de base de l'électrocardiogramme[7], permet de déduire trois cas : un cas normal, désigné par la variable acg.Normal, un cas indéterminé désigné par la variable ecg.equivocal (on ne peut pas confirmer s'il ya un trouble cardiaque ou pas) et un cas de crise cardiaque désigné par la variable ecgMI.

Pour lancer l'analyse de ces données sous RCHIC nous avons créé un fichier de type «.csv» (un extrait est montré dans la Figure 1) qui contient en lignes les identifiants des malades et en colonnes les variables. Certains variables sont suivies par la lettre « p ». Ceci précise que les variables sont à partitionner en un nombre fixe d'intervalles. Ensuite l'algorithme des nuées dynamiques Diday, E. (1971)constitue automatiquement les intervalles qui ont des limites distinctes. Nous avons choisi le partitionnement par défaut ce qui veut dire que RCHIC partitionne chaque intervalle en trois sous intervalles. Par exemple, la fréquence cardiaque de base bhr a été partitionnée en trois intervalles: élevée, moyenne et faible. 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. Pour les variables qui contiennent plus d'un sens, on a attribué une variable pour chaque sens par exemple à partir de la variable sexe on a créé deux variables « male » et « female »

	bhr p	basebp p	basedp p	pkhr p	sbp p	dp p	dose p	maxhr p	pctMphr p	mbp p	dpmaxdo p	dobdose p	age p	mal	female	baseEF p	dc
1	92	103	9476	114	86	9804	40	100	74	121	12100	40	85	1	0	27	
2	62	139	8618	120	158	18960	40	120	82	158	18960	40	73	1	0	39	
3	62	139	8618	120	157	18840	40	120	82	157	18840	40	73	1	0	39	
4	93	118	10974	118	105	12390	30	118	72	105	12390	30	57	0	1	42	
5	89	103	9167	129	173	22317	40	129	69	176	22704	40	34	1	0	45	
6	58	100	5800	123	140	17220	40	123	83	140	17220	40	71	1	0	46	
7	63	120	7560	98	130	12740	40	98	71	130	12740	40	81	0	1	48	
8	86	161	13846	144	157	22608	40	144	111	157	22608	40	90	0	1	50	
9	69	143	9867	115	118	13570	40	113	81	151	17063	40	81	0	1	52	
10	76	105	7980	126	125	15750	40	126	94	125	15750	40	86	1	0	52	
11	105	134	14070	171	182	31122	40	171	108	182	31122	40	61	0	1	52	
12	72	112	8064	127	95	12065	30	125	80	101	12625	20	63	1	0	53	
13	90	120	10800	169	184	31096	40	169	126	184	31096	40	86	1	0	54	
14	81	110	8910	110	130	14300	40	110	58	130	14300	40	29	0	1	55	
15	84	176	14784	110	194	21340	40	110	74	194	21340	40	71	0	1	55	

Figure 1 – Extrait du jeu de données

RCHIC calcul la valeur de confiance et d'indice d'implication pour chaque règle, et les enregistre dans un fichier appelé **transaction.out**. Le fichier transaction.out

[7]L'électrocardiographie (ECG) désigne l'examen permettant l'enregistrement du rythme cardiaque. L'ECG consiste à étudier précisément l'activité du cœur, grâce à des électrodes posées sur la poitrine, les poignets et les chevilles. Cette activité est mesurée en plusieurs points du cœur, appelés dérivations, et elle est enregistrée sous la forme d'une courbe pour chacune d'entre elles. L'ECG permet de découvrir des troubles du rythme cardiaque, des troubles de la conduction cardiaque, des signes de souffrance cardiaque.L'ECG peut être utilisé pour identifier une population de patients à faible risque de MI (crise cardiaque). Un ECG normal indique moins de 3% pour le risque de MI et moins de 6% pour le risque de décès dans l'année choisit

correspondant au jeu de données issu des données des échocardiographies de stress contient 3986 règles. Voici une partie de ce fichier :

hyp -> con	occurrence(	occurren	suppor	confidence	classical index	entropic index
maxhr.3 -> dose.2	128.000000000	125.000000	22.939068	28.90625000000	79.82639223337173	33.8972340076598540
dose.2 -> maxhr.3	125.000000000	128.000000	22.401433	29.59999999999	80.18773198127746	34.8368000960555620
maxhr.3 -> HxofCig.hea	128.000000000	122.000000	22.939068	27.34375000000	75.84681957960128	31.9113793722364250
HxofCig.heavy -> maxh	122.000000000	128.000000	21.863795	28.68852459016	76.52886509895324	33.7130189901168580
maxhr.3 -> bhr.3	128.000000000	108.000000	22.939068	38.28125000000	99.14471488445997	46.7913717142461320
bhr.3 -> maxhr.3	108.000000000	128.000000	19.354838	45.37037037037	99.60406208410859	58.2992773010986060
maxhr.3 -> dobEF.1	128.000000000	100.000000	22.939068	16.40625000000	42.49763488769531	18.6547383807354310
dobEF.1 -> maxhr.3	100.000000000	128.000000	17.921146	21.00000000000	41.25890731811523	24.4447615698129790
maxhr.3 -> dp.3	128.000000000	88.0000000	22.939068	39.06250000000	99.79559690691536	48.0348474151621050
dp.3 -> maxhr.3	88.0000000000	128.000000	15.770609	56.81818181818	99.99422854242922	80.2317373407175440
maxhr.3 -> basedp.3	128.000000000	88.0000000	22.939068	27.34375000000	92.31628403067588	32.0774828447578530
basedp.3 -> maxhr.3	88.0000000000	128.000000	15.770609	39.77272727272	96.39814011752605	49.5371237867437630
maxhr.3 -> hxofCABG	128.000000000	88.0000000	22.939068	15.62500000000	49.28393959999084	17.7773868543899950
hxofCABG -> maxhr.3	88.0000000000	128.000000	15.770609	22.72727272727	49.09715056419377	26.7144851954282670
maxhr.3 -> dpmaxdo.3	128.000000000	84.0000000	22.939068	38.28125000000	99.78225424420088	46.9033240263058960
dpmaxdo.3 -> maxhr.3	84.0000000000	128.000000	15.053763	58.33333333333	99.99615147714318	81.9021790430464590
maxhr.3 -> baseEF.1	128.000000000	80.0000000	22.939068	10.93750000000	33.88734459877014	12.3471268562275060
baseEF.1 -> maxhr.3	80.0000000000	128.000000	14.336917	17.50000000000	28.97271513938903	20.4322975199196330

Figure 2 – Extrait du fichier transaction.out correspondant au jeu de données.

Le graphe d'implication est constitué en ne représentant que les implications dont la valeur d'intensité d'implication est supérieure à un seuil choisi par l'utilisateur. Notre approche rajoute le seuil de confiance. Les règles en rouge dans la figure 2 correspondent à des règles d'implication très forte en termes d'intensité d'implication pour lesquelles la confiance est faible, c'est-à-dire que le nombre d'individus qui les réalisent est relativement faible. C'est ici qu'on peut voir l'intérêt de notre approche, qui donne la possibilité à l'expert premièrement d'avoir l'information sur la confiance de chaque règle et la possibilité de prendre ces règles ou de les laisser.

3.2 Résultats obtenus sans spécifier un seuil de confiance

Dans le graphe implicatif nous avons offert à l'utilisateur la possibilité d'afficher la valeur de confiance de chaque règle. Dans la Figure 3, vu le grand nombre de règles, nous avons montré qu'une partie du graphe d'implication. Dans les graphes d'implications qui suivent la flèche rouge désigne un seuil d'intensité d'implication égale à 99, La flèche bleueun seuil d'intensité d'implication égale à 95 et la flèche verte désigne un seuil d'intensité d'implication de 90.

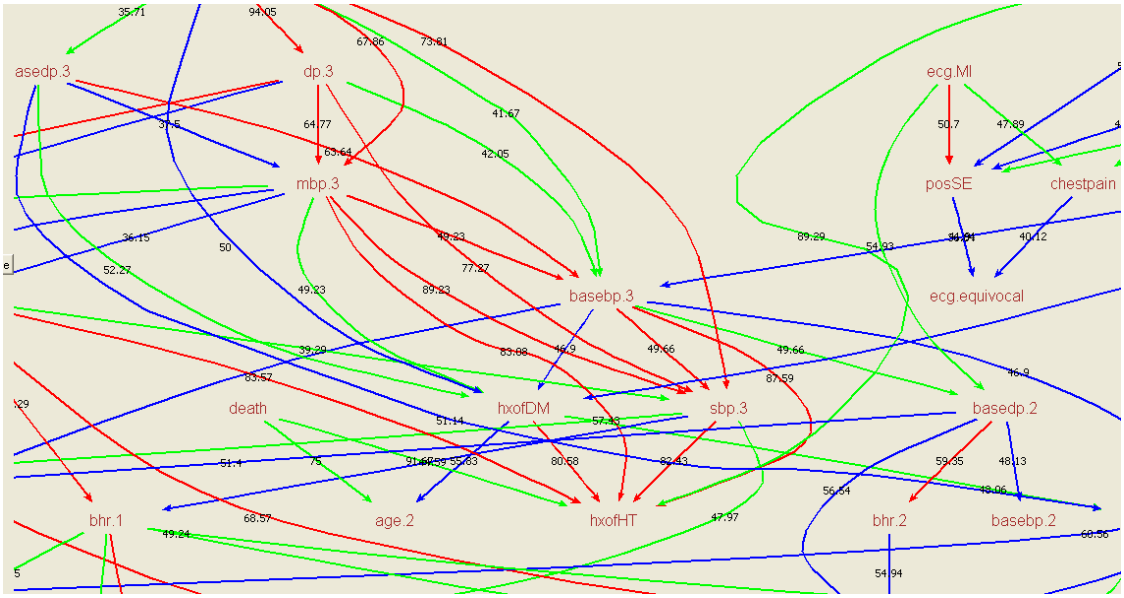


Figure 3 – Extrait du graphe d'implication avec l'affichage des valeurs de confiance

En affichant les valeurs de la confiance, l'utilisateur peut voir l'information sur le nombre de participation des individus à chaque règle, mais vu le grand nombre d'implications l'utilisateur aura des difficultés pour distinguer les règles.On a l'impression que tout est lié. Pour cela nous avons rajouté la possibilité de choisir un seuil de confiance qui nous donne la possibilité de voir le niveau d'importance des règles. Dans ce qui suit nous allons présenter le graphe d'implication en affichant juste les implications qui correspondent au seuil de confiance désiré.

3.3 Résultats obtenus en utilisant un seuil de confiance égale à 80

Le choix d'un indice de confiance élevé va nous permettre d'extraire les règles d'implications dont le taux de participation est très élevé, cela revient à identifier les facteurs les plus importants qui conduisent à un état de stress. Ci-dessous le graphe d'implication correspondant:

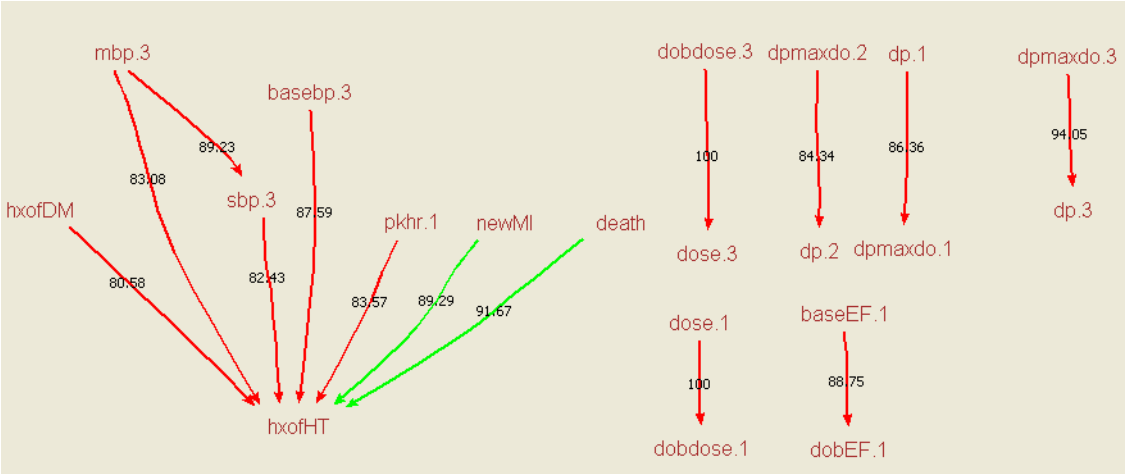


Figure 4 – Graphe implicatif avec un seuil de confiance égale à 80

Avec une valeur de confiance 80 nous remarquons que le graphe est plus lisible et que les résultats sont cohérents car les valeurs faibles impliquent les valeurs faibles et les valeurs élevées impliquent les valeurs élevées. Parmi les implications nous rencontrons les suivantes :

Si la personne est diabétique alors elle a un grand risque de se retrouver en état de stress.

Si la fréquence cardiaque de base est faible alors la personne a tendance à être en état de stress.

Si la pression sanguine maximale et la tension de base sont très élevées alors la personne risque de se retrouver en état de stress.

Si la personne subit une crise cardiaque c'est qu'elle a tendance à se retrouver en un état de stress.

À partir de ces implications nous constatons que les facteurs qui conduisent le plus à un état de stress sont le fait d'avoir:

Le diabète

Une fréquence cardiaque faible

Une pression sanguine maximale très élevée

Une tension de base élevée

Une nouvelle crise cardiaque ;

Car les valeurs de confiance sont très élevées.

3.4 Résultats obtenus en utilisant un seuil de confiance égale à 70

La figure ci-dessous montre le graphe d'implication correspondant aux résultats obtenus avec un seuil de confiance égale à 70

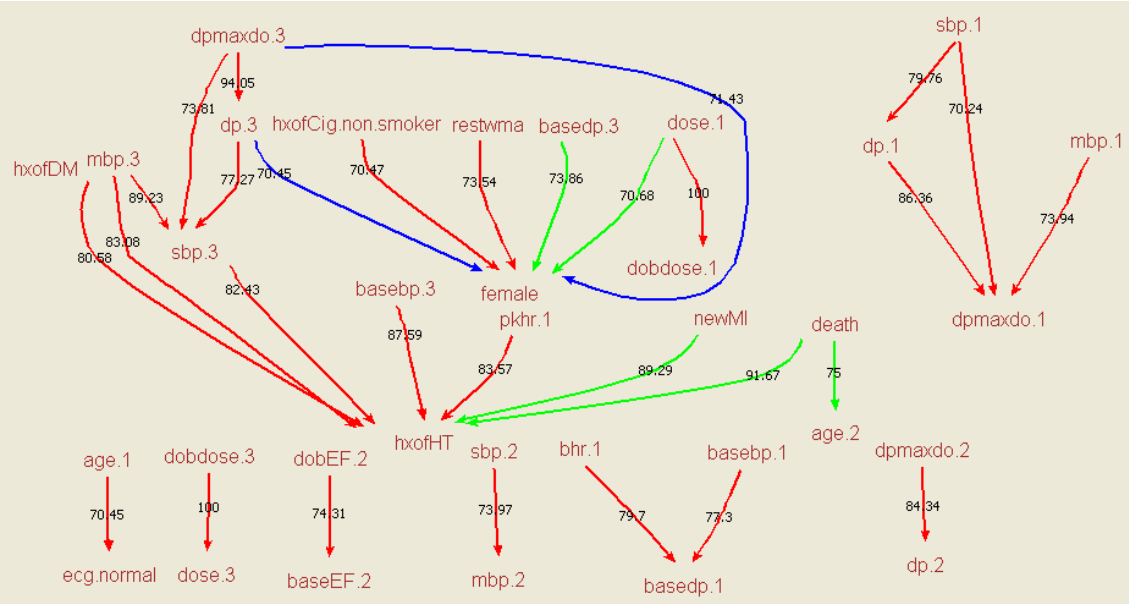


Figure 4 – Graphe implicatif avec un seuil de confiance égale à 70

Avec un seuil de confiance égale à 70, nous remarquons qu'on a, en plus des règles dans le graphe précédent, d'autres règles mais avec un seuil de confiance moins élevé. Ces implications correspondent aux suivantes :

- Les médecins ont utilisé une dose très élevée du double de dobutamine pour 71% des femmes.
- Dans le jeu de données utilisé, plus de 70 % des personnes non fumeurs sont des femmes.
- 73% des personnes qui ont l'anomalie de mouvement de la paroi au repos sont des femmes.
- Les personnes très jeunes ont un diagnostic normal.

3.5 Résultats obtenus avec un seuil de confiance égale à 65

La figure ci-dessous montre le graphe d'implication correspondant aux résultats obtenus avec un seuil de confiance égale à 65

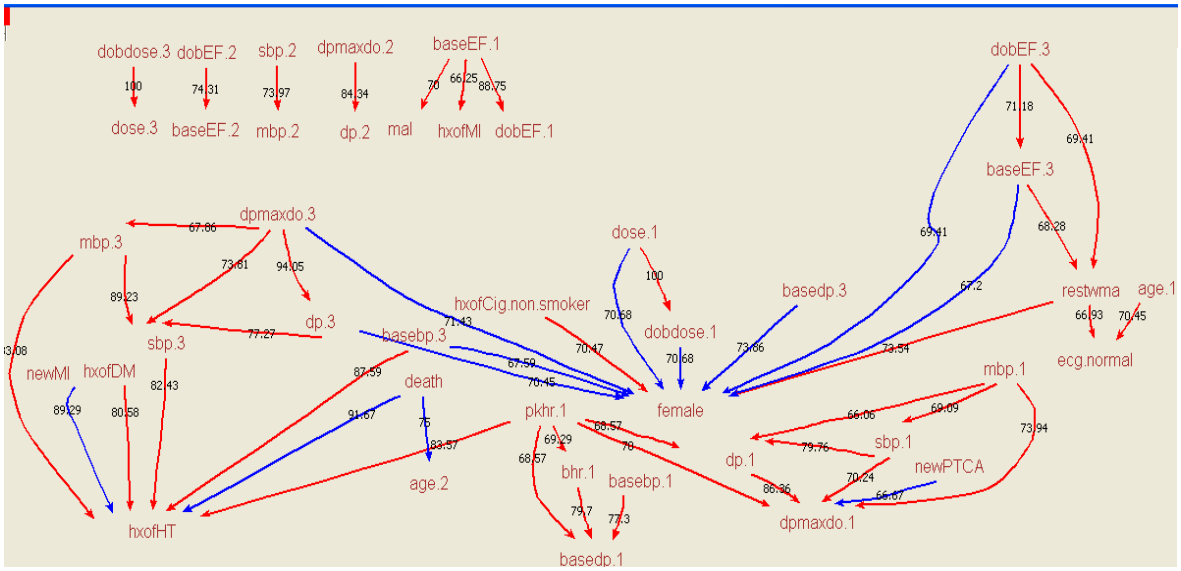


Figure 5 – Graphe implicatif avec un seuil de confiance égale à 65

En diminuant le seuil de confiance à 65, plus d'implications apparaissent parmi lesquelles on voit que:

66% des personnes qui ont l'anomalie de mouvement de la paroi au repos ont tendance à avoir un diagnostic ECG normal.

68 % des personnes qui ont la fraction d'injection cardiaque initiale très élevée et 69% qui ont la fraction d'injection cardiaque élevée sur dobutamine ont tendance à avoir l'anomalie de mouvement de la paroi au repos.

4 Conclusion

Dans ce papier nous avons étudié le couplage de l'indice d'implication avec la confiance afin d'améliorer de faciliter la recherche de règle avec la méthode d'analyse statistique implicative. Notre but est d'aider les experts et utilisateurs à utiliser le graphe d'implication. Pour cela, nous avons rajouté la valeur de la confiance pour chaque règle dans le graphe, ce qui permet de distinguer le niveau d'importance de chaque règle. Nous avons aussi rajouté un seuil de confiance afin de rendre le graphe plus lisible. Après avoir défini l'approche et l'avoir mise en œuvre dans le logiciel de statistique RCHIC, nous avons appliqué notre approche sur différents ensembles de données. Dans ce papier nous avons choisi les données issues des échocardiographies de stress. Les résultats obtenus nous ont permis d'avoir des graphes plus lisibles et d'identifier et de classer les facteurs les plus importants qui conduisent à un état de stress.

Comme perspective à ce travail nous envisageons d'utiliser d'autres mesures que la confiance parmi celles disponibles Guillet, F. et Hamilton, H. J. (2007). De plus, nous pourrions comparer les résultats présentés dans cet article à ceux obtenus en utilisant d'autres mesures telles que l'estimateur de Laplace, DIR Blanchard, J. (2005) ou IPEE Blanchard et al (2005).

Références

[1] Agrawal, R. and Srikant, R. (1994), Fast algorithms for mining association rules, *Proceedings of the 20th VLDB Conference*, Santiago, Chile.

[2] Freitas, A. (2000), Understanding the crucial difference between classification and discovery of association rules - a position paper. SIGKDD *Explorations, vol. 2, 1*, pp.65-69.

[3] Gras, R. (2000), Les Fondements De L'analyse Statistique Implicative, *Quaderni di Ricerca in Didattica*

[4] Couturier, R. (2007). CHIC : utilisation et fonctionnalités, Nouveaux Apports Théoriques à l'Analyse Statistique Implicative et Applications, p. 41-49.

[5] Couturier, R. (2008). CHIC: Cohesive Hierarchical Implicative Classification, In Statistical Implicative Analysis, volume 127 of *Studies in Computational Intelligence*, pages 41-52. Springer.

[6] Malaise, S. (2010), Classification hiérarchique de compétences par l'intermédiaire du logiciel CHIC fondée sur méthode d'analyse statistique implicative, *Actes du congrès de l'actualité de la recherche en éducation et en formation (AREF), université de Genève.*

[7] Montpied, P. et al (2010), La diversité des désirs d'apprendre : quelles informations pratiques l'Analyse Statistique Implicative fournit-elle en regard des curriculums et de l'enseignement des sciences, *A.S.I. 5 Proceedings* 5-7.

[8] Leroyer, L. (2013), Le rapport au support dans le travail de préparation en mathématiques des enseignants du premier degré. pp7-1.

[9] Gras, R. (2008), *Statistical Implicative Analysis Theory and Application*. Book Springer.

[10] Demeuse, M. (2005). Introduction aux théories et aux méthodes de la mesure en psychologie et en sciences de l'éducation. Liège : Les *éditions de l'Université de Liège*.

[11] Lallich, S et Teytaud, O. (2004), Évaluation et validation de l'intérêt des règlesd'association, *Revue des Nouvelles Technologies de l'Information (RNTI)*, Actes EGC-2003: 193-218.

[12] Diday, E. (1971), La méthode des nuées dynamiques,*Journal Revue de statistique appliquée*, volume 19, numéro 2, pages (19-34).

[13] Harrell, F.(2015),http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/CstressEcho.html

[14] Guillet, F. et Hamilton, H. J. (2007). Quality Measures in Data Mining. Springer.

[15] Blanchard, J. (2005).Un système de visualisation pour l'extraction, l'évaluation, et l'exploration interactives des règles d'association, Human-Computer InteractionUniversité de Nantes

[16] Blanchard, J. Guillet, F. Gras, R et Briand, H. (2005), Using information-theoretic measures to assess association rule interestingness», in Proceedings of the fifth IEEE international conference on data mining ICDM'05, IEEE Computer Society.

EXTENSION DE L'ANALYSE STATISTIQUE IMPLICATIVE AU CAS DES VARIABLES CONTINUES QUELCONQUES

Régis GRAS[1], **Jean-Claude REGNIER**[2]

EXTENSION OF THE IMPLICATIVE STATISTICAL ANALYSIS TO ANY CONTINUOUS VARIABLE

RESUME

Après avoir généralisé l'Analyse Statistique Implicative au cas où l'espace des sujets est continu, nous étendons son champ d'application au cas où cette fois les espaces des variables sont continus sur [0; 1]. Ainsi, les variables seront observées sur des intervalles munis d'une loi de répartition continue. Nous procédons, tout d'abord, à l'extension à partir du traitement connu en ASI des variables-intervalles. Puis, nous envisageons un cas particulier où les distributions sur les espaces des variables suivent une même loi uniforme. Enfin, nous traitons le cas général de l'extension aux espaces de variables munis de lois différentes et quelconques.

***Mots-clés** : variables, variable-intervalle, variable continue, densité, intensité d'implication, propension.*

ABSTRACT

After having generalized the implicative statistical analysis to the case in which the subjects' space is continuous, we extend its field of application to the case in which variables' spaces are continuous on [0 ; 1]. Thus, the variables will be observed as interval-variables. We present a specific case where the distributions of the variables' spaces follow the same uniform law. Finally, we treat the general case of the extension to variable spaces with various and undefined laws.

***Keywords:** variables, interval-variable, continuous variable, density, intensity of application, propensity.*

1 Introduction

Une matrice de données numériques croisant classiquement sujets et variables étant donnée, l'Analyse Statistique Implicative (ASI ou SIA en anglais) attribue une mesure de qualité à des énoncés du type « si un sujet satisfait la variable a alors, généralement, il satisfait la variable b ». Nous avons traité (Gras et Régnier, 2012) le cas où, en ASI, l'espace des sujets E est continu muni d'une loi de répartition donnée. Nous abordons ici la situation où, cette fois, les variables actives elles-mêmes sont continues. Quel est le sens de cette continuité ? Elle exprime que, dans la population-mère de sujets, les observations de la variable suivent une loi continue, par exemple une loi gaussienne ou

[1] Ecole Polytechnique de l'Université de Nantes, Équipe DUKE Data User Knowledge, Laboratoire d'Informatique de Nantes-Atlantique (LINA), UMR 6241, Site de la Chantrerie, rue C.Pauc, BP 44306, Nantes cedex 3, e-mail : regisgra@club-internet.fr,

[2] Laboratoire UMR 5191 ICAR – Université Lumière de Lyon –Lyon2, 86 rue Pasteur, 69635 LYON Cedex 07, e-mail : jean-claude.regnier@univ-lyon2.fr

une loi uniforme. La population E de la matrice des données est censée en être une fidèle réalisation.

Précisons. Le type de variable générique que nous étudierons est une variable numérique dont la distribution est continue, de densité donnée et dont les valeurs finies sont prises sur IR ou sur un intervalle de IR. Si sa nature initiale est qualitative, nous nous limiterons au cas où cette « qualité » est quantifiable et ordonnée. Citons, par exemple, des variables qui expriment des saveurs, des sensations, des phénomènes climatiques,… On rencontre en effet ce type de situation dans le cas de données sensorielles. Mais aussi, comme dans (Gras et al, 2001), à la faveur du recueil des notes attribuées à des élèves dans des disciplines variées et conduisant à des intervalles de IR.

L'ensemble V des variables a, b, c… de l'étude peut être constitué de telles variables continues de distributions souvent différentes, de densités de mesure : f_a, f_b, f_c,… et présenter des espaces de valeurs également différents que l'on peut tous ramener par une homothétie convenable, à l'intervalle [0 ; 1]. Si X est la variable aléatoire représentant la valeur de a inférieure à α, alors Prob [$X<\alpha$] = $\int_0^{\alpha} f_a(t)dt$ et de même : Prob [$X\geq\alpha$] = $\int_{\alpha}^{1} f_a(t)dt$ représente la probabilité pour que la valeur de a soit supérieure ou égale à α. Si a(s) (resp. b(s)) sont les valeurs observées chez le sujet s selon a (resp . b), nous évaluerons dans quelle mesure on observe « rarement » dans E : a(s) > b(s).

Pour chaque sujet s de E, espace des sujets supposé ici discret et fini, nous disposons de la valeur numérique réelle prise selon chacune des variables soit : a(s), b(s), c(s),…Ces valeurs traduisent, dans la plupart des cas sémantiques, un degré d'intensité d'adhésion du sujet s à la variable ou de satisfaction de celle-ci. Elles sont, rappelons-le, manifestement ordonnées sur IR ou un intervalle réel.

Dans cet article, nous revenons d'abord sur la situation traitée dans le cadre des variables-intervalles où chaque intervalle est pondéré par la fréquence de ses instanciations. Le découpage de l'ensemble des valeurs de chaque variable par une partition optimale permet de définir les implications d'intervalles. Dans la section suivante 2, nous étudions le cas où les variables sont continues et uniformément distribuées sur l'ensemble de leurs valeurs. Nous définissons alors un indice de propension entre ces variables. Dans les deux dernières sections 3 et 4, nous étendons l'étude au cas où les variables ont une distribution quelconque mais, possiblement, différentes entre elles. Dans un premier temps, nous traitons le problème par la méthode générale de l'implication statistique. Ensuite, nous envisageons la même situation mais par l'approche propensive comme dans la section 2.

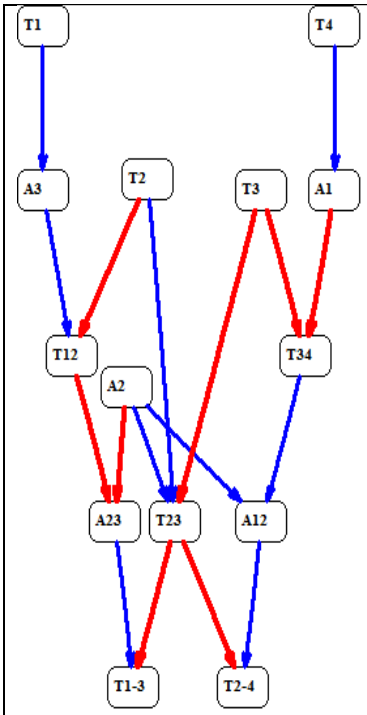
2 Première approche

Pour l'ensemble des variables a, b, c,…, nous examinons, sur leur espace de réalisation ramené à [0, 1], l'ensemble des valeurs prises par les sujets de E selon leurs densités respectives f_a, f_b, $f_{c,\ldots}$non explicitement définies. Notre objectif est de décomposer l'intervalle des valeurs de chaque variable sous forme de sous-intervalles réels. Pour ce faire, sur la base de la répartition des valeurs prises par les n sujets, nous

recherchons la meilleure partition de ces sous-intervalles en un nombre k pour la variable a (l pour b, m pour c, …) maximisant la variance interclasse pour un nombre fixé de sous-intervalles. Cette partition est effectuée selon la méthode inspirée et élargie à partir de celle des « Nuées dynamiques » de E. Diday (1972) et que nous avons utilisée précédemment pour définir l'intensité d'implication entre les variables-intervalles (Gras et al, 2001, 2013). Supposons la partition optimale obtenue constituée des sous-intervalles distincts A_1, A_2,…, A_k (resp. B_1, B_2,…B_l et C_1, C_2, …C_m), eux-mêmes avatars de sous-variables a_1, a_2, …, a_k (resp. b_1, b_2, …,b_l et c_1, c_2, …,c_m.) L'ensemble des valeurs prises par l'ensemble des sujets sur ces sous-intervalles est fini et est représentable par un intervalle de IR, réunion des A_i, que nous pouvons ramener à [0 ; 1] par une homothétie bijective. Nous convenons qu'à A_1 (resp. à A_k), par exemple, s'agrègeront les valeurs non observées « en bout » respectivement à gauche et à droite. Le tableau de données original est remplacé par un nouveau tableau à valeurs binaires. Sur la ligne du sujet x qui prend sa valeur en a selon la modalité a_i (ou de façon équivalente dans A_i) par exemple, on note la nouvelle valeur $1(A_i)$. (mesure de A_i). Ainsi, pour tout x qui satisfait a_i, alors $a_i(x) = 1$ et $a_j(x) = 0$ pour les autres sous-intervalles. Les différentes intensités d'implication des a_i sur les modalités b_j et vice-versa, sont calculées comme pour des variables binaires sous-intervalles par sous-intervalles. Un algorithme permet également de conjoindre, des sous-intervalles contigus d'une variable et d'évaluer les implications optimales d'une variable segmentée vers une autre. Puis les règles sont représentées par un graphe implicatif.

Exemple

On cherche à valider l'hypothèse (certes très vraisemblable) que, généralement, « être de grande taille » en 2013 implique « être jeune ». On dispose pour cela d'un échantillon, pris au hasard dans la population de français, de 77 sujets dont les âges varient de 19 ans à 60 ans et les tailles de 168 cm à 198 cm.



On obtient par C.H.I.C., logiciel de traitement de l'ASI, le graphe implicatif ci-contre lorsque l'on demande que les données numériques de la variable A (âges) soient partitionnées en 3 sous-intervalles optimaux et celles de la variable T (tailles) en 4 sous-intervalles optimaux :

Partitions optimales des données de la variable A :

A1 de 19 à 31 (soit moins de 31)

A2 de 32 à 43

A3 de 46 à 60 (soit plus de 46)

Partitions optimales des données de la variable T:

T1 de 168 à 172 (soit moins de 172 cm)

T2 de 173 à 176

T3 de 178 à 184

T4 de 185 à 198 (soit plus de 185 cm)

On observe que les grandes tailles impliquent (intensité d'implication 0.95) plutôt les jeunes : T4 => A1 (ce qui ne surprendra pas le lecteur) et que ces mêmes jeunes impliquent les tailles de 178 à 198 : A1 => T34

Remarques

1° La prise en compte des mesures des sous-intervalles permet de limiter l'intérêt aux implications à la hauteur de leur importance pondérale.

2° Par la finesse du découpage (k sous-intervalles) certes d'une part on multiplie le temps de calcul, mais d'autre part, on améliore la représentation des nuances de la distribution des instanciations.

3° Une autre approche du problème des variables continues, comparable à l'extension de l'espace des sujets au cas continu, sera à l'étude dans les trois sections suivantes.

4° Ainsi, la distribution a posteriori définie pour chaque variable telle que a à partir de la partition de A en sous-intervalles est explicitement donnée sous forme d'histogramme sur l'intervalle [0 ; 1]. Elle est une des réalisations contingentes de la distribution a priori de a soit f_a. De ce fait, l'implication entre intervalles pourrait être désignée comme implication entre histogrammes.

3 Cas où la variable est uniformément distribuée sur [0 ; 1].

Les extensions de traitements de données dans une approche ASI, développées à partir des premiers travaux de R. Gras, par M. Bailleul (1994) et J.B. Lagrange (1998),

considèrent les variables à valeurs sur [0 ; 1] comme des variables discrètes. Ce que nous souhaitons introduire maintenant, est la prise en considération de variables continues à valeurs sur l'intervalle [0 ; 1] ou s'y ramenant. Ici, nous allons, dans un premier temps, étudier le cas des variables continues sur [0 ; 1] dont la distribution de probabilité est uniforme sur cet intervalle et identique pour toutes les variables.

3.1 Propriétés de la variable Z, produit de deux variables aléatoires continues de loi uniforme sur [0 ; 1]

La problématique de l'étude présente s'exprime ainsi :

Exprimer par un nombre compris entre 0 et 1, dans quelle mesure l'observation, sur un ensemble de sujets E, des valeurs (numériques ou numérisées) prises par une variable continue a s'accompagne généralement de l'observation de valeurs plus grandes prises par une variable continue b.

Nous satisfaisons bien ainsi la philosophie de l'Analyse Statistique Implicative puisque nous voulons quantifier la tendance de la variable a à prendre des valeurs sur E inférieures aux valeurs prises selon la variable b. Cette quantification porte le nom de **propension**. Nous suivons la même procédure que celle suivie par J.-B. Lagrange à propos des variables modales discrètes. Soient X et Y variables aléatoires continues à valeurs dans [0 ; 1] de loi uniforme continue. Elles représentent les valeurs aléatoires des variables a et b qui se réaliseront dans l'observation. Comme il est coutumier en Analyse Statistique Implicative, les deux variables X et Y sont supposées indépendantes. Nous savons alors que la densité de probabilité de X ou de Y est donnée par la fonction $f(u)=1_{[0\,;\,1]}(u)$ de laquelle nous pouvons calculer la fonction de répartition :

$$F(u) = \begin{cases} 1 & u > 1 \\ u & 0 \le u \le 1 \\ 0 & u < 0 \end{cases}$$

Par ailleurs sans difficulté on montre que les espérances respectives de X et Y valent $E(X) = E(Y) = \frac{1}{2}$ et les variances $V(X) = V(Y) = \frac{1}{12}$.

De manière évidente, la variable $\overline{Y} = 1 - Y$ est encore une variable aléatoire à valeur dans [0 ; 1] de loi uniforme continue dont l'espérance et la variance sont celles de Y.

Poursuivant le chemin réalisé dans la construction de l'indice de propension établi avec des variables modales, nous posons Z = X(1-Y). Z est encore une variable aléatoire continue à valeurs dans [0 ; 1] mais qui n'admet plus une densité uniforme. On montre que la densité de probabilité de Z est $g(u) = -\ln(u)\, 1_{]0\,;\,1]}(u)$ ou encore que sa fonction de répartition est :

$$G(u) = \begin{cases} 1 & u > 1 \\ -u\ln(u) + u & 0 < u \leq 1 \\ 0 & u \leq 0 \end{cases}$$

À partir de ces informations, nous pouvons calculer l'espérance puis la variance de la variable aléatoire continue à valeurs sur [0; 1] de la manière suivante:

$E(Z) = \int_{0+}^{1} uf(u)du = \left[-\frac{1}{2}u^2 \ln(u) + \frac{1}{4}u^2 \right]_{0+}^{1} = \frac{1}{4}$ mais comme Z est le produit de deux variables indépendantes, nous pouvons obtenir directement : $E(Z) = E(X)\,E(1-Y) = \frac{1}{4}$

Pour ce qui est de la variance, nous savons que :

$V(Z) = E[(Z-E(Z))^2] = E(Z^2)-[E(Z)]^2$

Or $E(Z^2) = \int_{0+}^{1} u^2 f(u)du = \left[-\frac{1}{3}u^3 \ln(u) + \frac{1}{9}u^3 \right]_{0+}^{1} = \frac{1}{9}$

donc la variance de Z vaut $V(Z) = \left(\frac{\sqrt{7}}{12} \right)^2$

3.2 Indice de propension et intensité de propension pour des variables aléatoires continues de loi uniforme sur [0 ; 1]

Pour suivre la procédure de construction qui mène à la formule de l'indice de propension, nous considérons le n-uplet de variables iid $(Z_1, Z_2, \ldots, Z_n)$.

Posons $T_n = \frac{1}{n}\sum_{k=1}^{k=n} Z_k$. Il en résulte que $E(T_n) = \frac{1}{n}\sum_{k=1}^{k=n} E(Z_k) = \frac{1}{4}$ ainsi que $V(T_n) = \left(\frac{\sqrt{7}}{12\sqrt{n}} \right)^2$ puisque les variables Z_k sont indépendantes deux à deux. Nous sommes en mesure de fournir l'expression algébrique exacte de l'indice de propension entre les deux variables X et Y.

$$Q_u(X, \overline{Y}) = \frac{T_n - \frac{1}{4}}{\frac{\sqrt{7}}{12\sqrt{n}}} = \sqrt{n}\,\frac{T_n - \frac{1}{4}}{\frac{\sqrt{7}}{12}} = \sqrt{n}\,\frac{\left[\frac{1}{n}\sum_{k=1}^{k=n} X_k(1-Y_k)\right] - \frac{1}{4}}{\frac{\sqrt{7}}{12}}$$

On peut considérer que la variable aléatoire $Q_u(X, \overline{Y})$ indice de propension entre deux variables continues de loi uniforme sur [0 ; 1] suit approximativement une loi de Laplace-Gauss centrée réduite dans la mesure où sa construction même nous place dans les conditions d'un des théorèmes central-limite.

L'intensité d'implication se calcule donc de la façon suivante :

$$\varphi_u(X,Y) = 1 - \Pr ob\left[Q_u(X,\bar{Y}) \le q_u(X,\bar{Y})\right] = \frac{1}{\sqrt{2\pi}} \int_{q_u(X,\bar{Y})}^{\infty} e^{-\frac{t^2}{2}} dt$$

4 Cas général traité par la méthode ASI classique

4.1 Problématisation

Pour chaque variable observée, nous avons jusqu'alors examiné, sur son espace de réalisation, l'ensemble des valeurs prises sur l'ensemble des sujets E. Nos travaux sur les variables numériques et modales ont porté, comme dans la section 2, sur la définition d'une mesure de propension dont nous venons de rappeler une modélisation. A l'instar de J.B. Lagrange (1998) et de S. Guillaume (2000), nous utilisons encore l'expression : **propension** (ou tendance) de a vers b, dès lors que l'on rencontre généralement dans E peu de transactions s∈ E dans lesquelles a(s) > b(s)) pour la relation d'ordre sur [0 ; 1] où a(s) et b(s) sont respectivement les valeurs de a et b observées en s. Notons que si a(s) et b(s) sont des valeurs de rang ou d'intensité préférentielle, on peut exprimer l'inégalité a(s) > b(s) par : « a est préférée à b en s »

Notre exploration implicative, notée encore a=>b, dans le cas des variables continues est ici plus générale et s'exprime de la façon suivante :

Dans quelle mesure peut-on considérer que, pour x∈ E « ***si a(x) ≥ α, alors on peut affirmer que b(x) ≥ β*** » (F1) où α et β sont des réels de l'intervalle [0 ; 1] choisis dans l'analyse. Autrement dit, observe-t-on généralement que : « *si une réalisation pour x selon a dépasse la valeur α alors sa réalisation selon b dépasse aussi la valeur β* ? » (F2). Cette fois, contrairement à ce qui est envisagé dans la section 2, ce sont toutes les valeurs de deux intervalles, pondérés par des densités, qui sont mobilisées. De plus, les bornes de ces intervalles peuvent être indépendantes et de sémantique différentes. La réponse confère à la mesure associée une valeur prédictive.

Notons que si $\alpha = \beta$ égalité à laquelle on peut toujours ramener les deux ensembles de valeurs des deux variables, nous retrouvons la formulation de l'ensemble des contre-exemples, en tout point comparable à celle de la propension conceptualisée précédemment pour les variables numériques (Gras et al, 2009, 2013). En effet, un contre-exemple aux énoncés des règles (F1) et (F2) apparaît dès lors que a(s) > b(s). Comparable certes, mais pas identique épistémologiquement car le caractère continu demeure par la prise en compte des intervalles. Or, rappelons que dans le cadre des variables numériques a et b (Gras et Régnier, 2009, 2013), pour tout i∈ E, nous notons $\bar{b_i}$ le complément à 1 de bi : $\bar{b_i}$ = 1-bi, valeur de la variable b qui récuse b_i. On choisit 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 (Gras et Régnier, 2009). Toutefois, cet indice ne tient plus dans le cas qui nous intéresse maintenant où les variables sont continues.

4.2 Formalisation de l'intensité d'implication entre variables continues

La formalisation va s'inspirer de celle retenue dans le cas où l'espace des sujets est continu (Gras et Régnier, 2013, p. 165-175). Considérons les sous-ensembles de sujets A et B de E admettant les valeurs respectives selon les variables a et b et satisfaisant les inégalités suivantes : $A=\{s\in E\ /a(s) \geq \alpha\}$ et $B=\{s\in E\ /b(s) \geq \beta\}$ où α et β sont deux valeurs quelconques appartenant à l'intervalle [0 ; 1]. Les cardinaux de A et B sont respectivement n_a et n_b (avec $n_a \leq n_b$). Par conséquent, les sujets s du sous-ensemble $A \cap \overline{B}$ vérifient $a(s) \geq \alpha$ et $b(s) < \beta$ ou encore $1-\beta \leq 1-b(s)<1$

Comme nous le faisons dans le cas binaire, nous comparons le nombre de contre-exemples à l'implication a => b (au sens de la relation F1 indiquée dans la section 3.1) observés dans la contingence et celui que l'on obtiendrait si les variables a et b étaient indépendantes. Pour cela, choisissons respectivement au hasard, de façon indépendante, deux parties $\hat{X}$ et $\hat{Y}$ de E de mêmes cardinaux respectifs que A et B et vérifiant les mêmes inégalités que ces deux sous-ensembles. Cependant la probabilité associée au tirage de $\hat{X}$ et $\hat{Y}$ parmi les parties satisfaisant les propriétés cardinales doit être affectée de celle qui traduit les contraintes pesant sur les valeurs prises selon a et b par les sujets des sous-ensembles $\hat{X}$ et $\hat{Y}$. Cette seconde probabilité est calculée à partir de la distribution produit entre les lois respectives de a et de b. Or, sous l'hypothèse d'indépendance *a priori* de a et b, la loi produit est égale au produit des lois respectives de a et b de densités f_a et f_b nulles à l'extérieur de [0 ; 1]. Les variables a et $\overline{b}=1-b$ sont également indépendantes et $f_{\overline{b}}=1-f_b$ est la densité de probabilité de $\overline{b}$. Les fonctions de répartition associées sont respectivement F_a, F_b et $F_{\overline{b}}$.

Ainsi la probabilité qu'un sujet pris au hasard ait un comportement vis-à-vis des modalités continues de a et de celles de b qui soit un contre-exemple à a => b (c'est à dire que $a(s) \geq \alpha$ tandis que $0 \leq b(s) < \beta$ ou $1-\beta \leq \overline{b}(s) <1$) est

$$\left[\int_{\alpha}^{1} f_a(t)dt\right]\left[\int_{0}^{\beta} f_b(u)du\right]=\left[\int_{\alpha}^{1} f_a(t)dt\right]\left[\int_{1-\beta}^{1} f_{\overline{b}}(u)du\right]=\left[1-F_a(\alpha)\right]\left[1-F_{\overline{b}}(1-\beta)\right]$$

A chaque sujet s de $\hat{X}\cap\hat{\overline{Y}}$, nous associons une variable de Bernoulli de paramètre p, qui est la probabilité qu'il soit à la fois dans l'ensemble des contre-exemples et qu'il vérifie la double inégalité $a(s)\geq \alpha$ et $b(s)<\beta$. Ces quatre événements étant indépendants par hypothèse, la probabilité p s'écrit comme produit de leurs probabilités, à savoir :

$$p=\frac{n_a n_{\overline{b}}}{n^2}\left[\int_{\alpha}^{1} f_a(t)dt\right]\left[\int_{0}^{\beta} f_b(u)du\right]$$

De là, Card($\hat{X}\cap\hat{\overline{Y}}$), somme des variables de Bernoulli indépendantes associées à chaque sujet de $\hat{X}\cap\hat{\overline{Y}}$ peut être considérée comme une variable binomiale de paramètres n et p.

$$\Pr\left[\text{Card}(\hat{X}\cap\hat{\overline{Y}})=k\right]=C_n^k p^k (1-p)^{n-k}$$

Espérance et variance de cette variable aléatoire sont donc :

E[Card ($\hat{X}\cap\hat{\overline{Y}}$] = np et Var [Card ($\hat{X}\cap\hat{\overline{Y}}$)]= np(1-p)

Par suite,

$$\Pr[\mathrm{Card}(\hat{X} \cap \hat{\overline{Y}}) \leq \mathrm{Card}(A \cap \overline{B})] = \sum_{k=0}^{k=\mathrm{Card}(A \cap \overline{B})} C_n^k p^k (1-p)^{n-k}$$

Comme dans le cas de variables discrètes, nous définissons l'intensité d'implication de a sur b par :

$$\varphi(a,b) = 1 - \Pr\left[\mathrm{Card}(\hat{X} \cap \hat{\overline{Y}}) \leq \mathrm{Card}(A \cap \overline{B})\right]$$

Remarque 1

Cette approche a-t-elle un sens dans sa restriction au cas de variable binaires ? La réponse est oui. En effet si les variables a et b sont binaires, la relation d'implication a => b n'est pas satisfaite lorsque pour $s \in \hat{X} \cap \hat{\overline{Y}}$, a(s) = 1 et b(s) = 0 ou $\overline{b}(s) = 1$. Les densités de probabilité sont dégénérées. Pour que l'on ait Pr[{$s \in X$ | a(s)=1]=1, il est suffisant que α= 0. De même, pour que ait Pr[{$s \in \overline{Y}$ | b(s)=0]=1, il est également suffisant que β=1. Les deux intégrales définissant p sont alors égales à 1 et p devient égale à $\frac{n_a n_{\overline{b}}}{n^2}$, expression effectivement conforme à ce que nous avons dans la théorie de l'ASI dans le cas binaire.

Remarque 2

Comme nous le montrons dans (Gras et al, 2009, 2013), nous pouvons modéliser le cardinal aléatoire de $X \cap \overline{Y}$ par une variable de Poisson de paramètre

$$\lambda = \frac{n_a n_{\overline{b}}}{n} \left[\int_{\alpha}^{1} f_a(t)dt\right]\left[\int_{0}^{\beta} f_b(u)du\right]$$

Par suite $\forall s \in \{0,1,2,...,n\}$ $\Pr[\mathrm{Card}(\hat{X} \cap \hat{\overline{Y}}) = s] = \frac{\lambda^s}{s!} e^{-\lambda}$

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

Pour $\lambda \geq 5$, la variable « indice d'implication empirique », notée :

$$Q(a,\overline{b}) = \frac{\mathrm{Card}(X \cap \overline{Y}) - E\left[\mathrm{Card}(\hat{X} \cap \hat{\overline{Y}})\right]}{\mathrm{Var}\left[\mathrm{Card}(\hat{X} \cap \hat{\overline{Y}})\right]} = \frac{\mathrm{Card}(\hat{X} \cap \hat{\overline{Y}}) - \lambda}{\sqrt{\lambda}}$$

qui résulte du centrage-réduction de la variable de Poisson, $\mathrm{Card}(\hat{\overline{X}} \cap \hat{\overline{Y}})$, peut être approchée par la variable gaussienne centrée réduite N(0 ;1).

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

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

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, $\Pr[Q(a,\bar{b}) \le q(a,\bar{b})]$ 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 Cas général traité par la méthode ASI propensive

5.1 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** entre variables modales qui généralise l'indice d'implication entre variables binaires. En posant les conditions suivantes :

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

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

Définition 1: L'indice de propension de variables modales est :

$$\tilde{q}(a,\bar{b}) = \frac{\sum_{x \in E} a(x)\bar{b}(x) - \frac{n_a n_{\bar{b}}}{n}}{\sqrt{\frac{(n^2 s_a^2 + n_a^2)(n^2 s_{\bar{b}}^2 + n_{\bar{b}}^2)}{n^3}}}$$

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.

Tout en suivant une modélisation comparable à celle de la section 2.1 où la loi de chaque variable est celle de la variable continue uniforme sur [0 ; 1], nous envisageons le cas où les variable sont continues, mais de lois données différentes et à valeurs ramenées par normalisation sur l'intervalle [0 ; 1]. Par ailleurs, notre recherche porte, comme dans la section 3, sur la relation F 1 :

Dans quelle mesure peut-on considérer que, pour $x \in E$ « ***si a(x) ≥ α, alors on peut affirmer que b(x) ≥ β*** » ?

Reprenons les notations de la section 2 : X (resp. Y, resp . $\bar{Y}$) est une variable aléatoire de [0 ; 1], ensemble normalisé des valeurs a_i (resp. b_j et 1- b_j) de la variable a

(resp. b, resp. $\bar{b}$) mais de loi de densité de probabilité f_a (resp. f_b et f_{1-b}) quelconque. De cette manière nous avons Prob $[X \geq \alpha] = \left[\int_{\alpha}^{1} f_a(t)dt\right]$ et Prob$[Y < \beta] = \left[\int_{0}^{\beta} f_b(u)du\right]$.

Les variables aléatoires X et Y (donc également $\bar{Y}$) sont, par hypothèse commune en ASI, indépendantes. Leurs paramètres, espérance et variance, sont calculables et donnés par les formules habituelles.

Introduisons la variable aléatoire $Z = X\bar{Y} = X.(1-Y)$. Elle représente la conjonction d'une réalisation des variables a et 1-b. La loi de Z est le produit des lois de X et de 1-Y et sa densité de probabilité est le produit des celles des lois de probabilité de X et 1-Y. L'espérance de Z est alors connue par : $E(Z) = E(X)E(1-Y) = M_1$ et sa variance est : $V(Z) = E(Z^2) - [E(Z)]^2 = M_2 - M_1^2$

Nous considérons le n-uplet $(Z_1, Z_2, \ldots, Z_n)$ composé de n variables aléatoires indépendantes de même loi que Z et dont la réalisation est celle des n sujets de E. Nous définissons alors comme dans la section 2, la variable somme : $T_n = \frac{1}{n}\sum_{k=1}^{k=n} Z_k$. Il en résulte que l'espérance $E(T_n) = \frac{1}{n}\sum_{k=1}^{k=n} E(Z_k) = M_1$ ainsi que la variance $V(T_n) = \frac{1}{n}\sum_{k=1}^{k=n} V(Z_k) = \frac{1}{n}(M_2 - M_1^2)$ puisque les variables Z_k sont mutuellement indépendantes.

En utilisant le théorème de la limite centrale dit de Lindeberg-Lévy puisque l'indice de propension est une moyenne de variables aléatoires indépendantes identiquement distribuées, on démontre que T_n suit approximativement, pour n grand, la loi de Laplace-Gauss 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 »

$$\tilde{Q}(a, \bar{b}) = \frac{T_n - M_1}{\sqrt{\frac{1}{n}(M_2 - M_1^2)}}$$

est approximativement la loi gaussienne centrée réduite N(0 ; 1).

Dans une mise en pratique, la calcul d'une réalisation de la variable « indice de propension empirique », nécessite d'estimer l'espérance et la variance de la variable T_n. Pour ce faire, nous utilisons m_a la moyenne des observations a_i et m_b la moyenne des observations b_i ; qui nous permet d'obtenir une estimation de la moyenne M_1 avec $m_a(1-m_b)$. Ensuite nous réalisons une estimation de la variance de T_n à partir de la variance v_a des a_i et v_b celles des b_i, une estimation du moment M_2 d'ordre 2 est obtenue à partir des observations a_i et b_i par $(v_a + m_a^2)(v_b + m_{\bar{b}}^2)$ puisque les observations b_i et $\bar{b}_i$ 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 :

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

Nous avons démontré (Gras, Régnier, 2013) que la réduction de ces formules exprimées dans le cas continu à l'instar des variables numériques au cas binaire était parfaitement valide.

6 Conclusion

Nous avons examiné différentes situations relatives aux variables principales continues dans le cadre de l'Analyse Statistique implicative de l'A.S.I. Partant de la situation la plus élémentaire où ces variables sont uniformément distribuées, nous avons considéré les variables continues à distribution quelconque sur [0 ; 1]. Après un retour simplificateur par les variables continues par intervalle, nous avons établi des lois permettant de calculer le critère fondamental en A.S.I., l'intensité d'implication qui évalue la qualité implicative d'une variable sur une autre, leurs distributions pouvant être différentes. Nous pouvons retenir la méthodologie adaptée pour traiter le cas continu dans d'autres types d'analyse de données. De plus, nous retenons la capacité dont nous bénéficierons dans des applications où il s'agit de nuancer spécifiquement la connaissance des variables à prendre en compte dans l'analyse, comme par exemple nuancer l'information sur la complexité *a priori* de ces variables pour une population de sujets donnée.

Références

[1] 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

[2] Diday E, (1972), *Nouvelles méthodes et nouveaux concepts en classification automatique et reconnaissance des formes*, Thèse d'Etat, Université de Paris VI.

[3] Gras R., Régnier J.C. et Guillet F. (2009). *L'Analyse Statistique Implicative. Une méthode d'analyse de données pour la recherche de causalités*, RNTI-E-16, Cépaduès Editions

[4] Gras R., Diday E., Kuntz P et Couturier R. (2001), Variables sur intervalles et variables-intervalles en analyse statistique implicative, *Actes du 8ème Congrès de la Société Francophone de Classification, Université des Antilles-Guyane, 17-21 décembre 2000,*166-173

[5] Gras R. et Régnier J.-C. (2013). Extension de l'A.S.I. aux variables non binaires in *L'Analyse Statistique Implicative, Méthode exploratoire et confirmatoire à la recherche de causalités*, Cépaduès Editions, 70-77

[6] Lagrange J.-B., (1998), Analyse implicative d'un ensemble de variables numériques ; application au traitement d'un questionnaire à modalités modales ordonnées, *Revue de Statistique Appliquée*, XLVI, 71-93.

[7] Régnier, J.-C., & Gras, R. (2005) Statistique de rangs et Analyse Statistique Implicative. *Revue de Statistique Appliquée*. 53(1) p.5-38

Ouvrages de référence

[1] *L'implication statistique. Nouvelle méthode exploratoire de donnée*, sous la direction de R.Gras, et la collaboration de S. Ag Almouloud, M. Bailleul, A. Larher, M. Polo, H. Ratsimba-Rajohn, A.Totohasina, La Pensée Sauvage, Grenoble (1996)

[2] *Mesures de Qualité pour la Fouille de Données*, H.Briand, M.Sebag, R.Gras er F.Guillet eds, RNTI-E-1, Cépaduès, 2004

[3] *Quality Measures in Data Mining* , F.Guillet et H.Hamilton eds, Springer, 2007,

[4] *Statistical Implicative Analysis, Theory and Applications*, R.Gras, E. Suzuki, F. Guillet, F. Spagnolo, eds, Springer, 2008.

[5] *Analyse Statistique implicative. Une méthode d'analyse de données pour la recherche de causalités*, sous la direction de Régis Gras, réd. invités R. Gras, J.C. Régnier, F. Guillet, Cépaduès Ed. Toulouse, 2009.

[6] *Teoria y Aplicaciones del Analisis Estadistico Implicativo*, Eds : P.Orus, L.Zamora, P.Gregori, Universitat Jaume-1, Castellon (Espagne), ISBN : 978-84-692-3925-4, 2009..

[7] *L'Analyse Statistique Implicative : de l'exploratoire au confirmatoire*. Eds : J.C. Régnier, Marc Bailleul, Régis Gras, Université de Caen, ISBN : 978-2-7466-5256-9, 2012

[8] *L'analyse statistique implicative, Méthode exploratoire et confirmatoire à la recherche de causalités*, sous la direction de Gras R., eds Gras R., Régnier J.-C., Marinica C., Guillet F., Cépaduès Editions, 522 pages, ISBN 978.2.36493.056.8, 2013.

CLASSIFYING OBJECTIVE INTERESTINGNESS MEASURES BASED ON THE TENDENCY OF VALUE VARIATION

Nghia QUOC PHAN[1], Hiep XUAN HUYNH[2], Fabrice GUILLET[3], Régis GRAS[4]

CLASSIFICATION DES MESURES D'INTÉRÊT OBJECTIVES BASÉE SUR LA TENDANCE DE VARIATION DES VALEUR

RÉSUMÉ

Dans les dernières années, la recherche de découverte de connaissances à partir des données a intéressé à nombreux chercheurs et de nombreux résultats de recherche ont été utilisés efficacement dans de plusieurs domaines de la vie. La mesure d'intérêt joue un rôle important dans le domaine de recherche de découverte des connaissances. C'est pourquoi l'étude sur les mesures d'intérêt qui ont déjà développés est de plus en plus importante. La plupart des études se sont basées sur deux méthodes principales: classification basée sur les propriétés d'une mesure et de classification basée sur les comportements d'une mesure. Dans notre étude, nous allons étudier et examiner la tendance de variation de la valeur des mesures d'intérêt objectives qui répondent à la nature asymétrique en prenant la dérivée partielle de la fonction qui calcule la valeur des mesures d'intérêt selon le tableau de contingence 2x2. Nos résultats montrent que les mesures d'intérêt objectives asymétriques sont classées par la considération de la variation (l'augmentation, la réduction ou la stabilité) à partir des formules dérivées de chaque mesure.

***Mots-clés**: Mesures d'intérêt objectives, classification, dérivées partielles, tendance des valeurs de variation, intensité d'implication.*

ABSTRACT

In recent years, research in Knowledge Discovery in Databases has been a topic of increasing interest for many researchers. Currently, a variety of such research has effectively been used in many areas of life. Interestingness measures play an important role in the field of research in knowledge discovery. Therefore, the study of the classification of interestingness measures is also an important topic for researchers. Classification of measures is mostly based on two main methods: classification based on the properties of measures and classification based on the behavior of measures. In this study, we propose a new classification method focusing on the research and study of the value variation of the objective interestingness measures that satisfy the asymmetric nature, by taking the partial derivative of the function that calculates the value of interestingness measures according to the 2×2 contingency table. Our results show that asymmetrical objective interestingness measures are classified by considering the increasing, decreasing or stable derivative formula for each measures.

***Keywords**: objective interestingness measures, classification, partial derivative, tendency of value variation, implication intensity.*

[1] TraVinh University, 126 National Road 53, Tra Vinh City, Vietnam, nghiatvnt@gmail.com
[2] Can Tho University, 3/2 Street Ninh, Kieu District, Can Tho City, Vietnam, hxhiep@ctu.edu.vn
[3]University of Nantes, Polytechnic School of Nantes University, France, fabrice.guillet@univ-nantes.fr
[4]University of Nantes, Polytechnic School of Nantes University, France, regisgra@club-internet.fr

1 Introduction

Interestingness measures (Piatetsky-Shapiro and Matheus, 1994; Guillet and Hamilton, 2007), both subjective and objective measures (Silberschatz and Tuzhilin 1995), play an important role in term of evaluating the quality of knowledge according to association rules. In the past ten years, the study of interestingness measures (Jalali-Heravi and Zaïane, 2010; Jon Hills *et al.*, 2012; David Glass, 2013; Martínez-Ballesteros *et al.* 2014) has developed significantly. Researchers in this field only focus on two main directions: (1) proposing new measures (Liu, 2008; David Glass, 2013; Selvarangam and Ramesh Kumar, 2014); and (2) studying the properties, behaviors and trends of the variation of the measures to rank, cluster and classify these measures (Tan *et al.*, 2002; Lenca *et al.*, 2004; Blanchard *et al.*, 2009; Huynh *et al.*, 2005, 2006, 2007), which assists users to select appropriate measures for their specific application. Since the number of interestingness measures have been increased, the concern of researchers have been increased, especially for the classification of these measures as:

- Classification based on objective and subjective criteria (Silberschatz and Tuzhilin, 1995, 1996);
- Evaluation of a good measures based on three principles (Piatetsky-Shapiro, 1991) based on the key properties in order to evaluate the measures to conclude that each measures only satisfies a number of properties and possibly appiesl in some specific areas (Tan *et al.*, 2002, 2004);
- Classification of measures based on the development a multi-criteria system (Lenca *et al.*, 2004, 2008);
- Classification of measures based on three criteria: the subject, the scope and the nature of the measures (Blanchard *et al.*, 2009);
- Classification of 62 measures based on a definition of 12 properties and an assigniment values for first nine properties that supported users for selecting appropriate measures in their specific application (Maddouri and Gammoudi, 2007);
- Classification for 40 objective interestingness measures by examining the value of 6 properties: independence, balance, symmetry, variation, description and statistic (Huynh *et al.*, 2007, 2011);
- Sixty two measures are classified into seven classes by basing on 19 properties and applying classification techniques (Guillaume *et al.*, 2012);
- Classification of 35 measures by the examining of behavior (Huynh *et al.*, 2005, 2006, 2007);
- Classification of 61 measures by studying of over 110 data sets (Tew *et al.*, 2013);
- Examining of variable values of the individual measures by taking the partial derivative of the calculated function of measures (Gras and Kuntz, 2008).

The determination of the variability of the interestingness measures is one of the most important criteria in assessing the objective interestingness measures (Huynh *et al.*, 2011). Based on examining the variation of the value by taking the partial derivative of the calculated function of Implication index (Gras and Kuntz, 2008), a new method is proposed for the classification of objective interestingness measures. The proposed method has ability to determine the increasing, the decreasing or the independence of the interestingness measures for each input parameter. Accordingly, the measures are

classified into different classes, i.e., increasing measures, decreasing measures, and independent measures, for each input parameter of the measures calculated function. Results of this classification will help users select the appropriate measures for their application when they know the input parameters of the actual data set.

This paper is organized into five sections. Section 1 introduces the general objective interestingness measures and measures classification methods. Section 2 reviews three methods for measures classification: classification based on a study of the properties of measures; classification based on a examining of the behavior of measures; and classification based on an examining of the variable values of individual measures. Section 3 describes the details of the method for the behavior evaluation of the measures according to the tendency of value variation. Section 4 conducts a classification of the interestingness measures which satisfy the asymmetric with additional reviews and comments based on partial derivatives formula. The final section summarizes the important results achieved in this paper.

2 The classification methods of interestingness measures

Recently, the classification usually focuses on three main approaches, i.e., classification based on examining of the properties of the measures, classification based on examining of the behavior of the measures, and classification based on a examining of the variable values of individual measures.

2.1 Classification based on examining of measures properties

The classification method based on theoretical attributes is first used popularly to classify the measures. In the general level of classification, interestingness measures are divided into two separate classes, such as subjective interestingness measures and objective interestingness measures. The subjective interestingness measures evaluate knowledge models based on the target, the knowledge and the belief of users. The objective measures valuate knowledge models based on the structure of the statistical data (Silberschatz and Tuzhilin, 1996). In the detailed level of classification, there are three key principles for a good measures: (1) the law $a \rightarrow b$ is equal to 0 if a and b independent; (2) the monotonically increasing with $a \cap b$; (3) the monotonically decreasing with $a \; or \; b$ is set to a criterion to evaluate the interestingness measures (Piatetsky-Shapiro, 1991). In the experimental level of classification, several researchers have been proposed the number of measures by using different criteria and classes to categorize and to evaluate measures. For example, using nine criteria for evaluation, i.e, conciseness, coverage, reliability, peculiarity, diversity, novelty, surprisingness, utility, and actionability, are examined in order to classify 38 objective measures (Geng and Hamilton, 2006). Building a system of multi-criteria decision to classify 20 measures based on the theoretical properties and the experimental results on 10 datasets (Lenca *et al.*, 2004, 2008). Studying theoretical properties, such as independence, balance, symmetry, variability, description and statistic of the measures is used to divide 40 objective interestingness measures into five classes: balance, independence, description, statistic and others (Huynh *et al.*, 2007, 2011). Expanding with 19 properties is used to divide 61 measures into 7 separate classes (Guillaume *et al.*, 2012).

2.2 Classification based on the behavior of interestingness measures

The classification method based on a study of the behaviors of the measures is based on the experimental investigation of the interestingness measures on specific data sets. From the experimental results, the behaviors of the measures can be predicted for each specific application area for the classification of the measures. By doing so, users can select the appropriate interestingness measures for their specific application. Method of assessing the interestingness measures based on behavioral examination was first proposed to investigate the behavior of 35 measures based on calculating of the distance between interestingness measures by using two clustering methods, Agglomerative Hierarchical Clustering (AHC) and Partitioning Around Medoids (PAM). As a result, measures are classified into 16 separate clusters to help users choose measures that can discover the best knowledge (Huynh *et al.*, 2005). The behavior of the measures was studied by correlation graph method. The approach was conducted on two prototypical datasets with opposite characteristics: a strongly correlated one (mushroom dataset) and a lowly correlated one (synthetic dataset). The results showed that the correlation between the interestingness measures depended on the type of data and the ranking rules to partition 40 measures into 6 stable clusters (Huynh *et al.*, 2006). In their experimental results, the study examined the behavior of 61 objective interestingness measures are conducted on 110 different datasets. The findings of the study classified 61 measures into 21 clusters and confirmed that the selection of the measures depends on the dataset. However, this research has not given the criteria of how to select a good dataset for inclusion in their experiments (Tew *et al.*, 2013).

2.3 Classification based on value variation of each individual measures

The examining of value variation of each individual interestingness measures allows us to know the increasing or the decreasing tendency of them according to the parameters of the calculated function (Gras and Kuntz, 2008). This method is carried out by taking the partial derivative of the calculated function of measures for each parameter. From the obtained result of taking the partial derivative, the behavior of each measures can be considered. Specifically, based on the partial derivative formula of each measures, the relationship of the measures with each parameter and the variable correlation between parameters of the calculated function of measures both can be considered.

3 Evaluation of the behavior of objective interestingness measures according to variable value trends

3.1 Objective interestingness measures

Let an association rule $a \rightarrow b$ where a and b are two disjoint sets of items (called item set). Item set a (resp. b) is associated with a subset of transactions $A = T(a)$ (resp. $B = T(b)$) with $T(a) = \{t \in T, a \subseteq t\}$, and T is the set of all transactions. The rule can be described by four cardinalities $(n, n_A, n_B, n_{A\bar{B}})$ where: $n = |T|, n_A = |A|, n_B = |B|, n_{A\bar{B}} = |A \cap \bar{B}|$. The cardinal $n_{A\bar{B}}$ corresponds to the effective number of negative examples of the rule (Figure 1).

The interestingness value of an association rule based on an objective interestingness measures will then be calculated by using the cardinality of a rule $m(a \to b) = f(n, n_A, n_B, n_{A\bar{B}}) \in \mathrm{R}$. To calculate easily, the following equivalent transformations should be used: $\mathrm{n_{AB}} = \mathrm{n_A} - \mathrm{n_{A\bar{B}}}$, $\mathrm{n_{\bar{A}}} = \mathrm{n} - \mathrm{n_A}$, $\mathrm{n_{\bar{B}}} = \mathrm{n} - \mathrm{n_B}$, $\mathrm{n_{\bar{A}B}} = \mathrm{n_B} - \mathrm{n_A} + \mathrm{n_{A\bar{B}}}$, $\mathrm{n_{\bar{A}\bar{B}}} = \mathrm{n} - \mathrm{n_B} - \mathrm{n_{A\bar{B}}}$.

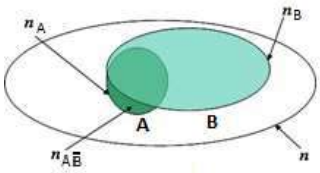


Figure 1.The cardinality of an association rule a ⟶ b

For example: two given sets A and B where A has 1 element and B has 3 elements. An association rule is in the form a ⟶ b. A = {Bread}, B = {Milk, Diappers, Beer} where $n = 100$, $n_A = 50$, $n_B = 80$ and $n_{A\bar{B}} = 10$.

The objective interestingness measures, Support Expectation, is identified by the formula:

$$m(a \to b) = f(n, n_A, n_B, n_{A\bar{B}}) = \frac{n_A(n_B - n_A + n_{A\bar{B}})}{n(n - n_A)}$$

Therefore, the interestingness value is:

$$m(a \to b) = \frac{50(80 - 50 + 10)}{100(100 - 50)} = 0.4$$

3.2 Statistical implicative analysis

The theory of statistical implicative analysis was developed to evaluate the learning behavior of pupils in the process of teaching algebra and geometry (Gras, 1979). This theory affects many areas such as pedagogy, psychology, and computer science. It has developed a unifying methodology and created synergy of scientific disciplines such as mathematics, statistics, psychology, education and data mining. It has also created a framework allowing for the evaluation of the strength of implications, formed through the acquisition of technical knowledge by natural or artificial processes. Especially in data mining, the extraction of knowledge from a large set of association rules is not conducted by humans to serve decision-making processes. Therefore, the development of interestingness measures plays an important role in the evaluation of association rules, and this is a clear success of statistical implicative analysis theory.

3.3 Examining the behavior of measures according to variable value trends

Stability analysis of implication index is to observe small variations of this measures in the surrounding space of parameters n, n_A, n_B and $n_{A\overline{B}}$ (Gras and Kuntz, 2008). To do this, taking the partial derivative for each parameter of the implication index formula is carried out. This method shows some abilities in supporting the research of the interestingness measures, i.e., the examining of increasing or decreasing variability of the measures, the relationship between the dependent variable parameters, n, n_A, n_B and $n_{A\overline{B}}$. However, the method also showed a limited role of parameters, n, n_A, n_B and $n_{A\overline{B}}$ to denote the increasing or decreasing variability of the measures.

Example 1. Let us examine the Added value measures (Sahar, 2003) with the following formula: $f = \frac{n_A - n_{A\bar{B}}}{n_A} - \frac{n_B}{n}$

We have the partial derivative formula for each parameters, $n, n_A, n_B, n_{A\bar{B}}$, of the function f as follows:

$$\frac{\partial f}{\partial n} = \frac{n_B}{n^2}\ ;\ \ \frac{\partial f}{\partial n_A} = \frac{n_{A\bar{B}}}{n_A^2}\ ;\ \ \frac{\partial f}{\partial n_B} = -\frac{1}{n}\ ;\ \ \frac{\partial f}{\partial n_{A\bar{B}}} = -\frac{1}{n_A}$$

where n = 100, $n_A = 20$, $n_B = 40$, $n_{A\bar{B}} = 4$, we calculate the relative values as follows:

$$f = 0.4\ ;\ \ \frac{\partial f}{\partial n} = 0.004\ ;\ \ \frac{\partial f}{\partial n_A} = 0.01\ ;\ \ \frac{\partial f}{\partial n_B} = -0.01\ ;\ \ \frac{\partial f}{\partial n_{A\bar{B}}} = -0.05$$

Thus, the variation of the Added value measures depends on two parameters, n and n_A and independent of two parameters, n_B and $n_{A\bar{B}}$. The interestingness value will increase as the number of n and n_A increases, whereas the interestingness value decreases when the number of parameters, n_B and $n_{A\bar{B}}$, increases. However, the declining rate of interestingness value does not depend on the increased rate of these parameters.

Example 2. Let us examine the Conviction measures (Brin et al., 1997) with the following formula: $f = \frac{n_A(n - n_B)}{n n_{A\bar{B}}}$

We have the partial derivative formula for each parameters, $n, n_A, n_B, n_{A\bar{B}}$ of the function f as follows:

$$\frac{\partial f}{\partial n} = \frac{n_A n_B}{n_{A\bar{B}} n^2}\ ;\ \ \frac{\partial f}{\partial n_A} = \frac{(n - n_B)}{n n_{A\bar{B}}}\ ;\ \ \frac{\partial f}{\partial n_B} = -\frac{n_A}{n n_{A\bar{B}}}\ ;\ \ \frac{\partial f}{\partial n_{A\bar{B}}} = -\frac{n_A(n - n_B)}{n n_{A\bar{B}}^2}$$

where n = 100, $n_A = 20$, $n_B = 40$, $n_{A\bar{B}} = 4$, we calculate the relative values as follows:

$$f = 3\ ;\ \ \frac{\partial f}{\partial n} = 0.02\ ;\ \ \frac{\partial f}{\partial n_A} = 0.15\ ;\ \ \frac{\partial f}{\partial n_B} = -0.05\ ;\ \ \frac{\partial f}{\partial n_{A\bar{B}}} = -0.75$$

Different from Added value measures, the variability of the Conviction measures depends on two parameters, n and $n_{A\bar{B}}$, and independent of two parameters, n_A $and n_B$. However, interestingness value will increase when the value of n, n_A rises and the increased rate of the value does not depend on the increased rate of the parameter n_A. Conversely, the interestingness value will decrease when the value of parameters, n_B and $n_{A\bar{B}}$ increases and the decreased rate of the interestingness value does not depend on the increase rate of the parameter n_B.

Example 3. Let us examine the Coverage measures (Geng and Hamilton, 2006) with the following formula: $f = \frac{n_A}{n}$

We have the partial derivative formula for each of the four parameters, $n, n_A, n_B, n_{A\bar{B}}$ of the function f as follows:

$$\frac{\partial f}{\partial n} = -\frac{n_A}{n^2}\ ;\ \ \frac{\partial f}{\partial n_A} = \frac{1}{n}\ ;\ \ \frac{\partial f}{\partial n_B} = 0\ ;\ \ \frac{\partial f}{\partial n_{A\bar{B}}} = 0$$

where n = 100, $n_A = 20$, $n_B = 40$, $n_{A\bar{B}} = 4$, we calculate the relative values as follows:

$$f = 0.2\ ;\ \ \frac{\partial f}{\partial n} = -0.002\ ;\ \ \frac{\partial f}{\partial n_A} = 0.01\ ;\ \ \frac{\partial f}{\partial n_B} = 0\ ;\ \ \frac{\partial f}{\partial n_{A\bar{B}}} = 0$$

Unlike these two above mentioned measures, i.e., Conviction measures and Added measures, the variation of the Coverage measures depends only on the parameter n and has a tendency of decreasing when the parameter n increases. In contrast, its variation has tendency of increasing and is independent of the parameter n_A. Two parameters, $n_B \; and n_{A\overline{B}}$, are not constributed completely on the demonstration of the variation in this measures. This is a good example to illustrate the limitation of the parameters in indicating the variation of the measures as fore-mentioned.

4 Classification based on examining the tendency of value variation

Based on a synthesic list of objective interestingness measures, which is applied to evaluate the quality of the knowledge as the form of the association rule $a \rightarrow b$ (Agrawal and Srikant, 1994), we have found that the group of measures satisfied the symmetry property ($m(a \rightarrow b) = m(b \rightarrow a)$) and the group of measures satisfied the asymmetry property ($m(a \rightarrow b) \neq m(b \rightarrow a)$), they offered an almost equal proportion of quantities. According to the five criteria are used to evaluate objective interestingness measures based on the distribution of the probability with the size of 2×2, a good measures must satisfy the symmetry property of the permutation of two variables (Tan *et al.*, 2002). It means that two association rules, $a \rightarrow b$ and $b \rightarrow a$, must have the same interestingness value. However, this is not true for many applications. The Confidence measures is an example. It is evaluated as a good measures by many researchers and is applied in many practical applications but it is an asymmetric measures. Therefore, in this paper, we focus on the study of the group of objective interestingness measures satisfying the asymmetric property. We calculate the partial derivative of the measures function according to 4 parameters $\boldsymbol{f}(n, n_A, n_B, n_{A\overline{B}})$ of this group with the purpose of analyzing the value variability of each measures according to the distribution of the probability with the size of 2×2. Accordingly, the group of asymmetric measures is divided into four different classes for each parameter. This classification helps researchers and users see the relationship between the value of each parameter and the value variation of the asymmetric measures from which they can select suitable measures for their study and application. The formula for calculating the asymmetric measures and the partial derivatives of measures based on 4 parameters are presented in the appendices.

In the next step, an examination of the derivative formula of each measures is conducted to classify them into groups with tendency of to increasing, decreasing or to determine that they are independent of parameters n, n_A, n_B, $n_{A\overline{B}}$.

Table 1: Findings of the variation of the measures based on partial derivatives under 4 parameters (1: increasing Variability; -1: decreasing Variability; 0: independent).

No	Name of interestingness measures	n	n_A	n_B	$n_{A\overline{B}}$
1	1-way Support	1	1	-1	-1
2	Added value, Pavillon, Centred Confidence, Dependency	1	1	-1	-1
3	Bayes factor, Odd multiplier	1	1	-1	-1
4	Causal-Confidence	1	1	-1	-1
5	Causal-Confirmed confidence	1	1	-1	-1
6	Loevinger, Certainty Factor, Satisfaction	1	1	-1	-1
7	Relative Risk, Class correlation ratio	1	1, 0, -1	-1	-1
8	Collective strength	1	1, 0, -1	-1	-1
9	Confidence	0	1	0	-1
10	Causal Confirm	1	1	-1	-1
11	Conviction	1	1	-1	-1
12	Coverage	-1	1	0	0
13	Descriptive Confirmed-Confidence, Ganascia Index	0	1	0	-1
14	Descriptive-Confirm	-1	1	0	-1
15	Entropic Implication Intensity 1	1	1	-1	-1
16	Entropic Implication Intensity 2	1	1	-1	-1
17	Examples and counter-examples rate	0	1	0	-1
18	Gain, Fukuda	-1	1	0	-1
19	Gini index	1	1	-1	-1, 0, 1
20	Goodman–Kruskal	1	1	-1	-1, 0, 1
21	Implication index	-1	-1	1	1
22	Implication Intensity (II)	1	1	-1	-1
23	Probabilistic measures of deviation from equilibrium (IPEE), Indice Probabiliste d'Ecart d'Equilibre	0	-1	0	0
24	Directed Information ratio (DIR)	0, 1, -1	1, 0, -1	0, 1, -1	-1, 0, -1
25	MGK, Ion	-1,1	1	-1,1	-1
26	J-measures	1	1	-1	-1, 0, 1
27	Klosgen	1	1	-1	-1
28	K-measures	1	1	-1	-1
29	Kulczynski index	0	1	-1	-1
30	Laplace	0	1	0	-1
31	Least contradiction	0	1	-1	-1
32	Leverage, Leverage 1	1	-1	-1	-1
33	Mutual Information MI, 2-way Support Variation	1	1	-1	-1, 0, 1
34	Prevalence	-1	0	1	0
35	Putative Causal Dependency	-1	1	-1	-1
36	Recall, Completeness	0	1	-1	-1
37	Sebag and Schoenauer	0	1	0	-1
38	Specificity 1, Negative Reliability	1	1	-1	-1
39	Zhang Zhang	1	1	-1	-1

Table 1 shows that most of the measures are variable to the tendency of increasing or decreasing according to four parameters n, n_A, n_B, $n_{A\overline{B}}$. However, the majority of the measures are of increasing variability with the parameter n_A. This reflects the general rule of the objective interestingness measures: the more examples do appear, the more reliability do increase. On the other hand, the decreasingly variable measures with the parameter $n_{A\overline{B}}$ are also prominent in Table 1. This also follows the general rule of interestingness measures. If the number of counter-examples increases, then the value of the interestingness measures decreases.

In the list of the analyzed measures, we have found that some measures having the variation is quite interesting. They do not follow the general variation rule of the measures is always increasing, decreasing and independent under parameters $n, n_A, n_B, n_{A\overline{B}}$. For example, Relative risk and Collective strength are two measures whose variation depends on the value of the parameter n_A. When examining these two measures, we see that the values of their partial derivative with respect to n_A is as follows: initial positive value, then proceed to 0, finally reaching negative values. This means that when the parameter n_A increases, the variability of those measures increases, decreases, and reaches extreme values. When we examine the Gini index, Goodman-Kruskal, J-measures, and Mutual information under the parameter $n_{A\overline{B}}$, we also see cases similar to Relative risk and Collective strength variation depend on the value of parameter n_A. In particular, Directed Information Ratio (DIR) and MGK measures have variability that depend on specific condition of $\frac{n_A - n_{A\overline{B}}}{n_A}$, $\frac{n - n_{A\overline{B}}}{n_A}$ and $\frac{n_B}{n}$.

Table 2: Classification of measures based on partial derivative under the parameter *n*

Decreases with n	Independent with n	Increases with n	Others
Coverage Descriptive-Confirm Gain, Fukuda Implication index Prevalence Putative Causal Dependency	Confidence Descriptive Confirmed-Confidence , Ganascia Index Examples and counter-examples rate Probabilistic measures of deviation from equilibrium (IPEE), Indice Probabiliste d'Ecart d'Equilibre Kulczynski index Laplace Least contradiction Recall, Completeness Sebag and Schoenauer	1-way Support Added value, Pavillon, Centred Confidence, Dependency Bayes factor, Odd multiplier Causal-Confidence Causal-Confirmed confidence Loevinger, Certainty Factor, Satisfaction Relative Risk , Class correlation ratio Collective strength Causal Confirm Conviction Entropic Implication Intensity 1 Entropic Implication Intensity 2 Implication Intensity Gini index Goodman–Kruskal J-measures Klosgen K-measures Leverage, Leverage 1	Directed Information ratio (DIR) MGK, Ion

		Mutual Information MI, 2-way Support Variation Specificity 1, Negative Reliability Zhang Zhang	

Table 2 shows that the number of measures are variable increasingly on the parameter n that gets over 50 percents of the studied measures. This reflects that the interestingnes value of measures depend on the size of the data used to examine them. It means that their rate of the variation depends on the rate of change of the parameter n. In contrast, the number of measures are variable decreasingly on the parameter n quite small. these measures are Coverage, Descriptive-Confirm, Gain, Fukuda, Implication index, Prevalence, and Putative Causal Dependency. Group of the independent measures on the parameter n is an especial class because most of them agrees with the descriptive property. If the measures satisfies the descriptive property, its interestingness value will not depend on the value of the parameter n (the size of the data). This result shows that the method of the study of the variation trend of measures based on partial derivative achieved results as accurate as other methods. DIR measures and MGK measures have the variation depending on the specific values of the parameters n.

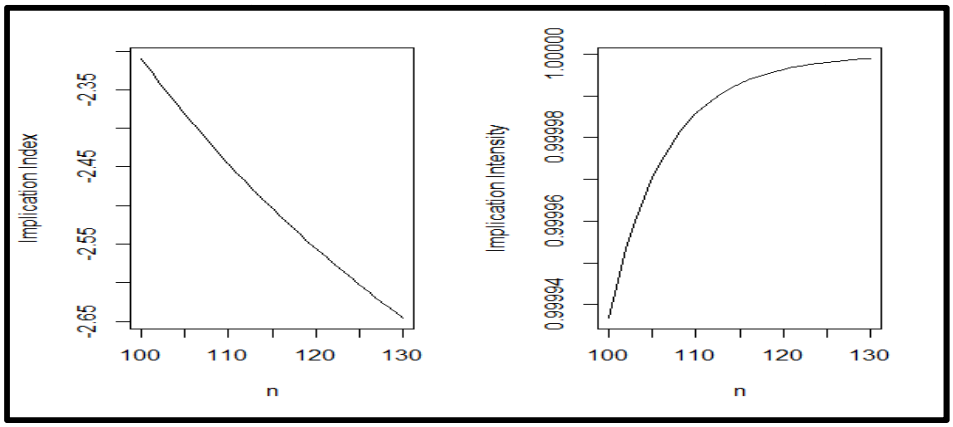


Figure 2: Compares the variation between Implication index and Implication intensity under the parameter n

Figure 2 shows the decreasing variability of Implication index and the increasing variability of Implication intensity under the parameter *n* on ARQAT tool implemented in R (Nghia and Hiep, 2014). They are two measures representing for two measures classes which varies on the parameter n (Implication index for increasing class and Implication intensity for decreasing class).

Table 3: Classification of measures based on partial derivative under the parameter n_A

Decreases with n_A	**Independent with n_A**	**Increases with n_A**	**Others**
Implication index Probabilistic measures of deviation from equilibrium (IPEE), Indice Probabiliste d'Ecart d'Equilibre Leverage, Leverage 1	Prevalence	1-way Support Added value, Pavillon, Centred Confidence, Dependency Bayes factor, Odd multiplier Causal-Confidence Causal-Confirmed confidence Loevinger, Certainty Factor, Satisfaction Confidence Causal Confirm Conviction Coverage Descriptive Confirmed-Confidence, Ganascia Index Descriptive-Confirm Entropic Implication Intensity 1 Entropic Implication Intensity 2 Examples and counter-examples rate Gain, Fukuda Gini index Goodman–Kruskal Implication Intensity MGK, Ion J-measures Klosgen K-measures Kulczynski index Laplace Least contradiction Mutual Information MI, 2-way Support Variation Putative Causal Dependency Recall, Completeness Sebag and Schoenauer Specificity 1, Negative Reliability Zhang Zhang	Relative Risk , Class correlation ratio Collective strength Directed Information ratio (DIR)

The results of Table 3 show that the class of objective interestingness measures varies in the increasing tendency under parameter n_A with relatively high proportion (31/39) and Prevalence measures is the only independent measurement parameter n_A. This result shows that the interestingness measures of the association rule $a \rightarrow b$ depend on the number of elements of the set A (n_A). As the parameter n_A is greater, the interestingness measures of the association rules reach the maximum value under the variability of the derivative. Particularly, the group of measures that is derived from Confidence measures belongs to the class of the increasing measures by the parameter n_A. This is consistent with the principle for determining the reliability of the association rule $a \rightarrow b$. The class of decreasing measures under the parameter n_A shares a small percentage (3 over all measures in our examining). They include measures such as Implication index, IPEE, Leverage. The group consisting of Relative risk, Class correlation ratio, Collective strength, Directed information ratio have variation value depending on the specific value of the parameter n_A.

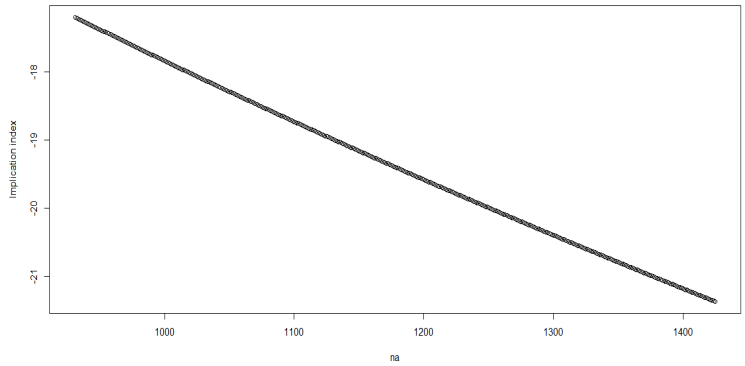


Figure 3: Show the decreasing variability of Implication index under the parameter *1521, ={931, 1425}, =1025, =4*

Implication index measures is the decreasing variability on the parameter . The decreasing variation of this measures is shown in Figure 3. This is a measures representing the class of the decreased measures by the parameter .

Table 4: Classification of measures based on partial derivative under the parameter

Decreases with	**Independent with**	**Increases with**	**Others**
1-way Support Added value, Pavillon, Centred Confidence, Dependency Bayes factor, Odd multiplier Causal-Confidence Causal-Confirmed confidence Loevinger, Certainty Factor, Satisfaction Relative Risk , Class correlation ratio Collective strength Causal Confirm Conviction Entropic Implication Intensity 1 Entropic Implication Intensity 2 Gini index Goodman–Kruskal Implication Intensity J-measures Klosgen K-measures Kulczynski index Least contradiction Leverage, Leverage 1 Mutual Information MI, 2-way Support Variation Putative Causal Dependency Recall, Completeness Specificity 1, Negative Reliability Zhang Zhang	Confidence Coverage Descriptive Confirmed-Confidence , Ganascia Index Descriptive-Confirm Examples and counter-examples rate Gain, Fukuda Probabilistic measures of deviation from equilibrium (IPEE), Indice Probabiliste d'Ecart d'Equilibre Laplace Sebag and Schoenauer	Implication index Prevalence	Directed Information ratio (DIR) MGK, Ion

Table 4 shows that most of the examining measures in groups of variable inverse measures with the parameter . According to this table, it can be confirmed that if the value of increases, the interestingness value will decrease. This is entirely consistent with the principles of the theory to determine the interestingness value of the association

rules because the studied measures are the asymmetry measures. Similarly, when we consider the measures with the parameters n, most of the independent variable measures with the parameter are the descriptive measures as Confidence described, Coverage, Laplace, Sebag and Schoenauer. The class of measures with variable covariates has only two measures: Implication index and Prevalence. It means that the interestingness value of those measures will increase when the value of increases. Like the examining results with parameters n, Directed Information Ratio, MGK have variation value depending on the specific value of the parameter .

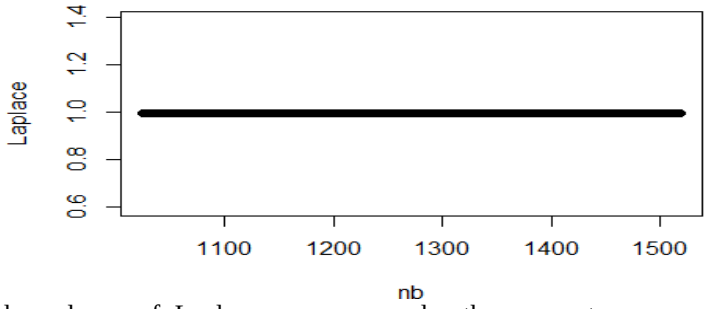


Figure 4: Show independence of Laplace measures under the parameter

Figure 4 shows the independence of Laplace measures under parameter on ARQAT tool implemented in R (Nghia and Hiep, 2014). This is a measures of the class of independent measures of the parameter .

Table 5: Classification of measures based on partial derivative under parameter –

Decreases with –	**Independent with –**	**Increases with –**	**Others**
1-way Support Added value, Pavillon, Centred Confidence, Dependency Bayes factor, Odd multiplier Causal-Confidence Causal-Confirmed confidence Loevinger, Certainty Factor, Satisfaction Relative Risk , Class correlation ratio Collective strength Confidence Causal Confirm Conviction Descriptive Confirmed-Confidence , Ganascia Index Descriptive-Confirm Entropic Implication Intensity 1 Entropic Implication Intensity 2 Examples and counter-examples rate Gain, Fukuda Implication Intensity MGK, Ion Klosgen K-measures Kulczynski index Laplace Least contradiction Leverage, Leverage 1	Coverage Probabilistic measures of deviation from equilibrium (IPEE), Indice Probabiliste d'Ecart d'Equilibre Prevalence	Implication index	Directed Information ratio (DIR) Gini index Goodman–Kruskal J-measures Mutual Information MI, 2-way Support Variation

Putative Causal Dependency Recall, Completeness Sebag and Schoenauer Specificity 1, Negative Reliability Zhang Zhang			

Table 5 shows that the class of objective interestingness measures in the decreasing variation with the parameter accounts for 71%. This reflects accurately the role of the parameter when determining the interestingness measures of the association rule . If the number of counter-examples increases, the interestingness value of the association rule decreases. In this class, most of measures are derived from Confidence measures. This result reflects exactly the formula of determining the reliability of an association rule as defined in the formula of Confidence measures (———). Therefore, the parameter is always inversely related to the reliability. The class of independent measures on the parameter has relatively small proportion of total of studied measures. It includes the three measures Coverage, IPEE, and Prevalence. The group of measures having increasing variation on the parameter is only one measures (Implication index). This is a paradoxical situation in determining the interestingness measures of the association rules because as the number of counter-examples increases interestingness value should decrease, but in this situation it has increased. The last subclass includes five measures, in which the variability of Directed Information ratio measures depends on the constraint of two expressions: ——— and —. The remaining measures have variation value depending on the specific value of the parameter .

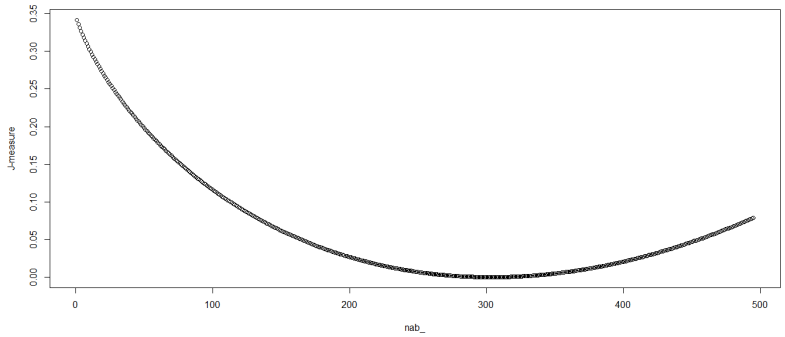


Figure 5: Show variable value depending on the specific value of the parameter of J-measures

The variation value of measures J-measures depends on the specific value of the parameter . The variation of this measures is shown in Figure 5. This measures represents the class of measures whose variation value depending on the specific value of the parameter .

5 Conclusion

Classification of the interestingness measures has attracted many researchers in the field of Knowledge Discovery in Databases. The study focuses on two main classification methods: one based on the properties of the measures and the other based

on the behaviors of the measures. In this paper, we examined the value variation of the asymmetric objective interestingness measures by taking their partial derivatives using the four parameters, $n, n_A, n_B, n_{A\overline{B}}$. The results of classification are considered in the increasing, decreasing or independent variation of the derivative formula of each measures, thus evaluating the variable tendency of the asymmetric objective interestingness measures. This classification helps researchers and users to obtain more insight into the group of asymmetric objective interestingness measures such as the increasing, decreasing variability of each measures, the relationship between the value variability and the statistical parameter values $n, n_A, n_B, n_{A\overline{B}}$, and the interdependence between these parameters in the formula calculating the interestingness value of the measures. This information is the basis for defining an appropriate measures in their specific application.

References

[1] A. Silberschatz and A. Tuzhilin (1995), On subjective measures of interestingness in knowledge discovery, *KDD'95 - Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, 275-281.

[2] A. Silberschatz and A. Tuzhilin (1996), What makes patterns interesting in knowledge discovery systems, *IEEE Transactions on Knowledge and Data Engineering 8(6)*, 970-974.

[3] C. Tew *et al.* (2013), Behavior-based clustering and analysis of interestingness measures for association rule mining, *Journal of Data Mining and Knowledge Discovery 28*, Springer-Verlag, 1004-1045.

[4] David H. Glass (2013), Confirmation measures of association rule interestingness, *Knowledge-Based Systems 44*, 65–77.

[5] F. Guillet and H. J. Hamilton (2007), Quality Measures in Data Mining - Series in Computational Intelligence 43, Springer-Verlag.

[6] G. Piatetsky-Shapiro (1991), Discovery, analysis, and presentation of strong rules, *Knowledge Discovery in Databases*, 229-248.

[7] G. Piatetsky-Shapiro and C. J. Matheus (1994), The interestingness of deviations, *AAAI'94 - Knowledge Discovery in Databases Workshop*, 25-36.

[8] H. X. Huynh *et al.* (2006), Evaluating Interestingness Measures with Linear Correlation Graph, *Advances in Applied Artificial Intelligence (LNCS 4031)*, Springer-Verlag, 312 – 321.

[9] H. X. Huynh *et al.* (2005), Data Analysis Approach for Evaluating the Behavior of Interestingness Measures, *Discovery Science (LNCS 3735)*, Springer-Verlag, 130-137.

[10] H. X. Huynh *et al.* (2006), Extracting representative measures for thepost-processing of association rules, *The 2006 IEEE International Conference on Research, Innovation and Vision for the Future (RIVF'08)*, 100-106.

[11] H. X. Huynh *et al.* (2007), A graph-based clustering approach to evaluate interestingness measures: a tool and a comparative study (Chapter 2), *Quality Measures in Data Mining*, Springer-Verlag, 25-50.

[12] H. X. Huynh *et al.* (2012), Classification of objective interestingness measures, *Journal of Can Tho University (2011:20a)*, 147 - 158.

[13] J. Blanchard *et al.* (2009), Semantics-based classification of rule interestingness measures, *Post-Mining of Association Rules: techniques for effective knowledge extraction" (Y. Zhao, C. Zhang, L. Cao editors), IGI Global*, 56-79.

[14] J. Liu et al. (2008), A New Interestingness Measures of Association Rules, The Second International Conference on Genetic and Evolutionary Computing (WGEC'08), 393 – 397.

[15] Jon Hills *et al.* (2012), Interestingness Measures for Fixed Consequent Rules, *Intelligent Data Engineering and Automated Learning - IDEAL 2012 (LNCS Volume 7435)*, Springer-Verlag, 68–75.

[16] K. Selvarangam and K. Ramesh Kumar (2014), Selecting Perfect Interestingness Measures By Coefficient Of Variation Based Ranking Algorithm, *Journal of Computer Science (Volume 10, Issue 9)*, 1672-1679.

[17] L. Geng and H. J. Hamilton (2006), Interestingness measures for data mining: A survey, *ACM Computing Surveys (Volume 38)*, 1-32.

[18] M. Jalali-Heravi and O. Zaïane (2010), A study on interestingness measures for associative classifiers, *SAC'10 - Proceedings of the 25th ACM Symposium on Applied Computing*, 1039 -1046.

[19] M. Martínez-Ballesteros *et al.* (2014), Selecting the best measures to discover quantitative association rule, *Neurocomputing (Volume 126)*, 3–14.

[20] Maddouri and Gammoudi (2007), On Semantic Properties of Interestingness Measures for Extracting Rules from Data, *ICANNGA (1) 2007, Springer-Verlag Berlin Heidelberg*, 148-158

[21] P. Lenca *et al.* (2004), A multicriteria decision aid for interestingness measures selection, LUSSI-TR-2004-01-EN, 1–27.

[22] P. Lenca *et al.* (2008), On selecting interestingness measures for association rules: user oriented description and multiple criteria decision aid, *European Journal of Operational Research (Volume 184, Issue 2)*, 610–626.

[23] P. N. Tan *et al.* (2002), Selecting the Right Interestingness Measures for Association Patterns, *SIGKDD '02 Edmonton, Alberta, Canada ACM 1-58113-567-X/02/0007*, 1-10.

[24] P. N. Tan *et al.* (2004), Selecting the right objective measures for association analysis, *Journal of Information Systems (Volume 29, Issue 4)*, 293-313.

[25] R. Agrawal and R. Srikant (1994), Fast algorithms for mining association rules, *VLDB'94 - Proceedings of the 20th International Conference on Very Large Data Bases*, 487-499.

[26] R. Gras (1979), Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs en didactique des mathématiques, *PhD thesis, Université de Rennes 1*.

[27] R. Gras and P. Kuntz (2008), An overview of the Statistical Implicative Analysis (SIA) development, *Statistical Implicative Analysis - Studies in Computational Intelligence (Volume 127)*, Springer-Verlag, 11-40.

[28] S. Brin *et al.* (1997), Dynamic itemset counting and implication rules for market basket data, *Proceedings of the ACM SIGMOD international conference on management of data*, 255–264.

[29] S. Guillaume et al. (2012), Categorization of interestingness measures for knowledge extraction, journals/corr/abs-1206-6741, 1 – 34.

[30] S. Sahar (2003), What is interesting: studies on interestingness in knowledge discovery. *PhD thesis, School of Computer Science, Tel-Aviv University*.

Appendix 1

Formula of asymmetric objective interestingness measures

No	Name of interestingness measures	$f(n, n_A, n_B, n_{A\bar{B}})$
1.	1-way Support	$\frac{n_A - n_{A\bar{B}}}{n_A} log_2 \frac{n(n_A - n_{A\bar{B}})}{n_A n_B}$
2.	Added value, Pavillon, Centred Confidence, Dependency	$\frac{n_A - n_{A\bar{B}}}{n_A} - \frac{n_B}{n}$
3.	Bayes factor, Odd multiplier	$\frac{nn_A - n_A n_B - nn_{A\bar{B}} + n_B n_{A\bar{B}}}{n_B n_{A\bar{B}}}$
4.	Causal-Confidence	$1 - \frac{1}{2}\left(\frac{1}{n_A} + \frac{1}{n - n_B}\right) n_{A\bar{B}}$
5.	Causal-Confirmed confidence	$1 - \frac{1}{2}\left(\frac{3}{n_A} + \frac{1}{n - n_B}\right) n_{A\bar{B}}$
6.	Loevinger, Certainty Factor, Satisfaction	$1 - \frac{nn_{A\bar{B}}}{n_A(n - n_B)}$
7.	Relative Risk , Class correlation ratio	$\frac{(n_A - n_{A\bar{B}})(n - n_A)}{n_A(n_B - n_A + n_{A\bar{B}})}$
8.	Collective strength	$\frac{(n_A - n_{A\bar{B}})(n - n_B - n_{A\bar{B}})(n_A(n - n_B) + n_B(n - n_A))}{((n - n_A)(n - n_B) + n_A n_B)(n_B - n_A + 2n_{A\bar{B}})}$
9.	Confidence	$\frac{n_A - n_{A\bar{B}}}{n_A}$
10.	Causal Confirm	$\frac{n + n_A - n_B - 4n_{A\bar{B}}}{n}$
11.	Conviction	$\frac{n_A(n - n_B)}{nn_{A\bar{B}}}$
12.	Coverage	$\frac{n_A}{n}$
13.	Descriptive Confirmed-Confidence, Ganascia Index	$1 - \frac{2n_{A\bar{B}}}{n_A}$
14.	Descriptive-Confirm	$\frac{n_A - 2n_{A\bar{B}}}{n}$
15.	Entropic Implication Intensity 1	$\sqrt{II\left(\left(1 - H_{B\|A}^{\propto}\right)\left(1 - H_{\bar{A}\|\bar{B}}^{\propto}\right)\right)^{\frac{1}{2\propto}}}$ with (α=1) and $H_{A\|B} = \frac{n_A - n_{A\bar{B}}}{n_B} log_2 \frac{n_{A\bar{B}} - n_A}{n_B} + \frac{n_A - n_B - n_{A\bar{B}}}{n_B} log_2 \frac{n_B - n_A + n_{A\bar{B}}}{n_B}$ Where II is Inplication intensity
16.	Entropic Implication Intensity 2	$\sqrt{II\left(\left(1 - H_{B\|A}^{\propto}\right)\left(1 - H_{\bar{A}\|\bar{B}}^{\propto}\right)\right)^{\frac{1}{2\propto}}}$ with (α=2) and $H_{A\|B} = \frac{n_A - n_{A\bar{B}}}{n_B} log_2 \frac{n_{A\bar{B}} - n_A}{n_B} + \frac{n_A - n_B - n_{A\bar{B}}}{n_B} log_2 \frac{n_B - n_A + n_{A\bar{B}}}{n_B}$

17.	Examples and counter-examples rate (Exam-Cex-rate, Excounterex rate)	$\frac{n_A - 2n_{A\bar{B}}}{n_A - n_{A\bar{B}}}$
18.	Gain, Fukuda	$\frac{n_A(1-\theta) - n_{A\bar{B}}}{n}$
19.	Gini index	$\frac{(n_A - n_{A\bar{B}})^2 + {n_{A\bar{B}}}^2}{nn_A} + \frac{(n_B - n_A + n_{A\bar{B}})^2 + (n - n_B - n_{A\bar{B}})^2}{n(n-n_A)} - \frac{{n_B}^2 + (n-n_B)^2}{n^2}$
20.	Goodman–Kruskal	$\frac{\alpha}{\beta}$ where $\alpha = max\left(\frac{n_A - n_{A\bar{B}}}{n}, \frac{n_{A\bar{B}}}{n}\right) + max\left(\frac{n_B - n_A + n_{A\bar{B}}}{n}, \frac{n - n_B - n_{A\bar{B}}}{n}\right) + max\left(\frac{n_A - n_{A\bar{B}}}{n}, \frac{n_B - n_A + n_{A\bar{B}}}{n}\right) + max\left(\frac{n_{A\bar{B}}}{n}, \frac{n - n_B - n_{A\bar{B}}}{n}\right) - max\left(\frac{n_A}{n}, \frac{n - n_A}{n}\right) - max\left(\frac{n_B}{n}, \frac{n - n_B}{n}\right)$ $\beta = 2 - max\left(\frac{n_A}{n}, \frac{n - n_A}{n}\right) - max\left(\frac{n_B}{n}, \frac{n - n_B}{n}\right)$
21.	Implication index	$q(A,\bar{B}) = \frac{n_{A\bar{B}} - \frac{n_A(n-n_B)}{n}}{\sqrt{\frac{n_A(n-n_B)}{n}}}$
22.	Implication Intensity (II)	$\frac{1}{\sqrt{2\pi}} \int_{q(A,\bar{B})}^{+\infty} e^{-\frac{t^2}{2}} dt$ **Or** $1 - \sum_{k=max\,(0,n_A-n_B)}^{n_{A\bar{B}}} \frac{C_{n_B}^{n_A-k} C_{(n-n_B)}^{k}}{C_n^{n_A}}$
23.	Probabilistic measures of deviation from equilibrium (IPEE),Indice Probabiliste d'Ecart d'Equilibre	$1 - \frac{2}{2^{n_A}} \sum_{k=0}^{n_{A\bar{B}}} C_{n_A}^{k}$
24.	Directed Information ratio(DIR)	$\begin{cases} -\infty & if\ \frac{n_B}{n} = 1 \\ 0 & if\ \frac{n_B}{n} \le \frac{1}{2}\ and\ \frac{n_A - n_{A\bar{B}}}{n_A} \le \frac{1}{2} \\ 1 + \frac{n_A - n_{A\bar{B}}}{n_A} log_2 \frac{n_A - n_{A\bar{B}}}{n_A} + \frac{n_{A\bar{B}}}{n_A} log_2 \frac{n_{A\bar{B}}}{n_A} & if\ \frac{n_B}{n} \le \frac{1}{2}\ and\ \frac{n_A - n_{A\bar{B}}}{n_A} > \frac{1}{2} \\ 1 + \frac{1}{\frac{n_B}{n} log_2 \frac{n_B}{n} + \frac{n-n_B}{n} log_2 \frac{n-n_B}{n}} & if\ \frac{n_B}{n} > \frac{1}{2}\ and\ \frac{n_A - n_{A\bar{B}}}{n_A} \le \frac{1}{2} \\ 1 - \frac{\frac{n_A - n_{A\bar{B}}}{n_A} log_2 \frac{n_A - n_{A\bar{B}}}{n_A} + \frac{n_{A\bar{B}}}{n_A} log_2 \frac{n_{A\bar{B}}}{n_A}}{\frac{n_B}{n} log_2 \frac{n_B}{n} + \frac{n-n_B}{n} log_2 \frac{n-n_B}{n}} & if\ \frac{n_B}{n} > \frac{1}{2}\ and\ \frac{n_A - n_{A\bar{B}}}{n_A} > \frac{1}{2} \end{cases}$
25.	MGK, Ion	$\begin{cases} 1 - \frac{nn_{A\bar{B}}}{n_A(n-n_B)} & if\ \frac{n - n_{A\bar{B}}}{n_A} > \frac{n_B}{n} \\ \frac{n(n_A - n_{A\bar{B}})}{n_A n_B} - 1 & otherwise \end{cases}$
26.	J-measures	$\frac{n_A - n_{A\bar{B}}}{n} log_2 \frac{n(n_A - n_{A\bar{B}})}{n_A n_B} + \frac{n_{A\bar{B}}}{n} log_2 \frac{nn_{A\bar{B}}}{n_A(n-n_B)}$

27.	Klosgen	$\sqrt{\frac{n_A - n_{A\bar{B}}}{n}}\left(\frac{n - n_B}{n} - \frac{n_{A\bar{B}}}{n_A}\right)$
28.	K-measures	$\left(\frac{n_A - n_{A\bar{B}}}{n_A} - \frac{n - n_B - n_{A\bar{B}}}{n - n_A}\right)(log_2(n - n_B) - log_2 n_B)$
29.	Kulczynski index	$\frac{(n_A - n_{A\bar{B}})}{2}\left(\frac{1}{n_A} + \frac{1}{n_B}\right)$
30.	Laplace	$\frac{n_A - n_{A\bar{B}} + 1}{n_A + 2}$
31.	Least contradiction	$\frac{n_A - 2n_{A\bar{B}}}{n_B}$
32.	Leverage, Leverage 1	$1 - \frac{n_{A\bar{B}}}{n_A} - \frac{n_A n_B}{n^2}$
33.	Mutual Information **MI,** 2-way Support Variation	$\frac{n_A - n_{A\bar{B}}}{n} log_2 \frac{n(n - n_{A\bar{B}})}{n_A n_B} + \frac{n_{A\bar{B}}}{n} log_2 \frac{n n_{A\bar{B}}}{n_A(n - n_B)} + \frac{n_B - n_A + n_{A\bar{B}}}{n} log_2 \frac{n(n_B - n_A + n_{A\bar{B}})}{(n - n_A)n_B} + \frac{n - n_B - n_{A\bar{B}}}{n} log_2 \frac{n(n - n_B - n_{A\bar{B}})}{(n - n_A)(n - n_B)}$
34.	Prevalence	$\frac{n_B}{n}$
35.	Putative Causal Dependency	$\frac{3}{2} + \frac{4n_A - 3n_B}{2n} - \left(\frac{3}{2n_A} + \frac{2}{n - n_B}\right) n_{A\bar{B}}$
36.	Recall, Completeness	$\frac{n_A - n_{A\bar{B}}}{n_B}$
37.	Sebag and Schoenauer	$\frac{n_A}{n_{A\bar{B}}} - 1$
38.	Specificity 1, Negative Reliability	$\frac{n - n_B - n_{A\bar{B}}}{n - n_A}$
39.	Zhang Zhang	$\frac{n n_A - n_A n_B - n n_{A\bar{B}}}{max\left((n_A - n_{A\bar{B}})(n - n_B), n_B n_{A\bar{B}}\right)}$

Appendix 2

Formula of partial derivative under the parameter n

No	$\frac{\partial Z}{\partial n}$
1	$\frac{n_A - n_{A\bar{B}}}{n n_A ln2}$
2	$\frac{n_B}{n^2}$
3	$\frac{n_A - n_{A\bar{B}}}{n_B n_{A\bar{B}}}$
4	$\frac{n_{A\bar{B}}}{2(n - n_B)^2}$
5	$\frac{n_{A\bar{B}}}{2(n - n_B)^2}$
6	$\frac{n_B n_{A\bar{B}}}{n_A(n - n_B)^2}$

7	$\frac{(n_A - n_{A\bar{B}})}{n_A(n_B - n_A + n_{A\bar{B}})}$
8	$\frac{(n_A - n_{A\bar{B}})\left(n_A(n-n_B)+n_B\left((n-n_A)\right)+(n-\ n_B-n_{A\bar{B}})(n_A+\ n_B)\right)\left((n-n_A)(n-n_B)+n_An_B\right)(n_B-\ n_A+2n_{A\bar{B}})-k}{\left(\left((n-n_A)(n-n_B)+n_An_B\right)(n_B-\ n_A+2n_{A\bar{B}})\right)^2}$ $k=(n_A-n_{A\bar{B}})\left(n_A(n-n_B)+n_B\left((n-n_A)\right)+(n-\ n_B-n_{A\bar{B}})(n_A+\ n_B)\right)((n-n_B)+(n-n_B)(n_B-\ n_A+2n_{A\bar{B}})$
9	0
10	$\frac{-n_A+n_B+4n_{A\bar{B}}}{n^2}$
11	$\frac{n_An_B}{n_{A\bar{B}}n^2}$
12	$-\frac{n_A}{n^2}$
13	0
14	$-\frac{n_A-2n_{A\bar{B}}}{n^2}$
15	$\frac{II'_n(h)^{\frac{1}{2}}+II\frac{1}{2}(h)^{-\frac{1}{2}}\left(1-\left(-\frac{n_A-n_{A\bar{B}}}{n_A}log_2\frac{n_A-n_{A\bar{B}}}{n_A}-\frac{n_{A\bar{B}}}{n_A}log_2\frac{n_{A\bar{B}}}{n_A}\right)\left(1-\left(-\frac{1}{n_B}(log_2\frac{n-n_B-n_{A\bar{B}}}{n_B}-\frac{1}{ln2})\right)\right)\right)}{2\sqrt{II(h)^{\frac{1}{2}}}}$ $h=1-\left(-\frac{n_A-n_{A\bar{B}}}{n_A}log_2\frac{n_A-n_{A\bar{B}}}{n_A}-\frac{n_{A\bar{B}}}{n_A}log_2\frac{n_{A\bar{B}}}{n_A}\right)\left(1-\left(-\frac{n_{A\bar{B}}}{n_B}log_2\frac{n_{A\bar{B}}}{n_B}-\frac{n-n_B-n_{A\bar{B}}}{n_B}log_2\frac{n-n_B-n_{A\bar{B}}}{n_B}\right)\right)$
16	$\frac{II'_n(h)^{\frac{1}{4}}+II\frac{1}{4}(h)^{-\frac{3}{4}}\left(1-\left(-\frac{n_A-n_{A\bar{B}}}{n_A}log_2\frac{n_A-n_{A\bar{B}}}{n_A}-\frac{n_{A\bar{B}}}{n_A}log_2\frac{n_{A\bar{B}}}{n_A}\right)^2\right)k}{2\sqrt{II(h)^{\frac{1}{4}}}}$ $h=1-\left(-\frac{n_A-n_{A\bar{B}}}{n_A}log_2\frac{n_A-n_{A\bar{B}}}{n_A}-\frac{n_{A\bar{B}}}{n_A}log_2\frac{n_{A\bar{B}}}{n_A}\right)^2\left(1-\left(-\frac{n_{A\bar{B}}}{n_B}log_2\frac{n_{A\bar{B}}}{n_B}-\frac{n-n_B-n_{A\bar{B}}}{n_B}log_2\frac{n-n_B-n_{A\bar{B}}}{n_B}\right)^2\right)$ $k=\left(-2\left(-\frac{n_{A\bar{B}}}{n_B}log_2\frac{n_{A\bar{B}}}{n_B}-\frac{n-n_B-n_{A\bar{B}}}{n_B}log_2\frac{n-n_B-n_{A\bar{B}}}{n_B}\right)(-\frac{1}{n_B}(log_2\frac{n-n_B-n_{A\bar{B}}}{n_B}+\frac{1}{ln2})\right)$
17	0
18	$-\frac{n_A(1-\theta)-n_{A\bar{B}}}{n^2}$
19	$-\frac{(n_A-n_{A\bar{B}})^2+{n_{A\bar{B}}}^2}{n^2{n^2}_A}+\frac{(n_B-n_A+n_{A\bar{B}})^2(2n-n_A)+2(n-n_B-n_{A\bar{B}})-(n-n_B-n_{A\bar{B}})^2(2n-n_A)}{(n(n-n_A))^2}-\frac{2n{n_B}^2+2(n-n_A)-2n(n-n_B)^2}{n^4}$
20	$\frac{\alpha}{\beta}$ where

	$\alpha = \left(\frac{n_A - n_{A\bar{B}}}{n^2} + \frac{n - n_B - n_{A\bar{B}}}{n^2} + \frac{n_A - n_{A\bar{B}}}{n^2} + \frac{n - n_B - n_{A\bar{B}}}{n^2} - \frac{n - n_A}{n^2} - \frac{n - n_B}{n^2}\right)\left(2 - max\left(\frac{n_A}{n}, \frac{n - n_A}{n}\right) - max\left(\frac{n_B}{n}, \frac{n - n_B}{n}\right)\right) - \left(max\left(\frac{n_A - n_{A\bar{B}}}{n}, \frac{n_{A\bar{B}}}{n}\right) + max\left(\frac{n_B - n_A + n_{A\bar{B}}}{n}, \frac{n - n_B - n_{A\bar{B}}}{n}\right) + max\left(\frac{n_A - n_{A\bar{B}}}{n}, \frac{n_B - n_A + n_{A\bar{B}}}{n}\right) + max\left(\frac{n_{A\bar{B}}}{n}, \frac{n - n_B - n_{A\bar{B}}}{n}\right) - max\left(\frac{n_A}{n}, \frac{n - n_A}{n}\right) - max\left(\frac{n_B}{n}, \frac{n - n_B}{n}\right)\right)\left(-\frac{n - n_A}{n^2} - \frac{n - n_B}{n^2}\right)$ $\beta = (2 - max\left(\frac{n_A}{n}, \frac{n - n_A}{n}\right) - max\left(\frac{n_B}{n}, \frac{n - n_B}{n}\right))^2$
21	$\frac{1}{2\sqrt{n}}\left(n_{A\bar{B}} + \frac{n_A(n - n_B)}{n}\right)$
22	$-\sum_{k=max\,(1, n_A - n_B)}^{n_{A\bar{B}}} \frac{n_B!}{((n_A - k)!\,(n_B - n_A + k)!} \frac{(n - n_B - 1)!}{k!\,(n - n_B - k - 1)!} \frac{n_A!\,(n - n_A - 1)!}{(n - 1)!}$
23	0
24	$\begin{cases} 0 & if\left(\frac{n_B}{n} \le \frac{1}{2}\ and\ \frac{n_A - n_{A\bar{B}}}{n_A} > \frac{1}{2}\right) \\ \frac{\frac{1}{n^2}\left(log_2 \frac{n_B}{n} + \frac{n_B}{ln2}\right) + \left(\frac{n_B}{n^2}\left(log_2 \frac{n - n_B}{n}\right) + \frac{1}{ln2}\right)}{\left(\frac{n_B}{n} log_2 \frac{n_B}{n} + \frac{n - n_B}{n} log_2 \frac{n - n_B}{n}\right)^2} & if\left(\frac{n_B}{n} > \frac{1}{2}\ and\ \frac{n_A - n_{A\bar{B}}}{n_A} \le \frac{1}{2}\right) \\ -\frac{\left(\frac{n_A - n_{A\bar{B}}}{n_A} log_2 \frac{n_A - n_{A\bar{B}}}{n_A} + \frac{n_{A\bar{B}}}{n_A} log_2 \frac{n_{A\bar{B}}}{n_A}\right)\left(\frac{1}{n^2}\left(log_2 \frac{n_B}{n} + \frac{n_B}{ln2}\right) + \left(\frac{n_B}{n^2}\left(log_2 \frac{n - n_B}{n}\right) + \frac{1}{ln2}\right)\right)}{\left(\frac{n_B}{n} log_2 \frac{n_B}{n} + \frac{n - n_B}{n} log_2 \frac{n - n_B}{n}\right)^2} & if\left(\frac{n_B}{n} > \frac{1}{2}\ and\ \frac{n_A - n_{A\bar{B}}}{n_A} > \frac{1}{2}\right) \end{cases}$
25	$\begin{cases} -\frac{n_{A\bar{B}}(nn_A - n_A n_B) - nn_A n_{A\bar{B}}}{(nn_A - n_A n_B)^2} & if\ \frac{n - n_{A\bar{B}}}{n_A} > \frac{n_B}{n} \\ \frac{(n_A - n_{A\bar{B}})}{n_A n_B} & otherwise \end{cases}$
26	$\frac{n_A - n_{A\bar{B}}}{n^2}\left(-log_2 \frac{n(n_A - n_{A\bar{B}})}{n_A n_B} + \frac{1}{ln2}\right) + \frac{n_{A\bar{B}}}{n^2}\left(log_2 \frac{nn_{A\bar{B}}}{n_A(n - n_B)} - \frac{n_B}{(n - n_B)ln2}\right)$
27	$\left(\frac{\frac{n_A - n_{A\bar{B}}}{n^2}}{2\sqrt{\frac{n_A - n_{A\bar{B}}}{n}}}\left(\frac{n - n_B}{n} - \frac{n_{A\bar{B}}}{n_A}\right)\right) + \left(\sqrt{\frac{n_A - n_{A\bar{B}}}{n}}\right)\left(\frac{n_B}{n^2}\right)$
28	$\left(-\frac{(n - n_A) - (n - n_B - n_{A\bar{B}})}{(n - n_A)^2}\right)(log_2(n - n_B) - log_2 n_B) + \left(\frac{n_A - n_{A\bar{B}}}{n_A} - \frac{n - n_B - n_{A\bar{B}}}{n - n_A}\right)\left(\frac{1}{(n - n_B)ln2}\right)$
29	0
30	0
31	0
32	$\frac{2nn_A n_B}{n^4}$

33	$\frac{n_A - n_{A\bar{B}}}{n^2}(-log_2\frac{n(n-n_{A\bar{B}})}{n_A n_B}+\frac{(2n-n_{A\bar{B}})}{n(n-n_{A\bar{B}})ln2}) + \frac{n_{A\bar{B}}}{n^2}(-log_2\frac{nn_{A\bar{B}}}{n_A(n-n_B)}-\frac{n_B}{(n-n_B)ln2}) + \frac{n_B-n_A+n_{A\bar{B}}}{n^2}(-log_2\frac{n(n_B-n_A+n_{A\bar{B}})}{(n-n_A)n_B} + \frac{(n_B-n_A+n_{A\bar{B}})(n-n_A)-n(n_B-n_A+n_{A\bar{B}})}{(n_B-n_A+n_{A\bar{B}})ln2}) + \frac{n-n_B-n_{A\bar{B}}}{n^2}(log_2\frac{n(n-n_B-n_{A\bar{B}})}{(n-n_A)(n-n_B)} + \frac{(2n-n_B-n_{A\bar{B}})(n-n_A)(n-n_B)-n(n-n_B-n_{A\bar{B}})(2n-n_B-n_A)}{(n-n_A)(n-n_B)(n-n_B-n_{A\bar{B}})ln2}$
34	$-\frac{n_B}{n^2}$
35	$-\frac{4n_A-3n_B}{2n^2}+\frac{2n_{A\bar{B}}}{(n-n_B)^2}$
36	0
37	0
38	$\frac{-n_A+n_B+n_{A\bar{B}}}{(n-n_A)^2}$
39	$\frac{(n_A-n_{A\bar{B}})\,max\big((n_A-n_{A\bar{B}})(n-n_B),n_Bn_{A\bar{B}}\big)-(nn_A-n_An_B-nn_{A\bar{B}})(n_A-n_{A\bar{B}})}{(max\big((n_A-n_{A\bar{B}})(n-n_B),n_Bn_{A\bar{B}}\big))^2}$

Appendix 3

Formula of partial derivative under the parameter n_A

No	$\frac{\partial Z}{\partial n_A}$
1	$\frac{n_{A\bar{B}}}{n^2{}_A}(log_2\frac{n(n_A-n_{A\bar{B}})}{n_An_B}+\frac{1}{nln2})$
2	$\frac{n_{A\bar{B}}}{n_A^2}$
3	$\frac{n-n_B}{n_Bn_{A\bar{B}}}$
4	$\frac{n_{A\bar{B}}}{2n_A{}^2}$
5	$\frac{3n_{A\bar{B}}}{2n_A{}^2}$
6	$\frac{1}{n_A^2}.\frac{nn_{A\bar{B}}}{(n-n_B)}$
7	$\frac{nn_A^2-n_A^2n_B+nn_Bn_{A\bar{B}}-2nn_An_{A\bar{B}}+nn_{A\bar{B}}^2}{(n_A(n_B-n_A+n_{A\bar{B}}))^2}$
8	$\frac{h+(n_A-n_{A\bar{B}})(n-n_B-n_{A\bar{B}})(n_A(n-n_B)+n_B(n-n_A)\big((n-n_A)(n-n_B)\big)(n_B-n_A+2n_{A\bar{B}})((n-n_A)(n-n_B)+n_An_B)}{\Big(\big((n-n_A)(n-n_B)+n_An_B\big)(n_B-n_A+2n_{A\bar{B}})\Big)^2}$ $h=\big((n-n_B-n_{A\bar{B}})(n_A(n-n_B)(n_A+n_B)\big)\big(n_B\big((n-n_A)\big)(n_A+n_{A\bar{B}})(n-n_B)-n_B\big)\big((n-n_A)(n-n_B)+n_An_B\big)(n_B-n_A+2n_{A\bar{B}})$
9	$\frac{n_{A\bar{B}}}{n_A^2}$
10	$\frac{1}{n}$

11	$\frac{(n-n_B)}{nn_{A\bar{B}}}$
12	$\frac{1}{n}$
13	$\frac{2n_{A\bar{B}}}{n_A^2}$
14	$\frac{1}{n}$
15	$\frac{II'_{n_A}(h)^{\frac{1}{2}} + II\frac{1}{2}(h)^{-\frac{1}{2}}\left(\left(\frac{n_{A\bar{B}}}{n_A^2}\left(log_2\frac{n_A-n_{A\bar{B}}}{n_A}+\frac{1}{ln2}\right)+\frac{n_{A\bar{B}}}{n_A^2}\left(log_2\frac{n_{A\bar{B}}}{n_A}+\frac{1}{ln2}\right)\right)\left(1-\left(-\frac{n_{A\bar{B}}}{n_B}log_2\frac{n_{A\bar{B}}}{n_B}-\frac{n-n_B-n_{A\bar{B}}}{n_B}log_2\frac{n-n_B-n_{A\bar{B}}}{n_B}\right)\right)\right)}{2\sqrt{II\left(1-\left(-\frac{n_A-n_{A\bar{B}}}{n_A}log_2\frac{n_A-n_{A\bar{B}}}{n_A}-\frac{n_{A\bar{B}}}{n_A}log_2\frac{n_{A\bar{B}}}{n_A}\right)\left(1-\left(-\frac{n_{A\bar{B}}}{n_B}log_2\frac{n_{A\bar{B}}}{n_B}-\frac{n-n_B-n_{A\bar{B}}}{n_B}log_2\frac{n-n_B-n_{A\bar{B}}}{n_B}\right)\right)\right)^{\frac{1}{2}}}}$ $h = 1-\left(-\frac{n_A-n_{A\bar{B}}}{n_A}log_2\frac{n_A-n_{A\bar{B}}}{n_A}-\frac{n_{A\bar{B}}}{n_A}log_2\frac{n_{A\bar{B}}}{n_A}\right)\left(1-\left(-\frac{n_{A\bar{B}}}{n_B}log_2\frac{n_{A\bar{B}}}{n_B}-\frac{n-n_B-n_{A\bar{B}}}{n_B}log_2\frac{n-n_B-n_{A\bar{B}}}{n_B}\right)\right)$
16	$\frac{II'_n(h)^{\frac{1}{4}} + II\frac{1}{4}(h)^{-\frac{3}{4}}\left(\left(1-\left(-\frac{n_{A\bar{B}}}{n_B}log_2\frac{n_{A\bar{B}}}{n_B}-\frac{n-n_B-n_{A\bar{B}}}{n_B}log_2\frac{n-n_B-n_{A\bar{B}}}{n_B}\right)^2\right)-k\right)}{2\sqrt{II(h)^{\frac{1}{4}}}}$ $h = 1-\left(-\frac{n_A-n_{A\bar{B}}}{n_A}log_2\frac{n_A-n_{A\bar{B}}}{n_A}-\frac{n_{A\bar{B}}}{n_A}log_2\frac{n_{A\bar{B}}}{n_A}\right)^2\left(1-\left(-\frac{n_{A\bar{B}}}{n_B}log_2\frac{n_{A\bar{B}}}{n_B}-\frac{n-n_B-n_{A\bar{B}}}{n_B}log_2\frac{n-n_B-n_{A\bar{B}}}{n_B}\right)^2\right)$ $k = 2\left(-\frac{n_A-n_{A\bar{B}}}{n_A}log_2\frac{n_A-n_{A\bar{B}}}{n_A}-\frac{n_{A\bar{B}}}{n_A}log_2\frac{n_{A\bar{B}}}{n_A}\right)\left(-\frac{n_{A\bar{B}}}{n_A^2}\left(log_2\frac{n_A-n_{A\bar{B}}}{n_A}+\frac{1}{ln2}\right)-\frac{1}{n_A^2}\left(log_2\frac{n_{A\bar{B}}}{n_A}-+\frac{1}{ln2}\right)\right)$
17	$\frac{n_{A\bar{B}}}{(n_A-n_{A\bar{B}})^2}$
18	$\frac{(1-\theta)}{n}$
19	$\frac{2nn_A(n_A-n_{A\bar{B}})-n_A((n_A-n_{A\bar{B}})^2+{n_{A\bar{B}}}^2)}{n^2{n^2}_A}$ $-\frac{2(n_B-n_A+n_{A\bar{B}})+((n_B-n_A+n_{A\bar{B}})^2+(n-n_B-n_{A\bar{B}})^2)(2n-n_A}{(n(n-n_A))^2}$
20	$\frac{\alpha}{\beta}$ where $\alpha = \left(\frac{3}{n}\right)\left(2-max\left(\frac{n_A}{n},\frac{n-n_A}{n}\right)-max\left(\frac{n_B}{n},\frac{n-n_B}{n}\right)\right)$ $-\left(max\left(\frac{n_A-n_{A\bar{B}}}{n},\frac{n_{A\bar{B}}}{n}\right)+max\left(\frac{n_B-n_A+n_{A\bar{B}}}{n},\frac{n-n_B-n_{A\bar{B}}}{n}\right)+max\left(\frac{n_A-n_{A\bar{B}}}{n},\frac{n_B-n_A+n_{A\bar{B}}}{n}\right)+max\left(\frac{n_{A\bar{B}}}{n},\frac{n-n_B-n_{A\bar{B}}}{n}\right)-max\left(\frac{n_A}{n},\frac{n-n_A}{n}\right)-max\left(\frac{n_B}{n},\frac{n-n_B}{n}\right)\right)\left(\frac{1}{n}\right)$ $\beta = (2-max\left(\frac{n_A}{n},\frac{n-n_A}{n}\right)-max\left(\frac{n_B}{n},\frac{n-n_B}{n}\right))^2$

21	$-\frac{1}{2\sqrt{\frac{n-n_B}{n}}}\frac{n_{A\bar{B}}}{n_A}\left(\frac{n}{n_A}\right)^{\frac{3}{2}}-\frac{1}{2}\sqrt{\frac{n-n_B}{n_A}}$
22	$-\sum_{k=max\,(1,n_A-n_B)}^{n_{A\bar{B}}}\frac{n_B!}{((n_A-k-1)!\,(n_B-n_A+k)!}\frac{(n-n_B)!}{k!\,(n-n_B-k)!}\frac{(n_A-1)!\,(n-n_A)!}{n!}$
23	$-\frac{2}{2^{n_A}ln2}\sum_{k=1}^{n_{A\bar{B}}}\frac{(n_A-1)!}{k!\,(n_A-k-1)!}$
24	$\begin{cases}\frac{n_{A\bar{B}}}{n^2{}_A}\left(log_2\frac{n_A-n_{A\bar{B}}}{n_A}+\frac{n_{A\bar{B}}}{ln2}\right)+\frac{1}{n^2{}_A}\left(log_2\frac{n_{A\bar{B}}}{n_A}+\frac{n_{A\bar{B}}}{ln2}\right) & if\left(\frac{n_B}{n}\leq\frac{1}{2}\ and\ \frac{n_A-n_{A\bar{B}}}{n_A}>\frac{1}{2}\right.\\ 0 & if\left(\frac{n_B}{n}>\frac{1}{2}\ and\ \frac{n_A-n_{A\bar{B}}}{n_A}\leq\frac{1}{2}\right.\\ -\frac{\frac{n_{A\bar{B}}}{n^2{}_A}\left(log_2\frac{n_A-n_{A\bar{B}}}{n_A}+\frac{n_{A\bar{B}}}{ln2}\right)+\frac{1}{n^2{}_A}\left(log_2\frac{n_{A\bar{B}}}{n_A}+\frac{n_{A\bar{B}}}{ln2}\right)}{\frac{n_B}{n}log_2\frac{n_B}{n}+\frac{n-n_B}{n}log_2\frac{n-n_B}{n}} & if\left(\frac{n_B}{n}>\frac{1}{2}\ and\ \frac{n_A-n_{A\bar{B}}}{n_A}>\frac{1}{2}\right.\end{cases}$
25	$\begin{cases}\frac{nn_{A\bar{B}}}{(n-n_B)n^2{}_A} & if\ \frac{n-n_{A\bar{B}}}{n_A}>\frac{n_B}{n}\\ \frac{(n_An_B(n_A-n_{A\bar{B}}))-nn_B(n_A-n_{A\bar{B}})}{n_Bn^2{}_A} & otherwise\end{cases}$
26	$\frac{1}{n}\left(log_2\frac{n(n_A-n_{A\bar{B}})}{n_An_B}-\frac{n_Bn_{A\bar{B}}}{ln2}\right)+\frac{n_{A\bar{B}}}{nn_Aln2}$
27	$\left(\frac{\frac{1}{n}}{2\sqrt{\frac{n_A-n_{A\bar{B}}}{n}}}\left(\frac{n-n_B}{n}-\frac{n_{A\bar{B}}}{n_A}\right)\right)-\sqrt{\frac{n_A-n_{A\bar{B}}}{n}}\left(\frac{n_{A\bar{B}}}{n^2{}_A}\right)$
28	$\left(\frac{n_{A\bar{B}}}{n^2{}_A}-\frac{n-n_B-n_{A\bar{B}}}{(n-n_A)^2}\right)(log_2(n-n_B)-log_2\,n_B)$
29	$\frac{1}{2}\left(\frac{1}{n_B}+\frac{n_{A\bar{B}}}{n_A^2}\right)$
30	$\frac{3+n_{A\bar{B}}}{(n_A+2)^2}$
31	$\frac{1}{n_B}$
32	$\frac{n_{A\bar{B}}}{n_A^2}-\frac{n_B}{n^2}$
33	$\frac{1}{n}\left(log_2\frac{n(n-n_{A\bar{B}})}{n_An_B}+\frac{n_A-n_{A\bar{B}}}{n_Aln2}\right)+\frac{n_{A\bar{B}}}{nn_Aln2}-\frac{1}{n}log_2\frac{n(n_B-n_A+n_{A\bar{B}})}{(n-n_A)n_B}+\frac{n_B-n_A+n_{A\bar{B}}}{n}\left(\frac{(n_B-n_A+n_{A\bar{B}})(n-n_A)}{(n-n_A)(n_B-n_A+n_{A\bar{B}})ln2}\right)+\frac{n-n_B-n_{A\bar{B}}}{n}\frac{\left(n(n-n_B-n_{A\bar{B}})\right)(n-n_B)}{((n-n_A)(n-n_B))^2ln2}$
34	0
35	$\frac{2}{n}+\frac{3n_{A\bar{B}}}{2n_A^2}$
36	$\frac{1}{n_B}$
37	$\frac{1}{n_{A\bar{B}}}$

38	$\frac{n-n_B-n_{A\bar{B}}}{(n-n_A)^2}$
39	$\frac{(n-n_B)\,max\big((n_A-n_{A\bar{B}})(n-n_B),n_Bn_{A\bar{B}}\big)-(nn_A-n_An_B-nn_{A\bar{B}})(n-n_B)}{(max\big((n_A-n_{A\bar{B}})(n-n_B),n_Bn_{A\bar{B}}\big))^2}$

Appendix 4

Formula of partial derivative under the parameter n_B

No	$\frac{\partial Z}{\partial n_B}$
1	$-\frac{n_A-n_{A\bar{B}}}{n_An_B.ln2}$
2	$-\frac{1}{n}$
3	$-\frac{n_A-n_{A\bar{B}}}{n_{A\bar{B}}}.\frac{n}{n_B^2}$
4	$\frac{n_{A\bar{B}}}{2(n-n_B)^2}$
5	$\frac{n_{A\bar{B}}}{2(n-n_B)^2}$
6	$\frac{nn_{A\bar{B}}}{n_A(n-n_B)^2}$
7	$-\frac{(n_A-n_{A\bar{B}})(n-n_A)}{n_A(n_B-n_A+n_{A\bar{B}})^2}$
8	$\frac{-\big((\,n_A-n_{A\bar{B}})(n_A(n-n_B)\big)+h+k+l}{\Big(\big((n-n_A)(n-n_B)+n_An_B\big)(n_B-\,n_A+2n_{A\bar{B}})\Big)^2}$ $h=\Big(n_B\big((n-n_A)\big)+(n-n_B-n_{A\bar{B}})\Big)\big((n-2n_A)\big)((n-n_A)(n-n_B)+n_An_B)(n_B-\,n_A+2n_{A\bar{B}})$ $k=\,(n_A-n_{A\bar{B}})(n-\,n_B-n_{A\bar{B}})\big(n_A(n-n_B)\big)$ $l=n_B(n-\,n_A)\big((n_B-n_A+2n_{A\bar{B}})(n-n_A)\big)(2n_{A\bar{B}})((n-n_A)n-n_B(n-n_B)+n_An_B)$
9	0
10	$-\frac{1}{n}$
11	$-\frac{n_A}{nn_{A\bar{B}}}$
12	0
13	0
14	0
15	$\frac{II'_{n_A}(h)^{\frac{1}{2}}+II\frac{1}{2}(h)^{-\frac{1}{2}}\left(1-\left(-\frac{n_A-n_{A\bar{B}}}{n_A}log_2\frac{n_A-n_{A\bar{B}}}{n_A}-\frac{n_{A\bar{B}}}{n_A}log_2\frac{n_{A\bar{B}}}{n_A}\right)\left(\left(\frac{n_{A\bar{B}}}{n_B^2}(log_2\frac{n_{A\bar{B}}}{n_B}+\frac{1}{ln2})-\frac{(n-n_{A\bar{B}})}{n_B^2}(log_2\frac{n-n_B-n_{A\bar{B}}}{n_B}+\frac{1}{ln2})\right)\right)}{2\sqrt{II\left(1-\left(-\frac{n_A-n_{A\bar{B}}}{n_A}log_2\frac{n_A-n_{A\bar{B}}}{n_A}-\frac{n_{A\bar{B}}}{n_A}log_2\frac{n_{A\bar{B}}}{n_A}\right)\left(1-\left(-\frac{n_{A\bar{B}}}{n_B}log_2\frac{n_{A\bar{B}}}{n_B}-\frac{n-n_B-n_{A\bar{B}}}{n_B}log_2\frac{n-n_B-n_{A\bar{B}}}{n_B}\right)\right)\right)^{\frac{1}{2}}}}$ $h=1-\left(-\frac{n_A-n_{A\overline{B}}}{n_A}log_2\frac{n_A-n_{A\overline{B}}}{n_A}-\frac{n_{A\overline{B}}}{n_A}log_2\frac{n_{A\overline{B}}}{n_A}\right)\left(1-\left(-\frac{n_{A\overline{B}}}{n_B}log_2\frac{n_{A\overline{B}}}{n_B}-\frac{n-n_B-n_{A\overline{B}}}{n_B}log_2\frac{n-n_B-n_{A\overline{B}}}{n_B}\right)\right)$

16	$\frac{II'_n(h)^{\frac{1}{4}} + II\frac{1}{4}(h)^{-\frac{3}{4}}\left(1-\left(-\frac{n_A-n_{A\bar{B}}}{n_A}log_2\frac{n_A-n_{A\bar{B}}}{n_A}-\frac{n_{A\bar{B}}}{n_A}log_2\frac{n_{A\bar{B}}}{n_A}\right)^2 l\right)}{2\sqrt{II\left(1-\left(-\frac{n_A-n_{A\bar{B}}}{n_A}log_2\frac{n_A-n_{A\bar{B}}}{n_A}-\frac{n_{A\bar{B}}}{n_A}log_2\frac{n_{A\bar{B}}}{n_A}\right)^2\left(1-\left(-\frac{n_{A\bar{B}}}{n_B}log_2\frac{n_{A\bar{B}}}{n_B}-\frac{n-n_B-n_{A\bar{B}}}{n_B}log_2\frac{n-n_B-n_{A\bar{B}}}{n_B}\right)^2\right)\right)}}$ $h = 1-\left(-\frac{n_A-n_{A\bar{B}}}{n_A}log_2\frac{n_A-n_{A\bar{B}}}{n_A}-\frac{n_{A\bar{B}}}{n_A}log_2\frac{n_{A\bar{B}}}{n_A}\right)^2\left(1-\left(-\frac{n_{A\bar{B}}}{n_B}log_2\frac{n_{A\bar{B}}}{n_B}-\frac{n-n_B-n_{A\bar{B}}}{n_B}log_2\frac{n-n_B-n_{A\bar{B}}}{n_B}\right)^2\right)$ $l = \left(-2\left(-\frac{n_{A\bar{B}}}{n_B}log_2\frac{n_{A\bar{B}}}{n_B}-\frac{n-n_B-n_{A\bar{B}}}{n_B}log_2\frac{n-n_B-n_{A\bar{B}}}{n_B}\right)\left(-\frac{n_{A\bar{B}}}{n_B^2}(log_2\frac{n_{A\bar{B}}}{n_B}+\frac{1}{ln2})-\frac{(n-n_{A\bar{B}})}{n_B^2}(log_2\frac{n-n_B-n_{A\bar{B}}}{n_B}+\frac{1}{ln2})\right)\right)$
17	0
18	0
19	$\frac{2(n_B-n_A+n_{A\bar{B}})-2(n-n_B-n_{A\bar{B}})}{n(n-n_A)}-\frac{2n_B+2(n-n_B)}{n^2}$
20	$\frac{\alpha}{\beta}$ *where* $\alpha = \left(\frac{2}{n}\right)\left(2-max\left(\frac{n_A}{n},\frac{n-n_A}{n}\right)-max\left(\frac{n_B}{n},\frac{n-n_B}{n}\right)\right)(max\left(\frac{n_A-n_{A\bar{B}}}{n},\frac{n_{A\bar{B}}}{n}\right)+max\left(\frac{n_B-n_A+n_{A\bar{B}}}{n},\frac{n-n_B-n_{A\bar{B}}}{n}\right)+max\left(\frac{n_A-n_{A\bar{B}}}{n},\frac{n_B-n_A+n_{A\bar{B}}}{n}\right)+max\left(\frac{n_{A\bar{B}}}{n},\frac{n-n_B-n_{A\bar{B}}}{n}\right)-max\left(\frac{n_A}{n},\frac{n-n_A}{n}\right)-max\left(\frac{n_B}{n},\frac{n-n_B}{n}\right))(\frac{1}{n})$ $\beta = (2-max\left(\frac{n_A}{n},\frac{n-n_A}{n}\right)-max\left(\frac{n_B}{n},\frac{n-n_B}{n}\right))^2$
21	$\frac{1}{2}n_{A\bar{B}}(\frac{n_A}{n})^{-\frac{1}{2}}(n-n_B)^{-\frac{3}{2}}+\frac{1}{2}(\frac{n_A}{n})^{\frac{1}{2}}(n-n_B)^{-\frac{1}{2}}$
22	$-\sum_{k=max\,(1,n_A-n_B)}^{n_{A\bar{B}}}\frac{(n_B-1)!}{((n_A-k)!\,(n_B-n_A+k-1)!}\frac{(n-n_B)!}{k!\,(n-n_B-k)!}\frac{n_A!\,(n-n_A)!}{n!}$
23	0
24	$\begin{cases} 0 & if(\frac{n_B}{n}\le\frac{1}{2}\ and\ \frac{n_A-n_{A\bar{B}}}{n_A}>\frac{1}{2} \\ \frac{\frac{1}{n}(log_2\frac{n_B}{n}+\frac{1}{ln2})-\frac{1}{n}log_2\frac{n-n_B}{n}-\frac{1}{(n-n_B)ln2}}{(\frac{n_B}{n}log_2\frac{n_B}{n}+\frac{n-n_B}{n}log_2\frac{n-n_B}{n})^2} & if(\frac{n_B}{n}>\frac{1}{2}\ and\ \frac{n_A-n_{A\bar{B}}}{n_A}\le\frac{1}{2} \\ -\frac{(\frac{n_A-n_{A\bar{B}}}{n_A}log_2\frac{n_A-n_{A\bar{B}}}{n_A}+\frac{n_{A\bar{B}}}{n_A}log_2\frac{n_{A\bar{B}}}{n_A})(\frac{1}{n}(log_2\frac{n_B}{n}+\frac{1}{ln2})-\frac{1}{n}log_2\frac{n-n_B}{n}-\frac{1}{(n-n_B)ln2})}{(\frac{n_B}{n}log_2\frac{n_B}{n}+\frac{n-n_B}{n}log_2\frac{n-n_B}{n})^2} & if(\frac{n_B}{n}>\frac{1}{2}\ and\ \frac{n_A-n_{A\bar{B}}}{n_A}>\frac{1}{2} \end{cases}$
25	$\begin{cases} -\frac{nn_An_{A\bar{B}}}{(nn_A-n_An_B)^2} & if\ \frac{n-n_{A\bar{B}}}{n_A}>\frac{n_B}{n} \\ \frac{n(n_A-n_{A\bar{B}})}{n_An^2{}_B} & otherwise \end{cases}$
26	$\frac{n_A-n_{A\bar{B}}}{nn_Bln2}+\frac{n_{A\bar{B}}}{n(n-n_B)ln2}$
27	$-\frac{1}{n}\sqrt{\frac{n_A-n_{A\bar{B}}}{n}}$
28	$\left(\frac{1}{n-n_A}\right)(log_2(n-n_B)-log_2\,n_B)+\left(\frac{n_A-n_{A\bar{B}}}{n_A}-\frac{n-n_B-n_{A\bar{B}}}{n-n_A}\right)\left(\frac{-1}{(n-n_B)ln2}-\frac{1}{n_Bln2}\right)$
29	$-\left(\frac{n_A-n_{A\bar{B}}}{2n_B^2}\right)$

30	0
31	$-\frac{n_A - 2n_{A\bar{B}}}{n_B^2}$
32	$-\frac{n_A}{n^2}$
33	$\frac{n_A - n_{A\bar{B}}}{nn_B ln2} + \frac{n_{A\bar{B}}}{n(n-n_B)ln2} + \frac{1}{n} log_2 \frac{n(n_B - n_A + n_{A\bar{B}})}{(n-n_A)n_B} + \frac{n_B - n_A + n_{A\bar{B}}}{n}\left(\frac{n_A - n_{A\bar{B}}}{n_B(n_B - n_A + n_{A\bar{B}})ln2}\right) - \frac{1}{n} log_2 \frac{n(n - n_B - n_{A\bar{B}})}{(n-n_A)(n-n_B)} - \frac{n - n_B - n_{A\bar{B}}}{n}\left(\frac{(n-n_B) + (n - n_B - n_{A\bar{B}})}{(n-n_B) + (n - n_B - n_{A\bar{B}})ln2}\right)$
34	$\frac{1}{n}$
35	$\frac{-3}{2n} - \frac{2n_{A\bar{B}}}{(n-n_B)^2}$
36	$-\frac{n_A - n_{A\bar{B}}}{n_B^2}$
37	0
38	$\frac{-1}{n - n_A}$
39	$\frac{(nn_A - n_A n_B - nn_{A\bar{B}})n_B}{(max\left((n_A - n_{A\bar{B}})(n-n_B), n_B n_{A\bar{B}}\right))^2}$

Appendix 5

Formula of partial derivative under the parameter $n_{A\bar{B}}$

No	$\frac{\partial Z}{\partial n_{A\bar{B}}}$
1	$-\frac{1}{n_A} log_2 \frac{n(n_A - n_{A\bar{B}})}{n_A n_B} - \frac{1}{(n_A - n_{A\bar{B}})ln2}$
2	$-\frac{1}{n_A}$
3	$-n_A \frac{n - n_B}{n_B n_{A\bar{B}}^2}$
4	$-\frac{1}{2}\left(\frac{1}{n_A} + \frac{1}{n - n_B}\right)$
5	$-\frac{1}{2}\left(\frac{3}{n_A} + \frac{1}{n - n_B}\right)$
6	$-\frac{n}{n_A(n-n_B)}$
7	$-\frac{(n-n_A)n_B}{n_A(n_B - n_A + n_{A\bar{B}})^2}$
8	$\frac{h - k + n_B(n - n_A) - l)(2\left((n-n_A)(n-n_B) + n_A n_B\right)}{\left(\left((n-n_A)(n-n_B) + n_A n_B\right)(n_B - n_A + 2n_{A\bar{B}})\right)^2}$ $h = -\left((n_A(n-n_B)\right) + \left(n_B\left((n-n_A)\right)\left((n - n_B - n_{A\bar{B}})\right)\right)$ $k = (n_A - n_{A\bar{B}})(n_A - n_{A\bar{B}})(n - n_B - n_{A\bar{B}})\left(n_A(n-n_B)\right)$ $l = (n_A - n_{A\bar{B}})(n - n_B - n_{A\bar{B}})(n_A(n-n_B) + n_B(n - n_A)$
9	$-\frac{1}{n_A}$

10	$-\frac{4}{n}$
11	$-\frac{n_A(n-n_B)}{nn_{A\bar{B}}^2}$
12	0
13	$-\frac{2}{n_A}$
14	$-\frac{2}{n}$
15	$\frac{II\frac{1}{2}(h)^{-\frac{1}{2}}\left(k+\left(1-\left(-\frac{n_A-n_{A\bar{B}}}{n_A}log_2\frac{n_A-n_{A\bar{B}}}{n_A}-\frac{n_{A\bar{B}}}{n_A}log_2\frac{n_{A\bar{B}}}{n_A}\right)\left(\left(\frac{1}{n_B}(log_2\frac{n_{A\bar{B}}}{n_B}+\frac{1}{ln2})-\frac{1}{n_B}(log_2\frac{n-n_B-n_{A\bar{B}}}{n_B}+\frac{1}{ln2})\right)\right)\right)\right)}{2\sqrt{II\left(1-\left(-\frac{n_A-n_{A\bar{B}}}{n_A}log_2\frac{n_A-n_{A\bar{B}}}{n_A}-\frac{n_{A\bar{B}}}{n_A}log_2\frac{n_{A\bar{B}}}{n_A}\right)\left(1-\left(-\frac{n_{A\bar{B}}}{n_B}log_2\frac{n_{A\bar{B}}}{n_B}-\frac{n-n_B-n_{A\bar{B}}}{n_B}log_2\frac{n-n_B-n_{A\bar{B}}}{n_B}\right)\right)\right)^{\frac{1}{2}}}}$ $h=1-\left(-\frac{n_A-n_{A\bar{B}}}{n_A}log_2\frac{n_A-n_{A\bar{B}}}{n_A}-\frac{n_{A\bar{B}}}{n_A}log_2\frac{n_{A\bar{B}}}{n_A}\right)\left(1-\left(-\frac{n_{A\bar{B}}}{n_B}log_2\frac{n_{A\bar{B}}}{n_B}-\frac{n-n_B-n_{A\bar{B}}}{n_B}log_2\frac{n-n_B-n_{A\bar{B}}}{n_B}\right)\right)$ $k=\left(\left(-\frac{1}{n_A}(log_2\frac{n_A-n_{A\bar{B}}}{n_A}+\frac{1}{ln2})+\frac{1}{n_A}(log_2\frac{n_{A\bar{B}}}{n_A}+\frac{1}{ln2})\right)\left(1-\left(-\frac{n_{A\bar{B}}}{n_B}log_2\frac{n_{A\bar{B}}}{n_B}-\frac{n-n_B-n_{A\bar{B}}}{n_B}log_2\frac{n-n_B-n_{A\bar{B}}}{n_B}\right)\right)\right)$
16	$\frac{II'_n(h)^{\frac{1}{4}}+II\frac{1}{4}(h)^{-\frac{3}{4}}(k)+\left(1-\left(-\frac{n_A-n_{A\bar{B}}}{n_A}log_2\frac{n_A-n_{A\bar{B}}}{n_A}-\frac{n_{A\bar{B}}}{n_A}log_2\frac{n_{A\bar{B}}}{n_A}\right)^2 l\right)}{2\sqrt{II\left(1-\left(-\frac{n_A-n_{A\bar{B}}}{n_A}log_2\frac{n_A-n_{A\bar{B}}}{n_A}-\frac{n_{A\bar{B}}}{n_A}log_2\frac{n_{A\bar{B}}}{n_A}\right)^2\left(1-\left(-\frac{n_{A\bar{B}}}{n_B}log_2\frac{n_{A\bar{B}}}{n_B}-\frac{n-n_B-n_{A\bar{B}}}{n_B}log_2\frac{n-n_B-n_{A\bar{B}}}{n_B}\right)^2\right)\right)^{\frac{1}{4}}}}$ $h=1-\left(-\frac{n_A-n_{A\bar{B}}}{n_A}log_2\frac{n_A-n_{A\bar{B}}}{n_A}-\frac{n_{A\bar{B}}}{n_A}log_2\frac{n_{A\bar{B}}}{n_A}\right)^2\left(1-\left(-\frac{n_{A\bar{B}}}{n_B}log_2\frac{n_{A\bar{B}}}{n_B}-\frac{n-n_B-n_{A\bar{B}}}{n_B}log_2\frac{n-n_B-n_{A\bar{B}}}{n_B}\right)^2\right)$ $k=-2\left(-\frac{n_A-n_{A\bar{B}}}{n_A}log_2\frac{n_A-n_{A\bar{B}}}{n_A}-\frac{n_{A\bar{B}}}{n_A}log_2\frac{n_{A\bar{B}}}{n_A}\right)\left(\frac{1}{n_A}(log_2\frac{n_A-n_{A\bar{B}}}{n_A}-\frac{1}{ln2})-\frac{1}{n_A}(log_2\frac{n_{A\bar{B}}}{n_A}+\frac{1}{ln2})\right)\left(1-\left(-\frac{n_{A\bar{B}}}{n_B}log_2\frac{n_{A\bar{B}}}{n_B}-\frac{n-n_B-n_{A\bar{B}}}{n_B}log_2\frac{n-n_B-n_{A\bar{B}}}{n_B}\right)^2\right)$ $l=\left(-2\left(-\frac{n_{A\bar{B}}}{n_B}log_2\frac{n_{A\bar{B}}}{n_B}-\frac{n-n_B-n_{A\bar{B}}}{n_B}log_2\frac{n-n_B-n_{A\bar{B}}}{n_B}\right)\left(-\frac{1}{n_B}(log_2\frac{n_{A\bar{B}}}{n_B}+\frac{1}{ln2})+\frac{1}{n_B}log_2\frac{n-n_B-n_{A\bar{B}}}{n_B}-\frac{1}{ln2})\right)\right)$
17	$-\frac{n_A}{(n_A-n_{A\bar{B}})^2}$
18	$-\frac{1}{n}$
19	$\frac{-2(n_A-n_{A\bar{B}})+2n_{A\bar{B}}}{nn_A}+\frac{2(n_B-n_A+n_{A\bar{B}})-2(n-n_B-n_{A\bar{B}})}{n(n-n_A)}$
20	$\frac{\alpha}{\beta}\ where$ $\alpha=\left(max\left(\frac{-1}{n},\frac{1}{n}\right)+max\left(\frac{1}{n},\frac{-1}{n}\right)+max\left(\frac{-1}{n},\frac{1}{n}\right)+max\left(\frac{1}{n},\frac{-1}{n}\right)\right)$ $\beta=2-max\left(\frac{n_A}{n},\frac{n-n_A}{n}\right)-\left(\frac{n_B}{n},\frac{n-n_B}{n}\right)$

21	$\frac{1}{\sqrt{\frac{n_A(n-n_B)}{n}}}$
22	$-\sum_{k=max\,(1,n_A-n_B)}^{n_{A\bar{B}}-1} \frac{n_B!}{((n_A-k)!\,(n_B-n_A+k)!}\frac{(n-n_B)!}{k!\,(n-n_B-k)!}\frac{(n_A)!\,(n-n_A)!}{n!}$
23	0
24	$\begin{cases} -\frac{1}{n_A}log_2\frac{n_A-n_{A\bar{B}}}{n_A}-\frac{1}{(n_A-n_{A\bar{B}})ln2}+\frac{1}{n_A}(log_2\frac{n_{A\bar{B}}}{n_A}+\frac{1}{ln2}) & if(\frac{n_B}{n}\le\frac{1}{2}\ and\ \frac{n_A-n_{A\bar{B}}}{n_A}>\frac{1}{2} \\ 0 & if(\frac{n_B}{n}>\frac{1}{2}\ and\ \frac{n_A-n_{A\bar{B}}}{n_A}\le\frac{1}{2} \\ -\frac{-\frac{1}{n_A}log_2\frac{n_A-n_{A\bar{B}}}{n_A}-\frac{1}{(n_A-n_{A\bar{B}})ln2}+\frac{1}{n_A}(log_2\frac{n_{A\bar{B}}}{n_A}+\frac{1}{ln2})}{\frac{n_B}{n}log_2\frac{n_B}{n}+\frac{n-n_B}{n}log_2\frac{n-n_B}{n}} & if(\frac{n_B}{n}>\frac{1}{2}\ and\ \frac{n_A-n_{A\bar{B}}}{n_A}>\frac{1}{2} \end{cases}$
25	$\begin{cases} -\frac{n}{n_A(n-n_B)} & if\ \frac{n-n_{A\bar{B}}}{n_A}>\frac{n_B}{n} \\ -\frac{n}{n_An_B} & otherwise \end{cases}$
26	$-\frac{1}{n}(log_2\frac{n(n_A-n_{A\bar{B}})}{n_An_B}+\frac{1}{ln2})\ +\frac{1}{n}(log_2\frac{nn_{A\bar{B}}}{n_A(n-n_B)}+\frac{1}{ln2})$
27	$\frac{-\frac{1}{n}}{2\sqrt{\frac{n_A-n_{A\bar{B}}}{n}}}\left(\frac{n-n_B}{n}-\frac{n_{A\bar{B}}}{n_A}\right)-\frac{1}{n_A}\sqrt{\frac{n_A-n_{A\bar{B}}}{n}}$
28	$-\left(\frac{1}{n_A}-\frac{1}{n-n_A}\right)(log_2(n-n_B)-log_2 n_B)$
29	$-\frac{1}{2}\left(\frac{1}{n_A}+\frac{1}{n_B}\right)$
30	$-\frac{1}{n_A+2}$
31	$-\frac{2}{n_B}$
32	$-\frac{1}{n_A}$
33	$-\frac{1}{n}log_2\frac{n(n-n_{A\bar{B}})}{n_An_B}+\frac{n_A-n_{A\bar{B}}}{(n-n_{A\bar{B}})ln2}$ $+\frac{1}{n}log_2\frac{nn_{A\bar{B}}}{n_A(n-n_B)}+\frac{1}{nln2}+\frac{1}{n}log_2\frac{n(n_B-n_A+n_{A\bar{B}})}{(n-n_A)n_B}$ $+\frac{n_B-n_A+n_{A\bar{B}}}{n(n_B-n_A+n_{A\bar{B}})ln2}-\frac{1}{n}log_2\frac{n(n-n_B-n_{A\bar{B}})}{(n-n_A)(n-n_B)}-\frac{n-n_B-n_{A\bar{B}}}{(n-n_B-n_{A\bar{B}})ln2}$
34	0
35	$-\left(\frac{3}{2n_A}+\frac{2}{n-n_B}\right)$
36	$-\frac{1}{n_B}$
37	$-\frac{n_A}{n_{A\bar{B}}^2}$
38	$\frac{-1}{n-n_A}$
39	$\frac{-n\,max\left((n_A-n_{A\bar{B}})(n-n_B),n_Bn_{A\bar{B}}\right)-(nn_A-n_An_B-nn_{A\bar{B}})max\,(-(n-n_B),n_B)}{(max\left((n_A-n_{A\bar{B}})(n-n_B),n_Bn_{A\bar{B}}\right))^2}$

MEDIDAS ALTERNATIVAS DE LOS GRADOS DE ADHESIÓN DE INDIVIDUOS A LA IMPLICACIÓN Y SIMILARIDAD PARA VARIABLES MODALES

Larisa ZAMORA[1] y Pablo GREGORI[2]

MESURES ALTERNATIVES DES DEGRES D'ADHESION DES INDIVIDUS A L'IMPLICATION ET SIMILARITE POUR DES VARIABLES MODALES

ALTERNATIVE MEASUREMENTS OF AGREEMENT LEVEL OF INDIVIDUALS TO IMPLICATION AND SIMILARITY FOR CATEGORICAL VARIABLES

RESUMEN

Se ofrecen tres propuestas alternativas a las introducidas en Bailleul y Gras (1994), para medir el grado de adhesión de los individuos a las implicaciones y similaridades, que afectan, a su vez, al cálculo de su tipicalidad y contribución a la formación de las clases del árbol jerárquico de cohesión y similaridad. Dos propuestas atienden principalmente al efecto que los individuos muestreados producen sobre los índices calculados, mientras que la tercera se basa en una muestra auxiliar.

Palabras clave : *tipicalidad, contribución, similaridad, cohesión, análisis implicativo.*

RÉSUMÉ

On présente trois propositions alternatives à celles introduites à Bailleul et Gras (1994), afin de mesurer le degré d'adhésion des individus aux implications et similarités, qui modifient, à son tour, le calcul de leur typicalité et contribution à la formation des classes de l'arbre hiérarchique de cohésion et similarité. Deux propositions partent de l'effet que les individus échantillonnés produisent sur les indices calculés, tandis que la troisième est basée sur un échantillon auxiliaire.

Mots-clés : *typicalité, contribution, similarité, cohésion, analyse implicative.*

ABSTRACT

We propose some alternative definitions to the one given in Bailleul and Gras (1994), for measuring the agreement level of individuals to implications and similarities, which alter, at the same time, the computation of their typicality and contribution to the formation of classes in the hierarchical tree. Some proposals are based on the effect of the sampled individuals on the computed index, as well as one of them starts from an auxiliary sample.

Keywords : *typicality, contribution, similarity, cohesion, implicative analysis.*

[1] Departamento de Matemática. Facultad de Matemática y Computación, Universidad de Oriente, Santiago de Cuba, Cuba, larisa@csd.uo.edu.cu

[2] Departamento de Matemáticas. Instituto de Matemáticas y Aplicaciones de Castellón, Universitat Jaume I de Castellón, Campus Riu Sec E-12071, Spain, gregori@uji.es

1 Introducción

El Análisis Estadístico Implicativo (ASI) recopila una serie de procedimientos de la estadística multivariante cuyo objetivo es estructurar un grupo de variables muestreadas sobre un conjunto de individuos. Nació en el contexto de la Didáctica de las Matemáticas, en la tesis doctoral Gras (1979), en su vertiente implicativa. Previamente, Lerman (1970) introdujo un análisis de similaridad, donde se usaban la independencia estadística y la probabilidad para medir lo próximas que estaban dos variables binarias. Y en base a esa forma original de medir la similaridad (diferente de las otras similaridades y distancias utilizadas en los análisis *cluster*, o de conglomerados), se creaba la estructura jerárquica de las variables.Trasponer esta situación simétrica, de la similaridad entre variables, a una no simétrica, como es la relación de implicación, donde hipótesis y tesis no son intercambiables, fue el objetivo de la tesis Gras (1979). En ella se definía la intensidad de una (cuasi-)implicación entre variables binarias, con la filosofía de Lerman (independencia + probabilidad), y se aplicaba a un estudio experimental sobre adquisiciones cognitivas. Aquello supuso el nacimiento del ASI que, por completitud, englobó al análisis de similaridades de Lerman en su repertorio.

A partir de entonces, se fue ampliando el rango de actuación de los procedimientos: nacido en el seno de las variables binarias (Lerman et al., 1981ab), que describían la presencia o ausencia de ciertos rasgos de interés didáctico, se ampliaron las definiciones a las variables modales, frecuenciales y de intervalo (Bailleul y Gras, 1994, Lagrange, 1998). De igual manera, se desarrolló la profundidad de dichos análisis. Por una parte, las intensidades de las implicaciones entre parejas de variables permitían representar un grafo dirigido en el que los investigadores pueden interpretar las relaciones implicativas. Por otro lado, tomando como ejemplo la estructura jerárquica a la que daba lugar el análisis de similaridad (en forma de árbol, en el que cada variable es una clase inicial, y a cada nivel de la jerarquía se reúnen las dos clases más similares del nivel precedente), se intentó encontrar una estructura jerárquica entre las implicaciones. Al igual que en el mundo matemático un teorema implica ciertos corolarios, se buscó la forma de crear esa estructura donde unas implicaciones « conducen » a otras.

Para ello, se definió la cohesión de una pareja (ordenada) de variables, como una transformación concreta de su intensidad implicativa. Y la de una clase (ordenada) de variables, como la media geométrica de las cohesiones de todas las parejas (ordenadas) de variables. Finalmente, la implicación de una clase de variables sobre otra, se definió como una fusión particular entre intensidades implicativas y las cohesiones de las clases. Con este último índice se pudo reproducir una estructura jerárquica, como en las similaridades, pero de relaciones no simétricas, que ordena las clases « implicativas ». A este análisis se le llama análisis de cohesiones, y su aplicación a datos reales conduce a conclusiones como « esta relación implicativa entre este grupo de variables *favorece* esta otra relación implicativa entre este otro grupo de variables ».

El ASI también enriquece la profundidad de estos dos análisis jerárquicos:

- destacando algunos de los niveles de la jerarquía (nodos significativos),
- destacando algunos individuos de la muestra, por el papel que juegan en la formación de cada nuevo nivel de la jerarquía.

Se llama nodo (o nudo o clase o nivel) significativo a aquel que contiene una serie de variables que guardan una especial consonancia respectiva, comparada con la de los

niveles de la jerarquía anterior y posterior. El investigador suele atender a alguno de estos niveles significativos a la hora de las interpretaciones, ya que sus variables constituyentes están más fuertemente ligadas que en otros niveles del árbol jerárquico.

En cada nivel de la jerarquía (sea un nodo significativo o no), se clasifican los individuos de la muestra. Por una parte se clasifican en torno a la « contribución » para la formación de ese nivel (o clase). Es decir, que para cada individuo se calcula un valor que indica lo mucho que sus datos han favorecido la formación de dicha clase. Por ejemplo, si una clase consiste en la similaridad de dos variables, un individuo que tiene valores distintos en dichas variables, no facilita en absoluto que esas dos variables se hayan reunido en una clase, como similaridad. Sin embargo, otro individuo cuyos datos sobre ambas variables coinciden, es quien más fuertemente favorece que dichas variables se hayan reunido en una clase. El reconocimiento de dichos individuos también puede ayudar a la interpretación de resultados, incluso si se usan individuos « ficticios » que representen perfiles fácilmente identificables por el investigador.

La otra clasificación que se realiza sobre los individuos se formaliza en torno a la sintonía que hay entre la clase formada y los datos del individuo sobre las variables de esa clase: si una clase tiene un índice de similaridad medio, un individuo « típico » no será aquel cuyos datos sean muy coincidentes para todas las variables de dicha clase (ya que esa idea lleva al concepto descrito anteriormente). Sería muy típico, más bien, aquel individuo cuyos datos coinciden en las variables muy similares y no tanto en variables menos similares (de ahí el nombre tipicalidad).

Estas clasificaciones se realizan a partir de comparar los índices de las clases con una versión « individualizada » de los mismos (a la que llaman « respeto » o « grado de adhesión »), y el ordenamiento posterior de los individuos, con los pesos (de contribución o tipicalidad) respectivos. En este trabajo repasamos las propuestas de la literatura y aportamos nuevos puntos de vista para su consideración.

2 Definiciones, notaciones y datos de ejemplo

En el ASI, se denota por V al conjunto de variables, entre las que se encuentran a, b, etc., y por E, al conjunto de individuos, entre los que se encuentran x, y, etc., de modo que a(x) es el dato del individuo x sobre la variable a, etc. Las variables originarias del ASI fueron binarias, pero su aplicación se amplió a más tipos de variables. En este trabajo nos centramos en las variables modales ordinales, con n categorías cada una, como las escalas Likert. Para su tratamiento por el ASI, estas variables deben recodificar sus categorías a los valores $0, \frac{1}{m-1}, \frac{2}{m-1}, \ldots, \frac{m-2}{m-1}$ y 1. Por ejemplo, para $m=5$ categorías, se tendrían las recodificaciones 0, 0.25, 0.5, 0.75 y 1.

En este trabajo vamos a usar un conjunto de datos de preferencias de estilos musicales (ver anexo), donde un grupo de individuos valora su nivel de preferencia sobre 15 estilos musicales. El ASI nos permite estructurar dichos estilos musicales (en base a las respuestas de los individuos) de dos modos: por similitud y por implicación. Por ejemplo, si los estilos HIP y PUN resultan muy similares para el ASI, no significará que sean parecidos a nivel armónico, lógicamente, sino que ambos estilos agrupan a audiencias relativamente cercanas. De igual modo, si la intensidad de implicación HIP → PUN es elevada, el número de contraejemplos, es decir, de individuos a los que

les gusta HIP y no les gusta PUN, es extraordinariamente bajo, para lo que ocurriría si esos estilos gustaran al público de manera independiente.

Aunque este trabajo aborda una temática que afecta por igual al análisis de similaridades y de cohesiones, debido a su estrecho paralelismo formal, escribiremos únicamente el caso de cohesiones, como es habitual en la literatura, y por ser ligeramente más abstracto, en razón a su carácter no simétrico. La intensidad de implicación entre las variables binarias a y b se denota por $\varphi(a,\bar{b})$ y mide cuán bajo es el número de contraejemplos a la regla $a \rightarrow b$, al compararlo con una situación de aleatoriedad total (independencia estadística) entre a y b. Se calcula por:

$$\varphi(a,\bar{b}) = P(N_{a\bar{b}} > n_{a\bar{b}})$$

donde $n_{a\bar{b}}$ representa el número de contraejemplos a la regla $a \rightarrow b$ observados en la muestra, y $N_{a\bar{b}}$ es una variable aleatoria que calcula el número de contraejemplos bajo independencia entre *a* y *b*, y su media y varianza depende de los efectivos *n* (tamaño de la muestra), n_a (número de individuos que cumplen *a*) y $n_{\bar{b}}$ (número de individuos que no cumplen *b*). Se puede calcular bajo varios modelos de probabilidad, a escoger según la forma del muestreo (Lerman *et al.*, 1981).

La intensidad de implicación generaliza su definición para variables frecuenciales (con valores en el intervalo [0,1]) en Gras y Larher (1992). Las variables modales, tras recodificar sus categorías a valores numéricos equidistantes en el intervalo [0,1], se pueden tratar como variables frecuenciales. Sin pérdida de generalidad, seguiremos llamando *a* y *b* a las variables recodificadas. La intensidad de implicación entre variables modales y frecuenciales se calcula de forma que mantiene coherencia con la definición sobre variables binarias. Basta con definir el nuevo número de contraejemplos como $n_{a\bar{b}} = \sum_{x\in E} a(x)(1-b(x))$, y los nuevos efectivos

$n_a = \sum_{x\in E} a(x)$ y $n_{\bar{b}} = \sum_{x\in E} (1-b(x))$, que coinciden en el caso binario.

Con las intensidades de implicación se puede representar el llamado grafo implicativo, en el que el investigador visualiza fácilmente las implicaciones que resultan con intensidad alta. En la Figura 1 se muestra el resultado para los datos de la música.

Por otro lado se tiene el análisis jerárquico de similaridades de Lerman y el de cohesiones de Gras. Como su estructura es muy paralela, se describe el de cohesiones por ser algo más abstracto.

Inicialmente, podemos decir que cada variable forma una clase por sí misma.

En el primer nivel del proceso jerárquico, se va a formar una nueva clase ordenada, llamémosla $C_1 = (C_{11}, C_{12})$, como la unión de las dos clases del nivel anterior (es decir, las variables) con mayor intensidad de implicación. Aunque en este nivel se torna trivial el concepto que se explica a continuación, se introduce aquí para facilitar su comprensión en los siguientes niveles:

– Se llama par genérico de la clase $C_1 = (C_{11}, C_{12})$ a la pareja de variables que maximiza la intensidad de implicación, de entre todas las parejas posibles, con una variable en C_{11} y otra en C_{12}. En este nivel el concepto es trivial pues sólo

hay una variable en C_{11} y otra en C_{12}, y por tanto dichas variables forman el par genérico.

- Se llama intensidad de implicación genérica de C_1, al valor de la intensidad implicativa para el par genérico.
- Se valora el « respeto » o « grado de adhesión » de los individuos *x* de *E* frente a dicha implicación genérica, de la siguiente forma :
  - Mínimo respeto: $\varphi_x(a,\bar{b}) = 0$, si *a(x)=1* y *b(x)=0*, por ser *x* un contraejemplo genuino de la regla $a \rightarrow b$.
  - Máximo respeto: $\varphi_x(a,\bar{b}) = 1$, si *b(x)=1*, por ser *x* un ejemplo que « respeta » la regla $a \rightarrow b$.
  - Respeto neutro: $\varphi_x(a,\bar{b}) = 0.5$, si *a(x)=b(x)=0*, por ser *x* un individuo que ni la refuerza ni la contradice.
- Estas valoraciones, del respeto de los individuos sobre la implicación genérica de la clase C_1, sirven para clasificarlos en torno a dos nuevos conceptos: la tipicalidad y la contribución. Se calcula entonces una lista con el valor de tipicalidad para cada individuo *x* respecto de la clase C_1, y otra lista similar con el valor de contribución. No describimos su formulación, pero indicamos su utilidad para enriquecer la información proporcionada por el análisis, pudiendo ser productiva para el investigador.

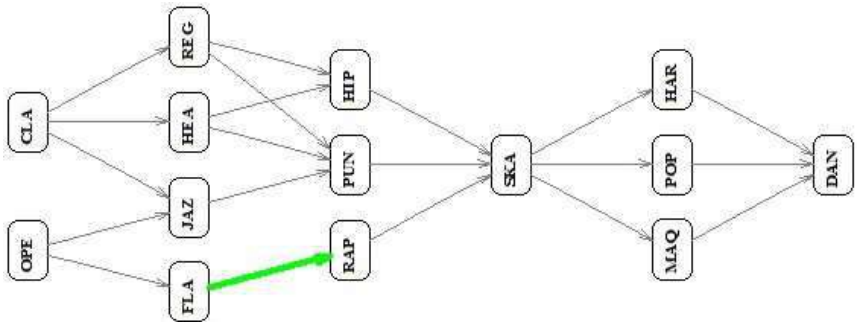


Figura 1 – Grafo implicativo (girado por conveniencia de edición), que expresa una estructura de relaciones de implicación. En este caso las intesidades resultan de muy bajo nivel, pero se representan a modo de ilustración. La única implicación significativa es FLA → RAP.

Para continuar con el nivel 2 (y posteriores) del algoritmo, se precisa definir la implicación entre grupos (o clases) ordenados de variables. Esta definición se proporciona en varias etapas:

1. Cohesión de una pareja (ordenada) de variables *(a,b)*, denotada por *c(a,b)* : se define como una transformación monótona de la intensidad de implicación :
   1. cuyo valor es nulo si $\varphi(a,\bar{b}) < 0.5$, por ser una implicación demasiado débil,
   2. y crece de forma estrictamente monótona con $\varphi(a,\bar{b}) \geq 0.5$. La forma elegida fue $c(a,b) = \sqrt{1 - E(\varphi(a,\bar{b}))^2}$, donde se usa la función de entropía $E(p) := -p\log_2 p - (1-p)\log_2(1-p)$.

2. Cohesión de una clase (ordenada) de variables $A=(a_1,\ldots, a_r)$, denotada por $c(A)$: se define como la media geométrica de las cohesiones de todos los pares de variables (a_i,a_j) donde $1 \le i < j \le r$.
3. Implicación de la clase (ordenada) A sobre la clase (ordenada) B, denotada por $\psi(A,B)$: se define teniendo en cuenta la intensidad máxima entre un elemento de A y otro de B, además de las cohesiones de ambas clases. En concreto se definió como $\psi(A,B)=\left[\max_{\substack{a\in A\\ b\in B}} \varphi(a,\bar{b})\right]^{rs} \sqrt{c(A)c(B)}$, donde r y s son las cantidades de variables en las respectivas clases A y B.

Así pues, en el segundo nivel de la jerarquía (y posteriores) se consideran todas las clases del nivel anterior (quitando las dos que se reunieron, y la nueva que se formó con ellas), y se reúnen las dos clases que ahora presenten la mayor implicación $\psi(A,B)$. Denotemos por $C_2 = (C_{21}, C_{22})$ la nueva clase. Entonces:

- los conceptos de par genérico e intensidad de implicación genérica conllevan buscar esa implicación máxima entre parejas de variables, la primera en C_{21} y la segunda en C_{22}.
- C_{21} y C_{22}, por venir del nivel anterior, deben tener identificados (si no son variables sencillas) sus respectivos pares genéricos, intensidades de implicación genéricas, y clasificados a todos los individuos de la muestra según el respeto sobre sí mismas, en el nivel en que se formó cada una. Esta información se debe guardar a cada nivel del algoritmo.
- Para la nueva clase C_2, se consideran todas las implicaciones genéricas de clases englobadas por C_2, definiendo el llamado vector potencia implicativa de C_2.
- Los conceptos de tipicalidad y contribución de los individuos x sobre la nueva clase C_2 se calcularán, como un promedio, a partir de las valoraciones del respeto de los individuos sobre sólo las implicaciones genéricas de las clases que engloba C_2. Y por tanto retoma los datos calculados en aquellos niveles anteriores, que en esta iteración queden contenidos en la clase C_2.

Estando, los conceptos de tipicalidad y contribución, basados en las valoraciones sobre el respeto de las implicaciones genéricas (solo esas) que contiene la clase, nuestro trabajo pretende aportar nuevos puntos de vista a las filosofías presentadas por Ratsimba-Rajohn (1991), sobre variables binarias, y Bailleul (1994), sobre variables modales.

A modo de ilustración, en la Figura 2 se muestran los resultados respectivos de aplicar los análisis de similaridades y cohesiones a los datos de las preferencias musicales.

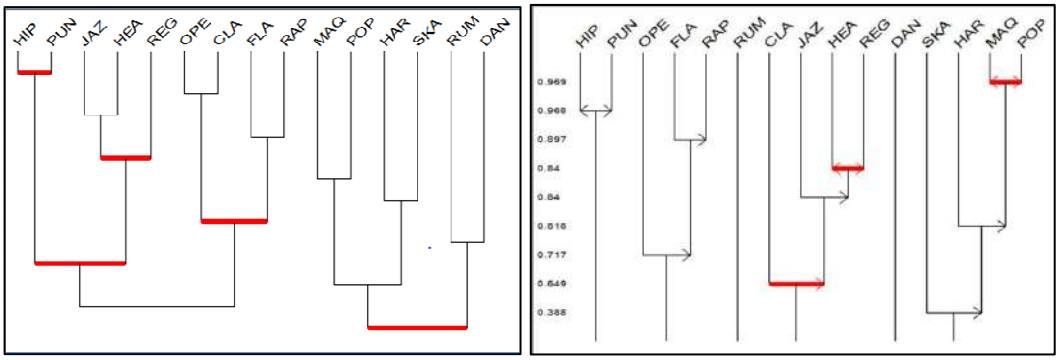


Figura 2 – Árboles jerárquicos, de similaridad (izquierda) y cohesión (derecha), para los datos de música. Los nodos en trazo grueso (y rojo) son los significativos, y se suelen escoger para la interpretación de resultados.

Como se puede ver en las figuras, en los árboles jerárquicos se señalan unos niveles especiales, los nodos significativos, que no trataremos en este trabajo, pero que indican nodos cuyas variables están más ligadas que las del nodo precedente y el posterior, motivo por el cual, el investigador puede indagar con más atención en busca de una interpretación práctica de los resultados, además de centrar en ellos la interpretación de los individuos en cuanto a tipicalidad y contribución.

3 Grado de adhesión de un individuo a una implicación en variable binaria y modal

Como se indicaba en la introducción, la base para las medidas de tipicalidad y contribución de los individuos a las clases formadas, es el « respeto » o « grado de adhesión » de los mismos a las implicaciones genéricas de la clase en cuestión.

En el caso binario, Ratsimba-Rajohn (1991) estableció el « respeto » de un individuo *x* hacia una implicación $a \rightarrow b$, tal y como hemos descrito en el primer nivel del algoritmo jerárquico de la sección anterior.

En la Tabla 1 presentamos dos cuadros: uno en el que ordenamos a los cuatro tipos de individuos (según sus repuestas a *a* y *b*, respectivamente) por orden de menor a mayor respeto, y otro en el que colocamos su « valor de respeto ».

Tabla 1 – A la izquierda, los 4 individuos posibles se ordenan de menor a mayor nivel de « respeto » por la implicacion $a \rightarrow b$. A la derecha, sus valores de « respeto » por la regla $a \rightarrow b$, según la filosofía de Ratsimba-Rajohn (1991).

Ordenamiento		
a / *b*	0	1
0	2	3
1	1	3

Valor de respeto		
a / *b*	0	1
0	0.5	1
1	0	1

En Bailleul y Gras (1994) se plantea la ambigüedad de los dos individuos que presentan el mismo orden (rango), tratándose de dos comportamientos distintos, además de plantear la situación paralela en el caso de variables modales, con *n* categorías cada una de ellas. Para resolverla, propone realizar dos pasos:

1. Ordenar el conjunto de parejas de respuestas posibles,
2. Vincular el valor que mide el « respeto » o la contradicción de la implicación, a la propia intensidad de dicha implicación.

Tomando la noción de implicación, extendida a variables modales, el valor paralelo al número de contraejemplos, es la suma $a(x)(1 - b(x))$ para todos los individuos x (ver definición en sección anterior).

Por ese motivo, los individuos x para los que $a(x) > b(x)$, respetan « menos » la regla $a \rightarrow b$, siendo el menos respetuoso de todos, aquel individuo x tal que $a(x)=1$ y $b(x)=0$. Y por contra, los individuos x para los que $a(x) \leq b(x)$ respetan « más » la regla $a \rightarrow b$, siendo el más respetuoso de todos, aquel individuo x tal que $a(x)=1$ y $b(x)=1$. Según la intuición y justificación, Bailleul y Gras (1994) establecen un ordenamiento, compatible con el de Ratsimba-Rajohn (1991) y que matiza el caso coincidente. En la Tabla 2 mostramos ejemplos de la filosofía de ordenamiento del « grado de adhesion » de Bailleul (1994) para distinto número de categorías de las variables modales.

Tabla 2 – Ordenamientos según el « grado de adhesión », de menor a mayor, de los individuos a la regla $a \rightarrow b$, para el caso de variables modales con $n=2,3$ y 5 categorías, y propuesto por Bailleul y Gras (1994).

a / b	0	1
0	2	3
1	1	4

a / b	0	0.5	1
0	4	5	6
0.5	3	7	8
1	1	2	9

a / b	0	0.25	0.5	0.75	1
0	11	12	13	14	15
0.25	10	16	17	18	19
0.5	6	9	20	21	22
0.75	3	5	8	23	24
1	1	2	4	7	25

Por no abundar en fórmulas innecesarias, indicamos que el ordenamiento se inicia en la pareja de modalidades que ocupa la casilla inferior izquierda (que sería la que menos respeta la implicación), continúa ascendiendo por las paralelas a la antidiagonal (desde abajo hacia arriba), sin llegar a la propia antidiagonal. Y a partir de entonces se ordenan los restantes casos por filas, de izquierda a derecha, y desde la fila superior hacia abajo, hasta llegar a la esquina inferior derecha (que es la que más respeta la implicación).

Una vez establecido el ordenamiento compatible con la filosofía de respeto o grado de adhesión de cada pareja de categorías a la implicación, se calcula un valor numérico para medir dicho grado de respeto, que se calcula a partir del ordenamiento, y del valor del índice de la implicación:

- El ordenamiento realizado se cambia de escala al intervalo [0,1]: restando 1 unidad de las posiciones y dividiendo todas ellas por la máxima resultante (m^2-1).
- Se usa el nuevo valor de la escala como factor multiplicativo de la implicación en cuestión, $\varphi(a, \bar{b})$.

En la Tabla 3 se muestran las tres etapas de la ideas de Bailleul y Gras (1994), para el caso de variables binarias, donde se aprecia la diferencia respecto de la propuesta de Ratsimba-Rajohn (1991), mostrada en la Tabla 1.

Tabla 3 – Proceso de cálculo del grado de adhesión de los individuos a la regla $a \rightarrow b$, para el caso de variables binarias, propuesto por Bailleul y Gras (1994).

Ordenamiento		
a / b	0	1
0	2	3
1	1	4

→

Orden re-escalado		
a / b	0	1
0	1/3	2/3
1	0	1

→

Grado de adhesión		
a / b	0	1
0	$\frac{1}{3}\varphi(a,\bar{b})$	$\frac{2}{3}\varphi(a,\bar{b})$
1	0	$\frac{1}{3}\varphi(a,\bar{b})$

Bailleul y Gras (1994) definen, a partir de este grado de adhesión de cada individuo x a una implicación, las fórmulas de contribución y tipicalidad de cada individuo x a cada nueva clase formada C, usando las implicaciones genéricas de dicha clase C, como se comentaba en la descripción del algoritmo jerárquico de la sección anterior.

4 Propuestas alternativas al grado de adhesión

Debido a que en la literatura no se habían encontrado expresiones para la determinación del índice de similaridad, los nodos significativos, la tipicalidad y contribución de los individuos a las clases para el caso de variables modales, en Zamora (2015) se propone una expresión para calcular el índice de similaridad para estas variables, así como para la determinación de la tipicalidad y contribución de los individuos a las clases que se forman durante la construcción del árbol de similaridad, basándose en los trabajos de Bailleul y Gras (1994) y Lagrange (1998), lo que permitió transponer el concepto de grado de adhesión de los individuos a la similaridad entre dos variables modales. Esto conllevó a los autores del presente trabajo a reformular el papel de este grado de adhesión incluso para la implicación, más consolidada en la literatura.

Por una parte, las propuestas de Ratsimba-Rajohn (1991) y Bailleul y Gras (1994) vienen debidamente justificadas por razonamientos basados en la definición de *regla* y *contraejemplo*. Sin embargo, el concepto de respeto o grado de adhesión a una regla, se puede cuestionar racionalmente desde otros puntos de vista. Por ejemplo:

1. ¿Qué efecto provocaría un nuevo individuo sobre la intensidad de una implicación ya valorada en una muestra? Si se añade, y provoca un aumento de la intensidad, se trataría de un individuo que la respeta (o se adhiere). Si, por el contrario, provoca una disminución de la intensidad, se trataría de un individuo que no la respeta.
2. ¿Qué efecto han provocado los individuos de la muestra sobre la intensidad de una implicación ya valorada? Si se extrae, y provoca un aumento de la intensidad, se trata de un individuo que no la respeta (pues la ha lastrado al incluirse en la muestra). Si, por el contrario, disminuye la intensidad, se trata de un individuo que la respeta (pues su ausencia debilita la implicación).
3. Consideremos una muestra ajena especial, en la que hay un individuo que representa a cada pareja de categorías de las variables *a* y *b*. Para esa muestra, las variables *a* y *b* serían, por lógica, estadísticamente independientes. Calcular

el incremento de cada individuo al índice también informa de su respeto a la implicación.

En los siguientes apartados desarrollamos estas tres ideas y reflexionamos sobre sus posibilidades, usando datos simulados sobre gustos musicales, bajo distintos números de categorías.

4.1 Propuesta 1 : añadir a la muestra un individuo ficticio de cada tipo

Para la muestra considerada, en las variables *a* y *b*, se procede a añadir cada vez un individuo con cada pareja de categorías posible. Se calcula el incremento que dicho individuo produce en el valor de la intensidad de implicación (de no estar a estar), y al final se ordenan estos individuos por el incremento producido en la intensidad de implicación, desde el que da un incremento más negativo (y sería el primero en el ordenamiento), hasta el que produce un incremento más positivo (que sería el último en el ordenamiento). Esta sería la primera alternativa al ordenamiento dado por Bailleul y Gras (1994).

Para visualizar los efectos de esta propuesta, usamos los datos binarios de la música, y consideramos dos implicaciones, que mostramos en la Tabla 4.

Tabla 4 – Aplicación de la propuesta 1 con 2 modalidades para dos implicaciones de la muestra: datos (izquierda), valores de intensidad de implicación al añadir cada individuo posible (centro), y ordenamiento según el incremento provocado en la intensidad de implicación. Las intensidades reales, sin individuos añadidos, son $\varphi(HIP, \overline{OPE}) = 0.3970013$ y $\varphi(\mathrm{HIP}, \overline{\mathrm{MAQ}}) = 0.5536308$, respectivamente.

Efectivos		
HIP/OPE	0	1
0	6	3
1	9	2

→

Intensidad al añadir 1 individuo		
HIP/OPE	0	1
0	0.4153317	0.3417477
1	0.3884028	0.4418189

→

Ordenamiento		
HIP/OPE	0	1
0	3	1
1	2	4

Efectivos		
HIP/MAQ	0	1
0	2	7
1	2	9

→

Intensidad al añadir 1 individuo		
HIP/MAQ	0	1
0	0.6489726	0.5262412
1	0.4663390	0.5749585

→

Ordenamiento		
HIP/MAQ	0	1
0	4	2
1	1	3

Se puede ver en la Tabla 4 que los ordenamientos son distintos, y que los individuos *extremos* (según los razonamientos a priori de Ratsimba-Rajohn y Bailleul) pueden dejar de serlo, al nivel práctico de "elevar" o "rebajar" la intensidad de la implicación.

Para visualizar los ordenamientos en el caso de más modalidades, simulamos aleatoriamente (ver Anexo) 1000 individuos en 2 variables, con 5 modalidades por variable, mostrando el resultado de la propuesta en la Tabla 5.

Tabla 5 – Aplicación de la propuesta 1 con 5 modalidades para una implicación de una muestra simulada con 1000 individuos y 5 modalidades por variable: valores de intensidad de implicación al añadir cada individuo posible (izquierda), y ordenamiento según el incremento provocado en la intensidad de implicación. La intensidad real, sin individuos añadidos, es $\varphi(V1,\overline{V}\,\overline{2}) = 0.3923378$. Los valores están redondeados por cuestión de espacio, pero no hay coincidencias (ejecutar código R de anexo para ver valores con más precisión).

Intensidad al añadir 1 individuo					
V1/V2	0	0.25	0.5	0.75	1
0	0.401	0.397	0.392	0.388	0.384
0.25	0.396	0.394	0.392	0.390	0.388
0.5	0.392	0.392	0.392	0.392	0.393
0.75	0.388	0.390	0.392	0.395	0.397
1	0.384	0.388	0.392	0.397	0.401

→

Ordenamiento					
V1/V2	0	0.25	0.5	0.75	1
0	24	21	14	5	2
0.25	20	18	12	8	6
0.5	9	10	11	16	17
0.75	3	7	13	19	23
1	1	4	15	22	25

Usando otras simulaciones se llega a ordenamientos ligeramente distintos. Sorprende en este caso el individuo de datos (V1=0;V2=1), que tiene un ordenamiento muy distinto al que le otorga el enfoque de Bailleul y Gras (1994).

En resumen, se puede concluir que, en el caso general, el ordenamiento de las parejas de modalidades no siempre es el mismo, y depende de los valores que tomen las variables en la muestra original, más cuanto menor sea el tamaño de la muestra.

4.2 Propuesta 2 : suprimir de la muestra un individuo de cada tipo

En este caso, se procede a suprimir de la muestra un individuo. De igual manera que en la primera propuesta, se ordenan estos individuos por el incremento producido en la intensidad de implicación (de no estar a estar), y lo establecemos como ordenamiento alternativo al dado por Bailleul y Gras (1994). Aunque este algoritmo recorre aparentemente toda la muestra de individuos, esencialmente sólo precisa calcularse para cada cruce de categorías de *a* y *b* que estén presentes en la muestra, y luego atribuir su valor a cada individuo de la muestra que corresponda. Es cierto que con esta propuesta no atribuimos orden a los individuos que no están presentes en la muestra, aunque no causa problemas por no ser necesario su grado de adhesión para los cálculos posteriores de tipicalidad y contribución. Para visualizar los efectos de esta propuesta, usamos los datos binarios de la música, y consideramos dos implicaciones, que mostramos en la Tabla 6. Se puede ver en la Tabla 6 que los ordenamientos que provoca esta propuesta coinciden con los de la propuesta 1, a pesar de variar los valores de las intensidades de implicación. Es presumible que haya datos para los que se encuentren diferencias entre las propuestas 1 y 2, aunque se supone serán muy ligeras, menores cuando el tamaño de la muestra aumente, pues los parámetros muestrales (media y varianza) serán más robustos frente al cambio de añadir o suprimir un individuo.

Tabla 6 – Aplicación de la propuesta 2 con 2 modalidades para dos implicaciones de la muestra: datos (izquierda), valores de intensidad de implicación al suprimir cada individuo posible (centro), y ordenamiento según el incremento provocado en la intensidad de implicación. Las intensidades reales, sin individuos añadidos, son $\varphi(HIP, \overline{OPE}) = 0.3970013$ y $\varphi(\text{HIP}, \overline{\text{MAQ}}) = 0.5536308$, respectivamente. Nótese que, a diferencia de la propuesta 1, el individuo de menor conformidad a la regla es aquel cuya supresión provoca el mayor valor del índice.

Efectivos		
HIP/OPE	0	1
0	6	3
1	9	2

→

Intensidad al suprimir 1 individuo		
HIP/OPE	0	1
0	0.3766556	0.4573310
1	0.4080088	0.3470280

→

Ordenamiento		
HIP/OPE	0	1
0	3	1
1	2	4

Efectivos		
HIP/MAQ	0	1
0	2	7
1	2	9

→

Intensidad al suprimir 1 individuo		
HIP/MAQ	0	1
0	0.4208651	0.5821959
1	0.6775072	0.5289169

→

Ordenamiento		
HIP/MAQ	0	1
0	4	2
1	1	3

Para visualizar los ordenamientos en el caso de más modalidades, usamos los datos simulados descritos en el apartado anterior, y mostramos los resultados en la Tabla 7. Aunque los valores de intensidad de implicación son distintos, los ordenamientos producidos resultan ser los mismos que en la propuesta 1, debido al gran tamaño de la muestra.

Tabla 7 – Aplicación de la propuesta 2 con 5 modalidades para una implicación de una muestra simulada con 1000 individuos y 5 modalidades por variable: valores de intensidad de implicación al suprimir cada individuo posible (izquierda), y ordenamiento según el incremento provocado en la intensidad de implicación. La intensidad real, sin individuos añadidos, es $\varphi(V1, \overline{V\,2}) = 0.3923378$. Los valores están redondeados por cuestión de espacio, pero no hay coincidencias (ejecutar código R de anexo para ver valores con más precisión). Nótese que, a diferencia de la propuesta 1, el individuo de menor conformidad a la regla es aquel cuya supresión provoca el mayor valor del índice.

Intensidad al suprimir un individuo tipo					
V1 /V2	0	0.25	0.5	0.75	1
0	0.384	0.388	0.392	0.396	0.401
0.25	0.388	0.390	0.392	0.394	0.396
0.5	0.392	0.392	0.392	0.392	0.392
0.75	0.397	0.394	0.392	0.390	0.388
1	0.401	0.397	0.392	0.388	0.384

→

Ordenamiento					
V1 /V2	0	0.25	0.5	0.75	1
0	24	21	14	5	2
0.25	20	18	12	8	6
0.5	9	10	11	16	17
0.75	3	7	13	19	23
1	1	4	15	22	25

En resumen, se puede concluir que, en el caso general, el ordenamiento de las parejas de modalidades no siempre es el mismo, y que dependa de los valores que tomen las variables en la muestra original, más cuanto menor sea el tamaño de la muestra.

4.3 Propuesta 3 : muestra auxiliar

Con una muestra auxiliar, que contiene un individuo, y sólo uno, de cada cruce de categorías, el ordenamiento es independiente de la muestra que cada uno esté analizando, al igual que ocurre con las propuestas de Bailleul y Gras (1994) y Ratsimba-Rajohn (1992). Esta muestra auxiliar, por simetría, representa una independencia estadística perfecta entre *a* y *b*, y por tanto una intensidad de implicación de 0.5. La supresión de cada individuo, provoca un desequilibrio en función de lo que aportaría dicho individuo a la implicación. Y el cálculo de ese incremento se puede interpretar como el apoyo o respeto a la regla. En la Tabla 8 se muestra la intensidad de implicación para la ausencia de cada individuo, y el ordenamiento inducido por esos cambios en la intensidad de implicación. Nótese cómo se escribe el ordenamiento en caso de "empate", pues es frecuente bajo esta propuesta.

Tabla 8 – Aplicación de la propuesta 3 con 2, 3 y 5 modalidades. Valores de intensidad de implicación al extraer de la muestra auxiliar cada individuo tipo, y ordenamiento según el incremento provocado. La intensidad con la muestra completa es 0.5.

Intensidad al suprimir un individuo tipo		
a / *b*	0	1
0	0.341	0.613
1	0.718	0.341

→

Ordenamiento		
a / *b*	0	1
0	3	2
1	1	3

Intensidad al suprimir un individuo tipo			
a / *b*	0	0.5	1
0	0.402	0.5	0.583
0.5	0.5	0.5	0.5
1	0.613	0.5	0.402

→

Ordenamiento			
a / *b*	0	0.5	1
0	8	3	2
0.5	3	3	3
1	1	3	8

Intensidad al suprimir un individuo tipo					
V1 /V2	0	0.25	0.5	0.75	1
0	0.442	0.472	0.5	0.527	0.554
0.25	0.471	0.485	0.5	0.513	0.527
0.5	0.5	0.5	0.5	0.5	0.5
0.75	0.529	0.514	0.5	0.485	0.472
1	0.560	0.529	0.5	0.471	0.442

→

Ordenamiento					
V1 /V2	0	0.25	0.5	0.75	1
0	24	20	9	5	2
0.25	22	18	9	8	5
0.5	9	9	9	9	9
0.75	3	7	9	18	20
1	1	3	9	22	24

En el caso de variables binarias, hay una diferencia de ordenamiento con las propuestas anteriores, pues el individuo con dato (0,0) es tan ventajoso de ser incluido en esta muestra auxiliar, como el individuo con datos (1,1). Lo es objetivamente, porque su introducción provoca el mismo incremento en la intensidad de la implicación.

5 Conclusiones

En el presente trabajo se ofrecen tres propuestas alternativas a las introducidas en Bailleul y Gras (1994) para medir el grado de adhesión de los individuos a las implicaciones entre variables modales, el cual sirve de base para obtener las tipicalidades y contribuciones de los individuos a las clases que se forman durante la construcción del árbol cohesitivo. Las propuestas vienen justificadas racionalmente, por el cambio que producen los individuos en el índice de la regla en cuestión. A partir de los ejemplos presentados se pudo evidenciar que corresponden a nuevas formas de ordenamiento de las parejas de modalidades, dependiendo de los valores que tomen las variables en los *n* individuos originales. Otro detalle importante, que no hemos entrado a valorar en esta propuesta, pero que vemos también influyente en una mejor clasificación de los individuos, es el de no usar un ordenamiento de posiciones, sino de los valores relativos. En el ordenamiento de Bailleul y Gras (1994), no hay "empates", y cada posición produce un salto fijo (de igual longitud) en el grado de adhesión a la regla. Con nuestra propuesta, el re-escalado desde las posiciones del ordenamiento al intervalo [0,1], se puede realizar como una homotecia, no desde los ordenamientos, sino desde los incrementos de intensidad de implicación calculados, de modo que dos individuos que producen el mismo incremento o muy similar, tendrían un mismo grado de adhesión, o muy similar, lo cual es más razonable que producir un salto de tamaño fijo de uno al siguiente. Una posible continuación de este trabajo sería la comparativa de grupos óptimos de individuos para la tipicalidad y contribución, entre la propuesta de Bailleul y Gras (1994) y las nuestras, usando muestras pequeñas, que se puedan interpretar de modo directo.

6 Agradecimientos

El segundo autor agradece el apoyo económico de los proyectos P1·1B2012-52 de la Universitat Jaume I de Castellón y MTM2013-43917-P del Ministerio de Economía y Competitividad de España.

Bibliografía

[1] 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, Université de Rennes 1.

[2] Bailleul, M. et R. Gras (1994), L'implication statistique entre variables modales, *Mathématiques et sciences humaines*, **128**, 41-57.

[3] Gras, R. (1979), *Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques*, Thèse de doctorat, Université de Rennes 1.

[4] Gras, R., S. Ag Almouloud, M. Bailleul, A. Larher, M. Polo, H. Ratsimba-Rajohn et A. Totohasina (1996), *L'implication statistique : nouvelle méthode exploratoire de données*, La Pensée Sauvage, Grenoble.

[5] Gras, R. et A. Larher (1992), L'implication statistique, une nouvelle méthode d'analyse de données, *Mathématiques et sciences humaines*, **120**, 5-31.

[6] 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*, **46**, 71-93.

[7] Lerman, I.C. (1970), Sur l'analyse des données préalable à une classification automatique (proposition d'une nouvelle mesure de similarité), *Mathématiques et sciences humaines*, **32**, 5-15.

[8] Lerman I.C., R. Gras R. et H. Rostam (1981), Elaboration et évaluation d'un indice d'implication pour des données binaires I, *Mathématiques et sciences humaines*, **74**, 5-35.

[9] Lerman I.C., R. Gras R. et H. Rostam (1981), Elaboration et évaluation d'un indice d'implication pour des données binaires II, *Mathématiques et sciences humaines*, **75**, 5-47.

[10] Pitarch, I. y Orús, P. (2002), *Estudio sobre la viabilidad y el interés didáctico del tratamiento de la información en la ESO*, Trabajo de Investigación, Tercer ciclo, Universidad Jaume I de Castellón.

[11] 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, Université de Rennes 1.

[12] Zamora L., J. Díaz y L. Portuondo (2015), Fundamental concepts on classification and statistical implicative analysis for modal variables, próxima publicación en *Revista Colombiana de Estadística*.

Anexo

6.1 Datos originales

La Tabla 9 contiene los datos originales sobre gustos musicales, extraídos de Pitarch y Orús (2002). Se lanza la pregunta *¿Te gusta la música…?* bajo 15 estilos musicales (hip-hop, ópera, máquina, jazz, flamenco, rumba, heavy, pop, clásica, dance, reggae, hardcore, rap, punk y ska), a 20 jóvenes de entre 15 y 20 años.

Tabla 9 – Datos de un estudio real sobre gustos musicales (Pitarch y Orús, 2002), que utilizamos para valorar nuestras propuestas sobre alternativas al grado de adhesión de los individuos a las implicaciones.

	HIP	OPE	MAQ	JAZ	FLA	RUM	HEA	POP	CLA	DAN	REG	HAR	RAP	PUN	SKA
al1	1	1	1	0	0	1	0	1	1	1	0	0	1	1	1
al2	1	0	1	1	0	1	1	1	0	1	1	1	0	1	1
al3	1	0	0	0	0	1	0	0	0	1	0	1	1	1	1
al4	0	1	0	1	1	1	1	0	1	0	1	0	1	0	1
al5	1	0	1	1	0	1	1	1	1	1	1	1	0	1	1
al6	0	0	1	0	1	1	0	1	0	1	0	1	1	0	0
al7	1	0	1	1	1	1	1	1	1	1	1	1	0	1	0
al8	0	1	1	1	1	1	0	1	0	1	0	1	1	0	1
al9	1	0	0	0	1	1	0	0	0	1	1	0	1	1	0
al10	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0
al11	0	0	1	1	1	1	1	1	0	1	1	1	1	1	1
al12	1	0	1	1	0	1	1	1	0	1	1	1	0	1	1
al13	1	0	1	0	0	1	1	1	0	1	0	1	0	1	1
al14	0	0	1	0	1	1	0	1	0	1	0	1	1	0	1
al15	0	1	1	1	1	1	0	1	0	1	0	1	1	0	1
al16	1	0	1	1	0	1	1	1	0	1	1	1	0	1	1
al17	1	0	1	0	1	1	0	1	0	1	1	1	1	0	1
al18	0	0	1	0	0	1	0	1	0	1	0	1	0	0	0
al19	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
al20	0	0	1	0	0	1	1	1	0	1	0	1	1	0	1

6.2 Códigos en lenguaje R para la computación

Código para calcular el ordenamiento de Bailleul y Gras (1994), que se muestra en la Tabla 2:

```
m = 5 # número de categorías
rg = array(0, dim=c(m,m), dimnames=list(categ, categ))
# preparamos matriz para poner rangos
for(i in 1:(m-1)) {
  for(j in 1:i) {
    a = as.character((m-j)/(m-1))
    b = as.character((i-j)/(m-1))
    rg[a,b] = i*(i-1)/2+j
  }
}
for(i in m:1) {
  for(j in 0:(i-1)) {
    a = as.character((m-i)/(m-1))
    b = as.character((m-i+j)/(m-1))
    rg[a,b] = m^2 - i*(i+1)/2 + j + 1
  }
}
```

```
rg
```

Código para calcular la intensidad de implicación de una pareja de variables modales:

```
phi = function(x) {
  # x : dos columnas de datos de variables modales escaladas
  # al intervalo [0,1]
  # devuelve phi con aproximación gaussiana
  n = dim(x)[1]
  mu.a = mean(x[,1])
  mu.nob = mean(1-x[,2])
  mu.anob = mean(x[,1]*(1-x[,2]))
  var.a = var(x[,1])*(n-1)/n
  var.b = var(x[,2])*(n-1)/n
  return( 1 - pnorm(q=mu.anob, mean=mu.a*mu.nob,
                    sd=sqrt((mu.a^2+var.a) *
                            (mu.nob^2+var.b)/n) )
        )
}
```

Código para generar los datos simulados, utilizados en las propuestas 1 y 2, cuyos resultados se muestran en las Tablas 5 y 7:

```
# semilla para reproducibilidad
set.seed(20151111) # fecha congreso ASI8
m=5 # número de modalidades
categ = (0:(m-1))/(m-1) # las modalidades 0, 1/(m-1),...,1
x = matrix(data=sample(x=categ, replace=TRUE, size=1000*2),
           ncol=2)
# x es la matriz de datos con 2 columnas y 1000 filas
```

Código para representar los ordenamientos de la Tabla 5 (propuesta 1), y que se pueden aplicar a cualquier conjunto de datos:

```
# propuesta 1: añadir individuo de cada tipo
# x contiene los datos los simulados anteriormente
# pero se puede poner cualquier muestra bivariante
phi0 = phi(x) # intensidad implicativa muestra de datos
m = 5 # número de modalidades de x (cotejar con x)
phi1 = matrix(data=0, ncol=m, nrow=m)
# phi1 es matriz que almacena la nueva intensidad implicativa
# al añadir un individuo de cada uno de los m^2 tipos
for(i in 1:m) { # para cada modalidad de V1
  for(j in 1:m) { # para cada modalidad de V2
    x2 = rbind(x, c(categ[i], categ[j])) # x con nuevo individuo
    phi1[i,j] = phi(x2)
    # calcula phi de la muestra con individuo añadido
  }
}
phi1 # matriz con nuevas intensidades de implicación
matrix(rank(phi1, ties="min"), ncol=m, dimnames=list(categ, categ))
# los ordenamientos resultantes de comparar
# (phi con individuo) – (phi sin individuo)
# phi1 – phi0
```

Código para representar los ordenamientos de la Tabla 7 (propuesta 2):

```
# propuesta 2: suprimir individuo de cada tipo
# x contiene los datos los simulados anteriormente
# pero se puede poner cualquier muestra bivariante
phi0 = phi(x) # intensidad implicativa muestra de datos
m = 5 # número de modalidades de x (cotejar con x)
phi2 = matrix(data=0, ncol=m, nrow=m)
# phi2 es matriz que almacena la nueva intensidad implicativa
# al suprimir un individuo de cada uno de los m^2 tipos
# si está presente en la muestra
n = dim(x)[1] # número de individuos
for(i in 1:m) { # para cada modalidad de V1
  for(j in 1:m) { # para cada modalidad de V2
    donde = (1:n)[categ[i]==x[,1] & categ[j]==x[,2]]
    # busca lugar(es) donde el individuo está en los datos
    existe = sum(donde) > 0
    # detecta si ese individuo existe en la muestra
    if(existe)  phi2[i,j] = phi((x[-donde[1], ]))
    else phi2[i,j] = NA
    # si existe, lo suprime y calcula phi
    # si no existe devuelve NA (no disponible)
  }
}
phi2 # matriz con nuevas intensidades de implicación
matrix(rank(-phi2, ties="min"), ncol=m,
       dimnames=list(categ, categ))
# los ordenamientos resultantes de comparar
# (phi con individuo) – (phi sin individuo)
# phi0 – phi2
```

Código para representar los ordenamientos de la Tabla 8 (propuesta 3):

```
# propuesta 3: muestra auxiliar con un individuo de cada tipo
# que se suprime y se compara el incremento en la intensidad
m=5 # número de modalidades
categ = (0:(m-1))/(m-1) # las modalidades 0, 1/(m-1),...,1
x = expand.grid(a=categ, b=categ) # crea muestra auxiliar completa
phi0 = phi(x) # phi de la muestra completa
phi3 = matrix(data=0, ncol=m, nrow=m)
# phi3 es matriz que almacena la nueva intensidad implicativa
# al suprimir un individuo de cada uno de los m^2 tipos
for(i in 1:m) {
  for(j in 1:m) {
    phi3[i,j] = phi(x[-(j+(i-1)*m),])
  }
}
phi3 # matriz con nuevas intensidades de implicación
matrix(rank(-phi3, ties="min"), ncol=m, byrow=TRUE,
       dimnames=list(categ, categ))
# los ordenamientos resultantes de comparar
# (phi con individuo) – (phi sin individuo)
# phi0 – phi3
```

VARIABILITY OF GRAS CLASSICAL IMPLICATION INDEX AND OTHER RULE QUALITY MEASURES UNDER SMALL SIZE SAMPLING FROM BIVARIATE BINARY PROCESSES

Pablo GREGORI[1] and Raphaël COUTURIER[2]

VARIABILITE DE L'INDICE D'IMPLICATION CLASSIQUE DE GRAS ET D'AUTRES MESURES DE QUALITE DE REGLES PAR ECHANTILLONAGE DE PROCESSUS BINAIRES BIVARIES DE PETITE TAILLE

ABSTRACT

Starting from a random bivariate Bernoulli process (A,B), we compute the variability in the (random) rule quality $A \rightarrow B$, measured by four common index (confidence, lift, Loevinger's H and classical Gras' φ). We also compute correlation coefficients among every couple of these measures. Computations are constrained to small size samples, and performed with R Statistical Software.

Keywords: *Statistical implicative analysis, association rules, confidence, lift, R language.*

RÉSUMÉ

A partir d'un processus aléatoire de Bernoulli bivarié (A,B), on calcule la variabilité dans la qualité de la règle (aléatoire) $A \rightarrow B$, mesuré par quatre indices habituels (confiance, lift, H de Loevinger et φ de Gras). On calcule aussi des coefficients de corrélation entre chaque couple de ces mesures. Les calculs sont restreints à des échantillons de petite taille, et ont été réalisés avec le logiciel statistique R.

Mots-clés : *Analyse statistique implicative, règles d'association, confiance, lift, langage R.*

1 Introduction

Association rules (AR) is a data mining technique initiated by Agrawal *et al.* (1993). It aims at turning data into knowledge, in the form of quasi-implications (or rules, i.e. sentences like « whenever a feature a is observed, then another feature b is usually observed »). Some of the most common statistical measures of the level of accomplishment of such rules are the *confidence*, *lift* and *support*, but there are dozens of such indicators (see, for instance, the review of Gras *et al.*, 2004, or Lalich *et al.*, 2007). It is a widely used technique, whose prototypical example and application field was the *market basket analysis*. The well known *apriori* algorithm (Agrawal *et al.*, 1993) and its subsequent improvements, together with the ease of interpretation of such measures, have been the keypoints for its spread use. The success of the algorithm (high speed in processing huge databases) is based upon the *a priori* rejection of candidate

[1] Departamento de Matemáticas. Instituto de Matemáticas y Aplicaciones de Castellón, Universitat Jaume I de Castellón, Campus Riu Sec E-12071, Spain, gregori@uji.es

[2] FEMTO-ST Institute, University of Bourgogne Franche-Comte, IUT Belfort-Montbéliard, Belfort, France, raphael.couturier@univ-fcomte.fr

rules of *low* support. That is the reason why this algorithm neglects rules involving a small fraction of the population, even when those rules might hide amazing and important results.

Statistical implicative analysis (SIA) was defined in order to help in the same kind of problems: measuring the quality of rules in the context of Didactics of Mathematics (Gras *et al.* 1996, 2013). The chosen statistical measure was the departure from independence between hypothesis and thesis, under some sampling scenarios. In this context, Loevinger's H index was already defined and also widely used in applications in Human Sciences, but for the author of SIA, it was not rich enough (see for instance Gras and Couturier, 2010).

Although SIA has been developing tools to deal with many types of data (Gras *et al.* 2008, 2013), we restrict ourselves to the study of the behavior of (the classical) Gras implication index for the simplest binary data: a bivariate Bernoulli process. We started this analysis in Gregori *et al.* (2013), providing an explicit computation of the probability distribution of the index, in terms of the parameters of the Bernoulli process. We also showed the effect of the conditional probability on the distribution of the expected value and quartiles of the implication index, fixing the other parameters and the sample size.

In this work, we relate the distribution of the Gras implication index to the distributions of other commonly used indexes, such as confidence, lift and Loevinger's H. We also stress the meaning of each index, in order to help researchers to better interpret them and choose the most suitable one for his or her particular work.

2 Definitions and notations

In SIA historical notation, at the very first definition of the implicative intensity, a set of variables is denoted by V and a couple of those variables are denoted by a and b. The set of sampled individuals is denoted by E, and each of those individuals is said to « hold property a » (or to be an example for property a) if a is « true » for the individual, and it is said « not to hold property a » (or to be a counterexample for property a) if a is « false » for the individual. Using logical operators (complementary, AND) we can also consider properties $\bar{b}$ and $a \wedge \bar{b}$ and check whether individuals hold those properties as true or false.

Afterwards, letters A and B denote, respectively, the subsets of individuals holding true respectively variable a and b, and finally, X and Y denote random sets of individuals, modelling the respective sets A and B, under some random sampling schemes. The surprisingness of the implication (or rule) $a \rightarrow b$ is then defined comparing the sample data against the statistical independence assumption.

In this paper, we do not remind the theory of SIA, and we use A and B to denote a couple of Bernoulli random variables that shall generate, at each instance, and for a given sample size $N = n$, a particular sample of n individuals for variables a and b, with the usual identification 0 (for false) and 1 (for true).

Example: Let (A,B) be a couple of Bernoulli variables, with $P(A=1)=P(B=1)=0.5$ and $P(B=1|A=1)=0.75$, and let $N=10$. Let us run a simulation of (A,B) and show it in Table 1.

Table 1 – A particular instance of (A,B) for $P(A=1)=P(B=1)=0.5$ and $P(B=1|A=1)=0.75$, and sample size $N=10$, seen as raw data (left), contingency table (middle), relative frequency table (right).

a	0	1	0	1	0	1	1	1	1	0
b	0	1	1	1	0	1	1	1	1	0

a / b	0	1
0	3	1
1	0	6

a / b	0	1
0	0.3	0.1
1	0	0.6

The notations for the contingency table and the relative frequency table are shown in Table 2.

Table 2 – Notations for contingency (left) and relative frequency tables (right).

a / b	0	1	Margin a
0	$n_{\bar{a}\bar{b}}$	$n_{\bar{a}b}$	$n_{\bar{a}}$
1	$n_{a\bar{b}}$	n_{ab}	n_a
Margin b	$n_{\bar{b}}$	n_b	n

a / b	0	1	Margin a
0	$f_{\bar{a}\bar{b}}$	$f_{\bar{a}b}$	$f_{\bar{a}}$
1	$f_{a\bar{b}}$	f_{ab}	f_a
Margin b	$f_{\bar{b}}$	f_b	1.00

Definition: For a given joint frequency table for (a,b), the quality index for rule $a \rightarrow b$ that we consider are defined as:

- Confidence : $C = \dfrac{f_{ab}}{f_a}$.
- Lift : $L = \dfrac{C}{f_b} = \dfrac{f_{ab}}{f_a \times f_b}$.
- Loevinger's H : $H = 1 - \dfrac{f_{a\bar{b}}}{f_a \times f_{\bar{b}}}$.
- Gras' φ : $\varphi = 1 - F\left(n_{a\bar{b}}\right)$, where F is the cumulative distribution function of the binomial model of size n and success probability $\dfrac{n_a \times n_{\bar{b}}}{n^2}$, in one of its possible modelisations, the one we choose for this contribution.

After the realisation (a,b) of (A,B) in the example, the quality of the sampled rule $a \rightarrow b$ shows the index values :

- $C = \dfrac{0.6}{0.6} = 1$, $L = \dfrac{0.6}{0.6 \times 0.6} = 1.4285714$, $H = 1 - \dfrac{0}{0.6 \times 0.3} = 1$.
- $\varphi = 1 - \mathrm{F}_{\mathrm{Bin}\left(10, \frac{6 \times 3}{10^2}\right)}(0) = 1 - 0.137448 = 0.862552$.

Obviously, the sampled realisation (a,b) is only one, and we want to show the « whole » picture for (A,B). For that purpose, we need intensive computation in order to

write all possible simulations, weighted by their probabilities, and compute all indexes. We shall analyse the variability in the rule quality under each index, and also the joint variability under each couple of indexes when the underlying process is known (by the given parameters).

Then, for (A,B) a bivariate Bernouilli random variable, determined by parameters $p_A := P(A=1)$, $p_B := P(B=1)$ and $p_{AB} := P(A=1,B=1)$ (or equivalently $p_{B|A} := P(B=1 \mid A=1) = p_{AB} / p_A$, whenever $p_A > 0$), we can represent it in the form of a joint probability table as in Table 3.

Table 3 – Notation for the parameters of a bivariate Bernoulli random variable (A,B).

A / B	0	1	Margin A
0	$p_{\bar{A}\bar{B}}$	$p_{\bar{A}B}$	$p_{\bar{A}}$
1	$p_{A\bar{B}}$	p_{AB}	p_A
Margin B	$p_{\bar{B}}$	p_B	1.00

We shall consider only simple random sampling (SRS) from (A,B) with « fixed » sample size n, and our goal is to show probabilistic properties of some quality (or interestingness) measures of the implication $A \rightarrow B$. To this end, and arising from the sampling process, we consider the random contingency table shown in Table 4, where all absolute frequencies (cases) are integer random variables.

Table 4 – Notation for the random contingency table arising from a SRS of size N from the bivariate Bernoulli variable (A,B).

A / B	0	1	Margin A
0	$N_{\bar{a}\bar{b}}$	$N_{\bar{a}b}$	$N_{\bar{a}}$
1	$N_{a\bar{b}}$	N_{ab}	N_a
Margin B	$N_{\bar{b}}$	N_b	N

Under our « fixed » sample size framework, N is not random, because $N = n$. Now we study the probability laws and relationships among the following *random* quality measures (seen as functions of random variables A and B):

– Confidence $C := \dfrac{N_{ab}/N}{N_a/N} = \dfrac{N_{ab}}{N_a} = 1 - \dfrac{N_{a\bar{b}}}{N_a}$.

– Lift $L := \dfrac{C}{N_b/N} = \dfrac{1 - N_{a\bar{b}}/N_a}{1 - N_{\bar{b}}/N}$.

– Loevinger's $H := 1 - \dfrac{N_{a\bar{b}}/N}{(N_a/N)(N_{\bar{b}}/N)} = 1 - \dfrac{N_{a\bar{b}}/N_a}{N_{\bar{b}}/N}$.

– Gras' $\varphi := 1 - F(N_{a\bar{b}})$, where F means the cumulative distribution function (cdf) of the binomial model $Bin\left(N, \frac{N_a \times N_{\bar{b}}}{N^2}\right)$ (for this « fixed sample size » case).

We have intentionally displayed the formulae in common terms (N, N_a, $N_{\bar{b}}$ and $N_{a\bar{b}}$) in order to better compare their analytical expressions. The implementation of these formulae in R language can be seen in the Appendix.

One can see that confidence is ill defined when $N_a = 0$, and so is lift when $N_a \times N_b = 0$, and Loevinger's H whenever $N_a \times N_{\bar{b}} = 0$. Gras' φ is well defined in any case, and then we shall exclude the problematic cases of the other indexes from computations.

As it was shown in the example before, and in usual practice, all these (random) quality measures are actually seen as statistics, computed from an observed sample, and they seize the quality of the instance rule $a \to b$ in different manners. It is clearly detailed in Vaillant *et al.* (2008) and we just stress it here for convenience:

– *Confidence* C stands for estimating the predictive quality of the rule (if property A is held by an individual of the sample, then an approximation to the probability for that individual to hold property B is C). The range of confidence lies within $[0,1]$. High values of confidence are attributed to a strong degree of association. However, for example, if A and B are any two statistically independent binary random variables, the law of large numbers suggests that C will be likely close to $p_{b|a} = p_b$. If one chooses B to be abundant (with p_b close to 1.0), then many samples of (A,B) will show large confidence values, even when $A = 1$ does not push the chances of $B = 1$.

– *Lift* L stands for measuring the improvement of predicting B once the rule is taken into account. For example, a value of 3.0 means that the probability of an individual to hold property B raises by a factor of 3.0 at the moment we know that the individual holds property A (i.e. $p_{b|a} = 3p_b$). In this case $A = 1$ would « lift » the chances of $B = 1$. Statistical independence corresponds to a lift of 1.0. The range of lift is within $[0, N]$.

– *Loevinger's* H measures a distance of the relative number of counter examples from independence. Statistical independence corresponds to a value of 0. A value of 1.0 means no counterexamples at all (hence the strongest association from A to B). The range of H lies within $[1 - N, 1]$.

– *Gras'* φ measures how extraordinarily small the number of counterexamples to the rule is, given the sampled amounts of $A = 1$ and $B = 1$, compared to what are expected under independence of A and B. The range of φ lies within $[0,1]$. Its purpose is to discover association even if only a small number of individuals is involved. In practice, users should pay special attention only to rules with indexes over 0.90 or 0.95, because they are more clearly away from independence. Rules with lower values (such as 0.80 and below) must be

analysed with care. We have found that many practitioners of SIA tend to interpret φ in the same way as C, and our intention is to show clearly that those indexes measure different aspects: not only by definition, but by computing their distributions and correlations. We stress that, *if the goal is to discover rules which cannot be explained by randomness, then confidence is not the tool, but* φ. Once the rule is targeted, even if the confidence were to be low, the association is very likely to exist. *If one seeks for predictive ability, then the confidence is the right tool*.

Finally, we shall mention that statistics C, L and H do not vary under the scaling of the contingency table. However φ does. Besides, a drawback of φ was highlighted in Bodin (1996): with very large sample sizes, it was given an extremely high value (0.9999) to a rule with a confidence as poor as 33%. Gras redefined φ (dubbed *entropic version*) in order to avoid such behavior. However, that behavior is not necessarily a sign of weakness: Gras index and confidence are measuring different effects indeed; 66% of counterexamples are unacceptable for the task of prediction, but it is a ridiculous fraction of counterexamples, in the eyes of probability, when checking for independence.

3 Probability distributions of absolute frequencies

Under the SRS with fixed sample size, N is not random, but takes the fixed value n. The 4-variate $\left(N_{\bar{a}\bar{b}}, N_{\bar{a}b}, N_{a\bar{b}}, N_{ab}\right)$ follows the multinomial distribution $Multinom\left(n; p_{\bar{A}\bar{B}}, p_{\bar{A}B}, p_{A\bar{B}}, p_{AB}\right)$. Recall that the probability function (pf) of a general $X \sim Multinom\left(n; p_1, p_2, \ldots, p_k\right)$ is

$$f_n(x_1, x_2, \ldots, x_k) = \begin{cases} \dfrac{n!}{x_1! x_2! \quad x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}, & \text{if } x_1 + x_2 + \cdots + x_k = n \\ 0, & \text{otherwise} \end{cases}$$

The distributions of the two bivariate margins are $\left(N_{\bar{a}}, N_a\right) \sim \text{Multinom}\left(n; p_{\bar{A}}, p_A\right)$ and $\left(N_{\bar{b}}, N_b\right) \sim Multinom\left(n; p_{\bar{B}}, p_B\right)$.

Although we have not implemented the SRS with a random sample size for computations, let us sketch up the theoretical distribution function, as a 2-step random procedure:

1. Sampling from $N \sim Po(\mu)$, for a given parameter μ, and
2. conditionned to $N = n$, sampling from $\left(N_{\bar{a}\bar{b}}, N_{\bar{a}b}, N_{a\bar{b}}, N_{ab}\right) \sim Multinom\left(n; p_{\bar{A}\bar{B}}, p_{\bar{A}B}, p_{A\bar{B}}, p_{AB}\right)$.

In this case, N is random, and it is easy to deduce that the (unconditional) pf of the 4-variate $\left(N_{\bar{a}\bar{b}}, N_{\bar{a}b}, N_{a\bar{b}}, N_{ab}\right)$ is:

$$
\begin{aligned}
&f\left(n_{\bar{a}\bar{b}}, n_{\bar{a}b}, n_{a\bar{b}}, n_{ab}\right) = \sum_{n=0}^{\infty} P(N=n)\, f_n\left(n_{\bar{a}\bar{b}}, n_{\bar{a}b}, n_{a\bar{b}}, n_{ab}\right) \\
&= P(N = n_{\bar{a}\bar{b}} + n_{\bar{a}b} + n_{a\bar{b}} + n_{ab})\, f_n\left(n_{\bar{a}\bar{b}}, n_{\bar{a}b}, n_{a\bar{b}}, n_{ab}\right) \\
&= e^{-\mu}\, \frac{\mu^{n_{\bar{a}\bar{b}}+n_{\bar{a}b}+n_{a\bar{b}}+n_{ab}}}{(n_{\bar{a}\bar{b}} + n_{\bar{a}b} + n_{a\bar{b}} + n_{ab})!} \cdot \frac{(n_{\bar{a}\bar{b}} + n_{\bar{a}b} + n_{a\bar{b}} + n_{ab})!}{n_{\bar{a}\bar{b}}!\, n_{\bar{a}b}!\, n_{a\bar{b}}!\, n_{ab}!}\, p_{\bar{A}\bar{B}}{}^{n_{\bar{a}\bar{b}}}\, p_{\bar{A}B}{}^{n_{\bar{a}b}}\, p_{A\bar{B}}{}^{n_{a\bar{b}}}\, p_{AB}{}^{n_{ab}} \\
&= \left(e^{-(\mu p_{\bar{A}\bar{B}})}\, \frac{(\mu p_{\bar{A}\bar{B}})^{n_{\bar{a}\bar{b}}}}{n_{\bar{a}\bar{b}}!}\right) \cdot \left(e^{-(\mu p_{\bar{A}B})}\, \frac{(\mu p_{\bar{A}B})^{n_{\bar{a}b}}}{n_{\bar{a}b}!}\right) \cdot \left(e^{-(\mu p_{A\bar{B}})}\, \frac{(\mu p_{A\bar{B}})^{n_{a\bar{b}}}}{n_{a\bar{b}}!}\right) \cdot \left(e^{-(\mu p_{AB})}\, \frac{(\mu p_{AB})^{n_{ab}}}{n_{ab}!}\right) \\
&= f_{Po(\mu p_{\bar{A}\bar{B}})}(n_{\bar{a}\bar{b}}) \cdot f_{Po(\mu p_{\bar{A}B})}(n_{\bar{a}b}) \cdot f_{Po(\mu p_{A\bar{B}})}(n_{a\bar{b}}) \cdot f_{Po(\mu p_{AB})}(n_{ab})
\end{aligned}
$$

It can be seen that all 4 random variables are mutually independent and have marginal Poisson distributions: $N_{\bar{a}\bar{b}} \sim Po(\mu p_{\bar{A}\bar{B}})$, $N_{\bar{a}b} \sim Po(\mu p_{\bar{A}B})$, $N_{a\bar{b}} \sim Po(\mu p_{A\bar{B}})$ and $N_{ab} \sim Po(\mu p_{AB})$.

For the two bivariate margins, one can proceed similarly, or by seeing them as sums of independent Poisson distributions. Then: $N_a \sim Po(\mu p_A)$ and $N_{\bar{a}} \sim Po(\mu p_{\bar{A}})$ are mutually independent, and so $N_b \sim Po(\mu p_{\bar{B}})$ and $N_{\bar{b}} \sim Po(\mu p_{\bar{B}})$ are. Be careful that N_a and N_b are not independent since they share a common term in the sum.

The pdf of the full 5-variate $(N_{\bar{a}\bar{b}}, N_{\bar{a}b}, N_{a\bar{b}}, N_{ab}, N)$ can be simplified to:

$$
f\left(n_{\bar{a}\bar{b}}, n_{\bar{a}b}, n_{a\bar{b}}, n_{ab}, n\right) = \begin{cases} e^{2\mu}\, \dfrac{\mu^n}{n!}\, \dfrac{\mu_{\bar{a}\bar{b}}{}^{n_{\bar{a}\bar{b}}}}{n_{\bar{a}\bar{b}}!}\, \dfrac{\mu_{\bar{a}b}{}^{n_{\bar{a}b}}}{n_{\bar{a}b}!}\, \dfrac{\mu_{a\bar{b}}{}^{n_{a\bar{b}}}}{n_{a\bar{b}}!}\, \dfrac{\mu_{ab}{}^{n_{ab}}}{n_{ab}!}, & \text{if } n_{\bar{a}\bar{b}} + n_{\bar{a}b} + n_{a\bar{b}} + n_{ab} = n \\ 0, & \text{otherwise} \end{cases}
$$

where $\mu_{\bar{a}\bar{b}} = \mu p_{\bar{A}\bar{B}}$, $\mu_{\bar{a}b} = \mu p_{\bar{A}B}$, $\mu_{a\bar{b}} = \mu p_{A\bar{B}}$ and $\mu_{ab} = \mu p_{AB}$.

4 Probability distributions of the quality measures

The probability distributions of the quality measures shall be computed directly from the 4-variate joint absolute frequencies. Exact analytical expressions are not available, then we shall show particular examples and restrict our computations to moderate values of n, since partitions of n in all variations of 4 integers are computationally expensive for large n. We are using the `partitions` package of R Statistical Software (Hankin, 2006).

The laws of C, L, H and φ can be computed as usual in transformation of discrete random variables:

– $P(C=c)=\sum_{c=\frac{n_{ab}}{n_a}}\left(N_{\bar{a}\bar{b}}=n_{\bar{a}\bar{b}},N_{\bar{a}b}=n_{\bar{a}b},N_{a\bar{b}}=n_{a\bar{b}},N_{ab}=n_{ab}\right)$, where the sum spans over all partitions of n into four non negative integers $\left(n_{\bar{a}\bar{b}},n_{\bar{a}b},n_{a\bar{b}},n_{ab}\right)$, and c attains the value $\frac{n_{ab}}{n_a}$, with $n_a := n_{a\bar{b}}+n_{ab}$.

– $P(L=l)=\sum_{l==\frac{nn_{ab}}{n_a n_b}}\left(N_{\bar{a}\bar{b}}=n_{\bar{a}\bar{b}},N_{\bar{a}b}=n_{\bar{a}b},N_{a\bar{b}}=n_{a\bar{b}},N_{ab}=n_{ab}\right)$, where the sum spans over all partitions of n into four non negative integers $\left(n_{\bar{a}\bar{b}},n_{\bar{a}b},n_{a\bar{b}},n_{ab}\right)$, and l attains the value $\frac{nn_{ab}}{n_a n_b}$, with $n_a := n_{a\bar{b}}+n_{ab}$ and $n_b := n_{\bar{a}b}+n_{ab}$.

– $P(H=h)=\sum_{h=1-\frac{nn_{a\bar{b}}}{n_a n_{\bar{b}}}}\left(N_{\bar{a}\bar{b}}=n_{\bar{a}\bar{b}},N_{\bar{a}b}=n_{\bar{a}b},N_{a\bar{b}}=n_{a\bar{b}},N_{ab}=n_{ab}\right)$, where the sum spans over all partitions of n into four non negative integers $\left(n_{\bar{a}\bar{b}},n_{\bar{a}b},n_{a\bar{b}},n_{ab}\right)$, and h attains the value $1-\frac{nn_{a\bar{b}}}{n_a n_{\bar{b}}}$, with $n_a := n_{a\bar{b}}+n_{ab}$ and $n_{\bar{b}} := n_{a\bar{b}}+n_{\bar{a}\bar{b}}$.

– $P(\varphi=f)=\sum_{f=1-F\left(n_{a\bar{b}}\right)}\left(N_{\bar{a}\bar{b}}=n_{\bar{a}\bar{b}},N_{\bar{a}b}=n_{\bar{a}b},N_{a\bar{b}}=n_{a\bar{b}},N_{ab}=n_{ab}\right)$, where the sum spans over all partitions of n into four non negative integers $\left(n_{\bar{a}\bar{b}},n_{\bar{a}b},n_{a\bar{b}},n_{ab}\right)$, f attains the value of the complement to one, to the cdf of the binomial model of parameters n and $n_a n_{\bar{b}}/n^2$, with $n_a := n_{a\bar{b}}+n_{ab}$ and $n_{\bar{b}} := n_{a\bar{b}}+n_{\bar{a}\bar{b}}$.

We show in Table 5 the way we have proceeded with a toy example: (1) a list of all compositions of n is created, using the R package `partitions`, (2) the probability of each composition is computed, using the R function `dmultinom()`, and (3) the computation of the four index for each composition.

In order to obtain the probability distribution of each index, seen as a random variable, one has to check the repeated values of the index under different compositions, and sum up all their respective probabilities. As one can see in the previous table, all index but Gras, have ill definition (NA value) with positive probability.

We show results in Figure 1, in the form of plots, for specific parameter values, but the scripts are provided, so that the interested researcher can modify and get computation and plots for other parameter values.

Table 5: A list of some compositions of $n = 5$, seen as the four joint frequency tables of the contingency table, together with their probability, under $p_{\bar{A}\bar{B}} = p_{\bar{A}B} = p_{A\bar{B}} = p_{AB} = 0.25$, and the computed values of the quality index. Ill definition occurs at some compositions for all index except Gras' φ.

$N_{\bar{a}\bar{b}}$	$N_{\bar{a}b}$	$N_{a\bar{b}}$	N_{ab}	Prob	C	L	H	φ
5	0	0	0	0.0009766	NA	NA	NA	0.0000000
4	1	0	0	0.0048828	NA	NA	NA	0.0000000
3	2	0	0	0.0097656	NA	NA	NA	0.0000000
2	3	0	0	0.0097656	NA	NA	NA	0.0000000
1	4	0	0	0.0048828	NA	NA	NA	0.0000000
0	5	0	0	0.0009766	NA	NA	NA	0.0000000
4	0	1	0	0.0048828	0.0	NA	0.0000000	0.2627200
3	1	1	0	0.0195312	0.0	0.0000000	-0.2500000	0.1834910
2	2	1	0	0.0292969	0.0	0.0000000	-0.6666667	0.1124509
1	3	1	0	0.0195312	0.0	0.0000000	-1.5000000	0.0543613
0	4	1	0	0.0048828	0.0	0.0000000	-4.0000000	0.0147580
3	0	2	0	0.0097656	0.0	NA	0.0000000	0.3174400
2	1	2	0	0.0292969	0.0	0.0000000	-0.2500000	0.1905263
1	2	2	0	0.0292969	0.0	0.0000000	-0.6666667	0.0932512
0	3	2	0	0.0097656	0.0	0.0000000	-1.5000000	0.0317587
2	0	3	0	0.0097656	0.0	NA	0.0000000	0.3369600
1	1	3	0	0.0195312	0.0	0.0000000	-0.2500000	0.1634992
0	2	3	0	0.0097656	0.0	0.0000000	-0.6666667	0.0597943
1	0	4	0	0.0048828	0.0	NA	0.0000000	0.3276800
0	1	4	0	0.0048828	0.0	0.0000000	-0.2500000	0.1073742
0	0	5	0	0.0009766	0.0	NA	0.0000000	0.0000000
4	0	0	1	0.0048828	1.0	5.0000000	1.0000000	0.5817881
3	1	0	1	0.0195312	1.0	2.5000000	1.0000000	0.4722681
2	2	0	1	0.0292969	1.0	1.6666667	1.0000000	0.3409185
1	3	0	1	0.0195312	1.0	1.2500000	1.0000000	0.1846273
0	4	0	1	0.0048828	1.0	1.0000000	NA	0.0000000
3	0	1	1	0.0195312	0.5	2.5000000	0.3750000	0.5125046
2	1	1	1	0.0585937	0.5	1.2500000	0.1666667	0.3461014
1	2	1	1	0.0585937	0.5	0.8333333	-0.2500000	0.1834910
0	3	1	1	0.0195312	0.5	0.6250000	-1.5000000	0.0543613

As one can see, all distributions attain many different values (all the fractions of the joint frequencies for all the possible compositions). All these probability functions are not piecewise monotone. This is an inconvenience if one wants to write sets of most likely index values. For instance, let us observe Figure 1, first row, second column (i.e., the confidence for $p_{B|A} = 0.5$). The most likely value of C is 0.5, and the second most likely value are 0.3333333 and 0.6666667. However, lots of values in the interval $[0.3333333,0.6666667]$ are very unlikely to happen. What to say, then, about that

interval? We can say that the probability of C to lie on that interval is 0.824941, but it contains very likely values, yet very unlikely ones too.

Another point is variance. Lift and H shows the largest variances in all cases, which is a negative point. The reason might be the range of these index, which is not bound to $[0,1]$. φ has a larger variance than C for $p_{B|A} = 0.5$ but, as mentioned before, and as we shall stress afterwards, confidence is not the *alter ego* of Gras' φ.

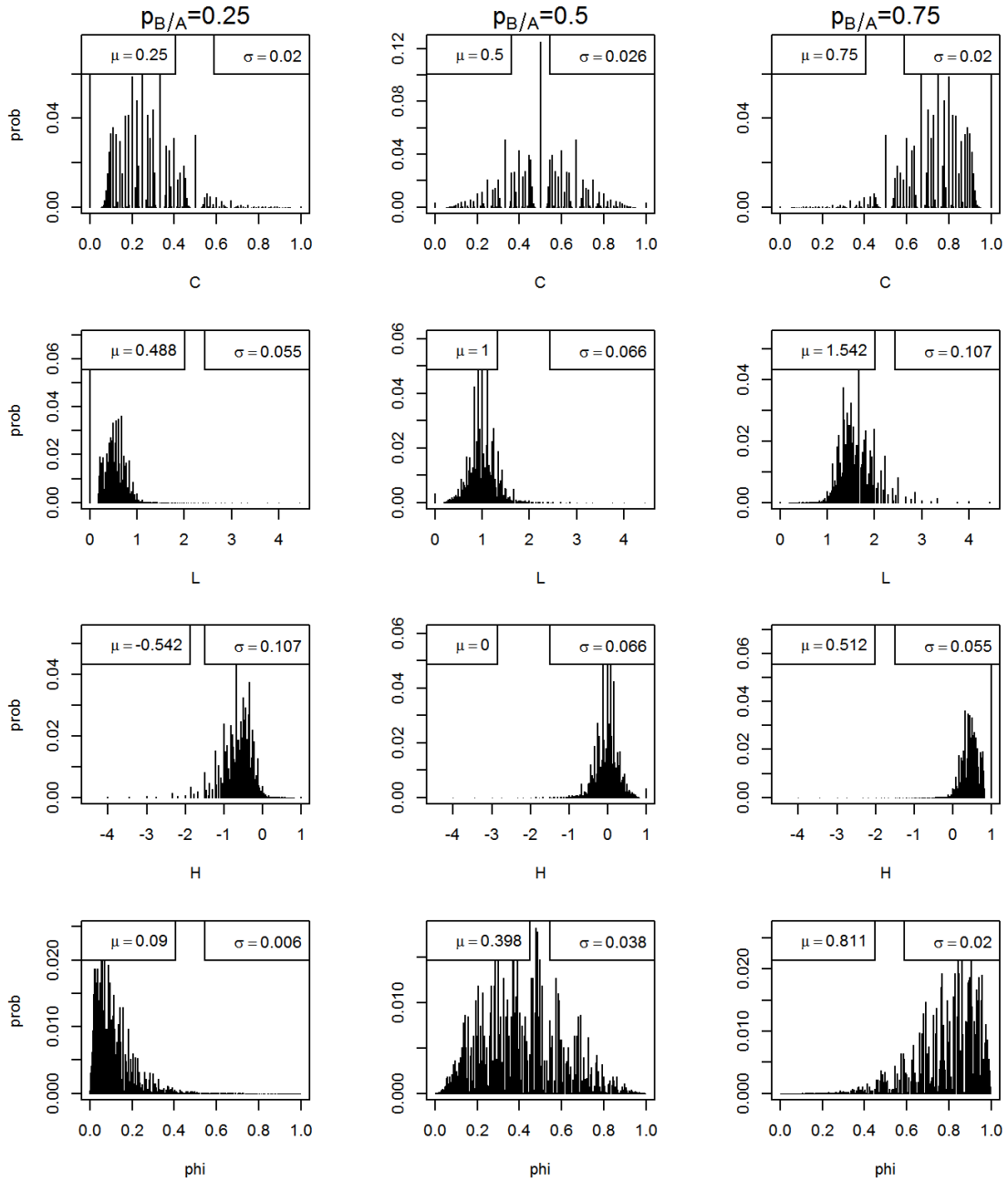


Figure 1 – Probability distribution, expectation and standard deviation of all the quality index for $n = 30$, $p_A = 0.5$, $p_B = 0.5$ and $p_{B|A} = 0.25,0.5,0.75$ (from left to right). Ranges of lift and Loevinger's H have been truncated for convenience.

5 Bivariate distributions : plots and correlations

A very interesting approach for the analysis of the relation among several different implication index was presented in Vaillant *et al.* (2008). In that work, a couple of large datasets served as a quarry of some thousands of rules, and several index were compared by plotting them in pairs against the quarry of rules. Some of those plots served as a motivation for the definition of new index.

In our setting, the quarry of rules is the complete bivariate random variable (A,B), with its specific parameters and the fixed sample size. And the goal is to understand the joint behavior of the considered quality measures.

To begin with, let us plot each couple's joint probability distribution, as in Figure 2. It is not complete, since the joint probabilities are not reflected in the plot, but any attempt to jump to 3D plots (with bars showing the joint probabilities) is not satisfactory neither for reflecting the joint distributions.

Another way to understand the relationship among index is to compute correlation coefficients (Pearson's r, Spearman's ρ and Kendall's τ). With these computations, we can see how similarly, or differently, the considered index are measuring the quality of the rules.

In general, such correlation coefficients are computed for two-variate samples of the type

$$\begin{array}{cccc} x_1 & x_2 & \dots & x_n \\ \hline y_1 & y_2 & \cdots & y_n \end{array}$$

by:

- Pearson's $r_{xy} := \dfrac{(1/n)\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{(1/n)\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\sqrt{(1/n)\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}}$ (linear correlation coefficient).
- Spearman's $\rho_{xy} := r_{zw}$, where z_i and w_i are the ranks of x_i and y_i respectively (i.e., both of them have values, from 1 to n, indicating the rank of original data, when sorted increasingly). Tied ranks are averaged.
- Kendall's $\tau_{xy} := \dfrac{\#\,concordant\ pairs - \#\,discordant\ pairs}{\#\ pairs}$, where concordant pairs are couples $\{(x_i,y_i),(x_j,y_j)\}$ such that $x_i < x_j$ and $y_i < y_j$, or conversely $x_i > x_j$ and $y_i > y_j$. Discordant pairs are those couples such that $x_i < x_j$ and $y_i > y_j$, or conversely $x_i > x_j$ and $y_i < y_j$. In case of ties ($x_i = x_j$ or $y_i = x_j$), those pairs are not concordant neither discordant, and some authors prefer to use modified versions of Kendall's τ (named τ_b and τ_c).

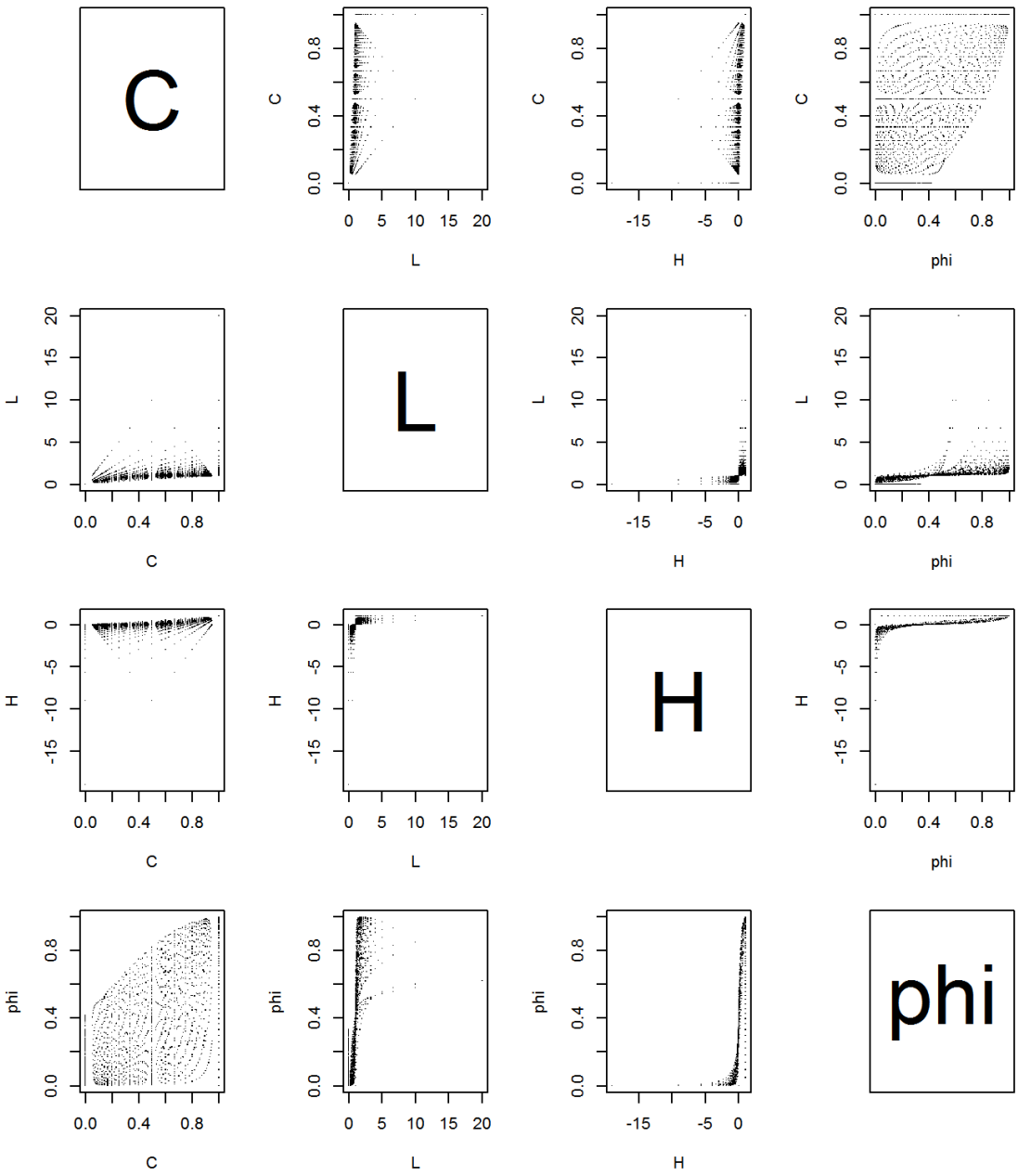


Figure 2 – Plot of joint values of all pairs of quality index, for samples of size $n = 20$, when parameters are $p_A = 0.5$, $p_B = 0.5$ and $p_{B|A} = 0.25$. Some couples of values have a very low probability and some other couples have a rather high probability, but it can hardly be reflected on the plots.

We provide the code to compute correlation coefficients among the random variables C, L, H and φ. Parameters p_A, p_B and $p_{B|A}$ of the binary process, and sample size n, need to be fixed beforehand. We display in Figures 3-5 a few cases, and invite the interested reader to use the code shared in Appendix for computations of any desired case (with moderate values of n since simulations need large memory and are time consuming).

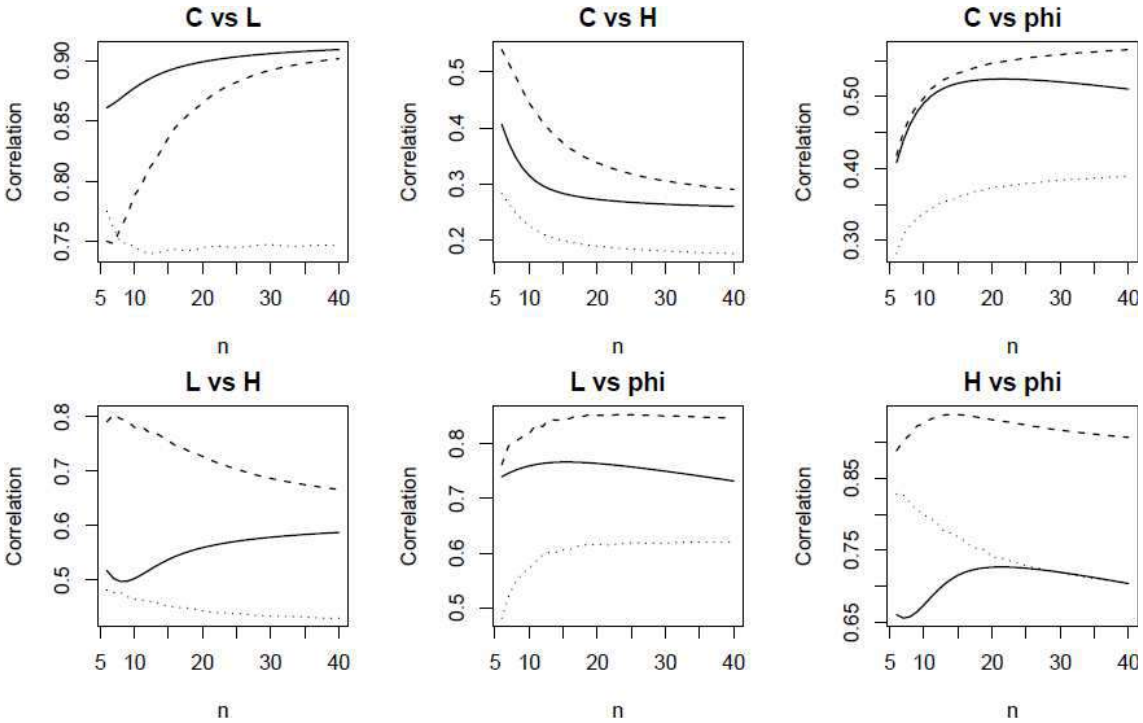


Figure 3 – Evolution, along the sample size, of the Pearson's (continuous line), Spearman's (dashed line) and Kendall's (dotted line) correlation coefficients among all pairs of quality index of the rule $A \rightarrow B$ where $p_A = 0.5$ and $p_B = 0.5$, and conditional probability $p_{B|A} = 0.25$.

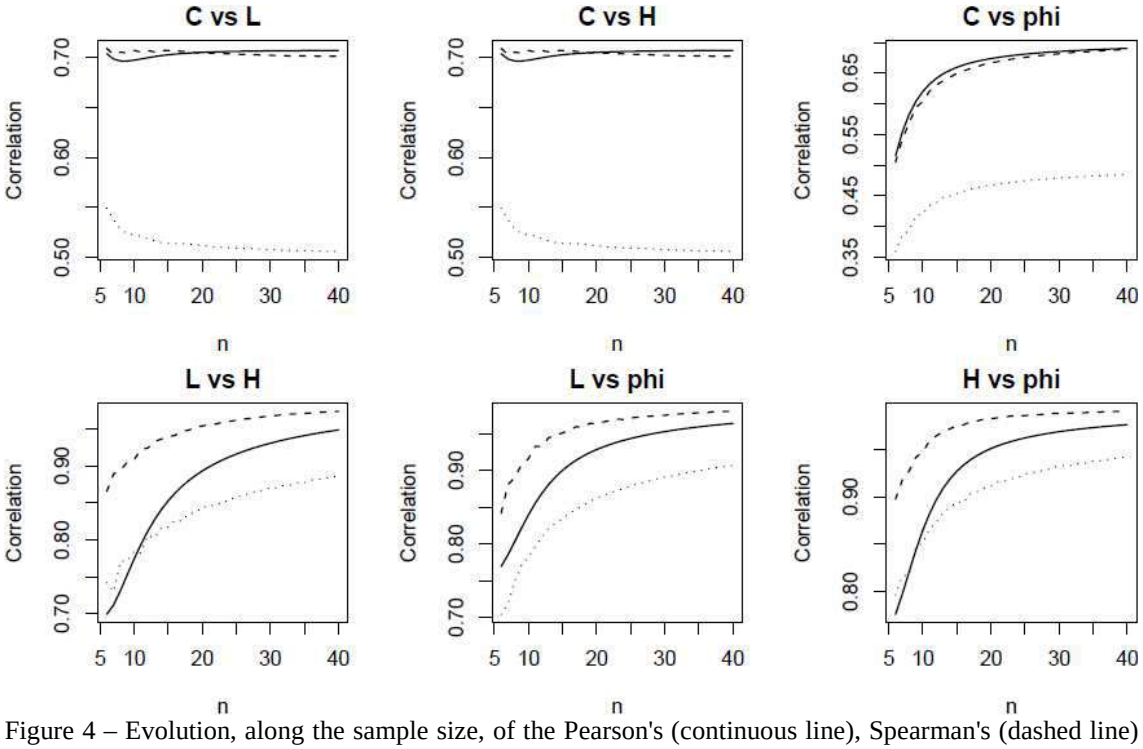


Figure 4 – Evolution, along the sample size, of the Pearson's (continuous line), Spearman's (dashed line) and Kendall's (dotted line) correlation coefficients among all pairs of quality index of the rule $A \rightarrow B$ where $p_A = 0.5$ and $p_B = 0.5$, and conditional probability $p_{B|A} = 0.5$.

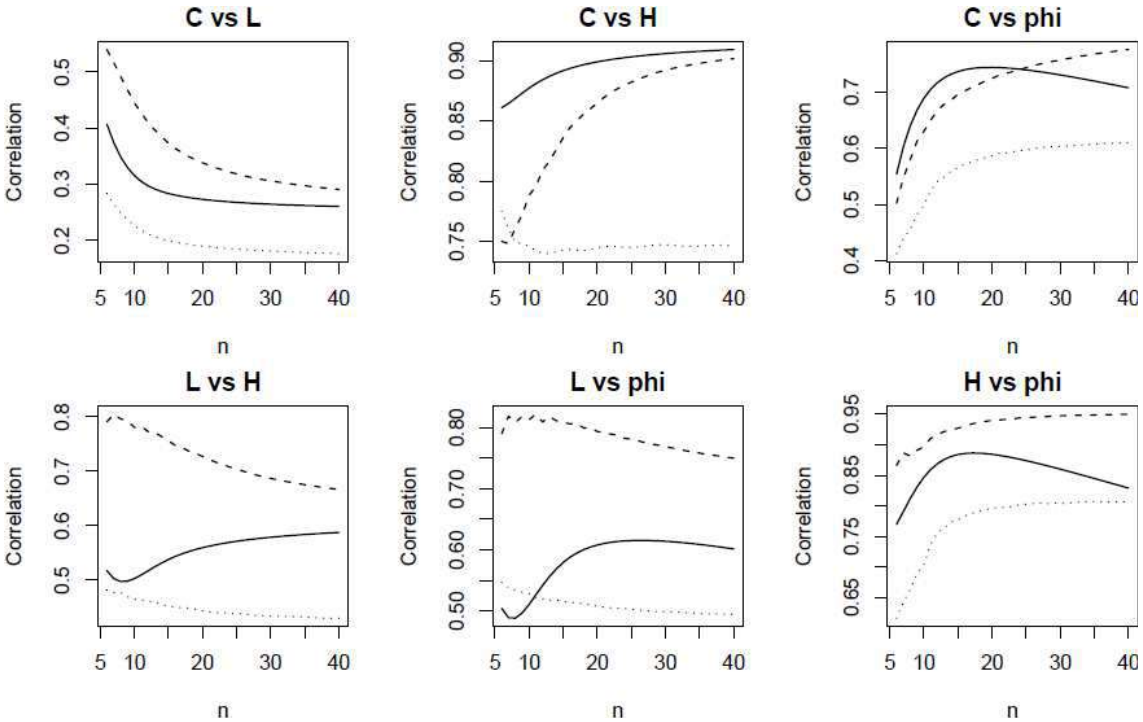


Figure 5 – Evolution, along the sample size, of the Pearson's (continuous line), Spearman's (dashed line) and Kendall's (dotted line) correlation coefficients among all pairs of quality index of the rule $A \rightarrow B$ where $p_A = 0.5$ and $p_B = 0.5$, and conditional probability $p_{B|A} = 0.75$.

Figures 3-5 show, for a range of values of the sample size, the correlation coefficients among all pairs of quality index measured on a bivariate Bernoulli process (A,B). We have used 3 different scenarios, tuned by the conditional probability $p_{B|A}$: low (0.25), average (0.5) and high (0.75). For the conclusions, one can discard the observations for low values of sample size : they are interesting in order to show the trend, but unrealistic cases for the real life applications. Besides, we display the three correlation coefficients in the plots, for the interester reader, but we shall base our discussion on the Spearman's ρ. After carefully examination we conclude that:

- When conditional probability is low ($p_{B|A} = 0.25$), confidence is unexpectedly highly correlated with lift, and (as expected) poorly correlated with Loevinger and Gras' index. The three indexes in the triplet (L, H, φ) keep rather high pairwise correlations, but not as high as one may expect. In most of cases (except for the smallest values) the larger the sample size the lower the correlation, apparently converging to a limiting value.
- When conditional probability is average ($p_{B|A} = 0.5$), the plots are more interesting. The three indexes in the triplet (L, H, φ) keep extraordinary high pairwise correlations, probably because they all measure the *distance from independence*. and this scenario is pure independence. The trend shown by the sample size suggests an increasing limiting *perfect* correlation (under Spearman's eyes). Confidence keeps a moderately good correlation with the rest of indexes.

- When conditional probability is high ($p_{B|A} = 0.75$), the triplet (L, H, φ) loses the extraordinary high pairwise correlations, although they remain rather high. In this case, confidence is surprisingly highly correlated to Loevinger's H, as well as H is with φ. The trend shown by sample size is unclear about stabilization of the values.
- In all cases, an asymptotical result ($n \to \infty$) is an interesting open question for future research.

6 Conclusions

For the simplest bivariate binary random variable (A,B), and moderate sample sizes n, we have provided scripts in R language for computing the probability functions of the rule quality measures confidence, lift, Loevinger's H and Gras' φ, seen as random variables.

We have shown, through plots for particular parameter values, the spread of those random variables with respect to their expected value, giving an idea of how close (or far away) is the value for a particular sample to other possible samples.

We have also computed correlation coefficients among the (random) quality indexes, in order to understand what these indexes are *measuring* indeed. We have found that, when (A,B) are close to statistical independence, lift, Loevinger and Gras are measuring a very similar concept (deviation from independence), and their values are strongly correlated. However, when (A,B) are far from statistical independence, their relation is direct but moderate. In contrast, confidence is in another plane, measuring a different concept. However, we have come across with a rather strong relation between confidence and lift (respectively Loevinger), when (A,B) are away from statistical independence.

With these exploratory results, we have more arguments to suggest that lift, Loevinger's H and Gras' φ do measure such a similar concept. Hence, we think that the researcher can choose freely one or the other without concerns. However we recommend that she, or he, should understand not only the interpretation of extreme values of the index, but also intermediate values:

- Lift is a very popular choice, because it has a very intuitive interpretation : a factor of improvement of prediction of b when a is present.
- However, we do not find an easy or intuitive interpretation of the intermediate values of Loevinger's H (except the trivial value $H = 0$, interpreted as pure independence).
- In our humble opinion, we prefer Gras' φ since any of its values is clearly interpreted from its definition, as a p-value of a hypothesis test against statistical independence.

7 Acknowledgements

The first author acknowledges financial support of research projects P1·1B2012-52 from *Universitat Jaume I de Castellón* and MTM2013-43917-P from *Ministerio de Economía y Competitividad* of Spain.

References

[1] Agrawal, R., T. Imielinski, and 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).

[2] Bodin, A. (1997), *Analyse implicative : modèles sous-jacents à l'analyse implicative et outils complémentaires*, Prepublication IRMAR No. 97-32, Université de Rennes.

[3] Gras, R., S. Ag Almouloud, M. Bailleul, A. Lahrer, M. Polo, H. Ratsimba-Rajohn, H. and A. Totohasina, (1996), *L'implication statistique*, La Pensée Sauvage, Grenoble.

[4] Gras, R. and R. Couturier (2010), ASpécificité de l'A.S.I. par rapport à d'autres mesures de qualité de règles d'association, *Quaderni di Ricerca in Didattica (Mathematics)*, **20**(supp.1), 175-200.

[5] Gras, R., R. Couturier, J. Blanchard, H. Briand, P. Kuntz and P. Peter (2004), Quelques critères pour une mesure de qualité de règles d'association. Un exemple: l'implication statistique, In H. Briand, M. Sebag, R. Gras and F. Guillet (eds.), *Mesures de qualité pour la fouille de données* (pp. 3-32), RNTI-E-1, Cépaduès éditions.

[6] Gras R., J.-C. Régnier, C. Marinica and F. Guillet (2013), *L'analyse statistique implicative : méthode exploratoire et confirmatoire à la recherche de causalités*, Cépaduès éditions.

[7] Gras R., S. Suzuki, F. Guillet and F. Spagnolo (2008), *Statistical Implicative Analysis, Theory and Applications*, Springer.

[8] Gregori, P., R. Couturier and R. Pazmiño-Mají (2014), On the probability distribution of the classical Gras implication index between two binary random variables, *Educação Matemática Pesquisa*, **16**(3), 969-980.

[9] Hankin, R.K.S. (2006), Additive integer partitions in R, *Journal of Statistical Software, code snippets*, **16**(1), 1-3.

[10] Lallich, S., B. Vaillant and P. Lenca (2007), A probabilistic framework towards the parameterization of association rule interestingness measures, *Methodology and Computing in Applied Probability*, **9**(3), 447-463.

[11] Loevinger, J. (1947), A systematic approach to the construction and evaluation of tests of ability, *Psychological Monographs*, **61**(4), i-49.

[12] Vaillant, B., S. Lallich, and P. Lenca (2008), On the behaviour of the generalisations of the intensity of implication: a data-driven comparative study, In

R. Gras, E. Suzuki, F. Guillet and F. Spagnolo (eds.), *Statistical Implicative Analysis: Theory and Applications* (pp. 421-448), Studies in Computational Intelligence, 127, Springer-Verlag, Berlin Heidelberg.

Appendix : code blocks

Code block for the computation of all considered index:

```
funindex = function(x, index) {
  # computes quality 'index' for a given contingency table,
  # Arguments:
  #   x: cell counts, in the form c(n00, n01, n10, n11)
  #   index: name of the desired index
  n00=x[1]; n01=x[2]; n10=x[3]; n11=x[4];
  if(index=='conf') { # confidence
    if(n11+n10 !=0 ) res = n11/(n11+n10)
    else res=NA
  }
  if(index=='lift') { # lift
    if(((n11+n10)*(n11+n01)) !=0 )
      res = sum(x)*n11/((n11+n10)*(n11+n01))
    else res=NA
  }
  if(index=='loev') { # Loevinger's H
    if(((n11+n10)*(n00+n10)) !=0 )
      res = 1 - sum(x)*n10/((n11+n10)*(n00+n10))
    else res=NA
  }
  if(index=='gras') { # Gras phi binomial model
    res = 1-pbinom(q=n10, size=sum(x),
                   prob=((n10+n11)*(n00+n10))/((sum(x))^2))
  }
  return(res)
}
```

Code block for the computation of the probability functions in fixed sample size.

```
pf.index = function(index='gras', p=rep(1,4),
                    pX=NULL, pY=NULL, pYgivenX=NULL,
                    size=4) {
  # It computes the probability function of any 'index'
  # Arguments:
  #   index: 'conf', 'lift', 'loev' or 'gras', for the confidence,
  #     lift, Loevinger's H or Gras index.
  #   p: vector of 4 probabilities c(p00, p01, p10, p11)
  #     or proportional.
  #   pX, pY, pYgivenX: alternative to 'p', marginal and
  #     conditional probability.
  #   size: sample size.
```

```
  # Value: a data frame with sample space and probability function
  #   of the random variable index, as a function of the random
  #   bivariate Bernoulli with parameters 'p' or pX, etc.
  require(partitions)
  N = size
  if( !(is.null(pX) & is.null(pY) & is.null(pYgivenX)) ) {
    p11 = pYgivenX * pX
    p10 = pX - p11
    p01 = pY - p11
    p00 = 1 - (p01+p10+p11)
    p = c(p00, p01, p10, p11)
  }
  NN = compositions(N,4) # all partition of N into 4 numbers
  # compute index value for all possible partition
  index.NN = apply(X=NN, MARGIN=2, FUN='funindex', index=index)
  # compute probability of all such partition
  pr.NN = apply(X=NN, MARGIN=2, FUN='dmultinom', size=N, prob=p)
  # REMARK: computed only for fixed sample size
  res = aggregate(x=pr.NN, by=list(index.NN), FUN='sum')
  # NA values (and their probability) are discarded
  names(res) = c(index, 'prob')
  return(res)
}
```

Code block for the computation of the bivariate probability functions in fixed sample size.

```
pf.2index = function(index=c('conf', 'conf'), p=rep(1,4),
                     pX=NULL, pY=NULL, pYgivenX=NULL,
                     size=4) {
  require(partitions)
  N = size
  if( !(is.null(pX) & is.null(pY) & is.null(pYgivenX)) ) {
    p11 = pYgivenX * pX
    p10 = pX - p11
    p01 = pY - p11
    p00 = 1 - (p01+p10+p11)
    p = c(p00, p01, p10, p11)
  }
  NN = compositions(N,4) # all partitions of N into 4 numbers
  # probability of all such partition
  pr.NN = apply(X=NN, MARGIN=2, FUN='dmultinom', size=N, prob=p)
  # REMARK: computed only for fixed sample size
  ind1 = apply(X=NN, MARGIN=2, FUN='funindex', index=index[1])
  ind2 = apply(X=NN, MARGIN=2, FUN='funindex', index=index[2])
  res = aggregate(x=pr.NN, by=list(ind1, ind2), FUN='sum')
  # NA values (and their probability) are discarded
  names(res) = c(index, 'prob')
  return(res)
}
```

Code block for the computation of the correlation coefficients.

```
cor.index = function(index = c('conf', 'conf'),
                     p=rep(1,4), pX=NULL, pY=NULL,
                     pYgivenX=NULL, size=4,
                     method='pearson') {
  if( !(is.null(pX) & is.null(pY) & is.null(pYgivenX)) ) {
    p11 = pYgivenX * pX
    p10 = pX - p11
    p01 = pY - p11
    p00 = 1 - (p01+p10+p11)
    p = c(p00, p01, p10, p11)
  }
  pf2 = pf.2index(index=index, p=p, size=size)
  ind1 = pf2[[ index[1] ]]
  ind2 = pf2[[ index[2] ]]
  pr.NN = pf2$prob/sum(pf2$prob)
  if( method=='pearson') {
    cor = ( sum(ind1*ind2*pr.NN) - sum(ind1*pr.NN)*sum(ind2*pr.NN) ) /
      ( sqrt( (sum(ind1^2*pr.NN) - sum(ind1*pr.NN)^2) *
                (sum(ind2^2*pr.NN) - sum(ind2*pr.NN)^2)) )
  }
  if( method=='kendall' ) {
    ind1.ij = outer(ind1, ind1, '-')
    ind2.ij = outer(ind2, ind2, '-')
    pr.ij = outer(pr.NN, pr.NN)
    conc.disc = sum(sign(ind1.ij * ind2.ij) * pr.ij)
    cor = sum(conc.disc)/(1-sum((pr.ij[ ind1.ij==0 & ind2.ij==0 ])))
  }
  if( method=='spearman') {
    ind1 = rank(ind1)
    ind2 = rank(ind2)
    cor = ( sum(ind1*ind2*pr.NN) - sum(ind1*pr.NN)*sum(ind2*pr.NN) ) /
      ( sqrt( (sum(ind1^2*pr.NN) - sum(ind1*pr.NN)^2) *
                (sum(ind2^2*pr.NN) - sum(ind2*pr.NN)^2)) )
  }
  return( list(cor=cor, method=method, index=index,
               size=size, p=p, cor=cor) )
}
```

POURQUOI ET COMMENT TRANSFORMER DES VARIABLES QUANTITATIVES EN CATEGORIELLES ? APPLICATION A L'INTONATION DE LA LANGUE FRANÇAISE

Martine CADOT[1] et Anne BONNEAU[2]

TRANSFORMING QUANTITATIVE VARIABLES INTO QUALITATIVE ONES: RATIONALE AND METHOD. APPLICATION TO FRENCH INTONATION.

RÉSUMÉ

L'interprétation d'une phrase en français dépend non seulement de ses mots, mais aussi de son intonation : interrogative, déclarative, dubitative, etc. Notre but est de construire un modèle statistique permettant de différencier les types intonatifs d'une phrase prononcée par un locuteur à partir des indices acoustiques issus de son enregistrement sonore. Pour étudier le lien entre la variable catégorielle « type de phrase » et les variables quantitatives d'indices acoustiques, il a fallu transformer ces dernières en variables catégorielles. Nous discutons ici des raisons statistiques qui peuvent imposer cette transformation et de la façon d'y procéder en nous plaçant d'un point de vue plus général et théorique avant d'appliquer ce formalisme aux données expérimentales recueillies.

Mots-clés : *Implication statistique, liaisons complexes entre variables, fouille robuste de données, recherche de modèle de discrimination, découpage de variables quantitatives, intonation, indices acoustiques, courbes mélodiques, phonétique.*

ABSTRACT

The interpretation of a French oral sentence depends not only on its word sequence but also on its intonation : interrogative, declarative, doubtful .... Our aim is build a statistical model able to differentiate various intonations and relying upon the acoustic cues extracted from sentences pronounced by French speakers. To study the connection between the categorical variable "type of sentence" and the quantitative variables stemming from the various acoustic cues considered in this study, it was necessary to transform the latter into qualitative variables. We discuss the statistical reasons that can impose this transformation and the way to carry it by considering a more general and theoretical point of view before applying this formalism to the collected experimental data..

Keywords : *statistical implication, complex links between variables, robust data mining, discriminant model research, quantitative versus qualitative variables, intonation, acoustic cues, melodic curves, phonetic.*

1 Introduction

Notre but est de construire un modèle statistique permettant de différencier les types intonatifs d'une phrase prononcée par un locuteur à partir des indices acoustiques de son enregistrement sonore. Le modèle que nous visons est de la forme X→Y, se lisant selon les cas : X implique Y, X cause Y, X explique Y, X produit Y, etc. (Gras *et al.*, 2013)

[1] LORIA/UdL, 54600 Villers-lès-Nancy, France, martine.cadot@loria.fr
[2] LORIA/CNRS, 54600 Villers-lès-Nancy, France, anne.bonneau@loria.fr

où X représente les indices acoustiques et Y le type intonatif de la phrase. Il est « asymétrique » dans la mesure où les variables n'ont pas toutes le même rôle[3] : la variable Y est la « variable à expliquer » (appelée également variable dépendante), et les variables acoustiques X sont les « variables explicatives » (appelées également variables indépendantes). Pour construire ce modèle nous avons utilisé une partie des données recueillies selon un « plan d'expérience » pré-établi, nous avons créé 29 variables acoustiques quantitatives X, 3 variables qualitatives (appelées également catégorielles) Z à contrôler (telles que le genre : M/F, dont il faut contrôler les effets sur le modèle pour les retirer le cas échéant[4]) et la variable catégorielle de contour mélodique Y dont nous n'avons gardé que 3 modalités.

Dans une première partie, plus théorique, nous réfléchissons à ce qu'est une liaison entre une variable catégorielle Y et un ensemble de variables quantitatives X, en nous restreignant toutefois à ce qui sera applicable à nos données et à notre problématique. Nous voyons comment cette liaison peut s'exprimer quantitativement de diverses façons, notamment après recodages des variables quantitatives, et être validée statistiquement. Pour rendre cet exposé plus lisible, nous l'avons illustré à l'aide d'un jeu de données bien connu, « les iris de Fisher » (Fisher 1936), formé des valeurs de 5 variables (X : 4 mesures de fleurs, et Y : la catégorie de la fleur parmi 3 catégories) recueillies sur 150 fleurs. À partir de cette réflexion nous donnons les raisons qui nous ont fait choisir l'ASI (Gras *et al.* 1996) et le logiciel CHIC (Version 6.0, Copyright© 2012) pour traiter nos données.

Dans la deuxième partie, plus applicative, nous exposons notre problématique, relative au rôle de l'intonation dans l'interprétation d'une phrase enregistrée en français, le plan d'expériences choisi ainsi que le prétraitement des données, puis nous montrons les résultats de l'utilisation de CHIC pour la mise au point de notre modèle sur ces données. A cette occasion, nous comparons les modèles obtenus avec CHIC en reprenant la réflexion de la première partie et notamment les divers recodages des données quantitatives.

A part CHIC pour l'ASI, et le tableur Excel pour la régression linéaire, le tableau de contingence et une partie des graphiques, tous les modèles testés l'ont été à l'aide de scripts écrits dans le langage/logiciel R (disponible gratuitement sous licence GNU http://www.r-project.org/), dont le code est mis en annexe.

2 Liaison statistique entre une variable catégorielle et des variables quantitatives

Dans cette section, nous présentons d'abord le jeu de données qui va nous permettre d'illustrer les modèles, puis les divers modèles existants, avant de choisir celui que nous utiliserons.

[3] Pour un exposé détaillé sur la planification des variables en psychologie expérimentale voir Hoc (1983), en agronomie voir Dagnelie.(2003).

[4] Pour un exposé détaillé sur la gestion des variables à contrôler en psychologie expérimentale voir Léon *et al.* (1977), en médecine voir Schwartz (1983).

2.1 Les iris de Fisher

Ce jeu de données est téléchargeable depuis UCI repository (https://archive.ics.uci.edu/ml/datasets/Iris). Il est formé d'une variable catégorielle Y (l'espèce) à 3 modalités (iris-setosa, iris-versicolor, iris-virginica), avec 50 fleurs de chaque espèce, et d'un ensemble de 4 variables numériques X, qui sont la longueur et la largeur moyennes des sépales et des pétales (Sepal_Length, Sepal_Width, Petal_L, Petal_W). Il a été utilisé par de nombreux chercheurs en analyse de données désirant éprouver une nouvelle méthode statistique, soit en classification pour retrouver les 3 groupes de fleurs, soit en discrimination pour trouver les règles d'attribution d'une fleur à un groupe. C'est dans ce deuxième cadre que nous nous plaçons. Dans la figure 1, nous avons représenté les 150 fleurs dans l'espace formé de 3 variables de X (Petal_L, Petal_W et Sepal_W), avec une couleur par espèce.

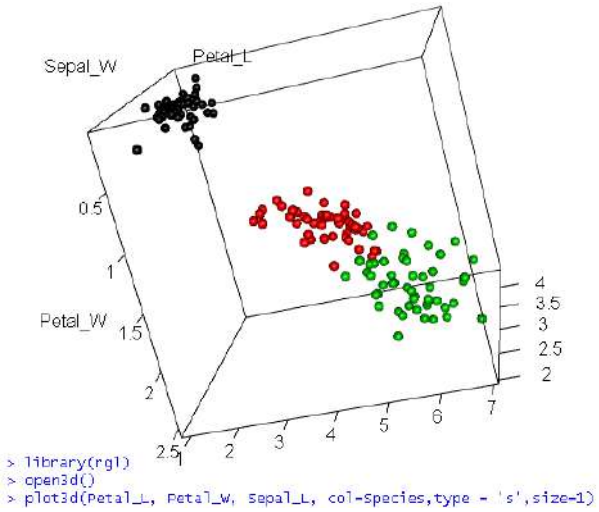


Figure 1 – Les 150 iris de Fisher dans l'espace de 3 variables, avec une couleur par espèce

On peut remarquer que la longueur et la largeur des pétales permettent d'affecter plus ou moins facilement une espèce à chaque iris : les plus petits pétales correspondent à l'espèce Setosa (en noir), les moyens à l'espèce Versicolor (en rouge) et les plus grands à l'espèce Virginica (en vert). Plus précisément, parmi les 3 nuages de points de couleurs différentes de la figure 1, seuls le nuage de points rouges et celui de points verts ont une petite partie commune d'une dizaine d'individus. Dans la mesure où un simple découpage en quelques sous-espaces de la figure 1 permet d'établir l'espèce des 150 iris avec une dizaine d'erreurs, on attend de chacune des méthodes de discrimination étudiées qu'elles atteignent des taux de reconnaissance de plus de 90%.

2.2 Les différents modèles possibles

Nous nous plaçons ici dans le cadre de la prédiction : connaissant la valeur de X et celle de Y sur un ensemble de données, nous recherchons une méthode qui permette de trouver Y_{pred} (prédiction de Y) à partir de X, en faisant le moins possible d'erreurs (Erreur=$Y-Y_{pred}$). Pour que cette méthode puisse être élevée au rang de modèle, il faut qu'elle soit suffisamment détachée des données de départ, simple et intelligible. Nous faisons un tour rapide des modèles prédictifs courants en donnant pour chacun une référence renvoyant à un manuel pédagogique dans lequel cette méthode est détaillée

parmi d'autres, une partie restreinte de ces références pouvant donc suffire pour couvrir l'ensemble des modèles décrits.

2.2.1 Les modèles statistiques

Un des modèles les plus utilisés est le ***modèle linéaire*** (Prum 1996). Sa version de base est le ***modèle de régression linéaire***, pour lequel X et Y sont deux variables quantitatives. Il se résume à la connaissance des 2 coefficients a et b de la ***droite de régression*** Y_{pred}=aX+b sur laquelle sont situés les points de coordonnées (X, Y_{pred}) dans le plan de X et Y (par exemple la première équation à droite de la figure 2 exprime la liaison entre la longueur prédite du pétale et la longueur du sépale pour les iris setosa, et elle est représentée par la droite bleue du graphique). Quand X est formé de p variables quantitatives (p>1), le modèle de régression linéaire s'écrit Y_{pred}=XA+B avec A et B des vecteurs de p valeurs fixées, et c'est l'équation d'un hyper-plan dans l'espace de dimension p+1 de X et Y.

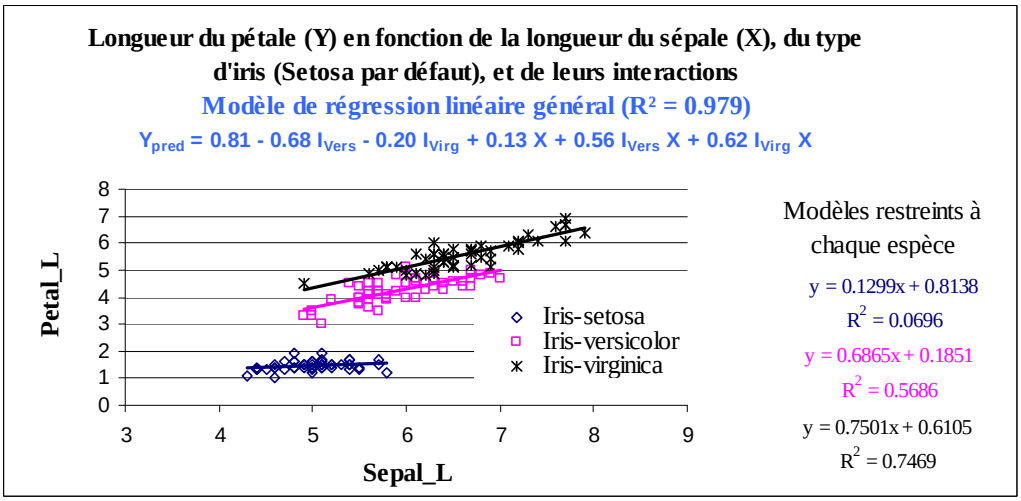


Figure 2 – Un modèle de régression de la longueur du pétale sur la longueur du sépale et l'espèce d'iris. En haut, en bleu le modèle global prenant en compte les 150 mesures, et à droite les 3 modèles spécifiques à chaque type d'iris, ne prenant en compte que 50 mesures chacun.

La régression linéaire peut s'étendre à des variables X non toutes quantitatives et on parlera plutôt de ***modèle linéaire généralisé*** (Baillargeon 2000). Par exemple en Figure 2, on a exprimé la dépendance de Y (longueur des pétales) en fonction de X (petal_L, la longueur des pétales, I_{vers} et I_{virg} les variables indicatrices de 2 types d'iris (respectivement iris-versicolor et iris-verginica, le troisième type étant la référence par défaut, obtenu quand les deux variables sont nulles simultanément). Si au lieu d'être quantitative, Y est binaire (1 : réussite, 0 : échec), on utilise le modèle ***logistique*** (Besse 2003) pour lequel l'équation devient logit(P_{est})=XA+B, où P_{est} est la probabilité que Y soit égal à 1, et logit(P)= log(P/(1-P)). Le modèle logistique peut encore s'étendre à une variable Y de comptage (Y : nombre de réussites) ou ordinale, mais plus difficilement à une variable catégorielle. Toutefois, si on « éclate » en p variables binaires la variable catégorielle Y à p catégories, on peut alors chercher p modèles logistiques, un par variable binaire. Une extension dans une autre direction est le modèle ***log-linéaire*** (Morineau, 1996), pour lequel toutes les variables sont catégorielles, X comme Y, et c'est alors chaque variable de X qu'il convient d'éclater en intervalles, ces intervalles devenant les catégories de la variable, et dans ce modèle, ce n'est plus Y qui dépend

linéairement de X mais log(Effectif$_{pred}$) qui est une fonction linéaire de X et Y. La ***discrimination linéaire*** (Nakache 2003) est un modèle un peu différent, elle consiste à trouver un changement de repère dans X pour lequel les catégories de Y sont séparées le mieux possible. En Figure 3, on a représenté la projection des 150 iris dans l'espace des deux hyperplans obtenus pour la discrimination linéaire des 3 espèces d'Iris (un seul aurait suffit, car il produit plus de 99% de la variance).

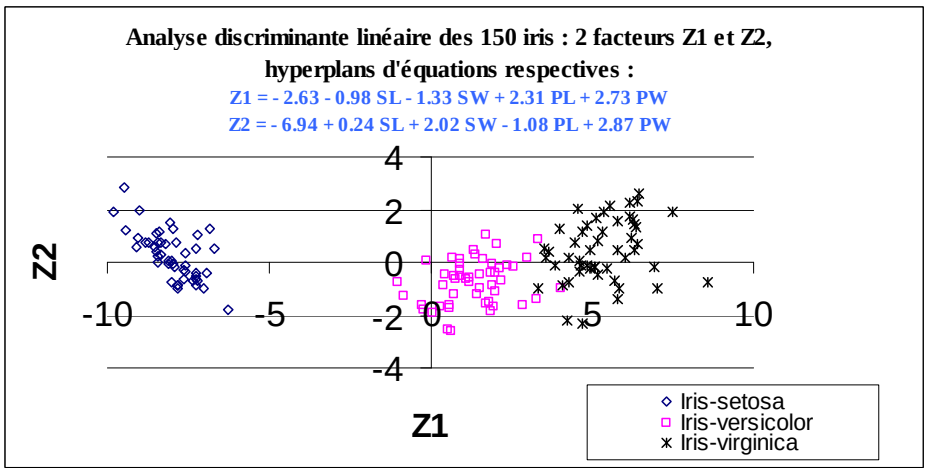


Figure 3 – Analyse discriminante des 150 iris : plus de 99% de variance selon Z1.

Le « modèle linéaire » et ses extensions, sont issus des statistiques classiques. Ils font partie des ***modèles paramétriques*** et sont assortis de conditions d'application qui permettent de garantir la qualité des estimations produites. Quand les conditions d'application ne sont pas vérifiées (par exemple la normalité des erreurs), des modèles non paramétriques (Siegel 1988) peuvent être utilisés, par exemple, si le nombre d'ex-aequo est réduit, en remplaçant les variables par leurs rangs (Droesbeke, 1996).

2.2.2 Les modèles issus de l'intelligence artificielle

A côté de ces méthodes qui, depuis plus d'un siècle, ont permis de produire des prédictions « garanties » à partir de mesures faites sur un seul échantillon choisi avec soin (Morin 1999), depuis quelques dizaines d'années de nouvelles méthodes de prédiction sont apparues avec l'avènement de l'informatique, qui se sont généralisées dernièrement avec le libre accès à des ressources sur Internet. Issues de « l'apprentissage automatique » (Mitchell 1997) ces méthodes se proposent d'évaluer la qualité des prédictions en partitionnant l'ensemble des données en plusieurs parties de diverses façons (par exemple k parties pour la ***validation croisée***, le modèle étant construit sur la réunion de k-1 parties, testé sur la partie restante, on itère le processus k fois en changeant la partie restante). Le nombre de ces méthodes est en constante augmentation (plus de 6500 packages R disponibles à ce jour (http://cran.r-project.org/), dont une grande partie développant des méthodes statistiques), dans la mesure où leur validité peut s'établir sans avoir besoin de théories statistiques ***asymptotiques*** (par exemple s'appuyant sur l'hypothèse de normalité), mais par une mise à l'épreuve sur des jeux de données caractéristiques d'un problème spécifique à résoudre.

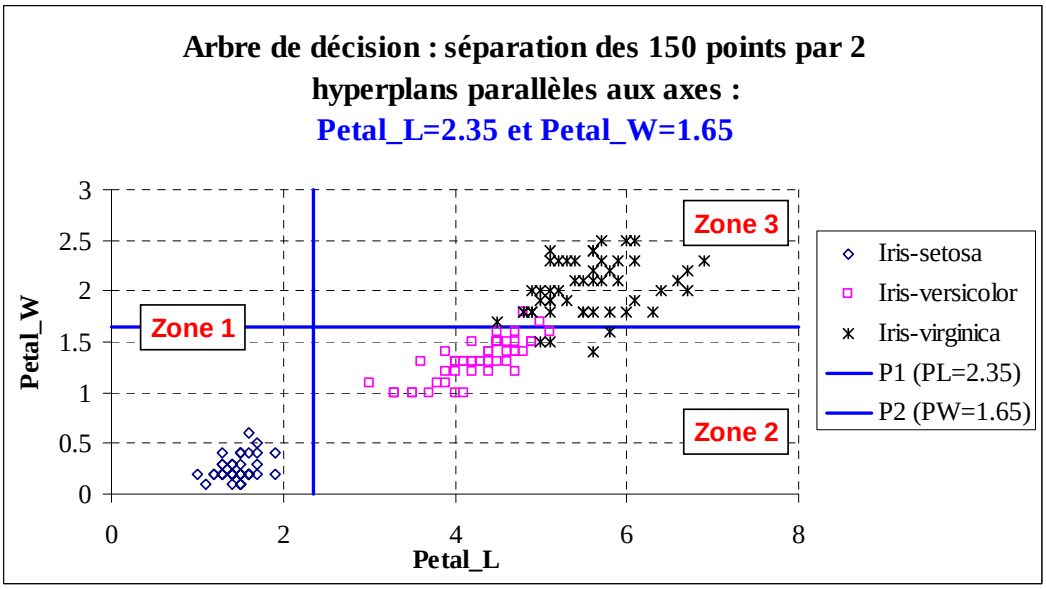


Figure 4 – Résultat sur les 150 iris d'un arbre de décision entraîné une fois sur 75 iris tirés au hasard (25 par classe) : 2 hyperplans séparateurs et 6 erreurs (2 pour la partie apprentissage et 4 pour l'autre).

Nous ne nous intéresserons ici qu'aux méthodes qui peuvent prédire une variable Y catégorielle de plus de 2 catégories à partir d'un ensemble de variables X quantitatives, et donc s'appliquer au jeu de données des iris de Fisher. Les plus pratiquées sont des extensions des réseaux neuronaux (Ripley 1996), les MSVM (Multiple Support Vector Machine, Weston 1998) et les arbres de décision (Breiman 1984). Dans la Figure 4, nous montrons le partage en régions proposé par un arbre de décision sur les données iris. Deux hyperplans parallèles aux axes ont suffi pour déterminer les 3 zones de points formées quasi-exclusivement d'une seule espèce d'Iris. La zone 1 correspond à Petal_L<2.35 et ne contient que les iris Setosa. La zone 2 correspond à Petal_L≥2.35 et Petal_W<1.65, et contient tous les iris Versicolor, sauf 1, ainsi que 5 iris Virginica. Et dernière zone correspondant à Petal_L≥2.35 et Petal_W≥1.65 contient la quasi-totalité des iris Virginica, ainsi qu'un iris Versicolor. Il y a 6 erreurs sur 150 iris, soit 4% d'erreurs. Du seul point de vue de l'interprétation graphique, on peut dire que les MSVM sont une extension des arbres de décision, avec des hyperplans frontières qui ne sont plus nécessairement parallèles aux axes, des frontières qui ne sont plus strictes, mais floues (ce sont des bandes), et dans un espace qui est une extension de X.

Signalons pour finir l'extraction ***des règles d'association***, pour lesquelles les méthodes de validation sont encore à l'état de recherches (Cadot 2006). Partant d'un jeu de données formé des valeurs de q variables catégorielles Xi sur un certain nombre de sujets, la méthode extrait automatiquement des règles de la forme (A et B et C)→D, où le membre de gauche contient un certain nombre d'affirmations (ici il y en a 3, A, B et C) du type Xi=Cik (la variable Xi a pour valeur sa catégorie Cik), celui de droite n'en contenant qu'une en général. Toute combinaison de variables et de catégories fournissant une règle, leur nombre augmente de façon exponentielle avec le nombre de variables, et le défi est de trouver le (ou les) indices de qualité permettant de ne garder que les meilleures. Si on ne garde que les règles ayant Y en partie droite, et si on a transformé les variables quantitatives en catégorielles pour les utiliser en partie gauche, cette méthode peut répondre à notre recherche en nous fournissant des règles de discrimination.

2.2.3 Où se situe l'ASI ?

Dans sa version originale, l'ASI s'applique à des variables binaires, qu'on obtient souvent par éclatement de variables catégorielles (la variable Genre avec 2 catégories F et M, donnera par éclatement 2 variables binaires F et M). Les règles fournies ressemblent donc aux règles d'association (RA) mais la méthode présente en outre 2 avantages : 1) l'interface graphique de CHIC permet de « voir » les règles et de les manipuler 2) seules les règles valides s'affichent, la validité des règles étant établie selon la théorie statistique asymptotique (comptages vérifiant la loi de poisson ou binomiale, selon un choix à préciser dans les options de CHIC).

Dans la version actuelle, on peut utiliser des variables quantitatives telles quelles, ou les faire recoder en variables catégorielles. Nous avons utilisé les données des iris pour confronter les réponses que l'ASI offre à notre problème de discrimination X→Y. Nous avons d'abord utilisé les variables sans les découper mais en les remplaçant par une valeur entre 0 et 1 selon la formule indiquée dans l'aide :

valeur_nouvelle=(valeur_ancienne-min_valeurs)/(max_valeurs-min_valeurs),

Puis nous les avons fait découper en 2, 3, ou 4 parties par le logiciel CHIC. Le graphe implicatif des variables quantitatives non recodées, seulement réajustées pour être dans l'intervalle [0 ; 1] est en figure 5, à gauche, et à droite on a celui obtenu avec les variables découpées en 3 par CHIC (les 2 autres découpages testés, qui se sont avérés moins bons, ont été mis en annexe).

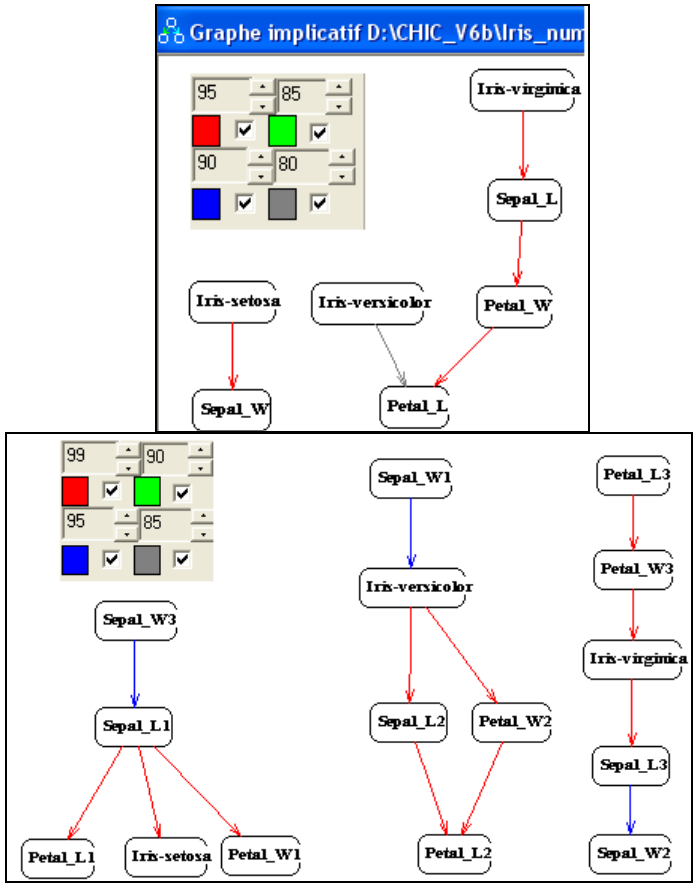


Figure 5 – Graphes implicatifs produits par CHIC sur les 150 iris. À gauche les 4 variables quantitatives à valeurs réajustées dans [0 ; 1], à droite recodées en 3 catégories chacune par CHIC.

La première différence entre les deux graphiques de la figure 5 porte sur les seuils. Comme nous voulions faire apparaître la totalité des 3 catégories, dans le graphique de gauche, nous avons dû descendre jusqu'à un seuil d'implication statistique de 0.80 pour faire apparaître Iris-versicolor, ce qui fournit donc une règle de piètre qualité pour cette espèce. La recherche de règles de bonne qualité pour chaque espèce d'iris est la première raison qui nous fait préférer le graphique de droite.

La deuxième raison qui nous fait préférer le graphique de droite est que chaque espèce apparaît dans un graphe séparé. Cette séparation exprime graphiquement ce que les autres modèles nous ont indiqué dans des tables de confusion croisant les espèces estimées par le modèle et les espèces observées : connaissant les mesures des pétales et des sépales, on peut en déduire avec peu d'erreurs l'espèce de l'iris. Comparons maintenant la figure 5 à la figure 2. La régression de Petal_L sur Sepal_L n'était significative que pour iris-virginica et iris-versicolor, pas pour iris-setosa, comme on peut le voir dans les équations à droite du graphique de la figure 2 et nous retrouvons un peu cela dans le graphique de gauche de la figure 5, dans la mesure où les 2 variables Petal_L et Sepal_L sont reliées entre elles fortement ($p>0.95$) et également avec les 2 espèces d'iris convenables. Le modèle de régression linéaire nous en dit plus, notamment, dans l'équation globale, les coefficients qui diffèrent significativement de 0 sont les seules interactions entre Sepal_L et chacune des 2 catégories. Mais ce qu'il dit est d'interprétation difficile, et si l'on faisait la régression pour chaque sous-ensemble de variables cela poserait des problèmes d'interprétation encore plus complexes. Si on prend maintenant le graphique de droite de la figure 5 nous pouvons dire que la liaison des variables Sepal_L et Petal_L est forte car chaque fois qu'elles sont reliées par une flèche, c'est avec la même catégorie (1, 2 ou 3), et à chaque catégorie correspond une espèce d'iris. Sauf pour l'espèce setosa, un chemin rouge (seuil=0.99) fait de 2 flèches qui se suivent joint l'espèce et les 2 catégories de Sepal_L et Petal_L, on retrouve ainsi les éléments déjà cité de la figure 2, mais en plus on a une information fine sur la relation entre les espèces et les variables Sepal_L et Petal_L : les valeurs les plus faibles des 2 variables sont pour iris-setosa, les valeurs intermédiaires pour iris-versicolor et les plus fortes pour iris-virginica. Et on voit dans ces 3 graphes que les niveaux de Petal_W sont associés aux espèces de la même façon. Par contre la variable Sepal_W se comporte différemment : le niveau le plus faible est associé à versicolor, le moyen à virgina et le plus fort à setosa, et ces associations sont plus faibles (entre 0.95 et 0.99).

Jusqu'ici, la balance penche fortement du côté du graphique de droite de la figure 5 : Il nous donne des informations qui rejoignent celles que nous avons trouvées avec les autres modèles donnés en exemples, ainsi que des informations supplémentaires que nous n'avons pas su extraire des autres modèles, pour autant qu'elles y soient.

Maintenant se pose la question de l'interprétation du sens des flèches, et notamment : pourquoi la position en hauteur de l'espèce diffère-t-elle d'un graphe à l'autre dans le graphique de droite de la figure 5. Faut-il interpréter différemment ces 3 graphes ? Ce qui nous renvoie à la sémantique de ces flèches. Dans l'ouvrage collectif sur l'ASI (Gras *et al,* 2013) nous lisons dans l'introduction (Gras et Regnier, section 4, p. 16-17) que cette flèche peut indiquer une inclusion plutôt qu'une causalité, ce que nous appellerons un « effet d'effectif ».

niveau	Sepal_L	Sepal_W	Petal_L	Petal_W
1	46	47	50	50
2	53	67	54	52
3	51	36	46	48

Tableau 1 – Effectifs selon les catégories obtenues en découpant chaque variable quantitative en 3 par CHIC (ces « informations sur le fichier » figurent sous cette dénominations dans les sorties de CHIC)

Avant donc d'essayer d'interpréter le sens des flèches pour les iris, examinons la relation entre l'ordre des effectifs et la hauteur dans le graphe. Dans le tableau 1 on a reporté les effectifs de chacune des 3 catégories des variables selon le découpage fait par CHIC. On voit que la variable Sepal_W a des catégories déséquilibrées, ce qui n'est pas le cas des 3 autres variables, dont les effectifs restent entre 46 et 54. Dans le niveau 3, si on range ces 3 variables ainsi que l'espèce par effectif croissant, on obtient (le symbole « < » se lira « précède »):

Petal_L (46) < Petal_W (48) < Iris-virginica (50) < Sepal_L (51)

C'est exactement dans cet ordre que les flèches successives joignent les items du 3ème graphe, figure 5, à droite. Pour le niveau 2, on a :

Iris-versicolor (50) < Petal_W (52) < Sepal_L (53) < Petal_L (54),

Et c'est quasiment dans cet ordre qu'on trouve le graphe 2, sauf que Petal_W et Sepal_L (53) ne sont pas l'un sous l'autre mais l'un à côté de l'autre. Pour le premier graphe, on retrouve exactement l'ordre des effectifs.

Dans notre exemple, le sens des flèches pourrait ainsi être un effet d'effectif. D'ailleurs, il paraît difficile de justifier sémantiquement un chemin de causalité différent pour chaque espèce d'iris, par exemple le fait d'avoir l'implication « sépales courtes → iris-setosa » et l'implication inverse « iris-virginica→sépales longues ». Nous décidons donc de ne pas essayer d'interpréter en termes de causalité le sens des flèches dans les graphes implicatifs.

Nous essayons maintenant d'en déduire des règles de discrimination. Pour cela, nous utilisons le fichier de données recodées après découpage en 3 des variables de X par CHIC, et nous en extrayons sous Excel un tableau de contingence (voir tableau 2) croisant la variable Y des catégories avec les 3 variables de X (Petal_L, Petal_W et Sepal_L) liées aux catégories de Y par des flèches rouges dans le graphe implicatif (Figure 5, à droite).

Nombre de Id			Species			
Petal_L	Petal_W	Sepal_L	Iris-setosa	Iris-versicolor	Iris-virginica	Total
1	1	1	40			40
		2	10			10
2	2	1		5		5
		2		29		29
		3		13		13
	3	1			1	1
		2		1	4	5
		3			1	1
3	2	2		1	2	3
		3			2	2
	3	2			6	6
		3		1	34	35
Total			50	50	50	150

Tableau 2 – Répartition des 150 iris selon leur espèce et 3 de leurs dimensions, en rouge les iris « mal classés ».

La lecture de ce tableau nous permet d'écrire le jeu de trois règles de discrimination suivant :

Si Petal_L1 alors iris_setosa (N=50, nbVrai=50, nbFaux=0)

Si Petal_L2 et Petal_W2 alors iris_versicolor (N=47, nbVrai=47, nbFaux=0)

Si (Petal_L2 et Petal_W3) ou (Petal_L3) alors iris_virginica (N=53, nbVrai=50, nbFaux=3)

Et en récupérant les intervalles de valeurs des variables figurant dans le journal de CHIC (voir tableau 3), on peut remplacer les modalités des variables quantitatives dans les règles par leur appartenance à un intervalle (par exemple Petal_L1 devient Petal_L $\in$ [1 ; 1.9]).

niveau	Sepal_L	Sepal_W	Petal_L	Petal_W
1	de 4.3 à 5.3	de 2 à 2.8	de 1 à 1.9	de 0.1 à 0.6
2	de 5.4 à 6.2	de 2.9 à 3.3	de 3 à 4.9	de 1 à 1.6
3	de 6.3 à 7.9	de 3.4 à 4.4	de 5 à 6.9	de 1.7 à 2.5

Tableau 3 – Intervalle de valeurs de chaque niveau des variables quantitatives donné par CHIC

Nous constatons que ce jeu de trois règles permet de discriminer les 150 iris avec seulement 3 erreurs. Ce qui nous permet d'affirmer que CHIC nous a fourni pour les iris un modèle tout à fait pertinent de la relation X→Y, où X est la matrice de 4 variables quantitatives et Y une variable catégorielle.

2.3 Pourquoi et comment choisit-on l'ASI

Comme on peut le constater dans la section 2.2, complétée par les annexes relatives aux sections 2.2.1 et 2.2.2, les diverses méthodes visant à modéliser X→Y arrivent presque toutes à discriminer les 3 catégories d'iris avec plus de 95% de réussite, ce qui n'est pas surprenant sachant que dans le sous-espace de X formé de la longueur et de la largeur des pétales (voir figure 1), une espèce se détache nettement des deux autres, ces dernières formant 2 blocs distincts juxtaposés avec une zone commune de moins de dix individus plus difficiles à catégoriser. C'est donc sur d'autres critères que les résultats quantitatifs que nous choisissons la méthode destinée à nous fournir un modèle.

2.3.1 Comparaison entre ASI et les autres méthodes de modélisation

Nous avons donné une liste de méthodes visant à produire un modèle simple et fiable de la relation X→Y. D'abord les méthodes statistiques linéaires pour lesquelles la fiabilité est assurée par la théorie statistique (tests d'hypothèses), et la simplicité par la linéarité, ce qui signifie que la valeur prédite de Y est combinaison linéaire des valeurs de X. C'est la formule la plus simple quand Y est quantitative, comme dans la figure 2 où Y est la longueur du pétale. Mais quand Y n'est pas quantitative, ce qui est le cas de l'espèce, il faut faire des transformations préalables des données pour conserver la linéarité, et le modèle se complexifie, même si les résultats peuvent s'exprimer par des équations et des graphiques comme en figure 3. Quand X contient peu de variables, on peut s'y retrouver, mais avec 29 variables comme dans les données d'intonation que nous voulons traiter, le choix et l'utilisation d'un tel modèle s'avère délicat. De plus les

conditions d'application exigées par ces méthodes statistiques linéaires ou dérivées sont vérifiées par les variables X des iris de Fisher, mais pas par celles des données d'intonation, qu'il faut transformer pour les adapter. Une fois ces étapes franchies, on dispose d'un modèle explicite utilisable pour la discrimination, et de la mesure de sa fiabilité, qui peut s'avérer plus ou moins forte. Quand elle est suffisamment forte, elle permet d'établir la théorie visée, d'où son intérêt.

Les méthodes de discrimination suivantes que nous avons vues (MSVM, arbres de décision) fournissent des résultats fiables sans exiger que les données vérifient autant de conditions que les méthodes statistiques précédentes, mais soit le modèle détaillé n'est pas fourni (fonctionnement des MSVM en « boîte noire »), soit il est simpliste (arbres de décision), la qualité se situant au niveau du résultat (taux élevé de prédictions réussies), pas du modèle[5]. Notons toutefois que ce n'est pas parce que ces méthodes n'exigent pas de conditions sur les distributions des données qu'elles pourront traiter correctement des données mal distribuées. Quant aux règles d'association, elles fournissent un modèle riche, mais complexe et qui reste attaché aux données traitées, par manque de méthode de validation statistique universellement reconnue. Leur avantage est de faire découvrir des relations locales inattendues, qui peuvent entrer dans la composition d'un modèle global qu'il conviendra de valider ensuite.

L'ASI est à la frontière entre les deux types de méthodes : elle fournit un modèle des données dont on peut évaluer la fiabilité grâce aux seuils indiqués dans l'interface graphique, mais sans avoir d'exigences sur la distribution des données. Son modèle sous forme de graphe implicatif permet d'exprimer des relations complexes entre les variables au sein d'un modèle global, mais comme elles ont toutes le même statut (pas de variable à expliquer ou explicative), la prédiction ne peut se faire directement dans CHIC. L'interface graphique permet de modifier les éléments du modèle (ajouter/retirer des variables, changer les seuils, déplacer des groupes de flèches, …) par simples clics.

Notre but étant d'extraire de nos données un modèle fiable, simple mais riche, nous avons choisi CHIC et ses graphes implicatifs, sachant qu'avec la facilité de manipulation procurée par l'interface graphique de CHIC, nous pouvons ajuster par clics jusqu'à obtenir le modèle qui nous convient le mieux sur les données d'intonation, sans être obligées de diagnostiquer leurs « défauts » pour les corriger au préalable (liaisons fortes entre certaines variables de X, distributions déséquilibrées, nombreux ex-aequo, valeurs extrêmes, etc.).

2.3.2 Découpage ou non des variables quantitatives

Transformer une mesure quantitative en 3 catégories fait perdre de la précision sur les valeurs (on ne dispose plus que de 3 valeurs non ordonnées), et on pourrait penser qu'il vaut mieux l'éviter pour avoir un modèle plus fiable. Nous avons vu que ce n'est pas le cas avec CHIC pour les iris, le modèle avec les variables découpées en 3 catégories ayant été préféré à celui avec les variables simplement recodées sur l'intervalle [0 ; 1]. Il semblerait donc que le recodage de variables quantitatives en variables catégorielles ne diminue par la précision du graphe implicatif. Et dans le cas des iris, il lui en a fait gagner. Cela s'explique par le fait que 3 dimensions sur 4

[5] A nos yeux, la qualité d'un modèle est fonction de son intelligibilité, de sa concision, opérationnalité, …

(Petal_L, Petal_W, Sepal_L) de la fleur d'iris sont très liées entre elles et ont tendance à varier dans le même sens, et si on ne les découpe pas, elles se retrouvent dans le même graphe et liées aux mêmes espèces d'iris (voir figure 5, à gauche), la dernière variable (Sepal_W) se trouvant dans un autre graphe avec les espèces restantes d'iris, ce qui ne fait que deux graphes pour 3 espèces, donc une mauvaise discrimination. Ce n'est qu'en découpant les variables qu'on a pu obtenir une discrimination fine de chaque catégorie, avec des graphes séparés (voir Figure 5, à droite).

Ayant établi que découper les variables quantitatives peut augmenter la précision du modèle global, il s'agit maintenant de définir la façon de procéder. La découpe d'une variable quantitative se fait automatiquement dans CHIC, une fois choisi le nombre de parties. C'est un algorithme qui détermine les seuils de découpe en optimisant le rapport entre variance inter et variance intra, indépendamment des autres variables, contrairement à l'algorithme des arbres de décision, qui optimise en fonction de la variable à discriminer. Ce choix permet à CHIC d'obtenir des niveaux de significativité non biaisés en conservant sa méthode d'estimation asymptotique, alors que les arbres de décision utilisent des méthodes de validation croisée pour éviter ce biais. L'algorithme de CHIC fonctionnant avec des variances, il faut que la distribution des valeurs soit suffisamment équilibrée, ce qui est le cas des iris. Dans le cas contraire, comme celui de certaines variables acoustiques, il est préférable de découper soi-même les variables selon d'autres critères, ne faisant pas intervenir la variable Y, pour éviter d'introduire de biais dans la discrimination, en utilisant par exemple des quantiles pour avoir des catégories d'effectifs proches.

Le découpage peut se faire selon une théorie (par exemple pour la tension artérielle, découpage selon les seuils de l'hypotension et de l'hypertension). Comme nous ne disposions pas de théorie sur les dimensions des fleurs d'iris, nous avons profité de l'algorithme de découpage automatique proposé par CHIC pour faire essayer divers découpages, en commençant par un découpage identique de toutes les variables, en deux, trois, puis quatre parties. Le découpage en trois parties, procuré par CHIC, a donné de meilleurs modèles : en demandant des seuils entre 0.80 et 0.99, nous avons obtenu 3 graphes séparés, un par espèce, et dans chaque graphe, plus d'une flèche rouge (indices>=0.99), une seule flèche bleue (0.95<=indice<0.99) et aucune flèche d'une autre couleur, donc des indices d'implication statistique tous supérieurs à 0.95. En effet, on peut voir en annexe qu'en fixant le nombre de parties à 2 ou 4 pour le découpage de toutes les variables, les espèces iris-virginica et iris_versicolor apparaissent dans le même graphe. Ensuite nous avons tenté deux découpages en un nombre différent de parties selon les variables des pétales et celles des sépales, qui sont joints en annexe de la partie 2.2.3. On peut y voir qu'un de ces deux graphes implicatifs l'emporte sur tous les autres découpages, y compris ceux avec toutes les variables découpées en 3 parties, c'est celui obtenu en découpant les pétales en trois et les sépales en deux, car il est formé de 3 graphes séparés (un par espèce), il ne contient que des flèches rouges (implication statistique>=0.99) et il est plus concis (moins de variables après découpage). Toutefois, en créant à partir de ce graphe le tableau et le jeu de règles de discrimination comme nous l'avons fait dans la section précédente pour le découpage de chaque variable en 3 parties, nous constatons que sa qualité de discrimination n'est pas meilleure (également 3 erreurs sur les 150 iris), nous privilégions donc le modèle issu du découpage de toutes les variables en trois parties, plus facile à justifier théoriquement ici (nombre de parties des variables de X égal au nombre de catégories de Y).

Ayant perdu de la qualité en passant le nombre de parties du découpage de trois à quatre, nous n'avons pas essayé au-delà de quatre parties.

3 Application aux données d'intonation

Malgré le grand nombre d'études consacrées à l'intonation du français, relativement peu d'entre elles ont porté sur une analyse fine des indices acoustiques et sur un grand nombre de locuteurs. En outre, la reconnaissance automatique des différentes intonations est actuellement faite par des systèmes de reconnaissance fondés sur des modèles cachés, qui ne dévoilent pas les paramètres utilisés pour leurs résultats, et qui ont pour l'instant des résultats assez faibles. L'interprétation objective des indices prosodiques, et tout particulièrement des courbes mélodiques est actuellement de plus en plus utilisée pour l'apprentissage des langues et, dans le domaine de la rééducation, l'apprentissage de la langue (De Bot, 1983), (Bonneau et Colotte, 2011).

Dans une première partie nous décrivons la problématique, dans la suivante les données et leur recodage, puis dans la troisième leur traitement par CHIC, et enfin l'interprétation des résultats.

3.1 La problématique

Lors d'une conversation en français entre deux personnes, il n'y a pas que les mots qui donnent du sens, il y a aussi l'intonation des phrases prononcées : par exemple si le ton du locuteur monte en fin de phrase, l'auditeur en déduit généralement qu'on lui pose une question. En effet, si la question ne comporte pas de marques grammaticales (inversion sujet-verbe « Faut-il », « As-tu » ou ajout en début de phrase de « Est-ce que », « Pourquoi », etc.) c'est par son ***contour mélodique*** qu'elle se distinguera par exemple d'une déclaration prononcée avec une intonation neutre : montant en fin de phrase pour la première et descendant pour la seconde. Delattre (1966) propose dix intonations de base du français, en distinguant quatre niveaux de hauteur de voix. Les schémas de l'article montrent clairement que pour l'auteur, les contours des intonations dites « de base » s'opposent non seulement par leur direction (montante vs descendante) et leurs niveaux de départ et d'arrivée mais également par leur forme : ainsi le contour des questions est caractérisé par un accroissement de hauteur plus fort en fin de contour (la ***vélocité*** est donc élevée en ce point), alors que pour les groupes intonatifs non terminaux (non en fin de phrase, voir Figure 8), l'accroissement est plus fort en début de contour (la vélocité est faible en fin de contour).

Plus précisément, les trois types de contour que nous étudierons dans cet article sont représentés ainsi à l'aide des 4 niveaux de Delattre :

les ***questions*** sans marques grammaticales, que nous noterons ***qis***, avec un contour montant qui part d'un niveau 2 pour aller à un niveau 4 (par exemple la question « Les agneaux ? » en réponse à l'affirmation « Je les ai vus »),
les ***phrases déclaratives***, que nous noterons ***dis***, avec un contour (final) descendant, partant du niveau 2 pour aller au niveau 1 (par exemple la réponse « Les agneaux. », à la question « Qu'as-tu vu ? »),

les ***syntagmes non terminaux***, que nous noterons ***cis***, avec un contour montant partant du niveau 1 pour aller au niveau 2 (par exemple « les agneaux » qui est au début de la phrase « Les agneaux ont vu leur mère », voir Figures 6 et 7).

Dans cette étude, notre objectif est donc de caractériser ces 3 types de contours prosodiques à l'aide d'indices acoustiques, notamment les niveaux de hauteur, la direction et l'allure générale de la pente mélodique en fin de phrase (ou de syntagme). En particulier, nous essaierons de vérifier si la hauteur atteinte en fin de question est effectivement plus élevée que celle des syntagmes non terminaux et des déclarations, et si les questions sont caractérisées par une accélération plus tardive que celle qu'on observe normalement en fin de syntagme non terminal. Enfin, nous essaierons de distinguer les syntagmes non terminaux des fins de phrases déclaratives, y compris quand celles-ci ont été prononcées avec une intonation particulière, comme le doute, qui a pour effet de modifier l'allure de la courbe qui tend à devenir légèrement montante (et non descendante, comme pour une intonation neutre).

3.2 Les données

Ce corpus est issu d'un corpus plus vaste, qui comprend notamment des phrases plus longues, et qui a été enregistré dans le cadre du projet Intonale (2009-2011) supporté par le Comite de Coordination et d'Orientation Scientifique Lorrain: Nancy University, Paul Verlaine University, CNRS,INRIA, INRA, INSERM, CHU.

Nous ne décrivons ici que les éléments nécessaires à notre étude. Celle-ci s'appuie sur 115 enregistrements sonores concernant 5 phrases simples prononcées de 3 façons différentes par 33 locuteurs différents (chaque locuteur n'a prononcé qu'une partie des phrases, comme c'est détaillé plus bas). Nous décrivons d'abord les conditions de l'expérience construite pour enregistrer la parole des locuteurs, puis comment nous avons créé les indices acoustiques et les variables à partir des enregistrements sonores.

3.2.1 Les conditions de l'expérience

Les éléments relatifs aux 115 enregistrements que nous avons analysés sont décrits en utilisant la terminologie des plans d'expérience (Hoc 1983).

Facteur P : « Phrase » :

Voici les cinq phrases (ou parties de phrases) différentes qui ont été étudiées :

P1 : Nos amis
P2 : Les élèves
P3 : Les lamas
P4 : La marelle
P5 : Les agneaux

Facteur T : « Type de phrase » :

Le contexte dans lequel les phrases ont été lues a été conçu pour que les locuteurs prononcent les phrases de manière neutre, sans émotion particulière, avec l'intonation qui respecte chacune des trois catégories qui nous intéressent ici. Certains locuteurs ont cependant parfois donné à certaines phrases une émotion particulière, notamment en fin de phrases déclaratives (expression du doute, par exemple). Nous avons conservé ces

phrases, et notre analyse a donc porté, sur les questions (***qis***), les syntagmes non terminaux (***cis***), et les déclarations (***dis***), ces dernières comportant un certain nombre de phrases non neutres.

C'est ce facteur « type de phrase »qui deviendra la variable Y de notre modèle X→Y

Facteur L : « Locuteur » :

Ce sont 6 hommes et 27 femmes, ayant le français comme langue maternelle. Le nombre de phrases que chaque locuteur a prononcées varie entre 1 et 5.

Facteur G : « Genre »

Ce facteur à 2 modalités H (Homme) et F (Femme).

Facteur S : « Rang du son » :

Pour rendre compte de la totalité du contour mélodique de chaque phrase, nous avons numéroté chaque partie de la phrase. Le numéro 1 correspond à ce qui est enregistré avant le début, le numéro 2 à la première consonne, le numéro 3 à la première voyelle et ainsi de suite. Pour la phrase interrogative « les agneaux ? » les numéros vont de 1 à 7, par contre pour les syntagmes non terminaux (de type « cis »), ils vont au-delà de 7 car pour forcer le locuteur à adopter l'intonation convenable pour « les agneaux », on lui a demandé de prononcer la phrase plus longue « les agneaux ont vu leur mère », et on a annoté le début de phrase incluant le début de la partie suivante pour être sûres d'avoir toute l'information pertinente (voir Figures 6 et 7).

Le plan d'expérience

Nous aurions aimé que ce soit un plan factoriel complet équilibré, qu'il y ait autant d'hommes que de femmes, que chacun prononce les 5 phrases dans les 3 intonations. Ce n'a pas été le cas, aucun locuteur n'a prononcé une même phrase avec des intonations différentes, et ceux qui ont prononcé 4 ou 5 phrases ont utilisé les 2 intonations « qis » (pour les phrases 1, 3 et 5, ou bien 2 et 4) et « dis » (pour les phrases restantes, la 3 étant parfois inutilisée) alors que les autres, qui ont prononcé seulement 2 ou 3 phrases, soit les phrases 1, 3 et 5, soit les phrases 2 et 4, ont utilisé uniquement l'intonation « cis ». Ce plan incomplet et déséquilibré s'explique par le poids de réalisation d'une telle expérience, autant pour ceux qui la subissent en tant que locuteur, que pour ceux qui la font passer et ceux qui transforment les enregistrements sonores en fichiers de données exploitables par des logiciels statistiques. En effet, dans le cadre du projet Intonale, chaque locuteur à prononcé une dizaine de phrases, dont nous n'avons extrait qu'une petite partie, les plus simples, pour initier notre analyse. Puis c'est de façon semi-automatique que nous avons créé le fichier destiné aux analyses statistiques, avec 115 lignes et une cinquantaine de colonnes : il nous a fallu annoter les 115 enregistrements sonores, en extraire les parties correspondant aux phrases visées, et les aligner (leur longueur dépendant de la rapidité d'élocution du locuteur), avant de créer les variables. Ce plan d'expérience n'est pas non plus un plan randomisé dans les règles de l'art, ce qui interdit d'utiliser des techniques paramétriques poussées visant à éliminer les effets des variables de contrôle telles que « locuteur », « phrase », « voyelle », etc. afin d'obtenir des valeurs épurées pour le seul facteur qui nous intéresse ici, le contour mélodique, que nous avons appelé type de phrase, ayant 3 modalités, ***cis***, ***dis*** et ***qis***.

Malgré cela, c'est une chance d'avoir pu obtenir de telles données, qui se sont avérées adaptées à notre problématique une fois choisis les indices acoustiques et la technique statistique appropriés.

Pour notre étude, tous les facteurs de ce plan d'expérience ont été pris en compte dans le modèle X→Y que nous cherchons à établir : ***Locuteur***, ***Rang du son*** dans la création des variables formant X, ***Phrase*** et ***Genre*** ont formé la matrice Z des variables à contrôler et ***Type de phrase*** est devenu la variable Y du modèle X→Y que nous cherchons à établir.

3.2.2 La construction des variables de X

Les variables de X dérivent des indices acoustiques, eux-mêmes construits à partir des enregistrements sonores.

Les indices acoustiques

Ils ont été extraits du fichier son en utilisant diverses techniques ainsi que le logiciel PRAAT (Boersma et Weenink, Version 5.4.10 du 27 June 2015, www.praat.org). Nous donnons ci-dessous quelques éléments de leur construction sans entrer dans les détails, le but étant seulement d'établir leur signification et de juger de leur précision pour pouvoir interpréter ensuite les résultats obtenus par CHIC.

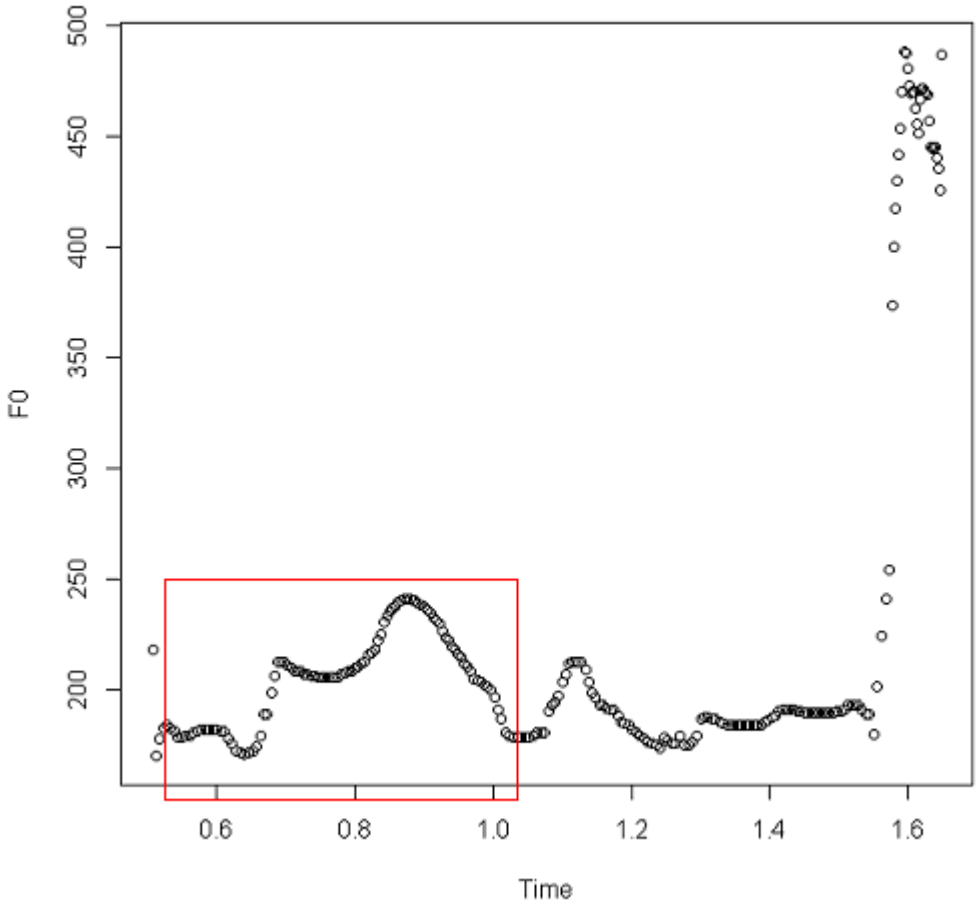


Figure 6 – Valeurs de F0 en fonction du temps pour la phrase « Les agneaux ont vu leur mère », donc de type « cis » prononcée par un locuteur. La partie de la phrase qui nous intéresse « Les agneaux », incluant la voyelle suivante « on » est encadrée de rouge.

Le premier de ces indices représente la fréquence fondamentale, F0, qui correspond à la hauteur de la voix.. Dans la figure 6, on a représenté les valeurs brutes de F0 en fonction du temps lors de la prononciation d'une phrase commençant par « les agneaux ». Puis on a repéré la partie du signal correspondant à chaque son, comme indiqué dans la figure 7. Bien sûr, les frontières verticales dessinées pour séparer les sons ne peuvent pas toujours être placées de manière précise, comme c'est le cas ici pour le passage entre les deux voyelles /o/ et /on/. Le travail de segmentation, qui demande beaucoup de temps et de rigueur, a été réalisé par l'une des auteures.

Trouver des indices qui mesurent l'évolution de F0 quand on passe d'un son à un autre (hauteur, croissance, décroissance, stabilité, vitesse, etc.) n'est pas simple, comme on peut le remarquer dans la figure 7. Les pentes des droites rouges, obtenues par régressions sur des points consécutifs, auraient pu fournir un coefficient d'accroissement de F0. D'autres indices exprimant l'accroissement ont été testés, mais finalement c'est la vélocité instantanée calculée par PRAAT, donc une valeur en chaque point, qui a été retenue à sa place et qui est notre deuxième indice.

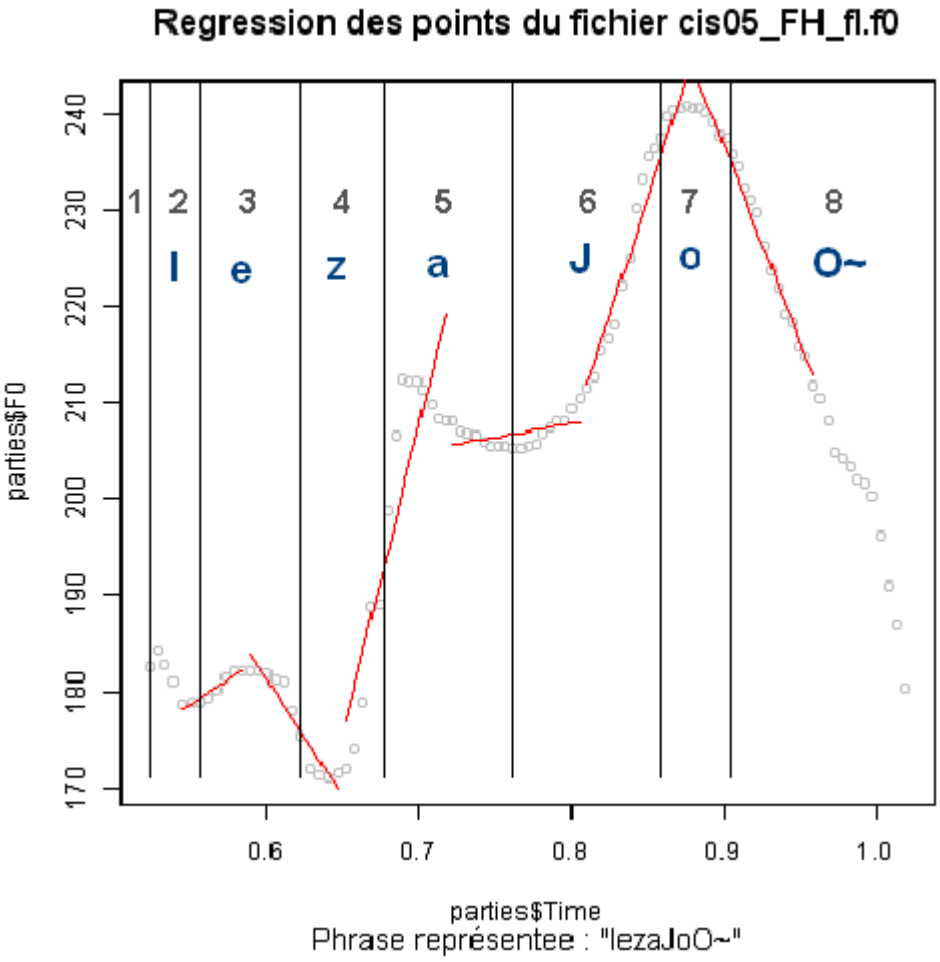


Figure 7 – La phrase a été découpée en sons par les traits verticaux, et les droites rouges ont été obtenues par des régressions sur les points allant du milieu d'une bande au milieu de la suivante.

Les 29 variables de X

Selon les enregistrements le nombre de valeurs des indices diffère pour chaque son (il y en a plus quand le locuteur parle lentement), ce qui rend difficile leur utilisation dans des modèles. Nous en avons calculé un seul par son et par locuteur en appliquant diverses fonctions sur l'ensemble des valeurs successives pour les fréquences F0 du son, max, min, moyenne (mean), final, amplitude (ex_size), durée, force/intensité (mean_int), et max ainsi que final pour la velocité. Les valeurs de ces derniers indices pour chacune des 3 positions de voyelles (numérotées 3, 5 et 7) ainsi que les différences de maxF0 entre 2 voyelles voisines (3 et 5, 5 et 7) ont donné les 29 variables suivantes :

3maxF0, 5maxF0, 7maxF0, 3minF0, 5minF0, 7minF0, 3meanF0, 5meanF0, 7meanF0, 3finalF0, 5finalF0, 3ex_size, 5ex_size, 7ex_size, 3duration, 5duration, 7duration, 3mean_int, 5mean_int, 7mean_int, 3max_vel, 5max_vel, 7max_vel, 3final_vel, 5final_vel, 7final_vel, 7finalF0, maxF0_v5_3, maxF0_v7_5.

Pour faciliter l'interprétation des fréquences F0, nous les avons transformées avant traitement statistique en « semi-tons » par la formule Semitone(F0) = 12 $\log_2$(F0/base) où nous avons choisi base = 100. Puis pour contrebalancer l'effet « locuteur », notamment les grandes différences de hauteur de voix (ici exprimée en semi-tons) femme/homme, nous avons normalisé en retirant à chaque mesure la moyenne du locuteur concerné.

3.3 Le traitement par CHIC

Nous recherchions un modèle explicite des données permettant de différencier les 3 catégories de phrases mélodiques.

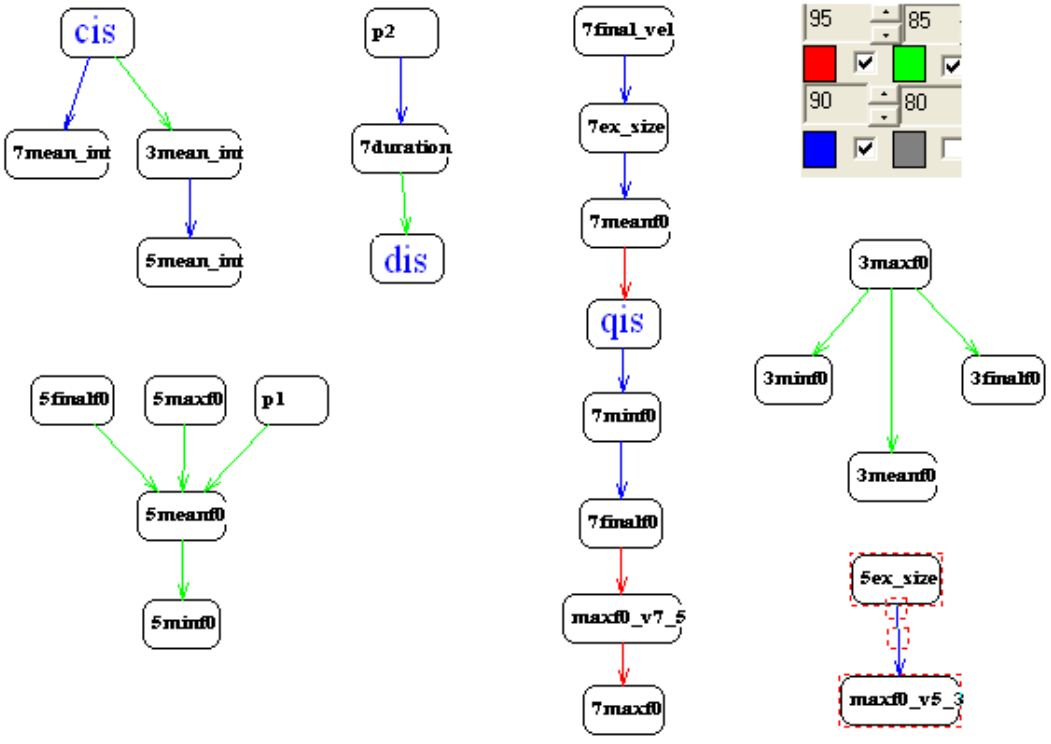


Figure 8 – Graphe implicatif des variables des données mélodiques : 29 variables acoustiques numériques recodées sur [0 ; 1], 3 catégories, 5 phrases, les couleurs correspondant aux seuils cochés en haut à droite.

Le grand nombre de variables « explicatives » que nous avons créées, et les liaisons fortes entre elles dues à leur mode de construction nous interdisaient sans correction préalable complexe des données l'utilisation de la plupart des méthodes de discrimination des statistiques classiques (décrites précédemment). Quant à l'arbre de décision que nous avons obtenu, il n'utilisait que 2 ou 3 variables parmi les 29 créées pour discriminer les 3 catégories, ce qui donnait un modèle trop pauvre en regard de la théorie de Delattre. Grâce à CHIC, nous avons pu utiliser l'ASI sur la totalité des variables afin d'extraire des modèles de contour mélodique à partir d'indices acoustiques. Nous décrivons ici ceux qui ont été obtenus à partir des graphes implicatifs en répétant la démarche suivie pour les iris de Fisher, d'abord, en Figure 8, avec la variable Y des 3 catégories (***qis, cis, dis***), les 29 variables quantitatives X recodées sur l'intervalle [0 ; 1], ainsi que 2 variables de contrôle Z (phrases P1 à P5, genre H/F), puis en Figure 9, avec chaque variable de X découpée en 3 catégories.

L'ASI sur les 29 variables recodées sur [0 ; 1]

Notre but premier étant de discriminer les 3 catégories de Y, pour le graphe implicatif de la figure 8, nous avons du « descendre » jusqu'au seuil de 0.85 pour faire apparaitre la catégorie 'dis'.

On peut voir que la catégorie ***qis*** (questions) est nettement discriminée par le son 7, qui correspond à la dernière voyelle de la phrase (par exemple le « i » de « nos amis »), qui doit avoir des valeurs élevées de F0, que ce soit le min, le max, la moyenne, la valeur finale, la différence entre le max et le min (7ex_size), la vélocité finale, ainsi que la différence entre le max sur cette voyelle et le max sur la voyelle précédente (maxF0_v7_5). Nous retrouvons donc bien pour les questions la forte vélocité en fin de contour signalée par Delattre, ainsi que la forte montée de tonalité pour la dernière voyelle. Par contre nous ne retrouvons pas pour ***dis*** et ***cis*** les hypothèses de Delattre, mais d'autres informations qui restent à expliquer : ***dis*** est discriminée par la durée importante de la dernière voyelle, et ***cis*** par l'importance de l'intensité avec laquelle chaque voyelle est prononcée. De plus, nous voyons apparaitre la spécificité de 2 phrases, indépendante de l'intonation, ***p1*** dont la deuxième voyelle (numéro 5) a des valeurs fortes pour F0, et ***p2*** dont la troisième et dernière voyelle (numéro 7) est plus longue. Pour p2, il s'agit de la variabilité inter-vocalique, le /e/ est déjà long et en plus il est devant une consonne allongeante. Pour trouver des relations plus intéressantes entre les catégories ***dis*** et ***cis*** et les 29 variables, il faut créer des variables correspondant aux valeurs faibles, donc les découper en intervalles.

L'ASI sur les 29 variables découpées en 3 catégories chacune

Nous avons scindé la distribution de chacune de ces variables en trois intervalles en regroupant dans le premier intervalle les valeurs les plus basses de la distribution, dans le deuxième les valeurs moyennes, et dans le troisième les valeurs les plus élevées.

Comme le montre la Figure 9, le graphe implicatif a spontanément séparé et regroupé entre eux les trois types de phrases analysées dans cette étude, et associé à chacun de ces types des groupes de variables (graphes numérotés de 1 à 3). L'importance de la voyelle finale, ici la troisième voyelle, bien connue dans le domaine

prosodique, apparaît de manière très claire puisque ce sont les valeurs fréquentielles de cette voyelle qui sont associées par le logiciel à chacun des trois types de phrase.

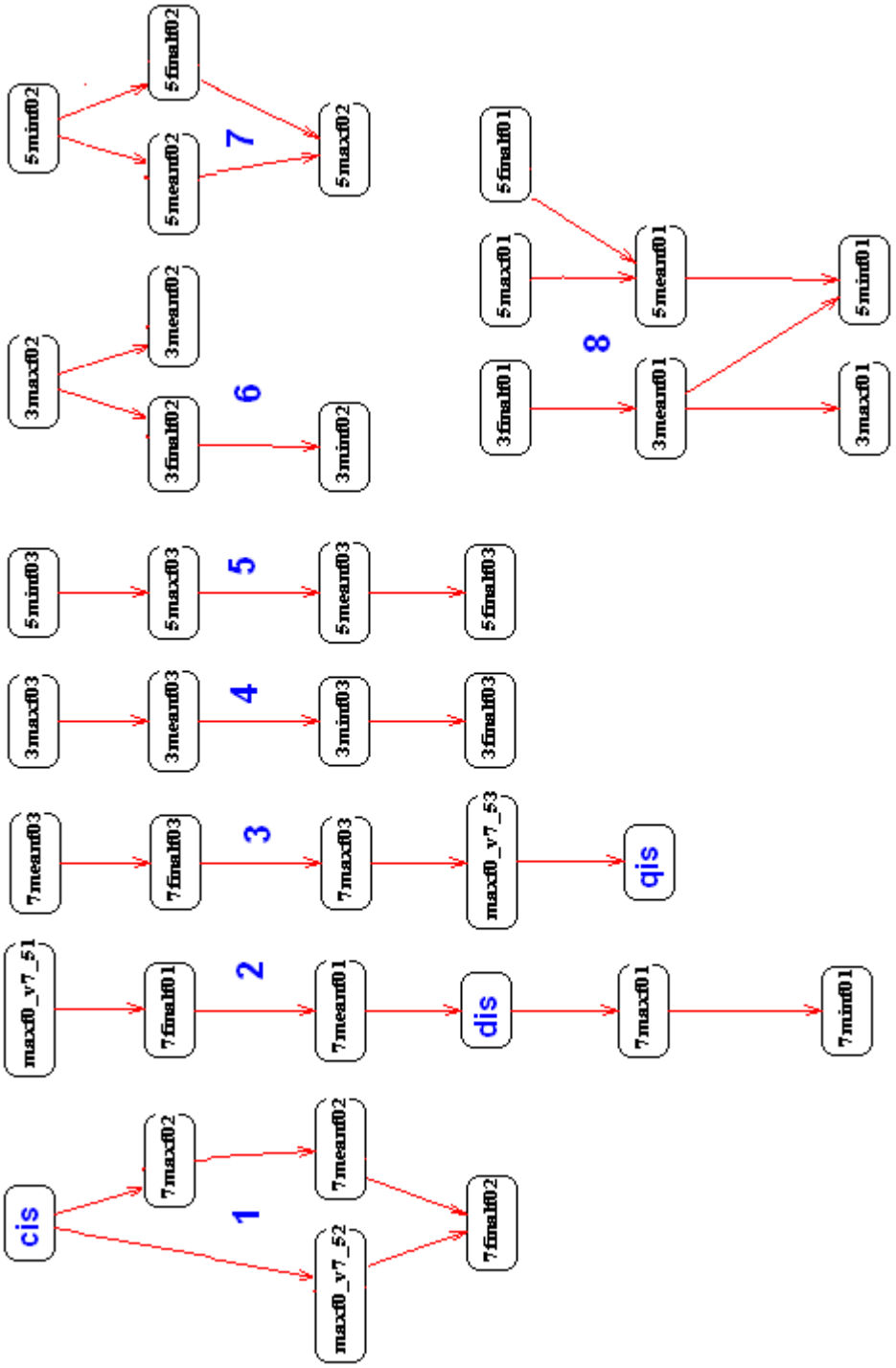


Figure 9 – Graphe implicatif comprenant toutes les valeurs de p>0.9999 (flèches rouges) pour les variables concernant F0. Chaque graphe séparé est numéroté en bleu.

Regardons maintenant quelles sont les variables précises et les régions fréquentielles (intervalles de distribution) qui sont associées par CHIC aux phrases. Les variables associées aux questions (qis) et aux déclarations (dis) sont la valeur moyenne,

maximale, finale, et minimale de la fréquence fondamentale pour la voyelle finale, ainsi que la différence entre la valeur maximale de la voyelle finale et celle de la pénultième (l'ordre des variables est légèrement différent selon les types de phrases). Pour les questions (qis), nous constatons que, quelle que soit la variable, ce sont les valeurs élevées de sa distribution qui sont significatives. Ce résultat est en adéquation avec les descriptions prosodiques des questions sans marque grammaticale (contour mélodique montant entre la pénultième et la voyelle finale, fréquences finales élevées). En ce qui concerne les déclarations (dis), ce sont les valeurs basses de la distribution qui sont significatives. Encore une fois ce résultat est en adéquation avec les descriptions prosodiques des phrases déclaratives (contour descendant, fréquences finales basses).

La caractérisation des groupes nominaux sujets (cis) est plus délicate. Les segments de ce type, situés à l'intérieur d'une phrase, possèdent en général un contour mélodique montant, mais de moindre envergure que celui des questions. Les valeurs fréquentielles caractéristiques de la dernière voyelle de ce groupe sont en général moins élevées que celles des questions et plus élevées que celles des déclarations. De manière intéressante, on voit également que ce sont les valeurs moyennes de la distribution des variables qui sont associées aux groupes nominaux sujets.

On s'aperçoit également que maxF0_V7_5, qui est la différence entre le max de la voyelle précédente et celui de la voyelle finale, est très important (troisième intervalle) pour les questions, moyen pour les groupes nominaux sujets, et faible pour les déclarations. Ce qui va encore une fois dans le sens de la description en niveaux de Delattre, qui indiquent des passages d'un niveau 2 à 4 pour les questions 1 à 2 pour les syntagmes non terminaux, et 2 à 1 pour les déclarations.

Nous continuons actuellement notre interprétation pour les graphes de 4 à 8, dans lesquelles interviennent d'autres variables que la catégorie des phrases. Ces analyses seront fournies ultérieurement sur demande.

4 Conclusion et perspectives

Comme nous l'avons souligné lors de l'analyse de nos résultats, le logiciel CHIC nous a permis 1) de confirmer les descriptions prosodiques théoriques, notamment celles de Delattre, à l'aide d'indices spécialement conçus pour vérifier ses schémas ; 2) de disposer d'éléments précis sur la manière dont ces indices codent les principales intonations du français

Le modèle que nous cherchions était un modèle de discrimination, du type X→Y, un modèle courant en statistique, qui pouvait être établi par des méthodes statistiques classiques mais également par des méthodes plus récentes de l'apprentissage automatique, et par l'ASI, à travers CHIC. Nous les avons testées et comparées sur les iris de Fisher, jeu de données ayant des caractéristiques proches du nôtre (variables de X quantitatives très liées, variable Y à 3 catégories, discrimination facile de Y à partir de X même pour un observateur inexpérimenté) et constaté que CHIC nous fournissait de façon plus ergonomique un modèle des données tout aussi fiable, mais plus simple, tout en étant plus riche, qui pouvait servir à la discrimination, moyennant toutefois une utilisation détournée de l'ASI (qui n'est pas une méthode de discrimination).

L'ASI est une méthode statistique de traitement de données catégorielles qui a été étendue aux variables quantitatives à condition de les recoder sur l'intervalle [0 ; 1], mais il est aussi possible de les transformer en variables catégorielles en les découpant en plusieurs intervalles. Nous avons comparé sur les iris de Fisher le modèle des données en recodant les variables sur [0 ; 1] puis en les découpant en plusieurs parties. Les différents modèles obtenus ne se recouvrent pas, chacun apporte des informations spécifiques, mais c'est un découpage de chaque variable de X selon 3 intervalles qui fournit le modèle de discrimination souhaité X→Y.

Le parallèle fait avec les iris nous a conduites à utiliser CHIC pour traiter nos données d'intonation. En effet, avec le plan d'expérience que nous avions, les locuteurs adaptaient leur intonation au type de phrase demandé, en produisant un contour mélodique mesuré par nos indices acoustiques. Et selon la théorie de Delattre, les indices acoustiques devaient permettre de discriminer les contours mélodiques. Nous étions donc dans un cas de discrimination « facile », comme pour les iris de Fisher, et on aurait pu choisir n'importe laquelle des diverses méthodes exposées dans la première partie pour une discrimination de bonne qualité. Aux raisons qui nous avaient fait choisir CHIC pour les iris : avoir facilement un modèle explicite et riche, s'en est ajoutée une autre : l'ensemble des variables acoustiques s'avérant plus complexes que les dimensions des fleurs d'iris, la plupart des autres méthodes s'appliquaient mal.

Nous avons étudié les 2 graphes implicatifs obtenus pour les données d'intonation, et c'est également celui avec découpage en 3 de chaque variable qui a permis de retrouver au mieux les hypothèses de Delattre. Mais l'autre graphe a apporté certaines informations supplémentaires, justes mais qui ne se sont pas avérées pertinentes car elles n'ajoutent rien à la théorie de Delattre sur les intonations de base du français.

Il nous reste à interpréter toutes les autres sorties produites par CHIC (jointes en annexe), refaire le traitement en découpant en 4 les variables issues de F0, pour voir si on peut retrouver les 4 niveaux de la théorie de Delattre, et estimer le pourcentage de schémas intonatifs qui peuvent être identifiés à l'aide des indices choisis.

Pour conclure, à la question du titre « Pourquoi et comment transformer des variables quantitatives en catégorielles ? », nous pouvons répondre ainsi :

« Pourquoi ? » : pour prédire une variable catégorielle, quand on dispose de variables explicatives quantitatives très liées entre elles, il est mieux de les transformer en variables catégorielles.

« Comment ? » : si on ne dispose pas de théorie donnant des seuils de coupes, et donc le nombre de parties pour chaque variable, il est mieux d'essayer d'abord un découpage de chaque variable quantitative en k intervalles pour prédire k catégories. Le choix de ces intervalles peut se faire avec des quantiles pour obtenir des effectifs voisins.

Mais c'est une heuristique que nous proposons ici, bien sûr, à éprouver sur d'autres données et d'autres problématiques.

Références

[1] Baillargeon, (2000), *La régression linéaire*, Edition des trois Sources, Quebec

[2] Besse, P. (2003), *Pratique de la modélisation Statistique*, Document interne du Laboratoire de Statistique et Probabilités, Université Paul Sabatier de Toulouse, http://www.math.univ-toulouse.fr/~besse/pub/modlin.pdf

[3] Bonneau, A. et Colotte, V (2011), Automatic Feedback for L2 Prosody Learning, in *Speech Technologies, Book 2*, I. IPSIC (editor), Intech, June 2011, 55-70, http://hal.inria.fr/inria-00579255/en

[4] Cadot, M ; (2006), Extraire et valider les relations complexes en sciences humaines : statistiques, motifs et règles d'associatio, Thèse de l'université de Franche-Comté.

[5] Dagnelie P., (2003), *Principes d'expérimentation : planification des expériences et analyse de leurs résultats*. Grenoble, Presses Agronomiques de Gembloux, Edition électronique : http://www.dagnelie.be

[6] Droesbeke J.-J., Fine J., éditeurs. (1996) *Inférence non paramétrique, les statistiques de rangs*. Journées d'Etude en Statistiques de l'Association pour la Statistique et ses Utilisations, Edition de l'Université de Bruxelles, Ellipses

[7] De Bot, K. (1983), Visual feedback on intonation I: Effectiveness and induced practice behaviour, *Language and Speech* , **Vol.6 , No.4**, 331–350

[8] Delattre, Pierre (1966), Les dix intonations de base du français, *The french Review*, **vol 40**, 1-14.

[9] Fisher,R.A. (1936), The use of multiple measurements in taxonomic problems *Annual Eugenics*, **7, Part II**, 179-188

[10] Gras R. et collaborateurs (1996), L'implication statistique, une nouvelle méthode exploratoire de données, La pensée sauvage, Grenoble

[11] Gras, R., Régnier, J.-C., Marinica, C., Guillet, F. (2013), L'analyse statistique implicative : méthode exploratoire et confirmatoire à la recherche de causalités, 2ème éd., Cépaduès, Toulouse

[12] Hoc J.-M. (1983), L'analyse planifiée des données en psychologie, PUF Paris

[13] Logiciel CHIC : Classification Hiérarchique Implicative et Cohésitive, Version 6.0, Copyright (c) 2012

[14] Mitchell, T. (1997). *Machine Learning*, McGraw Hill

[15] Morin H., (1999), *Théorie de l'échantillonnage*, Les presses de l'université, Laval

[16] Morineau, A., Nakache, J.-P., Krzyzanowski, C., (1996), *Le modèle log-linéaire et ses applications*, Cisia-Ceresta, Paris.

[17] Nakache, J.-P., Confais J.,(2003), Statistique explicative appliquée : analyse discriminante, modele logistique, segmentation. Editions Technip, Paris, France

[18] Prum, B., (1996), *Modèle linéaire. Comparaison de groupes et régression*, Eyrolles, Collection INSERM, Paris

[19] Breiman L., Friedman J. H., Olshen R. A., and Stone, C. J. (1984), *Classification and Regression Trees*. Wadsworth.

[20] Ripley, B. D. (1996), Pattern Recognition and Neural Networks. Cambridge.

[21] Schwartz, D., (1991), Méthodes statistiques à l'usage des médecins et des biologistes, Flammarion, Paris

[22] Siegel S., Castellan N.J. Jr.,(1988), *Nonparametric statistics for the behavorial sciences*, McGraw-Hill, London.

[23] Weston J. and Watkins, C. (1998), *Multi-class support vector machines*. Technica l Report CSD-TR-98-04,Royal Holloway, University of London, Department of Computer Science

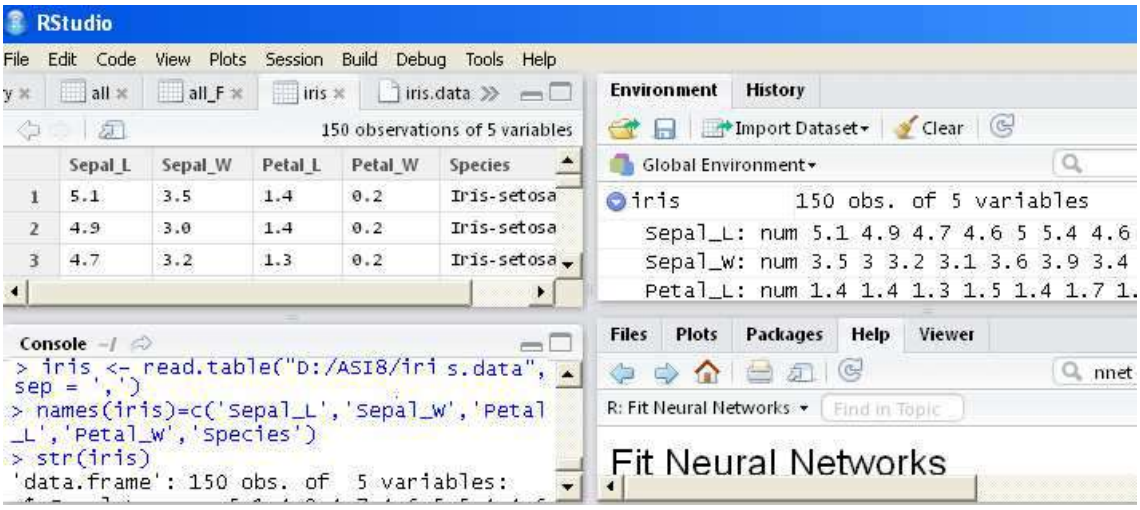
Annexe de la partie 2.2.1 : Scripts R et leur résultats

L'import des données dans l'environnement R

Les iris ont été récupérées dans UCI repository () sous forme d'un fichier iris.data, de 5 colonnes, une par variable, séparées par des virgules, et de 150 lignes, sans intitulés. On les importe dans R, puis on nomme les variables et on affiche leur structure de la façon suivante, avec en bleu les lignes de code écrites après le symbole d'invite « > » et en noir le résultat de la commande, quand il peut s'afficher.

```
> iris <- read.table("D:/ASI8/iris.data",sep = ',')
> names(iris)=c('Sepal_L','Sepal_W','Petal_L','Petal_W','Species')
> str(iris)
'data.frame':  150 obs. of  5 variables:
 $ Sepal_L: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal_W: num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal_L: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal_W: num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species: Factor w/ 3 levels "Iris-setosa",..: 1 1 1 1 1 1 1 1 1 1
...
```

On peut voir dans la copie d'écran ci-dessous de RStudio (environnement de programmation de R disponible sous Windows, http://www.rstudio.com/) en haut à gauche les données importées dans R sous forme tabulaire :



Indication pour les grands débutants en R : quand on veut utiliser une fonction d'un package, il faut le télécharger auparavant puis le charger en mémoire par la commande library(nom_package) tapée dans la console. Il n'y a plus qu'à copier-coller à la suite les commandes (en les validant une à une dans l'interface classique de R, ou en un seul bloc dans la console de RStudio), et une fois celles-ci validées (en enfonçant la touche « entrée »), observer les résultats. Attention à ne pas mettre des minuscules à la place de majuscules ou inversement, R est un langage sensible à la casse.

La régression linéaire généralisée

En R, comme dans beaucoup d'autres langages statistiques, c'est la fonction ***glm*** (Generalized Linear Model) qui permet d'obtenir les divers modèles, en faisant varier

les arguments. Mais bien sûr il y en a d'autres plus spécifiques adaptées à chaque cas, et il en apparaît régulièrement de nouvelles disponibles dans de nouveaux packages R (en un an, le nombre de packages R a plus que doublé).

Dans l'aide sur ***glm***, obtenue en tapant help(glm), on apprend que la fonction de base a été implémentée par Simon Davies, et réécrite et complétée régulièrement depuis.

```
glm(formula, family = gaussian, data, weights, subset,    na.action,
start = NULL, etastart, mustart, offset,    control = list(...), model
= TRUE, method = "glm.fit",    x = FALSE, y = TRUE, contrasts = NULL,
...)
```

Cette écriture compliquée se simplifie si on omet les arguments suivis d'un signe égale, ils sont alors remplacés par leur valeur par défaut, qui suit leur signe égale. Par exemple la droite de régression des iris-setosa, en bleu dans la figure 2, s'obtient par

```
> setosa=iris[iris$Species=='iris-Setosa',]
> model.setosa=glm(setosa$Petal_L~setosa$Sepal_L)
> summary(model.setosa)
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.81377    0.34390   2.366   0.0221 *
setosa$Sepal_L  0.12989    0.06853   1.895   0.0641 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

On a bien retrouvé les coefficients de l'équation de régression des iris-setosa. L'argument family a été omis, donc sa valeur est gaussian qui est la loi normale. Il convient de contrôler que les résidus suivent bien la loi normale :

```
> shapiro.test(model.setosa$residuals)
        Shapiro-Wilk normality test
data:  model.setosa$residuals
W = 0.9671, p-value = 0.1762
```

On peut juger que c'est le cas puisque puisque p>0.05

Le modèle logistique

Dans ce cas, on doit créer une variable dichotomique (ici versicolor) qui prend la valeur 1 pour l'espèce qu'on souhaite discriminer, et 0 sinon. L'argument family change pour indiquer que la fonction est un logit.

```
> mlogistic_versicolor=glm(versicolor~Petal_L+Petal_W+Sepal_L+Sepal_W,
data=iris, family=binomial(link=logit))
> summary(mlogistic_versicolor)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1185  -0.7929  -0.3835   0.7980   2.1149
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   7.3229     2.4980   2.932 0.003373 **
Petal_L       1.2993     0.6823   1.904 0.056864 .
Petal_W      -2.7043     1.1627  -2.326 0.020021 *
Sepal_L      -0.2527     0.6495  -0.389 0.697154
Sepal_W      -2.7794     0.7859  -3.537 0.000405 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Pour les modèles dérivés du modèle linéaire avec une variable Y à expliquer de plus de 2 catégories, nous renvoyons aux nombreux ouvrages en ligne sur les fonctions statistiques R. Les données des iris y sont traitées de multiples façons.

Annexe de la partie 2.2.2 : scripts R et leur résultats

Pour toutes ces méthodes issues de l'apprentissage automatique, le script de l'application aux iris a été adapté des exemples figurant dans l'aide des packages. L'ensemble des 150 iris numérotés de 1 à 150, espèce après espèce, a été découpé en deux parties, la partie destinée à l'entraînement a ses numéros dans l'ensemble « samp » obtenu en prenant au hasard 25 numéros de chaque espèce, l'ensemble « test » ayant les numéros dans l'ensemble complémentaire, codé « -samp ». La qualité de prévision se mesure uniquement sur l'ensemble test. Nous avons néanmoins fourni les tables de confusion croisant classe prédite et classe observée sur les deux ensembles afin de pouvoir confronter ces méthodes aux autres. Bien sûr, si on faisait un autre tirage de « samp », on pourrait obtenir des résultats légèrement différents, et également si on modifiait les paramètres des fonctions utilisées, en les adaptant aux caractéristiques des données, ce que nous n'avons pas fait ici, cette annexe visant seulement à illustrer la partie théorique correspondante.

Les données d'iris sont conservées de la partie précédente, sous la même forme. Voici le code permettant de générer les 25 numéros d'iris par espèce tirés au hasard (on a passé d'une couleur différente les numéros correspondant à des espèces différentes) :

```
> samp <- c(sample(1:50,25), sample(51:100,25), sample(101:150,25))
> samp
 [1]  21  48  31  14  44  15   9  36  40  16   2  39   6  24  35  33
8  10
[19]  27  26  12  41  37  11  46  65  84  56  99  91  82  52  93  79
77  85
[37]  60  90  75  64  95  59 100  58  87  57  63  74  94  96 119 106
104 114
[55] 150 122 123 111 107 103 131 145 115 117 120 147 141 110 121 132
127 148
[73] 133 125 134
```

Les réseaux neuronaux

Package nnet de R mis à jour le 29 juin 2015, auteur Brian Ripley, http://cran.r-project.org/web/packages/nnet/nnet.pdf

```
model.nnet <- nnet(Species ~ ., data = iris, subset = samp, size = 2, rang = 0.1,
decay = 5e-4, maxit = 200)
# table de confusion pour les données de test
table(iris$Species[-samp], predict(model.nnet, iris[-samp,], type = "class"))
# pour info table de confusion pour les données d'entraînement
table(iris$Species[samp], predict(model.nnet, iris[samp,], type = "class"))
```

Données test : 4 erreurs sur 75				Données d'entraînement			
	Iris-setosa	Iris-versicolor	Iris-virginica		Iris-setosa	Iris-versicolor	Iris-virginica
Iris-setosa	25	0	0	Iris-setosa	25	0	0
Iris-versicolor	0	21	4	Iris-versicolor	0	24	1
Iris-virginica	0	0	25	Iris-virginica	0	1	24

Les « Support Vector Machine » multiples

Package e1071 de R mis à jour le 19 février 2015, auteurs David Meyer et al., http://cran.r-project.org/web/packages/e1071/e1071.pdf

model.svm <- svm(iris[samp,1:4], iris[samp,5])
# table de confusion pour les données de test, puis d'entrainement
table(iris$Species[-samp], predict(model.svm, iris[-samp,1:4]))
table(iris$Species[samp], predict(model.svm, iris[samp,1:4]))

Données test : 3 erreurs sur 75				Données d'entraînement			
	Iris-setosa	Iris-versicolor	Iris-virginica		Iris-setosa	Iris-versicolor	Iris-virginica
Iris-setosa	25	0	0	Iris-setosa	25	0	0
Iris-versicolor	0	22	3	Iris-versicolor	0	24	1
Iris-virginica	0	0	25	Iris-virginica	0	1	24

Les arbres de décision

Package rpart de R mis à jour le 29 juin 2015, auteurs Terry Therneau et al., http://cran.r-project.org/web/packages/rpart/rpart.pdf

model.arbre_dec=rpart(Species ~ Sepal_L + Sepal_W + Petal_L + Petal_W, data=iris, subset=samp)
# table de confusion pour les données de test, puis d'entrainement
table(iris$Species[-samp], predict(model.arbre_dec, iris[-samp,], type = "class"))
table(iris$Species[-samp], predict(model.arbre_dec, iris[-samp,], type = "class"))

Données test : 4 erreurs sur 75				Données d'entraînement			
	Iris-setosa	Iris-versicolor	Iris-virginica		Iris-setosa	Iris-versicolor	Iris-virginica
Iris-setosa	25	0	0	Iris-setosa	25	0	0
Iris-versicolor	0	23	2	Iris-versicolor	0	25	0
Iris-virginica	0	2	23	Iris-virginica	0	2	23

```
# Voici le modèle
```
model.arbre_dec

n= 75
node), split, n, loss, yval, (yprob)

* denotes terminal node

1) root 75 50 Iris-setosa (0.33333333 0.33333333 0.33333333)
  2) Petal_L< 2.35 25 0 Iris-setosa (1.00000000 0.00000000 0.00000000) *
  3) Petal_L>=2.35 50 25 Iris-versicolor (0.00000000 0.50000000 0.50000000)
    6) Petal_W< 1.65 27 2 Iris-versicolor (0.00000000 0.92592593 0.07407407) *
    7) Petal_W>=1.65 23 0 Iris-virginica (0.00000000 0.00000000 1.00000000) *

Annexe de la partie 2.2.3 : autres résultats sur les iris avec CHIC

D'abord des éléments supplémentaires pour le découpage des variables en 3 parties, puis un rapide tour des autres essais de découpage : chaque variable en 2 parties, les dimensions des pétales en 2 parties et celles des sépales en 3 parties, puis chaque variable en 4 parties.

Découpage de chacune des 4 dimensions de la fleur en 3 parties

En figure 5, de la section 2.2.3, à droite, nous avons représenté le graphe implicatif avec 3 graphes bien séparés, un par type d'iris, qui nous montrait l'association forte entre les 4 variables de dimension de la fleur et le type d'iris. Ci-dessous figure une sortie de CHIC montrant les fréquences des couples de variables :

Nous avons encadré de rouge les nombres de co-occurences de chacune des 3 valeurs de largeur de pétale avec le type d'iris. On constate que 33% des 150 fleurs sont de l'espèce « iris-setosa » et que la largeur de leur pétale est petite (Petal_W1), que 32% sont de l'espèce « iris-versicolor » et que la largeur de leur pétale est moyenne (Petal_W2), et que 31% sont de l'espèce « iris-virginica » et que la largeur de leur pétale est grande (Petal_W3), soit un taux de ***coïncidence***[6] de 96% obtenu entre l'espèce d'iris et la mesure de 1 à 3 de la largeur de ses pétales. De la même façon, la coïncidence est de 94% pour la longueur des pétales (en bleu), de 73% pour la longueur des sépales (en orange), et seulement de 55% pour la largeur des sépales (en vert).

[6] Nous utilisons ici le terme de ***coïncidence*** à la place du terme plus statistique de ***prédiction*** dans la mesure où nous raisonnons à partir de comptages faits sur la totalité des iris, sans utiliser d'échantillonnage ou de découpage en plusieurs sous-ensembles (entraînement et test, etc.).

	Sepal_L1	Sepal_L2	Sepal_L3	Sepal_W1	Sepal_W2	Sepal_W3	Petal_L1	Petal_L2	Petal_L3	Petal_W1	Petal_W2	Petal_W3	Iris-setosa	Iris-versicolor	Iris-virginica
Sepal_L1		0.00	0.00	0.05	0.13	0.13	0.27	0.04	0.00	0.27	0.03	0.01	0.27	0.03	0.01
Sepal_L2			0.00	0.17	0.10	0.08	0.07	0.23	0.06	0.07	0.21	0.07	0.07	0.21	0.08
Sepal_L3				0.09	0.22	0.03	0.00	0.09	0.25	0.00	0.10	0.24	0.00	0.09	0.25
Sepal_W1					0.00	0.00	0.01	0.20	0.11	0.01	0.20	0.11	0.01	0.18	0.13
Sepal_W2						0.00	0.13	0.15	0.17	0.13	0.14	0.18	0.13	0.15	0.17
Sepal_W3							0.20	0.01	0.03	0.20	0.01	0.03	0.20	0.01	0.03
Petal_L1								0.00	0.00	0.33	0.00	0.00	0.33	0.00	0.00
Petal_L2									0.00	0.00	0.31	0.05	0.00	0.32	0.04
Petal_L3										0.00	0.03	0.27	0.00	0.01	0.29
Petal_W1											0.00	0.00	0.33	0.00	0.00
Petal_W2												0.00	0.00	0.32	0.03
Petal_W3													0.00	0.01	0.31
Iris-setosa														0.00	0.00
Iris-versicolor															0.00
Iris-virginica															

Ce tableau indique donc comment chaque dimension de la fleur prise séparément peut contribuer à prédire son espèce, alors que le graphe implicatif (figure 5 à droite) montre quelle combinaison de variables permet de prédire l'espèce, les détails se trouvant dans le tableau 2.

Un autre traitement de CHIC, « la réduction des variables », donne les résultats suivants :

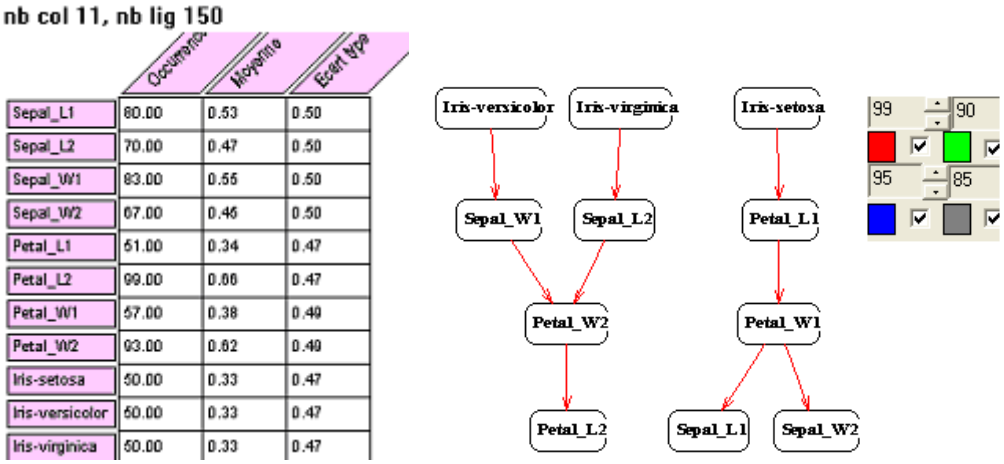
Calcul avec un seuil d'équivalence fixé à 0.9 :
Sepal_L1 Petal_W1 Sepal_W3 Iris-setosa Petal_L1 représentant de la classe : Petal_W1
Iris-versicolor Petal_W2 Petal_L2 Sepal_W1 Sepal_L2 représentant de la classe : Petal_W2
Petal_L3 Sepal_L3 Iris-virginica Petal_W3 représentant de la classe : Iris-virginica

Cette fois, toutes les variables sauf Sepal_W2 se retrouvent « équivalentes au seuil de 0.9 » à une espèce d'iris. Cette information, qui ne fait que conforter celle donnée par le graphe implicatif dans le cas des iris, peut s'avérer très utile quand on dispose d'un grand nombre de variables dont certaines très liées entre elles, pour obtenir un graphe implicatif avec un nombre réduit de « classes de variables ».

Découpage de chacune des 4 dimensions de la fleur en 2 parties

Le découpage automatique de petal_L, petal_W, sepal_L et Sepal_W en 2 se fait selon les bornes suivantes :

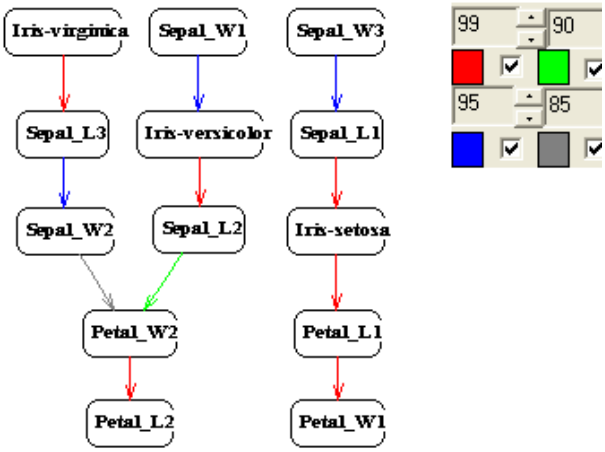
	Sepal_L	Sepal_W	Petal_L	Petal_W
Partie 1	de 4.3 à 5.8	de 2 à 3	de 1 à 3	de 0.1 à 1
Partie 2	de 5.9 à 7.9	de 3.1 à 4.4	de 3.3 à 6.9	de 1.1 à 2.5



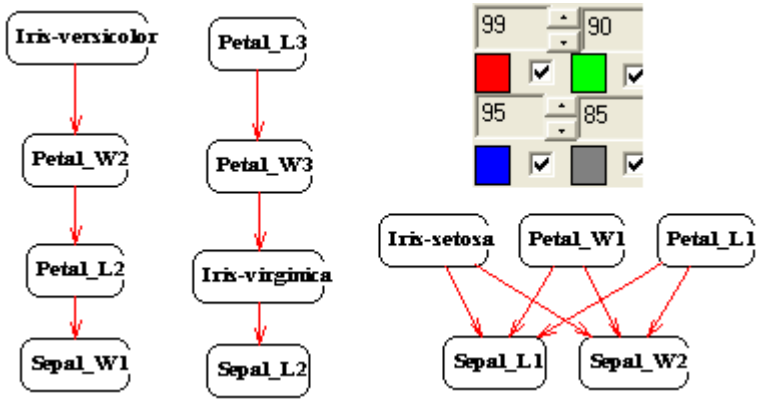
Ci-dessus, à gauche, les « informations sur le fichier » montrent que le choix de découpage fait par CHIC (selon la variance) produit pour certaines variables une répartition inégalitaire des effectifs (51 pour petal_L1 contre 99 pour pétal_L2). Ci-dessus, à droite, son graphe implicatif, qui ne permet pas de différencier Iris-versicolor de iris-verginica.

Nous avons ensuite testé trois autres découpages des variables, dont seul le deuxième a été jugé satisfaisant, car ayant fourni un graphe séparé par type d'iris :

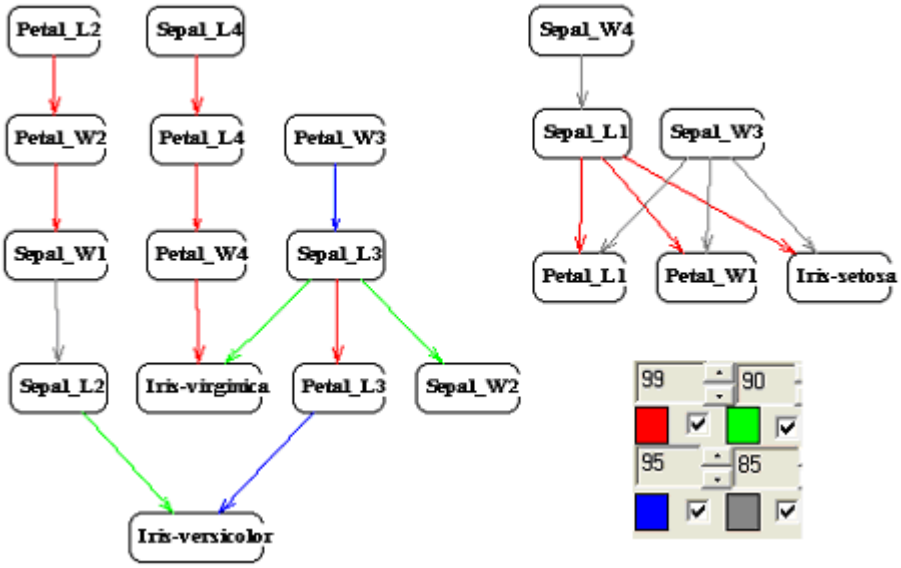
Découpage automatique de petal_L et petal_W en 2, de sepal_L et Sepal_W en 3



Découpage automatique de petal_L et petal_W en 3, de sepal_L et Sepal_W en 2



Découpage automatique de petal_L, petal_W, sepal_L et Sepal_W en 4

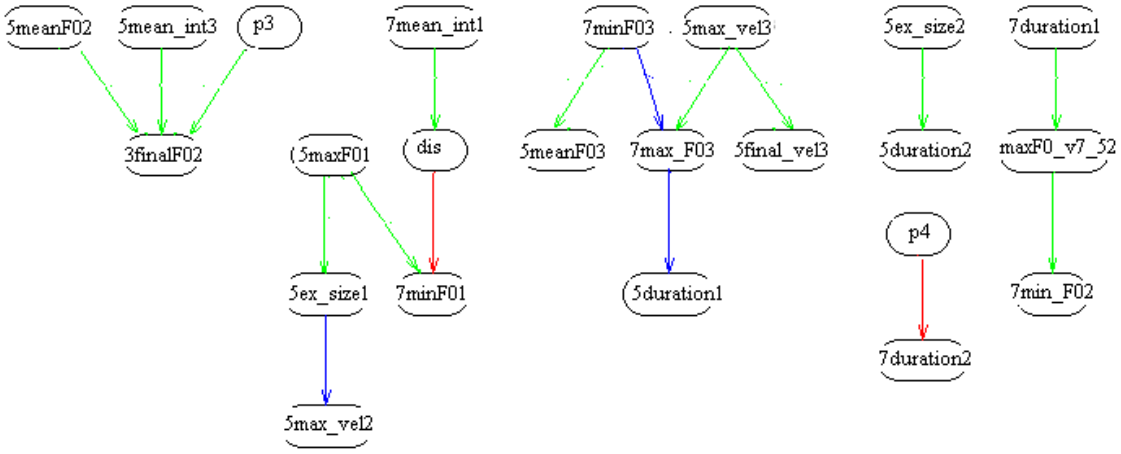


Annexe de la partie 3 : autres résultats sur l'intonation avec CHIC

Réduction des variables découpées en 3, avec recollements, au seuil de 0.9

3meanf01 3mean_int1 3finalf01 5finalf01 3maxf01 représentant de la classe : 3maxf01
maxf0_v7_51 7finalf01 7meanf01 dis 7final_vel1 7maxf01 représentant de la classe : dis
cis 3mean_int3 7mean_int3 5mean_int3 représentant de la classe : 5mean_int3
5duration2 7ex_size1 représentant de la classe : 5duration2
3minf03 5minf03 3maxf03 5maxf03 3meanf03 5meanf03 5finalf03 3finalf03 représentant de la classe : 5meanf03
qis maxf0_v7_53 7maxf03 7finalf03 7meanf03 7final_vel3 7ex_size3 représentant de la classe : 7maxf03
7meanf02 7maxf02 7final_vel2 maxf0_v7_52 7finalf02 représentant de la classe : maxf0_v7_52
3duration1 p4 représentant de la classe : p4
5maxf01 maxf0_v5_31 5meanf01 5minf01 représentant de la classe : 5maxf01
3maxf02 5finalf02 3meanf02 5meanf02 5minf02 5maxf02 représentant de la classe : 5meanf02
5final_vel3 5ex_size3 maxf0_v5_33 représentant de la classe : 5final_vel3
3finalf02 3minf02 représentant de la classe : 3finalf02
7duration3 p2 représentant de la classe : 7duration3
3final_vel3 3ex_size3 représentant de la classe : 3ex_size3
7ex_size2 7minf03 représentant de la classe : 7minf03

et le graphe implicatif avec les variables réduites et les seuils de 99.9, 99 et 95



APPORT DE LA COMBINAISON DE LA METHODE D'ANALYSE STATISTIQUE IMPLICATIVE (A.S.I.) AVEC LA THEORIE DE REPONSES AUX ITEMS (IRT).

Hayette KHALED[1], Raphael COUTURIER[2]

CONTRIBUTION OF STATISTICAL IMPLICATIVE ANALYSES (SIA) COMBINED WITH THE ITEM RESPONSE THEORY (IRT).

RÉSUMÉ

Dans ce papier, nous illustrons les avantages de la combinaison de la méthode d'Analyse Statistique Implicative (ASI) avec la Théorie de Réponses aux Items (IRT) à travers un exemple d'analyse des résultats des étudiants en Informatique de l'université A/Mira de Bejaia. L'ASI est utilisée pour découvrir et analyser les implications les plus pertinentes entre les différents modules de formation étudiés. L'IRT, quant à elle, permet l'analyse de la qualité des items et leur calibrage (difficulté) en permettant la construction d'échelles quasiment indépendantes des données d'évaluation (étudiants et modules). Une analyse des résultats est proposée ainsi qu'une comparaison avec les résultats que nous avons obtenus dans le papier Khaled *et al*. (2014) dans lequel nous avons appliqué l'ASI aux notes des étudiants.

Mots-clés : *ASI, IRT, Notes des Etudiants.*

ABSTRACT

In this paper we illustrate the benefits of combining Statistical Implicative Analysis (SIA) method with the Items Response Theory (IRT) throw a study of a sample of computer science students of Bejaia University. SIA is used to discover and analyze the most relevant implications between different studied modules. The IRT, in turn, allows the analysis of the items quality and their calibration (difficulty) by allowing the construction of scales independent from the test data (students and modules). An analysis of the results is proposed with a comparison with the results that we have obtained in the paper khaled*and al.*(2014) in which we applied the SIA on students' marks.

Keywords : *SIA, IRT, Students' Marks.*

1 Introduction

La théorie de réponses aux items (IRT) est un modèle statistique de la mesure relativement récent Pini (2012), elle permet de faire face à des problèmes auxquels la psychométrie classique n'apporte pas toujours des réponses et des solutions satisfaisantes, elle s'efforce de produire une estimation des propriétés de l'item quasiment indépendante d'un groupe particulier d'individus. En d'autres termes, elle cherche à élaborer des instruments de mesure dont les caractéristiques ne soient pas excessivement influencées par tel ou tel autre groupe de référence.

[1]U A/Mira Béjaia, Targa Ouzemour, hayette.khaled@univ-bejaia.dz ou aya_info6@yahoo.fr
[2]IUT Belfort-Montbéliard, 19 Avenue du Maréchal Juin BP 527, raphael.couturier@univ-fcompte.fr

L'IRT repose sur des modèles de type probabiliste. Ces modèles sont fondés sur le principe que la réponse d'un individu à l'item (et notamment sa probabilité de fournir une réponse correcte) est déterminée par deux sortes de facteurs d'une part, la compétence du sujet (qualifié de traits latents) ; d'autre part, les propriétés de l'item lui-même : son degré de difficulté, son pouvoir de discrimination ainsi que la "chance" qui peut jouer un rôle dans certains cas.

L'IRT a été proposée pour la première fois dans le domaine de la psychométrie, dans le but de l'évaluation de la capacité Xinming et Yiu-Fai. (2014). Elle est par la suite utilisée dans différents domaines comme la médecine Michael (2013), le tabagisme chez les adolescents Hedeker*et al*. (2006), etc. En outre elle est largement utilisée dans l'éducation pour calibrer et évaluer les items dans les tests ainsi que la capacité des sujets. Au cours des dernières décennies, l'évaluation pédagogique a utilisé de plus en plus les techniques à base d'IRT afin de développer des tests.

Par exemple, dans Johns *et al*. (2006) les auteurs ont conçu plusieurs expériences pour tester les modèles d'IRT qui ont été utilisés comme un outil pour la modélisation des élèves dans un système de tutorat intelligents. Les modèles ont été testés en utilisant des données réelles des élèves du secondaire en utilisant le système de tutorat Outpost Wayang. Un modèle de validation croisée a été développé et trois métriques de mesure de la précision de la prédiction ont été comparées. Les résultats ont montré que les modèles formés prédisent avec une précision de 72% si un étudiant répond correctement à un problème de choix multiples ainsi, nous pouvons dire que ces modèles ont un bon pouvoir de prédiction.

Dans Hamdare (2014) les auteurs proposent un système d'évaluation adaptif se basant sur l'IRT. Ce système permet de personnaliser dynamiquement les questions pour les élèves en fonction de leur réponse à la question précédente, mais en plus de cela, et ce qui est très intéressant, il permet aussi d'ajuster le degré de difficulté des questions d'évaluation en fonction de la capacité de l'élève ; ainsi cela aidera les enseignants à acquérir une mesure valide et fiable des compétences des étudiants.

Dans Rowan (2001) les chercheurs de l'Université de Michigan travaillent sur l'élaboration de mesures fondées sur des enquêtes concernant le contenu pédagogique des connaissances des enseignants, en identifiant trois dimensions à mesurer, à savoir : le contenu de la connaissance, la connaissance de la réflexion des élèves ainsi que la connaissance des stratégies pédagogiques ; afin d'analyser leurs effets sur la réussite des étudiants. Les chercheurs établissent des questionnaires pour les enseignants, concernant les trois dimensions citées auparavant puis appliquent le programme commercial informatique BILOG pour l'analyse et la notation des réponses aux questionnaires qui se basent sur l'IRT pour exclure les questions qui diminuent la fiabilité de l'ajustement.

L'équipe de recherche de l'IREM d'Aix-Marseille développe une méthode adaptive d'évaluation sur ordinateur Bodin *et al*. (2010) qu'elle teste sur le socle commun en termes de connaissances et de compétences. Elle cherche ainsi à répondre aux questions suivantes: quelles sont les connaissances et les compétences nécessaires à tous les individus, indépendamment du cursus de formation que chacun pourra avoir suivi ou non ? Comment est-il possible de positionner des individus par rapport aux demandes du socle ?

La méthode adaptive consiste à former un référentiel R qui contient l'ensemble des connaissances et des compétences à évaluer. Un ensemble de questions Q est associé au référentiel R ; et l'arbre implicatif cohésif de CHIC Couturier et Gras (2005) ou Couturier (2008) est utilisé afin de déterminer les questions qui correspondent à la même unité de compétence.

Le principe du test est de prendre au hasard la première question (q1,p) à poser à la personne p puis la question suivante sera choisie selon le graphe implicatif des questions de ASI. Le test est arrêté lorsque l'amplitude de l'intervalle de confiance du score estimé devient inférieure à un intervalle de confiance choisi.

Le test adaptatif sur R ici consiste à estimer le score d'une personne sur l'ensemble de questions Q à partir de ses réponses à un sous ensemble Q' en utilisant la fonction d'estimation probabiliste de l'IRT. Les auteurs montrent que la combinaison de l'ASI et de l'IRT permet d'envisager des tests très fiables.

Tous les travaux qui utilisent l'IRT montrent que cette dernière a des apports très intéressants, sur tous ceux qui s'inscrivent dans le domaine de l'éducation qui est notre domaine d'intérêt. Ces travaux s'intéressent en général au raffinement de questionnaires d'un système du tutorat ou d'un système web d'essai ou l'étude des connaissances des enseignants et leurs effets sur la réussite des étudiants.

Il y a peu de travaux qui traitent les notes de cursus des étudiants. Dans notre travail Khaled *et al*. (2014), nous utilisons l'analyse statistique implicative (ASI) Gras et al (2008) et Gras et al (2013) afin d'identifier les implications pertinentes entre les modules.

Notre objectif dans ce papier est de montrer les avantages et apports des deux méthodes ASI et IRT à travers un exemple qui porte sur les notes des étudiants de département informatique de l'université A/Mira de Bejaia. Nous avons cité auparavant un travail qui fait aussi la combinaison de ces deux méthodes sur des questions de tests dont les réponses représentent des données binaires. Dans notre cas nous traitons des données polytomiques, c'est-à-dire avec plusieurs modalités, obtenus en divisant les notes des étudiants en trois groupes par l'algorithme des nuées dynamique pour représenter les étudiants avec des résultats faibles, des résultats moyens et de bons résultats. Ainsi même si les notes des étudiants sont globalement élevées ou au contraire très faibles, nous aurons les 3 groupes précédemment énoncés.

Dans la suite du papier nous introduisons la théorie de réponses aux items (IRT) puis nous montrons notre expérience et ses résultats. Finalement nous présentons quelques résultats obtenus avec l'ASI présentés dans Khaled *et al*. (2014) ainsi qu'une comparaison entre l'IRT et l'ASI.

2 Théorie de Réponses aux Items (IRT) Pini (2012),Hamdare (2014) Dai-Trang (2013),Rizopoulo (2006)

Les modèles d'IRT sont des fonctions mathématiques appelées fonction caractéristique d'item qui spécifient la probabilité de répondre correctement à un item, en termes de paramètres d'items et d'individus. Ils sont représentés graphiquement par une courbe continue appelée la Courbe Caractéristique d'Item (CCI) (Figure 1), elle

décrit le lien qui existe entre la situation d'un individu par rapport au trait latent, représentée par θ et la probabilité que cet individu réussi l'item i.

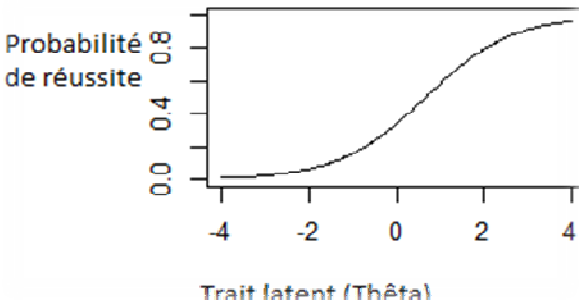


Figure 1 – La Courbe Caractéristique d'Item (CCI)

Il existe trois modèles d'IRT : le modèle à un seul paramètre (ou modèle de Rasch), le modèle à deux paramètres et le modèle à trois paramètres. Les trois modèles ont un paramètre de difficulté d'item représenté par b_i.

– l'équation du modèle à un seul paramètre est donnée comme suit :

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} \quad (1)$$

– l'équation du modèle à deux paramètres est donnée comme suit :

$$P_i(\theta) = \frac{e^{\alpha_i(\theta - b_i)}}{1 + e^{\alpha_i(\theta - b_i)}} \quad (2)$$

– l'équation du modèle à trois paramètres est donnée comme suit :

$$P_i(\theta) = C_i + \frac{(1 - C_i)e^{\alpha_i(\theta - b_i)}}{1 + e^{\alpha_i(\theta - b_i)}} \quad (3)$$

En plus, le modèle à deux paramètres et celui à trois paramètres possèdent un paramètre de discrimination noté α lié à la rapidité de changement de la probabilité avec les changements en capacité. Le modèle à trois paramètres contient un troisième paramètre, appelé paramètre de pseudo-chance noté *C* qui représente la probabilité qu'un sujet de très faible capacité répondra correctement à un item *i* en choisissant aléatoirement l'une des options proposées. Il y a deux branches principales de l'IRT : l'unidimensionnelle (UIRT) et la multidimensionnelle (MIRT). La première branche opère sur l'hypothèse qu'il existe une idée ou un message à travers le contenu de tous les items et les instruments de test sont mis au point pour mesurer cette idée unique. La deuxième branche, quant à elle, suppose que les items sont regroupés dans des

domaines différents, et les instruments sont utilisés pour mesurer ces multiples constructions de domaine Dai-Trang (2013). Les modèles les plus courants d'IRT sont des modèles dichotomiques avec deux catégories de réponses pour les items et les modèles polytomiques avec des catégories multiples et ordinales de réponses pour les items. Dans notre cas nous allons traiter les notes des étudiants qui sont entre zéro et vingt qui seront divisées en trois groupes donc elles représentent des données polytomiques. Elles incluent les données d'items à choix multiples, des questions mathématiques ouvertes, les échelles de Likert, les items ordinaux, les réponses d'évaluation d'échelles et les réponses graduées de test ou de questions d'enquête. Le type de modèle IRT utilisé pour décrire l'interaction entre les sujets et les items de test dépend de la nature des données collectées. Plusieurs modèles ont été utilisés pour les données polytomiques, tels que le modèle de crédit partiel (PCM), le modèle de crédit partiel généralisé (GPCM), le modèle de la réponse nominale (NRM) et le modèle de la réponse graduée (GRM). Ce dernier modèle est proposé par Samejima (1969), c'est le mieux approprié pour le traitement des notes des étudiants, sa forme logistique est exprimée comme suit:

$$P(Y_{ijk} = k|\theta_j, b_{ik}, a_i) = \frac{1}{1 + e^{-a_i(\theta_i - b_{i,k})}} - \frac{1}{1 + e^{-a_i(\theta_i - b_{i,k+1})}} \quad (4)$$

3 Analyse et résultats

3.1 Description des données

La population considérée représente les étudiants de licence 2 de département Informatique de l'université A/Mira de Bejaïa pendant les années 2010, 2011 et 2012. Les étudiants de licence 2 sont évalués sur les modules programmation linéaire (PL), théorie de langages (THL), logique mathématique (LogMat), analyse numérique (ANum), système d'exploitation (SE), probabilités et statistiques (P.S), traitement de signal (TSig), génie logiciel (GL), algorithme et structure de données (ALSTRD2), architecture (ARCH), bases de données (BDD) et structure de données (STRD1).

Les notes des étudiants sont représentées par un fichier de type « .csv » (un exemple est montré dans le tableau 1). Ce fichier contient en lignes les matricules des étudiants et en colonnes les modules suivis par ces derniers ; et la variable « p » qui suit chaque intitulé de module pour dire que les notes de modules sont à partitionner en un nombre fixe d'intervalles, nous utilisons l'algorithme des nuées dynamiques Diday (1971) qui constitue automatiquement les intervalles qui ont des limites distinctes.

	ARCH p	STRD1 p
0809TMI02	11,25	11,5
09MI0034	9,38	11,38
09MI0590	10,63	9,38
…	…	…

Tableau 1 – Exemple de données de type .csv

Pour traiter ces données avec l'IRT nous avons choisi le modèle de la réponse graduée (GRM) à deux paramètres qui permet de donner la difficulté des modules et une

bonne discrimination de ces derniers. Le choix de ce modèle est dû au fait que les notes des étudiants sont divisées en trois catégories (résultats faibles, moyens et bons) et elles représentent ainsi des données polytomiques qui peuvent être traitées par le modèle GRM.

L'analyse des résultats a été effectuée en traçant les courbes caractéristiques des modules comme montré sur la figure suivante :

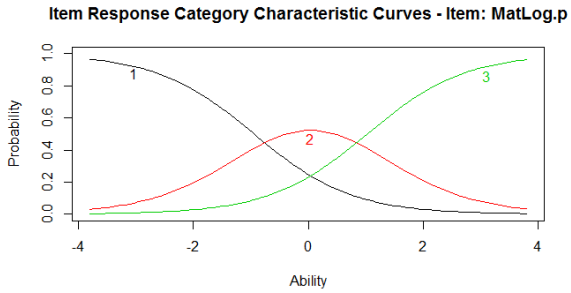


Figure 2 – Courbe Caractéristique d'Item obtenue par le modèle GRM.

La courbe 1 (en noir) représente la probabilité d'échouer le module en fonction de la capacité de l'étudiant, la courbe 2 (en rouge) représente la probabilité d'avoir un résultat moyen dans le module en fonction de la capacité de l'étudiant et la courbe 3 (en vert) représente la probabilité de réussir le module toujours en fonction de la capacité de l'étudiant. Cette figure montre que les étudiants avec de faibles résultats (capacité entre - 4 et - 2) ont une grande probabilité d'échouer le module et une probabilité presque nulle de réussir le module, contrairement aux étudiants avec de bons résultats (capacité entre 2 et 4) qui ont une grande probabilité de réussir le module et une probabilité presque nulle d'échouer au module.

3.2 Résultats du modèle IRT et analyse

Dans ce qui suit nous présentons quelques courbes caractéristiques d'items (CCI) (un item représente un module) en licence 2 durant les années scolaires 2010-2011 (Figure 3), 2011-2012 (Figure 4) et 2012-2013 (Figure 5).

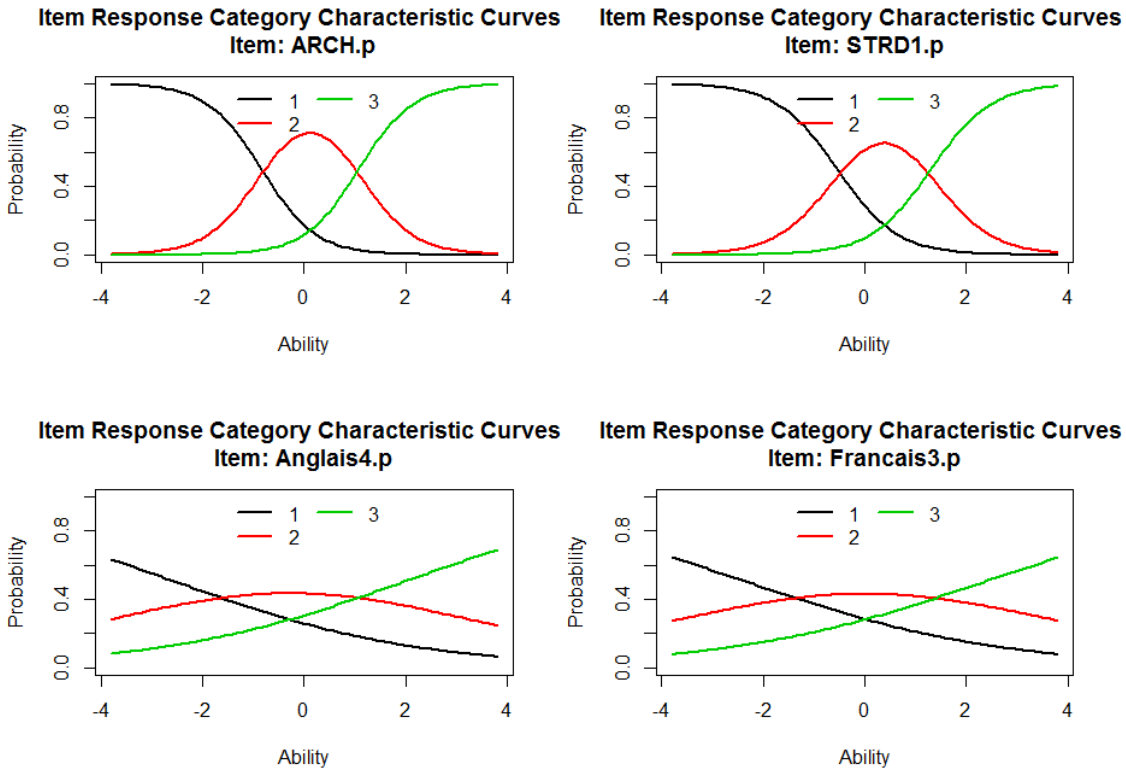


Figure 3 (suite) – CCI de licence 2 2010-2011

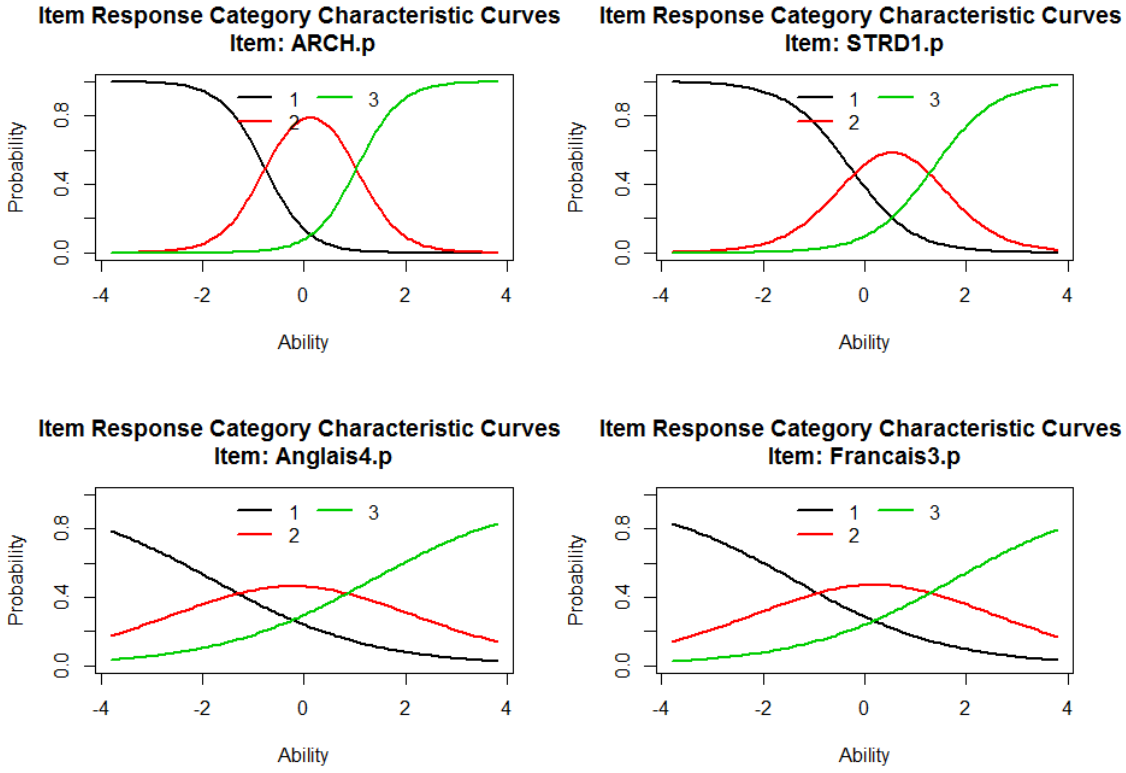


Figure 4 – CCI de licence 2 2011-2012

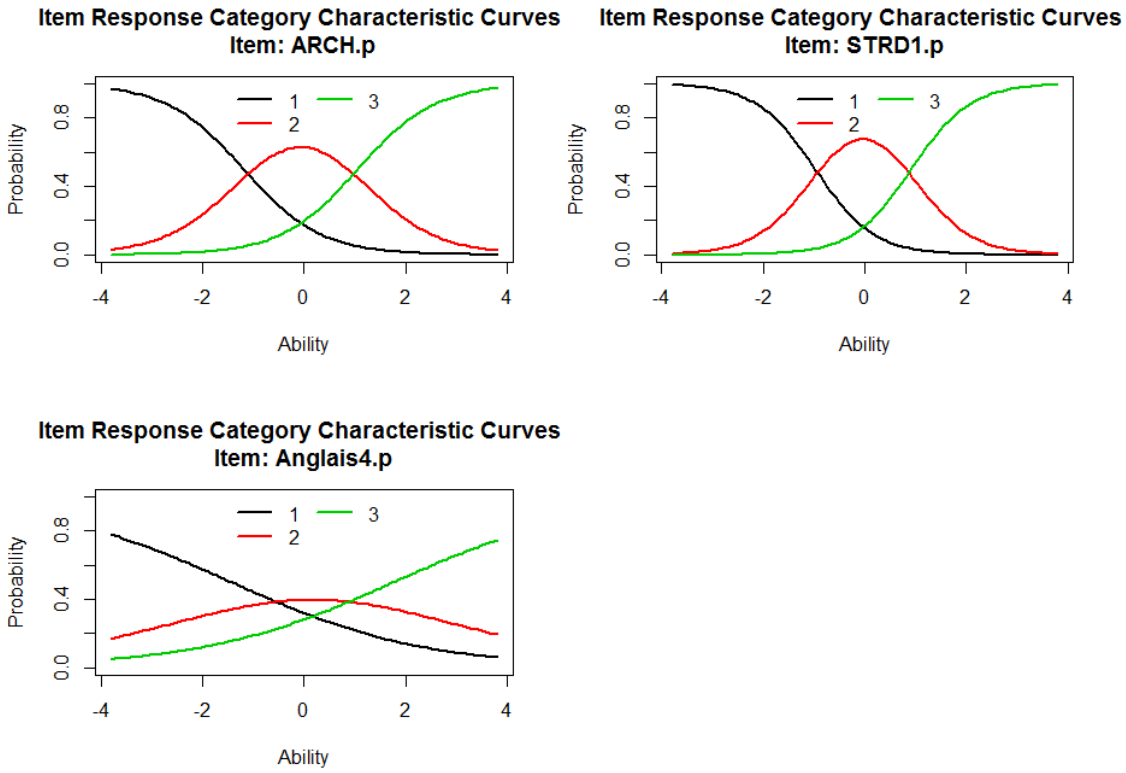


Figure 5 – CCI de licence 2 2012-2013

D'après les figures 3, 4 et 5 présentées précédemment et d'après les figures 6, 7 et 8 qui sont mises en annexe, nous observons que nous avons pratiquement les mêmes résultats pour les différentes années. Nous voyons bien que les modules de spécialité ARCH et STRD1(figures 3, 4 et 5), ANum, P.S, LogMat, TSig, ALSTRD2, BDD, SE, THL, PL et GL(figures 6, 7 et 8) sont les plus discriminants[3] (nous observons des courbes plus raides), autrement dit ces modules font bien la différence entre les étudiants avec de faibles résultats et ceux avec de bons résultats contrairement aux modules Anglais4 et Français3 (figures 3, 4 et 5), ainsi que SI, Anglais3 et Français4[4] (figures 6, 7 et 8) qui ne le sont pas.

Nous savons qu'en général les étudiants avec de bons résultats en informatique ont de bons résultats dans tous les modules de spécialité et les étudiants avec de faibles résultats en informatique ont des résultats similaires dans tous les modules de spécialité. Ceci est confirmé par les résultats présentés ci-dessus.

Les étudiants avec de bons résultats, c'est-à-dire avec une capacité supérieure à 2 par exemple ; ont une probabilité très proche de 1 de réussir les modules de spécialité et une probabilité très proche de 0 d'échouer aux modules de spécialité. Par contre les étudiants avec de faibles résultats, ceux qui ont une capacité inférieure à - 2 par

[3]Nous donnons entre parenthèses le paramètre de discrimination de chaque module :ARCH(1.867), STR1D(1.693), ANum(1.353), P.S(1.273), LogMat(1.202), TSig(1.287), ALSTRD2 (1.614), BDD (1.639), SE (1.466), THL (1.305), PL (1.272), GL (1.428)

[4]SI (0.483), Anglais3 (0.409), Anglais4 (0.421), Francais3 (0.400), Francais4 (0.062)

exemple, ont une probabilité très proche de 1 d'échouer aux modules de spécialité et une probabilité presque nulle de réussir les modules de spécialité.

Le fait d'être bon en informatique n'implique pas être bon en langues (français et anglais) et, de la même manière, le fait d'être faible en informatique n'entraîne pas d'être faible en langue. Nous observons exactement cela sur les résultats obtenus dans les figures 3, 4 et 5 et nous constatons que ces modules ne sont pas discriminants. Par rapport au module système d'information (SI)(figures 6, 7 et 8 en annexe), nous remarquons aussi qu'il n'est pas discriminant et cela s'explique par le fait que ce n'est pas un module de spécialité informatique, il porte sur l'organisation des entreprises, leur gestion, etc. Pour les étudiants qui ont des résultats moyens (dont la capacité est entre -2 et 2), les courbes numéros 2 (rouge) nous montrent qu'ils ont une probabilité moyenne (entre 0.5 et 0.7) de réussir les modules de spécialité.

3.3 Résultats de la méthode ASI et analyse

Dans ce qui suit nous présentons quelques graphes implicatifs résultants de l'application de l'analyse statistique implicative (ASI) sur les même notes d'étudiants présentée ci-dessus, ce travail a fait l'objet d'un papier Khaled *et al* (2014) publié à ISKO-Maghreb (2014).

Les intervalles des notes des étudiants ici sont également partagés en trois sous intervalles identifiant les étudiants avec des résultats faibles, moyens et bons. Par exemple, le module STRD1 a été partitionné selon les intervalles suivants :

STRD11 de 5.94 à 10.38, STRD12 de 10.5 à 13.38, STRD13 de 13.5 à 18.38

STRD11 reflète les étudiants qui sont faibles en STRD1, STRD12 les étudiants qui sont moyens et STRD13 les étudiant qui sont bons en STRD1.

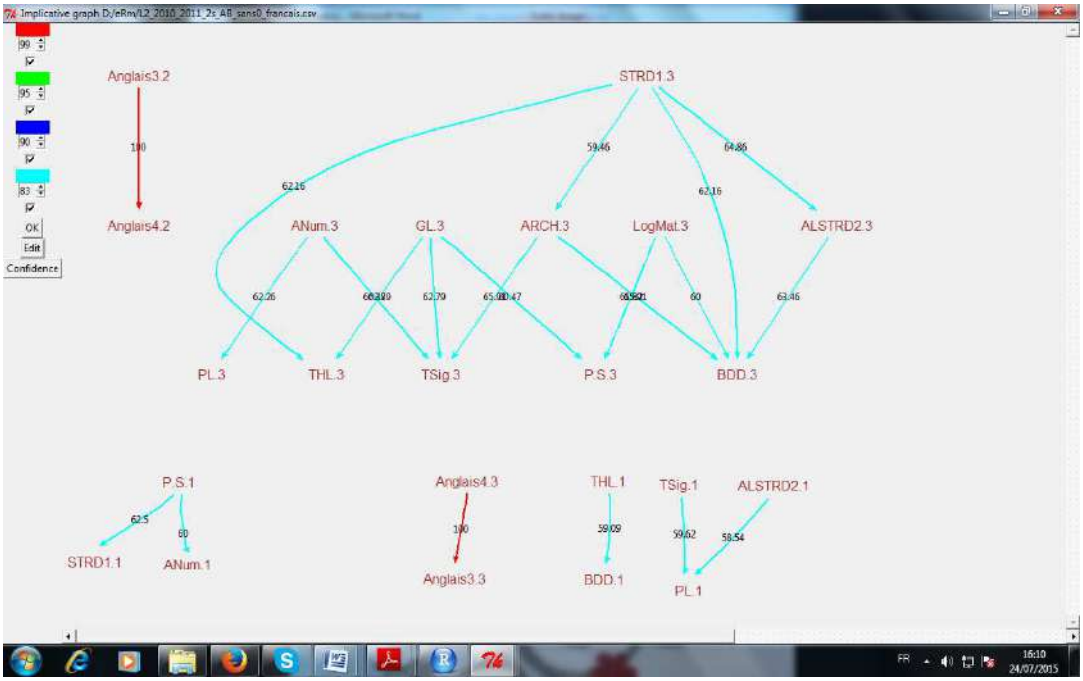


Figure 9– Graphe implicatif L2 2010-2011

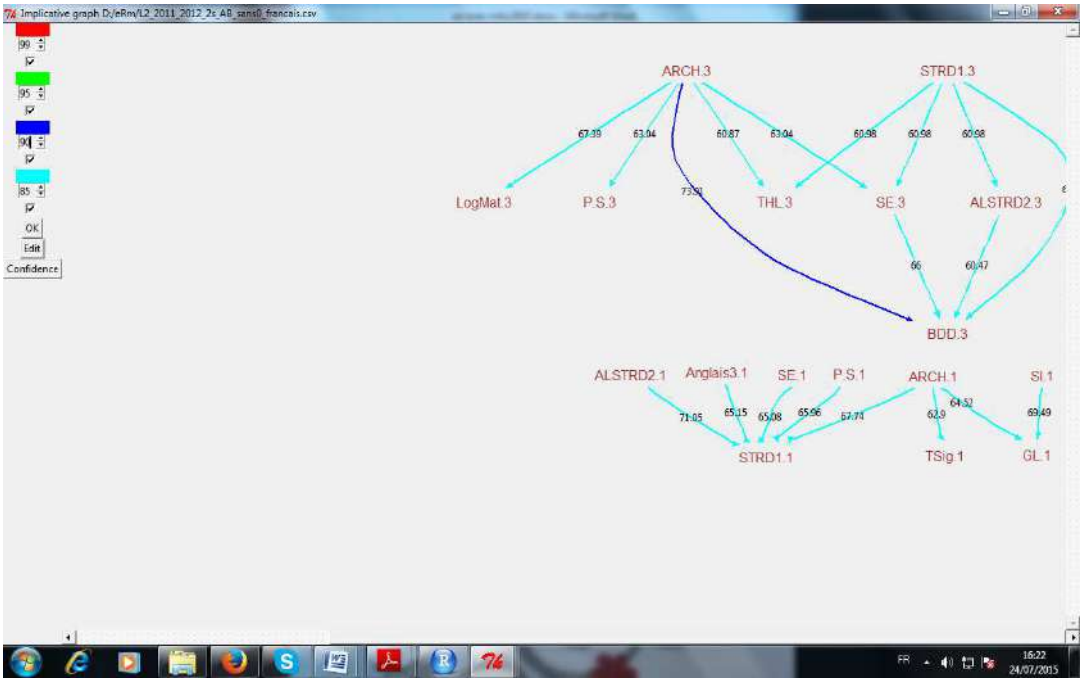


Figure 10 – Graph implicatif L2 2011-2012

Le critère utilisé pour la recherche des implications pour les graphes implicatifs est l'implifiance présenté dans GRAS *et al* (2015), ce dernier présente une combinaison ou un mariage entre les deux critères confiance et intensité d'implication.

Nous voyons sur ces figures que les arcs sont pondérés par un poids ce qui représente la « confidence en anglais », si nous prenons l'exemple de la confiance de 71,05 entre les modules ALSTRD2.1 et STRD1.1 dans la figure 10 ceci signifie que si un étudiant est faible en ALSTRD2 il a une chance de 71,05% d'être aussi faible en STRD1. Toutes les implications que nous avons obtenues sont intéressantes car elles ont presque toutes des confiances supérieures à 60.

Nous observons sur les deux figures 9 et 10 des implications qui se répètent sur deux années consécutives. Nous remarquons clairement qu'un module réussi par les bons élèves entraine d'autres modules réussis par les bons élèves (par exemple STRD1.3 -> ALSTRD2.3, STRD1.3 -> THL3, ALSTRD2.3 -> BDD3). Il en est de même pour les modules échoués par les élèves faibles qui impliquent d'autres modules échoués par des élèves faibles (par exemple PS1 -> STRD1.1). Nous ne voyons pas sur les figures les étudiants moyens, nous expliquons ceci par le fait qu'un étudiant qui est moyen dans un module peut être faible ou bon dans d'autres modules et il n'est pas forcément moyen dans tous les modules.

3.4 Comparaison des résultats de l'ASI et l'IRT

Dans le papier Khaled *et al* (2014) nous avons présenté les résultats de l'application de l'analyse statistique implicative (ASI) sur les notes des étudiants nous avons reporté quelques figures (9 et 10) dans le présent papier. Les résultats obtenus montrent que les implications entre les modules se répètent généralement sur les trois promotions (2010_2011, 2011_2012 et 2012_2013). Ces implications sont intéressantes car elles

correspondent aux réussites ou aux échecs des étudiants (les modules réussis impliquent d'autres modules réussis et les modules échoués par les étudiants impliquent d'autres modules échoués). Nous avons aussi noté des implications entre les modules qui sont liés par leur contenu ou par l'enseignant qui assure le cours.

Nous remarquons que l'ASI nous donne des implications très intéressantes entre les modules, elle nous permet de dire si un étudiant maitrise un module X alors il aura tendance à maitriser un autre module Y.

La Théorie de Réponse aux Items (IRT) quant à elle, nous montre quels sont les modules qui sont discriminants (en fonction de la capacité des étudiants) et de plus elle nous indique la difficulté des modules.

D'après les résultats obtenus dans Khaled *et al* (2014) et ceux de ce papier, nous constatons que l'Analyse Statistique Implicative et la Théorie de Réponses aux Items nous donnent des informations différentes. La première nous donne les implications entre les modules. La seconde nous informe sur la probabilité d'un étudiant de réussir ou non un module selon ses capacités et la difficulté de module. Ainsi l'ASI et l'IRT sont bien complémentaires et offrent des indications précieuses pour analyser les résultats des étudiants du département concerné. Nous encourageons les enseignants à analyser les résultats de leurs élèves ou étudiants avec l'ASI et l'IRT afin d'avoir des informations leur permettant de mieux comprendre les résultats des élèves ou étudiants.

4 Conclusion

Nous avons appliqué la méthode d'Analyse Statistique Implicative et la Théorie de Réponses aux Items sur les notes des étudiants de l'université A/Mira de Bejaia à travers les trois promotions 2010-2011, 2011-2012 et 2012-2013.

Les notes ont été partitionnées en trois intervalles en utilisant l'algorithme des nuées dynamiques pour avoir les étudiants avec de bons résultats, les étudiants avec des résultats moyens et les étudiants avec des résultats faibles. Ensuite nous avons utilisé les deux méthodes ASI et IRT pour pouvoir comparer leurs résultats.

Avec l'ASI nous avons utilisé le logiciel CHIC pour tracer les graphes implicatifs qui nous donne les implications pertinentes entre les différents modules. Pour l'IRT nous avons utilisé le modèle GRM à deux paramètres pour estimer et tracer les courbes caractéristiques d'items.

Les résultats des deux méthodes montrent que ces dernières sont complémentaires. Ainsi l'ASI nous donne les implications les plus pertinentes entre les modules ; alors que l'IRT nous donne la difficulté des modules et nous montre que les étudiants avec de bons résultats ont de bons résultats dans tous les modules de la spécialité. De même, les étudiants avec de faibles résultats sont également de faibles résultats dans tous les modules de la spécialité.

Nous conseillons aux enseignants d'utiliser ces méthodes pour analyser les notes de leurs étudiants afin de déterminer la difficulté des modules, afin d'améliorer les cours ou les sujets d'examens par exemple. En outre, ces deux méthodes peuvent être utilisées pour raffiner les modules d'une spécialité, elles permettent de déterminer les modules qui ont une relation avec la spécialité. Ainsi, il est possible d'effectuer des choix pourles modules d'une formation.

Dans un futur proche, nous comptons tester les deux méthodes sur des données plus volumineuses comme celles de Programme International pour le Suivi des Acquis des élèves (PISA). Ces données sont rassemblées dans le cadre de l'enquête internationale PISA qui est menée tous les trois ans depuis 2000 auprès des jeunes de 15 ans, dans la plupart des pays membres de l'OCDE et dans de nombreux pays partenaires. Cette enquête évalue l'acquisition de savoirs et savoir-faire essentiels à la vie quotidienne au terme de la scolarité obligatoire. Ce travail sera également l'occasion d'évaluer les performances d'autres méthodes d'apprentissage automatique à travers l'environnement statistique R.

Références

[1] Bodin, A et al (2010), Vers un test adaptatif - criterié, combinant l'utilisation de l'IRT et de l'ASI, pour l'évaluation du socle commun de connaissances et de compétences, *Quaderni di Ricerca in Didattica (Mathematics)*,n°20, 383-409.

[2] Couturier, R. (2008), Statistical Implicative Analysis. In CHIC: CohesiveHierarchical Implicative Classification, volume 127 of Studies in Computational Intelligence, Springer ; 41-52.

[3] Couturier, R. Gras, R. (2005), CHIC: Traitement de données avec l'analyse implicative, *Revue des nouvelles technologies de l'information*, n°3, 679-684.

[4] Dai-Trang, L. (2013), *Applying item responsetheorymodeling in educationalresearch*, These: Philosophie, Ames Iowa aux États-Unis : Université d'État de sciences et technologie de l'Iowa, 187 p.

[5] Diday, E. (1971), La méthode des nuées dynamiques, *Journal revue de statistique appliqué*, **n°** 19 (2), 19-34.

[6] GrasR., SuzukiS., GuilletF., Spagnolo F. (2008) F. Statistical Implicative Analysis, Theory and Applications, eds, Springer.

[7] Gras R., Régnier J.-C., Marinica C., Guillet F. (2013) L'analyse statistique implicative, Méthode exploratoire et confirmatoire à la recherche de causalités, sous la direction de Gras R., edsCépaduès Editions, 522 pages, ISBN 978.2.36493.056.8.

[8] Gras R., Couturier R., Gregori P. (2015), Un mariage arrangé entre l'implication et la confiance ?, *A.S.I.8*.

[9] Hamdare, S. (2014), An Adaptive Evaluation System to Test Student Caliber using Item Response Theory, *IJMTER*, 329-333.

[10] Hedeker, D et al. (2006), Application of Item Response Theory Models for Intensive Longitudinal Data,*in T.A. Walls & J.L. Schafer (Eds.), Models for Intensive Longitudinal Data*,Oxford University Press, New York, 84-108.

[11] Johns, J. Mahadevan, S. Woolf, B. (2006), Estimating Student Proficiency Using an Item Response Theory Model, In : Mitsuru Ikeda, Kevin D. Ashley, Tak-Wai Chan. Intelligent Tutoring Systems. Taiwan: Springer, 473-480.

[12] Khaled, H. Ghanem, S. Couturier, R. (2014), Analysis of Bejaia University Computer Science students' marks through the CHIC software and Statistical Implicative Analysis, *ISKO-Maghreb*, (n° 94).

[13] Michael, L. (2013), An application of item response theory to fMRI data: Prospects and pitfalls, *Psychiatry Research: Neuroimaging*, **n° du volume**212(n° 3), 167–174.

[14] À propos de la théorie des réponses aux items. Le cas d'items dichotomiques, http://www.irdp.ch/edumetrie/tri.htm, consulté le 28-04-2015.

[15] Rowan, B et al. (2001), Measuring Teachers' Pedagogical Content Knowledge in Surveys: An Exploratory Study, Philadelphia, *PA: Consortium for Policy Research in Education, Ministry of Education, Youth and Sports*.

[16] Rizopoulo, D. (2006), ltm: An R Package for Latent Variable Modeling and Item Response Theory Analyses, *Journal of Statistical Software*, **n° du volume** 17 (n°5), 1-25.

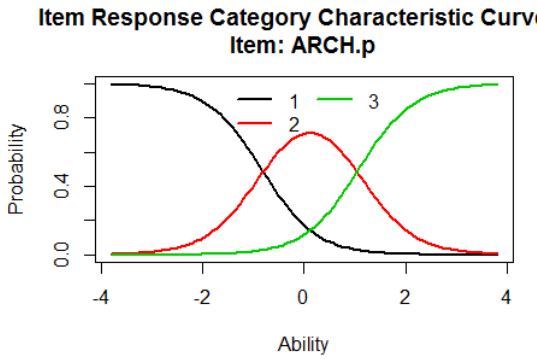
[17] Xinming, A et Yiu-Fai, Y. (2014), Item Response Theory: What It Is and How You Can Use the IRT Procedure to Apply It, *SAS Institute Inc*.

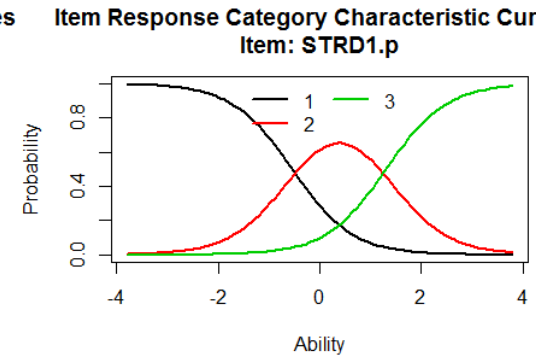
Annexe

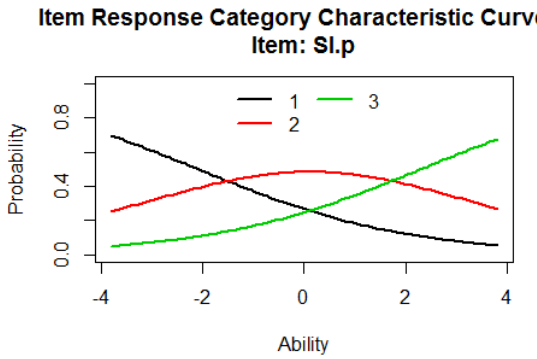
Dans cette annexe nous présentons les courbes caractéristiques d'items de tous les modules étudiés en licence 2 des trois promotions 2010_2011, 2011_2012 et 2012_2013 :

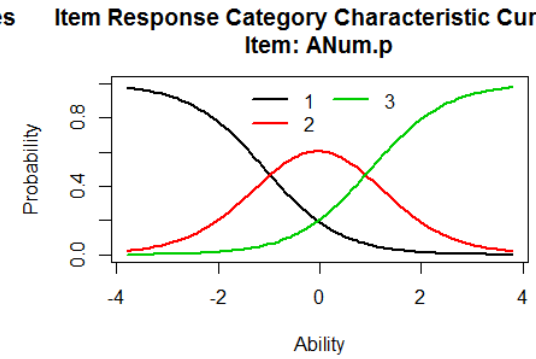
Dans les figures 6, 7 et 8 on voit les courbes caractéristiques d'items (CCI) de tous les modules assurés en licence 2 pour les trois promotions 2010_2011, 2011_2012 et 2012_2013. Chaque CCI contient trois courbes, la noir (numéro 1) représente les étudiants qui sont faibles dans le module, la rouge (numéro 2) représente ceux qui sont moyens et la verte (numéro 3) représente ceux qui sont faibles dans le module.

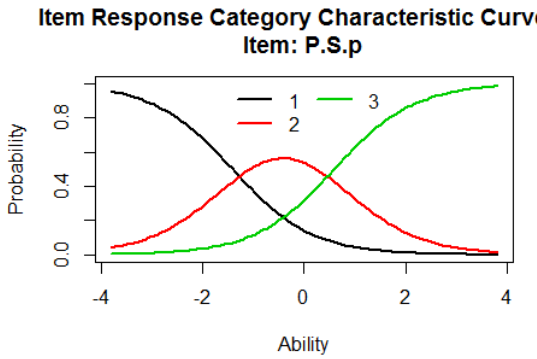
Nous remarquons sur les trois figures que les courbes caractéristiques d'items sont presque les mêmes entre les trois promotions. Nous constatons aussi que les modules sont divisés en deux groupes; ceux qui sont discriminants (ARCH, STRD1, ANum, P.S, LogMat, TSig, ALSTRD2, BDD, SE, THL, PL et GL); ils représentent les modules de spécialité et ceux qui ne le sont pas (Anglais4, Français3, SI, Anglais3 et Français4).

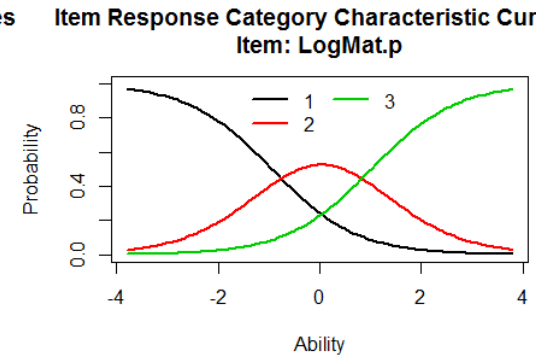

Item Response Category Characteristic Curves
Item: ARCH.p
Probability
Ability
1
2
3

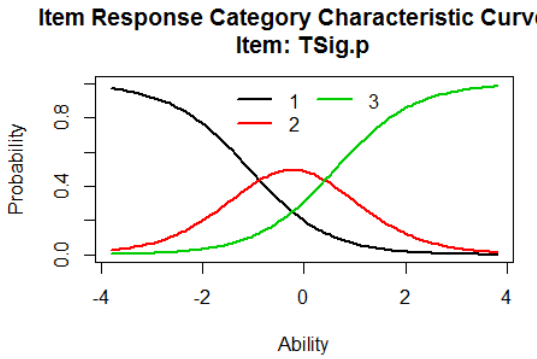

Item Response Category Characteristic Curves
Item: STRD1.p
Probability
Ability
1
2
3

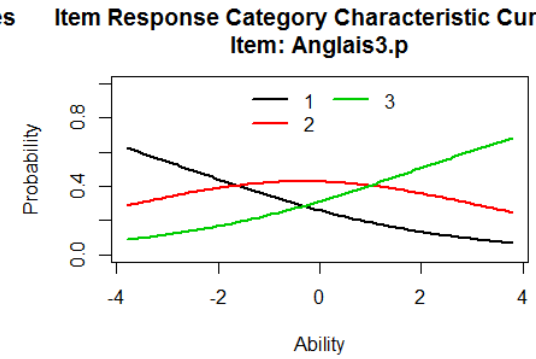

Item Response Category Characteristic Curves
Item: SI.p
Probability
Ability
1
2
3

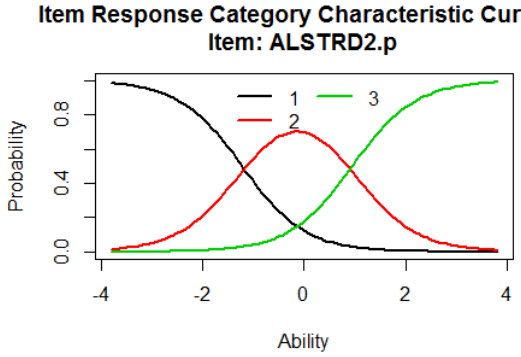

Item Response Category Characteristic Curves
Item: ANum.p
Probability
Ability
1
2
3

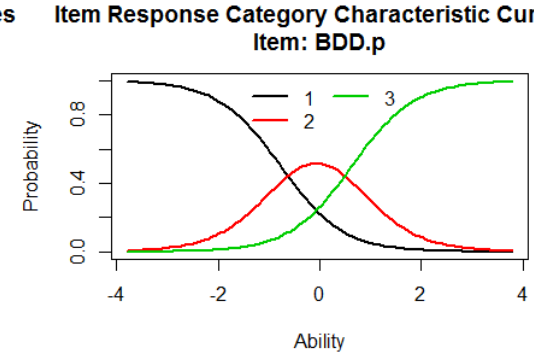

Item Response Category Characteristic Curves
Item: P.S.p
Probability
Ability
1
2
3

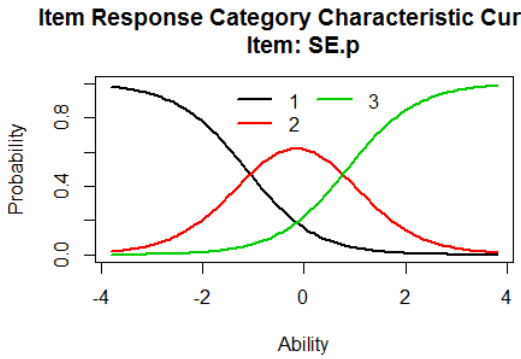

Item Response Category Characteristic Curves
Item: LogMat.p
Probability
Ability
1
2
3

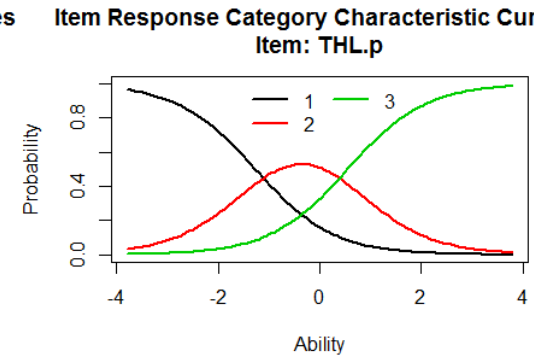

Item Response Category Characteristic Curves
Item: TSig.p
Probability
Ability
1
2
3

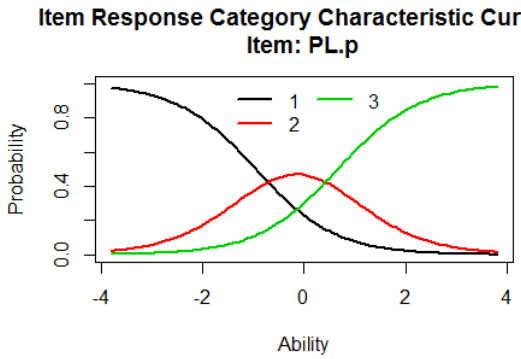

Item Response Category Characteristic Curves
Item: Anglais3.p
Probability
Ability
1
2
3

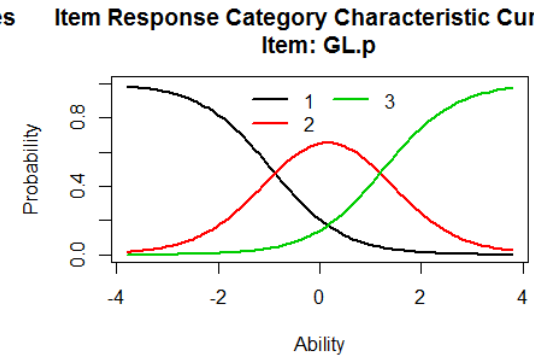

Item Response Category Characteristic Curves
Item: ALSTRD2.p
Probability
Ability
1
2
3

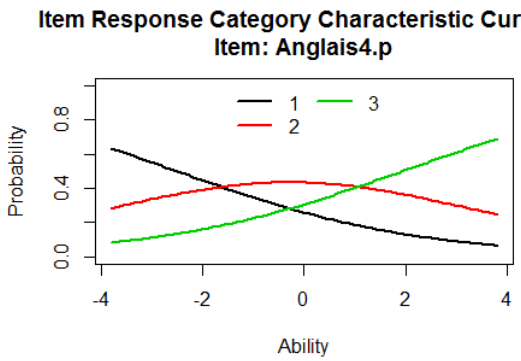

Item Response Category Characteristic Curves
Item: BDD.p
Probability
Ability
1
2
3

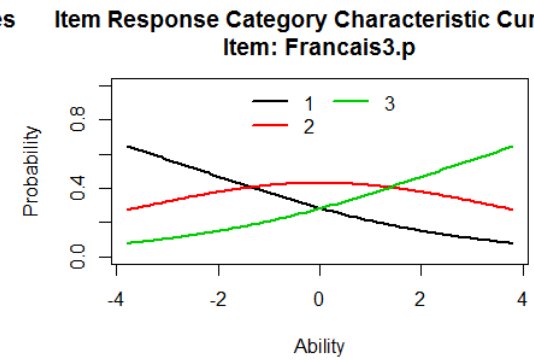

Item Response Category Characteristic Curves
Item: SE.p
Probability
Ability
1
2
3


Item Response Category Characteristic Curves
Item: THL.p
Probability
Ability
1
2
3


Item Response Category Characteristic Curves
Item: PL.p
Probability
Ability
1
2
3


Item Response Category Characteristic Curves
Item: GL.p
Probability
Ability
1
2
3


Item Response Category Characteristic Curves
Item: Anglais4.p
Probability
Ability
1
2
3


Item Response Category Characteristic Curves
Item: Francais3.p
Probability
Ability
1
2
3

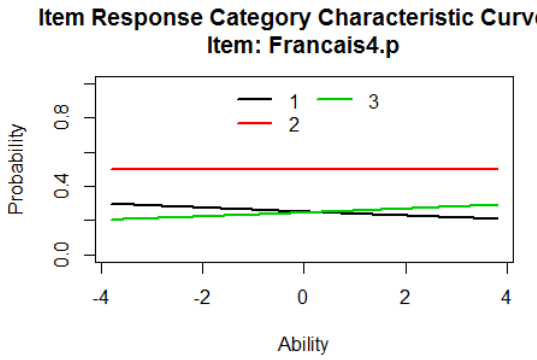
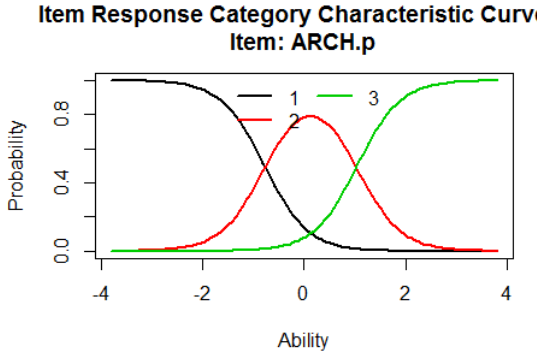
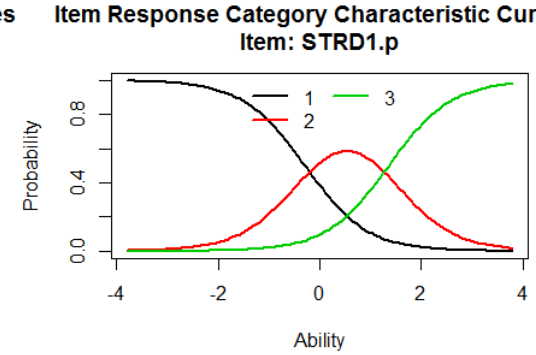
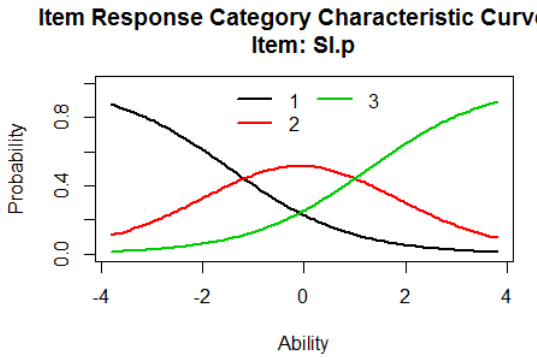
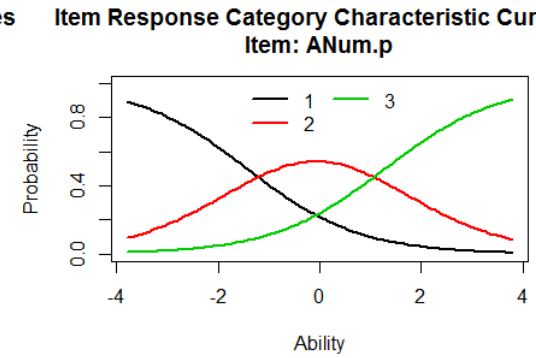


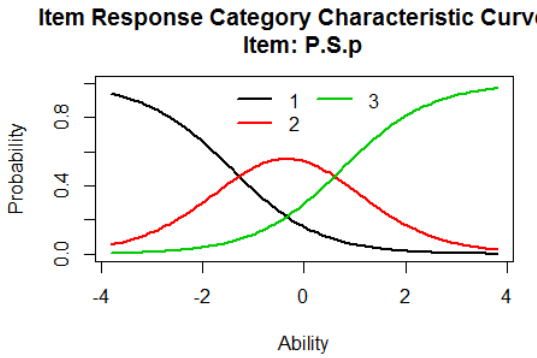
Figure 6 – CCI de licence 2 2010-2011

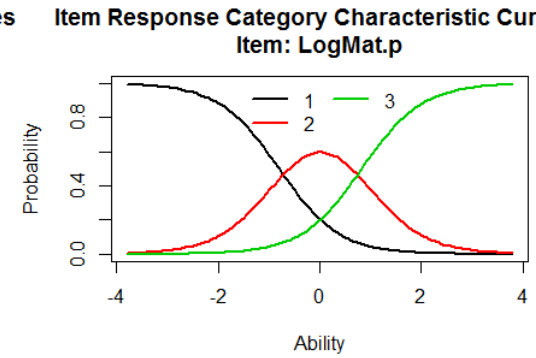


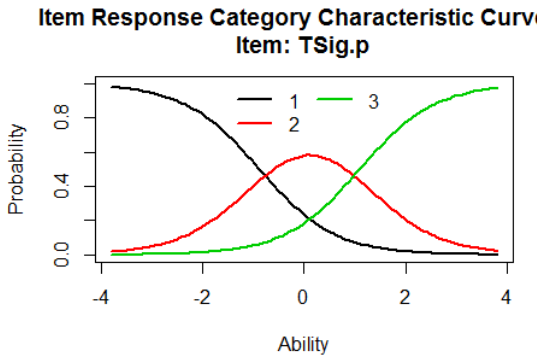


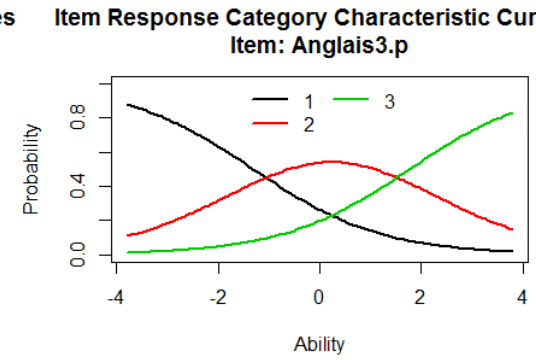


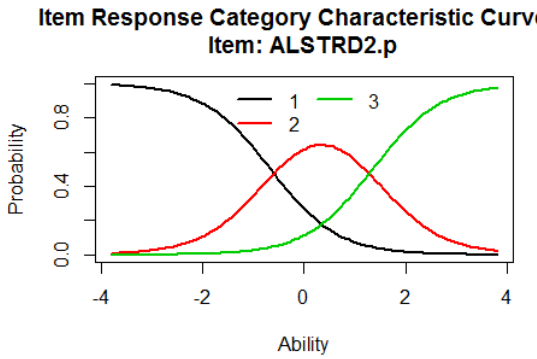


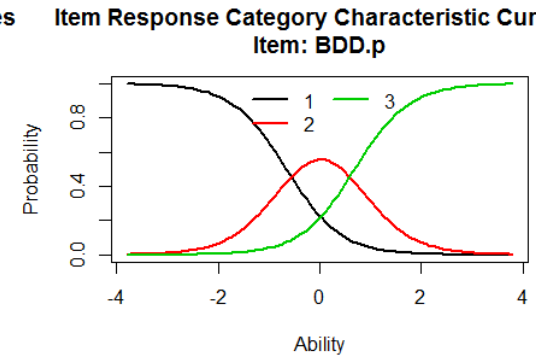


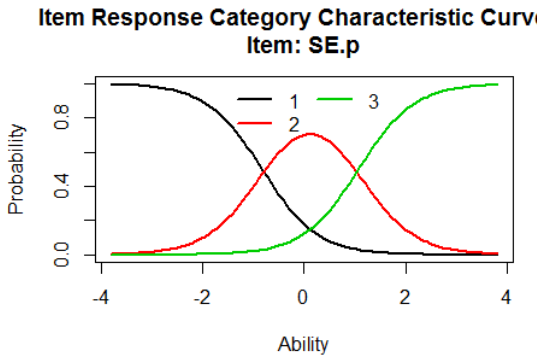


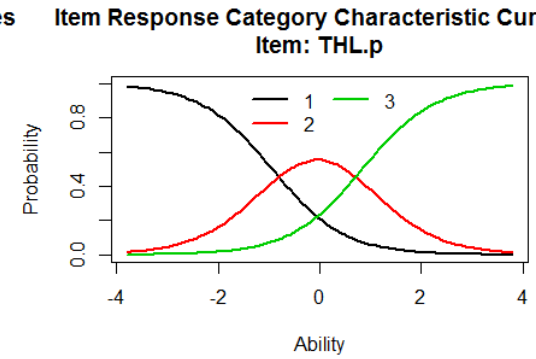

Item Response Category Characteristic Curves
Item: TSig.p
Probability
Ability
1
2
3

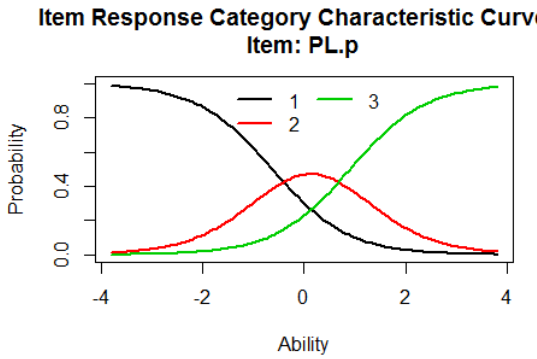

Item Response Category Characteristic Curves
Item: Anglais3.p
Probability
Ability
1
2
3

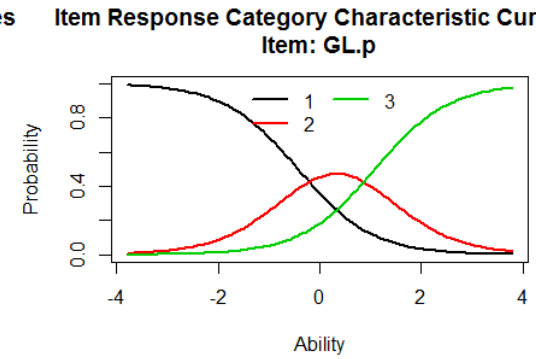

Item Response Category Characteristic Curves
Item: ALSTRD2.p
Probability
Ability
1
2
3


Item Response Category Characteristic Curves
Item: BDD.p
Probability
Ability
1
2
3


Item Response Category Characteristic Curves
Item: SE.p
Probability
Ability
1
2
3


Item Response Category Characteristic Curves
Item: THL.p
Probability
Ability
1
2
3


Item Response Category Characteristic Curves
Item: PL.p
Probability
Ability
1
2
3


Item Response Category Characteristic Curves
Item: GL.p
Probability
Ability
1
2
3

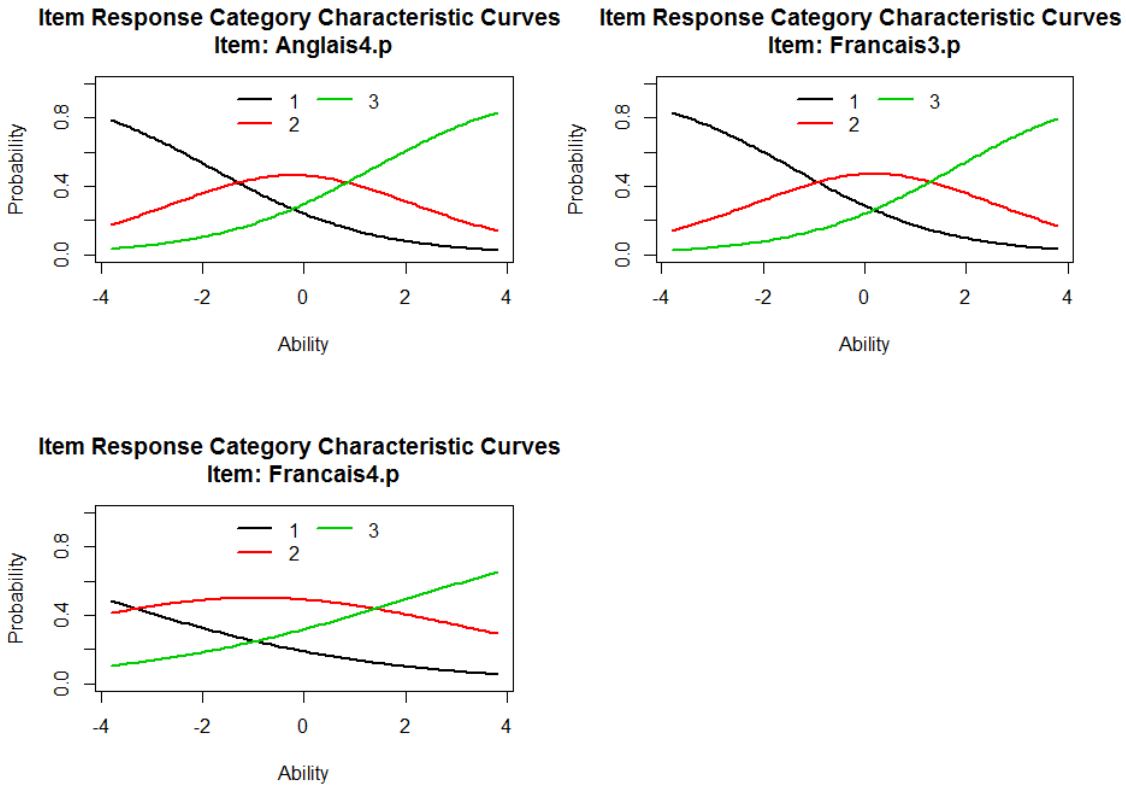
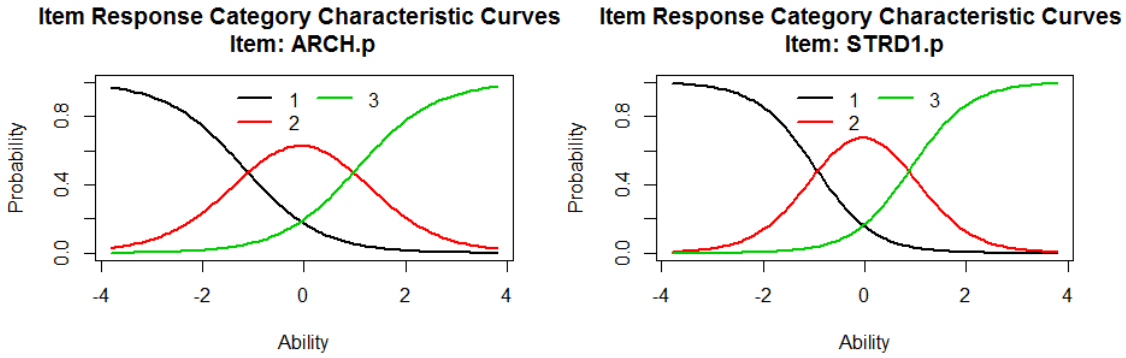
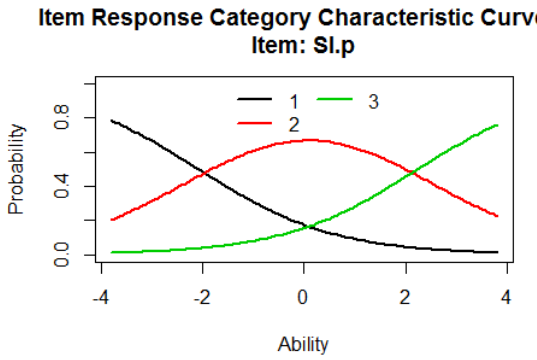
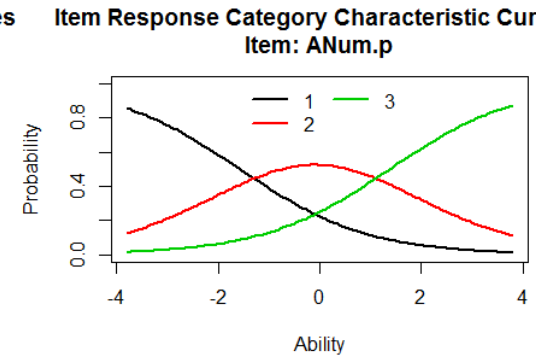
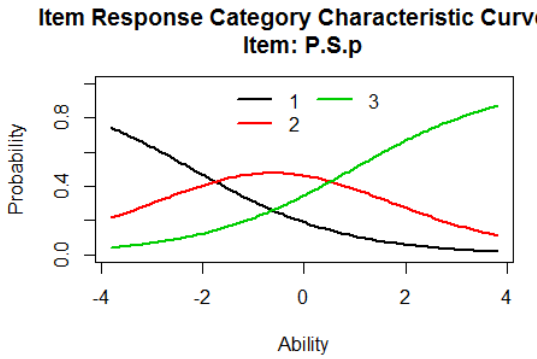


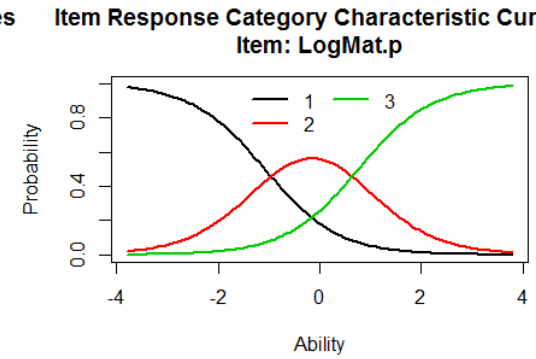
Figure 7 – CCI de licence 2 2011-2012

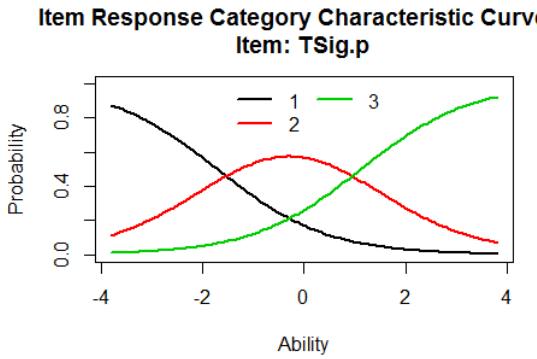


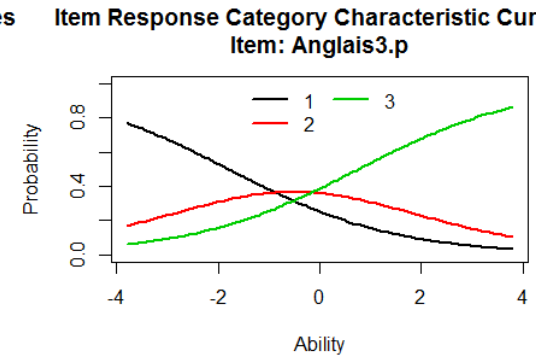

Item Response Category Characteristic Curves
Item: Sl.p
Probability
Ability
1
2
3

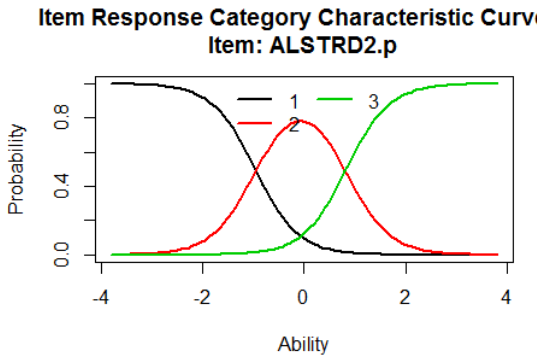

Item Response Category Characteristic Curves
Item: ANum.p
Probability
Ability
1
2
3

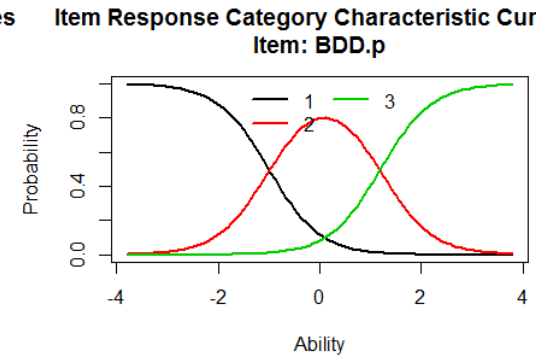

Item Response Category Characteristic Curves
Item: P.S.p
Probability
Ability
1
2
3


Item Response Category Characteristic Curves
Item: LogMat.p
Probability
Ability
1
2
3


Item Response Category Characteristic Curves
Item: TSig.p
Probability
Ability
1
2
3


Item Response Category Characteristic Curves
Item: Anglais3.p
Probability
Ability
1
2
3


Item Response Category Characteristic Curves
Item: ALSTRD2.p
Probability
Ability
1
2
3


Item Response Category Characteristic Curves
Item: BDD.p
Probability
Ability
1
2
3

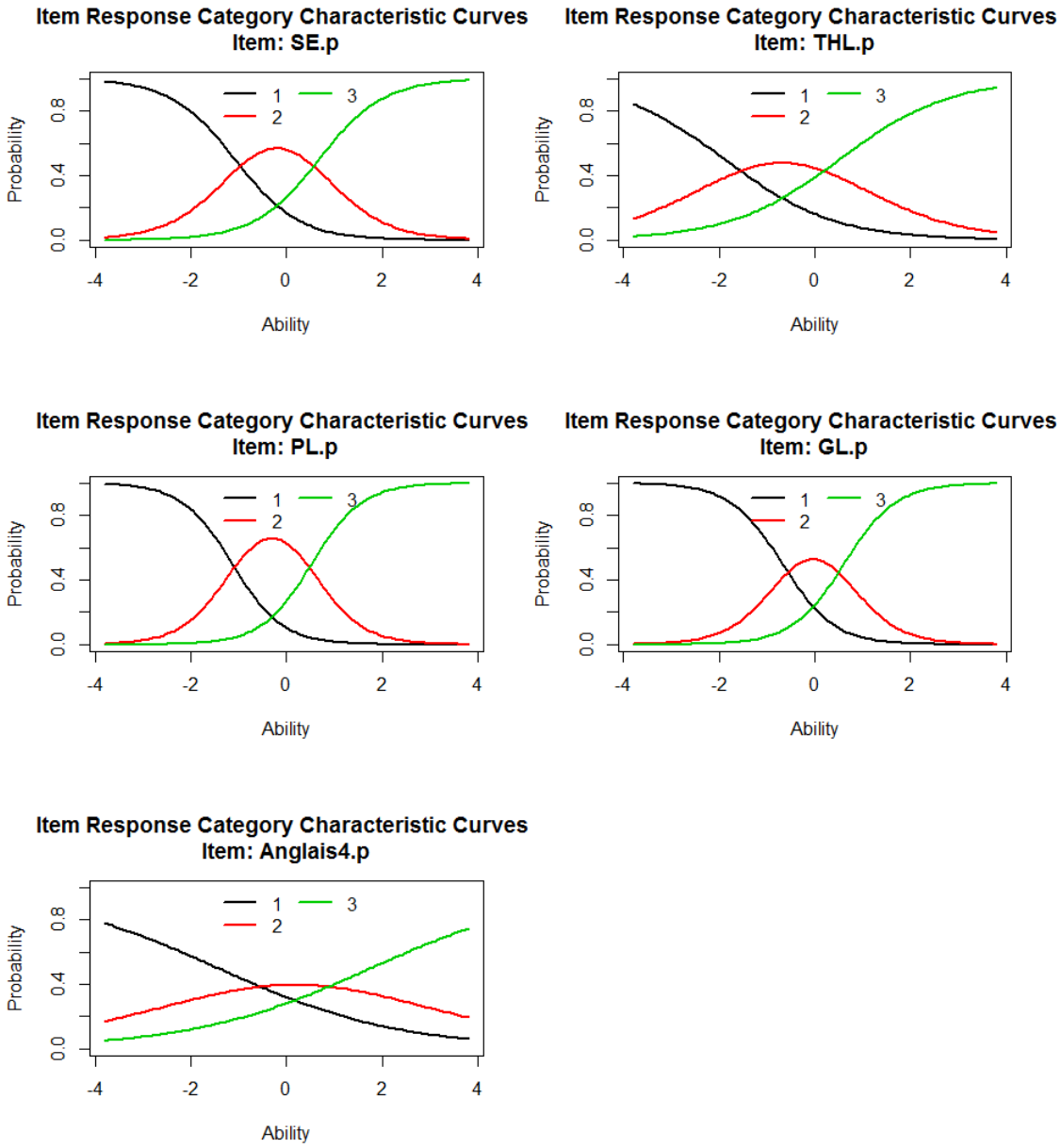


Figure 8 – CCI de licence 2 2012-2013

COORDINACIÓN DE LOS PROCESOS DE APROXIMACIÓN EN LA COMPRENSIÓN DEL LÍMITE DE UNA FUNCIÓN EN UN PUNTO. UNA APROXIMACIÓN A TRAVÉS DEL ANÁLISIS ESTADÍSTICO IMPLICATIVO

Joan PONS[1], Julia VALLS[2], Salvador LLINARES[3]

COORDINATION DES PROCESSUS D'APPROXIMATION DANS LA COMPRÉHENSION DE LA LIMITE D'UNE FONCTION EN UN POINT. UNE APPROCHE AU MOYEN DE L'ANALYSE STATISTIQUE IMPLICATIVE

COORDINATION OF THEAPPROXIMATION PROCESSES IN THE UNDERSTANDING OF THE LIMIT OF A FUNCTION AT A POINT. AN APPROCH THROUGE THE STATISTICAL IMPLICATIVE ANALISYS

RESUMEN

Esta investigación estudia la influencia de la comprensión de la coordinación de las aproximaciones en el dominio y en el rango en la comprensión de la concepción dinámica del límite en estudiantes de Bachillerato. El análisis se realizó usando el análisis implicativo (Gras et al., 2008). Los resultados indican que la comprensión de la concepción dinámica del límite se realiza mediante la coordinación de los procesos de aproximación en el dominio y en el rango, distinguiendo las aproximaciones laterales coincidentes de las no coincidentes en el rango. Además, nuestros resultados también indican que la coordinación de las aproximaciones se inicia en modo gráfico cuando las aproximaciones laterales coinciden y que los modos algebraico-numérico y numérico desempeñan un papel relevante en el desarrollo de la comprensión de la concepción dinámica de límite.

Palabras clave: *Comprensión de límite, coordinación de procesos de aproximación, modos de representación, análisis implicativo, estudiantes de bachillerato*

RÉSUMÉ

Cette recherche étudie l'influence de la compréhension de la coordination des approximations dans le domaine et dans le rang dans la construction de la conception dynamique de limite pour des élèves de la première et de la terminale. L'analyse a été effectuée en utilisant l'analyse implicative (Gras et al., 2008). Les résultats indiquent que la construction progressive de la conception dynamique de limite se fait par des processus d'approximation différenciées de la coordination des processus d'approximation dans le domaine et dans le rang et, en distinguant ceux des approximations latérales qui coïncident avec ceux qui ne coïncident pas. En outre, nos résultats indiquent également que la coordination des approximations commence en mode graphique lorsque les approximations latérales coïncident et que les modes algébriques-numériques et numériques jouent un rôle important dans le développement d'une compréhension de la notion dynamique de limite.

Mots-clés**:** *Compréhension des limites, coordination du processus d'approche, modes de représentation, analyse implicative, élèves de la première et de la terminale.*

[1] I.E.S Mutxamel, C/Mondúber, sn Mutxamel (Alicante), jpt11@alu.ua.es
[2] Universidad de Alicante, Sant Vicent del RAspeig (Alicante), julia.valls@ua.es
[3] Universidad de Alicante, Sant Vicent del RAspeig (Alicante), sllinares@ua.es

ABSTRACT

This research studies the influence of the understanding of the coordination of the approximations in the domain and range in the construction of the dynamic conception of the limit when teaching it to high school students. The analysis was carried out using the implicative analysis (Gras et al., 2008). The findings indicate that the gradual construction about the dynamic conception is made by processes that differentiate the coordination of approximation processes from those, in which the lateral approximation match of mismatched. In addition, our results indicate that the coordination of the approximations begins in graphic mode when the approximations coincide. To end, the algebraic-numeric and numeric modes play an important role in the development of the understanding of the dynamic conception of limit.

Keywords: *Understanding of limit concept, coordination of approximation processes, representation modes, implicative analysis, high school students.*

1 Introducción

La comprensión de la noción límite de una función real de variable real es un aspecto fundamental del análisis. La importancia del concepto de límite radica en que es una idea básica del análisis matemático por estar vinculada a conceptos como continuidad, derivada e integral. Las diferentes investigaciones indican que para los estudiantes el concepto de límite es una noción difícil (Sierpinska, 1985) que presenta dificultades cognitivas (Cornu, 1991). La influencia que tienen las diferentes representaciones en la comprensión del concepto de límite (Blázquez y Ortega, 2000; Lacasta y Wilhelmi, 2010; Monaghan, 2001) constituye, junto con el papel que desempeñan los obstáculos epistemológicos en el acceso al concepto de límite (Cornu, 1981; Garbin y Azcárate, 200; Moru, 2007; Sierpinska, 1985) y el papel de las concepciones espontáneas en las dificultades que presenta el propio concepto (Elia et al., 2009; Fernández-Plaza et al., 2013; Monaghan, 1991; Oehrtman, 2009; Szydlik, 2000; Williams, 1991), uno de los ámbitos en los que se han centrado las investigaciones sobre la comprensión que tienen los estudiantes del concepto de límite de una función

Una dificultad añadida a todo concepto matemático es la distinción entre la imagen del concepto y la definición del concepto (Tall y Vinner, 1981). Para estos autores la definición del concepto significa su definición formal tal como la entiende la comunidad matemática; la imagen del concepto significa toda estructura cognitiva asociada con el concepto e incluye todas las representaciones mentales con los procesos y propiedades asociadas. Pero debemos tener en cuenta que toda representación es cognitivamente parcial en relación a lo que representa, y, además, representaciones de registros diferentes no presentan los mismos aspectos de un mismo contenido conceptual (Duval, 1995). La búsqueda de soluciones, mediante enfoques didácticos, con la finalidad de superar las dificultades de los estudiantes en la comprensión del concepto de límite ha sido ampliamente investigada (Blázquez y Ortega, 2000; Engler et al., 2007; Kidron, 2010; Swinyard, 2011; Mamona-Downs, 2001; Mira et al., 2013), si bien no resulta fácil encontrar soluciones.

1.1 El límite de una función real de variable real como objeto de aprendizaje

Cauchy (1821) definió el limite como:"*cuando los valores atribuidos sucesivamente a la misma variable se aproximan indefinidamente a un valor fijo, de manera que terminen difiriendo de él tan poco como se quiera, este último valor se llaman el límite de todos los demás*" (p.4). Esta definición, al ser de carácter aritmético, hace de ella un concepto dinámico. Con posterioridad, Weierstrass, dio una definición métrica (1):

$$\lim_{x \to a} f(x) = L \Leftrightarrow \forall \varepsilon > 0, \exists \partial > 0 : \forall x, \ 0 < |x - a| < \delta \Rightarrow |f(x) - L| < \varepsilon \qquad (1)$$

que es la que actualmente se utiliza en los primeros cursos universitarios. En contraposición a la concepción dinámica, la definición métrica de límite representa una concepción estática.

Desde una perspectiva didáctica, Blázquez y Ortega (2002) consideran la concepción dinámica como una aproximación óptima de la siguiente forma: *"Sea "f" una función y "a" un número real, el número "L" es el límite de la función "f" en el punto "a",si cuando "x" se acerca al número "a" más que cualquier aproximación, sus imágenes "f(x)" se acercan a "L" más que cualquier otra aproximación fijada"*

Desde una perspectiva cognitiva, Cottrill et al. (1996) afirman que la concepción dinámica del límite debe visualizarse y, para ello, el estudiante debe construir "*un proceso en el dominio y un proceso en el rango y usar la función para coordinarlos*" (p.174). Sin embargo, también sugieren que la concepción dinámica del límite, es más complicada de lo que se había pensado, ya que en la comprensión de la coordinación de los procesos de aproximación desempeñan un papel determinante los diferentes modos de representación.

La controversia entre la concepción dinámica versus concepción métrica es ampliamente analizada (Cottrill et al., 1996; Williams, 1991). Sin embargo, investigaciones recientes han mostrado la necesidad de inducir en los estudiantes imágenes dinámicas que sean compatibles con la definición formal de límite (Roh, 2008), y han puesto de manifiesto mediante experimentos de enseñanza que la forma en la que los estudiantes pueden desarrollar una rica y sólida comprensión de la definición de límite se apoya en la habilidad para usar de forma flexible las dos perspectivas (Swinyard, 2011).

1.2 La influencia de las diferentes representaciones en la comprensión del concepto de límite de una función en un punto

Los resultados de las investigaciones realizados sobre la influencia de las diferentes representaciones en la comprensión del concepto de límite de una función indican la necesidad de introducir la noción de límite con distintas representaciones (Blázquez, 1999). La elección de los distintos modos de representación depende de los aspectos del límite que se quieran subrayar (Engler et al., 2007). Los resultados de la investigación realizada por Engler et al. (2007) sobre los diferentes papeles que juegan los modos de representación indican que a los estudiantes les resulta más sencillo identificar la tendencia de una función en forma numérica o gráfica que en la algebraica y que cometían más errores para encontrar a qué valor se aproxima la función que para determinar a qué valores se aproxima la variable independiente. Las dificultades que

para los estudiantes representa el modo de representación algebraico, al resolver problemas de límites, se producen cuando realizan manipulaciones algebraicas (Moru, 2009), porque estas manipulaciones pueden ser la llave o la cerradura (Bergsten, 2006), puesto que las manipulaciones algebraicas pueden iniciar o bloquear el concepto. Sin embargo, también hay investigaciones que indican que la representación gráfica es más sólida que la numérica en el sentido de que los estudiantes resuelven mejor las situaciones gráficas que las numéricas (Monaghan, 2001).

Las investigaciones centradas en buscar diferentes alternativas didácticas a la definición formal también centran su atención en el papel de los modos de representación. Una de ellas utiliza el modo de representación gráfico asociándolo a la definición formal (Mamona-Downs, 2001); otra utiliza el modo de representación numérico y presenta la conceptualización del límite como aproximación óptima (Blázquez y Ortega, 2000) al considerar, que la utilización de la conceptualización de límite como aproximación óptima es más sencilla que la conceptualización formal (Blázquez et al., 2006). Sin embargo, los estudiantes que utilizan la definición formal y los que utilizan la noción de límite de un modo informal cuando trabajan con "números grandes" y con "números pequeños" cometen los mismos errores (Parameswaran, 2007).

Las investigaciones anteriores nos han proporcionado cierta información sobre las características de la comprensión del concepto de límite en los estudiantes, y en particular el papel desempeñado por las traslaciones entre los diferentes modos de representación, pero no aportan información sobre el papel de la coordinación de los procesos de aproximación en el dominio y en rango en la comprensión de la concepción dinámica del límite. Por ello nos hemos planteado las siguientes preguntas:

¿Cuál es el papel que desempeña la comprensión de la coordinación de los procesos de aproximación en el dominio y en el rango en la construcción de la concepción dinámica del límite?

¿Cómo influyen los distintos modos de representación en el acceso a la concepción dinámica de límite?

2 Método

2.1 Participantes

En la investigación han participado 129 estudiantes de bachillerato de edades comprendidas entre 16-18 años (66 de 1º y 63 de 2º curso de Bachillerato de Ciencias de la Naturaleza y de la Salud). Estos estudiantes habían sido iniciados a "*una aproximación al concepto de límite*" trabajando con: el concepto de función, la idea intuitiva de límite, los límites laterales, el cálculo de límites, límites en el infinito, límites infinitos, el cálculo de asíntotas, la noción de continuidad, sucesiones y límites de sucesiones. A los estudiantes de 1º de Bachillerato se les había introducido la noción de límite de una función en un punto dos semanas antes de contestar el cuestionario. A los de 2º de Bachillerato les habían introducido la misma noción seis meses antes de que

realizaran el cuestionario. Los estudiantes no tenían ninguna característica especial y su participación fue voluntaria.

2.2 Instrumentos y procedimiento de recogida de datos

Los datos provienen de un cuestionario y de las entrevistas realizadas a 21 estudiantes. Para la elaboración del cuestionario tuvimos en cuenta los resultados de las investigaciones sobre el concepto de límite de una función, analizamos libros de texto y materiales curriculares. Para el diseño de las diferentes cuestiones caracterizamos las nociones de aproximación dinámica y métrica de límite de una función con el objetivo de identificar los elementos matemáticos que la constituyen (Figura 1).

- *Sea f una función y x_0 un número real. El valor de la función f en $x=x_0$, $f(x_0)$ si existe o no, (E0)*
- *Idea de aproximación*
  - *x se aproxima al número a (E1)*
  - *f(x) se aproxima a L (E1)*
- *Coordinación en la concepción dinámica: cuando x se aproxima al número a, sus imágenes f(x) se aproximan a L (E2)*
- *Coordinación en la concepción métrica: se puede encontrar para cada ocasión un x suficientemente cerca de a tal que el valor de f(x) sea tan próximo a L como se desee (E4)*
- *Formalización como una manifestación de ser consciente de la existencia del límite L de la función f(x) en el punto a, escribiendo lím f(x) = L (E3)*

Figura 1– Elementos matemáticos considerados en el esquema límite de una función

Los ítems del cuestionario tenían como finalidad obtener información sobre la influencia que la coincidencia o no de las aproximaciones laterales, en el rango, tiene en la comprensión que los estudiantes poseen de la coordinación de las aproximaciones y el papel que desempeñan los diferentes modos de representación en esta coordinación. Se diseñaron 10 tareas con un total de 34 ítems. En esta investigación solo vamos a estudiar los datos correspondientes a las tareas 1, 2, 3, 6, 7, y 8 (Figura 2) que hacen referencia a la concepción dinámica (E0, E1 y E2) en tres modos de representación: numérico (N), gráfico (G), y algebraico-numérico (AN) y considerando la coincidencia o no de las aproximaciones laterales en el rango.

Las preguntas son:

1. "¿A qué número se aproxima la x y la f(x)?" (Elemento E1);
2. "Describe el comportamiento de la función, f(x), con relación al comportamiento de la variable x" (Elemento E2);
3. "Di, si es posible, cuál es el límite de la función en x =a" (Elemento E3).

Las aproximaciones en el dominio (la x) siempre se plantean en modo numérico y siempre son coincidentes. Las aproximaciones en el rango (la f(x)) se presentan en modos numérico y algebraico-numérico y pueden ser coincidentes o no coincidentes. En las tareas presentadas en el modo gráfico, 2 y 7, solo se pregunta a qué se aproxima la función cuando la x toma determinados valores. La tarea 3 proviene de Engler et al. (2007). Las otras las hemos elaborado nosotros.

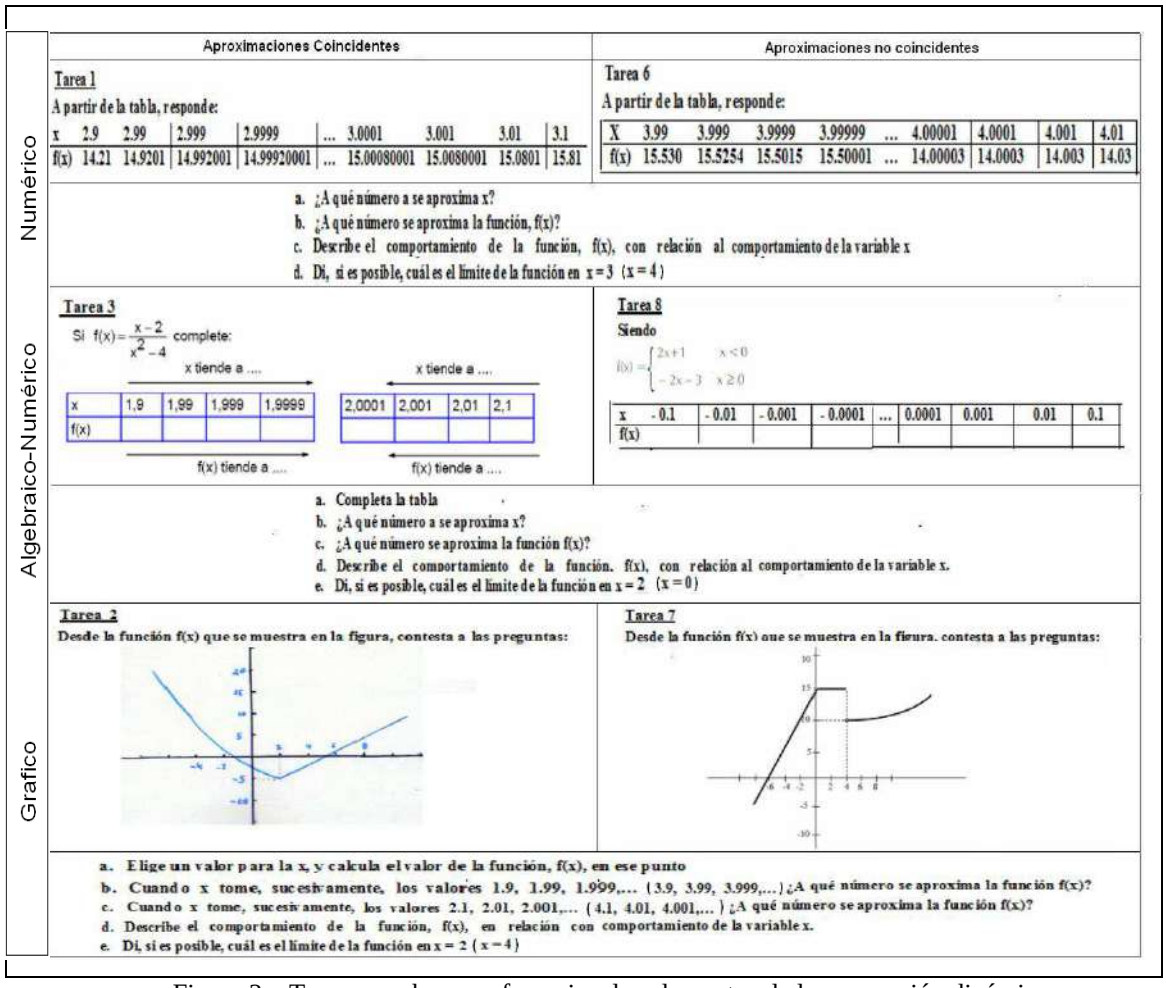


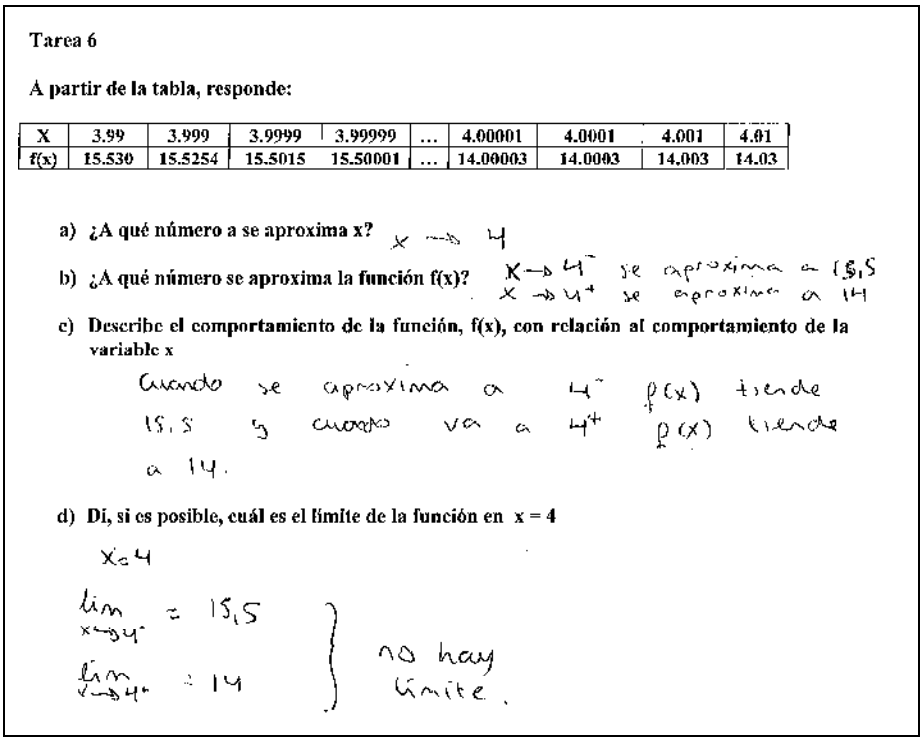
Figura 2 – Tareas que hacen referencia a los elementos de la concepción dinámica

2.3 Análisis

Cada ítem fue codificado indicando el elemento matemático que se pone de manifiesto en el ítem (Ei, i = 0, 1, 2, 3), el modo de representación usado en la presentación del problema (G, N, AN), la tarea a la que corresponde y su posición en la misma, a través de un número y una letra en minúscula. Por ejemplo, E1N6a, indica el ítem a de la tarea 6 presentado en modo Numérico y hace referencia al elemento matemático E1. Inicialmente, un grupo de tres investigadores realizaron una lectura conjunta de las respuestas a cada uno de los ítems de las tareas con el objetivo de generar criterios y unificar la puntuación dicotómica, (1 respuesta correcta, 0 incorrecta).

Ejemplificamos el análisis realizado a partir de las respuestas dadas a la tarea 6 por la estudiante EST36 (Figura 3). La respuesta a la tarea 6 fue puntuada como (1,1,1,1) dado que EST36 respondió de forma correcta ¿a qué número se aproxima la x y la f(x)? (ítems 6a y 6b) al asociar la aproximación a un número en el dominio a un número natural ($x \rightarrow 4$); la aproximación en el rango a diferentes números por la izquierda y por la derecha de x=4 (f(x) *se aproxima a 15,5 cuando* $x \rightarrow 4^-$ *y cuando* $x \rightarrow 4^+$ *se aproxima a 14*). También respondió correctamente al ítem 6c- Describe el comportamiento de la

función, f(x), con relación al comportamiento de de la variable x- al coordinar las aproximaciones en el dominio y el rango (*Cuando x se aproxima a 4^-, f(x) tiende a 15,5 y cuando se aproxima a 4^+, f(x) tiende a 14*). Finalmente, el ítem 1d se puntuó con un 1 al indicar que "*no hay límite*" después de diferenciar que los límites laterales de la función son distintos a la derecha y a la izquierda.



Tarea 6

A partir de la tabla, responde:

X	3.99	3.999	3.9999	3.99999	…	4.00001	4.0001	4.001	4.01
f(x)	15.530	15.5254	15.5015	15.50001	…	14.00003	14.0003	14.003	14.03

a) **¿A qué número a se aproxima x?** $x \to 4$

b) **¿A qué número se aproxima la función f(x)?** $x \to 4^-$ se aproxima a 15,5 $x \to 4^+$ se aproxima a 14

c) **Describe el comportamiento de la función, f(x), con relación al comportamiento de la variable x**

Cuando se aproxima a 4^- f(x) tiende 15,5 y cuando va a 4^+ f(x) tiende a 14.

d) **Dí, si es posible, cuál es el límite de la función en x = 4**

$x \to 4$

$\lim_{x \to 4^-} = 15,5$

$\lim_{x \to 4^+} = 14$

} no hay límite.

Figura 3 – Respuesta de la EST36 a la Tarea 6

De esta manera, cada estudiante venía definido por una 28-tuplas (variables) relativas a la valoración de las respuestas a las tareas centradas en la concepción dinámica (1, 2, 3, 4, 6, 7, 8).

A continuación realizamos un análisis estadístico implicativo utilizando el software CHIC (Classification Hiérarchique Implicative et Cohésitive) (Gras et al., 2008). En la estadística implicativa, Trigueros y Escandon (2008) señalan que dada una población E, los estudiantes, y un conjunto de variables, los ítems del cuestionario, "*se busca dar sentido estadístico a una implicación no estricta $a \Rightarrow b$*" (pág. 66). En esta metodología, "*la implicación $a \Rightarrow b$ será admisible en una experiencia si el número de individuos de E que la contradicen es muy pequeño, en términos probabilísticos, en relación con el número de individuos esperado bajo la hipótesis de ausencia de relación. Si esto ocurre, se puede decir que A, conjunto de observaciones que satisfacen la característica a, está "casi" contenido en B, conjunto de observaciones que satisfacen la característica b"* (pág. 67).

El análisis implicativo genera un gráfico jerárquico y no transitivo, que permite identificar variables que se unen a través de implicaciones binarias ($a \Rightarrow b$, "a" entonces "b") o mediante implicaciones no binarias ($a \Rightarrow (b \Rightarrow c)$, si "a" y "b" entonces "c") haciendo emerger una primera estructura conceptual entre diferentes elementos matemáticos que conforman la concepción dinámica de límite. Este análisis traduce gráficamente el conjunto de las relaciones cuasi-implicativas entre las variables en distintos niveles de significación.

En nuestro estudio la población es la muestra de estudiantes de educación postobligatoria que respondieron al cuestionario y las variables son los ítems de las diferentes tareas considerando el elemento matemático y el modo de representación en que se presenta la tarea. Las relaciones implicativas entre las diferentes variables nos indican en qué medida el uso de un elemento matemático en una tarea (planteada en un modo de representación concreto y desde aproximaciones coincidentes o no en el rango) está relacionado con el uso de ese mismo u otro elemento matemático en otra tarea (en ese u otro modo de representación y coincidiendo o no las aproximaciones laterales en el rango) con una determinada probabilidad. La medida de esta relación nos permite inferir cómo se está construyendo el acceso al concepto de límite de una función en un punto.

En el análisis estadístico implicativo realizado en este trabajo (al 99% de significación), hemos usado una de las posibilidades del programa CHIC para *"suprimir o centrarse"* solamente en determinadas variables (Couturier, 2009; Gras y Kuntz, 2009) con la finalidad de resaltar determinadas relaciones entre elementos matemáticos. En los gráficos implicativos generados, los ítems que inician las ramas indican que los estudiantes que responden adecuadamente a ellos también lo hacen a la mayoría de los ítems que aparecen en la parte inferior del gráfico (Bodi et al., 2009; Zamora et al., 2009)

3 Resultados

Los resultados están organizados en tres secciones. En la primera, describimos la influencia de los modos de representación en la manera en la que los estudiantes usaban la idea de aproximación a un número (E1). En la segunda, describimos cómo la no coincidencia de las aproximaciones laterales en el rango, determina un punto de referencia en la comprensión dinámica de límite. Finalmente, en la tercera, describiremos los resultados relativos a la influencia de los modos de representación en la comprensión de la coordinación dinámica de límite (E2).

3.1 La influencia de los modos de representación en el uso de la idea de aproximación a un número

Para poner de manifiesto la influencia de los modos de representación en la comprensión de la aproximación a un número (E1) seleccionamos las variables que hacen referencia a este elemento en los modos de representación numérico y algebraico-numérico. La selección realizada dio lugar al gráfico implicativo que se muestra en la figura 4.

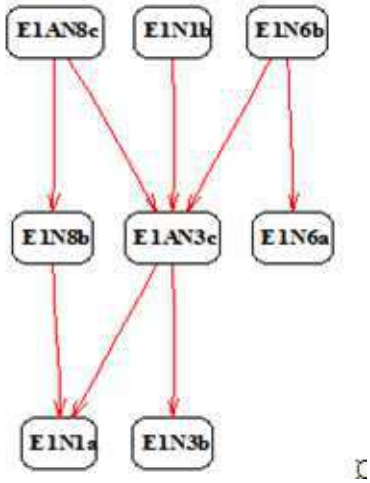


Figura 4 – Relaciones implicativas al 99% de las variables relativas al elemento E1

Desde este gráfico podemos inferir cuatro ideas relevantes relativas al papel de los modos de representación en relación a la comprensión de la idea de aproximación a un número.

La primera idea la inferimos a partir de las relaciones implicativas de la parte superior derecha de la figura 4:

- E1N1b $\Rightarrow$ E1AN3c ;
- E1N6b $\Rightarrow$ E1AN3c.

Estas dos relaciones implicativas sugieren que si los estudiantes comprenden la aproximación a un número en el rango, en modo numérico, tanto cuando las aproximaciones laterales son coincidentes (E1N1b) como cuando no son coincidentes (E1N6b), también comprenden las aproximaciones laterales coincidentes en modo algebraico-numérico (E1AN3c).

La segunda idea la inferimos a partir de las relaciones implicativas de la parte superior izquierda y central de la figura 4:

- E1AN3c $\Rightarrow$ E1N1a ;
- E1AN3c $\Rightarrow$ E1N3b ;
- E1AN8c $\Rightarrow$ E1N8b.

Estás tres relaciones implicativas indican que si los estudiantes comprenden la aproximación a un número en modo algebraico-numérico, tanto cuando las aproximaciones laterales son coincidentes (E1AN3c) como cuando no son coincidentes (E1AN8c), también comprenden las aproximaciones laterales coincidentes en el modo numérico (E1N1a, E1N3b, E1N8b).

Estas dos ideas sugieren que **los estudiantes utilizan indistintamente los modos de representación numérico y algebraico-numérico** para acceder a la idea de aproximación a un número.

La tercera idea la inferimos a partir de las relaciones implicativas que muestra la figura 5 (extraída de la figura 4) en las que las cuatro variables relacionadas hacen referencia a la idea de aproximación en el rango. En esta subestructura la idea de aproximación en modo algebraico-numérico cuando las aproximaciones laterales

coinciden (E1AN3c) desempeña un papel relevante al agrupar al resto de las variables que hacen referencia a la idea de aproximación en el rango.

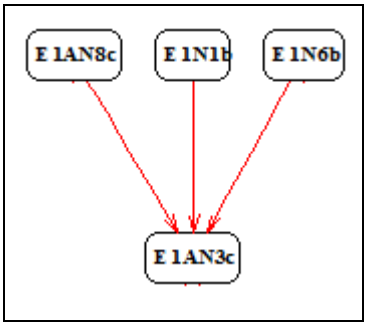


Figura5 – Relaciones al 99% entre las variables "idea de aproximación numérica", en el rango (E1)

Las relaciones implicativas:

- E1AN8c ⇒ E1AN3c ;
- E1N6b ⇒ E1AN3c ;

ponen de manifiesto que la comprensión de la aproximación a un número en el rango cuando las aproximaciones laterales no coinciden (E1AN8c y E1N6b) se apoya en la comprensión de la aproximación a un número en el rango cuando las aproximaciones laterales coinciden (E1AN3c). Es decir, **la comprensión de la aproximación a un número en el rango se inicia en modo algebraico-numérico con aproximaciones laterales coincidentes y se consolida cuando las aproximaciones laterales no coinciden (representadas en el modo numérico y numérico-algebraico).**

Globalmente consideradas, las implicaciones generadas parecen indicar que la comprensión de la aproximación en el rango (E1N1b, E1AN3c, E1N6b y E1AN8c) está vinculada a la comprensión previa de la aproximación en el dominio (E1N8b, E1N1a, E1N3b y E1N6a) pero la relación inversa no se establece. Es decir, la comprensión de la aproximación en el dominio no implica la comprensión de la aproximación en el rango. Este hecho parece establecer **una diferencia cognitiva entre la comprensión de los procesos de aproximación en el dominio y en el rango**.

3.2 La idea de aproximación a un número y la comprensión dinámica de límite. El papel de la no coincidencia de las aproximaciones laterales

Con la finalidad de analizar cómo la no coincidencia de las aproximaciones laterales determinan la comprensión dinámica del límite entendida como la coordinación de las aproximaciones en el dominio y en el rango" (E2), seleccionamos las variables que hacen referencia a la idea de aproximación y a la idea de coordinación de las aproximaciones Esta selección dio lugar al gráfico implicativo generado al 99% de significación que muestra la figura 6.

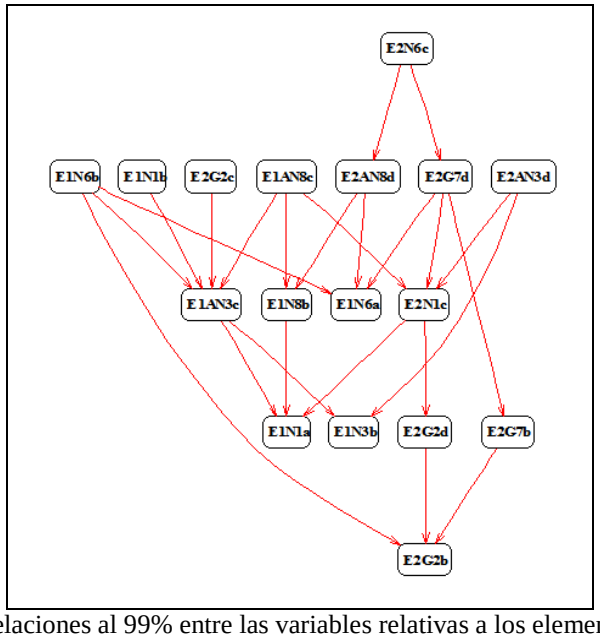


Figura 6 – Relaciones al 99% entre las variables relativas a los elementos E1 y E2

En este gráfico las relaciones en las que toma parte la idea de aproximación a un número en el rango cuando las aproximaciones laterales no coinciden permiten identificar dos ideas del papel que desempeña la idea de aproximación a un número en la comprensión dinámica de límite

- E1AN8c $\Rightarrow$ E2N1c ;
- E1N6b $\Rightarrow$ E2G2b.

En primer lugar, podemos inferir que si los estudiantes comprenden la aproximación a un número en el rango en modo algebraico-numérico cuando las aproximaciones laterales no coinciden (E1AN8c), serán capaces de identificar la coordinación de los procesos de aproximación en el dominio y en el rango en modo numérico cuando las aproximaciones laterales coinciden (E2N1c). En segundo lugar, si los estudiantes comprenden la aproximación a un número en el rango cuando las aproximaciones laterales no coinciden en modo numérico (E1N6b), coordinarán por la izquierda las aproximaciones coincidentes en el dominio y en el rango en modo gráfico (E2G2b).

Estos datos indican que comprender la aproximación no coincidente implica ser capaz de coordinar las aproximaciones coincidentes en el dominio y en el rango. Es decir, **la no coincidencia de las aproximaciones laterales a un número en el rango parece que determina la posibilidad de coordinar las aproximaciones entre el dominio y el rango que subyace a la concepción dinámica de limite**.

3.3 El papel de los modos de representación en la coordinación de las aproximaciones.

Para poner de manifiesto la influencia de los modos de representación en la comprensión de la coordinación de las aproximaciones (E2) seleccionamos las variables que hacen referencia a la coordinación en los distintos modos de representación, numérico, gráfico y algebraico-numérico. Del gráfico implicativo generado al 99% de significación que se muestra en la Figura 7, podemos extraer tres ideas relativas al papel

de los modos de representación comprensión de la coordinación de las aproximaciones en el dominio y en el rango.

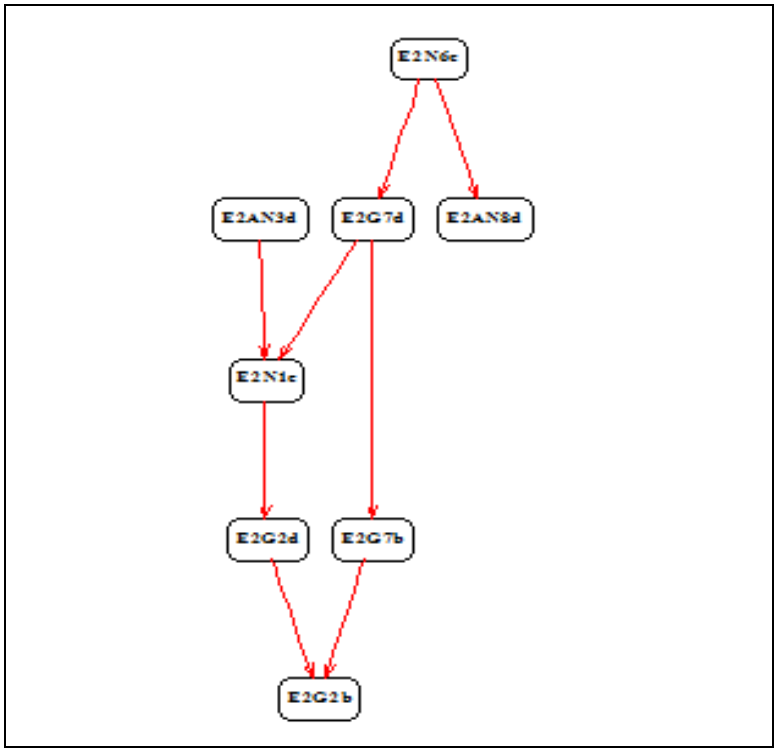


Figura 7 – Relaciones al 99% entre variables relativas al elemento E2

La primera idea la inferimos a partir de las relaciones implicativas (cuando las aproximaciones laterales coinciden) que se encuentran en la parte izquierda de la figura 7:

- E2AN3d ⇒ E2N1c ;
- E2N1c ⇒ E2G2d.

Estas dos relaciones implicativas indican que la coordinación de las aproximaciones cuando estas coinciden se inicia en modo gráfico (E2G2d), progresa en modo numérico (E2N1c) y se consolida en modo algebraico-numérico (E2AN3d). Es decir, **la comprensión de la coordinación de las aproximaciones coincidentes en el dominio y en el rango se inicia en modo gráfico, progresa en modo numérico y se consolida en modo algebraico-numérico**. Las relaciones implicativas entre estas variables indican que **si la comprensión que tienen los estudiantes de la coordinación de los procesos de aproximación en modo algebraico-numérico va acompañada de la comprensión de dicha coordinación en modo numérico, entonces también la comprenderán en modo gráfico**.

La segunda idea la inferimos a partir de las relaciones implicativas de la parte superior de la figura 7:

- E2N6c ⇒ E2AN8d ;
- E2N6c ⇒ E2G7d.

Estas dos relaciones implicativas indican que la comprensión de la coordinación de las aproximaciones no coincidentes en el rango se inicia indistintamente en modos gráfico (E2G7d) y algebraico-numérico (E2AN8d), pero se consolida en modo numérico (E2N6c).

La tercera idea la inferimos a partir de la relación implicativa que se encuentran en la parte central de la figura 7:

– E2G7d $\Rightarrow$ E2N1c.

Esta relación implicativa muestra que **el nexo de unión entre la coordinación de los procesos de aproximaciones no coincidentes en el rango y la coordinación de los procesos de aproximaciones coincidentes en el rango se realiza con los modos gráfico (G) y numérico (N).**

4 Conclusión y discusión

El objetivo de esta investigación ha sido aportar información sobre cómo los estudiantes de primero y segundo de bachillerato construyen el significado del concepto de límite de una función en un punto a partir de la coordinación de los procesos de aproximación (coincidentes o no en el rango) y qué papel desempeñan los modos de representación. Los resultados indican que la comprensión de la coincidencia o no coincidencia de las aproximaciones laterales en el rango es un elemento importante en el proceso de construcción del significado de límite de una función en un punto.

En un primer momento, la idea de aproximación en el rango se apoya en la comprensión de la idea de aproximación en el dominio indistintamente en los modos de representación numérico y algebraico-numérico. En este sentido, nuestros resultados apoyan los presentados por Engler et al. (2007) cuando indican que los estudiantes tienen más dificultades en encontrar a qué valor se aproxima la función que en determinar a qué valor se aproxima la variable independiente. Por otra parte, la idea de aproximación en el rango se inicia en modo algebraico-numérico cuando las aproximaciones laterales coinciden y se consolida en los modos numérico y algebraico-numérico cuando dichas aproximaciones no coinciden. Igualmente, la idea de aproximaciones no coincidentes a un número en el rango, en los modos numérico y algebraico-numérico, se apoya en la comprensión de la coordinación de las aproximaciones coincidentes en el dominio y en el rango. Es decir, la construcción paulatina de la concepción dinámica del límite se realiza mediante la comprensión de las aproximaciones, diferenciando las que se realizan en el dominio de las que se realizan en el rango, y dentro de estos últimos aquellos en los que las aproximaciones laterales coinciden de las que no coinciden. Sin embargo, un hecho importante puesto de manifiesto por nuestros resultados es que la comprensión de la idea de aproximaciones no coincidentes en el rango implica la comprensión de la coordinación de las aproximaciones coincidentes en el dominio y en el rango, pero no implica necesariamente el comprender la coordinación cuando las aproximaciones laterales no coinciden. Este hecho podría señalar la diferencia cognitiva que para el estudiante resulta tener el establecimiento de las aproximaciones a un número en el rango, en relación a que coincidan o no las aproximaciones laterales (Pons et al., 2011; Valls et al., 2011)

En segundo lugar, y en relación a la comprensión de la concepción dinámica de límite, nuestros resultados indican que la comprensión de la coordinación de las aproximaciones se inicia en modo gráfico cuando las aproximaciones laterales coinciden, progresa en modo numérico y se consolida en modo algebraico-numérico. Estos resultados apoyan los presentados por Monaghan (2001), Blázquez y Ortega (2001) y Engler et al. (2007) en el sentido de que los límites presentados en modo gráfico son resueltos por más estudiantes que los presentados en modo numérico. Sin embargo, un resultado relevante de nuestra investigación es que la coordinación de los procesos de aproximación, cuando las aproximaciones laterales no coinciden en el rango, se inicia indistintamente en modos gráfico o algebraico-numérico y se consolida en modo numérico. Además, del hecho de que la coordinación de los procesos de aproximaciones no coincidentes en modo gráfico se apoye en la coordinación de procesos de aproximaciones coincidentes en modo numérico, nos permite inferir que son los modos gráfico y numérico el nexo de unión entre la comprensión de las coordinaciones de las aproximaciones laterales no coincidentes y las coincidentes. Por otra parte, la comprensión de la coordinación de las aproximaciones coincidentes en el dominio y en el rango no implica necesariamente el ser capaz de coordinar los procesos de aproximaciones laterales no coinciden. La importancia de este hecho radica en que podría señalar la diferencia cognitiva que para el estudiante podría tener el establecimiento de la coordinación de las aproximaciones en el dominio y en el rango, en relación a que coincidan o no las aproximaciones laterales en el rango.

En relación a la comprensión del concepto de límite de una función nuestros resultados estarían en sintonía con los presentados por Blázquez (1999) y Blázquez y Ortega (2001) cuando afirman que la utilización de distintos registros de representación (algebraico, numérico, gráfico, verbal) mejora esta comprensión. Para estos autores un estudiante que utilice indistintamente diferentes modos de representación tendrá una sólida comprensión de la noción de límite de una función.

Las estructuras implicativas identificadas describen ciertas relaciones entre la comprensión que los estudiantes tienen de la coordinación de los procesos de aproximación en los diferentes modos de representación (numérico, gráfico y algebraico-numérico) y la coincidencia o no de las aproximaciones laterales. La información aportada puede ser útil para la organización del contenido de enseñanza y la planificación de secuencias de actividades dirigidas a desarrollar la comprensión en los estudiantes de la concepción dinámica del límite. Una implicación de este hecho es que al diseñar un proceso de enseñanza–aprendizaje de la noción dinámica de límite de una función debemos ser conscientes del papel que tienen los diferentes modos de representación para apoyar la comprensión de la coordinación de los procesos de aproximación, coincidan o no las aproximaciones laterales en el rango.

Referencias

[1] Bergsten, C. (2006). Trying to Reach the Limit. The Role of Algebra in Mathematical Reasoning. *Proceedings of the 30nd Conference of the International Group for the Psychology of Mathematics Education*, **2**(2), 153–160.

[2] Blázquez, S. (1999). *Noción de límite en Matemáticas Aplicadas a las Ciencias Sociales.* Tesis doctoral. Universidad de Valladolid. Valladolid. España.

[3] Blázquez, S. y Ortega, T. (2000). El concepto de límite en la Educación Secundaria. En *El futuro del cálculo infinitesimal.* México: Grupo Editorial Iberoamérica. S.A. de C.V.

[4] Blázquez, S. y Ortega, T. (2001). Los sistemas de representación en la enseñanza del límite. *Revista Latinoamericana de Investigación en Matemática Educativa,* **4**(3), 219- 236.

[5] Blázquez, S. y Ortega, T. (2002). Nueva definición de límite funcional. *UNO,* **30,** 67–83.

[6] Blázquez, S., Ortega, T., Gatica, S. y Benegas, J. (2006). Una conceptualización de límite para el aprendizaje inicial de análisis matemático en la universidad. *Revista Latinoamericana de Investigación en Matemática Educativa,* **9**(2), 189–209.

[7] Bodi, S.D., Valls, J. y Llinares, S. (2009). La comprensión de la divisibilidad en N. Un análisis implicativo. En P. Orús, L. Zamora, P. Gregori (ed), *Teoría y aplicaciones del Análisis Estadístico Implicativo* (215–233). Castellón, España: Innovació Digital Castelló.

[8] Cauchy, A. (1821). *Cours d'Analyse de l'École Royale Polytechnique, (Primer Partie. Analyse Algébrique).* Sevilla: Sociedad Andaluza de Educación Matemática "THALES". Edición facsímil.

[9] Couturier, R. (2009). CHIC: utilización y funcionalidades. En P. Orús, L. Zamora, P. Gregori (ed), *Teoría y aplicaciones del Análisis Estadístico Implicativo* (51–63). Castellón, España: Innovacio Digital Castelló.

[10] Cornu, B. (1991). Limits. En D. Tall (Ed), *Advanced Mathematical Thinking* (153-166). Dordrecht: Kluver.

[11] Cottrill, J., Dubinsky, E., Nichols, D., Schwingendorf, K., Thomas, K. y Vidakovic, D. (1996). Understanding the Limit Concept: Beginning whit a Coordinated Process Scheme. *Journal of Mathematical Behavior,* **15,** 167–192.

[12] Duval, R. (1995). Sémiosis et pensée humaine. Registres sémiotiques et apprentissages intellectuels. Nuchatel: Peter Lang.

[13] Elia, A., Gagatssi, A., Panaoura, A., Zachariades, T. y Zoulinaki, F. (2009). Geometric and algebraic approaches in the concept of "Limit" and the impact of the "Didactic Contract". *International Journal of Science and Mathematics Education*. Published online: 20 de February 2009.

[14] Engler, A., Vrancken, S., Hecklein, D., Müller, D. y Gregorini, M.I. (2007). Análisis de una propuesta didáctica para la enseñanza de límite finito de variable finita. *UNIÓN*, **11,** 113–132.

[15] Fernández-Plaza, J.A., Rico, L. y Ruiz-Hidalgo, J.F. (2013). Variación de las concepciones individuales sobre límite finito de una función en un punto. En A: Berciano, G. Gutiérrez, A. Estepa y N. Climent (Eds.) *Investigación en Educación Matemática XVII* (253–261). Bilbao: SEIEM.

[16] Garbin, S. y Azcarate, C, (2002). Infinito actual e inconsistencias: acerca de las incoherencias en los esquemas conceptuales de alumnos de 16–17 años. *Enseñanza de las Ciencias,* **20**(1), 87–113.

[17] Gras, R., Suzuki, E., Guillet, F. y Spagnolo, F. (Eds.) (2008). *Statistical Implicative analysis.* Theory and Applications. London: Springer.

[18] Gras, R. y Kuntz, P. (2009). El análisis estadístico implicativo (ASI) en respuesta a problemas que le dieron origen. En P. Orús, L. Zamora, P. Gregori (ed), *Teoría y aplicaciones del Análisis Estadístico Implicativo* (3–49). Castellón, España: Innovació Digital Castelló.

[19] Kidron, I. (2010). Constructing knowledge about the notion of limit in the definition of the horizontal asymptote. *International Journal of Science and Mathematics Education*, **9**(6), 695–717

[20] Lacasta, E. y Wilhelmi, M.R. (2010). Deslizamiento metadidáctico en profesores de secundaria. El caso del límite de funciones. En M.M. Moreno, A. Estrada, J. Carrillo y T.A. Sierra (Eds.), *Investigación en Educación Matemática XIV* (379–394). Lleida. SEIEM.

[21] Mamona–Downs, J. (2001). Letting the intuitive bear on the formal: A didactical approach for the understanding of the limit of a sequence. *Educational Studies in Mathematics,* **48**, 259–288.

[22] Mira, M., Valls, J. y Llinares, S. (2013). Un experimento de enseñanza sobre el límite de una función. Factores determinantes en una trayectoria de aprendizaje. *UNIÓN*, **36**, 89-107.

[23] Monaghan, J. (1991). Problems whit the language of limits. *For the learning of Mathematics,* **11**(3), 20–21.

[24] Monaghan, J. (2001). Young peoples' ideas of infinity. *Educational Studies in Mathematics,* **48**, 239–257.

[25] Moru, E. K. (2007). Talking with the literature on epistemological obstacles. *For the learning of Mathematics,* **27**(3), 34–37.

[26] Moru, E.K. (2009). Epistemological obstacles in coming to understand the limit of a function at undergraduate level: A case from the National University of Lesotho. *International Journal of Science and Mathematics Education,* **7**, 431–454.

[27] Oehrtman, M. (2009). Collapsing Dimensions, Physical Limitation, and Other Students Metaphors for Limit Concepts. *Journal for Research in Mathematics Education*, **40** (4), 396–426.

[28] Parameswaran, R. (2007). On understanding the notion of limits and infinitesimal quantities. *International Journal of Science and Mathematics Education*, **5**, 193–216.

[29] Pons, J., Valls, J., y Llinares, S. (2011). Coordination of Approximations in Secondary School Students' Understanding of Limit Concept. *Proceedings of the 35nd Conference of the International Group for the Psychology of Mathematics Education*, **4**(3), 393-400.

[30] Roh, K. H. (2008). Students' Images and their Understanding of Definitions of the Limit of a Sequence. *Educational Studies in Mathematics*, **69**, 217–233.

[31] Sierpinska, A. (1985). Obstacles épistémologiques relatives à la notion de limite. *Recherches en Didactique des Mathématiques*, **6** (1), 5–67.

[32] Szydlik, J.E. (2000). Mathematical beliefs and conceptual understanding of the limit of a function. *Journal for Research in Mathematics Education*, **31**(3), 258–276.

[33] Swinyard,C. (2011). Reinventing the formal definition of limit: The case of Amy and Mike. *Journal of Mathematical Behavior* (2011), doi:10.1016/j.jmathb.2011.01.001

[34] Tall, D. y Vinner, S. (1981). Concept image and concept definition in mathematics with particular reference to limit and continuity. *Educational Studies in Mathematics*, **12**, 151–169.

[35] Trigueros, M. y Escandon, C. (2008). Los conceptos relevantes en el aprendizaje de la graficación. *Revista Mexicana de Investigación Educativa*, **13**(36), 59–85.

[36] Valls, J., Pons, J., y Llinares, S. (2011). Coordinación de los procesos de aproximación en la comprensión del límite de una función. *Enseñanza de las Ciencias*, **29**(3), 325-338.

[37] Williams, S.R. (1991). Models of limit held by college calculus students. *Journal for Research in Mathematics Education*, 22(3), 219–236.

[38] Zamora, L., Gregori, P. y Orús, P. (2009). Conceptos fundamentales del Análisis Estadístico Implicativo (ASI) y su soporte computacional CHIC. En P. Orús, L. Zamora, P. Gregori (ed), *Teoría y aplicaciones del Análisis Estadístico Implicativo* (65–101). Castellón, España: Innovació Digital Castelló.

LES SUPPORTS D'ENSEIGNEMENT DANS LA REPRESENTATION DU METIER CHEZ DES PROFESSEURS DES ECOLES DEBUTANTS

Marc BAILLEUL[1] et Laurence LEROYER[2]

SUPPORTS OF TEACHING IN THE REPRESENTATION OF TEACHING AMONG BEGINNING TEACHERS

RÉSUMÉ

Les supports d'enseignement sont au cœur de l'activité des enseignants, tant dans la phase de conception des séquences didactiques que dans celle de leurs pratiques de classe. Nous allons étudier dans ce texte la place que donnent des professeurs des écoles débutants à ces supports dans leurs représentations du métier. Nous allons pour cela mobiliser des résultats extraits de deux recherches sur la professionnalisation des enseignants, recherches menées respectivement en 2007 et 2012. En comparant ces résultats, nous mettrons en évidence des effets du changement de logique de formation voulu par l'institution entre ces deux dates.

Mots-clés : *supports d'enseignement, professionnalisation des enseignants, analyse statistique implicative.*

ABSTRACT

The teaching materials are at the heart of the activity of teachers, both in the design phase of didactic sequences than in their classroom practices. We will study in this paper the place that give novice teachers to these materials in the representations of their occupation. For this, we mobilize extracted results of two studies on the professionalization of teachers, research conducted respectively in 2007 and 2012. By comparing these results, we will highlight the effects of a change of training logic, desired by the institution, between these two dates.

Keywords : *supports of teaching, teachers professionalization, statistical implicative analysis.*

1 Une recherche sur la professionnalisation des enseignants

1.1 L'objet de recherche : la construction de la professionnalité chez les enseignants débutants

Notre intention consiste à mieux comprendre la façon dont les enseignants élaborent leur professionnalité en cours de formation (en IUFM[3], devenu ESPE4 en septembre 2011) en prenant en compte les différents contextes d'exercice (écoles, collèges, lycées) et les disciplines enseignées. Ce travail est réalisé par une équipe de chercheurs représentant plusieurs disciplines (sociologie, sciences de l'éducation, analyse du travail,…) et quatre institutions universitaires normandes, deux d'entre elles à vocation

1 CERSE, Université de Caen, marc.bailleul@unicaen.fr
2 ESPE, Université de Caen, 186 rue de la Délivrande, laurence.leroyer01@unicaen.fr
3 Institut Universitaire de Formation des Maîtres
4 École Supérieure du Professorat et de l'Éducation

formatrice (les ESPE de Basse et de Haute-Normandie), les deux autres centrées sur la recherche en sciences de l'éducation (les laboratoires CIVIIC[5] de l'Université de Rouen et CERSE[6] de Université de Caen Basse-Normandie).

Cette recherche a pour originalité de se démarquer sensiblement des recherches conduites jusqu'à ce jour sur la thématique de la professionnalisation des enseignants sous un double aspect :

- d'une part, nous n'abordons pas seulement la professionnalisation sous l'angle de la formation formelle (en IUFM puis ESPE) mais sous l'angle des transitions formation-travail et sous l'angle de la perception des apprentissages en cours d'activité professionnelle. L'enjeu consiste ici à mieux comprendre comment les enseignants parlent de et vivent leurs apprentissages professionnels par et dans la conduite des activités professionnelles ;
- d'autre part, nous croisons des grilles conceptuelles différentes pour comprendre de façon globale et non compartimentée, une réalité complexe (les conditions du développement professionnel).

1.2 Le cadre conceptuel de ces recherches

Nos travaux s'inscrivent dans le champ des réflexions sur la professionnalisation des enseignants (Bourdoncle, 1993). Nous entrons dans cette thématique en mettant en tension l'offre de professionnalisation (côté institution) et la dynamique de la construction initiale de la professionnalité (côté sujets), première étape de ce que nous appellerons le développement professionnel.

Tardif et Lessard (2000, p. 401), parlant des enseignants, insistent sur l'idée que « les connaissances issues de la pratique (connaissances ouvragées) semblent les fondements de la pratique du métier et de la compétence professionnelle […] Les routines (des enseignants) sont des moyens de gérer la complexité des situations d'interactions et de diminuer l'investissement cognitif de l'enseignant dans le contrôle des événements […] La force et la stabilité de ce contrôle ne peuvent pas dépendre de décisions volontaires de choix, de projets, mais bien de l'intériorisation des règles implicites d'action acquises avec et dans l'expérience de l'action… » Pour ces auteurs, l'activité enseignante présente ainsi un caractère syncrétique, « le syncrétisme signifie le fait que l'enseignement exige du travailleur une capacité à utiliser dans l'action quotidienne un large spectre de connaissances et de savoir-faire composites » (opus cité, p. 369). Au final, pour Tardif et Lessard, ce qui caractérise l'enseignant, c'est la production de savoirs d'expérience qui ont pour propriété d'être « ouvragés », liés aux tâches, façonnés, pratiques, « interactifs », « syncrétiques » et « pluriels », hétérogènes, complexes et « non analytiques », ouverts, « poreux », « perméables », « existentiels et faiblement formalisés », « temporels » et « évolutifs ».

Avec l'objectif de mieux comprendre la construction expérientielle des sujets dans l'activité (de travail et/ou de formation), nous proposons de différencier ce qui relève d'une logique de professionnalisation de ce qui relève d'une logique de développement professionnel.

---

[5] Centre de recherches Interdisciplinaires sur les Valeurs, les Idées, les Identités et les Compétences

[6] Centre d'Études et de Recherche en Sciences de l'Éducation

Cette notion de développement professionnel n'est pas, à proprement parler, construite. Elle surgit depuis quelques décennies dans des travaux qui s'inspirent des théories du développement (Piaget, Léontiev, Vygotski) pour étudier, comme le fait par exemple la didactique professionnelle, les apprentissages réalisés par des adultes en situation professionnelle. Les sensibilités sont alors différentes selon la référence théorique choisie.

Pour être plus précis, dans les travaux classiques, trois options semblent exister pour expliquer le développement : il se réalise par l'**activité** (thèse de Léontiev, 1984), par le **langage et l'interaction** (Vygotski, 1985) ou par la **pensée** (Piaget, 1970). On retrouve ces débats dans les sensibilités aujourd'hui présentes : la clinique de l'activité (Clot, Prot, Werthe, 2001) emprunte aux thèses de Léontiev et Vygotskiquand la didactique professionnelle (Pastré, Vergnaud, Mayen, 2006) est influencée par la thèse constructiviste de Piaget,…

Pour notre part, nous proposons la distinction suivante entre professionnalisation et développement professionnel (distinction largement développée dans Wittorski, 2007) :

- la professionnalisation aurait à voir avec une intention sociale de professionnalisation émanant d'une institution ou d'une organisation ; soit l'injonction faite aux individus en formation, d'engager un processus de construction-transformation de leurs activités au service d'une efficacité et d'une lisibilité plus grandes du travail. Cette intention se traduit par la proposition de dispositifs de travail et de formation particuliers ;
- le développement professionnel concernerait un mouvement conjoint de développement d'une activité, de stratégies et de dynamiques identitaires accompagnant la construction-transformation d'une professionnalité au niveau d'un individu agissant.

Notre groupe de recherche considère que le processus de développement professionnel des enseignants s'effectue dans les espaces de la formation (en IUFM), de la transition formation-travail (IUFM et établissement scolaire) et du travail investi (en établissement scolaire). Il correspond à trois dynamiques :

- une dynamique de transformation des connaissances et des compétences ;
- une dynamique de transaction individuelle entre une offre institutionnelle de professionnalisation, un mode de contrôle des professionnalités et des modalités effectives de construction de celles-ci comme enseignant stagiaire puis titulaire ;
- une dynamique de transaction individuelle entre l'identité héritée et l'identité vécue ou en construction/transformation.

1.3 Problématique générale de la recherche

Nous considérons ainsi que les enseignants en formation, en livrant, via un questionnaire, leur vécu et ressenti à propos de leur parcours de formation et de travail, révèlent dans le même temps la proximité ou la distance qu'ils ont vis à vis de l'offre-injonction institutionnelle de professionnalisation (discours de la formation et du terrain). Il y aurait ainsi une tension entre le projet de l'institution et les vécus des enseignants en formation et au travail. Le projet de l'institution est lisible à la fois dans les textes qui régissent la formation (de ce point de vue la prescription est forte), mais

aussi dans les pratiques de formation proposées et les modalités d'évaluation professionnelle (visites de formateurs et/ou d'inspecteurs notamment). Pour leur part, les enseignants vivent des expériences personnelles contrastées dans leur établissement d'exercice, ils sont confrontés à la fois à des discours venant de leurs pairs qui n'entrent pas toujours en résonance avec le projet de l'institution et à des réactions des élèves et des parents qui leur renvoient une image d'eux-mêmes et de leur métier qui les interroge. Dès lors, une tension est ressentie par les enseignants en formation qui est à l'origine du développement de stratégies identitaires qui nous semblent être au cœur du processus de développement professionnel des nouveaux enseignants. Nous faisons l'hypothèse que les réponses au questionnaire constituent des traces de ces stratégies.

En comparant les résultats issus de deux campagnes de collecte de données réalisées respectivement en 2007 et 2012, auprès de deux promotions ayant suivi des formations conçues sur la base de logiques de formation différentes, nous faisons l'hypothèse que les représentations globales du métier ne seront pas structurées de la même façon relativement à ces deux corpus de données.

De plus, nous chercherons à valider cette hypothèse sur une question particulière ; la place des supports d'enseignement dans ces représentations. Au cœur de l'activité enseignante, tant dans la phase de conception des séquences didactiques que dans celle de leurs pratiques de classe (Leroyer, 2010), il nous semble possible d'identifier ici des différences de conception du métier liées aux modalités de formation.

1.4 Méthodologie mise en œuvre

1.4.1 Le questionnaire

Le cahier des charges de la formation des enseignants (2007) définit les compétences attendues des enseignants en général, que ceux-ci exercent dans le premier degré ou le second degré. Ces compétences sont regroupées en dix grands domaines. Les 28 items[7] qui structurent le questionnaire renvoient à ce référentiel, avec la répartition suivante :

Concevoir son enseignement : 5 items (8-10-11[8]-13-14)
Travail en partenariat : 5 items (1-7-24-25-26)
Gérer la classe : 4 items (15-16-17-28)
Prendre en compte la diversité des élèves : 4 items (2-3-12-18)
Évaluer : 3 items (19-20-21)
TIC : 2 items (22-23)
Agir en fonctionnaire de l'État de manière éthique et responsable : 2 items (4-5)
Recherche et innovation : 1 item (27)
Maîtrise de la discipline : 1 item (6)
Maîtrise de la langue française : 1 item (9)

Pour chacun des items, trois jugements successifs sur une échelle en 6 points sont demandés (1 = non, pas du tout, 6 = oui, tout à fait). La compétence évoquée dans l'item est-elle importante ? Appartient-elle au domaine du « faisable » ? Est-elle en construction dans la pratique ? Chaque stagiaire (statut des répondants en 2007) ou

[7] Voir annexes 1 et 2.

[8] C'est l'item 11 qui est centré sur la place des supports d'enseignement ; « Être enseignant, pour vous, c'est… avoir un regard critique sur les supports d'enseignement utilisés en classe (manuels, logiciels, autres) »

étudiant (statut des répondants en 2010) était donc amené à produire une liste de 84 jugements (28 x 3). En outre, il est demandé avec qui ces compétences se construisent. Deux choix sont permis parmi six possibles : les autres professeurs débutants, les formateurs, les tuteurs terrain ou conseillers pédagogiques, les collègues de l'établissement, la famille ou les amis, seul dans la classe. La dernière page du questionnaire est consacrée aux variables qu'on peut désigner comme « supplémentaires », caractérisant les répondants : genre, ancienneté, parcours universitaire, disciplines enseignée pour les PLC[9], etc. L'objectif n'est en aucun cas de demander aux enseignants stagiaires/étudiants d'évaluer la formation qu'ils ont suivie ni même de comprendre quels apprentissages particuliers ils ont développé, mais plutôt d'analyser leur perception de ce qui permet de devenir un professionnel de l'enseignement d'un double point de vue :

- perception des compétences requises, « atteignables » et plus ou moins acquises à l'issue de l'année,
- modalités d'apprentissage perçues comme étant les plus efficaces pour acquérir ces compétences.

Les jugements sur l'importance des items dessinent une image du métier idéalisé. Leur analyse permet de mesurer l'impact de l'injonction de professionnalisation. En demandant des jugements sur la « faisabilité » ; nous accédons aux représentations du métier possible, en relation avec les expériences personnelles dans les établissements d'exercice. Les jugements sur des constructions de compétences nous renseignent sur la perception d'apprentissages réalisés au cours de l'année. Les écarts entre ces trois séries de jugements doivent permettre de repérer où portent, où agissent les tensions ressenties par les enseignants en formation, confrontés à des discours partiellement dissonants.

1.4.2 Le recueil des données

Un double recueil de données a été mené. Les effectifs se répartissent ainsi :

- Première phase de recueil des données (mai 2007) : PE[10] : 496 ; PLC : 350 ; PLP[11] : 54, soit un total de 900 stagiaires ayant répondu au questionnaire.
- Seconde phase de recueil des données (mai 2012) : PE : 150 ; PLC/PLP : 75.

Nous ne nous intéresserons dans ce texte qu'aux résultats concernant les PE, les effectifs PLC/PLP étant trop faibles en 2012.

2 Un résultat majeur

Nous nous centrerons dans ce texte sur le seul point de vue « Important ». L'analyse statistique implicative nous permet-elle de révéler des logiques d'acteurs différentes entre 2007 et 2012 ?

[9] Professeur des Lycées et Collèges
[10] Professeur des Écoles
[11] Professeur de Lycée Professionnel

On trouvera les graphes implicatifs issus des analyses menées en 2007 et 2012 respectivement dans les annexes 3 et 4[12]. Le graphe de 2007 est très clairement structuré en deux réseaux disjoints. Celui qui figure sur la gauche de la figure (respectivement sur la droite) regroupe des items choisis en position 5 ou 6 (respectivement des items choisis en position 1 ou 2). nous interprétons ce graphe comme révélant deux rapport différents au projet institutionnel de professionnalisation, présent à travers les items du questionnaire : l'adhésion à ce projet (choix 5 ou 6) d'un côté, le rejet de ce projet de l'autre choix 1 ou 2). Il nous faut apporter ici une importante précision relative à la lecture du graphe : la lisibilité a été privilégié par le logiciel CHIC (Couturier, Almouloud, 2013 ; Ratsimba-Rajohn, 2013) lors de la réalisation du graphe, il ne faut donc pas tenir compte de la position des items, ni de la « hauteur » des réseaux, sur l'axe vertical pour leur accorder un poids correspondant à leur position sur cet axe. Pour information les items du « réseau du rejet » (à droite de la figure) ont des poids (pourcentage de choix) compris entre 4 et 6 % alors que ceux du « réseau de l'adhésion » ont des poids compris entre 30 et 93 % (variable attractrice 28 : être capable d'établir une relation de confiance avec les élèves). Il est intéressant de noter qu'à l'intérieur même du réseau de l'adhésion, il est possible de mettre en évidence deux sous-réseaux (voir annexe 5) dont l'un est structuré par les items du domaine didactique[13] (R2), l'autre réseau (R1) regroupant des items relevant d'autres problématiques plus transversales (citoyenneté, autorité, différenciation, dimension collective du travail).

Le réseau R2 est articulé autour des items suivants, qu'on peut qualifier de variables « nodales »[14] :

- item 20 (ajuster sa progression en fonction des résultats des évaluations) ;
- item 9 (participer à l'apprentissage de la langue française chez ses élèves à travers son enseignement) ;
- item 16 (susciter l'implication des élèves dans le travail) ;
- item 8 (aider les élèves à acquérir des méthodes de travail) ;
- item 15 (créer les conditions favorables à la mise au travail des élèves).

Les variables 11 (avoir un regard critique sur les supports d'enseignement utilisés en classe (manuels, logiciels, autres) et 21 (consacrer un temps, avant chaque évaluation, à l'explicitation des attentes et des critères) sont aussi, à un degré moindre (parce qu'alimentées chacune par des implications issues de l'autre sous-réseau) des variables nodales de ce sous-réseau R2.

Le graphe de 2012, à la différence du précédent au même seuil, ne fait pas apparaître de réseaux disjoints mais des chemins implicatifs entremêlés, aboutissant sur dix (!) variables attractrices et dont les origines sont, pour certains, des positions de rejet des items choisis ! Nous interprétons ce résultat comme la trace d'une représentation du métier très éclatée, non encore structurée suite à une année de formation professionnelle

[12] Dans ces graphes, il faut lire les noms des variables de la façon suivante : vi12.27 : item 27 choisi en position 1 ou 2, vi34.27 : item 27 en position 3 ou 4, vi56.27 : item 27 en position 5 ou 6.

[13] Nous qualifions de « domaine didactique » le regroupement des items suivants : 6, 8, 9, 10, 11, 12, 14, 15, 16, 19, 20 et 21.

[14] Voir Lahanier-Reuter, Gras, Bailleul dans cet ouvrage.

qui n'a pas permis aux étudiants interrogés de percevoir la complexité du métier autrement que comme une juxtaposition de préoccupations.

3 La variable 11, variable nodale des deux graphes

3.1 La variable 11 dans le graphe de 2007

Nous avons explicité dans la figure ci-dessous les items constituant le « cône 11 » pour l'année 2007. La participation à des collectifs de réflexion contribue fortement à l'analyse critique des supports d'enseignement qui constitue une ressource pour la prise en compte des items suivants :

- ajuster sa progression en fonction des résultats des évaluations (à travers la constitution d'une « banque de scénarios » potentiellement mobilisables à différents degrés de maîtrise des connaissances visées),
- créer les conditions favorables à la mise au travail des élèves (en proposant aux élèves des scénarios motivants et adaptés à leurs profils),
- participer à l'apprentissage de la langue française chez ses élèves à travers son enseignement (en s'appropriant des scénarios construits pas des « spécialistes » : auteurs de manuels, créateurs de ressources en ligne, etc.),
- être capable d'établir une relation de confiance avec les élèves (en assurant ainsi, d'une certaine façon, la « qualité » de sa prestation didactique, ce qui permet alors de « faire.autorité »)

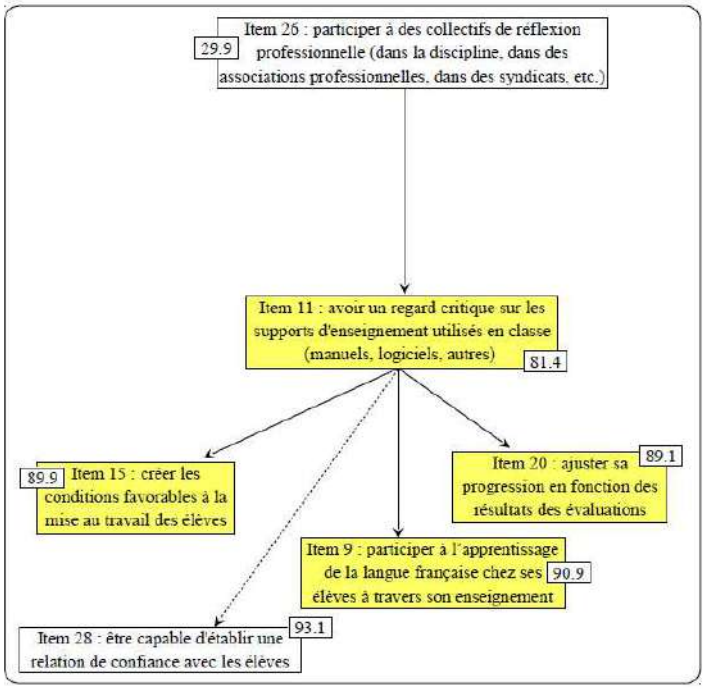


Figure 1 : La variable nodale 11 (avoir un regard critique sur les supports d'enseignement utilisés en classe (manuels, logiciels, autres) dans le graphe de 2007 NB : le grisé sur une case marque son appartenance au domaine didactique

Les pointillés entre les variables 11 et 28, dans le mode cône du logiciel CHIC, indiquent que cette implication « passe » par des chemins transitifs (11 → 9 → 28 ; 11

→ 15 → 28 ; 11 → 20 → 28).Notons le rôle essentiellement didactique de la variable 11.

3.2 La variable 11 dans le graphe de 2012

Nous avons explicité dans la figure de la page suivante les items constituant le « cône 11 » pour l'année 2012. Par simple perception visuelle, une différence apparaît très nettement entre les deux figures : l'item 11 a perdu, entre 2007 et 2012, de son importance didactique puisque seul un autre item de ce domaine (l'item 8 : aider les élèves à acquérir des méthodes de travail) figure encore dans le cône en 2012.

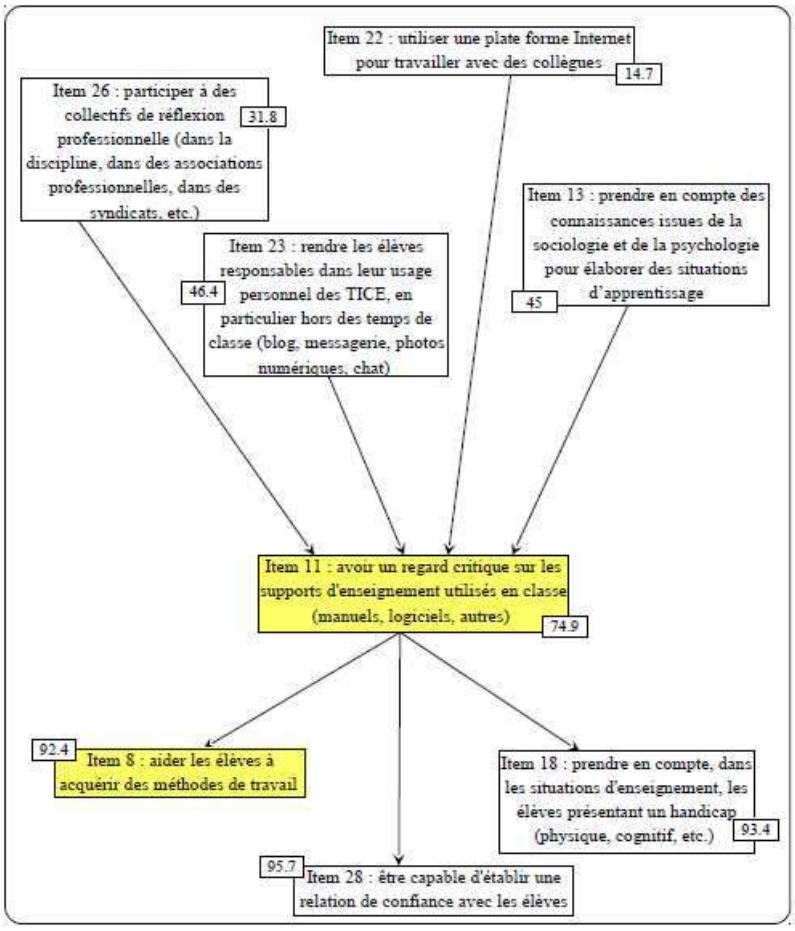


Figure 2 : La variable nodale 11 (avoir un regard critique sur les supports d'enseignement utilisés en classe (manuels, logiciels, autres) dans le graphe de 2012 NB : le grisé sur une case marque son appartenance au domaine didactique

3.3 Regard complémentaire sur les différences

Pour aller au-delà de cette différence perceptive, nous avons mobilisé le concept de variance implicative (Gras & Régnier, 2013). Nous avons calculé les variances implicatives intra-grappe de tous les cônes des deux graphes (11 pour le graphe de 2007 et 9 pour celui de 2012). Rappelons que la variance implicative n'étant pas un indice compris entre 0 et 1, nous ne pouvons que lui attribuer une valeur relative. Néanmoins nous pouvons comparer les places obtenues, pour chacune des deux années, par les cônes ayant pour variable nodale la variable 11, relativement à l'ensemble des autres

cônes. En 2007, le « cône 11 » est classé en position 6, sur 11 alors qu'il est en position 8 sur 9 en 2012. Non seulement les cônes étudiés ne concernent pas les mêmes items mais, de plus, le flux implicatif qui traverse chacun des deux cônes est donc plus significatif en 2007 qu'en 2012.

3.4 Discussion des résultats

Construire sa propre représentation de l'activité enseignante n'est pas chose simple pour un professeur en formation. Nous considérons l'activité comme la résultante d'un système de ressources et de contraintes composé de trois sous-systèmes en interaction que sont : l'enseignant exerçant, le contexte dans lequel il exerce et le genre professionnel (Leroyer, 2013). L'enseignant exerçant est considéré comme un sujet caractérisé par des connaissances, des valeurs et croyances, des émotions, des capacités physiques, etc. Le contexte est composé d'un environnement organisationnel, relationnel et matériel (Chatigny et Vézina, 2008). Enfin, le genre professionnel est constitué des formes communes de la vie professionnelle (Clot, Faïta, 2000).

L'existence en 2007, la non existence en 2012, d'une organisation des réponses au questionnaire à travers un graphe implicatif fortement structuré en 2007, et passablement éclaté en 2012, nous donnent à voir les effets de deux logiques de formation complètement différentes. Mai 2012 voit sortir des IUFM la première promotion qui a été formée selon les modalités de formation mises en œuvre suite à la réforme dite « de la mastérisation ». Une des principales différences concernait la confrontation au terrain à l'occasion des stages. Elle n'avait pas du tout la même importance : plus développée en 2007 qu'en 2012 et aussi beaucoup plus soucieuse de la mise en synergie théorie-pratique. Nous pensons que c'est cette synergie des propos entre les formateurs universitaires et les formateurs de terrain, à travers des co-interventions par exemple, qui permet aux professeurs débutants de mettre en cohérence les différents éléments d'un discours qui apparaît alors comme unique et représentant les attentes de l'institution. L'éviction, sur la période 2010-2012, des formateurs de terrain des interventions à l'IUFM et leur « cantonnement » aux classes, révélateurs d'une juxtaposition des savoirs et savoir-faire et non de leur inter dépendance, peut être avancée comme explication à l'éclatement de la représentation du métier. Les conditions de découverte du système de ressources et de contraintes évoqué plus haut sont, en 2012 et à la différence de 2007, défavorables puisque favorisant la construction d'une image « biface » du métier, assez souvent relayées par le discours des médias grand public.

L'étude, dans ce contexte, du rôle particulier de la variable 11 (avoir un regard critique sur les supports d'enseignement utilisés en classe (manuels, logiciels, autres) nous a semblé être prioritaire dans la mesure où, le temps consacré à la dimension *stricto sensu* professionnelle de la formation diminuant, il est urgent d'outiller les professeurs débutants de savoirs et de savoir faire centrés sur l'analyse critique des supports d'enseignement. Les résultats confirment, pour ce micro point de vue, l'analyse générale ci-dessus. La problématique du choix et de l'utilisation des supports d'enseignement n'est plus intégrée à une réflexion didactique elle-même partie prenante de l'activité enseignante. Elle n'apparaît, en 2012, que comme un des éléments qu'il est possible de mobiliser pour atteindre certains des multiples, ou du moins perçus comme tels, objectifs assignés aux professeurs en classe. La comparaison des variances

implicatives des cônes extraits des deux graphes confirme, pour cet item, la détérioration de la structuration des représentations du métier entre 2007 et 2012.

4 Conclusion

Nous avons, dans ce texte, comparé les résultats de deux recherches menées à cinq ans d'intervalle, avec un même support d'enquête et un même outil d'analyse statistique, sur deux promotions de professeurs des écoles en fin de formation, promotions qui ont suivi des parcours différents suite à la mise en place de la réforme dite « de la masterisation » de la formation des enseignants.

Le premier niveau de comparaison a été global : les deux graphes implicatifs issus des analyses menées avec le logiciel CHIC sont très différents, l'un fortement structuré, l'autre non.

Le deuxième niveau que nous avons étudié était un micro point de vue : le rôle d'une variable spécifique parce que perçue par nous comme particulièrement importante dans le contexte actuel de l'évolution de la formation des enseignants, l'item de l'enquête ciblé sur le trait de compétence « avoir un regard critique sur les supports d'enseignement utilisés en classe (manuels, logiciels, autres) ». Là encore, un effet différenciateur a été mis en évidence rendant compte d'une détérioration de la perception de cet élément comme central dans une réflexion didactique au cœur de la pratique.

D'autres membres de l'équipe ont travaillé à une autre dimension mise en avant par l'institution comme une compétence clé des futurs enseignants : le travail en équipe (Tavignot, Thémines, Buhot, 2014).

L'organisation de la formation des enseignants a, une nouvelle fois, été remaniée avec la création des Écoles Supérieures du Professorat et de l'Éducation (ESPE) en 2013. C'est pourquoi nous allons reconduire cette enquête en mai 2015, faisant l'hypothèse de nouvelles transformations des représentations induites par les modalités de formation auxquelles sont confrontés les professeurs débutants. L'histoire se poursuivra donc dans un prochain colloque A.S.I. 9

Références

[1] Bourdoncle, R., 1993. La professionnalisation des enseignants : les limites d'un mythe. *Revue Française de Pédagogie*, 105, 83-114.

[2] Chatigny, C., & Vezina, N. 2008. L'analyse ergonomique de l'activité de travail : un outil pour développer les dispositifs de formation et d'enseignement. In Y. Lenoir & P. Pastré (dir.), *Didactique professionnelle et didactiques disciplinaires en débats, un enjeu pour la professionnalisation des enseignants,* Toulouse : Octarès, 269-284.

[3] Clot, Y., & Faïta, D. (2000). Genres et styles en analyse du travail. Concepts et méthodes,Vol. 4, Travailler.

[4] Clot Y., Prot B., Werthe C., 2001. Clinique de l'activité et pouvoir d'agir, *Éducation Permanente,* 46, 12-37.

[5] Couturier, R., Ag Almouloud, S., 2013. Historique et fonctionnalités de CH.I.C. in Gras, R. (dir.) *Analyse Statistique Implicative, Méthode exploratoire et confirmatoire pour la recherche de causalités*, 2ème édition revue et augmentée, RNTI E-16, Toulouse : Cépaduès éditions, 313-326.

[6] Gras, R. & Régnier, J.-C., 2013. Qualité d'un graphe implicatif : variance implicative, in Gras, R. (dir.) *Analyse Statistique Implicative, Méthode exploratoire et confirmatoire pour la recherche de causalités*, 2ème édition revue et augmentée, RNTI E-16, Toulouse : Cépaduès éditions, 207-218.

[7] Leontiev, A. 1984. *Activité, conscience, personnalité*. Moscou : Editions du progrès.

[8] Leroyer, L., 2013. Le rapport au support dans le travail de préparation en mathématiques des enseignants du premier degré, *Education & didactiques*, vol. 7, n° 1, 147-164.

[9] Ministère de l'Education Nationale, Cahier des charges de la formation des maîtres en institut universitaire de formation des maîtres, A. du 19-12-2006 JO du 28-12-2006, BOEN n°1, 4 janvier 2007.

[10] **Pastré** P., Mayen P., **Vergnaud** G. 2006. La **didactique professionnelle**. *Revue française de pédagogie*, 154, 145-198.

[11] Piaget, J., 1970. *L'épistémologie génétique*, Paris : PUF.

[12] Ratsimba-Rajohn, H., 2013. Guide d'utilisation des principales fonctionnalités du logiciel CHIC, Gras, R. (dir.) *Analyse Statistique Implicative, Méthode exploratoire et confirmatoire pour la recherche de causalités*, 2ème édition revue et augmentée, RNTI E-16, Toulouse : Cépaduès éditions, 327-348.

[13] Tardif, M. et Lessard, C., 2000. *Le travail enseignant au quotidien*. Louvain : De Boeck.

[14] Tavignot, P., Thémines, J.-F., Buhot, E. 2014. Intentions de formation et dimension collective : enseignants débutants, in *Actes du colloque Convisciencia*, Toulouse : ENFA.

[15] Vygotski L.S., 1985. *Pensée et Langage*, Terrains Editions Sociales.

[16] Wittorski, R. 2007. *Professionnalisation et développement professionnel*. Paris : L'Harmattan.

Annexe 1 : Le questionnaire, extrait de la première page.

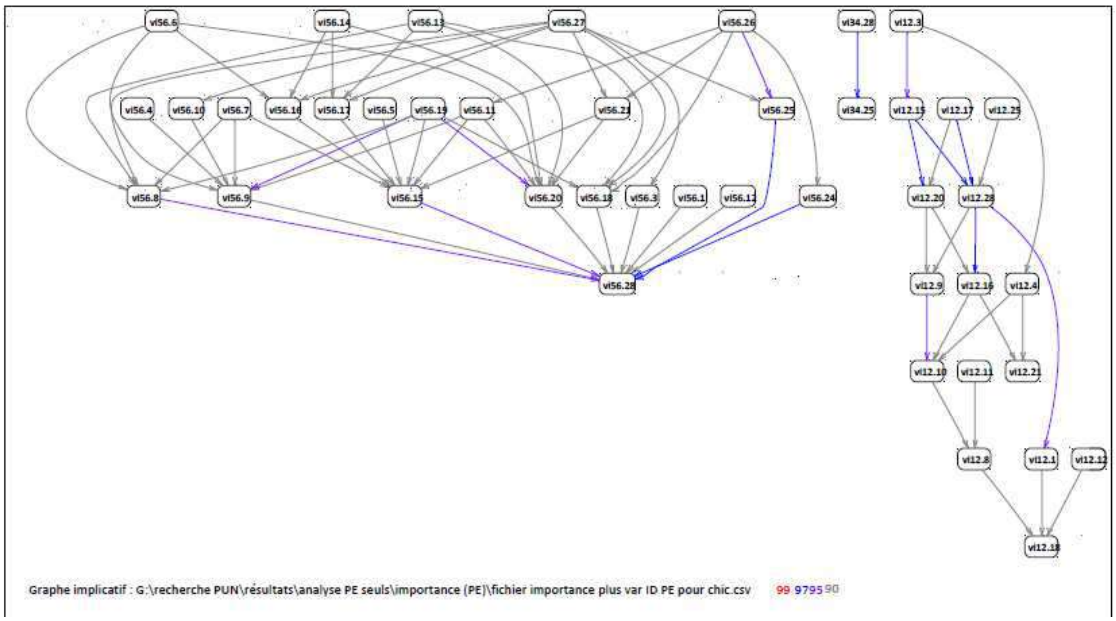
Être enseignant, pour vous, c'est…		Pour vous, la **construction** de cette compétence est…			Selon vous, cette compétence **se construit avec qui** ? (Cochez deux cases maximum par item)							
		importante	faisable	en cours dans votre pratique		avec vos collègues stagiaires	avec vos conseillers pédagogiques et tuteurs	avec vos formateurs	avec vos collègues dans l'établissement	seul dans l'établissement	avec des proches (parents, amis, …)	
Item 1	… pouvoir discuter avec des collègues des difficultés rencontrées dans la classe	-- ++	-- ++	-- ++								
Item 2	… consacrer du temps à s'entretenir avec les élèves en dehors de la classe	-- ++	-- ++	-- ++								
Item 3	… aider l'élève à se construire en tant que personne	-- ++	-- ++	-- ++								

Annexe 2 : Les 28 items du questionnaire

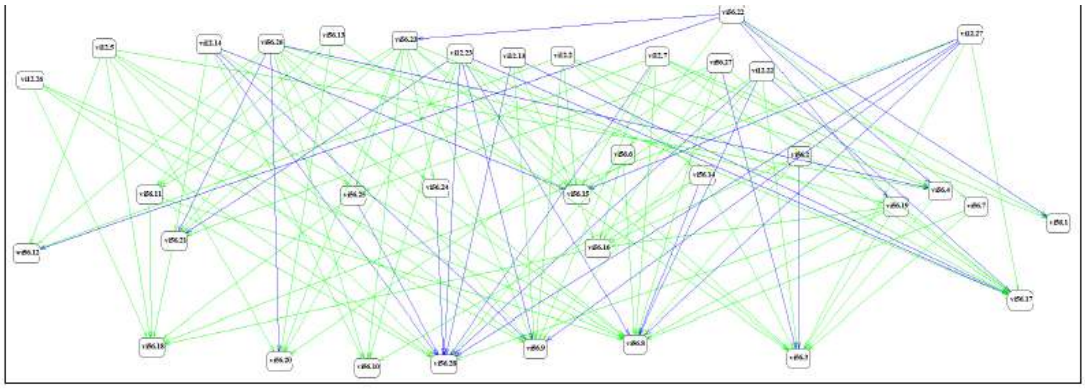
Être enseignant, pour vous, c'est…

Item 1	… pouvoir discuter avec des collègues des difficultés rencontrées dans la classe
Item 2	… consacrer du temps à s'entretenir avec les élèves en dehors de la classe
Item 3	… aider l'élève à se construire en tant que personne
Item 4	… former les élèves à une citoyenneté active
Item 5	… assumer une fonction d'autorité auprès des élèves
Item 6	… s'informer de l'actualité scientifique et / ou didactique dans sa ou ses disciplines d'enseignement
Item 7	… aider les élèves à faire leurs choix d'orientation
Item 8	… aider les élèves à acquérir des méthodes de travail
Item 9	… participer à l'apprentissage de la langue française chez ses élèves à travers son enseignement
Item 10	… concevoir une progression dans les notions à acquérir
Item 11	… avoir un regard critique sur les supports d'enseignement utilisés en classe (manuels, logiciels, autres)
Item 12	… aménager le programme pour s'adapter aux élèves
Item 13	… prendre en compte des connaissances issues de la sociologie et de la psychologie pour élaborer des situations d'apprentissage
Item 14	… développer des approches pluridisciplinaires et transversales (IDD, TPE, Education à..., PPCP, ECJS, etc.)
Item 15	… créer les conditions favorables à la mise au travail des élèves
Item 16	… susciter l'implication des élèves dans le travail
Item 17	… instaurer un climat de classe propice à la coopération entre élèves
Item 18	… prendre en compte, dans les situations d'enseignement, les élèves présentant un handicap (physique, cognitif, etc.)
Item 19	… diversifier les formes d'appréciation du travail des élèves
Item 20	… ajuster sa progression en fonction des résultats des évaluations
Item 21	… consacrer un temps, avant chaque évaluation, à l'explicitation des attentes et des critères
Item 22	… utiliser une plate forme Internet pour travailler avec des collègues
Item 23	… rendre les élèves responsables dans leur usage personnel des TICE, en particulier hors des temps de classe (blog, messagerie, photos numériques, chat)
Item 24	… être à l'écoute des parents qui rencontrent des difficultés dans l'éducation de leurs enfants
Item 25	… participer à des temps réguliers d'échanges entre professeurs dans l'établissement
Item 26	… participer à des collectifs de réflexion professionnelle (dans la discipline, dans des associations professionnelles, dans des syndicats, etc.)
Item 27	… se tenir informé des résultats de la recherche en éducation
Item 28	… être capable d'établir une relation de confiance avec les élèves

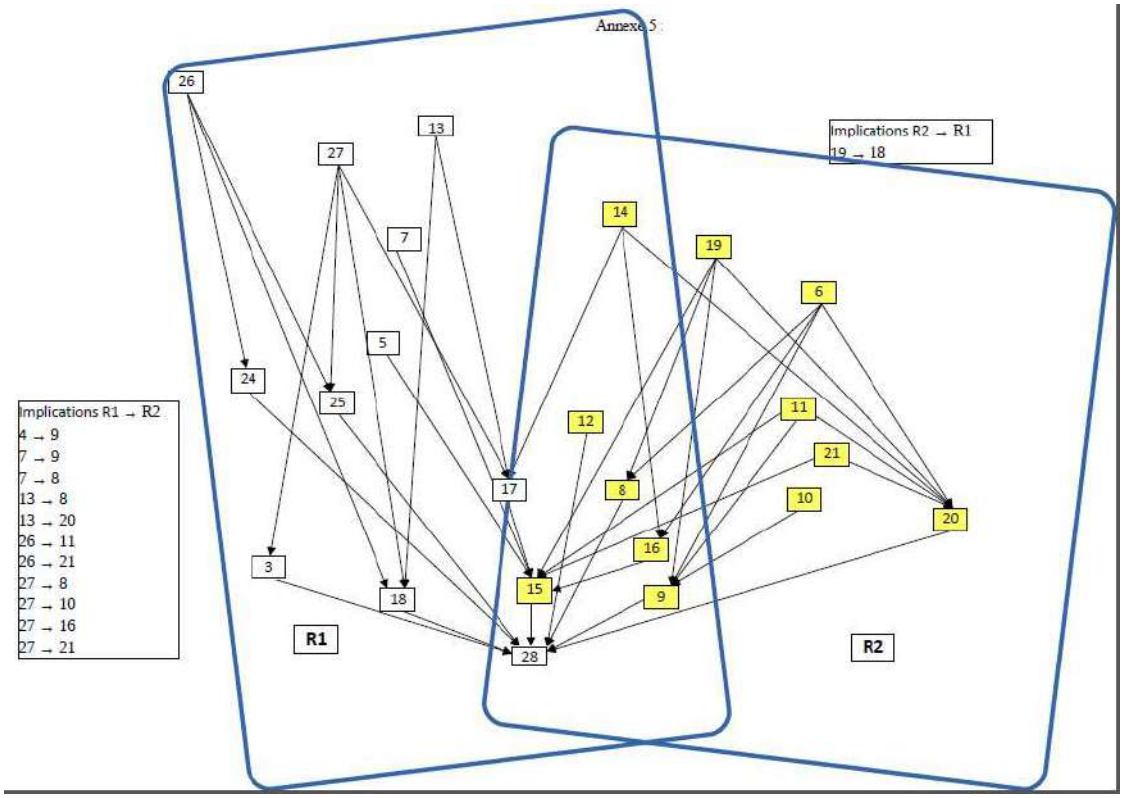
Annexe 3 : 2007, le graphe implicatif PE au seuil de .90



Annexe 4 : 2012, le graphe implicatif PE au seuil de .90



Annexe 5 : 2007, le réseau de l'adhésion et ses deux sous-réseaux



UMA ANÁLISE DAS DIFERENTES PRAXEOLOGIAS NO ENSINO DAS EQUAÇÕES DE SEGUNDO GRAU. UM OLHAR A PARTIR DA QUADRO TEORICO DA A.S.I[1]

Marcus Bessa DE MENEZES[2], Marcelo Câmara DOS SANTOS[3]

UNE ANALYSE DES PRAXÉOLOGIES DIFFÉRENTES DANS L'ENSEIGNEMENT DES ÉQUATIONS DU SECOND DEGRÉ. UN REGARD A PARTIR DE L'A.S.I.

AN ANALYSIS OF THE DIFFERENT PRAXEOLOGY IN THE TEACHING OF SECOND DEGREE EQUATIONS, A LOOK FROM THE S.I.A.

RESUMO

Essa investigação caracteriza-se por um novo olhar no trabalho de Tese defendida pelo proponente dessa pesquisa. A tese de doutoramento em questão tratou das diferentes praxeologias apresentadas pelo professor e pelos alunos durante o ensino das equações de segundo grau. Durante a investigação, pudemos observar como se comportam os saberes em sala de aula, o saber-fazer, que valoriza a técnica de resolução, e um saber de cunho mais científico que avança na consolidação do conhecimento. No entanto, algumas perguntas não foram respondidas por não ser o foco da pesquisa ou por falta de ferramentas que pudessem nos elencar informações sobre os eventos ocorridos. Com isso, em nossa nova pesquisa, propomos buscar elementos que respondam algumas perguntas abertas e promover um novo olhar sobre os dados que foram coletados a partir do quadro da A.S.I. Acreditamos que esse trabalho poderá fomentar discussões importantes para a formação do professor de Matemática quanto à gestão do saber em sala de aula.

Palavras chave: *Gestão da sala de aula; Fenômenos Didáticos; Teoria Antropológica do Didático; ASI.*

RÉSUMÉ

Cette recherche se caractérise par un nouveau regard sur le travail de thèse de l'auteur, sous la direction du coauteur de ce texte. La thèse de doctorat en question a traité des différentes praxéologies présentées par l'enseignant et ses étudiants pendant l'enseignement des équations du second degré. Au cours de la recherche, nous avons vu comment se comportent les connaissances dans la classe, le savoir-faire qui valorise la technique de résolution et une connaissance du point de vue plus scientifique, qui avance dans la consolidation des connaissances. Toutefois, certaines questions n'ont pas été abordées, soit pour n'être pas objet de la thèse, soit par le manque d'outils qui pouvaient nous fournir des informations sur les événements qui ont eu lieu dans le cours. Ainsi, dans cette nouvelle recherche, on a récupèré des éléments qui puissent permettre de répondre à des questions encore ouvertes et promouvoir un nouveau regard sur les données qui ont été recueillies,

[1] Esta pesquisa contou com o apoio da CAPES através da bolsa PVE do programa PPGEC-UFRPE com o Professor Jean-Claude Régnier UMR 5191 ICAR – Université Lyon2

[2] Universidade Federal de Campina Grande, Campus de Sumé-PB, Rua Luiz Grande S/N – Frei Damião, Sumé-PB – CEP 58540-000, marcusbessa@gmail.com

[3] Universidade Federal de Pernambuco, Colégio de Aplicação da UFPE, Avenida dos Funcionários, S/N - Cidade Universitária, Recife – PE – CEP 50740-550, marcelocamaraufpe@yahoo.com.br

dans le cadre de l'A.S.I.. Nous croyons que ce travail peut stimuler des discussions importantes pour la formation de l'enseignant de mathématiques au sujet de la gestion des connaissances dans la classe.

***Mots-clés**: Gestion de la classe ; Phénomènes didactiques ; Théorie anthropologique du didactique ; A.S.I.*

ABSTRACT

This investigation is characterized by a new look on work Theses defended by the proposer of this research. The doctoral thesis in question dealt with the different praxeologias presented by the teacher and the students during the teaching of second degree equations. During the investigation, we have seen how to behave in the classroom knowledge, know-how, who values the resolution technique, and a knowledge of more scientific slant that advances in the consolidation of knowledge. However, some questions have not been answered by not being the focus of research or by lack of tools that could list information about the events that occurred. With that, in our new survey, we get elements that respond some open-ended questions and promote a new look at the data that was collected from the frame of the S.I.A. We believe that this work can stimulate important discussions for the formation of the math teacher regarding the management of knowledge in the classroom.

***Keywords:** Classroom management; Didactic Phenomena; Anthropological theory of Didactic; S.I.A.*

1. Introdução

Procurando avançar em algumas questões abertas na tese intitulada: "Praxeologia do professor e do aluno: uma análise das diferenças no ensino de equações do segundo grau", estamos propondo um novo olhar para os dados que foram coletados durante a tese. Para isso, utilizaremos o quadro metodológico da A.S.I. como uma ferramenta de análise, a partir do software CHIC[4].

Hoje em dia, a grande maioria das escolas apresenta um discurso construtivista para a construção do conhecimento. A ideia construtivista na Educação é a de que o conhecimento é construído a partir de um processo de ensino-aprendizagem infinito, ou seja, a relação com o objeto de saber (rapport au savoir) vai se modificando com o tempo, e não se restringe aos conteúdos trabalhados em sala de aula, pois há que se considerar o meio social, o contexto, a cultura local. César Coll (2006) afirma que o Construtivismo não é uma teoria, mas uma referência explicativa, um auxílio na reflexão sobre a prática docente. Pois há a necessidade de compreender que o aluno é um aprendiz social e o professor um agente mediador entre o indivíduo e a sociedade.

Apesar desses discursos, percebemos em sala de aula que muitos alunos não conseguem explicar os mecanismos ou as técnicas que utilizam para resolverem as atividades ou tarefas a eles propostas. É como se soubessem fazer, no entanto, não sabem muito ao certo porque fazem, ou seja, repetem o que o professor fez, o que está no Livro Didático que é adotado pela escola ou, até mesmo, a forma como um parente ou amigo realiza essas tarefas. Enfim, percebemos a existência de um saber-fazer em detrimento de saber de cunho mais científico, o qual forneceria condições aos alunos de explicarem exatamente suas escolhas para a realização de suas tarefas.

[4] Classification Hiérarchique Implicative et Cohésitive.

Em nossa tese, observamos que alguns alunos realizam suas atividades ou tarefas de formas diferentes das que seus professores apresentam, porém, não sabem justificar suas escolhas, ficando limitados ao saber fazer. Todavia, não era o foco da tese saber se existia uma estabilidade por parte desses alunos na realização dessas tarefas e quais os motivos que levariam a essa invariabilidade, ficando assim como uma questão em aberto. Será a partir dessa questão que iremos propor nosso trabalho de pesquisa, o qual irá buscar elementos que indiquem a estabilidade dos alunos na resolução de suas tarefas e qual a influência dos professores para que isso ocorra. Utilizaremos o quadro metodológico da A.S.I.[5] para nos apontar essa possibilidade (ou não) de estabilidade nas escolhas dos alunos para a realização de uma tarefa.

Com os dados encontrados nessa pesquisa, acreditamos que poderemos levantar uma discussão nos Cursos de formação de professores de Matemática: qual o saber que é valorizado em sala de aula pelo professor? Pensando no construtivismo como a concepção de ensino que é defendida pelos pesquisadores em Educação Matemática, essa discussão é extremamente importante para a formação do aluno, pois, no enfoque construtivista, se atribui ao sujeito um papel ativo, o aluno é responsável por seu próprio processo de aprendizagem, cabendo ao professor criar situações e implementar condições para que se desenrole o processo de construção.

Acreditamos ser necessário a apresentação de algumas teorias que orientaram nosso olhar para a sala de aula durante o desenvolvimento dessa pesquisa.

2. Fundamentação Teórica

2.1 Um breve histórico sobre a A.S.I.

É a partir dos trabalhos de Régis Gras na década de 1970 que se inicia e impulsiona desenvolvimento da análise estatística implicativa, sendo esses trabalhos amplamente apoiados pelas reuniões internacionais realizadas sucessivamente no IUFM de Caen (França), em 2000 e 2012; na PUC de São Paulo (Brasil), em 2003; na Universidade de Palermo (Itália), em 2005 e 2010; e na Universidade de Castellón (Espanha), em 2007. O pesquisador Saddo Ag Almouloud, nos aponta que "o desenvolvimento da A.S.I. vem, antes de tudo, do fruto da interação prévia e posterior das aplicações em diversas áreas, tais como as Ciências da Educação, Psicologia, Sociologia, Didática, Bioinformática, Medicina, História da Arte etc." (ALMOULOUD, 2014, p. 624).

Ainda sobre a A.S.I, o pesquisador afirma o seguinte: "a Análise Estatística Implicativa (A.S.I.) visa à extração de conhecimentos, invariantes, regras indutivas não simétricas consistentes, e atribui uma medida para proposições do tipo 'quando a é escolhido, tende-se a escolher b'." (ALMOULOUD, 2014, p. 625).

Com intuito de buscar a explicitação de uma forma mais simples quanto possível, a intenção e o significado da análise estatística implicativa pode ser enunciada afirmando que a A.S.I. é um marco teórico para a análise de dados com base em uma relação não-simétrica. Sobre esse olhar, Gras & Regnier (2009) avançam:

> (...) um campo teórico, centrado no conceito de implicação estatística ou, mais especificamente, sobre o conceito de quase-implicação para

[5] Analyse Statistique Implicative.

> distingui-lo da implicação lógica das áreas de lógica e matemática. O estudo do conceito de quase-implicação como objeto matemático, nas áreas de probabilidade e estatística, permitiu construir ferramentas teóricas que instrumentam um método de análise de dados (GRAS, REGNIER, 2009, p. 12).

Em todos os colóquios realizados sobre a A.S.I., a abordagem foi realizada a partir de uma ferramenta fundamental, o software CHIC.

2.1.1 A ferramenta fundamental: o software CHIC

De acordo com Saddo Ag Almouloud (2014) O software, cuja sigla CHIC significa Classificação Hierárquica Implicativa e Coesitiva, foi elaborado na década de 1980 por Régis Gras, nos anos 90 por Saddo Ag Almouloud; posteriormente, por Harrison Ratsimba-Rajon, e, finalmente, por Raphaël Couturier nos dias de hoje.

O CHIC encontra-se na versão 6.d, podendo ser trabalhado em cinco idiomas: inglês, espanhol, francês, italiano, e português. Além de permitir, segundo Almouloud (2014): - tratar diferentes tipos de variáveis (binária, modais, frequênciasfrequências vetoriais, intervalos, fuzzy), - quantificar a significância dos valores atribuídos à qualidade, consistência da regra associada à relação implicativa entre variáveis, entre classes ordenadas de regras, a tipicidade e a contribuição de sujeitos ou categorias de sujeitos a determinadas regras, - representar por um gráfico, para um limiar de qualidade escolhido, os caminhos de regras e, por hierarquia, regras de entre regras (também chamadas de regras generalizadas), - excluir, adicionar, fazer conjunção de variáveis, - representar por uma hierarquia ascendente as classes de similaridade de variáveis.

Em sua última versão, um novo recurso, denominado '*Redução'*, foi introduzido, e é parte integrante da teoria da implicação estatística (A.S.I.) de R. Gras. Ainda segundo Almouloud, com esse novo recurso, "o CHIC permitirá reduzir um conjunto muito grande de variáveis a um subconjunto menor equivalente ao primeiro, no sentido de A.S.I., que permita ao usuário controle do limiar de redução, na tentativa de minimizar as informações ocultadas" (ALMOULOUD, 2014, p. 626).

2.2 Transposição Didática

A produção e a comunicação dos saberes de referência são necessidades sociais. O pesquisador, no mundo acadêmico/científico, sofre pressões internas e externas (ARSAC, 1989) para que comunique suas 'descobertas', suas 'teses'. As pressões internas aparecem quando a própria comunidade científica exige que tais saberes sejam comunicados, pois, a partir deles, novos saberes serão produzidos.

Por outro lado, existem, também, as pressões externas para a apresentação desse saber à sociedade. Os saberes comunicados, inicialmente no mundo acadêmico e científico, necessitam de um novo tratamento, no sentido de que sua roupagem mais acadêmica seja retirada e que ele possa, após essa primeira "transformação", ser comunicado, compreendido e, se possível, utilizado socialmente num período breve.

Acreditamos ser importante a identificação das diferenças entre os saberes que estarão envolvidos em nossa pesquisa, o primeiro deles será o saber científico. O saber

científico está associado à vida acadêmica, porém, devemos lembrar que não são todas as produções acadêmicas que serão saberes científicos. O saber científico é um saber criado nas universidades que irá servir de parâmetro para os saberes que irão chegar ao ensino básico, mas não está necessariamente vinculado a ele (ensino básico). A linguagem é uma das diferenças entre o saber científico e os outros saberes. Ela possui características diferentes nos outros saberes, visto que atende a um público específico, a comunidade científica, e assim sendo, não poderíamos levar esse tipo de linguagem para a sala de aula, pois dificilmente conseguiríamos auxiliar na compreensão e entendimento de nossos alunos.

Para identificarmos o próximo saber, o saber a ser ensinado, também chamado saber escolar, recorreremos a Luiz Carlos Pais (2001) que avança na seguinte explicação:

> O saber escolar representa o conjunto dos conteúdos previstos na estrutura curricular das várias disciplinas escolares valorizadas no contexto da história da educação. Por exemplo, no ensino da matemática, uma parte dos conteúdos tem suas raízes na matemática grega, de onde provém boa parte de sua caracterização. (PAIS, 2001, p 22)

Assim sendo, poderemos entender o saber a ser ensinado como todos os saberes eleitos para comporem a grade curricular de uma determinada disciplina. Será na "produção" do saber a ser ensinado que irão ser evidenciadas as diferenças, como avança Pais (2001), ao afirmar que na passagem do saber científico ao saber previsto na educação escolar ocorre a criação de vários recursos didáticos, cujo resultado prático ultrapassa os limites conceituais do saber matemático. A partir do surgimento desses recursos, surgem também as criações didáticas que fornecem o essencial da intenção de ensino da disciplina.

Outro ponto de diferença entre os saberes até aqui apresentados, está no seu aspecto. Enquanto o saber científico aparece a partir de artigos, teses, livros e relatórios o saber a ser ensinado se apresenta por meio de livros didáticos, programas e de outros materiais, o que ratifica a necessidade de uma linguagem diferente entre eles, tendo em vista o público ao qual são apresentados.

Segundo Ravel (2003), o saber preparado é o saber apresentado no plano de aula do professor, um saber que está envolvido com as expectativas que este professor tem com relação aos alunos, e ao saber a ser ensinado. Esse saber terá uma particularidade de que, normalmente, se apresenta de forma própria, pois as expectativas poderão ser diferentes para cada professor em relação ao grupo de alunos, que estão envolvidos no cenário didático.

O saber ensinado resulta das mudanças ocorridas durante a aplicação do que estava previsto no plano de aula (saber preparado) para o que efetivamente ocorre na sala de aula, ou seja, a realização, ou não, das expectativas. Esse saber será impregnado, principalmente, pela relação existente entre o professor e o saber a ser ensinado, o que irá orientar as mudanças que ocorrerão no processo de "produção" deste saber (saber ensinado), como avança Bessa de Menezes (2004):

> Um outro ponto está nas expectativas que os professores tinham em relação ao saber, fazendo, assim, com que esse objeto de ensino recebesse uma maior ou menor relevância no seu discurso em sala de aula, criando, desta forma, discursos diferentes para esses saberes em

> função dessas expectativas, as quais se apresentaram distintas para cada professor (BESSA DE MENEZES, 2004, p 131)

O saber aprendido seria o último saber dentro desse processo de apropriação do saber que ocorre em sala de aula. Diferente do que o nome dado a este saber possa parecer, principalmente para a área da psicologia da educação, iremos definir este saber como sendo todo e qualquer saber "retornado" pelo aluno, após esse saber ter sido "apresentado" em sala de aula. Sabemos que o mesmo (saber aprendido) não é somente formado pelo que é apresentado em sala de aula, ou seja, somente através do que é "ensinado"; temos consciência de que outras relações fora da sala de aula, na família, na comunidade em que vive, nos clubes, enfim, em outros locais onde pode aparecer esse saber em jogo, fazem com que nossos alunos tenham outras fontes para transformar este saber.

Para cada mudança no saber, nesse processo de intencionalidade do ensino, iremos ter as fases (ou etapas) da transposição didática, a saber: transposição didática externa e a transposição didática interna.

2.3 A Teoria Antropológica do Didático (TAD)

Segundo Chevallard, a sua teorização proposta na Teoria Antropológica do Didático (TAD) deve "(...) ser encarada como um desenvolvimento e uma articulação das noções cuja elaboração visa permitir pensar de maneira unificada um grande número de fenômenos didáticos, que surgem no final de múltiplas análises." (1998, p. 92)

Assim, podemos ver a TAD funcionando como uma forma de explicar a transposição didática (TD) no ecossistema[6] da sala de aula, ou melhor dizendo, um prolongamento da teoria da transposição didática, no momento em que amplia estes ecossistemas para relações, entre objetos de ensino, que irão além da sala de aula.

> Na prática, as primeiras análises propostas em *la transposition didactique*[7] limitavam-se a distinguir objetos «matemáticos», «paramatemáticas» e «protomatemáticos». O alargamento do quadro, levado a cabo por necessidades de análise, conduziu-me a propor uma teorização em que qualquer «objeto» pudesse aparecer : a função logarítmica é, evidentemente, um objeto («matemático»), mas existe igualmente o objeto «escola», o objeto «professor», o objeto «aprender», o objeto «saber», o objeto «dor de dente», o objeto «fazer xixi», etc. (CHEVALLARD, 1998, p.92)

O autor afirma que para começar sua teorização são necessários três conceitos primitivos: os objetos O, as pessoas X e as instituições I; e que outros virão a ser acrescentados subsequentemente.

O objeto O tomará uma posição privilegiada em relação aos outros temas, em virtude do mesmo ser o "material de base" da construção teórica. Tudo será objeto. Chevallard faz uma analogia com o universo matemático contemporâneo, o qual é fundado na teoria dos conjuntos, tudo é um conjunto. Assim também será na sua teoria,

[6] Entendemos ecossistema como sendo o local onde se desenvolve um determinado sistema que possui uma ecologia própria, no caso em estudo, o sistema didático.

[7] Ver Chevallard 1991.

"todas as coisas serão objetos", as pessoas X e as instituições I também são objetos, assim como as outras entidades que serão introduzidas.

O objeto irá existir no momento em que for reconhecido como existente por uma pessoa X ou instituição I. Com isso, aparecerão a relação pessoal de X com O, que será denotada por R(X, O), e a relação institucional de I com O, R(I, O). Ou seja, o objeto irá existir caso seja reconhecido por, pelo menos, uma pessoa X ou instituição I.

> Do ponto de vista da «semântica» da teoria, qualquer coisa pode ser um objeto. Um objeto existe a partir do momento em que uma pessoa X ou uma instituição I o reconhece como *existente* (para ela). Mais precisamente, podemos dizer que o objeto O existe para X (respectivamente, para I) se existir um objeto, que denotarei por R(X, O) (resp. R_1(O)), a que chamarei de *relação pessoal de X com O* (resp. *relação institucional de I com O*)[8]. (CHEVALLARD, 1998, p 93)

Chegamos a um ponto em que necessitamos evidenciar o que são as instituições. Segundo Chevallard (1998), "(...) uma instituição pode ser quase o que quer que seja". Devido à natureza da palavra, poderíamos dar uma conotação própria a esse personagem, ou seja: "Associação ou organização de caráter social, educativo, religioso, de ensino, etc." (Kury, 2002), porém, não devemos nos surpreender ao vermos, em certos momentos, objetos tomarem o status de instituição. Uma escola é certamente uma instituição, que possui outras instituições a ela agregada, uma sala de aula, por exemplo.

O conceito de Instituição pode ser explicitado como sendo um dispositivo social, total ou parcial, que impõe aos seus sujeitos formas de fazer e de pensar, que são próprias a cada "tipo ou forma" de instituição. Para avançarmos ainda mais sobre o conceito de instituição I, devemos percebê-la não como uma estrutura homogênea, mas sim heterogênea, em que existem várias relações de pessoas X com objetos O que pertencem a I.

Mas de que forma se relacionam os objetos O e instituição I? A cada instituição I está associado um conjunto de objetos O que são conhecidos por I, ou seja, existe uma relação institucional R(I, O).

> (...) A cada instituição I está associado um conjunto de objetos 0_1, chamado conjunto dos objetos *institucionais* (para I), que é o conjunto dos objetos O que I conhece, ou seja, para os quais *existe* uma relação institucional R_1(O). Um objeto O é institucional para I ou, dito de outro modo, existe para I, quando I define uma relação (institucional) com O.[9] (CHEVALLARD, 1999, p 225)

O objeto O se relaciona com a instituição I através de suas características próprias, por exemplo, a noção de porcentagem para uma instituição financeira (um banco) pode representar taxas e lucros, enquanto para a engenharia civil pode representar proporcionalidade entre partes de uma mistura (um traço de concreto). Assim sendo, o objeto O pode estabelecer diferentes formas de relações de acordo com a instituição R1(O), R2(O), R3(O), etc. Da mesma forma, seu desenvolvimento, dentro destas instituições, pode vir a ser modificado com o passar do tempo, ou seja, evoluir, envelhecer ou até mesmo desaparecer.

[8] Grifos do autor.
[9] Grifos do autor.

Essas relações são permeadas por outro fenômeno didático que surge nas relações dos sujeitos X com os objetos O da instituição I, fenômeno este que se estabelece devido às expectativas que existem dentro das relações, o contrato didático.

Apesar de já termos diversas vezes citado o próximo conceito fundamental da TAD, a pessoa, e não termos, ainda, definido o seu conceito, chegou a sua vez. Para isso, iniciaremos diferenciando alguns estágios deste conceito, a saber: o indivíduo, o sujeito e a pessoa. Podemos dizer que o estágio mais primitivo seria o de Indivíduo, visto que, ele não se sujeita, nem muda com as relações cotidianas com objetos e instituições. Chevallard afirma que:

> Bem entendido, no curso do tempo, o sistema das relações pessoais de X evolui; objetos que não existem para ele passam a existir; outros deixam de existir; para outros enfim a relação pessoal de X muda. Nesta evolução, o invariante é o indivíduo; o que muda é a pessoa (CHEVALLARD, 1999, 226).

O indivíduo se torna um sujeito quando se relaciona com uma Instituição I qualquer, ou melhor dizendo, quando se sujeita a uma Instituição I, sob suas demandas, hábitos, formas; enfim, se sujeitando a esta relação.

É por meio das várias relações que o indivíduo tem com instituições diferentes que se constitui a pessoa, ou seja, o conjunto de sujeitos do indivíduo é que forma a *pessoa* X, a qual irá mudando conforme estabelece suas relações com as instituições, as quais toma conhecimento com o passar do tempo.

> Uma pessoa X está sujeita a uma série de instituições. Introduzo aqui o axioma segundo o qual uma pessoa não é, na realidade, mais do que *a emergência de um complexo de sujeições institucionais*. Aquilo que se chama de «liberdade» da pessoa surge então com o efeito obtido em conseqüência *de uma ou de várias sujeições institucionais contra outras*.[10] (CHEVALLARD, 1999, p. 227)

Uma pessoa X entra para uma instituição I, e existe um objeto O para I, que é chamado de objeto institucional. Assim, X ao entrar em I, começa a viver uma relação com O sob a influência da relação institucional, ou seja, a relação R(X, O) irá se alterar ou se construir mediante a relação R(I, O), e, de forma mais ampliada, sob o constrangimento do contrato institucional C.

Devemos deixar claro que O poderia ou não existir para X antes de sua entrada para I (que analogamente podemos sugerir como conjunto vazio, sem existência), porém, independente desse fato, a relação R(X, O) irá alterar-se Daí, então, Chevallard dirá que há aprendizagem de X em relação a O. Ou seja, havendo alteração em R(X, O) então haverá aprendizagem da pessoa X sobre o objeto O. De forma análoga, caso R(X, O) não se altere, podemos afirmar que nada aprendeu. Devemos observar que não há nada de didático até agora, pois a instituição I não se manifestou com intencionalidade de fazer com que R(X, O) se altere ou modifique.

Para que a instituição I manifeste uma intencionalidade de fazer uma modificação ou uma alteração na relação R(X, O), é necessário que se introduza uma nova noção

[10] Grifos do autor.

primitiva, a do sujeito adequado. Com isso, uma pessoa X se tornará um sujeito da instituição I, relativamente ao objeto O, quando as relações R(X, O) e R(I, O) estão em conformidade. Ou seja, o sujeito está cumprindo as expectativas desejadas pela Instituição, está conforme "deseja" a Instituição. Caso isso não esteja ocorrendo, é considerado que o sujeito está inadequado em relação ao contrato institucional C.

Assim, entra em cena um desenvolvimento relativo à avaliação institucional. Segundo Chevallard (1999), essa avaliação é um dos mecanismos segundo os quais I é levada a pronunciar, através de alguns dos seus agentes, um veredicto de conformidade (ou de não conformidade) R(X, O) com R(I, O).". Ainda sobre esse assunto, o autor afirma que:

> A este respeito, as instituições são sempre «vigarizadas» (trapaceadas) pelos seus sujeitos. Quando esperam encontrar *sujeitos puros*, que julgam ser inteiramente moldados por elas, deparam-se com *pessoas*, que lhes aparecem sempre, de uma forma ou de outra, como *sujeitos desadequados*. (CHEVALLARD, 1999, p. 227)

É preciso ressaltar que, em nosso trabalho, desejamos observar a Instituição Sistema Didático, pois será lá o local em que poderemos verificar as transposições que serão efetuadas pelo aluno enquanto pertencente a esta instituição. Em outra instituição, com outros contratos estabelecidos, em virtude dos ecossistemas serem distintos, identificaríamos variantes que não poderíamos analisar no âmbito desta tese.

Assim sendo, podemos pensar que a instituição sala de aula – a qual chamaremos, a partir deste momento, de I1 – tem em seus sujeitos X1 – os alunos –, objetos O1 – saberes em jogo – e seus agentes que irão regular a conformidade, ou a não conformidade, com a instituição I1, de acordo com a intencionalidade estabelecida – são os professores, o contrato didático e o contrato institucional estabelecidos, as avaliações, entre outros, que aparecerão de acordo com o momento necessário.

A avaliação, como um dos elementos controladores da conformidade, ou não conformidade, na Instituição I1, pode, nesse sentido, vir (ao contrário do que se espera) a podar todo esse interesse pelo objeto O1, fazendo com que o sujeito X1 se preocupe somente com a conformidade, ou seja, quais são os conjuntos de sequência que deve realizar com intuito de ter a "adequação" esperada. Não podemos esquecer que essa avaliação é estabelecida por meio de um contrato pedagógico e um contrato didático definidos, que, de certa forma, dá sua importância dentro de I1. Assim sendo, isso poderá comprometer a formação dos conceitos desse objeto O1 em jogo no cenário didático.

Essas alterações nas relações entre o sujeito X1 e o objeto O1, (ou as transformações que o aluno faz do saber, quando tratamos da questão da Transposição Discente), vão muito além de uma questão epistemológica do objeto O1 (saber) ou de uma questão metodológica. Elas partem, também, de uma intencionalidade vinculada ao contrato que é estabelecido. Não significa que deixamos de fora esses outros fatores. Porém, é extremamente necessário, quando olhamos para o saber aprendido, a relação entre os contratos (pedagógico e didático) estabelecidos que têm, se assim podemos dizer, um peso maior nas escolhas realizadas pelos sujeitos X1 (alunos).

Algumas relações entre sujeitos, objetos e instituição são permeadas por intencionalidades diversas, tanto por parte dos sujeitos como, também, das instituições perante os objetos em jogo nessa relação. Fazendo um paralelo com a sala de aula,

podemos identificar vários fenômenos didáticos que ocorrem devido a essas intencionalidades, mediante as relações entre alunos e professores diante do saber a ser ensinado.

2.3.1 A Organização Praxeológica ou Praxeologia

Podemos entender uma organização praxeológica, ou praxeologia, como a realização de certo tipo de tarefas (T) através de um modo de fazer, que Chevallard (1999) chama de técnica (t). Essa associação tarefa-técnica (T-t) irá definir um saber-fazer próprio para esse tipo de tarefa. Porém, ela (T-t), não se mantém em estado isolado, ou seja, não se sustentará por si só. A T-t necessita de um amparo tecnológico-teórico (ou saber), que é formado por uma tecnologia (θ), que irá dar uma racionalidade e uma sustentação inteligível à técnica (t) aplicada, e uma teoria (Θ) que irá justificar e esclarecer a tecnologia (θ).

Assim sendo, a organização praxeológica ou praxeologia (que a partir desse momento iremos tratar somente como praxeologia) será composta por quatro elementos, a saber: tipo de tarefa (T), técnica (t), tecnologia (θ) e teoria (Θ); articulados a partir de um bloco prático-técnico (gerando o saber-fazer) e um bloco tecnológico-teórico (amparado no saber).

Segundo Chevallard (1998), a noção de tarefa, ou de tipos de tarefas, se encontra na raiz da noção de praxeologia. Podemos entender como tarefa (T), de acordo com a TAD, como todo e qualquer objeto que não encontramos sua existência diretamente na natureza, ou seja, será necessário realizar procedimentos próprios, no caso de nosso estudo: matemáticos, para encontrá-lo. Quando uma tarefa (t) ressalta de um tipo de tarefa T, escreveremos então: (t) ϵ T.

Para Chevallard (1998) podemos, ainda, diferenciar o gênero de tarefa do tipo de tarefa ou tarefa propriamente dita. O gênero de tarefa seria caracterizado por um verbo, como, por exemplo, montar, levar, calcular, etc., sendo expresso de forma mais ampla e conteúdo não definido. Já o tipo de tarefa, ou tarefa, tem seu conteúdo estritamente especificado. Assim sendo, para exemplificarmos um tipo de tarefa, temos resolver uma equação de primeiro grau, encontrar a altura de um triângulo isósceles, etc.

Chevallard (1998) observa ainda que um determinado tipo de técnica (t) não é universal para todas as instituições I. Em certos casos, algumas instituições não estão em conformidade com determinados tipos de técnicas, e assim sendo, não reconhecerão e contestarão a validade desta técnica (t).

Para poder dar um suporte racional e justificar a técnica (t) aplicada para a realização de uma tarefa (T) é necessário a introdução da noção de tecnologia (θ), a qual é definida por Chevallard (1998) como sendo:

> (...) um discurso racional (*logos*) sobre a técnica – a tekhnê – t, discurso tendo por objetivo primeiro de justificar 'racionalmente' a técnica t, e nos assegurar que ela permite o bom cumprimento das tarefas do tipo T, isto quer dizer realizar o que é pretendido. (CHEVALLARD, 1998, p 93)

Ainda sobre tecnologia (θ), Chevallard afirma que em dada instituição I uma técnica (t) para a realização de um tipo de tarefa (T) vem, frequentemente, acompanhada de vestígios ou embriões de tecnologia (θ), e, em diversos casos, na técnica (t), certos elementos tecnológicos vêm incorporados. O autor avança ainda, ao afirmar que quando em uma instituição I existe, em princípio, somente uma técnica (t) que é reverenciada, reconhecida e empregada, esta técnica adquire um papel de "autotecnológica", ou seja, não irá necessitar de justificativas, pois esta é a melhor maneira de se fazer nesta instituição I.

Para assegurarmos o funcionamento regular de uma tecnologia (θ) em uma instituição I, necessitamos de uma nova noção que explique e justifique esta tecnologia (θ). Esse fato nos leva com a noção de Teoria (Θ) que é a especulação abstrata da tecnologia; no plano teórico encontram-se as definições, os teoremas, as noções mais abrangentes que servem para explicar, justificar e produzir novas tecnologias.

Segundo Chevallard (1998), poderíamos chegar a uma regressão absurda, na qual sempre teríamos que justificar uma coisa atrás da outra, ou seja, a técnica justificada por uma tecnologia, que é justificada por uma teoria, que seria justificada por outra teoria, por outra e outra... Porém, o autor afirma que "(...) a descrição em três níveis (técnica/tecnologia/teoria), em geral, é o suficiente para dar conta da atividade a analisar" (CHEVALLARD, 1998, p 94).

Para analisarmos a prática docente, devemos observar as seguintes questões: Como realizar a tarefa do tipo T? Ou ainda, Como realizar melhor esta tarefa? Essas questões invocam uma produção de técnicas, e, portanto, de praxeologias.

Sendo os tipos de tarefa T, acima citados, objetos matemáticos O para serem tratados em uma instituição I (uma sala de aula qualquer), podemos considerar essa análise em duas classes distintas: a) observando o primeiro questionamento com um viés pela realidade matemática, poderemos construir uma realidade como uma praxeologia matemática ou organização matemática, a qual denominaremos como OM; b) ao observarmos o segundo questionamento, teremos um olhar sobre a didática, ou seja, de que forma encaminharemos a realidade matemática estabelecida na OM. Assim, essa realidade se denominará uma praxeologia didática ou uma organização didática OD.

Chamaremos de praxeologia matemática ou organização matemática, toda realidade matemática que está envolvida na resolução de um tipo de tarefa matemática T. Para isso, serão exigidas técnicas t, amparadas por um conjunto teórico-tecnológico [θ; Θ].

A organização matemática tem sua origem nas análises, efetuadas pelos professores[11], dos documentos oficiais existentes (tais como programas e manuais escolares, além do livro didático), dos quais saem os saberes matemáticos escolhidos a serem ensinados.

Outro ponto da prática docente será de como conduzir esta praxeologia matemática, agora estabelecida, para a sala de aula. Isto é, como transpor da realidade matemática para a realidade didática. Segundo Chevallard (1999), a construção da praxeologia se inicia em uma falta de técnica para a resolução de um determinado tipo de tarefa. Assim sendo, podemos pensar no exemplo dado anteriormente: "Como encontrar as raízes de

[11] Lembramos que, nesse momento em particular, estamos fazendo um olhar pela prática docente.

uma equação de 2º grau?", e fazer agora a seguinte questão: "Como ensinar a encontrar as raízes de uma equação de 2º grau?" Dar resposta a esta nova questão nos leva a elaborar um novo tipo de praxeologia, a praxeologia didática.

A organização (ou praxeologia) didática surge na intenção de pôr em prática, ou de conduzir, uma organização matemática qualquer. Será ela, a OD, que irá dar conta da (re)construção ou transposição de uma determinada OM. Assim como toda praxeologia, a OD é composta de tipos de tarefas que serão resolvidas por técnicas, as quais serão explicadas pelas tecnologias e justificadas por teorias.

Para identificarmos as praxeologias, do professor e dos alunos, e fazermos a análise das diferenças, e elencar elementos para análise a partir do quadro metodológico da ASI, foram realizadas algumas etapas, as quais apresentaremos a seguir na metodologia utilizada.

3 Metodologia

3.1 Caracterização dos sujeitos e ações seguidas

Tomamos como sujeito um professor do 9º ano, com formação em Licenciatura em Matemática, e seus 24 alunos. O saber matemático que será observado são as equações do segundo grau. Isso se deve por ser nesse ano (9º) que essas equações são introduzidas formalmente no domínio algébrico no Brasil. A escolha pela álgebra se dá pelo fato de considerarmos que a transição do domínio aritmético para o algébrico marca uma das mais importantes rupturas no ensino de matemática: de uma matemática mais 'concreta' a um campo que exige um nível maior de abstração e generalização.

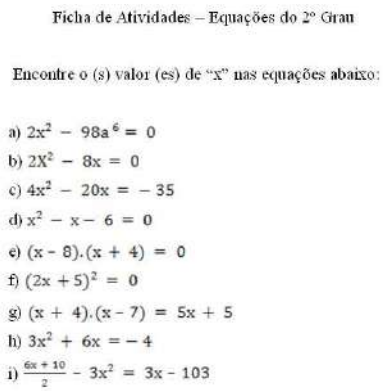
Uma questão que achamos pertinente destacar aqui, diz respeito à escolha da equação do segundo grau como conteúdo matemático a ser contemplado nesse estudo. As pesquisas voltadas para a Educação Matemática apontam para uma ruptura existente na passagem da aritmética à álgebra (VERGNAUD e CORTES, 1986; VERGNAUD, CORTES e FAVRE-ARTIGUE, 1987; KIERAN, 1992; BOOTH, 1995; USISKIN, 1995, dentre outros). Entretanto, entendemos que dentro da própria álgebra o aluno também se depara com outra ruptura, ao passar das equações de primeiro grau, para as equações de segundo grau. Enquanto que, no primeiro tipo de equações (1º grau), o aluno elege o procedimento de resolução (transposição de um membro para outro da igualdade, realizando a operação inversa, por exemplo), nas equações de 2º grau ele precisa lançar mão de outros procedimentos, como fatorar a equação ou mesmo utilizar a fórmula de Bhaskara.

Durante a coleta de dados para a tese (que é a base dessa nossa nova pesquisa), tivemos como referencial para observar a dinâmica da sala de aula a teoria antropológica do didático (TAD) (CHEVALLARD, 1999), a metodologia consistiu de uma análise das atividades propostas pelo professor em sala de aula, quando analisamos a sua prática sob o olhar da praxeologia, e comparamos com as atividades realizadas pelos alunos, também sob a ótica da praxeologia. Nas análises buscamos evidenciar os elementos que nos apontaram para as diferenças nas praxeológias do professor e dos alunos (técnicas, tecnologias e teorias) que vieram a aparecer.

Para identificarmos a praxeologia do professor e do aluno tivemos que realizar algumas ações. Primeiramente, realizamos uma descrição e análise do livro didático utilizado, a primeira ação. Com isso, buscamos estabelecer contato com o saber a ensinar, a partir da análise do livro didático utilizado, especificamente conteúdo de equação do segundo grau. Considerando que este tem se mostrado como uma espécie de "texto do saber", conforme a conceituação de Chevallard (1991). Posteriormente, observamos o professor, a segunda ação a ser realizada foi de identificar os tipos de tarefas que foram propostas. Para identificarmos esses tipos de tarefas, filmamos as aulas do professor sobre o conteúdo escolhido, equações de segundo grau, além de identificarmos elementos ostensivos e não-ostensivos que apareceram durante essa filmagem. A partir daí, partimos para a terceira ação que foi fazer uma análise da Organização Matemática, na qual identificamos as técnicas, os elementos tecnológicos e teóricos que se apresentam para a realização da tarefa. Com isso, identificamos a Organização Matemática proposta pelo professor.

Após a identificação dos elementos componentes da Organização Matemática, realizamos a quarta ação que consistiu em identificar quais são as técnicas, tecnologias e teorias mobilizadas pelo professor em sala de aula, também através das filmagens feitas anteriormente, para a realização do tipo de tarefa proposta, ou seja, a Organização Didática. Diferentemente da Organização Matemática, que tem um olhar sobre a realidade matemática envolvida, a Organização Didática foca em como pôr em prática o conteúdo matemático, nesse caso a equação do segundo grau. Com a intenção de facilitar o entendimento da Organização Didática realizada pelo professor, realizamos uma quinta ação que foi uma entrevista semiestruturada com o professor, na qual buscamos identificar suas escolhas na OD. A escolha por esse tipo de entrevista se deu porque ela pode nos permitir que, através do conteúdo manifesto da fala, alcançar respostas mais ricas e complexas, o que acreditamos ser mais proveitoso para o nosso trabalho, uma vez que reconhecemos que apenas com as observações não iríamos conseguir as informações necessárias, relativas ao interesse pela álgebra, especificamente no conteúdo de equações de segundo grau, o que como já citamos anteriormente, é um elemento importante para a formação do saber ensinado. Visto que, pode revelar indícios de sua relação (do professor) com o saber em jogo (equação do segundo grau). Para a identificarmos os elementos da praxeologia do aluno, elaboramos uma lista de atividades que continha os mesmos subtipos de tarefa que foram apresentados pelo professor em sala de aula. Nessa lista, buscamos as técnicas, tecnologias e teorias que foram mobilizadas pelos alunos participantes da pesquisa na realização desses subtipos de tarefas. Essa foi a nossa sexta ação. Será a partir desses dados coletados que iniciamos nossa pesquisa com a A.S.I. Diante das devolutivas dos alunos na atividade proposta, elaboramos nossas variáveis e geramos as tabelas necessárias para a análise com o software CHIC. A lista de atividades possui nove itens, os quais serviram de variáveis para análise do tipo de técnicas que foram utilizadas para a resolução de cada item, bem como, para identificarmos a praxeologia que foi utilizada pelos alunos. Os itens 'c' e 'h' da lista, não foram observados na pesquisa por erros de digitação na sua confecção, assim sendo, foram analisados sete itens.

A seguir, apresentamos a Lista de Atividades que foi aplicada aos alunos.



Ficha de Atividades – Equações do 2° Grau

Encontre o (s) valor (es) de "x" nas equações abaixo:

a) $2x^2 - 98a^6 = 0$
b) $2X^2 - 8x = 0$
c) $4x^2 - 20x = -35$
d) $x^2 - x - 6 = 0$
e) $(x - 8).(x + 4) = 0$
f) $(2x + 5)^2 = 0$
g) $(x + 4).(x - 7) = 5x + 5$
h) $3x^2 + 6x = -4$
i) $\frac{6x+10}{2} - 3x^2 = 3x - 103$

Figura 1 – Lista de Atividades aplicada aos alunos.

3.2 Tarefa em jogo, Variáveis e as delimitações para o uso do quadro metodológico da A.SI..

A tarefa T em jogo foi de *Resolver equações de segundo grau*. Durante as aulas foram apresentadas, pelo professor, 07 (sete) subtipos de tarefas relativas a T, nas quais identificamos as seguintes estruturas:

Subtipos de tarefas identificadas

T_1: $ax^2 + c = 0$
T_2: $ax^2 + bx = 0$
T_3: $(ax + c)^2$
T_4: $(x + a).(x + b) = 0$
T_5: $(x + a).(x + b) = cx + d$
T_6: $\frac{(ax+b)^2}{c} + dx^2 = ex + f$
T_7: $ax^2 + bx + c = 0$

Figura 2 – Subtipos de tarefas identificadas nas durante as aulas

Na tarefa T, em nenhum momento, se propõe a identificação ou a conceituação de equações do segundo grau. Todo o trabalho do professor gira em torno da resolução destas equações. Isso é percebido, em sala de aula, no momento em que as aplicações se voltam a simplesmente resolver tipos de equações. O trabalho do professor finda em sala de aula com a resolução de exercícios. O professor apresenta alguns subtipos de tarefa, os quais identificamos sete, anteriormente enumerados em T_1, T_2, T_3... T_7, e os resolve utilizando-se de técnicas por ele (o professor) apresentadas. Podemos colocar, em linhas gerais, que o procedimento do professor em sala de aula foi de apresentar técnicas para os subtipos da tarefa T que foram expostas em sala de aula e fazer aplicação dessas técnicas com base em exercícios. O professor reafirma essa hipótese durante a entrevista, quando anuncia que,

> **Prof**: (...) Então, eles iam começar a fazer sem usar a fórmula de Bhaskara, usando apenas as propriedades, quando eles... quando não tinham o valor de "b", resolviam por radicais, e quando tinham o valor de "b" e não tinham o de "c", faziam por fatoração. A gente passou bastante tempo fazendo isso para

> eles reconhecerem melhor. Aí a gente, depois, entrou no trinômio quadrado perfeito, para eles identificarem quando elas forem completas, aí resolveram completando quadrados. Aí, depois que eles trabalharam nestas fases, aí eu entrei na equação do segundo grau, com a fórmula de Báskara. (...) A partir disso foi identificar a fórmula de Báskara saber o que é o valor de "a", de "b" e de "c", e resolver, parte mecânica mesmo, só substituindo, aí depois, nós fomos para a parte de problemas".

Como dito anteriormente, pedimos aos alunos que resolvessem uma lista de atividade a qual continha os sete subtipos de tarefas. Em posse da devolutiva dos alunos, pudemos identificar as técnicas utilizadas pelos alunos, assim como, as praxeologias utilizadas.

Quanto às técnicas utilizadas, foram identificadas as seguintes técnicas:

Nomenclatura da Técnica	**Sigla utilizada no CHIC**
Não Fez	NF
Bháskara	BHAS
Propriedade do Produto Nulo	PN
Transposição dos Termos (Conjunto dos Números Reais)	TT
Fatoração[12]	FAT
Radiciação	RAD
Tentativa	TENT

Quadro 1 – Técnicas utilizadas pelos alunos para resolução da Lista de Atividades

Para cada item da atividade foram elaboradas as possibilidades de cada evento, assim sendo, tivemos as seguintes variáveis: **V1NF** (Item 1 – Não Fez); **V1BHAS** (Item 1 – Bháskara); **V1PN**(Item 1 – Propriedade do Produto Nulo); **V1TT**(Item 1 – Transposição dos Termos); **V1FAT**(Item 1 – Fatoração); **V1RAD**(Item 1 – Radiciação); **V1TENT** (Item 1 – Tentativa); **V2NF** (Item 2 – Não Fez); **V2BHAS** (Item 2 – Bháskara); e assim sucessivamente nos sete itens que foram analisados.

Quanto às praxeologias, elaboramos duas possibilidades para cada item da Lista de Atividades, ou o aluno seguia a praxeologia que o professor apresenta em sala de aula, a qual denominamos Praxeologia do Professor, ou utilizou uma praxeologia própria, a qual iremos tratar como Praxeologia do Aluno. Assim sendo, temos as seguintes variáveis quanto às praxeologias: **PROF_T1** (Item 1 – Praxeologia do Professor); **ALUNO_T1** (Item 1 – Praxeologia do Aluno); **PROF_T2** (Item 2 – Praxeologia do Professor); **ALUNO_T2** (Item 2 – Praxeologia do Aluno); **PROF_T3** (Item 3 – Praxeologia do Professor); **ALUNO_T3** (Item 3 – Praxeologia do Aluno); **PROF_T4** (Item 4 – Praxeologia do Professor); **ALUNO_T4** (Item 4 – Praxeologia do Aluno); **PROF_T5** (Item 5 – Praxeologia do Professor); **ALUNO_T5** (Item 5 – Praxeologia do Aluno); **PROF_T6** (Item 6 – Praxeologia do Professor); **ALUNO_T6** (Item 6 – Praxeologia do Aluno); e **PROF_T7** (Item 7 – Praxeologia do Professor); **ALUNO_T7** (Item 7 – Praxeologia do Aluno).

[12] Estamos chamando de *Fatoração* a regra da soma e do produto das raízes da equação, e não o conceito matemático. Fizemos essa opção porque alguns professores se reportam a regra da soma e do produto como fatoração, assim como o professor dessa pesquisa também o fez.

Quanto às delimitações do uso do quadro metodológico da A.S.I., para a análise dos dados coletados na devolutiva dos alunos com o software CHIC, utilizamos como *Tipo de Implicação* segundo ***Método Entrópico*** (GRAS ET AL, 2009, 2013), por melhor satisfazer ao objetivo de modelagem da inclusão conjuntista, além de ser mais severa no diz respeito à intensidade de implicação. Quanto ao *Tipo de Lei* fizemos a opção pela ***Lei Binomial***, tendo em vista termos um número baixo para as tentativas (n=24 alunos). Caso tivéssemos uma amostra igual ou superior a 30 (n≥30), poderíamos usar a Lei de Poisson como uma aproximação da distribuição Binomial.

Faremos, a partir de agora, uma análise dos sete subtipos de tarefa que apareceram durante as aulas. Mais adiante, iremos apresentar uma síntese com as expectativas, diante as apresentações das técnicas pelo professor, de resolução para cada item da Lista de Atividades.

3.3 Subtipos de Tarefas

3.3.1 Subtipo de Tarefa T_1: $ax^2 + c = 0$

No primeiro subtipo de tarefa T1 foram reunidos os exercícios que levavam os alunos a resolver equações de segundo grau apresentadas na forma $ax^2 + c = 0$. Parece-nos que a intenção do professor, nesse subtipo de tarefa, era de aproximar ao máximo das resoluções de equações de primeiro grau.

Assim sendo, para realizar esse subtipo de tarefa, o professor começa fazendo uma breve explanação sobre os graus da equação. Para isso, escreve no quadro a seguinte equação: $x^3 + y^x + 12 + 10 = 0$, e faz o seguinte comentário: "(...) se for em relação a 'x' a equação será do segundo grau, em relação a 'y' será de grau 'x'". Em seguida, o professor utiliza um exercício do livro adotado pela escola: $3x^2 - 75a^4 = 0$. Durante a resolução, o professor isolou o $3x^2$, transpondo o termo $75a^4$ invertendo as operações. Essa técnica, transpor os termos e inverter as operações, é a escolhida pelo professor como a técnica principal para a resolução de exercícios desse subtipo de tarefa T_1: $ax^2 + c = 0$. Após a transposição do termo $75a^4$, o professor passou o número três dividindo e reduziu a expressão para $x^2 = 25a^4$, nesse momento fez uma referência ao conteúdo de radicais: "(...) estão lembrados do que trabalhamos com radicais? Quando passamos o 2 (índice) para o outro lado... teremos a raiz! (os alunos repetiram junto com o professor)" extraiu a raiz e encontrou o valor da incógnita $5a^2$. É importante salientar que o professor não considera a raiz negativa $(-5a^2)$.

A subtécnica adotada pelo professor, para a resolução desse subtipo de tarefa, foi desenvolver ou reduzir expressões, técnicas que os alunos aprenderam para resolver equações de primeiro grau. Pudemos identificar, também, apesar de não serem explicitados, os elementos tecnológicos que deram a suporte à técnica. Foram eles: as propriedades das operações inversas em R (conjunto dos números reais) ou leis de transposição de termos e as propriedades da radiciação.

De certa forma, o professor pretendia começar por um subtipo mais próximo ao que os alunos já tinham trabalhado (equações de primeiro grau) até chegar a um subtipo que não seria mais possível, com as técnicas "até agora conhecidas", chegar a uma solução

da equação. No entanto, o professor não deixa claro que estava buscando uma "evolução das técnicas". A reação dos alunos era de passividade, repetiam uníssonos as "deixas" do professor, observando cada passo na resolução do exercício feito no quadro.

3.3.2 Subtipo de tarefa T_2: $ax^2 + bx = 0$

Neste subtipo de tarefa o professor utiliza como técnica principal, ou primária, a fatoração, deixando a técnica do produto do produto nulo e de transposição dos termos como subtécnicas (auxiliares ou secundárias). Os elementos tecnológicos, novamente, não foram evidenciados pelo professor. Porém, pudemos identificar as propriedades distributiva da multiplicação, do produto nulo e das operações inversas em R (conjunto dos números reais) ou leis da transposição de termos, que serviram para dar uma sustentação inteligível a técnica e as subtécnicas utilizadas nesse subtipo de tarefa.

O exercício utilizado pelo professor para esse subtipo foi a equação $3x^2 - 24x = 0$. O diálogo que ocorre na sala de aula é o seguinte:

> **Prof**: *Como é que iremos resolver essa?* (apontando para a equação escrita no quadro)
> Os alunos começam a conversar entre si e o professor aguarda o silêncio da turma. Quando o silêncio é estabelecido, o professor se volta para a turma e indaga:
> **Prof***: E aí?* (alguns alunos tentam fornecer algumas ideias, porém falam todos juntos sem que haja uma sincronia. Então o professor diz) *Vamos lá! Isolo o 'x' e tenho: 'x' vezes* $(3x - 24)$ *igual a zero (fatoração - técnica principal). Quando tenho um produto de dois números que dá zero, então é porque um deles é zero ou os dois são zero. (produto nulo - subtécnica)".*

O professor se volta para o quadro e iguala os termos a zero desenvolvendo as expressões, encontrando as raízes zero e oito. A partir da utilização da técnica de fatoração, por parte do professor, os alunos permanecem em silêncio acompanhando todo o desenvolvimento da resolução do exercício.

3.3.3 Subtipo de tarefa T_3: $(ax + c)^2$

O subtipo T3 apresenta uma pequena variação dos subtipos anteriores. Para sua resolução o professor extraiu a raiz de zero, isolou a incógnita e inverteu as operações. Para este subtipo, o professor não considerou a possibilidade de resolver a partir da propriedade da potenciação $(a^2 = a.a)$ e, em seguida, utilizar a subtécnica do produto nulo, utilizada anteriormente. Caso os alunos desenvolvessem o produto de $(ax + c).(ax + c) = 0$, aplicando a propriedade distributiva da multiplicação, poderia causar um bloqueio, pois eles ainda não possuíam técnicas que dessem conta, visto que, os alunos, teoricamente, ainda não foram apresentados à fórmula de Bhaskara. Um outro elemento que não foi considerado pelo professor, foi ressaltar a existência de duas raízes, apesar de serem iguais.

E, por fim, o professor não demonstrou que a técnica aplicada também "funciona" quando a equação é igualada a outro número qualquer diferente de zero, tendo em vista que em todos os exercícios apresentados, neste subtipo de tarefa, a expressão aparece igualada a zero. Podemos analisar este fato com um elemento indicador de transposição

didática interna, já que no livro utilizado, esse subtipo de tarefa, apresenta equações igualadas a números diferentes de zero. Acreditamos que a estratégia do professor era, de certa forma, apresentar as equações em condições, teoricamente, mais simples para a resolução, focando, principalmente, na aplicação da técnica principal.

Neste subtipo de tarefa o professor utilizou o seguinte exercício do Livro: $(z + 5)^2 = 0$. O professor inicia a resolução perguntando a turma:

> **Prof**: *Como é que iremos resolver essa equação aqui?* (Aguarda durante um tempo alguma resposta dos alunos, que estão agitados, com vários falando ao mesmo tempo, até que ele continua)
> **Prof**: *Eu posso fazer desse jeito aqui!* (a turma faz silêncio e ouve o professor) *'z' mais cinco igual... Se é potencia passa pra lá como?.*
> **Alunos**: *Raiz*! (O professor se volta para o quadro e escreve o sinal da raiz no zero. É nesse momento que o professor apresenta a técnica principal, a extração da raiz quadrada. Após, ele volta a falar)
> **Prof**: *z mais cinco é igual... Quanto é a raiz de zero?*
> **Alunos**: *Zero*! (O professor termina de resolver a equação dizendo que a raiz é cinco, para isso, utilizou como subtécnica a transposição dos termos, invertendo as operações. Como foi dito anteriormente, o professor não fez nenhuma alusão às raízes iguais).

Em nenhum dos subtipos de tarefa os elementos tecnológicos são evidenciados de forma clara. Em um momento, ou outro, se faz uma explicação da técnica ou subtécnica utilizada, mas sem explicitar a tecnologia e a teoria que, de certa forma, deram uma racionalidade a essas técnicas e subtécnicas. Porém, em nossa tese, buscamos identificar, em todos os subtipos de tarefa, os elementos tecnológicos que compõem o bloco teórico-tecnológico. Para esse subtipo de tarefa, as propriedades da radiciação e das operações inversas em R (conjunto dos números reais) ou leis da transposição de termos, como elementos que constituem esse bloco teórico-tecnológico.

3.3.4 Subtipo de tarefa T_4: $(x + a).(x + b) = 0$

O subtipo de tarefa T4, diferentemente de T3, gera duas raízes distintas. Nesse subtipo de tarefa, a técnica utilizada é a do produto nulo que ganha um status, nesse subtipo de tarefa, de técnica principal ou primária, deixando para a transposição de termos e desenvolvimento da expressão o papel de auxiliares ou secundárias. Os elementos tecnológicos observados foram as propriedades do produto nulo e as operações inversas em R (conjunto dos números reais) ou leis da transposição de termos.

A equação $(x - 17).(x + 11) = 0$ é utilizada como exemplo para este subtipo de tarefa. Ao iniciar a resolução o professor relembra aos alunos que a técnica do produto nulo já foi aplicada, afirmando que: "Na aula passada, nós já fizemos algumas questões parecidas com essa, que era o seguinte: Eu tenho aqui dois números que quando eu multiplico dá zero, e aí?", os alunos respondem: "Um deles é zero!".

As questões parecidas, mencionadas pelo professor, são as do subtipo de tarefa T_2 $(ax^2 + bx = 0)$, nas quais ele, como descrito anteriormente, utiliza-se da fatoração como técnica principal (colocando em evidência o termo comum "x"), para depois

aplicar as subtécnicas do produto nulo (igualando os termos do produto a zero), da transposição dos termos e do desenvolvimento das expressões.

Continuando a resolução, o professor avança: “É isso aí! Um dos dois vai ser zero!”. O professor começa a escrever no quadro igualando os termos a zero, encontrando as duas raízes da equação: 17 e -11. Após a transposição dos termos e o desenvolvimento das expressões, o professor alerta para a existência das duas raízes dizendo: “Então, as raízes dessa equação será 17 e -11”.

Assim como nos subtipos de tarefas anteriores, os alunos se portam de modo indiferente, observando a resolução do professor e respondendo juntos as indagações do professor.

3.3.5 Subtipo de tarefa T_5: $(x + a).(x + b) = cx + d$

Da mesma forma que foi feito no subtipo de tarefa anterior, no subtipo T_5, o professor utiliza subtécnicas preliminares para chegar a uma equação de 2º grau, são elas, desenvolvimento ou redução de expressões e a transposição de termos invertendo as operações, chegando a um subtipo de tarefa já conhecido pelos alunos T_1: $(ax^2 + b = 0)$. Com isso, tivemos como técnica principal uma das subtécnicas preliminares, a de desenvolvimento ou redução de expressões. Ou seja, a técnica mudou de status no desenvolvimento da resolução do subtipo de tarefa, devido a ser novamente utilizada durante a atividade.

Para esse subtipo de tarefa, identificamos como elementos tecnológicos as propriedades distributiva da multiplicação, das operações inversas em R (conjunto dos números reais) ou leis da transposição de termos e da radiciação.

Para esta resolução, é necessário que o valor de c seja igual à soma de a e b $(c = a + b)$, para que possa eliminar o termo que é multiplicado por x. E assim, voltar a um modelo T_1: $(ax^2 + b = 0)$. Novamente, a equação é preparada para se chegar a uma equação incompleta. Assim, a partir do desenvolvimento da multiplicação entre $(x + a).(x + b)$, da transposição e da redução dos termos, além da inversão das operações, o professor chegará ao valor da incógnita.

A equação utilizada pelo professor foi $(2x + 1).(x + 3) = 7x + 11$. Mais uma vez, o professor aguarda o silêncio da turma para começar a resolver a equação. Assim que é feito o silêncio, o professor pergunta:

> **Prof**.: *E aí? Como vamos resolver essa?* (Os alunos ficam em silêncio, sem darem nenhuma opinião. Então, o professor começa a resolver)
> **Prof**.: *Multiplica isso daqui* (apontando para o primeiro termo da equação), *então vamos lá.* (O professor utiliza a propriedade distributiva da multiplicação, passa os termos do segundo membro $(7x + 11)$ para o primeiro membro mudando o sinal, verbalizando)
> **Prof**.: *Eu já posso passar esses pra cá* (apontando para o segundo termo da equação. Em seguida, ele reduz os termos semelhantes) *Vamos agora agrupar e cortar os iguais,* e continua *e o que temos agora?* $2x^2 - 8 = 0$. (A partir desse momento o professor desenvolve a equação de modo similar ao subtipo de tarefa T1, no qual aplica a técnica de transposição dos termos e desenvolvimento de expressões).

Nesse subtipo de tarefa, o professor, novamente, não considerou a raiz negativa da equação, tratando somente da raiz positiva.

3.3.6 Subtipo de tarefa $T_7: \frac{(ax+b)}{c} + dx^2 = ex + f$

O subtipo de tarefa T_7 demanda a determinação do MMC, o desenvolvimento de expressões e a transposição de termos como subtécnicas preliminares. O professor não considera a técnica da simplificação ou a multiplicação de todos os termos pelo mesmo número para a resolução desse subtipo de tarefa.

Assim como ocorreu no subtipo de tarefa T_6, uma subtécnica muda de status passando para técnica principal. Novamente, a equação é preparada para se chegar em uma equação incompleta, um modelo já conhecido pelos alunos, o $T_1: (ax^2 + b = 0)$, após a utilização das subtécnicas preliminares.

Quanto aos elementos tecnológicos identificamos as seguintes propriedades relativas aos números racionais, as operações inversas em R (conjunto dos números reais) ou leis da transposição de termos e a da radiciação.

O professor utiliza a equação $\frac{(2x+3)}{2} - x^2 = x + 6$ para iniciar esse subtipo de tarefa. O professor começa a resolver a equação chamando a atenção dos alunos

> **Prof**.: *Silêncio! Essa equação aqui, como é que eu começo a resolver? Quando eu tenho essa situação* (ele está falando do denominador) *o que é que eu tenho que fazer? Vou tirar o MMC gente! Toda vez que eu tiver fração... Silêncio! Toda vez que eu tiver fração, eu tenho que tirar o MMC!* (Em seguida, ele se volta para o quadro e começa a narrar todos os passos que está fazendo)
> **Prof**.: *Eu retirei o MMC e agora vou colocar esses termos para o outro lado..., até chegar à equação* $-2x^2 - 9 = 0$. (A partir desse momento, ele desenvolve conforme fez no subtipo de tarefa $T_1: (ax^2 + b = 0)$, transpondo termos e desenvolvendo e reduzindo expressões, sempre voltado para o quadro narrando todas as etapas que são efetuadas).

No final, o professor encontra uma raiz que não existe para o conjunto dos números reais $\sqrt{\frac{-9}{2}}$. O professor pergunta a turma:

> **Prof**.: *Quanto é que é essa raiz aqui?*
> **Alunos**: *É a raiz de quatro vírgula cinco!*
> **Prof**.: *Eu posso achar a raiz desse número aqui? Não! Vocês lembram que quando o índice aqui for par, não pode ser negativo aqui dentro! Então, isso aqui não tem raiz nos reais. Certo?* (Em seguida, ele escreve no quadro) *Não existe raiz nos reais.*

3.3.7 Subtipo de Tarefa $T_7: ax^2 + bx + c = 0$

O subtipo T_8 corresponde às atividades agrupadas com a seguinte forma: $ax^2 + bx + c = 0$. Esse subtipo já vai apresentar a equação de segundo grau

completa e, com isso, o professor começará utilizando como técnica principal ou primária a resolução completando quadrados, até, enfim, introduzir outra técnica que irá substituí-la, a fórmula da Bhaskara. Essa técnica já era uma expectativa de alguns alunos, pois, como afirmou o professor durante a entrevista, alguns já a conheciam. Vale a pena salientar que nenhum dos alunos utilizou a técnica de completar quadrados durante a resolução da lista de exercícios, como iremos ver posteriormente.

A estratégia elaborada pelo professor para mostrar "todos os tipos de formas" de resolução (técnicas) das equações de segundo grau, foi de começar apresentando equações incompletas, passando por casos particulares de equações completas (completar quadrados) culminando no *grand finale* com uma fórmula para todos os tipos de tarefas: a fórmula de Bhaskara. Como pudemos observar durante seu discurso em sala de aula:

> **Prof**.: *Gente! Pronto! É o seguinte, a gente antes estava resolvendo as equações de segundo grau de várias maneiras... Essa daqui foi criada* (apontando para a fórmula de Bhaskara que estava escrita no quadro) *para resolver todos os tipos de equação de segundo grau.*

Após analisarmos todos os subtipos de tarefas, pudemos verificar, durante a resolução dos alunos, que, apesar de o professor cumprir exatamente o que programou para as suas aulas, essa organização matemática levou-os a uma possível dificuldade. Nas tentativas de encontrar as soluções dos exercícios, os alunos, normalmente, encontravam uma única raiz para equação. Acreditamos que esse comportamento dos alunos foi devido à associação feita com as equações de primeiro grau, que apresentam uma única raiz. Outro fator que pode ter contribuído para esse comportamento, foi que o professor, em algumas resoluções, como evidenciamos durante a análise dos subtipos de tarefa, não alertou para a existência das duas raízes.

Outro detalhe que observamos durante as aulas, foi que o professor não apresentava os subtipos de tarefas como novos desafios, de forma que com as técnicas até então "conhecidas", não conseguiríamos realizar a nova atividade proposta, ou seja, deveríamos buscar novas técnicas que iriam superar a(s) anterior(es). Tal fato é importante para que o aluno não fique desestimulado e desista de realizar as tarefas propostas, bem como não fique parado, aguardando que o professor apresente a nova técnica. Sobre essa evolução de técnicas, Chevallard (1999) afirma que quando uma "maneira de fazer" tem êxito somente sobre uma parte de um tipo de tarefa T, a qual é relativa, essa parte se denomina alcance da técnica. Essa técnica será "substituída" por outra que dê conta, se não sobre toda tarefa T, ao menos sobre parte dela. Nessa visão, podemos dizer que uma técnica é superior à outra.

3.4 Expectativas de resolução da Lista de Atividades

Foi elaborado um quadro (2) em que são apresentadas as expectativas de resolução de cada item da Lista de Atividades[13]. Essas Expectativas foram elaboradas diante a apresentação das técnicas durante as aulas do conteúdo de equações de 2º grau, e servirá para identificar se o aluno seguiu ou não as técnicas apresentadas pelo professor o que irá caracterizar sua praxeologia.

[13] Os itens 'c' e 'h' foram descartados para análise, por erro de digitação durante a confecção da Ficha de Atividades. Com isso, os itens não possuem raízes reais, o que não era objeto de estudo da pesquisa.

Variável	Item da Lista	Técnica esperada
1	a) $2x^2 - 98a^6 = 0$	**RAD** – Radiciação
2	b) $2X^2 - 8x = 0$	**FAT** – Fatoração
3	d) $x^2 - x - 6 = 0$	**BHAS** – Bháskara
4	e) $(x - 8).(x + 4) = 0$	**PN** – Produto Nulo
5	f) $(2x + 5)^2 = 0$	**RAD** – Radiciação
6	g) $(x + 4).(x - 7) = 5x + 5$	**TT** – Transposição dos Termos
7	i) $\frac{6x + 10}{2} - 3x^2 = 3x - 103$	**TT** – Transposição dos Termos

Quadro 2 – Expectativas de resolução

Será a partir dessas expectativas que veremos se ocorre ou não uma uniformidade na ação do aluno durante a resolução de atividades em sala de aula. Ou seja, se esses alunos indicam ter uma preferência por determinada técnica ou se seguem as técnicas apresentadas pelo professor. Diante das análises dos gráficos fornecidos polo CHIC e com os dados que foram coletados durante o estudo da tese que serve de base para essa pesquisa, poderemos avançar nas análises apresentadas na conclusão da tese, e levantar novas hipóteses sobre o que aconteceu em sala de aula.

4. Análise com o software CHIC

Para iniciarmos nossas análises com o software CHIC, elaboramos duas tabelas com os dados coletados. Uma denominada Técnica x Aluno (**Tabela 1**), em que identificamos quais as técnicas que os alunos utilizaram em cada um dos itens da Lista de Atividades, e outra denominada Praxeologia x aluno (**Tabela 2**), em que identificamos se o aluno acompanhou ou não a técnica do professor em cada um dos itens da atividade.

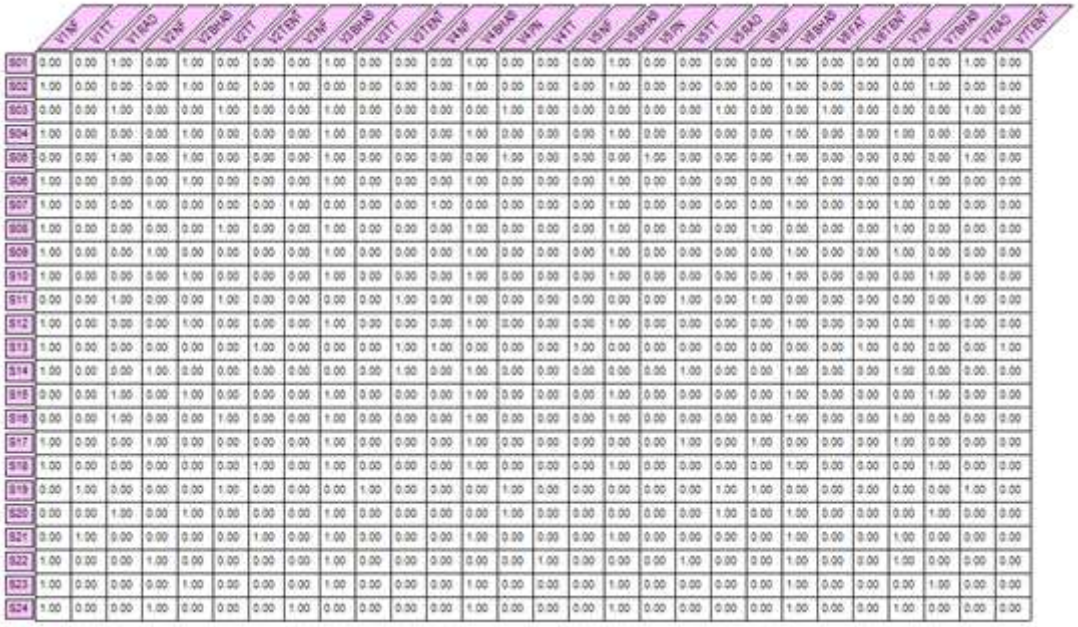


Tabela 1 – Técnicas - Alunos

	PROF_T1_	ALUNO_T1_	PROF_T2_	ALUNO_T2_	PROF_T3_	ALUNO_T3_	PROF_T4_	ALUNO_T4_	PROF_T5_	ALUNO_T5_	PROF_T6_	ALUNO_T6_	PROF_T7_	ALUNO_T7_
S01	1.00	0.00	0.00	1.00	1.00	0.00	0.00	1.00	0.00	1.00	0.00	1.00	1.00	0.00
S02	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
S03	1.00	0.00	0.00	1.00	1.00	0.00	1.00	0.00	1.00	0.00	0.00	1.00	1.00	0.00
S04	0.00	0.00	0.00	1.00	1.00	0.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	0.00
S05	1.00	0.00	0.00	1.00	1.00	0.00	1.00	0.00	0.00	1.00	0.00	1.00	1.00	0.00
S06	0.00	0.00	0.00	1.00	1.00	0.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
S07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	1.00	0.00	0.00
S08	0.00	0.00	0.00	1.00	1.00	0.00	0.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00
S09	0.00	0.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	0.00
S10	0.00	0.00	0.00	1.00	1.00	0.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
S11	1.00	0.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	0.00	1.00	0.00
S12	0.00	0.00	0.00	1.00	1.00	0.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
S13	0.00	0.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
S14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00
S15	1.00	0.00	0.00	1.00	1.00	0.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
S16	1.00	0.00	0.00	1.00	1.00	0.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	0.00
S17	0.00	0.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00
S18	0.00	0.00	0.00	1.00	1.00	0.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
S19	1.00	0.00	0.00	1.00	0.00	1.00	1.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00
S20	1.00	0.00	0.00	1.00	1.00	0.00	1.00	0.00	1.00	0.00	0.00	1.00	0.00	1.00
S21	1.00	0.00	0.00	1.00	1.00	0.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	0.00
S22	0.00	0.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	0.00
S23	0.00	0.00	0.00	1.00	1.00	0.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
S24	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	0.00

Tabela 2 - Praxeologias

Como já foi explicitado anteriormente, quanto ao uso do quadro metodológico da ASI, para a análise dos dados coletados com o software Chic, utilizamos como *Tipo de Implicação* a **Implicação segundo ao método entrópico** e o *Tipo de Lei* foi a **Lei Binomial**. Para a elaboração dos gráficos implicativos utilizamos um nível de

implicação de no mínimo de 0,80, por considerarmos um nível alto para análise, e que nos daria *quase-implicações* com grau altamente significativos.

4.1 Análise das técnicas utilizadas pelos alunos – Tabela 1

Ao analisarmos os gráficos apresentados pelo CHIC, podemos perceber que há uma implicação no uso das mesmas técnicas para resolver determinados itens da Lista de Atividades. Como já foi informado, para a construção do gráfico, utilizamos um nível mínimo de 0,80; o que nos permite apresentar uma estrutura extremamente significativa do ponto de vista da estatística. As implicações com a cor azul até 0.95; a cor verde até 0,90; e a cinza até 0,80.

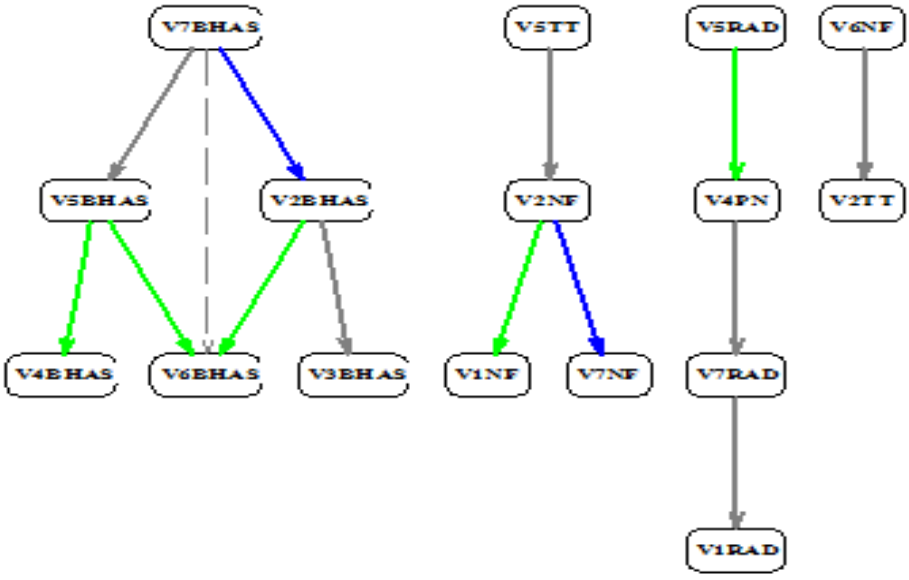


Gráfico 1 – Gráfico implicativo das técnicas x aluno

De um modo geral as variáveis de mesma técnica possuem uma forte implicação como podemos perceber no Gráfico 1, gerando o que chamaremos de classes apresentadas no quadro a seguir.

Classes	Técnicas Implicadas
1ª	V2BHAS – V3BHAS – V4BHAS – V5BHAS – V6BHAS – V7BHAS
2ª	V5TT – V1NF – V2NF – V7NF
3ª	V1RAD – V5RAD – V7RAD – V4PN
4ª	V6NF – V2TT

Quadro 3 – Classes geradas pelo Gráfico 1

No que denominamos de *classe 1*, percebemos uma forte implicação entre a técnicas BHAS. Esse fato reforça a hipótese que apresentamos na tese (BESSA DE MENEZES, 2010), em que os dados nos levam a crer que técnica de Bhaskara ocorra, principalmente, pelo status que a fórmula possui quanto à resolução de equações de 2º grau. Esse status dá a Bhaskara uma garantia na resolução dessas equações, o que coloca o aluno em conformidade com a Instituição Matemática, na qual, historicamente, é valorizado o resultado encontrado.

Ainda refletindo sobre a hipótese, a necessidade de conformidade, por parte dos alunos, com a Instituição Matemática, faz com que eles, apesar de as técnicas propostas pelo professor darem a impressão de serem "*mais simples*" de aplicá-las na resolução das tarefas, não percorram esse caminho, ou seja, nos dão indícios de que eles querem a "*segurança*" que Bhaskara fornece. Os resultados encontrados no tratamento dos dados no CHIC nos reforçam essa ideia.

Na *classe 2*, percebemos a implicação entre os alunos que não realizaram alguns itens (NF), esse fato já era esperado devido os itens apresentarem ou equações incompletas do 2º grau (Itens 1 e 2) ou em forma pouco usual como no item 7. A diferente apresentação pode ter levado aos alunos a não realizarem a tarefa.

Quanto à implicação no item 5 (transposição de termos – V5TT), com os alunos que não realizaram os itens 1, 2 e 7; podemos levantar a hipótese que isso ocorra devido o professor iniciar o conteúdo de equações do 2º grau, a partir dos conhecimentos dos alunos em equações do 1º grau, e no caso particular do item 5, não há o expoente 2 em nenhum valor de '*x*' [$(2x + 5)^2 = 0$]. Essa situação nos leva a supor que o aluno entendeu como uma equação de grau 1 e, por isso, se sentiu com possibilidades de resolvê-la, o que não ocorre nos itens 1, 2 e 7, em que se observa a presença do expoente 2 para a incógnita '*x*'.

A *classe 3* apresenta resoluções a partir de técnicas que foram apresentadas pelo professor em sala de aula, o que nos leva a identificar um alinhamento do que foi estabelecido como "*modelo de resolução*" pelo professor, pelos alunos dessa classe.

Quanto à *classe 4*, verificamos que ao desenvolver a equação do item 6, o aluno se depara com uma completa do 2º grau, já no item 2, ele encontra uma incompleta. Tal fato, nos leva ao indício de que os alunos dessa classe, não conseguem resolver uma equação completa, somente a incompleta.

O árvore de similaridade, gerada pelo CHIC, reforça a ideia de um agrupamento no uso de técnicas, como podemos perceber no gráfico 3 e no quadro de similaridade quadro 3). Nele, podemos perceber que os 'galhos' da árvore possuem uma tendência de ter as mesmas técnicas.

Tal fato nos reforça a ideia de que o aluno sempre busca estar em conformidade com a Instituição Matemática, pois utiliza de uma técnica que provavelmente tem confiança ou uma '*boa relação ao saber*'[14] para resolver as tarefas (itens) que são propostos pelo professor em sala de aula.

Essa relação é, em alguns casos, determinante para a escolha de técnicas para resolução de tarefas no dia a dia. Vejamos o gráfico a seguir.

[14] Entendemos como 'boa relação ao saber' a proximidade do indivíduo (I) ao objeto do saber (O). Quanto maior proximidade com o saber, melhor sua relação.

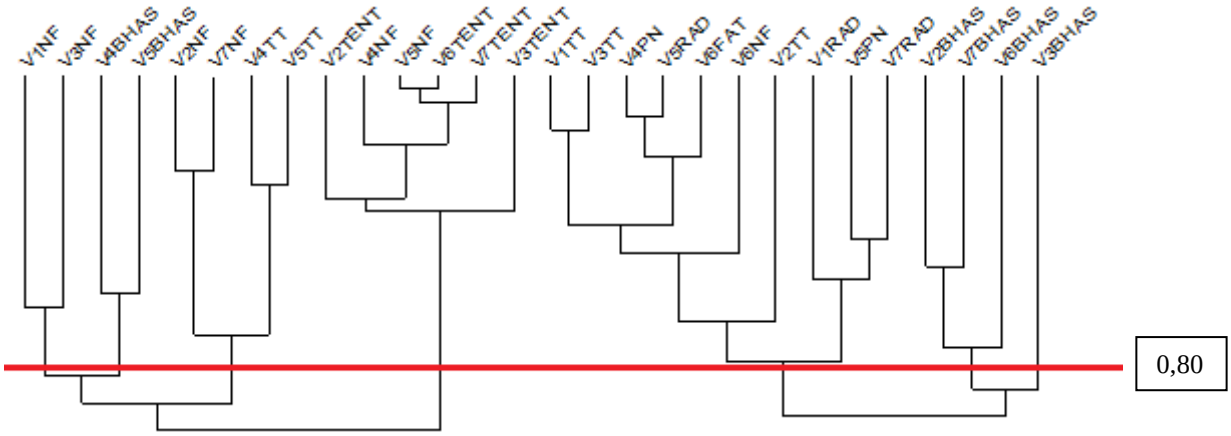


Gráfico 2 – Gráfico de similaridade das técnicas utilizadas pelos alunos

Classificação ao nível: 1 : (V5NF V6TENT) similaridade : 0.999999
Classificação ao nível: 2 : ((V5NF V6TENT) V7TENT) similaridade : 0.999999
Classificação ao nível: 3 : (V4PN V5RAD) similaridade : 0.999898
Classificação ao nível: 4 : (V1TT V3TT) similaridade : 0.999626
Classificação ao nível: 5 : (V4NF ((V5NF V6TENT) V7TENT)) similaridade : 0.998878
Classificação ao nível: 6 : ((V4PN V5RAD) V6FAT) similaridade : 0.993336
Classificação ao nível: 7 : (V2NF V7NF) similaridade : 0.993263
Classificação ao nível: 8 : (V4TT V5TT) similaridade : 0.98964
Classificação ao nível: 9 : (V2TENT (V4NF ((V5NF V6TENT) V7TENT))) similaridade : 0.986716
Classificação ao nível: 10 : ((V2TENT (V4NF ((V5NF V6TENT) V7TENT))) V3TENT) similaridade : 0.983423
Classificação ao nível: 11 : ((V1TT V3TT) ((V4PN V5RAD) V6FAT)) similaridade : 0.98014
Classificação ao nível: 12 : (V5PN V7RAD) similaridade : 0.979072
Classificação ao nível: 13 : (((V1TT V3TT) ((V4PN V5RAD) V6FAT)) V6NF) similaridade : 0.949261
Classificação ao nível: 14 : (V2BHAS V7BHAS) similaridade : 0.905895
Classificação ao nível: 15 : (V1RAD (V5PN V7RAD)) similaridade : 0.905169
Classificação ao nível: 16 : (V4BHAS V5BHAS) similaridade : 0.891594
Classificação ao nível: 17 : (V1NF V3NF) similaridade : 0.891259
Classificação ao nível: 18 : ((((V1TT V3TT) ((V4PN V5RAD) V6FAT)) V6NF) V2TT) similaridade : 0.880819
Classificação ao nível: 19 : ((V2NF V7NF) (V4TT V5TT)) similaridade : 0.870849
Classificação ao nível: 20 : ((V2BHAS V7BHAS) V6BHAS) similaridade : 0.819345
Classificação ao nível: 21 : (((((V1TT V3TT) ((V4PN V5RAD) V6FAT)) V6NF) V2TT) (V1RAD (V5PN V7RAD))) similaridade : 0.803561

Quadro 4 – Classificação ao nível de similaridade das técnicas utilizadas pelos alunos

Outro ponto de análise de nossa pesquisa envolve as praxeologias que pudemos observar na tese, que serve de base para essa pesquisa, explicitada na devolutiva dos alunos na resolução da Lista de Atividades. Foram identificadas as praxeologias do Professor e dos Alunos.

1.1 4.2 Análise das "*praxeologias*"[15] do Professor e do Aluno – Tabela 2

Ao analisarmos os gráficos gerados a partir do CHIC para a tabela 2, em que buscamos identificar se o aluno segue ou não a praxeologias do Professor, identificamos uma linearidade, ou seja, aqueles que optam por seguir a práxis do professor, tendem a acompanhá-lo na resolução dos itens. No entanto, aqueles que optam por um outro caminho (ou outra práxis), também, tendem a fazê-lo na resolução dos itens. Assim como trabalhamos na construção dos gráficos da Tabela 1, utilizamos um nível mínimo

[15] O que iremos observar na realidade são as técnicas que o professor utilizou e quais os alunos utilizaram, o que podem caracterizar uma diferença de praxeologias entre o professor e os alunos.

de 0,80; o que nos permite apresentar uma estrutura extremamente significativa do ponto de vista da estatística. As implicações com a cor azul até 0.95; a cor verde até 0,90; e a cinza até 0,80.

Vejamos o gráfico implicativo da tabela 1, a seguir.

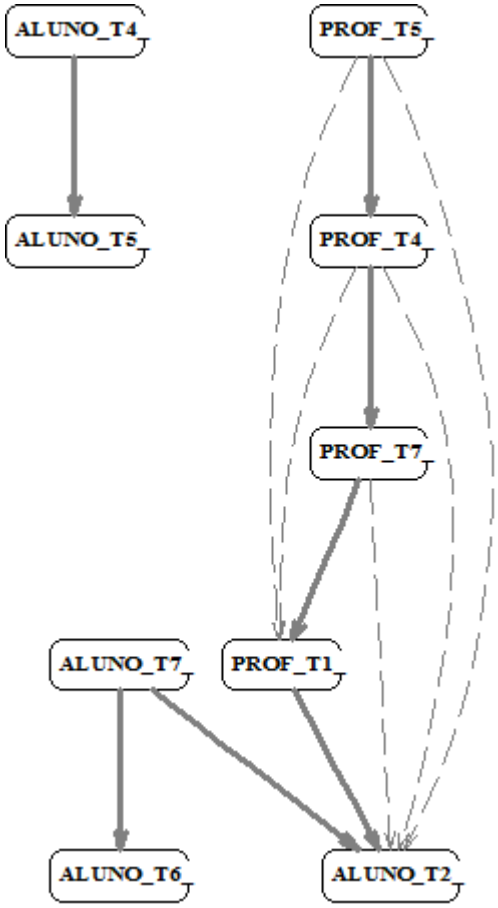


Gráfico 3 – Gráfico implicativo das praxeologias x aluno

De forma natural existe uma expectativa que o aluno siga a técnica de resolução do professor, pois existe uma regra de contrato didático implícito de que "*se fizer como o professor faz está certo*". O gráfico 3 nos dá uma confirmação dessa expectativa. Os itens 1, 4, 5 e 7, não estão apresentados na forma reduzida de uma equação de 2º grau completa ($ax^2 + bx + c = 0$), com isso, devido o professor ter iniciado esse conteúdo a partir de equações incompletas fazendo relação com o assunto trabalhado nas equações do 1º grau, acreditamos que os alunos se sentiram mais seguros em resolver com as técnicas apresentadas pelo professor. Um fato que reforça nossa hipótese é de que a maioria dos alunos encontra somente uma raiz na resolução dos itens em que as equações não estão completas, o que pode estar relacionado diretamente com um efeito da transposição do saber que foi realizada em sala de aula, um possível efeito de transposição.

A implicação que ocorre entre as técnicas do professor (PROF) nos itens 1, 4, 5 e 7 e a técnica ALUNO_T2, pode ser explicada a partir da escolha dos alunos pela transposição de termos (TT), que os aproxima das técnicas utilizadas para resolução de equações de grau 1; como também, pela escolha da fórmula de Bhaskara (BHAS), a

qual já identificamos na tabela 1 como tendo um status muito forte para resolução de equações de grau 2.

Podemos notar que as implicações entre as variáveis ALUNO_T4 e ALUNO_T5 ocorrem devido à escolha dos alunos em realizar as questões T4 $(x - 8).(x + 4) = 0$ e T5 $(2x + 5)^2 = 0$, pela fórmula de Bhaskara. Assim sendo, eles desenvolvem o produto entre os fatores e aplicam a fórmula. Novamente, pensamos na ideia de se estar adequado à instituição de ensino ou ao professor por parte do aluno.

O gráfico e a classificação de similaridade nos apontam para o mesmo caminho, ou seja, a uma regularidade nas escolhas das técnicas para a resolução dos itens. Considerando o índice de 0,80 (estatisticamente mais significativo), temos três classes (nível 1, 2 e 3) em todas temos um agrupamento segundo a similaridade das técnicas do professor com professor e do aluno com aluno.

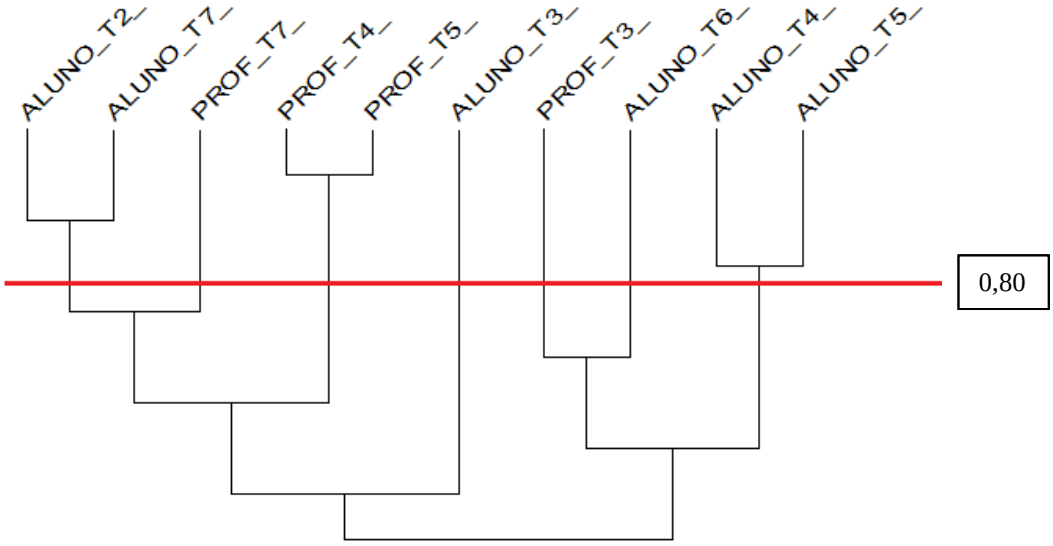


Gráfico 4 – Gráfico de Similaridade das Praxeologias

Classificação ao nível: 1 : (PROF_T4_ PROF_T5_) similaridade : 0.999898 Classificação ao nível: 2 : (ALUNO_T2_ ALUNO_T7_) similaridade : 0.898199 Classificação ao nível: 3 : (ALUNO_T4_ ALUNO_T5_) similaridade : 0.883579 Classificação ao nível: 4 : ((ALUNO_T2_ ALUNO_T7_) PROF_T7_) similaridade : 0.740698 Classificação ao nível: 5 : (PROF_T3_ ALUNO_T6_) similaridade : 0.707813

Quadro 4 - Classificação ao nível de similaridade das Praxeologias

Após a análise dos gráficos gerados a partir do CHIC, das tabelas 1 e 2, buscaremos fazer uma retomada da pesquisa e apontar algumas considerações que foram possíveis a partir do uso da A.S.I.

5 Considerações Finais

Nossa pesquisa tinha a como objetivo avançar em um ponto que não foi observado durante a realização da nossa tese que tratou das diferentes praxeologias apresentadas pelo professor e pelos alunos durante o ensino das equações de segundo grau, que seria a estabilidade na escolha de técnicas na resolução de tarefas propostas em sala de aula. Para isso, recorremos ao uso do quadro metodológico da A.S.I.

A partir da análise dos dados percebemos que há uma estabilidade nas opções dos alunos para resolverem tarefas propostas, seja para seguir as técnicas apresentadas pelo professor, seja para *caminhar com seus próprios passos*.

Quanto a *caminhar com seus próprios passos*, o gráfico de implicação e de similaridade da Tabela 1, gerados segundo o software CHIC, nos aponta para essa estabilidade, e nos permite reforçar algumas hipóteses que foram levantadas durante a tese, porém, não tínhamos elementos para afirmá-las. Pudemos perceber uma implicação significativa na escolha da técnica de Bhaskara (BHAS) para a resolução de atividades que poderiam ser solucionadas por técnicas mais simples com um número menor de passos para se chegar ao resultado. Esse fato pode estar ligado ao fato da busca pela conformidade.

Chevallard (1998) trata o aluno como um sujeito em cada instituição (família, escola, bairro, trabalho...) que ele participa em sua vida. Sujeito no sentido de se sujeitar, de estar adequado, de estar em conformidade, com essas instituições. A ação de se sujeitar à instituição poderá ser de forma indiferente ou ativa. Na forma indiferente, o sujeito irá se adequar às normas e regras da instituição sem contestá-las. No entanto, quando se sujeita de forma ativa, ele se transformará e será transformado pela instituição. O aluno, como sujeito de várias instituições, carrega em si elementos de cada uma dessas relações institucionais, as quais irão transformá-lo na pessoa que é. Com isso, a partir dessas relações, o aluno reconstrói o conhecimento para si, dando a ele características particulares, que ele traz das diversas instituições a que pertence.

Essa conformidade está presente, no momento em que os alunos buscam estratégias que os mantenham seguros quanto à resolução de um tipo de tarefa. Podemos pensar que os alunos não queriam errar, e simplesmente ter a certeza de que resolveriam a questão, e, com isso, se manterem em conformidade com as instituições escolares. O próprio livro didático, utilizado pelos alunos, valoriza a técnica de Bhaskara, ao citá-la como “Essa célebre fórmula aplica-se a todas as equações...”.

Quanto a *acompanhar as técnicas utilizadas pelo professor*, também remonta o que foi dito anteriormente sobre a conformidade. Foi possível perceber apoiado pelo gráfico de implicação e similaridade da Tabela 2. No entanto, sobre outro olhar: o do contrato didático que é estabelecido.

O professor, de certa maneira, impõe aos sujeitos (alunos) formas de fazer e de pensar, que são próprias a cada professor. Quanto a estas formas de fazer e de pensar, muitos pesquisadores já avançaram a partir do momento em que perceberam que o professor “cria” um metatexto do saber a ser ensinado em sala de aula (BESSA DE MENEZES, 2004; CÂMARA DOS SANTOS, 2002; BRITO DE MENEZES, 2006). Com isso, podemos, sim, entender o professor como uma Instituição durante sua atividade docente. Ou, na pior das hipóteses, adquirir status de Instituição durante sua atividade em sala de aula. Com isso, o aluno tende a repetir as técnicas que são utilizadas pelo professor, e, assim, garantem que não errarão, assegurando a sua conformidade com a instituição escolar ou a Instituição Professor.

A nossa pesquisa gera uma reflexão sobre a prática docente. O professor deverá estar atento à busca pela conformidade institucional, por parte do discente, a partir de escolhas das técnicas de resolução de tarefas, seja por influência do professor ou por caminhos próprios. A conformidade pode não representar um aprendizado significativo.

Referências

[1] Almouloud, S., Gras. R. & Régnier, J-C. (2014), *Especial ASI*. Revista Educação Matemática Pesquisa, Programa de Estudos Pós-graduados em Educação Matemática – PUCSP. São Paulo, v.16, n.3, pp.623-1087.

[2] Arsac, G.; Develay. M. & Tiberghien, A. (1989), *La transposition didactique en mathématiques, en physique, en biologie*. Lyon: IREM.

[3] Bessa de Menezes, M. (2004), *Investigando o processo de transposição didática interna: o caso dos quadriláteros*. 184f. Dissertação (Mestrado em Educação) – Universidade Federal de Pernambuco, Recife.

[4] Bessa de Menezes, M. (2010), *Praxeologia do Professor e do Aluno: Um análise das diferenças no ensino de Equações do Segundo Grau*. 177f. Tese (Doutorado em Educação) – Universidade Federal de Pernambuco, Recife.

[5] Booth, L. (1995), Dificuldades das crianças que se iniciam em álgebra. Em: Coxford, A. & Shulte, A (Orgs.), 1995. *As Idéias da Álgebra*. São Paulo, SP: Atual Editora. p. 23-37.

[6] Brito Menezes, A.P. (2006), *Contrato Didático e Transposição Didática: Inter-relações entre os fenômenos didáticos na iniciação à álgebra na 6ª série do Ensino Fundamental*. 410f. Tese (Doutorado em Educação). Universidade Federal de Pernambuco, Recife.

[7] Câmara dos Santos, M. (2002), Algumas concepções sobre o ensino-aprendizagem de matemática. In: *Educação Matemática em Revista*. nº 12. São Paulo: SBEM.

[8] Chevallard, Y. (1991), *La transposition didactique.* Grenoble: La pensée Sauvage.

[9] Chevallard, Y. (1998) Analyse des pratiques enseignantes et didactique des mathématiques: l'approche anthropologique. In : *L'UNIVERSITE D'ETE*, p.91-118. Actes de l'Université d'été La Rochelle, IREM, Clermont-Ferrand, France, 1998.

[10] Chevallard, Y. (1999), L'analyse des pratiques enseignantes en Théorie Anthropologie Didactique. In: *Recherches en Didactiques des Mathématiques*, p. 221-266.

[11] Coll, C. et All. (2006), *O construtivismo na sala de aula*. São Paulo: Ática.

[12] Gras R., Briand H., Peter P., Philippe J. (1997) : *Implicative statistical analysis, Proceedings of International Congress I.F.C.S.*, 96, Kobe, Springer-Verlag, Tokyo.

[13] Gras R., Régnier J.-C., Guillet F. (Eds) (2009) *Analyse Statistique Implicative. Une méthode d'analyse de données pour la recherche de causalités*. RNTI-E-16 Toulouse: Cépaduès.

[14] Gras R., Régnier J.-C., Marinica C., Guillet F. (Eds) (2013) *Analyse Statistique Implicative. Méthode exploratoire et confirmatoire à la recherche de causalités.* Toulouse: Cépaduès.

[15] Kieran, C. (1992), *The learning and teaching of school algebra. Handbook of research on mathematics teaching and learning*. National Council of Teachers of Mathematics; New York. p. 390-419.

[16] Kury, M. G. (2002), *Minidicionário Gama **Kury** da Língua Portuguesa*. 1ª Edição. São Paulo: FTD.

[17] Pais, L. C. (2001) *Didática da Matemática: Uma Análise da Influência Francesa*. Belo Horizonte: Autêntica.

[18] Ravel, L. (2003) *Des programmes a la classse: etude de la transposition didactique interne.* Tese de Doutorado não-publicada. Université Joseph Fourier – Grenoble I.

[19] Usiskin, Z. (1995), Concepções sobre a álgebra da escola média e utilizações das variáveis. In: Coxford, A. & Shulte, A (Orgs.), 1995. *As Ideias da Álgebra*. São Paulo, SP: Atual Editora. p. 9-22.

[20] Vergnaud, G., Cortes, A. (1986), *Introducing Algebra to "Low-level" Eighth and Nineth graders*. Proceedings of the Xth International Conference of Psychology of Mathematics Education. London. p. 319-324.

[21] Vergnaud, G., Cortes, A. & Favre-Artigue, P. (1987), Introduction de l'algèbre auprès de débutants faibles: problèmes épistémologiques et didactiques. In: Vergnaud, G., Brousseau, G. & Hulin, M. (Orgs.) 1987. *Didatique et aquisition des connaissances scientifiques: Actes du Colloque de Sèvres*. Sèvres, La Pensée Sauvage. p. 259-280.