UFR SLHS : Sciences du Langage de l'Homme et de la Société

École doctorale Langages, Espaces, Temps, Société

Extraire et valider les relations complexes en sciences humaines :

statistiques, motifs et règles d'association

THÈSE

présentée et soutenue publiquement le 12 décembre 2006

pour l'obtention du

Doctorat de l'université de Franche-Comté – Besançon

(spécialité informatique)

par

Martine CADOT

sous la direction du Professeur Alain LELU

Membres du jury

Michel ARMATTE, Maître de Conférences à l'université de Paris-Dauphine
Alexandre AUSSEM, Professeur à l'université de Lyon I
Alain LELU, Professeur à l'université de Franche-Comté
Engelbert Mephu NGUIFO, , Professeur à l'université d'Artois
Gilbert RITSCHARD, Professeur à l'université de Genève, Rapporteur
Djamel A. ZIGHED, Professeur à l'université de Lyon II, Rapporteur

LAboratoire de SÉmio-Linguistique, Didactique et Informatique — EA 2281

Mis en page avec la classe thloria.

Remerciements

Ce travail n'aurait pu avoir lieu sans l'aide de nombreuses personnes que je tiens à remercier ici.

D'abord, à titre posthume, mes ascendants : ma grand-mère qui devant mes yeux émerveillés de petite fille trouvait la solution de mes exercices de mathématiques par de mystérieux algorithmes de calcul datant d'avant la guerre 14-18 ; et mes parents qui ont fait en sorte que je puisse aller dans la voie tracée par ma grand-mère.

Les personnes qui m'ont aidée à toutes les étapes de mon cheminement dans la recherche, par leurs discussions, leurs conseils, leur soutien logistique, ou tout simplement leur écoute, que ce soit à l'Irem de Dijon, à l'Iredu, à l'université de Bourgogne, ou plus récemment, au LPA, à l'université de Reims, au Loria, à l'université de Nancy, au Laseldi de l'université de Franche-Comté, sont trop nombreuses pour être citées ici. Que celles qui se reconnaissent soient remerciées ! Ainsi que les autres.

Mes remerciements vont plus particulièrement aux chercheurs avec lesquels j'ai eu les échanges les plus fructueux pour la problématique de cette thèse, Nicole Battaglia, Viviane Nahama, André Masson, Claire François, Xavier Polanco et Pascal Cuxac.

Et mes plus vifs remerciements vont à ceux qui m'ont aidée au moment où j'en avais le plus besoin et m'ont fait confiance : Yves Laprie, qui m'a fait une place dans son équipe de recherche au Loria, et Alain Lelu, qui a pris à coeur son rôle de directeur de thèse et fait mûrir ce travail ; sans oublier Lina, Tarek, Rayan et Milo, mes enfants et petits-enfants, pour leur soutien chaleureux et inconditionnel.

0

Table des matières

Table des figures

Résumé

Cette thèse concerne la "fouille de données" en sciences humaines. La fouille de données (en anglais : *Data Mining*) connaît un essor grandissant depuis son apparition il y a une vingtaine d'années. Cette branche de l'intelligence artificielle consiste en un ensemble de méthodes visant l'extractraction de connaissance (en anglais : *Knowledge Discovery*) à partir de données stockées sur des supports informatiques, données qui peuvent être complexes et/ou volumineuses. Pour transformer les données en connaissance, on procède en plusieurs étapes :

1. codage des données
2. découverte des relations existant dans les données
3. interprétation des relations trouvées

Ces mêmes étapes sont celles que suivent les chercheurs en sciences humaines lorsqu'ils utilisent la méthodologie statistique. Depuis l'époque des pionniers de la statistique, il y a une centaine d'années, de nombreuses difficultés tant théoriques que pratiques du traitement des données ont été rencontrées et identifiées, et sont à l'origine de son développement dans de multiples directions.

Nous confrontons d'abord les deux méthodologies que sont la statistique et la fouille de données dans le but de répondre aux questions suivantes :

1. A quels besoins répond l'utilisation de la méthodologie statistique chez les chercheurs en sciences humaines ?
2. Quelles limitations rencontrent-ils lors de son utilisation sur les données stockées actuellement disponibles ?
3. Quelles méthodes en fouille de données pourraient-il utiliser ?
4. Comment peut-on les améliorer pour qu'elles répondent à leurs besoins ?

La conclusion que l'on tire de cette première partie est que les chercheurs en sciences humaines désirent établir, avec un degré de certitude quantifiable, des relations entre les données pour construire une représentation symbolique des données, c'est-à-dire permettant de raisonner sur les phénomènes à l'origine des données. Les méthodes statistiques habituelles leur permettent de le faire tant que les caractéristiques des données peuvent être traduites par des variables constituant une structure de taille raisonnable (moins de 100 variables) et relativement simple (interactions entre moins de 10 variables, voire pas d'interactions du tout, linéarité des effets, répartitions normales, de Poisson, multinomiales, ...), mais s'avèrent insuffisantes dès que la structure des données se complexifie. Parmi les méthodes de fouille de données actuelles, l'extraction de motifs et de règles d'association est la méthode de traitement des données à base de comptage de cooccurrences de leurs caractéristiques qui permet de représenter au mieux une structure complexe de données de façon symbolique. Toutefois ce modèle informatique des données n'est pas directement utilisable par les chercheurs en sciences humaines en son état actuel : il est

essentiellement dédié aux données dichotomiques (un objet possède ou non une propriété), ses résultats directs, très morcelés, sont difficiles à interpréter, et sa validité peut paraître douteuse aux chercheurs habitués à la démarche statistique. Actuellement, des recherches visant à réduire ces difficultés se font dans la communauté de fouille de données. C'est dans cet axe de recherche que s'inscrivent nos propositions.

Voici nos propositions :

1. Nous développons une technique statistique pour établir qu'une règle d'association A->B, extraite des données et concernant l'association de deux de leurs caractéristiques (ou propriétés) A et B, est "significative", c'est-à-dire ne peut pas être "due au hasard". Le test statistique que nous construisons, à base de simulations créant des matrices aléatoires de mêmes sommes en ligne et colonne que la matrice d'origine, est très robuste : il peut s'appliquer à toutes les données susceptibles de se représenter par une matrice de type sujets x propriétés, formée de zéros (le sujet ne vérifie pas la propriété) et de uns (il la vérifie), quelle que soit leur distribution de valeurs et il prend en compte les propriétés dans leur ensemble, quel que soit leur nombre, contrairement aux méthodes statistiques classiques. Ce test est susceptible d'éclairer à terme de nombreux problèmes de l'analyse des données en sciences humaines : sélection des variables, significativité des liens entre variables ou des similarités entre observations, construction de nouvelles variables susceptibles d'incarner les phénomènes d'interaction et d'effets non-linéaires, etc.
2. Nous étendons la méthode d'extraction de motifs et de règles d'association à des données numériques en "fuzzifiant", en rendant floues, toutes les étapes de cette méthode. Les co-occurrences sont remplacées par des covariations de telle façon que les règles d'association floues extraites des données coïncident avec les règles d'association strictes quand les données sont dichotomiques. Nos règles d'association se trouvent ainsi coïncider avec des règles floues définies par ailleurs par des chercheurs continuant les travaux initiés par Zadeh.
3. Nous créons des méta-règles pour nettoyer le jeu de règles d'association de ses principales contradictions et redondances. Les incohérences qu'il contient ne sont pas dues à des erreurs d'extraction mais à l'enchevêtrement des relations entre les caractéristiques des données, et ne peuvent se dénouer qu'une fois choisi un niveau d'interprétation (par exemple local ou global). Nos méta-règles contiennent des paramètres que l'utilisateur peut fixer selon le niveau d'interprétation qu'il désire privilégier afin de faire du jeu de règles nettoyé une représentation symbolique des données qui lui convienne. A terme, notre démarche conduit à ne présenter à l'utilisateur, ou aux traitements suivants qu'il aura mis en place, qu'un jeu de motifs ou règles minimal, valide statistiquement et débarrassé de toute redondance et incohérence.

A cela s'ajoute un début d'unification de ces propositions séparées au sein d'une chaîne de traitement dont l'élément moteur est l'algorithme MIDOVA. Cet algorithme et la chaîne de traitement sont en cours d'amélioration. Une première version d'essai en est décrite dans la dernière partie de cette thèse, ainsi que ses résultats prometteurs sur un jeu de données réelles.

Mots-clés : fouille de données, fouille de textes, apprentissage artificiel, motifs, règles d'association, motifs flous, règles floues, interaction statistique, significativité statistique, test de randomisation, nettoyage et prétraitement des données.

Keywords : Data Mining, Text Mining, Machine Learning, Itemsets, Association Rules, Fuzzy Itemsets, Fuzzy rules, Statistical Interaction, Statistical Significance, Randomisation Test, Data Cleaning and Preprocessing.

Introduction

1 Quelle proposition pour quel besoin ?

But et cadre de la thèse

Construire de la connaissance à partir d'observations est ce que nous faisons chaque jour. C'est une activité naturelle, mais complexe, qui combine les nouvelles observations aux connaissances déjà acquises de telle façon que l'ensemble forme un système suffisamment bien structuré pour qu'on puisse oublier les observations et ne garder que la connaissance à laquelle elles ont contribué. La qualité de la connaissance s'apprécie au moment où on l'utilise. Si elle produit les effets attendus, elle est validée, et dans le cas contraire, elle est réajustée. La *fouille de données* (appelée plus noblement ECD : *Extraction de Connaissances à partir de Donnée*s) ou *Data Mining* en anglais (mais aussi KDD : *Knowledge Discovery in Databases*) est l'automatisation de ce processus. En cela, elle fait partie de l'*apprentissage artificiel*, en anglais ML : *Machine Learning*, et plus généralement de l'*intelligence artificielle* (IA et en anglais AI) cette dernière faisant partie de l'informatique.

Cette thèse s'inscrit dans le cadre de la fouille de données. L'utopie ultime, dans cette perspective, serait de créer une méthode automatique transformant un ensemble de données de sciences humaines en un modèle causal symbolique. Avec un tel modèle les chercheurs en sciences humaines pourraient faire des raisonnements en utilisant l'"essence" de leurs données et non en effectuant les manipulations "de bas étage" des données elles-mêmes - ce qui ne les dispenserait pas du travail intelligent de sélection et construction de ces données en fonction de leur problématique. Bien sûr, le travail de recherche relaté dans ce mémoire reste très loin de ce but ambitieux, mais essaie plus modestement de poser quelques jalons sur le chemin y menant.

Les anciennes habitudes des chercheurs en sciences humaines

On peut trouver prétentieuse une telle visée. Les chercheurs en sciences humaines savent mieux que personne construire des modèles dans leur domaine de recherche. Ils utilisent depuis longtemps des méthodes statistiques, créées dans la plupart des cas par des chercheurs de leur propre communauté. Ces méthodes ont pour noms : bibliométrie, économétrie, psychométrie, sociométrie, écologie numérique, etc. selon la branche des sciences qu'elles essaient d'éclairer. En suivant la démarche de recherche apprise dans sa discipline, le chercheur est capable de valider les modèles symboliques dont il a besoin. Cette démarche est rigoureuse, et peut même paraître rigide, mais elle protège le chercheur des erreurs de raisonnement statistique les plus courantes dans son domaine. Elle définit le type d'observations à faire, le modèle statistique à utiliser, la façon de l'utiliser : selon les cas, le chercheur peut faire des "interviews", monter des expériences, préparer des questionnaires, utiliser des instruments divers de mesure, etc. pour obtenir les données, puis les données recueillies sont codées selon le modèle statistique approprié

aux hypothèses qu'il cherche à prouver, et une fois les calculs faits, le chercheur interprète les résultats obtenus.

Les nouveaux besoins des chercheurs en sciences humaines

L'avènement de l'informatique a bouleversé les habitudes de ces chercheurs. D'abord l'utilisation des logiciels de statistiques s'est généralisée aux laboratoires de sciences humaines depuis une dizaine d'années, ce qui a permis aux chercheurs de ne plus se préoccuper des calculs statistiques, et donc de tenter une modélisation plus perfectionnée. Puis la mise à disposition plus récente de données sur Internet, que ce soit par d'autres chercheurs ou par des organismes officiels, a rendu possible une collecte des données beaucoup plus rapide et moins coûteuse qu'auparavant. En psychologie par exemple, les chercheurs qui peinaient à trouver suffisamment de sujets voulant bien répondre à leur questionnaire, ou servir de "cobaye" dans une expérience, sont maintenant face à des données de tous horizons. Certains préféreront continuer à l'identique les démarches de leurs prédécesseurs, et construire des données plus susceptibles de répondre à leur questionnement, mais pour les autres la fouille de données peut s'avérer une méthodologie fructueuse de recherche en psychologie. Plus généralement Internet et les nouvelles technologies de la communication mettent à disposition des chercheurs en sciences humaines beaucoup de nouveaux territoires d'observation : on peut explorer les réseaux sociaux par le biais des messageries et forums, consulter en ligne les rapports d'entreprises, les banques de données scientifiques, explorer la "mémoire" d'une entreprise, d'une institution, d'un projet, etc. Cette disponibilité de données variées ne peut que susciter de nouvelles problématiques en sciences humaines.

Les anciennes méthodes statistiques ne sont plus adaptées

Les méthodes statistiques précédentes qui avaient été mises au point pour des possibilités d'observation plus restreintes (enquêtes, expériences, etc.) ne sont plus à la mesure des nouvelles possibilités pour de nombreuses raisons :

- Les données collectées sur Internet ont un niveau de complexité bien supérieur à celui qui était pris en compte précédemment. Par exemple, en éducation, l'imbrication en partie hiérarchique des différents niveaux d'influence sur la réussite d'un élève (cours, classe, cursus, famille) ne rendait pas aisé le choix d'un modèle statistique. Avec les pages Web, comportant à côté du texte des animations, hyperliens, photos, etc. on monte d'un degré dans la complexité.
- Le nombre d'objets envisagé dans les manuels statistiques est bien modeste à l'échelle d'Internet : concernant les tests d'hypothèses, on y lit qu'on peut faire des approximations quand on dispose d'un "grand" échantillon, c'est-à-dire dont la taille dépasse 30, nombre qui parait bien petit à l'échelle du Web ! Et il est souvent conseillé au lecteur d'augmenter la taille de ses échantillons (c'est-à-dire du nombre d'objets) pour obtenir un test de meilleure qualité, par exemple pour comparer les moyennes d'une propriété (ou variable) mesurée sur deux groupes d'objets. Mais une petite différence entre les deux moyennes a tendance à devenir significative dès que la taille dépasse un certain seuil.
- Les statistiques permettent de travailler avec un nombre de variables qui dépasse rarement la centaine, en pratique, en cas d'analyse exploratoire comme l'*analyse en composantes principales* (ACP), la vingtaine en cas de *pistes causales* et la demi-douzaine dans la *régression linéaire*, alors que le nombre de variables pouvant être considérées dans les données accessibles sur Internet [1] est impressionnant. De plus, il peut dépasser le nombre

[1] L'analyse d'une simple page web peut requérir un grand nombre de descripteurs.

d'objets, ce qui n'est pas envisageable en statistique classique.
- Parmi les relations entre variables considérées par un modèle statistique, les plus complexes sont l'interaction linéaire entre deux, trois voire quatre variables, ou des relations non linéaires entre quelques variables. Dans les données accessibles sur Internet l'enchevêtrement des relations entre les nombreuses variables ne peut pas toujours se démêler aussi simplement.
- La plupart des tests statistiques ont des conditions d'application qui ne sont pas toujours respectées par les données d'Internet (normalité des lois, ou lois courantes en statistique paramétrique, faible nombre d'ex-aequo en cas de tests de rangs, etc.)

Quelle solution existe-t-il actuellement ?

L'avènement de l'informatique a également permis la réactualisation d'anciennes méthodes reléguées à cause des nombreux calculs qu'elles nécessitaient, comme le *test exact de Fisher*, qui est à l'origine des *tests de permutation*. Et de nouvelles méthodes de traitement des données se démarquant des statistiques sont apparues dans le champ de la fouille de données, s'appuyant sur la puissance de calcul des ordinateurs ou sur un mode de raisonnement plus informatique que statistique. On a vu que le chercheur en sciences humaines suivait une démarche rigoureuse et contraignante pour assurer par les statistiques la validité de ses résultats. Avec les nouvelles méthodes de la fouille de données, cette assurance attachée à l'usage du formalisme statistique disparaît. Il faut la reconstruire en dehors des statistiques habituelles, afin que le chercheur en sciences humaines puisse utiliser les nouvelles méthodes de fouille de données à la place des anciennes sans prendre le risque d'être désavoué par sa communauté.

Parmi les méthodes actuelles de fouille de données l'*extraction des motifs et des règles d'association* nous a parue la plus apte à donner une représentation des liaisons complexes entre variables car les motifs et les règles d'association permettent de décrire de façon plus souple que les statistiques un réseau de variables. Et notre travail de thèse a consisté à essayer de la rendre plus rigoureuse selon les exigences des chercheurs en sciences humaines.

Avantages et inconvénients des règles d'association

Les motifs sont les groupes de variables présentant de fortes associations, et les liens orientés au sein de ces motifs entre deux sous-groupes de variables sont exprimés par des règles d'association.

Cette méthode de fouille de données a de nombreux avantages :
- avec les motifs sont extraits dans le même processus les objets concernés,
- les motifs réunissant un grand nombre de variables ne sont pas "oubliés" : l'utilisateur n'a pas à spécifier la taille maximale des motifs, mais seulement leur *support* minimal (le nombre minimum d'objets devant les vérifier),
- cette méthode prévue pour décrire de grosses bases de données fonctionne également avec très peu de données.

Ses inconvénients sont nombreux également :
- les variables doivent être dichotomiques [2]
- la quantité de motifs et règles d'association extraits dépend du seuil de support choisi par l'utilisateur. Si ce seuil est trop grand, seuls les liens "évidents" sont exprimés, et s'il est

[2] Si la variable est un mot, sa valeur pour un texte est de 1 ou 0 selon qu'il contient ou non ce mot. Même type de relation d'*incidence* entre les articles d'un supermarché et les tickets de caisse dans lesquels ils peuvent figurer ou non, les symptômes de maladie et les patients d'un service médical qui peuvent les présenter ou non, etc.

trop petit, le nombre de liens exprimés est tel qu'on ne peut en étudier qu'une infime partie,
- l'interprétation des règles d'association est délicate :
  - il est difficile de faire un choix parmi la cinquantaine d'indices exprimant la qualité des règles. Mais c'est obligatoire pour limiter l'interprétation aux meilleure règles.
  - la relation entre plus de 2 variables n'est pas très claire,
  - les règles prises dans leur ensemble peuvent s'avérer redondantes, incohérentes, voire contradictoires.

Les apports de cette thèse

Nous avons choisi cette méthode pour ses avantages, mais il convient de corriger ses défauts. Voici dans quelles directions nous avons travaillé pour essayer de rajouter à cette méthode de fouille de données la rigueur exigée par la démarche scientifique en sciences humaines.

Nous avons introduit des statistiques pour évaluer une règle entre 2 variables tout en prenant en compte l'ensemble des variables, au moyen d'une procédure de validation utilisant des versions permutées du tableau initial. Ce processus de décision globale échappe ainsi aux problèmes posés par les *comparaisons multiples*, les *hypothèses composites* et permet d'assurer la significativité d'une règle entre 2 variables compte tenu des autres variables (simulation orientée contexte) indépendamment des lois de probabilités. Cela a pour effet un nettoyage des règles sur la base de leur pouvoir de généralisation, donc plus adapté aux données qu'un nettoyage fait en prenant le même seuil (par ex. 0.8 de confiance) pour toutes. Le nombre de variables n'est pas limité. Notre procédure, publiée en 2005, de randomisation par échanges rectangulaires pour [in]valider les motifs extraits paraît plus satisfaisante que les approches antérieures.

Nous proposons une version floue de l'extraction des motifs et règles d'association qui permet de la généraliser aux variables quantitatives, et d'éviter ainsi la perte par dichotomisation d'une éventuelle liaison graduelle fine (croissante ou décroissante) entre certaines d'entre elles.

Enfin nous proposons un principe de nettoyage, par des *méta-règles* [3], des incohérences du jeu de règles d'association extrait. Ce nettoyage des incohérences des règles d'association suit la logique du formalisme des SGBD (Système de Gestion de Bases de Données) qui assure la cohérence des règles de stockage des données. Ces méta-règles doivent être pilotées par l'utilisateur qui choisit alors la façon de dénouer certaines liaisons complexes gênant la cohérence tout en gardant une trace de ses choix.

L'interface de pilotage de ce nettoyage n'a pas encore été réalisée. Pas plus que n'a été terminée la dernière mise au point de l'algorithme MIDOVA de repérage des liaisons complexes que sont les interactions. Un premier effet appréciable de cet algorithme est sa capacité à n'extraire à chaque niveau que les motifs exprimant un différentiel de liaison par rapport aux sous-motifs les constituant. Son utilisation réduit de façon drastique le nombre de motifs et permet, en étendant la définition des motifs à ceux de support nul, de découvrir toutes les interactions non négligeables, qu'elles soient "positives" ou "négatives" (comme le XOR du calcul booléen [186]). Une chaîne de traitement associant les éléments de ce mémoire est en cours de développement, et les premiers résultats prometteurs de son passage sur un petit corpus de textes sont décrits dans la partie "bilan" de ce mémoire. Il nous reste encore à faire de nombreuses mises au point et expérimentations pour que ce prototype devienne un véritable logiciel de fouille de données destiné aux chercheurs en sciences humaines.

[3] Nos méta-règles sont formées chacune de deux éléments : d'un côté la spécification du lien à repérer entre un certain nombre de règles et des diverses corrections possibles (en supprimant une partie des règles), et de l'autre côté le processus parcourant le jeu de règles pour détecter ces liens et les corrigeant selon la façon choisie par l'utilisateur.

2 D'où vient la problématique de cette thèse ?

Il est certainement utile de signaler au lecteur les étapes de ma trajectoire personnelle qui ont inspiré la problématique de cette thèse, car c'est à travers des contacts quotidiens et concrets avec plusieurs milieux de la recherche en sciences humaines, en mathématiques appliquées et en informatique, qu'elle a puisé ses motivations et pris sa forme définitive. C'est en premier lieu lors de ma collaboration à des recherches en didactique au sein de l'IREM[4] / Dijon que j'ai rencontré la problématique de la causalité dans le cadre de l'apprentissage (humain !) et de l'évaluation des acquis. Le séminaire de Régis Gras m'a alors fourni les bases indispensables pour m'intéresser au problème de l'implication statistique. Mon travail de DEA[5] d'analyse politique et économique de l'université de Dijon, dans le cadre de l'Institut de Recherche en Économie de l'Éducation (IREDU) - de sujet "Modélisation mathématique des cohortes scolaires", à travers la comparaison des réussites et échecs de deux établissements scolaires de ZEP[6] - m'a mise en contact avec la rigueur statistique et les contraintes complexes du modèle linéaire utilisé par les économètres ; en particulier je salue ici Pietro Balestra qui m'a orientée vers les tests de permutation, Bootstrap et Jackknife, quand les données sortent du cadre contraignant des lois statistiques habituelles. Le travail effectué lors de la réalisation de mon mémoire de DEA m'a aussi mise en contact avec les réalités ingrates du recueil et du recodage des données, de la fabrication et de la publication d'indicateurs... Ensuite dans mon travail au laboratoire de psychologie de l'université de Reims, j'ai découvert la problématique des petits échantillons, quand le statisticien est obligé de "faire avec ce qu'il a" du fait de contraintes drastiques sur le recueil des données : les mémoires de mes étudiants pouvaient porter sur l'observation de 10 bébés, ou de 6 autistes... J'ai découvert et pratiqué à cette occasion l'analyse factorielle classique "à l'anglo-saxonne", sur laquelle j'ai testé la stabilité et validé les axes par Bootstrap. J'ai aussi découvert la modélisation de l'interaction - hors de portée des analyses factorielles - et de la causalité par le logiciel LISREL[135], mais aussi les limites de son usage : nécessité d'échantillons importants, nécessité de construire un modèle a priori, à raffiner ensuite par itérations successives, quitte à bousculer quelque peu l'orthodoxie statistique. La comparaison avec les travaux de Régis Gras s'imposait : celui-ci construisait des enchaînements asymétriques de causes et d'effets entre variables manifestes, à partir des règles d'association tirées directement des données, sans modèle a priori, mais sans construire de variables latentes, et avec des limitations importantes sur les données traitées : pas plus d'une quarantaine d'individus et d'une quinzaine de variables. Enfin mon arrivée au LORIA[7] à Nancy m'a permis d'élargir mes perspectives et de développer mes idées, mon but étant de parvenir à des résultats voisins de LISREL - prise en compte de l'interaction, construction de concepts en sciences humaines et de réseaux de causalité - mais sans modèle a priori, à partir de la seule extraction de motifs et de règles d'association, pour exploiter les nouveaux gisements de données considérables à disposition des sciences humaines. Ceci n'était possible qu'à la condition d'élaguer les trop nombreuses règles extraites sous le double point de vue de la validation statistique et de la suppression des incohérences et redondances. Ma participation à plusieurs défis nationaux et internationaux en extraction de connaissance et fouille de textes [8] a été l'occasion de tester la solidité de mes conceptions et activités de réalisation dans le domaine - connexe - de l'apprentissage automatique.

[4] IREM : Institut de Recherche en Enseignement des Mathématiques.

[5] DEA : Diplôme d'Etudes Approfondies (bac+5) de l'Education Nationale.

[6] ZEP : Zône d'Education Prioritaire de l'Education Nationale.

[7] LORIA : Laboratoire LOrrain de Recherche en Informatique et ses Applications

[8] Les informations sur les défis DEFT, KDDCup, TREC se trouvent aisément sur Internet. Nous avons détaillé notre travail pour deux d'entre eux dans [40, 160].

3 Plan du mémoire

Le mémoire est formé de quatre parties suivies d'annexes. Cette thèse ayant comme but de développer une méthodologie de recherche de liaisons complexes dans les données pouvant s'automatiser et être appliquée par des chercheurs en sciences humaines, la première partie fait un état de l'art des différents domaines impliqués, ce qui explique sa longueur. Puis les deux parties suivantes sont un exposé des constructions théoriques qui ont été réalisées pour avancer dans cette voie à partir d'une méthode existant déjà en informatique : la partie II expose des constructions statistiques permettant de renforcer la robustesse de cette méthode, la partie III expose des techniques permettant d'affiner sa sémantique. La quatrième partie concerne le développement et la validation de cette méthodologie et contient les conclusions. Les annexes regroupent des preuves un peu longues auxquelles renvoient certaines sections. Les deux dernières annexes peuvent être lues séparément de ces sections.

Contenu de la première partie

La partie I est un état de l'art formé de quatre chapitres. Nous nous penchons dans le premier chapitre sur le codage des données en variables qui représente le premier travail d'un processus de fouille de données. C'est celui qui demande le plus de temps et d'énergie (en général 60% du processus de fouille d'après Han et al. [113]) et sa qualité et les choix qui y sont faits sont déterminants pour la suite du processus. Cette étape est encore plus importante pour la méthode de fouille de données que nous avons choisie. L'extraction des motifs et des règles d'association se faisant avec des variables dichotomiques, cela nous oblige à transformer les données en les modifiant de façon plus ou moins importante selon leur nature. Par exemple deux variables quantitatives recodées automatiquement en un certain nombre de variables dichotomiques peuvent ne pas produire d'associations alors qu'elles sont liées par une liaison graduelle croissante ou décroissante. Notre proposition de "fuzzifier" le processus d'extraction de motifs et de règles d'association est issue de ce constat.

Puis nous examinons les ingrédients qui font la rigueur de la démarche en sciences humaines. Les statistiques y ont une place fondamentale. La première raison est qu'elles proposent diverses modélisations des liaisons entre variables s'appuyant sur leur nature. Par exemple, la conception d'un plan d'expérience permet de décomposer les effets de plusieurs variables sur une autre en tenant compte de leur rôles respectifs (variables contrôlées, emboîtements, interactions, etc.). La deuxième raison de la place centrale des statistiques au coeur de la démarche en sciences humaines est qu'elles proposent de repérer des liaisons à l'aide de différents coefficients pour la plupart enrichis de tests assortis eux-mêmes de conditions d'utilisation. C'est par de tels tests que les chercheurs "prouvent" leurs hypothèses. Ces tests ne peuvent pas fonctionner si les données ne respectent pas leurs conditions, ce qui est le cas général des données que nous traitons. L'examen un peu détaillé du modèle linéaire, de sa plus simple expression par une seule équation à celle plus complexe par des équations simultanées avec variables latentes et manifestes, est assorti de nombreux exemples afin de montrer comment les chercheurs en sciences humaines combinent tous ces éléments statistiques pour établir leur preuve. Notre but est de les transposer dans l'extraction de règles d'association sans qu'ils perdent leur force de persuasion.

Le troisième chapitre de cette partie fait le tour des nouvelles méthodes apparues avec l'avènement de l'informatique et l'augmentation exponentielle et continue sur 50 ans de ses capacités de traitement. L'examen plus ou moins détaillé de chacune, fait apparaître que leur nature ne leur permet pas de rendre compte des liaisons complexes que nous voulons représenter. Exception faite de la dernière que nous avons choisie.

Le quatrième chapitre décrit le cadre complexe de cette méthode d'extraction de motifs et de règles d'association, issue de trois courants différents, selon trois logiques irréconciliables. Nous avons dû choisir une logique, mais la connaissance des autres nous permet d'y greffer leurs apports s'ils sont compatibles, et de faire en sorte que nos apports respectent le plus possible les autres logiques.

Contenu de la deuxième partie

La partie II est formée de deux chapitres visant à rendre plus robuste la méthode d'extraction des motifs et des règles d'association par l'utilisation des méthodes de statistiques inférentielles. Dans le chapitre 5, nous partons d'un ensemble de textes de biologie et de leurs mots-clés dont ont été extraits les motifs et les règles d'association selon la méthode la plus courante : un seuil de support a été imposé. Pour tester si les motifs sont ou non dûs au hasard, nous créons un test de randomisation pour les raisons suivantes :

- Les données suivent une loi zipfienne et nous ne connaissons aucun test pour ces lois.
- Nous préférons faire un test qui n'exige pas de spécifier la loi et ses paramètres, qui pourra être utilisé avec d'autres types de données
- Nous voulons un test qui considère l'ensemble des mots et des textes comme un tout
- Nous voulons éviter le problème des comparaisons multiples [131].

Nous constatons sur l'exemple que les permutations les plus adaptées à nos exigences sont celles qui gardent inchangées les marges de la matrice d'incidence mots×textes.

Dans le chapitre 6, nous définissons les permutations gardant les marges inchangées qui sont nécessaires pour faire fonctionner notre test de randomisation. Pour les construire, nous développons une technique à base d'échanges dans une matrice de 0 et de 1 afin de produire à volonté des matrices de mêmes marges mais dont les valeurs sont bouleversées de façon aléatoire. Puis nous essayons de nous donner quelques méthodes pour contrôler que la suite des matrices engendrées respecte bien les lois du hasard.

Contenu de la troisième partie

La partie III contient trois chapitres qui portent sur les relations complexes. Dans le chapitre 7 nous confrontons les liaisons complexes des sciences humaines avec les motifs et les règles en faisant porter notre effort essentiellement sur l'interaction. Nous sommes assurée ainsi de pouvoir trouver toutes les interactions grâce aux motifs. L'algorithme MIDOVA qui découle de cela est en phase de vérification et figure donc dans la partie " bilan et perspectives de ce mémoire.

Dans le chapitre 8, nous mettons au point des méta-règles de nettoyage (une définition rapide en est donnée dans la note de bas de page n°3) dont le but est de supprimer les incohérences d'un jeu de règles. Comme il n'y a pas de solution idéale dans certains cas pour "lever" les incohérences, nous paramétrons ces méta-règles afin que l'utilisateur puisse les piloter en fonction de ses besoins. Nous arrivons à nettoyer des incohérences qu'un utilisateur non averti risquerait de ne pas remarquer : ce sont celles qui proviennent de la dichotomisation des variables catégorielles à plus de deux modalités.

Dans le chapitre 9 nous prolongeons l'extraction des motifs et des règles d'association à des variables numériques en essayant de garder le plus d'éléments possible lors du prolongement (par exemple, nous avons pu garder la structure de treillis sur laquelle s'appuient certains algorithmes, mais pas la dualité entre variables et sujets). Les règles d'association floues ainsi construites se sont trouvées correspondre à des règles floues établies par des chercheurs n'appartenant pas à la communauté de fouille de données. Dans le cas particulier où les variables numériques se

trouvent être binaires, l'extraction des motifs et règles d'association floues produit les mêmes motifs et règles d'association que la méthode habituelle, ce qui fait de cette méthode floue une généralisation de la méthode classique à des données mixtes (binaires et quantitatives).

Contenu de la quatrième partie

La partie IV est un bilan du travail réalisé et des perspectives qu'il laisse entrevoir. Il est formé de quatre sections.

La première section relate les implémentations qui ont été réalisées. Elles sont la première étape de la création d'une plate-forme informatique de l'extraction des liaisons complexes entre variables à partir de données, mais de nombreuses étapes sont encore nécessaires pour rendre ces outils utilisables par un expert. La méthodologie construite dans ce travail de thèse a été mise en pratique sur ordinateur au fur et à mesure de sa construction : chaque élément de cette méthodologie a donné lieu à la construction d'un algorithme qui a été implémenté en un programme qu'on a fait "tourner" sur de nombreux petits ensembles de données créés pour l'occasion, et sur des données réelles dont les spécificités étaient bien connues. Toutefois certains programmes informatiques ont été également été développés pour traiter de nouvelles données, dans le cadre de projets ou de défis. Ce sont ces développements qui sont décrits dans la deuxième section.

Dans la troisième section est décrit l'état actuel de l'algorithme *par niveau* MIDOVA, visant à n'extraire des données, à chaque étape, que les liaisons entre variables complètant l'information apportée par les liaisons extraites à l'étape précédente. L'information différentielle à laquelle on se réfère est attachée à la notion d'intraction et se trouve au sein des motifs, comme détaillé dans la partie III de ce document. L'inclusion de cet algorithme dans une chaîne de traitement contenant d'autres éléments de cette thèse, limités pour l'instant à TourneBool, algorithme de repérage des 2motifs (motifs de deux variables) statistiquement significatifs, est également décrite. Comme on peut le voir dans cette section, l'effet de la chaîne de traitement appliquée à un jeu de données réelles ouvre de nombreuses perspectives.

La dernière section est la conclusion générale de la thèse.

Contenu des annexes

En annexe A, nous donnons quelques détails de preuves pour le chapitre 9 (section 9.4.1). Le problème traité est celui de l'extraction de motifs dont le support est compris entre deux seuils, un seuil inférieur et un seuil supérieur. Nous n'avons pas trouvé dans la littérature d'exposé théorique des effets que le choix d'un seuil inférieur de support pouvait avoir sur la structure de treillis. Le seuil supérieur, quant à lui, n'apparaît nulle part, à notre connaissance, dans les publications en fouille de données, bien qu'en pratique certains utilisateurs aient l'habitude d'en fixer un pour filtrer les règles à interpréter. Nous avons choisi de construire une méthode de réduction du treillis prenant en compte ces deux types de seuillage et pouvant résister à des utilisations variées :

- la construction des treillis réduits se fait sur des motifs et des treillis flous qui suivent notre "fuzzification" exposée dans le chapitre 9, ce qui fait que leur application à des variables binaires donne une construction également valable pour les motifs et treillis classiques.
- les deux seuils peuvent avoir des valeurs extrêmes, ce qui permet de s'intéresser uniquement aux motifs fréquents, ou rares (y compris de support nul), ou ni rares ni fréquents.
- la construction de ces treillis réduits préserve la cohérence avec les négations des variables telle que définie dans le chapitre 9.

L'annexe B est également l'exposé d'une preuve d'un chapitre du mémoire (chapitre 7, section 7.2.3), mais l'écriture formelle est très simplifiée. Nous montrons d'abord sur un petit exemple à trois propriétés que le paradoxe de Simpson provient de l'interprétation de calculs "illicites" de proportions. Ces calculs se font à partir de tableaux de contingence croisant deux variables binaires, comme se font les calculs des indices de qualité des règles d'association. Ce qui nous permet ensuite de nous placer dans le cas d'un utilisateur désirant avoir un jeu de règles d'association sur les trois variables binaires représentant ces trois propriétés. Notre but ici est d'obtenir un jeu de règles exempt de contradictions lors de l'interprétation à l'aide d'une dizaine d'indices de qualité parmi les plus courants. A cette occasion, nous montrons que ces indices se répartissent en deux groupes selon leur façon de choisir entre deux règles simples contradictoires ; et que ces deux types de choix ne peuvent pas déboucher sur un quelconque paradoxe. Toutefois, tout nouvel indice qui fonctionnerait différemment pourrait donner lieu à un jeu de règles d'association contenant des paradoxes.

L'annexe C a pour but de montrer qu'un jeu de règles d'association exprime les *interactions* de tous niveaux entre des variables binaires, et que, de ce fait, il ne peut qu'être contradictoire, redondant, incohérent. Pour cela, le modèle loglinéaire des statistiques est confronté aux règles d'association, La comparaison se fait à l'aide de nombreux exemples, qui explorent toutes les facettes de la liaison entre trois variables que peut exprimer ce modèle. La formalisation se fait à travers une écriture mathématique. Ce modèle loglinéaire fait partie des statistiques inférentielles, et pour chaque exemple, tous les coefficients indiqués dans les équations "diffèrent significativement de zéro", afin de pouvoir convaincre les utilisateurs férus de statistique que les effets repérés ne peuvent être négligés. Mais l'aspect inférentiel a été totalement "gommé", pour éviter une complication inutile dans le cadre de cette comparaison : son seul but est de montrer en détail comment les interactions entre trois variables repérées à travers le modèle loglinéaire se traduisent en des règles d'association formant un ensemble incohérent par nature.

Première partie

Problématique et état de l'art

Introduction de la partie I

En sciences humaines ou en sciences du vivant, le traitement des données est motivé par deux buts : expliquer et prédire les réactions, transformations, évènements et phénomènes psychologiques, biologiques, sociaux, écologiques, économiques à travers un modèle de la réalité. La transformation des données par codage en variables est la première étape menant à ce modèle. Les principes de codage des données n'ont pas énormément évolué avec les progrès de l'informatique, mais la complexité croissante des données a rendu cette étape de codage de plus en plus importante dans le processus de fouille de données. Plutôt que codage, on l'appelle alors plus noblement *nettoyage et pré-traitement* (en anglais "Cleaning and Preprocessing"). Cette étape est décrite dans le chapitre 1.

La seconde étape consiste en la découverte ou l'établissement des liens entre les variables qu'on a créées à la première étape. Dans le chapitre 2, c'est la méthode statistique classique employée par les chercheurs qui est décrite, celle qui les protège des erreurs de raisonnement les plus courantes. Traditionnellement, deux branches différentes de la statistique sont utilisées pour l'étude de ces liens, les statistiques exploratoires (statistiques descriptives, analyse de données) pour les découvrir, et les statistiques inférentielles (théorie de l'échantillonnage, des tests, modèle linéaire) pour les établir. Il peut arriver que la recherche d'un modèle adapté nécessite une succession d'analyses statistiques de ces deux types, mais cette succession est soumise à des règles strictes, comme par exemple l'utilisation de deux ensembles distincts d'objets, l'un pour la découverte d'un modèle candidat et l'autre pour l'évaluation de son adéquation à la réalité. Pour rendre cette modélisation statistique de la réalité plus facile à manipuler, on la spécifie le plus souvent par des formules mathématiques liant les variables, comme par exemple les équations de l'offre et de la demande en économie ou celles de l'évolution d'un système proie-prédateur en écologie. Toutefois, l'acceptation d'écarts imprévisibles à l'égalité (appelés aussi erreurs ou résidus) fait que, malgré son écriture mathématique, un tel modèle n'est pas déterministe. Quand l'écriture du modèle statistique par des équations comportant des erreurs aléatoires n'est pas possible, on rappelle l'existence de ces écarts au modèle déterministe par l'emploi de termes comme valeur observée/attendue, variable manifeste/latente, moyenne/espérance, variance empirique/théorique. Avec ce chapitre, nous espérons avoir fait le tour de ce que les statistiques apportent aux chercheurs en sciences humaines, et comment elles le font, afin de pouvoir garder ce qui peut l'être, et de modifier au mieux ce qui ne peut plus fonctionner avec les données actuelles.

Dans le chapitre 3, c'est le traitement des données tel qu'on peut le faire depuis que les ordinateurs ont remplacé la calculatrice. D'anciennes méthodes statistiques trop gourmandes en calculs pour s'être développées auparavant ont réapparus, pendant que d'autres ont été créées en marge des statistiques sur de nouvelles idées transformées en algorithmes, ou à partir de modèles informatiques. Elles sont décrites plus ou moins en détail selon la contribution qu'elles peuvent nous apporter dans notre recherche de modélisation des liaisons complexes entre variables. La

dernière est celle que nous choisissons car c'est la plus souple. Malgré sa simplicité apparente, elle a eu droit à un chapitre complet dans cet état de l'art, afin de la positionner au mieux au sein des trois courants dont elle est issue.

Le chapitre 4 est donc consacré à l'extraction des motifs et des règles d'association. La version informatique de cette méthode est très connue depuis une dizaine d'années dans le milieu économique de la grande distribution, où elle est utilisée pour faire des études sur le "panier de la ménagère". Ce n'est pas un modèle explicatif que recherchent les économistes en l'utilisant, mais un modèle prédictif trouvant des "pépites" de connaissance au milieu d'un grand nombre de données. Guigues et Duquenne [105] ont défini des *implications*, qui sont aussi des règles d'association, mais dans un but différent : établir un modèle explicatif permettant une représentation de données en concepts. Ces concepts sont organisés selon une structure de treillis, dans laquelle les variables et les objets jouent un rôle dual. Quant à Régis Gras [99], il a défini des *règles d'implication statistique* pour en faire un modèle des relations de causes à effets entre les variables, dans lequel les sujets n'étaient pas différenciés, comme en statistiques. Le point commun à ces deux pionniers est qu'ils se contentaient de petits nombres de variables, dérisoires à côté du nombre de tickets de caisse d'un supermarché qui peut être stocké dans une base de données. La version informatique a intégré à sa façon quelques éléments des deux versions des pionniers. Mais elle en a laissé d'autres de côté, notamment l'aspect causal des relations qui se trouve dans la version de Régis Gras. Ce chapitre va nous permettre d'examiner les raisons de cet "oubli" et comment y remédier. Il va aussi nous permettre de poser les bases de nos diverses contributions visant à mettre plus de rigueur dans l'extraction des motifs et des règles d'association

1

Du codage de l'information au traitement des données en sciences humaines ou du vivant

Sommaire

1.1 La diversité des données

Les données traditionnelles en sciences humaines sont issues de mesures ou d'observations plus ou moins simples à décrire. Si on prend comme exemple des données de psychologie obtenues à l'issue d'un questionnaire de psychologie, et qu'on essaie de ranger ces "mesures" selon leur niveau de complexité, en premier figurent le sexe de la personne interrogée, son âge, sa taille, son poids, qui peuvent être établis avec une certitude raisonnable, suivis de son intelligence, ses comportements, sentiments, impressions, attitudes, opinions, plus complexes à mesurer. Ces observations peuvent également être réparties en mesures *subjectives* quand elles sont faites par la personne elle-même, ou *objectives* quand elles sont faites par un spécialiste du domaine qui remplit

tous les questionnaires à la place des personnes interrogées. Elles peuvent aussi être classées selon leur précision et leur fiabilité qui dépendent de nombreux éléments comme la qualité de la balance, en cas de poids, l'expertise du spécialiste en cas de mesure d'un comportement. Ce ne sont pas les seuls aspects de ces mesures. La mise au point des questionnaires en psychologie sociale et des protocoles d'expérience en psychologie expérimentale a permis de pointer d'autres éléments permettant d'assurer une meilleure qualité des résultats obtenus par leur analyse statistique. Nous renvoyons le lecteur intéressé aux ouvrages dans ces domaines [163, 151, 65] .
Les données issues de textes diffèrent aussi selon leur plus ou moins grande simplicité. Compter le nombre de fois que le mot "je" apparaît dans un texte, ou est prononcé dans un discours est simple. Il est plus délicat de résumer un texte, un discours, d'en extraire le style, les idées, les mots-clés, et la difficulté augmente quand il s'agit d'un journal papier avec sa mise en page, ses textes, images, ou d'un site internet qui contient en plus de la navigation, de l'animation d'images, voire du son. On parle d'ailleurs dans ce cas de données complexes, parmi lesquelles on fait figurer également les images satellitaires avec leurs indications, les dossiers médicaux de patients hospitalisés contenant des données variées comme résultats d'analyses, courbes de températures, radiographies, commentaires des soignants, prescriptions. La "fouille de données complexes" fait actuellement l'objet de recherches actives dans la communauté de fouille de données [9].

1.2 Le premier codage des données : création de variables

Une fois établie la liste de ces mesures, elles sont appliquées à chaque objet d'investigation, c'est-à-dire à chaque personne sondée s'il s'agit d'une enquête, à chaque dossier médical d'un hôpital, à chaque texte d'une oeuvre littéraire, à chaque article traitant d'un élément scientifique, à chaque ticket de caisse d'un supermarché. Leur application nécessite l'utilisation d'une échelle de mesure qui peut être de plusieurs types, allant du type catégoriel (ou qualitatif) au type numérique (ou quantitatif), en passant par le type ordinal. Voici un exemple de ces types dans l'extrait de questionnaire de la figure 1.1 :

Les réponses attendues aux questions sont catégorielles pour $Q_{1.1}$, $Q_{1.3}$, $Q_{3.a}$, $Q_{3.b}$, ordinales pour $Q_{1.2}$ et Q_4 , numériques de $Q_{2.a}$ à $Q_{2.e}$. Le type numérique permet de faire des opérations arithmétiques telles que moyenne, écart-type. Le type ordinal permet d'ordonner les réponses de la plus petite à la plus grande (relation d'ordre total), mais la distance entre deux échelons consécutifs n'est pas nécessairement la même tout au long de l'échelle, contrairement à l'échelle numérique. On ne peut pas faire de moyenne mais on peut utiliser les statistiques de "rang", par exemple médiane et quartiles. Quant au type catégoriel, on ne peut que faire des opérations de comptage du nombre d'objets de chaque catégorie (ou modalité). Ces divers types d'échelles peuvent être encore subdivisés et complétés par d'autres types qui peuvent intervenir dans l'interprétation des résultats de la fouille de données, dont le détail peut être trouvé dans [55, 65] (échelle d'intervalle, de rapports, de Thurstone, Likert) et [66] (données symboliques). Pour ce qui est du traitement en lui-même, nous n'envisagerons que 4 types de données, les trois décrits précédemment, ainsi que les données dichotomiques (ou binaires), qui sont du type Vrai/Faux, Présence/absence, Oui/Non, et qui font partie des données catégorielles (2 modalités), mais peuvent en plus, quand on les code par 1/0, bénéficier d'un traitement identique aux données numériques (par exemple moyenne), alors que les données catégorielles à plus de deux modalités, même recodées par des nombres, ne le peuvent pas. Dans le questionnaire, les questions $Q_{1.1}$ et

[9] Chaque année, un atelier d'une journée sur la fouille de données complexes a lieu lors la conférence française EGC (Extraction et Gestion de Données). Voici le lien vers EGC 2006 : http ://www.rech.enic.fr/egc2006.

On a interrogé des étudiants d'informatique par un questionnaire anonyme. Voici un extrait des questions posées

Q1 : 1. Disposez-vous d'un ordinateur personnel, c'est-à-dire sur lequel vous pouvez décider de faire des installations de logiciels, langages, voire systèmes d'exploitation. : ☐ oui ☐ non
2. Si oui, depuis combien de temps ☐ moins d'1 mois ☐ entre 1 et 3 mois ☐ entre 3 mois et 1 an ☐ plus d'un an
3 . Quel type d'ordinateur avez-vous actuellement ☐ portable ☐ fixe.

Q2 : Donner la répartition en pourcentage de temps passé sur les ordinateurs (domicile ou autre) pour chacune de ces activités :

a) Travail universitaire (calcul scientifique, programmation)	b) Communication (Chat, mails,...)	c) Navigation Internet	d) Distractions (jeux, films, photos, musique,...)	e) Autre	f) Total
					100

Q3 :Vous arrive-t-il de faire du développement d'applications ☐ oui ☐ non. Si oui, en quel langage ? ☐

Q4 : Pensez-vous que les méthodes vues en cours vous seront utiles plus tard ? ☐ pas du tout ☐ un peu ☐ beaucoup

Voici les réponses de 3 étudiants à ces questions :

Num	Q1			Q2					Q3		Q4
	1	2	3	a	b	c	d	e	a	b	
1	oui	moins d'1 mois	portable	10	10	30	40	10	non		beaucoup
2	non			40	30	10	20	0	non		un peu
3	oui	plus d'un an	fixe	60	20	10	10	0	oui	Java	beaucoup

FIG. 1.1 – Quelques questions d'un questionnaire

$Q_{3.a}$ sont dichotomiques.
Après dépouillement des questionnaires remplis, on peut obtenir un tableau ressemblant à celui en bas de la figure 1.1, avec une ligne par objet (ici par étudiant), et dans chaque ligne, au plus une valeur par colonne. On a rajouté une colonne d'identifiants (ici des numéros), afin de pouvoir retrouver les questionnaires en cas de doute sur une valeur du tableau. On a ainsi une représentation des données par un ensemble d'objets, de variables, et de valeurs associées à chaque couple (objet, variable) Dans notre exemple, il y a un ensemble de trois objets o_1, o_2, o_3, identifiés par leur numéro, un ensemble de 11 variables, et pour chaque couple (objet, variable), il y a au plus une valeur, qui peut être une "chaîne de caractères" ou un nombre. En général, l'échelle ordinale sera représentée par un nombre, mais ce peut être aussi le cas de l'échelle catégorielle. Il convient alors de garder en mémoire le type de chaque variable afin de ne pas faire de traitement inadapté (la plupart des logiciels prévoient une possibilité de signaler qu'une variable codée par un nombre est catégorielle, afin d'éviter qu'elle ne soit traitée comme une variable numérique).

1.3 Les distributions de valeurs des variables

Une variable statistique est un vecteur, une variable probabiliste est une fonction. Dans un cas on a des valeurs pour n sujets, dans l'autre cas on a une valeur pour un sujet potentiel, celle-ci pouvant osciller autour d'une valeur théorique. Ces oscillations sont décrites par sa loi de probabilités (densité). On passe du vecteur à la variable aléatoire en faisant tendre n vers l'infini.

Cette transformation des données en variables de divers types que nous venons d'exposer est la première étape du processus de formalisation. Elle est suivie d'un examen pour chaque variable de ses valeurs pour l'ensemble des objets, appelé "tri à plat". Si la variable est catégorielle, on relève l'*effectif* (le nombre d'objets) de chaque catégorie. Si elle est quantitative, le même type

de comptage peut se faire en remplaçant les catégories par des intervalles de valeurs[10] qui est un guide précieux pour les diverses représentations de données. Si la variable est ordinale, selon que son nombre de valeurs différentes est petit ou grand, c'est le premier type de comptage ou le second qui est privilégié.

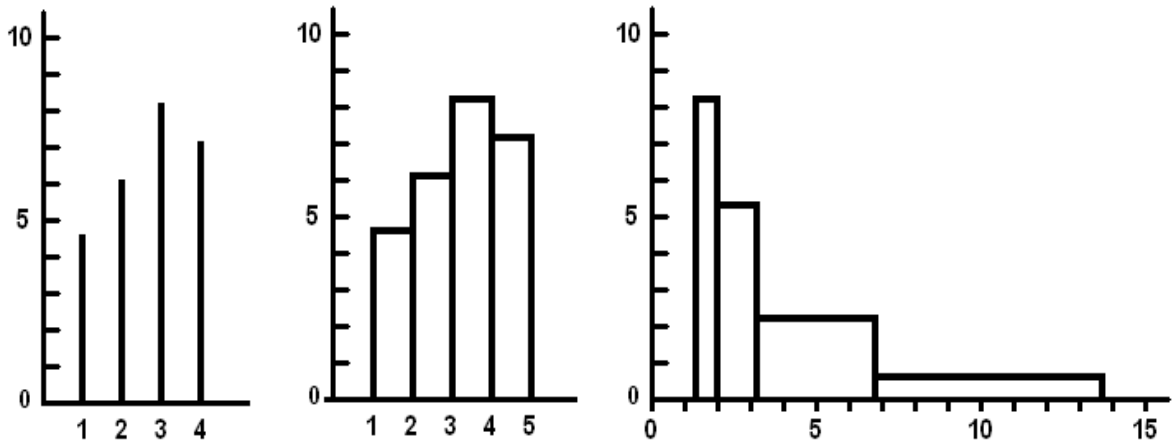


FIG. 1.2 – Diagramme à bâtons à gauche, histogrammes au milieu et à droite

Le tableau de nombres obtenu par ce comptage peut être représenté graphiquement (voir la figure 1.2) en exprimant les effectifs en fonction des valeurs. C'est le *diagramme à bâtons* dans le cas catégoriel (chaque catégorie indiquée en abscisse est surmontée d'un trait proportionnel à son effectif, graphique à gauche de la figure 1.2) et l'*histogramme* dans le cas quantitatif (sur chaque intervalle de valeurs en abscisse est construit un rectangle dont l'aire est proportionnelle à son effectif, graphiques au milieu et à droite de la figure 1.2). Avec un tel graphique, on peut repérer les caractéristiques de la variable, et notamment émettre une hypothèse sur la loi de probabilité qu'elle suit. Par exemple, si l'histogramme a la forme d'une cloche (comme au centre de la figure 1.2), on envisagera que la variable puisse suivre la *loi normale*[11], ce qui permettra de la caractériser par la connaissance de deux nombres seulement, moyenne et écart-type (ce sont les *paramètres* de la loi) et de se placer dans les conditions idéales d'application de nombreux tests statistiques ("modèle linéaire" par exemple). Si on retient cette hypothèse, la variable prend alors le statut de *variable aléatoire*, et une partie de son traitement peut se faire dans le cadre de la théorie des probabilités. La manipulation des données s'en trouve banalisée, mais les données deviennent plus abstraites. Ce ne sont plus les données réelles, correspondant à un nombre fini d'objets tous différents, bien déterminés, mais un modèle de ces données, correspondant à un nombre quelconque, le plus souvent infini, d'objets indifférenciés. Le passage dans ce cadre impose le choix d'un modèle adapté aux traitements que l'on désire faire et au type de données dont on dispose. Nous précisons ci-dessous ce que sont ces deux types de choix.

[10] Dans le cas le plus courant, ces intervalles sont choisis de même amplitude (graphique au milieu de la figure 1.2). Mais si on attend une distribution inégalitaire de valeurs, il vaut mieux les choisir de grande amplitude là où il y a peu de valeurs et de petite amplitude là où il y en a beaucoup (graphique à droite de la figure 1.2). En cas de difficulté de détermination de ces intervalles, on peut se reporter à l'ouvrage du "groupe Chadule" [50]

[11] on pourra conforter cette impression par le calcul d'indicateurs tels que kurtosis (aplatissement), et asymétrie, assortis de tests comme indiqué dans [59]

1.3.1 Lois de répartition classiques

Loi de Bernouilli : 2 modalités (jet d'une pièce non équilibrée, probabilité de pile ou face : si pièce équilibrée, identique, sinon différents).

Loi binomiale : n épreuves de Bernouilli independantes : on tire n boules successivement avec remise, et on compte combien de noires (ou de blanches) : 2 modalités noire ou blanche, succes /échec. Ce nombre est entre 0 et n.

Loi de Poisson : dans un nombre indéfini d'épreuves, le nombre d'apparition d'un évènement de faible probabilité (2 modalités se produisent ou non) le nombre est entre 0 et l'infini, mais plus la valeur s'approche de l'infini plus la probabilité est petite.

Loi multinomiale : nombre de modalités k>=2 (boules de k couleurs différentes) fermés, répartition de probabilité de chaque modalité donnée à l'avance, on compte le nombre de boules de chaque sorte quand on en tire n.

1.3.2 Choix d'une loi de distribution dictée par les valeurs :

Dans son ouvrage de statistique, Dagnelie [59] expose en détail les calculs statistiques permettant de caractériser une distribution de valeurs. Notamment, il décrit page 244 le "système de Pearson" qui connaissant la forme approximative de la densité d'une variable numérique , permet de déterminer à partir des seuls 4 premiers moments[12] de la variable (respectivement espérance, variance, asymétrie et aplatissement) l'équation de cette densité[13]. Bien sûr, ce ne sont que les lois de probabilités de densités "ordinaires" (en forme de dôme comme au milieu de la figure 1.2, de "i" comme à droite de la figure 1.2 ou de "j", le symétrique du "i") dont il est question ici. En sont exclues par exemple les lois bimodales (comme le serait la taille des individus de deux populations l'une formée de pygmées et l'autre de géants), les lois de variance infinie (comme le cours de certaines actions en bourse).

1.3.3 Choix d'une loi de distribution dictée par le domaine des données

Selon l'origine des données, on attend plutôt certaines répartitions que d'autres. Par exemple, si le phénomène quantitatif est dû à de nombreuses petites causes indépendantes dont les effets s'ajoutent, c'est la *loi normale*. Si les effets se multiplient, c'est la loi *lognormale*. Le nombre d'arrivées des clients à une file d'attente suit la *loi de Poisson* si l'arrivée de chacun est indépendante de celle des autres - cette loi est entièrement déterminée par la connaissance d'un seul paramètre, son espérance.

1.3.4 Les lois de puissance

Mais revenons sur les effets d'une somme de nombreuses petites causes indépendantes : François Bavaud [16] éclaire la contradiction apparente entre l'existence de distributions à valeurs extrêmes et les conséquence du théorème de la limite centrale, qui dit que la somme d'une multitude de distributions quelconques finit par produire une distribution normale, quelles que soient ces distributions : ceci suppose en fait qu'elles soient de *variance finie*, ce qui est loin d'être

[12] Le moment d'ordre n d'une variable est la somme des nièmes puissances de ses valeurs, divisée par leur nombre. Dans le cas d'une variable théorique numérique continue, il s'agit d'une intégrale. Le moment *centré* d'ordre $n \geqslant 2$ est obtenu en prenant les puissances des écarts entre chaque valeur et sa moyenne/espérance. Ici, il s'agit en fait du moment d'ordre 1 et des moments centrés d'ordres 2, 3 et 4

[13] Effectivement on a déjà vu que la loi normale était entièrement caractérisée par ses deux premiers moments, le troisième étant nul et le quatrième égal à 3.

le cas dans nombre de phénomènes auxquels sont confrontées les sciences humaines ! Ainsi ces dernières années la prise de conscience s'est faite jour de l'existence, pour décrire globalement de nombreux phénomènes d'auto-organisation, de "lois de puissance" de la forme $p(x) \sim x^{-\alpha}$ (où $\sim$ signifie proportionnel et α est une constante positive) dans lesquelles l'apparition de valeurs exceptionnelles en queues de distribution n'est pas négligeable et empêche de stabiliser leur variance. Quand on fait croître de telles distributions et qu'on en fait la somme, la variance du tout est dans un rapport fixe $n^{1-\frac{1}{\alpha}}$ à la variance des parties (propriété des "distributions stables d'ordre α", où $\alpha \in [0, 2[$), ce qui rattache aussi ces lois aux phénomènes d'auto-similarité :

- Dans les graphes dits "petits mondes" [80] - par exemple ceux qui traduisent les relations sociales d'un groupe étendu d'individus, les compositions de conseils d'administration de sociétés, les liens entre pages Web, la structure d'un réseau téléphonique, les citations entre articles scientifiques...- la répartition des degrés des noeuds du graphe suit une loi de puissance.
- En économie, les répartitions inégalitaires de biens suivent également des lois de puissance, comme la loi de Pareto pour les revenus, .
- Il est connu depuis longtemps que la répartition des mots dans un corpus de taille quelconque et de langue quelconque suit approximativement une loi de puissance dont l'exposant dépend de la langue et du type de textes ; c'est la loi d'Estoup-Zipf, précisée par Benoît Mandelbrot [14] [172]).

Risquons ici un concept inédit à notre connaissance : puisque le rang d'un mot dans un corpus est une variable aléatoire de variance empirique non bornée, toujours susceptible de prendre des valeurs plus grandes à l'introduction de textes nouveaux (mots spécialisés, néologismes...), on pourrait dire que les mots constituent une variable catégorielle *ouverte*, toujours susceptible d'accueillir des éléments nouveaux, bien que rares - type de variable à ajouter aux types répertoriés plus haut. A noter que ceci introduirait la notion d'infini dans les variables catégorielles, à l'instar des variables numériques où cette notion est présente implicitement.

1.4 Les recodages courants des données : transformation des variables

1.4.1 Recoder pour mieux coller à un modèle :

Selon les traitements de données envisagés, on peut avoir besoin de recoder les variables de base en de nouvelles variables plus adaptées. Cela arrive quand la distribution diffère de celle souhaitée pour le traitement ultérieur, ou quand l'examen de leurs distributions de valeurs fait apparaître des anomalies par rapport à la distribution attendue. Nous développons les anomalies les plus importantes, qui sont l'existence de valeurs manquantes ou d'*outliers* (valeurs extrêmes).

La présence de **valeurs manquantes** est souvent gênante dans certains traitements, et sa prise en compte dans les logiciels peut s'avérer désastreuse (ainsi des problèmes de lecture de fichier entraînent des décalages de toutes les valeurs suivantes, ou la suppression de tous les objets ayant une valeur manquante pour une variable). Par exemple, dans le tableau 1.1 des réponses des 3 étudiants, la question $Q_{3.b}$ n'a fait l'objet de réponse que de la part de l'étudiant 3, car c'est le seul à avoir répondu "oui" à la question $Q_{3.a}$. Pour éviter toute mauvaise interprétation de cette

[14] $f \sim (r-\rho)^{-\alpha}$, où f est la fréquence du mot de rang r, α est une "température informationnelle" généralement supérieure à 1, ρ est une constante de "correction d'aplatissement" de la courbe f=fonction de r dans la zone des fortes fréquences, K une constante = somme de toutes les fréquences / somme de tous les termes $(r-\rho)^{-\alpha}$.

non-réponse, qui n'est pas un refus de réponse, mais un effet de la structure de la question, on peut décider d'un code de non-réponse, qui diffère de celui du refus de réponse. Mais on peut aussi décider de recoder la question $Q_{3.a}$ et $Q_{3.b}$ en une seule variable contenant le nom du langage en cas de réponse "oui", et la valeur "aucun" en cas de réponse "non". Une autre façon de corriger, plus automatique, consiste à affecter à l'objet n'ayant pas de valeur pour une variable une valeur "neutre" (par exemple la valeur la plus courante en cas de variable catégorielle, ou la médiane en cas de variable ordinale, ou la moyenne en cas de variable quantitative), ou même la valeur la plus "probable", c'est-à-dire "proche" de celles des objets ayant une valeur proche pour les autres variables[15]. Bien sûr cette correction des données (on parle même d'*enrichissement des données*) peut avoir des conséquences gênantes sur les résultats ultérieurs du traitement si elle diffère trop de la réalité. Par exemple, selon que les non-réponses sont l'indicateur de valeurs extrêmes que les personnes interrogées n'ont pas souhaité divulguer, ou au contraire de valeurs beaucoup plus courantes que celles attendues pour ces personnes au vu de leurs réponses aux autres questions, l'utilisation de tests statistiques basés sur des distributions normales, des calculs de moyennes, et de variance risquent de donner une conclusion "fausse" quand les valeurs manquantes ont été corrigées automatiquement.

En présence d'**outliers**, un retour au données brutes, s'il est possible, s'impose. Car il convient de s'assurer que ces données extrêmes ne sont pas dues à des erreurs de cotation, de mesure, ou tout simplement, de saisie. Dans ce cas, la donnée est corrigée par sa vraie valeur si on peut la retrouver, ou sinon transformée en donnée manquante, avec le problème que pose sa prise en compte. Dans le cas contraire, si la valeur est jugée correcte et qu'elle gène le traitement initialement prévu, c'est que la distribution de valeurs n'est pas celle attendue, auquel cas on peut essayer de changer la distribution de valeurs par recodage.

1.4.2 Recoder pour mieux coller à la sémantique des données :

A côté de ces problèmes courants pouvant nécessiter un recodage des données, d'autres plus spécifiques aux données peuvent se poser. Par exemple, les questions $Q_{2.a}$ à $Q_{2.e}$ posent un autre type de problème : on se doute bien que la proportion de temps n'est pas très réelle. Elle indique surtout une préférence pour certaines occupations. Si la personne traitant les données est plutôt intéressée par une ou deux occupations parmi les quatre proposées, on peut se contenter de recoder seulement celles-ci en variables, en mettant une valeur d'intérêt, qualitative ou ordinale, pour l'occupation (par exemple 3 si elle est choisie en premier, 2 en second, 1 en troisième ou au-delà et 0 si elle n'a pas été choisie).

1.4.3 Diagramme des recodages possibles d'une variable

Ces recodages font partie du "nettoyage des données" qui représente plus de la moitié du travail en fouille de données [113] et qui ne peut se faire sans une connaissance approfondie du processus de collecte des données. De sa qualité dépend l'efficacité de la fouille. Une fois la variable choisie, on peut encore changer son type comme indiqué dans le schéma de la figure 1.3.

Dans ce schéma les flèches en traits pleins indiquent le changement d'un type à un autre et se justifient aisément qu'elles soient à un même niveau ou dans le sens descendant.

1. numérique vers :

[15]Par exemple s'il manque le poids d'une personne dont on connaît le sexe, l'âge, la catégorie socioprofessionnelle, on lui attribuera la moyenne des poids des personnes de même sexe, et d'âge comme de catégorie socio-professionnelle proche

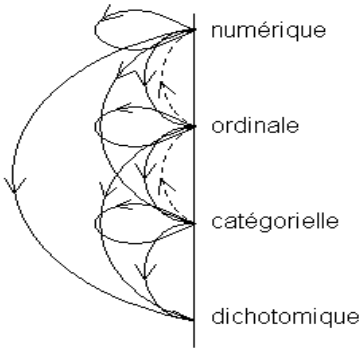


FIG. 1.3 – Les changements possibles entre types élémentaires de données

- numérique : standardisation [16], normalisation [17] pour obtenir une loi normale afin de simplifier par exemple les calculs d'intervalles de confiance.
- ordinale : l'étendue des valeurs est découpée selon des seuils donnés en plusieurs intervalles consécutifs, puis on numérote les intervalles dans le sens des valeurs croissantes[18]
- catégorielle, dichotomique : on procède comme précédemment, mais au lieu de numéroter les intervalles, on leur donne un nom de catégorie.

2. ordinale vers :
   - ordinale : on regroupe des échelons successifs afin de garder l'ordre initial des valeurs
   - catégorielle : on oublie l'ordre
   - dichotomique : par exemple on renomme les échelons intermédiaires d'une échelle médicale en "normaux" et on regroupe les autres (hypertendus, hypotendus) en "anormaux"
3. catégorielle vers catégorielle ou dichotomique
   - regroupement de catégories (par exemple pour obtenir des effectifs plus importants respectant les conditions d'utilisation du test du Chi2)

Les flèches sont plus problématiques pour le sens ascendant. Il y en a deux indiquées par des traits en pointillés :

1. du catégoriel vers l'ordinal : possible dans le cas des échelles de Guttman[55, 229], ce qui consiste à faire apparaître un ordre "par vote" (par exemple, si tous ceux qui préfèrent A à C et C à B préfèrent également A à B [19], l'ordre qui s'impose est $A \succ C \succ B$),

[16] Se fait surtout pour les variables supposées suivre une loi normale. On centre en remplaçant chaque valeur par son écart à la moyenne, puis on réduit en divisant par l'écart-type, la variable obtenue a pour moyenne 0 et pour écart-type 1.

[17] Se fait surtout quand on a une distribution étirée. Par exemple, si la proportion d'objets ayant leurs valeurs comprises entre 0 et 5 est satisfaisante, et qu'elle est trop petite pour les intervalles à partir de 5, il suffit de remplacer toutes les valeurs x au dessus de 5 par des valeurs plus petites, comme $5 + (x - 5)/2$.

[18] Les seuils peuvent être choisis de diverses façons, notamment en fonction des habitudes des praticiens, comme les stens ou les stanines [208] utilisés lors de la constuction des tests de psychologie, ou comme les échelles en trois points hypo/normal/hyper obtenus à l'issue d'analyses biologiques (tension par exemple).

[19] Il est rare que cela se passe vraiment ainsi, non seulement parce que la personne interrogée peut répondre "n'importe comment", mais aussi parce que les comparaisons de A avec C, de C avec B et de A avec B peuvent se faire sur trois critères différents, par exemple, on peut préférer une Vespa à une bicyclette parce que cela rend le déplacement moins fatiguant, la bicyclette à la voiture pour préserver l'environnement, et la voiture à la Vespa car elle permet d'emmener plus de passagers. Ces phénomènes sont au coeur de plusieurs paradoxes : paradoxe du tournoi, paradoxe de Condorcet ...

2. de l'ordinal vers le numérique (on considère que les valeurs sont à peu près équidistantes, les écarts à cette norme étant considérés comme une des nombreuses causes de la variabilité courante)

Signalons la possibilité de transformer des données dichotomiques en données numériques selon un modèle plus complexe, qui est celui de la théorie de la réponse à l'item (IRT : *Item Response Theory* [112]), à condition qu'elles vérifient les contraintes imposées par ce modèle de type probabiliste. Ce type de modèle connaît un engouement croissant depuis une dizaine d'années dans les domaines de sciences humaines utilisant des tests d'évaluation, comme l'édumétrie, la didactique. D'après D. Laveault et J. Grégoire [151], sa forme la plus courante est basée sur la le modèle de Rasch (1966) qui est, d'après Marc Demeuse [63], une version probabiliste du modèle de Guttman que nous venons d'exposer. L'utilisation de ce modèle très gourmande en ressources (nombre important de données, et capacité de traitement importante) est maintenant possible grâce à des logiciels disponibles sur Internet (comme QUEST de R.J. Adams et S.T. KHOO, 1993).

1.4.4 Recodage courant d'une variable en plusieurs

Dans la figure 1.4, on peut voir les codages les plus courants de la réponse à une question du genre "donnez votre opinion sur la déclaration suivante : il faut interdire toute circulation automobile dans les villes".

a	a'	a"	a_1	a_2	a_3	a_4	a_5	b_1	b_2	b_3	c_5	c_4	c_3	d_1	d_2
Tout à fait contre	-2	1	1	0	0	0	0	0	0	0	1	1	1	0	1
Plutôt contre	-1	2	0	1	0	0	0	1	0	0	1	1	1	1	0
Ni pour, ni contre	0	3	0	0	1	0	0	1	1	0	1	1	0	0	0
Plutôt pour	1	4	0	0	0	1	0	1	1	1	1	0	0	1	0
Tout à fait pour	2	5	0	0	0	0	1	1	1	1	0	0	0	0	1

FIG. 1.4 – Les codages d'une échelle de Likert en 5 points

En première colonne figure la réponse a proposée (case à cocher dans le questionnaire), puis deux codages quantitatifs équivalents de la réponse, par les variables a' et a", le premier étant une traduction plus "fidèle" du texte des réponses proposées que le second. Les suivants sont des codages binaires. Le codage en 5 variables notées de a_1 à a_5, appelé *codage par dichotomisation*, transforme toute variable catégorielle en autant de variables dichotomiques qu'elle a de modalités. Si on se réfère à la variable a", a_i prend la valeur 1 si $a" = i$ et la valeur 0 sinon. C'est le codage le plus utilisé en fouille de données pour traduire de façon numérique une variable dichotomique, sans la transformer en véritable variable numérique toutefois. Les codages suivants sont plus rares et viennent parfois compléter ou remplacer le premier codage binaire. Les variables b_i, pour i variant de 1 à 3 sont égales à 1 quand $a" > i$ et à 0 sinon, elles permettent de cumuler les opinions de personnes dans le sens de l'accord, les variables c_i faisant la même opération dans le sens contraire. La variable d permet de regrouper les opinions selon leur extrémalité, les valeurs "tout à fait contre" et "tout à fait pour" exprimant une conviction plus forte que les valeurs "plutôt contre" et "plutôt pour". Un codage prenant tous ces effets en compte revient à remplacer la variable ordinale de départ par 13 variables, et même par plus si on veut traduire aussi les négations des variables a_2, a_3 et a_4 (celles des variables a_1 et a_5 sont déjà représentées par b_1 et c_5). L'avantage d'un tel codage est essentiellement de permettre l'utilisation d'algorithmes simples sur ces variables pour exprimer des liaisons fines, mais il y a un inconvénient qui est celui de l'explosion combinatoire. Précisons que ce type de codage des

données préalable au traitement fait partie des traditions en analyse des données, le plus connu étant celui qui permet de transformer des données catégorielles afin de les soumettre à une analyse prévue pour des variables quantitatives (il s'agit de "l'analyse en composantes principales" qui devient, formellement,"l'analyse factorielle des correspondances" par un simple changement de métrique [19]) alors qu'en statistique inférentielle, la recherche de relations plus fines se fait plutôt au niveau du traitement, comme par exemple dans les calculs des contrastes du modèle linéaire [1, 236]).

Ce type de codage dichotomique d'une variable en plusieurs peut être étendu en partie à une variable quantitative, ou à une variable catégorielle. Pour la première, on la transforme d'abord en variable ordinale en choisissant plusieurs points de coupure qui pourront être des quantiles ou obtenus à partir des paramètres d'une loi de probabilité théorique (comme les stanines [208]), par connaissance experte (seuils d'hypertension, d'hypotension), ou empiriquement par observation des changements de pente de la courbe des fréquences cumulées des valeurs [50], puis on procède de la même façon qu'indiqué dans la figure. Les résultats s'expriment avec des intervalles (a_1, a_5, les b_i et c_i), ou des réunions d'intervalles de valeurs (a_2, a_3, a_4, d_1 et d_2).

Pour ce qui est des variables catégorielles, a priori seule la dichotomisation en variables a_i se pratique, les éventuels regroupements de modalités ayant lieu par connaissance experte du domaine des données, les modalités ne vérifiant a priori aucune relation d'ordre.

1.5 Autres recodages des données

En plus des méthodes classiques de recodage que nous venons de voir, dictées par un traitement ultérieur des données à l'aide des techniques statistiques exploratoires et inférentielles, d'autres modèles des données plus informatiques que statistiques peuvent guider le recodage. Par exemple le codage flou (figure 1.5), permet de recoder les variables ordinales ou numériques d'une façon intermédiaire entre le codage numérique par une seule variable et le codage dichotomique par un certain nombre de variables, qui peut être assez lourd [240].

On peut aussi s'inspirer des "portes logiques" de l'électronique en opérant un changement à l'intérieur d'un groupe de variables par la création de variables synthétiques. Par exemple : remplacer les deux variables A, B binaires par les quatre variables (A et B), (A et non B), (non A et B) et (non A et non B). Le modèle algébrique des variables qui est sous-jacent à ce recodage est à rapprocher de celui présent dans les plans d'expériences des statisticiens [60].

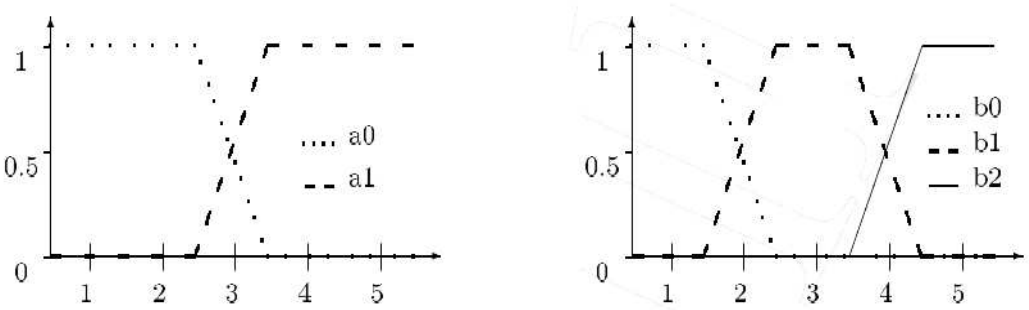
A côté de ces nombreuses transformations plus ou moins automatiques des variables, il existe tous les recodages "à la main" réalisés en utilisant une connaissance experte des données. Par exemple, le challenge de la KDD Cup 2006[20] auquel nous avons participé, consistait à prédire des thromboses pulmonaires connaissant la valeurs de variables mesurées sur des points (appelés "candidats") de radiographies de patients. Parmi les nombreux recodages proposés par un membre de l'équipe après une étude minutieuse des lésions des poumons de chaque patient d'après les images 3D qu'il a reconstituées à partir des données fournies, citons celui de la variable appelée "x". Cette variable, présentée par les organisateurs comme une coordonnée dans l'espace à 3 dimensions des points candidats, avait une répartition bimodale (voir dans la figure 1.6), mais nulle part n'était évoquée la prise en compte des 2 poumons.

Une nouvelle variable nommée "poumon droit" a alors été créée, qui valait 1 quand x était supérieur à 275 et 0 sinon. La création de cette variable permettait d'éviter de décider que 2 points aux valeurs de (x,y,z) proches étaient voisins alors qu'ils étaient situés dans 2 poumons différents

[20] Résultats sur <http ://www.cs.unm.edu/kdd_cup_2006>. Sur 68 équipes notre équipe a été classée 21ème à la tâche 1, 18ème à la tâche 2, 4ème à la tâche 3.

a	a'	a1	a0	b2	b1	b0
1 : pas du tout	0	0	1	0	0	1
2 : un peu	0	0	1	0	0,5	0,5
3 : assez	0,5	0,5	0,5	0	1	0
4 : beaucoup	1	1	0	0,5	0,5	0
5 : énormément	1	1	0	1	0	0

Transformation de la propriété "a" codée selon une échelle de Lickert à 5 points en 1, 2 ou 3 propriétés floues.



Transformation de la propriété "a" codée de 1 à 5 en 2 ou 3 propriétés floues.

FIG. 1.5 – Les codages flous d'une variable ordinale ou quantitative

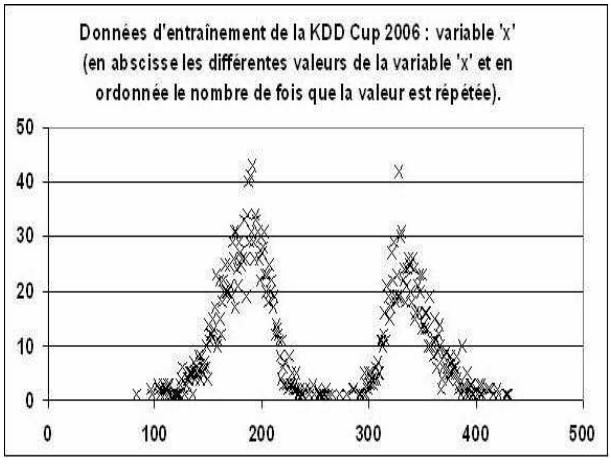


FIG. 1.6 – Répartition des valeurs de la variable "x" de la Kdd Cup 2006 pour les données d'entraînement

(ce qui était le cas par exemple si les deux valeurs de x étaient 270 et 280). La robustesse de notre mesure de ressemblance entre candidats qui en a découlé nous a permis d'améliorer notre stratégie de décision sur l'existence ou non d'une lésion pour les données de test. En effet pour l'évaluation d'une tâche, il suffisait de trouver une seule lésion par zone du poumon atteinte, alors que les faux positifs étaient lourdement sanctionnés. L'étape ultérieure du codage, que nous n'avons pas eu le temps de faire, consistait à créer une deuxième variable en "repliant" le deuxième dôme de x sur le premier après l'avoir éventuellement étiré en hauteur et en largeur pour obtenir une distribution unimodale. Cette transformation permettait de garder à la variable son statut de variable quantitative, tout en la transformant en variable unimodale. Cette transformation visant à normaliser une variable bimodale par repliage est bien sûr uniquement adaptée à la nature des données dont nous disposions. Une contraction des valeurs de la variable "x" autour de la valeur de 275, tout à fait exclue pour notre traitement des données utilisant une mesure de proximité spatiale, aurait pu normaliser "x" dans un autre contexte.

1.6 Conclusion sur le codage : rôle, avantages et limites

Pour conclure sur les différentes façons de créer une variable, nous avons vu qu'elle peut se limiter à la transcription des données après nettoyage, ou devenir un pré-traitement de celles-ci permettant un traitement ultérieur "aveugle" des données avec des algorithmes appropriés. Notre thèse portant principalement sur le traitement automatique des données, nos exemples ont essentiellement montré les transformations opérées quand on dispose déjà de trois ensembles, un ensemble d'objets, un ensemble de variables, et un ensemble des valeurs associées à chaque couple (objet, variable). Cette description, qui fait jouer le même rôle aux variables et aux objets, est appropriée quand les variables sont toutes de même type, en particulier quand elles sont de type numérique ou booléen. Si ce n'est pas le cas, un recodage des variables permet de s'y ramener, comme nous venons de l'exposer.

1.6.1 La place du codage dans le processus de fouille de données

Mais les données ne sont pas toujours fournies sous cette forme. Nous avons participé à la tâche 2 du challenge de la KDD Cup 2003[21] qui consistait à déterminer une matrice des cocitations (liste des liens entre auteurs citants et auteurs cités). Le corpus du challenge contenait environ un million d'articles en anglais soumis dans le domaine de la physique et fournis sous forme de fichiers "latex". La difficulté spécifique venait de l'origine variée des textes (les informations sur la revue dans laquelle ils avaient été soumis avaient été retirées), et de leur manque d'adéquation à une norme latex unique [22]. Dans l'impossibilité où nous étions de repérer automatiquement la position du champ des auteurs, des titres, et des références, ainsi que les revues de la bibliographie qui étaient indiquées par des codes variés selon les articles (en général quelques lettres prises dans les titres des revues, ce qui donnait parfois des codes différents pour une même revue et des codes identiques pour deux revues différentes), le premier travail a été de créer des variables artificielles afin d'extraire des types de "styles latex". Nous avons ensuite établi à partir de celles-ci une typologie des articles, ce qui nous a permis d'appliquer un traitement identique à tous les articles d'une même classe, qui consistait entre autres choses en l'extraction du titre, des auteurs, des références à des années quelle que soit leur position dans le texte, des références

[21] Résultats sur <http ://www.cs.cornell.edu/projects/kddcup/results.html> notre équipe a été classée 3ème.

[22] La moitié seulement des textes ont pu être compilés en une trentaine d'essais, et certains textes ne contenaient pratiquement aucune balise *latex* autre que la mise en forme de l'article, alors que certains définissaient des macros *latex* qu'ils utilisaient par la suite notamment pour leurs références bibliographiques

bibliographiques avec réécriture en clair du nom de la revue pour chacune des références de la bibliographie chaque fois que c'était possible. Nous sommes ainsi arrivés à créer deux ensembles de données, l'un dont chaque objet était un article, avec des variables telles que la liste des auteurs, l'année au dessus de laquelle l'article a été soumis (on prenait le maximum des années auxquelles l'article se référait), etc. et l'autre dont chaque objet était une référence bibliographique d'un article, avec des variables telles que le nom du premier auteur, son prénom, la liste éventuelle des autres, le nom de la revue, l'année, le numéro, etc... leur variable commune étant le numéro d'article. Pour les détails voir [40].

On voit sur cet exemple que la construction de nouvelles variables permettant de construire des données exploitables pour des traitements automatiques est liée à une connaissance experte du domaine, et qu'elle nécessite une prise de distance permettant de les écrire sous une autre forme. Cette gymnastique intellectuelle qui consiste à établir un modèle des données en les considérant d'une autre façon grâce à notre connaissance du domaine s'avère nécessaire pour les données non structurées, dont font partie les données textuelles. Elle fait également partie des pratiques du statisticien qui doit construire un modèle probabiliste des données afin de pouvoir se raccrocher aux lois de probabilité courantes. Un de nos buts dans cette thèse est de faire émerger des modèles de données en repérant des liaisons entre les variables, et sans un bon codage, il est difficile de construire un bon modèle. Et l'apparition de certaines liaisons nous permet d'approfondir notre connaissance des données et peut se traduire par la création de nouvelles variables suivie d'un nouveau traitement.

De façon plus générale, l'analyse descriptive des données peut être considérée comme un processus de création de nouvelles variables, plus pertinentes, de "plus haut niveau", à partir de variables proches de l'observation brute. Par exemple, les composantes obtenues par ACP (Analyse en Composantes Principale) ou ICA (Analyse en Composantes Indépendantes), les facteurs obtenus par AFC (Analyse Factorielle des Correspondances) sont des variables quantitatives ou ordinales. Les clusters obtenus par *analyse typologique* et autres méthodes de classification non supervisée forment des variables catégorielles. Ces méthodes peuvent remplacer la transformation "à la main" des réponses à des questionnaires à partir d'un premier codage des réponses en utilisant par exemple des regroupements d'items en catégories, des créations d'échelles par sommation de réponses cotées à des questions. Il existe donc un continuum entre recodage manuel et création informatisée de nouvelles variables.

A signaler aussi que l'apprentissage artificiel est une tentative d'approcher automatiquement par le calcul la valeur d'une variable d'intérêt (généralement catégorielle) attribuée humainement : par exemple des diagnostics médicaux (KDD Cup 2006), rupture de thèmes dans des discours (Défi Fouille de Textes DEFT 2006).

1.6.2 La perte de relations entre les données pouvant résulter de ce codage

Toutefois, le processus de codage, ou de recodage, qui a l'avantage de formaliser les données pour permettre l'exécution d'algorithmes de traitement des données, a l'inconvénient d'occulter certains liens difficiles à transcrire dans ce formalisme. Il faut alors les exprimer à part pour en tenir compte lors du traitement. Par exemple, la réponse à la question Q_2 de la figure 1.1, a été transformée par simple recopiage des réponses en cinq variables numériques $Q_{2.a}$ à $Q_{2.e}$, ces variables étant liées par l'égalité $Q_{2.a} + Q_{2.b} + Q_{2.c} + Q_{2.d} + Q_{2.e} = 100$ due au fait qu'elles expriment la répartition d'un tout en ses parties[23]. Méconnaître cette relation aurait pour effet

[23]On peut avoir des données de préférence (on demande à une personne d'indiquer ses préférences entre des objets pris 2 à 2) de classement (faire un classement de l'ensemble des candidats). Dans le dernier cas, le total est

des résultats aberrants à certains tests statistiques (dans le modèle linéaire, certains calculs nécessitent l'inversion d'une matrice, qui se trouverait alors être non inversible), et pléthore de résultats inintéressants (par exemple, lors de l'extraction des règles d'association, au premier rang apparaissent toutes les règles du type "si $Q_{2.a} < 10$ et $Q_{2.b} < 10$ et $Q_{2.c} < 10$ et $Q_{2.d} < 10$ alors $Q_{2.e} > 60$" ce qui rassure d'abord, en constatant que les algorithmes permettent de retrouver les propriétés connues des données, mais rend longue et pénible la recherche de nouvelles relations enfouies sous un grand nombre de relations sans intérêt. On sait que ces liens mathématiques, induits soit par les données, soit par leur codage (la somme des variables a_i de la figure 1.4 est 1), peuvent s'exprimer à travers un langage adapté dans les SGBD, sur lequel on peut s'appuyer pour supprimer leurs effets des résultats. Il en existe nombre d'autres qui, lorsqu'ils sont négligés lors du traitement des données, produisent des résultats tout aussi désastreux, bien que moins visibles. Ils sont répertoriés depuis longtemps en statistique et sont séparément à l'origine du développement de plusieurs branches de cette discipline. Par exemple, la prise en compte approfondie de la dépendance spatiale et/ou temporelle des objets a débouché sur les modèles ARIMA de la finance [69], les modèles spatiaux de la géographie [111], les modèles de durée des populations et de diffusion en épidémiologie [68]), les modèles hiérarchiques de l'économie de l'éducation (ex. : effet famille/école/région), et les modèles "capture/recapture" de l'écologie (un même individu peut être comptabilisé plusieurs fois). Mais ces modèles permettant d'exprimer séparément certaines particularités des données se sont avérés insuffisants pour le traitement des données issues de nombreuses branches de sciences humaines et biologiques (par ex. en agronomie [54], écologie numérique [157], psychologie du développement [24])). Le besoin d'une compréhension globale de ces effets variés est à l'origine du développement du *modèle expérimental*, formalisé en *plan d'expériences* à l'aide d'un langage de description de type algébrique, puis traduit en termes statistiques en "modèle linéaire" de la théorie des tests. L'exposé que nous allons faire du modèle expérimental et du modèle linéaire a pour but de définir de façon opérationnelle les liens présents dans les données et de permettre l'extraction des seuls liens apportant de la connaissance par une méthodologie construite sur l'extraction des règles d'association, en remplacement du modèle linéaire mal adapté à ces données.

fixé - en général (n*(n+1)) si les éventuels ex aequo sont codés de la façon appropriée.

2

L'implication en sciences humaines

Deux démarches sont principalement utilisées en sciences humaines pour "établir" des relations de causes à effets : la démarche expérimentale, qui consiste à déclencher les causes pour en observer les effets, et la démarche exploratoire, qui consiste à s'appuyer sur les covariations repérées sur des observations variées pour en déduire des relations de causes à effets. Nous décrivons en première partie sur un exemple imaginé toutes les étapes-types de la construction d'une expérience, en insistant plus particulièrement sur ses deux points cruciaux que sont la neutralisation des effets parasites, et le choix du modèle statistique approprié. En deuxième partie, nous décrivons quatre exemples de modèles s'appuyant sur l'exploration des covariations de données collectées dans différents champs de sciences humaines (en économie, sociologie, psychologie). Leurs techniques d'investigation et de validation des relations causales diffèrent bien qu'elles s'appuient toutes sur une même modélisation statistique : un système d'équations linéaires.

Une fois cette revue faite des principaux types de preuves en sciences humaines, nous faisons le tour des instruments de la preuve qu'elles ont utilisés. Nous revenons d'abord sur les différents statuts des variables, liés à leur signification, qui ont guidé le raisonnement dans ces exemples. Nous détaillons particulièrement le concept d'*interaction* entre variables, selon lequel l'action conjointe de deux causes ne se réduit pas toujours au cumul de leurs deux actions séparées, qu'elle peut les renforcer en amplifiant leurs effets séparés, mais également aller à l'encontre de l'une ou de l'autre, voire des deux, en amenuisant leurs effets, en les annihilant, ou même les inversant. Dans les deux parties suivantes, nous exposons les techniques statistiques permettant d'établir les liaisons entre les variables. Nous commençons par définir les liaisons en statistiques descriptives puis nous continuons avec les statistiques inférentielles, cet ordre étant justifié par le fait que certains tests de la dernière partie s'appuient sur les coefficients de la première partie.

Puis nous terminons par une description de liaisons complexes qui ne peuvent être représentées par le modèle linéaire. comme le célèbre "paradoxe de Simpson".

Dans ce chapitre, certains passages ont une expression formelle mathématique ou statistique. Ils sont là pour témoigner de la complexité de la démarche du chercheur en sciences humaines qui veut établir des relations de causes à effets, quelle que soit sa méthodologie. Non seulement il doit s'appuyer sur toute une expertise liée à son domaine de recherche qui lui fait privilégier certaines hypothèses plutôt que d'autres, choisir comment les opérationnaliser (monter une expérience, construire un questionnaire, écrire des requêtes pour des moteurs de recherche, etc..) et interpréter les résultats du traitement des données produits par les logiciels, mais il doit également construire des variables à partir de ses données afin qu'elles puissent entrer dans des modèles formalisés mathématiquement, et vérifier les conditions d'utilisation des tests statistiques permettant d'éprouver ces modèles.

Ces tests sont aussi là pour nous montrer que tout ce travail complexe ne peut être remplacé par l'exécution d'un simple algorithme d'extraction de règles d'association sur des données brutes, voire même déjà codées en variables.

Sommaire

2.1 La démarche expérimentale

Les agronomes, les psychologues, les biologistes, les économistes, et en général tous les chercheurs des sciences humaines qui veulent établir des règles du type A→B font, chaque fois qu'ils le peuvent, des expérimentations. La démarche expérimentale qu'ils suivent s'est complexifiée au fil des ans, intégrant des corrections et ajouts au fur et à mesure que certaines règles établies précédemment s'avéraient imprécises, incomplètes, voire inexactes. Nous en déroulons les étapes clefs sur un exemple imaginé.

2.1.1 Le principe

Lors d'une expérimentation, on "prouve" que A a un effet sur B si quand on modifie la valeur de A, toutes choses égales par ailleurs, celle de B se modifie. L'expérience se fait selon un protocole expérimental rigoureux : on définit l'hypothèse de travail qu'on désire prouver, on cherche tous les paramètres pouvant intervenir, les différentes valeurs à leur donner, comment faire varier la variable A, et on spécifie le sens de la variation correspondante de B, les tests à utiliser, le seuil de significativité. On réalise alors l'expérience, et si on obtient ce qu'on avait prévu, on déclare que A a un effet sur B et qu'on l'a prouvé. Toutefois, on sait que la règle n'est jamais prouvée définitivement, car un paramètre oublié peut la remettre en cause. D'après Hume et Popper [197], seule la réfutation (*falsification*) de la règle peut être assurée de manière définitive, dès qu'un chercheur découvre une condition dans laquelle elle ne se vérifie pas.

2.1.2 Une petite expérience

Imaginons par exemple qu'on désire prouver que le sectarisme est une caractéristique modifiable de l'individu. Pour cela, il est nécessaire de définir de façon opérationnelle le sectarisme afin d'en établir une mesure. La mesure idéale pour les calculs est quantitative, la moins bonne est binaire (oui/non). En effet, si la personne change "un peu" d'attitude, on doit pouvoir le repérer, et cela se fait facilement en regardant le signe de la différence. Notons que la "bonne" mesure vérifie un nombre important d'autres conditions qui sont décrites dans tous les manuels de sciences expérimentales, notamment "mesure-t-elle ce qu'elle est censée mesurer ?". Nous renvoyons le lecteur intéressé à ces ouvrages [65], et retournons au problème de la mesure de l'attitude sectaire. Si on peut mesurer en temps (par exemple le temps que la personne met à répondre à certaines questions, le temps de fixation de la pupille de l'oeil), en intensité (par exemple le rythme cardiaque, la composition en certains éléments de la salive), on dispose d'une mesure objective. Sinon on élabore un questionnaire en suivant de nombreuses règles qu'on trouve exposées dans tous les manuels traitant des questionnaires [151], afin que la mesure soit la plus objective possible. Une fois qu'on a réuni ces conditions, on peut établir la mesure O de l'attitude. Mais l'expérience X n'est pas encore montée. Il faut maintenant trouver comment changer cette attitude. Imaginons qu'on passe un film montrant une personne victime de sectarisme. Posons l'hypothèse suivante : "La visualisation du film diminue l'attitude sectaire des personnes l'ayant vu"[24]. Appelons donc X la visualisation de ce film. On peut supposer que si on mesure l'attitude O1 de quelqu'un avant le film, puis O2 de la même personne après le film, O2 est inférieur à O1, le sectarisme ayant diminué. Bien sûr nous faisons l'expérience sur plusieurs personnes, et toutes les personnes ne réagissent pas de la même façon, au point que certaines personnes ont même une valeur de O2 supérieure à O1.

2.1.3 Premières réflexions suite à l'examen des données collectées

Pour fixer les idées, imaginons que l'on ait réuni un groupe de 10 personnes, et que l'on dispose pour chacune d'elle de la mesure O1 de son attitude avant la visualisation et O2 après, comme indiqué dans le tableau de la figure 2.1. Pour évaluer l'évolution de l'attitude de chaque personne, on calcule la différence $O_2 - O_1$. Cette différence n'est pas constante, mais elle varie selon les sujets. On peut voir sur le tableau qu'elle est plutôt négative, cette impression étant renforcée par le graphique de la figure 2.2. Dans ce graphique, l'évolution de chaque sujet est

[24] Cet exemple est une vision simpliste des choses qui n'a rien à voir avec une théorie quelconque en psychologie

	O1	O2	O2-O1
s1	5.5	2.1	-3.4
s2	10.2	8.4	-1.8
s3	7.2	4.1	-3.1
s4	8.3	8.6	0.3
s5	9.8	6.5	-3.3
s6	14.5	12	-2.5
s7	8.5	7	-1.5
s8	8.3	7.4	-0.9
s9	9.4	5.4	-4
s10	11.3	11.5	0.2

FIG. 2.1 – Les deux mesures O1 et O2 des 10 sujets et leurs différences

indiquée par des traits, dont l'inclinaison permet de déterminer le sens et l'amplitude de cette variation. On peut ainsi constater que deux sujets seulement, s4 et s10, ont une augmentation de leur attitude sectaire (indiquée par un trait montant) alors que tous les autres ont une diminution de celle-ci. L'évolution moyenne est indiquée par le trait en pointillés, qui est descendant, ce qui indique une diminution moyenne de l'attitude sectaire.

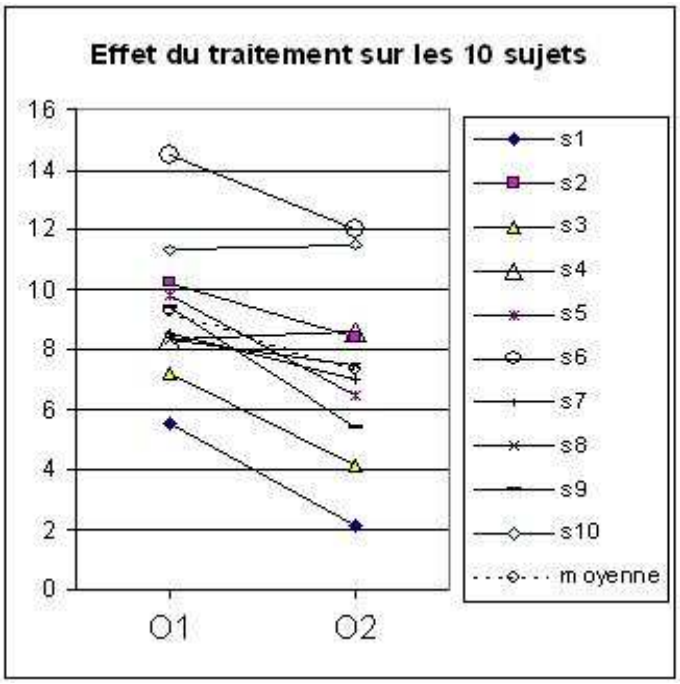


FIG. 2.2 – Les évolutions de chacun des 10 sujets

Toutefois, le fait que l'évolution diffère selon les sujets, au point même que deux personnes aient leur valeur de O2 supérieure à celle de O1, invite à se poser quelques questions avant de conclure que l'hypothèse "La visualisation du film diminue bien l'attitude sectaire" est vérifiée. Notamment cette question "Quel résultat attendait-on ?". On aurait certainement aimé trouver l'équation $O_2 - O_1 = t$, où t est un nombre négatif exprimant l'effet du traitement, ici la visualisation du film. Une telle équation est un modèle déterministe de la réalité (selon[16]), comme l'est la mécanique classique qui donne l'équation du mouvement des planètes de notre

système solaire[25]. Ce modèle peut aider à se la représenter, et à appuyer des raisonnements sur des calculs, comme ceux ayant permis d'établir l'existence de la planète Pluton et sa position, alors que la faible puissance des télescopes de l'époque, ne permettait pas de la trouver. Mais à côté de ces modèles déterministes en existent d'autres, plus adaptés aux objets d'études moins prévisibles que les planètes, comme les êtres vivants, en particulier les humains. Dans le cas de notre expérience, l'équation $O_2 - O_1 = t + \epsilon$ où la différence n'est pas exactement égale à t, mais peut fluctuer autour de cette valeur, représente beaucoup mieux cette réalité, ainsi qu'on le voit dans le graphique de la figure 2.3.

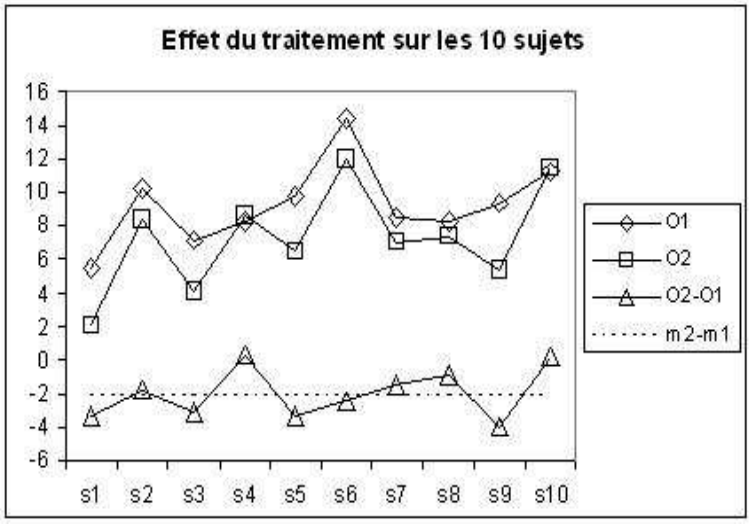


FIG. 2.3 – Les différences de mesure pour chacun des 10 sujets

Dans ce graphique, on a représenté pour chaque sujet les valeurs de O_1, O_2, $O_2 - O_1$, ainsi que la différence des moyennes m1 et m2 respectives de O1 et O2 pour tous les sujets, indiquée par un trait en pointillés. Ainsi, si on estime la valeur de t par m2-m1, les fluctuations ϵ de $O_2 - O_1$ autour de t sont indiquées par les écarts entre les points symbolisés par des triangles et le trait pointillé. Les écarts à la valeur estimée de t ont une apparence tout à fait aléatoire, et sa valeur égale à -2 est bien inférieure à 0, ce qui peut rassurer définitivement l'expérimentateur sur la véracité de l'hypothèse. Mais pour la communauté scientifique des psychologues, ce n'est pas une preuve suffisante de celle-ci. L'effet mis en évidence est certes important, mais il faut maintenant prouver qu'il est "significatif", c'est-à-dire qu'il n'est pas dû au seul hasard. Cette preuve fait appel aux statistiques, qui s'appuient sur une modélisation probabiliste de la réalité. Dans les cas les plus généraux, le modèle utilisé est "le modèle linéaire" des statistiques, car il représente une approximation raisonnable de la réalité, comme l'est localement la tangente à la courbe représentant graphiquement une fonction quand il s'agit d'établir le sens des variations de la fonction. Nous allons exposer son utilisation dans ce cas très simple, ce qui va nous permettre de montrer les principes de son fonctionnement, et nous le formaliserons ultérieurement.

[25] Toute cette analyse de la modélisation de la réalité est inspirée de la lecture de l'ouvrage de cet auteur, auquel nous renvoyons le lecteur pour plus de détails.

2.1.4 Utilisation d'un modèle statistique nécessaire à la "preuve"

Description du modèle

Modèle M1 : $O_2 = O_1 + t + \varepsilon$ H0 : t = 0

valeur estimée : t_{est} = -2

sources de variations	SC	ddl	CM	F	p
traitement	40	1	40	17.4	0.002
résidus	20.74	9	2.304		
totale	60.74	10			

FIG. 2.4 – Modèle linaire pour tester l'effet traitement

Dans la figure 2.4 est présenté le modèle linéaire qu'on désire tester. C'est l'équation $O_2 = O_1 + t + \epsilon$ selon laquelle la mesure O_2 (après traitement) pour un sujet est égale à la somme de sa mesure O_1 (avant traitement), d'une valeur t qui représente l'effet traitement, et d'une valeur de ϵ qui est aléatoire. On impose à cette variable aléatoire d'être indépendante de O_1, d'espérance nulle et de variance fixe, Pour estimer la valeur de t (qui est supposée ne pas dépendre des sujets), on peut écrire l'équation pour chaque sujet en remplaçant O_1 et O_2 par les valeurs du sujet et on obtient autant d'équations que de sujets. Il faut alors résoudre le système pour trouver la valeur de t_{est}. Dans la figure 2.4, on peut lire le résultat : -2, qui indique que le sectarisme d'un sujet moyen diminue de 2 après la visualisation du film[26].

L'hypothèse testée H0 est que l'effet traitement (visionnement d'un film) est nul. Ce qui ne semble pas être le cas. Toutefois, ce n'est pas parce que t est estimé à -2 que sa valeur théorique est différente de 0. Tout dépend des valeurs de ϵ. Si elles varient entre -20 et 20, une valeur de -2 peut être jugée proche de zéro, alors que si elles varient entre -1 et 1, la valeur de -2 est jugée éloignée de 0. Les valeurs de ϵ sont appelées les résidus dans le tableau. Et ce ne sont pas les valeurs que l'on compare dans ce modèle, mais les sommes des carrés, car elles ont de bonnes propriétés mathématiques. La conclusion se trouve en interprétant la valeur de p en haut à droite du tableau. Elle est égale à 0.002. c'est la probabilité qu'on puisse trouver une valeur de -2 alors qu'en théorie elle est de O. Cette probabilité est très petite, notamment inférieure au seuil de risque $\alpha = 0.05$ qui est habituellement pris. Ainsi, nous rejetons la possibilité que t soit nul, et nous acceptons l'hypothèse alternative H1, que t est négatif. C'est précisément le résultat que nous escomptions. Nous détaillons maintenant les calculs qui ont permis d'obtenir de résultat.

Calculs aboutissant aux résultats du tableau de la figure 2.4

La simplicité de cette expérience permet de faire les calculs "à la main", c'est-à-dire sans utiliser de logiciel spécialisé, une calculatrice scientifique, ou un tableur comme Excel ou comme celui de la suite OpenOffice suffisant largement. La valeur totale de la somme des carrés (SC) est obtenue en sommant les carrés des différences $O_2 - O_1$ pour chaque sujet. Comme il y a 10 sujets, et que la liberté de variation de leurs valeurs respectives à la différence $O_2 - O_1$ n'est pas restreinte par une liaison entre elles, le nombre de degrés de liberté correspondant est de 10.

[26] Toutes les valeurs données dans le tableau s'obtiennent en utilisant un logiciel de statistique, et le chercheur en sciences humaines se contente habituellement de lire le résultat qui est la valeur de p. Le calcul de p se fait à partir des sommes des carrés (SC) et des *degrés de liberté* (ddl) en utilisant la loi de Fisher-Snedecor (F). Ces nombres sont indiqués dans le tableau pour permettre au chercheur de contrôler éventuellement les calculs.

Puis on décompose cette somme de carrés en deux parties, la partie due au traitement t, dont la valeur estimée est -2 (l'estimation dans ce modèle, se fait généralement par la méthode des moindres carrés, mais dans ce cas très simple, c'est tout simplement la moyenne des différences), et la partie résiduelle, les doubles produits disparaissant dans les cas les plus simples comme ici. La première somme des carrés est 40, obtenue en faisant le produit de t_{est}^2 par 10, et la seconde est la somme des carrés des écarts $(O_2 - O_1) - t_{est}$. Les degrés de liberté se décomposent de la même façon, mais la justification de leur valeur est plus délicate ici. Les résidus sont liés par la relation qui leur impose une somme nulle, ce qui fait que neuf d'entre eux seulement peuvent varier librement, le dixième étant égal à l'opposé de leur somme. Les degrés de liberté des résidus sont donc égaux à 9, et comme ceux du total sont de 10, il en reste un pour le modèle. Les carrés moyens (CM) sont obtenus en faisant le quotient des SC avec les ddl correspondants, et le quotient des deux CM qui vaut 17.4 suit, dans le cas de H0, la loi de Fisher-Snedecor, avec ddl1=1 et ddl2=9, ce qui donne une probabilité de 0.002 que F soit supérieur à 17.4.

Les diverses acceptions de "modèle linéaire"

Précisons le sens du mot "linéaire" que nous avons utilisé dans cet exemple : il s'agit de la linéarité au sens des espaces vectoriels, qui permet le calcul matriciel. Sa forme la plus courante est une équation ayant en membre gauche la variable "à expliquer", et en membre droit une somme dont les termes sont des variables "explicatives" multipliées par des coefficients auxquels on a ajouté un terme d'erreur. On l'appelle généralement *équation de régression linéaire* (pour des raisons historiques). Les *paramètres* que l'on cherche à estimer sont les coefficients de l'équation, on teste en général leur nullité par un tableau comme celui de la figure 2.5. Pour les trouver on exige que soient vérifiées un certain nombre de conditions sur les résidus (similaires à celles données dans l'exemple précédent). Si le membre droit de l'équation contient des termes avec des produits de variables explicatives, le modèle est encore considéré comme linéaire par la plupart des auteurs, la méthode de calcul ne changeant pas. En effet, la linéarité concerne ces paramètres qui sont les inconnues de cette équation (et non les variables dont les valeurs sont connues). Les produits de variables indiquent leurs interactions. Et on teste la nullité de leur coefficient. Si elle est retenue, on dit que l'interaction est nulle, ou qu'il n'y a pas d'interaction entre les variables considérées.

On peut aussi avoir plusieurs variables à expliquer et plusieurs équations. Dans ce cas, les calculs peuvent différer, surtout si certaines variables figurent dans le membre gauche d'une équation et dans le membre droit d'une autre, comme c'est le cas de l'exemple d'économie traité plus loin.

2.1.5 Utilisation du "modèle linéaire" habituel

Description du modèle

De nombreuses formes du modèle linéaire pourraient être écrites pour notre exemple, et testées. Voici dans la figure 2.5 la plus courante selon laquelle les valeurs de $O_2 - m_2$ sont proportionnelles à celles de $O_1 - m_1$. Cette équation peut aussi s'écrire sous la forme équivalente $O_2 = a_1 O_1 + a_0$. C'est l'équation de régression simple. le résultat du test de ce modèle peut se lire dans la figure 2.5 au même endroit que dans la figure précédente. L'hypothèse nulle testée ici est que le coefficient a_1 est nul, et le but du test est d'établir que cette hypothèse est fausse, c'est-à-dire que a_1 est différent de 0. C'est bien le cas puisque on peut lire sur le tableau que p vaut 0.001, qui est encore inférieur au seuil de 0.05.

Modèle M2 : $(O_2 - m_2) = a_1(O_1 - m_1) + \varepsilon$ H0 : $a_1 = 0$

valeur estimée : $a_{1\,est} = 1.09$

sources de variations	SC	ddl	CM	F	p
régression	63.98	1	63.98	25.2	0.001
résidus	20.28	8	2.535		
totale	84.26	9			

FIG. 2.5 – Modèle linaire pour tester la régression de O2 sur O1

Calculs aboutissant aux résultats du tableau de la figure 2.5

L'écriture la plus courante du modèle de régression de O2 sur O1 est sous la forme $O_2 = a_1O_1 + a_0 + \epsilon$, mais pour faciliter l'exposé des calculs sur cet exemple, on écrira le modèle sous la forme $O_2 - m_2 = a_1(O_1 - m_1) + \epsilon$, où $m_1 = 9.3$ et $m_2 = 7.3$ sont les moyennes respectives de O1 et O2. Cette fois, la somme des carrés totale est égale à la somme des écarts des valeurs de O2 à leur moyenne m2, et la somme des écarts étant nulle, on perd un degré de liberté par rapport aux 10 de la somme précédente, qui ici sont de 9. Pour évaluer la somme des carrés des écarts au modèle, il faut auparavant avoir calculé une estimation de a_1, qui est le quotient de la somme des produits croisés des écarts de O1 à sa moyenne et de O2 à la sienne, par la somme des carrés des écarts de O1 à sa moyenne ($\frac{\Sigma(O_1-m_1)(O_2-m_2)}{\Sigma(O_1-m_1)^2}$). Pour obtenir la somme des carrés résiduelles, il suffit de calculer pour chacun des sujets sa valeur estimée pour O2 en remplaçant O1 par sa valeur dans la formule $O_2 - m_2 = a_1(O_1 - m_1)$ et de faire la somme des carrés des écarts entre la valeur estimée de O2 et sa valeur observée. Le nombre de degrés de liberté de cette somme diminue encore de 1, et devient 8 car s'ajoute la contrainte que leur somme soit nulle. L'effet dû à la régression se trouve en multipliant la somme des carrés des écarts de O1 à sa moyenne par l'estimation du coefficient de régression. Le nombre de degrés de libertés associés à cette somme est 1, comme dans le modèle précédent. Puis on termine le calcul comme précédemment, et la valeur de 0.001 pour p, bien inférieure à 0.05, permet de conclure que H0 est fausse et que les variations de O2 sont bien liées à celles de O1 de façon significative.

2.1.6 Le choix d'un modèle statistique approprié

Nous venons de voir deux formes du modèle linéaire. Avec la première forme l'effet du traitement t a une valeur de -2 qui s'ajoute à la première mesure pour former la seconde tirée à la lecture du tableau de la figure 2.4.. Cette forme est une écriture mathématique de l'effet que l'expérience a pour objectif d'établir. La seconde forme, qui est celle de la régression linéaire, est assez utilisée pour examiner les covariations de deux variables observées comme indiqué dans la figure 2.6. Dans notre cas elle permet d'établir que les variations de O2 autour de sa moyenne s'obtiennent approximativement en multipliant celles de O1 autour de la sienne par un coefficient de 1.09, donc proche de 1. Ce qui ne prouve pas du tout l'effet du traitement, mais peut rassurer sur le choix qui a été fait dans le modèle précédent. Il aurait été plus intéressant de tester l'hypothèse nulle H0 : $a_1 = 1$, ce qui peut être fait également en utilisant le modèle linéaire, et qui aboutit à une acceptation de cette hypothèse H0. Le manque d'intérêt de ce deuxième modèle, produisant pourtant un résultat très significatif, nous montre que l'écriture d'un modèle statistique a moins de chances d'aboutir à des conclusions intéressantes quand elle n'est pas guidée par la connaissance de l'expérience. Les modèles linéaires qu'on peut tester sont trop nombreux

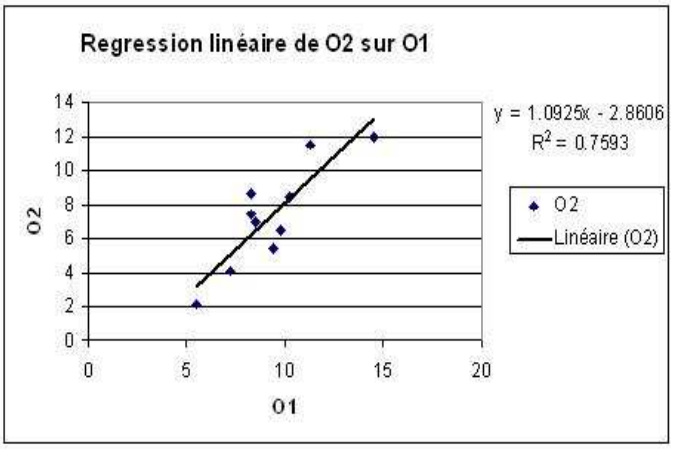


FIG. 2.6 – La droite de régression de O2 sur O1

pour être tous essayés de façon aveugle[27].

2.1.7 Un "bon" test statistique ne suffit pas à la "preuve"

Maintenant l'hypothèse postulée que la visualisation du film permet de diminuer l'attitude sectaire paraît prouvée car le premier modèle linéaire montre un changement très significatif de la mesure après traitement. Mais cela ne suffira pas a convaincre la communauté scientifique des chercheurs en psychologie. L'établissement de la preuve est encore sujet à des critiques. Certes, il y a une différence avant/après qui est importante et statistiquement significative. Mais peut-on être sûr que cette différence n'est pas due à d'autres causes que la visualisation du film ? Parmi les causes possibles, n'y aurait-il pas eu entre les deux mesures une émission à la télévision évoquant une conséquence grave du sectarisme, ce qui aurait eu pour effet de diminuer l'attitude sectaire tirée à la lecture du tableau de la figure 2.4.des personnes l'ayant regardée ? Le plan expérimental aurait dû prendre en compte cette variable. Et puis, une mesure ce n'est pas toujours neutre. Peut-être que pour certaines personnes, le fait d'avoir répondu à un questionnaire leur a fait prendre conscience de leur sectarisme, qu'ils se sont efforcés de corriger. Cette diminution n'est alors pas l'effet de la visualisation, mais l'effet de la participation à l'expérience. Tous ces effets sont bien connus par les psychologues, et d'ailleurs des autres sciences expérimentales comme l'agronomie ou la médecine (parcelle témoin, groupe Placebo, [60, 212]) et le dispositif expérimental peut permettre de les contrôler.

2.1.8 Modification du plan expérimental pour renforcer la "preuve"

Voici un exemple de plan expérimental proposé par A. Léon [163] qui permet de corriger ces deux effets. C'est le plan à 4 groupes de Salomon :

$G1$	(R)	$O1$	X	$O2$	$(experimental)$
$G2$	(R)	$O3$		$O4$	$(Controle1)$
$G3$	(R)		X	$O5$	$(Controle2)$
$G4$	(R)			$O6$	$(Controle3)$

[27] En tester un grand nombre est de toute façon exclu, la succession de tests de modèles étant soumise à des règles strictes, comme nous l'exposons dans le paragraphe portant sur les principes du modèle linéaire.

La première ligne se lit ainsi : les sujets du groupe G1 ont d'abord fait l'objet d'une mesure, O1 étant la suite de leurs résultats à cette mesure, ensuite ils ont été soumis à une expérience X, puis ils ont fait l'objet d'une mesure O2. La dernière ligne se lit ainsi : les sujets du groupe G4 ont fait une seule mesure, qui a eu lieu en même temps que la mesure O2 du groupe G1.

- O représente la mesure de la variable qu'on désire modifier par l'expérience X, Cette variable qu'on désire mesurer est appelée "variable dépendante". C'est pour notre exemple une attitude sectaire.
- X représente l'expérience qu'on fait agir sur la variable dépendante, (pour notre exemple la visualisation d'un film), appelée "le traitement".
- R représente la randomisation, qui assure que les sujets de chacun des 4 groupes ont été prélevés de façon aléatoire dans l'ensemble total des sujets. Cette précaution permet de limiter l'influence parasite sur la variable dépendante de toutes les causes extérieures en contrebalançant leurs effets.
- La présence du groupe G2 permet de vérifier que la modification observée n'est pas due à certains évènements extérieurs au traitement, (pour notre exemple tels qu'une émission à la télévision sur le sectarisme), ou tout simplement la maturation, c'est-à-dire à l'histoire qui a eu lieu entre deux mesures non simultanées.
- Le groupe G3 assure que la modification de l'attitude n'est pas tout simplement causée par la mesure initiale et non par l'expérience.
- L'ajout du groupe G4 permet de mesurer finement l'effet "pur" du traitement. On l'obtient en retirant du changement total non seulement l'effet dû à l'histoire, mais encore celui dû à l'interaction entre le traitement et l'histoire qui agit en amplifiant ou en réduisant le changement (ce type d'effet est décrit en détails dans un paragraphe suivant).

L'expérience que nous avons décrite précédemment ne contenait que le groupe G1, ce qui avait pour conséquence que la différence de mesures mise en évidence pouvait être attribuée non seulement au traitement, mais aussi à d'autres causes. Avec les 3 groupes supplémentaires, on va pouvoir contrôler les effets éventuels des deux causes concurrentes de variation que sont l'histoire et la mesure. L'écriture du modèle linéaire prenant en compte tous ces effets va permettre cette fois d'évaluer l'effet du traitement en retirant les effets parasites. Bien sûr le modèle va être plus complexe, les contraintes d'utilisation plus nombreuses, et les calculs plus difficiles, mais les principes sont identiques : on estime tous les paramètres nécessaires, puis on décompose la somme des carrés de la mesure faite en fin d'expérience en autant de parts indépendantes que d'effets entrant en compte, ce qui permet notamment de tester l'effet du traitement conditionnellement à l'effet de l'histoire et de la mesure. Et si l'effet du traitement s'avère significatif, alors on aura fait un pas supplémentaire dans la direction de la preuve. Mais bien sûr, on ne peut jamais prouver définitivement que l'effet repéré est bien dû au traitement et non à une cause inconnue qu'on aurait négligée.

2.2 La démarche exploratoire

La démarche exploratoire peut être choisie pour 2 raisons

1. Pour prouver une hypothèse unique, de type causal, comme remplacement de la démarche expérimentale, car l'expérience n'a pas pu être montée pour diverses raisons (raisons techniques, éthiques, etc.)
2. Quand on ne dispose pas d'une théorie permettant d'établir une hypothèse, ou quand on est en présence de plusieurs théories concurrentes.

2.2.1 Comment prouver une hypothèse quand on ne peut pas faire d'expérience

Si aucune expérience n'a pu être montée pour prouver la règle A→B, par exemple parce qu'il a été impossible de réunir les conditions matérielles à une telle expérimentation, on peut utiliser des données recueillies pour un autre usage si elles contiennent des informations sur nos deux variables A et B. Toutefois, la preuve de la règle est plus difficile à établir par cette démarche exploratoire. Non seulement, comme dans la situation expérimentale, on n'est pas assuré d'avoir recueilli toutes les mesures intéressantes de tous les paramètres importants, mais encore la condition "toutes choses égales par ailleurs" ne peut plus être invoquée car il y a très peu de chances que les données respectent tous les réglages fins des paramètres du protocole expérimental qui permettaient cela. Si on cherche une liaison entre les variations de A et celles de B, il va alors falloir évaluer la part de cette liaison qui n'est pas due à d'autres phénomènes, car c'est cette part qui nous intéresse. Chaque fois que cela est possible, on essaie de se replacer dans le cadre de l'analyse de variance, en prenant quelques précautions supplémentaires[28] et on procède de la même façon que dans la démarche expérimentale, la baisse de qualité en découlant étant relativement faible.

2.2.2 Dégager les causes et les effets ?

Alors que dans la démarche expérimentale classique, on cherche à tester une seule hypothèse du type "Les modifications de B sont dues à celles de A, toutes choses étant égales par ailleurs"[29], la démarche exploratoire permet de plus larges investigations dans l'ensemble des variables. En voici 4 exemples, le premier en sociologie, le second en économie et les deux derniers en psychologie différentielle[30] . Ils sont tous basés sur une modélisation de ces covariations à l'aide d'un système d'équations linéaires, mais leur mode de preuve diffère. Le premier construit trois variables à partir de données sociologiques recueillies pour un grand nombre de comtés et utilise les corrélations linéaires pour démêler les effets des causes. Sa preuve est faite par un jeu de déductions confrontant les différents liens de causalité entre trois variables qu'un système linéaire peut représenter (en excluant toutefois les causalités réciproques) à trois théories sociologiques à la lumière des valeurs trouvées sur les données. Pour le second, on raisonne sur trois variables économiques dont les valeurs sont (ou plutôt seraient car il s'agit d'un exercice théorique) collectées à intervalles réguliers avant la date T. Le problème ici n'est pas de déterminer les causes et les effets, qui sont imposées par le modèle choisi (celui de la théorie keynésienne de l'équilibre du marché), mais de se mettre dans les conditions permettant de faire une prédiction fiable pour une date T'>T. L'exemple suivant montre la construction de pistes causales à travers une ensemble de variables recueillies sur des enfants et leurs parents après d'une expérience éducative. Le but est ici de démêler l'enchaînement des faits qui peuvent influer sur la construction de l'intelligence d'un enfant et leur importance respective.

Le dernier exemple, tiré du même domaine de la psychologie différentielle que le troisième, ajoute un élément supplémentaire, qui est la présence de variables latentes, nouveaux concepts créés à partir des données.

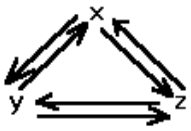
Exemple 1. Une recherche des causes et des effets en sociologie :

[28] Il faut décider si les différentes modalités d'une variable sont fixes ou aléatoires [151], choisir des tests a posteriori plutôt qu'a priori [126], comme détaillé dans la section 2.3.

[29] A peut être composée de plusieurs variables, agissant indépendamment ou en interaction, comme dans la *régression multivariée* et l'*Anova* du modèle linéaire, mais B également comme dans la *régression multidimensionnelle*, et de la *Manova*.

[30] Ces exemples forment une grande partie d'un article paru dans [47].

Hayward R. Alker, dans son ouvrage de "sociologie mathématique", expose différents liens de causalité pouvant intervenir entre 3 variables X, Y et Z de façon totalement abstraite. Il les formalise par le graphique à gauche de la figure 2.7, et par le système d'équations linéaires à droite où X, Y et Z sont des variables aléatoires centrées réduites, et u_1, u_2 et u_3 les résidus du modèle[31]. La signification de la première équation est que Y et Z sont les causes de X [32], et dans le graphique cela correspond aux deux flèches qui arrivent en X. De même pour la deuxième et la troisième qui signifient respectivement que X et Z sont les causes de Y, et que X et Y sont les causes de Z. La coexistence des flèches de X vers Y et de Y vers X signifierait l'existence d'une "causalité réciproque".



$$x + a_{12}y + a_{13}z = u_1$$
$$a_{21}x + y + a_{23}z = u_2$$
$$a_{31}x + a_{32}y + z = u_3$$

FIG. 2.7 – Les relations possibles entre 3 variables et le système linéaire associé.

Selon les valeurs trouvées pour les coefficients de ces équations, la causalité peut s'exprimer différemment. En excluant les causes "réciproques", l'auteur propose cinq types de causalité différents, qu'il représente dans la figure 2.8 de 3 façons chacun ; dans la première colonne le graphe des relations causales, chaque flèche allant de la cause vers l'effet, dans la seconde le système d'équations associé qu'il convient d'établir pour que le modèle soit validé, et dans la dernière la règle que vérifient dans ce cas les coefficients de corrélation linéaire entre les variables prises deux à deux.

Détaillons la façon dont il les utilise pour trancher entre trois hypothèses issues de théories sociologiques expliquant les effets et les causes de l'importance électorale des Noirs aux États-Unis en 1948 (voir figure 2.9). Ces trois théories sont les suivantes :

1. Théorie de l'*oppression par les blancs* : l'augmentation de l'hostilité raciale contre les Noirs et l'existence d'organisations racistes blanches font baisser la participation des Noirs aux élections.
2. Théorie de l'oncle Tom : Une diminution de la participation des Noirs aux élections fait baisser l'hostilité raciale envers eux et le nombre d'organisations racistes blanches.
3. Théorie du Ku Klux Klan : l'apparition d'organisations racistes blanches fait baisser la participation des Noirs aux élections et augmenter l'hostilité raciale envers eux.

Les trois variables X, Y et Z qu'il crée pour mesurer respectivement *l'existence d'organisations racistes blanches*, *l'importance électorale des Noirs*, et *l'hostilité raciale contre les Noirs* sont issues des données socio-économiques de l'époque pour plus de 300 comtés. X est une variable dichotomique notée WRO dont la valeur est 1 ou 0 selon qu'il existe ou non des organisations racistes blanches dans le comté. La valeur de Y pour chaque comté est obtenue en faisant le rapport noté N/WR du nombre d'électeurs noirs du comté sur le nombre d'électeurs blancs.

[31] Dans ce système, les variables aléatoires X, Y et Z peuvent être vues comme 3 vecteurs ainsi que les résidus u_1, u_2 et u_3, les paramètres $a - 12$, $a - 13$, etc. comme des scalaires.

[32] le fait que le coefficient de X est 1 permet d'obtenir une estimation de X à partir des valeurs de Y et de Z une fois les coefficients a_{12} et a_{13} trouvés même s'ils sont nuls, ce qui n'est pas le cas avec les deux autres équations où a_{21} et a_{31} tirée à la lecture du tableau de la figure 2.4. peuvent être nuls. Rappelons que l'hypothèse que l'on teste le plus souvent porte précisément sur leur nullité

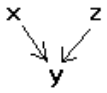
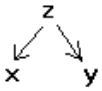
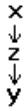
Description	Diagramme	Equations linéaires	Prédictions
Double causalité		$x = u_1$ $a_{21}x + y + a_{31}z = u_2$ $z = u_3$	si $r_{zy} \neq 0$ et $r_{xy} \neq 0$ $\Rightarrow r_{xz} = 0$
Chaîne de causalité		$x + a_{13}z = u_1$ $a_{21}x + y = u_2$ $z = u_3$	si $r_{zx} \neq 0$ et $r_{xy} \neq 0$ $\Rightarrow r_{zy} = r_{zx}r_{xy}$
Double effet		$x = u_1$ $a_{21}x + y = u_2$ $a_{31}x + z = u_3$	si $r_{zx} \neq 0$ et $r_{xy} \neq 0$ $\Rightarrow r_{zy} = r_{zx}r_{xy}$
Corrélation fallacieuse		$x + a_{13}z = u_1$ $y + a_{23}z = u_2$ $z = u_3$	si $r_{zx} \neq 0$ et $r_{zy} \neq 0$ $\Rightarrow r_{xy} = r_{zx}r_{zy}$
Chaîne de causalité avec variable intermédiaire		$x = u_1$ $y + a_{23}z = u_2$ $a_{31}x + z = u_3$	si $r_{zx} \neq 0$ et $r_{zy} \neq 0$ $\Rightarrow r_{xy} = r_{zx}r_{zy}$

FIG. 2.8 – Les divers types de "causalité" pour expliquer que y dépend de x

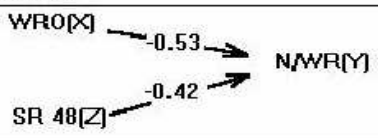
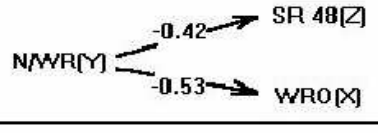
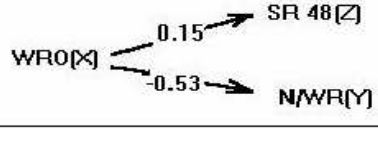
	1 – Théorie de l'oppression par les blancs		
Modèle	Prédiction	Résultat	
	$r_{xz} = 0$	$r_{xz} = 0.15$	L'ajustement est assez bon, on ne peut rejeter totalement
	2 – Théorie de l'oncle Tom		
	$r_{xz} = 0.22$	$r_{xz} = 0.15$	meilleur ajustement, mais avec des signes opposés
	3 – Théorie du Ku Klux Klan		
	$r_{yz} = -0.08$	$r_{yz} = -0.42$	Ajustement assez mauvais, rejeter même si la théorie est assez plausible

FIG. 2.9 – Les 3 théories à l'épreuve des nombres ; comparaison des coefficients de corrélation déduits des modèles (prédiction) à ceux déduits des données (résultats) et conclusion.

Celle de Z est le pourcentage des votes du comté en faveur du "parti pour les droits des États" au cours de l'élection de 1948 et noté SR 48. Et Les valeurs des corrélations trouvées sur les données sont de -0.53 pour r_{XY}, corrélation entre X et Y, de 0.15 pour celle X et Z et de -0.42 pour Y et Z.

La simple lecture des valeurs de ces trois coefficients ne suffit pas à tirer des conclusions sur les liens de cause à effet qui unissent ces variables. Le coefficient de corrélation est seulement l'indicateur d'une liaison linéaire entre deux variables sans qu'il soit possible de lui attacher une interprétation causale. Par exemple l'interprétation causale d'une forte liaison négative entre entre Y et Z sans tenir compte de X peut se faire de deux façons différentes :

1. Z→Y : l'accroissement de l'hostilité raciale contre les Noirs a tendance à faire diminuer la participation électorale de ces derniers
2. Y→Z : l'augmentation de la participation électorale des Noirs a tendance à faire diminuer l'hostilité raciale envers eux

Et la prise en compte de la variable X multiplie les possibilités d'interprétation,

Toutefois si on se limite aux cinq types de causalité de la figure 2.8, nous voyons dans la dernière colonne qu'on peut en déduire des relations de deux types : soit les trois coefficients de corrélation sont tels que l'un est le produit des deux autres, soit l'un est nul alors que les deux autres ne le sont pas. Ce qui permet à l'auteur d'évaluer la qualité de l'ajustement des données au modèle par l'écart des valeurs observées sur les données aux valeurs théoriques imposées par le modèle, celles-ci étant calculées selon cette dernière colonne. Sa démonstration peut se lire dans la figure 2.9 :

1. Théorie de l'*oppression par les blancs* : X et Z causent Y. D'après la dernière colonne du tableau 2.8 relative à la "double causalité", on doit avoir une corrélation nulle entre les deux causes X et Z, soit $r_{XZ} = 0$, les deux autres corrélations étant non nulles, . Ces exigences portant sur des valeurs théoriques, on peut s'attendre à ce que les valeurs observées en diffèrent légèrement. C'est ainsi que l'auteur déclare la théorie plausible, 0.15 étant jugé proche de 0.
2. Théorie de l'oncle Tom : Y cause X et Z. En se référant au tableau de la figure 2.8, on constate que le "double effet" de Y sur X et Z doit avoir pour conséquence que r_{XZ} est le produit de r_{XY} et r_{YZ}. Comme le produit de -0.42 et -0.53 est égal à 0.22, l'auteur conclut que la valeur observée de 0.15 est un bon ajustement de 0.22. Par contre, il fait remarquer que les coefficients de corrélation négatifs obligent à renverser le sens de la cause sur les deux effets : c'est une augmentation (et non une diminution) de la participation des Noirs aux élections qui fait baisser l'hostilité raciale envers eux et le nombre d'organisations racistes blanches.
3. Théorie du Ku Klux Klan : X cause Y et Z. On est encore dans le cas d'un double effet, mais r_{YZ} serait le produit de r_{XY} et r_{XZ}, donc égal à -0.08. L'auteur juge l'ajustement assez mauvais car la valeur observée de r_{YZ} est -0.42

L'auteur justifie ses choix par un raisonnement de type exploratoire basé sur le modèle linéaire également utilisé dans la démarche expérimentale, mais sans évocation de ses diverses conditions de fonctionnement (elles sont détaillées dans l'exemple suivant). Aucun test non plus n'est fait pour juger de la nullité de l'écart entre la valeur observée 0.15 de r_{XZ} et sa valeur théorique de 0 selon le modèle 2 ou de 0.22 selon le modèle 3. L'auteur n'essaie pas de valider à l'aide des statistiques inférentielles (qui seraient ici les tests du modèle linéaire, ou bien le test de l'égalité du coefficient de corrélation linéaire à une valeur théorique donnée) chacune des trois hypothèses données séparément mais de ranger de la plus probable à la moins probable plusieurs hypothèses

issues des courants sociologiques de l'époque. Et il arrive ainsi à convaincre le lecteur qu'il a trouvé parmi les 3 variables X, Y et Z lesquelles sont les causes et lesquelles sont les effets : c'est certainement la participation électorale accrue des Noirs (Y) qui est la cause d'une diminution du racisme envers eux (X et Z).

Exemple 2. Système d'équations d'offre et de demande en économie :

équation d'offre : $Q_t = a\,P_{t-1} + e_{1t}$

équation de demande : $P_t = b\,Q_t^* + e_{2t}$

équation d'équilibre du marché: $Q_t = Q_t^*$

où Q_t est la quantité offerte au cours de l'année t
Q_t^* est la quantité demandée au cours de l'année t
P_t est le prix pratiqué au cours de l'année t
P_{t-1} est le prix pratiqué au cours de l'année t-1

FIG. 2.10 – Modèle économique en 3 équations de l'équilibre du marché d'une denrée

Comme les données sociales de l'exemple précédent, les données économiques peuvent se modéliser à l'aide de systèmes d'équations linéaires. Toutefois, le but des économistes est plus de faire des prédictions sur les phénomènes futurs que de trouver des explications aux phénomènes passés ou actuels. L'importance de la chronologie dans cette discipline fait que le temps est une variable particulière qui, même si elle ne figure pas expressement dans les modèles, doit être prise en compte. Le problème de l'antériorité entre deux évènements se substitue alors souvent à celui de décider lequel est la cause et lequel est l'effet. Nous reprenons l'exemple que propose Luciole Batola page 229 de son ouvrage sur l'économétrie [14] où elle représente l'équilibre de l'offre et de la demande d'une denrée par un système de trois équations dépendant du temps (voir figure 2.10). La première est déterminée par les producteurs/fournisseurs qui proposent une quantité d'autant plus importante de cette marchandise qu'elle s'est vendue cher auparavant. La seconde est déterminée par les acheteurs qui en achètent d'autant plus qu'elle est bon marché actuellement. La troisième est la relation d'équilibre du marché qui fait qu'on en vend autant qu'on en achète.

$$(Q\ P)\begin{pmatrix} 1 & -b \\ 0 & 1 \end{pmatrix} + P_{-1}(-a\ \ 0) = (e_1\ e_2)$$

où a et b sont des scalaires fixes, inconnus, à estimer,
Q, P, P_{-1}, des vecteurs aléatoires, observés, de dimension n-1
e1 et e2 des vecteurs aléatoires, inconnus, à estimer, de dimension n-1
(si le nombre de périodes est n)

FIG. 2.11 – Modèle statistique en une équation matricielle.

La réécriture de ce modèle proposée par l'auteur est dans la figure 2.11, où P est le vecteur des prix observés P_t au cours du temps, P_{-1}, le vecteur de ces prix au temps $t-1$, Q le vecteur des quantités observées Q_t au cours du temps, $(Q\ P)$ la matrice à deux colonnes obtenue en juxtaposant les vecteurs Q et P, et la matrice $(e_1\ e_2)$ étant obtenue par juxtaposition des deux vecteurs résidus au cours du temps. Du fait de la présence de l'indice $t-1$, ces vecteurs sont tous de dimension T-1, où T est le nombre de périodes consécutives pour lequel on a des observations. Ce modèle fait partie des modèles linéaires de l'économétrie appelés modèles "structuraux"

(également appelés modèles d'*équations linéaires simultanées*) qui sont de la forme matricielle[33] $YB + X\Gamma = E$. Dans le cas particulier de son exemple, l'auteur ajoute la condition supplémentaire que les résidus suivent la loi normale, ce qui lui permet de trouver une solution au système d'équations proposé en appliquant le théorème de Gauss Markov qui est le suivant :

Si les hypothèses statistiques suivantes sont vérifiées :

- Il peut se mettre sous la forme[34] $Y = X\Pi + V$.
- L'espérance de toutes les perturbations V est nulle.
- Les perturbations sont homoscédastiques (de même variance) , non corrélées dans le temps.
- Les régresseurs (vecteurs constituants de X) sont observés sans erreurs.
- La matrice X est de plein rang.

alors, le meilleur estimateur de A est $(X'X)^{-1}X'Y$ (où X' est la transposée de X).

Cet estimateur est appelé l'estimateur des MCO[35] (Moindres Carrés Ordinaires), et sa forme simplifiée (calcul de la pente et de l'ordonnée à l'origine de la droite de régression d'une série statistique double) figure dans la plupart des formulaires et calculatrices scientifiques à la suite de celles de la moyenne, de l'écart-type et de la corrélation. Quand les cinq conditions ne sont pas toutes réalisées, il est possible de trouver d'autres estimateurs de A (selon le cas échéant, ce peut être celui des MCD, Moindres Carrés Doubles, MCT, triples, etc..comme indiqué dans cet ouvrage, et dans les ouvrages d'économétrie de base, par exemple [134]).

Exemple 3. les pistes causales en psychologie différentielle :

Les modèles de psychologie différentielle consacrés aux acquisitions de connaissances par l'être humain ont de nombreux points communs avec les modèles socio-économiques. Le temps y est aussi un facteur important, la recherche des causes et des effets également, et à cela s'ajoute la difficulté d'évaluer les variables étudiées. Nous empruntons à Maurice Reuchlin l'exemple (voir figure 2.12) qu'il cite dans son ouvrage traitant des différences individuelles dans le développement cognitif de l'enfant [202]. C'est un modèle d'explication du QI non verbal de 69 enfants mexico-américains de 6 à 8 ans par R.J. McGowan et D.L.Johnson (1984), dans lequel la plupart des variables ont été construites à l'aide de questionnaires, et non mesurées directement. Chaque variable est représentée par une ellipse, et les flèches et nombres associés qui aboutissent à cette ellipse correspondent à une équation de régression. Au membre à gauche de cette équation figure la variable dans l'ellipse. Les flèches aboutissant dans cette ellipse ont partent presque toutes d'autres ellipses. Chacune de ces ellipses fournit la variable qu'elle contient en membre droit de l'équation multipliée par un coefficient qui est le nombre figurant sur la flèche joignant les deux ellipses. Quand une flèche ne part d'aucune forme (ellipse ou autre comme dans l'exemple suivant), la valeur indiquée dessus est alors celle de la variance résiduelle de l'équation. Par exemple, à la variable à droite du graphique, qui est le QI à 7-8 ans, aboutissent 4 flèches qui se transcrivent en l'équation de régression linéaire suivante :

$$QI_{\ 7-8\ ans} = 0.12 \times langage_{\ 3\ ans} + 0.27 \times QI_{\ 3\ ans} + 0.30 \times Home_{\ 3\ ans} + e$$

avec les résidus e vérifiant $var(e) = 0.88$. Si on admet que ces relations sont causales, on peut augmenter le QI des enfants en augmentant le niveau d'éducation des mères (si la variable "home à 3 ans" augmente de 1 point, cela devrait entraîner une élévation moyenne du QI de l'enfant de

[33] Ici Y est la matrice (Q P) de dimension (2,T), B la matrice $\begin{pmatrix} 1 & -b \\ 0 & 1 \end{pmatrix}$ de dimension (2,2), X le vecteur P_{-1} de dimension T, Γ le vecteur ligne (-a 0), et E la matrice $(e_1\ e_2)$ de dimension (2,T).

[34] Cette forme est appelée *forme réduite* du système. Les principes justifiant cette transformation peuvent être trouvés dans [109].

[35] L'auteur nous signale en page 123 de son ouvrage [14] qu'il a été inventé en 1821 simultanément par Gauss et Legendre

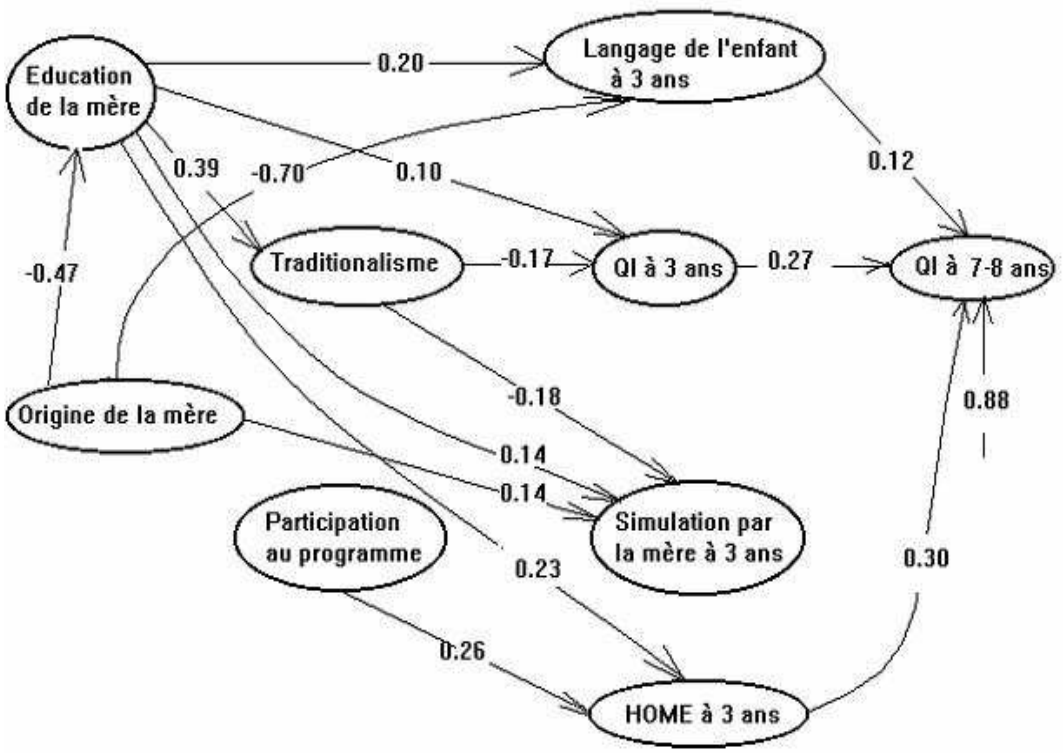


FIG. 2.12 – Un modèle explicatif du QI bâti sur un questionnaire.

0.38 point), en les faisant par exemple participer à un programme de formation ("Participation au programme"). Cette "participation au programme" fait partie des causes indirectes de l'élévation du QI de l'enfant, car elle se fait à travers son influence sur le "home d'enfant à 3 ans", ce dernier faisant quant à lui partie des causes directes.

La construction des variables dans ce modèle s'est faite d'après les réponses à des questionnaires en partie créés pour cet usage. Cette technique de recueil de données se situe entre deux extrêmes. D'un côté, elles sont recueillies à l'issue d'expériences où les valeurs de la plupart des variables sont connues au préalable (ex. : temps qu'il fait), contrôlées (ex. : répartition des sexes dans les groupes) même fixées par l'expérimentateur (groupe témoin/placebo/test), comme dans l'exemple cité dans le cadre du modèle expérimental. De l'autre on les obtient directement auprès d'organismes les produisant (par exemple les instituts officiels de la statistique). La fiabilité des résultats de ce type de modèle construit sur des réponses à des questionaires dépend de la technique de sondage utilisée (le choix des personnes interrogées doit suivre certaines règles de représentativité de la population [7]) ; elle dépend aussi de la qualité des questionnaires [151]. Ce modèle fait partie des "modèles structuraux" des psychologues dans lequel les liens entre variables s'interprètent tous en terme de causalité, leur ensemble formant des "pistes causales". Mais il y en a de plus complexes. C'est le cas du modèle de l'exemple de Frome et al. [89] représenté dans la figure 2.13[36].

Exemple 4. Modèles structuraux avec variables manifestes et latentes : Dans l'exemple précédent des pistes causales (figure 2.12), les variables n'étaient pas différenciées

[36]Nous remercions Tarek Ziadé qui nous a aidé à traduire les termes anglais relatifs au système scolaire anglo-saxon dans leurs équivalents français les plus proches

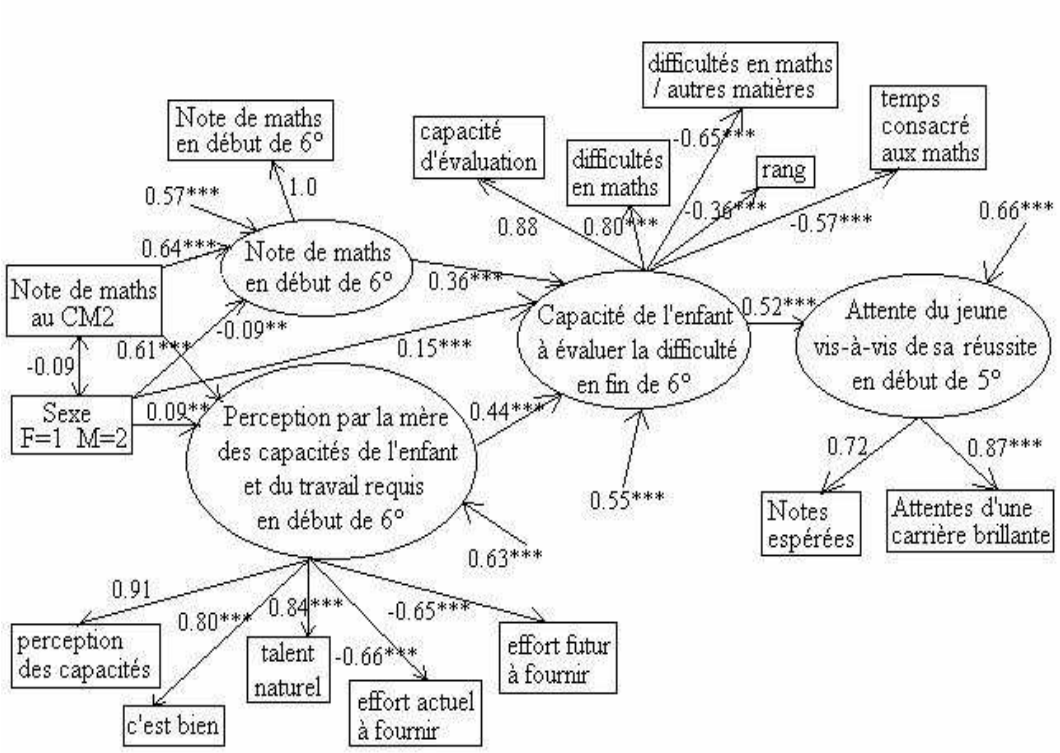


FIG. 2.13 – Un modèle explicatif scolaire

selon leur mode de création, étant produites toutes de la même façon (probablement issues du recodage "manuel" par un expert des réponses à un questionnaire). Dans ce nouvel exemple (figure 2.13), les variables sont représentées dans des formes rectangulaires ou ovales selon leur mode de création respectif. Dans les rectangles, ce sont les variables *manifestes*, créées à partir des questionnaires, de façon directe (comme le sexe, la note de mathématiques en CM2, à gauche de la figure) ou de façon plus élaborée (comme le "talent naturel" ou "l'effort actuel à fournir", en bas de la figure). Les variables *latentes* des formes ovales n'ont pas de valeurs préexistantes au modèle, mais leurs valeurs sont créées lors de l'estimation de ce dernier, en même temps que sont calculées les valeurs des paramètres (les coefficients de régression figurant sur les flèches joignant deux variables, et les variances résiduelles au bout des flèches qui ne partent d'aucune variable).

Ce modèle est la conclusion d'une étude longitudinale sur 914 enfants américains et leurs parents faite par Frome, P.M.et Eccles J.F. afin de mettre en évidence les relations entre les perceptions des parents, celles des enfants et les résultats scolaires de ces derniers. La validité de ce modèle est bonne (Goodness of Fit Index : GFI=0.91), et beaucoup de valeurs sont "très significatives"[37].

La conclusion que l'on tire des nombres figurant dans ce modèle porte essentiellement sur les relations entre les 4 facteurs latents, et les 2 facteurs manifestes (sexe et note de maths en CM2). Par exemple, le facteur latent "capacités de l'enfant à évaluer la difficulté en fin de 6°" est "expliqué" (au sens habituel de la régression linéaire, en suivant la même lecture des flèches et des nombres que dans le graphique de l'exemple précédent) par les deux facteurs latents que sont

[37]Les habitudes en sciences humaines sont de rajouter une, deux, trois étoiles, au coefficient selon que la probabilité qu'il soit nul dans la population est inférieure à 1 chance sur 20, 100 ou 1000.

la "perception par la mère des capacités de l'enfant et du travail requis en fin de 6°" (coefficient de régression de 0.44) et "la note de mathématiques en début de 6°" (0.36), et en moindre mesure par le facteur manifeste qu'est le "sexe de l'enfant" (0.15). D'autres facteurs extérieurs au modèle interviennent également dans la perception par l'enfant des difficultés en fin de 6° puisqu'il reste 55% de la variance non expliquée par les notes de mathématiques en début de 6°.

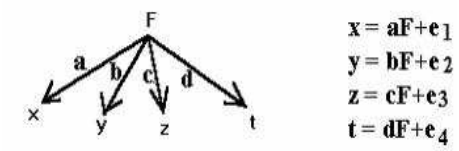


FIG. 2.14 – Une variable latente F obtenue à partir de quatre variables manifestes, toutes ces variables ainsi que les erreurs constituent des vecteurs

L'ajustement de ces modèles structuraux nécessite que soient fournies en entrée la liste des variables manifestes et leurs valeurs (en fait leur matrice de covariance suffit pour la plupart des logiciels traitant ces modèles[38], ainsi que la liste des variables latentes et toutes les équations de régression qui forment le modèle. Ces équations sont de deux types principaux : celles qui unissent les variables latentes aux variables manifestes, et celles qui unissent les variables latentes entre elles. Les premières sont toujours du même type, un exemple en étant donné dans la figure 2.14, où F est une variable latente représentant un facteur commun (si on se place dans le cas le plus courant d'une analyse de type factoriel permettant de faire émerger les facteurs communs aux réponses d'un questionnaire comme par exemple dans [175]) à des variables manifestes, ici au nombre de quatre, x, y, z et t. Les secondes sont des équations du même type que dans l'exemple précédent de pistes causales, contenant en partie gauche une variable latente et en partie droite autant de variables latentes que l'on désire. Ainsi une variable latente peut à la fois être l'effet de certaines variables latentes et faire partie des causes d'autres, ce qui permet de modéliser des situations complexes où des effets s'enchaînent, avec des causes directes, indirectes comme dans l'exemple précédent, voire même des effets de bouclage. A côté de cela, il reste la possibilité d'autoriser des corrélations entre variables (on voit sur la gauche de la figure 2.13 qu'une corrélation a été autorisée entre les deux variables "notes de maths en CM2" et "sexe") qui ne figurent pas dans les liens spécifiés par les équations du modèle, donc parmi les relations de cause à effet qu'on désire établir.

Les conditions que doivent vérifier les données pour qu'on puisse appliquer ce modèle sont encore plus contraignantes que celles des modèles précédents, et même quand on dispose d'un logiciel avec une interface graphique permettant un examen et recodage assisté des variables afin de les normaliser par exemple avant traitement, et de corriger facilement le modèle testé en cas de non convergence, comme c'est le cas pour LISREL [135], il reste d'une utilisation difficile si les hypothèses à tester n'ont pas été spécifiées en détail (notamment, décider si une variable manifeste participe à un facteur latent ou à plusieurs, si on autorise ou non des corrélations entre certaines variables, si on fixe à 1 le coefficient d'une des variables manifestes constituant un facteur latent, si on autorise à un facteur latent d'être à la fois cause directe et indirecte d'un autre, etc.). En effet, plus le nombre de variables manifestes augmente, plus le nombre de

[38] LISREL est un logiciel essentiellement dédié aux équations structurelles (pour une description de ses fonctionnalités, voir [135] et pour sa comparaison avec d'autres [207]), et on peut trouver dans des logiciels généralistes de statistiques comme SAS par exemple des fonctions permettant d'estimer ces modèles.

coefficients à estimer augmente également, et plus il faut de données. Dans l'exemple, il y a 26 coefficients à estimer auxquels s'ajoutent les 14 estimations des variances résiduelles des variables manifestes alors qu'on ne s'intéresse vraiment qu'à ceux qui lient entre elles les variables latentes. La résolution d'un système statistique, donc non déterministe, exige qu'il y ait plus d'un sujet par coefficient à estimer, afin que les coefficients ne soient pas les solutions d'un système d'équations où il y a autant d'inconnues que d'équations [39], sinon la solution devient exacte ou impossible. En augmentant suffisamment le nombre de sujets, on reste protégé de ce genre de problème, même s'il y en a un certain nombre de sujets avec des valeurs identiques. Avec 5 sujets par coefficient, ce modèle de la figure 2.13 nécessite 200 sujets. Kline signale dans le chapitre intitulé "Confirmatory factor analysis and path analysis" de son livre [141] combien ce modèle paraît exigeant en nombre de données, Nunnally ayant estimé en 1978 un ratio de 20 sujets souhaitable par variable, et Loehlin en 1987 qu'un effectif total de 500 sujets n'était pas excessif.

2.3 La signification des variables peut intervenir à chaque étape du traitement

Dans ces deux démarches traditionnelles de raisonnement en siences humaines à partir des données que sont la démarche expérimentale et la démarche exploratoire, nous avons vu que les variables avaient des statuts différents, qui entraînaient des traitements différenciés. Dans la première, la connaissance experte du chercheur n'intervient que lors de la création de l'hypothèse à tester et du montage de l'expérience. Le recueil des données, leur traitement, le test à appliquer, et l'explicitation de la conclusion à en tirer selon le résultat trouvé sont définis de façon détaillée lors de la construction de l'expérience, et ne peuvent être modifiés une fois que l'expérience a commencé sous peine de disqualifier la conclusion. Dans la seconde, la connaissance experte peut intervenir également en amont du processus de recueil de données, comme pour le choix des questions posées, des personnes interrogées dans le cas d'une enquête, ou pour le choix des données récupérées sur Internet. Mais elle intervient principalement lors du traitement, quand il s'agit de choisir le type et l'orientation des liens entre variables ou les différentes contraintes du modèle, (comme dans les exemples de sociologie, d'économie et de psychologie différentielle que l'on vient de voir). Et elle peut encore intervenir lors de la conclusion, pour choisir les causes et effets les plus probables.

Nous décrivons d'abord en détail comment les divers statuts des variables interviennent lors de la construction d'une expérience, car c'est dans ce cadre que la connaissance de l'expert y afférant a été le plus structurée, puis nous montrons aussi sur un petit exemple que même quand cette connaissance n'est pas codifiée dans le modèle, on ne peut s'en abstraire sous peine de tirer des conclusions aberrantes.

2.3.1 Typologie et structure des variables dans le montage d'une expérience

Dans les plans expérimentaux, on sépare la variable "à expliquer" (ou *dépendante*) dont on étudie les variations, des variables "explicatives" (ou *indépendantes* [40]). Ces variables elles-mêmes se subdivisent en "facteurs principaux" qui sont l'objet d'étude et dont l'expérimentateur fait

[39] Une telle situation provoque sous LISREL l'affichage du commentaire ironique "Fit is perfect" et un GFI de 1 !

[40] L'appellation "variables indépendantes" utilisée ici a pour rôle essentiel d'opposer le statut de ces variables que l'expérimentateur peut directement "manipuler" ou contrôler, à celui de la "variable dépendante", à laquelle il n'a pas d'accès direct. Cette "indépendance" ne préjuge pas de leurs possibilités d'interaction sur la variable dépendante

varier les valeurs (on les appelle *manipulés*), des "facteurs secondaires" dont les effets sur la variable dépendante sont sans intérêt pour l'expérimentateur qui les neutralise (on les appelle *contrôlés*). Une fois les effets de ces derniers disparus, les effets des facteurs principaux sont décomposés de diverses façons en s'appuyant sur leur structure.

La neutralisation des facteurs secondaires

Reprenons l'exemple de la section 2.1 qui relatait une expérience imaginée. Son but était d'établir l'effet de la visualisation d'un film sur le sectarisme. A coté du facteur principal V, "visualisation du film" présent dans la section 2.1.2, sont apparus ensuite dans la section 2.1.8 deux autres facteurs qui sont connus pour intervenir dans les expériences portant sur des sujets humains. Le premier est P, "l'effet de la mesure", (proche de l'effet "Placebo" en médecine, ou de l'effet "Participation à une expérience Pilote" en éducation), qui produit une modification sur précisément ce qu'on veut mesurer, et le second est M : "Maturation", qui prend en compte toutes les modifications liées à l'environnement entre la première et la deuxième mesure. Le plan expérimental de Solomon utilisé dans notre exemple contenait 4 groupes de sujets différents G_1, G_2, G_3 et G_4 et produisait 6 mesures O1 à O6.

Nous les rappelons :

- G_1. deux mesures O1 : V=0, P=0, M=0, puis O2 : V=1, P=1, M=1,
- G_2. deux mesures O3 : V=0, P=0, M=0, puis O4 : V=0, P=1, M=1,
- G_3. une mesure O5 : V=1, P=0, M=1,
- G_4. une mesure O6 : V=0, P=0, M=1,

Des comparaisons entre les mesures O2, O5 vérifiant V=1 (film visualisé) et les autres mesures O1, O3, O4, O6 vérifiant V=0 (film non visualisé), permettent de s'assurer que le facteur principal V, visualisation du film, a bien un effet qui ne se réduit pas à celui des facteurs secondaires P et M. Nous renvoyons le lecteur intéresse par le détail des différentes comparaisons permettant d'éliminer les effets de P et de M aux pages 129 à 132 de l'ouvrage de Léon [163]. A côté de ces deux facteurs, d'autres encore peuvent intervenir comme le sexe, l'âge, la catégorie socio-professionnelle, qui sont observables directement, ou comme l'influençabilité, les opinions politiques, religieuses qui nécessitent des investigations plus profondes. La *randomisation*, qui consiste à affecter au hasard chaque sujet à un des quatre groupes, permet de les neutraliser en répartissant leurs diverses modalités de façon proche dans chaque groupe, dès que le nombre de sujets est assez important. D'autres méthodes de contrôle des variables parasites sont relatées dans l'ouvrage de Léon, comme le *contre-balancement* pour contrôler l'effet de l'ordre en cas de succession d'épreuves.

Sur cet exemple, nous avons montré quelques méthodes différentes de contrôle des variables parasites. Trouver toutes les variables parasites et choisir pour chacune la méthode qui permet d'éliminer son effet si elle a déjà été répertoriée, ou créer une nouvelle méthode dans le cas contraire, est une des étapes cruciales du montage d'une expérience. Une autre étape importante est de décomposer les variations de la variable dépendante en autant de parties que d'effets non négligeables dûs aux facteurs principaux. Pour faire cette décomposition de telle façon qu'on puisse mesurer puis tester l'effet de chaque partie, on s'appuie sur la structure qui lie ces facteurs.

Structures de croisement et d'emboîtement

Quand l'expérimentateur agit sur deux facteurs principaux pour faire varier la variable dépendante, le facteur A pouvant prendre a valeurs et le facteur B b valeurs, et s'il peut agir indépendamment sur les deux facteurs, alors il dispose de $a \times b$ valeurs pour les deux. C'est une

structure de *croisement* pour A et B qui se note A*B. Par contre s'il ne peut choisir les valeurs de A qu'une fois connues celles de B car l'ensemble des valeurs de A peut être partitionné selon les valeurs de B, on dit que A est *emboîté* dans B et on note A<B> [41].

Illustrons cela sur l'exemple de la visualisation du film que nous venons de rappeler dans le paragraphe précédent, où on décide de donner à P le même statut de facteur principal que V afin d'étudier son effet sur la variable dépendante. On a une structure de croisement entre les facteurs V et P, car ces deux facteurs ont deux modalités chacun, ce qui en donne quatre pour leur croisement. Les six mesures dont la répartition est la suivante :

- V=0 et P=0 pour O1, O3 et O6
- V=0 et P=1 pour O4
- V=1 et P=0 pour O5
- V=1 et P=1 pour O2

permettent de définir un facteur O emboîté dans le croisement V*P. Si on suppose qu'on a recruté 40 sujets pour cette expérience, et qu'ils ont été répartis à raison de 10 par groupe, le facteur sujet S (qui prend 40 valeurs) est emboîté dans le facteur groupe G, les sujets ayant été répartis par groupes. Mais le facteur S ne peut pas former de structure d'emboîtement avec le facteur O, une partie des sujets faisant l'objet des 2 mesures O1 et O2 pour les uns et O3, O4 pour les autres, contre une seule mesure, O5 ou O6, pour l'autre partie.

Pour conclure sur la combinaison des facteurs de cet exemple au sein de structures d'emboîtement et de croisements, on peut écrire S<G>, O<V*P>, O<G>, mais on ne peut associer pour cet exemple dans une même formule S, G, O, V, P, et encore moins M. Certains plans expérimentaux le permettent, et quand on a la possibilité matérielle de monter une expérience qui se place dans leur cadre, les calculs s'en trouvent facilités. Par exemple, si on se place dans le cadre du modèle linéaire et si on arrive à mettre une structure de *plan complet équilibré* [120] sur les facteurs, on peut alors utiliser l'*orthogonalité* sous-jacente à ce plan pour décomposer la variance de la variable dépendante en somme de parties élémentaires, et tester la nullité de chaque effet séparément. Parmi ces parties élémentaires figurent les *interactions*, qui, si elles s'avèrent significativement non nulles (i.e. leur nullité a été rejetée lors du test), rendent l'interprétation des autres effets délicate. Nous détaillons dans le paragraphe suivant ce qu'est cette interaction qu'on ne peut négliger sous peine de tirer des conclusions erronées.

L'interaction

Supposons que nous voulions éprouver une théorie affirmant qu'une variable A à deux modalités a_1 et a_2 a un effet sur une variable numérique V. On attend par exemple comme effet une valeur de V plus élevée chez les sujets ayant la valeur a_1 que chez ceux ayant la valeur a_2. Si nous disposons de 20 sujets et que nous pouvons attribuer à chacun indifféremment la valeur a_1 de A ou sa valeur a_2, l'expérience la plus simple consiste à les répartir au hasard en deux groupes G1 et G2 de 10 sujets chacun, l'un auquel on applique a_1 et l'autre a_2. Une visualisation de la

[41] Ces deux notations sont largement utilisées dans les écrits francophones traitant des plans d'expériences en sciences humaines [120, 205, 1], mais les logiciels de statistiques généralistes peuvent avoir des notations légèrement différentes. Dans SAS, pour les procédures "proc GLM" (Modèle linéaire généralisé) et CATMOD (toutes les variables, y compris la variable dépendante, si elle est spécifiée, sont à modalités catégorielles) par exemple, l'utilisateur qui veut tester le modèle où Y est une variable dépendante, A, B des variables indépendantes l'écrit $Y = A\ B$, avec un espace entre A et B si A et B sont croisés mais s'il pense que l'interaction entre A et B est inexistante, avec le symbole | s'il désire tester tous les effets y compris l'*interaction* (nous la définissons dans le paragraphe suivant), le symbole $*$ étant réservé à la seule interaction.

valeur de V qui en résulte pour les sujets de chaque groupe[42] permet de s'assurer qu'on a bien le résultat attendu, et la comparaison du quotient des carrés moyens à un seuil de la statistique F de Fisher Snedecor (voir la figure 2.4) permet de conclure [43] par le rejet ou l'acceptation de l'hypothèse nulle H0 : $\mu_1 - \mu_2 = 0$ d'égalité des moyennes théoriques des deux groupes.

Si nous voulons par la même occasion éprouver l'effet d'une autre variable B à deux modalités b_1 et b_2, le plan d'expérience le plus courant consiste à répartir les sujets en quatre groupes équivalents G1, G2, G3 et G4 de cinq sujets, auxquels on applique les modalités respectives a_1b_1, a_1b_2, a_2b_1 et a_2b_2 du croisement de A et de B. Ce plan s'écrit $S_5 < A_2 * B_2 >$. Son utilisation nécessite qu'on précise au préalable l'effet attendu pour B relativement à celui de A en choisissant l'une de ces deux alternatives :

- *l'interaction entre A et B est nulle* : l'effet de B est le même selon les diverses modalités de A.
- *l'interaction entre A et B est non nulle* : l'effet de B diffère selon les modalités de A.

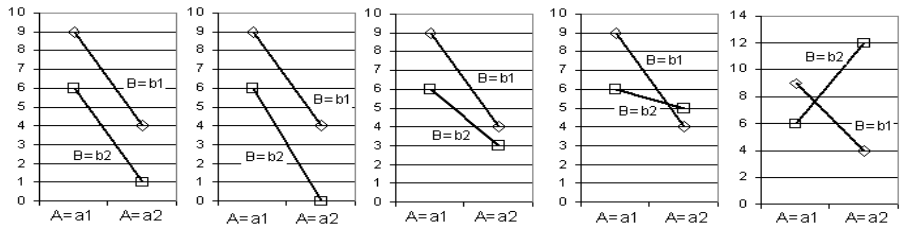


FIG. 2.15 – Interaction de B avec A, nulle à gauche puis en allant vers la droite augmentant l'effet de A, le diminuant, le faisant presque disparaître, le faisant changer de sens.

Si on est sûr que l'interaction est nulle, on peut procéder pour B comme on a procédé pour A. Dans le cas contraire, on doit détailler l'action de B sur V en fonction des modalités de A. Ainsi, l'ajout d'un deuxième facteur dans le modèle augmente sa complexité par l'ajout de calculs en cas d'interaction non nulle, et aussi en cas d'interaction nulle quand on choisit de s'assurer de sa nullité par un test. Mais l'ajout de ce facteur B a un effet beaucoup plus gênant qui est celui de la remise en cause de l'interprétation de l'effet de A sur V faite auparavant. En effet l'interaction entre deux facteurs est une opération *symétrique*, c'est-à-dire que la définition que nous avons donnée pour l'interaction entre A et B peut être lue en échangeant les rôles de A et de B : l'effet de A sur V diffère selon les modalités de B en cas d'interaction non nulle. Face à ces difficultés, l'expert s'aide souvent de graphiques pour pouvoir interpréter tous les effets de façon correcte. La représentation graphique la plus courante s'obtient en représentant les valeurs de V pour chaque croisement de modalités, soit ici pour les quatre groupes. On se contente souvent de ne représenter que les moyennes (appelons m1, m2, m3 et m4 les moyennes respectives dans groupes G1, G2, G3 et G4) comme on peut le voir dans chacun des cinq graphiques de la figure 2.15 avec les modalités d'un facteur (ici A) en axe des x, les valeurs de V en axe des y, les

[42] On peut faire une représentation comme celle de la figure 2.1 correspondant à l'exemple de la section 2.1.2, mais sans joindre les points. Les sujets de nos deux groupes ne sont pas "appariés", c'est-à-dire qu'aucun lien de similitude ne joint un sujet d'un groupe avec un sujet d'un autre groupe, contrairement à l'exemple de la section 2.1.2 où le trait joignait deux mesures prises sur un même sujet.

[43] Les calculs peuvent se disposer dans une tableau selon le même principe que le tableau de la figure 2.3. Toutefois le calcul des SC et des ddl est différent car la structure liant le facteur sujet S à l'autre facteur est ici un emboîtement S<A> (cas des "échantillons indépendants") alors que c'est un croisement S*O (cas des "échantillons appariés") dans la section 2.1.2.

modalités identiques du deuxième facteur (ici B) étant jointes par des traits. Si les deux traits sont parallèles, comme c'est le cas du graphique à gauche de la figure, il n'y a pas d'interaction entre A et B. Dans les autres cas, l'interaction est non nulle. C'est quand il y a croisement entre les deux traits[44], que l'interprétation s'avère délicate pour au moins un des facteurs, voire les deux. C'est le cas des deux graphiques à droite de la figure, le premier étant associé à un effet de B s'inversant selon les modalités de A, le second au même problème à la fois pour A et B.

Pour faciliter l'explicitation de ce problème d'interprétation, nous fixons les moyennes des trois premiers groupes pour V, m1=9, m2=6, m3=4, et nous choisissons les valeurs successives suivantes pour la moyenne du dernier groupe m4= 1, 0, 3, 5 et 12, qui correspondent aux graphiques de la figure 2.15 pris successivement de gauche à droite. Le modèle complet peut s'écrire sous la forme

$$V = \alpha + \beta(A = a_2) + \gamma(B = b_2) + \delta(A = a_2)(B = b_2) + \epsilon$$

qui lie les valeurs V, A et B de tout individu, $A = a_2$ valant 0 ou 1 selon que sa valeur de A vaut a_1 (valeur choisie par défaut) ou a_2, et pareillement pour $B = b_2$, ϵ étant l'écart à la moyenne de son groupe. Avec ces définitions[45], α correspond à la valeur pour laquelle $A = a_1$ et $B = b_1$, c'est-à-dire celle du groupe G1, le groupe par défaut de l'individu, β à l'effet sur V de A quand il passe de a_1 à a_2, sans que B change (il reste à b_1), γ est l'effet sur V de B quand il passe de b_1 à b_2, sans que A change (il reste à a_1), et δ est l'effet d'interaction sur V de A et de B qui se rajoute aux deux effets précédents de A et de B quand A passe de a_1 à a_2 en même temps que B passe de b_1 à b_2.

Ce qui donne l'équation

$$\begin{aligned} V \;=\; m_1 \;& +(m_3 - m_1)(A = a_2) \\ & +(m_2 - m_1)(B = b_2) \\ & +[(m_4 - m_2) - (m_3 - m_1)](A = a_2)(B = b_2) \\ & +\epsilon \end{aligned}$$

soit ici

$$V \;=\; 9 - 5(A = a_2) - 3(B = b_2) + \delta(A = a_2)(B = b_2) + \epsilon$$

avec $\delta \;=\; m_4 - 1$.

Ainsi l'interaction δ est nulle quand m4=1, ce qui correspond bien au graphique de gauche de la figure 2.15, puis il prend les valeurs successives de -1, 2, 4 et 11. Les valeurs de -1 et 2 ne sont pas très gênantes, car elles ne compensent pas les valeurs de β et de γ. Par contre la valeur 4 compense γ, la conséquence en étant que le passage de b_1 à b_2, augmente la valeur de V de γ=-3 si $A = a_1$ et de $\gamma+\delta$=-3+4 donc de 1 si $A = a_2$, et en moyenne, donc si on ne tient pas compte de la valeur de A, de (-3+1)/2, soit -1. en réunissant les groupes G1 et G3, ainsi que les groupes G2 et G4. Ainsi l'effet de B dans ce cas est délicat à interpréter. En moyenne, il peut paraître faiblement négatif, mais si l'on détaille selon les valeurs de A, on peut dire qu'il est fortement négatif pour A=a_1 et faiblement positif pour A=a_2. Ce sont donc deux actions contraires qu'il a sur V. Par contre, comme δ ne compense pas β, le problème est moins important pour A, bien que l'effet de A sur V soit fortement négatif pour B=b_1 et faiblement négatif pour B=b_2. Et dans le dernier cas, correspondant à une valeur de δ=11, en moyenne A et B ont des effets positifs sur

[44] Si on choisit d'inverser les modes de représentation des deux facteurs (ici en mettant les modalités de B sur l'axe des x, et en joignant les modalités identiques de A), le parallélisme des deux traits reste, mais pas toujours le croisement. On peut le vérifier en prenant le cas de l'avant dernier graphique à droite de la figure avec m1=9, m2=6, m3=4 et m4=5.

[45] La formule la plus utilisée diffère de celle-ci : ici les valeurs par défaut sont celles de G1 alors qu'habituellement ce sont les valeurs moyennes

V, assez faible pour A (0,5), mais assez important pour B (2,5), alors que relativement à chaque valeur de l'autre variable ils ont de forts effets qui se contredisent.

Pour tester la nullité de l'interaction, il faut contrôler que pour chaque groupe, les écarts des valeurs des individus à la moyenne de leur groupe suivent une loi normale d'espérance nulle et de même variance. Puis on décompose la somme des carrés des écarts à la moyenne en 4 parties, l'une due à l'interaction, la deuxième à l'effet *principal* de A (c'est proche de ce qu'on a appelé l'effet *en moyenne* précédemment), la troisième à l'effet principal de B, et la résiduelle. Ce qui produit un tableau proche de celui de la figure 2.4, mais comportant deux lignes de plus. Chaque effet peut alors être testé séparément. Toutefois, en cas d'interaction significativement non nulle, les tests des effets *principaux* de A et de B ne sont pas très informatifs, car ce sont les effets d'un facteur pour chaque modalité de l'autre qui sont interprétés.

En cas de croisement de plus de deux facteurs, le tableau d'ANOVA (ANalysis Of Variance, cf. [126, 59]) peut contenir jusqu'à huit lignes : une ligne pour tester l'interaction des trois facteurs ensemble, trois lignes pour tester l'interaction des trois facteurs pris deux à deux, trois lignes pour tester les trois effets principaux, et une ligne pour les résidus. On suit le même ordre que dans le cas du croisement de deux facteurs : si l'interaction de niveau trois est non nulle, on évite de tester les interactions de niveau 2 et les effets principaux. mais on interprète les interactions de niveau deux selon les modalités du facteur omis. Ces effets sont parfois difficile à exprimer simplement, les graphiques n'étant pas d'un grand secours. Et les effets de chaque variable en fixant les modalités de l'une ou de l'autre des variables ne sont pas toujours faciles à explorer non plus. Si l'interaction de niveau 3 est nulle, on teste alors chaque interaction de niveau deux, en procédant comme précédemment.

On voit là toute la difficulté de l'interprétation des effets d'un facteur qui est à la merci de l'apparition d'un nouveau facteur avec lequel il pourrait interagir.

Conclusion

Nous venons de décrire, parmi beaucoup d'autres possibles, trois types de structures inter-variables qui ne peuvent être déterminés que par une expertise sur le problème posé. Parmi les autres on peut trouver les *contrastes* (on peut en voir un exemple dans la section 2.6.2), les *modalités aléatoires* [236, 1]. Le calcul du F de Fisher-Snedecor est déterminé par leur connaissance. Les ignorer peut entraîner des résultats erronés. Le modèle linéaire n'est pas le seul à provoquer ces problèmes [46], mais c'est celui qui permet le mieux de les formuler dans les statistiques classiques.

2.3.2 De l'importance de l'expertise dans le processus de fouille de données

L'exploration aveugle des données peut déboucher sur des conclusions fantaisistes. Dans l'exemple de notre petite expérience, on aurait très bien pu envisager que l'attitude sectaire soit plus importante chez les personnes âgées que chez les jeunes (ou inversement), ou bien que la visualisation du film n'ait pas la même influence selon les tranches d'âges. Ce qui pourrait s'écrire pour la première par "si l'âge augmente alors l'attitude sectaire augmente", et pour la seconde par

$$Sect = Age + Visual + Age * Visual$$

.

[46] On les rencontre également dans tous les modèles dérivés du modèle linéaire, dont le modèle loglinéaire [181], que nous avons comparé aux règles d'association dans l'annexe C de ce mémoire.

Par exemple si on établit l'équation du modèle linéaire

$$Sect = 4,3 + 0,5.Age - 2,1.Visual + 1,3.Age * Visual + \epsilon$$

,

Age étant une variable binaire valant 1 ou 0 selon que la personne est âgée ou non, et $Visual$ prenant la valeur 1 si la personne a vu le film et 0 sinon. Les valeurs des 4 coefficients, dans le cas idéal où on dispose d'autant de personnes, pour chacun des 4 cas (visualisation ou non, jeune ou âgé) ont été obtenues par simples moyennes, et si toutes les conditions statistiques sont réunies (représentativité de l'échantillon, normalité des résidus, etc.) en cas de non-nullité de ces coefficients, la conclusion s'écrit aisément : "le sectarisme est plus important chez les personnes plus âgées, mais on peut le faire diminuer par la visualisation d'un film, cette diminution étant plus importante chez les jeunes que chez les personnes âgées".

Imaginons maintenant que les données ne soient pas issues d'une expérience, mais par exemple récupérées à l'issue d'une enquête, et qu'on dispose de ces trois variables, Age, $Sect$, $Visual$ parmi d'autres. L'hypothèse n'est plus préalable aux données, mais une fois celle-ci établie (avec plus de difficultés, on l'a vu, pour cette méthode exploratoire que pour la méthode expérimentale), l'interprétation ne pose pas non plus de difficultés si on connaît la signification des variables, et si on dispose de quelques connaissances sur les théories existantes. Par contre, si on ne connaît pas leur sens, on peut en déduire des règles surprenantes, comme par exemple "le sectarisme et la visualisation du film ont augmenté l'âge des spectateurs". Il y a loin de la covariation à la causalité !

2.4 La liaison entre variables en statistiques descriptives

2.4.1 Un exemple de liaison entre deux variables

Appelons A et B ces deux variables, dont on connaît la valeur pour n objets. Par exemple prenons comme objets des sphères métalliques, et comme variable A leur poids et B leur volume. Si on se place d'abord dans le cas où toutes les sphères ont la même composition en métal, on établit facilement un modèle de la liaison entre A et B : c'est une équation de la forme $A = \rho B$. En effet, cette équation fait partie des lois de la physique classique qui est déterministe, contrairement à la physique quantique. On peut donc l'obtenir, en principe, à partir des valeurs d'une seule sphère. En fait, les imprécisions possibles des mesures font qu'il faut envisager des écarts à l'égalité. C'est leur prise en compte qui est à l'origine de la construction d'une méthodologie de calcul des incertitudes en physique. Si on admet à présent que les sphères peuvent être de compositions variées, le même modèle peut encore être envisagé, car il signifie que plus les sphères sont grosses, plus elles sont lourdes. Toutefois le modèle devient statistique par l'ajout à l'équation d'un terme d'erreur pour prendre en compte l'existence éventuelle de deux sphères de même volume mais qui n'auraient pas le même poids du fait de leurs compositions différentes. Mais dans ce cas, l'écriture du modèle sous forme de l'équation $A = \rho B + \epsilon$ n'est plus justifiée, la recherche de la valeur de ρ n'ayant plus le sens physique qu'elle avait dans le cas précédent (c'était le poids volumique, valeur caractéristique de la composition du métal). On préfère établir de façon plus générale la présence d'une relation linéaire entre les variations de A et celles de B autour de leurs moyennes respectives, ce qui se fait en calculant le coefficient de corrélation linéaire de Bravais-Pearson entre les deux variables.

2.4.2 Le coefficient de corrélation linéaire de Bravais-Pearson

Sa formule est $r_{A,B} = \frac{\sum_{i=1}^{n}(A_i-\overline{A})(B_i-\overline{B})}{\sqrt{\sum_{i=1}^{n}(A_i-\overline{A})^2}\sqrt{\sum_{i=1}^{n}(B_i-\overline{B})^2}}$, autrement dit le quotient de la covariance de A et B et des écarts-types de A et de B. La valeur absolue de ce coefficient de corrélation est de 1 quand pour chaque objet i, l'écart de A entre sa valeur A_i et la valeur moyenne $\overline{A}$ est proportionnel à l'écart correspondant de B, comme c'était le cas dans l'exemple précédent, avec le coefficient de proportionnalité égal à ρ, n étant le nombre d'objets. Quand la valeur est nulle, cela signifie que les variations de A au-dessus de sa moyenne sont associées à des variations de B qui peuvent être au-dessus comme au-dessous de sa moyenne et que les covariations se compensent. Quand il est positif, cela indique que les covariations de A et de B ne se compensent pas, et que leur tendance est d'aller dans le même sens, et ceci d'autant plus que le coefficient de corrélation linéaire est proche de 1. Et quand il est négatif, c'est qu'elles varient de façon similaire, mais en sens contraire. Nous en donnons quelques exemples dans la figure 2.16, où les valeurs de la variable A sont données pour 10 objets par ordre décroissant, ce qui fait que les écarts à la moyenne mA=0.1 sont positifs pour les 5 premiers objets et négatifs pour les 5 autres. Dans les colonnes suivantes, sont données les valeurs de quatre autres variables B, C, D et E et de leurs écarts à leur moyenne. Les écarts qui sont dans le sens contraire à ceux de A sont passés en gras. Le nombre de ces derniers et leur importance globale croît quand on passe de B à C, D puis E, en même temps que la corrélation diminue en passant de 0.445 pour AB à 0 pour AC, -0.053 pour AD et -0.806 pour AE.

	A	A-mA	B	B-mB	C	C-mC	D	D-mD	E	E-mE
o1	1.5	1.4	2.1	0.3	2.1	0.3	3	1.2	1.2	**-0.6**
o2	1	0.9	2.1	0.3	1.3	**-0.5**	2	0.2	1.4	**-0.4**
o3	0.7	0.6	1.8	**0**	2	0.2	1.6	**-0.2**	1.6	**-0.2**
o4	0.2	0.1	1.2	**-0.6**	1.1	**-0.7**	1.1	**-0.7**	1.1	**-0.7**
o5	0.2	0.1	2.6	0.8	3	1.2	0.4	**-1.4**	2	0.2
o6	0	-0.1	2.5	**0.7**	0.7	-1.1	0.5	-1.3	1.7	-0.1
o7	-0.3	-0.4	1.1	-0.7	1.8	**0**	1.8	**0**	1.8	**0**
o8	-0.5	-0.6	1.6	-0.2	2.5	**0.7**	2.5	**0.7**	2.5	**0.7**
o9	-0.8	-0.9	1.8	**0**	2.1	**0.3**	2.1	**0.3**	2.1	**0.3**
o10	-1	-1.1	1.2	-0.6	1.4	-0.4	3	**1.2**	2.6	**0.8**
Moyenne	0.1	0	1.8	0	1.8	0	1.8	0	1.8	0

FIG. 2.16 – Comparaisons entre les variations de B, C, D et E et celles de A

Le coefficient de corrélation de Bravais-Pearson est très utilisé pour évaluer la présence et la force d'une relation entre les variations de deux variables, mais il n'est pas toujours approprié. Sa popularité s'explique par le fait qu'en absence d'information a priori sur le lien possible entre deux variables, la linéarité est examinée en premier car c'est la plus simple des relations de dépendance. Et ce coefficient est une bonne mesure de l'adéquation du lien à la linéarité. Si on représente graphiquement chaque objet dans le plan des deux variables par un point ayant pour coordonnées les valeurs de l'objet aux variables, on obtient un nuage de points qui a une allure oblique (montante ou descendante) d'autant plus linéaire que le coefficient de corrélation est proche de 1 ou de -1. Quand il atteint une de ces deux valeurs extrêmes, les points sont tous alignés sur une droite oblique passant par le point ayant pour coordonnées les valeurs moyennes des deux variables, dont l'équation produit une relation linéaire entre A et B, comme celle que nous avions précédemment entre le poids et le volume des sphères d'une même composition métallique. Dans ce cas le modèle se réduit à cette équation, sans qu'on ait besoin de lui ajouter

un terme d'erreur. Par contre, quand le coefficient de corrélation a sa valeur absolue qui diminue, à l'équation de la droite s'ajoute un terme d'erreur dont l'amplitude de variation augmente en conséquence, et ce terme indique comment le nuage de points est éloigné de la droite. Mais le fort éloignement d'une droite oblique indiqué par une valeur faible (en valeur absolue) du coefficient de corrélation linéaire peut exprimer une absence totale de liaison, ou alors une forte liaison non linéaire. Dans la figure 2.17 sont représentés les nuages de points respectifs de A avec B, C, D et E correspondants aux valeurs du tableau de la figure 2.16. Si les corrélations assez importantes de A avec B ($r = 0.445$) et avec E ($r = -0.806$) sont bien associées à des nuages assez linéaires, celles très proches de zéro de A avec C et de A avec D correspondent pour la première à une absence de liaison visible entre A et C, et pour la seconde à une assez forte liaison parabolique entre A et D.

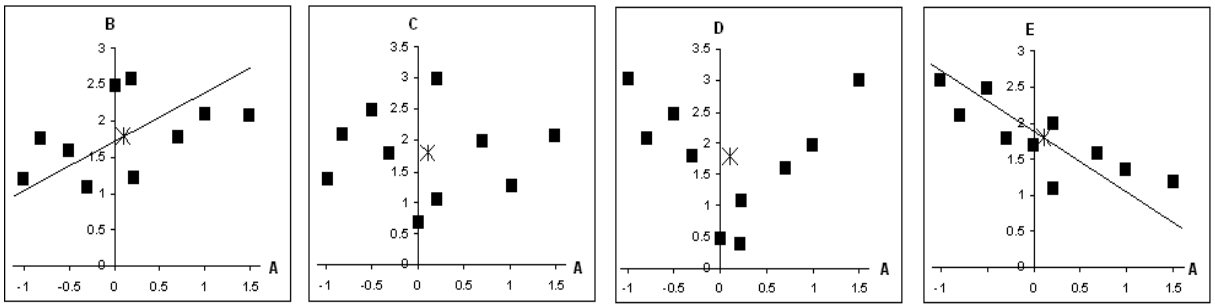


FIG. 2.17 – Nuages de points correspondants au tableau de la figure 2.16

C'est là la première limite de son utilisation, qui est d'autant plus délicate qu'il est proche de zéro. L'absence de lien linéaire (nous verrons plus loin que le mot linéaire peut avoir de nombreux sens) apporte en effet peu d'information sur l'existence d'un lien quelconque quand les données ne suggèrent pas spécialement la linéarité. Pour définir l'ICA (Analyse en Composantes Indépendantes), Hérault J., Jutten C. and Ans B., [117] ont utilisé non seulement la nullité de la statistique d'ordre 2 qu'est la covariance pour établir l'absence de lien mais encore celles de toutes les statistiques d'ordre supérieur. Avec une telle définition de l'indépendance, la liaison parabolique entre les deux variables correspondant au troisième graphique est mise en évidence alors qu'elle ne l'est pas avec la simple covariance.

Une deuxième limite est l'interprétation qu'on peut donner à sa valeur quand elle est élevée : trouver une corrélation de 1 ou de -1 entre deux variables définies sur deux objets n'a rien de remarquable, dans la mesure où par deux points distincts passe une seule droite, c'est plutôt le contraire qui est surprenant. Par contre une corrélation de 1 ou de -1 entre deux variables définies pour plus de deux objets est plus inattendue, car cela signifie qu'on peut trouver une droite à laquelle les points représentant ces objets appartiennent tous. La signification de l'importance de la liaison linéaire doit donc s'appuyer sur d'autres éléments que la seule valeur de ce coefficient, par exemple en prenant en compte l'importance du nombre d'objets. Une troisième limite de ce coefficient est qu'il s'utilise pour des données numériques quantitatives, pour lesquelles la proportion des écarts a un sens (échelles de rapport). Pour les données d'autres types, il convient de le remplacer par d'autres coefficients. Une quatrième limite est qu'un coefficient de corrélation élevé entre deux variables prises en dehors du contexte, et notamment sans envisager les autres variables peut être artificielle.

2.4.3 Les autres coefficients de liaison

Si les données sont ordinales, le coefficient de rangs de Spearman se calcule avec la même formule que le coefficient de corrélation de Bravais-Pearson en remplaçant les valeurs par leurs rangs. On peut aussi utiliser le coefficient τ de corrélation de rangs de Kendall calculé à partir de toutes les paires d'objets en comparant leurs différences de valeurs pour chacune des deux variables (les paires sont dites concordantes si les différences sont de même signe, et la valeur du coefficient est proportionnelle à l'écart entre le nombre n_c de paires concordantes et le nombre n_d de paires discordantes : $\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$). Ce dernier coefficient est tout à fait adapté aux données de jugement, si les deux variables correspondent à une notation des objets par deux juges. Ces deux coefficients s'interprètent de façon similaire au coefficient de corrélation de Bravais-Pearson : ils sont nuls en cas d'absence de liaison, et dans le cas contraire, la liaison est d'autant plus forte que leur valeur est élevée ; si leur valeur est positive, c'est que les variations des deux variables vont dans le même sens, sinon qu'elles s'opposent.

Pour les données catégorielles, le coefficient de liaison le plus courant est le coefficient du "Chi2 de Pearson", noté également Khi2 ou χ^2. Son calcul se fait à partir des effectifs des sujets répartis selon le croisement des modalités. Nous le décrivons dans le paragraphe de la partie 2.5.2 qui traite du test du Chi2. En effet, sa valeur est difficilement appréciable sans test car elle est d'autant plus grande que les variables sont liées mais aussi que l'effectif total est grand. Il existe des variantes de formules qui le corrigent. Pour les données catégorielles, dans le cas où elles sont binaires, le coefficient de *corrélation tétrachorique* peut être obtenu par la même formule que le coefficient de corrélation linéaire, et il s'interprète de la même façon que ce dernier [47] . Toutefois il est plus facile de le calculer d'après un tableau de contingence, comme les autres coefficients de liaison des données catégorielles, sa formule étant $\phi = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2}}}$ [48].

La liaison entre une variable quantitative et une variable ordinale n'ayant qu'un nombre restreint de modalités peut s'établir à l'aide du coefficient η^2.

2.4.4 Conclusion sur les coefficients de liaison entre deux variables

Avec ces coefficients, on s'affranchit de la troisième limite du coefficient de Bravais-Pearson. Les statistiques inférentielles aident à dépasser la deuxième limite, qui est l'interprétation du sens d'une valeur élevée d'un coefficient de liaison. On peut également dépasser la première limite, qui est celle d'une ambiguïté d'interprétation en cas de valeur faible de ce coefficient grâce aux définitions détaillées de nombreuses variétés de dépendance linéaire et même non linéaire, qui ont fait l'objet de développements dans le cadre des statistiques inférentielles (voir la section 2.5). Nous allons montrer maintenant l'interprétation de la liaison entre deux variables relativement aux autres variables, afin de nous affranchir de la quatrième limite, c'est-à-dire de pouvoir établir un lien entre deux variables, "toutes choses égales par ailleurs", puis de pouvoir établir des groupes de liens.

[47] Pour que le signe du coefficient soit interprétable de la même façon, il faut toutefois que les codes 0 et 1 des deux variables soient attribués selon la même logique, comme 0 : 'absence' et 1 : 'présence' proche de la logique numérique du coefficient de corrélation linéaire

[48] Notation : si les deux variables A et B ont chacune deux modalités (A1 et A2, B1 et B2), n_{ij} est le nombre d'objets qui ont les valeurs respectives Ai de A et Bj de B, soit A1 et B2 pour n_{12}. Si l'un des deux indices est remplacé par un point, c'est qu'on "somme" sur cet indice en prenant la valeur marginale : par exemple $n_{1\cdot}$ est la somme de n_{11}, nombre d'objets vérifiant la modalité A1 de A et la modalité B1 de B, et de n_{12}, nombre d'objets vérifiant la modalité A1 de A et la modalité B2 de B.

2.4.5 La liaison entre deux variables conditionnellement aux autres variables

Le calcul du coefficient de corrélation entre deux variables A et B se fait dans le but d'expliquer ou de prédire les variations d'une variable par celles de l'autre. Par exemple en sociologie, si la variable B est le taux de suicide, le taux de mort-nés, d'illettrés, la recherche d'une corrélation entre A et B a pour but d'agir sur A pour faire diminuer les problèmes sociaux représentés par B. La corrélation sert alors à une recherche de causes. On fait aussi ce travail de recherche de corrélations en écologie ou en économie pour prédire par exemple les catastrophes naturelles, les montées ou baisses de cours d'une action, afin de s'en prémunir pour les premières et d'optimiser la gestion d'un portefeuille d'actions pour la seconde. L'action envisagée dans les deux cas n'est pas la même. Dans le premier cas, on agira sur la valeur de A pour modifier celle de B, et dans le second, on observera les variations de A, afin d'agir avant ou au moment où B va varier, cette action ne se faisant ni sur A, ni sur B. Si le choix est possible entre deux variables A, l'une très fortement corrélée à B, et l'autre faiblement corrélée à B, on choisira la première pour prédire B, mais la seconde pour expliquer B si on pense que la faible corrélation de celle-ci avec B peut s'interpréter en terme de causalité de A vers B alors que la forte corrélation de la première avec B ne peut pas être interprétée en ces termes. Une forte corrélation de ce type se produit par exemple quand A et B ont une même cause C. Quand on dispose des valeurs de nombreuses autres variables que A et B pour les objets, le coefficient de corrélation partielle ([126, 201]) permet, en cas de doute sur le sens à donner à une corrélation élevée entre A et B, de s'assurer qu'elle n'est pas due aux autres variables.

2.4.6 Le groupement d'un grand nombre de variables à partir de leur liaison 2 à 2

Il est possible à partir des liaisons deux à deux d'un grand nombre de variables, qu'elles soient liées par des coefficients de corrélation, des distances ou des similarités de construire de nouvelles variables numériques par *analyse factorielle*, ou variables catégorielles par *classification non supervisée* sur les variables [19].

On reviendra dans le chapitre 4 sur ces méthodes et leurs limites.

2.5 La liaison entre variables en statistiques inférentielles

Pour l'établissement des liens entre les deux variables, la démarche classique des statistiques inférentielles diffère de celle des statistiques exploratoires par la mise au point préalable du modèle. Par exemple, pour évaluer l'effet d'une variable A sur une variable B, on prépare à l'avance les conditions de l'expérience, c'est-à-dire l'instrument pour mesurer les variations de A et de B (on peut évaluer l'importance du stress par des mesures physiologiques, par un questionnaire d'auto évaluation, par un examen clinique des symptômes, on peut évaluer l'importance des capacités cognitives mobilisables par une série de petits problèmes logiques à résoudre), les méthodes précises pour faire varier A (on peut augmenter la difficulté de reconnaissance auditive d'une phrase en rajoutant du bruit), et les liens attendus entre les variations de A et de B (on peut s'attendre à ce que l'augmentation du stress entraîne une diminution des capacités cognitives mobilisables) selon le modèle linéaire par défaut $B = aA + b + \epsilon$, pour lequel les valeurs de ϵ suivent une loi de probabilité donnée et la valeur de a appartient à un intervalle donné fourni par l'utilisateur du modèle (la valeur de b ne présente pas d'intérêt en général, comme nous l'avons vu précédemment). Puis on choisit au hasard des sujets, on réalise l'expérience, on effectue les mesures de A et de B, on calcule la valeur de a, les valeurs de ϵ que l'on confronte aux résultats

attendus, et on conclut par la présence ou l'absence de l'effet attendu de A sur B. La théorie n'affirme pas que les conclusions tirées sont justes, mais elle fournit une évaluation du risque de se tromper selon la conclusion choisie. L'équation linéaire $B = aA + b + \epsilon$ utilisée dans cet exemple est probabiliste dans la mesure où les sujets sont choisis au hasard et où les résidus suivent une loi de probabilité donnée. Nous allons décrire la "mise en probabilité" que requiert l'utilisation du modèle linéaire des statistiques pour évaluer la liaison entre deux variables, les types de calculs qui sont faits selon les différents cas.

2.5.1 Les hypothèses probabilistes des statistiques inférentielles

Un jeu d'hypothèses.

Selon F. Bavaud [16], un modèle probabiliste contient, par opposition à un modèle déterministe :

- un Univers (ou population) Ω, qui constitue l'ensemble de tous les événements élémentaires (ou individus) ω possibles, cet ensemble étant fini ou infini.
- des variables aléatoires $X_i(\omega)$ dont les valeurs dépendent de l'événement ω.
- une distribution de probabilité d'évènements sur Ω, définissant les probabilités d'apparition des évènements élémentaires. Cette distribution est discrète ou continue.

Il permet de déterminer la distribution des valeurs des variables aléatoires X_i, qui est la probabilité $P(X_i = x)$ que la variable X_i prenne la valeur x.

Ceci montre la rigueur du formalisme nécessaire pour pouvoir faire des tests sur les liaisons entre variables, inutile dans l'approche descriptive précédente. En statistiques inférentielles on veut pouvoir prouver à ses pairs scientifiques qu'une liaison est "significativement" non nulle, c'est-à-dire qu'elle ne peut pas être due au hasard, alors qu'en statistiques descriptives seule l'appréciation de l'utilisateur distingue les valeurs importante des valeurs négligeables.

Un exemple.

Prenons l'exemple d'un sondage d'opinion. S'il se fait chez des personnes de 20 à 30 ans résidant en France, Ω représente l'ensemble des personnes de cet âge. Si chaque personne a la même probabilité d'être sondée, la distribution de probabilité est la distribution uniforme. Les variables aléatoires peuvent être l'âge, le sexe et les réponses aux diverses questions. Si la variable A correspond à la question "Irez-vous voter aux prochaines élections présidentielles ? ", la distribution de probabilité de A est formée par exemple des 3 valeurs Proba(A= "oui"), Proba(A= "non"), Proba(A= "ne sait pas"). Si la variable B est la réponse à la question "Combien pensez-vous qu'il y aura de candidats au premier tour ?", la distribution de probabilité de B est donnée par les P(B=n) où n est un entier positif. On peut également se donner la loi de distribution de probabilité du couple de variables (A,B), appelée distribution multivariée (ou plutôt bivariée car on a ici deux variables), qui contient par exemple P(A= "oui" et B=5).

Des postulats nuancés.

En théorie, il est très simple d'obtenir la distribution de probabilité des variables, par exemple P(A="oui") s'obtient en sommant les probabilités des personnes ayant répondu "oui", ce qui revient, en cas de répartition uniforme, à calculer la proportion des personnes ayant répondu "oui". En pratique il est souvent impossible de disposer de ces informations pour toute la population, notamment quand sa taille est importante. On prend donc un sous-ensemble restreint d'individus

de cette population, chacun étant choisi au hasard selon la loi de probabilité définie sur la population. C'est ce qu'on appelle "tirer un échantillon représentatif de la population". La théorie de l'échantillonnage ([7, 183]) propose diverses stratégies de tirage d'un échantillon pour lui assurer une bonne représentativité. Du fait des fluctuations d'échantillonnage, les probabilités des variables calculées sur deux échantillons distincts diffèrent entre elles, et donc des valeurs qu'on aurait pu trouver sur la population. On parle alors de valeurs empiriques (ou observées) quand les calculs sont faits à partir de l'échantillon et de valeurs théoriques (ou attendues) pour celles de la population, qui ne peuvent pas être connues, mais seulement estimées. On pourrait penser que cette différence n'existe plus dans le cas d'un "échantillon exhaustif", c'est-à-dire quand on dispose des valeurs des variables pour tous les individus de la population, et que celles-ci ne sont plus aléatoires, mais déterministes. F. Bavaud ([16]) signale que dans de nombreuses branches des sciences, on choisit de garder dans ce cas le cadre d'un modèle probabiliste, et non déterministe, la valeur d'une variable pour un individu donné étant considérée comme aléatoire du fait des erreurs de mesure, imprécisions, fluctuations possibles, et il qualifie les variables "d'intrinsèquement aléatoires", pour les différencier des variables dont la nature aléatoire vient du processus de sélection d'un échantillon. Excepté pour les sondages, on se place souvent dans ce second cadre théorique en sciences humaines, comme par exemple en psychométrie où on distingue le "score vrai" du "score apparent" d'un sujet, sans disposer toutefois en général d'un échantillon exhaustif. Il convient alors de prendre un certain nombre de précautions si on veut généraliser les résultats obtenus à toute une population. On montrera sur un exemple le montage d'une expérience prenant ce type de problème en compte.

2.5.2 L'indépendance entre deux variables

La définition de l'indépendance.

On dit que deux variables aléatoires X et Y sont indépendantes si pour tout couple de valeurs x et y, la probabilité conjointe P(X=x et Y=y) est égale au produit des probabilités P(X=x) et P(Y=y). Ces deux dernières probabilités sont appelées probabilités marginales car on ne fixe qu'une des deux variables, l'autre pouvant prendre toutes les valeurs possibles. Cela revient à la condition que les distributions de probabilité d'une variable "conditionnellement" aux différentes valeurs de l'autre sont identiques. Si on reprend l'exemple du sondage, où A est l'intention d'aller voter et B le nombre de candidats attendu, l'indépendance entre A et B signifie que la distribution du nombre de candidats attendu B est la même pour les personnes qui ont répondu "oui" à la question A, que pour celles qui ont répondu "non", et celles qui ont répondu "ne sait pas", ces trois distributions de B "conditionnelles" à A sont identiques entre elles, et à la distribution marginale de B. Et inversement la distribution de probabilité marginale de la variable A, c'est-à-dire l'ensemble des 3 valeurs des probabilités de répondre "oui", "non", et "ne sait pas" sans tenir compte des réponses à la question B est identique aux lois de probabilité de A conditionnellement aux valeurs des réponses à la question B, c'est-à-dire à l'ensemble des 3 valeurs des probabilités de répondre "oui", "non", et "ne sait pas" pour ceux qui pensent qu'il y aura 2 candidats, ceux qui pensent qu'il y en aura 3, et ainsi pour chaque nombre de candidats envisagé.

Si on reprend l'exemple de la liaison parabolique, dans la figure 2.18, on voit qu'elle diffère de l'indépendance.

Comment doit-on établir l'indépendance.

D'après la définition, pour établir l'indépendance entre 2 variables X et Y, il faudrait calculer pour toutes les valeurs (x,y) du couple (X,Y) les différences entre la probabilité conjointe et

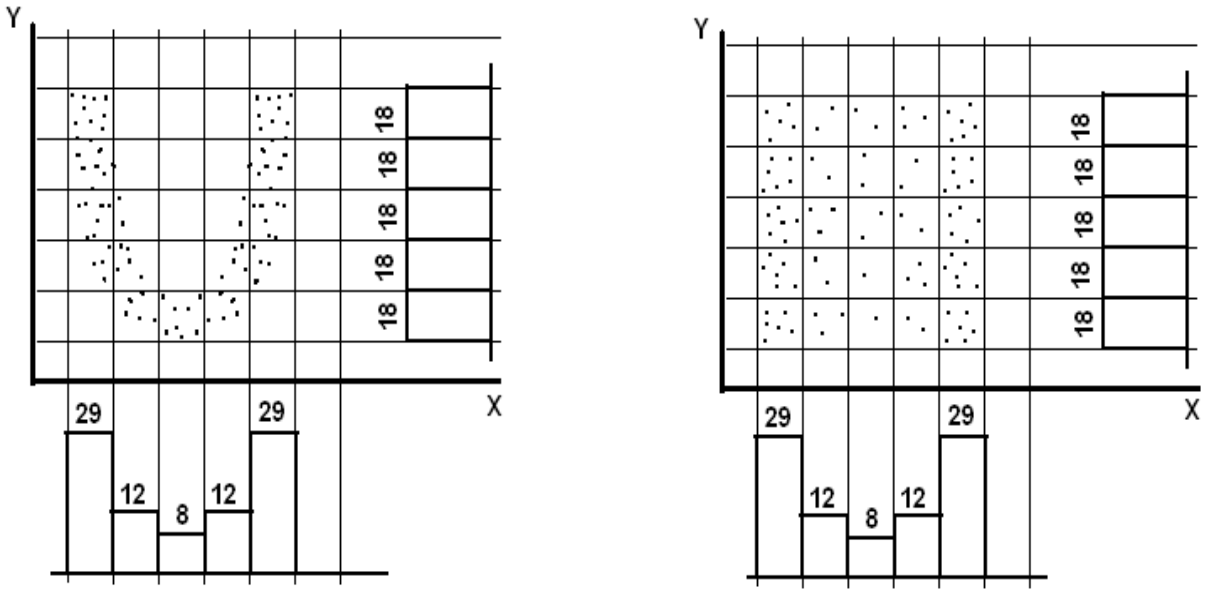


FIG. 2.18 – A gauche une liaison parabolique, à droite indépendance, à distributions marginales égales

le produit des probabilités marginales. Dans notre exemple, A n'a que 3 valeurs, mais B en a beaucoup plus, et même un nombre indéterminé. A est en effet une variable catégorielle bien déterminée, alors que B est une variable quantitative qui peut prendre a priori toute valeur entières positives. De plus, les probabilités sont théoriques donc inconnues. Il faut donc faire des hypothèses afin de déterminer la loi suivie par les écarts entre les différences calculées et la valeur. En effet ceux-ci ont très peu de chances d'être nuls, mais on désire savoir à partir de quelle valeur on peut estimer qu'ils s'éloignent trop de 0 pour qu'on puisse continuer à accepter l'indépendance. Les deux tests les plus utilisés sont *le test du Chi2* et le test du coefficient de corrélation linéaire nul.

Le test d'indépendance du Chi2.

On rappelle qu'il se fait à partir d'un tableau d'effectifs observés, croisant les valeurs des deux variables catégorielles. Ce tableau est appelé *tableau de contingence*. Puis on ajoute les effectifs des lignes et des colonnes et on met les résultats en marge, ce qui donne les effectifs marginaux de chaque variable, c'est-à-dire sans considérer l'autre. Les probabilités des loi théoriques marginales s'obtiennent en divisant ces effectifs par l'effectif total, et leurs produits donnent une estimation des probabilités de la loi conjointe en cas d'indépendance. On calcule alors les écarts entre les probabilités en cas d'indépendance et les probabilités de la loi conjointe, estimées à partir des effectifs dans le tableau. Par exemple, si on a un effectif total de 100, et que 70 personnes ont répondu "oui", on estime la probabilité marginale P(A="oui") par 0.7. Si 30 personnes ont déclaré qu'il y aurait 8 candidats, on estime P(B=8) par 0.3. et si 3 personnes ont répondu "oui" et déclaré qu'il y aurait 8 candidats, on estime la probabilité conjointe P(A="oui" et B=8) par 0.03. Comme le produit P(A="oui")P(B=8) est 0.21, la valeur de 0.03 présente un écart de 0.18 avec celle attendue de 0.21 en cas d'indépendance. On calcule ainsi les écarts pour toutes les cases du tableau, qui sont au nombre de $n \times p$ s'il y a n catégories pour A et p pour B. Et le

coefficient du Chi2 est la somme des carrés de ces écarts. Si tous les écarts étaient nuls, cette somme serait nulle. Bien sûr, cela n'arrive que dans des *cas d'école*.

L'utilisation du test nécessite le vérification d'hypothèses : si A et B suivent chacun une loi multinomiale (c'est ce qui a permis notamment d'estimer les probabilités marginales), si les effectifs théoriques (obtenus en multipliant la probabilité conjointe par l'effectif total, soit ici 21 pour A="oui" et B=8) ne sont pas trop petits (ils doivent tous dépasser 1, et en majorité dépasser 3, voire 5 selon les auteurs), alors ce coefficient suit la loi du Chi2 à (n-1)(p-1) degrés de libertés. On peut alors voir si la valeur qu'on a obtenue pour ce coefficient est excessive ou non par rapport aux valeurs de la loi (lue sur une table, ou donnée par un tableur) et si elle l'est, on rejette l'hypothèse H0 d'indépendance.

Mais ces hypothèses sont parfois difficiles à vérifier. Pour retourner à notre exemple, il est tout à fait vraisemblable que A soit une loi multinomiale, car elle correspond à 3 catégories. Mais il faudrait que B soit aussi une loi multinomiale, ce qui signifie que ses valeurs devraient être des catégories en nombre p fini, sans aucune relation d'ordre entre elles. De plus s'il n'y a que 100 personnes interrogées, cela signifie que les effectifs vont se répartir en 3p cases, et si p vaut 11, il sera difficile d'obtenir que la plupart des 33 cases aient un effectif théorique supérieur à 3, surtout s'il y en a déjà une avec un effectif de 21. Tous ces problèmes d'application du test du Chi2 sont bien connus, et des corrections sont proposées pour remédier à chacun. Par exemple pour les effectifs trop petits, il suffit de regrouper des valeurs consécutives de B et de recommencer les calculs. Mais on s'éloigne alors de la nature de B. Et il reste encore d'autres difficultés. Par exemple en cas d'effectif total important ce test a tendance à conclure systématiquement à une dépendance entre les variables.

La nullité du coefficient de corrélation linéaire.

L'indépendance théorique des deux variables aléatoires s'établit en calculant un coefficient de liaison adapté à leur type (quantitatif, ordinal, catégoriel) sur les valeurs observées, et en le comparant à sa valeur théorique en cas d'indépendance (en général nulle). Les écarts entre les valeurs observées et les valeurs théoriques sont aléatoires. On peut faire un certain nombre de suppositions sur les distributions de ces écarts, la plus courante étant qu'ils suivent une loi normale et qu'ils ne dépendent pas les uns des autres (ils sont i.i.d. c'est-à-dire identiquement et indépendamment distribués). Quand ces conditions sont remplies alors le coefficient de liaison est une variable aléatoire dont on connaît la distribution de probabilités, et une fois choisi le niveau de risque (appelé risque α, ou risque de première espèce), on peut obtenir un intervalle de valeurs du coefficient autour de sa valeur à l'indépendance, appelé zone d'acceptation. Si la valeur empirique appartient à cette zone, alors on décide qu'elle est approximativement égale à sa valeur à l'indépendance (on dit qu'on accepte l'hypothèse nulle H0), l'écart à la valeur attendue étant dû au hasard. Sinon, on rejette H0, et on décide que les variables ne sont pas indépendantes, avec l'assurance que le risque de se tromper dans cette conclusion est inférieur au seuil α, si toutes les hypothèses qu'on a faites par ailleurs sont vérifiées. Le fait de rejeter H0 correspond à accepter une hypothèse alternative, notée H1 qui peut avoir des formes variées. Par exemple, les plus courantes pour $H0 : \rho = 0$, où ρ est le coefficient théorique de corrélation linéaire entre les deux variables, sont l'hypothèse bilatère $H1 : \rho \neq 0$ et les deux hypothèses unilatères $H1 : \rho > 0$, et $H1 : \rho < 0$.

Reprenons le coefficient de corrélation linéaire qui a déjà été exposé dans la partie précédente et dont les valeurs ont été calculées pour le petit exemple de la figure 2.16. On a défini 5 variables A, B, C D et E, et on a observé leurs valeurs sur 10 objets. On suppose que la distribution des 5 variables suit la loi normale. Si on se place dans le cadre d'un échantillon représentatif, on

suppose que les objets ont été tirés indépendamment et au hasard dans la population, chacun ayant la même probabilité d'être choisi, et si on se place dans le cadre d'un échantillon exhaustif, on suppose qu'ils ne sont pas liés entre eux par des relations de dépendance (hiérarchique, spatiale ou temporelle). En cas d'indépendance entre deux de ces cinq variables le coefficient de corrélation théorique ρ est nul et la statistique $F_{\nu_1,\nu_2}=\frac{r^2\nu_2}{(1-r)^2\nu_1}$ calculée à partir du coefficient de corrélation empirique r suit alors une loi de Fisher-Snedecor dont les paramètres sont $\nu_1 = 1$ et $\nu_2 = 8$. En prenant $H0 : \rho = 0$ et $H1 : \rho \neq 0$ et un niveau de significativité de 0.05 (c'est-à-dire en acceptant un risque α inférieur à 5% de se tromper en rejetant à tort H0), la zone d'acceptation de H0 est $[0; 5.32[$, comme nous l'apprend la table statistique de la loi F , et la zone de rejet de H0 est $[5.32; +\infty[$. En remplaçant F par sa valeur en fonction de r, on trouve que pour rejeter H0, il faut que r soit supérieur en valeur absolue à 0.632. Les coefficients de corrélation empiriques respectifs de A avec B, C, D et E étant respectivement de 0.445, 0, -0.053, et -0.806, on décide que A est linéairement indépendant de B, de C et de D, mais pas de E, et comme on vient d'établir que le coefficient de corrélation entre A et E n'est pas nul, on peut alors interpréter sa valeur négative en terme d'opposition ou de répulsion selon le domaine de provenance des données, ou tout simplement que la liaison est négative si on reste dans une interprétation purement statistique. Ainsi, la valeur de 0.445 qui semblait indiquer un lien positif assez fort entre A et B, a été jugée "non significative", c'est-à-dire due au hasard. Cela ne signifie pas que A et B ne sont pas liés linéairement, mais seulement qu'il a été impossible de conclure que la valeur de leur coefficient de corrélation linéaire était différente de zéro. La même valeur de 0.445 avec un échantillon de plus de 17 objets aurait abouti au rejet de H0. On voit ainsi l'importance du choix de la taille d'échantillon, comme en atteste la présence dans certains manuels de statistiques appliquées ([210, 182]) de formules permettant de la calculer dans divers cas courants. Une expérimentation sur l'évolution de personnes ou plus généralement d'être vivants menée sur plusieurs années avec les mêmes sujets peut en effet aboutir à des résultats inexploitables car non significatifs si la taille de l'échantillon a été sous-évaluée au départ.

Coefficient de corrélation et indépendance

Dans la partie précédente, à l'aide du coefficient de corrélation linéaire, nous avons établi une liaison linéaire positive importante entre A et B, négative encore plus importante entre A et D, nulle entre A et C et quasi-nulle entre A et D, tout en remarquant toutefois la présence d'une liaison parabolique qu'on pourrait quantifier. Dans cette partie, en utilisant ce même coefficient de corrélation linéaire, nous avons conclu à l'existence d'une relation significative entre A et D, et à une indépendance entre A et B, A et C, et A et D, sous la condition toutefois que ces variables suivent une loi normale. Nous n'avons pas vérifié le bien-fondé de cette assertion de normalité, car elle a peu de chances d'être réfutée pour un si petit nombre de valeurs. Le fait d'exiger des lois normales a permis de se contenter de contrôler la nullité d'une seule valeur (qui se ramène ici à la différence entre le produit des moyennes et la moyenne des produits), au lieu de contrôler la nullité de la différence entre le produit de probabilités et la probabilité produit pour chaque valeur, comme spécifié dans la définition de l'indépendance. La simplification obtenue dans le cadre de la normalité n'est pas surprenante car on a déjà vu qu'une distribution normale ne dépendait que de deux paramètres, au lieu des 10 que sont les objets de cet exemple. Ici, il ne s'agit plus d'une loi normale simple, définie sur une seule variable, mais d'une loi normale double définie sur 2 variables, et elle est caractérisée par 5 paramètres qui sont les deux espérances des variables, leurs deux variances, et leur corrélation. Notons toutefois que dans ce cadre, une fois l'indépendance entre A et C décidée, on ne va pas examiner une liaison parabolique du genre $C = aA^2$, le carré d'une variable normale n'en étant plus une. Notons que seule la loi normale

permet une telle simplification des calculs.

Pour conclure sur les tests d'indépendance

Nous avons vu les deux tests les plus utilisés pour établir l'indépendance de deux variables aléatoires. Le premier se fait sur les variables catégorielles suivant la loi multinomiale et le second sur des variables quantitatives suivant la loi normale. Nous avons vu également qu'ils nécessitent tous deux que soient vérifiées des conditions d'application assez contraignantes. Quand ces conditions ne sont pas vérifiées, il y a toujours la possibilité d'utiliser d'autres tests en remplacement de ceux-ci. Si les distributions ne sont pas normales, ou même si elles sont ordinales, on peut utiliser des tests issus des statistiques non paramétriques [220]. Ce sont par exemple les tests de rangs (on peut notamment tester la nullité du coefficient de rangs de Spearman ou de Kendall[49]). Toutefois elles devront également vérifier des conditions d'application (par exemple, les tests basés sur les rangs acceptent mal les ex-aequo) Et le test du Chi2 admet des variantes comme le test de la médiane, de McNemar.

Mentionnons aussi le test exact de Fisher en cas de petits effectifs. Historiquement c'est le premier test de permutation[50], qui ne se faisait qu'en cas de tous petits échantillons, sans supposer de lois particulières, en acceptant des cases d'effectifs nuls, mais qui entraînait trop de calculs vu les possibilités informatiques de l'époque. Il a été remplacé par les tests asymptotiques qui sont basés sur des hypothèses théoriques permettant des résolutions analytiques donnant des formules simples à appliquer, assorties de tables de lois simples à utiliser, et de conditions d'applications simples à oublier.

2.6 Relations complexes et causalité en sciences humaines

A part les liaisons classiques entre variables que nous venons de voir, qui s'appuient toutes sur une modélisation de type linéaire, parmi lesquelles l'interaction peut être considérée comme complexe, il en existe, en sciences humaines, de plus complexes encore qui échappent à cette modélisation, et font actuellement l'objet de peu de recherches et d'applications en fouille de données. Parmi elles la plus connue est le paradoxe de Simpson,

2.6.1 Un exemple historique de liaison complexe : le paradoxe de Simpson

Le "paradoxe de Simpson" rend compte du changement de sens d'une relation entre 2 variables lors de l'intervention d'une troisième variable. Simpson a détaillé le problème en 1951 [217], mais il était déjà connu précédemment des statisticiens, Yule [239] notamment, comme le signale Pearl [192]. L'exemple le plus courant met en jeu les 2 variables "sexe" et "réussite à un examen" d'un ensemble d'étudiants issus d'une zone géographique donnée, par exemple d'une ville. On s'aperçoit que les étudiants d'un sexe réussissent mieux l'examen que ceux du sexe opposé. Puis on détaille les résultats par sous-zones homogènes, par exemple par établissement d'enseignement, et on constate que pour chacune de ces sous-zones, c'est l'autre sexe qui réussit le mieux à l'examen. L'intervention d'une troisième variable, ici "établissement", contredit donc la conclusion tirée précédemment sur la relation entre sexe et réussite. D'où le paradoxe, développé formellement en annexe B. Les chercheurs en sciences humaines connaissent bien ce problème qui fait que toute

[49] Le coefficient tétrachorique des données binaires, vu précédemment dans la section 2.4.3, se teste aussi de la même façon

[50] Il fait l'objet d'une description détaillée dans la section du chapitre suivant portant sur les nouveaux tests de validation

conclusion sur des relations entre variables peut être remise en cause par l'intervention d'une variable "oubliée" (Howell [126]).

2.6.2 Les liaisons complexes

Nous venons d'exposer un exemple mettant en jeu le paradoxe de Simpson pour montrer en quoi il peut gêner l'interprétation de règles. Nous allons maintenant décrire le cadre plus général des liaisons complexes en le remettant dans le contexte de "l'analyse multivariée causale" de J. Herman qui expose dans son ouvrage [118] comment établir une "preuve" quand on dispose de données binaires.
Établir une preuve est pour lui est une procédure qui consiste à introduire une variable-test pour examiner son effet sur l'hypothèse de départ. Par exemple si l'hypothèse est la règle *tabac* → *cancer*, on prend la variable-test "alcool" afin de vérifier si cette règle est valable indépendemment de la consommation d'alcool. Pour cela on partitionne l'ensemble de sujets en différents groupes à raison d'un groupe par modalité de cette variable, et on examine si la règle est valable sur chaque groupe. Cela peut se faire en décomposant l'association entre "tabac" et "cancer" en deux parties : l'association "partielle" intra-groupes, et l'association "différentielle" inter-groupe, pondérées par des coefficients liés aux différences d'effectifs des groupes. Les calculs diffèrent selon le coefficient d'association choisi. Appelons cette hypothèse A→B, et en reprenant les notations de J. Herman, T la variable test, s un seuil d'intensité significative de l'association, Z l'association totale "d'ordre zéro", P la partielle, et D la différentielle selon T. L'auteur propose alors une classification en quatre issues de l'intervention de la variable test qui sont :

- La *corroboration* quand Z,P>s et $Z \approx P$. La prise en compte de la variable test n'a pas modifié l'association entre A et B.
- L'*explication* quand Z>s et $P \approx 0$. L'association entre A et B en prenant en compte la variable test est nulle. Ce qui fait que l'association qu'on avait repérée entre A et B n'était due qu'à la présence de T. Cela peut se passer de deux façons différentes :
  1. La relation entre A et B est *fallacieuse* (falsification de Popper [197]). Ce qui signifie qu'une modification de A n'a aucune raison d'entraîner une modification de B. On peut même avoir un cas d'"hyper-falsification" quand au lieu d'être nulle, l'association partielle est de signe contraire à l'association globale.
  2. La relation entre A et B est *indirecte*. A→T→B. Bien que A ne soit pas la cause de B, une modification A entraîne une modification de T, qui lui-même entraîne une modification de B.
- La *contribution* quand Z,P>s et P<Z. Quand on examine l'association entre A et B en prenant en compte la variable T, elle est moins forte que quand on ne la prend pas en compte.
- L'*atténuation* quand Z,P>s et P>Z. Quand on examine l'association entre A et B en prenant en compte la variable T, elle est plus forte que quand on ne la prend pas en compte.

A tout cela vient se rajouter une possible interaction[51], qui exprime le fait que l'association entre A et B n'est pas la même selon les différentes valeurs de T. Ainsi une même valeur de P, qui est une combinaison de ces associations partielles peut provenir de cas bien différents, par exemple de deux associations de même grandeur, de même sens mais de grandeurs différentes, l'une pouvant être négligeable, ou même de sens contraires. Ainsi d'après Herman, "une association bivariée globalement corroborée peut être localement hyper-spécifiée... Dans l'un des sous-groupes

[51] dont le principe a déjà été exposé dans la section 2.3.1

il y aura un effet local de contribution et dans l'autre un effet local d'atténuation de l'association-clé".

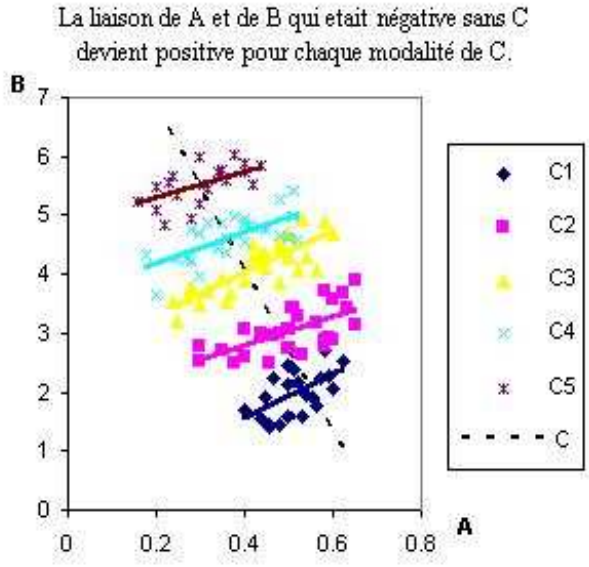


FIG. 2.19 – La corrélation entre A et B est de -0.41 pour l'ensemble, et elle varie de 0,58 à 0,83 à Ci fixé.

Parmi tous ces cas, les plus gênants pour le sens commun sont la relation fallacieuse et les interactions. Notons qu'ils n'apparaissent pas seulement dans les tableaux de données binaires, les interactions étant susceptibles de se présenter dans tous les modèles traitant de la liaison de plus de deux variables et la relation fallacieuse apparaissant également dans les modèles de corrélation/régression, comme on peut le voir dans le graphique de la figure 2.19, qui illustre l'hyper-falsification également appelée paradoxe de Simpson.

Dans ce graphique, A et B sont deux variables quantitatives, et C est une variable prenant 5 modalités de C1 à C5. Chaque point est la représentation d'un sujet, son abscisse étant sa valeur pour la variable A, son ordonnée celle pour B, et sa valeur pour C est représentée par sa couleur et sa forme, par exemple le sujet représenté par un losange bleu foncé a la modalité C1 de la variable C. On a représenté les droites de régression de B selon A pour chacun des nuages partiels, afin de montrer que leurs pentes sont positives et de valeurs proches, ce qui indique que si on fixe la valeur de C, quelle que soit sa valeur, quand A croît B croît également de façon similaire. Si on regarde maintenant le nuage de points dans son ensemble, on voit qu'il suit la direction indiquée par la droite noire en pointillés, qui est la direction inverse de celle à C fixé. C'est-à-dire que cette fois quand A croît B décroît. En économie, ce phénomène est bien connu, qui fait qu'on ne peut pas généraliser des lois de la micro-économie à celles de la macro-économie sans risquer de "contresens".

Cet effet diffère de l'interaction dont les bases ont été exposées précédemment au moyen d'un exemple basique. Nous rappelons ce qu'est cette liaison au moyen d'un exemple présentant un niveau supérieur de complexité. Le graphique de la figure 2.20 représente une interaction croisée entre deux variables, selon un modèle d'analyse de la variance[52], pour lequel B est une

[52]Comme dans la représentation de la figure 1, on ne parlera pas des tests associés, car c'est une simple illustration, mais on peut imaginer que dans le modèle B=A*C+A+C, où la variance expliquée de B se décompose

variable quantitative à expliquer, et A et C sont des variables qualitatives explicatives l'une à trois modalités A1, A2 et A3, et l'autre à deux modalités, C1 et C2.

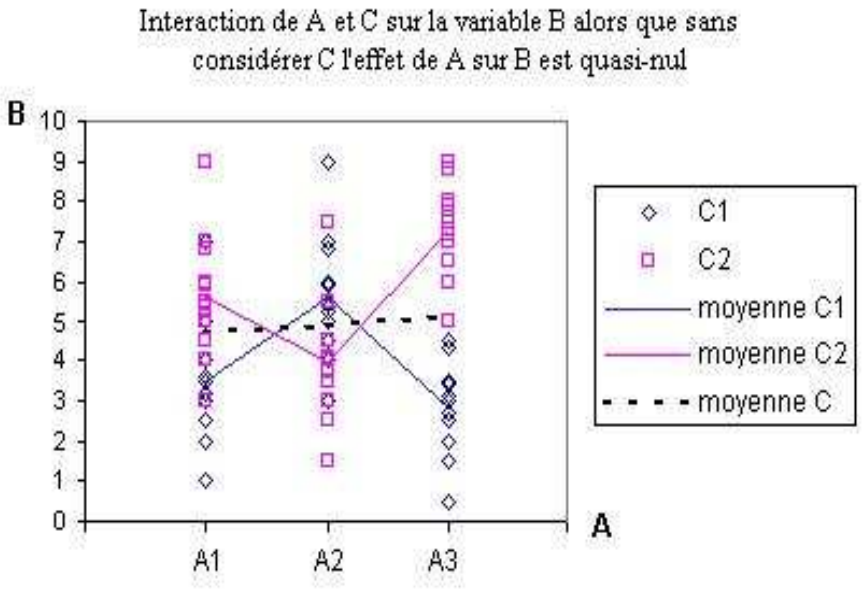


FIG. 2.20 – Les moyennes de B selon A sont 3,48, 5,61 et 2,82 pour C1, 5,61, 3,98 et 7,32 pour C2 et 4,64, 4,87 et 5,07 pour C1 et C2.

Chaque point représente un sujet, il a pour abscisse sa valeur selon A, pour ordonnée sa valeur selon B et sa forme et sa couleur indiquent sa valeur selon C. Les moyennes partielles de B selon les 6 cas A1C1, A1C2, ..., A3C2 sont jointes par des lignes de couleur. Les sujets ayant la valeur 1 de C ont des valeurs en moyenne plus petites pour A1 que pour A2 alors que c'est l'inverse pour les sujets ayant la valeur 2 de C. L'effet de A sur B est modifié par la valeur de C, c'est en cela qu'on l'appelle une interaction. Cette interaction est "croisée" car les lignes de couleur correspondant à ces 6 moyennes se croisent. C'est l'effet le plus marquant, car on voit que la ligne noire qui joint les moyennes de B selon A1 et A2 est pratiquement horizontale, les moyennes étant très proches. Si bien qu'en ne considérant pas l'effet de C on peut arriver à la conclusion que A n'a pas d'effet sur B, alors que son effet est important quand on fixe C. Cette interaction oppose les effets de A1 et A3 (qui vont dans le même sens) à ceux de A2, en sens contraire. L'opposition entre ces effets au sein d'une même variable (ici A) fait partie des *contrastes* qui permettent de décrire encore plus finement la complexité des relation entre variables (pour plus de détails sur ces décompositions, voir [1, 236, 126, 120]).

D'autres liaisons causales plus complexes ont été mises à jour en systémique, comme par exemple la rétroaction [6]. Pour mettre ces effets en évidence, il est mieux de disposer de tableaux de données sur plusieurs instants. Dans notre thèse, nous en restons à un seul tableau de données, collectées à un seul moment, dans lequel la variable temps, si elle est présente, n'est pas différenciée des autres variables. Mais cela peut en être un prolongement intéressant, compte tenu de l'essor que ces modèles ont depuis plus d'une dizaine d'années en sciences humaines [127] grâce à la diffusion de logiciels qui permettent de tester aisément des modèles à base de systèmes

en trois parties, celle correspondant à l'interaction A*C est significative alors que dans le modèle B=A, où C ne figure pas, l'effet de A n'est pas significatif

d'équations structurelles [213]. Un exposé très détaillé de tous ces liens complexes entre variables, de leurs définitions et de leurs applications aux phénomènes sociaux est fait dans le livre de R. Boudon [26].

3

Possibilités liées à l'augmentation de puissance des ordinateurs pour l'extraction de liaisons entre variables

Dans ce chapitre nous faisons le tour des méthodes qui se sont développées avec l'avènement de l'informatique. On y trouve à la fois des versions améliorées des méthodes classiques des statistiques décrites dans le chapitre précédent, mais également de nouvelles méthodes qui sont apparues depuis. Parmi celles-ci des méthodes d'investigation "locales" dont l'*extraction de règles d'association*, qui sera décrite sommairement ici, le chapitre suivant lui étant entièrement consacré.

Par contre les méthodes du type STATIS [90], ARIMA [67], les modèles de Markov [178], ou les méthodes de traitement du signal (ondelettes,..) ne seront pas examinées dans ce chapitre car cette thèse est centrée sur les liaisons causales et non temporelles ou séquentielles.

Sommaire

3.1 Les nouvelles approches descriptives

3.1.1 La nombreuse descendance de l'analyse factorielle.

L'analyse factorielle (Spearman 1904 [219], Pearson 1901 [193]) et l'analyse en composantes principales (ACP Hotelling 1933 [125]) ont été conçues dans le premier tiers du 20ème siècle, mais sont restées confinées à des applications de taille très limitée, principalement en psychologie, avant l'arrivée de l'informatique dans les années 1950 et 1960. Leur principe est simple :

- chaque vecteur-donnée $\mathbf{x}$ (individu, ou observation), à I dimensions, autant que de variables, est exprimé comme une somme pondérée de K (où $K \leqslant I$) "composantes" $\mathbf{w}_k$, appelées aussi facteurs communs ; chaque composante traduit une variable "latente", cachée dans les données :
  - $\mathbf{x} = \mathbf{W}\mathbf{y} + \mathbf{e}$ (analyse factorielle)
  - $\mathbf{x} = \mathbf{W}\mathbf{y}$ (cas de l'ACP avec K=I)

  où $\mathbf{y}$ est le vecteur des coordonnées factorielles (*factor score*) de l'individu $\mathbf{x}$, $\mathbf{W}$, la matrice formée par l'ensemble des vecteurs $\mathbf{w}_k$ et $\mathbf{e}$ un vecteur bruit spécifique de cet individu.

La conséquence qui nous intéresse ici est que l'effet de 2 (ou n) variables dont la forte valeur simultanée aurait un effet différent de la somme de leurs effets individuels c'est à dire de 2 (ou n) variables en interaction, n'est pas pris en compte, structurellement, dans ce modèle fondamentalement additif.

Il est à noter que l'interaction peut cependant être prise en compte en créant une grande quantité de nouvelles variables, à savoir toutes les combinaisons 2 à 2, 3 à 3, ... de toutes les variables d'origine, mais ceci poserait des problèmes de multiplication exponentielle de la taille des données, ou obligerait à utiliser, pour remplacer les produits scalaires, des fonctions noyaux polynomiales, comme le font les *Support Vector Machines* pour l'apprentissage supervisé [56], au prix de la perte de l'explicitation[53] des combinaisons de variables intervenant dans telle ou telle valeur factorielle d'individu. Cette perte est acceptable dans le domaine de l'apprentissage supervisé, dans le cadre d'une démarche d'ingénierie où seul le résultat compte, mais pas dans celui des sciences humaines où il est important d'expliciter au maximum le " pourquoi " des résultats trouvés.

Ce modèle est décliné sous de nombreuses formes, a eu et continue d'avoir une riche descendance au fur et à mesure que la puissance informatique disponible augmente.

- L'**analyse en composantes principales** est une méthode devenue standard dans de nombreux domaines scientifiques, où elle porte parfois des noms différents (transformée de Karhunen-Loeve...). Le tableau de données $\mathbf{X}$ comporte I variables centrées et N observations, et la décomposition $\mathbf{X} = \mathbf{W}\mathbf{D}^{\frac{1}{2}}\mathbf{Y}$, où $\mathbf{D}$ est la matrice diagonale des I valeurs propres obtenue à partir de la décomposition spectrale de la matrice de variance-covariance des données $\frac{1}{N}\mathbf{X}\mathbf{X}'$ en composantes orthonormales non-corrélées :
  - $\mathbf{XX'} = \mathbf{WDW'}$ (formule de reconstitution de la variance-covariance)
  - $\mathbf{W} = \mathbf{XYD}^{-\frac{1}{2}}$ (formule de transition)

On s'intéresse généralement aux éléments propres des k premiers rangs, qui donnent souvent lieu à des cartes représentant soit les individus, soit les variables, soit les deux, les autres éléments propres étant considérés représenter le "bruit" dans les données :

$\mathbf{X} = \mathbf{W}_k\mathbf{D}_k\mathbf{Y}_k$

[53] Effet diabolique du "kernel trick" qui permet à la fois la prise en compte de l'interaction dans les données à grande échelle, et en interdit l'explicitation!

Une variante importante en est l'**analyse factorielle des correspondances**, qui utilise la métrique du Chi2 et permet d'analyser les tableaux de contingence.

- L'**analyse sémantique latente** (LSA, *Latent Semantic Analysis*) est utilisée dans le domaine de la recherche d'information textuelle. Elle procède par décomposition aux valeurs singulières de la matrice (mots × textes) brute, en conservant le plus souvent les quelques centaines de composantes les plus importantes, qu'on ne cherche pas à interpréter, le but étant une réduction " technique " du nombre de dimensions et du bruit pour des calculs de distances entre textes et entre mots. Cette décomposition s'exprime de la même façon que l'ACP ci-dessus, sans la contrainte de centrage-réduction des variables.

- Les **analyses factorielles** sont utilisées principalement en psychologie et géologie. Après centrage et réduction des variables, le modèle général de décomposition est $\mathbf{X} = \mathbf{W}\ \mathbf{Y} + \mathbf{E}$ où $\mathbf{E}$ est une matrice modélisant le bruit, $\mathbf{W}$ et $\mathbf{Y}$ les matrices formées d'éléments orthogonaux, à savoir les facteurs-variables (*factor loadings*) et les facteurs-individus (*factor scores*). De multiples variantes existent, selon le modèle de bruit utilisé et la méthode de détermination des facteurs [115, 124] - les facteurs étant indéterminés en règle générale, on a conçu diverses méthodes de rotation d'axes sur des critères variés (visuels, ou optimisant un critère tel que Varimax, Promax, ...[115] pour y parvenir. La contrainte d'orthogonalité des facteurs peut également être levée (rotations Oblimax), afin de pointer librement vers les zones de densité élevée des données, et tendre au maximum vers l'idéal de la " structure simple " où les composantes des facteurs sont les plus " contrastées " possible - seules quelques valeurs ressortant par rapport à une grande majorité de valeurs négligeables.

- Les **réseaux neuronaux non supervisés** : ce formalisme recouvre une vaste famille d'algorithmes qui régissent l'évolution et l'interaction de " cellules " élémentaires dites neurones. Chaque neurone est caractérisé par un vecteur " poids synaptiques ", à raison d'un poids attribué à chaque " entrée " (= variable) ; la présentation à ce neurone d'un vecteur-individu entraîne une valeur de sortie, fonction croissante de l' " activité " du neurone (produit scalaire du vecteur-individu et du vecteur-poids) [fig. 1], et une modification des poids, dite " apprentissage ".

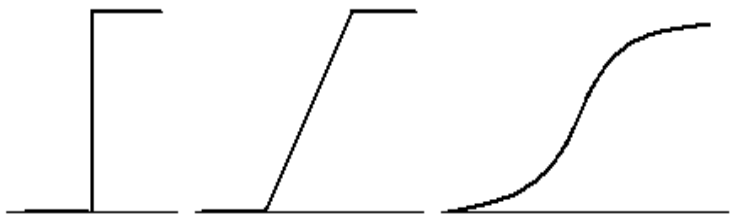


FIG. 3.1 – Fonctions de transfert de modèles neuronaux - différentes formes de courbes, η'=f(η) (sortie η' en fonction de l'activité η)

Cet apprentissage est généralement de type Hebbien, c'est-à-dire qu'il consiste, pour un neurone isolé, sans contrainte, en une montée en gradient sur une fonction objectif, par exemple ici la somme, pour tous les vecteurs-individus, des carrés des sorties :

$m(t+1) = m(t) + \alpha\eta$' x,

où α est une constante petite, avec une normalisation périodique de m.

Les fonctions de transfert " sortie en fonction de l'activité " peuvent prendre diverses formes [figure 3.1] :

- fonction identité : on démontre [188] que le vecteur-poids converge alors vers le premier vecteur singulier de la matrice des données (fonction objectif : inertie = somme des η^2).
- fonction à seuil, par exemple $\eta' = \begin{cases} \eta - \eta_0 & si\ \eta > \eta_0 \\ 0 & sinon \end{cases}$, où η_0 est une valeur de seuil.
  Cette fonction est présente dans le modèle *Analyse en Composantes Locales* [158, 159] (à fonction objectif : inertie locale = somme des η'_2).; cf. aussi [189].
- fonction logistique (saturation) (cf. [136] pour la séparation aveugle de signaux]
- etc.

De ce point de vue, la *décomposition aux valeurs singulières* peut être formalisée [203] comme un ensemble de neurones simples (sortie = activité) en interaction unidirectionnelle deux à deux (" inhibition " = empêcher l'apprentissage) selon une structure hiérarchisée de type :

- **N**1 inhibe tous les K-1 autres neurones,
- **N**2 inhibe tous les autres neurones sauf **N**1,
- **N**3 inhibe tous les autres neurones sauf **N**1 et **N**2
- etc.

Des structures d'inhibition/excitation particulières (en grilles 2D à mailles carrées, triangulaires,...) caractérisent le modèle très utilisé et étudié de Kohonen [144, 128], qui réalise ainsi simultanément l'apprentissage des données et la cartographie d'ensemble positionnant les neurones entre eux.

- La **poursuite par projection** (" *projection-pursuit* ") On recherche ici une ou plusieurs directions " intéressantes " sur lesquelles projeter le nuage de points, après centrage, réduction et " sphérage " des données par ACP (même variance unité dans toutes les directions). Une direction étant d'autant plus intéressante que la répartition des projections s'éloigne de la loi normale, on maximise un indice de non-gaussianité, par exemple la kurtosis (aplatissement) : $\kappa = E(\eta^4) - 3$, de valeur nulle pour une répartition de Gauss.

- L'**analyse en composantes indépendantes** (ICA, Independant Component Analysis) Il s'agit ici de reconstituer un modèle explicatif d'un ensemble d'observations d'individus (ou de mesures temporelles) par le mélange de K composantes latentes indépendantes. Dans le cas non-bruité, si X est la matrice des données centrées-réduites et sphérées comme ci-dessus, et si K=I (problème " cocktail-party " : autant de micros que de conversations à démêler) :

$\mathbf{F} = \mathbf{W}\ \mathbf{X}$, où **W** est la matrice (K,I) de mélange des composantes indépendantes.

On estime alors **A** tel que $\mathbf{X} = \mathbf{A}\ \mathbf{F}$ par un algorithme itératif maximisant, pour l'ensemble des composantes, un indice de non-gaussianité, la kurtosis par exemple ; **W** est alors l'inverse de **A**, ou sa pseudo-inverse si **A** n'est pas de rang plein.

- D'autres approches, comme **NNMF** (Non-Negative Matrix Factorization, [155]) ou **PLSA** (Probabilistic Latent Semantic Analysis [122]) relâchent les contraintes de centrage-réduction des données, ou d'orthogonalité des composantes. Cette famille de méthodes débouche sur des modèles plus complexes explicitant par exemple un processus de choix de mots dans un ensemble de textes appartenant de façon floue à plusieurs thémes [35], où l'estimation des composantes se fait au moyen d'algorithmes de type EM (Expectation Maximization).

Les analyses factorielles étant fondées sur l'analyse des matrices de variance-covariance, elles reposent sur les seuls liens deux à deux entre variables. C'est la raison pour laquelle l'interaction, qui met en jeu par définition des liaisons complexes entre plusieurs variables, ne peut pas être prise en compte directement pas ces méthodes, et de façon plus générale les relations complexes entre variables.

3.1.2 Les modèles de proximité : clustering, graphes

Les nombreuses méthodes de **classification non-supervisées** (ascendantes/descendantes hiérarchiques, à centres mobiles, floues...) sont toutes basées sur le calcul d'un indice de similarité entre individus pris deux à deux, ou entre individus et " individus idéal-types " définissant chaque classe. Cet indice relève, dans la quasi-totalité des cas, du modèle additif (distance, ou produit scalaire entre vecteurs), et ne prend pas en compte l'interaction entre variables. L'utilisation de noyaux polynomiaux à la place de produits scalaires le permettrait, mais interdirait du même coup toute explicitation du contenu des classes en termes de variables, ce qui est précisément un des résultats préférés des utilisateurs de ces méthodes, pour lesquels il est souvent aussi important de connaître le pourquoi d'un groupement d'individus que le groupement lui-même.

D'autre part, une matrice de similarité (individus $\times$ individus) peut être construite à partir de toute définition de la similarité, et traduite sous forme de **graphe** dont les noeuds sont les individus et les arêtes des fonctions des valeurs de similarité. De nombreuses méthodes permettent de calculer les propriétés globales d'un graphe, ou de le partitionner [30]. Mais dans le cas général ces techniques sont intrinsèquement incapables de prendre en compte les phénomènes d'interaction, sauf bien sûr à utiliser des fonctions noyaux, sans possibilité alors d'expliciter le contenu des liens.

3.1.3 Conclusion

Pour conclure cette section, on observera que malgré la prolifération de nombreuses formes du modèle factoriel additif, et l'existence de modèles de proximité variés, et même de formes non linéaires comme le "MDS" (Multidimensional Scaling [146], en français "échelonnement multidimensionnel" [229]) ces nouvelles approches descriptives, restent toutes limitées par leur non-prise en compte des phénomènes d'interaction entre les variables - ou du moins par leur non-explicitation quand elles sont l'objet d'extensions au moyen de méthodes à noyaux, tous inconvénients rédhibitoires pour les applications aux sciences humaines.

Les liaisons complexes sont étudiées à travers des modèles qui contiennent quelques variables, rarement plus de cinq, car les modèles statistiques permettant de le faire ne permettent pas mieux. En effet avec 5 variables à 2 modalités, on a déjà 2^5, soit 32 cas, et dans un modèle d'Anova, on doit avoir pour chaque cas des variances égales, un nombre d'individus proche, des lois normales afin de pouvoir conclure si les moyennes sont suffisamment différentes pour que le test soit significatif. Le problème est que si les effectifs sont importants, les différences même petites ont tendance à devenir significatives, comme le constate S. James Press[54], alors que c'est l'inverse si les effectifs sont petits. En d'autres termes ces tests ne sont pas utilisables autrement que sur des effectifs "raisonnables". On peut sélectionner ces variables par des procédures automatiques de choix mais on se heurte à un autre problème qui est celui des *hypothèses multiples* [131] en statistiques. Ce problème se rencontre aussi au niveau de l'interprétation quand il s'agit de comparer deux des 32 cas, ou plusieurs, ce qui s'appelle rechercher des contrastes. Ces modèles ne peuvent pas s'étendre aisément à une recherche exploratoire dans de grandes bases de données.

[54] dans [196] : « *As a result, in Data Mining, likelihood ratio methods of hypothesis testing (for large samples), and p-value significance level testing (for large samples) will tend to make the tiniest effects appear to be significant. Bayesian methods are preferable because they are more conservative in rejecting null hypothesis (Berger & Selke, 1987)* »

3.2 Les nouveaux tests de validation : Monte-Carlo, *bootstrap*, *jackknife*, permutation, randomisation.

Les tests d'hypothèses des statistiques classiques sont construits sur un mode très mathématique (des hypothèses précises, des démonstrations, des conclusions). Il en découle une qualité supérieure de preuve scientifique, mais cela impose aux chercheurs des sciences humaines de s'adapter aux modèles statistiques disponibles, ou d'en chercher de nouveaux, ce qui peut freiner l'avancée de leurs réflexions dans leur propre discipline. D'autres tests sont possibles, nécessitant un moindre formalisme mathématique, mais plus gourmands en capacité informatique, ce sont les tests basés sur des simulations. D'un cote le *bootstrap* et le *jackknife* donnent des versions "approximatives" des variantes qui "auraient pu être", alors que la randomisation et la permutation produisent des données qui n'ont "strictement rien a voir". Dans le premier cas on recherche les relations qui résistent à la variation (pour les garder), dans l'autre celles qui résistent à la randomisation et la permutation (pour les éliminer).

3.2.1 Un modèle proche des données, versus des données proches d'un modèle

Nous avons vu dans le chapitre précédent que les chercheurs en sciences humaines ont recours à des méthodes statistiques quand ils veulent convaincre leur communauté scientifique de l'action de certaines variables sur d'autres. En statistique inférentielle, leur choix porte essentiellement sur le modèle linéaire. Ce modèle, même dans sa version la plus simple qu'est la régression, nécessite la réunion d'un certain nombre de conditions pour que les résultats obtenus soient valides. S'affranchir de la condition de normalité des distributions est possible en utilisant par exemple des tests basés sur les rangs [220, 70], mais ces tests ont également leurs conditions d'application (ils fonctionnent généralement mal quand les ex-aequo sont trop nombreux). Une autre stratégie consiste à recoder les variables pour qu'elles soient plus adaptées aux modèles existants, en les normalisant[55] par exemple. Ce n'est pas toujours adapté, et cela peut même être impossible[56]. Mais surtout, nous souhaitons développer des méthodes de traitement de données qui peuvent être utilisées dans les différents domaines des sciences humaines, donc avec des données ayant des distributions variées de valeurs, ou même qui peuvent se modifier au cours du temps (flux de données). Les nouveaux tests de validation à base de simulation permettent de remplacer la comparaison des valeurs observées à des valeurs théoriques par leur comparaison à des valeurs simulées.

3.2.2 Les principes des différents types de tests de simulation

La simulation de données permet, dans le cas d'un modèle pour lesquels des solutions explicites, analytiques, n'ont pas été développées, de pallier cette difficulté par la création de données artificielles correspondant au modèle, et l'observation de leurs propriétés. Ce modèle peut être théorique (on pourrait vouloir construire un test d'indépendance entre deux variables dont les distributions de valeurs suivent des lois peu courantes, par exemple une loi zipfienne et une loi de Poisson, et spécifié indépendamment des données, bien que leur ressemblant. C'est le cadre

[55] Ce terme est pris dans le sens statistique tel que décrit dans la section 1.4.3.

[56] Nous avons été confrontés lors de la KDD Cup 2006 à des variables pour lesquelles plus de la moitié des individus avaient une valeur identique (égale au maximum ou au minimum), les autres ayant des valeurs d'autant plus rares qu'on s'éloignait de cet extremum (histogramme en forme de j ou de i). Normaliser une telle distribution nécessiterait de séparer les ex-aequo en leur attribuant des valeurs différentes de façon artificielle. Il nous parait souhaitable d'éviter une telle modification des données.

des *simulations de Monte-Carlo*. Mais on peut aussi faire des copies modifiées des données : Dans le cas du *bootstrap* et du *jackknife* on tire parmi les données des échantillons d'individus (sujets, objets, enregistrements, observations) sans changer leur valeur pour les variables (leurs propriétés) alors que dans le cas des *tests de permutation*, et de *randomisation*, on tire des valeurs nouvelles pour chaque variable sans changer leur distribution d'origine. Nous allons essayer de donner quelques éléments sur la genèse de ces diverses méthodes et sur leurs utilisations actuelles.

3.2.3 Les simulations de Monte-Carlo

D'après Malvin H. Kalos et Paula A. Whitlock [137], ce sont des chercheurs travaillant dans le domaine nucléaire américain dans les années 1940 qui ont utilisé les premiers le nom de "méthode de Monte Carlo" en référence aux jeux de hasard. Les méthodes mathématiques ainsi désignées avaient pour but de trouver des valeurs approchées de quantités numériques en utilisant des simulations du hasard. L'exemple le plus connu est celui du calcul de la valeur approchée d'une intégrale quand on ne peut calculer sa valeur exacte. Il remonterait à Buffon, auteur également en 1777 de l'expérience de l'"aiguille de Buffon"[57]. Dans leur premier chapitre intitulé "What is Monte Carlo", les auteurs choisissent la définition suivante d'une "méthode de Monte Carlo" :

« C'est une méthode de calcul qui requiert une utilisation délibérée de nombres tirés au hasard selon un processus stochastique ».

Ils définissent un tel processus comme une succession d'états dont l'évolution est déterminée par des évènements aléatoires. C'est donc plus qu'une simple génération de nombres aléatoires. La distinction que certains chercheurs font entre cette méthode de résolution approchée et la seule transcription informatique d'un processus stochastique naturel que serait la *simulation* proprement dite, leur semble théoriquement justifiée mais difficile à maintenir dans la pratique, tant les deux paraissent liées.

Reuven Y. Rubinstein, au début de son ouvrage intitulé "Simulation and the Monte Carlo method" [209], donne trois éléments qui peuvent aider à distinguer les méthodes de Monte Carlo des simulations. Ce sont le rôle plus important du temps dans les secondes, l'indépendance entre observations dans les premières ainsi que leur facilité d'écriture de la "réponse" en une fonction des observations simulées. Dans le reste de son ouvrage, il fait une revue détaillée des divers types d'utilisation des simulations de Monte-Carlo et de leurs fondements. Il décrit la résolution approchée d'équations linéaires variées (simultanées, intégrales, différentielles) par simulations de chaînes de Markov à temps discret ou continu, la résolution approchée de problèmes d'optimisation, mais également les méthodes de génération de nombres aléatoires vérifiant des distributions de probabilité variées qui sont davantage dans notre champ d'investigations.

Nous allons nous attarder sur les méthodes de simulation de Monte-Carlo permettant de tirer "au hasard" une série statistique vérifiant une loi de probabilité donnée, qu'elle soit unidimensionnelle ou multidimensionnelle. D'après Reuven Y. Rubinstein, ces méthodes ont vu le jour dans les années 1950, les plus anciennes étant la *méthode de la transformation inverse* et la *méthode d'acceptation-rejet*, due à Von Neumann (1951)[58]. Prenons l'exemple simple d'une

[57] Elle consiste à calculer une valeur approchée de π obtenue en comptant le nombre de fois qu'une aiguille de longueur L jetée de façon répétée sur un parquet formé de lames de largeur $l < L$ tombe à cheval sur deux lames, sachant que la probabilité que cet évènement arrive est $P = \frac{2L}{\pi l}$.

[58] L'auteur cite également la *méthode de composition*, due à Butler (1956), ramenant une loi de probabilité complexe à un mélange de lois simples, à travers l'expression de sa fonction de répartition F ou de densité f.

variable X continue "triangulaire" dont la densité de probabilité est définie par

$$f(x) = \begin{cases} \frac{x}{8} & si\ 0 \leqslant x \leqslant 4 \\ 0 & sinon \end{cases}$$ et la loi de répartition par $$F(x) = \begin{cases} 0 & si\ x < t0 \\ \int_0^x \frac{v}{8} dv = \frac{x^2}{16} & si\ 0 \leqslant x \leqslant 4 \\ 1 & si\ x > 4 \end{cases}$$

dont on peut voir une représentation graphique dans la figure 3.2.

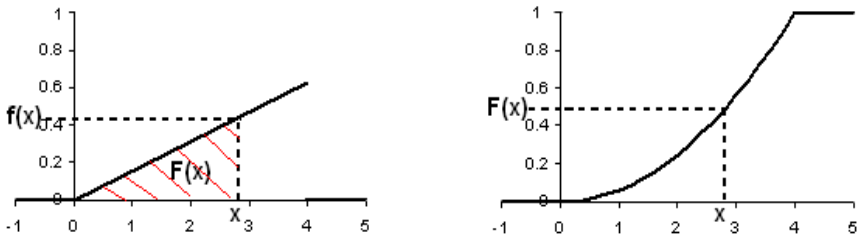


FIG. 3.2 – Les éléments caractéristiques de la loi de probabilité de X : f sa densité, F sa fonction de répartition

Dans le premier cas (méthode de la transformation inverse) la simulation d'une valeur X se fait de la façon suivante :

1. on tire un nombre U au hasard entre 0 et 1
2. on calcule l'antécédent X de U par F, donc en résolvant l'équation[59] F(X)=U

Appliquons cette procédure à la création d'une valeur X concernant la densité f triangulaire citée précédemment. Imaginons qu'on a tiré la valeur $U = 0,25$. X est obtenu en résolvant l'équation $\frac{x^2}{16} = 0.25$ sur l'intervalle (0, 4), ce qui donne X=2.

Dans le second cas (méthode d'acceptation-rejet), voici une version très simplifiée de la procédure de tirage d'une valeur (dans l'intervalle (a, b) pour lequel la fonction de densité n'est pas nulle) selon l'algorithme d'acceptation-rejet :

1. on tire deux valeurs U et V au hasard entre 0 et 1,
2. on pose $X = a + V(b - a)$ (ce qui revient à tirer X au hasard entre a et b)
3. si $U \leqslant f(X)$, on accepte la valeur X, sinon on la rejette,

Cette procédure revient à choisir un point(U,V) au hasard à l'intérieur du rectangle contenant la courbe de la densité, donc avec U dans l'intervalle (0, 1) et X dans (a, b). Si le point est au dessus de la courbe, il et rejeté, sinon il est gardé. Ainsi les points gardés se trouvent répartis au hasard sous la courbe de densité. Reprenons notre même exemple et la valeur tirée U=0.25 avec cette seconde méthode. On tire alors une deuxième valeur V. Si V=0.4, comme a=0 et b=4, $X = a + V(b - a) = 0 + 0.4(4 - 0)$, soit 1.6, et $f(X) = \frac{x}{8} = \frac{1.6}{8} = 0.2$. L'inégalité $U \leqslant f(X)$ n'est pas vérifiée, le point de coordonnées (1.6, 0.25) se trouvant au-dessus de la courbe. On

[59] Cette résolution se fait de façon exacte, c'est-à-dire en utilisant la forme analytique de la fonction F^{-1} réciproque de F. Ce qui restreint l'utilisation de cette procédure aux lois, peu nombreuses, pour lesquelles on sait la calculer.

ne garde pas cette valeur de X. On voit qu'il faudra dans notre cas[60] beaucoup d'essais "ratés" pour simuler des valeurs correctes, car on prend au hasard des points dans un rectangle d'aire 4 fois supérieure à celle sous la courbe (qui est 1 par définition de la fonction de densité). On aura donc besoin d'environ 400 tirages pour obtenir 100 valeurs. Cet algorithme s'étend sans difficulté théorique au cas d'une variable multidimensionnelle, en tirant comme ici une valeur de U dans l'intervalle (0,1), et autant de valeurs que de dimensions afin de se trouver dans un hyper-rectangle englobant toutes les valeurs de probabilité non nulles.

Les deux procédures que nous venons de décrire sont les plus anciennes. Dans la première, aucune valeur tirée n'est rejetée, mais il faut connaître non seulement la forme analytique de la fonction de densité, mais aussi celle de sa réciproque. La seconde, dans sa version très simplifiée, ne nécessite que la connaissance de la forme de sa fonction de densité, mais produit une bonne quantité d'essais infructueux. Toutefois la proportion de rejets peut être diminuée par l'ajout de connaissance experte dans le modèle. Depuis une cinquantaine d'année, les chercheurs ont fait preuve d'ingéniosité pour développer des procédures de simulation efficaces répondant à leurs besoins spécifiques (en biologie [173], en écologie [157]). Mais comme on vient de le voir pour ces deux procédures, augmenter l'efficacité d'une procédure de simulation a souvent pour effet connexe de restreindre son champ d'application. Pour notre part, nous ne pouvons pas accepter la méthode analytique, qui nous obligerait à spécifier la loi de distribution. Nous pouvons utiliser la méthode d'acceptation-rejet d'origine uniquement si la proportion de rejets reste raisonnable, ce qui n'a pas été le cas dans nos simulations (cf. partie II de ce mémoire).

3.2.4 *Bootstrap* et *jackknife*

D'après Bertail [21], ces méthodes ont été introduites par Quenouille (1949) pour le *jackknife*, repris par Tukey (1958), et par Efron en 1974 pour le *bootstrap* (également appelé *bootstrapping*). Ces deux méthodes font partie des *méthodes de rééchantillonnage* dont le but est d'évaluer la confiance qu'on peut avoir dans les valeurs de paramètres (comme la moyenne, la variance) calculés sur les données. Pour cela on tire dans les données un nouvel échantillon d'individus pour lequel on calcule le paramètre d'étude, et on répète cela pour un certain nombre d'échantillons, en observant les variations du paramètre.

D'après Christopher Z. Mooney et Robert Duval [180], le *bootstrap* et les statistiques inférentielles ont le même but : estimer la valeur de certains paramètres (comme la moyenne, la variance) de la population à partir de la distribution de ses valeurs, qu'on déduit de l'échantillon. La différence vient de la façon dont on la déduit. Pour les statistiques inférentielles c'est en utilisant des hypothèses sur la façon dont les valeurs sont distribuées dans la population alors que pour le *bootstrap*, on suppose qu'elles sont distribuées comme dans les données dont on dispose. C'est en ce sens qu'on peut dire que le *bootstrap* fait partie des méthodes non paramétriques.

Les échantillons ne sont pas construits de la même façon pour ces deux méthodes de rééchantillonnage. Pour le *jackknife* on construit les échantillons de la façon suivante :

1. On range les observations dans un ordre quelconque (l'ordre peut être celui de saisie des données), et on choisit la première observation
2. on crée l'échantillon formé de toutes les observations sauf celle choisie.
3. on choisit l'observation suivante s'il en reste au moins une et on retourne au point 2, sinon on arrête

[60] La procédure de Von Neumann n'utilisait pas le rectangle, mais une surface plus petite englobant celle sous la fonction de densité, ce qui produisait moins de rejets mais nécessitait de connaître au préalable la forme de la densité.

On dispose à la fin d'autant d'échantillons qu'on avait d'observations, sans répétitions.

Un échantillon *bootstrap* contient également des observations tirées des données sans modification mais elles sont tirées successivement et indépendamment "au hasard et avec remise", ce qui signifie qu'on autorise les répétitions. Les échantillons tirés contiennent autant d'observations que les données de départ, mais le nombre d'échantillons n'est plus imposé. Il est souvent de 100, mais il peut être plus grand, et on peut recommencer plusieurs fois ce tirage d'échantillons, car on trouvera des résultats a priori différents. On peut tirer de ces échantillons des intervalles de confiance des paramètres de la population à partir de l'ensemble des valeurs prises par le paramètre dans les échantillons. L'intervalle à 95% de confiance peut s'obtenir en excluant les valeurs les plus extrêmes afin de n'en garder que 95%. Ce "bootstrap naïf" à base de percentile a fait l'objet de nombreuses corrections et améliorations depuis (calcul différent des intervalles de confiance dans le *bootstrap-t*, modification des valeurs obtenues par pondération, choix des valeurs à rééchantillonner, etc. voir [173])

En dehors de la construction de versions "bootstrappées" des tests classiques des statistiques inférentielles, qui est à leur origine, ces méthodes sont utilisées comme techniques de validation (contrôle de la stabilité d'une ACP [22], validation des règles d'affectation produites par une analyse discriminante dans les logiciels SPAD et SAS [184]). Cette utilisation plus récente a été rendue possible par l'augmentation de puissance des ordinateurs car chaque tirage d'échantillon doit être suivi d'une phase de traitement identique à celle qui a été faite sur les données, la validation à l'aide d'un tirage de n échantillons pouvant ainsi être n fois plus longue que le traitement qu'on veut valider.

3.2.5 Les tests de permutation et de randomisation

On peut voir les tests de permutation comme une alternative combinatoire aux tests d'hypothèses des statistiques classiques quand les conditions exigées par ces derniers ne sont pas vérifiées par les données. Phillip Good, dans son ouvrage consacré à ces tests [96], va plus loin en affirmant que ce sont les tests d'hypothèses classiques qui sont une alternative théorique aux tests de permutation quand la puissance de calcul est insuffisante. Citons-le (page 10 de son ouvrage, dans la section intitulé "History") : « World war II provided impetus for developping a theorical basis for parametric procedures that would "serve" in place of the correct but computationnaly demanding permutations.»

Les tests de permutation consistent à tester une hypothèse (l'hypothèse nulle H0) sur des données en calculant la proportion de données simulées qui sont au moins aussi extrêmes que les données d'origine selon cette hypothèse. Cette proportion correspond au risque α de se tromper en rejetant H0. Les données simulées sont créées à partir des données d'origine en gardant toutes les observations, toutes les variables et toutes les valeurs présentes dans les données observées, mais en réaffectant différemment les valeurs des variables aux observations sous l'hypothèse H0. Dans la mesure où chaque variable garde globalement le même ensemble de valeurs, il s'agit d'une permutation entre ces dernières. Toutes les permutations possibles sous H0 sont envisagées et comparées aux données d'origine. Le plus ancien de ces tests de permutation, remonte à Fisher (1932), et selon Good ces tests sont redécouverts régulièrement depuis, notamment les tests utilisant les rangs [214, 220, 70] font partie des tests de permutation car ils procèdent par échange des rangs des valeurs.

Un prédécesseur : le test exact de Fisher

Le test de permutation le plus connu est le test exact de Fisher (1932) que l'on retrouve dans les ouvrages de statistiques de base, à la suite du test du Chi2 d'indépendance [206, 214]. On l'illustre habituellement par un tableau de contingence croisant deux propriétés à deux modalités chacune, donc de quatre cases, avec un effectif très faible dans l'une de cases, et des effectifs un peu plus élevés pour les trois autres, le total n'étant pas très élevé (cf. exemple du tableau 3.1). La question qui se pose est alors de décider si la valeur de l'effectif le plus faible est due à un lien entre les deux propriétés ou au seulhasard. Pour l'établir, on part des effectifs marginaux, comme dans le test du Chi2, mais au lieu de calculer les effectifs théoriques en cas d'indépendance, on fait le compte de toutes les distributions de valeurs conjointes qui auraient pu aboutir aux mêmes effectifs marginaux. Reprenons l'exemple que citent Henri Rouanet, Jean-Marc Bernard et Brigitte le Roux dans leur ouvrage [206] : 5 sujets pour lesquels on aurait relevé les valeurs suivantes de taille et de poids relativement à une taille de 171 cm et un poids de 82 kgs selon le tableau 3.1 :

numéro	s1	s2	s3	s4	s5
Taille	-	-	+	+	+
Poids	-	-	-	+	+

Taille/Poids	+	-	total
+	2	1	3
-	0	2	2
Total	2	3	5

TAB. 3.1 – Test exact de Fisher : à gauche, les valeurs de 5 sujets pour deux variables, à droite la répartition des 5 sujets selon leurs valeurs.

On remarque dans la partie droite de ce tableau un effectif nul : il n'y a pas de sujets pour lesquels la taille est inférieure à 171 cm (codée '-') et le poids supérieur ou égal à 82 kgs (codé +). Ce qui peut surprendre si on fait l'hypothèse H0 d'indépendance entre les poids et les tailles. Dans le cadre de cette hypothèse nulle, les valeurs des poids peuvent être interchangées sans problème si on ne change pas les tailles[61] Le nombre d'échantillons différents obtenus par permutation est donc 10 (choix des 2 sujets qui auront un '+' parmi les 5, ou ce qui revient au même, des 3 qui auront un '-' parmi les 5),

Dans le tableau 3.2 on a donné les valeurs des poids pour les 10 échantillons, et on a calculé pour chacun le nombre de sujets ayant une petite taille (T='-') et un grand poids (P='+') En première ligne figurent les données réelles. On constate que seulement les 3 premiers échantillons ech1, ech2 et ech3 n'ont aucun sujet de petite taille et de grand poids, alors que 6 échantillons en ont un, le dernier échantillon en ayant deux. Ainsi la probabilité d'avoir si peu de sujets de ce type est de 3/10, bien supérieure au seuil de 5% (et même 10%) en dessous duquel on rejette habituellement l'hypothèse nulle. On ne peut donc rejeter l'hypothèse d'indépendance entre la taille et le poids et on attribue au hasard cet effectif nul. Ce résultat n'est pas étonnant au vu de la faiblesse de l'effectif : en répartissant 5 sujets dans 4 cases, avec des marges équilibrées, on peut s'attendre à obtenir des effectifs nuls plus souvent qu'avec dix fois plus de sujets. L'intérêt de ce petit exemple est de montrer qu'on peut tester l'indépendance[62] par simple comptage au

[61] On pourrait échanger les valeurs des tailles au lieu de celles des poids, cela donnerait le même nombre d'échantillons, et les mêmes échantillons, à la renumérotation des sujets près. On pourrait aussi échanger simultanément les tailles et les poids, cela multiplierait les échantillons, (par 10 dans notre cas) mais chacun serait dupliqué à la renumérotation des sujets près ce même nombre de fois, et on aboutirait au même résultat

[62] Et pas seulement l'indépendance dans cet exemple, comme le signalent les auteurs, mais également la comparaison d'une fréquence à une valeur donnée et l'homogénéité de deux fréquences pour deux groupes indépendants.

Sujet	s1	s2	s3	s4	s5
Taille	-	-	+	+	+

	Poids					Effectif pour
Échantillon	s1	s2	s3	s4	s5	T=- et P=+
ech1	-	-	-	+	+	**0**
ech2	-	-	+	-	+	0
ech3	-	-	+	+	-	0
ech4	-	+	-	-	+	1
ech5	-	+	-	+	-	1
ech6	-	+	+	-	-	1
ech7	+	-	-	-	+	1
ech8	+	-	-	+	-	1
ech9	+	-	+	-	-	1
ech10	+	+	-	-	-	2

TAB. 3.2 – En haut la taille fixée, en dessous les poids selon les 10 échantillons, à droite le nombre de sujets pour lesquels T='-' et P='+', en gras les données d'origine

lieu d'utiliser des formules complexes valables uniquement dans des conditions spécifiques. Ici les effectifs trop petits sont une violation rédhibitoire pour l'application du test du Chi2.

L'art de construire un bon test de permutation

Selon Phillip Good [96], tout test d'hypothèse des statistiques classiques a son équivalent en test de permutation, ce dernier étant dans tous les cas de qualité supérieure ou égale à celui qu'il remplace. Marc Hallin écrit dans le chapitre intitulé "tests sans biais, tests de permutation, tests invariants, tests de rangs" pages 101 à 127 de l'ouvrage [70], que c'est leur absence de biais qui fait leur principal attrait, et que de ce fait, "ils devraient recevoir beaucoup plus d'attention de la part des praticiens". Il existe toutefois une difficulté théorique : comment échanger les valeurs sous H0 ? En effet le but d'un test n'est pas de valider H0, mais de la rejeter pour prouver son alternative H1, en utilisant la valeur extrême (ou supposée comme telle) d'une "statistique" calculée sur les données vérifiant H1[63]. Les permutations ont pour rôle de "rétablir" le hasard en détruisant le lien dû à H1. Pour pouvoir juger de l'extrémalité de la valeur de la statistique correspondant aux données d'origine, il faut calculer les valeurs pour toutes les permutations. Si cette position extrême est confirmée, elle pourra être attribuée à H1 à condition que les permutations n'aient pas cassé d'autres liens dans les données. Et c'est la toute la difficulté de ces tests : la construction d'une bonne permutation requiert autant d'ingéniosité que celle d'une bonne expérience (voir la partie 2.1.8 de ce document qui expose les conditions d'une bonne expérience). A cette difficulté s'ajoute une difficulté pratique due à l'explosion combinatoire du nombre de permutations quand la taille des données augmente. Là encore l'ingéniosité peut permettre de réduire le coût informatique. Pour chaque test proposé dans l'ouvrage de Phillip Good [96], la méthode de permutation est décrite en détail ainsi que ses alternatives moins coûteuses

[63] Dans l'exemple décrit juste avant, la statistique est le nombre de sujets de petite taille et de grand poids, l'hypothèse nulle H0 est l'indépendance entre taille et poids, l'hypothèse alternative H1 étant que la taille et le poids sont liés positivement, les permutations se faisaient entre les 5 valeurs de la variable "poids" en échangeant seulement des '+' avec des '-', soit 10 permutations au lieu des 120 possibles.

en temps machine quand elles existent. On peut notamment tirer au hasard des permutations parmi l'ensemble de toutes celles possibles. Cette variante des tests de permutation fait partie des améliorations proposées par l'auteur à la fin de son ouvrage et pour Eugène S. Edgington [73] et Bryan E.J. Manly [173], c'est une autre interprétation de ces tests, qu'ils appellent "test de randomisation".

Notons que l'utilisation de ces tirages au hasard impose des précautions supplémentaires. La génération de séquences de nombres par ordinateur pour simuler des tirages au hasard a fait l'objet de nombreux perfectionnements depuis un demi-siècle [176], et des procédures variées ont été définies pour contrôler la qualité de ce "pseudo-hasard". En cas de doute[64] il suffit de faire les tests préconisés afin de choisir la fonction de hasard plus appropriée. Mais un autre problème plus difficile est de s'assurer que les permutations que nous tirons au hasard sont bien représentatives de l'ensemble des permutations possibles. Si la permutation a été construite en relation avec une hypothèse complexe et/ou si elle est elle-même complexe, une vérification de la représentativité s'impose à notre avis une fois les tirages effectués. N'ayant pas trouvé, dans les ouvrages traitant de randomisation, de méthodes de contrôle de ce type de qualité, pas plus que dans les articles lus, nous avons tenté d'en construire nous-mêmes dans la partie II, et c'est à la lumière des résultats décevants de nos premiers essais que nous avons pu améliorer notre méthode de randomisation. *bootstrap*

3.2.6 Conclusion

Les méthodes que nous venons d'exposer permettent à un chercheur en sciences humaines disposant d'un ordinateur et d'une certaine connaissance de l'informatique (allant de l'utilisation approfondie d'un tableur à une petite pratique d'un langage de programmation) d'accompagner les inductions qu'il fait à partir de ses données de toute une panoplie de vérifications possibles. Il peut éprouver la valeur de ses estimations ou la stabilité de ses analyses par *bootstrap* ou *jackknife*. Il peut éprouver les relations qu'il suppute en créant par des simulations de Monte Carlo des données artificielles selon ses hypothèses et en les confrontant à ses propres données ; il peut aussi les éprouver en confrontant ses données à des copies randomisées de celles-ci dans lesquelles les relations ont été supprimées par permutation. Bien sûr, il peut également utiliser les logiciels de traitement des données qui intègrent au fur et à mesure ces nouveaux tests si ses données ou ses hypothèses s'y prêtent. Mais quand il a des données ou des hypothèses très particulières, il peut dorénavant créer lui-même les outils de traitement adaptés à ses données. C'est ce qui se fait depuis un certain temps en écologie [157], mais à notre connaissance, mis à part le *bootstrap* et le *jackknife*, ces méthodes ne se sont pas encore généralisées à la psychologie.

Ce sont ces raisons qui sont à l'origine de la création d'une méthode à base d'échanges que nous exposons dans la partie II de ce document. Le test de randomisation utilisant cette méthode de permutation a plusieurs buts : traiter des données spécifiques (l'exemple traité porte sur des distributions de valeur zipfiennes, la loi de probabilité d'Estoup-Zipf n'étant prise en compte dans aucun test à notre connaissance), mais également tester des types particuliers de liaisons entre deux variables (les indices de qualité des règles d'association ne disposent pas de tests associés), et échapper au problème des comparaisons multiples [131, 126] (qui rend incertain le calcul des valeurs des tests usuels quand on a une dizaine de variables, et inutilisable quand elles atteignent ou dépassent la centaine)

[64] "Maurice Clerc invite à "se méfier du hasard" dans les pages 71 à 79 de son ouvrage sur "l'optimisation par essaims particulaires" [53]. Toutefois les procédures d'optimisation sont certainement, de par leur nature, plus sensibles à la qualité du "hasard" que les procédures de randomisation.

3.3 Les nouvelles méthodes de discrimination

On dit aussi classification supervisée, ou classement. Si on dispose, pour un ensemble d'individus ou d'observations - dit *ensemble d'apprentissage* - à la fois des valeurs d'un ensemble de variables dites explicatives, et des valeurs d'une variable catégorielle à expliquer, le problème de la discrimination consiste à induire les valeurs de la variable à expliquer pour un ensemble de nouvelles observations ou individus sur lesquels on ne dispose que des valeurs des variables explicatives. En pratique, la majorité des problèmes relevant de l'*apprentissage automatique* (machine learning) ou de l'*extraction de connaissances* (knowledge discovery) se ramènent à ce problème de discrimination, dit encore *classification supervisée*.

Des indicateurs de qualité de discrimination (sensibilité et spécificité, dénommés aussi rappel et précision, Fscore, ...) permettent d'évaluer les méthodes pour y parvenir, ou la difficulté des tâches, à méthode égale. Des défis (*Challenges*) ont été mis en place dans divers domaines d'application (fouille de textes, de données biomédicales, ...) pour comparer les méthodes et le travail des équipes participantes, et faire progresser ce nouveau domaine d'ingénierie - domaine où les statistiques tiennent une place de choix pour maximiser la capacité de généralisation du dispositif (stratégies de sélection et rotation d'échantillons d'entraînement, de mise au point, de test, ...).

La discrimination représente une application directe des méthodes d'extraction de règles d'association significatives qui sont le sujet de cette thèse, en ce sens qu'il suffit d'extraire les seules règles dont la partie droite est constituée par la variable à expliquer. Nous avons ainsi participé, plutôt avec succès, à plusieurs de ces défis : KDD Cup 2003 [65], 2004 [66], 2006 [67], Défi Fouille de Textes 2006[68].

La plupart des méthodes utilisées dans ce domaine sont d'inspiration plus numériques que symboliques :

- *Analyse factorielle discriminante*, linéaire à l'origine [82], aujourd'hui pourvue d'extensions non-linéaires [58].
- *Arbres de décision* [246], Forêts aléatoires[69].
- *Réseaux neuronaux supervisés*, en particulier le plus connu d'entre eux : le *Perceptron* [204], linéaire à l'origine et critiqué pour son incapacité à prendre en compte un effet d'interaction simple, le XOR, puis renaissant sous sa forme multicouche, non-linéaire dans les années 80 [154].
- *Séparateurs à Vastes Marges* (SVM, Support Vector Machines) [104], dans le cadre desquels a été conçu le " kernel trick " consistant à remplacer la brique de base 'produit scalaire' par un noyau (polynomial, gaussien, ...), pour passer de surfaces séparatrices linéaires à surfaces non-linéaires prenant en compte le phénomène d'interaction ; mais on perd au passage la transparence des méthodes linéaires qui permettait d'illustrer en termes de variables le processus de discrimination.

Ces méthodes peuvent être perfectionnées, par exemple la procédure dite de *boosting* [184] consiste à donner un poids croissant aux éléments permettant d'augmenter le rappel et la précision pour une méthode d'apprentissage automatique donnée. On peut aussi combiner différentes méthodes (*bagging* [184]).

[65] http ://www.cs.cornell.edu/projects/kddcup/
[66] http ://kodiak.cs.cornell.edu/kddcup/
[67] http ://www.kdd2006.com/kddcup.html
[68] http ://www.lri.fr/ia/fdt/DEFT06/
[69] On peut trouver actuellement sur Internet à l'adresse http ://stat-www.berkeley.edu/users/breiman un petit logiciel de Leo Breiman et Adele Cutler accompagné de la documentation concernant leur méthodologie.

Notre expérience des tâches de discrimination, par exemple à l'occasion de DEFT'06 [160], et celles d'autres, par exemple celle des vainqueurs DEFT'05 [75], montre qu'il peut être efficace de préférer à des méthodes automatiques et aveugles, des méthodes " informées " explicitant la connaissance contenue dans le processus de discrimination, donc les variables en jeu, voire utilisant une connaissance extérieure. C'est pourquoi nous pensons que l'apport de cette thèse en matière d'extraction minimale et de validation statistique des motifs contenus dans un ensemble d'apprentissage est important pour le problème de la discrimination : il permet d'expliciter les variables en jeu et leurs interactions, quel que soit l'ordre de ces interactions.

3.4 Les nouvelles méthodes d'investigation des données

Les méthodes que nous venons de décrire visent une appréhension globale des données. De nouvelles méthodes sont apparues visant à mettre à jour une structure plus locale des données. Les deux formes principales de ces méthodes sont les *réseaux bayésiens* et l'extraction de *motifs* et de *règles d'association*. Les réseaux bayésiens permettent une automatisation du raisonnement humain par navigation à travers l'enchevêtrement des relations de causes à effets entre les variables à partir de "noeuds" privilégiés. L'extraction de motifs permet de trouver des "amas" particuliers de variables au sein des données, et l'extraction des règles d'association de les transformer en "pépites" (nuggets) de connaissances. Nous décrivons ces deux méthodes dans les deux sous-sections suivantes.

3.4.1 Les réseaux bayésiens

Un exemple : raisonner comme Sherlock Holmes

Un réseau bayésien est une représentation par un graphe de liens probabilistes entre faits. Reprenons l'exemple très pédagogique que donne Jensen dans son introduction aux réseaux bayésiens [133] en nous autorisant toutefois une certaine liberté d'adaptation de l'histoire qu'il raconte :

Sherlock Holmes sort de chez lui le matin pour aller travailler. En arrivant à sa voiture, il constate que sa pelouse est mouillée alors qu'il ne pleut pas. Il se dit qu'il a dû oublier de couper son système d'arrosage la veille au soir. Il se dirige vers la cave pour aller l'arrêter quand il jette un coup d'oeil sur la pelouse de son voisin Watson : elle est mouillée. Il rebrousse alors chemin et monte dans sa voiture pour se rendre à son travail.

Son raisonnement peut être représenté comme un parcours dans un réseau causal comportant quatre faits et trois relations,

Voici les quatre faits qui peuvent prendre la valeur "Vrai" ou "Faux" :

- H : La pelouse de Holmes est mouillée
- W : La pelouse de Watson est mouillée
- A : Le système d'arrosage de Holmes n'a pas été coupé
- P : Il a plu

Et voici les trois relations causales qui les lient :

1. P–>H : s'il a plu, la pelouse de Holmes est mouillée
2. P–>W : s'il a plu, la pelouse de Watson est mouillée
3. A–>H : si le système d'arrosage de Holmes n'a pas été coupé alors la pelouse de Holmes est mouillée

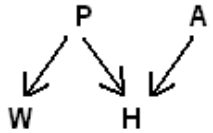


FIG. 3.3 – Petit exemple de réseau bayésien emprunté à Jensen [133]

Holmes utilise l'information au fur et à mesure qu'elle arrive pour évaluer la probabilité qu'il ait oublié de couper l'arrosage. Ainsi, il prend trois décisions successives contradictoires :

1. le risque que A soit vrai est a priori faible. Il va directement à sa voiture
2. Il constate que H est vrai. Cela augmente suffisamment à ses yeux le risque que A soit vrai pour justifier un détour par la cave
3. Il constate que W est vrai. Cela diminue suffisamment à ses yeux le risque que A soit vrai pour ne plus justifier un détour par la cave. Il rebrousse chemin et monte dans sa voiture.

Pour modéliser son comportement, il faut rajouter des probabilités. On associe à chaque fait ayant plusieurs causes une table de probabilités donnant toutes les probabilités de ce fait conditionnellement aux autres. Les probabilités de ces faits, selon l'auteur, sont dans les tableaux 3.3.

P=1	0.2
P=0	0.8

P(P)

A=1	0.1
A=0	0.9

P(A)

	P=1	P=0
W=1	1	0.2
W=0	0	0.8

P(W|P)

	P=1		P=0	
	A=1	A=0	A=1	A=0
H=1	1	1	0.9	0
H=0	0	0	0.1	1

P(H|P,A)

TAB. 3.3 – Les probabilités conditionnelles des 4 faits de la figure 3.3

Explicitons ces tableaux. Celui de gauche n'a que deux lignes car aucune flèche n'arrive à P (voir la figure 3.3, ce qui signifie qu'il n'est l'effet d'aucune cause dans ce modèle. On a en première ligne la probabilité 0.2 que P soit vrai ($P(P = 1) = 0.2$[70]), donc qu'il ait plu cette nuit. Et dans la ligne suivante 0.8, qui est la probabilité qu'il n'ait pas plu. Pour toutes les colonnes de ces tableau la somme sera 1. A côté on a le même type de tableau pour A. Notons que pour une valeur de P(A=1) de 0.1, Holmes ne fait pas de détour pas la cave. Le tableau de P(W|P) contient 4 valeurs. En haut à gauche, c'est la probabilité que W soit vraie, sachant que P est vraie. Elle est de 1, ce qui signifie que s'il a plu, on a la certitude que la pelouse de Watson est mouillée. Par contre s'il n'a pas plu, il y a quand même une probabilité de 0.2 pour que la pelouse de Watson soit mouillée, certainement parce qu'il l'arrose parfois lui aussi. Le dernier tableau s'interprète de la même façon. Cette fois, on a deux causes qui sont A et P. Que Holmes ait oublié ou non le système d'arrosage ouvert, s'il a plu, on est certain que sa pelouse est mouillée. Par contre s'il n'a pas plu, la probabilité pour que la pelouse soit mouillée est nulle s'il n'a pas arrosé et de 0.9 dans le cas contraire.

[70] On note Vrai par 1 et Faux par 0

La formule de décomposition des probabilités nous permet de calculer la loi de probabilité conjointe en fonction des probabilités conditionnelles

$P(W,P) = P(W|P)P(P)$ ainsi que $P(H,P,A) = P(H|P,A)P(P,A)$

En prenant en compte l'indépendance de P et de A, on peut remplacer P(P,A) par le produit de P(P) et de P(A), ce qui donne

$P(H,P,A) = P(H|P,A)P(P)P(A)$.

On obtient alors les 4 tableaux des lois conjointes de la table 3.4 auxquels on a rajouté les deux lois marginales P(W) et P(H) obtenues en sommant les cellules des lignes des lois conjointes.

P=1	0.2
P=0	0.8

P(P)

A=1	**0.1**
A=0	0.9

P(A)

	P=1	P=0	
W=1	0.2	0.16	0.36
W=0	0	0.64	0.64

P(W,P) P(W)

	P=1		P=0		
	A=1	A=0	A=1	A=0	
H=1	0.02	0.18	0.07	0	0.27
H=0	0	0.1	0.01	0.72	0.73

P(H,P,A) P(H)

TAB. 3.4 – Les 4 probabilités marginales et les 2 conjointes déduites du tableau 3.3

On vient de voir comment on est parti des probabilités a priori sur A et P, et des probabilités conditionnelles de W et H pour obtenir les probabilités conjointes P(W,P) et P(H,P,A), et en déduire les probabilités marginales de W et H. C'est un raisonnement de déduction classique qui se fait en suivant les flèches du modèle dans le graphique 3.3. Maintenant on dispose de la connaissance supplémentaire que H est vrai. On peut alors remonter les flèches de la conséquence H à la cause A en utilisant la formule de Bayes (aussi appelée *de probabilités des causes*)[71]. Pratiquement cela consiste à multiplier les probabilités conjointes de P(H,P,A) par un coefficient pour que la première ligne ait pour somme 1 (c'est le quotient de la nouvelle valeur sur l'ancienne de P(H=1)), et pareillement pour que la deuxième soit de somme nulle. Une fois les valeurs de P(H,P,A) réactualisées, on réactualise aussi celles de P(A) et P(P) par sommation et on reprend le sens "normal" de la flèche pour réactualiser P(W,P), ce qui donne les valeurs du tableau 3.5.

P=1	0.74
P=0	0.26

P(P)

A=1	**0.34**
A=0	0.66

P(A)

	P=1	P=0	
W=1	0.74	0.05	0.79
W=0	0	0.21	0.21

P(W,P) P(W)

	P=1		P=0		
	A=1	A=0	A=1	A=0	
H=1	0.07	0.66	0.26	0	1
H=0	0	0	0	0	0

P(H,P,A) P(H)

TAB. 3.5 – Modification des valeurs du tableau 3.4 prenant en compte la certitude que H est vrai

On constate à la lecture de ce tableau que P(A=1) est passé à 0.34, alors qu'il était de 0.16 dans le précédent. Ainsi la prise en compte de l'information sur H a augmenté la probabilité que A soit vrai, ce qui a poussé Holmes a faire un détour par sa cave.

Il reste une nouvelle information à prendre en compte, c'est la certitude que W est vrai. On met à jour les valeurs de la même façon que précédemment en commençant par la loi conjointe

[71] D'après cette formule, si X et Y sont deux évènements, on peut écrire $P(X|Y) = \frac{P(Y|X)P(X)}{P(X,Y)}$. Pour le détail de la formule concernant les trois évènements et des calculs associés, nous renvoyons le lecteur intéressé à l'ouvrage de Jensen [133]

de P(W,P) (voir tableau 3.6). La probabilité de A[72] redescend à 0.16, et Holmes juge le risque suffisamment faible que A soit vrai pour repartir sans passer par la cave.

P=1	0.93
P=0	0.07

P(P)

A=1	**0.16**
A=0	0.84

P(A)

	P=1	P=0
W=1	0.93	0.07
W=0	0	0

P(W,P)

1
0

P(W)

	P=1		P=0	
	A=1	A=0	A=1	A=0
H=1	0.09	0.84	0.07	0
H=0	0	0	0	0

P(H,P,A)

1
0

P(H)

TAB. 3.6 – Modification des valeurs du tableau 3.5 prenant en compte la certitude que W est vrai

Pour conclure sur l'exemple

Le raisonnement utilisé dans ce petit exemple est basé sur des faits (pouvant prendre la valeur vrai ou faux) et sur des relations de causes à effets. On parcourt ce réseau de relations représenté dans la figure 3.3 en suivant le sens des flèches, mais également en sens inverse. Cette "navigation entre faits" rappelle celle qu'utilise Alker dans l'exemple de la partie 2.2.2 pour rechercher les causes et les effets parmi trois variables sociologiques concernant la ségrégation raciale.

Les modèles de ces deux exemples diffèrent de par leur nature et leur utilisation. Le modèle du réseau bayésien est de type probabiliste (avec des équations de probabilités donnant des solutions exactes) alors que l'autre est de type statistique (avec des équations comportant des termes d'erreur). Les deux modèles s'utilisent différemment sur des données réelles : l'un permet d'utiliser à bon escient l'information apportée par les données en suivant de façon automatique le cheminement du raisonnement de l'expert entre les diverses règles régissant les données, l'autre permet de découvrir de nouvelles règles régissant les données en confrontant l'information qu'il contient à divers modèles tant statistiques que sociaux, selon un cheminement propre à l'auteur.

Les principes de fonctionnement d'un réseau bayésien

Le réseau bayésien est une écriture probabiliste de la connaissance d'un expert à un moment donné. Elle est formée de deux parties : la structure du réseau constituée des relations de causes à effets et codée par des probabilités conditionnelles, et les valeurs de probabilités des évènements (ou faits, propriétés, ou variables). La structure n'est pas susceptible de modification alors que les probabilités des évènements sont toutes remises à jour dès que l'une d'elles est modifiée pour intégrer une nouvelle connaissance.

Nous avons vu l'utilisation de ce type de réseau pour faire des raisonnements "à la Sherlock Holmes". D'un coté trois règles de type causal, immuables, et de l'autre quatre faits dont les valeurs de probabilité changent au fur et à mesure que de nouvelles informations arrivent. L'un de ces faits intéresse plus particulièrement Holmes, c'est A : "le système d'arrosage n'a pas été coupé". La mise à jour suite à l'intégration d'une nouvelle information se propage à partir du fait dont la probabilité vient de changer (d'abord H puis W dans l'exemple) selon un cheminement dans le réseau entièrement déterminé par la structure. Pour cet exemple, il n'était que de deux

[72] En fait il s'agit ici de P(A=1|W=1,H=1) et non P(A=1) comme indiqué dans le tableau, de la même façon, il faudrait écrire dans le tableau 3.5 P(A=1|H=1) au lieu de P(A=1)

types : suivre les flèches ou les remonter, mais cela suffisait pour simuler un raisonnement humain assez "futé".

L'évolution des réseaux bayésiens

La création de ce type de réseau répond, d'après Olfa Ben Naceur-Mourali et Christophe Gonzales [18], à un besoin d'assouplissement des systèmes experts à base de règles, ces derniers ne fonctionnant qu'avec des faits certains. D'après les auteurs, MYCIN, créé en 1976 par Shortliffe, fut une première tentative d'introduction de l'incertitude au moyen de "facteurs de certitude", mais la première formalisation de cette incertitude véritablement opérationnelle date de Pearl (1988) : ce sont les réseaux bayésiens qui traduisent l'incertitude par des probabilités.

Depuis, les possibilités de ces réseaux se sont étendues, les algorithmes se sont développés : on a augmenté le nombre de types de cheminements possibles (le réseau peut même contenir des cycles), on a étendu le modèle aux variables continues et pris en compte le problème des données manquantes, tout cela sur des données de taille de plus en plus importante [73].

Depuis une quinzaine d'années, des développements sont en cours pour adapter au mieux la structure du réseau bayesien au problème qu'il traite. Divers algorithmes permettent la modification, voire la création de la structure du réseau, certains prenant en compte des variables cachées afin de renforcer l'expression de la causalité par le réseau. Philippe Leray signale dans [165] la difficulté de construction de cette structure quand on dispose de nombreuses variables. En effet cette construction se base sur un test statistique d'indépendance conditionnelle, qui, quand on l'utilise en grande dimension, donne des résultats peu fiables[74]. Il cite des algorithmes conçus par Chen et al. entre 1997 et 2002 qui permettent de ne faire qu'un nombre limité de ces tests. Pour l'estimation de l'adéquation du réseau aux données se pose également un problème statistique classique : si on ne prend en compte que l'écart entre les données estimées par le modèle et les données réelles pour juger de sa qualité, le "meilleur modèle" sera celui qui codera le plus d'informations des données (ici un grand nombre de flèches dans le graphe) et sera donc complexe et "collera" trop aux données. Une des solutions retenues est de pénaliser ce score par un terme dépendant de la taille du réseau (critères AIC de Akaike en 1970 ou BIC de Schwartz en 1978 qu'on peut trouver détaillés dans [98]).

Conclusion sur les réseaux bayésiens

La conception d'un réseau bayésien permet la prise en compte automatisée d'un grand nombre de relations de causes à effets enchevêtrées pour établir par exemple un diagnostic, ou un comportement de robot, en imitant un raisonnement humain assez complexe. La force de ces réseaux consiste en la variété des possibilités de circulation de l'information, qui peut se faire se faire localement dans des sous-groupes de variables. La difficulté essentielle est d'écrire au préalable la structure de ce réseau bayésien, en fixant toutes les valeurs de probabilités conditionnelles. Des solutions sont proposées pour apprendre cette structure à partir des données, les plus récentes semblant pouvoir traiter le cas de gros réseaux, notamment en les décomposant en zones assez indépendantes les unes des autres sur chacune desquelles une structure est apprise. A notre connaissance, l'interaction n'est pas modélisée dans cette structure, et il paraît difficile de le faire. En effet les informations circulent essentiellement selon des formules de probabilités bayésiennes, dans lesquelles l'interaction n'est pas formalisée.

[73] Philippe Leray évoque dans [165] une application médicale, le système Pathfinder de 1992 qui traite 130 symptômes, 60 diagnostics : 75000 probabilités ont dû être spécifiées.

[74] Ce problème de passage à grande échelle des tests statistiques est un des points, centraux dans notre thèse, que nous avons évoqués dans les parties précédentes.

3.4.2 L'extraction de motifs et de règles d'association

L'extraction de motifs et de règles d'association faisant partie des méthodes de la fouille de données (Data Mining). Nous essayons d'abord de donner une définition de la fouille de données puis nous précisons ce qu'est l'extraction de motifs et de règles d'association.

Fouille de données et extraction de connaissances

Pour Jiawei Han et Micheline Kamber, dans leur ouvrage sur le Data Mining [113], les techniques de fouille de données (Data Mining) se sont développées à partir des années 1990 et sont liées à l'évolution des "bases de données". Créées dans les années 1960, les bases de données (databases) se sont complexifiées pour faire face à l'afflux de plus en plus important de données de plus en plus complexes intégrant notamment multimedia et Web. Le modèle relationnel apparu dans les années 1970 permettait le stockage très structuré de données de taille et de complexité raisonnables, et les langages de requêtes comme SQL suffisaient à leur exploration. La croissance informatique a permis l'avènement des entrepôts de données (Data Warehousing), la disponibilité de grosses bases de données (génomique, astronomie, multimedia, Web, etc.) et en même temps un foisonnement de nouvelles formes d'investigation des données pour en extraire de la connaissance. Jiawei Han et Micheline Kamber définissent [113] la fouille de données comme équivalente à l'Extraction de Connaissances dans les bases de Données (ECD ou KDD : Knowledge Discovery in Databases), et ils précisent sa nature par la phrase suivante reprise de [78] (les mots soulignés par les auteurs ont été mis ici en gras) :

« *Extraction of interesting (**non-trivial**, **implicit**, **previously unknown** and **potentially useful**) information or patterns from data in **large databases***»

La fouille de données est selon eux un processus complet comportant l'étude du problème posé, la sélection des données, leur nettoyage et pré-traitement, leur réduction et transformation, le choix du traitement, de l'algorithme, la recherche des "patterns" d'intérêt, l'évaluation et la mise en forme des résultats, et l'utilisation de la connaissance découverte. Selon leur définition, tout ce qui a été décrit dans cette partie s'inscrit dans la fouille de données. La seule exception qu'ils font est :

« *What is not data mining ?*
- *(Deductive) query process*
- *Expert systems or small ML/statistical programs* »

Cette définition d'une fouille de données qui engloberait toutes les connaissances et pratiques connues précédemment pour être connexes au traitement des données, est aussi celle de Cornuéjols [56] pour lequel l'apprentissage artificiel (ML : Machine Learning) en est l'étape qui permet de construire un modèle à partir des données. Selon lui la recherche d'*associations* fait partie des méthodes d'*apprentissage symbolique non supervisé.*

Règles d'association et fouille de données

Jiawei Han et Micheline Kamber consacrent un chapitre de leur livre à l'extraction des règles d'association qui est « probablement la contribution la plus significative de la communauté des bases de données à la fouille de données ». Selon eux, extraire des associations signifient trouver des cooccurrences, des corrélations ou des structures causales. A notre connaissance, c'est un des rares ouvrages sur la fouille de données mentionnant l'utilisation de cette méthode pour découvrir des relations de causes à effets. La description de cette méthode se limite souvent à quelques exemples et définitions dans les ouvrages de data mining, excepté ceux destinés aux gestionnaires [156]. Dans ces derniers, elle est plus détaillée, faisant partie des méthodes

qu'ils apprécient. Toutefois leur intérêt pour cette méthode porte sur la prédiction et non sur l'explication.

Rapide présentation des règles d'association

L'extraction de motifs et de règles d'association sera décrite en détail dans le prochain chapitre car c'est la méthode sur laquelle ce mémoire est centré. Mais donnons dès à présent un exemple de règle d'association. Nous choisissons l'exemple-type d'une règle du style "panier de la ménagère" qui pourrait être extraite d'une base de données de la grande distribution à partir de l'examen des tickets de caisse d'un samedi soir à la l'heure de la fermeture :

"Si achat de bière et de charcuterie, alors achat de chips", support[75]=5%, confiance=30%,

ce qui indique que 5% des tickets de caisse comportent l'indication d'un achat de bière, de charcuterie et de chips et qu'ils forment 30% des tickets qui comportent l'indication d'un achat de bière et de charcuterie.

La découverte de cette règle a nécessité de compter tous les tickets de caisse comportant les trois achats, et ceux ne comportant que les 2 premiers. Cette opération est l'extraction de motifs préalable à l'extraction de règles. Ici les deux motifs extraits sont "achat de bière, de charcuterie et de chips" et "achat de bière et de charcuterie".

Bien que nous ayons donné un exemple avec trois variables, ce sont souvent des exemples de règles avec deux variables qui font l'objet d'interprétation, et les règles portant sur plusieurs variables ont presque toujours une seule variable dans leur partie droite (conclusion) et plusieurs dans leur partie gauche (prémisse).

Pour conclure sur cette méthode d'extraction d'association, elle permet de trouver des pépites locales aux données, cette localité étant relative aux variables, comme avec le réseau bayésien, mais aussi aux individus [150].

Induction par des règles

Quand en membre droit des règles d'association on trouve toujours la même variable, ses modalités seules changeant, les règles d'association peuvent alors permettent de faire de l'induction, comme les *règles de décision* [246]. Les différences essentielles entre les deux types de règles portent sur la nature des variables traitées, l'algorithme utilisé pour l'extraction, et la représentation des résultats de l'extraction. Les variables sont catégorielles pour les règles d'association mais peuvent être numériques pour les règles de décision, excepté la variable cible (c'est-à-dire figurant dans la conclusion) qui est en principe catégorielle. L'extraction des règles de décision n'est pas précédée d'une extraction préalable des motifs. L'algorithme utilisé permet de découper l'intervalle de valeurs de la variable numérique en deux parties, donc à la recoder en variable catégorielle, ce recodage étant toutefois local à la règle. A la fin de l'extraction, un arbre de décision est construit, qui fait de l'induction par règle de décision une des méthodes les plus appréciées par les praticiens devant prendre des décisions, par exemple en médecine. La complexité des jeux de règles d'association rebute au contraire les praticiens, vite débordés par le nombre impressionnant de règles à examiner, par leur interprétation difficile et leur généralisabilité difficile à évaluer.

Daniel T. Larose, dans son ouvrage d'introduction au data mining [150], signale qu'il existe une version des règles d'association pour lesquelles on n'exige des variables catégorielles qu'en

[75]Selon les cas, le support peut indiquer le nombre total de tickets de caisse comportant ces trois achats simultanément ou comme ici le pourcentage de ceux-ci rapporté au nombre total de tickets de caisse.

conclusion, celles en prémisse pouvant être numériques. Il s'agit de l'Induction de Règles Généralisée (GRI) présentée par Smyth et Goodman en 1992. Toutefois l'extraction de ces dernières se fait selon une méthode plus proche de celle des règles de décision que des règles d'association.

4

État de l'art des règles A→B associées à des données

Ce chapitre décrit les principes théoriques de l'extraction des règles d'association, qui proviennent de trois champs disciplinaires différents, et répondent donc à des logiques différentes, en visant des buts également différents. Cela fait la richesse des possibilités de cette méthode, mais aussi sa complexité d'utilisation. Elle est présente depuis peu dans la boîte à outils des logiciels permettant de faire du Data Mining, comme SAS (Enterprise Miner), Weka, SPSS (Clémentine), etc., ce qui la rend d'utilisation courante sous sa forme basique. Nous décrivons en première section ce type d'utilisation courante des règles d'association par un exemple fictif situé dans le milieu qui l'a rendue célèbre, celui de la gestion d'un commerce, ici un petit supermarché de proximité. Cela nous permettra de pointer les trois aspects essentiels qui font de cette méthode un outil de gestion des données utilisable par un non spécialiste en fouille de données. Nous exposons à travers cet exemple les principales difficultés de l'utilisation d'un jeu de règles d'association et celles pour lesquelles nous avons cherché et parfois trouvé des solutions.

Puis nous détaillons dans la deuxième et troisième sections les deux formalismes les plus anciens à la base des règles d'association. Il correspondent à deux points de vue différents. Le premier, selon lequel le raisonnement à partir de règles doit suivre la logique mathématique, est à l'origine des implications de Guigues et Duquenne [105] et des treillis de concepts [235]. Selon le second, les règles sur les données représentent une approximation de la réalité, la qualité de cette approximation étant mesurée par des indices statistiques comme le support et la confiance et d'autres qui sont décrits dans [107]. Leur confrontation et l'examen de la possibilité de les concilier est en quatrième section.

Les deux sections suivantes abordent les fondements les plus récents des règles d'association qui suivent une logique informatique. La quatrième section fait un tour rapide des algorithmes actuels d'extraction de règles d'association. La cinquième compare les règles d'association au modèle informatique proche que sont les *dépendances fonctionnelles*. Ce modèle informatique est issu des bases de données et porte sur les contraintes imposées à la fois aux données et aux relations entre les données. La transposition de ce formalisme informatique a permis de développer des techniques de codage, d'élagage [245, 231, 51, 232], de navigation [25] dédiées aux règles d'association. A l'occasion de cette comparaison, nous évoquons les positions respectives de l'informaticien et de l'expert dans le domaine scientifique dont sont issues les données, qui doivent collaborer pour que l'extraction de règles d'association construite par le premier produise de la connaissance utile au second.

Nous concluons par une description très rapide de nos contributions qui ont pour ambition d'améliorer cette méthode dans certaines directions afin que les chercheurs en sciences humaines puissent l'utiliser pour la recherche des relations complexes entre les variables.

Sommaire

4.1 Exemple d'utilisation d'un jeu de règles d'association

Le gérant d'un petit commerce observe les achats de ses clients, et tire régulièrement un bilan de ses ventes afin de calculer son bénéfice. L'"analyse comptable" permet au gérant formé dans cette discipline de naviguer au milieu de tous ses comptes pour trouver comment faire prospérer son commerce. Mais il peut encore tirer de l'information supplémentaire des observations qu'il a faites sur les achats de ses clients. Il peut faire des statistiques jour après jour sur certains articles pour suivre l'évolution de leurs ventes, et même combiner plusieurs articles afin de savoir s'ils se vendent de façon concurrente (la lessive A peut être remplacée par la lessive B), complémentaires (le client achetant du fromage achète du pain) ou indépendante (l'achat d'un stylo n'est pas lié à celui d'une boite de conserve). De ses observations il déduit ainsi des règles. Et il les utilise

pour passer ses commandes (il commande la lessive qui lui rapporte le plus gros bénéfice), pour disposer les articles dans les rayons (il met le pain vers le fromage, et s'arrange pour que le passage de l'entrée à la caisse passe par les articles de papeterie ainsi que par les boîtes de conserves). Cette production de règles s'appuie sur l'examen des achats en caisse, mais aussi sur les remarques des clients, ou même sur leur façon de regarder les articles en rayon. C'est de l'artisanat tout à fait adapté à la taille du commerce.

Imaginons maintenant que le petit commerce est devenu un supermarché. On voit mal le gérant parcourir les rayons pour discuter avec les clients, ou examiner le contenu des caddies afin de trouver des règles. Il utilise les données enregistrées par toutes les caisses pour découvrir des règles. C'est de la "fouille de données". Et les règles qu'il obtient associant plusieurs articles s'appellent des "règles d'association" (nous les définissons rigoureusement dans la section 4.2 de ce chapitre). Comme le gérant du petit commerce, il utilise ces règles afin de réorganiser ses commandes, ses rayons, sa publicité. Mais à l'échelle d'un supermarché, cette réorganisation a un coût et le gérant n'a pas droit à l'erreur. Il faut qu'une augmentation de bénéfice découle de cette nouvelle gestion et permette au minimum de récupérer les frais. Il lui faut donc des règles de bonne qualité.

Nous décrivons maintenant ce que nous appelons "qualité" des règles.

4.1.1 Qualité individuelle d'une règle

Reprenons la règle "si achat de fromage alors achat de pain". Si le fait de l'utiliser (en mettant le pain près du fromage, ou en rajoutant au dessus du rayon de fromages de la publicité pour du pain) n'a pas modifié les ventes, cette règle n'est pas de bonne qualité. L'idéal serait de tester cette règle avant de l'appliquer en montant une petite expérimentation. Ce n'est pas simple, comme nous l'avons vu dans le chapitre 3 portant sur l'implication en sciences humaines, et cela a aussi un coût. Nous avons vu également dans ce chapitre que les autres techniques habituelles, telles que les statistiques ou l'analyse de données ne permettent pas de mesurer la valeur de chacune de ces règles. Nous devons donc l'établir autrement. Si de nombreux clients ont acheté simultanément du pain et du fromage, si parmi ceux qui ont acheté du fromage, il y en a un pourcentage important qui ont aussi acheté du pain, on peut supposer que la règle est de bonne qualité. Ces deux mesures de la règle s'appellent respectivement le "support" et la "confiance". Ce sont des indices de qualité d'une règle. Mais cela ne suffit pas. Si l'on découvre que le pourcentage de clients achetant du pain est supérieur à celui des clients achetant du pain parmi ceux qui ont acheté du fromage, il serait peut-être mieux d'écrire la règle "si pas d'achat de fromage alors achat de pain". Des réflexions de ce genre sont à l'origine des nombreux indices que nous décrivons dans la section 4.3 de ce chapitre.

Plutôt que d'aider le gérant de supermarché à faire un choix parmi tous ces indices, puis un choix de leurs valeurs selon des méthodes d'arbitrage plus ou moins complexes, décrites également dans ce chapitre de l'état de l'art, nous proposons de faire des simulations qui permettront de décider si une règle est ou non due au hasard. Dans le chapitre 5, nous montrons que des données provenant du hasard fournissent aussi des jeux de règles, et qu'il convient de retirer ces règles sans valeur du jeu final. Le problème est de définir un type de hasard qui prenne en compte les lois de probabilités que suivent les données. En effet celles-ci varient fortement selon les domaines dont sont issues les données (loi Zipf-like et matrices binaires creuses, essentiellement remplies de 0 pour les données du texte et du Web,). Nous n'avons pas trouvé de simulations de ce type, nous en construisons donc une dans le chapitre 6.

4.1.2 Qualité d'un groupe de règles

Le gérant va utiliser plusieurs règles. Il s'attend donc à ce qu'elles soient compatibles. Si, en suivant les règles, il doit mettre les boîtes de conserve au dessus des paquets de sucre, les paquets de sucre au dessus du chocolat et le chocolat au dessus des boîtes de conserve, il va vite s'apercevoir qu'il y a un problème dans les règles obtenues, et abandonner la réorganisation du rayon. S'il découvre d'autres règles contradictoires, il va finir par rejeter toutes les règles. Le jeu de règles doit donc être structuré de façon logique. Nous décrivons dans la section 4.2 de ce chapitre les "treillis de Galois" qui sont une structuration des données produisant des jeux de règles de ce type. Malheureusement, ils ne sont pas utilisables pour les données approximatives de notre gérant, comme nous l'expliquons dans la section 4.4 de ce chapitre sur l'état de l'art. Nous décrivons également dans ce chapitre (section 4.6) les dépendances fonctionnelles qui, avec les règles d'inférence associées, permettent une normalisation des bases de données en les rendant quasiment exemptes de toutes ces incohérences. Elles non plus ne s'accordent pas avec nos données approximatives. Et puis toutes nos exigences de logique ne sont pas assurées par ces modèles. Nous voulons une logique du "sens commun" utilisable pour faire des raisonnements courants utilisant certes la transitivité, mais également la négation. Nous aimerions également faire une représentation des données du gérant qui pourrait permettre la navigation, comme celle utilisée dans le logiciel Chic[76]. La visualisation des données de fait sous la forme de points (ici les articles) joints par des flèches (les règles d'association), comme nous le décrivons dans la section 4.3 de ce chapitre. Mais alors que dans le treillis de Galois la propriété AB obtenue par fusion des propriétés A et B ne se fait que sur les valeurs 1, ce qui assure les règles d'inférence, dans Chic, les propriétés ne sont pas fusionnées, mais des règles d'inférence sont utilisées après coup pour retirer les incohérences. On trouve également ce genre de nettoyage de la "redondance" dans les travaux de chercheurs en bases de données qui travaillent sur les "cubes OLAP" (On line Analytical Processing) ou des incohérences comme indiqué dans la section 8.6. C'est une correction de ce type que nous proposons. Notre contribution se situe essentiellement au niveau de l'exposé de ce que sont les liaisons complexes dans le chapitre 7, et dans la méthode de correction du chapitre 8 utilisant des méta-règles de type algorithmique accessibles à l'expert des données, comme le gérant du supermarché par exemple. En effet, comme ce sont des règles approximatives, il peut tolérer des incohérences dans les règles, mais il veut qu'on les lui signale plutôt que de les découvrir par hasard, et il apprécie qu'on lui laisse le choix du type de nettoyage qu'il fera. Comme il aime qu'on lui signale les liaisons gênantes, il apprécie qu'on lui garde les liaisons simples. Par exemple, par le codage inapproprié qui est fait en codant le poids de la tablette de chocolat achetée en trois propriétés "pas de chocolat", "un peu de chocolat", "beaucoup de chocolat", et pareillement pour les gâteaux, au lieu d'obtenir la règle simple "si achat de chocolat alors achat de gâteaux" on va obtenir quelques-unes des 9 règles "partielles" possibles comme "si pas d'achat de chocolat alors pas d'achat de gâteaux" "si achat d'un peu de chocolat alors achat de beaucoup de gâteaux" sans pouvoir remonter à la règle d'origine. Comme nous le décrivons dans le chapitre de l'état de l'art, ce problème n'arrive pas avec les dépendances fonctionnelles du fait de leur nature. Pour éviter ce problème d'émiettement nous proposons dans le chapitre 9 de faire un codage flou de ces deux articles permettant de définir la règle floue : "si achat de chocolat alors achat de gâteaux".

[76] http ://www.ardm.asso.fr/CHIC.html réalisé par Régis Gras [101] avec son équipe.

4.1.3 Rapidité de l'extraction des règles

La qualité du jeu de règles dépend également de la rapidité de son extraction. Si le gérant doit attendre un mois pour pouvoir appliquer les règles, elles ne seront peut-être plus valables. En effet, il y a des modes, des saisons, ce qui entraîne des changements dans les achats des clients. Nous exposons ces algorithmes dans la section 4.5 de ce chapitre. Pour qu'ils fonctionnent de façon optimale, ils doivent s'appuyer sur une représentation informatique adaptée des données. Nous exposons sommairement ces algorithmes et leur fonctionnement, notre but n'étant pas d'en écrire des concurrents mais de créer des règles d'association floues qui puissent s'extraire avec les plus courants de ces algorithmes moyennant de légères modifications.

4.2 Les règles d'association comme ensemble logique

Dans cette section, nous donnons la définition d'une *règle d'association* en réactualisant la définition de *l'implication* choisie par Guigues et Duquenne [105], grâce notamment à l'utilisation des termes tels que *motif* et *support* que nous définissons également dans cette partie. Nous en profitons pour indiquer les principales notations des règles d'association, et nous illustrons toutes ces notions par un petit exemple repris par de nombreux auteurs en fouille de données. Le cadre formel des règles d'association étant ainsi posé, nous exposons deux représentations condensées de l'ensemble des règles d'association qui sont les *bases de règles* et les *treillis de concepts*, le parallèle entre ces deux représentations étant illustré par l'exemple.

4.2.1 Des implications selon Guigues et Duquenne aux règles d'association

Guigues et Duquenne [105] se placent dans un *contexte* $(\mathcal{S}, \mathcal{T}, \mathcal{R})$, où $\ll \mathcal{S}$ peut être considéré comme un ensemble de sujets, $\mathcal{T}$ comme un ensemble de traits dichotomiques décrivant les sujets moyennant la relation $\mathcal{R}$. Pour tout couple (s,t) de $\mathcal{S} \times \mathcal{R}$, on peut interpréter la relation s$\mathcal{R}$t comme "le sujet s satisfait le trait t" et définir une notion d'*implication.* $\gg$

Définition 4.2.1. *Implication résultant d'un contexte ($\mathcal{S}$,$\mathcal{T}$,$\mathcal{R}$) selon Guigues et Duquenne.*
A et B étant 2 parties de $\mathcal{T}$, on dira que A implique B (on notera A→B) lorsqu'un sujet satisfaisant tous les traits de A satisfait également ceux de B. Cette implication sera appelée informative si B n'est pas inclus dans A.

La notion de règle *informative* permet d'éliminer les règles n'apportant aucune information sur les données, c'est-à-dire sur la relation entre les sujets et les propriétés. Par exemple, si on considère les ensembles de propriétés A={a, b, c} et B={b}, comme B est inclus dans A, l'implication A→B n'est pas *informative*. En effet, si on sait qu'un sujet satisfait les propriétés a, b et c, il est inutile d'examiner les données pour en déduire qu'il possède la propriété b. Bien que n'étant pas porteuses d'information, ces règles peuvent avoir leur utilité. Par exemple, la règle A→A, où A est un sous ensemble quelconque de $\mathcal{T}$ est toujours vraie, comme Guigues et Duquenne l'exposent à la page 11 de leur article [105]. Elle n'est bien sûr pas informative, mais bien pratique pour alimenter des inférences.

En réécrivant cette définition de la règle d'*implication informative* avec des termes plus courants en fouille de données, nous obtenons la définition d'une *règle d'association* suivante :

Le contexte $(\mathcal{S}, \mathcal{P}, \mathcal{R})$ est formé de deux ensembles finis non vides, l'ensemble $\mathcal{S}$ des N sujets, l'ensemble des propriétés booléennes $\mathcal{P}$ (dont les valeurs sont *Vrai, Faux* et qui correspondent aux traits), et de $\mathcal{R}$ la relation entre les deux. Les parties de $\mathcal{P}$ sont appelées *motifs*. Le *support*

d'un motif est le nombre de sujets vérifiant les propriétés qui le composent. Un motif aura 2 notations possibles, soit par une lettre majuscule, qui représente alors l'ensemble des propriétés, soit par une suite de lettres minuscules, représentant la liste des propriétés. Ainsi si on a A={b, c, d} et B={c, e}, la réunion de A et de B se notera selon les cas A∪B ou *bcde*. Si tous les sujets vérifiant les propriétés de A vérifient celles de B, on aura la *règle d' association exacte* A→B, qui se trouve être également, selon la définition de Guigues et Duquenne, la *règle d'implication informative* A→A∪B. Puis on définit une règle d'association approximative en tolérant que des sujets vérifient A sans vérifier B. Toutefois, pour rester cohérente avec le formalisme de Guigues et Duquenne et garder ainsi tout leur sens aux bases de règles, nous notons dans cette partie les règles d'association par A→A∪B au lieu de A→B, mais nous rajoutons à chaque fois que nous utilisons cette notation le symbole **GD** pour montrer au lecteur que ce n'est pas la notation courante, et nous acceptons les motifs de support nul.

Reprenons un exemple proposé par Stumme et al. [223] et repris dans d'autres articles en fouille de données qui nous permettra d'illustrer ensuite les notions de "bases de règles".

Exemple 1. *Exemple figurant dans [57, 64, 223].*
$\mathcal{S}$ est l'ensemble de sujets {s1,s2,s3,s4,s5}, $\mathcal{P}$ l'ensemble de propriétés {a,b,c,d,e}, et $\mathcal{R}$ la relation liant les deux, selon la matrice d'incidence figurant dans le tableau 4.1.

	a	b	c	d	e
s1	x		x	x	
s2		x	x		x
s3	x	x	x		x
s4		x			x
s5	x	x	x		x

TAB. 4.1 – Exemple de Stumme et al. [223]

.

On déduit de ce contexte les 20 motifs suivants, dont les supports sont donnés entre parenthèses, abce (2)[77]*, abc (2), abe (2), acd (1), ace (2), bce (3), ab (2), ac (3), ad (1), ae (2), bc (3), be (4), cd (1), ce (3), a (3), b(4), c (4), d (1), e (1), ∅ (5)*[78]*. Les règles d'association* **GD** *comportent à droite un motif, et à gauche un motif strictement inclus dans celui de droite*[79]*. Par exemple, du motif abce on tire ainsi autant de règles d'association* **GD** *qu'il y a de parties strictement incluses, soit 15 règles : abc → abce, abe → abce,..., ∅ → abce. Le jeu obtenu contient 79 règles, dont 19 règles exactes et 60 règles approximatives. Par exemple, la règle d'association* **GD** *ab → abc est exacte, car les supports des 2 motifs ab et abc sont égaux à 2 (les 2 sujets sont s3 et s5), et la règle b → ab est approximative, le support de b étant 4, et celui de ab seulement 2 car deux sujets vérifient b sans vérifier ab (ce sont les sujets s2 et s4).*

A partir de ce petit exemple, on voit que le jeu de règles d'association peut difficilement apporter de la connaissance de qualité, étant de taille importante, et non structuré. Nous allons voir comment il peut être réduit à une "base de règles" sans perte d'information.

[77] Le support est 2 car les sujets ayant en commun les propriétés a, b, c et e, sont au nombre de 2, ce sont s3 et s5.

[78] Le support est 5 car les sujets qui ont en commun les propriétés de l'ensemble vide sont au nombre de 5, ce sont tous les sujets de $\mathcal{S}$.

[79] Par l'ajout de **GD**, nous voulons indiquer que cette notation provisoire des règles d'association reprenant en partie droite tous les éléments de la partie gauche est celle utilisée dans l'article fondateur de Guigues et Duquenne [105].

4.2.2 Les bases de règles

Un ensemble de règles étant donné, la base est une partie de cet ensemble, à partir de laquelle on peut régénérer l'ensemble total de règles grâce à des règles d'inférence. La règle d'inférence la plus répandue dans le "sens commun" est la transitivité qui à partir des 2 règles suivantes tirées de la connaissance du climat de notre région :

"Si c'est l'hiver, alors il fait froid"

"S'il fait froid, alors les arbres perdent leurs feuilles"

permet de déduire la règle :

"Si c'est l'hiver alors les arbres perdent leurs feuilles"

Les bases de règles les plus utilisées en fouille de données sont la base de Guigues et Duquenne [105] et celle de Luxenburger [170]. Les deux s'appuient sur les motifs *fermés*, c'est-à-dire les motifs tels que l'ajout d'une propriété dans le motif diminue son support. Pour l'exemple figurant dans le tableau 4.1, *bce* est un fermé, de support 3 (les sujets s2, s3 et s4 vérifient les 3 propriétés) : il est fermé car si on lui ajoute la propriété a, on obtient le motif *abce* de support 2 (le sujet s2 ne vérifie pas a), et si on ajoute la propriété d, le motif *bcde* a un support nul. Au contraire, d n'est pas un motif fermé, le seul sujet le vérifiant vérifie également a et c. Donc l'ajout de a et c ne modifie pas son support, le motif *acd*, de support 1 est appelé la *fermeture du motif d*. Nous obtenons ainsi les 7 fermés, dont le support est indiqué entre parenthèses : *abce* (2), *acd* (1), *bce* (3), *ac* (3), *be* (4), c (4), $\emptyset$ (5), auquel nous ajoutons un huitième qui est le motif fermé *abcde* (0).

Les règles de la *base de Guigues et Duquenne* comportent en partie gauche un motif non fermé (*pseudo-fermé*), et en partie droite un motif le contenant de même fermeture. Mais toutes les règles vérifiant cette propriété ne sont pas mises dans la base. Elles doivent vérifier des propriétés supplémentaires [105], ceci afin d'obtenir une base de taille minimale. Pour notre exemple, cela donne les règles GD1[80] : $b \rightarrow be$, GD2 : $e \rightarrow be$, GD3 : $a \rightarrow ac$ et GD4 : $d \rightarrow ad$.

Définition 4.2.2. *Règles d'inférence associées à la base de Guigues et Duquenne.*
Soient A, B, C et D des parties de $\mathcal{P}$ telles que les règles **GD**[81] *A→B,* **GD** *B→C,* **GD** *C→D sont des implications informatives, on définit les 3 règles d'inférence suivantes*

$mrGD_1$: ((A→B) et (C→D)) ⊢ (A∪C→B∪D)

$mrGD'_1$: (A→B) ⊢ (A∪C→B∪C)

$mrGD_2$: ((A→B) et (B→C)) ⊢ (A→C)

Ces 3 règles d'inférence sont acceptables pour la logique du "sens commun", on reconnaît notamment la transitivité dans $mrGD_2$. Avec cette base et ces 3 règles d'inférence[82], on peut reconstruire l'ensemble de toutes les règles exactes. Pour notre exemple, les 4 règles de la base permettent de reconstruire les 19 règles exactes du jeu de règles. Ainsi, en faisant agir la règle d'inférence $mrGD'_1$ avec A={b}, B={b, e}, C={a}, elle produit la règle **GD** $ab \rightarrow abe$ à partir

[80] On notera GDi la règle n°i de la base de Guigues et Duquenne, et de la même façon Li pour la base de Luxenburger.

[81] Rappel : on note par **GD** une règle A→B selon la notation de Guigues et Duquenne, c'est-à-dire pour laquelle B=A∪C, au lieu de A→C qui est sa notation plus habituelle.

[82] On a ajouté la règle d'inférence $mrGD'_1$ aux deux règles d'inférence de la définition intiale d'implication [57] alors qu'elle se déduisait de la règle $mrGD_1$ en remplaçant la règle **GD** (C→D) par la règle **GD** (C→C) afin de l'adapter aux seules implications informatives.

de la règle GD1 ($b \rightarrow bc$)[83].

Les règles de la *base de Luxenburger* comportent en partie gauche un motif fermé, et en partie droite un motif fermé obtenu en ajoutant des propriétés à celui de gauche. Ce qui donne dans notre exemple les règles suivantes pour lesquelles on a indiqué entre parenthèses les supports respectifs des parties gauche et droite L1 : $\emptyset \rightarrow c$ (5 ;4), L2 : $\emptyset \rightarrow be$ (5 ;4), L3 : $c \rightarrow ac$ (4 ;3), L4 : $c \rightarrow bce$ (4 ;3), L5 : $be \rightarrow bce$ (4 ;3), L6 : $ac \rightarrow acd$ (3 ;1) L7 : $ac \rightarrow abce$ (3 ;2), L8 : $bce \rightarrow abce$ (3 ;2). Comme pour la base précédente, les règles construites ainsi ne figurent pas toutes dans cette base [170]. Les règles d'inférence permettant de générer à partir de cette base l'ensemble des règles d'association sont la transitivité (mrL_1) et la règle d'inférence mrL_2 qui, à la règle **GD** A→B de la base (donc pour laquelle A et B sont deux motifs fermés), et aux motifs C et D de fermetures respectives A et B et tels que C soit inclus dans D, associe la règle **GD** C→D. Pour notre exemple, de la règle L5 ($be \rightarrow bce$), en prenant b, de fermeture be et bc de fermeture bce, on peut déduire la règle **GD** $b \rightarrow bc$. On arrive ainsi avec ces 2 règles d'inférence à obtenir les 60 règles approximatives à partir des 8 règles de la base. Pour générer le jeu complet de règles d'association, il faut donc ces deux bases et leurs règles d'inférence associées, qui ne relèvent pas toutes de la logique du sens commun.

Pour conclure sur ces bases de règles, avec celle de Guigues et Duquenne on peut retrouver toutes les règles en utilisant quelques règles d'inférence de "bon sens". On peut aboutir à des règles qui ne sont vérifiées par aucun sujet, mais on interdit celles contredites par au moins un sujet. Avec la base de Luxemburger, en utilisant des règles d'inférence beaucoup moins naturelles, nécessitant la connaissance des ensembles fréquents fermés, on obtient toutes les règles approximatives mais seulement ces règles. Il existe d'autres sortes de bases de règles, qu'on peut trouver notamment dans Cristofor et Simovici [57], qui en proposent une permettant de générer à la fois les règles exactes et les règles approximatives, mais avec des règles d'inférence d'interprétation difficile selon la logique du "sens commun". En glissant des *implications informatives* aux *règles d'association*, on est passé d'une base manipulable par des utilisateurs doués de bon sens à une base manipulable par programmation.

Nous verrons dans le chapitre portant sur le "Nettoyage par des méta-règles des incohérences dues aux liaisons complexes" notre proposition visant à rendre à l'utilisateur un jeu de règles réduit respectant autant que possible la logique du "sens commun". Mais pour l'instant nous allons continuer d'examiner les formalismes liés aux règles d'association. Les bases que nous venons de voir, que ce soient la première orientée utilisateur ou les suivantes orientées programmation, s'appuient sur une structure particulière des données, appelée "treillis de Galois". Nous en faisons une description rapide car les "règles d'association floues" que nous proposons plus loin, permettant d'éviter la perte de liaisons qui se produit lors des recodages classiques de propriétés numériques en propriétés booléennes, ont été construites pour conserver le plus possible la structure de treillis, afin de faciliter leur programmation.

4.2.3 Les treillis de Galois

Sous ce nom, on désigne une structure double de treillis issue du tableau Sujets×Propriétés grâce à une "connexion de Galois". Nous allons montrer ci-dessous ce que c'est, et comment elle

[83] Notons toutefois que la règle **GD** $abcd \rightarrow abcde$, obtenue également à partir de GD1 avec $mrGD'_1$ n'est pas valide selon une définition de la règle d'association qui imposerait un seuil de support, le support du motif $abcde$ étant nul.

exprime d'une façon condensée la relation liant les Sujets et les Propriétés en donnant seulement quelques définitions et en renvoyant les lecteurs intéressés par les théories algébriques sous-jacentes aux ouvrages cités dans la bibliographie [23], [77], [61].

Reprenons la définition que donne G. Birkhoff en page 56 de son ouvrage intitulé "Lattice Theory" [23] :

Définition 4.2.3. *Connexion de Galois.*
Soit P et Q deux ensembles partiellement ordonnés, f une application de P vers Q, et g une application de Q vers P tels que

1. *si* $x \succeq x_1$ *dans P, alors* $f(x) \preceq f(x_1)$ *dans Q*
2. *si* $y \succeq y_1$ *dans Q, alors* $g(y) \preceq g(y_1)$ *dans P*
3. $x \preceq g(f(x))$ *pour tout x de P,* $y \preceq f(g(y))$ *pour tout y de Q*

On dit alors que f et g définissent une connexion de Galois entre P et Q.

Les "treillis de Galois" sont obtenus par Barbut et Monjardet [11] à partir d'un contexte ($\mathcal{S}$,$\mathcal{P}$,$\mathcal{R}$) en prenant pour ensembles P et Q de la définition précédente, les ensembles respectifs de parties de l'ensemble des propriétés $\mathcal{P}$ et de l'ensemble des sujets $\mathcal{S}$, pour relation d'ordre l'inclusion ensembliste, pour f l'application qui à tout ensemble de propriétés associe l'ensemble de tous les sujets qui les vérifient toutes (c'est-à-dire qui sont en relation par $\mathcal{R}$), et pour g celle qui à tout ensemble de sujets associe l'ensemble de toutes les propriétés qu'ils vérifient tous. Les *fermés* étant les parties x de $\mathcal{P}$ telles que g(f(x))=x, et celles y de $\mathcal{S}$ telles que f(g(y))=y, ils établissent dans leur ouvrage [11] que l'ensemble des parties fermées de $\mathcal{P}$ muni de la relation d'ordre $\subseteq$ et l'ensemble des parties fermées de $\mathcal{S}$ muni de la relation d'ordre $\supseteq$ sont deux treillis isomorphes[84]. Ce sont ces deux treillis qu'ils appellent "treillis de Galois". Wille [235] définit de la même façon ses "treillis de concepts", où un concept est formé d'une partie fermée de propriétés et de la partie fermée de sujets correspondante.

Donnons un exemple de concept afin d'illustrer ces définitions. Si on considère l'ensemble A formé de la seule propriété a, A={a}, f(A) est l'ensemble de tous les sujets vérifiant a, donc f(A)={s1, s3, s5}. Appelons B cet ensemble, g(B) est l'ensemble de toutes les propriétés vérifiées par s1, s3 et s5. On obtient g(B)={a, c}. On a donc bien {a}$\subseteq$g(f(a))={a, c} comme indiqué dans le troisième alinéa de la définition de la connexion de Galois, mais comme on n'a pas {a}=g(f(a)), {a} n'est pas un fermé [85]. L'ensemble {a} ne figure donc pas dans le treillis de Galois, alors que l'ensemble {a, c} y figure, représenté le plus souvent par l'écriture ac. Comme il est associé à l'ensemble de sujets B={s1, s3, s5}, qui est forcément fermé[86], le motif ac et l'ensemble de sujets {s1, s3, s5} forment un concept, et quand on représente le treillis des sous-ensembles fermés de propriétés par un diagramme dans lequel les noeuds sont les motifs et les flèches indiquent l'inclusion, si on rajoute dans chaque noeud l'ensemble de sujets associé au motif, le diagramme représente également le treillis des sous-ensembles fermés de sujets. Ainsi un même diagramme

[84] Un treillis est un ensemble partiellement ordonné dans lequel tout couple d'éléments admet à la fois une borne inférieure et une borne supérieure. Par exemple, l'ensemble des parties d'un ensemble muni de l'inclusion est un treillis car l'inclusion est une relation d'ordre partielle, l'intersection de deux parties étant la plus grande partie incluse dans chacune, et la réunion étant la plus petite partie les contenant toutes deux [77].

[85] Nous avions défini auparavant un fermé par le fait qu'on ne pouvait pas lui rajouter de propriété sans diminuer son support, c'est-à-dire le nombre de sujets vérifiant ses propriétés. Ce n'est pas le cas de {a} car on peut lui adjoindre la propriété c sans changer son support qui est 3. Il est clair que ces deux façons de définir des fermés sont équivalentes.

[86] Puisque B est l'image par f de l'ensemble de propriétés A, B=f(A), donc g(B)=g(f(A)), et g(f(A)) contient A d'après l'alinéa 3 de la définition, donc g(B)$\supseteq$A, et d'après l'alinéa 1, f(g(B))$\subseteq$f(A)=B. Comme f(g(B))$\supseteq$B d'après l'alinéa 3, on a donc f(g(B))=B.

représente simultanément les deux treillis de fermés liés par la connexion de Galois. Le diagramme le plus pratique est le diagramme de Hasse, qui représente de façon verticale les concepts, les plus "petits" dans le sens de l'ordre étant en bas, les plus "grands" au dessus, chaque inclusion entre un élément x et celui immédiatement plus grand y, donc mis plus haut que x, étant indiquée par un trait entre x et y. On peut voir dans la figure 4.1 le diagramme de Hasse de l'exemple.

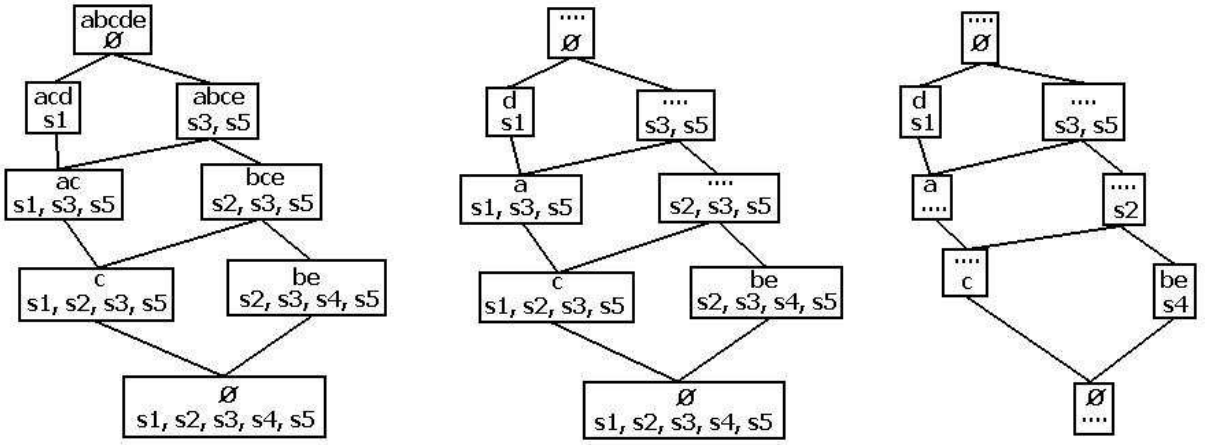


FIG. 4.1 – Treillis des concepts de l'exemple du tableau 4.1

On peut le construire "à la main" pour mieux comprendre les définitions. On place tout en bas le motif vide, formé d'aucune propriété. Tous les sujets le vérifient, on lui associe donc l'ensemble de tous les sujets, et son support est 5. On est sûr qu'il est fermé, puisque si lui ajoute une propriété, son support diminue. En effet toutes les propriétés ont un support inférieur à 5. Puis on lui adjoint la propriété a, ce qui donne le motif a (car $\emptyset \cup \{a\} = \{a\}$), associé aux sujets s1, s3 et s5. Toutefois, comme on vient de le voir, ce motif n'est pas fermé, c'est le motif ac qui est fermé, on écrit donc le concept (ac,{s1, s3, s5}) au dessus du concept ($\emptyset$,{s1, s2, s3, s4, s5}). Entre les deux, on a le motif c, qui est vérifié par tous les sujets sauf s4. Il est fermé car on ne peut pas lui adjoindre d'autres motifs sans diminuer son support de 4. On écrit entre ces deux concepts le concept (c, {s1, s2, s3, s5}) qu'on joint à celui du dessous par un trait et à celui du dessus par un autre trait. On ne rajoute pas de trait entre celui du bas et celui du haut, car comme on pourrait obtenir ce lien en appliquant la transitivité de la relation sur les deux liens précédents, il est superflu. En suivant ces liens du haut vers le bas, on trouve que $\emptyset \subseteq \{c\} \subseteq \{a, c\}$ et que {s1, s2, s3, s4, s5}$\supseteq${s1, s2, s3, s5}$\supseteq${s1, s3, s5}. Puis on repart du bas, et on ajoute la propriété b à l'ensemble vide. On obtient le motif b. Il est vérifié par tous les sujets sauf s1. Mais il n'est pas le seul. Il n'est donc pas fermé. C'est le motif be qui est fermé et qui produit le concept (be,{s2, s3, s4, s5}). On le joint par un trait à celui du bas, puis on considère la propriété d. Le motif d, vérifié par le sujet s1 seul, n'est pas fermé, on peut lui adjoindre les propriétés a et c, ce qui forme le motif acd, qui est fermé et le concept correspondant est le concept (acd, {s1}). Il se place au dessus du concept (ac,{s1, s3, s5}) et on le joint par un trait à ce dernier. Ensuite on peut voir sur le diagramme que les concepts (c, {s1, s2, s3, s5}) et (be,{s2, s3, s4, s5}) ont en commun les sujets s2, s3 et s5, ce qui donne le concept (bce,{s2, s3, s5}), qui représente leur borne supérieure, obtenu en faisant la réunion des ensembles de propriétés et l'intersection des ensembles de sujets, et on l'écrit donc au dessus de ces deux concepts puis on le joint à eux par deux traits. De la même façon, ce concept et le concept (ac,{s1, s3, s5}) produisent le concept ($abce$,{s3, s5}) qu'on écrit au dessus et qu'on joint à eux par deux traits, et on termine par le

concept (*abcde*,∅)[87]. On a ainsi obtenu un treillis avec 8 concepts alors qu'il y a $2^5 = 32$ parties possibles de l'ensemble $\mathcal{P}$, et de même pour $\mathcal{S}$. Ce qui explique que cette structure de treillis contenant l'information du tableau de façon structurée et condensée ait été privilégiée pour y adapter des algorithmes performants de construction de motifs.

Une écriture allégée de ce diagramme permet encore de simplifier le stockage en mémoire du treillis des concepts [94]. Pour alléger les motifs, on part du bas, et en suivant les traits on remonte en effaçant les propriétés qui ont déjà été rencontrées. Ainsi, dans le concept (*ac*,{s1, s3, s5}), on peut supprimer *c* qui apparaît dans un concept inférieur. A la fin de cette simplification, chaque propriété est écrite une seule fois dans le concept situé le plus bas parmi ceux auxquels elle appartient. C'est sur ce diagramme simplifié que se lisent la base de Guigues et Duquenne et celle de Luxemburger en suivant le long des traits du diagramme comme indiqué dans [222], l'une en descendant, et l'autre en montant. On peut encore alléger en effaçant les sujets de la même façon en descendant cette fois dans le diagramme. On peut voir dans le graphique de la figure 4.1 le diagramme de départ, et les versions allégées d'abord par effacement des propriétés, puis des sujets.

4.3 Les indices de qualité des règles

A partir de maintenant on reprend la notation la plus courante d'une règle en fouille de données, en ne mettant en partie droite que les propriétés ne figurant pas en partie gauche, et en n'acceptant pas de construire une règle à partir d'un motif de support nul.

Dans la partie précédente, on s'intéressait principalement à l'ensemble des règles d'association extrait des données. On acceptait de garder certaines règles "inintéressantes" (ayant par exemple le motif vide en partie gauche) afin de pouvoir construire des raisonnements en combinant entre elles les règles d'association, que ce soit par des règles d'inférence ou en parcourant le diagramme du treillis des concepts. Dans cette partie, nous allons voir comment la qualité d'une règle d'association prise isolément est mesurée par des indices. Nous allons d'abord nous focaliser sur la sémantique de cette qualité et montrer comment on en est arrivé à un grand nombre d'indices, puis nous exposerons leur mode de calcul, et nous terminerons par une synthèse.

4.3.1 Le sens général d'un indice de qualité

Les indices de qualité des règles servent à mesurer la qualité individuelle d'une règle A→B selon différents aspects orientés utilisateur. Il y en a actuellement une cinquantaine. Les plus utilisés sont le support et la confiance. Nous les étudions de façon détaillée, puis nous en examinons d'autres, et nous dressons un inventaire de leurs caractéristiques.

Le support

Le *support* [113] est le nombre de sujets qui vérifient la règle A→B, c'est-à-dire toutes les propriétés de A et de B. On exige en général qu'il soit positif, car ce qui prime ici n'est pas la logique formelle comme dans le paragraphe précédent, mais la valeur d'usage : une bonne règle n'est pas une règle pour laquelle il n'y a pas de contre-exemple, mais une règle exprimant des

[87] Notons que ce concept est indispensable pour qu'on obtienne bien un treillis, alors que les motifs de support nul sont rarement gardés. Le motif vide doit être conservé aussi si aucun sujet ne vérifie simultanément toutes les propriétés.

cooccurrences de propriétés constatées sur des sujets. En général, on accepte un degré d'imprécision plus ou moins élevé selon le domaine d'où sont issues les données. Par exemple, la saisie peut être entachée d'erreurs de frappe, mais la mesure peut aussi être approximative, notamment quand elle provient de questionnaires d'opinion. Ce qui fait qu'on préfère avoir un support supérieur à un seuil donné afin que les règles extraites ne soient pas des conséquences des seules erreurs, mais expriment une relation tangible. Plus le seuil de support est élevé, moins on obtient de règles, et plus les règles obtenues sont solides. Elles sont d'ailleurs parfois tellement triviales qu'elles n'apportent aucune connaissance nouvelle. En effet, si on a par exemple N sujets et deux propriétés a et b qui sont vérifiées par tous les sujets sauf 1, on obtient un motif ab dont le support est N-1 si le sujet qui ne vérifie pas a ne vérifie pas non plus b, ou N-2 dans le cas contraire, ainsi les règles $a \rightarrow b$ et $b \rightarrow a$ ont leur support élevé ($\geq$N-2) de façon "mécanique". Si un certain nombre de propriétés sont dans ce cas, augmenter le seuil de support pour diminuer le nombre de règles produit comme effet supplémentaire que ces règles inintéressantes deviennent majoritaires. On ne peut donc pas se contenter du support pour mesurer la qualité d'une règle. Certains auteurs appellent support ce que nous appelons fréquence, qui est le rapport du support au nombre total de sujets. Cette fréquence est plus utilisée que le support dès que le nombre de sujets est important, ou tout simplement si on veut comparer les règles extraites à celles établies par d'autres chercheurs sur des données de même type, mais ne comportant pas nécessairement le même nombre de sujets.

La confiance

La *confiance* [113] d'une règle A→B est la proportion de sujets qui vérifient les propriétés de B parmi ceux qui vérifient celles de A, donc le quotient du support des deux motifs correspondants. On a vu que les règles de confiance 1 avaient un statut spécial, ce sont les règles exactes, qui ne sont contredites par aucun sujet. A l'autre extrême, la confiance peut être proche de zéro sans toutefois atteindre 0, puisque nous avons décidé de ne construire que des règles vérifiées par au moins un sujet. A supports de A et de B constants, la règle A→B est de meilleure qualité quand sa confiance est plus grande, car en augmentant ce coefficient, on augmente le nombre de sujets vérifiant les propriétés de B parmi ceux qui vérifient celles de A, tout en diminuant le nombre de ceux qui vérifient B sans vérifier A. Cela fait apparaître une influence grandissante de A sur B. On extrait très souvent toutes les règles dont le support et la confiance dépassent des seuils fixés par l'utilisateur. Mais on n'est plus dans le cas de règles A→B où A et B ont des supports constants, ce qui complique la comparaison. Et choisir parmi deux règles de même support celle de plus grande confiance peut conduire à garder la plus "douteuse" des deux comme l'indique l'exemple suivant : supposons que les propriétés a, b, c et d sont vérifiées respectivement par 20, 24, 25 et 16 parmi les 30 sujets de l'ensemble $\mathcal{S}$, que la règle $a \rightarrow b$ est vérifiée par 15 sujets, ce qui fait une confiance de 15/20=0,75 , et la règle $c \rightarrow d$ est vérifiée également par 15 sujets, ce qui fait une confiance de 15/25=0,60. Les deux règles ayant même support, on aurait tendance à préférer la règle $a \rightarrow b$ à la règle $c \rightarrow d$ du fait de sa plus grande confiance. Mais en regardant de plus près, on voit que la propriété b est vérifiée par 24/30=80% des sujets de $\mathcal{S}$. Elle est donc proportionnellement plus rare chez les sujets vérifiant a (c'était 75%) que chez les autres. Alors que la propriété d est vérifiée par 16/30=53% des sujets de $\mathcal{S}$, donc proportionnellement plus fréquente chez les sujets vérifiant c (c'était 60%) que chez les autres. Du coup, la règle $a \rightarrow b$ paraît moins être une relation de "type causal" que la règle $c \rightarrow d$. C'est avec un exemple de ce genre que J. Han [113] montre l'insuffisance du support et de la confiance pour assurer la qualité d'une règle.

Les autres indices

Ce qui invite à créer un nouvel indice qui prend cette remarque en compte, comme la *différence* entre la confiance de la règle A→B et la proportion de sujets vérifiant les propriétés de B parmi tous les sujets, qui est négative dans le premier cas (0,75-0,80=-0,05) et positive dans le second (0,60-0,50=0,10). La règle est considérée d'autant meilleure qu'il est plus élevé.

Nous avons montré le sens du support, et de la confiance. Puis que ces deux indices se révèlent insuffisants pour exprimer toutes les facettes de la qualité d'une règle. La différence corrige une de ces insuffisances, mais il y en d'autres qu'elle ne corrige pas. De nombreux indices ont été créés pour cela, et nous renvoyons le lecteur intéressé par leur sens particulier à [52, 152, 142].

Signalons toutefois l'indice d'"implication statistique" qui provient d'une autre démarche initiée par R. Gras [99] bien avant l'avènement de la fouille de données. Ce travail s'inscrivait alors dans le domaine de la didactique et avait pour but de trouver des implications entre apprentissages d'atomes de connaissances chez des élèves en partant de leurs résultats à des contrôles de connaissances. Une valeur élevée de cet indice (ou de ses diverses améliorations [102, 106]) correspond à un degré élevé de certitude qu'elle n'est pas due à une configuration particulière des effectifs, comme comme dans le dernier exemple cité dans le paragraphe précédent. Il a créé cet indice à partir de l'écart entre la configuration d'effectifs observée et celle arrivant par hasard selon des modèles statistiques développés avec l'aide d'I.C. Lerman [166] (lois binomiales, de Poisson, normales indépendance). Malgré cette approche différente, cet indice peut être utilisé comme les autres pour classer les règles d'association de la meilleure à la moins bonne.

Les indices de qualité de la règle A→B que nous venons de décrire se calculent à partir de quatre effectifs trouvés dans la base de données, auxquels s'ajoutent éventuellement des éléments fournis par l'utilisateur tels par exemple que la loi de Poisson pour l'indice d'implication statistique, ou des pondérations différentes pour les propriétés de la partie gauche de la règle comme celles proposées par A. Freitas [86][88]. Nous indiquons maintenant comment se font les calculs.

4.3.2 Le calcul d'un indice de qualité

Supposons que nous avons 2 propriétés A et B, et N sujets numérotés de 1 à N qui vérifient ou non chacune de ces 2 propriétés. Les données peuvent être représentées par le tableau 4.2 dans lequel on a une matrice booléenne de N lignes et 2 colonnes, c'est-à-dire contenant uniquement les valeurs 0 et 1. Par exemple, si A est le sexe masculin, B est la réussite à l'examen, et les sujets sont des étudiants, on voit dans le tableau 1 que l'étudiant s1 est une fille (A=0) qui a échoué (B=0), que l'étudiant s2 est un garçon (A=1) qui a échoué, etc...

Sujet	A	B
s1	0	0
s2	1	0
s3	1	1
s4	1	0
s5	0	1
...	...	...
sN	0	1

TAB. 4.2 – Matrice booléenne de données

[88] Ses contributions sont proposées pour des règles de classement (la propriété de droite est la même pour toutes les règles, c'est la propriété de classement, seule sa valeur change) mais peuvent être transposées sans problèmes aux règles d'association.

On ne s'intéresse pas aux sujets en eux-mêmes, mais au fait qu'ils vérifient ou non les propriétés. Aussi le sujet s2 et le sujet s4 sont-ils considérés comme identiques car ils ont même valeur pour les propriétés A=1 et B=0. Cette démarche est de type statistique dans la mesure ou dans l'ensemble Sujets×Propriétés, les sujets ne nous intéressent que par les propriétés qu'ils vérifient, et sont considérés comme étant des représentants d'une population infinie dont on a tiré un échantillon fini S[89]. Cela permet de synthétiser cette information en un tableau de contingence, comme indiqué dans la table 4.3. Ces deux sujets contribuent à la valeur c du tableau de contingence.

AxB	B=0	B=1	total
A=0	a	b	a+b
A=1	c	d	c+d
total	a+c	b+d	N=a+b+c+d

TAB. 4.3 – Tableau de contingence des 2 variables observées sur N sujets

Dans ce tableau de contingence figurent les effectifs des 4 cas possibles de valeurs du couple de variables (A,B), auxquels on a rajouté une colonne et une ligne de totaux, qui sont appelées les marges du tableau. On notera les modalités 0 et 1 de la variable A par A0 et A1, et de même pour B. On se limitera aux tableaux de contingence n'ayant aucune marge nulle, c'est-à-dire que les variables A et B ont réellement deux modalités chacune. Voici les formules de quelques indices de qualité de la règle A0→B0 :

- le *support* a, qui est le nombre d'objets vérifiant simultanément A0 et B0, et la fréquence $\frac{a}{N}$ est le pourcentage correspondant
- la *confiance* (observée), $\frac{a}{a+b}$ est la proportion des objets vérifiant B0 parmi ceux qui vérifient A0. On appelle confiance attendue la proportion d'objets vérifiant B0 parmi tous les objets : $\frac{a+c}{N}$
- la *différence* (entre la confiance observée et la confiance attendue) : $\frac{a}{a+b} - \frac{a+c}{N} = \frac{ad-bc}{N(a+b)}$
- l'*intérêt* $\frac{aN}{(a+b)(a+c)}$ qui est le rapport entre la proportion d'objets vérifiant A0 et B0 par rapport à ceux vérifiant A0 et la proportion d'objets vérifiant B0 et également le rapport entre la proportion d'objets vérifiant A0 et B0 par rapport à ceux vérifiant B0 et la proportion d'objets vérifiant A0
- la *nouveauté* : $\frac{a}{N} - \frac{(a+b)(a+c)}{N^2} = \frac{ad-bc}{N^2}$, qui est la différence entre la proportion observée d'objets vérifiant A0 et B0 et la proportion attendue en cas d'indépendance entre A et B.
- la *satisfaction* : $1 - \frac{bN}{(a+b)(b+d)} = \frac{ad-bc}{(a+b)(b+d)}$, qui est le complément à 1 du rapport entre la proportion d'objets vérifiant A0 et B1 par rapport à ceux vérifiant A0 et la proportion d'objets vérifiant B1
- l'*étonnement*, $\frac{a-b}{a+c}$, qui est la différence entre le nombre d'objets vérifiant A0 et B0 et celui vérifiant A0 et B1, relativement au nombre d'objets vérifiant B0
- la *conviction*, $\frac{(a+b)(b+d)}{bN}$ (si b=0, cet indice n'est pas défini), qui est l'inverse du rapport entre la proportion d'objets vérifiant A0 et B1 par rapport à ceux vérifiant A0 et la proportion d'objets vérifiant B1
- l'*intensité d'implication*, $1 - Proba(X \leq b)$ où X est une variable aléatoire qui suit la loi de Poisson de paramètre $\lambda = \frac{(a+b)(b+d)}{N}$ qui donne le complément à 1 de la probabilité que

[89] Cette approche n'est pas du tout celle du premier paragraphe traitant des règles d'association comme ensemble logique, où la dualité complète entre les ensembles de sujets et de propriétés permet de définir des règles d'association entre sous-ensembles de sujets comme on l'a fait entre sous-ensembles de propriétés.

le nombre observé d'objets de A0B1 soit "si petit" par rapport au nombre attendu en cas d'indépendance [102], la loi de Poisson étant obtenue en définissant le mode statistique de tirage des objets. On notera f(b,λ) cette intensité.

Comme nous l'avons dit précédemment, ces formules ont toutes en commun une utilisation des données réduite aux 4 effectifs a, b, c et d, de total N. Nous n'avons donné qu'une partie des formules des indices de ce type et nous renvoyons le lecteur désirant en découvrir d'autres au didacticiel de Fabrice Guillet [107]. Les qualités numériques de ces indices sont diverses. Une proposition de normalisation a été faite par Piatetsky-Shapiro [195] qui donne ainsi les trois premiers critères de qualité d'une mesure m de la règle A→B, auxquels Major et Mangano [171] ont ajouté un quatrième critère :

1. m=0 si A et B statistiquement indépendants
2. m est fonction croissante du support de la règle à autres paramètres fixés
3. m est fonction décroissante du support de A ou de B à autres paramètres fixés
4. m est fonction croissante du support de A quand la confiance est fixée supérieure à la fréquence de B

D'autres critères ont été proposés par la suite, comme celui de Kamber et Shingal [138], mais sans avoir le même impact.

4.3.3 Synthèse sur les indices de qualité

On vient de voir que parmi les indices de qualité des règles, un grand nombre exprime par une formule fonction de quatre effectifs une facette particulière de la qualité qui n'est pas prise en compte par le support et/ou la confiance. L'utilisateur se trouve alors confronté à un dilemme : lequel choisir ? Nous allons d'abord examiner le problème posé par leur utilisation, puis voir d'autres façons plus globales d'aborder la qualité d'une règle.

Les difficultés d'utilisation des indices

Pour choisir un indice adapté à ses besoins, l'utilisateur peut s'appuyer sur la sémantique de ces indices [148]. S'il en trouve un parfaitement adapté, il n'a plus qu'à déterminer un seuil afin de ne garder que les règles dont la valeur pour cet indice dépasse le seuil, ce qui peut se faire par tâtonnement afin d'avoir un jeu de règles de la taille attendue. S'il en prend deux, c'est déjà plus difficile. En effet, les meilleures règles selon un indice ne sont pas nécessairement les meilleures selon l'autre. Par exemple, si on prend le support et la confiance, on a plusieurs possibilités de choix des deux seuils pour un jeu de règles d'une taille donnée, et ces choix ne vont pas donner les mêmes jeux de règles. Et quand l'utilisateur choisit plus d'indices, le problème devient rapidement complexe. Il peut alors s'aider, des techniques d'analyse multi-critères, ou d'autres méthodes, comme le propose P. Lenca [162].

Des utilisations plus globales

Notons que certains indices ne sont pas utilisés seuls, mais associés à des règles d'inférence, comme la transitivité et la contraposition[90]. Par exemple R. Gras [100] et les personnes qui ont

[90]Rappelons que la transposition est une règle d'inférence permettant de déduire de la règle A→B la règle nonB→nonA. Cette règle fait partie des règles de la logique du "sens commun" utilisant la négation, au même titre que le *raisonnement par l'absurde* qui consiste à prouver la règle A→B en établissant l'impossibilité d'avoir simultanément A et nonB.

travaillé à ses cotés ne se sont pas contentés de définir un indice d'*implication statistique* sur une règle. Dans le logiciel Chic[91], les règles sont représentées en réseau de telle façon qu'elles respectent la transitivité. Et un travail a été fait pour que les effets de la négation soient le plus possible pris en compte. Ainsi, bien que l'indice d'implication, ou ses extensions comme l'*indice d'implication ordinale* de S. Guillaume [106], soient au départ construits sur une règle avec seulement une propriété à gauche et une à droite, un examen approfondi des diverses possibilités avec les règles ayant des propriétés en commun a été fait. Par exemple, dans la représentation qui est faite dans Chic du réseau de règles d'association, ils évitent les cycles avec des règles telles que A→B, B→C et C→A quand ils ne contiennent pas les réciproques. De plus, après avoir signalé que la contraposition est une règle d'inférence valable dans leur ensemble de règles sur deux propriétés, ils mènent une réflexion sur l'action de la négation sur un ensemble de règles ayant plusieurs propriétés en commun avec notamment, la règle d'inférence suivante :

$$(ab \rightarrow c) \vdash (a\overline{c} \rightarrow \overline{b}) \text{ et } (b\overline{c} \rightarrow \overline{a}).$$

Un travail du même genre est mené pour prendre en compte l'effet des valeurs différentes d'une même propriété sur le réseau des règles [106].

Les mesures *subjectives* permettent également de prendre en compte d'autres informations que celles spécifiques à la règle pour évaluer sa qualité. La règle produite n'est pas comparée aux autres règles extraites en même temps, comme dans le cas précédent, mais à des règles fournies par l'expert. L'*intérêt* de A. Silbershatz et A. Tuzhilin [216] en fait partie. Les auteurs quantifient le gain de croyance produit par l'apport d'une nouvelle connaissance E dans la croyance a s'appuyant sur une ancienne connaissance e. La formule de ce gain est $\sum_a \frac{p(a|E,e)-p(a|e)}{p(a|e)}$, il peut être positif ou négatif.

Ce qui manque aux indices de qualité

Ces indices de qualité sont associés à une règle unique. Et la règle est considérée comme l'association d'une partie gauche et d'une partie droite, ces deux parties étant formées de propriétés. Il y a eu un développement important des indices considérant les parties gauches et droites comme un tout indivisible, mais, à notre connaissance, dans le calcul d'aucun de ces indices n'a été considérée la composition en propriétés de ces deux parties. Cela pose le problème de la prise en compte des relations entre les propriétés avant leur fusion dans la partie gauche ou droite. Montrons ce problème sur un exemple. Imaginons que nous avons 100 sujets dont 10 seulement vérifient simultanément trois propriétés a, b, et c, ainsi que les deux propriétés a et b, et que 20 vérifient la propriété c. On n'a pas besoin de plus d'informations pour calculer la plupart des indices de qualité de la règle $ab \rightarrow c$ (Par exemple, le support est 10, la confiance est 1, la différence est 0,8). On a ainsi des indices de qualité égaux pour des cas qui différent beaucoup, comme celui où 90% des sujets ne possèdent aucune des deux propriétés a ou b, cas totalement différent de celui où 90% des sujets possèdent l'une sans posséder l'autre. Cet exemple n'est qu'un aperçu des relations entre propriétés qui sont ignorées par les indices de qualité des règles. Dans la dernière section de ce chapitre, elles sont exposées de façon plus approfondie. cours

[91] http ://www.ardm.asso.fr/CHIC.html

4.4 Difficulté de concilier des règles de bonne qualité avec une structure logique

Nous allons voir maintenant qu'une fusion de ces deux points de vue que sont : augmenter la qualité de chaque règle en s'appuyant sur les indices, et augmenter la qualité de l'ensemble du jeu de règles en s'appuyant sur sa structure logique, est difficile. On va d'abord montrer sur un petit exemple que la transitivité, règle d'inférence du premier point de vue, n'est pas vérifiée en général dans un jeu de "bonne qualité" selon le second de point de vue, c'est-à-dire dont le support et la confiance dépassent des seuils donnés à l'avance, puis montrer que la négation des propriétés, qui est prise en compte selon le second point de vue grâce à certains indices comme l'indice d'implication statistique, pose des problèmes selon le premier.

4.4.1 Transitivité et règles d'association

Supposons qu'on dispose du tableau des valeurs de 3 propriétés A, B et C pour 20 sujets, dont on a extrait les règles de support supérieur ou égal à 4 et de confiance supérieure ou égale à 0,5. On a notamment la règle A→B et la règle B→C. L'utilisateur peut-il encore inférer selon la transitivité que la règle A→C fait également partie des règles extraites ? La réponse est non. Et même si on lui donne le support et la confiance de chacune de ces deux règles, il ne peut pas deviner sauf cas extrêmement particulier le support et la confiance de la règle A→C. Dans cet exemple, si la règle A→B a pour support 4 et confiance 0,5, et la règle B→C a pour support 6 et confiance 0,5, le support s de la règle A→C peut prendre toutes les valeurs de l'ensemble 0,1,...,8 tandis que la confiance pour s non nul est égale à s/8. Nous avons représenté dans la figure par un diagramme de Venn[92] deux cas extrêmes pouvant se produire, le cas le moins "surprenant" étant celui ou les trois ensembles se rencontrent, comme dans la figure 4.3.
Détaillons le diagramme de la partie gauche de la figure 4.2. La propriété A est vérifiée par 8 sujets, 4 qui vérifient également la propriété B, et 4 qui ne vérifient que A (ni B, ni C). L'intersection entre A et B contient 4 sujets, c'est-à-dire que ces 4 sujets vérifient à la fois A et B. Le support de la règle A→B est égal au support du motif AB, c'est-à-dire au nombre de sujets qui vérifient simultanément A et B. On a vu que ce nombre est 4. Et la confiance est le quotient de ce nombre par le nombre d'éléments qui vérifient A, soit 8, ce qui fait 0,5. On voit que A est disposé de telle façon qu'aucun de ses sujets ne vérifie C. C'est pour cela que le support de la règle A→C est 0. Et on n'a pas calculé la confiance, car en général on élimine les motifs de support nul.
Dans la partie droite, A est passé de l'autre coté, il est entièrement à l'intérieur de C, ce qui explique que la confiance de la règle A→C soit 1. Le cas le plus courant est celui de la figure 4.3, où A rencontre de façon "proche" B et C et leur intersection. On peut voir qu'alors la règle A→C a un support et une confiance proches de ceux des autres règles. Cela correspond ici à une forme d'indépendance[93] entre A, B et C.

Pour conclure sur cet exemple, partant d'un jeu de règles d'association extrait selon le deuxième point de vue en utilisant des seuils de support de 4 et de confiance de 0,5, si on combine deux règles A→B et B→C de ce jeu selon le premier point de vue par la transitivité,

[92] C'est une représentation pratique des éléments de moins de quatre ensembles faisant apparaître leurs éléments communs. Chaque ensemble est représenté par une ligne fermée, ses éléments peuvent être représentés chacun par une croix ou tous par leur effectif écrit à l'intérieur de cette forme. Les éléments figurant à l'intérieur de plusieurs formes sont communs à celles-ci.

[93] La définition de l'indépendance entre deux variables correspond à des notions assez intuitives sur l'absence de liaison. Quand on passe à plus de deux variables, cela se complique, comme on peut le voir dans [16].

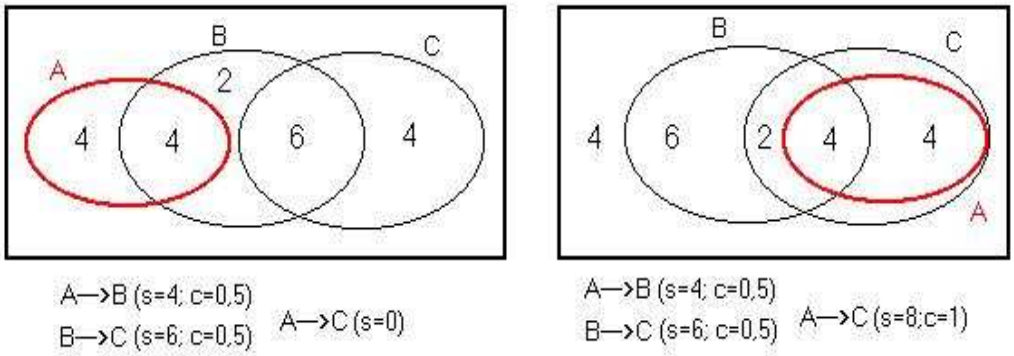


FIG. 4.2 – transitivité 1 .

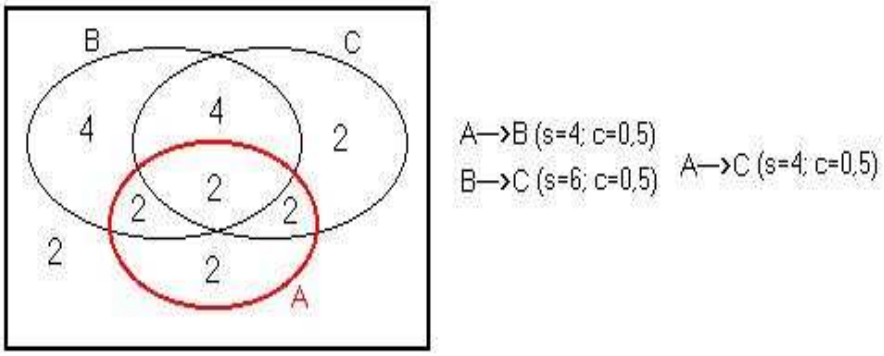


FIG. 4.3 – transitivité 2 .

en la règle A→C, elle ne fait pas toujours partie du jeu, dans la mesure où son support peut être en dessous de 4 et/ou sa confiance en dessous de 0,5, comme c'est le cas du graphique de gauche de la figure 4.2. Certes, en cas d'indépendance entre A, B et C, comme dans la figure 4.3 la propriété est vérifiée, mais le but des règles d'association est plutôt de faire apparaître des liaisons entre propriétés que leur indépendance. Ainsi le jeu de règles trouvé selon le deuxième point de vue peut être rejeté si on se place selon le premier point de vue. Notons au passage que cet exemple montre bien qu'une liaison entre trois variables est complexe et ne peut pas en général se déduire des liaisons entre les variables prises deux à deux.

4.4.2 Négation et treillis de Galois

La négation des propriétés est souvent utilisée dans le raisonnement du "sens commun". Un raisonnement complexe peut être démonté par un simple contre-exemple. Un jeu de règles soumis à interprétation se doit, à notre avis, d'être compatible avec la négation. La négation est prise en compte par certains indices de qualité comme nous le détaillerons plus loin, donc selon le deuxième point de vue. Nous allons examiner ce qui se passe en cas de négation de propriétés selon le premier point de vue, d'abord en utilisant le treillis des concepts, puis les bases de règles.

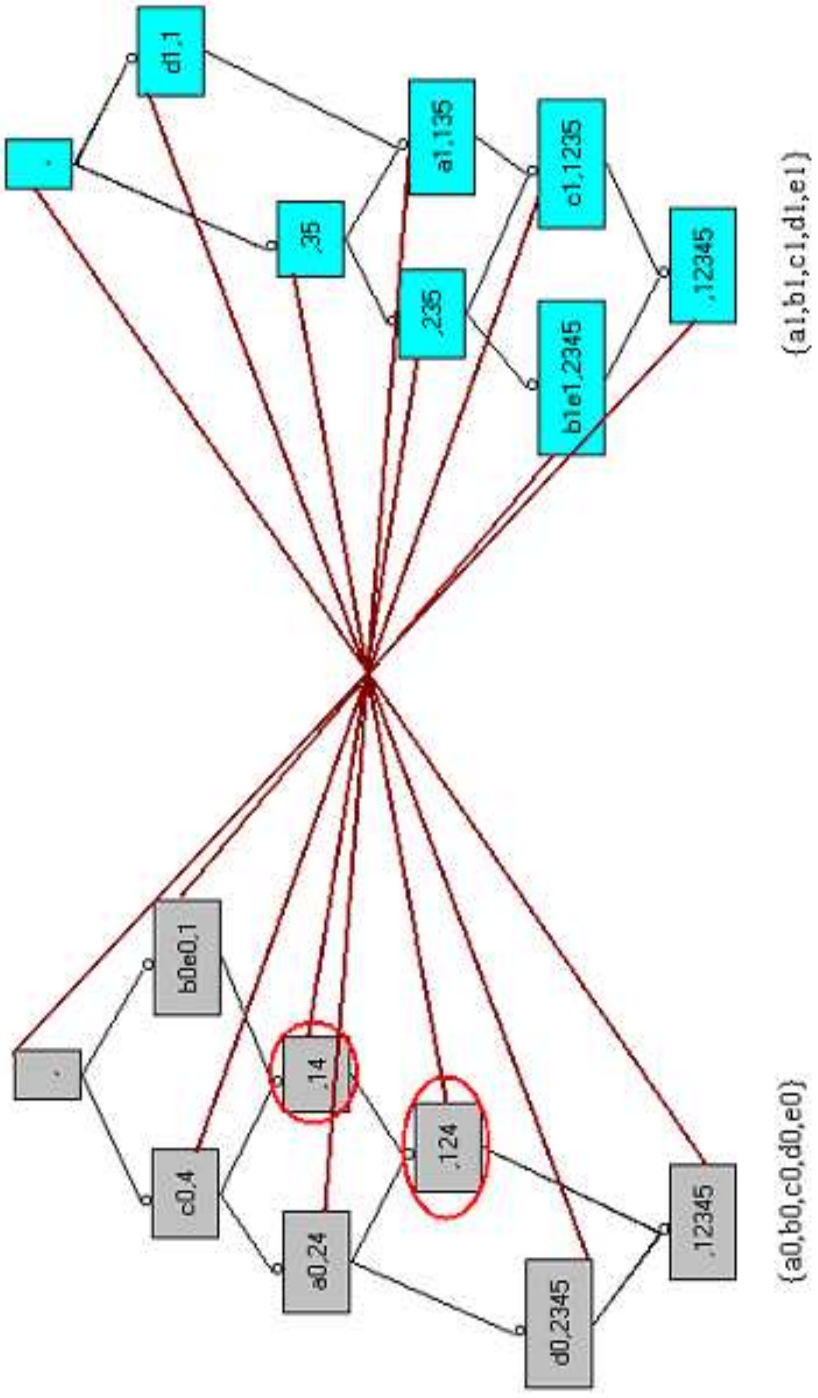


FIG. 4.4 – Le treillis des propriétés positives et négatives séparées

sujets	a	b	c	d	e
1	1	0	1	1	0
2	0	1	1	0	1
3	1	1	1	0	1
4	0	1	0	0	1
5	1	1	1	0	1

TAB. 4.4 – Le tableau T des propriétés

Au tableau T de la table 4.4 correspondent les 8 concepts (∅, 12345), (c, 1235), (a(c), 135)[94], (be, 2345), ((ac)d, 1), ((bce), 235), ((abce), 35), ((abcde), ∅), ainsi qu'on peut le voir dans la partie droite du graphique de la figure 4.4 (le diagramme de Hasse du treillis a été simplifié pour les propriétés, mais pas pour les sujets) où on a noté les propriétés a, b, c, d et e de façon positive[95] par a1, b1, c1, d1 et e1.

sujets	a0	a1	b0	b1	c0	c1	d0	d1	e0	e1
1	0	1	1	0	0	1	0	1	1	0
2	1	0	0	1	0	1	1	0	0	1
3	0	1	0	1	0	1	1	0	0	1
4	1	0	0	1	1	0	1	0	0	1
5	0	1	0	1	0	1	1	0	0	1

TAB. 4.5 – Le tableau T' des propriétés dédoublées du tableau T

Si on considère maintenant les propriétés négatives, notées a0, b0, c0, d0 et e0 obtenues en échangeant les valeurs 0 et 1 du tableau T, qui figurent dans le tableau T' de la table 4.5, on peut construire le graphe semi-simplifié des concepts de la figure 4.4 (dans chaque concept, on a le motif et les sujets séparés par une virgule, les motifs étant simplifiés mais pas les sujets) et en transformant celui des propriétés positives par symétrie de la façon suivante : on transforme toutes les lettres du motif simplifié en remplaçant 1 par 0, et on fait le complément de l'ensemble des sujets. Par exemple, au concept (a(c), 135), qui est écrit (a1, 135) on fait correspondre le concept (a0, 24). En fait il n'y a que 6 concepts (∅, 12345), (d0, 2345), (a0(d0), 24), (b0e0, 1), ((a0d0)c0, 4), ((a0b0c0d0e0), ∅), les deux candidats (∅, 124), (∅, 14), obtenus à partir des concepts ((abce), 35) et ((bce), 235) n'en sont pas[96]. Ceci nous montre que la négation n'est pas compatible avec le treillis des concepts, puisque partant d'un treillis de 8 concepts, on obtient par négation un treillis de 6 concepts.

Regardons maintenant les bases de Guigues et Duquenne associées. La base des propriétés négatives du tableau (c0→a0, a0→d0, b0<→e0) s'obtient en prenant les contraposées de celle des propriétés positives (a1→c1, d1→c1, e1<→b1). Ce qui semble montrer que la négation est

[94] On indique par a(c) le fait que le motif est ac, mais que sa forme simplifiée est a.

[95] La notation a indique que seule la propriété a est considérée, et pas sa négation. Quand on la note par a1, cela signifie que la propriété a va être dédoublée en deux propriétés a=a1, et sa négation a0. On dit que c'est une notation positive, par analogie avec la façon dont on écrit le nombre 3 de façon positive en +3 quand on envisage l'existence de nombres négatifs comme -3.

[96] Le lecteur désireux de s'en assurer peut construire directement ce diagramme selon les indications figurant dans le paragraphe consacré aux treillis de Galois.

compatible avec la formalisation en base de Guigues et Duquenne alors qu'elle ne l'est pas avec les treillis de concepts[97]. Toutefois, cette base n'est elle-même pas totalement compatible avec la négation, car dès qu'il y a plus de 2 propriétés dans la règle, on ne peut plus déduire de cette façon automatique la base des règles entre les propriétés négatives de celle entre les propriétés positives. Par exemple, si dans la première base, il y avait une règle a1f1→b1, elle ne pourrait pas se transformer en la règle b0→a0f0. En effet l'ensemble des sujets vérifiant a1 et f1 est obtenu par intersection des ensembles de sujets vérifiant chacune des deux propriétés, il n'est pas le complémentaire de l'intersection des ensembles de sujets vérifiant chacune des deux propriétés a0 et f0, mais de leur réunion[98](lois de Morgan[99]). Par contre R. Gras et al. [102] le prennent en compte dans l'utilisation de leurs indices : ils signalent que cela donne les règles a1b0→f0 et b0f1→a0.

Nous voyons ainsi que la négation n'est pas prise en compte selon le premier point de vue, c'est-à-dire dans le modèle logique des règles d'association, alors qu'elle est un peu plus prise en compte selon le second point de vue, notamment par l'utilisation faite de l'indice d'implication [102, 224] .

4.4.3 Peut-on se passer de la structure logique de l'ensemble de règles

On a vu qu'un jeu de règles de bonne qualité peut être apprécié selon deux points de vue difficiles à concilier. On peut se demander si un seul point de vue ne serait pas suffisant, et alors lequel choisir. L'importance du second point de vue n'est plus à démontrer. Une règle doit être de bonne qualité afin de servir à un utilisateur. Nous avons vu que ce point de vue est à l'origine de nombreuses recherches qui ont déjà produit une quantité d'indices, puis de recherches sur la classification de ces indices. Puisque ce second point de vue est si important, peut-on oublier le premier ? Il a fait également l'objet de recherches, mais dans le sens d'optimisation des algorithmes, pas dans le sens de l'amélioration de la qualité pour un utilisateur non informaticien. Nous allons montrer sur un exemple qu'on ne peut pas se satisfaire d'un jeu de règles qui ne respecterait pas la logique du "sens commun".

Cet exemple est tiré du livre de J. Pearl [192] sur la causalité. Il évoque le problème d'interprétation délicat posé par le "paradoxe de Simpson" sur un exemple médical. On donne à un médecin les résultats observés sur un groupe de malades. La relation complexe entre les trois propriétés que sont l'administration d'un médicament (C), la guérison d'un malade (E) et le sexe de ce malade (F) est indiquée par les règles suivantes[100] :

1. Tous sexes confondus, la guérison E est de 40% quand on n'administre pas le médicament (non C), et de 50% dans le cas contraire (C).
2. Si le malade est une femme (F) la guérison E est de 30% quand on n'administre pas le médicament (non C), et de 20% dans le cas contraire (C).
3. Si le malade est un homme (non F) la guérison E est de 70% quand on n'administre pas le médicament (non C), et de 60% dans le cas contraire (C).

[97] Notons au passage qu'il n'y a pas donc pas équivalence entre le treillis des concepts et la base de Guigues et Duquenne

[98] On peut s'en convaincre en associant à la propriété a1 vérifiée par les sujets s1, s3 et s5, une propriété f1 vérifiée uniquement pour les sujets s2, s3 et s5. le motif "a1f1" est uniquement vérifié pour s3 et s5. Comme la propriété f0 est vérifiée pour les sujets s1 et s4, le motif "a0f0" est vérifié pour s4. Comme b est vérifié également pour les s3 et s5, on a la règle af→b, mais b0 étant vérifié pour s1, on n'a pas la règle b0→a0f0.

[99] Ces lois sont relatives aux propriétés liant la négation, l'intersection et la réunion d'ensembles, et se trouvent dans tous les livres d'algèbre "moderne", c'est-à-dire traitant de la théorie des ensembles [93].

[100] Les conditions numériques permettant l'apparition de ce paradoxe sont détaillées en annexe.

J. Pearl reprend alors le raisonnement de Lindley et Novick [167] qui sont à l'origine de cet exemple :
"Si un patient entre dans le cabinet du médecin et que c'est un homme ou une femme, le médecin ne lui prescrit pas le médicament, en suivant les probabilités à sexe donné, alors que s'il doit laisser une prescription pour un patient dont il ne connaît pas le sexe, il lui prescrira ce médicament."

Et ils concluent que c'est un raisonnement stupide, qu'en fait il ne prescrira pas non plus le médicament car il a connaissance des tables par sexe. Ils proposent alors de reprendre le raisonnement en considérant F non comme le sexe mais comme une pression sanguine basse. Dans ce cas le médecin prescrit le médicament car il ne regarde que la table des effets combinés, et pas les tables selon le niveau de pression sanguine. La différence de raisonnement est que dans le premier cas, comme F est le sexe, le médecin lui attribue un rôle causal (en sciences humaines on n'imagine pas qu'un traitement peut changer le sexe d'un individu, donc chaque fois que le sexe est présent dans un schéma explicatif, il fait partie des causes) alors que dans le second, quand F est la pression sanguine il la considère comme un effet du traitement, et il n'a aucune raison d'examiner l'effet du traitement C sur la guérison E du patient en segmentant selon l'effet du traitement sur F.

Et après avoir établi par plusieurs exemples qu'il n'y aucune connection logique entre la vue statistique et la vue causale, J. Pearl établit deux définitions mathématiques différentes de ce qu'il appelle le concept de "confounding", une causale, et l'autre selon un critère associatif qu'il définit comme non purement statistique car prenant en compte également un présupposé causal.

Nous n'irons pas dans ce sens, notre but n'étant pas de faire nous-mêmes un raisonnement à partir de la base de données, mais de fournir un jeu de règles d'association qui sera utilisé par des experts pour faire un raisonnement. De cet exemple nous tirons tout simplement l'enseignement que laisser un jeu de règles avec trois règles aussi délicates à interpréter sans les avoir pointées, ou sans les avoir éliminées est très gênant. En effet, vu le nombre de règles obtenues, si l'utilisateur prend une décision au vu de l'une des règles et qu'il découvre après coup une autre règle qui lui aurait fait prendre la décision contraire, il risque de rejeter en bloc le jeu de règles d'association.

Après avoir ainsi montré qu'on ne peut ignorer impunément le premier point de vue en ne regardant pas la qualité globale des règles, nous allons aborder l'aspect informatique des règles d'association en faisant non seulement le tour des algorithmes d'extraction automatique des règles d'association dans le prochain paragraphe, mais également en étudiant dans un autre paragraphe un formalisme proche de celui des règles d'association, qui est celui des dépendances fonctionnelles. Bien que son but soit différent - il ne s'agit pas dans ce cas d'extraire des règles de données, mais de ranger les données pour qu'elles respectent des règles- il y a moyen de tirer des enseignements sur la façon dont sont définies ces dépendances fonctionnelles et dont opèrent leurs règles d'inférence, ainsi que des problèmes rencontrés lors de cette formalisation.

4.5 Les règles d'association extraites des grosses bases de données

Pour extraire les règles d'association des grosses bases de données, on procède principalement en deux temps. On extrait d'abord tous les motifs fréquents, c'est-à-dire dont la fréquence dépasse un seuil fixé par l'utilisateur, ou seulement un nombre réduit de ces motifs fréquents permettant de générer toutes les règles issues des motifs fréquents [101]. Puis on extrait de ces motifs fréquents

[101] Nous allons voir qu'en fait on obtient souvent une partie seulement de ces règles les autres pouvant en principe s'en déduire en appliquant des règles d'inférence

les règles dépassant un seuil donné de confiance.

Nous ne nous intéressons ici qu'aux algorithmes générant un jeu de règles d'association représentant la totalité des données, notre but étant leur exploration à travers cette méthodologie. Nous renvoyons le lecteur intéressé par les algorithmes permettant de chercher des règles d'association en utilisant un langage d'interrogation à l'article de M. Botta, J.-F. Boulicaut, C. Masson et R. Meo [25]. Nous décrivons d'abord les premiers algorithmes Apriori et GenRules [2], puis nous exposons rapidement les améliorations proposées par leurs successeurs.

4.5.1 L'algorithme A priori

C'est un des premiers algorithmes [2] permettant d'extraire les motifs fréquents. Il se fait par étapes successives, les données pouvant être représentées par un tableau T booléens Sujets×Propriétés de dimension n×p, un seuil de support a sont donnés.

- étape 1 : Initialisation
  On cherche F_1 l'ensemble de tous les motifs fréquents de longueur 1 dont le support est supérieur ou égal au seuil. F, l'ensemble de tous les fréquents est initialisé à F_1
- étape k>1
  1. Génération de C_k par examen de F_{k-1}
     On génère l'ensemble C_k de tous les motifs de longueur k candidats, c'est-à-dire susceptibles d'être reconnus comme motifs fréquents, en concaténant deux motifs distincts de Fk-1 qui ne diffèrent que d'une propriété. Cette étape peut se faire ainsi grâce à l'antimonotonie du support/l'inclusion des sous-ensembles de propriétés.
  2. F_k obtenu par élagage de C_k L'ensemble F_k est obtenu en éliminant de C_k tous les motifs dont le support est en dessous du seuil. On ajoute alors à F les éléments de F_k.

  On passe de l'étape k à l'étape k+1 si F_k est non vide.

Cet algorithme admet des variantes qui sont décrites en détail dans la thèse de N. Pasquier [191], mais on y retrouve le même enchaînement des étapes, ce qui impose qu'on accède non seulement au support de chaque motif fréquent ou candidat mais également aux sujets correspondants, et ceci pour chaque étape.

4.5.2 L'extraction des règles qui s'ensuit

Une fois extraits les motifs fréquents, on parcourt leur ensemble F en décomposant chaque motif en deux parties qui sont la partie gauche et la partie droite d'une règle d'association. Ainsi pour un motif d'une longueur donnée, on obtient autant de règles d'association que de couples de sous-motifs différents. Quand on désire extraire les seules règles dépassant un seuil donné de confiance, des versions plus rapides de cet algorithme existent, qui s'appuient sur les propriétés de la confiance. C'est le cas de l'algorithme "Genrules" proposé par R. Agrawal et R. Srikant [2] qui diminuent le nombre de recherches en explorant les sous-motifs de chaque motif fréquent selon leurs longueurs décroissantes. Cela permet d'éviter d'explorer systématiquement toutes les parties d'un motif en utilisant la propriété suivante de la confiance : si un sous-motif m' d'un motif m produit des règles de confiance inférieure au seuil, alors il en est de même de tout sous-motif de m'. Par exemple si la règle ABC→D construite sur le motif fréquent ABCD a une confiance inférieure à un seuil c, c'est le cas également des règles AB→CD, AC→BD, BC→AD, A→BCD, B→ACD et C→ABD.

Le pire des cas a lieu quand toutes les propriétés sont vérifiées par tous les sujets, ce qui se produit quand le tableau booléen sujetsXpropriétés ne contient que des 1. Il y a alors p étapes,

et chaque ensemble F_k est formé d'autant d'éléments que de parties à k éléments de l'ensemble de propriétés, soit C(k,n), et il y a en tout 2^p motifs fréquents de support n si on inclut le motif vide vérifié par tous les éléments, et $\sum_{k=1}^{p} 2^k$ (soit $2^{p+1} - 2$) règles d'association de support n et de confiance 1, si on accepte les règles d'association contenant un motif vide en partie gauche ou droite. A l'autre extrême, si le tableau T n'est formé que de 0, F1 est vide, et on n'obtient aucun motif fréquent en dehors du motif vide, donc aucune règle d'association. Entre ces deux extrêmes, le nombre de fréquents et de règles d'association varie non seulement en fonction de la proportion de 1 présents dans le tableau, mais également en fonction de la distribution des valeurs 1 dans ce tableau [243].

4.5.3 Les algorithmes suivants

Les algorithmes proposés par la suite sont des améliorations de l'algorithme Apriori dans plusieurs directions. On a essayé de réduire le nombre de données explorées en même temps, par une lecture à travers une fenêtre glissant sur les données [33], par tirage au sort d'un partie des données (algorithme "Sampling" de Toinoven [228]), par partition de l'ensemble des sujets (algorithme "Partition" de Savasere [211]). On a aussi pris en compte la structure de treillis des sous-ensembles de propriétés afin de se limiter à une partie des motifs fréquents, comme les *maximaux* dans l'algorithme de Godin et Missaoui [95], et dans ClusterApr, Eclat, MaxEclat, Clique, MaxClique, TopDown, algorithmes décrits dans la thèse de N. Pasquier [191], les *fermés* dans Close [191] et Titanic [222], les *free sets* dans l'algorithme de J.-F. Boulicaut et al. [28], les *clés* dans Pascal [13]. Et les algorithmes Charm [243], FP-Tree, FP-Grow [114] et dEclat [244] s'appuient sur une représentation des données par un arbre construit en explorant simultanément l'espace des propriétés et celui des sujets. Le parcours de l'arbre permet au premier algorithme d'éviter la création de sous-motifs, et aux derniers d'éviter la création de motifs candidats.

La plupart de ces améliorations visent à rendre plus efficace l'extraction des règles construites à partir des motifs fréquents. Signalons que les algorithmes d'extraction de règles d'association utilisés après extraction d'une partie réduite de motifs fréquents ne fournissent souvent qu'une partie des règles d'association. C'est le cas notamment de l'algorithme JEN [83], opérant sur l'ensemble des motifs fermés fréquents générés par CHARM après exécution d'un algorithme naïf pour avoir l'ordre complet des motifs fermés fréquents. Il permet d'obtenir des règles de deux types : "RIE exactes" et "RIA approximatives", les premières étant la base de Guigues et Duquenne, et les secondes étant la base de Luxemburger qui ont été décrites dans l'état de l'art (paragraphe intitulé "les bases de règles").

Chaque nouvel algorithme est comparé à un ensemble d'autres contenant souvent Apriori, en faisant varier les seuils de support, de confiance, et les tableaux de valeurs, ces derniers étant pris dans les bases de données de l'"UCI Knowledge Discovery in Databases Archives" (http ://kdd.ics.uci.edu/), pour lesquelles les dimensions des tableaux sont variées, ainsi que les liens entre les colonnes, ou dans les bases de données synthétiques de [221]. Notons que sont essentiellement pris en compte des critères de temps d'exécution, et que sont souvent passées sous silence les étapes de pré-traitement (prise en compte des valeurs manquantes, recodage des propriétés non booléennes). De plus ces comparaisons fournissent des classements différents selon les propriétés statistiques des bases de données. M.J. Zaki et C.-J Hsiao [243] montrent ainsi que la différence de qualité des algorithmes n'est pas la même selon la distribution statistique des bases de données de l'UCI utilisées habituellement pour ce genre de tests, en particulier selon la distribution de la longueur de leurs motifs fermés, ces distributions pouvant être bimodales ("Mushroom", "T40"), symétriques ("chess", "pumsb", "connect"), asymétriques ("gazelle", "T10").

4.5.4 Conclusion

Les algorithmes créés pour extraire les règles d'association des grosses bases de données n'utilisent que deux indices de qualité des règles qui sont la fréquence et la confiance, ces dernières ne devant pas passer en dessous des seuils fournis par l'utilisateur. Comme ces algorithmes sont très rapides, l'utilisateur désirant mesurer d'une autre façon la qualité des règles peut les modifier légèrement si sa mesure de qualité le permet, mais il peut aussi choisir de les utiliser pour dégrossir les données et créer un algorithme explorant leurs résultats pour en extraire les règles qu'il trouve les meilleures. Nous avons procédé dans notre recherche de la première façon pour créer des règles d'association floues, et de la seconde en créant des méta-règles d'élagage pour obtenir un jeu de règles dont la qualité se mesure à la petitesse de son nombre d'incohérences.

Le fait d'essayer d'introduire dans les algorithmes une variabilité des seuils de fréquences des motifs [169], de noter que les comparaisons entre rapidité des algorithmes dépendent des caractéristiques statistiques des données [243], ou de chercher à estimer les valeurs des fréquences des motifs après avoir calculé leurs valeurs pour un nombre réduit d'entre eux [179] montre que les chercheurs de la communauté de fouille de données ont essayé à plusieurs reprises de prendre en compte des éléments statistiques dans les algorithmes. Il semble raisonnable d'envisager, ce qui n'est pas encore fait à notre connaissance, des versions de ces algorithmes contenant une phase d'initialisation qui ferait une investigation des caractéristiques statistiques des données, afin de fixer en fonction de celles-ci les principaux paramètres, comme le seuil de support par exemple.

4.6 Les dépendances fonctionnelles dans les bases de données

Nous exposons d'abord dans ce paragraphe ce que sont les dépendances fonctionnelles des bases de données, et leurs ressemblances et différences avec les règles d'association du point de vue du principe, puis les règles d'inférence qui permettent de combiner les dépendances fonctionnelles, et finalement nous regardons leurs différences du point de vue technique.

4.6.1 Le stockage des données : les bases de données

Historiquement, la fouille de données intervient a posteriori sur des données qui ont été collectées dans des buts comptables (tickets de caisse de supermarché) ou gestionnaires (fichiers des organismes de santé, d'assurance), et mises ensuite à disposition des chercheurs pour des investigations plus poussées. On peut disposer ainsi sur Internet de nombreuses bases de données). On s'intéressera toutefois en partie au mode de collecte et de stockage des données car ils ont des conséquences sur le type de traitement qu'on peut faire sur les données, et sur les conclusions à tirer des résultats, ainsi que leur mode d'interprétation.

Une "base de données" informatique a pour rôle de stocker de l'information. Pour écrire de l'information sur un support quelconque (par exemple le disque dur d'un ordinateur) il convient de la coder. Nous n'allons pas nous intéresser ici à la transcription physique de l'information en suites de 0 et de 1, mais à sa transcription logique en "données" Ce terme de données désigne une information non seulement codée, mais assez structurée pour qu'on puisse lui appliquer des traitements automatiques comportant souvent des formulations mathématiques. Le formalisme des premières bases des données était plutôt dirigé vers des opérations d'ordre comptable, la structuration des données devant permettre l'émission de bulletins de salaire, de factures, de commandes. Les bases de données scientifiques sont plutôt l'objet de traitements issus des statistiques, comme "l'analyse des données", ou de l'informatique, comme les "réseaux neuronaux" de

l'intelligence artificielle, ou de la "fouille de données" au carrefour de l'informatique et des statistiques. Le premier type de motivation est à l'origine d'une structuration actuelle des données dans les SGBD relationnels (Systèmes de Gestion de Bases de Données) selon une normalisation complexe, alors que les bases de données scientifiques ont une structure beaucoup plus simple en général. C'est cette dernière structure que nous allons décrire d'abord car elle suffit pour développer les premiers éléments de notre problématique. Certains aspects liés à la normalisation plus complexe de la première structure seront décrits plus loin pour aider à préciser notre problématique.

4.6.2 Les principes

Nous allons comparer les deux types de règles, dépendances fonctionnelles et règles d'association selon leur place dans l'utilisation et la gestion de bases de données.

Les dépendances fonctionnelles

Une base de données relationnelle est un ensemble de relations exprimées par des "tables" qui sont des tableaux de type Sujets×Propriétés et un ensemble de contraintes sur ces tables. Considérons par exemple la base de données d'un magasin. Elle contient plusieurs tables dont la table "Clients" où figure la liste des clients avec des renseignements les concernant et la table "Commandes" dans laquelle on note pour chaque commande la liste des produits commandés, avec les quantités.

Les contraintes ont pour rôle de maintenir la cohérence de la base de données lors des mises à jour, qui consistent à modifier, enlever des éléments ou à en rajouter. Par exemple, si on veut intégrer la commande d'un nouveau client dans la table "Commandes", l'ajout de la ligne correspondante nécessitera l'ajout d'une ligne dans la table "clients" avec toutes les indications nécessaires . Les programmes de saisie d'une nouvelle commande doivent donc s'appuyer sur une structure de la base de données décrivant formellement ces liens entre les attributs des tables.

C'est pour ces raisons [27] qu'une normalisation des bases de données a été proposée il y a une trentaine d'années par E.F. Codd. Aux trois "formes normales" qu'il a définies en 1972, se sont ensuite ajoutées d'autres formes normales, la plus courante étant celle de Boyce-Codd en 1974. Ces formes normales sont une description des liens souhaitables entre les attributs des tables, le but premier de cette normalisation étant d'éviter toute redondance. La redondance en base de données signifie qu'un même renseignement figure en plusieurs endroits. Par exemple, l'adresse pourrait figurer dans la table "Commandes" et dans la table "Clients", ce qui est gênant lors des mises à jour, car si le client déménage, on risque de changer l'adresse dans la table "Clients" en oubliant de la changer dans la table "Commandes" et la commande sera alors livrée à la mauvaise adresse. Pour éviter cela, on ne fait figurer l'attribut "adresse de livraison" que dans la table "Clients", et on indique dans l'attribut "adresse de livraison" de la table "Commandes" le "numéro d'identification" du client dans la table "Clients". Cela impose donc de déterminer un statut particulier pour l'attribut "numéro d'identification" dans la table client, qui devient une "clé", car on ne peut accéder aux autres attributs du client qu'en passant par celui-ci.

Ces exigences de qualité d'une base de données vont en grande partie s'exprimer par un jeu de règles entre attributs, et par des règles d'inférence sur ce jeu de règles. Grâce aux premières règles, qui sont appelées "dépendances fonctionnelles", on peut écrire des programmes d'interfaçage avec la base de données (saisie, mise à jour, établissement de facture, comptabilité) qui prennent en compte sa structure, et avec les secondes, on peut écrire des programmes qui modifient la structure de la base de données, et donc le jeu de règles pour passer par exemple de deuxième

en troisième forme normale.

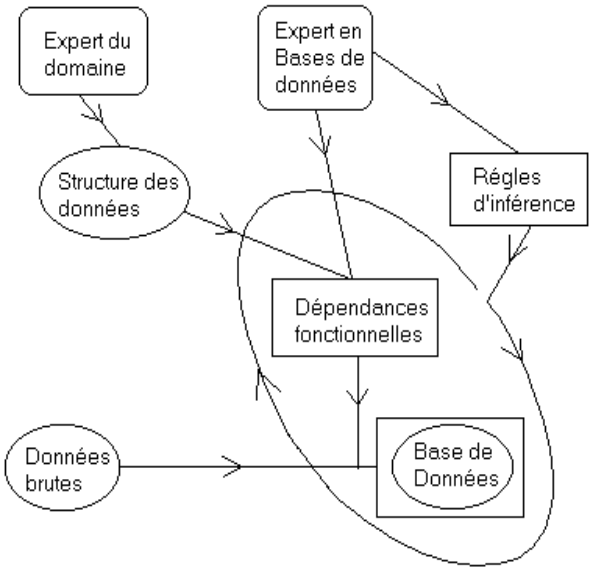


FIG. 4.5 – Production d'une base de données à l'aide de dépendances fonctionnelles

Dans la figure 4.5, nous représentons un schéma montrant le rôle des dépendances fonctionnelles dans la production d'une base de données. Imaginons par exemple que l'on désire créer la base de données d'une entreprise. Les acteurs sont en haut du schéma, ce sont l'expert du domaine, ici le gestionnaire de l'entreprise et l'expert en bases de données. On récupère des données brutes, comme les livres de comptabilité, les fichiers d'adresses des clients, les bons de livraisons, qui n'ont pas toujours de sens pour l'expert des bases de données, mais qui en ont un pour l'expert du domaine. Celui-ci connaît en effet la structure de ces documents, et peut la décrire à l'expert en bases de données. Leurs échanges vont permettre d'établir les dépendances fonctionnelles, et avec celles-ci on va pouvoir établir la structure de la base de données dans laquelle seront rangées les informations correspondant aux documents. Une fois ce prototype réalisé, l'expert en bases de données va le tester et peut-être réaliser que la normalisation n'est pas suffisante. En effet lors des échanges avec le gestionnaire, tout ne lui a pas été dit sur les propriétés des données. Ces omissions, voire ces incompréhensions, sont très courantes lors d'échanges entre deux spécialistes de domaines différents. Il va donc rajouter des contraintes sous forme de dépendances fonctionnelles, et avec l'aide des règles d'inférence, vérifier que la base de données est au moins en troisième forme normale. Une découverte de nouvelles dépendances fonctionnelles peut se reproduire chaque fois qu'il augmente la base de données, par de nouvelles saisies, ou par des modules supplémentaire, comme l'impression des commandes, ou l'envoi de publicités à des clients n'ayant pas passé de commande depuis longtemps. Et l'apparition de nouvelles contraintes est examinée à la lumière des règles d'inférence et peut avoir comme effet une remise en cause de la structure complète de la base de données. Cette étape d'optimisation de la base de données est indiquée par une ellipse actionnée par les règles d'inférence.

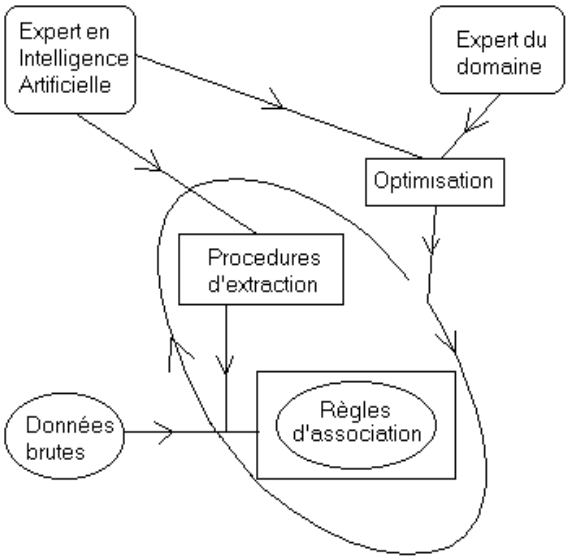


FIG. 4.6 – Production d'un jeu de règles d'association

Les règles d'association

Dans la figure 4.6, on a représenté un schéma semblable à celui de la figure 4.5, mais montrant le rôle des règles d'association. On part encore des données, brutes si elles sont simples, comme les tickets de caisse des supermarchés, ou figurant déjà dans des tableaux de données, mais cette fois l'expert en intelligence artificielle opère seul sur ces données, en leur appliquant des procédures d'extraction de règles d'association. Ces procédures produisent un jeu de règles d'association. L'expert du domaine intervient alors pour interpréter ce jeu de règles, et c'est à ce moment que les échanges ont lieu entre les deux experts, l'expert en intelligence artificielle montrant à l'expert des données comment se servir de ce jeu de règles. Cet échange peut être réel entre deux personnes, ou virtuel, à travers des interfaces de navigation ou des choix d'un sous-ensemble de règles qui sont faits pour aider l'utilisateur final à se servir du jeu de règles d'association. Cette mise au point du jeu de règles d'association, qui peut nécessiter de recommencer l'extraction avec des paramètres différents pour tenir compte des besoins des utilisateurs est indiquée par une ellipse d'optimisation qui peut être parcourue plusieurs fois.

La comparaison

On voit que les ensembles de règles d'association et les ensembles de dépendances fonctionnelles ont des points communs. Tous deux opèrent sur les mêmes objets qui sont des relations de type sujetsXattributs, dans des bases de données, ils expriment des relations entre attributs, et on attend que les règles puissent s'enchaîner selon des règles d'inférence. Mais ils diffèrent par leur mode de création et par leur utilisation. Les règles d'association sont découvertes sur des données déjà collectées, et apportent de nouvelles informations sur ces données alors que les dépendances fonctionnelles aident à ranger des informations préalables à des données selon une structure facilitant leur exploitation. Dans le premier cas, la connaissance découle des règles découvertes et c'est l'inverse dans le second, la connaissance sur les données permet d'écrire les

règles. Le jeu de règles d'association est optimisé selon des critères qui peuvent intervenir avant son extraction, ou après, et le jeu de dépendances fonctionnelles est examiné à l'aide de règles d'inférence qui permettent de voir à quelle normalisation correspond la structure et de l'optimiser en augmentant le rang de la forme normale.

Le produit délivré à l'utilisateur est dans le premier cas le jeu de règles, alors que dans le second c'est la structure de la base de données. Comme les procédures d'extraction de règles d'association, les dépendances fonctionnelles sont les outils de l'informaticien et peuvent être ignorées des utilisateurs. Toutefois, il arrive qu'on en informe l'utilisateur averti quand il s'étonne du comportement du produit délivré, afin de l'aider à mieux appréhender son fonctionnement. Aucune information n'est délivrée à l'utilisateur sur les règles d'inférence car elles représentent pour l'informaticien sa connaissance sur l'usage de ses outils, et n'a aucun intérêt pour l'utilisateur, même averti. Par contre l'optimisation du jeu de règles d'association dépend de l'utilisateur. C'est avec ses propres règles d'inférence qu'il va utiliser les règles d'association. A moins de lui fournir une interface qui lui permette de naviguer dans les règles d'association en intégrant à la représentation qu'il se fait de ses données les seules règles qui l'intéressent, il va falloir choisir un ensemble de règles qui puisse être utilisable par tout utilisateur non spécialisé en informatique. Cet utilisateur peut être l'expert du domaine, qui dans ce cas a des besoins spécifiques, comme de rechercher de nouveaux segments de marché si c'est du marketing, ou au contraire une vue générale des principales tendances, mais ce peut être aussi un utilisateur qui n'est pas connu au moment de l'extraction des règles. Nous appelons raisonnement selon le "sens commun" le raisonnement de l'utilisateur non spécialiste.

4.6.3 Les règles d'inférence des dépendances fonctionnelles

Pour décrire ces règles d'inférence, C. Carrez [49] fait la différence entre le schéma de relation R, qui représente la structure d'une table (son intention) et un exemplaire de la relation qui en est une instance (son extension). Le schéma de la relation est exprimée par 4 éléments qui sont le nom de la relation, la liste des propriétés P et de leurs valeurs possibles, la signification de la relation et la liste des contraintes imposées dont les dépendances fonctionnelles font partie. L'exemplaire de la relation est représenté par le tableau T donnant pour chaque sujet, élément de S sa valeur pour chaque propriété de P. Une définition de dépendances fonctionnelles plus adaptée à ce formalisme est la suivante :

Définition 4.6.1. *Dépendance Fonctionnelle.*
Soit R un schéma de relation sur l'ensemble de propriétés P, A et B deux sous-ensembles de P, on a la dépendance fonctionnelle A→B si pour tout ensemble de sujets S vérifiant un exemplaire de R, deux sujets de S ayant les mêmes valeurs pour toutes les propriétés de A ont également les même valeurs pour toutes les propriétés de B.

Nous ne reprenons pas ici la notation AB courante en bases de données pour représenter l'union des ensembles, mais nous notons A∪B, comme le fait N. Boudjlida [27], afin de rester cohérente avec les notations de ce mémoire. Toutefois, quand on a un ensemble réduit à une seule propriété, on écrira l'ensemble sans les accolades, et on gardera la concaténation pour les seules propriétés. Si on dispose d'un schéma de relation R, défini sur un ensemble de propriétés P, vérifiant un ensemble F de dépendances fonctionnelles, on peut déduire de F de nouvelles dépendances fonctionnelles qui en sont les "conséquences logiques". Par exemple, si on a la dépendances fonctionnelles A→B, où A et B sont deux sous-ensembles de P, et si C est un sous-ensemble de P contenant A, la connaissance pour un sujet s de toutes les valeurs des propriétés

de C permet de connaître celles de A, donc celles de B, et on en déduit ainsi la dépendance fonctionnelle C→B. Plus généralement nous reprenons ici les définitions de C. Carrez [49]

Définition 4.6.2. *Définition de la conséquence logique d'une dépendance fonctionnelle.*
Soit F un ensemble de dépendances fonctionnelles pour un schéma de relation R, et A→B une dépendance fonctionnelle. On dit que A→B est la conséquence logique de F, et on l'écrit F |– A→B si tout exemplaire de relation r de R qui satisfait les dépendances de F satisfait aussi A→B.

De cette définition découle celle de clôture :

Définition 4.6.3. *Clôture d'un ensemble de dépendances fonctionnelles.*
On appelle clôture de F l'ensemble $F^+ = \{X \rightarrow Y | F \vdash X \rightarrow Y\}$.

Ce qui va permettre de dire que deux ensembles de dépendances fonctionnelles sont *équivalents* quand ils ont même clôture, et de chercher parmi ces ensembles ceux formés des dépendances fonctionnelles les plus élémentaires, les autres s'en déduisant par conséquences logiques. Une dépendance fonctionnelle *élémentaire* A→B est telle que l'ensemble B des attributs de droite est formé d'un seul attribut qui n'est pas inclus dans l'ensemble A des attributs de gauche, et et qu'on ne peut pas trouver un sous-ensemble C de A tel que C→B, c'est-à-dire qu'elle ne peut pas se déduire d'une autre dépendance fonctionnelle par *augmentation* de son membre de gauche.
Pour trouver la clôture d'un ensemble F, on peut appliquer de façon itérative les règles d'inférence suivantes appelées "axiomes d'Amstrong" :

Définition 4.6.4. *Axiomes d'Amstrong.*
si F est un ensemble de dépendances fonctionnelles d'un schéma d'une relation R définie sur un ensemble P de propriétés, on peut trouver d'autres dépendances fonctionnelles de R en appliquant les trois règles suivantes :

1. *Réflexivité : si* $B \subset A \subset P$, *alors* $(A \rightarrow B) \in F$
2. *Transitivité : si* $(A \rightarrow B)$, $(B \rightarrow C) \in F$ *alors* $(A \rightarrow C) \in$
3. *Augmentation : si* $(A \rightarrow B) \in F$ $C \subset P$ *alors* $(A \cup C \rightarrow B \cup C) \in F$

Trouver la clôture d'un ensemble de dépendances fonctionnelles peut être long avec ces axiomes. D'autres règles d'inférence découlant de ces axiomes peuvent s'avérer plus efficaces comme les suivantes :

1. Union : *si* $(A \rightarrow B)$, $(A \rightarrow C) \in F$ *alors* $(A \rightarrow B \cup C) \in F$
2. Pseudotransitivité : *si* $(A \rightarrow B)$, $(B \cup C \rightarrow D) \in F$ *alors* $(A \cup C \rightarrow D \cup C) \in F$
3. Décomposition : *si* $(A \rightarrow B) \in F$, $C \subset B \subset P$*alors* $(A \rightarrow C) \in F$

La recherche d'un ensemble minimal de dépendances fonctionnelles permettant de retrouver F à partir de ces axiomes s'appuie sur les notions suivantes :

Définition 4.6.5. *Couverture d'un ensemble de dépendances fonctionnelles et irredondance de celle-ci selon M. Léonard [164].*
Si F est un ensemble de dépendances fonctionnelles d'un schéma de relation R, un ensemble G de dépendances fonctionnelles de R est une couverture de F si G est équivalent à F (c'est-à-dire a même clôture). Cette couverture est irredondante si aucune partie de G n'est couverture de F.

Une base de F est ainsi un ensemble de dépendances fonctionnelles élémentaires formant une couverture de F. Des algorithmes permettent de trouver une base de F. Notons toutefois qu'il n'y a pas une base unique, et que toutes les bases ne sont pas irredondantes. Nous en donnons un exemple dans l'introduction de la troisième partie de ce mémoire.

4.6.4 Les détails techniques

Nous allons maintenant regarder en détail ce qu'est une dépendance fonctionnelle, afin de voir si elle s'exprime par les mêmes calculs qu'une règle d'association.

Un exemple de table présente dans les schémas de bases de données

Pour cela nous reprenons un exemple du cours de M. Léonard [164]. D'après lui, les dépendances fonctionnelles "permettent aux concepteurs d'une structure de données de repérer les propriétés des objets qui jouent un rôle déterminant (un numéro de machine et une heure permettent de déterminer exactement une entité de l'emploi du temps)" Dans la relation EmploiDuTemps(NoMachine, NoHeure, NoProduit, NoLot), la dépendance fonctionnelle NoMachine,NoHeure→NoProduit,NoLot exprime la contrainte "à une heure donnée, une machine donnée fabrique un seul lot d'un même produit".

Appelons C cette contrainte, E, M, H, P et L les attributs EmploiDuTemps, NoMachine, NoHeure, NoProduit, NoLot et imaginons qu'on a saisi les données observées selon le tableau 4.6.

E	M	H	P	L
e1	m1	h1	p2	l1
e2	m1	h1	p2	l1
e3	m1	h1	p2	l1
e4	m1	h2	p1	l1
e5	m1	h2	p1	l1
e6	m2	h2	p3	l2
e7	m2	h2	p3	l2
e8	m3	h2	p2	l1
e9	m3	h2	p2	l1
e10	m3	h2	p2	l1

TAB. 4.6 – Une table T en 22ème forme normale, avec la dépendance fonctionnelle MH→PL

Dans cette table, on voit que la contrainte C qui est MH→PL est vérifiée car les 4 valeurs différentes de MH qui sont m1h1, m1h2, m2h2 et m3h2 déterminent entièrement les valeurs de P et L qui sont respectivement p2l1, p1l1, p3l2 et p1l1. La table est en deuxième forme normale car elle est en première forme normale (chaque ligne a pour chaque attribut des valeurs simples, et non composites c'est-à-dire pour H, h1 ou h2, mais pas h1h2) et de plus les contraintes E→M, E→H, E→P et E→L sont assurées de par la définition de E qui est une clé de la table. Par contre, elle n'est pas en troisième forme normale pour la raison suivante : il est possible de saisir une nouvelle ligne de données du moment que sa valeur de e est différente de celle des autres. Si on choisit e11, m3, h2, p3, l2, e11 aura les mêmes valeurs que e9 pour M et H, mais pas pour P et L. La contrainte MH→PL n'est plus vérifiée. Donc si cette relation doit être respectée, la saisie de la ligne telle que nous la proposons est erronée. Il convient donc de l'interdire. Pour cela, on peut passer en troisième forme normale en découpant le tableau en deux tables comme indiqué dans le tableau 4.7.

Désormais la première table assure les contraintes E→M, E→H et la deuxième table ayant comme clé MH assure la contrainte MH→PL et on retrouve les deux contraintes E→P et E→L en appliquant des règles d'inférence sur le jeu des 3 règles {E→M, E→H, MH→PL}. En pratique,

E	M	H
e1	m1	h1
e2	m1	h1
e3	m1	h1
e4	m1	h2
e5	m1	h2
e6	m2	h2
e7	m2	h2
e8	m3	h2
e9	m3	h2
e10	m3	h2

M	H	P	L
m1	h1	p2	l1
m1	h2	p1	l1
m2	h2	p3	l2
m3	h2	p2	l1

TAB. 4.7 – La transformation en 33ème forme normale : T est remplacé par deux tables T1 et T2

les attributs M et H de la première et ceux de la seconde ne sont pas les mêmes, un lien entre les deux tables permet de passer des attributs M et H de la première à ceux de la seconde. Ainsi, on ne peut pas créer dans la première table un enregistrement de type e11, m3, h3 sans avoir ajouté une ligne dans la seconde commençant par m3, h3. La table sous cette forme permet de garder la cohérence de la structure.

La table modifiée pour en extraire les règles d'association

Voyons le jeu de règles d'association extrait de cette table. Pour cela, il convient de remplacer chaque propriété par autant de propriétés qu'il y a de valeurs, exception faite de la propriété E. Celle-ci étant l'identifiant de la table correspond aux sujets habituels des tableaux de données. Cela nous donne le tableau suivant :

sujets	m1	m2	m3	h1	h2	p1	p2	p3	l1	l2
e1	1	0	0	1	0	0	1	0	1	0
e2	1	0	0	1	0	0	1	0	1	0
e3	1	0	0	1	0	0	1	0	1	0
e4	1	0	0	0	1	1	0	0	1	0
e5	1	0	0	0	1	1	0	0	1	0
e6	0	1	0	0	1	0	0	1	0	1
e7	0	1	0	0	1	0	0	1	0	1
e8	0	0	1	0	1	0	1	0	1	0
e9	0	0	1	0	1	0	1	0	1	0
e10	0	0	1	0	1	0	1	0	1	0

TAB. 4.8 – Tableau T' de données booléennes correspondant à la table T

Dans le tableau T', on retrouve la dépendance fonctionnelle MH→PL sous la forme de 4 règles d'association exactes qui sont m1h1→p2l1, m1h2→p1l1, m2h2→p3l2 et m3h2→p2l1. Mais ces règles d'association ne sont pas les seules règles du jeu de règles exactes extrait de ce tableau. Il y en a beaucoup plus qui se déduisent toutes de la base lisible sur le diagramme simplifié en figure 4.7du treillis de Galois associé à ce tableau.

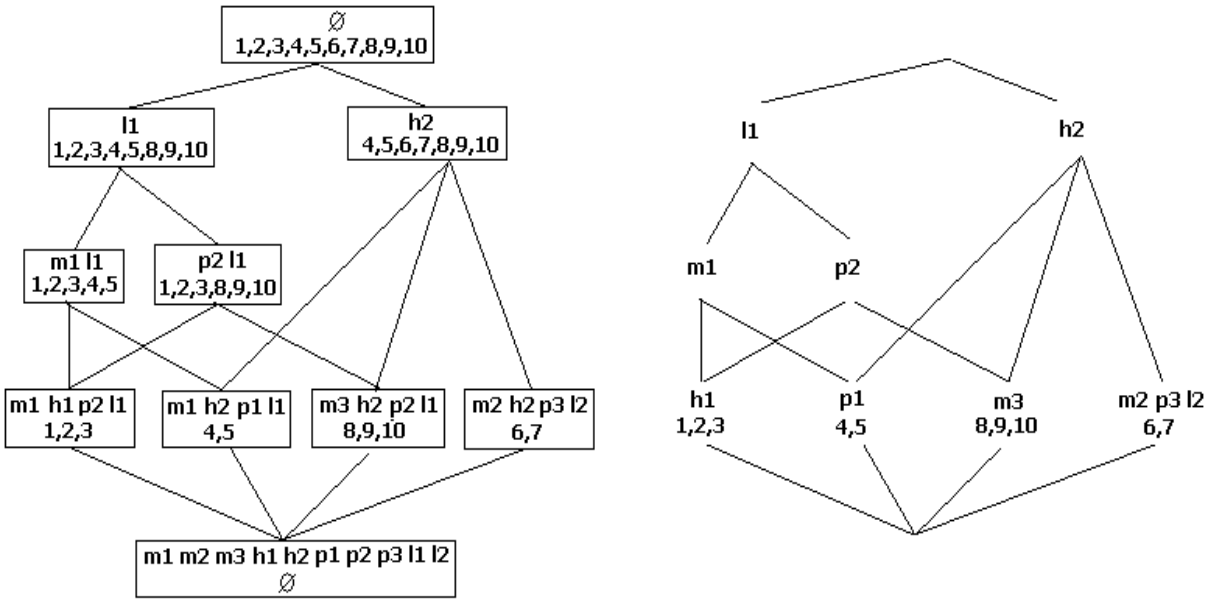


FIG. 4.7 – Diagramme de Hasse et diagramme simplifié du treillis des concepts du tableau T'

Et retrouver ces 4 règles parmi les autres nécessite qu'on soupçonne un lien entre MH et PL. Mais la découverte des 4 règles ne suffit pas à nous l'assurer. En effet, il va ensuite falloir se convaincre que pour les 6 valeurs possibles du croisement des 3 valeurs de M et des 2 valeurs de H, il est normal qu'il n'y ait que 4 règles sur 6 dans le jeu de règles. C'est-à-dire que si aucune règle n'est apparue ayant en partie gauche m2h1 ou m3h1 et en partie droite une valeur du croisement de PL, c'est que ces deux motifs n'apparaissent jamais. On peut s'en assurer en vérifiant qu'on n'a pas les règles m2h1→∅ et m3h1→∅ pour autant qu'on fasse apparaître des règles avec ∅ dans le jeu fourni. On voit donc qu'on a beaucoup de mal à retrouver une dépendance fonctionnelle à partir d'un jeu de règles d'association. Cela vient en partie du fait qu'on n'a pas maintenu le lien entre toutes les variables créées à partir d'une même variable. Et la restauration après coup de ce lien détruit est nécessaire pour pouvoir extraire la dépendance fonctionnelle MH→PL du jeu de règles d'association.

Comparaison des deux formalismes sur cette table

Pour conclure, notons que les règles d'association exactes font découvrir des liens comme par exemple ici l'équivalence entre m2, p2 et l2, qui pourraient peut-être s'exprimer en dépendances fonctionnelles moyennant une restructuration de la base de données en d'autres tables, et d'autres variables, comme c'est le cas de la table T2 qui exprime la dépendance fonctionnelle MH→PL, mais que le passage des règles d'association aux dépendances fonctionnelles ne peut pas se faire directement, sans ajout de connaissance supplémentaire sur la base de données.

Dans la figure 4.8 on a représenté une dépendance fonctionnelle simple, et une règle d'association simple, pour montrer leurs différences structurelles. Dans le graphique de gauche, A a 5 modalités, qui sont a1, a2, a3, a4 et a5, et B a 3 modalités qui sont b1, b2, et b3. Au croisement de a1 et de b1, il y a 20 sujets, indiqués par un gros point, surmonté du nombre 20, et dans les autres croisements de a1 avec B, il n'y a aucun sujet, ce qu'on a indiqué par une croix surmontée du nombre 0. La définition de la dépendance fonctionnelle exige que pour pour chaque modalité

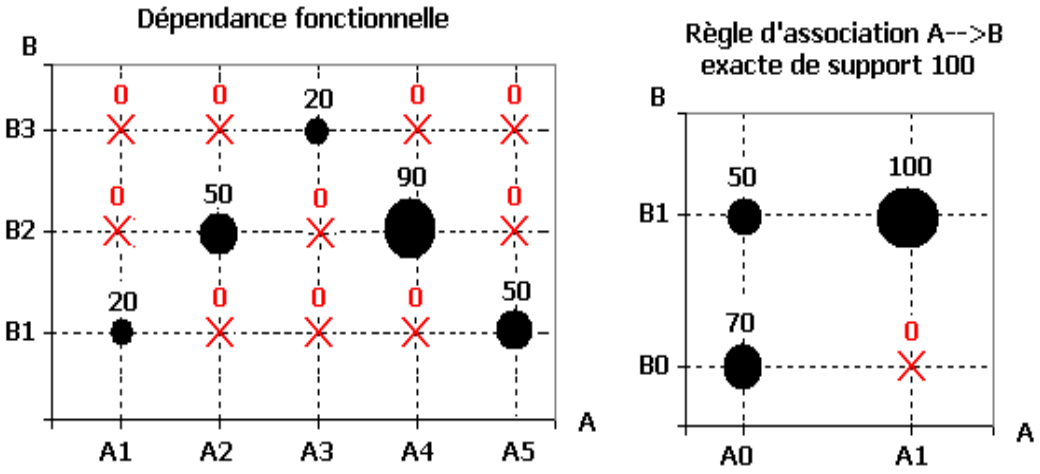


FIG. 4.8 – Une dépendance fonctionnelle et une règle d'association exacte

de A il y ait une seule modalité de B avec un nombre de sujets non nuls. Cela a pour conséquence que la classification des individus selon les diverses modalités de A est plus fine que celle selon les diverses modalités de B. Les 230 sujets se répartissent en 70 sujets pour lesquels la valeur de B est b1, 140 sujets pour lesquels elle est b2, et 20 sujets pour lesquels elle est b3. Puis les 70 sujets de la classe b1 se répartissent à leur tour en 20 sujets qui ont la valeur a1 de A et 50 qui ont la valeur a5, les 140 sujets de la classe b2 se répartissent en 50 sujets de la classe a2 et 90 de la classe a4, les 20 sujets de la classe b3 allant tous dans la classe a3. Ainsi la classification (b1 b2 b3) devient ((a1 a5)(a2 a4)(a3)).

Ce n'est pas la même chose pour la règle d'association. D'abord, A et B n'ont que deux modalités, la modalité 1, qui indique une présence, et la modalité 0 qui indique une absence. Et la règle d'association exacte exige une seule chose, c'est qu'il n'y ait aucun sujet ayant simultanément a1 et b0, ou autrement dit, ayant a1 sans avoir b1. On n'a pas cette fois de classification plus fine pour A, car on peut avoir a0 avec b0 et a0 avec b1.

4.6.5 Ce que nous apportent les dépendances fonctionnelles

Les données arrivent en principe une fois que les dépendances fonctionnelles ont été crées. Ces dernières doivent donc prendre en compte une information dont on ne dispose pas afin de la ranger de façon efficace quand elle arrivera. Pour cela, l'expert du domaine doit coopérer avec l'expert en Informatique. Pour que la structure des données soit la plus efficace possible, le spécialiste en bases de données dispose d'une normalisation assortie de règles d'inférences. Cela ne se passe pas de la même façon en extraction de règles d'association. Les règles ne sont découvertes qu'une fois les données collectées et débouchent rarement sur une restructuration de la base de données. En effet ce sont essentiellement des règles approximatives ayant donc des contre-exemples, ce que précisément la formalisation des bases de données essaie d'éliminer. Quant à l'expert du domaine, il a plutôt un rôle de consommateur que de collaborateur car dans la plupart des cas, nous ne le connaissons pas, et nous essayons de lui délivrer un produit fini en imaginant ses besoins. Mais il y a un point commun entre les règles d'association et les dépendances fonctionnelles, c'est leur syntaxe, ce qui permet de reprendre sans problème toute la construction des règles d'inférence sur les dépendances fonctionnelles pour la reporter

sur les règles d'association exactes mais pas approximatives. En effet, on a vu sur un exemple que la transitivité, règle d'inférence des dépendances fonctionnelles, ne peut pas s'appliquer aux règles approximatives. C'est en suivant ce formalisme que des techniques d'élagages des règles d'association ont été développées au sein de l'Université Simon Fraser des États-Unis d'Amérique et décrites dans les thèses de [245, 231, 51, 232].

On a vu dans la partie technique qu'on peut partir d'une dépendance fonctionnelle et arriver à une série de règles d'association entre les divers croisements des valeurs des propriétés intervenant dans la dépendance fonctionnelle, mais qu'il est difficile de remonter des règles d'association à la dépendance fonctionnelle. Pour cela on a recodé le tableau T en un tableau T'. Ce recodage produit des règles d'association ayant en général moins de valeur informative que la dépendance fonctionnelle. Il y a un autre inconvénient de ce type de recodage qui est d'ordre numérique. En effet le nombre de propriétés augmente vite si chaque propriété est recodée en plusieurs, et les sujets sont tellement émiettés dans les différents croisements de ces nombreuses variables qu'on est obligé d'abaisser le seuil de support, ce qui n'arrange rien. Nous proposons d'éviter cette perte d'information et ce surcoût numérique dans le cas de propriétés numériques en recodant celles-ci de façon floue. On a vu dans la partie technique que le sens mathématique d'une dépendance fonctionnelle diffère du sens mathématique d'une règle d'association. Nous considérons que la règle d'association A→B sur des propriétés numériques indique une croissance.

4.7 Conclusion

Dans cet état de l'art consacré à l'extraction des règles d'association, nous avons examiné ce que la communauté de fouille de données propose pour renforcer la qualité des jeux de règles d'association : des algorithmes de plus en plus performants, des indices de qualité de plus en plus nombreux et variés, des jeux de règles exactes structurés selon une logique formelle grâce aux treillis de Galois, et un formalisme de règles d'inférence sur les dépendances fonctionnelles. Malheureusement les jeux de règles produits par les algorithmes ne débouchent pas automatiquement sur un jeu bien structuré de règles de bonne qualité. Les exigences de qualité ne peuvent pas à la fois porter sur chaque règle et sur la structure du jeu. Nous avons vu en effet que dès qu'on ne se limite plus aux règles exactes, la transitivité n'est pas conservée et que la négation pose des problèmes dans des jeux de règles exactes issus de treillis de Galois. La démarche la plus courante adoptée dans la communauté de fouille de données consiste à privilégier la qualité de chaque règle par un choix approprié d'indices, et d'assurer ensuite la qualité du jeu de règles ainsi formé par la suppression de celles qui compromettent sa structure.

Mais nous avons constaté que le choix des indices est difficile, et qu'une fois ce choix fait, nous n'avons pas pour autant l'assurance qu'ils nous permettent de sélectionner les bonnes règles. En effet dès que plus de deux propriétés sont en jeu pour une règle, ces indices opèrent en partie sur des effectifs fusionnés et ne tiennent donc pas compte des relations complexes entre les propriétés. Cette complexité peut déboucher sur des incohérences dans les jeux de règles. Il convient de les repérer afin de les nettoyer.

Conclusion de la partie I

Nous proposons de contribuer à l'amélioration de la méthode d'extraction de motifs et de règles d'association de la façon suivante :

1. Pour décider si une règle est de bonne qualité, nous proposons de ne pas créer de nouvel indice, mais de s'assurer que la règle ne peut pas être "due au hasard". Pour cela nous montrons sur un exemple l'effet que peut avoir le hasard sur un jeu de règles, puis nous proposons une technique de simulation de ce hasard aisée à implémenter.
2. Nous faisons apparaître l'effet des relations complexes entre propriétés sur le jeu de règles d'association, et nous proposons de le corriger dans deux directions. Suivant la première, nous rétablissons une relation détruite par le codage binaire habituel, et suivant la seconde, nous supprimons l'effet gênant de certaines relations complexes sur le jeu de règles.
   - Les différentes propriétés a_1, a_2, ..., a_p créées par l'éclatement d'une propriété numérique a lors d'un codage binaire sont liées par des relations particulières qui ne sont pas visibles dans le jeu de règles. Nous rétablissons ces relations en recodant a de façon floue et en créant à la place des diverses règles $a_i \rightarrow b$ une seule règle d'association floue $a \rightarrow b$. Nous construisons les règles d'association floue en "fuzzifiant" toutes les étapes courantes d'extraction des règles.
   - Nous proposons de repérer et nettoyer par des méta-règles de nettoyage les incohérences et redondances d'un jeu de règles d'association dues à certaines relations complexes entre propriétés.

Bien sûr, la dernière contribution n'est pas sensée fournir une structure logique complète à un jeu de règles d'association avec des règles d'inférence valides [102], vu les difficultés que nous avons constatées, mais simplement traduire en termes appropriés au formalisme des règles d'association les incohérences et redonner par élagage une apparence la plus acceptable possible au jeu de règles.

[102] Notre algorithme MIDOVA, actuellement en phase d'achèvement, permet d'extraire des données booléennes un ensemble de motifs ne contenant aucune des incohérences dues aux liaisons complexes particulières que sont les interactions. Il est décrit dans la section 10.3 de la partie "Bilan et perspectives" de ce document. Sa version floue est en cours de développement.

Deuxième partie

Simuler pour valider

5

Une évaluation statistique à base de simulations d'un jeu de règles dérivé de données réelles

Sommaire

Ce chapitre présente un essai d'optimisation du jeu de règles extrait par la technique des motifs fréquents. Partant d'un tableau de données tiré d'un corpus de résumés d'articles scientifiques dans le domaine de la biologie moléculaire, on utilise les techniques usuelles d'extraction de règles d'association pour construire le jeu de règles associé aux données. On définit ensuite des règles «fortuites» par des techniques de simulation. On discute alors du choix de celles qu'il convient de supprimer afin d'optimiser le jeu de règles de départ. Les indices associés à des règles extraites de données s'appuient généralement sur le support et la confiance. On mentionne dans l'article les résultats obtenus avec d'autres indices de qualité utilisés actuellement en fouille de données. Enfin, on se réfère aux propriétés statistiques des données afin de préparer la voie à une optimisation des jeux de règles extraits de bases de données variées, ce qui donne des pistes de prolongement à ce travail.

5.1 Introduction

L'extraction de règles d'association fait partie des techniques de fouille de données visant à construire de la connaissance à partir de bases de données. Même appliquées à de petites bases de données, ces techniques fournissent des jeux de règles de grande taille. La qualité de ces règles est mesurée par des indices, les plus utilisés étant le support et la confiance. En imposant des seuils pour certains de ces indices, on réduit le nombre de règles, en gardant celles de meilleure qualité (selon ces indices). Bien que le choix de ces seuils soit rarement controversé au sein de la communauté des chercheurs en fouille de données (l'un des plus utilisés pour le support est 10 correspondant à 10 objets, quand le nombre total d'objets se compte en centaines, et le plus courant pour la confiance est 80 %), de plus en plus de chercheurs ne s'en contentent pas, constatant qu'ils obtiennent, parmi les règles de support et de confiance élevés, des règles n'apportant aucune connaissance. Après fixation des seuils de support et de confiance, ils essaient d'optimiser le jeu de règles obtenu de diverses façons. Certains (Kodratoff, [142] ; Cherfi et al. [52]) cherchent et utilisent de nouveaux indices de qualité des règles en ne gardant que les règles ayant des valeurs extrêmes pour ceux-ci. D'autres essaient d'améliorer le choix de seuils des indices courants, soit en examinant leur sensibilité à des perturbations des données (Azé et al, [9]), soit en établissant pour ces indices un intervalle de confiance basé sur un modèle probabiliste des données (Teytaud et al [226].

La présente partie vise également l'optimisation d'un jeu de règles qui décrit les liaisons entre les mots-clés d'un corpus de textes. Nous simulons l'indépendance entre ces mots-clés afin d'observer le comportement de divers indices de qualité des règles. Nous établissons notamment des intervalles de confiance des indices sous cette hypothèse H0 d'indépendance, et les mettons en relation avec les valeurs observées des indices des règles obtenues à partir de la base de données réelle. Notre travail a de nombreux points communs avec celui des deux derniers auteurs cités, comme cela est évoqué et discuté plus loin. L'extraction de règles d'association se fait sur des données structurées de telle sorte que l'on puisse directement les représenter sous forme de matrice objets×attributs. Les textes sont des données particulières, non structurées (Han et al.,[113]). Pour leur appliquer cette technique, il faut les munir d'une structure. Nous partons ici d'un corpus de résumés d'articles scientifiques dans le domaine de la biologie moléculaire textes×mots-clés (Cherfi et al., [52]). En premier lieu, nous exposons les principes de l'extraction de règles utilisée couramment en fouille de données, sans entrer dans les détails des algorithmes utilisés (Agrawal et al., [3]), et nous détaillons les caractéristiques du jeu de règles obtenu par ces techniques sur ces données. Ensuite nous décrivons les simulations qui nous permettent d'imputer certaines de ces règles en partie au hasard, règles que nous appelons « règles fortuites », et nous discutons de l'optimisation du jeu de règles par leur suppression. Nous nous focalisons dans cette partie sur deux indices, celui d'implication statistique de Régis Gras [99] et l'indice de confiance. Puis nous terminons par quelques pistes sur une « mise en probabilités » du problème qui permettrait d'obtenir un élagage des règles, de qualité légèrement inférieure à celui obtenu par simulations, mais réalisable en temps réel.

5.2 Le principe d'extraction de règles à partir d'un tableau de données

5.2.1 Le principe

Une relation entre un ensemble d'objets et un ensemble d'attributs étant donnée par un tableau booléen rectangulaire objets×attributs, nous mettons en œuvre la méthode d'extraction de

règles entre attributs qui s'appuie sur les cooccurrences d'attributs pour un nombre suffisamment important d'objets. Elle consiste en les deux étapes suivantes :

- recherche des *motifs fréquents* de longueur supérieure à 1 : ce sont les ensembles d'au moins deux attributs possédés simultanément par un nombre d'objets supérieur à un seuil a donné (on prend habituellement $a = 10$ dès que le nombre d'objets est suffisamment grand, (Bastide, [13])).
- extraction des règles de plus grande confiance : elles sont obtenues par décomposition de chaque motif fréquent en deux parties complémentaires. On ne garde que les règles pour lesquelles la proportion d'objets vérifiant les attributs de la deuxième partie parmi ceux qui vérifient les attributs de la première partie dépasse un seuil b donné (habituellement de 0.80).

5.2.2 Définitions, premières propriétés, notations, exemple

Les mots *attribut* et *propriété* seront employés indifféremment, ainsi que les mots *objet* et *sujet*. On dira qu'un sujet possède une propriété, ou qu'il la vérifie, chaque fois que dans le tableau booléen on trouve un 1 à l'intersection de la ligne et de la colonne correspondant au sujet et à la propriété.
Nous ne rappelons pas ici ce que sont un *sous-motif*, la *longueur*, *le support*, la *fréquence* d'un motif, pas plus que le *support* et la *confiance* d'une règles d'association, ces notions ayant été définies dans le chapitre 4.

5.2.3 Les règles extraites d'un fichier de données

Description du «corpus bio» et de son indexation

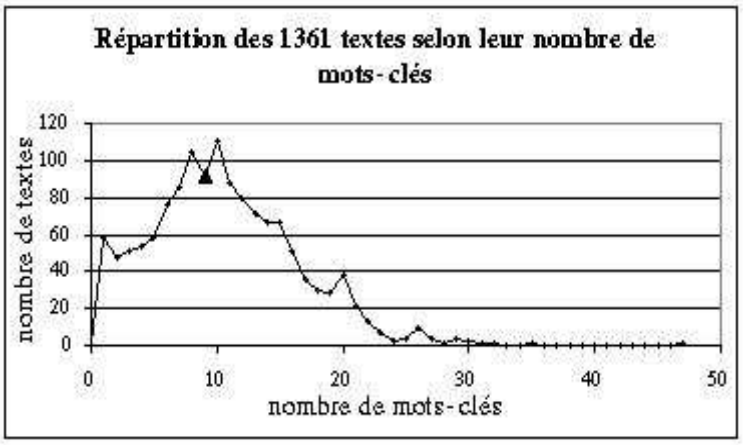


FIG. 5.1 – Les 1 361 textes répartis selon leur nombre de mots-clés.

C'est un ensemble de 1361 textes de biologie, provenant de l'INIST[103]. L'analyse de ces textes par des experts de cette discipline a fourni un certain nombre de mots-clés par texte. L'ensemble des mots-clés du corpus a été revu en plusieurs étapes, afin de ne garder que les plus pertinents

[103]Institut National de l'Information Scientifique et Technique à Vandœuvre-lès-Nancy.

(Cherfi et al., [52]). Le nombre de mots-clés des textes du "corpus bio" varie de 1 à 47. Leur répartition est illustrée[104] en figure 5.1.

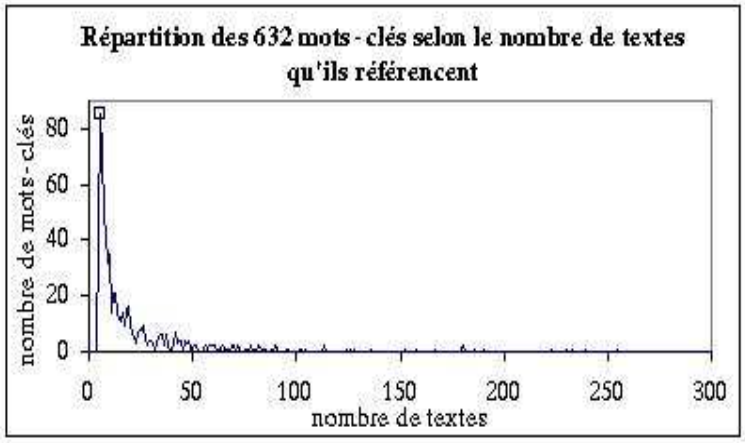


FIG. 5.2 – Les 632 mots-clés répartis selon le nombre de textes qu'ils référencent.

La fréquence des mots-clés de ce corpus est variable. Le plus courant est présent dans 255 textes alors que les moins courants figurent dans un seul texte. Pour limiter le nombre de mots-clés, on n'a gardé que les mots-clés présents dans au moins 5 textes [105]. La figure 5.2 représente la répartition de ces 632 mots-clés[106].

Nombre de motifs fréquents et de règles trouvés

Nous avons pris comme attributs les mots-clés, et comme objets les textes. Un motif est donc une association de mots-clés. Un motif fréquent est un motif que l'on trouve dans au moins 10 textes, ce qui signifie que le seuil de support est 10 et que le seuil de fréquence correspondant est α=10/632.

corpus bio		
longueur	nb de motifs	nb de règles
1	353	
2	1157	2314
3	424	2544
4	86	1204
5	14	420
$\geq$1	2034	6482

TAB. 5.1 – Répartition des motifs de support$\geq$10, et nombre de règles correspondantes

[104]Indication de lecture de la figure 5.1 : le point représenté par un triangle est le point de coordonnées (9 ; 92), ce qui signifie que le nombre de textes ayant exactement 9 mots-clés est de 92.

[105]Ce nombre de mots-clés passera à 353 mots-clés quand on se limitera aux mots-clés qui figurent dans au moins 10 textes, pour construire les motifs fréquents de support au moins égal à 10.

[106]Indication de lecture de la figure 5.2 : le point représenté par un carré est le point de coordonnées (5, 85), ce qui signifie que les mots-clés figurant dans 5 textes seulement sont au nombre de 85.

Les indices de qualité des règles

longueur des motifs	Intervalles de variation de la confiance de type [a ; b [											
	0 ; 0,1	0,1 ; 0,2	0,2 ; 0,3	0,3 ; 0,4	0,4 ; 0,5	0,5 ; 0,6	0,6 ; 0,7	0,7 ; 0,8	0,8 ; 0,9	0,9 ; 1	=1	total
5	19	28	46	82	45	66	37	33	23	14	27	420
4	114	126	146	188	124	148	101	85	71	40	61	1204
3	630	403	335	316	237	173	125	108	82	62	73	2544
2	641	680	455	218	122	67	39	34	21	19	18	2314
≥2	1404	1237	982	804	528	454	302	260	197	135	179	6482

TAB. 5.2 – Répartition de l'indice de confiance des règles selon la longueur des motifs correspondants

On a calculé la confiance de chacune des règles, et dans le tableau 5.2 on a indiqué pour chaque longueur de motif et pour chaque intervalle de confiance le nombre de règles correspondantes. Par exemple, on peut voir que sur les 511 règles de confiance supérieure ou égale à 0,8, 179 sont exactes et 197+135, soit 332, sont partielles. Les similitudes de variation des divers indices, tels que la confiance, la différence, l'intérêt, la nouveauté, la satisfaction, l'étonnement, la conviction, et l'implication peuvent se constater à la lecture des tableaux 5.10, 5.11 et 5.12 en appendice[107]. On se focalisera ici sur la confiance, qui sert habituellement de référence pour l'extraction du jeu de règles.

5.3 Les simulations

Les simulations portent sur le "corpus bio", décrit dans la partie précédente. Elles sont de trois types qui ont en commun l'ensemble des mots-clés avec leurs fréquences respectives fixées, et qui diffèrent par le type d'attribution aléatoire des 1 361 textes à ces mots-clés. Nous exposons d'abord le principe à l'origine de ces simulations. Ensuite nous comparons à partir de quelques simulations les trois types choisis avant d'examiner les résultats obtenus pour 100 simulations de chaque type.

5.3.1 Le principe

Dans le corpus figure un mot-clé présent dans 255 des 1 361 textes. Appelons-le "a" . Et appelons "b" le mot-clé présent dans 55 des 1 361 textes[108]. Supposons l'absence de liens entre ces deux mots : il n'y a a priori aucune raison que la présence de l'un implique la présence ou l'absence de l'autre. Si nous appliquions ici le modèle probabiliste d'indépendance[109] , nous nous attendrions à un résultat d'environ 10.3 textes ($255 \times 55/1361$) contenant le motif "ab", ce qui, avec le seuil de support de 10 des motifs du corpus, ferait de l'association de ces 2 mots-clés un motif fréquent "par hasard". Plutôt que de construire un test d'absence de liens

[107] Les tableaux de l'écart type et des 3 quartiles pourront être consultés par les lecteurs intéressés dans (Cadot et al.[45]).

[108] "a" est l'item 189 du corpus nommé «escherichia coli», "b" est l'item 422, nommé «point mutation», le motif "ab" a un support de 12, la règle "a→b" a une confiance de 0.031 et la règle "b→a" une confiance de 0,218.

[109] On s'appuie, pour ce calcul, sur une hypothèse d'indépendance entre les deux lois marginales d'un tableau de contingence, alors que nous avons un tableau de booléens.

entre mots-clés en modélisant par les probabilités notre grand tableau de booléens textes×mots-clés, nous avons préféré le construire par simulations. Nous générons donc de façon aléatoire des tableaux permettant d'estimer les divers indices des règles en cas d'absence de liens. Nous pouvons alors comparer, pour chaque règle, les indices calculés sur le corpus aux intervalles de confiance correspondants, trouvés par simulations, et en conclure que la règle est ou n'est pas due au hasard, selon que les valeurs calculées sont ou non dans l'intervalle de confiance Pour cela, nous avons généré aléatoirement, de façon identique et indépendante, cent tableaux de booléens ne différant de celui du corpus que par l'absence de liens entre les mots-clés. Ces tableaux booléens sont de même taille que le tableau issu du corpus, et la fréquence de chaque mot-clé est fixée, identique à celle qu'il possède dans le corpus. Pour chaque mot-clé, on tire au hasard, sans répétition, autant de numéros de texte qu'il en a dans le corpus. Nous avons choisi trois façons de le faire qui sont les suivantes :

1. Tirage de type 1 : pour un mot-clé donné, à chaque tirage d'un numéro de texte, tous les numéros ont la même probabilité d'être tirés.
2. Tirage de type 2 : pour un mot-clé donné, à chaque tirage d'un numéro de texte, chaque numéro de texte a une probabilité d'être tiré qui est une fonction croissante de son nombre de mots-clés dans le corpus. On distingue 4 sous-types différant selon le poids p affecté à un texte, qui est fonction de son nombre x de mots-clés. Cette fonction est p(x)=x pour le sous-type 1, p(x)=x+1 pour le sous-type 2, p(x)=x+2 pour le sous-type 3 et p(x)=3 si x<2, p(x)=2 sinon, pour le sous-type 4.
3. Tirage de type 3 : pour le premier mot-clé, chaque numéro de texte a une probabilité d'être tiré proportionnelle à son nombre de mots-clés dans le corpus, puis, chaque fois qu'il est tiré pour un mot-clé, sa probabilité d'être tiré pour les mots-clés suivants diminue afin qu'en fin de simulation, chaque texte ait été tiré autant de fois qu'il a de mots-clés dans le corpus.

les 100 supports du motif "ab" selon le type de simulation						
indices statistiques	type 1	type 2-1	type 2-2	type 2-3	type 2-4	type 3
moyenne	10.41	13.5	12.68	11.88	12.52	12.84
max	17	22	23	18	21	21
min	5	5	3	7	6	6
95ème centile	16	19	18	17	18	17

TAB. 5.3 – Résultats des simulations pour le support du motif "ab"

Pour chaque motif, nous calculons la série des 100 supports, et nous choisissons comme seuil de support de ce motif le 95e centile de la série. Ainsi tout motif du corpus qui dépasse son seuil de support a approximativement moins de 5 % de chances d'être dû au hasard. Nous comptabilisons alors tous les motifs fréquents du corpus qui sont dus au hasard, ou qui contiennent des sous-motifs dus au hasard, et toutes les règles qui en sont déduites. Par exemple, dans le tableau 5.3, on voit que la moyenne issue de la simulation de type 1 est proche, pour le motif "ab" de la valeur que nous avons estimée selon le « modèle probabiliste d'indépendance », alors qu'elle s'en éloigne pour les deux autres types de simulation. On peut remarquer également que selon le type de simulation choisi, le seuil de support du motif "ab" varie de 16 à 19. Comme il est de 12 dans le corpus, on peut juger que c'est un motif fréquent "par hasard", quel que soit le type de simulation choisi.

5.3.2 Les premiers essais

Simulations de type 1

corpus bio	données simulées no1		données simulées no2		données simulées no3	
long. motif	nb motifs	nb regles	nb motifs	nb regles	nb motifs	nb regles
1	353		353		353	
2	569	1138	576	1152	581	1162
3	2	12	3	18	7	42
≥1	924	1150	932	1170	941	1204

TAB. 5.4 – Nombre de motifs et de règles trouvées dans une simulation de type 1.

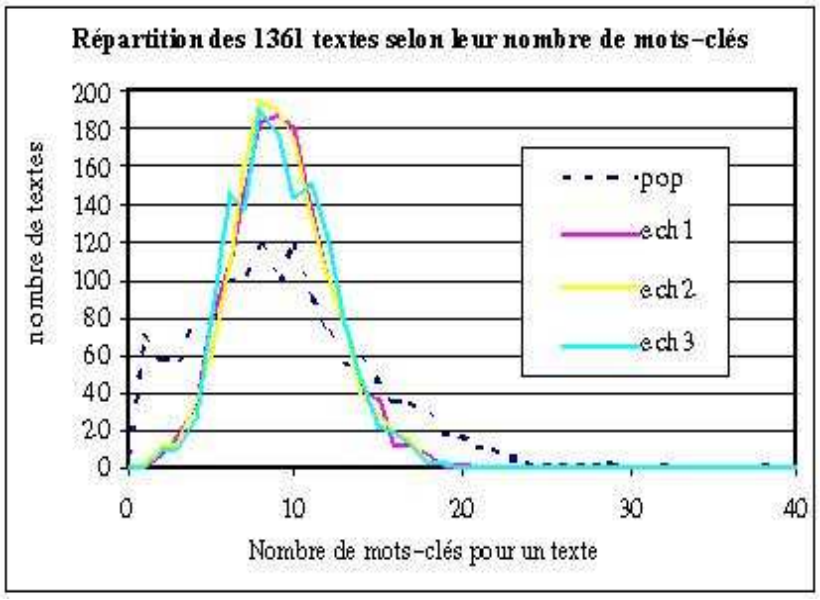


FIG. 5.3 – Histogramme de répartition des textes selon leur nombre de mots-clés pour 3 simulations de type 1.

Nous générons un échantillon de 1361 textes, contenant les 353 mots-clés en respectant exactement la répartition de ceux-ci dans le corpus : par exemple, le mot-clé "a" sera présent dans 255 textes différents dont les numéros ont été tirés au hasard sans remise entre 0 et 1 360, et le mot-clé "b" dans 55 textes différents, tirés au hasard de la même façon, donc indépendamment de ceux où est le mot-clé précédent, et ainsi de suite pour les 353 mots clés. Du tableau booléen obtenu, nous extrayons tous les motifs fréquents, et les règles correspondantes, puis nous recommençons avec 2 autres échantillons. Les résultats figurent dans le tableau 5.4.

Nous avons comparé les supports des motifs du corpus avec les supports des mêmes motifs chaque fois qu'ils apparaissent dans l'un des trois échantillons. 236 motifs des 1 157 motifs fréquents de longueur 2 du corpus ont un support inférieur ou égal à celui qu'ils ont dans l'un des trois échantillons, et apparaissent dans 1 068 règles sur les 6 482 règles du corpus, dont 46 sur les 511 de confiance $\geq 0{,}8$. On examine maintenant la distribution des textes selon leur nombre

de mots-clés, afin de contrôler qu'elle ressemble bien à celle du corpus. Un examen de la figure 5.3 nous montre que le corpus, noté *pop*, et les échantillons notés *ech1, ech2, ech3,* ne sont pas formés de la même façon. Dans cette simulation de type 1, nous n'avons pas pris en compte le fait que certains textes ont plus de mots-clés que d'autres. Nous essayons de corriger cela en organisant une nouvelle simulation.

Simulations de type 2

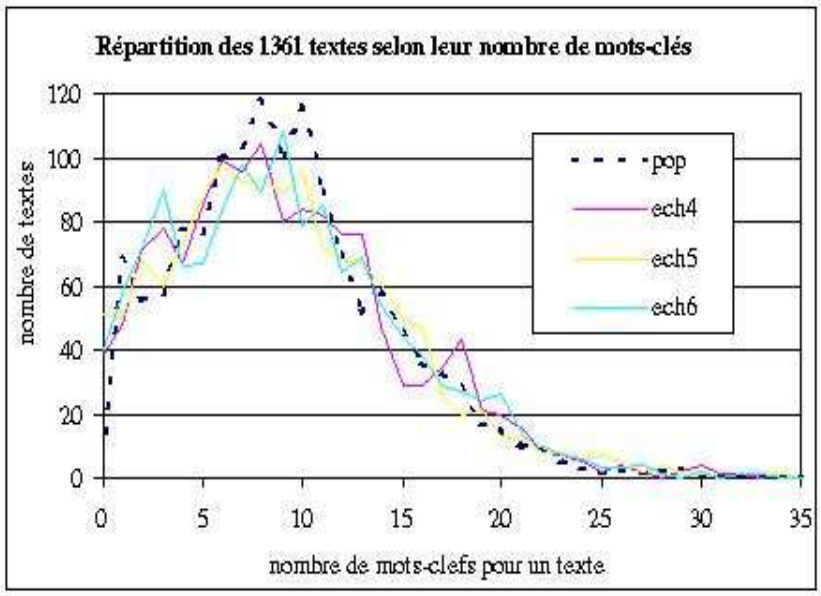


FIG. 5.4 – histogramme de la répartition des textes selon leur nombre de mots-clés pour 3 simulations de type 2.

corpus bio	données simulées no1		données simulées no2		données simulées no3	
long. motif	nb motifs	nb regles	nb motifs	nb regles	nb motifs	nb regles
1	353		353		353	
2	911	1822	966	1932	917	1834
3	104	624	104	624	97	582
≥ 1	1368	2446	1367	2556	1367	2416

TAB. 5.5 – Nombre de motifs et de règles trouvées dans une simulation de type 2.

Nous tirons 3 échantillons, en donnant à chaque numéro de texte une probabilité d'être tiré proportionnelle au nombre de mots-clés qu'il possède dans le corpus. Le tirage d'un numéro de texte se fait toujours sans remise pour un mot-clé donné. Et d'un mot-clé à l'autre, les probabilités de tirer les numéros de textes ne varient pas. Les motifs et règles générés sont représentés dans le tableau 5.5. La lecture des figures 5.4 et 5.5 permet d'apprécier une plus grande ressemblance entre les courbes des simulations et des données de départ que pour la simulation de type 1.

En procédant comme pour les échantillons de type 1, nous trouvons 393 motifs sur les 1 157 motifs fréquents de longueur 2 du corpus qui ont des supports "trop petits", ce qui porte le

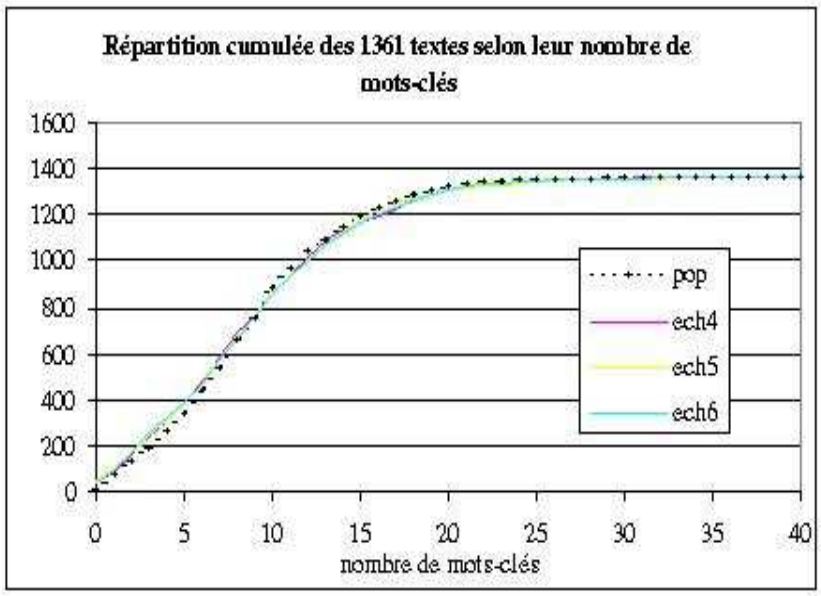


FIG. 5.5 – Fréquences cumulées de la répartition des textes selon leur nombre de mots-clés pour 3 simulations de type 2.

nombre de règles "fortuites" du corpus à 2 510 sur 6 482 dont 158 sur 511 de confiance supérieure ou égale à 0,8.

Un examen attentif de la figure 5.5 montre que quand les 4 courbes ne coïncident pas, les courbes des 3 échantillons débordent toujours du même côté de la courbe de la population. L'explication se trouve dans la figure 5.4, où l'on remarque à gauche du graphique que le nombre de textes n'ayant aucun mot-clé est bien supérieur au nombre correspondant de textes dans la population. Et c'est l'inverse pour les textes contenant une dizaine de mots-clés : la courbe de la population culmine à près de 120 dans cette région, alors qu'elle atteint tout juste 110 pour les échantillons. Nous avons essayé de corriger ces effets en ajoutant un même nombre à tous les poids afin d'accorder une probabilité non nulle aux numéros des textes sans mots-clés d'être tirés au hasard, et les 3 courbes se rapprochent de la courbe du corpus. Toutefois, même en faisant varier le nombre ajouté selon les poids, nous n'arrivons pas à obtenir que la distribution de l'un des 3 échantillons soit plus proche de celle du corpus que de celle des autres échantillons. Nous réalisons alors un troisième type de simulation avec des échantillons respectant pour chaque texte, le nombre de mots-clés, et pour chaque mot-clé, le nombre de textes.

Simulations de type 3

Pour tirer un échantillon selon ce troisième type, nous avons créé une liste où figure chaque numéro de texte dupliqué autant de fois qu'il possède de mots-clés dans le corpus. Puis nous avons tiré au hasard, sans remise dans cette liste, tous les numéros de textes du premier mot-clé, puis du second, et ainsi de suite jusqu'au dernier. Lorsque les 353 mots-clés ont obtenu leurs numéros de textes, on a comptabilisé pour chaque mot-clé les répétitions de numéros de textes, et chaque fois qu'un numéro de textes était répété pour un mot-clé, nous avons échangé ce numéro avec celui d'un autre mot-clé qui était également répété, quand cela ne rajoutait à aucun des deux de nouvelle répétition. Au bout d'une dizaine de manipulations, les quelque 400 répétitions se

réduisaient à environ une dizaine. On procédait alors par échange avec des mots-clés qui n'avaient pas de répétition. Au bout d'une dizaine d'échanges, on venait à bout des répétitions[110].

corpus bio	données simulées no1		données simulées no2		données simulées no3	
long. motif	nb motifs	nb regles	nb motifs	nb regles	nb motifs	nb regles
1	353		353		353	
2	835	1670	841	1682	844	1688
3	75	450	83	498	98	588
$\geq$1	924	2120	932	2180	941	2276

TAB. 5.6 – Nombre de motifs et de règles trouvées dans une simulation de type 3.

Nous ne représentons pas graphiquement la répartition des textes selon leur nombre de mots-clés car il y a identité des 4 courbes, celle des 3 échantillons et celle des données réelles. Nous trouvons alors 348 motifs de supports "trop petits" sur les 1 157 motifs de longueur 2 du corpus bio, ce qui porte le nombre de règles "fortuites" à 1 956 sur 6 482 dont 94 sur 511 de confianc e$\geq$ 0,8.

5.3.3 Les résultats

nombre de motifs trouvés	sur les 100 échantillons			sur le
	type1	type2	type3	corpus
de longueur 2 (excluant1)	2182 (334)	3218 (507)	2937 (475)	1157
de longueur 3 (excluant1, 2)	152 (0+4)	1028 (1+38)	987 (2+35)	424
de longueur 4 (excluant1, 2)	0	12 (0+0)	26 (0+0)	86
de longueur 5 (excluant1, 2)	0	0	0	14
tous (excluant1+2)	2334 (338)	4528 (546)	3950(512)	2034

TAB. 5.7 – Résultats des 3 types de simulations avec 100 échantillons.

Nous réalisons 100 simulations et non 3, et nous ne prenons plus comme seuil de support d'un motif le maximum de l'ensemble des supports de ce motif issu des simulations, mais le 95e centile, ce qui correspond à un risque inférieur à 5 % de rejeter à tort l'effet du hasard dans l'apparition du motif fréquent (Howel, [126]). Chaque fois qu'un motif trouvé par simulation se retrouve dans le corpus avec un support inférieur ou égal au seuil ainsi défini, il est appelé motif "excluant", avec l'indice 1 s'il ne contient pas de sous-motif excluant, et l'indice 2 sinon. Par exemple, si le motif "abc" (resp. abd, ab, bc, bd, ad) a comme support 15 (resp . 17, 30, 25, 28, 32) et comme seuil 16 (resp. 18, 28, 26, 27, 30), "abd" est un motif excluant.1 car aucun de ses sous-motifs n'est excluant, et "abc", contenant le sous-motif excluant "bc" est un motif excluant.2. Le tableau 5.7 récapitule selon les trois types de hasard le nombre de motifs fréquents trouvés sur au moins un échantillon, et entre parenthèses, parmi ceux-ci, ceux qui se sont trouvé excluants. Un résultat appréciable de ces essais est que les motifs excluants obtenus par la simulation de type 1 font tous partie des motifs excluants de la simulation de type 3. Par contre 40 (resp. 6) motifs excluants

[110]Un relecteur anonyme nous propose d'arriver au même résultat par une méthode plus directe consistant à itérer la permutation suivante : choisir deux documents contenant un nombre voisin de mots-clés, choisir deux mots-clés de fréquences comparables, échanger.

de type 2 et de longueur 2 (resp. 3) ne figurent pas parmi les motifs excluants de type 3, et 8 (resp. 4) des motifs excluants de type 3 ne se retrouvent pas parmi ceux de type 2. Toutefois, si le motif est excluant pour l'un de ces 2 types (type 2 ou 3), et non excluant pour l'autre, la différence entre les seuils dûs aux 2 motifs est 1 dans la majorité des cas, atteignant une seule fois 3, et une seule fois 4 sur les 58. Ce qui rend les conclusions tirées des simulations de type 2 et 3 très proches, les différences pouvant être attribuées aux fluctuations d'échantillonnage. En prenant le maximum au lieu du 95e centile, on ne retrouve pas une telle convergence entre les différents types de simulations. Le tableau 5.8 détaille les résultats de la simulation de type 3[111] pour 2 indices, l'indice de confiance et celui d'implication [112]. On voit que par ces simulations, 3 000 des 6 484 règles trouvées sur le corpus ont été jugées "fortuites". La répartition de ces règles selon la longueur du motif dont elles sont issues permet de faire quelques remarques.

corpus bio	hasard : non					le sous-motif de AB dû au hasard est de							Tous
règles A→B						longueur 2				lg>2	long3	Tous	
longueur AB	2	3	4	5	Tous	2	3	4	5	3	3		
Nbre de règles	1364	1248	631	239	3482	950	1284	574	180	12	222	3000	6482
Max conf obs	1,00	1,00	1,00	1,00	1,00	0,36	1,00	1,00	1,00	0,36	0,48	1,00	1,00
Moyenne conf	0,29	0,40	0,49	0,52	0,38	0,14	0,27	0,41	0,47	0,16	0,17	0,27	0,33
Min conf obs	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,04
Max impl	1,00	1,00	1,00	1,00	1,00	0,95	1,00	1,00	1,00	0,81	0,96	1,00	1,00
Moyenne impl	0,69	0,81	0,92	0,92	0,79	0,52	0,67	0,86	0,92	0,56	0,56	0,67	0,74
Min impl	0,00	0,00	0,07	0,13	0,00	0,05	0,06	0,11	0,19	0,32	0,21	0,05	0,00

TAB. 5.8 – Résultats de simulation de type 3 avec 100 échantillons.

Examinons en premier lieu les 4 colonnes de résultats obtenus pour les règles «fortuites» construites sur des motifs de longueur 2. Les 950 règles de longueur 2 qui ont été exclues ont un indice de confiance inférieur à 0,363 et un indice d'implication inférieur à 0,949. Ce qui signifie que ces règles sont éliminées par la suite, qu'on prenne un seuil de confiance de 0,8 (on pourrait même prendre un seuil bien inférieur) ou un seuil d'implication de 0,95. Par contre, pour ce qui est des règles éliminées parce que le motif de longueur supérieure à 2 qui les constitue contient un sous-motif de longueur 2 excluant, elles peuvent prendre tous les niveaux de confiance et d'indice d'implication possibles. Les deux colonnes suivantes concernent les règles fortuites contenant un motif excluant de longueur 3. Ces 37 motifs ont généré uniquement des règles de longueur 3, et pas de longueur supérieure, soit 222 règles en tout (2 de ces motifs ne contiennent pas de sous-motif excluant de longueur 2, et ont généré 12 règles). On voit que ces règles également auraient été éliminées, en prenant un indice de confiance supérieur à 0,8, ou d'implication supérieur à 0,95.

La première conclusion tirée de ces remarques est que le seuillage d'indice paraît une technique appropriée pour rejeter des règles bâties sur des motifs (et non des sous-motifs) dûs au hasard. En particulier, le seuil d'implication de 0,95, construit sur des modèles probabilistes pour un risque de 1ère espèce de 5 %, permet de rejeter toutes ces règles comme dues au hasard. On voit qu'on pourrait prendre un seuil pour la confiance qui aurait le même rôle, mais que 0,8 est trop élevé pour notre corpus, un seuil de 0,5 paraît largement suffisant pour un risque de première

[111] Les tableaux 5.13 et 5.14 en appendice, contiennent les mêmes informations pour les simulations de types 1 et 2.

[112] Les résultats concernant les autres indices tels que les 4 types de support, la différence, l'intérêt, la conviction, l'étonnement, la nouveauté et la satisfaction peuvent être consultés dans (Cadot M. et al.,[45]).

espèce de 5 %. Un coup d'oeil sur le tableau 5.15, figurant en appendice, pour une simulation de type 3 où le 95e centile a été remplacé par le max, donc correspondant à un risque de première espèce inférieur à 1 %, nous conforte dans notre opinion par un seuil d'implication de 0,997, et de confiance de moins de 0,7.
La deuxième remarque est que les règles "fortuites" ne contenant pas de motif excluant, mais des sous-motifs excluants, ne s'éliminent pas directement par ces techniques de seuillage. De plus on peut s'interroger sur la légitimité de leur exclusion. En effet, prenons un motif "abc" non excluant, contenant le sous-motif "ab" excluant. Nous estimons légitime, d'après les remarques précédentes, d'éliminer les règles a→b et b→a. Il peut être gênant de retirer également les 6 règles générées par le motif abc : a→bc, b→ac, c→ab, ab→c, ac→b et bc→a.

Illustrons cela par une règle tirée du corpus et jugée fortuite à la suite d'une simulation de type 3, où l'on a choisi comme seuil de support le 95e centile. La règle "153 ; 401 → 452" de support 12, de confiance 1, a été jugée fortuite car bien que son motif "153 ; 401 ; 452" ne soit pas excluant, un de ses sous-motifs l'est. Voici les supports, et les seuils pour certains, de ses divers motifs et sous-motifs :

- motif "153 ; 401 ; 452"[113], de support 12, avec un seuil de 0
- sous-motif "153 ; 401", de support 12, avec un seuil de 11
- sous-motif "153 ; 452", de support 22 avec un seuil de 23
- sous-motif "401 ; 452", de support 30 avec un seuil de 6
- motifs "153", "401", "452" de supports respectifs 239, 36 et 80

S'il est clair que la règle 401→452 de support 30 et de confiance 0,833 doit être gardée, au vu de la distance qui sépare son support de son seuil, le statut des autres règles est moins net, et nous pensons qu'on pourrait supprimer sans inconvénient la règle exacte "153 ; 401 → 452" . Elle s'appuie sur la règle "401→452", en y ajoutant la règle "153→452", qui provient du motif excluant, donc que nous supprimons sans hésitation, et sur le motif "153 ; 401" qui dans cette simulation n'est pas excluant.

Dans le tableau 5.9 figure le nombre de règles "fortuites", c'est à dire construites sur des sous-motifs excluants, On a également essayé de voir le nombre de règles "fortuites" qu'on peut atteindre en prenant le max au lieu du 95e centile (voir tableau 5.15 en appendice). Il ne croît pas de façon exceptionnelle. Par exemple, pour les simulations de type 3, nous en trouvons 3 934, dont 199 de confiance supérieure ou égale à 0,8.

nombre de règles	rendues "fortuites" par simulations de			du
	type1	type2	type3	corpus
de longueur 2 (de confiance≥0,8)	676 (0)	1014 (0)	950 (0)	2314 (58)
de longueur 3 (de confiance≥0,8)	660 (33)	1272 (62)	1296 (58)	2544 (217)
de longueur 4 (de confiance≥0,8)	224 (31)	420 (62)	574 (70)	1024 (172)
de longueur 5 (de confiance≥0,8)	30 (6)	120 (25)	180 (28)	420 (64)
toutes (de confiance≥0,8)	1590 (68)	2826 (149)	3000 (156)	6482 (511)

TAB. 5.9 – Résultats des 3 types de simulations avec 100 échantillons.

[113]Les items 153, 401, 452 correspondent aux mots-clés respectifs suivants : *dna, parc gene, quinolone.*

5.4 Bilan, discussion et perspectives

Les simulations ont consisté à remplir un tableau d'objets (1 361 textes) et d'attributs (632 mots-clés) de façon aléatoire avec des zéros et des uns, en respectant le nombre total de 0 et de 1 par mot-clé. Pour la simulation de type 1, nous n'avons ajouté aucune contrainte. Pour la simulation de type 2, nous avons exigé que la distribution des textes selon leur nombre de mots-clés soit proche de celle du corpus, et pour la simulation de type 3, nous avons exigé que chaque texte ait exactement le même nombre de mots-clés que le texte correspondant du corpus. Les résultats de ces simulations nous ont permis de définir des motifs excluants, en utilisant le 95e centile, plutôt que le max, moins stable, et d'en déduire des règles fortuites parmi les règles de notre corpus. Le statut de « règle fortuite » nous a été confirmé par l'indice d'implication de Régis Gras pour les règles issues de motifs excluants. Il est moins affirmé pour les règles issues de motifs non excluants, mais contenant un sous-motif excluant. Nous pensons que toutes les règles correspondant au premier cas peuvent être supprimées, alors que dans le second cas, il nous semble raisonnable de n'en supprimer qu'une partie, selon des métarègles à établir, qui pourraient être une première extension de ce travail. La simulation de type 3 nous paraît la plus sûre, car c'est celle qui « colle » le plus aux données. La simulation de type 2 est productive de façon équivalente[114], tout en exigeant une programmation moins lourde que la simulation de type 3. Et la simulation de type 1 est la moins productive, mais la plus facile à mettre en œuvre. De plus, elle paraît fournir des résultats solides, puisque toutes les règles fortuites selon cette dernière font partie des règles fortuites selon la simulation de type 3, simulation que nous avons qualifiée de « sûre ». Lors de ces simulations, nous avons remarqué que l'utilisation d'un seuil de confiance pour éliminer des règles paraissait appropriée à certaines règles, mais que la valeur de 0,8 était certainement surévaluée. On pourrait envisager comme continuation de ce travail, la recherche de seuils de confiance plus appropriés, par des techniques permettant de calculer le risque de première espèce en contrôlant le risque de deuxième espèce (Howel, [126]), qui est de considérer à tort qu'un motif est fréquent par "hasard". Ce travail d'optimisation peut être reconduit à l'identique sur d'autres tableaux booléens de données. On peut aussi essayer de produire des résultats qui seraient valables pour tous les tableaux booléens ayant des caractéristiques communes. Les répartitions des mots-clés dans les textes peuvent se réduire à quelques paramètres statistiques, dès lors qu'elles ne s'éloignent pas trop d'une loi statistique donnée.

Par exemple, dans les figures 5.6 et 5.7, nous avons essayé de mesurer l'adéquation de la loi de répartition des mots-clés de notre corpus à 2 types de lois, une loi lognormale (Armatte, [8]), et une loi "Zipf-like" (Breslau, [32]) où la valeur du paramètre a ne serait pas 1, comme pour la loi d'Estoup-Zipf (Guiraud, [108]), mais un nombre entre 1,28 et 1,38. Par un travail de statisticien, il devrait être possible d'établir des seuils de l'indice de confiance, mais peut-être également des seuils pour les autres indices tels que la différence, l'intérêt, la conviction, l'étonnement, la nouveauté, la satisfaction pour discriminer les bonnes règles sans recommencer les simulations.

5.5 Reflexions sur des travaux similaires

Le type de règles que nous extrayons fait partie d'un de ceux étudiés dans (Teytaud et al. [226]), leurs parties gauche et droite étant des conjonctions d'assertions de la forme « x possède

[114] La simulation de type 2 fournit 2 826 règles fortuites dont 1 248 (1 014+234, voir le tableau 5.14) règles à supprimer de façon certaine contre 3000 règles fortuites dont 1 172 (950+222, voir tableau 5.7) à supprimer de façon certaine, pour la simulation de type 3 et 1 590 fortuites dont 700 (676+24 voir tableau 5.13) à supprimer pour la simulation de type 1.

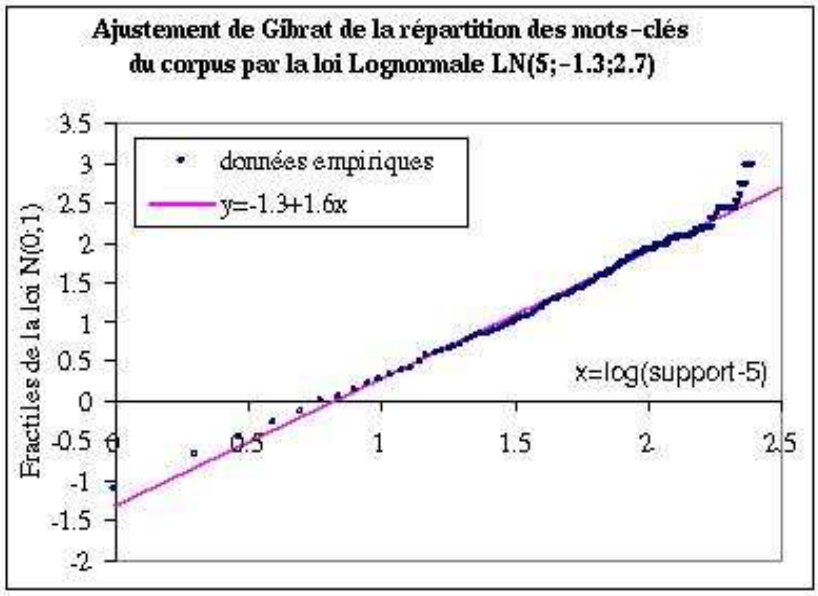


FIG. 5.6 – La loi Lognormale LN(5 ;-1,3 ;2,7) pourrait être une bonne approximation de la distribution des mots-clés.

l'attribut a ». Ces auteurs trouvent des intervalles de confiance théoriques des indices de qualité des règles, en utilisant notamment la VC-dimension de l'ensemble des règles à tester. Ils proposent également, compte tenu de la difficulté d'obtenir des intervalles de confiance suffisamment petits par ces techniques, d'utiliser les techniques de *bootstrap*, qui donnent des intervalles plus petits, bien qu'asymptotiques. Notre travail consistant en partie à chercher des intervalles de confiance des indices de qualité des règles, il peut être vu comme une implémentation possible du travail de ces auteurs. Toutefois, le modèle probabiliste utilisé par ces auteurs ne prend pas en compte l'hypothèse d'indépendance entre les attributs qui est au coeur de notre travail. C'est dans le but de tester celle-ci que, par simulation, nous estimons les indices des règles obtenues en cas d'indépendance, sans disposer d'une estimation ponctuelle préalable. Pour ce qui est de la technique de simulation, notre travail est proche de celui de (Azé et al [9]) . Comme ces auteurs, nous changeons à cent reprises de façon aléatoire, indépendante et identique, des 0 et des 1 de la matrice objets×attributs des données, puis nous examinons les niveaux d'indices des règles qui apparaissent et disparaissent. Leurs critères de changement ne sont pas les mêmes que les nôtres, mais la différence essentielle entre notre travail et celui de ces auteurs porte encore sur l'hypothèse d'indépendance. En cela, notre travail est plus proche d'un travail portant sur des arbres de décision (Oates et al, [187]). Ayant fait la constatation que la taille des arbres de décision augmente avec la taille de la base de données, même si la structure de celle-ci reste inchangée, ces auteurs remettent en cause une technique d'élagage (C4.5) de certaines branches qui consiste à ne pas élaguer dès qu'un indice de qualité de l'arbre de décision est plus petit sur l'arbre élagué que sur l'arbre non élagué. Ils prennent comme hypothèse nulle le fait qu'il n'y a pas à garder de sous-arbre pour le noeud, hypothèse qu'ils opérationnalisent en indépendance entre la variable de décision et les variables explicatives, ces variables étant restreintes aux objets du noeud considéré. Ils construisent alors par simulation un intervalle de confiance de l'indice de qualité de l'arbre de décision en répétant des permutations aléatoires des valeurs de la variable de décision pour les objets du noeud. Et ils décident d'élaguer si l'indice de qualité pour ce noeud

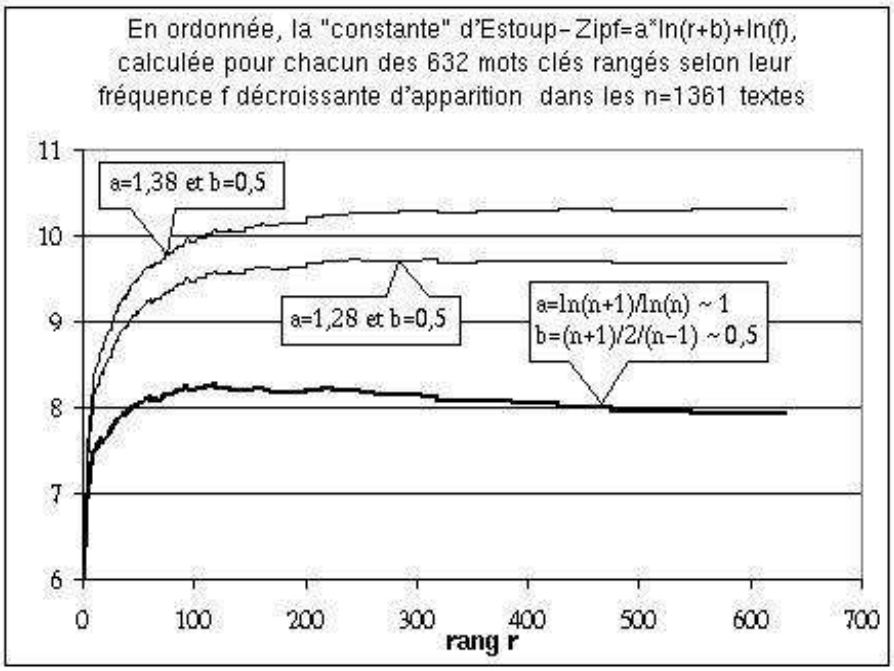


FIG. 5.7 – La répartition des mots-clés dans les textes ne suit pas exactement une loi d'Estoup-Zip (a=1), mais plutôt une loi "Zipf-like" ($1,28 \leq a \leq 1,38$).

est dans l'intervalle de confiance estimé. L'utilisation de l'hypothèse nulle faite pas ces auteurs est la même que la nôtre, bien que leurs simulations se répètent pour chaque noeud, alors que nous les réalisons une seule fois pour la base de données complète. Ils utilisent une technique de simulation par permutation d'une colonne, celle de la variable de décision. C'est également une technique de permutation, mais portant sur toutes les colonnes que nous utilisons dans notre troisième type de simulation.

5.6 Conclusion

Nous nous sommes essayée à montrer qu'il est possible d'optimiser l'extraction de règles en tenant compte des caractéristiques statistiques des données. Nous nous sommes appuyée sur celles-ci pour construire par simulations des motifs "fréquents par hasard" afin d'étudier le comportement de certains indices de qualité des règles, essentiellement la confiance et l'indice d'implication (Gras et al. [102]. Nous avons opérationnalisé l'hypothèse d'indépendance entre attributs de notre base de données de trois façons qui produisent des résultats voisins. Ces résultats nous ont permis de confirmer notre intuition, à savoir :

- de nombreuses règles sont dues en partie au hasard, et leur confiance peut être élevée, et même atteindre la valeur 1 [115].
- de nombreuses règles ne sont pas dues au hasard et leur confiance peut être faible.

Ainsi, si on fixe un seuil de confiance arbitraire pour ne garder que les règles le dépassant, on risque de rejeter des règles qui ne sont pas dues au hasard, et de garder des règles dues en partie

[115] Toutefois pour les règles construites sur des motifs de longueur 2, leur confiance maximum est de 0,36 dans nos simulations.

au hasard[116] .

Avec notre technique de simulation, on peut éliminer les seules règles dues au hasard.

Cette technique peut encore être optimisée selon un critère prenant en compte non seulement la pertinence des règles, mais également les ressources matérielles (temps de calcul et utilisation de mémoire vive). Nous avons exposé quelques éléments qui devraient permettre cette optimisation s'appuyant sur des principes probabilistes.

5.7 Appendice

max	supA B	sup A	sup B	supA, nB	conf obs	diffé rence	inte ret	convi ction	étonn ement	nouve aute	satisF action	impli cation
tous	0,060	0,187	0,187	0,18	1,0	0,991	113,4	33,9	0,944	0,05	1,0	1,0
long2	0,060	0,187	0,187	0,18	1,0	0,991	113,4	24,2	0,944	0,05	1,0	1,0
long3	0,026	0,187	0,187	0,18	1,0	0,991	113,4	33,9	0,882	0,02	1,0	1,0
long4	0,015	0,187	0,187	0,18	1,0	0,988	80,06	20	0,706	0,01	1,0	1,0
long5	0,010	0,176	0,176	0,168	1,0	0,978	65,12	13,7	0,474	0,01	1,0	1,0

TAB. 5.10 – Maximum de chaque indice pour l'ensemble des règles correspondant à un motif de longueur indiquée

min	supA B	sup A	sup B	supA, nB	conf obs	diffé rence	inte ret	convi ction	étonn ement	nouve aute	satisF action	impli cation
tous	0,007	0,007	0,007	0,0	0,039	-0,11	0,339	0,88	-21,3	0,0	-0,10	0,00
long2	0,007	0,007	0,007	0,0	0,039	-0,11	0,339	0,88	-21,3	0,0	-0,10	0,00
long3	0,007	0,007	0,007	0,0	0,039	-0,03	0,811	0,96	-19,2	0,0	0	0,00
long4	0,007	0,007	0,007	0,0	0,039	0,016	1,631	1,02	-20,3	0,0	0,02	0,07
long5	0,007	0,007	0,007	0,0	0,042	0,027	2,847	1,03	-18,5	0,0	0,03	0,13

TAB. 5.11 – Minimum de chaque indice pour l'ensemble des règles correspondant à un motif de longueur indiquée

[116] Toutefois si on se limite aux règles construites sur des motifs de longueur 2, d'après nos simulations, un seuil de confiance supérieur à 0,36 permettrait d'éliminer toutes celles dues au hasard, mais en éliminerait également d'autres. On est loin du seuil de confiance de 0.8 souvent utilisé !

moyenne	supAB	sup A	sup B	supA, nB	conf obs	différence	interet	conviction	étonnement	nouveauté	satisFaction	implication
tous	0,010	0,062	0,062	0,052	0,328	0,266	9,987	1,83	-1,72	0,01	0,28	0,74
long2	0,012	0,084	0,084	0,072	0,226	0,142	4,314	1,38	-1,47	0,01	0,15	0,62
long3	0,010	0,061	0,061	0,051	0,334	0,274	8,877	1,89	-2,13	0,01	0,29	0,74
long4	0,009	0,036	0,036	0,027	0,454	0,418	18,08	2,39	-1,56	0,01	0,44	0,89
long5	0,008	0,026	0,026	0,018	0,496	0,47	24,76	2,47	-1,07	0,01	0,49	0,94

TAB. 5.12 – Moyenne de chaque indice pour l'ensemble des règles correspondant à un motif de longueur indiquée

simulations de type 1	hasard : non					le sous-motif de AB dû au hasard						Tous
						est de longueur 2				long3	Tous	
longueur AB	2	3	4	5	Tous	2	3	4	5	3		
nombre de règles	1638	1884	980	390	4892	676	660	224	30	24	1590	6482
Max conf obs	1,00	1,00	1,00	1,00	1,00	0,31	1,00	1,00	1,00	0,42	1,00	1,00
Moyenne conf obs	0,27	0,36	0,47	0,50	0,36	0,12	0,26	0,40	0,45	0,14	0,23	0,33
Min conf obs	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,04
Max impl	1,00	1,00	1,00	1,00	1,00	0,95	1,00	1,00	1,00	0,89	1,00	1,00
Moyenne impl	0,66	0,77	0,90	0,94	0,77	0,52	0,53	0,86	0,92	0,53	0,67	0,74
Min impl	0,00	0,00	0,07	0,13	0,00	0,21	0,34	0,11	0,19	0,34	0,05	0,00

TAB. 5.13 – Résultats de simulations de type 1 avec 100 échantillons, seuil :95e centile.

simulations de type 2	hasard : non					le sous-motif de AB dû au hasard est de							Tous
						longueur 2				lg≠2	lg=3	Tous	
longueur AB	2	3	4	5	Tous	2	3	4	5	3	3		
NB de règles	1300	1272	784	300	3656	1014	1266	420	120	6	234	2826	6482
Max conf obs	1,00	1,00	1,00	1,00	1,00	0,39	1,00	1,00	1,00	0,28	0,48	1,00	1,00
Moyenne conf	0,30	0,40	0,47	0,00	0,39	0,14	0,27	0,42	0,48	0,15	0,17	0,25	0,33
Min conf obs	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,05	0,04	0,04	0,04
Max impl	1,00	1,00	1,00	1,00	1,00	0,95	1,00	1,00	1,00	0,77	0,95	1,00	1,00
Moyenne impl	0,69	0,81	0,92	0,95	0,79	0,52	0,67	0,86	0,92	0,55	0,56	0,67	0,74
Min impl	0,00	0,00	0,07	0,13	0,00	0,08	0,06	0,17	0,19	0,35	0,21	0,05	0,00

TAB. 5.14 – Résultats de simulations de type 2 avec 100 échantillons, seuil :95e centile.

simulations de type 3	hasard : non					le sous-motif de AB dû au hasard						Tous
						est de longueur 2				long3	Tous	
longueur AB	2	3	4	5	Tous	2	3	4	5	3		
NB de règles	1034	930	434	150	2548	1280	1614	770	270	414	3934	6482
Max conf obs	1,00	1,00	1,00	1,00	1,00	0,68	1,00	1,00	1,00	0,67	1,00	1,00
Moyenne conf	0,32	0,43	0,51	0,54	0,41	0,15	0,28	0,42	0,47	0,19	0,28	0,33
Min conf obs	0,04	0,04	0,04	0,05	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,04
Max impl	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,99	1,00	1,00
Moyenne impl	0,72	0,84	0,93	0,96	0,81	0,53	0,68	0,87	0,92	0,57	0,69	0,74
Min impl	0,00	0,00	0,07	0,13	0,00	0,01	0,03	0,07	0,13	0,16	0,01	0,00

TAB. 5.15 – Résultats de simulation de type 3 avec 100 échantillons, seuil : max.

6

L'échange en cascade pour valider motifs et règles

Dans le chapitre précédent, nous avons établi sur un exemple la nécessité de prise en compte des lois de distribution des données pour l'extraction des règles d'association les plus pertinentes. Par des simulations nous avons en effet constaté que même dans le cas où les propriétés n'ont pas d'autres liens que ceux dus à leurs distributions respectives (par exemple deux propriétés courantes se rencontrent plus souvent chez les même sujets que deux propriétés rares), elles peuvent être à l'origine de règles d'association ayant de bons indices de qualité. De telles règles d'association sont donc dues au hasard et doivent être éliminées. Pour faire ces simulations, nous avons privilégié une méthode de simulation (appelée méthode de type 3) qui est "orientée contexte" dans la mesure où le nombre de propriétés vérifiées par chaque sujet est conservé, ainsi que le nombre de sujets vérifiant chaque propriété. Ce chapitre est une formalisation de cette méthode de simulation permettant d'obtenir les données simulées à partir des données réelles.

Sommaire

6.1 Introduction

Un tableau binaire sujets×propriétés peut être vu comme un ensemble de p colonnes juxtaposées représentant les propriétés, chacune comportant n valeurs égales à 0 ou 1. Le total de chaque colonne est le nombre de sujets vérifiant la propriété. Si ce total est élevé, cela signifie que c'est une propriété courante chez les sujets considérés, et s'il est faible, qu'il s'agit d'une propriété rare. Trouver une règle d'association A–>B entre les propriétés A et B signifie qu'un certain nombre de sujets vérifiant B vérifient également A. La qualité de la règle est mesurée par de nombreux indices mettant en jeu le nombre total de sujets, le nombre de sujets vérifiant A, de ceux vérifiant B, et de ceux vérifiant A et B. Si on suppose fixé le total des deux colonnes, la qualité de la règle ne dépend plus que du nombre de sujets vérifiant les deux propriétés, donc de la façon dont les valeurs 1 des deux colonnes sont associés. Si les deux propriétés sont fréquentes, par exemple chez les 2/3 des sujets pour chaque propriété, la proportion de sujets qui vérifient simultanément les deux propriétés peut varier entre 1/3 et 2/3, et en cas d'absence de lien on attend sous certaines conditions, une proportion de 4/9 (carré de 2/3). Toutefois la proportion observée en cas d'absence de lien diffère souvent plus ou moins de la proportion attendue, la différence étant due au hasard, comme la différence entre le nombre de piles et de faces d'une pièce jetée en l'air plusieurs fois. On peut établir par le calcul des probabilités un intervalle de confiance à 95% autour de la proportion attendue, et ainsi rejeter l'effet du hasard pour les proportions en dehors de cet intervalle en prenant un risque réduit de se tromper (pas plus de 5% des cas par exemple). On peut également l'établir par simulations en permutant au hasard et de façon indépendante les colonnes de A et de B un grand nombre de fois, et en calculant à chaque fois la proportion de sujets vérifiant les deux propriétés, l'intervalle de confiance étant extrait de l'ensemble des proportions trouvées. C'est cette seconde voie que nous avons choisie, car elle nous a semblé plus facile à généraliser au cas où les totaux par lignes sont également fixés. Toutefois, la première généralisation que nous avons faite, qui consistait à créer un tableau en permutant toutes les colonnes du tableau de départ, et à le garder si les totaux par lignes étaient ceux attendus (procédure d'acceptation/rejet) n'a pas convenu. En effet le pourcentage de tableaux gardés devenait tellement petit dès que la taille du tableau de départ augmentait qu'après des heures de calcul, il pouvait ne pas y en avoir alors qu'il nous en faut au moins une centaine. Nous proposons donc une autre méthode qui consiste à choisir au hasard une transformation du tableau de départ gardant rigoureusement inchangés les totaux de lignes et de colonnes. Et nous répétons ce choix autant de fois que nous désirons de tableaux dûs au hasard. Nous montrons ci-dessous que toute transformation de ce type peut s'exprimer comme composition finie d'"échanges rectangulaires", puis nous examinons les résultats des simulations faites selon ces principes sur deux petits exemples. Nous concluons par un bilan sur ce que peut apporter cette méthode, et donnons quelques pistes d'améliorations possibles.

6.2 Des tableaux aux matrices

Dans toute cette partie, ce qu'on appelle matrice[117] est un tableau à n lignes et p colonnes de valeurs 0 ou 1, n et p étant deux nombres entiers supérieurs à 1. On appelle *marges* de la matrice la colonne (*colonne marginale*) des totaux de chaque ligne et la ligne (*ligne marginale*) des totaux de chaque colonne. Le total général s'écrit à l'intersection de la ligne et de la colonne marginales. On appelle *sommes marginales* les nombres contenus dans les marges. Les marges

[117]Dans tout ce chapitre, on dira matrice au lieu de matrice booléenne dans la mesure où toutes les matrices de ce chapitre sont booléennes.

ne font pas partie de la matrice. Voici dans le tableau 1 un exemple de matrice pour n=4 et p=3. Elle correspond à un tableau sujets×propriétés de 3 propriétés A, B, C et 4 sujets s1, s2, s3, s4. La matrice est le tableau de 12 valeurs situé au milieu et entouré d'un double trait. La colonne marginale est à droite et contient les totaux des lignes, la ligne marginale est en bas et contient les totaux des colonnes, et la dernière case en bas à droite contient le total général, donc le nombre de valeurs égales à 1 de la matrice.

Sujet	A	B	C	Total
s1	0	1	1	2
s2	1	0	0	1
s3	1	1	1	3
s4	1	0	1	2
Total	3	2	3	8

TAB. 6.1 – Un tableau de 3 propriétés et 4 sujets, la matrice associée et ses marges

6.3 Premières définitions

Définition 6.3.1. *Matrices S-équivalentes.*
On dit que deux matrices sont S-équivalentes si elles ont mêmes sommes marginales.

Par exemple, les matrices du tableau 2 sont S-équivalentes à celles du tableau 1.

1	0	1
0	1	0
1	1	1
1	0	1

1	1	0
0	0	1
1	1	1
1	0	1

1	0	1
1	0	0
1	1	1
0	1	1

1	0	1
0	0	1
1	1	1
1	1	0

TAB. 6.2 – Des matrices S-équivalentes à celle du tableau 1

Propriété 6.3.1. *La S-équivalence est une relation d'équivalence.*
Cette relation définie dans l'ensemble des matrices est réflexive, symétrique, transitive .

Preuve. *La preuve découle immédiatement de la définition.*

Ayant obtenu une relation d'équivalence, il convient maintenant de nous donner les moyens de trouver la classe d'équivalence de chaque matrice. Par exemple la classe d'équivalence de la matrice du tableau 1 contient 5 matrices qui sont la matrice d'origine et les quatre matrices du tableau 2. Dans le paragraphe suivant, nous allons définir une classe de transformations permettant de passer d'une matrice à une matrice S-équivalente, et établir qu'avec ces transformations nous générons bien tous les éléments de la classe d'équivalence d'une matrice. Puis nous en donnerons une expression plus simple du point de vue algorithmique qui nous permettra d'engendrer un grand nombre de ces matrices selon une répartition proche du hasard.

6.4 Transformations opérant sur la classe d'équivalence d'une matrice

6.4.1 Une transformation simple : l'échange rectangulaire

Si dans une matrice, on échange un 0 et un 1 d'une ligne, donc les valeurs sur deux colonnes, la somme sur les lignes ne change pas, mais la somme sur les 2 colonnes concernées change. Toutefois, si sur une autre ligne, on peut faire l'échange inverse sur les colonnes, les sommes sur ces colonnes redeviennent les sommes initiales. Cela signifie qu'on peut dessiner un rectangle dans la matrice pour lequel deux sommets consécutifs ont deux valeurs distinctes 0 et 1. Nous appelons "échange rectangulaire" la transformation qui consiste à changer ces 4 valeurs en leurs compléments à 1.

Définition 6.4.1. *Échange rectangulaire*
A étant une matrice, i_1 et i_2 deux valeurs distinctes de $I = \{0, 1, ..., n-1\}$ et j_1 et j_2 deux valeurs distinctes de $J = \{0, 1, ..., p-1\}$ telles que $A(i_1, j_1) = A(i_2, j_2) = 1 - A(i_1, j_2) = 1 - A(i_2, j_1)$ on appelle échange rectangulaire $E_{(i_1,i_2),(j_1,j_2)}$ la transformation de A en B qui échange les 4 valeurs à l'intersection des deux lignes et des deux colonnes :

$$B(i,j) = \begin{cases} 1 - A(i,j) & si\ (i,j) \in \{(i_1,j_1),(i_1,j_2),(i_2,j_1),(i_2,j_2)\} \\ A(i,j) & si\ (i,j) \in I \times J - \{(i_1,j_1),(i_1,j_2),(i_2,j_1),(i_2,j_2)\} \end{cases}$$

Par exemple, prenons comme matrice A la matrice du tableau 1. On a $I = \{0, 1, 2, 3\}$, $J = \{0, 1, 2\}$. On peut dessiner sur cette matrice 3 rectangles pour lesquels les valeurs 0 et 1 alternent. Cela nous donne les échanges rectangulaires $E_{(0,1),(0,1)}$, $E_{(0,1),(0,2)}$ et $E_{(0,3),(0,1)}$. Les 3 premières matrices du tableau 2 sont les résultats des transformations correspondantes de la matrice A.

	Colonnes		
	0	1	**2**
Ligne 0	**0**	1	**1**
Ligne 1	**1**	0	**0**
Ligne 2	1	1	1
Ligne 3	1	0	1

Colonnes		
0	1	**2**
1	1	**0**
0	0	**1**
1	1	1
1	0	1

TAB. 6.3 – Passage de la matrice A à la matrice B par l'échange rectangulaire $E_{(0,1),(0,2)}$

Détaillons le deuxième de ces échanges $E_{(0,1),(0,2)}$. Le premier couple (0,1) indique qu'on prend les lignes 0 et 1, qui sont les deux premières lignes. Le second couple (0,2) indique qu'on prend la première et la dernière colonne. On considère ainsi les quatre valeurs 0, 1, 0, 1 indiquées en gras dans la matrice de gauche du tableau 3, leurs coordonnées respectives (0,0), (0,2), (1,0), (1,2) formant un rectangle. Lors de la transformation, chacune de ces quatre valeurs est remplacée par son complément à 1, pour former la matrice à droite du tableau 3, les autres restant inchangées.

Propriété 6.4.1. *La transformée B d'une matrice A par un échange rectangulaire est une matrice S-équivalente.*

Preuve. *Soit E un échange rectangulaire opérant sur les deux lignes d'indices i_1 et i_2 et sur les deux colonnes d'indices j_1 et j_2.*

1. *Montrons d'abord que les sommes des lignes n'ont pas changé lors de la transformation E.*
   - *Pour chaque indice i différent de i_1 et i_2, la ligne n'a pas changé, donc la somme de ses valeurs non plus.*
   - *Dans la ligne d'indice i_1, deux termes ont changé, ce sont les termes des colonnes j_1 et j_2. Si l'échange a pu avoir lieu, c'est que l'un de ces 2 termes avait pour valeur 0 et l'autre pour valeur 1 avant l'échange. Leur somme était donc 1 avant l'échange. Après l'échange le terme de valeur 1 a pris comme nouvelle valeur 0, et inversement pour l'autre terme. Leur somme est toujours 1. La somme totale de la ligne i_1 n'a donc pas changé.*
   - *On ferait le même raisonnement pour la ligne i_2.*

   On a ainsi montré que les sommes des lignes de la matrice avant et après l'échange rectangulaire sont identiques.
2. *On montrerait de la même façon que les sommes des colonnes ne changent pas non plus lors de la transformation.*

La matrice B est donc S-équivalente à la matrice A.

Les échanges de zéros et de uns selon un rectangle permettent, nous venons de le voir, de garder les marges inchangées. Ce ne sont pas les seuls, les échanges pouvant aussi se faire en cascade. En effet, une fois deux valeurs 0 et 1 échangées sur une ligne, on a vu précédemment que les sommes sur les deux colonnes avaient changé et devaient être rétablies. Dans l'échange rectangulaire, on utilise une deuxième ligne avec ces mêmes colonnes. Mais on peut très bien utiliser pour cela une troisième colonne et deux autres lignes. On rétablira la somme de la première colonne à l'aide de la deuxième ligne et de la troisième colonne, et la somme de la deuxième colonne à l'aide de la troisième ligne et de la troisième colonne. Du coup la troisième colonne, ayant été modifiée deux fois, aura sa somme inchangée. Et on peut procéder ainsi de façon plus générale avec q colonnes et q lignes.

6.4.2 Une transformation plus complexe : l'échange en cascade

Comme pour l'échange rectangulaire, l'échange en cascade ne peut se faire sur une matrice que si les valeurs 0 et 1 de la matrice sur les lignes et colonnes considérées se succèdent correctement. C'est cette disposition particulière que nous appelons "suite diagonale".

Définition 6.4.2. *Suite diagonale*
On appelle suite diagonale d'une matrice A un triplet (I, J, val) où $I = (i_1, i_2, ..., i_q)$ est une suite ordonnée d'indices de lignes, $J = (j_1, j_2, ..., j_r)$ une suite ordonnée d'indices de colonnes, val une valeur égale à 0 ou 1, tels qu'on ait la suite d'égalités suivante :

$$\begin{array}{ll} A(i_1, j_1) = val & A(i_1, j_2) = 1 - val \\ A(i_2, j_2) = val & A(i_2, j_3) = 1 - val \\ & \dots \\ A(i_q, j_r) = val & A(i_q, j_1) = 1 - val \end{array}$$

La suite I n'a pas d'éléments répétés, pas plus que la suite J. Ce sont toutes deux des suites d'indices distincts.

Propriété 6.4.2. *Dans une suite diagonale (I,J,val), les suites d'indices I et J ont le même nombre d'éléments. Ce nombre est appelé longueur de la suite.*

Preuve. *Cela vient du fait que la suite s'écrit avec des égalités doubles $A(i_k, j_k) = val$ et $A(i_k, j_{k+1}) = 1 - val$, où les valeurs de k se succèdent de 1 en 1 modulo q (ou r) donc on a $q = r$.*

Une suite diagonale a donc une longueur comprise entre 2 et min(n,p), n et p étant les dimensions de la matrice. On l'a appelée suite diagonale car si on réordonne les lignes et les colonnes de la matrice selon les deux ensembles d'indices, on obtient une bande de largeur 2 sur la diagonale de la matrice. Et les extrémités de cette bande se recollent selon un tore.

Définition 6.4.3. *Échange en cascade*
On appelle échange en cascade sur une matrice A la transformation associée à une suite diagonale sur A qui change toutes les valeurs de A concernées par la suite diagonale. On appelle longueur de l'échange en cascade la longueur de la suite diagonale associée.

Propriété 6.4.3. *Échange en cascade de longueur 2*
C'est un échange rectangulaire.

Preuve. *La preuve est immédiate :*
Quand la longueur de la suite est 2, il n'y a plus que 4 équations qui sont :
$A(i_1, j_1) = val \quad A(i_1, j_2) = 1 - val$
$A(i_2, j_2) = val \quad A(i_2, j_1) = 1 - val$
la suite diagonale se réduit à un rectangle où les valeurs des sommets alternent, donc l'échange en cascade associé est l'échange rectangulaire $E_{(i_1,i_2),(j_1,j_2)}$.

Par exemple, la matrice du tableau 1 ne peut contenir que des suites diagonales de longueur 2 ou 3. Nous avons vu que les trois matrices de gauche du tableau 2 sont obtenues chacune à partir de la matrice du tableau 1 par un échange rectangulaire. Ces trois échanges rectangulaires $E_{(0,1),(0,1)}$, $E_{(0,1),(0,2)}$ et $E_{(0,3),(0,1)}$ sont les échanges en cascade associés aux suites diagonales de longueur 2 suivantes :((0,1),(0,1),0), ((0,1),(0,2),0), ((0,3),(0,1),0).

	Colonnes		
	0	**1**	**2**
Ligne 0	**0**	**1**	1
Ligne 1	**1**	0	**0**
Ligne 2	1	1	1
Ligne 3	1	**0**	**1**

	Colonnes		
	0	**1**	**2**
Ligne 0	**1**	**0**	1
Ligne 1	**0**	0	**1**
Ligne 2	1	1	1
Ligne 3	1	**1**	**0**

TAB. 6.4 – Passage de la matrice A à la matrice B par l'échange en cascade $E_{((0,1,3),(0,2,1),0)}$

L'élément de droite du tableau 2 est produit par l'échange en cascade associé à une suite diagonale de longueur 3 qui est ((0,3,1),(0,1,2),0) car :
$A(0, 0) = 0 \quad A(0, 1) = 1$
$A(3, 1) = 0 \quad A(3, 2) = 1$
$A(1, 2) = 0 \quad A(1, 0) = 1$
Cette suite diagonale de longueur 3 est indiquée en gras dans la matrice à gauche du tableau 4. L'échange en cascade associé à la suite diagonale qui change en 1 les valeurs 0 et en 0 les valeurs 1 indiquées par la suite produit la matrice à droite du tableau 4.

La suite diagonale telle qu'elle a été définie permet de trouver sans ambiguïté les positions dans la matrice des valeurs à changer avec l'échange en cascade associé. Mais l'écriture proposée n'est pas unique. Par exemple, l'échange en cascade de l'exemple, qui s'écrit ((0,3,1),(0,1,2),0), correspond aux trois paires d'égalités :

égalités 1	A(0,0)=0	A(0,1)=1
égalités 2	A(3,1)=0	A(3,2)=1
égalités 3	A(1,2)=0	A(1,0)=1

Si on écrit dans un ordre différent ces trois paires d'égalités comme par exemple

égalités 3	A(1,0)=1	A(1,2)=0
égalités 2	A(3,2)=1	A(3,1)=0
égalités 1	A(0,1)=1	A(0,0)=0

on obtient une suite diagonale qui s'écrit ((1,3,0),(0,2,1),1). Du point de vue de la définition, ce sont deux suites diagonales différentes, alors qu'elles ont le même ensemble d'égalités, et le même échange en cascade associé. On peut se fixer une règle afin d'obtenir une seule suite diagonale par échange en cascade.

Définition 6.4.4. *Règle de réécriture pour la suite diagonale (I,J,val).*
Nous proposons la procédure suivante :

- *dans la suite des indices de I, on écrit en premier le plus petit*
- *dans la suite des indices de J on écrit en premier le plus petit des deux indices de J associés au plus petit des indices de I*
- *La valeur de val est la valeur associée au plus petit indice de I et plus petit indice associé de J*

Cela nous donne la première paire d'égalité. Il suffit alors d'écrire les autres paires d'égalités dans le bon ordre, en ayant auparavant mis dans chaque paire d'égalités celle contenant la valeur val à gauche. Puis on lit les indices dans l'ordre des égalités.

Selon cette règle, la suite diagonale ((1,3,0),(0,2,1),1) se réécrit bien ((0,3,1),(0,1,2),0).

Propriété 6.4.4. *Unicité de la réécriture.*
Cette règle de réécriture aboutit à une écriture unique pour toutes les suites diagonales associées à un même échange en cascade sur une matrice donnée.

Preuve. *Elle découle de la définition d'une suite diagonale (I,J,val). Si on suppose que la longueur de la liste est k, on a vu qu'on pouvait lui associer k couples d'égalités écrites dans un certain ordre. La règle de réécriture impose un ordre unique pour cette suite de couples d'égalités, nous le décrivons ci-dessous.*

- *Choix du premier couple d'égalité :*
  La suite I d'indices de ligne ne comporte pas d'indices égaux, donc le choix du plus petit i est unique, et il en est de même pour le choix de l'indice j de colonne associé. La valeur val est alors imposée par le choix des deux premiers indices de I et de J. Cela nous donne la première égalité $A(i,j) = val$. Elle n'appartient qu'à un couple d'égalité, nous avons ainsi le premier couple.
- *Choix des premiers éléments de chaque couple :*
  On permute si nécessaire les couples d'égalités pour avoir toujours à gauche celle qui contient "= val".
- *Choix des couples d'égalités suivants :*
  Ayant choisi la première paire qui contient le couple d'indices (i,j), avec l'égalité $A(i,j) = val$ à gauche, l'égalité de droite nous impose le choix de l'indice suivant de J. Cet indice existe dans l'élément de gauche d'un seul couple d'égalité. Cet élément nous fournit l'indice suivant de ligne, et l'élément de droite nous fournit l'indice suivant de colonne. Et ainsi de suite.

Théorème 6.4.1. *: Décomposition d'un échange en cascade.*
Tout échange en cascade de longueur k>=2 peut se décomposer en k-1 échanges rectangulaires successifs. Cette décomposition n'est pas unique.

Preuve. *Montrons cette propriété par récurrence sur k.*

1. *pour k=2, la propriété est immédiate, car un échange en cascade de longueur 2 est un échange rectangulaire de longueur 2.*
2. *Supposons maintenant la propriété démontrée pour k=r (avec r>=2), et montrons la pour k=r+1.*
   Considérons un échange en cascade de longueur r+1. A cet échange en cascade correspond une suite diagonale (I,J,val) où I est une suite de r+1 indices $I = \{i_1, i_2, .., i_{r+1}\}$ *et de même* $J = \{j_1, j_2, .., j_{r+1}\}$*. D'après la définition de cette suite diagonale, on a* $A(i_1, j_1) = val$*,* $A(i_1, j_2) = 1 - val$*,* $A(i_2, j_2) = val$*, mais on ne peut pas trouver à l'aide de cette suite la valeur de* $A(i_2, j_1)$*. En effet, la suite diagonale est au moins de longueur 3, donc l'autre valeur de la ligne d'indice* i_2 *de la suite a pour indice de colonne* j_3*, qui est différent de* j_1 *par définition. On a deux possibilités pour* $A(i_2, j_1)$*, et on va décomposer l'échange en cascade de longueur r+1 de façon différente selon ces deux possibilités.*
   - *On suppose que* $A(i_2, j_1) = 1 - val$*.*
     On rajoute cette égalité (en gras) aux égalités de la suite diagonale de longueur r+1 :

     $A(i_1, j_1) = val$ $\quad A(i_1, j_2) = 1 - val$
     $\mathbf{A(i_2, j_1) = 1 - val}$ $\quad A(i_2, j_2) = val$ $\quad A(i_2, j_3) = 1 - val$
     $A(i_3, j_3) = val$ $\quad .....$
     ..... $\quad A(i_{r+1}, j_1) = 1 - val$

     La matrice A contient la suite diagonale de longueur 2 $((i_1, i_2), (j_1, j_2), val)$ *. On applique alors sur A l'échange rectangulaire associé à cette suite et on obtient une matrice B, qui vérifie les égalités suivantes (les changements de valeurs quand on est passé de A à B sont indiqués en gras ci-dessous) :*

     $\mathbf{B(i_1, j_1) = 1 - val}$ $\quad \mathbf{B(i_1, j_2) = val}$
     $\mathbf{B(i_2, j_1) = val}$ $\quad \mathbf{B(i_2, j_2) = 1 - val}$ $\quad B(i_2, j_3) = 1 - val$
     $B(i_3, j_3) = val$ $\quad .....$
     ..... $\quad B(i_{r+1}, j_1) = 1 - val$

     Puis on constate que sur cette matrice B, on a la suite diagonale de longueur r qui est $((i_2, i_3, .., i_{r+1}), (j_1, j_3, .., j_{r+1}), val)$*. En appliquant maintenant l'échange en cascade associé à cette suite diagonale sur B, on obtient une matrice C qui vérifie les égalités suivantes dans lesquelles on a mis en gras les valeurs qui ont changé quand on est passé de B à C :*

     $C(i_1, j_1) = 1 - val$ $\quad C(i_1, j_2) = val$
     $\mathbf{C(i_2, j_1) = 1 - val}$ $\quad C(i_2, j_2) = 1 - val$ $\quad \mathbf{C(i_2, j_3) = val}$
     $\mathbf{C(i_3, j_3) = 1 - val}$ $\quad .....$
     ..... $\quad \mathbf{C(i_{r+1}, j_1) = val}$

     Cette matrice C, obtenue par l'application successive de 2 échanges en cascade de longueurs respectives 2 et r sur la matrice A est identique à celle obtenue en appliquant directement l'échange en cascade de longueur r+1 sur la matrice A. En effet si on compare les égalités de A et de C, on voit que la seule valeur qui n'a pas changé de A à C est celle d'indice de ligne i_2 *et d'indice de colonne* j_1*, alors que toutes les valeurs concernées par la suite associée à l'échange en cascade de longueur r+1 ont changé. On a donc montré que dans ce cas, l'échange en cascade de longueur r+1 se décomposait en*

un produit de 2 échanges de longueurs respectives 2 et r.

- *Quand $A(i_2, j_1) = val$, on obtient une décomposition en deux échanges identiques aux précédents, mais composés dans l'ordre inverse.*
  En effet, la valeur de $A(i_2, j_1)$ ayant changé, on n'a plus besoin d'appliquer l'échange rectangulaire associé à la suite $((i_1, i_2), (j_1, j_2), val)$, on a directement dans la matrice A la suite diagonale de longueur r qu'on avait précédemment dans la matrice B qui est $((i_2, i_3, .., i_{r+1}), (j_1, j_3, .., j_{r+1}), val)$. On applique alors sur A l'échange en cascade associé, ce qui nous donne la matrice D, dans laquelle la valeur de $D(i_2, j_1)$ est $1 - val$, ce qui nous permet d'appliquer l'échange rectangulaire précédent qui rétablit la valeur initiale de $A(i_2, j_1) = val$ et on obtient ainsi comme dans le cas précédent, la même matrice que celle qu'on aurait obtenue en appliquant directement l'échange en cascade de longueur r+1.

On a décomposé cet échange de longueur r+1 en deux échanges de longueurs respectives 2 et r, c'est-à-dire en un échange rectangulaire et un échange de longueur r. En appliquant l'hypothèse de récurrence, on peut décomposer l'échange de longueur r obtenu en r-1 échanges rectangulaires. On obtient donc que l'échange de longueur r+1 est le produit d'un échange rectangulaire et d'un produit de r-1 échanges rectangulaires. C'est donc un produit de r échanges rectangulaires. On a bien démontré que la propriété est vraie pour k=r+1.

Comme la propriété est vraie pour 2, et que si elle est vraie pour k=r>=2, elle est vraie pour k=r+1, elle est donc vraie pour tout k>=2.

Cette décomposition n'est pas unique. En effet, on a pris comme "pivot" dans ce raisonnement la valeur située à la position (i_2, j_1), mais toute position qui forme un rectangle avec trois autres positions de la suite diagonale aurait convenu. La décomposition aurait alors été différente.

Par exemple, l'échange en cascade lié à la suite ((0,3,1),(0,1,2),0) sur la matrice du tableau 1 peut se décomposer de 3 façons différentes, selon que le pivot est pris en A(0,2), A(3,0) ou A(1,1), qui sont $E_{(0,3),(1,2)} \circ E_{(0,1),(0,2)}$, $E_{(1,3),(0,2)} \circ E_{(0,3),(0,1)}$ ou $E_{(1,3),(1,2)} \circ E_{(0,1),(0,1)}$. On peut voir la première décomposition sur le tableau 5. On a indiqué en gras dans la matrice A les éléments qui changent quand on applique l'échange rectangulaire $E_{(0,1),(0,2)}$. Le résultat est la matrice B. Puis on a indiqué en gras dans la matrice B les éléments qui changent quand on applique l'échange rectangulaire $E_{(0,1),(0,2)}$. Le résultat est la matrice C. Et dans C on a indiqué en gras les éléments qui ont changé par rapport à la matrice A, et en italique le pivot, qui a changé deux fois, donc qui a retrouvé sa valeur de départ.

matrice A		
0	1	**1**
1	0	**0**
1	1	1
1	0	1

matrice B		
1	**1**	**0**
0	0	1
1	1	1
1	**0**	**1**

matrice C		
1	**0**	*1*
0	0	**1**
1	1	1
1	**1**	**0**

TAB. 6.5 – La matrice A est transformée en B par $E_{(0,1),(0,2)}$ puis en C par $E_{(0,3),(1,2)}$

Propriété 6.4.5. *La transformée B d'une matrice A par un échange en cascade est une matrice S-équivalente.*

Preuve. *On a vu précédemment que la transformée d'une matrice par un échange rectangulaire est une matrice S-équivalente. Or un échange en cascade de longueur k est un produit de k-1*

échanges rectangulaires. Si on appelle A_i la matrice transformée de A_{i-1} par le i-ème échange rectangulaire, elles sont S-équivalentes et on obtient une suite de matrices (A_0, A_1, .. , A_{k-1}) avec A=A_0 et B=A_{k-1} qui permet de passer de A à B. Les matrices étant 2 à deux liées par la relation d'équivalence, il en est de même de A et de B, d'après la transitivité de cette relation. A et B sont donc S-équivalentes.

Nous avons montré que les suites en diagonales d'une matrice permettent de définir des transformations de cette matrice, les échanges en cascade, qui produisent des matrices équivalentes, c'est-à-dire qui conservent les sommes marginales du tableau de données associé. Nous allons maintenant montrer que toute transformation d'une matrice en une matrice S-équivalente peut s'écrire comme un produit d'échanges en cascades.

6.4.3 Les échanges en cascade permettent d'obtenir la classe d'équivalence d'une matrice

On a vu précédemment qu'il y a plusieurs façons de décomposer un échange en cascade de longueur k en une succession de k-1 échanges rectangulaires, et que cette composition n'est pas commutative. Notons que seul E, le premier de ces échanges rectangulaires se fait sur la matrice A, et pas les autres. En effet, le suivant F se fait sur l'image B de A par E et si la suite associée à E est sur A, celle associée à F est sur B et contient une position dont les valeurs ont été modifiées par l'échange E, cette position est celle que nous avons appelée pivot. Donc cette suite n'est pas une suite sur A, pas plus que les suivantes.

Si nous regardons maintenant comment passer d'une matrice A à une matrice B S-équivalente en utilisant seulement des suites diagonales sur A, cela revient à essayer de décomposer la transformation qui fait passer de A à B en un produit d'échanges en cascade. Toutefois, les suites étant sur A, pour qu'on puisse composer les échanges en cascade, il faut qu'ils portent sur des suites n'ayant aucune position en commun. En effet si une valeur de A est modifiée par un échange, elle ne peut plus faire partie d'une autre suite sur A. Nous allons établir l'existence de cette décomposition par un théorème qui en assure la construction.

Pour faciliter l'écriture de la démonstration de ce théorème, établissons d'abord le lemme suivant :

Lemme. *Si deux matrices distinctes A et B sont S-équivalentes, il existe une suite diagonale sur A telle que A et B diffèrent sur toutes les positions de cette suite.*

Preuve. *Si A et B diffèrent on peut trouver un indice de ligne i_1, et un indice de colonne j_1 tels que $A(i_1, j_1) = 1 - B(i_1, j_1)$. Appelons val la valeur $A(i_1, j_1)$. Nous allons produire une suite diagonale de la matrice A depuis cette position (i_1, j_1). Pour cela, nous allons procéder par étape, chaque étape construisant 3 positions de la suite selon un demi-rectangle, un coté sur une ligne, et l'autre sur une colonne, les demi-rectangles de deux étapes consécutives ayant un sommet en commun. La suite sera terminée quand le dernier demi-rectangle aura son dernier sommet à la première position.*

1. *étape 1 : recherche du premier demi-rectangle.*
   - *Comme $A(i_1, j_1) = val$ et $B(i_1, j_1) = 1 - val$, la première position est (i_1, j_1).*
   - *La ligne i_1 de la matrice A diffère de la même ligne i_1 de la matrice B car $A(i_1, j_1) = 1 - B(i_1, j_1)$. Comme le total des 2 lignes est le même et qu'elles sont toutes deux constituées de 0 et de 1, le nombre de colonnes pour lesquelles les valeurs diffèrent est pair. Par exemple si la ligne i_1 est 0 1 0 0 1 0 pour la matrice A et 1 0 0 1 0 1 pour la matrice B, les deux lignes diffèrent sur les quatre colonnes d'indices j=0, 1, 3, 4. On peut donc*

choisir un indice de colonne j_2 tel que $A(i_1, j_2) = 1 - val$, et $B(i_1, j_2) = val$. La deuxième position est (i_1, j_2).

- *Prenons la colonne j_2. D'après le même raisonnement que pour la ligne i_1, il y a un nombre pair d'indices de lignes pour lesquels les valeurs de la colonne j_2 de A et j_2 de B diffèrent. Ayant déjà l'indice i_1 pour lequel $A(i_1, j_2) \neq B(i_1, j_2)$, on choisit un indice de ligne i_2 pour lequel $A(i_2, j_2) = val$ et $B(i_2, j_2) = 1 - val$. La troisième position est (i_2, j_2).*

On a ainsi trouvé 3 positions successives (i_1, j_1), (i_1, j_2) et (i_2, j_2) formant un demi-rectangle telles que leurs valeurs respectives sur A soient val, $1 - val$, val et sur B $1 - val$, val, $1 - val$. On passe à l'étape 2.

2. *étape k : recherche du k-ème demi-rectangle. à l'étape précédente, on a trouvé 3 positions successives (i_k, j_k), (i_k, j_{k+1}) et (i_{k+1}, j_{k+1}) formant un demi-rectangle telles que leurs valeurs respectives soient val, $1 - val$, val sur A et $1 - val$, val, $1 - val$ sur B.*
   - *Comme $A(i_k, j_k) = val$ et $B(i_k, j_{k+1}) = 1 - val$, la première position est (i_{k+1}, j_{k+1}).*
   - *La ligne i_k de la matrice A diffère de la même ligne i_k de la matrice B car $A(i_k, j_k) = 1 - B(i_k, j_k)$. Comme le total des 2 lignes est le même et qu'elles sont toutes deux constituées de 0 et de 1, elles diffèrent d'un nombre pair de valeurs. On peut donc trouver un indice de colonne i_{k+1} tel que $A(i_k, j_{k+1}) = 1 - val$ et $B(i_k, j_{k+1}) = val$. Cet indice peut être différent de l'indice j_1 ou non. Mais on le choisit tel que la position (i_k, j_{k+1}) n'ait pas déjà été rencontrée. La deuxième position est (i_k, j_{k+1}).*
   - *La colonne j_{k+1} de la matrice A diffère de la même colonne j_{k+1} de la matrice B car $A(i_k, j_{k+1}) = 1 - B(i_k, j_{k+1})$. On peut trouver un indice de ligne i_{k+1} tel que $A(i_{k+1}, j_{k+1}) = val$ et $B(i_{k+1}, j_{k+1}) = 1 - val$. Si l'indice j_{k+1} diffère de j_1, on choisit l'indice de ligne i_{k+1} pour ne pas rencontrer une position déjà rencontrée. Sinon, on peut choisir pour i_{k+1} égal à i_1. La troisième position est (i_{k+1}, j_{k+1}).*

   Si j_{k+1} est égal à j_1, il se peut que la seule possibilité pour l'indice i_{k+1} soit i_1, auquel cas il n'y a pas d'étape suivante, la suite est trouvée. Sinon, on passe à l'étape k+1 qui est identique à l'étape k dans laquelle on remplace partout k par k+1.

Ce processus fournit une suite qui ne se recoupe pas, car on a exclu la possibilité d'avoir deux fois la même position. Comme la suite ne se recoupe pas, le nombre d'étapes est nécessairement fini. Toutefois, la suite obtenue par ces étapes n'est pas nécessairement une suite diagonale car il peut y avoir un indice de ligne ou de colonne qui se répète. Pour obtenir une suite diagonale, il faut l'extraire de la suite obtenue, ce qui se fait simplement[118]*. Ainsi, si on a deux indices de ligne qui se répètent, on retire toutes les positions qui ont été trouvées entre ces deux indices, ainsi qu'une des deux positions correspondant à ceux-ci. Par exemple, la suite ((0,1,2,1),(2,1,3,0),1) se réduit en ((0,1),(2,1),1) ou en ((2,1),(3,0),1).*

Théorème 6.4.2. *: Si A et B sont deux matrices S-équivalentes, il existe un nombre fini r>=0 de suites diagonales disjointes définies sur A telles que le produit des échanges en cascade associés transforme A en B.*

Preuve. *Nous procédons en r étapes.*

1. *étape 0*
   - *Si les deux matrices A et B sont identiques, la démonstration est terminée, et on a r=0.*

[118] On aurait pu faire un autre choix pour produire directement des suites diagonales, en interdisant de reprendre un indice de ligne qui a déjà été pris. Dans ce cas, les étapes pourraient ne pas aboutir pour certains choix d'indices ce qui imposerait des retours en arrière.

- *Si les deux matrices A et B diffèrent, d'après le lemme, on peut trouver une suite diagonale S_0 pour laquelle tous les éléments de A diffèrent de ceux de B. Appelons A_1 la matrice obtenue en transformant A à partir de l'échange en cascade correspondant à la suite diagonale S_0. Et passons à l'étape 1.*

2. *étape 1*
   - *Si la matrice A_1 est égale à la matrice B, la démonstration est terminée, et on a r=1*
   - *Sinon, d'après le lemme, nous pouvons trouver une suite diagonale S_1 sur A_1 telle que tous les éléments de A_1 diffèrent de ceux de B. S_1 est une suite diagonale sur A_1, mais également sur A_0. En effet, par construction, les valeurs de A_1 et de A ne diffèrent que pour les positions de la suite S_0, et les suites S_0 et S_1 n'ont aucune position en commun, donc les matrices A_1 et de A ont les mêmes valeurs pour les positions de S_1[119]. Appelons A_2 la matrice obtenue en transformant A_1 à partir de l'échange en cascade correspondant à la suite diagonale S_1. On peut également obtenir A_2 en transformant A à partir du produit des échanges en cascade correspondant aux suites diagonales disjointes de A qui sont S_0 et S_1. Et passons à l'étape 2.*

3. *étape k>=2*
   On a obtenu à l'étape k-1 une matrice A_k, transformée de la matrice A à partir d'un produit de k échanges en cascades correspondant aux k suites diagonales disjointes S_0, S_1,..,S_{k-1} sur A.
   - *Si la matrice A_k est égale à la matrice B, la démonstration est terminée, et on a r=k*
   - *Sinon, nous recommençons le raisonnement fait dans l'étape 1 en remplaçant A_1 par A_k, S_1 par S_k, et S_0 par les k suites diagonales S_0, S_1,..,S_{k-1} sur A. Et on peut passer à l'étape k+1.*

Nous voyons qu'on peut n'avoir qu'une étape, l'étape 0, quand les deux matrices sont identiques, deux étapes, qui sont l'étape 0 et 1 si la matrice A et la matrice B ne diffèrent que d'une suite, et r+1 étapes si leurs différences peuvent s'exprimer avec r suites. Le nombre r est nécessairement fini car les suites sont disjointes, et chacune couvre un nombre d'éléments égal à deux fois sa longueur, la plus petite suite en couvrant 4. Donc r varie entre 0 et $n \times p/4$.

Cette décomposition du passage d'une matrice à une autre S-équivalente en un ensemble de suites diagonales disjointes sur la première matrice est unique à des réécritures près, comme l'indique la propriété suivante :

Propriété 6.4.6. *: Le produit d'échanges en cascade associé à des suites diagonales sur la matrice A qui permet de passer de A à B est commutatif, et l'ensemble des suites diagonales associées est unique aux réécritures près.*

Preuve.
- *La commutativité du produit vient du fait que les suites sont disjointes. Lors d'un échange en cascade, on change les valeurs de la matrice pour la suite correspondant à l'échange sans toucher aux valeurs des autres suites. Ce qui fait que chaque valeur de la matrice est changée par un seul échange en cascade, quand elle l'est. Le résultat final des échanges sur la matrice est donc le même quel que soit l'ordre dans lequel on fait les échanges.*
- *Montrons l'unicité par l'absurde. Si on suppose qu'on peut trouver deux produits distincts d'échanges en cascade, comme ces produits sont commutatifs, c'est qu'ils diffèrent par au*

[119] Pour s'en convaincre, imaginons qu'une position (i,j) soit commune aux deux suites. Comme (i,j) fait partie de la suite S, on a $A(i,j) = 1 - B(i,j)$, et on a également par construction de A_1 l'égalité $A(i,j) = 1 - A_1(i,j)$. Comme (i,j) fait partie de la suite S_1, on a $A_1(i,j) = 1 - B(i,j)$. On voit que ces 3 inégalités sont contradictoires.

moins un échange. Supposons donc qu'un échange en cascade E appartienne au premier produit et pas au second. Il est associé à une suite diagonale sur A qui porte sur des positions dont les valeurs diffèrent de A à B. Nécessairement, ces positions font partie de suites diagonales correspondant à des échanges du second produit. Si E ne fait pas partie de ce produit, c'est que ces positions sont soit une partie stricte de celles d'une suite du second produit, soit qu'elles font partie d'au moins deux suites de ce produit.

Ce théorème permet de construire la classe d'équivalence complète d'une matrice A. Pour cela, il suffit de trouver toutes les suites diagonales sur A, et de faire tous les produits des échanges en cascade correspondants. On a alors autant de matrices S-équivalentes distinctes de A que de produits.

Si on reprend l'exemple de la matrice A du tableau 1, elle contient quatre suites diagonales, trois de longueur 2 et une de longueur 3. Ces 4 suites ont en commun une position, qui est $(0,0)$, ce qui fait qu'on ne peut pas faire le produit de 2 ou plus des échanges associés. La classe d'équivalence de A contient ainsi 5 matrices qui sont la matrice A elle-même, obtenue en appliquant la transformation identité sur A, produit de 0 échange, et les quatre matrices obtenues en appliquant l'échange en cascade associé à chacune des quatre suites, donc produit de 1 échange.

6.4.4 Avec des échanges rectangulaires successifs on obtient la classe d'équivalence d'une matrice

La technique précédente permet de trouver de façon exhaustive les matrices S-équivalentes à une matrice donnée. Si on calcule en plus les probabilités associées à l'apparition de chaque matrice, on peut évaluer la qualité des algorithmes de simulations de Monte-Carlo permettant de générer des matrices S-équivalentes. En effet, un algorithme de simulation qui "oublierait" systématiquement de générer un cas très particulier de probabilité 0,05 serait douteux. Et on s'attend à ce que la répartition des résultats des simulations tende vers les probabilités calculées. Ce qu'on peut vérifier avec de petits exemples. Mais ces simulations de Monte-Carlo ne peuvent pas s'appuyer sur une technique comme celle-ci. Nous proposons pour les réaliser une technique s'appuyant sur le théorème suivant :

Théorème 6.4.3. *: Si A et B sont deux matrices S-équivalentes, on peut passer de A à B par une suite finie d'échange rectangulaires.*

Preuve. *La preuve découle des théorèmes précédents. Comme on peut passer de A à B par un produit fini d'échanges en cascades, et que chaque échange en cascade se décompose en un produit fini d'échanges rectangulaires, l'associativité de la composition des transformations nous permet de passer de A à B par un produit fini d'échanges rectangulaires.*

Bien sûr, plus aucune unicité n'est assurée, mais nous avons ici une procédure très simple de génération de matrices S-équivalentes à une matrice donnée.

6.5 Deux exemples de tirages aléatoires

Nous avons exposé précédemment les principes de génération de matrices S-équivalentes à une matrice donnée. Ces principes ont été adoptés suite à l'échec de la procédure d'acceptation-rejet, qui ne produit pas suffisamment de matrices S-équivalentes. Après avoir exposé rapidement les résultats de cette première procédure à l'aide de quelques exemples, nous allons décrire la

procédure à base d'échanges rectangulaires qui suit les principes détaillés dans le paragraphe précédent.

6.5.1 Par acceptation/rejet

Nous reprenons la matrice du tableau 1, qui a 4 lignes et 3 colonnes. Nous avons vu qu'elle admet cinq matrices S-équivalentes, elle-même et les 4 matrices du tableau 2. Une matrice est produite en permutant au hasard chacune des 3 colonnes. De cette façon, les totaux des colonnes ne changent pas. Mais les totaux des lignes peuvent changer. Par exemple on peut échanger dans la colonne B du tableau 1 les valeurs 0 et 1 des sujets s2 et s3, ce qui garde inchangés les totaux par colonnes, alors que les totaux respectifs des lignes de s2 et de s3 deviennent 2.

On a ainsi généré de façon indépendante 3000 matrices. En répétant 10 fois ce tirage de 3000 matrices, on a obtenu respectivement 1855, 1812, 1866, 1861, 1882, 1879, 1828, 1882, 1867, 1867 matrices dont les totaux par lignes coïncident avec la matrice de départ. Ces matrices sont les 5 matrices attendues, reproduites en des nombres d'exemplaires variables.

1	0	0	0	0
1	0	1	0	0
1	0	0	1	1
1	1	1	1	0
1	1	1	0	1
0	0	0	1	1
0	0	0	1	1
1	0	1	0	0
1	1	0	1	1
1	1	1	1	1

Tab. 6.6 – Un matrice à 10 lignes et 5 colonnes

Puis nous avons augmenté la taille de la matrice qui est celle du tableau 6. La procédure d'acceptation/rejet a fourni avec 10 tirages entre 53 et 76 matrices S-équivalentes sur 3000 générées, alors qu'on a établi par ailleurs que cette matrice admet plus de 1800 matrices S-équivalentes différentes. Un dernier essai sur une matrice avec 33 lignes et 6 colonnes a fourni une seule matrice S-équivalente pour 40 000 simulations et montre que cette procédure n'est pas utilisable sur des données réelles.

Nous avons alors essayé de corriger les matrices obtenues par ces simulations en faisant glisser des valeurs 1 sur des colonnes depuis les lignes à total trop grand vers les lignes à total trop petit. Plus on acceptait d'enchaîner des corrections sur une matrice, moins elle avait de chances d'être rejetée. Mais plus il était difficile de mesurer les risques de s'éloigner du hasard qu'on cherchait à simuler. Et nous avons effectivement constaté qu'avec la méthode d'acceptation/rejet sans correction sur une matrice à 5 lignes et 3 colonnes, on obtient quelque 300 matrices sur les 1000 générées, chacune des 5 matrices S-équivalentes ayant une fréquence d'apparition voisine de 0,2, alors que si accepte un unique niveau de correction, on en obtient presque 800 mais l'une a une fréquence à peine supérieure à 0,19, et avec un ou 2 niveaux de corrections, les 1000 matrices sont toutes corrigées mais la fréquence de la matrice la plus rare se stabilise autour de 0,18.

Nous n'avons pas retrouvé ces défauts dans la procédure que nous décrivons maintenant, qui s'appuie sur les théorèmes démontrés précédemment.

6.5.2 Par produits d'échanges rectangulaires

La technique de simulation s'appuyant sur les échanges rectangulaires se fait ainsi :

- Une simulation est la création d'une matrice B à partir d'une matrice A par une succession de k échanges, selon cette procédure :
  1. Initialisation de B : nous créons une copie B de A.
  2. Transformation de B : nous prenons au hasard et indépendamment quatre nombres, deux entre 0 et n-1, et deux entre 0 et p-1, n étant le nombre de lignes et $J = (0, 1, .., p-1)$ et p le nombre de colonnes de la matrice B, ce qui nous donne deux indices de lignes et deux indices de colonnes. S'il est possible de faire un échange rectangulaire avec ces quatre positions, nous le faisons, en changeant ainsi le cas échéant le contenu de la matrice B.
  3. Nous recommençons k fois la transformation de B.
- Les simulations successives peuvent se faire en reprenant à chaque fois comme matrice A la matrice de départ, ou en prenant la matrice B obtenue lors de la simulation précédente. Nous avons appelé génération de type 1 la première façon de procéder, le type 0 étant l'autre façon.

Ceci est la méthode générale. On peut faire quelques améliorations de de cette méthode, par exemple en éliminant des indices pouvant être tirés, les indices des lignes ou colonnes ne pouvant jamais donner lieu à des échanges, comme les lignes ou colonnes constantes. Nous reprenons la matrice du tableau 1, qui a 4 lignes et 3 colonnes. Nous avons vu qu'il y a trois échanges rectangulaires possibles sur cette matrice, et une suite diagonale de longueur 3, qui peut être décomposée de trois façons différentes en produit de deux échanges rectangulaires.

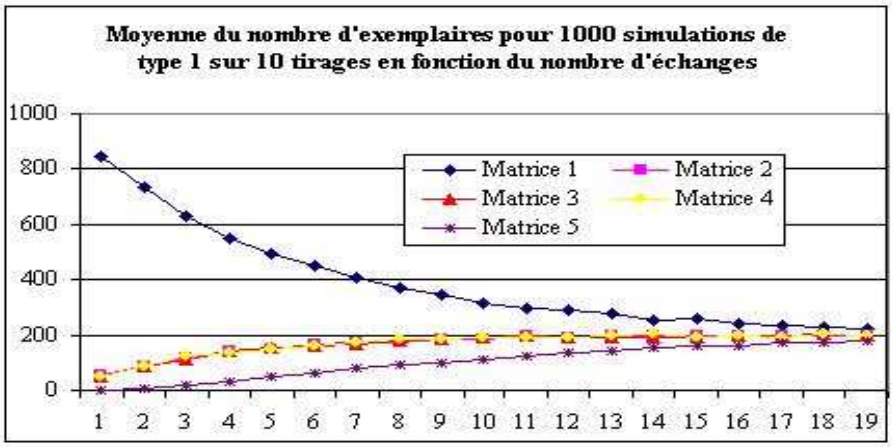
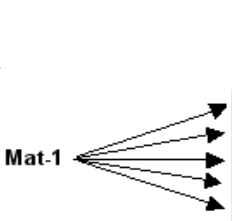


FIG. 6.1 – Le nombre moyen d'exemplaires de la matrice du tableau 1 obtenus par des simulations de type 1.

Dans la figure 1 sont représentés les résultats de 10 tirages de 1000 simulations de type 1 de cette matrice en fonction du nombre d'échanges successifs choisis. Examinons d'abord le cas où chaque simulation se fait avec 1 seul échange rectangulaire sur la matrice du tableau 1. C'est donc l'une des trois transformations ((0,1),(0,1),0) [120], ((0,1),(0,2),0), ((0,3),(0,1),0), correspon-

[120] Nous rappelons que cette notation est une de celles choisies pour un échange rectangulaire entre les lignes 0 et 1, et les colonnes 0 et 1, la position (0,0) ayant la valeur 0.

dant aux trois matrices de gauche du tableau 2, matrices nommées "Matrice 2", "Matrice 3" et "Matrice 4" dans la figure 1. Comme la ligne d'indice 2 est une ligne constante, elle n'est pas examinée. Nous tirons ainsi au sort à deux reprises un numéro de ligne qui peut être 0, 1 ou 3 et un numéro de colonne qui peut être 0, 1 ou 2 (soit $3^4 = 81$ quadruplets d'indices).



	Nombre de répétitions				
	1	2	10	100	250
Probabilité d'apparition de chaque matrice					
Mat-1	0.852	0.733	0.318	0.200	0.200
Mat-2	0.049	0.087	0.188	0.200	0.200
Mat-3	0.049	0.087	0.188	0.200	0.200
Mat-4	0.049	0.087	0.188	0.200	0.200
Mat-5	0.000	0.007	0.117	0.200	0.200
Plus grand écart à 0.2					
	0.652	0.533	0.118	5E-08	4E-16

FIG. 6.2 – Les probabilités associées au différentes matrices pour des simulations de type 1.

Si nous tirons au sort les indices de lignes 0 et 1 et les indices de colonnes 0 et 2, c'est l'échange ((0,1),(0,2),0) qui se fait et nous obtenons la matrice 3 (il peut se produire pour 4 tirages sur 81, soit une probabilité de 0.49 comme on peut le lire dans la figure 2 à l'intersection de la ligne $Mat - 3$ et de la colonne "nombre de répétitions : 1", car on peut permuter de 4 façons possibles les paires d'indices de lignes et de colonnes). Si nous tirons les indices de lignes 1 et 2, et les indices de colonne 0 et 2, aucun échange ne se fait, nous obtenons donc la matrice 1, ainsi que si nous tirons deux indices de lignes égaux, ou deux indices de colonnes égaux. Ce qui explique que la matrice 1 se retrouve à plus de 800 exemplaires (premier point sur la courbe de la matrice 1, en haut à gauche du graphique de la figure 1, probabilité de 0,853 comme on peut le lire dans la figure 2 à l'intersection de la ligne $Mat - 1$ et de la colonne 1) quand il n'y a qu'un échange, et ceci pour chacun des 10 tirages. Et la matrice 5 ne se trouve jamais (point d'abcisse 1 et d'ordonnée 0 en bas à gauche du graphique de la figure 1) car elle est obtenue à partir d'un échange en cascade de longueur 2, donc produit de deux échanges rectangulaires, et non à partir d'un unique échange rectangulaire. Les autres matrices se partagent les exemplaires restants de façon à peu près égale.

Si nous demandons maintenant des simulations à k échanges où le nombre k est 2, le nombre d'exemplaires de la matrice 1 diminue, puisque la probabilité de ne pas avoir d'échange possible deux fois de suite est nécessairement inférieure à celle de ne pas l'avoir la première fois, et il n'est pas compensé par la probabilité d'avoir deux fois de suite le même échange rectangulaire . Et on commence à obtenir quelques exemplaires de la matrice 5, car on a trois successions différentes possibles de deux échanges qui peuvent la produire. Quant aux autres matrices, leur probabilité d'apparaître a presque doublé, car elles peuvent provenir de la succession de leur transformation associée avec l'identité, ou inversement - le cas où leur transformation associée est suivie d'elle-même ou bien d'une autre transformation étant de probabilité très faible Et avec k=19, on peut voir dans la figure 1 que les fréquences des matrices 1 et 5 n'ont toujours pas rejoint celles des autres matrices. Dans la figure 2, on a indiqué les probabilités de chaque matrice selon quelques valeurs de k en colonnes afin de montrer qu'il y a bien convergence vers la valeur 0.2. En bas de la figure sont notés pour chaque valeur de k les écarts maximaux entre les probabilités d'apparition des 5 matrices et la valeur 0.2. Pour k=250, les probabilités sont à moins de 4.E-16 de la valeur 0.2.

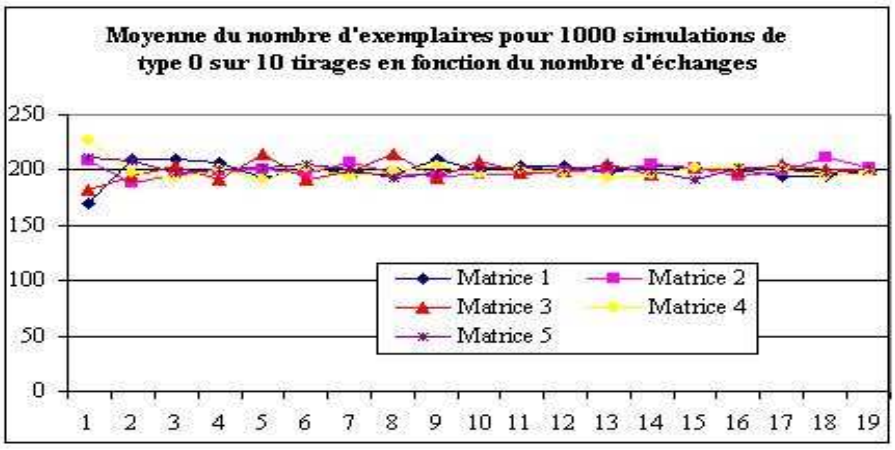


FIG. 6.3 – Le nombre moyen d'exemplaires de la matrice du tableau 1 obtenus par des simulations de type 0.

Sur la figure 3, on peut voir que les résultats sont meilleurs pour les simulations de type 0, c'est-à-dire quand on fait chaque simulation en utilisant la matrice obtenue à la simulation précédente. Si on complète l'examen des résultats des simulations de type 0 par le graphique des écart-type, on voit que plusieurs échanges successifs permettent d'obtenir une meilleure simulation. Bien sûr, le meilleur résultat est avec 19 échanges successifs possibles, pour lesquels on a le plus petit écart-type et des moyennes proches de 0.2.

Examinons maintenant la figure 5 qui montre le nombre de matrices produites par les simulations de type 0 sur la matrice du tableau 8. Nous voyons que si on augmente le nombre d'échanges, le nombre de matrices différentes obtenues s'approche de 1000. Toutefois, le nombre effectif de matrices différentes qu'on peut obtenir est voisin de 2000. Il n'est donc pas étonnant qu'on ne puisse pas atteindre 1000, certaines matrices se répétant. On aurait plus de chances d'obtenir un tirage par hasard de 1000 matrices toutes différentes si la taille de la classe d'équivalence de cette matrice était bien au delà de 2000.

Cette procédure nous permet donc de générer en grand nombre des matrices S-équivalentes à une matrice donnée, la répartition de leur fréquence ressemblant aux probabilités attendues. On peut l'améliorer de multiples façons. Par exemple, si on décide de ne pendre que des indices de lignes différents et des indices de colonnes différents, pour la matrice du tableau 1, on obtient avec des simulations à 10 échanges un résultat aussi bon qu'avec les 19 échanges quand on acceptait d'avoir deux lignes ou deux colonnes identiques dans les tirages.

6.6 Bilan et perspectives

Nous avons décrit dans ce chapitre une technique de simulation essayant de "coller" le plus possible aux lois du hasard. Nous avons montré dans le chapitre précédent une façon d'utiliser ces simulations. Nous la rappelons brièvement avant de proposer d'autres pistes.

6.6.1 Significativité d'une règle

A partir d'un tableau de données réelles, nous avons créé des motifs fréquents et un jeu de règles d'association. Nous avons ensuite simulé des tableaux de données dus au hasard selon

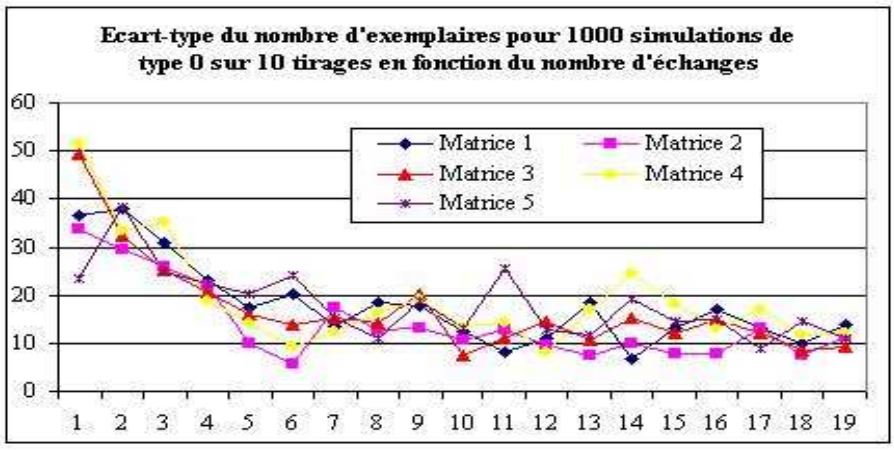


FIG. 6.4 – L'écart-type du nombre d'exemplaires de la matrice du tableau 6 obtenus par des simulations de type 0.

plusieurs techniques et créé les motifs fréquents de tous ces tableaux. Nous avons alors décidé que les motifs fréquents du tableau initial qui apparaissaient parmi les motifs fréquents d'au moins 5% de ces tableaux avec une fréquence supérieure ou égale à celle d'origine étaient dus au hasard. Puis nous avons appelé règle "douteuse" toute règle construite sur un motif contenant un sous-motif dû au hasard. Nous avons ainsi examiné une à une chaque règle et proposé d'éliminer les règles douteuses du jeu de règles initial, et de ne garder ainsi que les règles "significatives".

Ce nettoyage peut être amélioré dans deux directions que nous exposons ci-dessous.

6.6.2 Prise en compte de la multidimensionnalité des motifs

La recherche des règles douteuses que nous avons faite permet d'estimer les supports des motifs de longueur 2, et les confiances des règles construites sur ces motifs, ayant donc une propriété à gauche et une à droite. Nous éliminons ainsi de façon appropriée les règles dues au hasard construites sur des motifs de longueur 2. Mais au fur et à mesure que le nombre de propriétés présentes dans la règle augmente, nos performances de nettoyage se dégradent.

En effet si on obtient avec les données d'origine un motif de longueur 5, alors que dans les matrices S-équivalentes il y en a très peu, cela peut être dû à des sous-motifs de longueur 2 de supports exceptionnellement grands. On pourrait essayer de corriger cela en augmentant la matrice de départ par l'ajout de colonnes avec ces motifs fréquents de longueur 2 qui ne sont pas dus au hasard, et générer des matrices S-équivalentes pour déterminer quels sont les motifs de longueur 3 apparaissant par hasard, une fois la fréquence élevée du sous-motif de longueur 2 qu'ils contiennent prise ainsi en compte, et ainsi de suite en augmentant la longueur des motifs jusqu'à ce qu'il n'y ait plus de motifs significatifs d'une longueur donnée. On ne peut pas reproduire exactement les simulations que nous venons de faire en se bornant à remplacer la matrice de départ par cette nouvelle matrice En effet, il faudrait que les matrices générées respectent des contraintes supplémentaires : elles devraient non seulement avoir même marges que la matrice d'origine, mais encore reproduire les liens entre les motifs de longueur 2 et les propriétés les constituant[121]. Les échanges rectangulaires que nous avons définis ne permettent pas de produire de telles matrices. Il faudrait donc proceder autrement.

[121] Tout changement de la colonne 0 doit être répercuté sur la colonne 01.

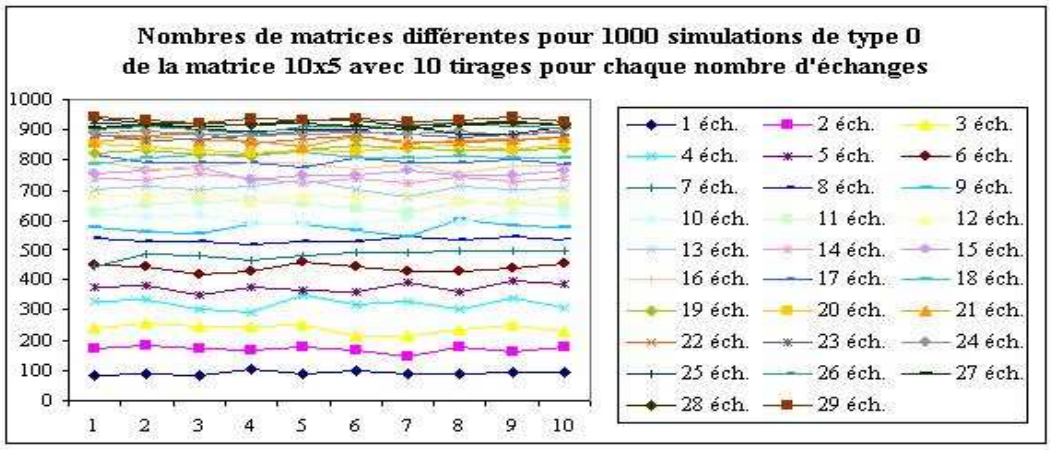


FIG. 6.5 – Le nombre moyen d'exemplaires de la matrice du tableau 6 obtenus par des simulations de type 0.

6.6.3 Un jeu de règles significatif

Bien que le tableau soit généré dans son ensemble, chaque règle est ensuite jugée douteuse ou non indépendemment des autres règles. On pourrait envisager une façon plus globale de le faire. Nous allons l'exposer sur un petit exemple.

Reprenons la matrice du tableau 1 et les quatre matrices du tableau 2 pour comparer leur jeux de règles d'association. En nommant A, B et C leurs propriétés, nous obtenons les supports suivants des motifs :

	Motifs							
Matrice n°	∅	A	B	C	AB	AC	BC	ABC
1	4	3	2	3	1	2	2	1
2	4	3	2	3	1	3	1	1
3	4	3	2	3	2	2	1	1
4	4	3	2	3	1	2	2	1
5	4	3	2	3	2	2	1	1

TAB. 6.7 – Les motifs des 5 matrices de la classe d'équivalence de la matrice du tableau 1

On constate dans le tableau 1 que les motifs des matrices 1 et 4 ont tous les mêmes supports. Si on retourne aux tableaux de données 1 et 2, on voit que la seule différence entre ces deux matrices est que les réponses des sujets s1 et s4 ont été échangées. Et entre les matrices 3 et 5, le même phénomène s'est produit. Si on considère que la numérotation des sujets peut se faire de façon quelconque, on peut alors dire qu'il n'y a pas de différence entre les matrices 1 et 4, et entre les matrices 3 et 5, et fusionner les matrices en ajoutant leurs probabilités ce qui donne les matrices 1&4, 2, et 3&5 de probabilités respectives 0,4, 0,2 et 0,4.

Si on se limite aux règles exactes de support non nul, les 3 jeux de règles produits sont {C–>B, AB–>C}, {A–>C, C–>A, AB–>C, BC–>A }, {A–>B, BC–>A} avec des probabilités respectives de 0,4, 0,2 et 0,4. Le premier jeu de règles étant issu de la matrice d'origine, si on examine chacune de ses deux règles, comme dans le chapitre précédent, on décide que la première

est plus significative que la seconde, car elle ne se trouve que dans le premier jeu de règles c'est-à-dire avec une probabilité de 0,4 et la seconde dans le premier et le deuxième, avec une probabilité de 0,6[122].

Mais on peut également examiner la significativité du jeu de règles pris dans son ensemble, sa probabilité étant de 0,4, si on estime que les propriétés ne peuvent pas être échangées, comme les sujets l'ont été. Ce point de vue est le plus courant, mais le point de vue autorisant à échanger les propriétés existe aussi[123]. En comparant dans le tableau de motifs les lignes des matrices 1, 3, 4 et 5, on voit qu'elles ne diffèrent que par les valeurs des colonnes AB et BC. On vérifie dans les tableaux 1 et 2 qu'on peut passer de la matrice 1 à la matrice 3 par échange des colonnes A et C, et de la matrice 4 à la matrice 5 de la même façon. Si on accepte de renommer les propriétés, on peut fusionner les 4 matrices 1, 3, 4 et 5, et ainsi le jeu de règles de la première obtient une probabilité de 0,8, ce qui le rend encore moins significatif.

Cette façon d'étendre la significativité des règles aux jeux de règles, nécessite non seulement des modifications de nos simulations, mais encore la construction de jeux de règles d'association sur les tableaux simulés, et d'une distance entre jeux de règles d'association.

[122] L'agrégation des matrices ne change pas la significativité des règles car elle s'obtient par addition des probabilités, ou des fréquences quand on simule.

[123] Dans les plans d'expériences cela correspond à des facettes aléatoires.

Conclusion de la partie II

Le fait de se référer au hasard pour renforcer après coup le pouvoir d'induction d'une règle n'est pas nouveau en intelligence artificielle. Il est notamment utilisé pour faire de l'"apprentissage" sur les données lors de l'élaboration des arbres de décision. Jensen [131] signale que bien qu'il y ait une variable à expliquer (dans le cas de l'arbre de décision, c'est le classement du sujet) on ne cherche pas à tester une hypothèse préalable, mais à la trouver, et qu'il est impossible de le faire en testant toutes les hypothèses possibles car on met alors en défaut les hypothèses préalables à l'utilisation de la quasi-totalité des tests statistiques :
≪*...some investigators use conventional software "inductively". They examine tens or hundreds of models while searching for useful generalizations. This invalidates the assumptions of nearly all statistical tests, and these investigators are often chastised for their statistical* naiveté.≫
Il conclut cette remarque en disant que le chercheur désire un système qui l'assiste à la recherche de modèles, et qui teste si ces modèles ne sont pas dus au hasard.
Puis ayant indiqué par là qu'on ne peut pas utiliser les tests statistiques d'hypothèses, il décrit la méthode alternative, qui assure un processus inductif, comme formée de deux étapes. La première est de générer un ou plusieurs modèles, c'est-à-dire les trouver et ajuster leurs paramètres, la seconde est de mesurer les diverses qualités de ces modèles (on suppose qu'il y a toujours un modèle naïf au départ). En principe, la qualité du modèle est estimée sur de nouvelles données. Si elle correspond à la qualité trouvée sur les données qui ont permis de construire le modèle, tout va bien, mais c'est en général moins bon, c'est ce qu'on appelle le "sur-apprentissage". La qualité d'un modèle inductif peut se tester en calculant pour chaque point l'écart entre la valeur observée et la valeur prédite. La difficulté de tester sur les données de départ les divers classements (problème dit des *comparaisons multiples* [131]) conduit naturellement à tester ces modèles sur de nouvelles données.

Nous voyons que la démarche que nous avons choisie pour les règles d'association ressemble à celle que nous venons de décrire. Dans les deux cas, le pouvoir inductif d'une règle trouvée sur un tableau T ne peut plus être assuré par les statistiques inférentielles, qui "légifèrent le hasard" ; il convient donc de vérifier que la règle n'est pas spécifique aux sujets du tableau T, et on le vérifie en comparant la qualité de la règle sur les sujets à partir desquels elle a été construite à celle qu'elle a sur d'autres sujets. La façon de "mettre le hasard" est la différence essentielle entre ces deux techniques. D'un côté, on définit un hasard lié au contexte, un seuil de significativité α, en général 0,10 0,05 ou 0,01, des indices de qualité avec leurs seuils, et on recommence 100 fois des tirages indépendants ce qui donne 100 exemplaires "aléatoires" de ce tableau. Une règle donnée est reconnue comme significative si elle apparaît dans plus de 100xα tableaux avec des indices de qualité dépassant les seuils. De l'autre on coupe de façon aléatoire le tableau T en deux tableaux T1 et T2 contenant les mêmes propriétés mais pas les mêmes sujets, et une règle trouvée sur T1 est reconnue comme valable si elle apparaît dans T2 avec les indices de qualité dépassant les

seuils [124]. La première méthode est plus lourde, mais on obtient plus de sûreté dans le jugement qu'avec la seconde, car on dispose non d'une valeur d'un indice, mais de sa loi de distribution sur nos données. Le risque de cette méthode est qu'elle peut fournir des résultats biaisés, comme certaines estimations obtenues par un bootstrap naïf [74]. Si le nombre de sujets est petit par rapport au nombre de propriétés, on préférera la première méthode. S'il est très grand et a du mal à tenir en mémoire centrale et/ou sur le disque dur, on préférera la seconde. Entre ces deux extrêmes, si on désire comparer l'efficacité de plusieurs indices de qualité sur des données ou trouver des seuils de décision pour ces indices adaptés aux données on préférera la première.

[124] Pour les tâches de classement, on peut recommencer plusieurs fois la manipulation qui consiste à couper T en deux. Mais on doit alors extraire à chaque fois les règles sur T1 et les valider sur T2. Ce qui risque de ne pas redonner les mêmes règles. Ce n'est pas important pour une tâche de classement qui consiste à prédire la classe d'un sujet. La qualité de la tâche de classement est en effet liée à la prédiction, pas au jeu de règles qui peuvent changer d'une manipulation à une autre.

Troisième partie

La prise en compte des liaisons complexes : position du problème et proposition de solutions

7

Les difficultés d'interprétation d'une règle

Nous avons déjà exposé précédemment (voir chapitre 2) les divers types de problèmes rencontrés dans le raisonnement courant en sciences humaines lorsqu'on passe de l'association de deux propriétés à celle de trois propriétés, qu'on appellera indifféremment variables ou attributs comme c'est l'habitude en sciences humaines. Nous constatons d'abord que ces problèmes se posent également dans certains modèles de l'intelligence artificielle. Nous décrivons ensuite les effets les plus courants de l'arrivée d'une troisième variable sur la liaison entre deux variables pour les règles d'association. Pour rendre ces effets plus faciles à appréhender, nous les représentons par des nuages de points, donc avec des variables quantitatives A et B, en utilisant le modèle de la *corrélation* et de la *régression* [10], puis nous transformons A en variables binaires, ce qui donne une représentation selon un modèle d'*analyse de la variance* [121], pour finalement coder B également de façon binaire et arriver ainsi au modèle *loglinéaire* [181] qui est une représentation des règles d'association. Puis nous terminons l'exposé de ces problèmes par un bilan de l'importance de leur effet sur les règles d'association.

Sommaire

7.1 Les problèmes posés par les liaisons complexes en IA

Nous avons déjà évoqué précédemment le problème de codage dans la partie du chapitre de l'état de l'art concernant les dépendances fonctionnelles. Les chercheurs en intelligence artificielle rencontrent également des problèmes liés aux relations complexes entre attributs dans les tâches de classement, la plus connue étant la construction d'un arbre de décision [246].

Les algorithmes de classement par apprentissage procèdent par sélection d'un ensemble de variables qui forment la partie gauche de la règle [140]. La sélection de ces variables est parfois

délicate comme en attestent de nombreux articles. Citons notamment l'introduction [110] au numéro spécial de Machine Learning consacré à ce problème, ainsi que l'étude du biais de sélection dû aux liens des bases de données relationnelles [132]. Les relations complexes entre attributs ont été mises en évidence lors de comparaisons entre des performances de classifieurs. Par exemple, on a constaté que ce qui rendait l'algorithme Relief [143] plus efficace était sa résistance aux interactions, due à un choix local des attributs. Si le choix des variables intervenant dans les règles ne tient pas compte de ces relations complexes, c'est-à-dire se fait de façon "myope" d'après le terme employé par I. Konenko [145], les performances de certains classifieurs diminuent. Ce sont d'après A. Jakulin [129] les techniques de classement par apprentissage qui utilisent des fonctions linéaires telles que les arbres de décision, la régression logistique, le classifieur naïf bayésien, les Support Vector Machines, le perceptron qui en pâtissent le plus. Parmi les relations complexes gênant la discrimination, il pointe non seulement l'interaction, repérée par de nombreux spécialistes des règles de décision, mais également l'effet Simpson mis en évidence dans les données de l'UCI Repository par Fabris C.C. et A.A. Freitas [76] - cf. annexe B de ce mémoire.

Pour définir ces liens complexes entre 3 variables, ou plutôt entre deux variables explicatives A et B et une variable de classement C, il considère l'association entre l'indépendance marginale entre A et B (quand on ignore C), et l'indépendance conditionnelle entre A et B (pour chaque valeur de C), l'indépendance étant définie de la façon probabiliste : $p_{AB} = p_A p_B$). Et il reprend les 4 possibilités de [5] :

marginal	conditionnel	commentaires :
indépendance	indépendance	inintéressant
indépendance	dépendance	dépendance conditionnelle
dépendance	indépendance	indépendance conditionnelle
dépendance	dépendance	dépendance conditionnelle

TAB. 7.1 – Liens de dépendances entre deux variables booléennes sachant une troisième selon [5]

Et il met l'"interaction" dans le deuxième cas, qui présente un danger en classement quand on sélectionne A et B de façon qu'il appelle "myope" c'est-à-dire sans tenir compte de C, et le paradoxe de Simpson dans le dernier cas, quand la dépendance marginale et la dépendance conditionnelle s'opposent.

7.2 Les problèmes des relations complexes dans les règles d'association

Les relations complexes entre variables sont plus difficiles à repérer quand les variables sont binaires que quand elles sont quantitatives. Pour mieux les appréhender, nous avons d'abord créé des tableaux avec des valeurs numériques pour que les variables vérifient différents types de liaisons.

Pour représenter une liaison entre deux variables quantitatives A et B, nous avons choisi un nuage de points, traversé par deux droites, l'une exprimant la dépendance de B par rapport à A, et l'autre celle A par rapport à B, l'angle des deux droites indiquant la force de la liaison : plus l'angle est petit, plus la liaison est forte ; plus il est proche de 90°, plus elle est faible. Et la liaison est positive (les valeurs selon A et B croissent ou décroissent ensemble) quand les droites montent de la gauche vers la droite et négative (A croit quand B décroît et inversement) dans

le cas contraire. Nous renvoyons le lecteur intéressé par la justification de cette modélisation et par les calculs des équations des droites aux ouvrages traitant des modèles de corrélation et de régression[10, 121].

Un fois cette liaison identifiée, nous avons transformé A en variable binaire, en remplaçant toutes ses valeurs supérieures à 0,5 par 1 et les autres par 0. Le nuage de points se sépare alors en deux droites, une pour A=1 et l'autre pour A=0. La liaison entre A et B se représente cette fois par la droite qui joint les centres de gravité des deux nuages. Plus sa pente est élevée, plus la liaison est forte, et plus elle s'approche de l'horizontale, plus elle est faible, le sens de la liaison étant repéré comme précédemment. Nous renvoyons le lecteur intéressé également par la mise en oeuvre de ce modèle aux ouvrages de statistique.

Puis nous avons codé de la même façon B. Cette fois la seule représentation possible est formée de 4 points formant un carré, tous ceux ayant une valeur de A supérieure à 0.5 et une valeur de B supérieure à 0.5 se retrouvant confondus avec le point de coordonnées A=1 et B=1. Représenter ces 4 points n'apporte rien ; c'est pourquoi nous avons dressé le tableau du nombre de points se trouvant confondus avec chacun des 4 points (A=0, B=0), (A=0, B=1), (A=1, B=0), (A=1, B=1). Ce qui constitue le tableau de contingence des deux variables A et B déjà rencontré quand nous avons défini les indices de qualité de la règle A→B. Nous passons ainsi d'une liaison bien identifiée grâce à sa représentation graphique, à l'ensemble formé des 4 effectifs associés à des règles d'association, dans lequel nous cherchons des indicateurs de cette liaison. Bien sûr, une fois ces indicateurs découverts au sein des règles d'association, il conviendra de les définir de façon formelle. Mais ce n'est pas l'objet de ce paragraphe dans lequel on se contente de faire le tour des liaisons reconnues par les scientifiques comme gênant le raisonnement, en les récrivant simplement en termes de règles d'association.

7.2.1 L'indépendance entre A et B, et une liaison positive

a) A et B sont des propriétés quantitatives

Ce sont les graphiques à gauche de la figure 7.1 pour lesquels A et B sont des propriétés quantitatives. Dans ces trois graphiques, on a représenté les valeurs prises par des sujets selon les propriétés A et B. Chaque sujet est représenté par un losange de coordonnées A et B. Dans le graphique du haut, les deux droites forment un angle proche de 90 degrés, ce qui exprime l'indépendance entre A et B. Dans le graphique du dessous, les droites ont à peine bougé, mais il paraît difficile de dire encore que A et B sont indépendants. En effet, les points se retrouvant proches du cercle, leurs valeurs de A et de B vérifient à peu de choses près l'équation $A^2+B^2=1$. On voit ainsi que ces deux droites nous aident à repérer un seul type de liaison, qui est appelée liaison linéaire, et ce que nous appelons le plus souvent indépendance est en fait une absence de liaison linéaire.

b) B est une propriété quantitative et A une propriété dichotomique

Les graphiques de droite de la figure 7.1 montrent les mêmes points quand A est codée de façon binaire, comme on l'a indiqué plus haut. Dans celui du haut, la valeur du centre de gravité du nuage A=0 est de 0,49 pour B, alors que celle du nuage A=1 est de 0,51 pour B. La droite est quasi-horizontale. Et de même pour le graphique du milieu. Par contre, elle monte pour celui du bas. On voit que la liaison linéaire s'exprime bien par la pente de cette droite. Par contre, la liaison complexe entre A et B qui apparaissait clairement dans le graphique de gauche au centre de la figure 7.1 n'est plus visible dans le graphique de droite. C'est l'effet du codage, qui a fait disparaître des informations qu'on pourra difficilement récupérer en cas de besoin.

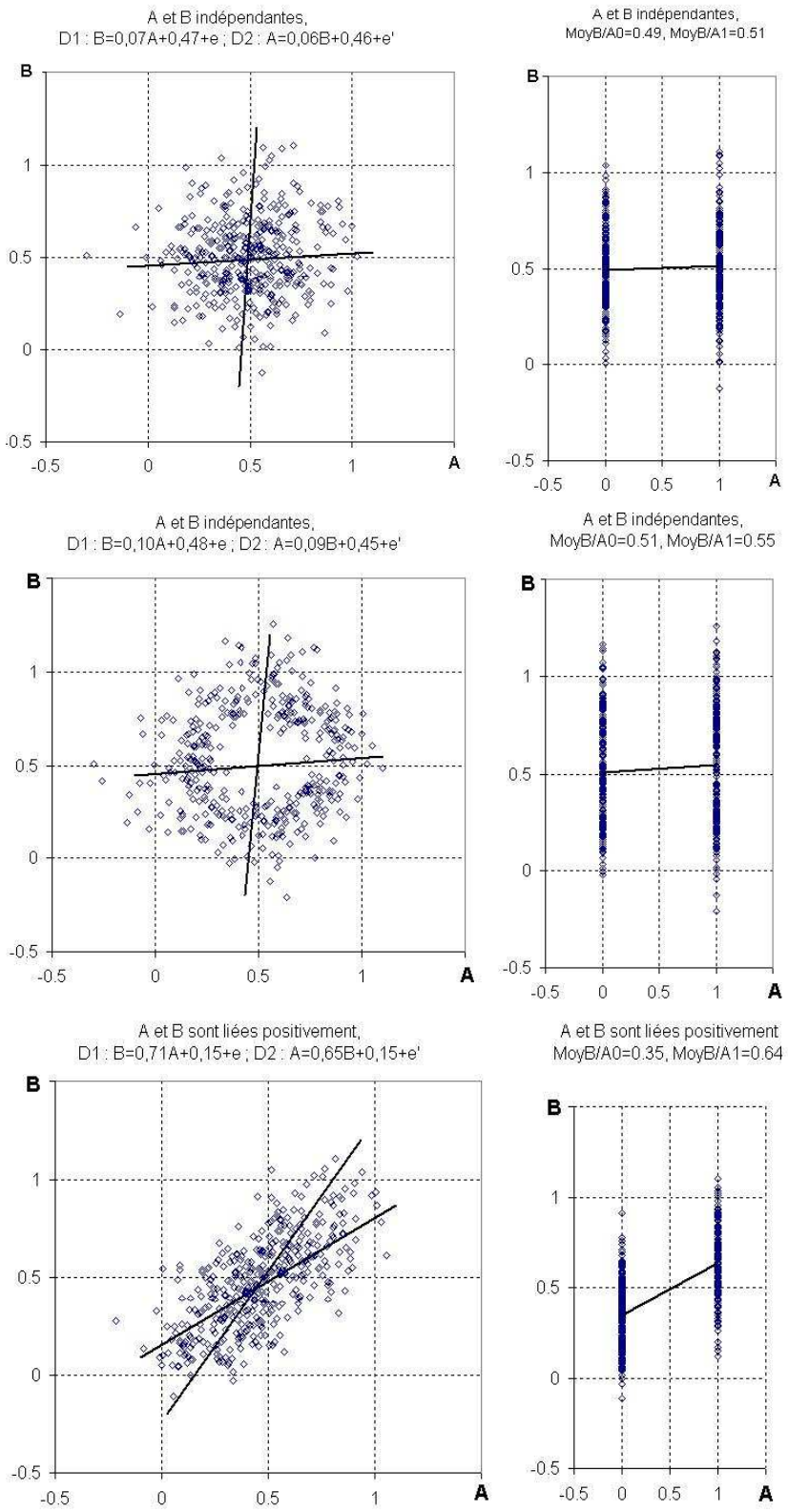


FIG. 7.1 – En haut, l'indépendance "vraie", en dessous, l'indépendance "fausse", en bas une "vraie" liaison positive.

c) A et B sont deux propriétés dichotomiques

Maintenant, on code B comme on a codé A, et dans la table 7.2, on donne les résultats de ce codage pour les données correspondant à chacun des 3 graphiques de la figure 7.1.

Effectifs	B		
A	0	1	Total
0	105	100	205
1	92	103	195
Total	197	203	400

Effectifs	B		
A	0	1	Total
0	98	94	192
1	95	113	208
Total	193	207	400

Effectifs	B		
A	0	1	Total
0	174	58	232
1	36	132	168
Total	210	190	400

TAB. 7.2 – Les données de la figure 7.1 après codage binaire de A et de B.

Ce passage des trois nuages de 400 points (graphiques de gauche) de la figure 7.1 aux 4 effectifs des trois tableaux de la table 7.2 représente le comptage des points situés dans chaque quadrant des graphiques obtenus par découpage selon la verticale et l'horizontale passant par le centre du nuage (ce centre a une valeur de A et de B égale à 0,5). En cas d'indépendance, il y en a autant dans chaque quadrant, et en cas de liaison positive, il y en a beaucoup plus dans les 2 quadrants où A et B sont du même coté de 0,5 (soit, après codage, pour lesquels A=1 et B=1 ou A=0 et B=0) que dans les deux autres.

Précisons que nous nous sommes placée dans un cas où les valeurs de A et de B sont concentrées autour de 0,5 de façon équilibrée, c'est-à-dire qu'il y a à peu près autant de valeurs de chaque côté de 0,5 pour A comme pour B. On attend donc en cas d'indépendance entre A et B un effectif de 100 dans chacune des 4 cases des tableaux de la table 7.2, et en cas de liaison positive, on s'attend non pas à ce que les distributions de valeurs de A ou de B prises séparément changent (les effectifs pour A=1, A=0, B=1, B=0, appelés effectifs marginaux sont fixés), mais que l'effectif des sujets pour lesquels A et B valent 1 augmente ; et ceci d'autant plus que la liaison entre A et B est forte. Comme les effectifs marginaux sont fixés et que ce sont les sommes des 4 cases du tableau prises deux à deux, l'augmentation de cet effectif d'un nombre x entraîne par compensation la diminution d'autant des effectifs des deux cases qui sont voisines de la sienne et l'augmentation d'autant de celui de la case qui lui est opposée [125]. Tout cela apparaît clairement dans la table 7.2. Les deux tableaux de gauche montrent une répartition équilibrée des effectifs des 4 modalités de AxB, chacune étant voisine de 100, alors que dans le dernier tableau les cases "positives" c'est-à-dire correspondant à des valeurs égales de A et de B ont un effectif bien supérieur à celui des cases "négatives", pour lesquelles A et B ont des valeurs complémentaires (c'est-à-dire de somme 1).
L'indépendance de A et de B (ou du moins une faible liaison) se lit ainsi dans les deux tableaux de gauche, et une liaison positive importante dans celui de droite.

d) Linéarité et log-linéarité

La mesure de la liaison repérée dans le tableau croisant les valeurs de A et de B n'est pas de même type que celle correspondant aux deux types de graphiques de la figure 7.1. On a déjà dit que ces modèles sont linéaires. Précisons-le. Dans le premier, on peut exprimer A en fonction de B ou inversement par une équation du type $A = \alpha B + \beta + \epsilon$, ou $B = \alpha A + \beta + \epsilon$ [10], comme c'est indiqué au dessus de chacun des graphiques de gauche de la figure 7.1. Dans le second, en utilisant les moyennes indiquées en haut des graphiques de droite de la figure 7.1, on peut

[125] C'est ce qu'on exprime en disant que le tableau d'effectifs de 4 cases a 1 degré de liberté quand on fixe les marges.

également exprimer B en fonction de A par une équation du type $B = \alpha A + \beta + \epsilon$ [126] [121]. Par contre, pour les tableaux, on utilise un modèle log-linéaire [181] pour exprimer non pas les valeurs de A ou de B, mais celles des logarithmes des effectifs des cases n_{ij} en fonction des marges du tableau. L'intervention des logarithmes permet de transformer un modèle de type multiplicatif en un modèle de type additif comme le modèle linéaire. Malheureusement, le modèle log-linéaire ne peut pas s'exprimer à travers des graphiques aussi expressifs que ceux que nous avons vus pour le modèle linéaire. On peut toutefois exprimer simplement la liaison entre A et B indiquée par ces effectifs en faisant le rapport des deux quotients, en ligne ou en colonne, ce qui donne pour chacun des trois tableaux de la table 7.2 les valeurs 1,18=(103/100)/(92/105), 1,24=(113/94)/(95/98), 11=(132/58)/(36/174). Les valeurs proches de 1 indiquent l'indépendance (mais comme on l'a vu, un certain type linéaire d'indépendance) les valeurs supérieures à 1 représentent une liaison positive et celles entre 0 et 1 une liaison négative, qui correspond au cas où le déséquilibre des cases désavantage celles où A et B ont mêmes valeurs au profit des deux autres. On peut passer au logarithme et on obtient alors les trois valeurs suivantes : 0,16 0,22 et 2,4 ; les deux premières valeurs sont proches de 0, alors que la suivante est nettement positive. Et si la valeur du rapport était comprise entre 0 et 1, le logarithme serait négatif.

Il y a une remarque à faire au sujet de ce rapport. Faire des quotients ou passer au logarithme peut être gênant quand il y a une case avec un effectif nul, comme c'est le cas pour les règles d'association exactes. On peut éviter ce problème en calculant le "produit en croix", c'est-à-dire en remplaçant par exemple (103/100)/(92/105) par 103x105-100x92, ce produit étant voisin de 0 quand le rapport est voisin de 1, positif quand le rapport est supérieur à 1, négatif sinon. Le problème est que ce produit en croix ne se généralise pas aussi bien que le rapport des quotients quand on passe de deux à trois propriétés.

e) Indépendance et "proximité"

Faisons une deuxième remarque. Depuis le début de cette section, nous parlons de valeurs proches : des angles proches de 90 degrés, des droites de pente quasi-nulle, des rapports proches de 1, la proximité représentant l'indépendance entre A et B, les autres valeurs représentant une liaison linéaire, ou log-linéaire. Nous n'avons pas exprimé rigoureusement ce que voulait dire ce terme. Il existe un concept de même usage, en statistique, qui est "ne diffère pas significativement de". Ce concept est défini par des tests d'hypothèses [121][127]. Pour la comparaison de deux moyennes (graphiques de droite de la figure 7.1), le calcul des valeurs "testées" s'appuie non seulement sur les moyennes de valeurs, mais également sur les écart-types des valeurs, et si un certain nombre de conditions sont vérifiées (on suppose que les distributions des valeurs de B selon A=0 et de B selon A=1 ne diffèrent que par leurs moyennes, et qu'elles sont de préférence "normales" , c'est-à-dire qu'elles suivent la loi de probabilité de Laplace-Gauss), la décision de "différence significative ou non" s'obtient en comparant le résultat du calcul à un seuil choisi dans une table. Ces tests sont robustes dans la mesure où une petite entorse aux conditions d'applications ne change pas beaucoup le résultat. Mais ils ne sont plus utilisables pour les données traitées habituellement en fouille de données pour diverses raisons que nous avons exposées dans le chapitre 2. De la même façon d'ailleurs les graphiques ne seront plus utilisables si nous voulons visualiser les liaisons de plusieurs variables. Si nous voulons alors définir le terme "proche" de façon plus rigoureuse sans utiliser les tests d'hypothèses statistiques, il faut définir un seuil de différence entre deux valeurs, que ce soit des mesures d'angle, ou des rapports de quotient. Comme nous

[126] les modèles pour les trois graphiques de droite de la figure 3 sont respectivement $B = 0,02A + 0,49 + \epsilon$, $B = 0,04A + 0,51 + \epsilon$, $B = 0,29A + 0,35 + \epsilon$.

[127] Ces tests font partie des statistiques inférentielles, et leurs principes ont été exposés dans le chapitre 2.

utilisons les graphiques pour illustrer seulement les diverses relations complexes entre propriétés avant de les dichotomiser, nous n'essayons pas de définir ici cette proximité. Nous le faisons dans le chapitre suivant qui concerne le "Nettoyage par des méta-règles des incohérences dues aux relations complexes", les algorithmes de nettoyage nécessitant des définitions rigoureuses.

f) Règle d'association A→B

Avec les valeurs du tableau 7.2, nous pouvons calculer les indices de qualités de la règle A→B dans les 3 cas, qu'on trouve dans le tableau 7.3.

Règle A→B	Tableau de gauche	Tableau du milieu	Tableau de droite
Support	103	113	132
Fréquence	0,26	0,28	0,33
Confiance	0,53	0,54	0,79
Différence	0,02	0,03	0,31

TAB. 7.3 – Les indices des règles A→B extraites des tableaux de la table 7.2.

Nous constatons sur la table 7.3 que pour chaque indice, les valeurs pour les deux premiers tableaux sont assez proches alors que celles pour le dernier sont plus élevées, cet écart étant plus net pour la confiance et la différence que pour le support et la fréquence. Examinons plus précisément chacun de ces indices.

Le support est le nombre de valeurs qui vérifient simultanément A et B, donc ici A=1 et B=1. Nous avons vu précédemment qu'en cas d'indépendance, on attend une valeur de 100, et qu'une augmentation de x indique une liaison d'autant plus forte que x est grand. Cela a pour conséquence que la fréquence de la règle A→B augmente de x/400, que sa confiance passe de 100/200 à (100+x)/200, soit de 0,5 à 0,5+0,01x. Quant à sa différence, elle est obtenue en faisant la différence entre la confiance et le pourcentage de sujets vérifiant B, soit 100/200, et passe de 0 à 0,01x. Notons au passage que du fait de la parfaite symétrie entre A et B dans notre cas, ce qui arrive pour la règle A→B arrive exactement pour les règles B→A, nonA→nonB, nonB→A, qui ont donc même support, même confiance et différence que la règle A→B. Et les règles A→nonB, nonA→B, B→nonA, nonB→A ont des indices égaux variant dans l'autre sens c'est-à-dire passant respectivement de 100 à 100-x pour le support, de 0,25 à 0,25-x/400 pour la fréquence, de 0,5 à 0,5-0,01x pour la confiance et de 0 à -0,01x pour la différence.

Ainsi nous venons d'expliquer que dans le cas d'une distribution marginale donnée de A et de B, l'indépendance entre A et B fait que la règle A→B a des indices de support, fréquence, confiance et différence de valeurs respectives 100 0,25 0,5 et 0, et que la liaison positive entre A et B fait qu'ils augmentent tous. Si on retourne maintenant au tableau 7.3, on peut conclure que les augmentations des indices de la règle A→B entre la première colonne et la seconde, correspondant à des propriétés indépendantes, doivent être jugées négligeables, alors que les augmentations entre ces deux colonnes et la dernière doivent être jugées importantes. Notons au passage la difficulté de prendre une telle décision (indépendant/lié, écart négligeable/important) sans s'appuyer sur les tests d'hypothèses en statistique.

7.2.2 L'ajout de C ne modifie rien à la règle A→B

a) L'absence d'interaction apparaît sur les graphiques

Dans la figure 7.2 le graphique du haut contient les mêmes données que celui du haut de la figure 7.1 et ceux du bas sont associés de la même façon. Celui du milieu n'a pas été repris. On a

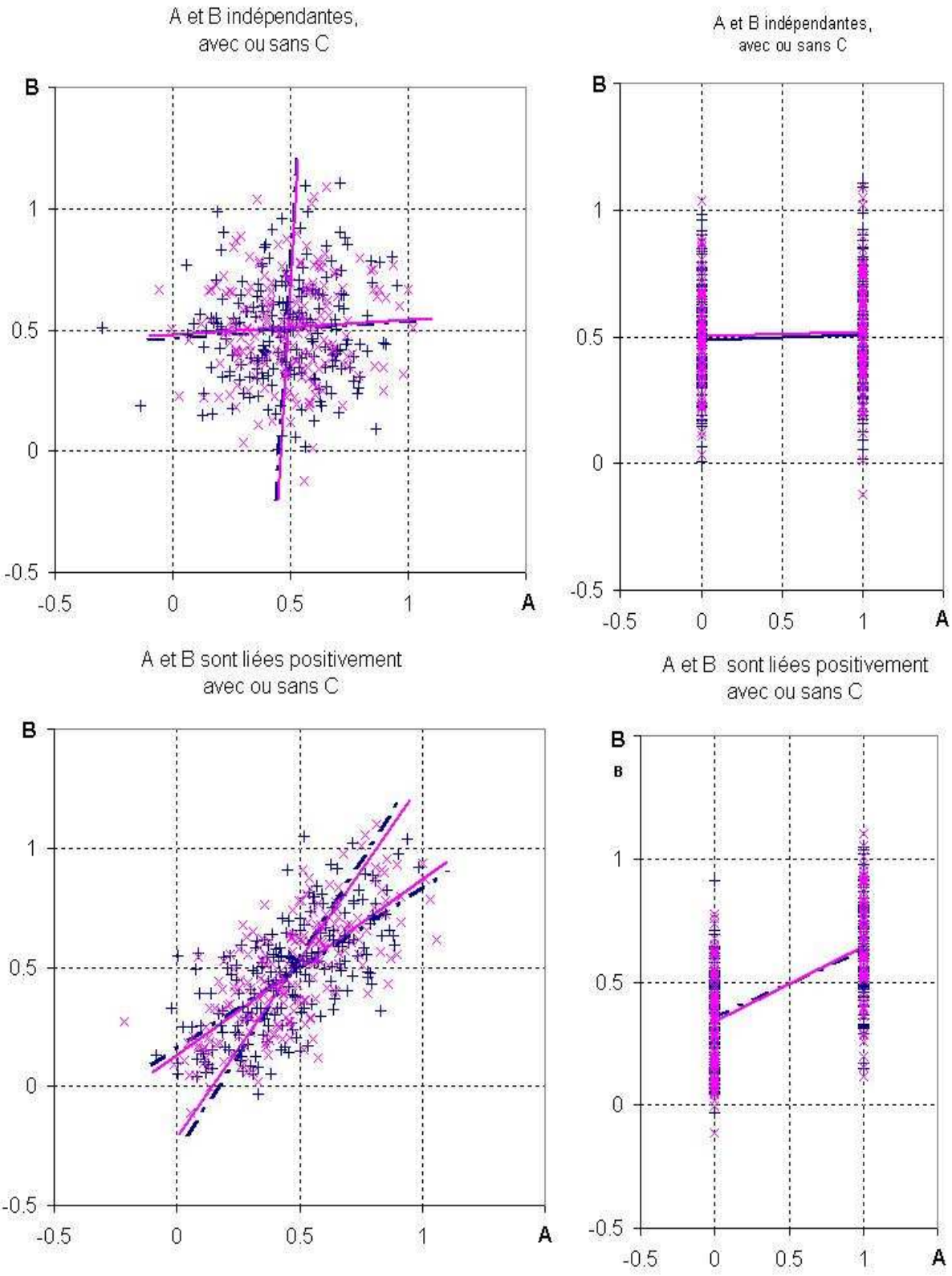


FIG. 7.2 – C ne modifie pas la liaison entre A et B

ajouté une nouvelle propriété C pour tous ces points avec deux valeurs C=0 (points indiqués par le symbole 'x' rouge clair) et C=1 (points indiqués par le symbole '+' bleu foncé) de telle façon que la répartition de ces valeurs de C soient indépendantes de celles de A et de B. On a ainsi obtenu deux nuages de points superposés dans chaque graphique de gauche. Pour celui du haut, les 2 droites de régression de chaque nuage, le clair et le foncé, dessinent un angle presque droit, et pour celui du bas, l'angle des droites de chacun de ces deux nuages est bien inférieur à 90 degrés. De plus, les angles des droites claires et des droites foncées sont très proches, et voisins de ceux qu'on a obtenu quand il n'y avait pas de variable C (dans la figure 7.1) ce qui permet de dire que l'intervention de C n'a pas changé la liaison entre A et B. On voit la même chose dans les graphiques de droite de la figure 7.2. Les effectifs correspondants se trouvent dans le tableau 7.4.

A	B	C	Graphique du haut	Graphique du bas
0	0	0	52	88
0	0	1	53	86
0	1	0	48	26
0	1	1	52	32
1	0	0	44	15
1	0	1	48	21
1	1	0	56	71
1	1	1	47	61
Total			400	400

TAB. 7.4 – Les effectifs des graphiques de la figure 7.2.

b) L'absence d'interaction apparaît sur les règles A→B, AC→B, A nonC→B

Examinons maintenant l'incidence de C sur la règle A→B. Dans le tableau 7.5 figurent les indices des règles qu'on peut extraire des données du tableau 7.4.

	Graphique du haut			Graphique du bas		
Règle	Fréquence	confiance	Différence	Fréquence	confiance	Différence
A → B	0,26	0,53	0,02	0,33	0,79	0,31
AC → B	0,12	0,49	-0,01	0,15	0,74	0,27
A nonC → B	0,14	0,56	0,05	0,18	0,83	0,35

TAB. 7.5 – Les indices des règles extraites des tableaux de la table 7.4.

Pour le graphique du haut comme pour celui du bas, on voit dans la table 7.5 que la règle A→B a pour fréquence, confiance et différence les mêmes valeurs qu'elle avait avant qu'on ajoute la propriété C, ce qui correspond à une indépendance entre A et B sans tenir compte de C. Examinons la règle AC→B pour le graphique du haut qui correspond à l'indépendance entre A et B. La fréquence correspond à la proportion de sujets vérifiant A, B et C. On a vu qu'elle est environ de 0,25 pour AB en cas d'indépendance de A et de B. Comme les valeurs de C=0 et C=1 sont réparties régulièrement, A, B et C étant indépendants, on attend pour ABC une fréquence de $0,125 = 0,5^3$. Et pour la confiance, qui est le rapport des valeurs pour ABC et AC, c'est une valeur de $0,5 = 0,5^3/0,5^2$ qu'on attend. On voit que ce sont bien des valeurs proches de celles que nous avons. Par rapport à la règle A→B, le support a été divisé par 2 et la confiance et la

différence sont à peu près égales. On obtient le même type de relation entre la règle A→B et la règle AC→B pour le graphique du bas, et pour la règle A nonC→B également.

On constate ainsi que l'ajout d'une variable indépendante C, distribuée de façon équilibrée, ne modifie pas les indices de qualité de la règle A→B, mais on constate que pour la règle AC→B, comme pour la règle A nonC→B, la fréquence diminue de moitié alors que la confiance et la différence restent identiques. Cela a comme effet que si on n'extrait que les règles ayant un support supérieur à un seuil proche de celui de la règle A→B, on a très peu de chances de garder la règle AC→B. Mais si la proportion de points du support de AB (c'est-à-dire vérifiant A=1 et B=1) qui ont également 1 pour valeur de C augmente, le support de la règle AC→B se rapproche de celui de la règle A→B, et cela d'autant plus que que cette proportion se rapproche de 100%, la confiance et la différence pouvant également augmenter, alors que c'est l'inverse pour les indices de la règle A nonC→B. Cette augmentation (ou diminution) peut se produire en cas de déséquilibre de la distribution des effectifs de C, ou en cas de liaison de C avec AB. Et cette fois, si on ne garde que les règles dont le support et la confiance dépassent des seuils, il est possible que la règle AC→B soit extraite alors que la règle AC→B ne l'est pas.

7.2.3 L'ajout de C modifie la règle A→B

a) La présence d'interaction apparaît sur les graphiques

A gauche de la figure 7.3, on a représenté en haut et en bas les mêmes types de graphiques que dans les deux figures précédentes, mais la propriété C n'a pas été rajoutée de façon indépendante des deux autres propriétés. Le plan contient quatre quadrants, qu'on numérote en général dans le sens contraire des aiguilles d'une montre, le premier contenant tous les points pour lesquels A=1 et B=1, le second les points tels que A=0 et B=1, le troisième A=0 et B=0 et le quatrième A=1 et B=0. Les points pour lesquels C=1 (symbole "+" bleu foncé) se trouvent essentiellement dans les quadrants 1 et 3, alors que les points pour lesquels C=0 (symbole "x" rouge clair) sont essentiellement dans les quadrants 2 et 4. Les droites de régression foncées font un angle bien inférieur à 90 degrés, ce qui montre une liaison entre A et B pour C=1, et l'angle est de même valeur entre les droites claires, ce qui montre une liaison entre A et B pour C=0. Toutefois, le type de liaison n'est pas le même, compte tenu de l'orientation des droites. Pour C=1, la liaison est positive, alors que pour C=0, elle est négative. On a représenté en pointillés les droites de régression représentant la liaison entre A et B sans tenir compte de C ; leur angle étant proche de 90 degrés, ils sont encore indépendants. Ce graphique fait apparaître l'influence de C sur la relation entre A et B. Sur le graphique de droite elle apparaît de façon tout aussi nette. La droite claire descend alors que la droite foncée monte, la droite en pointillés étant presque horizontale. Ces deux graphiques en haut de la figure 7.3 expriment une liaison complexe entre A et B. Cette liaison n'agit pas sur A tout seul, car il y a à peu près autant de points pour lesquels A=0 et C=1 que pour lesquels A=1 et C=1, de même pour C=0. Elle n'agit pas non plus sur B seul, elle agit sur le croisement de A et de B. Cet effet particulier fait partie des interactions entre variables, dont nous avons déjà parlé dans ce chapitre[128].

b) La présence d'interaction apparaît sur les règles A→B, AC→B, A nonC→B

Dans le tableau 7.6, on a écrit les effectifs après le codage binaire habituel de A et de B, qui nous ont permis de calculer les indices de qualité des règles A→B et AC→B. Dans le tableau 7.7 on constate que pour le graphique du haut, la règle A→B a encore les même valeurs, montrant

[128] Les interactions sont exposées de façon plus théorique en annexe C.

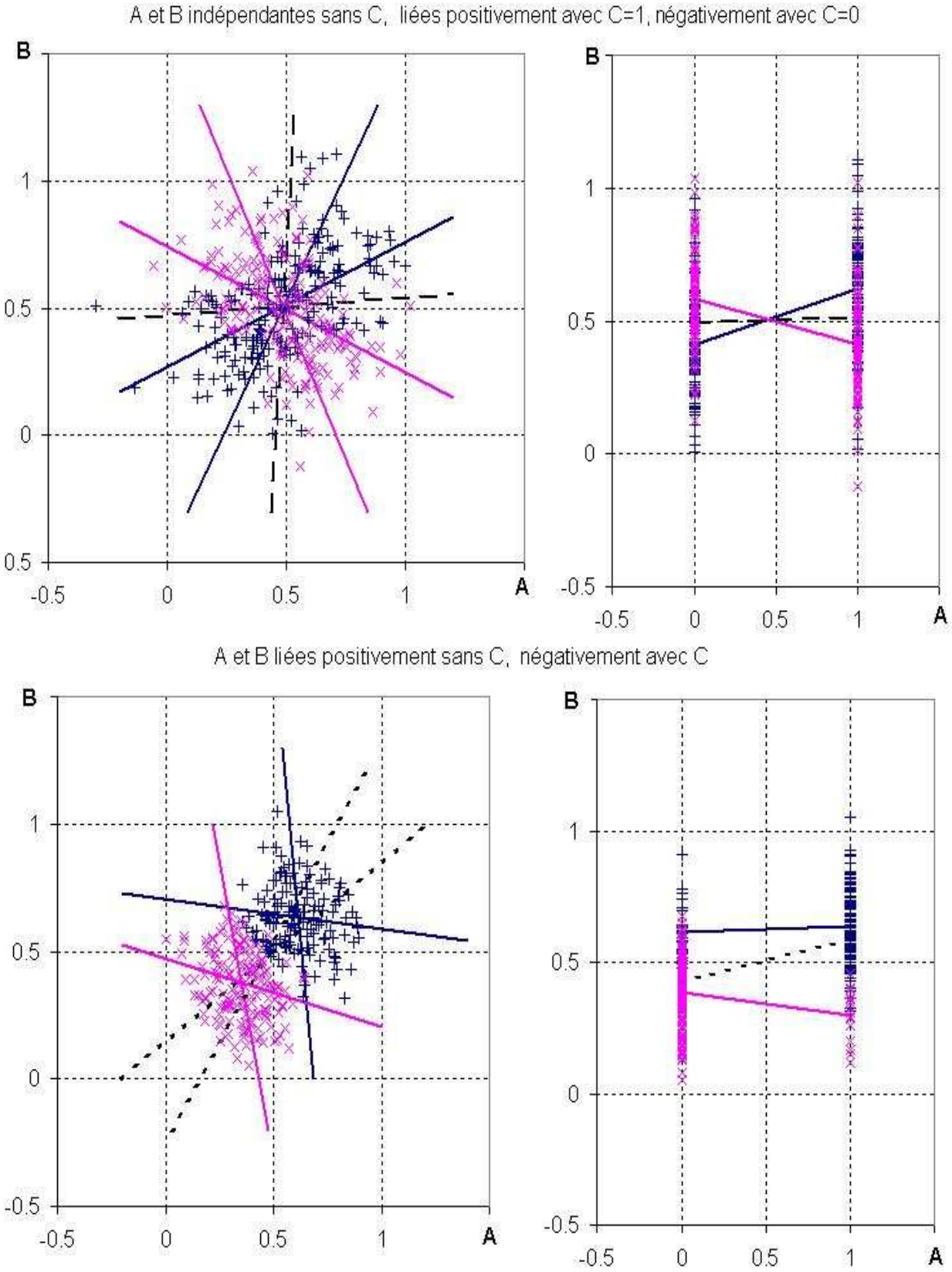


FIG. 7.3 – C modifie la liaison entre A et B

A	B	C	Graphique du haut
0	0	0	34
0	0	1	71
0	1	0	65
0	1	1	35
1	0	0	66
1	0	1	26
1	1	0	32
1	1	1	71
Total			400

Graphique du bas
119
2
36
27
19
17
0
105
325

TAB. 7.6 – Les effectifs des graphiques de la figure 7.3.

	Graphique du haut			Graphique du bas		
Règle	Fréquence	confiance	Différence	Fréquence	confiance	Différence
A → B	0,26	0,53	0,02	0,32	0,74	0,23
AC → B	0,18	0,73	0,22	0,32	0,86	0,34
A nonC → B	0,08	0,33	-0,18	0,00	0,00	-0,50

TAB. 7.7 – Les indices des règles extraites des tableaux de la table 7.6.

l'indépendance entre A et B sans tenir compte de C, la règle AC→B a des valeurs pour tous les indices supérieures à celles qu'elle avait quand C était indépendant de A et de B (dans le tableau 7.5), et la règle A nonC→B a des valeurs pour tous les indices inférieures à celles qu'elle avait quand C était indépendant de A et de B. Cela correspond à la lecture qu'on a faite des graphiques et des indicateurs de liaison associés tels que l'angle des droites de régression, et la pente de la droite des moyennes. L'effet des interactions sur le jeu de règles d'association est montré pour trois variables dans le cas général dans l'annexe C intitulée "Interactions entre variables binaires et Règles d'Association".

c) Le paradoxe de Simpson apparaît sur les graphiques

Dans le graphique du bas à gauche de la figure 7.3, l'effet est beaucoup plus surprenant. Les deux nuages de points, celui pour C=0 et celui pour C=1 montrent une liaison négative entre A et B, alors que le nuage total montre une liaison positive entre A et B (droites en pointillés), et on retrouve un effet surprenant dans le graphique en bas à droite de la même figure, une pente négative pour C=0, une pente quasi-nulle pour C=1, et une pente positive quand on ne tient pas compte de C. C'est le paradoxe de Simpson, qui ne se produit que pour des valeurs d'effectifs déséquilibrées comme on le voit dans les deux tableaux de droite de la table 7.8[129] . Contrairement à l'interaction, il n'apparaît pas dans la plupart des modèles statistiques, qui l'éliminent en posant des conditions sur les effectifs, sur les distributions des résidus aux modèles. S'il apparaît dans les règles d'association, on s'attend à avoir A→B avec des valeurs d'indices plus élevées que celles en cas d'indépendance, et AC→B et A nonC→B avec des valeurs d'indices plus faibles.

d) Les indices courants ne permettent pas l'apparition du paradoxe de Simpson dans les règles A→B, AC→B, A nonC→B

[129]Le "déséquilibre" d'effectifs produisant l'effet Simpson est exposé de façon plus formelle en annexe B

Dans la partie droite du tableau 7.7 on peut voir que la règle AC→B a tous ses indices supérieurs ou égaux à ceux de la règle A→B. Ainsi le paradoxe de Simpson n'apparaît pas dans ces règles d'association, dont la qualité est mesurée par la fréquence, la confiance et la différence. Il apparaît toutefois un effet voisin de celui qu'on a vu dans la partie gauche du tableau 7.7 correspondant à une interaction, c'est-à-dire que les indices des règles AC→B et A nonC→B expriment des liaisons différentes. Si la règle AnonC→B va bien dans le sens attendu d'une liaison négative, on peut se demander pourquoi la règle AC→B indique une liaison contraire à celle attendue. Nous montrons dans les tables 7.8 et 7.9 d'où provient cette différence[130].

	B0	B1
A0C0	119	36
A0C1	2	27
A1C0	19	0
A1C1	17	105

A→B

	B0	B1
A0C0,A0C1	121	63
A1C0,A1C1	36	105

A→B/C=1

	B0	B1
A0C1	2	27
A1C1	17	105

A→B/C=0

	B0	B1
A0C0	119	36
A1C0	19	0

TAB. 7.8 – Décomposition et recomposition des effectifs selon la règle totale et des règles partielles.

AC→B

	B0	B1
A0C0,A0C1,A1C0	140	63
A1C1	17	105

AnonC→B

	B0	B1
A0C0,A0C1,A1C1	138	168
A1C0	19	0

TAB. 7.9 – Recomposition des effectifs selon les règles classiques à 3 variables.

Dans le tableau de gauche de la table 7.8, nous avons recopié les 8 effectifs du tableau de droite de la table 7.6, en utilisant la notation A0C1, par exemple vérifiée par les sujets ne vérifiant pas A (A0) et vérifiant C (C1). A l'intersection de la ligne A0C1 et B1, il y a 27 sujets, ce sont ceux qui vérifient le motif A0C1B1. Pour la règle A→B, ces sujets sont regroupés en "oubliant" C, c'est-à-dire que les lignes A0C0 et A0C1 ont été regroupées en totalisant les sujets pour B0, comme pour B1 ans le tableau suivant de la table 7.8. Ainsi nous avons 36+27=63 sujets pour la ligne A0C0, A0C1 et la colonne B1, c'est-à-dire pour A0B0, sans tenir compte de C. L'indice de liaison de ce tableau est le rapport (105/63)/(36/121)=5,6, de logarithme 1,72 qui indique une liaison positive. Les 8 effectifs du tableau de gauche sont séparés en 4 effectifs pour C=1, et 4 effectifs pour C=0, ce qui donne les deux tableaux de droite, l'un pour la règle A→B sachant C=1, et l'autre pour la règle A→B sachant C=0. Les indices correspondants de fréquence| confiance| différence se calculent d'après ces tableaux et on obtient 0,7| 0,86| -0,01 pour C=1 et 0| 0| -0,2 pour le second. On ne peut plus comparer aux indices habituels de la règle en cas d'indépendance, car on a coupé l'effectif total en 2 parties selon les valeurs de C. Si on compare aux effectifs calculés de la façon habituelle en cas d'indépendance 0,70| 0,87| 0 pour la première et 0,02| 0,21| 0 pour la seconde, on trouve qu'ils sont proches pour la première règle et inférieurs pour la seconde, donc qu'il indiquent l'indépendance dans le cas C=1 et une liaison négative dans le cas C=0, alors que la liaison est positive si on ne tient pas compte de C. Le paradoxe apparaît donc avec ce mode de calcul, même s'il est moins accentué que pour le graphique de gauche de la

[130] Pour le lecteur préférant un exposé plus général, nous avons démontré en annexe B qu'avec les indices courants on ne peut pas aboutir à un jeu de règles du type AC→nonB, A nonC→nonB et A→B.

figure 7.3 (il l'est quand même plus que dans le graphique de droite de cette figure). Mais ce n'est pas ainsi qu'on calcule habituellement les indices de qualité des règles. Le calcul habituel s'appuie sur un répartition différente des effectifs qui est donnée dans la table 7.9 et qui contient tous les sujets. Pour cela, on oppose une ligne du tableau de gauche de la table 7.8 à la fusion des trois autres lignes. Ce qui donne pour les trois indices les valeurs trouvées précédemment qui indiquent une liaison positive pour la règle AC→B et une liaison négative pour la règle A nonC→B. Pour ces données, issues d'un unique tableau Sujets×Propriétés, le paradoxe de Simpson disparaît et laisse place à un effet ressemblant à une interaction. Nous avons montré en annexe B que cette disparition du paradoxe de Simpson a lieu dans le cas général pour la plupart des indices de qualité actuels des règles. Par contre, si on dispose de deux tableaux distincts de données, il peut apparaître dans le jeu de règles obtenu par fusion. Nous ne détaillons pas cette apparition car nous nous limitons dans notre mémoire à un seul tableau de données.

7.3 Le type de liaison indiqué par une règle d'association

Nous avons vu que si on part de deux variables quantitatives A et B linéairement indépendantes distribuées de façon "équilibrée", on retrouve cette indépendance dans le tableau de contingence de celles qu'on a déduites de A et B par codage binaire, et dans les indices les plus courants des règles d'association. De même si les variables sont liées linéairement. Ainsi, dans des cas de variables bien distribuées, les indices des règles d'association peuvent permettre de repérer certaines liaisons linéaires. Le fait de ne garder, comme il est habituel, que les règles au dessus d'un certain seuil de support, autorise le fait que, à distributions de valeurs équivalentes et en cas d'absence d'interaction, seules les règles entre deux variables liées par une liaison positive suffisamment forte sont conservées. Ces hypothèses sont celles à la base des analyses factorielles et de la typologie, méthodologies les plus utilisées de l'analyse de données[19, 118, 205, 84], qui permettent de regrouper des variables ayant de fortes cooccurrences deux à deux.

Quand on ajoute une troisième variable C, de façon équilibrée, la règle AC→B est de moindre qualité que la règle A→B si C est indépendante de A et B, mais qu'elle peut être de meilleure qualité si C est en interaction positive avec A et B, c'est-à-dire renforce en quelque sorte les valeurs de A et de B identiques. L'ajout d'une troisième variable a pour effet également de faire apparaître des difficultés d'interprétation de l'ensemble des deux règles {A→B, AC→ B} en cas d'interaction et d'effet Simpson. A part les chercheurs qui ont consacré une partie de leurs recherches à ce problème, comme c'est le cas de [245, 231, 51, 232], de l'Université Simon Fraser des États-Unis et de Régis Gras et ses collaborateurs dans le cadre de leur définition d'indices de qualité, ce problème n'est pas signalé, à notre connaissance, au sein des jeux de règles d'association par les autres chercheurs en fouille de données. Cela peut s'expliquer par le fait que l'utilisation des propriétés négatives dans les règles est rare, et les règles construites sur plus de deux propriétés sont plus facilement éliminées que les autres par le choix d'un seuil de support. Ainsi les difficultés d'interprétation disparaissent, si quand la règle A→B est extraite, les règles qui risquaient de la contredire telles que AC→nonB ou A nonC→nonB, ou de la rendre sans valeur telles que AC→B ou A nonC→B, ne sont pas extraites. Si on choisit un tel mode d'extraction, il convient, à notre avis, de contrôler chaque règle avant de l'interpréter, sinon le jeu de règles que nous fournissons à l'expert du domaine dont sont issues les données risque fort d'être une représentation de celles-ci entachée d'erreurs. C'est dans ce but que nous avons créé des méta-règles de nettoyage, qui sont exposées dans le chapitre suivant de ce mémoire.

Et bien sûr, nous n'avons pas fait le tour des problèmes que pose l'interprétation d'un jeu de règles d'association. Non seulement nous n'avons pas étudié tous les cas possibles pour 3

variables, mais nous n'avons pas abordé le cas de plus de trois variables. De plus les données sont très souvent déséquilibrées, comme c'est le cas en fouille de textes [43]. Et il est fort probable que d'autres problèmes d'interprétation que ceux du modèle linéaire ou loglinéaire n'aient pas encore été repérés du fait de la jeunesse de ce nouveau modèle des données qu'est le jeu de règles d'association[131].

[131] Je tiens à remercier au passage mon collègue psychiatre André M. Masson de l'Université de Liège, pour les longues heures qu'il a passées à essayer d'interpréter les jeux de règles d'association issus de ses données [174]. Sans ses remarques pertinentes, une grande partie de ces problèmes d'interprétation m'aurait échappé

8

Une solution : nettoyage par des méta-règles des incohérences dues aux liaisons complexes

Sommaire

8.1 Introduction

Nous avons vu dans les chapitres précédents qu'un jeu de règles d'association extrait de données peut contenir des incohérences du point de vue de "la logique du sens commun". Ces incohérences sont par exemple dues à des défauts de transitivité quand on extrait des règles approximatives, à la présence de propriétés négatives à coté des propriétés positives, aux interactions, etc. Un jeu de règles d'association ne peut pas être fourni à un expert s'il comporte de tels défauts. Nous proposons quelques méta-règles permettant de repérer et d'éliminer certains de ces défauts par élagage, c'est-à-dire en retirant des règles qui, combinées aux règles restantes, font apparaître ces défauts.

Toutes ces notions seront illustrées sur le tableau 4.1 (chapitre 4)

8.2 La technique de nettoyage

Nous allons nous limiter ici à quelques-unes des contradictions et redondances qui peuvent apparaître en considérant un sous-ensemble donné {r1,r2} de 2 règles prises dans l'ensemble $\mathcal{R}$

de toutes les règles. Pour chaque type de problème, nous créons une méta-règle de nettoyage mr_i de la façon suivante : mr_i est une application de l'ensemble des parties à 2 éléments de $\mathcal{R}$, que nous noterons $\mathbb{P}_2(\mathcal{R})$ vers l'ensemble $\mathbb{P}_{0,1,2}(\mathcal{R})$, l'ensemble des parties d'au plus 2 éléments de $\mathcal{R}$, qui à tout élément {r1,r2} associe l'ensemble des règles rejetées , c'est-à-dire {r1,r2}, {r1}, {r2}, $\emptyset$, selon qu'on élimine les 2 règles, l'une des deux, ou aucune.
Pour les 3 méta-règles exposées dans cet article, c'est la règle de moindre qualité qui est éliminée, quand on peut la déterminer. Ce qui n'est pas toujours le cas. En effet, nous utilisons plusieurs indices pour mesurer la qualité des règles, et quand ils ne s'accordent pas, nous ne pouvons pas comparer les règles, elles sont dites "non-comparables". De plus, quand les indices des deux règles diffèrent de peu, nous ne cherchons pas à établir que l'une l'emporte sur l'autre, nous disons qu'elles sont "similaires". Le choix des règles à éliminer dans ces deux derniers cas, varie selon le problème à traiter, et selon l'exigence de l'expert. Pour l'expert exigeant, les règles similaires et non comparables sont nettoyées par la méta-règle de façon identique (ce cas est noté *exigence* dans la définition de la méta-règle), alors que pour l'expert moins exigeant, la méta-règle nettoiera les règles similaires, mais pas les règles non comparables (ce cas est noté *non exigence*).

Nous allons d'abord exposer notre façon de comparer deux règles, puis les trois méta-règles choisies, et enfin comment nous les avons fait fonctionner ensemble sur le jeu de règles.

8.2.1 Comparaison de 2 règles

On compare deux règles à l'aide de leurs indices de qualité. On prend ces indices parmi les divers indices décrits dans l'état de l'art ci-dessus[132]. Appelons $\mathcal{I}$ ce sous-ensemble d'indices. A chaque indice I de $\mathcal{I}$, nous associons un écart e(I), qui peut être fixé au départ (la plus petite distance visible) où dépendre de l'ensemble des données (ex. l'écart-type des valeurs prises par cet indice)[133].

Définition 8.2.1. *Différence entre deux règles r1 et r2 selon un indice I élément de $\mathcal{I}$ auquel est associé un écart constant e(I) élément de $\mathbb{R}^+$:*

$$d_I(r1,r2) = \begin{cases} -1 & \text{si} \quad I(r1) - I(r2) < -e(I) \\ 0 & \text{si} \quad -e(I) \leq I(r1) - I(r2) \leq e(I) \\ 1 & \text{si} \quad e(I) < I(r1) - I(r2) \end{cases}$$

Tous les indices considérés ici sont fonction croissante de la qualité de la règle, ce qui nous permet de ne pas les distinguer dans $\mathcal{I}$.

Définition 8.2.2. *Comparaison des qualités de 2 règles selon l'ensemble d'indices :*

1. *r1 et r2 sont* ***similaires*** *si les différences sont nulles pour tous les indices de $\mathcal{I}$.*
2. *r1 l'****emporte*** *sur r2 si les différences $d_I(r1,r2)$ sont positives ou nulles pour tous les indices et si leur somme est supérieure ou égale à 1. (et r2 sur r1 en cas de négativité)*
3. *r1 et r2 sont* ***non-comparables*** *si pour au moins deux indices les différences sont de signes contraires.*

[132] Ce choix d'indices sera en fait laissé à l'expert.

[133] On pourrait même, pour le calcul de cet écart, prendre en compte la distribution statistique des indices sur l'ensemble des règles extraites, ce qui donnerait un écart variable, par exemple pour la confiance prendre un écart plus petit au voisinage de 1, et plus grand au voisinage de 0.

Propriété 8.2.1. *La relation de similarité est réflexive, alors que celles de supériorité, et de non-comparabilité sont antiréflexives. Les relations de similarité et de non-comparabilité sont symétriques, alors que celle de supériorité est antisymétrique*[134].

La preuve de cette propriété découle de façon immédiate des définitions des relations de comparaison.

Propriété 8.2.2. *Les relations de similarité, de supériorité, et de non-comparabilité ne sont pas transitives.*

En voici la preuve sur un exemple de 5 règles :

Exemple 2. *Comparaisons 2 à 2 de 5 règles selon 3 indices.*
Si l'on considère un ensemble l'ensemble $\mathcal{I}$ de 3 indices qui pourraient être le support, la confiance, la différence[135]*, d'écarts respectifs 10 ; 0,1 ; 0,2 et les 5 règles suivantes avec les valeurs correspondantes des indices :*
$r1(10; 0,77; -0,08)$, $r2(17; 0,70; 0,10)$, $r3(25; 0,59; 0,26)$, $r4(29; 0,65; 0)$, $r5(15; 0,72; -0,15)$
on obtient les résultats suivants figurant dans le tableau n°2 :

	r1	r2	r3	r4	r5
r1	(0 ;0 ;0)	(0 ;0 ;0)	(-1 ;1 ;-1)	(-1 ;1 ;0)	(0 ;0 ;0)
r2	(0 ;0 ;0)	(0 ;0 ;0)	(0 ;1 ;0)	(-1 ;0 ;0)	(0 ;0 ;1)
r3	(1 ;-1 ;1)	(0 ;-1 ;0)	(0 ;0 ;0)	(0 ;0 ;1)	(0 ;-1 ;1)
r4	(1 ;-1 ;0)	(1 ;0 ;0)	(0 ;0 ;-1)	(0 ;0 ;0)	(1 ;0 ;0)
r5	(0 ;0 ;0)	(0 ;0 ;-1)	(0 ;1 ;-1)	(-1 ;0 ;0)	(0 ;0 ;0)

TAB. 8.1 – Comparaison des 5 règles 2 à 2

On peut observer à la lecture de ce tableau la non-transitivité des 3 relations :
- *r2 et r1 sont similaires, ainsi que r1 et r5, mais r2 et r5 ne sont pas similaires.*
- *r2 l'emporte sur r3, r3 sur r4, mais r2 ne l'emporte pas sur r4, au contraire, c'est r4 qui l'emporte sur r2.*
- *r3 et r1 sont non comparables, ainsi que r1 et r4, mais r3 l'emporte sur r4.*

8.2.2 Méta-règle n°1 : contradiction d'ordre 1

Imaginons un jeu de règles dans lequel nous avons les 2 règles suivantes "Les filles de la ville réussissent et s'inscrivent" et "Les filles de la ville échouent et s'inscrivent". Toute personne douée du bon sens élémentaire relève là une contradiction, du fait du changement de modalité d'un attribut de la partie droite pendant que les autres ne changent pas. La méta-règle mr_1 a pour but de supprimer ce type de contradiction.

Nous formalisons ainsi l'écriture de ces deux règles, liant quatre attributs de 2 modalités chacun :

r1 : G1L1→R1I1 r2 : G1L1→R0I1

les quatre attributs étant les suivants :

[134] Par relation antisymétrique, nous entendons ici une relation R telle qu'on ne puisse jamais avoir simultanément aRb et bRa.

[135] La "différence" de la règle A→B est l'écart entre sa confiance et le pourcentage de sujets vérifiant B parmi l'ensemble des N sujets

- genre : G1="fille", G0="garçon"
- lieu : L1="ville", L0="campagne"
- réussite : R1="réussite", R0="échec"
- inscription : I1="s'inscrivent", I0="ne s'inscrivent pas"

Définition 8.2.3. *Soit {r1,r2} un élément de $\mathbb{P}_2(\mathcal{R})$, on dit qu'il est incohérent pour la méta-règle mr_1 si r1 et r2 ne diffèrent que par la modalité d'un attribut en partie droite. Il est cohérent pour mr_1 dans le cas contraire. L'ensemble correspondant $mr_1(\{r1, r2\})$ de règles à supprimer est défini par :*

$$\begin{cases} \emptyset & \text{si } \{\text{r1, r2}\} \textit{ cohérent} \\ \emptyset & \text{si } (\{\text{r1, r2}\} \textit{ incohérent}) \text{ et } (\text{r1 et r2 } \textit{non comparables}) \text{ et } (\textit{non exigence}) \\ \{r1, r2\} & \text{si } (\{\text{r1, r2}\} \textit{ incohérent}) \text{ et } (\text{r1 et r2 } \textit{similaires}) \\ \{r1, r2\} & \text{si } (\{\text{r1, r2}\} \textit{ incohérent}) \text{ et } (\text{r1 et r2 } \textit{non comparables}) \text{ et } (\textit{exigence}) \\ \{r1\} & \text{si } (\{\text{r1, r2}\} \textit{ incohérent}) \text{ et } (\text{r2 } \textit{l'emporte sur } \text{r1}) \\ \{r2\} & \text{si } (\{\text{r1, r2}\} \textit{ incohérent}) \text{ et } (\text{r1 } \textit{l'emporte sur } \text{r2}) \end{cases}$$

Par exemple, si l'on considère les règles r1 : G1L1→R1I1, r2 : G1L1→R0I1, r3 : G1L1→R1I0, r4 : G1L1→R0I0, alors on aura comme ensembles incohérents pour la méta-règle n°1 les 4 ensembles {r1,r2} , {r1,r3}, {r2,r4}, {r3,r4}. Si l'expert est peu exigeant et si le résultat de leurs comparaisons 2 à 2 est celui figurant dans le tableau n°2, r1 et r2 seront supprimées, car similaires, donc $mr_1(\{r1, r2\}) = \{r1, r2\}$, r1 et r3 sont non comparables, donc $mr_1(\{r1, r3\}) = \emptyset$, r4 l'emporte sur r2, donc $mr_1(\{r2, r4\}) = \{r2\}$, et r3 l'emporte sur r4, donc $mr_1(\{r3, r4\}) = \{r4\}$. En réunissant ces ensembles, on constate que les règles r1, r2 et r4 sont éliminées.

8.2.3 Méta-règle n°2 : redondance d'ordre 1

Considérons un ensemble de règles contenant de 2 règles comme celles-ci :
r1 : G1L1→R1I1, "Les filles de la ville réussissent et s'inscrivent"
r2 : G1L0→R1I1 "Les filles de la campagne réussissent et s'inscrivent".

La règle r1 seule apporte une information, mais l'ajout de la règle r2, au lieu d'apporter une information supplémentaire, fait perdre une partie de la valeur de cette information, celle qui porte sur la localisation. La méta-règle mr_2 a pour but de supprimer ce type de redondance.

Définition 8.2.4. *la définition de la méta-règle mr_2 reprend exactement celle de la méta-règle mr_1 dans laquelle on a remplacé "partie droite" par "partie gauche".*

8.2.4 Méta-règle n°3 : contradiction local/global d'ordre ≥1

Prenons l'exemple des deux règles suivantes :
r1 : G1L1→R1I1, "Les filles de la ville réussissent et s'inscrivent"
r2 : G1→R0I1, "Les filles échouent et s'inscrivent".
La règle r1 apporte une information dans une situation particulière (le lieu étant la ville), et la règle r2, plus générale, la contredit. Pour apporter une correction, nous allons nous inspirer de la façon de faire des statisticiens face au problème de l'interprétation des interactions[136] entre variables [181]. Chez ces derniers, les interactions "non significatives" ne sont pas interprétées.

[136] Le lien entre le modèle log-linéaire des statisticiens et les règles d'association est détaillé dans lannexe C.

Quand une interaction entre n variables[137] est significative, elle n'est considérée pour interprétation que si les autres variables n'interagissent pas de façon significative avec ces n variables prises dans leur ensemble.

Nous traduisons ainsi : quand une règle r1 l'emporte sur une règle r2, cela correspond à une relation significative pour r1 et non significative pour r2, et dans ce cas, nous éliminons r2. Quand les règles sont similaires, nous estimons qu'elles sont toutes deux significatives, nous choisissons alors d'éliminer la règle comportant le moins d'attributs, donc la plus "générale".

Définition 8.2.5. *Soit {r1,r2} un élément de $\mathbb{P}_2(\mathcal{R})$, on dit qu'il est incohérent pour la méta-règle mr_3 si l'ensemble des attributs de la partie gauche de l'une est strictement inclus dans l'ensemble des attributs de la partie gauche de l'autre, avec les mêmes modalités, et si les parties droites ont les mêmes attributs, avec les mêmes modalités pour tous sauf 1.*

Si r2 a moins d'attributs que r1 en partie gauche, donc est plus générale, $mr_3(\{r1, r2\})$ est défini par

$$\begin{cases} \emptyset & \text{si } \{\text{r1}, \text{r2}\} \textit{ cohérent} \\ \emptyset & \text{si } (\{\text{r1}, \text{r2}\} \textit{ incohérent}) \text{ et } (\text{r1 et r2 } \textit{non comparables}) \text{ et } (\textit{non exigence}) \\ \{r2\} & \text{si } (\{\text{r1}, \text{r2}\} \textit{ incohérent}) \text{ et } (\text{r1 et r2 } \textit{similaires}) \\ \{r2\} & \text{si } (\{\text{r1}, \text{r2}\} \textit{ incohérent}) \text{ et } (\text{r1 et r2 } \textit{non comparables}) \text{ et } (\textit{exigence}) \\ \{r1\} & \text{si } (\{\text{r1}, \text{r2}\} \textit{ incohérent}) \text{ et } (\text{r2 } \textit{l'emporte sur } \text{r1}) \\ \{r2\} & \text{si } (\{\text{r1}, \text{r2}\} \textit{ incohérent}) \text{ et } (\text{r1 } \textit{l'emporte sur } \text{r2}) \end{cases}$$

8.2.5 L'action des méta-règles

L'ensemble de règles nettoyé est

$$\mathcal{R} - \bigcup_{i=1,2,3} \Big(\cup mr_i(\mathbb{P}_2(\mathcal{R})) \Big)$$

où $\cup mr_i(\mathbb{P}_2(\mathcal{R}))$ désigne l'union des éléments de l'ensemble $mr_i(\mathbb{P}_2(\mathcal{R}))$.

Cet ensemble nettoyé a été obtenu de façon séquentielle : pour la méta-règle n°1, on a parcouru le jeu de règles à la recherche d'ensembles incohérents de 2 règles, mais les ensembles de règles à éliminer n'ont pas été retirés au fur à mesure du jeu, les règles correspondantes ont seulement été pointées. Puis pour la méta-règle n°2, le jeu de règles complet a été parcouru à nouveau, de la même façon, et nous avons terminé par la méta-règle n°3 de façon identique.

Ce mode de fonctionnement fait que le résultat obtenu est indépendant du sens de parcours du jeu de règles. Une règle, même éliminée à une étape, peut contribuer à en éliminer une autre à l'étape suivante. Si on reprend l'exemple 1, et si on imagine que pour chacun des ensembles {r2,r3}, {r3,r4}, {r2,r4}, on peut trouver une méta-règle pour laquelle ils sont incohérents, comme r2 l'emporte sur r3, r3 l'emporte sur r4, et r4 l'emporte sur r2, les 3 règles seront finalement éliminées. Par contre, si une règle éliminée ne pouvait plus être comparée avec un autre, le résultat obtenu varierait selon l'ordre d'examen des 3 paires de règles : il pourrait n'en rester aucune, ou n'importe laquelle des trois. Par exemple, si les 3 ensembles sont rencontrés dans l'ordre ci-dessus, alors l'examen de {r2,r3} fait que r3 serait éliminée, donc {r3,r4} ne serait pas examiné, et l'examen de {r2,r4} fait que r2 serait éliminée. Donc r4 resterait, comme si elle l'emportait sur les deux autres, ce qui n'a pas de sens, du fait de la non-transitivité de nos relations de comparaison.

[137] Nous prenons n≥1. Quand n=1, ce que nous appelons interaction "entre" 1 variable A est son effet principal.

8.3 Exemple

Nous reprenons ici l'exemple du tableau 4.1 afin de voir quelle action peuvent avoir nos méta-règles de nettoyage sur un jeu de règles déjà réduit par la technique de seuillage courante. Le jeu de règles que nous essayons de nettoyer contient toutes les règles d'association de support≥2 et de confiance ≥0,5. Il y en a 59.

	a1	b1	c1	d1	e1	a0	b0	c0	d0	e0
s1	1	0	1	1	0	0	1	0	0	1
s2	0	1	1	0	1	1	0	0	1	0
s3	1	1	1	0	1	0	0	0	1	0
s4	0	1	0	0	1	1	0	1	1	0
s5	1	1	1	0	1	0	0	0	1	0

TAB. 8.2 – Attributs positifs et négatifs de l'exemple du tableau 4.1.

Nous complétons le tableau 4.1 en ajoutant les attributs contraires, (a0, b0, c0, d0, e0) des attributs de départ que nous notons maintenant (a1, b1, c1, d1, e1), ce qui nous donne le tableau n°3. On extrait alors le jeu complet de règles d'association, il en contient 709. Puis nous ne calculons sur chaque règle que les indices qui ont été utilisés pour le seuillage de départ, ce sont le support et la confiance, afin de montrer que le gain de réduction n'est dû qu'à la seule action des méta-règles. Pour la comparaison de deux règles, nous utilisons comme écart un demi-écart-type de la distribution de l'indice sur l'ensemble des 709 règles, soit 0,37 pour le support et 0,15 pour la confiance, et choisissons un expert exigeant.

Voici pour chaque méta-règle un exemple remettant en cause une règle du jeu initial :
Selon mr_1, les règles b1e1→a1 (supp=2 ;conf=0,5) et b1e1→a0 (supp=2 ;conf=0,5) sont incohérentes, la contradiction portant sur a. Les différences entre leurs supports et leurs confiances étant nulles, donc inférieures aux écarts respectifs de 0,37 et 0,5, elles sont similaires et supprimées toutes deux.

Selon mr_2, les règles a1→b1e1 (supp=2 ; conf=0,67) et a0→b1e1 (supp=2 ; conf=1) sont incohérentes, la redondance étant due à l'attribut a. La différence entre les supports étant nulle, et celle entre les confiances de 0,33, donc supérieure à l'écart de 0,15, la deuxième règle l'emporte sur la première qui est supprimée.

Selon mr_3, les règles e1→a1b1 (supp=2 ; conf=0.5) et d0e1→a0b1 (supp=2 ; conf=0.5) sont incohérentes, la contradiction venant de l'attribut a dont la modalité est à 1 pour l'effet global (sans l'attribut d) et à 0 pour l'effet local (avec d0). Elles sont similaires, car de même support et de même confiance. On supprime la première, qui est la plus générale.

Finalement, l'action sur le jeu complet de 709 règles, de mr1 (resp. mr2, mr3, l'ensemble des 3 méta-règles) a permis d'éliminer 196 (resp. 183, 250, 382) règles, et sur le jeu initial de 59 règles, il y a 5 (resp. 19, 20, 31[138]) règles supprimées par mr1 (resp. mr2, mr3, l'ensemble des 3 méta-règles), soit dans les 2 cas plus de 50% des règles en utilisant les 3 méta-règles.

[138]Les 31 règles supprimées des 59 figurant dans [57] sont les suivantes : ∅ →e, ∅ →b, ∅ →a, ∅ →ce, ∅ →bc, a→b, b→a, b→c, c→b, a→e, e→a, c→e, e→c, b→ac, bc→a, a→bc, c→ab, ac→b, a→be, b→ae, e→ab, ab→e, ae→b, be→a, a→ce, c→ae, e→ac, ac→e, ce→a, c→be, ac→be.

8.4 Bilan et perspectives

Nos méta-règles sont utilisables par tous, car elles s'appuient sur la logique du sens commun. Elles partent du jeu complet de règles extrait des données, la seule exigence pour ces règles étant d'avoir un support non nul. Et elles réduisent ce jeu en éliminant des contradictions et redondances. Bien sûr, ces 3 méta-règles ne peuvent pas réduire le jeu complet de règles de façon suffisante. Une perspective de ce travail est d'en construire un ensemble cohérent, de les rendre plus générales, et d'optimiser les algorithmes correspondants.

On peut reprocher à cette technique le fait d'utiliser des propriétés avec négation, ce qui alourdit énormément les algorithmes de recherche des motifs fréquents et de règles d'association si les propriétés avec négation n'étaient pas prévues au départ. Mais il existe de nombreux cas où le codage fait que la propriété et sa négation existent toutes deux dans la base. Il y a des cas où le codage produit une série de propriétés formant une partition des sujets selon les diverses modalités d'une même propriété. C'est une extension du cas de la propriété et de sa négation que nous venons de traiter dans ce chapitre, et qui produit également des incohérences dans le jeu de règles, quand on ne peut pas le régler par un recodage flou, comme indiqué dans le chapitre suivant. Les méta-règles doivent alors être réécrites de façon plus générale. C'est une autre extension de ce travail.

Cette technique peut également permettre de nettoyer des jeux de règles obtenus par fusion de jeux extraits à partir des valeurs de différents ensembles de sujets pour les mêmes propriétés, ou proposés par un expert.

8.5 Une revue des méthodes proches

Nous avons cité dans le chapitre précédent plusieurs thèses relatives aux entrepôts de données [245, 231, 232, 91, 51] de l'Université Simon Fraser (Canada) dans lesquelles le nettoyage par méta-règles est également utilisé.

Certains informaticiens ayant la même orientation "bases de données", ont orienté leurs travaux vers la recherche d'exceptions, comme les *règles d'exception* de Suzuki [224], ou en se référant au paradoxe de Simpson [76], qui ne s'exprime pas nécessairement au sein du jeu de règles d'association, comme nous l'avons montré en annexe.

Certains travaux se concentrent sur le cas particulier des règles de classement, qui sont des règles d'association ayant toutes la même variable de classement en partie droite. Nous en détaillons deux dans les sous-sections suivantes.

Élagage de règles de classement

Dans [227], il est proposé un algorithme (RuleCover) d'élagage d'un ensemble de règles de type X→Y où Y est un motif fixé, à partir des lignes du tableau des sujets vérifiant ces règles. Le principe est que l'ensemble de règles à l'arrivée couvre les mêmes sujets qu'au départ tout en contenant moins de règles. Pour réaliser cela, on sélectionne les règles à conserver en parcourant l'ensemble initial de règles dans l'ordre décroissant du nombre d'éléments qui les vérifient encore après avoir retiré les éléments déjà pris en compte. Ce type d'élagage "orienté-sujet" ne prend pas en compte la composition en propriétés des règles, pas plus que leur support ou leur confiance. Ce qui compte n'est pas la qualité individuelle de chaque règle, mais celle de l'ensemble. Si on avait dans l'ensemble de départ deux règles de type A→B et C→B, où A est un sous-motif de C, la règle plus générale A→B a plus de chances d'être gardée, et la plus spécifique éliminée, dans la mesure ou la couverture de la première englobe celle de la seconde. Pourtant, ce n'est pas

systématique, car chaque fois qu'une règle est retenue, ses éléments sont éliminés de la couverture de chaque règle restante. A cela s'ajoute un algorithme plus traditionnel qui élimine les règles de type C→B à chaque fois qu'il rencontre une règle A→B telle que A soit un sous-motif de B. Pour celui-ci il est inutile de prendre en compte les sujets. Les auteurs donnent un exemple où il font fonctionner le deuxième algorithme, puis le premier, sur un ensemble de 1461 règles ayant une même propriété en partie droite, de seuil de support 2% et de seuil de confiance 90% . Cet ensemble passe ainsi de 1461 règles à 20 puis 5. Ils conseillent d'utiliser un seuil de confiance élevé, leurs algorithmes risquant sinon de supprimer une partie des règles de confiance élevée au profit de règles de faible confiance.

Remplacement de variables impliquées dans des liaisons complexes

Plutôt que de renforcer le choix des attributs à l'aide d'indicateurs comme dans [86] au moyen un calcul de coût algorithmique, A. Jakulin [129] transforme l'ensemble des attributs afin de faire disparaître ces relations complexes, d'une façon voisine de celle des chercheurs en psychologie [120], à savoir un pré-traitement. Il s'intéresse particulièrement aux relations complexes liant deux ou trois attributs et le classifieur, relations qu'il appelle interactions. Une interaction est pour lui significative si sa disparition augmente la qualité du classement, d'où sa volonté de faire disparaître les une partie des interactions .

Ce qu'il appelle interaction "vraie" est de type XOR, donc concerne des attributs à deux modalités. Elle fait partie des interactions croisées des sciences humaines analysées par l'Anova. Son interaction "fausse" est ce que nous appelerions la "redondance", car les attributs ont des valeurs identiques pour chaque classe dans le cas extrême. L'interaction vraie est corrigée en remplaçant les deux propriétés par leur produit AB, c'est-à-dire par l'attribut aux quatre modalités A0B0, A0B1, A1B0 et A1B1, et les interactions fausses sont susceptibles de différentes corrections, allant de la même correction que pour les interactions vraies au remplacement de ces deux attributs par un attribut latent à deux modalités. A cela il ajoute une interaction conditionnelle, c'est à dire en partie vraie et en partie fausse, ce qui est le cas par exemple d'une variable A à quatre modalités dont deux donnent une interaction vraie avec une variable B pour le classifieur C, et les deux autres donnent une interaction fausse, qu'il traite en remplaçant A par deux nouvelles variables. Ainsi avec deux attributs booléens, on obtient un attribut à quatre modalités, et ce chaque fois qu'on rencontre des interactions entre ces deux attributs et le classifieur.

Il donne l'exemple de la base de données "Car" de l'UCI, pour laquelle on dispose des valeurs d'objets pour 7 attributs qui sont : "car" (l'attribut de classement), "doors", "persons", "lug-boot", "safety", "buying" et "maint". Puis il crée de nouveaux attributs en combinant les anciens afin de faire disparaître certaines interactions :

- les attributs "doors", "persons", "lug-boot" sont remplacés par un attribut qu'il appelle "comfort"
- les attributs "comfort" et "safety" sont remplacés par un attribut qu'il appelle "tech"
- les attributs "buying" et "maint" sont remplacés par un attribut qu'il appelle "price"

Sa discrimination de la classe "car" se fait alors sur les deux seuls attributs "price" et "tech".

Ces deux façons de traiter ce problème se retrouvent dans l'Anova, il s'agit des "contrastes simples d'interaction" pour le premier, et de la "confusion" pour le dernier [120].

9

Une solution pour les propriétés numériques : motifs et règles d'association flous

Sommaire

9.1 Introduction

Les règles d'association sont construites à partir de matrices de type sujets×propriétés. Ces matrices sont constituées de 0 et de 1. Si à l'intersection de la ligne du sujet s et de la propriété p on a la valeur 1, cela signifie que le sujet s possède la propriété p, alors que si on a la valeur 0, cela signifie qu'il ne la possède pas. On extrait d'abord toutes les combinaisons de propriétés que les sujets peuvent posséder simultanément. Ce sont les *motifs*. Puis on scinde ces motifs en deux sous-motifs A et B qui forment la partie gauche et la partie droite de la règle, et on obtient ainsi la règle d'association "si A, alors B". Quand les matrices explorées sont petites, le nombre de règles est déjà important puisque pour une cooccurrence de n propriétés, c'est-à-dire un motif de longueur n, on a $2^n - 2$ règles ayant une partie gauche et une partie droite non vide. Mais les bases de données sont de plus en plus volumineuses. les matrices explorées sont de plus en plus grandes, et les données arrivent parfois en flot continu, ce qui nécessite de les synthétiser en connaissance rapidement assimilable. La construction d'un petit jeu de règles d'association est une réponse adaptée à ces contraintes, et couramment employée en fouille de données. Pour obtenir ce petit jeu, on essaie de limiter le nombre de motifs, en choisissant ceux qui sont vérifiés par un grand nombre de sujets, puis le nombre de règles en gardant celles qui ont les valeurs les plus élevées à certains indices de qualité, ou en éliminant celles qui apportent des incohérences au jeu de règles. Et pour agir de façon plus rapide, et mobiliser le moins de ressources informatiques, on optimise les algorithmes de recherche des motifs, en s'appuyant par exemple sur le formalisme du treillis des motifs fermés.

Quand la matrice n'est plus formée de données binaires, mais d'entiers, qui représentent par exemple la force d'une adhésion à une opinion (pas du tout d'accord, désapprouve, approuve, tout a fait d'accord) ou l'intensité d'un sentiment (pas du tout, un peu, beaucoup, énormément), on peut se ramener à une matrice booléenne en recodant chaque opinion, chaque sentiment, en autant de variables qu'il y a de modalités. En faisant ainsi, on multiplie de façon importante le nombre de variables, et au lieu d'obtenir une règle comme "si on a peur des araignées, alors on a peur des serpents" on obtient un nombre plus ou moins important de règles du genre "si on a *beaucoup* peur des araignées, alors on a *un peu* peur des serpents", "si on a *beaucoup* peur des araignées, alors on a *énormément* peur des serpents", etc. Dans le cas de 4 modalités pour l'une comme pour l'autre de ces deux propriétés, on est confronté à 16 règles possibles au lieu d'une seule, sans aucun moyen de les relier l'une à l'autre et de remonter à une règle plus générale. Et si plusieurs propriétés du tableau sont dans ce cas, le jeu de règles devient vite ingérable de par sa taille.

En fait, les données dont on dispose comme réponses des sujets à ces échelles d'opinions ou de sentiments sont subjectives et imprécises. Et construire 16 règles précises sur des réponses imprécises nous semble inadapté. Nous préférons construire une seule règle en acceptant qu'elle soit moins précise que les règles d'association habituelles. Pour cela, nous reprenons les étapes de création du jeu de règles d'association à partir de la matrice sujets×attributs en remplaçant toutes les définitions adaptées aux données booléennes par des définitions sur des données floues, en utilisant le formalisme de Zadeh et de Lukasiewicz le plus adapté à notre problème, à l'éclairage des analyses qui en sont faites dans l'ouvrage de Dubois et Prade [71] et dans l'ouvrage collectif sous la direction de Zadeh [241]. Puis nous appliquons cette technique à l'extraction d'un jeu de règles sur des données issues de réponses à un questionnaire.

9.2 Ensembles flous

On reprend les définitions de Zadeh[240] citées dans l'ouvrage de D. Dubois et H. Prade [71].

Définition 9.2.1. *Ensemble flou*
Un ensemble flou F est la donnée d'un référentiel Ω et d'une application μ_F de Ω dans [0,1], cette application étant interprétée comme le degré d'appartenance des éléments de Ω à F.

Si les seules valeurs d'appartenance sont 0 et 1, on retrouve les ensembles habituels, contenant tous les éléments du référentiel de degré d'appartenance égal à 1, et aucun élément de degré d'appartenance égal à 0.
Les opérations ensemblistes habituelles sont étendues aux ensembles flous de la façon suivante :

Définition 9.2.2. *Opérations sur les ensembles flous*
Si F et G sont deux ensembles flous de fonctions d'appartenance μ_F et μ_G sur le référentiel Ω, on pose :

Cardinal :		$card(F) = \sum_{\omega \in \Omega} \mu_F(\omega)$
Egalité :	$F = G$	*si* $\forall \omega \in \Omega,\ \mu_F(\omega) = \mu_G(\omega)$
Inclusion :	$F \subseteq G$	*si* $\forall \omega \in \Omega,\ \mu_F(\omega) \leq \mu_G(\omega)$
Complémentation	$\overline{F}$:	$\forall \omega \in \Omega,\ \mu_{\overline{F}}(\omega) = 1 - \mu_F(\omega)$
Intersection	$F \cap G$:	$\forall \omega \in \Omega,\ \mu_{F \cap G}(\omega) = min(\mu_F(\omega),\ \mu_F(\omega))$
Réunion	$F \cup G$:	$\forall \omega \in \Omega,\ \mu_{F \cup G}(\omega) = max(\mu_F(\omega), \mu_F(\omega))$

Nous avons choisi ces définitions car, d'après D. Dubois et H. Prade[71], parmi toutes les définitions possibles, ce sont les seules qui permettent de mettre une structure de treillis sur l'ensemble $[0,1]^\Omega$ des ensembles flous muni des opérations de complémentation, d'intersection et de réunion, comme celle qui existe sur l'ensemble $\{0,1\}^\Omega$ des ensembles classiques muni de ces mêmes opérations. Cette structure de treillis est exploitée par les algorithmes de recherche de motifs et de règles d'associations.

Nous allons maintenant essayer de donner des définitions floues aux éléments que nous utilisons pour les règles d'association, afin que le maximum de propriétés utilisées dans les algorithmes de génération des jeux de règles d'association soient conservées, tout en nous approchant le plus possible du formalisme flou le plus couramment utilisé.

Définition 9.2.3. *Propriété floue*
Une propriété floue P sur un ensemble $\mathcal{S}$ est une application de $\mathcal{S}$ vers l'ensemble [0,1]. L'ensemble flou F formé du référentiel $\mathcal{S}$ et de la fonction d'appartenance μ_F qui à tout élément s de $\mathcal{S}$ associe la valeur μ_F(s)=P(s) est noté P' et est appelé l'extension de P. Et on dit que l'élément s vérifie la propriété P dès que cette valeur est strictement positive.

Prenons l'exemple du questionnaire des peurs (voir Annexe), où $\mathcal{S}$ est un ensemble de sujets, $\mathcal{P}$ un ensemble de 89 peurs, avec pour chaque sujet sa réponse pour chaque peur selon l'échelle suivante : 1 :"pas du tout", 2 :"un peu", 3 :"assez", 4 :"beaucoup", 5 :"énormément". Si un sujet a coché 1 pour la peur des araignées "a", on peut être sûr qu'il n'en a pas peur. S'il a coché 5, il est certain qu'il en a peur. Par contre s'il a coché 2, 3 ou 4, on a beaucoup moins de certitude sur la présence ou l'absence de cette peur. Voici un codage de cette peur en deux propriétés floues a1 : peur élevée, et a0 peur faible, et un codage en trois propriétés floues b2 : peur importante, b1 : peur moyenne et b0 : peur très faible.

a
1 : pas du tout
2 : un peu
3 : assez
4 : beaucoup
5 : énormément

a1	a0
0	1
0	1
0,5	0,5
1	0
1	0

b2	b1	b0
0	0	1
0	0,5	0,5
0	1	0
0,5	0,5	0
1	0	0

TAB. 9.1 – Transformation de la propriété "a" codée selon une échelle de Lickert à 5 points en 2 ou 3 propriétés floues.

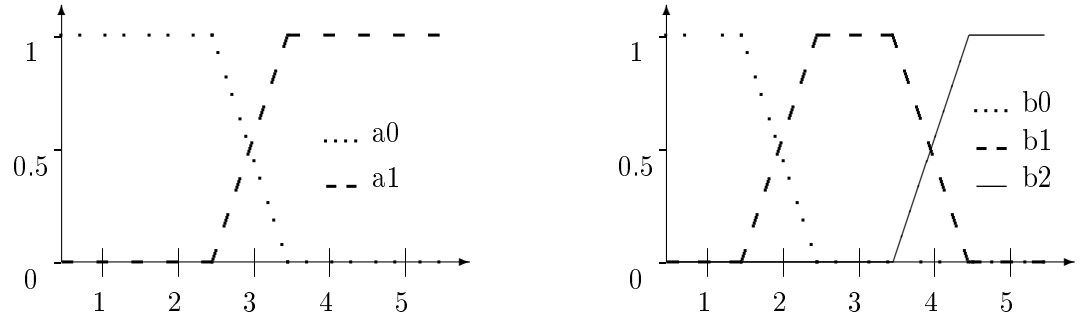


FIG. 9.1 – Transformation de la propriété "a" codée de 1 à 5 en 2 ou 3 propriétés floues.

Il y a plusieurs codages flous possibles en utilisant deux propriétés floues, mais également plus de deux propriétés floues, ou même une seule propriété floue.

La matrice sujets×propriétés n'est plus booléenne comme précédemment. C'est une matrice de nombres compris entre 0 et 1. la relation qui lie l'ensemble $\mathcal{S}$ des sujets et l'ensemble $\mathcal{P}$ des propriétés est devenue une *relation floue* qui peut être définie à l'aide de la fonction f suivante :

$f : \mathcal{S} \times \mathcal{P} \rightarrow [0,1]$

$(s , P) \mapsto f(s,P)$

Selon le sens de lecture de cette relation, "un sujet vérifie de façon floue une propriété" ou "une propriété est vérifiée de façon floue par un sujet", on peut s'intéresser non seulement aux ensembles flous de sujets, mais encore aux ensembles flous de propriétés. Cette dualité qui a permis d'associer à une relation ordinaire un treillis de concepts ne sera pas conservée dans la mesure ou les deux treillis ne coïncident pas en général, comme nous le verrons plus loin.

Définition 9.2.4. *Négation d'une propriété floue.*
La négation $\overline{a}$ de la propriété a est obtenue en remplaçant la valeur de chaque sujet de $\mathcal{S}$ par son complément à 1 :

$\forall s \in \mathcal{S},\ \mu_{\overline{a}}(s) = 1 - \mu_a(s)$

Selon cette définition, l'ensemble des sujets vérifiant la négation d'une propriété est le complémentaire de l'ensemble flou des sujets vérifiant la propriété. C'est donc une propriété floue que l'on obtient ainsi. Par exemple, dans le tableau 1, la propriété a1 est la négation de la propriété a0.

Exemple 3. *Dans le tableau 2 est représenté un exemple fictif de relation floue liant un ensemble de 2 propriétés $\mathcal{P}$={a,b} à un ensemble $\mathcal{S}$ de 5 sujets. Ceci afin d'illustrer quelques opérations*

sur les ensembles flous. Nous nous intéressons dans cet exemple aux ensembles flous de sujets que sont les extensions des propriétés floues.

sujet	a1	b1	a0	b0	a1b1	a1b0	a0b1	a0b0	∅
s1	0	0	1	1	0	0	0	1	1
s2	0,3	0,4	0,7	0,6	0,3	0,3	0,4	0,6	1
s3	1	1	0	0	1	0	0	0	1
s4	0,7	1	0,3	0	0,7	0	0,3	0	1
s5	0	0,8	1	0,2	0	0	0,8	0,2	1
total	2	3,2	3	1,8	2	0,3	1,5	1,8	5

TAB. 9.2 – 2 propriétés à valeurs floues sur un ensemble de 5 sujets

Dans cet exemple, on suppose que la propriété a, ayant un certain nombre de valeurs sur $\mathcal{S}$ a été rendue floue selon les techniques indiquées précédemment, et est devenue la propriété a1. La propriété a0 est la négation de la propriété a1, Et on a procédé de la même façon pour la propriété b, qui a fourni la propriété "positive" b1, et la propriété "négative" b0.
Un référentiel $\mathcal{S}$ étant donné, on a défini pour les ensembles classiques un motif comme réunion d'un ensemble P de propriétés sur $\mathcal{S}$. Et constaté que les éléments vérifiant ce motif forment l'intersection des ensembles d'éléments vérifiant chaque propriété. L'extension de l'intersection aux ensembles flous permet de généraliser ainsi la notion de motif aux ensembles flous :

Définition 9.2.5. *Motif flou*
Un motif flou sur un ensemble $\mathcal{S}$ est une réunion de propriétés floues sur cet ensemble. L'extension de ce motif est obtenue en faisant l'intersection des extensions des propriétés le composant. Pour le motif vide, chaque élément de $\mathcal{S}$ a une valeur de 1. On dit qu'un élément de $\mathcal{S}$ vérifie le motif flou s'il vérifie toutes les propriétés de ce motif, ou si ce motif est le motif vide.

Dans le tableau 2, pour le motif a1b1, on a fait l'intersection des ensembles flous a1' et b1', en prenant pour chaque sujet le minimum de ses valeurs pour a1 et b1. Nous avons obtenu de la même façon les motifs a1b0, a0b1 et a0b0. Les propriétés a1 et b1 sont telles que a1'⊆b1', car pour chaque sujet les valeurs de a1 sont inférieures à celles de b1, et on peut vérifier qu'on a bien, comme pour les ensembles classiques b0'⊆a0'[139].

On n'a pas indiqué dans ce tableau les valeurs des sujets pour les motifs a0a1, b0b1 pas plus que pour les motifs a0a1b0, a0a1b1, a0b0b1, a1b0b1, a0a1b0b1. On peut remarquer toutefois qu'elles ne sont pas toutes nulles. En effet, les sujets, comme s2 par exemple, qui ont une valeur strictement comprise entre 0 et 1 pour a1, ont également une valeur strictement comprise entre 0 et 1 pour a0, ce qui fait que le minimum de ces deux valeurs n'est pas nul. Ce fait signifie que la loi de "non-contradiction" des ensembles classiques n'est pas conservée quand on passe aux ensembles flous. On peut lire dans D. Dubois et H. Prade[71] qu'avec la loi du "tiers-exclu", ce sont les deux seules propriétés ensemblistes qui ne se généralisent pas aux ensembles flous muni des opérations définies précédemment. Toutefois, on peut corriger cet effet car les valeurs d'un sujet pour le motif a0a1 ou tout motif formé de deux propriétés contraires, ne dépassent jamais

[139]Cette propriété qui lie l'inclusion et la complémentation fait partie des nombreuses propriétés que les définitions choisies permettent de garder. Pour une liste plus détaillée, voir [71].

0,5[140]. En éliminant les motifs pour lesquels aucun sujet n'a de valeur supérieure à 0,5 [141] nous évitons ainsi de garder des motifs risquant de générer des règles vides de sens. En supprimant les motifs vérifiant cette règle, on est amené à éliminer également le motif a1b0 car sa seule valeur non nulle est 0,3. Cela n'est pas gênant, au contraire. Le fait d'éliminer le motif a1b0, même si ses valeurs ne sont pas toutes nulles, permet de garder une certaine cohérence avec les inclusions a1'⊆b1' et b0'⊆a0'. En effet, pour les ensembles classiques, l'inclusion de a1' dans b1' signifie qu'aucun élément vérifiant la propriété a1 ne vérifie la propriété b0, donc que le motif a1b0 est de support nul. Pour les ensembles flous, la propriété est affaiblie : le motif a1b0 n'a aucune valeur qui dépasse 0,5[142].

9.3 Règles d'association et ensembles flous

On a défini précédemment une règle comme couple de deux parties complémentaires d'un motif classique de support non nul, le support étant le cardinal de l'extension du motif. Il s'agit maintenant d'étendre cette définition aux motifs flous. Nous avons vu dans la partie précédente qu'il était préférable de ne pas garder les motifs pour lesquels les valeurs des sujets sont toutes inférieures ou égales à 0,5. Nous pourrions résoudre notre problème en changeant le calcul des valeurs des sujets pour l'intersection de deux ensembles flous[143]. Mais certains algorithmes de recherche de règles sont optimisés pour les parcours de treillis, et comme nous l'avons signalé plus haut, la définition de l'intersection proposée permet de garder la structure de treillis associée à la relation. Nous allons utiliser le support pour éliminer ces motifs en posant la définition suivante :

Définition 9.3.1. *Support d'un motif flou M sur un ensemble $\mathcal{S}$.*
Si toutes les valeurs des sujets de $\mathcal{S}$ pour le motif M sont inférieures ou égales à 0,5, le support de M est 0, sinon il est égal au cardinal de l'extension M' du motif M, c'est-à-dire à la somme des valeurs des sujets.

Lors de son extension aux motifs flous, le support a changé de nature . Ce n'est plus le nombre d'éléments vérifiant les propriétés du motif, mais la somme de leurs valeurs. Et cela a modifié ses propriétés. Par exemple, considérons les trois motifs A, B et C définis sur $\mathcal{S}$={x1, x2, x3} par leurs valeurs respectives {0 ; 0,6 ; 1}, {0,3 ; 1 ; 0,3} et {0 ; 0,9 ; 0,6}. Leurs supports respectifs sont 1,6 , 1,6 et 1,5. Les supports de A et B sont égaux alors que A est vérifié par 2 sujets et B par 3 sujets. A et C sont vérifiés par les mêmes sujets, mais le support de C est inférieur à celui de A, ce qui ne signifie pas d'ailleurs que C' est inclus dans A'.
Il convient donc de vérifier que la définition de la règle d'association à partir d'un motif de support non nul garde un sens. Pour définir une règle A→B, on considère un motif flou C de support non nul sur le référentiel Ω, donc pour lequel il y a au moins un élément ω qui a une valeur $\mu_{C'}(\omega)$ strictement supérieure à 0,5, et on le partage en deux motifs flous A et B tels

[140] Si x est la valeur d'un sujet à la propriété a1, 1-x est sa valeur à la propriété a0, et leur moyenne est $\frac{x+(1-x)}{2} = 0,5$. Quand l'un est inférieur à 0,5, l'autre est donc supérieur à 0,5, et le minimum des deux est forcément inférieur ou égal à 0,5. Il atteint la valeur 0,5 seulement quand x=0,5.

[141] Dans leur article[72], D. Dubois et H. Prade se sont même limités aux ensembles ayant au moins une valeur égale à 1

[142] Si le sujet s a une valeur de x pour a1, de y pour b0, sa valeur pour a1b0 est min(x,y). Supposons qu'elle soit supérieure strictement à 0,5, alors x et y sont également supérieurs à 0,5, et 1-y est ainsi inférieur à 0,5, donc à x. La valeur de ce sujet s pour b1 est inférieure strictement à sa valeur pour a1. Ceci est en contradiction avec l'inclusion de a1' dans b1', qui exige que pour tous les sujets, la valeur pour a1 soit inférieure à celle pour b1.

[143] Si F et G sont deux ensembles flous sur un référentiel Ω, on pourrait modifier la définition de l'intersection de F et G en remplaçant la valeur $min(\mu_F(\omega),\ \mu_F(\omega))$ par 0 pour les éléments ω de Ω pour lesquels elle est inférieure ou égale à 0,5. Mais la propriété $\overline{A \cup B} = \overline{A} \cap \overline{B}$ n'est alors plus valable sur les ensembles flous.

que C=$A \cup B$, et $A \cap B = \emptyset$. D'après la définition des motifs flous, C'=$A' \cap B'$ donc d'après la définition de l'intersection des ensembles flous, cette valeur est le minimum de $\mu_{A'}(\omega)$ et $\mu_{B'}(\omega)$. Ces deux expressions sont donc supérieures à 0,5 et la règle A→B est ainsi vérifiée par au moins un élément, qui est ω, et a en partie gauche comme en partie droite un motif de support non nul.

D'où la définition suivante :

Définition 9.3.2. *Règle d'association floue*[144]
On appelle règle d'association floue A→B sur un ensemble $\mathcal{S}$ un couple formé de 2 parties complémentaires A et B d'un motif flou sur $\mathcal{S}$ de support non nul. On dit qu'un élément de $\mathcal{S}$ vérifie la règle d'association floue s'il vérifie la partie gauche et la partie droite de la règle, donc le motif flou A∪B.

La plupart des indices des règles d'association A→B classiques, c'est-à-dire construites sur des matrices booléennes sujets×propriétés sont des fonctions de 4 nombres : le nombre d'éléments du référentiel et les supports sAB, sA et sB, sans ajout explicite d'hypothèses statistiques. Nous les reprenons donc tels quels pour calculer la qualité des règles d'association floue.

Exemple 4. *Dans le tableau 3 figure un exemple fictif avec 3 propriétés floues sur 10 sujets. On a calculé les valeurs de tous les motifs flous construits sur les seules propriétés positives. Puis les règles de la forme A→B sont obtenues en coupant chaque motif de support non nul en*

sujet	a	b	c	ab	ac	bc	abc	∅
s1	0	0	0,2	0	0	0	0	1
s2	0,8	0,4	1	0,4	0,8	0,4	0,4	1
s3	1	0,8	1	0,8	1	0,8	0,8	1
s4	1	1	1	1	1	1	1	1
s5	0,6	1	0,6	0,6	0,6	0,6	0,6	1
s6	0,2	1	0,2	0,2	0,2	0,2	0,2	1
s7	0,2	0,5	0,2	0,2	0,2	0,2	0,2	1
s8	0	0	0,2	0	0	0	0	1
s9	0	0,2	0	0	0	0	0	1
s10	0	1	0	0	0	0	0	1
support	3,8	5,9	4,4	3,2	3,8	3,2	3,2	10

TAB. 9.3 – Les motifs flous construits sur 3 propriétés à valeurs floues sur $\mathcal{S}$

deux parties A et B. Elles sont données dans le tableau 4. Dans ce tableau, on a indiqué pour chaque motif de support non nul les règles qu'il engendre. Les cinq dernières colonnes du tableau contiennent sAB, le support du motif de départ, qu'on appelle support de la règle, les supports sA et sB des parties gauche et droite de la règle, la confiance de la règle obtenue en divisant le support sAB par sA, et la différence entre la confiance de la règle A→B et celle de la règle ∅ →B, qui est sB/10.

Pour établir les propriétés de la confiance des règles d'association floue, nous allons utiliser la propriété suivante des motifs flous :

[144] Dans cette expression, c'est en fait l'association qui est floue, et non la règle. L'expression "règle floue" a été gardée pour les règles définies par D. Dubois et H. Prade

numéro	motif	A	→	B	sAB	sA	sB	confiance	différence
1	abc	abc	→	∅	3,2	3,2	10	1	0
2	abc	ab	→	c	3,2	3,2	4,4	1	0,56
3	abc	ac	→	b	3,2	3,8	5,9	0,84	0,25
4	abc	bc	→	a	3,2	3,2	3,8	1	0,62
5	abc	a	→	bc	3,2	3,8	3,2	0,84	0,52
6	abc	b	→	ac	3,2	5,9	3,8	0,54	0,16
7	abc	c	→	ab	3,2	4,4	3,2	0,73	0,41
8	abc	∅	→	abc	3,2	10	3,2	0,32	0
9	ab	ab	→	∅	3,2	3,2	10	1	0
10	ab	a	→	b	3,2	3,8	5,9	0,84	0,25
11	ab	b	→	a	3,2	5,9	3,8	0,54	0,16
12	ab	∅	→	ab	3,2	10	3,2	0,32	0
13	ac	ac	→	∅	3,8	3,8	10	1	0
14	ac	a	→	c	3,8	3,8	4,4	1	0,56
15	ac	c	→	a	3,8	4,4	3,8	0,86	0,48
16	ac	∅	→	ac	3,8	10	3,8	0,38	0
17	bc	bc	→	∅	3,2	3,2	10	1	0
18	bc	b	→	c	3,2	5,9	4,4	0,54	0,1
19	bc	c	→	b	3,2	4,4	5,9	0,73	0,14
20	bc	∅	→	bc	3,2	10	3,2	0,32	0
21	a	a	→	∅	3,8	3,8	10	1	0
22	a	∅	→	a	3,8	10	3,8	0,38	0
23	b	b	→	∅	5,9	5,9	10	1	0
24	b	∅	→	b	5,9	10	5,9	0,59	0
25	c	c	→	∅	4,4	4,4	10	1	0
26	c	∅	→	c	4,4	10	4,4	0,44	0
27	c	∅	→	∅	10	10	10	1	0

TAB. 9.4 – Les règles d'association floue construites sur les données du tableau 3

Propriété 9.3.1. *Propriétés des motifs flous emboîtés.*
Si deux motifs flous A et B, B étant de support non nul, sont tels que $A \subseteq B$, alors on a l'inclusion $B' \subseteq A'$ et l'inégalité $support(B) \leq support(A)$, l'égalité $A' = B'$ étant vérifiée si et seulement si les supports de A et de B sont égaux.

Preuve. *si A et B sont deux motifs de supports respectifs sA et sB, tels que $A \subseteq B$ c'est qu'on a adjoint à A des propriétés pour obtenir B. La valeur de chaque sujet pour le motif B a ainsi été obtenue en faisant le minimum de la valeur qu'il avait pour A, et de celles qu'il a pour les autres propriétés. La valeur de chaque sujet pour B est donc inférieure à la valeur de ce même sujet pour A. Le support de B étant non nul, il y a au moins un sujet qui a une valeur pour B supérieure à 0,5. Ce sujet a donc également sa valeur pour A supérieure à 0,5. A a ainsi un support non nul comme B, donc les deux supports sont obtenus en additionnant les valeurs pour tous les sujets et ainsi on a $sB \leq sA$.*
Ces sommes ne peuvent être égales que si les termes qui les composent sont égaux, chaque terme de l'une étant inférieur ou égal au terme correspondant de l'autre. On a donc l'égalité entre ces

deux supports uniquement si les valeurs de tous les sujets sont les mêmes pour A que pour B.

La réciproque n'est pas vraie : on peut avoir l'égalité des extensions de deux motifs sans avoir l'inclusion d'un motif dans l'autre. On peut le voir avec les motifs "ab" et "bc" du tableau 3.

Propriété 9.3.2. *Propriétés de la confiance d'une règle d'association floue.*

1. *La confiance d'une règle d'association floue A→B varie entre 0 et 1.*
2. *La valeur 0 n'est jamais atteinte.*
3. *La valeur 1 est atteinte si et seulement si A'⊆B'.*

Preuve. *Appelons C le motif sur lequel la règle A→B est définie, donc C=A∪B.*

1. *Par définition, le support du motif C sur lequel la règle a été construite n'est pas nul, et le motif A de la partie gauche est inclus dans le motif C. D'après la propriété précédente, le support de A est inférieur à celui de C. La confiance, qui est le rapport du support de C par celui de A est donc inférieure à 1. Elle est bien sûr positive, toutes les valeurs considérées étant positives.*
2. *La règle est construite sur un motif C de support non nul. La confiance est le résultat de la division de ce support par celui de A, qui est fini. Elle est donc non nulle.*
3. *D'après la propriété précédente, et la première partie de cette même preuve, la confiance atteint 1 seulement quand les supports sont égaux c'est-à-dire quand C'=A'. Comme le motif C est la réunion des motifs A et B, son extension C' est l'intersection de leurs extensions A' et C'. On obtient ainsi l'égalité A'∩B'=A', ce qui est équivalent à A'⊆B' (Cette propriété liant l'intersection et l'inclusion des ensembles classiques est conservée pour les ensembles flous*[145]*).*

On voit dans le tableau 4 que les règles ayant comme confiance 1 sont toutes celles qui ont l'extension du motif de gauche incluse dans celle du motif de droite. Il y en a 11 : les 8 règles ayant le motif vide à droite, ce qui est sans intérêt, car le motif vide est vérifié par tous les sujets, et les règles 2 : ab→c, 4 : bc→a, 14 : a→c. La lecture du tableau 3 permet de s'assurer que les inclusions correspondantes (ab)'⊆c', (bc)'⊆a', et a'⊆c' sont bien vraies en constatant que pour chaque sujet, la valeurs de la partie gauche de chaque inclusion est inférieure à celle de sa partie droite.

Examinons les règles d'association floue A→B ayant en partie droite l'ensemble vide. Comme A∪B=A∪∅=A, il y en a autant que de motifs flous A de supports non nuls. Pour toutes ces règles, on a support(A∪B)=support(A) et support(B)=support(∅)=N où N est le nombre d'éléments de $\mathcal{S}$. Donc la confiance est égale à 1 (support(A∪B)/support(A)) et la différence à 0 (confiance(A→B)-support(B)/N). On retrouve donc exactement ce qui se passe pour les règles d'association classiques. Bien que la confiance soit égale à 1, la différence nulle fait que la règle est sans intérêt.

Pour les règles d'association floue A→B ayant en partie gauche l'ensemble vide, elles sont également construites sur le motif A∪∅=A et leur confiance est égale au quotient du support de A par le support de ∅ qui est N. Leur différence est donc nulle (confiance(A→B)-support(A)/N).

On voit donc que, comme pour les règles d'association classiques, les règles d'association floue ayant en partie gauche ou en partie droite l'ensemble vide, n'apportent aucune connaissance.

[145] Preuve de l'équivalence entre C⊆D et C∩D=C pour les ensembles flous C et D :

(a) Si C⊆D, par définition de l'inclusion des ensembles flous, pour chaque élément de $\mathcal{S}$, sa valeur pour C est inférieure à sa valeur pour D, donc sa valeur pour l'intersection est celle qu'il a pour C. donc C∩D=C.

(b) Si C∩D=C, c'est que pour chaque sujet de $\mathcal{S}$, la valeur pour l'intersection est la même que celle pour C, comme c'est la plus petite des deux, c'est que la valeur pour C est la plus petite, donc on a bien C⊆D.

9.4 Le treillis des motifs flous

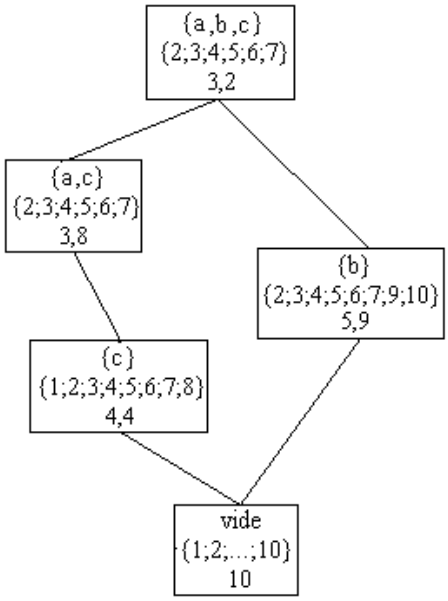


FIG. 9.2 – Le treillis des motifs fermés flous du tableau 3.

Comme on a défini les opérations ensemblistes afin de respecter la structure de treillis, on peut représenter les motifs issus de l'ensemble des 3 propriétés du tableau 3 par un treillis. Ce treillis est dessiné dans la figure 2. Il y a 5 motifs fermés associés aux 5 fermés $\emptyset$, b1, c1, a1c1 et a1b1c1. Chaque motif fermé figure dans un rectangle de 3 lignes. les propriétés du motif flou fermé sont dans la première, les sujets vérifiant le motif, donc pour lequel ils ont une valeur non nulle, dans la deuxième ligne, et le support dans la troisième. Si on regarde les motifs fermés correspondant aux motifs ac et abc, on constate qu'ils sont vérifiés par le même ensemble de sujets : {s2 ; s3 ; s4 ; s5 ; s7}, ce qui peut surprendre. En effet, dans un concept classique, l'ensemble des propriétés comme l'ensemble des sujets doivent être des ensembles fermés, donc deux motifs fermés différents ne peuvent avoir un même ensemble de sujets[146]

En fait, comme il s'agit ici d'ensembles flous, on ne peut avoir d'égalité entre deux ensembles flous que si les valeurs de tous leurs éléments sont les mêmes. On peut voir dans le tableau 3 que les valeurs pour abc et ac ne sont pas les mêmes pour tous les sujets, bien que ce soient les mêmes sujets qui aient les valeurs nulles pour les deux motifs. D'après la propriété 3.1, comme le motif flou ab est inclus dans abc, l'ensemble flou des sujets (abc)' est inclus dans l'ensemble flou (ac)', et ces ensembles flous ne sont égaux que si les supports des motifs sont égaux. Ce qui n'est pas le cas ici, le motif abc ayant pour support 3,2, alors que le motif ac a un support de 3,8. Par contre les extensions des motifs a et ac sont les mêmes, donc a n'est pas un motif fermé, et comme on ne peut pas trouver dans le tableau de motif contenant ac de même extension, on en déduit que ac est un motif fermé.

[146]Preuve : Appelons A et B deux ensembles différents de propriétés classiques sur un ensemble $\mathcal{S}$ de sujets, l'un pouvant être inclus strictement dans l'autre. Si A'=B'=S, c'est que tous les sujets de S vérifient les propriétés de A, celles de B, et celles de A∪B. Donc l'intention S' de S, qui est l'ensemble des propriétés vérifiées par les éléments de S, contient A∪B. S' est la fermeture de A et celle de B. Comme les deux ensembles diffèrent, ils ne peuvent tous deux être égaux à S', donc l'un des deux au moins n'est pas fermé.

On voit donc que le choix que nous avons fait de définitions floues a pour conséquence qu'on n'obtient pas le treillis de Galois de Barbut et Monjardet[11], ou, ce qui revient au même, le treillis des concepts de Wille [235]. En effet, en général, nos deux treillis sur une relation floue, celui des motifs flous de propriétés et celui des motifs flous de sujets, ne coïncident pas[147]

9.4.1 Un exemple

Dans le tableau 5 figure une relation floue liant 20 sujets et 9 propriétés. Avec ces 9 propriétés floues, on peut former 2^9, soit 512 motifs de support non nul[148]. Il fournissent 54 motifs fermés.

sujet	a	b	c	d	e	f	g	h	i
1	0,8	0,2	0,6	1	1	1	0,6	0,2	0
2	1	0	1	0,6	1	0,2	0,6	0	0,4
3	1	0,8	0,4	0,6	0,8	0	0,4	0,8	0,2
4	0,2	0,2	0,2	0,6	1	0,6	0,2	0,2	0
5	0,4	0	0,6	0,8	0,6	0	0,6	0	0
6	1	1	0	1	0,2	0	0	1	0,4
7	0	0	0,4	1	0,6	0	0,4	0	0,4
8	0	0	0	1	0,6	0,4	0	0	0
9	0	0	0,4	0,6	1	1	0,4	0	0,4
10	0	0	0,4	0,6	0,8	0,8	0,4	0	0,4
11	0,6	0,6	0,2	1	1	0,4	0,2	0,6	0,2
12	0,8	0	1	0,2	0,6	0	0,2	0	0,4
13	1	0,8	0	1	0,2	0,4	0	0,8	0,4
14	1	1	1	1	1	1	1	1	1
15	0	0	0,4	0,6	1	1	0,4	0	0,4
16	0	0	0,4	0,6	1	1	0,4	0	0,4
17	0	0	0	1	0,6	0,4	0	0	0
18	1	1	0	1	0,2	0	0	1	0,4
19	0	0	0	1	0,6	0,4	0	0	0
20	0	0	0,4	0,6	1	1	0,4	0	0,4
$>$0,5	9	6	5	19	17	8	4	6	1
$\leq$0,5	11	14	15	1	3	12	16	14	19

TAB. 9.5 – Un exemple de 9 propriétés floues définies sur 20 sujets.

Nous avons défini[149] un seuil γ qui permet d'exclure les motifs trop rares, c'est-à-dire ne contenant pas plus de γ sujets possédant beaucoup les propriétés du motif (support$>$0,5), et le seuil δ qui permet d'exclure les motifs trop courants, c'est-à-dire ne contenant pas plus de δ sujets possédant peu les propriétés du motif (support$\leq$0,5). La disparition de ces motifs entraîne forcément la disparition d'autres motifs. Afin de préserver la structure de treillis, nous avons défini une relation d'équivalence basée sur ces seuils qui est décrite en annexe.

[147] D'autres définitions floues permettent de générer un treillis de Galois des concepts sur une relation floue. Pour un exposé de ces définitions, voir l'article de R. Belohlavek [17]

[148] Leur support est même supérieure ou égal à 1, le sujet 14 ayant la valeur 1 pour tous les motifs

[149] Voir les définitions en annexe qui s'appuient sur un passage au quotient du treillis par une relation d'équiva-

Nous avons représenté dans la figure 9.3 le diagramme des 54 motifs fermés correspondants. le symbole # en bas du diagramme représente le motif vide, c'est à dire vérifié par tous les sujets, puis on a le motif formé de la seule propriété d, pour lequel tous les sujets sauf un, qui est le sujet 12, ont une valeur supérieure à 0,5, puis le motif e pour lequel 3 sujets seulement ne dépassent pas la valeur de 0,5. Au dessus, ce sont tous les motifs fermés pour lesquels il y a plus de 3 sujets qui ont une valeur inférieure à 0,5. Les zones correspondant aux premiers seuils possibles de δ sont bornées par des courbes, et étiquetées par les valeurs de δ. On voit que dans la zone de δ=0 des motifs pour lequel aucun sujet n'a une valeur inférieure ou égale à 0,5, il n'y a que le motif vide, si δ=1, à ce motif fermé s'ajoute celui correspondant au motif d, car $\emptyset$ et d sont les deux seuls motifs pour lesquels au plus 1 sujet a une valeur inférieure à 0,5. Pour δ=2, il n'y a en pas d'autre, et pour δ=3, il y en a un de plus, qui est celui de motif e. En haut du diagramme, on a les motifs pour lesquels très peu de sujets ont des valeurs supérieures à 0,5. Par exemple, le motif i, pour lequel seul le sujet 14 dépasse la valeur 0,5, est dans la première zone en haut délimitée par une courbe, et étiquetée par γ=1. Dans cette zone, on trouve forcément tous les motifs résultant d'une union avec ce motif, car pour chaque sujet leurs valeurs sont forcément inférieures à celles du motif i. Il n'y a pas de zone étiquetée par γ=0 car le sujet 14 a pour tous les motifs une valeur supérieure à 0,5. On a également représenté la zone correspondant à γ=2.

Pour les valeurs de γ=0 et δ=0, si on représente le diagramme des motifs fermés des classes d'équivalence des motifs, on a retrouve un diagramme identique à celui de la figure 9.3. En effet, le seul motif de type 1 forme une classe et comme deux motifs de type 0 ne peuvent être dans la même classe que si tous les sujets ont les mêmes valeurs pour ces motifs, on a également une classe par motif de type 0. Par contre, on peut voir dans la figure 9.4, que le diagramme des motifs fermés des classes pour γ=1 et δ=0 en contient deux fois moins. Les 28 motifs fermés de la zone étiquetée γ=1 dans la figure 3, sont maintenant de type -1, alors que 26 autres sont restés de type 0 et 1. Les 28 motifs fermés ont été remplacés par une seule classe, qui est notée abcdefghi, ce qui fait 1+26=27 motifs fermés. Le sens de cette opération est d'éliminer les motifs trop faiblement représentés, c'est-à-dire dans ce cas pour lesquels pas plus d'un sujet n'a obtenu de valeur supérieure à 0,5. Avec des seuils γ=2 et δ=0, on en obtient 23. En effet, les 4 motifs de la zone marquée γ=2 ont exactement 2 valeurs supérieures à 0,5. Et leur type étant -1, ils sont dans la classe de abcdefghi.

Dans la figure 9.5, on peut voir qu'il ne reste plus que 9 motifs fermés quand on prend les seuils γ=2 et δ=3. Le rôle du seuil δ=3 est d'éliminer les motifs trop fortement représentés, c'est-à-dire dans ce cas pour lesquels pas plus d'un sujet n'a obtenu de valeur inférieure ou égale à 0,5. Ainsi, non seulement les motifs e et d sont de type 1, comme $\emptyset$, mais également leur réunion, par définition de notre relation d'équivalence. Ce qui fait que ces 4 motifs forment une seule classe, notée "de" dans la figure 5. De plus, les 18 motifs de type 0 se réduisent à 7 classes. En effet, ceux qui sont formés par réunion avec d ou e (comme ad, abdeh) sont dans la même classe que les motifs obtenus par suppression d'une ou de ces deux lettres. Ainsi la classe "cde" de la figure 9.5 a pour éléments les motifs c, et ce de la figure 9.4, la classe "abdeh" de la figure 9.5 a pour éléments les motifs abh, abdh, abeh et abdeh de la figure 9.4. On voit ainsi que la fixation d'un seuil δ non nul permet également de fusionner des motifs qui différent peu, tout en préservant la structure de treillis.

Notons toutefois que la prise en compte de γ et de δ dans les calculs n'est pas la même. Par exemple, si on considère l'algorithme classique d'extraction de motifs par niveau, qui s'exécute des motifs de longueur 0 vers les motifs de longueur plus grande, à chaque niveau de cet algorithme,

lence.

il faut prendre en compte γ pour décider de garder ou non le motif extrait, alors qu'on peut ne prendre en compte δ qu'à l'étape 0 en éliminant simplement des données les propriétés en dessous du seuil choisi, et en faisant à la fin la réunion des motifs obtenus avec les propriétés éliminées au départ.

9.5 Comparaison du codage flou à une binarisation par seuil

Le codage flou est adapté au cas où on a une échelle ordonnée de modalités pour les propriétés. Dans ce cas on peut procéder également d'autres façons Si on désire ne garder qu'une propriété binaire, on peut décider que toutes les valeurs supérieures à un seuil sont égales à 1, et les autres égales égales à 0. Nous avons repris l'exemple et extrait les règles d'association correspondant à la binarisation des propriétés avec les seuils respectifs de 0,1 0,3 0,5 0,7 et 0,9. On voit dans le tableau 6 que les indices des règles ne sont pas les mêmes selon les seuils, mais qu'il n'y a pas de grandes contradictions. Les règles ayant l'ensemble vide à gauche ou à droite sont tout aussi inintéressantes dans tous les cas. On remarque que Les règles qui avaient comme confiance 1 dans un codage peuvent avoir une confiance inférieure dans un autre codage, et inversement, mais elles ne s'éloignent pas de façon importante. Dans l'exemple proposé du tableau 6, on remarque même que pour les trois indices calculés, la valeur pour le codage flou est toujours strictement comprise entre les valeurs extrêmes obtenues pour les autres codages (sauf bien sur, si pour tous les codages par seuil, ces valeurs sont égales, auquel cas le codage flou donne également cette valeur commune).
Dans le cas général, l'appartenance stricte de l'indice pour le codage flou dans l'intervalle de valeurs de tous les codages possibles par seuil est valable pour le support, mais non pour la confiance et la différence . Pour le démontrer, nous allons d'abord poser une définition et quelques propriétés.

Définition 9.5.1. *α-coupe[150] d'une propriété floue :*
Un nombre réel α étant donné, on appelle α-coupe d'une propriété floue la propriété obtenue en remplaçant toutes les valeurs supérieures ou égales au seuil α par 1, et les valeurs inférieures à α par 0.

Cela permet de remplacer une propriété floue à valeurs dans [0,1] par une propriété ordinaire à valeurs dans {0,1}. Selon le même principe on peut remplacer un motif flou par un motif ordinaire, et ceci de deux façons équivalentes comme le montre la propriété suivante :

Propriété 9.5.1. *α-coupe d'un motif flou.*
Pour un seuil donné α, le motif formé de propriétés obtenues par α-coupes de propriétés floues peut être obtenu directement par α-coupe du motif formé des propriété floues.

Preuve. *Pour un sujet donné s de $\mathcal{S}$, le motif flou A formé des q propriétés p_1, p_2, ... p_q a la valeur $a=min(a_1, a_2, ... a_q)$, les a_i étant les valeurs de s pour les propriétés a_i.*
Si on fait une α-coupe de ces propriétés, les a_i inférieurs à α deviennent 0, et les autres 1. Si au moins un a_i était inférieur à α, il est devenu nul, et la valeur a de s pour le motif est également nulle. Dans le cas contraire, la valeur de a est 1.
Inversement, si on fait une α-coupe du motif flou A, la valeur a du sujet s devient 0 si elle est inférieure à α, et 1 dans le cas contraire. Si elle est inférieure à α, c'est que le minimum des valeurs des propriétés pour le sujet s était inférieur à α, donc que pour au moins une des propriétés p_i, sa valeur pour s est inférieure à α.

[150] Cette définition est reprise de la définition de [71] pour les ensembles flous

TAB. 9.6 – Comparaison de 3 indices des règles du tableau 6 selon plusieurs codages

no	Règle			flou			seuil 0,1			seuil 0,3			seuil 0,5			seuil 0,7			seuil 0,9		
	A	→	B	sup	conf	diff	s	conf	diff	s	conf	diff	s	conf	diff	s	conf	diff	s	conf	diff
1	abc	→	∅	3,2	1	0	6	1	0	4	1	0	3	1	0	2	1	0	1	1	0
2	ab	→	c	3,2	1	0,56	6	1	0,2	4	1	0,6	3	1	0,6	2	1	0,7	1	1	0,7
3	ac	→	b	3,2	0,84	0,25	6	1	0,2	4	1	0,3	3	0,75	0,25	2	0,67	0,17	1	0,5	0,1
4	bc	→	a	3,2	1	0,62	6	1	0,4	4	1	0,6	3	1	0,6	2	1	0,7	1	1	0,8
5	a	→	bc	3,2	0,84	0,52	6	1	0,4	4	1	0,6	3	0,75	0,45	2	0,67	0,47	1	0,5	0,4
6	b	→	ac	3,2	0,54	0,16	6	0,75	0,15	4	0,57	0,17	3	0,6	0,2	2	0,4	0,1	1	0,25	0,03
7	c	→	ab	3,2	0,73	0,41	6	0,75	0,15	4	1	0,6	3	0,75	0,45	2	0,67	0,47	1	0,33	0,23
8	∅	→	abc	3,2	0,32	0	6	0,6	0	4	0,4	0	3	0,3	0	2	0,2	0	1	0,1	0
9	ab	→	∅	3,2	1	0	6	1	0	4	1	0	3	1	0	2	1	0	1	1	0
10	a	→	b	3,2	0,84	0,25	6	1	0,2	4	1	0,3	3	0,75	0,25	2	0,67	0,17	1	0,5	0,1
11	b	→	a	3,2	0,54	0,16	6	0,75	0,15	4	0,57	0,17	3	0,6	0,2	2	0,4	0,1	1	0,25	0,05
12	∅	→	ab	3,2	0,32	0	6	0,6	0	4	0,4	0	3	0,3	0	2	0,2	0	1	0,1	0
13	ac	→	∅	3,8	1	0	6	1	0	4	1	0	4	1	0	3	1	0	2	1	0
14	a	→	c	3,8	1	0,56	6	1	0,2	4	1	0,6	4	1	0,6	3	1	0,7	2	1	0,7
15	c	→	a	3,8	0,86	0,48	6	0,75	0,15	4	1	0,6	4	1	0,6	3	1	0,7	2	0,67	0,47
16	∅	→	ac	3,8	0,38	0	6	0,6	0	4	0,4	0	4	0,4	0	3	0,3	0	2	0,2	0
17	bc	→	∅	3,2	1	0	6	1	0	4	1	0	3	1	0	2	1	0	1	1	0
18	b	→	c	3,2	0,54	0,10	6	0,75	-0,05	4	0,57	0,17	3	0,6	0,2	2	0,4	0,1	1	0,25	-0,05
19	c	→	b	3,2	0,73	0,14	6	0,75	-0,05	4	1	0,3	3	0,75	0,25	2	0,67	0,17	1	0,33	-0,07
20	∅	→	bc	3,2	0,32	0	6	0,6	0	4	0,4	0	3	0,3	0	2	0,2	0	1	0,1	0
21	a	→	∅	3,8	1	0	6	1	0	4	1	0	4	1	0	3	1	0	2	1	0
22	∅	→	a	3,8	0,38	0	6	0,6	0	4	0,4	0	4	0,4	0	3	0,3	0	2	0,2	0
23	b	→	∅	5,9	1	0	8	1	0	7	1	0	5	1	0	5	1	0	4	1	0
24	∅	→	b	5,9	0,59	0	8	0,8	0	7	0,7	0	5	0,5	0	5	0,5	0	4	0,4	0
25	c	→	∅	4,4	1	0	8	1	0	4	1	0	4	1	0	3	1	0	3	1	0
26	∅	→	c	4,4	0,44	0	8	0,8	0	4	0,4	0	4	0,4	0	3	0,3	0	3	0,3	0
27	∅	→	∅	10	1	0	10	1	0	10	1	0	10	1	0	10	1	0	10	1	0

Voyons maintenant l'effet d'une α-coupe d'un motif flou sur le support de ce motif.

Propriété 9.5.2. *L'intervalle de R ayant pour extrémités les valeurs extrêmes des supports des α-coupes d'un motif flou quand α décrit]0,1] contient le support du motif flou. De plus, si ce support n'est pas nul, il n'atteint les bornes de l'intervalle que lorsqu'elles sont confondues.*

Preuve. *Soit A un motif flou sur l'ensemble $\mathcal{S}$ des N sujets. Appelons n_0 (resp. n_1, n_2) l'ensemble des sujets pour lesquels la valeur est 0 (resp. 1, strictement comprise entre 0 et 1). Soient m le minimum des supports des α-coupes du motif, et M leur maximum.*

1. *S'il n'y a aucun sujet pour lequel la valeur dépasse 0,5, le support du motif flou est 0 par définition. Une α-coupe avec α=0,6 ne produit que des valeurs nulles. Une telle α-coupe a un support nul, donc le minimum m est nul. Le maximum est différent de 0 dès qu'il y a au moins un sujet pour lequel la valeur n'est pas nulle. En prenant pour pour α cette valeur, ce sujet obtient une valeur de 1 dans l'α-coupe du motif. Et il peut y avoir de 1 à N sujets dans ce cas. Donc le maximum peut être n'importe quel nombre entier de 0 à N. Et le support du motif flou est la borne gauche de l'intervalle [m,M], mais cet intervalle n'est pas réduit à 1 point en général.*
2. *S'il y a au moins un sujet pour lequel la valeur dépasse 0,5, le support est la somme des valeurs des sujets. Si on prend α=1, l'α-coupe du motif remplace toutes les valeurs différentes de 1 par 0. Donc les sujets de n_0, n_2 ont pour valeur 0 et les sujets de n_1 ont pour valeur 1. Le support de cette α-coupe est donc égal au nombre d'éléments de n_1. Si on prend pour α une valeur non nulle inférieure ou égale à la plus petite valeur non nulle, l'α-coupe du motif remplace toutes les valeurs différentes de 0 par 1. Le support de cette α-coupe est donc égal au nombre d'éléments de n_1 et de n_2. Comme les éléments de n_2 ont des valeurs comprises strictement entre 0 et 1, la somme de ces valeurs est comprise strictement entre 0 et le nombre d'éléments de n_2, s'il n'est pas nul. Le support du motif flou est la somme des éléments de n_2 et des éléments de n_1, donc sa valeur est comprise strictement entre le nombre d'éléments de n_1 et le nombre d'éléments de n_1 et n_2 si n_2 n'est pas nul. Et s'il est nul, ces deux valeurs sont égales au support du motif flou. On a ainsi montré que le support est compris entre les supports de deux α-coupes, pour α=0 et α=1. Reste à montrer que ces deux supports sont les valeurs extrêmes des supports des α-coupes. Ceci vient du fait que le support d'une α-coupe est fonction décroissante de α. En effet pour une valeur donnée de α les éléments de n_2 qui sont supérieurs ou égaux à α prennent la valeur 1, et les autres la valeur 0, les éléments de n_1 et de n_0 ne changeant pas de valeurs. Si α diminue, le nombre d'éléments de n_2 dont la valeur est supérieure ou égale à α augmente, ou reste stationnaire, alors que c'est l'inverse pour ceux de n_2 dont la valeur est inférieure à α. Comme le support de l'alpha-coupe du motif est formé de deux parties, l'une ne dépendant pas de α, et l'autre égale au nombre d'éléments de n_2 égaux à 1, et que ce nombre est fonction décroissante d'α, nous avons bien montré que le support d'une α-coupe est fonction décroissante de α. Donc les valeurs extrêmes des supports des α-coupes sont les deux valeurs que nous avons calculé précédemment.*

La même propriété existe pour la confiance (la démonstration, qui ne figure pas dans ce document, se fait de façon similaire).

Mais cette propriété n'est pas vérifiée pour la différence, dont la formule contient les 3 supports. Un seul exemple suffit pour le prouver, il figure dans les tableaux 7 et 8.

motif	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10	support
A	0,2	0,4	1	1	0,2	0,2	0,2	1	0	1	5,2
B	0	0,2	0,2	0,2	0,4	0,2	0,2	0	0	1	2,4
A∪B	0	0,2	0,2	0,2	0,2	0,2	0,2	0	0	1	2,2

TAB. 9.7 – Valeurs des motifs A et B pour 10 sujets

On voit dans le tableau 8 que la différence pour l'ensemble flou est de 0,18, alors que les valeurs extrêmes pour les α-coupes sont 0 et 0,15.

		valeur de α		
indice	flou	]0, ;0,2]	]0,2 ;0,4]	]0,4 ;1]
support de A	5,2	9	5	4
support de B	2,4	7	2	1
support de A∪B	2,2	7	1	1
confiance	0,42	0,78	0,2	0,25
différence	0,18	0,08	0	0,15

TAB. 9.8 – Valeurs de la différence pour la règle A→B sur le motif flou et ses α-coupes

9.6 Comparaison des règles d'association floues et des règles floues initiées par Zadeh

Dans leur article[72], D. Dubois et H. Prade, exposent leurs définitions de règles floues[151]. Il définissent déjà les règles non floues, de la forme "si/alors", qu'on appellera strictes (pour traduire le terme anglais "crisp" qui oppose "crisp rules" et "fuzzy rules"), de la façon suivante :

si $\mathcal{R}$ est la règle "si x est A alors y est B"

où A et B sont des ensembles classiques, x et y sont des variables à valeurs dans les domaines de U et V, cela se traduit, avec l'aide des fonctions μ_A et μ_B, qui sont, pour ces règles strictes, à valeurs dans {0,1}, par

si μ_A(u)=1 alors μ_A(v)=1

et les paires de valeurs (u,v) de (x,y) compatibles avec la règle $\mathcal{R}$ forment un ensemble R tel que

$$\mu_A(u) \wedge \mu_B(v) \leq \mu_R(u,v) \leq n(\mu_A(u)) \vee \mu_B(v)$$

où $\wedge$, $\vee$ et n() représentent respectivement les opérateurs booléens de conjonction, de disjonction et de négation.

Leur interprétation de l'inégalité gauche est le modèle de règle basé sur la conjonction (ou encore possibilité), alors que l'inégalité droite représente le modèle basé sur l'implication (ou encore nécessité). En effet si on a simultanément $\mu_A(u)$=1 et $\mu_B(v)$=1, alors la partie gauche de cette double inégalité est égale à 1, donc $\mu_R(u,v)$ est forcé à 1 et la règle est vraie pour (u,v). Par contre si on a $\mu_A(u)$=1 et $\mu_B(v)$=0, alors la partie droite de cette double inégalité est égale à 0, donc $\mu_R(u,v)$ est forcé à 0 et la règle est fausse pour (u,v). Dans cette inégalité, on ne spécifie

[151] Notons que les ensembles flous qu'ils utilisent dans cet article sont des ensembles qu'ils appellent "normalisés", c'est-à-dire des ensembles F pour lesquels on peut trouver au moins un élément ω du référentiel Ω tel que $\mu_F(\omega)$=1.

pas la valeur associée aux couples qui ne vérifient pas la partie gauche de la règle, qu'ils vérifient ou non sa partie droite.
Si nous transposons cette définition à nos propriétés de $\mathcal{P}$ vérifiées de façon certaine par des sujets de $\mathcal{S}$, nous obtenons que si A et B sont deux motifs sur $\mathcal{S}$, alors l'ensemble R des sujets s compatibles avec la règle $\mathcal{R}$: A→B sont tels que :

$\mu_{A'}(s) \wedge \mu_{B'}(s) \leq \mu_R(s) \leq n(\mu_{A'}(s)) \vee \mu_{B'}(s)$.

Si s décrit l'ensemble des sujets, quand on additionne toutes les inégalités obtenues, on obtient l'inégalité suivante :

$card(A' \cap B') \leq card(R) \leq card(\overline{A'} \cup B')$

soit :

$support(A \cup B) \leq card(R) \leq N - card(A' \cap \overline{B'})$

Dans cette inégalité, on voit que le nombre d'éléments qui vérifient la règle R est compris entre le nombre d'éléments qui vérifient sa partie gauche et sa partie droite, et le nombre d'éléments qui ne la contredisent pas. La règle prise au sens implicatif correspond à la borne droite, c'est-à-dire est vérifiée par les éléments qui vérifient sa partie gauche et sa partie droite mais également par tous ceux qui ne vérifient pas sa partie gauche, qu'ils vérifient ou non sa partie droite.

L'extension de ce formalisme des règles aux ensembles flous s'est fait de multiples façons. Voici quelques formules extraites de [233] permettant de calculer la valeur d'une règle d'implication floue :

Définition 9.6.1. *Valeur d'une règle d'implication floue A→B*
Si A et B sont deux motifs flous sur un ensemble $\mathcal{S}$, s un élément quelconque de $\mathcal{S}$, a et b les valeurs respectives $\mu_{A'}(s)$ et $\mu_{B'}(s)$ de s pour A et B, rf_i une règle d'implication floue, la valeur $\mu_{rfi}(s)$ est

- $\mu_{rf1}(s) = min(1, 1 - a + b)$ *si rf1 est la règle floue de Lukasiewicz*
- $\mu_{rf2}(s) = 1 - a + a\dot{b}$ *si rf2 est la règle floue "probabiliste"*
- $\mu_{rf3}(s) = max(1 - a, b)$ *si rf3 est la règle floue de Kleene-Dienes*
- $\mu_{rf4}(s) = \begin{cases} 1 & si\ a \leq b \\ b & sinon \end{cases}$ *si rf4 est la règle floue de Brouwer-Gödel*
- $\mu_{rf5}(s) = \begin{cases} 1 & si\ a \leq b \\ b/a & sinon \end{cases}$ *si rf5 est la règle floue "quotient"*
- $\mu_{rf6}(s) = 1 - a + a^2\dot{b}$ *si rf6 est la règle floue "quadratique"*
- $\mu_{rf7}(s) = max(1 - a, min(a, b))$ *si rf7 est la règle floue "Early Zadeh"*

Ces règles coïncident toutes, pour les relations non floues, avec la règle d'implication classique. Il est indiqué dans [233] que pour les règles rf1, rf4 et rf5, la valeur de la règle floue est 1 dès que a≤b, alors qu'elle ne vaut 1 que si a=0 ou b=1 pour 4 autres règles, et que la valeur est 0 dès que a>0 et b=0 pour les règles rf6 et rf7, alors qu'elle ne vaut 0 que pour a=1 et b=0 pour les autres règles. On peut trouver dans [139] l'examen de 17 propriétés caractérisant les règles d'implication floues pour les 3 règles rf1, rf3 et rf4. Les résultats de cet examen nous font préférer la première des trois si on désire se rapprocher le plus possible de la règle d'implication stricte.
Une règle floue ainsi définie par ses valeurs pour tous les sujets est une nouvelle propriété floue dont nous pouvons calculons le support, en additionnant les valeurs pour tous les sujets. A titre d'illustration, dans le tableau 9 figurent les sept supports des règles d'association construites sur le tableau 3.

Avant d'examiner de plus près les résultats, remarquons que le choix d'une règle floue de ce type n'est pas compatible avec les algorithmes de recherche de règles que nous utilisons. En effet,

num	motif	A	B	sAB	sA	sB	conf	diff	rf1	rf2	rf3	rf4	rf5	rf6	rf7
1	abc	abc	∅	3.2	3.2	10	1.00	0.00	10	10.0	10	10	10	9.0	8.6
2	abc	ab	c	3.2	3.2	4.4	1.00	0.56	10	9.4	9.2	10	10	8.8	8.6
3	abc	ac	b	3.2	3.8	5.9	0.84	0.25	9.4	9.2	9	9.2	9.3	8.7	8.4
4	abc	bc	a	3.2	3.2	3.8	1.00	0.62	10	9.4	9	10	10	8.8	8.6
5	abc	a	bc	3.2	3.8	3.2	0.84	0.52	9.4	8.8	8.4	9.2	9.3	8.5	8.4
6	abc	b	ac	3.2	5.9	3.8	0.54	0.16	7.3	7.1	6.9	6	6.2	6.7	6.5
7	abc	c	ab	3.2	4.4	3.2	0.73	0.41	8.8	8.2	8	7.2	7.2	8.0	8
8	abc	∅	abc	3.2	10	3.2	0.32	0.00	3.2	3.2	3.2	3.2	3.2	3.2	3.2
9	ab	ab	∅	3.2	3.2	10	1.00	0.00	10	10.0	10	10	10	9.0	8.6
10	ab	a	b	3.2	3.8	5.9	0.84	0.25	9.4	9.2	9	9.2	9.3	8.7	8.4
11	ab	b	a	3.2	5.9	3.8	0.54	0.16	7.3	7.1	6.9	6	6.2	6.7	6.5
12	ab	∅	ab	3.2	10	3.2	0.32	0.00	3.2	3.2	3.2	3.2	3.2	3.2	3.2
13	ac	ac	∅	3.8	3.8	10	1.00	0.00	10	10.0	10	10	10	9.3	9
14	ac	a	c	3.8	3.8	4.4	1.00	0.56	10	9.4	9.2	10	10	9.1	9
15	ac	c	a	3.8	4.4	3.8	0.86	0.48	9.4	8.8	8.6	7.8	7.8	8.6	8.6
16	ac	∅	ac	3.8	10	3.8	0.38	0.00	3.8	3.8	3.8	3.8	3.8	3.8	3.8
17	bc	bc	∅	3.2	3.2	10	1.00	0.00	10	10.0	10	10	10	9.0	8.6
18	bc	b	c	3.2	5.9	4.4	0.54	0.10	7.3	7.2	7.1	6	6.2	6.8	6.5
19	bc	c	b	3.2	4.4	5.9	0.73	0.14	8.8	8.7	8.6	7.2	7.2	8.2	8
20	bc	∅	bc	3.2	10	3.2	0.32	0.00	3.2	3.2	3.2	3.2	3.2	3.2	3.2
21	a	a	∅	3.8	3.8	10	1.00	0.00	10	10.0	10	10	10	9.3	9
22	a	∅	a	3.8	10	3.8	0.38	0.00	3.8	3.8	3.8	3.8	3.8	3.8	3.8
23	b	b	∅	5.9	5.9	10	1.00	0.00	10	10.0	10	10	10	9.2	8.7
24	b	∅	b	5.9	10	5.9	0.59	0.00	5.9	5.9	5.9	5.9	5.9	5.9	5.9
25	c	c	∅	4.4	4.4	10	1.00	0.00	10	10.0	10	10	10	9.1	8.8
26	c	∅	c	4.4	10	4.4	0.44	0.00	4.4	4.4	4.4	4.4	4.4	4.4	4.4
27	∅	∅	∅	10	10	10	1.00	0.00	4.4	4.4	4.4	4.4	4.4	4.4	4.4

TAB. 9.9 – Les supports des sept règles floues correspondant aux règles du tableau 4

une règle est générée en coupant un motif en deux parties. Lors de cette opération, on ne connaît que les valeurs des supports des motifs, et non leurs valeurs pour chaque sujet qui ont permis de calculer le support. Bien sûr, on peut à tout moment, si une règle paraît intéressante, retourner voir les valeurs des sujets afin de calculer pour chacun la valeur x de la partie gauche et y de la partie droite. On fait alors la somme de ces valeurs pour chaque sujet et on obtient $\mu_R(s)$. Par contre, retrouver ce nombre à partir des supports des motifs A, B et A∪B est impossible en général. Il faudrait stocker pour chaque motif toutes les informations nécessaires au calcul des valeurs correspondant à une définition de règle floue donnée. Ceci est difficilement réalisable sur des bases de données importantes. Ces définitions de règles floues sont plus adaptées à la validation de règles qu'à leur découverte automatique.
Cette remarque étant faite, on peut toutefois noter que le support de la règle rf1 vérifie une propriété remarquable :

Propriété 9.6.1. *Le support de rf1 est support(A→B)=N+support(A∪B)-support(A)*

Preuve. *on a*

$$\begin{aligned} support(A \rightarrow B) &= \sum_{s \in \mathcal{S}} \mu_{rf1}(s) \\ &= \sum_{s \in \mathcal{S}} min(1, 1 - \mu_{A'}(s) + \mu_{B'}(s)) \\ &= \sum_{s \in \mathcal{S}} min(\mu_{A'}(s), \mu_{B'}(s)) + 1 - \mu_{A'}(s) \\ &= \sum_{s \in \mathcal{S}} min(\mu_{A'}(s), \mu_{B'}(s)) + \sum_{s \in \mathcal{S}} 1 - \sum_{s \in \mathcal{S}} \mu_{A'}(s) \\ &= support(A \cup B) + N - support(A) \end{aligned}$$

On peut donc en rajoutant cet indice, N+support(A∪B)-support(A), avoir la valeur de la règle floue de Lukasiewicz à partir des supports des motifs A∪B et A sans retourner voir les valeurs de chaque sujet. Vu les nombreuses qualités de la règle floue de Lukasiewicz que nous avons évoquées (pas moins de 17), et vu les problèmes calculatoires que poseraient les autres règles floues, pour faire de l'extraction automatique de règles à partir des matrices sujets×propriétés à valeurs dans [0,1], nous la choisissons comme seule définition de règle floue. Cela revient à reprendre tout simplement les règles strictes que nous avons définies à partir de relations floues et que nous avons appelées règles d'association floue, et rajouter un indice de qualité qui est le support de la règle défini par support(A→B)=N+support(A∪B)-support(A). Et nous pouvons désormais l'appeler *règle floue d'association.*

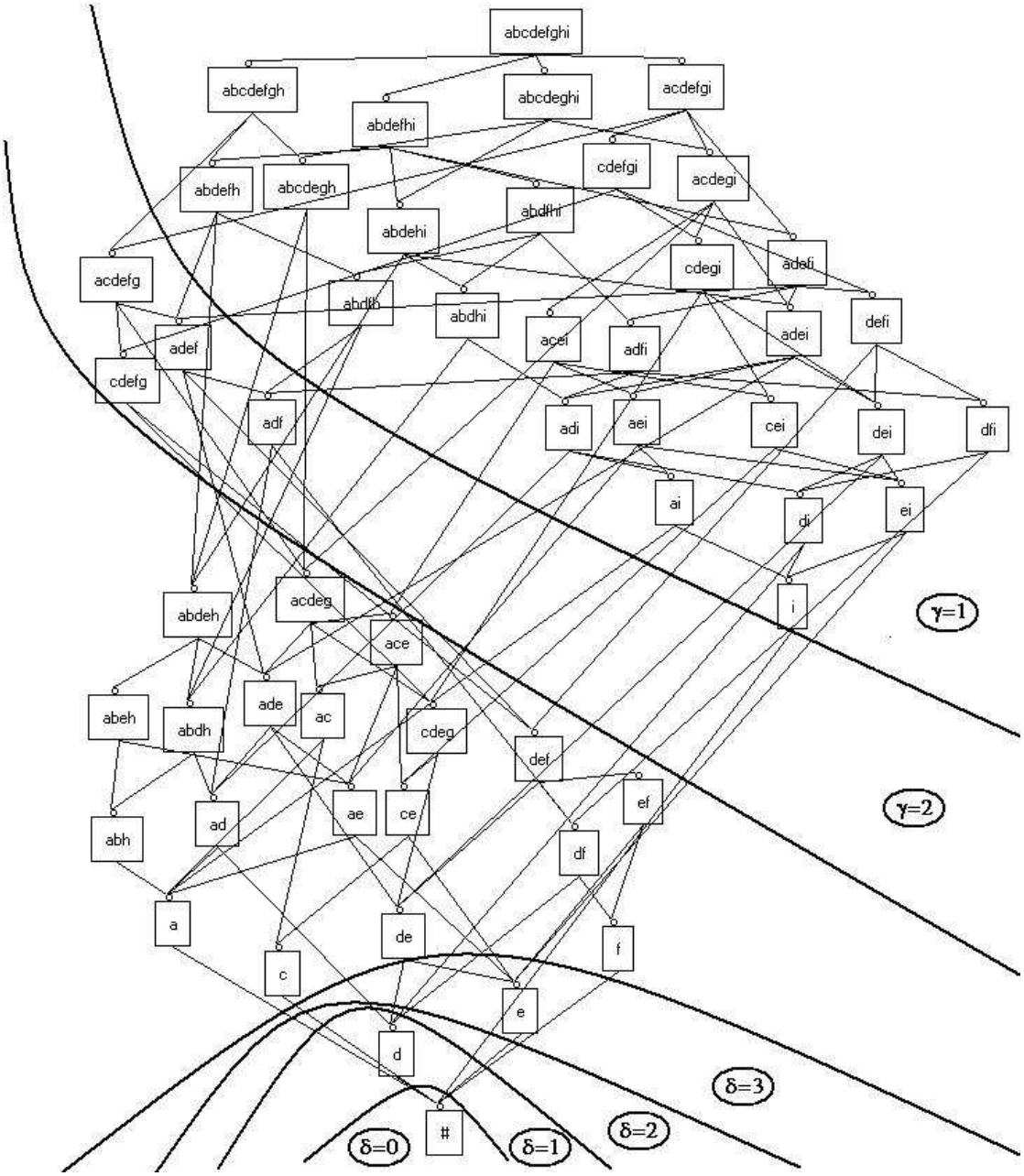


FIG. 9.3 – Le treillis des motifs fermés du tableau 5 avec γ=0 et δ=0.

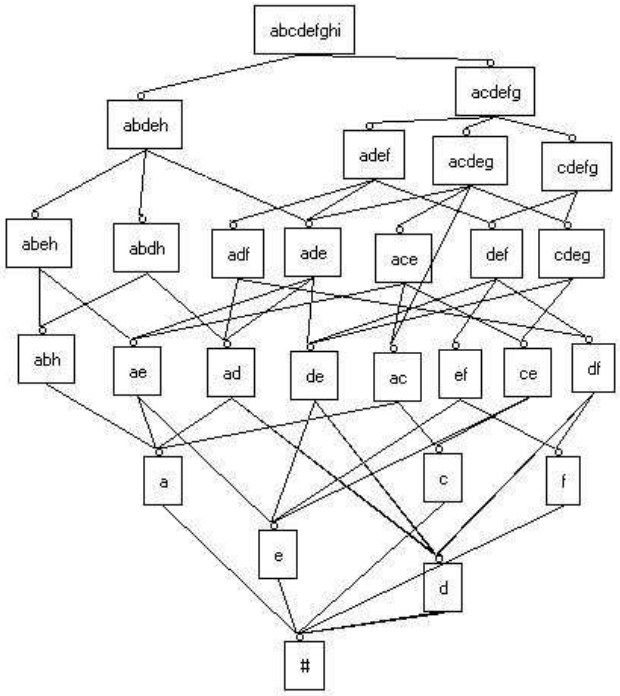


FIG. 9.4 – Le treillis des motifs fermés du tableau 5 avec γ=1 et δ=0.

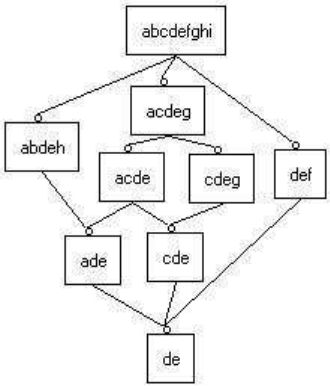


FIG. 9.5 – Le treillis des motifs fermés du tableau 5 avec γ=2 et δ=3.

9.7 Comparaison avec des méthodes proches

Nous venons de proposer une méthode de construction de règles d'association utilisant un codage flou d'une propriété ordinale afin d'éviter la perte de qualité résultant d'un recodage binaire. Nous allons la comparer avec deux méthodes proches proposées par des chercheurs, l'une qui étend les dépendances fonctionnelles sur des variables qualitatives à des variables quantitatives, et l'autre qui généralise les règles d'association binaires à des règles ordinales en remplaçant l'indice d'implication statistique de R. Gras par un indice d'implication statistique ordinal.

9.7.1 Dépendances fonctionnelles floues

Jean-Marc Petit [194] propose de relaxer la définition de la dépendance fonctionnelle définie sur des groupes de propriétés qualitatives à des groupes de variables quantitatives de la façon suivante :

Définition 9.7.1. *Dépendance fonctionnelle.*
Si on dispose d'une relation R sur liant un ensemble de sujets S et un ensemble de propriétés P, si A et B sont deux ensembles de propriétés de P, on dit qu'une dépendance fonctionnelle A→B est valide dans R si

$$\forall s_1, s_2 \in S, si\ \forall X \in A, X(s_1) = X(s_2)\ alors\ \forall Y \in B, Y(s_1) = Y(s_2)$$

Définition 9.7.2. *Dépendance fonctionnelle avec relaxation de l'égalité.*
Une dépendance fonctionnelle A→B est valide dans T si

$$\forall s_1, s_2 \in S, si\ \forall X \in A, \epsilon_1 \leq |X(s_1) - X(s_2)| \leq \epsilon_2\ alors\ \forall Y \in B, \epsilon_1 \leq |Y(s_1) - Y(s_2)| \leq \epsilon_2$$

Voyons les conséquences que cela a sur l'établissement de la règles A→B, A et B étant deux propriétés dont les valeurs varient entre 0 et 1 selon le tableau U .

sujets	A	B
s1	0	0,4
s2	0,09	0,6
s3	0,11	0,6
s4	0,3	0,7
s5	0,4	0,7
s6	0,5	1
s7	0,5	0,9
s8	0,7	0,6
s9	0,8	0,6
s10	1	0,3

a1	a2	a3	a4	a5	b1	b2
1	0	0	0	0	1	0
1	1	0	0	0	1	1
1	1	0	0	0	1	1
1	1	1	0	0	1	1
1	1	1	1	0	1	1
0	1	1	1	0	0	1
0	1	1	1	0	0	1
0	0	1	1	1	1	1
0	0	0	1	1	1	1
0	0	0	0	1	1	0

TAB. 9.10 – A gauche, le tableau U de deux propriétés quantitatives A et B, à droite les classes selon A et B avec $\epsilon_1 = 0$ et $\epsilon_2 = 0,4$

Dans le tableau U, il y a presque autant de valeurs différentes de A que de sujets, soit 9 classes pour 10 sujets. De plus, les deux seuls sujets qui sont dans la même classe, les sujets s6 et s7 se trouvent avoir des valeurs très proches de B, qui sont 1 et 0,9, mais pas égales, ce qui rend la dépendance fonctionnelle A→B invalide selon la première définition. Et même si elle avait été valide, avec par exemple les valeurs respectives de 0,51 et 0,49 de A pour s6 et s7, cette

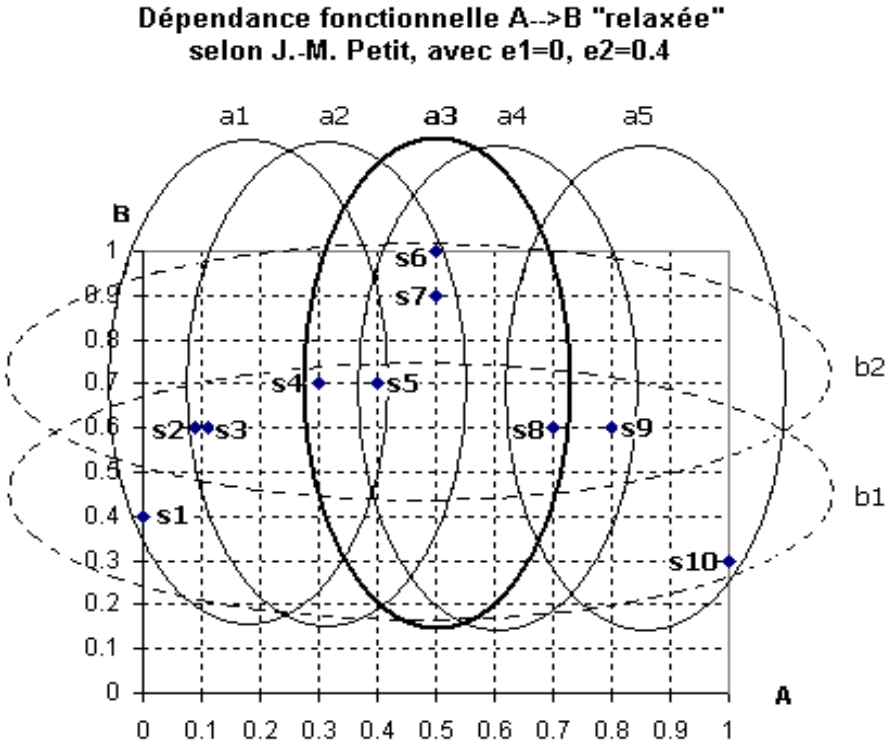


FIG. 9.6 – Les 10 sujets de la relation R du tableau U dans le plan AxB et leur répartition dans les classes selon A et B avec $\epsilon 1 = 0$ et $\epsilon 1 = 0,4$

dépendance fonctionnelle n'aurait pas apporté d'information, car une classification à 10 classes est forcément plus fine qu'une autre classification de ces 10 éléments, donc on aurait A→X pour toute propriété X.

A côté de ce tableau, on a représenté le plus petit nombre de classes selon A obtenues de telle façon que la propriété C est une classe selon A si pour tout sujet s_i et s_j vérifiant $C(s_i) = C(s_j)$, on a $\epsilon_1 \leq |A(s_i) - A(s_j)| \leq \epsilon_2$. On a obtenu ainsi 5 classes pour A, notées de a1 à a5, et en procédant de la même façon, 2 classes pour B, notées b1 et b2.

On peut voir sur le graphique 7 la représentation des 10 sujets dans l'espace des 2 propriétés A et B, et la répartition des sujets selon les classes. C'est cette répartition qui fait qu'on a la dépendance fonctionnelle A→B valide selon la définition avec relaxation des égalités. Les ellipses indiquent les classes. Celles en traits pleins sont les classes selon A, et celles en pointillés sont les classes selon B. La classe a3 (en trait gras) contient les 5 sujets de s4 à s8 dont les valeurs pour A varient de 0,3 à 0,7, et ils se retrouvent tous dans la classe b2, leurs valeurs pour B allant de 0,6 à 1. Les éléments de la classe a1 se retrouvent tous dans la classe b1, comme ceux de la classe a5, ceux des classes a2 et a4 se retrouvant tous à la fois dans la classe b1 et dans la classe b2. La dépendance fonctionnelle A→B ainsi définie assure une liaison de proche en proche entre les valeurs de A et celles de B. Toutefois cette liaison n'est pas la liaison "croissante" qu'on retrouve habituellement dans les règles d'association, la règle A→B signifiant que B est une fonction croissante de A, quand on dichotomise les propriétés A et B par seuillage, ou quand on étend les règles les règles d'association à des valeurs autres que binaires [106].

On peut voir dans dans le graphique suivant une dépendance fonctionnelle valide selon la deuxième définition qui ne parait pas du tout croissante. Chaque sujet est représenté par un losange. Si on déplace une fenêtre de largeur 0,4 dans ce graphique (et de hauteur illimitée), on peut remarquer que les points figurant dans cette fenêtre forment une sous-nuage entièrement inclus dans un zone carrée de côté 0,4. Toutefois, le point x en forme de losange, de valeurs A(x)=0,45 et B(x)=0,35 ne figure pas dans cette zone, car si on le compare avec le point y tel que A(y)=0,39 et B(y)=0,77, il lui est distant de moins de 0,4 selon A et de plus de 0,4 selon B. Donc si la donnée correspondante appartenait à la base de données, la dépendance fonctionnelle "relaxée" A→B ne serait plus valide.

Dans le cas quantitatif, on obtient ainsi moins de classes avec la deuxième définition, deux sujets pouvant être dans la même classe même s'ils ont des valeurs différentes pour une même propriété de A. Mais l'inconvénient est que les classes ne sont plus disjointes.

Cette "fuzzification" des dépendances fonctionnelles permet de créer un ensemble de dépendances fonctionnelles "relaxées" sur lequel agissent les règles d'inférence des dépendances fonctionnelles habituelles (reflexivité, transitivité, augmentation).

9.7.2 L'indice d'implication ordinal

Dans le même but d'étendre l'extraction de règles binaires à des règles ordinale, S. Guillaume définit dans sa thèse un indice [106] d'implication ordinale mesurant la qualité d'une règle d'association sur des propriétés qui ne sont pas binaires, ni formées de modalités incomparables, comme les divers pays, les couleurs, mais de modalités ordonnées, comme "pas du tout", "un peu", "beaucoup". Cette généralisation de l'indice d'implication de R. Gras est illustrée sur le graphique par une règle donnée dans sa thèse comme exemple. Nous voyons que cette règle "exacte" selon cet indice ressemble à notre règle d'association floue dans la mesure où la proportion de points sous la première bissectrice diminue quand on se rapproche du coté inférieur droit du graphique. Toutefois, la règle obtenue ainsi ne prend en compte des variables quantitatives qu'une fois celle-ci recodées en variables ordinales, ce codage étant proposé avant toute

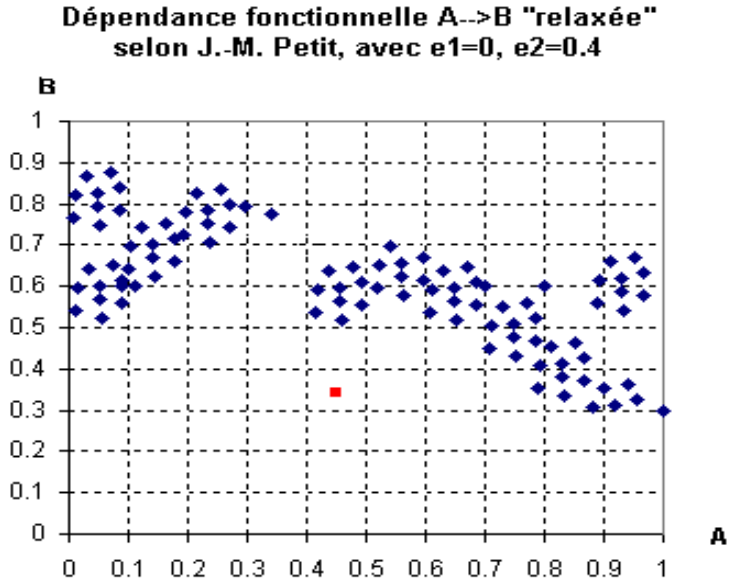


FIG. 9.7 – Une dépendance fonctionnelle A→B avec $\epsilon 1 = 0$ et $\epsilon 1 = 0,4$

extraction.

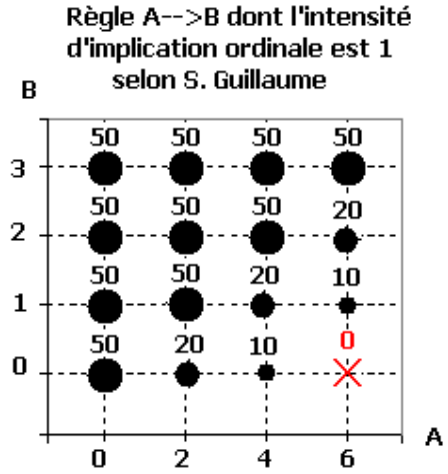


FIG. 9.8 – Une règle d'implication ordinale

Notons que l'usage de cet indice ne se limite pas à son calcul et à sa comparaison avec un seuil. Il y a utilisation de règles d'inférences pour assurer la cohérence du jeu de règles extraites avec cet indice. Pour assurer cette cohérence, on crée des propriétés A' en restreignant l'intervalle des valeurs prises par A. Les règles A'→B ainsi définies sont appelées règles partielles, la règle A→B étant la règle totale. Voici le principe d'extraction de la la règle A→B :
si la règle totale A→B est reconnue comme valide (son indice d'implication ordinal est supérieur à un seuil donné à l'avance), on contrôle les règles partielles A'→B afin de signaler celles qui contredisent la règle générale (leur indice est inférieur au seuil), et si la règle générale n'est pas

valide, on recherche une règle partielle valide.

Par contre, aucune technique n'est proposée pour faire une règle avec plus de deux propriétés ordinales, ni pour voir la validité d'autre règles d'inférence que la contraposition sur le jeu de règles ordinales.

9.7.3 Retour sur les règles d'association floues

La façon de rendre floues les dépendances fonctionnelle est très proche de celle que nous avons choisie pour les règles d'association. On peut le voir en comparant le graphique 10 des dépendances fonctionnelles classiques du chapitre 2, qui contient également le graphique des règles d'association classiques, à celui des dépendances fonctionnelles floues du paragraphe précédant, et à celui-ci représentant une règle d'association floue. Comme dans le graphique de la dépendance fonctionnelle floue, chaque point est représenté par un losange. La règle est exacte A→B car comme tous les losanges figurent au dessus de la bissectrice du plan AxB, sa confiance est 1. Si le point en forme de carré ayant pour valeur de A 0,53 et pour valeur de B 0,36 figurait dans les données, la règle aurait une confiance à peine inférieure à 1. Elle ne serait plus exacte mais encore de très bonne qualité. Son support est d'autant plus grand qu'il y a de points dans la zone en haut à droite du graphique. En s'inspirant du graphique, on peut donner la définition suivante

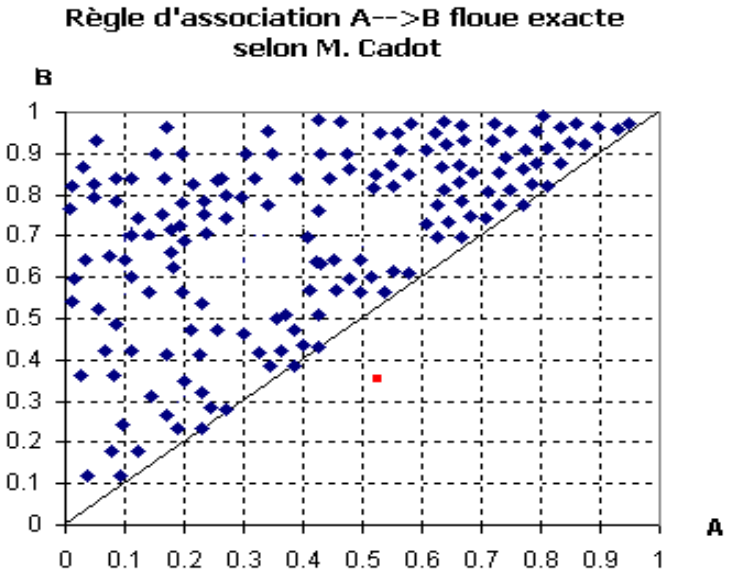


FIG. 9.9 – Une règle d'association floue A→B de confiance 1

d'une règle d'implication floue exacte quand on est dans un tableau de propriétés prenant leurs valeurs dans l'intervalle (0,1) :

Définition 9.7.3. *Règle d'association floue exacte généralisée.*
Une règle A→B est valide dans T si

$$\forall s \in S, si\ \forall X \in A,\ \forall Y \in B, X(s) \leq Y(s)$$

Avec une telle définition, qui n'évoque pas le support, on obtient un jeu de règles sur lequel peuvent agir les règles d'inférences, mais contenant des règles de support nul.

Toutefois, il y a une différence essentielle entre la relation de dépendance fonctionnelle floue que nous venons de voir et la règle d'association floue. La première n'indique pas une croissance de B en fonction de A, contrairement à la seconde. En ce sens, la règle d'association ordinale de S. Guillaume ressemble plus à la règle d'association floue que la dépendance fonctionnelle floue. L'ajout qu'elle fait de règles partielles permet même de trouver des règles quand la fonction est croissante par morceaux, à condition toutefois d'avoir fait le précodage convenable, mais son indice n'a pas été généralisé à des règles comportant plus d'une propriété à gauche (ou a droite).

9.8 Utilisation sur des données réelles

Les règles d'association floues ont été implémentées et utilisées sur un corpus de textes de géologie de l'Inist, dans le cadre d'un projet "IDL-IDST Axe veille"[152]. Elles ont montré leur efficacité pour trouver des liens entre différentes classifications du corpus de textes selon les mots-clés. Les résultats trouvés par les règles d'association floues dépassant un seuil de support et un seuil de confiance donnés au préalable ont permis d'établir les liens de type $A_i \rightarrow B_j$ où A et B représentent deux classifications, telles que A a moins de classes que B, et i et j sont des indices de classes. La quasi-identité des liens établis "à la main" par l'expert du domaine avec ceux trouvés par extraction de règles d'association a permis de valider les définitions choisies pour les règles d'association floues.

Projet ILD-ISTC Axe Veille - LORIA/INIST - 2004-2005 - Données mots-clés, classes, coefficients

	nb classes :	133		nb mots_cles	1330				
mots_cles	c5_00	c5_01	c5_02	c5_03	c5_04	c5_05	c5_06	c5_07	c5_08	c5_09	c
Aéroport	0.06	0.03	0.09	0.21	0.15	0	0	0	0.03	0	
Abrasion	0	0	0	0	0	0	0.11	0	0	0	
Absorption	0	0	0	0	0.24	0	0.25	0.07	0.2	0.06	
Absorption eau	0	0	0	0.17	0.13	0.03	0	0	0	0	
Accélérom&egrave	0	0	0.13	0	0	0	0	0	0	0.59	
Acidification	0	0	0	0	0.14	0.25	0.23	0	0	0	
Action chaleur	0	0.97	0	0	0.23	0	0	0.12	0	0	
Action climatique	0	0	0	0	0.02	0	0.01	0.1	0.02	0.08	
Action gel	0	0	0	0	0.31	0.11	0.08	0	0.43	0	
Action séisme	0	0.1	1.07	0.1	0.02	0	0	0	0.11	0.05	
Action végétation	0	0	0	0	0	0	0	0	0.02	0	
Action vague	0	0	0.32	0.09	0	0.14	0	0	0.01	0	
Action vent	0	0	0	0	0	0	0	0	0.15	0	
Action vibration	0	0	0	0	0.01	0	0	0	0	0.03	
Activité microbienne	0	0.59	0	0	0.34	0.15	0	0	0.03	0	
Adhérence	0	0	0	0	0.07	0	0	0	0.14	0	
Adoucissement mécaniqu	0.05	0.05	0	0	0	0.11	0	0	0	0.05	

FIG. 9.10 – Les données de géologie

[152] Je remercie toutes les personnes impliquées dans ce projet pour l'aide qu'elles m'ont apportée, Pascal Cuxac, Claire François, Patricia Gautier, Alain Lelu, Xavier Polanco, et les stagiaires de l'Esial Marie Hubert et David Racodon pour la réalisation de l'implémentation des règles floues et leur visualisation sous forme de graphique

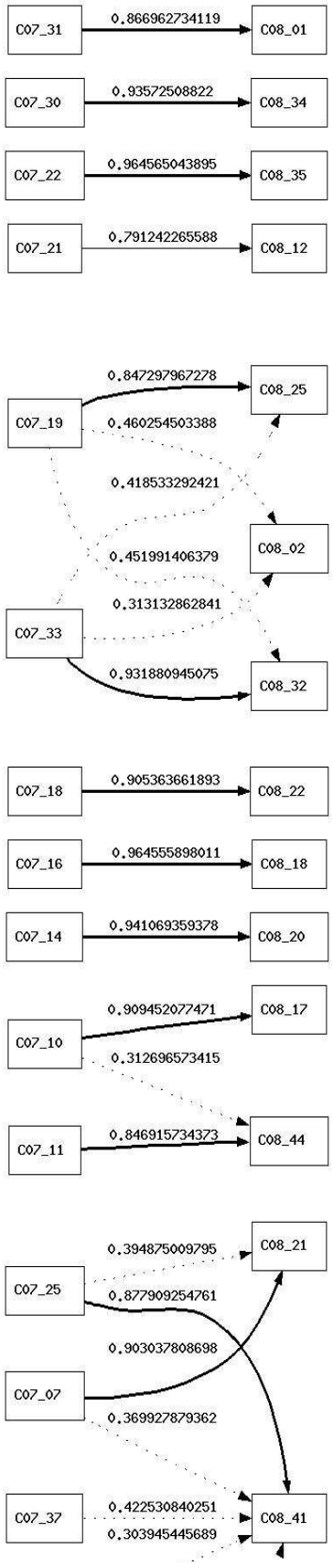


FIG. 9.11 – Quelques règles d'association floues extraites des données de géologie

On peut voir dans la figure 10 un extrait des données sur lesquelles les règles de classification floues ont été découvertes. Il y a 1330 mots-clés et 133 classes composant 4 classifications distinctes. Les coefficients figurant à l'intersection de la ligne d'un mot-clé et de la colonne d'une classe indique la participation du mot-clé à la formation de la classe. Par exemple le coefficient 0.97 figurant à l'intersection de la ligne "Action Chaleur" et de la colonne c5-01 indique que ce mot-clé a fortement participé à la création de la classe 1 de la classification en 20 classes appelée C5. Dans la figure 11 l'extrait du graphique montre une partie des règles d'association floues découvertes entre les classes de 2 classifications. L'épaisseur du trait et le coefficient indiquent la valeur de la confiance. Par exemple, le trait entre C07-31 et c08-01 de 0.86 indique qu'il y a une règle d'association entre la classe 31 de la classification C7 à 40 classes et la classe 1 de la classification C8 à 50 classes, et que cette règle a pour confiance 0.86.

Il reste maintenant à trouver une sémantique permettant d'interpréter des liens mettant en jeu plus de deux classes. Pour établir cette sémantique, la connaissance de l'expert des données est indispensable, car elle est liée au contexte, et il peut même y en avoir plusieurs comme le montrent les recherches de J.-M. Petit [194] au sein du groupe de recherche "biopuces". Dans ce projet, cette sémantique doit permettre de mesurer l'évolution d'une classification à l'aide du jeu de règles d'association floues extraites des classifications des textes d'un corpus. De façon plus générale, on devrait pouvoir résoudre avec une sémantique du même type des problèmes de choix du nombre de facteurs dans une analyse factorielle, non pas pour choisir le meilleur, mais pour éliminer les factorisations intermédiaires et ne garder que celles suffisamment stables pour exprimer des classification d'items en concepts selon différents champs[153].

[153]Par exemple, la violence peut s'exprimer selon 2 facteurs qui sont la violence physique, la violence verbale, mais également selon 4 facteurs qui sont la violence pathologique, la violence exprimée contre les autres, la violence contre soi-même et la violence étouffée, et la factorisation en 3 facteurs peut être intermédiaire.

Quatrième partie

Bilan et perspectives

10

Mise en oeuvre, conclusions

10.1 Les implémentations des méthodes et de la démarche globale

Dans notre exposé, nous avons décrit trois méthodes indépendantes visant à rendre un jeu de règles d'association plus pertinent. Nous les avons implémentées et testées séparément sur des jeux de données divers. La première était à base de simulation de Monte-Carlo. Nous avons d'abord montré sur des données réelles que certaines règles d'association peuvent être dues au hasard. Notre souci étant de faire des rapports circonstanciés à toutes les étapes de cet effet du hasard, par graphiques et tableaux croisés, nous avons choisi de faire cette implémentation en Excel+VBA. Une fois établi le rôle du hasard, nous avons construit les simulations de façon rigoureuse afin qu'elles soient programmables de façon aisée en n'importe quel langage et nous les avons implémentées en Python afin de tester leur robustesse. La deuxième méthode décrite concerne les règles d'association floues. Nous les avons exprimées de façon mathématique en "fuzzifiant" toutes les étapes de la construction la plus basique parmi celles existant actuellement : recherche des ensembles fréquents, fermés, puis définition des règles d'association par découpage de chaque motif fréquent flou, avec le calcul d'une confiance. Le fait d'avoir ainsi défini les règles d'association faisait qu'une implémentation "légère" suffisait à valider cette méthode. Cela a été écrit en Python lors d'un stage d'étudiants de l'ESIAL, sur des données de géologie de l'INIST. La troisième méthode concerne le nettoyage par des méta-règles. La mise au point de cette méthode nécessitait, comme la première méthode, des bilans partiels clairs suivis d'un ajustement des paramètres, soit une certaine interactivité. Nous l'avons réalisée avec les données du challenge PKDD 2003 en Excel+VBA.

L'avantage de proposer des méthodes indépendantes est qu'on peut les valider séparément et les réutiliser comme composants dans d'autres techniques. Toutefois, il est légitime de se demander si on peut les unifier dans une démarche globale de fouille de données. Répondre à ce genre de question fait partie des enjeux actuels de la fouille de données. J.F. Boulicaut [29] notamment en a fait un de ses axes de recherche dans la cadre de l'"Action Spécifique Discochallenge "[154]. J'ai pu constater lors des réunions de ce groupe auxquelles j'ai eu la chance d'assister que la tâche n'est pas simple, et je n'aurais pas la prétention d'y répondre par cette thèse. Toutefois, de façon pragmatique, j'ai voulu me rendre compte si la démarche que j'avais utilisée, formée de trois méthodes établies en prolongement de travaux d'autres chercheurs, et validées sur des données réelles, pouvait également être validée globalement. Pour cela j'ai profité d'un défi international de classement proposé à l'occasion de la conférence KDD2004. Ce n'était pas un jeu de règles d'association qu'il fallait produire, mais un jeu de règles de classement, ce qui nécessitait de créer

[154] http ://users.info.unicaen.fr/ bruno/asdisco.

d'autres méthodes que les trois précédentes adaptées aux règles d'association. C'est ce qui a été fait, en suivant la même logique que celle qui a abouti à ces trois méthodes. Des algorithmes ont été rapidement développés, implémentés en Python, et les résultats de la tâche de classement sur les données fournies ont été envoyés aux organisateurs avant la date limite du 14 juillet 2004. Notre classement honorable [155] parmi les équipes en compétition est une validation partielle de la démarche globale de ce travail de thèse. Certes, un jeu de règles de classement est bien plus simple à concevoir qu'un jeu de règles d'association. Mais les techniques de classement sont enrichies chaque jour de nouvelles méthodes[156] alors que nos quelques algorithmes ont été conçus en deux mois de temps. Nous en concluons que notre classement honorable ne peut être dû qu'à la prise en compte des liaisons complexes entre propriétés, au sein d'une démarche globale cohérente. Cette démarche globale est une contribution à la recherche en fouille de données qui vient s'ajouter aux trois méthodes développées dans ce mémoire. Nous allons décrire sommairement la démarche suivie pour créer des règles de classement pour ce défi, puis proposer quelques pistes d'un prolongement possible de ce travail de thèse à travers l'algorithme MIDOVA avant de conclure.

10.2 KDDCup 2004 : tâche de discrimination entre 2 classes.

10.2.1 Description

Depuis plusieurs années, la conférence KDD propose un défi de fouille de données. Nous avons participé en 2003 [40] et nous avons pu apprécier l'organisation qui en est faite[157]. Les données et les questions sont disponibles fin avril, et l'évaluation se fait sur les fichiers de résultats qu'envoient les participants de façon anonyme 2 à 3 mois plus tard. Un Workshop est consacré à ce défi lors de la conférence, puis les gagnants exposent leurs méthodes. Les questions sont simples et sans ambiguïté, et les réponses demandées sont tout aussi simples. La simplicité des règles du jeu et la diffusion internationale de ce défi sont un atout important pour la recherche en fouille de données. Cela m'a notamment permis de créer un groupe "KDDCup 2004"[158] de 7 personnes comportant des enseignants, chercheurs, étudiants, et un professionnel en informatique dont la plupart étaient novices en fouille de données.

L'année 2004, il s'agissait d'une tâche de discrimination sur des données de physique quantique ou des données de biopuces[159]. Nous avons choisi les données de biopuces. Un tableau de données "bio-train" comportant environ 140 000 lignes et 77 colonnes était fourni, avec une colonne de valeurs 0 et 1, et un autre fichier "bio-test" de la même taille où la colonne de 0 et de 1 était remplacée par une colonne de '?'. Nous devions rendre la colonne de '?' après avoir remplacé chaque occurrence de ce symbole par la valeur 0 ou 1. Et nous avions à notre disposition les outils d'évaluation qui seraient utilisés pour évaluer notre résultat. Il s'agissait de 4 indices pour lesquels une définition et un petit logiciel de calcul nous était fourni.

[155] Les résultats sont indiqués dans la section 10.2.4.

[156] On peut trouver actuellement sur Internet à l'adresse http ://stat-www.berkeley.edu/users/breiman un petit logiciel de Leo Breiman et Adele Cutler utilisant des forêts aléatoires pour ces tâches de classement dont la dernier mise à jour date de mars 2004. Il semblerait que ce soit à ce jour la méthode de classification la plus prometteuse parmi celles développées par les chercheurs en classification.

[157] Merci encore aux organisateurs pour l'énorme travail produit.

[158] Je remercie Nicolas Baumgarten, Joseph di Martino, Renaud Lifchitz, Laurent Pierron, Joseph Rouyer, Tarek Ziadé pour leur participation à cette équipe.

[159] La tâche et les données de la KDDCup 2004 se trouvent à l'adresse http ://kodiak.cs.cornell.edu/kddcup

10.2.2 Démarche utilisée

Les données du tableau "bio-train", qu'on appellera T, comportaient 2 variables qualitatives d'identification (V1 et V2), la variable binaire de classe (V3) et 74 variables numériques (V4 à V77). Quelques examens statistiques rapides nous ont laissé supposer qu'il y avait des liaisons complexes entre la variable de classe et les 74 autres variables. Nous avons choisi de générer des règles du genre "si x appartient à un produit d'intervalles sur plusieurs variables, alors x est de classe 1", et pareillement pour la classe 0 selon les 3 étapes suivantes :

1. Pour éliminer les règles dues au hasard, nous avons procédé par apprentissage plutôt que par simulations. En effet, compte tenu le grand nombre de lignes dans le tableau T, il n'était pas gênant d'extraire les règles en se limitant à une partie de ces lignes (T1), et d'éliminer celles ne se vérifiant pas sur une autre partie (T2). Alors que des simulations de variables quantitatives auraient exigé une étude approfondie des lois de probabilités suivies par la distribution des valeurs dans le tableau.
2. Il convenait ensuite de coder chaque variable pour construire des règles d'association. Mais toutes les règles devant comporter à droite la variable de classe binaire, les règles d'association floues risquaient d'être inadaptées. En effet elles mettent en évidence une liaison en partie croissante entre A et B, et cette liaison perd une grande partie de son sens si une des deux variables est binaire. Nous avons choisi de ne garder qu'un intervalle de la variable, en remplaçant les valeurs dans cet intervalle par 1, et les autres par 0. Toutefois, cet intervalle n'a pas été défini à priori, mais engendré lors de la création de la règle[160]. De ce fait, d'une règle à l'autre, les intervalles d'une même variable diffèrent en général.
3. Il fallait ensuite rendre compte des liaisons complexes entre variables. Notre but n'était pas d'enchaîner des règles, puisque toutes avaient à droite la même variable, ni de les interpréter. La tâche était seulement une tâche de prédiction. La redondance ne gênait pas, le seul problème étant d'avoir deux règles qui prédisent deux classes différentes pour un même sujet de T2. Nous avons réglé en partie ce problème en deux temps : 1) On construisait chaque règle localement : par exemple, pour un sujet de la classe 1 de l'ensemble T1 on a construit des règles ayant la classe 1 à droite, de confiance 1 et de support supérieur ou égal à un seuil fixé à l'avance. 2) On évaluait ces règles sur T2, et on éliminait toutes celles qui ne donnaient pas assez d'éléments de la bonne classe, ou trop d'éléments de la mauvaise classe, selon des seuils fixés à l'avance, puis on leur attribuait une valeur de qualité fonction de leur performance (support, confiance), mais également de leur composition.

   Alors que les règles sur T1 ne pouvaient pas se contredire, elles le pouvaient sur T2, n'étant plus de confiance 1. Et si on supprimait ces incohérences en enlevant de règles, il y avait beaucoup moins d'éléments reconnus. Cela n'empêchait pas certains éléments de T2 d'être encore reconnus à tort, les règles ayant été crées sur T1. Comme ce qui comptait était la qualité de la prédiction selon les 4 indices fournis et non la cohérence du jeu de règles nous avons gardé cette source d'incohérences. Notons que la possibilité d'avoir différents intervalles pour une même variable augmente la complexité des liaisons entre plusieurs variables prises en compte dans le premier temps.

La démarche était alors terminée. Il ne restait plus qu'à l'évaluer. Pour cela, il suffisait de prédire avec cet ensemble de règles la classe d'une partie T3 des éléments de T, et de mesurer les performances de cette prédiction avec les quatre indices proposés sur T3. Cette prédiction s'obtenait en combinant pour chaque sujet les différentes valeurs de qualité des règles qu'il vérifiait

[160] Chaque intervalle est obtenu par une succession d'ajustements de ses bornes à partir de leurs valeurs initiales.

en une valeur indiquant la probabilité estimée qu'il soit de classe 1. Puis nous avons amélioré les performances en modifiant les paramètres de la construction des règles. Et dans le peu de temps restant avant l'échéance, nous avons créé de nouvelles règles sur T2 et T3 ajustées sur T1 afin d'augmenter les performances, avant de prédire les classes des éléments du tableau "bio-test". Malheureusement nous n'avons pas eu le temps de recommencer le partage aléatoire de T en trois parties plusieurs fois, comme cela se fait habituellement en apprentissage[161]. L'évaluation de nos prédictions sur le fichier "bio-test" par les organisateurs du défi sur les 4 indices s'est avérée à peine inférieure à celle que nous avions obtenue sur le fichier "bio-train".

10.2.3 Un exemple de règle extraite

La règle $(var53 \leq -129)$ *et* $(var9 \leq 943.4)$ *et* $(var4 \geq 89.55) \rightarrow classe = 1$ a 162 éléments de T qui vérifient sa partie gauche, tous sauf 1 sont de classe 1, ce qui fait un support de 161 et une confiance de 0,99.

On peut voir dans les trois graphiques de la figure 10.1 les associations entre deux des trois variables var4, var9 et var53 et la classe (Var3). Les croix plus claires indiquent les éléments de classe 1, dessinés après ceux de classe 0, ce qui explique qu'on arrive à les voir (il y en a moins de 2000 de classe 1 et plus de 140 000 de classe 0). Dans chaque graphique, on a représenté les droites d'équations $var53 = -129$, $var9 = 943.4$ et $var4 = 89.55$. Bien que le quart de plan représentant les projections dans chaque plan de l'ensemble $(var53 \leq -129)$ *et* $(var9 \leq 943.4)$ *et* $(var4 \geq 89.55)$ paraisse plus chargé en points de classe 1 que les autres quarts, il aurait fallu que les trois règles associées soient de meilleure qualité pour être créées lors du processus d'extraction. En effet pour la règle $(var53 \leq -129)$ *et* $(var9 \leq 943.4) \rightarrow classe = 1$ on a 230 éléments de T qui vérifient sa partie gauche, dont 179 sont de classe 1, pour la règle $(var53 \leq -129)$ *et* $(var4 \geq 89.55) \rightarrow classe = 1$ on a 1631 éléments de T qui vérifient sa partie gauche, dont 284 sont de classe 1, et pour la règle $(var9 \leq 943.4)$ *et* $(var4 \geq 89.55) \rightarrow classe = 1$, on a 1430 éléments de T vérifiant sa partie gauche, dont 399 sont de classe 1. Nous ne décrirons pas ici les algorithmes ayant permis de créer ces règles de classement, car ils ne sont pas directement transposables aux règles d'association qui sont notre sujet premier.

10.2.4 Résultats

Les résultats des 59 participants pour les "données bio" sont indiqués dans les graphiques de la figure 2, avec une croix plus grosse pour chacune des deux techniques proposées par des personnes du groupe "KDDCup. Pour chaque indice, la technique que nous venons de décrire, qui est à base de règles d'association, donne de meilleurs résultats que l'autre[162] à base de Kmeans avec une distance de Mahalanobis, alors que cette dernière disposait d'une partie apprentissage plus robuste. Pour le Top1[163], elle a permis de reconnaître le premier élément de type 1 de 130 souches sur 150 (0.86667, le premier ayant 0.92000, soit 138 sur 150, et le dernier 0.02, soit 3 sur 150), pour le RMSE, le score a été de 0.04499 contre 0.03501 pour le premier et 0.99076 pour le dernier ; pour le RKL le score a été de 59.74 contre 45.62 pour le premier et 854.46 pour le dernier, et pour le APR, 0.79728 contre 0.84118 pour le premier et 0.01453 pour le dernier. Comme on peut le voir, nos résultats ne sont pas excessivement éloignés de ceux des premiers. Ce qui nous fait penser que cette technique qui prend en compte les liaisons complexes entre

[161] Ces trois parties sont T1, l'ensemble d'apprentissage, T2, l'ensemble de mise au point, T3, l'ensemble de test

[162] Les graphiques sont orientés de telle sorte que le premier est toujours à droite et le dernier à gauche.

[163] Le détail de tous ces indices est à l'adresse http ://kodiak.cs.cornell.edu/kddcup/metrics.html

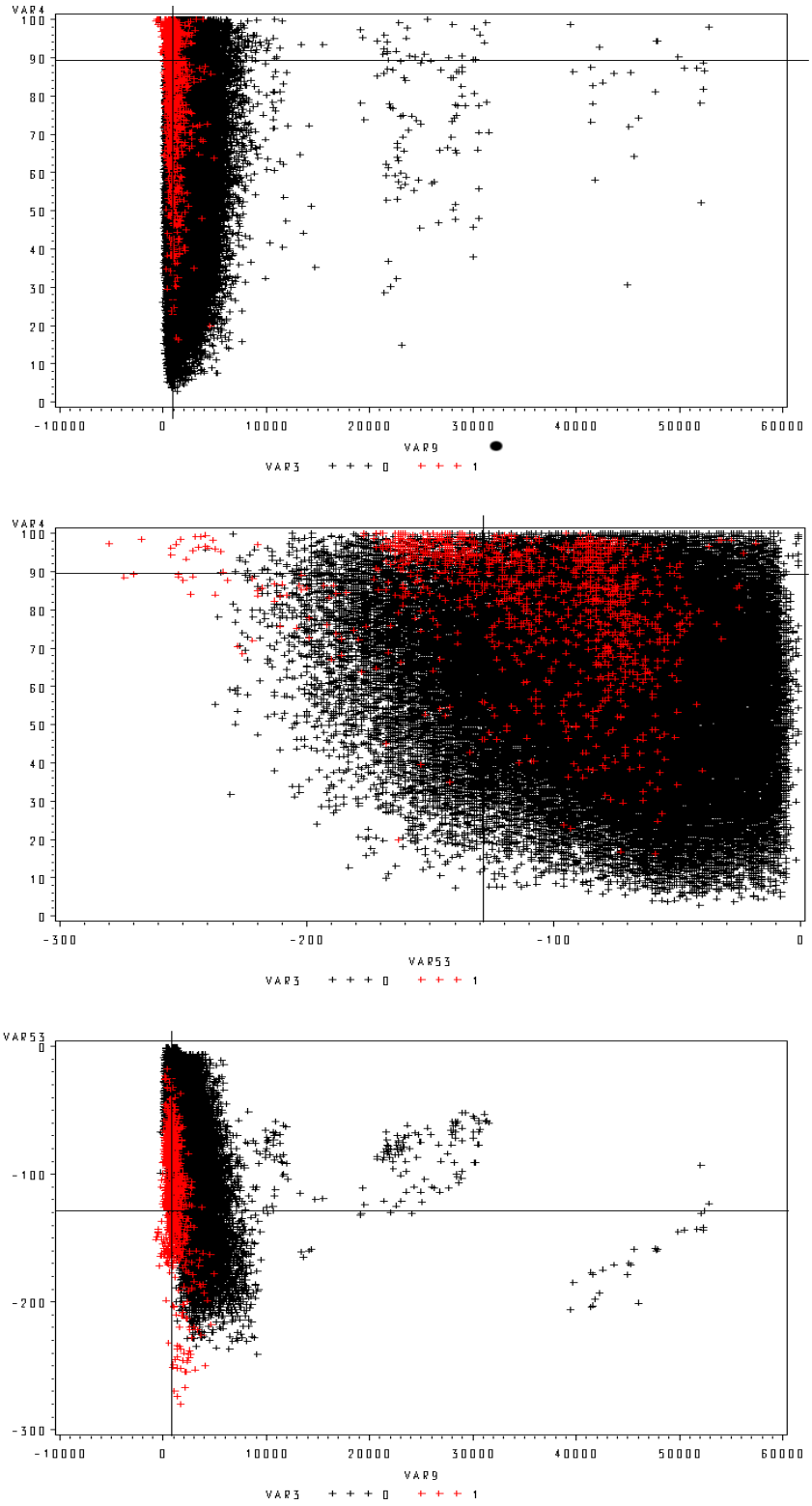


FIG. 10.1 – Les variables V4, V9 et V53.

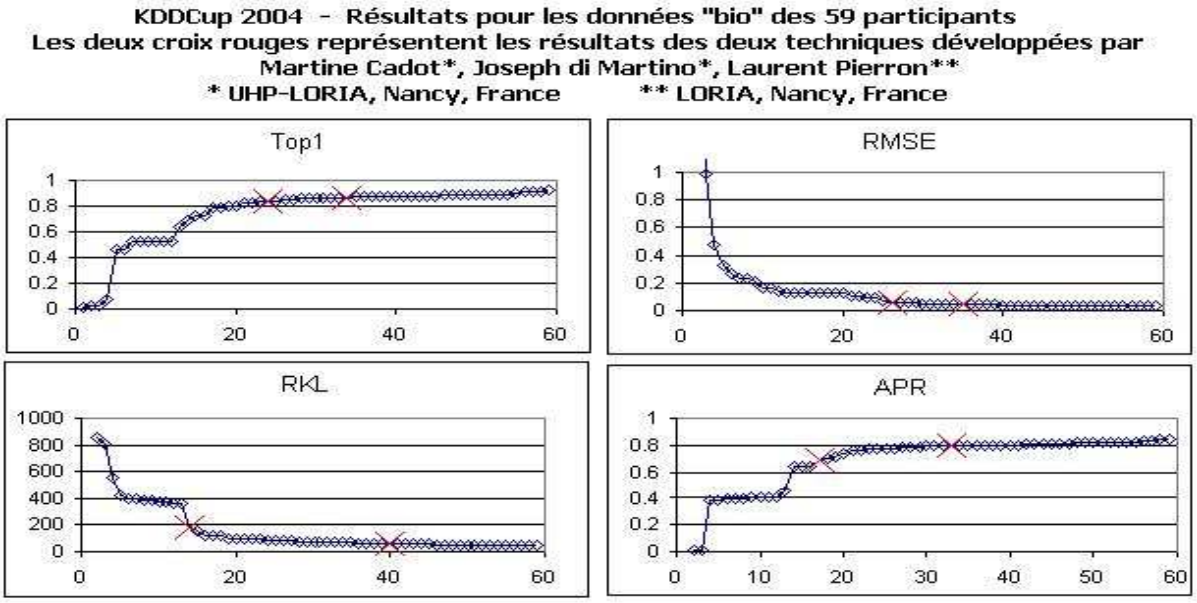


FIG. 10.2 – Les résultats de l'équipe "KDDCup2004".

propriétés pourrait être intégrée dans des logiciels de classement si elle corrige un certain nombre de défauts. Même si les algorithmes ne peuvent être optimisés dans le sens de la rapidité, elle reste pour nous la preuve que la prise en compte des liaisons complexes entre attributs est indispensable dans l'induction "orientée- attributs" dont fait partie l'extration de règles d'association.

10.3 MIDOVA : l'algorithme et ses indicateurs, son rôle dans une chaîne globale de validation et réduction d'information

10.3.1 Les liaisons entre variables exprimées par les motifs

Nous exposons dans cette sous-section la position du problème, en reprenant sommairement des éléments abordés dans différentes parties de ce document.

L'extraction habituelle de motifs

Si on dispose d'une base de données pouvant se représenter par une matrice booléenne "objets X variables", par exemple "textes X mots-clés" ou "sujets X propriétés", la conjonction de variables, appelée motif, permet d'exprimer des liaisons entres celles-ci. L'étude de ces liaisons passe essentiellement par le comptage de leurs cooccurrences, appelé support du motif. Si on dispose de p variables, les motifs peuvent être formés de k variables (on les appellera motifs de longueur k, ou k-motifs), k variant entre 0 et p. Le nombre de ces motifs est de l'ordre de 2^p, ce qui rend leur interprétation impossible. La stratégie la plus courante pour limiter cette explosion combinatoire consiste à choisir un seuil de support, et à n'extraire que les motifs dont le support dépasse ce seuil. Pour les interpréter en terme de liaisons entre variables, on extrait de ces motifs des règles d'association, qu'on range selon leurs valeurs à divers indices de qualité (DKQ2004, DKQ2005), afin de se limiter à celles de meilleure qualité selon la sémantique couverte par ces indices.

Ses inconvénients, auxquels nous souhaitons échapper

Le choix préalable d'un seuil de support a des inconvénients gênants pour notre but. Il fait disparaître les oppositions entre variables, quels que soient leurs supports respectifs (le support du motif les liant étant faible, voire nul dans le cas de variables exclusives l'une de l'autre). Il fait également disparaître les associations positives entre variables rares, sa valeur étant fixée indépendamment du support des variables constituant les motifs. A cela s'ajoutent des inconvénients dus à l'extraction proprement dite, qui produit des motifs pour lesquels aucun souvenir des détails de leur composition n'est conservé, compromettant ainsi toute interprétation fine postérieure, et dus à la redondance en cas de variables avec des valeurs très proches (si A, B, C et D sont recouvrantes, les motifs AB, AC, AD, BD seront extraits ainsi que ABC, ABD, ACD, BCD et ABCD). Nous désirons une extraction de motifs sans ces inconvénients, mais fournissant un nombre raisonnable de motifs. Pour cela, nous allons transposer à l'extraction de motifs la méthode statistique utilisée dans le modèle linéaire vu plus haut dans notre mémoire, car elle permet une analyse des liaisons entre variables plus proche de nos besoins, mais qui ne peut être appliquée à nos types de données.

10.3.2 Définition du gain

Le principe . Pour mesurer le gain d'information d'un motif, nous nous appuyons sur les variations possibles du support de ce motif. On impose à ces variations de se faire en laissant les supports des sous-motifs de M inchangés. Ainsi ce gain mesure ce que l'association de toutes les propriétés le composant apporte de plus que l'ensemble des diverses associations d'une partie de ces propriétés.

Recherche de l'intervalle de variation. Pour obtenir que le support de M augmente d'une unité, à supports constants des sous-motifs, on choisit un sujet qui vérifie toutes les propriétés sauf une, et on lui rajoute cette propriété. Cela a pour conséquence que le support de chaque sous-motif de M contenant cette propriété est également augmenté de une unité. On compense cette augmentation en faisant de nouveaux changements élémentaires, qui vont également devoir être compensés. Si ce processus peut se réaliser, il s'arrête nécessairement au bout de 2^{L-1} changements, L étant la longueur du motif M, c'est-à-dire son nombre de propriétés. Et le support de M peut augmenter d'autant d'unités que de réitérations possibles de ce processus. Pour le faire diminuer, on procède de même en faisant les changements inverses de sujets. On crée ainsi l'intervalle de variation du support de M. L'exemple développé plus loin permettra de concrétiser ce principe.

Choix de la valeur du gain. L'intervalle obtenu a un centre à partir duquel le support des motifs peut augmenter ou diminuer. Nous décidons que le gain d'information correspondant aux motifs de support central est nul, et cela reste valable si l'intervalle est réduit à une seule valeur. Puis plus le support du motif s'éloigne de ce centre en se rapprochant des bornes de l'intervalle, plus la valeur absolue du gain augmente. Selon que le support est à droite du centre ou à gauche, le gain est positif ou négatif. Nous sommes conscient que cette décision de placer le centre au milieu de l'intervalle de variation n'est justifiée qu'en première approximation, tant qu'aucun autre élément, théorique ou empirique, ne nous invite à en décider autrement - le point important est qu'une forte association se traduit par une forte valeur positive du gain, une forte "répulsion" par une forte valeur négative. Nous mesurons la valeur de ce gain en nombre de sujets

dont on doit changer la valeur d'une propriété pour obtenir ce motif en partant d'un motif de support central.

L'indice MIDOVA-g. Pour mesurer le gain d'information g d'une règle, nous calculons le support s du motif M sur lequel elle s'appuie, la longueur L de ce motif (le nombre de propriétés le constituant), et le centre c de l'intervalle qui peut être décrit par le support de M sans que changent ses sous-motifs. Pour la valeur de ce gain, les considérations détaillées ci-après nous amènent à choisir la fonction $g = 2^{L-1}(s - c)$.

L'indice MIDOVA-r. Une autre caractéristique essentielle dans notre optique est le "reste", défini comme 2^{L-1} fois la différence entre le support s et la borne (inférieure ou supérieure) la plus proche de s. Sa valeur, indicatrice du "potentiel de variation" du support au niveau supérieur commande la poursuite de l'algorithme ou son arrêt.

10.3.3 Le principe des algorithmes d'extraction de motifs par niveau

Un motif de longueur k, ou k-motif, est constitué de k variables binaires et peut être considéré comme une nouvelle variable obtenue par fusion des k variables en multipliant leurs valeurs pour chaque objet. Le nombre de ces valeurs est le support du motif. On peut également considérer un motif M de longueur k comme issu de l'adjonction d'une variable à un motif M' de longueur k-1, ce dernier étant un de ses sous-motifs (c'est-à-dire constitué d'une partie des variables du motif). Si ce sous-motif M' a un support s, c'est qu'il a s valeurs égales à 1, et le résultat de son produit avec une autre variable formée de 0 et de 1 qu'est le motif M ne peut contenir plus de uns. Ainsi le support du motif ne peut pas augmenter quand on lui adjoint de nouvelles variables. Au contraire, il a tendance à diminuer jusqu'à s'annuler, sauf dans les cas exceptionnels où des objets ont des valeurs de un pour toutes les variables. Les algorithmes par niveau, comme A-Priori, ont pour but d'extraire les motifs dont le support est supérieur à un seuil, donc au minimum non nuls. Si la plupart des objets vérifient une quantité faible de variables, il converge très vite. C'est le cas des données du type "panier de la ménagère", où le nombre d'articles (ce sont les variables) en vente est considérable, mais chaque "panier" (ce sont les objets) en contient un nombre restreint. En cas de seuil de support 1, le motif le plus long correspond au panier le plus rempli (en terme de variété d'articles, pas de quantité) et son support est 1 en général (une seul panier de ce type), mais si le seuil de support s est supérieur à 1, le motif le plus long correspond au plus grand nombre d'articles communs à au moins s paniers, et sa longueur a tendance à diminuer quand le seuil s de support augmente. En choisissant ces algorithmes par niveau, on privilégie les plus longs motifs et/ou les plus grands supports. Ils ne peuvent être utilisés pour extraire les motifs de supports extrêmes (petits et grands)

10.3.4 Le principe de fonctionnement de notre algorithme MIDOVA par niveau

Considérons les propriétés binaires A, B, C et D, définies sur un ensemble d'objets, et voyons comment on peut passer de A à AB, ABC et ABCD :

1. Si le motif A a pour support s, on attend pour le support AB un support compris entre 0 et s. Trois cas peuvent alors se produire :
   - le support de AB est proche de 0, il est intéressant à interpréter en terme d'opposition entre A et B

- le support de AB est proche de s, il est intéressant à interpréter en terme d'attirance entre A et B
- le support de AB est proche de $s/2$, il n'est pas intéressant à interpréter

Plaçons-nous dans le cas où le motif AB a un support proche de s, par exemple $s - r$, r étant un "petit" nombre positif ou nul, qu'on pourrait déclarer négligeable.

2. Si le motif AC a un support de s', on attend pour le motif ABC un support compris entre $s' - r$ et s', disons $s' - r/2$. Même si le support de ABC s'éloigne au maximum de sa valeur attendue "neutre" $s' - r/2$, cet éloignement est de $r/2$, donc négligeable. Ce qui rend ABC inintéressant à interpréter. Plaçons-nous dans le cas où le motif ABC a un support de $s' - r'$, avec $r' < r/2$.
3. Si le motif BCD a un support de s", on attend pour le motif ABCD un support compris entre $s" - r'$ et $s"$, disons $s" - r'/2$. En reprenant le raisonnement au point 2), on constate que ABCD est encore moins intéressant que ABC.

Inutile de continuer à diviser des quantités négligeables par 2, aucun motif "intéressant" de longueur supérieure à 2 ne peut contenir AB. Le motif AB doit être éliminé avant la construction des motifs de longueur 3. On ferait le même raisonnement en se plaçant au point 1) dans le cas où AB a son support proche de zéro. Dans ces deux cas, AB a consommé quasiment toute la part de variation du support que lui avait laissé A. Cela le rend intéressant à interpréter, mais il n'a laissé aucune part de variation du support pour ses sur-motifs ABC et ABCD.

Exemple :

Prenons le cas de 60 sujets pour lesquels nous connaissons les valeurs de 3 propriétés A, B et C. Les supports respectifs de A, B, C, AB, AC, BC, ABC sont 35, 28, 40, 20, 27, 22 et 15. Les valeurs des 60 sujets pour les 3 propriétés sont représentées dans la figure 10.3 par un tableau d'incidence et par un diagramme de Venn. Comme il y a 3 propriétés, le tableau contient $2^3 = 8$

sujets	A	B	C	eff
s1, s2, s3, s4, s5, s6, s7, s8, s9, s10, s11,s12, s13, s14, s15	x	x	x	15
s16, s17, s18, s19, s20	x	x		5
s21, s22, s23, s24, s25, s26, s27, s28, s29, s30, s31, s32	x		x	12
s33, s34, s35, s36, s37, s38, s39		x	x	7
s40 s41, s42	x			3
s43		x		1
s44, s45, s46, s47, s48, s49			x	6
s50, s51, s52, s53, s54, s55, s56, s57, s58, s59, s60				11

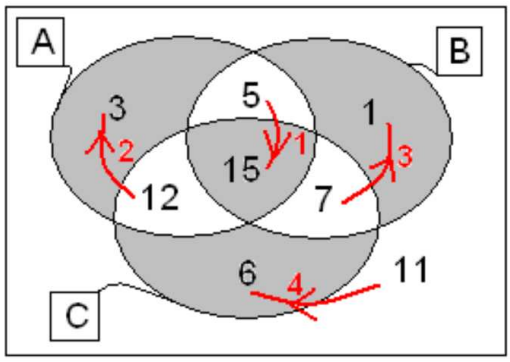


FIG. 10.3 – Répartition de 60 sujets selon 3 propriétés.

lignes, et le diagramme de Venn 8 zones. La zone où les trois propriétés sont simultanément vérifiées est grisée, ainsi que les zones dont le nombre de propriétés est également impair, donc ici vérifiant une seule propriété. Les quatre zones restantes (blanches) sont celles où un nombre pair de propriétés (0 ou 2) sont vérifiées. Pour chercher l'intervalle de variation du support de ABC à support constant de ses sous-motifs, on essaie d'abord d'augmenter son support. En suivant la flèche 1, on déplace un sujet (par exemple $s16$) en lui ajoutant la propriété C qu'il

n'avait pas. Il passe ainsi d'une zone non grisée à une zone grisée. Lors de ce déplacement, le support de AB ne change pas. Par contre ceux de AC et de BC augmentent d'une unité chacun. On compense cette augmentation en suivant les flèches 2 et 3, qui déplacent par exemple les sujets $s21$ et $s33$ en leur retirant la propriété C. Ce déplacement a pour conséquence une diminution du support de C d'une unité, qu'on compense en déplaçant par exemple le sujet $s50$ selon la flèche 4. Ainsi, si on désire augmenter le support de ABC, qui est dans une zone grisée, sans modifier les supports de ses sous-motifs, il faut augmenter d'autant les effectifs des 3 autres zones grisées et diminuer d'autant chaque effectif d'une zone non grisée. Comme le plus petit effectif des zones non grisées est 5, le support de ABC ne peut pas augmenter de plus de 5. Et pour faire diminuer le support de ABC, on procède de façon inverse, ce qui fait qu'il ne peut pas diminuer de plus de 1, minimum des effectifs des zones grisées. Le support de M varie ainsi entre 14 et 20, sa valeur centrale étant 17. Il ne reste plus qu'à calculer g en remplaçant le support s par 15, la longueur L par 3, et le centre c par 17, ce qui donne $g = 2^{L-1}(s - c) = 4(15 - 17)$, soit -8, ce qui veut dire qu'il faut déplacer 8 sujets pour faire passer le support de ABC de 17 à 15 sans changer les supports de ses sous-motifs. En résumé :

- Borne inf = s - min(zones d'arité impaire)
- Borne sup = s + min(zones d'arité paire)
- c = (Borne inf + Borne sup)/2

10.3.5 Propriétés du gain d'un motif

La condition préalable au calcul du gain de motifs est qu'ils soient construits avec des propriétés dont on connaît les valeurs pour chacun des N sujets d'un ensemble donné. Bien qu'on utilise par la suite le gain pour des règles à prémisse composée, donc construites sur des motifs de longueur au moins égale à 3, le principe de calcul du gain s'étend sans problème à des motifs de longueur 2 et la formule à des motifs de longueur 1 (notons toutefois que dans ce dernier cas $g = s - N/2$, ce qui fait qu'on peut obtenir un gain avec des moitiés de sujets).

- prop 1 : Le gain d'un motif M ne peut pas dépasser $N/2$ en valeur absolue.
- prop 2 : le gain d'un motif de longueur L est un nombre entier de fois 2^{L-2}
- prop 3 : Si a est l'amplitude de l'intervalle de variation du gain g d'un motif M de longueur L, l'intervalle de variation du gain de ses sur-motifs de longueur $L + 1$ a une amplitude inférieure ou égale à $a - 2|g|$. La valeur de a pour les motifs de longueur 1 est N, l'intervalle étant [-N/2 ;N/2].

Ces trois propriétés permettent de limiter le coût machine de la recherche du gain d'un motif. Avec les propriétés 1 et 2, dès que sa longueur est telle que $2^(L - 2)$ dépasse $N/2$ (soit $L > 1 + log(N)/log(2)$), le gain est nul. Avec la propriété 3, chaque fois que le gain d'un motif de longueur L est différent de 0, cela réduit l'intervalle de variation des motifs qui le contiennent. Ainsi, au fur et à mesure que la longueur du motif augmente par ajout de propriétés, ses possibilités de variation diminuent ou restent constantes, ce qui limite sa valeur possible de gain. Cet effet est accentué par la propriété 2. Cela est en adéquation avec le fait que dans le cas le plus courant, une fois que l'information essentielle est apportée par quelques propriétés, au fur et à mesure qu'on ajoute de nouvelles propriétés, l'information supplémentaire qui en résulte est de plus en plus petite.

Illustration des propriétés sur notre exemple

Nous mettons dans le tableau ci-dessous les sous-motifs de longueur 2 de ABC et le motif ABC lui-même. Quand le motif M=XY a pour longueur 2, les 2 propriétés X et Y déterminent

4 parties dans le diagramme de Venn, qui sont XY lui-même quand les deux propriétés sont vérifiées et nonXnonY quand aucune n'est vérifiée, que nous intitulons de parité 1 (leur nombre de propriétés est pair comme celui de M), et XnonY et YnonX que nous intitulons de parité -1 car elles ont un nombre de propriétés impair, contrairement à M. Pour le motif M=ABC, les parties de parité 1 sont les parties grisées (avec 1 ou 3 propriétés, donc de la même parité que M), et les autres les parties non grisées. De façon générale, à tout motif de longueur $L > 1$ on peut ainsi faire correspondre 2^{L-2} parties de parité 1 (comme M), et autant de parité -1. Une fois choisies ces parties, le calcul suit les étapes du tableau 1.

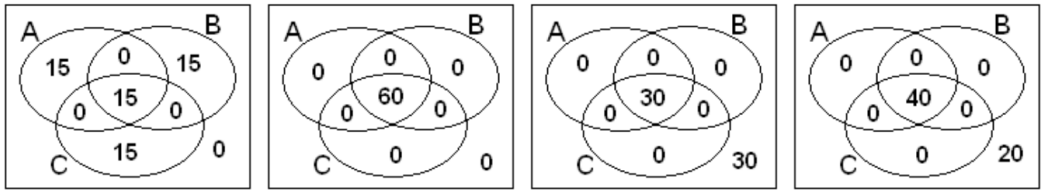
Motif M	M=AB	M=AC	M=BC	M=ABC
Longueur de M (nombre de propriétés)	L=2	L=2	L=2	L=3
Support de M (nombre de sujets)	s=20	s=27	s=22	s=15
Effectif des parties de parité 1 (comme M)	20 ; 17	27 ; 12	22 ;14	15 ; 1 ; 3 ; 6
Effectif des parties de parité -1 (contraire de M)	8 ;15	13 ; 8	18 ; 6	5 ; 7 ; 12 ; 11
m1 : minimum des effectifs de parité 1	17	12	14	1
m2 : minimum des effectifs de parité -1	8	8	6	5
Intervalle du support ($b1 = s - m1$, $b2 = s + m2$)	(3 ;28)	(15 ;35)	(8 ;28)	(14 ;20)
Amplitude de l'intervalle du support : b	25	20	20	6
Centre c de l'intervalle du support	15.5	25	18	17
Gain $g = 2^{L-1}(s - c)$	9	14	8	-8
Amplitude de l'intervalle du gain $a = ^{L-1}$b	50	40	40	24
Amplitudes restantes : - de l'intervalle du support, $\min(b2 - s, s - b1)$ - du gain $a - 2\|g\|$ ("MIDOVA-r")	 8 32	 8 32	 6 24	 1 8

TAB. 10.1 – Etapes de calcul du gain de ABC et de ses sous-motifs de longueur 2

Pour les motifs de longueur 1, le calcul se fait de façon similaire une fois choisi l'intervalle de variation du support. Comme N=60, on prend pour intervalle de variation de leur support [0 ;60], (0 pour la propriété fausse pour tous les sujets, et 60 pour celle vraie pour tous) de centre 30, et l'intervalle de gain est [-30 ;30]. Les intervalles de support et de gain ont ainsi même amplitude 60. Ces motifs A, B, C, ont pour supports respectifs 35, 28 et 40 et pour gains leurs écarts à 30, soit 5, -2 et 10. Les supports de A, de B et de C étant chacun différents du centre de leur intervalle de variation, cela laisse pour leurs sur-motifs respectifs des intervalles de variation d'amplitude inférieure à 60/2, qui sont respectivement de 25, 28 et 20. L'amplitude restante de leurs intervalles de gain respectifs est 50, 56 et 40. Cela a pour conséquence, pour AB par exemple, qui est un sur-motif de A et de B, que l'intervalle de variation de son support a une amplitude qui ne peut dépasser ni 25 ni 28, donc leur minimum 25, et pour le gain 50, minimum de 50 et 56. On voit dans cet exemple, que chaque fois qu'on passe d'un motif à un sur-motif, l'intervalle de variation du gain diminue bien. Il était au départ de 60, pour A, B, C, puis il est passé à 50 pour AB, 40 pour AC et 40 pour BC, puis à 24 pour ABC. Si on rajoute une propriété D, il reste pour le motif ABCD un intervalle de gain d'amplitude maximum 8, ce qui peut donner -4, 0 ou 4 comme gain. Et si on rajoute une cinquième propriété, on obtiendra nécessairement un gain nul, qui le restera quelles que soient les propriétés ajoutées à nouveau.

Comportement particulier du gain

On a vu que dans le cas le plus courant illustré par l'exemple précédent, l'intervalle de variation du gain d'un motif a tendance à diminuer petit à petit quand le motif s'agrandit. Toutefois, d'après la propriété 3, si le gain d'un motif est N/2 en valeur absolue, les gains de tous les autres sur-motifs ou sous-motifs de ce motif sont nuls. Cela est également vrai si la somme des valeurs absolues des gains de motifs emboîtés atteint N/2. Voici deux cas pour lesquels ce phénomène se produit, avec pour conséquence que l'amplitude de l'intervalle de variation du gain tombe d'un seul coup à 0 quand le motif s'agrandit d'une propriété. Prenons un motif M formé de L propriétés, pour un ensemble de N sujets où N est un multiple de $2^(L-1$, et répartissons ces sujets de façon égale dans les seules parties ayant le nombre de propriétés de même parité que le motif M. Le gain est alors égal à N/2, et d'après la propriété 3, tous les sous-motifs et sur-motifs éventuels de M ont un gain nul. On peut aussi mettre les sujets dans les seules parties de parité -1, et on obtient alors des gains opposés. Ce cas correspond à un effet particulier de l'association des L propriétés, qui ne se produit que quand elles sont toutes réunies. Ce qui fait par exemple que 4 propriétés qui ne sont pas liées 2 à 2, ni 3 à 3 peuvent être liées malgré tout. Prenons maintenant un motif M contenant tous les N sujets. Son gain est nul, ainsi que celui de tous ses sous-motifs de longueur >1. Et ses sous-motifs de longueur 1 ont un gain de N/2. Ce cas correspond à des propriétés toujours vraies, donc identiques. Si on se contente de les prendre toutes identiques, mais avec la moitié des sujets pour lesquels elles sont vraies, et l'autre pour lesquels elles sont fausses, tous les motifs de longueur 2 ont un gain de N/2, et les autres sous-motifs sont de gain nul. Si on augmente de a le nombre de sujets qui vérifient toutes les propriétés (en diminuant d'autant ceux qui n'en vérifient pas) on obtient un gain de a pour chaque motif de longueur 1, de N/2-a pour ceux de longueur 2, et de 0 pour les autres. Voici (fig. 10.4) les diagrammes associés à ces cas quand le motif considéré est M=ABC, de longueur 3, avec N=60. Le gain maximal de 30 est soit celui de ABC, soit celui des sous-motifs de ABC de longueur 1, soit de ceux de longueur 2, soit partagé entre une suite croissante de sous-motifs emboîtés comme c'est le cas du dernier graphique. Dans tous ces cas, tout sur-motif de M est de gain nul. Ces cas d'école expriment des liens particuliers qu'on peut rencontrer de façon moins accentuée dans les données et que le gain permet de repérer.



g(A)=g(B)=g(C)=0	g(A)=g(B)=g(C)=30	g(A)=g(B)=g(C)=0	g(A)=g(B)=g(C)=10
g(AB)=g(AC)=g(BC)=0	g(AB)=g(AC)=g(BC)=0	g(AB)=g(AC)=g(BC)=30	g(AB)=g(AC)=g(BC)=20
g(ABC)=30	g(ABC)=0	g(ABC)=0	g(ABC)=0
g(M)=0 si M > ABC	g(M)=0 si M > ABC	g(M)=0 si M > ABC	g(M)=0 si M > ABC

FIG. 10.4 – Valeurs remarquables du gain pour des cas particuliers

Il est à noter que le premier cas de la figure 10.4 n'est autre que la différence symétrique, généralisation du XOR à plus de 2 propriétés. Dans ce cas, alors que les motifs A, B, C, AB, AC, BC semblaient dus au hasard, c'est à l'étape 3 qu'apparaît la liaison remarquable entre A, B, et C, liaison qui échappe aux algorithmes à seuils de support, de type A-Priori, quand le nombre de propriétés est grand.

10.3.6 L'algorithme MIDOVA

- 0) Initialisation : on choisit une valeur négligeable e pour les écarts du support, dont l'unité est l'objet, donc en nombre d'objets. Cette valeur peut être 0, 1, 2 ou plus.
- 1) Au niveau 1 (motifs réduits à une variable), les motifs ont des supports qui peuvent varier de 0 à N (nombre d'objets sur lesquels les variables sont définies). On s'attend donc[164] à une valeur de $N/2$. On calcule leur écart au centre $N/2$ de l'intervalle, et la part de variabilité qu'ils laissent à leurs sur-motifs, qui est le support s lui-même si $s > N/2$, ou $N - s$ dans le cas contraire. Les motifs pour lesquels elle est négligeable (inférieure ou égale à e), ce qui est le cas des motifs ayant un support s proche de zéro ($s <= e$), ou proche de N ($s >= N - e$), ont épuisé leur part de variabilité. On les élimine des motifs à fusionner pour le niveau suivant.
- 3) Niveau k (tant qu'il reste des motifs) On combine les motifs du niveau précédent qui sont combinables en un motif de niveau k, et on en déduit son support et l'intervalle de variation de celui-ci (sg ; sd). On calcule son écart au centre de l'intervalle, et la part de variabilité qu'il laisse à ses sur-motifs (c'est le reste MIDOVA-r : l'écart entre sa valeur et la borne la plus proche sg ou sd). Si elle est négligeable, il a épuisé sa part de variabilité, on l'élimine des motifs à combiner à l'étape suivante.

Lorsque l'algorithme a convergé (ce qui se produit d'autant plus rapidement que e est grand), on interprète les motifs obtenus en terme de gain : positif si interaction positive, négatif dans le cas contraire d'exclusion entre la présence des variables.

10.3.7 Application

Les données

Les données traduisent la présence de 888 mots dans les résumés des 193 premiers livres de la collection Gallimard-jeunesse, qui forment une encyclopédie touchant des sujets très variés ; Ces présences/absences peuvent être représentées par une matrice booléenne Docs X Mots : si la valeur correspondant au document i et au mot j est égale à 1, cela signifie que le résumé du document i contient le mot j, et elle est égale à 0 dans le cas contraire. Le nombre de uns de la matrice est de 6559.

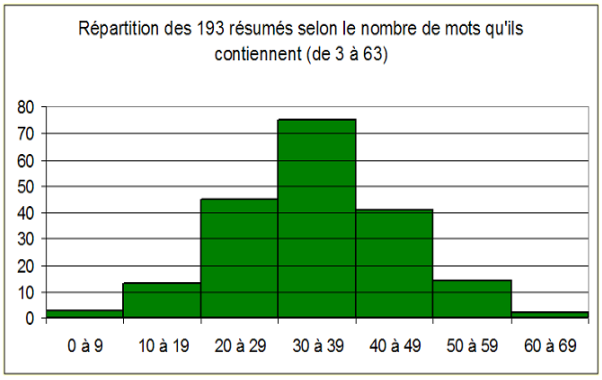


FIG. 10.5 – Répartition des mots et des résumés (1/2)

Le nombre de mots par document varie entre 3 et 63, la répartition des documents selon

[164] Sous la réserve exprimée plus haut en matière de "centre" de l'intervalle de variation du gain.

leur nombre de mots suivant une distribution approximativement binomiale, avec beaucoup de documents ayant entre 30 et 40 mots (cf. figure 10.5). Le nombre de documents par mot varie entre 1 et 62, la répartition des mots selon leur fréquence se faisant selon une distribution inégalitaire, plus de 90% des mots figurant dans moins de 15 textes chacun (cf. figure 10.6).

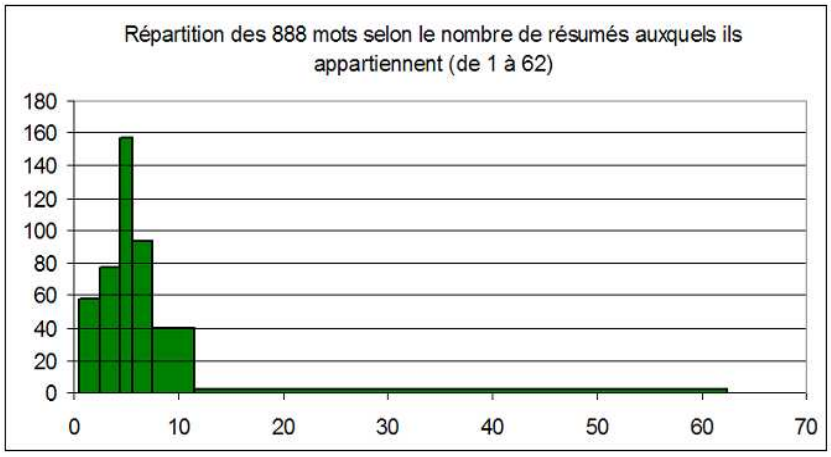


FIG. 10.6 – Répartition des mots et des résumés (2/2

Le test de significativité des 2-motifs

Nous nous sommes d'abord intéressés aux associations de deux mots. Sur la base de 888 mots, il y en a 888*887/2 soit 393 828. La seule information pertinente que nous avons retenue sur ces associations de deux mots est le nombre de documents dans lesquels ils apparaissent simultanément. Ce nombre peut varier entre 0 (ils n'apparaissent jamais ensemble dans un même texte) et le minimum des nombres de textes où chacun apparaît, qui est supérieur à 0, tous les mots apparaissant dans au moins un texte.

Nous avons décidé de ne garder que les associations de deux mots (2-motifs) figurant dans plus de textes qu'attendu par hasard, et celles dans moins de textes qu'attendu par hasard, en prenant un risque alpha inférieur ou égal à 5% (risque de se tromper en estimant qu'une association n'est pas due au hasard) dans cette décision. Ainsi c'est un test bilatéral que nous faisons, permettant d'établir un intervalle de confiance à 95% des valeurs du support en cas d'absence de liaison, les 2,5% à gauche de cet intervalle représentant les support trop petits pour être dus au hasard et les 2,5% à droite ceux trop grands pour être dus au hasard. Pour le réaliser nous générons au hasard et de façon indépendante 200 matrices booléennes ayant mêmes sommes marginales que la matrice de données. Et pour chaque association de deux mots, nous cherchons dans chaque matrice simulée combien de fois elle apparaît dans des documents. Nous disposons ainsi du support réel du motif de longueur 2 et de la liste de ses supports dans les matrices simulées.

Il y a peu de 2-motifs moins fréquents qu'attendus, mais bien davantage de plus fréquents qu'attendu. Par exemple le 2-motif famille, rêve a un support de 0 dans les données d'origine, ce qui signifie que ces 2 mots ne sont jamais dans un même texte, alors que dans 95% des matrices simulées, il apparaît avec un support compris entre 1 et 7. Ce qui indique une opposition significative entre ces deux mots dans notre corpus. De même, le 2-motif *peintre, ville* a un support de 2 dans les données d'origine, ce qui est peu par rapport au support de chacun (respectivement

de 26 et 44) comme l'indiquent les données simulées qui font apparaître un intervalle à 95% de confiance de (3, 10) : dans cette collection encyclopédique, ces deux thèmes fréquents et distincts ont peu de recouvrement. Par contre le 2-motif *François 1er, Charles Quint* est peu fréquent (support de 3), mais plus fréquent qu'attendu (intervalle de confiance de (0,1)), ce qui s'explique par la rareté de chacun de ces deux mots (supports respectifs de 5 et 4).

Après la phase de test statistique d'échanges en cascade, il reste 4 000 motifs de longueur 2 significatifs avec un risque a=5%, de support s avec 0<=s<=46.

A	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	34	37	38	44	46	51	53	57	62	B
A	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	0	0	0	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2
2	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	2	1	2	2	2	2	2	2	2	2	2	2	2	2	2	3
3	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	4
4	0	0	0	0	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	3	2	2	3	3	3	3	4	4	4	6	5
5	0	0	0	0	0	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3	4	4	4	4	4	7	7	6
6	0	0	0	0	0	0	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3	3	4	3	4	4	4	5	4	5	8
7	0	0	0	0	0	0	0	1	2	2	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	4	4	4	4	5	5	5	5	9
8	0	0	0	0	0	0	0	0	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4	5	5	5	5	5	6	6	10
9	0	0	0	0	0	0	0	0	0	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4	5	5	6	6	6	6	11	11
10	0	0	0	0	0	0	0	0	0	0	2	2	2	3	3	3	3	3	3	3	3	3	4	4	3	4	4	4	4	4	5	5	5	5	5	6	6	6	6	7	12	12
11	0	0	0	0	0	0	0	0	0	0	0	2	3	3	3	3	3	3	3	4	4	4	3	4	4	4	4	4	5	5	5	5	5	5	6	6	6	6	7	7	13	13
12	0	0	0	0	0	0	0	0	0	0	0	0	3	3	3	3	3	3	4	4	4	4	4	4	4	4	5	5	5	5	5	5	6	6	7	7	7	8	14	14	14	14
13	0	0	0	0	0	0	0	0	0	0	0	0	0	3	3	3	3	4	4	4	4	4	4	4	5	5	5	5	5	6	6	6	7	7	8	8	8	15	15	15	15	15
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	4	4	4	4	4	4	4	4	5	5	5	5	5	6	6	6	7	7	8	8	8	9	16	16	16	16	16
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	4	4	5	5	4	5	5	5	5	6	6	6	6	6	6	8	8	7	9	9	9	9	17	17	17	17
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	4	4	4	5	4	5	5	5	5	5	6	5	6	6	6	7	7	8	8	8	8	10	18	18	18
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	4	5	5	5	5	5	5	6	5	6	6	6	6	6	7	7	8	8	9	9	10	19	19	19
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	5	5	5	5	6	5	6	6	6	6	6	7	7	8	7	9	9	9	10	10	11	20	20
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	5	5	5	6	6	6	7	7	7	7	7	8	8	8	8	9	9	10	10	11	21	21
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	5	6	6	6	6	6	7	7	7	7	7	8	8	8	9	9	10	10	11	12	22
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	6	6	6	7	7	7	7	7	8	8	8	10	9	11	10	11	12	23	23	23
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	6	6	6	7	7	8	7	8	8	9	9	9	10	10	11	11	13	24	24
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	7	7	7	7	7	8	8	9	9	9	11	10	11	11	12	13	25	25	25
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	7	0	7	7	8	8	8	8	8	9	9	10	10	12	12	12	13	26	26	26
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	7	0	7	8	8	8	8	9	9	9	11	11	12	11	13	14	27	27	27
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	0	8	8	9	9	10	9	10	11	12	14	13	13	14	28	28	28	28
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	0	8	9	8	8	9	9	10	11	12	12	13	14	14	29	29	29
28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	9	9	10	11	11	12	12	13	13	14	15	30	30	30	30
29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	2	1	1	0	9	9	9	10	10	12	12	13	15	14	16	31	31
30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	0	9	10	11	11	12	12	13	14	14	16	32	32	32
31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1	1	1	1	2	1	2	1	2	2	2	0	0	11	12	13	13	14	16	15	16	34	34	34
32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1	1	1	1	2	2	2	2	2	2	3	0	0	0	13	13	14	15	16	17	18	37	37	37
34	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	1	1	2	1	2	1	2	2	2	2	2	2	0	0	0	0	14	15	15	15	16	17	38	38	38
37	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	2	2	1	2	2	2	2	3	2	3	3	3	0	0	0	16	18	18	19	20	44	44	44
38	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	1	1	1	2	2	2	2	2	2	2	2	2	3	3	3	0	0	0	18	19	20	20	46	46	46
44	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	2	2	2	2	2	2	2	3	3	2	3	3	4	4	4	5	5	0	0	20	21	23	51	51	51
46	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	2	2	2	2	2	3	2	3	3	3	3	3	4	4	4	5	6	6	0	0	22	24	53	53	53
51	0	0	0	0	0	0	0	0	1	1	1	1	1	1	2	2	2	2	2	2	3	3	3	3	4	4	4	4	5	5	5	6	6	5	7	8	0	0	23	24	57	57
53	0	0	0	0	0	0	0	1	1	1	1	1	1	2	2	2	1	2	2	2	3	3	3	4	4	4	4	4	4	5	5	5	5	6	6	7	9	9	0	0	0	0
57	0	0	0	0	0	0	0	1	1	1	1	1	2	2	2	2	3	3	3	3	3	4	4	4	4	5	4	5	5	6	6	6	7	7	8	9	11	11	12	0	0	0
62	0	0	0	0	0	0	0	1	1	1	1	1	2	2	2	2	2	3	3	3	3	4	4	5	5	5	5	5	6	6	6	7	7	8	10	10	12	12	14	0	0	0
B	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	34	37	38	44	46	51	53	57	62	B

FIG. 10.7 – Tables des seuils de confiance du support au risque $\alpha = 5\%$

Un sous-produit de cette approche est la possibilité de créer des tables statistiques caractéristiques de ce corpus et non dédiées à un type donné de répartition, comme les tables statistiques universelles. Ces tables spécifiques permettent de déterminer pour chaque couple de mots (A,B) la borne inférieure (partie triangulaire inférieure) et la borne supérieure (l'autre partie triangulaire) de l'intervalle de confiance à 95% (tab. 10.2) et à 90% (tab. 10.3) connaissant leurs supports respectifs. Cette approche permettrait, dans le cas de corpus importants, d'établir par simple consultation l'intervalle de confiance du support de toutes les associations de mots alors

que la table pourrait n'être établie qu'à partir d'un simple échantillon représentatif des mots de ce corpus.

FIG. 10.8 – Tables des seuils de confiance du support au risque $\alpha = 10\%$

En bas de la table des seuils de confiance du support au risque $\alpha = 5\%$, le seuil gauche et en haut le seuil droit. Par exemple, pour trouver l'intervalle de confiance du motif AB avec A de support 20 et B de support 50 (en fait on inverse les supports de A et B pour la lecture dans la table), on lit 2 dans la partie triangulaire en bas et 9 dans celle du haut. Ainsi l'intervalle est de (2, 9) , et si on trouve une valeur de 0 ou de 1 pour ce support dans les données réelles, on en déduit que le motif AB indique une opposition "inattendue" entre les variables A et B connaissant le support de chacune, alors que s'il dépasse 9, on en déduit une attraction inattendue. Il est alors gardé comme motif intéressant. S'il est compris entre 2 et 9, il est considéré comme dû au hasard et rejeté. Dans la table à 10%, l'intervalle est (2 ; 8,5), donc plus petit, ce qui fait que plus de motifs sont rejetés car attribués au hasard par le seuil à 5% que par celui à 10%.

Construction des motifs d'ordre supérieur valides avec l'algorithme Midova

Parmi les 4000 motifs de longueur 2, 3686 ont une valeur de Mr (reste selon Midova) supérieure à 1. Ils se combinent en 2276 motifs de longueur 3, dont 587 de Mr supérieur à 1, ces derniers créant 41 motifs de longueur 4, dont 2 de Mr supérieur à 1, trop peu nombreux pour produire des motifs de longueur supérieure. Ces motifs peuvent s'interpréter selon leur valeur de gain Mg. Voici quelques exemples d'interprétation.

Le 4-motif *archéologie, fouille, légende, site* a un indice Mr de 0, ce qui indique qu'il ne peut plus contribuer à un 5-motif. Son indice Mg est de 4, ce qui indique une liaison positive entre ces 4 mots plus informative que la liaison entre ces mots pris 3 à 3 et 2 à 2. Parmi ceux-ci, le 2-motif *fouille, légende* a un indice Mg de -2 qui indique une faible liaison négative. De même le 3-motif *pouvoir, puissance, Jérusalem* a un indice Mr de 0 et un indice Mg de 8 , et les 2 motifs *cinéma, film* a un indice Mr de 4 et un indice Mg de 9. Les plus fortes valeurs de Mg apparaissent pour les associations entre mots composés et leurs composants, par exemple *XXe siècle* et *XXe*, ou *chef* et *chef d'oeuvre*, *guerre mondiale* et *seconde guerre mondiale*. L'opposition maximale a lieu entre *Etats-Unis* et *Moyen-Age* avec Mg=-24.

Efficacité de la chaîne de traitement proposée

L'action combinée du test statistique et de Midova a réduit l'explosion de la pyramide des motifs en largeur (2-motifs) par un facteur 144 et en hauteur par un facteur d'au moins 2.

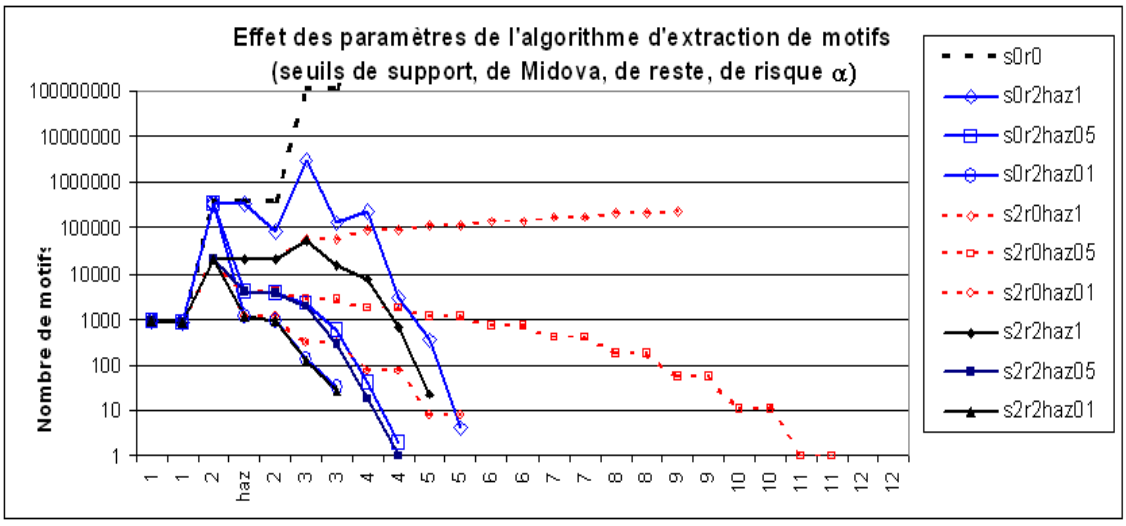


FIG. 10.9 – Comparaison des efficacités respectives de l'utilisation combinée de Midova et du test d'échanges en cascade avec la méthode classique de type A Priori, avec seuil de support

Dans la figure 10.9, on trouve en abscisse les longueurs des motifs en en ordonnée leur nombre. Chaque longueur est présente deux fois, la première pour les motifs extraits, la seconde pour ceux dont le reste Mr est non nul. La courbe pointillée est le résultat de l'extraction de toutes les associations de 2 mots, que ceux-ci figurent ou non dans un même texte. Elle indique clairement le caractère exponentiel de cette démarche naïve. Les 3 courbes rouges montrent le résultat de l'approche A Priori (le seuil de support est de 2), combinée ou non avec un filtrage par le test d'échange en cascade (aux deux seuils de risque 5% et 1%). Les 3 courbes bleues représentent

les résultats de l'approche Midova (Mr>=2) combinée ou non avec le test. Les 3 courbes noires combinent l'approche A-Priori et Midova, mais ne permettent pas d'obtenir les oppositions. Il en ressort clairement que notre approche permet d'obtenir les oppositions partielles ou totales que ne ressortent pas des autres méthodes, ainsi qu'une condensation non redondante des liaisons entre variables. Notre approche fournit une quantité de motifs bien inférieure et de moindre complexité, valides statistiquement, et d'interprétation plus riche.

Et ces principes ne sont pas limités aux données binaires. Nous avons ainsi commencé à définir un gain pour les règles d'association floue qui prolonge celui que nous venons de définir pour les RA classiques (Cuxac et al. 2005).

Il reste encore beaucoup de travail pour affiner les fondements de cette méthode, notamment au niveau de la définition du gain. Il en reste aussi pour aider à l'interprétation de ses résultats, particulièrement pour les motifs de gain extrême négatif, et pour les règles d'association construites sur ces motifs.

10.4 Conclusions

Nous avons essayé de défendre dans ce mémoire la thèse selon laquelle la masse de données accessibles depuis l'avènement d'Internet pouvait être exploitée par les chercheurs en sciences humaines au moyen de la méthodologie de fouille de données. Cette branche de l'informatique est consacrée à la construction de connaissances à partir des données. Les deux étapes importantes de la démarche de fouille de données sont le codage des données et leur traitement. Alors que les principes de codage de la démarche informatique ont été directement repris des sciences humaines et n'ont évolué depuis que dans le sens d'une certaine systématisation, le traitement informatique des données n'a pas pu bénéficier de ce transfert total depuis les sciences humaines. En revanche de nouvelles méthodes sont apparues ne reposant pas sur la seule modélisation statistique, mais sur des modélisations issues d'autres domaines scientifiques, dont l'informatique. Ainsi, parmi les méthodes de traitement de données des sciences humaines qui n'ont pu être transposées à la fouille de données figurent essentiellement celles qui permettent aux chercheurs en sciences humaines d'établir des preuves à partir d'observations sur les données, notamment la plupart des techniques de statistiques inférentielles, A l'opposé, parmi les nouvelles méthodes apparues, l'extraction de motifs et de règles d'association est celle qui nous paraît la plus apte à fournir de façon automatique une représentation symbolique des données pouvant rivaliser avec la modélisation statistique classique. Sa seule exigence de fonctionnement, dans sa forme la plus courante, porte sur le format des données qui doivent être exprimées par un tableau booléen objets×variables. Ses points forts sont sa capacité à explorer tant localement que globalement l'intégralité du réseau des liaisons entre variables, et le symbolisme de sa représentation par un ensemble de motifs et un jeu de règles d'association. Cette méthode a gagné une place privilégiée dans les domaines d'économie appliquée que sont par exemple la grande distribution et le milieu des assurances, toujours à la recherche de nouvelles pépites de connaissance, dont l'exploitation peut leur permettre de se démarquer de leurs concurrents, davantage que d'une modélisation de leur domaine plus générale donc plus partagée. Par rapport aux méthodes statistiques classiques, qui se concentrent sur une vision globale des liaisons entre variables, cette méthode informatique d'extraction peut paraître moins puissante aux chercheurs en sciences humaines désireux de brosser un tableau synthétique de la réalité à travers un modèle habituel liant quelques variables par des équations linéaires. La puissance explicative liée à une telle modélisation simplificatrice de la réalité par les statistiques se fait au prix d'hypothèses fortes faites sur les données, tant au niveau de chaque variable qu'au niveau de leurs liaisons. L'extraction informatique des liaisons se fait au contraire sans

autre simplification préalable que le codage adapté des données, ce qui produit un ensemble de motifs et un jeu de règles d'association reproduisant plus fidèlement l'enchevêtrement des liaisons existant dans les données. Cette richesse a un coût qui est la difficulté d'appréhension d'une telle représentation des données par des chercheurs en sciences humaines habitués à modéliser la réalité à l'aide d'un ensemble réduit de relations "de causes à effets", construit sur leurs données et habitués à justifier leur modélisation auprès de leur communauté.

Consciente de ces difficultés, depuis une dizaine d'années la communauté de fouille de données a fait principalement porter ses efforts sur l'aide à l'interprétation du jeu de règles d'association. Cela a produit notamment une cinquantaine d'indices mesurant la qualité d'une règle. Outre le fait que s'ajoute le choix délicat de l'indice (ou des indices, et dans ce cas de leur utilisation conjointe) aux difficultés d'utilisation du jeu de règles d'association extrait des données, cela ne résout pas le problème de la "preuve" de la valeur des liaisons trouvées, qui était auparavant assurée par les statistiques inférentielles. Notre contribution en ce sens, exposée dans la partie II de ce mémoire est la construction d'un test de randomisation par échanges en cascade permettant d'assurer la validité des relations trouvées. Il fait partie des tests d'hypothèses des statistiques inférentielles, mais sans certaines caractéristiques gênantes de ceux-ci : ce n'est pas un test nécessitant des hypothèses sur les données, ni un test d'hypothèse simple permettant d'assurer la validité d'une règle particulière entre deux variables prises indépendamment du contexte. Pour éviter ces écueils, nous avons justifié dans le chapitre 5 que les simulations devaient se faire en créant des tableaux booléens de données de même taille que le tableau d'origine et ayant les mêmes marges que ce dernier. Et dans le chapitre 6 nous avons construit une méthode rigoureuse permettant de le faire, à partir d'échanges en cascade, et démontré que cela revenait à faire se succéder des échanges rectangulaires sur le tableau d'origine selon certaines règles. Avec ce test, seuls les motifs significatifs sont gardés et c'est sur ceux-ci que sont construites les règles d'association. Les chercheurs en sciences humaines ont ainsi l'assurance que les liaisons repérées entre variables sont "significatives" selon le sens habituel statistique. Notons au passage que ce test peut être utilisé pour extraire toute liaison entre variables d'une base de données, que cette liaison soit "positive" ou "négative". Nous avons utilisé ce test ici pour extraire des liaisons "positives", c'est-à-dire exprimées par de "trop fortes" cooccurrences entre variables pour qu'elles puissent être attribuées au hasard. Nous aurions pu également extraire des liaisons "négatives" c'est-à-dire exprimées par de "trop faibles" cooccurrences entre variables pour qu'elles puissent être attribuées au hasard, ou même des liaisons "extrêmes" entre les variables, c'est-à-dire sans préjuger de leur positivité ou négativité. En effet ce test revient au calcul d'un nombre s réel permettant de définir un intervalle de confiance du support d'un motif de la forme (-∞, s). On peut le modifier en remplaçant le calcul de s par celui d'un nombre réel s' associé à un intervalle de confiance de la forme (s', ∞) ou même le modifier par le calcul de deux nombres réels s' et s associés à un intervalle de confiance (s', s). Dans les deux premiers cas le test est appelé *unilatère*, et dans le dernier cas *bilatère*.

Le deuxième besoin des chercheurs en sciences humaines concerne la cohérence du jeu de règles extraites. Quelques travaux de recherche ont été menés pour répondre à ce besoin, initiés d'un côté par Régis Gras [99] en didactique, et de l'autre, plus récemment par Jiawei Han [113] dans le domaine des bases de données. Bien que répondant à deux logiques différentes, ces travaux ont en commun le postulat qu'on peut définir a priori une méthode de nettoyage du jeu de règles extrait afin qu'il soit une représentation cohérente des données, et corriger ainsi "l'illogisme" de la procédure d'extraction. Cette incohérence, à notre avis, n'est pas une faiblesse de la procédure d'extraction, mais participe à la richesse de la représentation qu'elle fait des données. En effet les règles d'association expriment une grande partie voire la totalité des liaisons complexes entre

variables. Et les chercheurs en sciences humaines connaissent bien la difficulté de prise en compte simultanée des liaisons qui se situent à des niveaux de description différents, comme le sont les liaisons globales et locales, l'interaction et les effets principaux, etc. Le choix que nous avons fait, exposé dans la partie III de ce mémoire, est de spécifier divers types d'incohérences et de proposer pour chacun une méta-règle de nettoyage paramétrable par l'utilisateur en fonction du niveau de description qu'il choisit, et en lui laissant la trace des décisions prises au fur et à mesure par la méta-règle (les sous-ensembles de règles incohérents, et les règles éliminées ainsi que celles conservées). Il nous reste à réaliser une interface qui permette à l'utilisateur de piloter le nettoyage du jeu de règles afin d'en faire un jeu cohérent pour son usage spécifique, voire plusieurs jeux correspondant à des usages différents.

La troisième contribution de notre travail de thèse à l'amélioration de l'extraction des motifs et des règles d'association figurant également dans la partie III de ce mémoire a pour but d'empêcher un effet malencontreux du codage binaire des données numériques qui fait que certains types de liaisons ne peuvent plus apparaître dans les motifs extraits et donc dans les règles d'association. En "fuzzifiant" toutes les étapes du processus d'extraction, nous pouvons l'appliquer directement aux données numériques et préserver ainsi ces liaisons. L'idée de créer des règles d'association floues n'est pas nouvelle, mais notre façon de le faire permet de garder la structure de treillis nécessaire à certains algorithmes d'extraction, tout en produisant des règles d'association correspondant à des règles définies par ailleurs par les chercheurs poursuivant les travaux dans la direction initiée par Zadeh.

La mise en oeuvre de ces différentes contributions n'est pas terminée. Cependant, diverses réalisations ont été décrites dans la dernière partie de ce document afin de montrer au lecteur que l'extraction des relations complexes est réalisable selon les principes exposés dans ce document, et qu'elle a un intêret certain pour les chercheurs en sciences humaines.

Annexes

A

Treillis des motifs flous, compléments du chapitre 9

Lors de l'utilisation de la méthode d'extraction de motifs et de règles d'association, l'utilisateur est généralement invité à fixer un seuil de support, les seuls motifs dépassant ce seuil étant extraits, et les règles d'association extraites sont ainsi valables pour un nombre de sujets dépassant un minimum (égal à ce seuil), ce qui leur confère une certaine généralité. Toutefois l'utilisateur qui est plutôt à la recherche de "pépites de connaissances" se désintéresse des motifs à support élevé, qu'il suppose correspondre à une connaissance des données qu'il a déjà. Dans cette annexe, nous essayons de définir rigoureusement[165] cette élimination de motifs flous "trop" ou "trop peu" fréquents à partir de seuils. L'illustration de ces principes figure dans la partie 9.4.1 où on a représenté un treillis flou avant et après élimination des motifs "extrêmes"..

Dans la figure 1, on a représenté le treillis obtenu à partir des trois propriétés du tableau 9.3 (chapitre 9), notées a1,b1 et c1 et de leurs négations notées a0, b0 et c0, en apportant quelques modifications. Il manque en effet tous les motifs fermés correspondant à des motifs de support nul, c'est-à-dire n'ayant aucune valeur supérieure à 0,5, qui ont été identifiés au motif fermé a1b1c1a0b0c0. Comme on l'a dit précédemment, ce choix a été fait afin de ne pas laisser dans ce treillis des ensembles de propriétés contenant deux propriétés contraires, du type {a1, a0}, ou {a1,a0,b1}, etc... mais également pour éliminer des ensembles de propriétés remettant en cause les inclusions, comme {a1, c0} qui remet en cause l'inclusion de a1' dans c1', donc de a' dans c', inclusion qui est illustrée dans la figure 2 du chapitre 9 par l'absence du motif {a}, comme dans la figure 1 par l'absence du motif {a1}[166]. L'extraction de motifs classiques se limite généralement à ceux dont le support dépasse un seuil (voir [222], [52], [64], [223], [225], [57]) afin de produire des treillis des motifs fermés moins importants et/ou des algorithmes de recherche de motifs et de règles plus rapides. Ce seuil est choisi le plus souvent de façon arbitraire, les motifs de petits supports étant jugés de moindre intérêt. Nous désirons rendre plus souple ce choix de seuil et élargir son champ d'action aux motifs de support trop élevés [167]. Nous définissons deux seuils

[165] Nous désirons notamment préserver la structure de treillis flou. Cette façon de procéder peut être restreinte aux motifs classiques.

[166] Toutefois, on peut constater sur le treillis représenté en figure 1 que ce choix ne suffit pas à assurer un jeu de règles cohérent avec toutes les propriétés des règles d'implication de la logique "du sens courant". Par exemple, la règle b1c1→a1 de confiance 1, qui se lit dans le diagramme, devrait donner les règles de "démonstration par l'absurde" que sont b1a0→c0 et c1a0→b0, comme indiqué dans [106]. Ce n'est pas le cas pour la première, comme l'indique la présence du motif a0b1 dans le diagramme, alors qu'il aurait dû être confondu avec le motif a0b1c0.

[167] Ces motifs peuvent correspondre à des redondances dans les données, voire à des *tautologies* (propriétés vraies pour tous les sujets), et peuvent alors être éliminées par certains indices de qualité des règles comme c'est le cas de

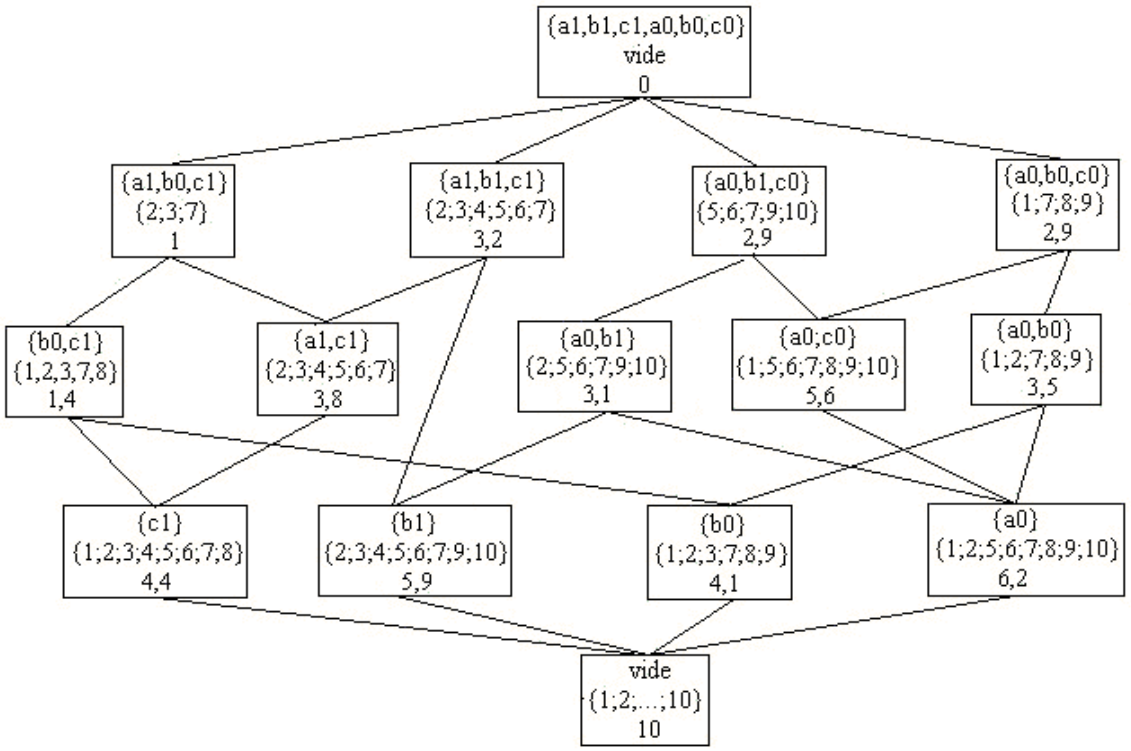


FIG. A.1 – Le treillis flou des motifs associés aux 3 propriétés du tableau 9.3, et à leurs négations.

γ et δ, l'un permettant d'éliminer les motifs "trop faiblement vérifiés" et l'autre permettant d'éliminer les motifs "trop fortement vérifiés". Plutôt que les éliminer, nous préférons les rendre équivalents, ceci afin de préserver, autant que possible, la structure de treillis.

Définition A.0.1. *(γ, δ)-type d'un motif flou.*
Si γ et δ sont deux nombres entiers petits devant le nombre N total de sujets de $\mathcal{S}$, le (γ, δ)-type d'un motif flou A représentant un ensemble de propriétés de $\mathcal{P}$ sur $\mathcal{S}$ est :

$$type_{\gamma\delta}(A) = \begin{cases} -1 & si\ card(\{s \in \mu_{\mathcal{A}}(s) > 0{,}5\}) \leq \gamma \\ 1 & si\ card(\{s \in \mu_{\mathcal{A}}(s) \leq 0{,}5\}) \leq \delta \\ 0 & sinon \end{cases}$$

Les motifs de type -1 sont les motifs "faiblement" vérifiés, les motifs de type 1 sont les motifs "fortement" vérifiés, et les motifs de type nul sont les motifs vérifiés de façon "satisfaisante", soit ni trop, ni trop peu.
Si on choisit $\gamma<0$, il n'y a pas de motif de type -1 , et si on choisit $\delta<0$, il n'y a pas de motif de type 1. Avec $\gamma=0$, un motif de type -1 sera un motif qui ne possède aucune valeur supérieure à 0,5. Parmi les motifs de ce type on trouve tous ceux qui contiennent 2 attributs contraires (cela correspond au choix que nous avons fait quand nous avons proposé la formule du support). Si on prend $\delta=0$, un motif de type 1 sera un motif qui possède au plus 0 (donc aucune) valeur inférieure ou égale à 0,5. C'est le cas du motif $\emptyset$ pour lequel toutes les valeurs sont égales à 1. Les motifs de type 0 pour ces deux valeurs $\delta=\gamma=0$ sont les motifs dont au moins une valeur est

notre indice MIDOVA, actuellement en fin de développement. Ils peuvent aussi correspondre à des "évidences" et dans ce cas le travail de nettoyage est souvent fait par l'expert quand il ne cherche que de nouvelles connaissances.

inférieure à 0,5 et dont au moins une valeur est supérieure ou égale à 0,5. Plus on augmente ces deux seuils, moins il y a de motifs de type 0, et plus il y en a de types -1 et 1.
Si on reprend les motifs du tableau 9.3, on voit qu'avec les 2 seuils à 0, le motif vide est de type 1, et les autres de type 0. Et il faudrait atteindre un seuil γ de 2 pour obtenir des motifs de type -1. Ce seraient les motifs ab, bc, abc.

Définition A.0.2. *(γ, δ)-égalité entre des motifs flous.*
On dit que deux motifs flous A et B formés de propriétés de $\mathcal{P}$ définies sur un ensemble $\mathcal{S}$ de N sujets sont (γ, δ)-égaux dans un des trois cas suivants :

$$A\ eq_{\gamma\delta}\ B\ si \begin{cases} ou\ bien\ type_{\gamma\delta}(A) = type_{\gamma\delta}(B) = -1 \\ ou\ bien\ type_{\gamma\delta}(A) = type_{\gamma\delta}(B) = 1 \\ ou\ bien\ type_{\gamma\delta}(A) = type_{\gamma\delta}(B) = 0\ et\ \forall s \in \mathcal{S},\ \mu_{\mathcal{A}}(s) = \mu_{\mathcal{B}}(s) \end{cases}$$

Propriété A.0.1. *La relation (γ, δ)-égalité associée à une relation floue sur $\mathcal{S} \times \mathcal{P}$ est une relation d'équivalence sur les motifs flous formés des propriétés de $\mathcal{P}$.*

La preuve découle immédiatement de la définition de la relation.

Notation 1. *La classe d'équivalence d'un motif a par la relation de (γ, δ)-égalité, c'est-à-dire l'ensemble des motifs qui sont en relation avec le motif a est notée $cl_{\gamma,\delta}(a)$. Les motifs de type 1 sont tous dans la même classe, qui est celle du motif vide . On la note $cl^1_{\gamma,\delta}$. Les motifs de type -1, s'il en existe, sont tous dans la même classe qui est notée $cl^{-1}_{\gamma,\delta}$*

Alors qu'il y a 0 ou 1 classe pour les motifs de type -1, et 1 classe pour les motifs de type 1, les motifs de type 0 forment en général de nombreuses classes, deux motifs de type 0 étant dans la même classe s'ils ont les mêmes valeurs pour tous les sujets, c'est-à-dire s'ils ont la même extension. On a ainsi pour le tableau 9.3, avec les deux seuils à 1 : aucune classe de type -1, une classe de type 1, qui est {$\emptyset$}, et les 4 classes de type 0 {a, ac}, {b}, {c}, {ab, bc, abc}. A chacune de ces classes correspond un motif fermé de la figure 9.2.
Maintenant que nous avons établi la cohérence de cette classification avec les treillis de motifs fermés habituels, et avant d'étendre formellement les opérations sur les motifs à ces classes de motifs, nous contrôlons ce que deviennent les négations des propriétés, utilisées dans la figure 1, quand on remplace les propriétés par leurs classes.

Propriété A.0.2. *(γ, δ)-égalité et négation.*
Si a et b sont deux propriétés floues sur un ensemble $\mathcal{S}$, de négations respectives $\overline{a}$ et $\overline{b}$[168] *et si on choisit pour γ et δ la même valeur m, on a l'équivalence suivante*

$$\{a\} eq_{\gamma\delta} \{b\} \Leftrightarrow \{\overline{a}\} eq_{\gamma\delta} \{\overline{b}\}$$

Preuve. *Si le motif a est de type -1, c'est que le nombre de sujets ayant une valeur de asupérieure à 0,5 est inférieur ou égal à γ, qui est m. Pour ces sujets, les valeurs pour $\overline{a}$ sont inférieures ou égales à 0,5, étant les compléments à 1 de leurs valeurs pour a. Ainsi le nombre de sujets ayant une valeur inférieure ou égale à 0,5 pour $\overline{a}$ est inférieur ou égal à m. Comme on a choisi δ=m, $\overline{a}$ est de type 1. Avec un raisonnement similaire, on obtient que si a est de type 1, $\overline{a}$ est de type -1, et par différence que si a est de type 0, $\overline{a}$ est de type 0.*

1. *Supposons maintenant que a est (γ, δ)-égal à b. Cela signifie que a et b sont de même type, et que si c'est le type 0, ils ont la même valeur pour chaque sujet. Si a est de type 1, d'après ce que nous venons de voir, $\overline{a}$ est de type -1. Comme b dans ce cas est de même type que*

[168] On pourrait bien sûr utiliser des négations de motifs flous, mais nous n'en avons pas l'usage habituellement pour la recherche de règles d'association.

a, il est de type 1, et $\overline{b}$ est de type -1. On voit donc que $\overline{a}$ et $\overline{b}$ sont de même type. On montrerait de même que si a est de type -1, $\overline{a}$ et $\overline{b}$ sont de type -1, et par différence que si a est de type 0, ils sont tous de type 0.
Quand $\overline{a}$ et $\overline{b}$ sont de type non nul, ils sont (γ, δ)-égaux, et quand ils sont de type 0, comme les valeurs pour a et b sont les mêmes pour tous les sujets, leurs compléments à 1 également, et donc les valeurs de $\overline{a}$ et $\overline{b}$ aussi.
On a donc montré que $\overline{a}$ et $\overline{b}$ sont de même type, et que si c'est le type 0, ils ont la même valeur pour chaque sujet. Il sont donc bien (γ, δ)-égaux. L'implication dans le sens direct est ainsi démontrée.

2. *Pour montrer la réciproque, il suffit de noter que $\overline{\overline{a}}$=a, et d'utiliser l'implication démontrée en remplaçant a par $\overline{a}$ et b par $\overline{b}$.*

Si on choisit des seuils γ et δ différents, on n'a plus cette relation de compatibilité entre la relation d'équivalence et la négation. En effet, on peut avoir par exemple une propriété moyennement vérifiée alors que sa négation est faiblement ou au contraire fortement vérifiée. Dans les articles précédemment cités [222, 52, 64, 223, 225, 57] la valeur de γ peut varier mais la valeur de δ est toujours nulle. Il faut dire que les négations sont rarement prises en compte dans les articles sur la fouille de données. Et, à notre connaissance, même quand les négations sont considérées ([222, 99, 107]) les relations particulières qui en découlent sont constatées dans les conclusions, mais ne servent pas à optimiser la construction du treillis des motifs fermés et/ou l'optimisation des algorithmes de recherche de motif. Si on désire prendre en compte dans le treillis des motifs fermés une relation particulière comme par exemple celle existant entre une propriété et sa négation, il faudra vérifier si cela nécessite ou non la compatibilité de la négation avec la relation d'équivalence. Si c'est le cas, il suffira de choisir les deux seuils égaux pour assurer cette compatibilité.

Définition A.0.3. *Ordre strict sur les classes d'équivalences de motifs.*
a et b étant deux motifs sur $\mathcal{S}$, on dit que la classe du motif a précède strictement celle du motif b et on note $cl_{\gamma,\delta}(a) \prec_{\gamma,\delta} cl_{\gamma,\delta}(b)$ quand on est dans l'un des deux cas suivants.
- *soit $type_{\gamma\delta}(a) < type_{\gamma\delta}(b)$*
- *soit $type_{\gamma\delta}(a) = type_{\gamma\delta}(b) = 0$ et $\forall s \in \mathcal{S},\ \mu_a(s) \leq \mu_b(s)$ et $\exists s \in \mathcal{S},\ \mu_a(s) \neq \mu_b(s)$*

Le sens choisi de la relation d'ordre n'indique pas une inclusion entre les deux ensembles de propriétés qui composent les motifs, mais entre les deux ensembles de sujets que sont leurs extensions. La relation de précédence ainsi définie entre classes d'équivalence reprend pour les éléments de type 0 le cas où a'⊂b', et l'étend au cas où a est de type -1 et b de type 0 ou de type 1, et aux cas où a est de type 0 et b de type 1. Dans ces trois derniers cas on n'avait pas nécessairement a'⊂b'. En effet, quand par exemple a est de type -1 et b de type 1, on peut avoir pour un sujet donné une valeur supérieure à 0,5 pour a et inférieure pour b, ceci si γ et δ sont tous deux supérieurs ou égaux à 1. Dans ce cas, les deux classes sont ordonnées sans qu'il y ait inclusion entre leurs extensions. Cette définition permet de garder une relation d'ordre sur les classes de motifs. Mais ce n'est pas, malgré tout, une relation d'ordre total. Si deux classes de type 0 sont engendrées par deux motifs non comparables par l'inclusion entre leurs extensions, elles ne seront pas non plus comparables.
Il convient de vérifier la cohérence de cette définition qui établit une relation d'ordre strict sur les classes d'équivalence à partir d'un seul élément de chaque classe. C'est le but de la propriété suivante :

Propriété A.0.3. *Cohérence de la définition de la précédence stricte.*

Si a et b sont deux motifs tels que $cl_{\gamma,\delta}(a) \prec_{\gamma,\delta} cl_{\gamma,\delta}(b)$ et si c et d sont deux motifs tels que a $eq_{\gamma\delta}$ c et b $eq_{\gamma\delta}$ d, alors $cl_{\gamma,\delta}(c) \prec_{\gamma,\delta} cl_{\gamma,\delta}(d)$

Preuve. *La preuve ne comporte aucune difficulté. Nous la donnons rapidement :*

- *Si a et b sont de types différents, pour vérifier la précédence, le type de a doit être inférieur à celui de b. Comme a et c sont dans la même classe, le type de c est égal à celui de a, et d et b étant dans la même classe, le type de d est égal à celui de b, donc le type de c est inférieur à celui de d.*
- *Si a et b sont de type 0, pour vérifier la précédence, tous les sujets doivent avoir une valeur pour a inférieure à celle pour b. Comme c est de même classe que a, il est de type 0 et tous les sujets ont la même valeur pour c que pour a, et on montrerait de même qu'ils ont la même valeur pour d que pour b, donc leur valeur pour c est inférieure à leur valeur pour d.*

Propriété A.0.4. *Propriétés de la précédence stricte.*
La relation de précédence stricte définie sur les classes d'équivalence des motifs est antiréflexive[169] *asymétrique*[170] *et transitive.*

Preuve. *L'antiréflexivité et l'asymétrie sont immédiates.*
Pour montrer la transitivité, on suppose que a, b et c sont trois motifs tels que $cl_{\gamma,\delta}(a) \prec_{\gamma,\delta} cl_{\gamma,\delta}(b)$ et $cl_{\gamma,\delta}(b) \prec_{\gamma,\delta} cl_{\gamma,\delta}(c)$.

- *soit a est de type -1, alors b est de type 0, donc c est de type 0 ou 1, et on a bien $cl_{\gamma,\delta}(a) \prec_{\gamma,\delta} cl_{\gamma,\delta}(c)$.*
- *soit a est de type 0, b de type 0, et c est de type 1, et on a bien $cl_{\gamma,\delta}(a) \prec_{\gamma,\delta} cl_{\gamma,\delta}(c)$*
- *soit a, b et c sont de type 0, alors a'⊂b' et b'⊂c', donc a'⊂c', et on a bien $cl_{\gamma,\delta}(a) \prec_{\gamma,\delta} cl_{\gamma,\delta}(c)$.*

Définition A.0.4. *Ordre large sur les classes d'équivalences de motifs.*
a et b étant deux motifs sur $\mathcal{S}$, on dit que la classe du motif a précède celle du motif b et on note $cl_{\gamma,\delta}(a) \preccurlyeq_{\gamma,\delta} cl_{\gamma,\delta}(b)$ quand on est dans l'un des deux cas suivants.

- *soit $cl_{\gamma,\delta}(a) \prec_{\gamma,\delta} cl_{\gamma,\delta}(b)$*
- *soit $cl_{\gamma,\delta}(a) = cl_{\gamma,\delta}(b)$*

Propriété A.0.5. *Propriétés de la précédence large :*
La relation de précédence large définie sur les classes d'équivalence des motifs est réflexive, antisymétrique et transitive. C'est donc une relation d'ordre.

Preuve.

- *La réflexivité est immédiate car on a $cl_{\gamma,\delta}(a) = cl_{\gamma,\delta}(a)$.*
- *Pour l'antisymétrie, on suppose que a et b sont tels que $cl_{\gamma,\delta}(a) \preccurlyeq_{\gamma,\delta} cl_{\gamma,\delta}(b)$ et $cl_{\gamma,\delta}(b) \preccurlyeq_{\gamma,\delta} cl_{\gamma,\delta}(a)$.*
  la relation $cl_{\gamma,\delta}(a) \prec_{\gamma,\delta} cl_{\gamma,\delta}(b)$ n'est compatible ni avec $cl_{\gamma,\delta}(a) \prec_{\gamma,\delta} cl_{\gamma,\delta}(b)$ car la relation stricte est antisymétrique, ni avec $cl_{\gamma,\delta}(a) = cl_{\gamma,\delta}(b)$ car elle est antiréflexive. On est donc dans le cas où $cl_{\gamma,\delta}(a) = cl_{\gamma,\delta}(b)$.
- *Pour la transitivité, on suppose qu'on a trois classes telles que $cl_{\gamma,\delta}(a) \preccurlyeq_{\gamma,\delta} cl_{\gamma,\delta}(b)$ et $cl_{\gamma,\delta}(b) \preccurlyeq_{\gamma,\delta} cl_{\gamma,\delta}(c)$, et on doit prouver que $cl_{\gamma,\delta}(a) \preccurlyeq_{\gamma,\delta} cl_{\gamma,\delta}(c)$. On a 3 cas possibles : soit les deux relations sont la précédence stricte, soit l'une est la précédence stricte et l'autre est l'égalité des classes, soit les deux sont l'égalité des classes. Pour le premier cas, on déduit le résultat attendu de la transitivité de la précédence stricte. Pour le deuxième cas, cela vient de la cohérence entre la précédence stricte et la relation d'équivalence. et pour le dernier cela provient des propriétés de classes d'équivalence.*

[169] Pour aucune classe x, on n'a x $\prec_{\gamma,\delta}$ x.
[170] Pour aucun couple de classes {x, y}, on n'a simultanément x $\prec_{\gamma,\delta}$ y et y $\prec_{\gamma,\delta}$ x.

La relation d'ordre que nous venons de définir sur l'ensemble des classes de motifs associés à une relation floue, reste une relation d'ordre quand on la restreint à l'ensemble des classes des propriétés floues. D'après H. Rasiowa et N. Cat Ho [200], on peut faire de l'ensemble des classes des propriétés floues muni de la relation d'ordre, une "expansion LT" qui est le treillis complet des ensembles sous la relation d'inclusion, contenant les éléments I suivants :

$$I = \bigcup_{t \in \mathcal{T}} \{I(t) : I(t) \preccurlyeq_{\gamma,\delta} I\}$$

Par exemple, si $\mathcal{P}$={a, b, c} est l'ensemble des trois propriétés du tableau 9.3, si on prend γ=δ=1, les propriétés sont toutes trois de type 0 et sont dans trois classes différentes, vérifiant la seule relation $cl_{\gamma,\delta}(c) \preccurlyeq_{\gamma,\delta} cl_{\gamma,\delta}(c)$. On identifie les trois classes à leurs propriétés, et on obtient que LT est l'ensemble des parties de $\mathcal{P}$ auquel on a retiré {a}, qui a été confondu avec {a,c} et {a,b} qui a été confondu avec {a,b,c}, soit LT={$\emptyset$,{b}, {c}, {a,c}, {b,c},{a,b,c}}. On constate qu'on a un motif fermé de trop, le motif fermé correspondant {b,c}, qui ne figure pas dans le diagramme de la figure 9.2. En effet, ce motif fermé est confondu avec celui de {a,b,c}, les motifs bc et abc ayant les mêmes éléments. Cette façon de faire, "automatique", à partir des classes de propriétés munies de l'ordre ne prend pas en compte le fait que bc $eq_{\gamma\delta}$ abc, car c'est une relation entre motifs contenant plus d'une propriété.
Pour obtenir le véritable treillis des motifs fermés, il faudrait fusionner tous les motifs fermés associés à des motifs $\gamma\delta$-égaux. Mais cette fusion ne peut conserver la structure de treillis que si la relation de $\gamma\delta$-égalité est compatible avec les opérateurs du treillis (inf et sup en général, ici intersection et réunion). C'est ce que nous allons contrôler maintenant.

La réunion de deux motifs flous A et B est obtenue de la même façon que la réunion de propriétés floues, c'est-à-dire par l'intersection des extensions A' et B'. Et l'intersection des motifs est obtenue en réunissant des extensions. Les différents résultats que donnent ces opérations en fonction du type de A et de B sont dans le tableau 1.

Type de la réunion de A et B			
Types	A : -1	A : 0	A : 1
B : -1	-1	-1	-1
B : 0	-1	0 ou -1	0 ou -1[1]
B : 1	-1	0 ou -1[1]	0[1] ou 1

Type de l'intersection de A et B			
Types	A : -1	A : 0	A : 1
B : -1	0[2] ou -1	0 ou 1[2]	1
B : 0	0 ou 1[2]	0 ou 1	1
B : 1	1	1	1

TAB. A.1 – Type du résultat de la réunion et de l'intersection de deux motifs.

Nous choisissons deux seuils entiers γ et δ positifs ou nuls, petits devant le nombre total de sujets N. Quand δ est nul, et quand la relation n'est pas floue, nous retrouvons les résultats habituels d'extraction de règles avec seuil de support non nul figurant dans les articles de fouille de données déjà cités. Si les deux sont nuls, nous retombons sur la définition courante du treillis des motifs fermés, que nous avons généralisée ci-dessus aux relations floues sans difficulté.
Nous allons d'abord établir les éléments figurant dans le tableau 1 pour la réunion. La valeur obtenue pour un sujet lors de la réunion des deux motifs est le minimum des valeurs obtenues pour chaque motif. Pour faciliter la lecture de la justification, nous noterons G- (resp. G+) les sujets de $\mathcal{S}$ ayant une valeur inférieure ou égale (respectivement supérieure) à 0,5 pour une propriété

[1] Un résultat de ce type est possible seulement si $\gamma \neq 0$.
[2] Un résultat de ce type est possible seulement si $\delta \neq 0$.

spécifiée.

- A de type -1, B de type quelconque : pour A, il y a au plus γ sujets qui sont dans le groupe G+, tous les autres étant dans le groupe G-. Pour la réunion, les sujets ont une valeur qui est le minimum de celle qu'ils ont pour A et de celle qu'ils ont pour B, elle est donc inférieure ou égale à celle qu'ils ont pour A. Ceux qui étaient dans G- y restent, alors que ceux qui étaient dans G+ peuvent rester si leur valeur n'est pas passée en dessous de 0,5, sinon ils passent dans G-. Il y a donc autant ou moins de sujets dans G+, donc au plus γ sujets. Ce qui fait que la réunion est de type -1.
- B de type -1, A de type quelconque : on déduit par le même raisonnement que la réunion est de type -1.
- A de type 0, B de type 0 : pour A, il y plus de γ sujets qui sont dans le groupe G+, et plus de δ sujets qui sont dans le groupe G-. Quand on fait le minimum des valeurs d'un sujet pour A et pour B, il ne peut pas quitter le groupe G-, par contre, il peut quitter le groupe G+ si sa valeur pour B est inférieure à 0,5, ce qui peut arriver car B est de type 0. Ainsi le résultat est de type 0 ou -1 selon la façon dont le groupe G+ a diminué.
- A de type 0, B de type 1 : la seule différence avec le cas précédent est que si δ est nul, le groupe G+ ne peut pas diminuer, et donc le résultat est de type 0. Alors que si δ est non nul, le résultat peut être des deux types 0 ou -1.
- B de type 0, A de type 1 : même cas que précédemment.
- A de type 1, B de type 1 : si $\delta>0$, pour A, le nombre de sujets qui sont dans G- est inférieur ou égal à δ, et il ne peut qu'augmenter, car si des sujets de G+ pour A sont dans G- pour B, ils passent dans G- pour la réunion. Mais c'est le cas pour au plus δ sujets, donc il y aura au plus $2\dot{\delta}$ sujets dans le groupe G-, et au moins N-$2\dot{\delta}$ sujets dans le groupe G+. Le résultat peut donc être de type 0 ou 1, mais pas de type -1, car γ et δ sont petits devant N, donc N-$2\dot{\delta} > \gamma$. Et si δ=0, il n'y a pas de sujet dans G- pour A, et comme il n'y en a pas non plus pour B, il n'y en a pas pour le résultat. Donc on obtient un résultat de type 1.

Pour l'intersection, on reprend la justification précédente en remplaçant minimum par maximum, et en échangeant δ et γ ainsi que 1 et -1.

Examinons la compatibilité de la réunion avec la relation d'équivalence. Pour cela nous prenons 4 motifs a, b, c et d tels que a et c sont (γ, δ)-égaux, b et d sont (γ, δ)-égaux, et nous nous interrogeons sur la (γ, δ)-égalité de a$\cup$b et c$\cup$d.

- si a ou b est de type -1, la réunion l'est également. Comme c et d sont respectivement équivalents à a et b, ils sont de type -1, et leur réunion aussi, donc les réunions sont dans la même classe, qui est $cl^{-1}_{\gamma\delta}$.
- si a et b sont de type 0, c et d le sont également. Et toutes les valeurs des sujets sont les mêmes pour a (resp. b) et c (resp. d). Donc les 2 réunions ont les mêmes valeurs, donc sont dans la même classe.
- si a est de type 0 et b de type 1, c est alors de type 0 et d de type 1. Les valeurs de tous les sujets pour a et pour c sont les mêmes, mais pas celles pour b et pour d. Et si les réunions sont toutes deux de type 0, les valeurs pour tous les sujets ont peu de chances d'être égales, ce qui fait que les réunions seront dans deux classes différentes dans la plupart des cas.
- si a et b sont de type 1, c et d le sont également, et si δ est non nul, et que la réunion de a et de b, ou celle de b et de c, est de type 0, elles ont très peu de chances d'être dans la

même classe, pour le même raison que dans le cas précédent. Par contre si δ est nul, on a alors $cl_{\gamma\delta}(\mathrm{a}) = cl_{\gamma\delta}(\mathrm{b}) = cl_{\gamma\delta}(\mathrm{c}) = cl_{\gamma\delta}(\mathrm{d}) = cl_{\gamma\delta}(\emptyset) = cl^1_{\gamma\delta}$, et les réunions sont toutes dans la classe de type 1.

Voici des définitions de la réunion et de l'intersection de classes de motifs qui restent valables quand on change le représentant de la classe sur lequel elles s'appuient. Comme les classes de motifs sont des ensembles, pour ne pas confondre la réunion que nous définissons ici entre les classes de motifs avec la réunion ensembliste classique, nous la notons $\cup_{\gamma\delta}$, et l'intersection de classes posant le même problème, nous la notons $\cap_{\gamma\delta}$:

Définition A.0.5. *Réunion de deux classes.*
Soient A et B deux motifs flous sur un ensemble de sujets $\mathcal{S}$ de taille N, γ et δ deux entiers positifs ou nuls petits devant N, on définit la réunion des classes de A et de B de la façon suivante :

$$cl_{\gamma\delta}(A) \cup_{\gamma\delta} cl_{\gamma\delta}(B) = \begin{cases} cl^{-1}_{\gamma\delta} & \text{si } A \text{ ou } B \text{ est de type } -1. \\ cl_{\gamma\delta}(A \cup B) & \text{si } A \text{ et } B \text{ sont de type } 0. \\ cl_{\gamma\delta}(B) & \text{si } A \text{ est de type } 1, \text{ et } B \text{ de type } 0. \\ cl_{\gamma\delta}(A) & \text{si } A \text{ est de type } 0, \text{ et } B \text{ de type } 1. \\ cl^1_{\gamma\delta} & \text{si } A \text{ et } B \text{ sont de type } 1. \end{cases}$$

Et pour l'intersection on prend la définition similaire :

Définition A.0.6. *Intersection de deux classes.*
Soient A et B deux motifs flous sur un ensemble de sujets $\mathcal{S}$ de taille N, γ et δ deux entiers positifs ou nuls petits devant N, on définit l'intersection des classes de A et de B de la façon suivante :

$$cl_{\gamma\delta}(A) \cap_{\gamma\delta} cl_{\gamma\delta}(B) = \begin{cases} cl^1_{\gamma\delta} & \text{si } A \text{ ou } B \text{ est de type } 1. \\ cl_{\gamma\delta}(A \cap B) & \text{si } A \text{ et } B \text{ sont de type } 0. \\ cl_{\gamma\delta}(B) & \text{si } A \text{ est de type } -1, \text{ et } B \text{ de type } 0. \\ cl_{\gamma\delta}(A) & \text{si } A \text{ est de type } 0, \text{ et } B \text{ de type } -1. \\ cl^{-1}_{\gamma\delta} & \text{si } A \text{ et } B \text{ sont de type } -1. \end{cases}$$

Les différents résultats que produisent ces opérations sur les classes de motifs $cl_{\gamma\delta}(\mathrm{A})$ et $cl_{\gamma\delta}(\mathrm{B})$ en fonction du type de A et de B se trouvent dans le tableau 2.

Valeur de $cl_{\gamma\delta}(A) \cup_{\gamma\delta} cl_{\gamma\delta}(B)$			
Types	A : -1	A : 0	A : 1
B : -1	$cl^{-1}_{\gamma\delta}$	$cl^{-1}_{\gamma\delta}$	$cl^{-1}_{\gamma\delta}$
B : 0	$cl^{-1}_{\gamma\delta}$	$cl_{\gamma\delta}(A \cup B)$	$cl_{\gamma\delta}(B)$
B : 1	$cl^{-1}_{\gamma\delta}$	$cl_{\gamma\delta}(A)$	$cl^1_{\gamma\delta}$

Valeur de $cl_{\gamma\delta}(A) \cap_{\gamma\delta} cl_{\gamma\delta}(B)$			
Types	A : -1	A : 0	A : 1
B : -1	$cl^{-1}_{\gamma\delta}$	$cl_{\gamma\delta}(A)$	$cl^1_{\gamma\delta}$
B : 0	$cl_{\gamma\delta}(B)$	$cl_{\gamma\delta}(A \cap B)$	$cl^1_{\gamma\delta}$
B : 1	$cl^1_{\gamma\delta}$	$cl^1_{\gamma\delta}$	$cl^1_{\gamma\delta}$

TAB. A.2 – Résultat de la réunion et de l'intersection de deux classes selon le type des motifs.

Propriété A.0.6. *Propriétés de la réunion et de l'intersection des classes de motifs.*
Étant donnés un ensemble fini $\mathcal{S}$ de taille N, un ensemble fini $\mathcal{P}$ de propriétés, une relation floue sur $\mathcal{S} \times \mathcal{P}$, γ et δ deux entiers positifs ou nuls petits devant N, la structure formée de l'ensemble des classes de motifs flous sur $\mathcal{N}$ associées à la relation d'équivalence $eq_{\gamma\delta}$, de l'opération de réunion et de l'opération d'intersection qu'on vient de définir, est un treillis. Ce treillis a deux extrema qui sont la classe du motif vide et la classe du motif plein $\mathcal{P}$. L'ordre canonique associé à cette structure de treillis est l'inclusion entre classes précédemment définie.

Preuve. *La commutativité et l'associativité de l'intersection et de la réunion découle immédiatement de leurs définitions. Montrons que si A et B sont deux motifs flous, on a les propriétés* $cl_{\gamma\delta}(A) \cup (cl_{\gamma\delta}(A) \cap cl_{\gamma\delta}(B)) = cl_{\gamma\delta}(A)$ *et* $cl_{\gamma\delta}(A) \cap (cl_{\gamma\delta}(A) \cup cl_{\gamma\delta}(B)) = cl_{\gamma\delta}(A)$*. Nous montrons la première, l'autre s'en déduit par dualité. Calculons* $cl_{\gamma\delta}(A) \cap (cl_{\gamma\delta}(A) \cup cl_{\gamma\delta}(B))$

- *Si A est de type -1, c'est-à-dire si* $cl_{\gamma\delta}(A) = cl^{-1}_{\gamma\delta}$*, cette classe étant un élément absorbant pour la réunion, et un élément neutre pour l'intersection, on a*
$$\begin{aligned} cl_{\gamma\delta}(A) \cap (cl_{\gamma\delta}(A) \cup cl_{\gamma\delta}(B)) &= cl^{-1}_{\gamma\delta} \cap (cl^{-1}_{\gamma\delta} \cup cl_{\gamma\delta}(B)) \\ &= cl^{-1}_{\gamma\delta} \cap cl^{-1}_{\gamma\delta} \\ &= cl^{-1}_{\gamma\delta} \end{aligned}$$
- *Si A est de type 1, c'est-à-dire si* $cl_{\gamma\delta}(A) = cl^{1}_{\gamma\delta}$*, cette classe étant absorbante pour l'intersection, on obtient*
$$\begin{aligned} cl_{\gamma\delta}(A) \cap (cl_{\gamma\delta}(A) \cup cl_{\gamma\delta}(B)) &= cl^{1}_{\gamma\delta} \cap (cl_{\gamma\delta}(A) \cup cl_{\gamma\delta}(B)) \\ &= cl^{1}_{\gamma\delta} \end{aligned}$$
- *Si B est de type -1, c'est-à-dire si* $cl_{\gamma\delta}(B) = cl^{-1}_{\gamma\delta}$*, cette classe étant absorbante pour la réunion, et élément neutre pour l'intersection, on obtient*
$$\begin{aligned} cl_{\gamma\delta}(A) \cap (cl_{\gamma\delta}(A) \cup cl_{\gamma\delta}(B)) &= cl^{A}_{\gamma\delta} \cap (cl_{\gamma\delta}(A) \cup cl^{-1}_{\gamma\delta}) \\ &= cl^{A}_{\gamma\delta} \cap cl^{-1}_{\gamma\delta} \\ &= cl^{A}_{\gamma\delta} \end{aligned}$$
- *Si B est de type 1, c'est-à-dire si* $cl_{\gamma\delta}(B) = cl^{1}_{\gamma\delta}$*, cette classe étant neutre pour la réunion, on obtient*
$$\begin{aligned} cl_{\gamma\delta}(A) \cap (cl_{\gamma\delta}(A) \cup cl_{\gamma\delta}(B)) &= cl_{\gamma\delta}(A) \cap (cl_{\gamma\delta}(A) \cup cl^{1}_{\gamma\delta}) \\ &= cl^{A}_{\gamma\delta} \cap cl^{A}_{\gamma\delta} \\ &= cl_{\gamma\delta}(A) \end{aligned}$$
- *Si A est de type 0 et B de type 0,* $A \cup B$ *est de type 0 ou -1*
  - *si* $A \cup B$ *est de type 0, on obtient*
$$\begin{aligned} cl_{\gamma\delta}(A) \cap (cl_{\gamma\delta}(A) \cup cl_{\gamma\delta}(B)) &= cl_{\gamma\delta}(A) \cap cl_{\gamma\delta}(A \cup B) \\ &= cl_{\gamma\delta}(A \cap (A \cup B)) \ , \quad \textit{les motifs formant un treillis} \\ &= cl_{\gamma\delta}(A) \end{aligned}$$
  - *si* $A \cup B$ *est de type -1, alors on obtient*
$$\begin{aligned} cl_{\gamma\delta}(A) \cap (cl_{\gamma\delta}(A) \cup cl_{\gamma\delta}(B)) &= cl_{\gamma\delta}(A) \cap cl^{-1}_{\gamma\delta} \\ &= cl_{\gamma\delta}(A) \end{aligned}$$

B

Le paradoxe de Simpson à l'épreuve des règles d'association

Sommaire

Nous avons exposé dans le chapitre 7 que les motifs et les règles d'association sont particulièrement bien adaptés pour représenter les liaisons complexes que sont les interactions entre variables (la définition d'une interaction figure dans le chapitre 2 de l'état de l'art, section 2.3.1). Nos méta-règles de nettoyage, exposées dans le chapitre 8, permettent à l'utilisateur de faire plus ou moins disparaître ces interactions au profit des effets principaux, selon le niveau de description plus ou moins détaillé auquel il choisit de se placer. Nous avons exposé ce qu'est le paradoxe de Simpson également dans le chapitre 2 (section 2.6) et signalé qu'il diffère de par sa nature de l'interaction. Puis nous avons signalé dans le chapitre 7 (section 7.2.3) qu'il ne se distingue pas de l'interaction par son effet sur le jeu de règles d'association. Le but de cette annexe est de le démontrer.

B.1 Introduction

Après avoir défini le paradoxe de Simpson, nous étudions les conditions que les données doivent vérifier pour que ce paradoxe apparaisse, puis, après avoir examiné les effets de ces conditions dans chacune des étapes de l'extraction des règles d'association, nous évaluons l'incidence de ce paradoxe sur l'interprétation des règles.

B.2 Le paradoxe

Nous allons prendre un exemple de données fictives vérifiant ce paradoxe, afin d'opposer la relation attendue à celle observée, puis nous l'exposerons en termes plus mathématiques.

B.2.1 Un exemple

Voici des données fictives qui illustrent le paradoxe de Simpson : on connaît la valeur de 3 variables A, B et C pour 2000 étudiants d'une ville donnée. Par exemple A est "être de sexe masculin", B est "réussir à l'examen", et C est "appartenir à l'établissement scolaire C". Le triplet des 3 variables peut prendre 8 valeurs données. Dans le tableau 1 figure la répartition des 2000 étudiants selon ces 8 valeurs. On peut lire en première ligne que 970 étudiants sont de sexe féminin (A=0), ont échoué à l'examen(B=0) et ne font pas partie de l'établissement C (C=0).

A	B	C	Effectifs
0	0	0	970
0	0	1	20
0	1	0	250
0	1	1	250
1	0	0	250
1	0	1	30
1	1	0	30
1	1	1	200
		total	2000

TAB. B.1 – Les données sous forme de répartition des 2000 étudiants selon les 8 modalités du triplet de variables (A,B,C)

Premier cas : A et B sans C

Pour étudier la relation entre A et B, on construit le tableau 2 en croisant les variables A et B, sans tenir compte de la variable C.

AxB	B=0	B=1	total
A=0	990	500	1490
A=1	280	230	510
total	1270	730	2000

TAB. B.2 – Tableau de contingence des variables A et B

Le nombre 990, à l'intersection de la ligne A0 et de la colonne B0, est le nombre d'étudiants de la ville qui sont de sexe féminin (A=0) et ont échoué à l'examen (B=0). Ce sont les 970 étudiants qu'on vient d'évoquer plus haut, qui ne sont pas de l'établissement C, auxquels on a ajouté les 20 étudiants de la ligne suivante du tableau 1, ceux-ci étant également de sexe féminin et ayant échoué à l'examen, mais provenant de l'établissement C (C=1). Pour comparer la réussite des filles à celle des garçons, on compare le tableau 2 d'effectifs "observés" au tableau 3 d'effectifs "attendus" en cas d'indépendance entre les deux variables, "sexe" et "réussite" [171].

[171] On obtient l'effectif attendu de chaque cellule en multipliant le total de sa ligne par le total de sa colonne, et

AxB	B=0	B=1	total
A=0	946.15	543.85	1490
A=1	323.85	186.15	510
total	1270	730	2000

TAB. B.3 – Tableau de contingence des variables A et B en cas d'indépendance

On remarque notamment que le nombre de filles ayant échoué à l'examen (990) est plus grand que celui qu'on attendait (946.15) en cas d'indépendance, alors que c'est l'inverse pour les garçons.

On peut mesurer l'écart de ce tableau à l'indépendance par la différence des produits diagonaux rapportée à l'effectif total, ici $\frac{990\times230-500\times280}{2000} = 43.85$, qui est nulle en cas d'indépendance, mais on peut également comparer le rapport du nombre de filles ayant échoué sur celui de filles ayant réussi au rapport correspondant pour les garçons en les divisant, ici $\frac{990}{500}/\frac{280}{230} = 1.63$. Ce nombre, appelé *odd-ratio*, est égal à 1 en cas d'indépendance. On préfère souvent utiliser le logarithme de ce nombre, ici 0.49, qui devient nul en cas d'indépendance [181]. Le logarithme de l'*odd-ratio* étant positif, (comme la différence des produits diagonaux), on dira que la liaison entre les variables A et B est positive [172].

Prenons maintenant en compte la variable C

On reprend le tableau 1, d'où on extrait d'abord les 4 lignes correspondant à C=0, pour construire la partie gauche du tableau 4 croisant A et B pour les étudiants qui ne sont pas dans l'établissement C puis on construit la partie droite du tableau 4 de la même façon avec les 4 lignes du tableau 1 correspondant aux étudiants de l'établissement C ($C = 1$). Si on examine la relation entre A et B, en se restreignant à C=0, on obtient une liaison "négative" entre A et B, et de même si on se restreint à C=1. Dans la partie gauche du tableau 4, pour les 1500 étudiants ne faisant pas partie de l'établissement C ($C = 0$), le logarithme de l'*odd-ratio* est de -0.76, et il est de -0.63 dans la partie droite, pour les 500 étudiants de l'établissement C. Ce qui signifie que *dans chacun de ces 2 établissements les filles réussissent mieux que les garçons, alors que pour les 2 établissements réunis, c'est l'inverse.*

Liaison entre A et B sachant C=0 : -0.76			
AxB	B=0	B=1	total
A=0	970	250	1220
A=1	250	30	280
total	1220	280	1500

Liaison entre A et B sachant C=1 : -0.63			
AxB	B=0	B=1	total
A=0	20	250	270
A=1	30	200	230
total	50	450	500

TAB. B.4 – Tableaux de contingence des variables A et B pour les 2 valeurs de C

en divisant le tout par le total général. Par exemple le nombre attendu d'étudiants de sexe féminin ($A = 0$) ayant réussi l'examen ($B = 1$) est $543.85 = \frac{730\times1490}{2000}$

[172] Le mot "positif" est employé par similarité avec le coefficient de corrélation de deux variables quantitatives. Dans le cas des deux variables booléennes, cela signifie qu'il y a plus d'étudiants vérifiant simultanément A=0 et B=0, ou A=1 et B=1 que d'étudiants vérifiant A=0 et B=1, ou A=1 et B=0.

B.2.2 Expression mathématique du paradoxe

On reprend dans le tableau 5 les données du tableau 2, en les généralisant.

AxB	B0	B1	total
A0	a	b	a+b
A1	c	d	c+d
total	a+c	b+d	N=a+b+c+d

TAB. B.5 – Tableau de contingence de 2 variables observées sur N objets

Puis on fait intervenir la troisième variable C, c'est-à-dire qu'on découpe chacun des 4 effectifs a, b, c et d du tableau 5 en 2 parties, l'une correspondant à la valeur C0, et l'autre à la valeur C1. Les effectifs correspondant à C0 sont a', b', c' et d', et ceux correspondant à C1 sont a", b", c" et d" répartis comme indiqué le tableau 6. On a donc les relations suivantes : $a = a' + a"$, $b = b' + b"$, $c = c' + c"$, $d = d' + d"$.

Liaison entre A et B sachant C=0			
AxB	B0	B1	total
A0	a'	b'	a'+b'
A1	c'	d'	c'+d'
total	a'+c'	b'+d'	N=a'+b'+c'+d'

Liaison entre A et B sachant C=1			
AxB	B0	B1	total
A0	a"	b"	a"+b"
A1	c"	d"	c"+d"
total	a"+c"	b"+d"	N=a"+b"+c"+d"

TAB. B.6 – Tableau de contingence de 2 variables observées sur N objets pour C=0 et C=1

Le paradoxe de Simpson se traduit alors par le fait que si $\frac{a'}{b'} < \frac{c'}{d'}$ et $\frac{a"}{b"} < \frac{c"}{d"}$ on n'a pas nécessairement $\frac{a'+a"}{b'+b"} < \frac{c'+c"}{d'+d"}$, c'est-à-dire que la somme "terme à terme" de 2 inégalités de proportions ne donne pas nécessairement une inégalité de même sens. C'est le cas de l'exemple pour lequel on a

- pour $C = 0$: $3.88 = \frac{970}{250} < \frac{250}{30} = 8.33$
- pour $C = 1$: $0.08 = \frac{20}{250} < \frac{30}{200} = 0.15$
- sans tenir compte de C : $\frac{970+20}{250+250} > \frac{250+30}{30+200}$ soit $1.98 = \frac{990}{500} > \frac{280}{230} = 1.22$

ce qui donne deux inégalités de même sens pour C fixé, et une de sens contraire quand on ne tient pas compte de C[173].

On peut faire apparaître aisement cet effet en exprimant $ad-bc = (a'+a")(d'+d")-(b'+b")(c'+c")$ sous la forme $(a'd' - b'c') + (a"d" - b"c") + (a'd" + a"d' - b'c" - b"c')$. Ce qui montre que le paradoxe de Simpson apparaît quand $a'd'-b'c' < 0$, $a"d"-b"c" < 0$ et que $a'd"+a"d'-b'c"-b"c'$ est assez grand pour qu'on ait $ad - bc > 0$.

B.3 Les règles d'associations

B.3.1 Extraction des règles à partir de deux variables

On dispose d'une matrice booléenne de N lignes et 2 colonnes, c'est-à-dire contenant uniquement les valeurs 0 et 1 des propriétés A et B pour les N sujets, comme figurant dans le tableau

[173] Le même effet est visible pour les odd-ratios et leurs logarithmes : on obtient également des odd-ratios inférieurs à 1 (0.47 et 0.53) pour C fixé, et supérieur à 1 (1.63) quand on ne tient pas compte de C, les logarithmes correspondants étant -0.76 et -0.63 à C fixé et 0.49 sinon.

4.2. On extrait les règles[174] dont le support dépasse un seuil donné (on ne garde pas les règles de support nul). Et on classe les règles extraites selon la valeur d'autres indices, comme ceux pour lesquels on a donné la définition dans le paragraphe 2.5.2 . Pour ces indices, plus leur valeur est élevée, plus la qualité de la règle est grande.

On dira que la règle A0→B0 est de meilleure qualité que la règle A0→B1 pour un indice donné E si l'expression $E(A0 \rightarrow B0) - E(A0 \rightarrow B1)$ est positive. Nous avons établi (voir preuves en appendice, dernière section de cette annexe) que ces indices se divisent en 2 groupes. Pour le premier groupe, qui contient le support, la fréquence, la confiance et l'étonnement, cette expression est de même signe que $D = a - b$, et pour le second, qui contient la différence, l'intérêt, la satisfaction, la nouveauté, la conviction et l'implication[175], cette expression est de même signe que $P = ad - bc$, donc que le log-ratio du tableau 5. Si on reprend notre exemple, le deuxième groupe d'indices privilégie donc les relations positives entre A et B, qui correspondent aux 4 règles A0→B0, A1→B1, B0→A0 et B1→A1 par rapport aux 4 règles négatives A0→B1, A1→B0, B0→A1 et B1→A0. Le premier groupe d'indices donne les mêmes résultats pour la moitié des règles seulement (les règles A0→B0 ($990 > 500$) et B0→A0 ($990 > 280$)), et pour les 2 autres, privilégie les règles négatives A1→B0 ($280 > 230$) et B1→A0 ($500 > 230$). Quand les indices donnent des informations contradictoires, on peut choisir de rejeter les 2 règles correspondantes, ou bien de garder celles maximisant la valeur d'un indice particulier, adapté aux besoins spécifiques de la personne qui a collecté les données. Une autre façon de faire assez courante est d'imposer un seuil à la confiance, de façon arbitraire ou sur des bases statistiques [43], puis d'examiner les valeurs les plus élevées de certains autres indices [52] choisis pour leur sémantique. Si le seuil de confiance dépasse 0.5, comme c'est généralement le cas quand il est choisi arbitrairement (le plus souvent 0.8), la seule des deux règles qui dépasse ce seuil est gardée[176]. Si ce seuil est plus bas, les deux règles peuvent le dépasser, et dans ce cas les autres indices peuvent intervenir dans le choix de la meilleure. Nous allons voir l'effet que ces diverses façons de décider produisent sur les règles obtenues à partir des trois variables, afin de repérer les éventuelles contradictions dues au paradoxe de Simpson.

B.3.2 L'extraction des règles issues de 3 variables

Les motifs extraits sont donnés sans que soient détaillés les algorithmes qu'on pourra trouver dans [13].

Les motifs extraits

On part d'un tableau de N lignes et de 6 colonnes, de la forme du tableau 7. Puis on l'utilise pour construire les motifs de support non nul.

Dans le tableau 8 figurent ceux les motifs que nous obtenons pour les données de l'exemple (tableau 1). Bien que les 3 variables du tableau 1 aient 6 modalités, on ne peut pas avoir dans un motif plus d'une modalité par variable, les modalités d'une même variable s'excluant, ce qui fait que la longueur maximale d'un motif est 3. Comme aucun effectif du tableau 1 n'est inférieur à 20, en choisissant 20 pour seuil de support, on obtient tous les motifs possibles de longueur 3, mais également tous ceux de longueur 2, et de longueur 1, dont les supports ne peuvent être inférieurs à ceux de longueur 3.

[174] Notre but étant seulement de contrôler si le paradoxe de Simpson apparaît, on ne considérera pas les règles contenant le motif vide.

[175] Pour l'implication, il faut toutefois que la valeur de $\frac{ad-bc}{N}$ soit assez grande en valeur absolue

[176] Seule une des deux confiances peut dépasser 0.5 car : confiance(A0→B0)+confiance(A0→B1)$= \frac{a}{a+b} + \frac{b}{a+b} = 1$.

Numéro de ligne	A1	A0	B1	B0	C1	C0
1	0	1	0	1	0	1
2	1	0	0	1	1	0
3	1	0	1	0	0	1
4	1	0	0	1	0	1
...	...	...				

TAB. B.7 – Matrice booléenne de données

motif	longueur	support
A0 B0 C0	3	970
A0 B0 C1	3	20
A0 B1 C0	3	250
A0 B1 C1	3	250
A1 B0 C0	3	250
A1 B0 C1	3	30
A1 B1 C0	3	30
A1 B1 C1	3	200

motif	longueur	support
A0 B0	2	990
A0 B1	2	500
A0 C0	2	1220
A0 C1	2	270
A1 B0	2	280
A1 B1	2	230
A1 C0	2	280
A1 C1	2	230
B0 C0	2	1220
B0 C1	2	50
B1 C0	2	280
B1 C1	2	450

motif	longueur	support
A0	1	1490
A1	1	510
B0	1	1270
B1	1	730
C0	1	1500
C1	1	500

TAB. B.8 – Les motifs et leur support

Les règles obtenues pour cet exemple

Nous construisons ensuite toutes les règles engendrées par ces motifs. Nous en avons 24 engendrées par les 12 motifs de longueur 2, et 48 engendrées par les 8 motifs de longueur 3, soit 72 en tout. Nous retrouvons bien sûr les 8 règles construites sur les motifs de longueur 2 contenant une modalité de A et une modalité de B (voir tableau 9).

No	Partie gauche		Partie droite	Frequ ence	Conf iance	Diffé rence	Inté rêt	Convic tion	Etonne ment	Nouv eauté	Satis faction	Impli cation
1	A0	→	B0	0.50	0.66	0.03	1.0	1.1	0.4	0.02	0.08	0.97
2	A0	→	B1	0.25	0.34	-0.03	0.9	1.0	-0.7	-0.02	-0.05	0.08
3	A1	→	B1	0.12	0.45	0.09	1.2	1.2	-0.1	0.02	0.14	0.99
4	A1	→	B0	0.14	0.55	-0.09	0.9	0.8	0.0	-0.02	-0.24	0.00
5	B0	→	A0	0.50	0.78	0.03	1.0	1.2	0.5	0.02	0.14	0.99
6	B0	→	A1	0.14	0.22	-0.03	0.9	1.0	-1.4	-0.02	-0.05	0.08
7	B1	→	A1	0.12	0.32	0.06	1.2	1.1	-0.5	0.02	0.08	0.97
8	B1	→	A0	0.25	0.68	-0.06	0.9	0.8	0.2	-0.02	-0.24	0.00

TAB. B.9 – Les règles construites sur un motif de longueur 2 contenant une modalité de A et une modalité de B

La règle numéro 1, (A0→ B0), qui représente une liaison positive entre A et B, et que nous appellerons règle "positive" est suivie de la règle "négative" numéro 2 (A0→ B1) ayant la même partie gauche, mais la partie droite contraire. Pour faciliter le choix entre ces 2 règles, chaque règle positive, de numéro impair, est suivie de la règle négative ainsi associée. Et nous retrouvons

les résultats énoncés précédemment, à savoir que sur les couples de règles (1,2) et (5,6) tous les indices s'accordent à donner la préférence à la règle positive. Sur les 2 autres couples, les indices de fréquence, de confiance et d'étonnement privilégient les règles négatives, alors que les autres indices privilégient les règles positives.

Examinons les 16 règles du tableau 10 contenant une modalité de A en partie gauche et une modalité de B en partie droite (ou inversement), avec en plus une modalité de C en partie gauche. Elles sont disposées comme dans le tableau précédent par couples, d'abord la règle positive[177], puis la règle négative associée obtenue en changeant la partie droite. Les 8 premières contiennent C0 et les 8 dernières contiennent C1.

No	Partie gauche		Partie droite	Frequence	Confiance	Différence	Intérêt	Conviction	Etonnement	Nouveauté	Satisfaction	Implication
1	A0 C0	→	B0	0.49	0.80	0.16	1.3	1.8	0.6	0.10	0.44	1.00
2	A0 C0	→	B1	0.13	0.20	-0.16	0.6	0.8	-1.0	-0.10	-0.25	0.00
3	A1 C0	→	B1	0.02	0.11	-0.26	0.3	0.7	-0.3	-0.04	-0.41	0.00
4	A1 C0	→	B0	0.13	0.89	0.26	1.4	3.4	0.2	0.04	0.71	1.00
5	B0 C0	→	A0	0.49	0.80	0.05	1.1	1.2	0.5	0.03	0.20	1.00
6	B0 C0	→	A1	0.13	0.20	-0.05	0.8	0.9	-1.4	-0.03	-0.07	0.02
7	B1 C0	→	A1	0.02	0.11	-0.15	0.4	0.8	-0.4	-0.02	-0.20	0.00
8	B1 C0	→	A0	0.13	0.89	0.15	1.2	2.4	0.1	0.02	0.58	1.00
9	A0 C1	→	B0	0.01	0.07	-0.56	0.1	0.4	-0.2	-0.08	-1.54	0.00
10	A0 C1	→	B1	0.13	0.93	0.56	2.5	8.6	0.3	0.08	0.88	1.00
11	A1 C1	→	B1	0.10	0.87	0.50	2.4	4.9	0.2	0.06	0.79	1.00
12	A1 C1	→	B0	0.02	0.13	-0.50	0.2	0.4	-0.1	-0.06	-1.38	0.00
13	B0 C1	→	A0	0.01	0.40	-0.35	0.5	0.4	0.0	-0.01	-1.35	0.00
14	B0 C1	→	A1	0.02	0.60	0.35	2.4	1.9	0.0	0.01	0.46	1.00
15	B1 C1	→	A1	0.10	0.44	0.19	1.7	1.3	-0.1	0.04	0.25	1.00
16	B1 C1	→	A0	0.13	0.56	-0.19	0.7	0.6	0.0	-0.04	-0.74	0.00

TAB. B.10 – Les règles construites sur un motif de longueur 3 avec une partie contenant une modalité de A et une modalité de B, avec en partie gauche un motif de longueur 2 contenant une modalité de C

Pour C0, les meilleures règles des couples (1,2), (3,4), (5,6) et (7,8) sont respectivement les règles 1, 4, 5 et 8, soit 2 règles positives et 2 règles négatives. Et pour C1, on obtient pour les couples (9,10), (11,12) et (13,14), les meilleures règles respectives 10, 11, 14, et pour le couple (15,16), la règle 15 ou la règle 16 selon qu'on privilégie la confiance et l'étonnement, ou les autres indices. Ce qu'on résume par le tableau 11.

C0		
A0	→	B0
A1	→	B0
B0	→	A0
B1	→	A0

C1		
A0	→	B1
A1	→	B1
B0	→	A1
B1	→	A1

TAB. B.11 – Les règles liant A et B pour chaque modalité de C

On voit que même en privilégiant certains indices, on ne retrouve pas l'effet attendu, c'est-à-dire que les relations entre A et B seraient d'un même sens pour C0 et C1, et du sens contraire

[177] le signe de la liaison, positif ou négatif, évoqué dans cette partie est relatif aux seules variables A et B

si on ne prend pas C en compte. Nous allons essayer de contrôler par des calculs littéraux que le paradoxe de Simpson ne peut pas apparaître sur les règles d'association quand on examine leur qualité avec les indices que nous avons choisis.

Le paradoxe de Simpson et les règles dans le cas général

Pour comparer les deux règles A0 C0→B0 et A0 C0→B1, au lieu de calculer pour chaque indice E la différence E(A0 C0→B0)-E(A0 C0→B1), nous nous contentons, comme dans la partie précédente, de calculer les 2 différences D' et P' sur le tableau de contingence correspondant, qui est le tableau 12. On trouve $D' = a' - b'$ et $P' = a'(d' + b" + d") - b'(c' + a" + c") = (a'd' - b'c') + (a'b" + a'd" - a"b' - b'c")$

AxB	B0	B1	total
A0 C0 A0 C1, A1 C0, A1 C1	a' c'+a"+c"	b' d'+b"+d"	a'+b' c'+a"+c"+d'+b"+d"
total	a'+c'+a"+c"	b'+d'+b"+d"	N=a'+b'+c'+d'+a"+b"+c"+d"

TAB. B.12 – Tableau de contingence pour la règle A0 C0→ B0

On voit déjà sur cette valeur de P' que si a'd'-b'c' est négatif, ce qui correspond à une liaison négative entre A et B pour C=0, alors la valeur de P' peut être positive dès que $(a'b" + a'd" - a"b' - b'c")$ est assez grand, ce qui peut faire préférer la règle A0 C0→B0 à la règle A0 C0→B1. Cela explique les résultats à priori surprenants du tableau 12.

Pour le tableau permettant de comparer la règle A0 C1→ B0 à la règle A0 C1→ B1, il suffit de remplacer les symboles primes par des symboles secondes et inversement. On a donc $D" = a" - b"$, et $P" = a"(d" + b' + d') - b"(c" + a' + c') = (a"d" - b"c") + (a"b' + a"d' - a'b" - b"c')$.

On constate alors que :
$D = a - b = (a + a") - (b + b") = (a' - b') + (a" - b") = D' + D"$
$P = ad - bc = (a'd' - b'c') + (a"d" - b"c") + (a'd" + a"d' - b'c" - b"c') = P' + P"$.

Si D est positif, il ne peut pas se décomposer en une somme de deux nombres négatifs, et la même chose pour P. Ainsi le paradoxe de Simpson ne peut pas se retrouver dans les règles, qu'on prenne le premier groupe ou le second groupe d'indices. C'est-à-dire que pour l'un comme pour l'autre de ces indices, on ne peut pas avoir la règle A0→ B0 qui l'emporte sur la règle A0→ B1, alors que la règle A0C0→ B1 l'emporte sur la règle A0C0→ B0, et que la règle A0C1→ B1 l'emporte sur la règle A0C1→B0. Ainsi, quand on se limite à ces indices, le jeu de règles d'association échappe au paradoxe de Simpson.

Toutefois, il n'est pas exempt de contradictions. Il est en effet tout à fait possible que D positif se décompose en un nombre positif et un nombre négatif, et de même pour P, comme on vient de le voir dans l'exemple, ce qui comporte également une contradiction, même si elle est moins spectaculaire que le paradoxe de Simpson.

B.4 Conclusion

Nous avons montré dans cet article que les jeux de règles d'association de la fouille de données obtenus en sélectionnant les meilleures règles grâce à une utilisation courante de certains indices laissent apparaître des contradictions, mais qu'ils échappent au paradoxe de Simpson, qui est

la plus spectaculaire de ces contradictions. Toutefois, cette démonstration s'appuie sur les deux façons dont opèrent les indices dans le tableau de contingence. L'utilisation d'un autre indice qui agirait selon une troisième façon pourrait produire un jeu de règles d'association présentant des formes de ce paradoxe.

B.5 Appendice : preuves

Pour un indice E de qualité d'une règle, on note par Δ(E) la différence E(A0 C0→B0)-E(A0 C0→B1). Ce qui donne, en reprenant du paragraphe 2.5.2 les formules des indices en fonction des effectifs a, b, c et d :

- Δ(support)=a-b
- Δ(fréquence)= $\frac{a}{N} - \frac{b}{N} = \frac{a-b}{N}$
- Δ(confiance)= $\frac{a}{a+b} - \frac{b}{a+b} = \frac{a-b}{a+b}$
- Δ(étonnement)= $\frac{a-b}{a+c} - \frac{b-a}{b+d} = (a-b)[\frac{1}{a+c} + \frac{1}{b+d}]$
- Δ(différence)= $\frac{ad-bc}{N(a+b)} - \frac{bc-ad}{N(a+b)} = \frac{2(ad-bc)}{N(a+b)}$
- Δ(intérêt)= $\frac{aN}{(a+b)(a+c)} - \frac{bN}{(a+b)(b+d)} = \frac{[a(b+d)-b(a+c)]N}{(a+b)(a+c)(b+d)} = \frac{(ad-bc)N}{(a+b)(a+c)(b+d)}$
- Δ(nouveauté)= $\frac{ad-bc}{N^2} - \frac{bc-ad}{N^2} = \frac{2(ad-bc)}{N^2}$
- Δ(satisfaction)= $\frac{ad-bc}{(a+b)/(b+d)} - \frac{bc-ad}{(a+b)(a+c)} = \frac{(ad-bc)N}{(a+b)(b+d)(a+c)}$
- Δ(conviction)= $\frac{(a+b)(b+d)}{bN} - \frac{(a+b)(a+c)}{aN} = \frac{(a+b)[a(b+d)-b(a+c)]}{abN} = \frac{ad-bc}{abN}$
- Δ(implication)= $f(b,\lambda) - f(a,\lambda')$ avec $\lambda = \frac{(a+b)(b+d)}{N}$ et $\lambda' = \frac{(a+b)(a+c)}{N}$.

On constate bien que pour le support la confiance et l'étonnement, la différence Δ est du signe de a-b, alors qu'elle est du signe de ad-bc pour la différence, l'intérêt, la nouveauté, la satisfaction et la conviction.

Pour l'implication, la différence n'est pas toujours du signe de $ad - bc$. Toutefois, comme $a - \lambda' = \frac{ad-bc}{N} = \lambda - b$, dès que $\frac{ad-bc}{N}$ est suffisamment grand en valeur absolue, b et a se retrouvent suffisamment éloignés du côté opposé non seulement de l'espérance, mais également de la médiane de leur loi respective, pour que f(b,λ) et f(a,λ') aient des valeurs de part et d'autre de 0.5. Dans ce cas, quand ad-bc est positif, $\lambda - b$ l'est également, donc b se trouve être inférieur à l'espérance λ de sa loi, et également à sa médiane, donc $P(X \leq b) < 0.5$ et $f(b,\lambda) > 0.5$. On a de la même façon, a-λ' qui est positif et suffisamment grand pour que a soit supérieur à la médiane de sa loi, et donc que $P(X \leq a) > 0.5$, et de ce fait, $f(a,\lambda') < 0.5$. Et donc $f(b,\lambda) > f(a,\lambda')$ et Δ(implication) est positif. On ferait de même pour prouver que Δ(implication) est négatif quand ad-bc est négatif, et suffisamment grand en valeur absolue. Par exemple, si N=100, parmi les 174 108 tableaux à marges non nulles obtenus en faisant varier les valeurs de a, b, c et d, seuls 1826 n'ont pas leur valeur Δ(implication) du signe de ad-bc. Ces 1826 tableaux faisant exception à la règle, ont une valeur de $\frac{ad-bc}{N}$ comprise entre -0.57 et 0.57, donc négligeable. Et parmi les 11 148 tableaux pour lesquels la valeur de $\frac{ad-bc}{N}$ est comprise entre ces deux bornes, ce sont les seuls à avoir une valeur de Δ(E) dont le signe s'oppose à celui de ad-bc.

C

Interactions entre variables binaires et règles d'Association

Sommaire

C.1 Introduction

Dans cette partie, nous mettons en relation deux formalimes permettant de décrire les liens entre des propriétés binaires vérifiées par des sujets. Le premier provient du modèle log-linéaire utilisé en statistique et le second de l'extraction des règles d'association. Du premier, nous ne reprenons pas les tests statistiques, qui ne pourront être réutilisés pour les règles d'association, mais les exemples ont été choisis de telle façon que le lecteur féru de statistique pourra remplacer "effet nul" par "effet non significatif", et "effet important" ou "effet non nul" par "effet significatif", en prenant comme risque de première espèce α=0,05. A travers le modèle log-linéaire, nous exposons ce qu'est l'interaction, et les précautions d'interprétation des liens entre variables qui en découlent. Puis nous examinons le formalisme d'extraction automatique de règles d'association, afin de voir si les problèmes d'interprétation liés à l'existence d'interactions se retrouvent, et nécessitent alors une correction du jeu de règles extrait.

C.2 Deux variables

On dispose de 2 propriétés A et B, et N sujets numérotés de 1 à N pour lesquels on connaît l'absence (0) ou la présence (1) de chacune de ces 2 propriétés, comme indiqué dans le paragraphe 2.5.2. On reprend également de ce paragraphe le tableau de contingence 4.3 avec ses éléments a, b, c et d et on se limitera aussi dans cette partie aux tableaux de contingence n'ayant aucune

marge nulle, en se plaçant dans le cas où les variables A et B ont réellement deux modalités chacune.

C.2.1 Le modèle log-linéaire

Définitions

Si on note n_{ij} le nombre de sujets pour lesquels on a $A = i$ et $B = j$, où $i, j \in \{0; 1\}$, et si aucun des quatre effectifs a, b, c et d n'est nul, on peut écrire l'équation suivante :

$$log(n_{ij}) = \alpha + \beta i + \gamma j + \delta ij \; avec \; \alpha = log(a), \; \beta = log(c/a), \; \gamma = log(b/a), \; \delta = log(ad/bc) \quad (C.1)$$

Les logarithmes permettent de transformer les propriétés liées aux proportions en additions. Dans le cas où $\delta = 0$, on dira qu'il n'y a pas d'interaction entre A et B, ce qui correspond au cas où $a/b = c/d = (a + c)/(b + d)$, c'est-à-dire où les colonnes sont proportionnelles, et où $a/c = b/d = (a + b)/(c + d)$, c'est-à-dire où les lignes sont proportionnelles[178]. Dans ce cas d'interaction nulle, on s'intéressera aux valeurs de β et de γ, qu'on appellera respectivement l'effet de A, et l'effet de B. Si β est nul, cela signifie que c=a et b=d, donc que les lignes A0 et A1 sont identiques, et que l'effet de A est nul. Si β est positif, cela signifie que la ligne de A1 l'emporte sur celle de A0, et inversement si β est négatif. De la même façon, on examine la valeur de γ en cas d'interaction nulle, pour tirer des conclusions sur l'effet de B. Quand l'interaction n'est pas nulle, comme on n'a plus proportionnalité entre les colonnes, on peut avoir dans l'une A0 qui l'emporte sur A1, et inversement dans l'autre, aussi bien que A0 qui l'emporte sur A1 dans les 2 colonnes, ou inversement. Et de même pour les lignes. Les coefficients β et γ ne sont plus alors interprétables.

Remarque : le modèle log-linéaire testé en statistique s'écrit, d'après Morineau [181].

$$log(n_{ij}) = \alpha' + \beta'_i + \gamma'_j + \delta'_{ij} \quad (C.2)$$

avec

$$\alpha' = (log(a) + log(b) + log(c) + log(d))/4$$
$$\beta'_0 = (log(a) + log(b))/2 - \alpha' = -\beta'_1 = -((log(c) + log(d))/2 - \alpha')$$
$$\gamma'_0 = (log(a) + log(c)/2) - \alpha' = -\gamma'_1 = -((log(b) + log(d)/2) - \alpha')$$
$$\delta'_{00} = (log(a) - log(b) - log(c) + log(d))/4 = \delta'_{11} = -\delta'_{01} = -\delta'_{10}$$

Dans ce modèle, on teste la nullité des divers coefficients, l'interaction, représentée par le coefficient δ', l'effet de A par le coefficient β', et l'effet de B par le coefficient γ'. Comme nous l'avons signalé plus haut, notre but n'est pas de tester ces effets, mais de repérer si les coefficients correspondants se retrouvent dans les indices des règles d'association. On se limitera donc dans cette étude à la nullité des coefficients, et non pas à une valeur qui ne diffère pas significativement de zéro.

On peut réécrire l'équation 2 sous la forme suivante

$$log(n_{ij}) = \alpha' + \beta'_1(2i - 1) + \gamma'_1(2j - 1) + \delta'_{11}(2i - 1)(2j - 1) \quad (C.3)$$

Ce qui permet de passer de l'équation 3 à l'équation 1 en s'aidant des formules de passages suivantes : $\alpha' = (4\alpha + 2\beta + 2\gamma + \delta)/4$, $\beta'_1 = (2\beta + \delta)/4$, $\gamma'_1 = (2\gamma + \delta)/4$, $\delta'_{11} = \delta/4$.
On voit donc que la nullité de l'interaction δ'_{11} est équivalente à celle de δ, et que quand δ et δ'_1 sont nuls, la nullité de β'_1 est équivalente à celle de β, même chose pour γ.

[178] On parle également d'indépendance entre les 2 variables A et B.

Exemples sans interaction

En appliquant cette décomposition aux exemples figurant dans la table 1, on obtient les 3 équations suivantes (le logarithme choisi est le logarithme népérien) :

pas d'effet principal			
AxB	B=0	B=1	total
A=0	250	250	500
A=1	250	250	500
total	500	500	1000

1 effet principal			
AxB	B=0	B=1	total
A=0	40	460	500
A=1	40	460	500
total	80	920	1000

2 effets principaux			
AxB	B=0	B=1	total
A=0	60	540	600
A=1	40	360	400
total	100	900	1000

TAB. C.1 – Tableau de contingence de 2 variables A et B sans interaction

$$\text{tableau de gauche} \qquad \forall i,j \in \{0;1\}, \quad ln(n_{ij}) = 5.52 \tag{C.4}$$

$$\text{tableau du milieu} \qquad \forall i,j \in \{0;1\}, \quad ln(n_{ij}) = 3.69 + 2.44j \tag{C.5}$$

$$\text{tableau de droite} \qquad \forall i,j \in \{0;1\}, \quad ln(n_{ij}) = 4.09 - 0.41i + 2.20j \tag{C.6}$$

Dans le premier tableau, toutes les cellules ont un effectif identique. On appelle "modèle constant" un tel modèle. Il ne comporte que le coefficient $\alpha = ln(250) = 5.52$. L'effet de A, correspondant au coefficient β est nul, les lignes étant identiques[179], et l'effet de B, correspondant au coefficient γ est nul, les colonnes étant identiques.

Le second tableau de la table 1 contient 2 lignes identiques, indiquant que les 2 modalités de A se comportent de la même façon (pas d'effet principal pour A, le coefficient β est nul), alors que les 2 colonnes ne sont pas identiques, mais proportionnelles. Il y a un effet principal pour B, B1 étant plus fréquent que B0, représenté par le coefficient γ . En passant de la colonne B=0 à la colonne B=1, les effectifs sont passés de 40 à 460. Il ont ainsi été multipliés par 11.5, donc leur logarithme, qui était $ln(40) = 3.69$ a été augmenté de $\gamma = ln(11.5) = 2.44$ et est devenu $ln(460) = 6.13$.

Dans le dernier tableau, les lignes et les colonnes sont proportionnelles. Quand on passe de la ligne A0 à la ligne A1, les effectifs sont multipliés par 2/3, qui correspond au coefficient $\beta = ln(2/3) = -0.41$. C'est l'effet de A. Et de la colonne B0 à la colonne B1, les effectifs sont multipliés par 9, ce qui correspond au coefficient $\gamma = ln(9) = 2.20$. C'est l'effet de B.

On remarque que dans aucune des 3 équations numérotées de 4 à 6 ne figure le coefficient δ. Nous sommes dans le cas où il n'y a pas d'interaction entre les variables A et B, les lignes sont égales ou proportionnelles entre elles, ainsi que les colonnes .

Exemples d'interaction contrariante ou non

Voyons maintenant ce qui se passe en cas d'interaction. Au milieu de la table 2 figure un tableau sans interaction. Les deux autres tableaux ont été créés à partir de celui-là en gardant les marges, et en modifiant donc les 4 cellules internes du tableau en ajoutant ou en retranchant un même nombre. Pour le tableau de gauche, on a augmenté l'effectif de la première cellule (A=0 et B=0) de 70, et pour celui de droite, on a retranché 10 à cet effectif. On obtient les 3 équations suivantes pour ces tableaux

[179] Pour passer de la ligne A0 à la ligne A1, on multiplie les effectifs par le coefficient 1, donc le coefficient β, qui est son logarithme, est nul.

interaction contrariante			
AxB	B=0	B=1	total
A=0	790	110	900
A=1	10	90	100
total	800	200	1000

pas d'interaction			
AxB	B=0	B=1	total
A=0	720	180	900
A=1	80	20	100
total	800	200	1000

interaction non contrariante			
AxB	B=0	B=1	total
A=0	710	190	900
A=0	90	10	100
total	800	200	1000

TAB. C.2 – Tableau de contingence de 2 variables A et B avec interaction de divers types

$$\text{tableau du milieu} \qquad ln(n_{ij}) = 6.58 - 2.2i - 1.39j \tag{C.7}$$

$$\text{tableau de gauche} \qquad ln(n_{ij}) = 6.67 - 4.37i - 1.97j + 4.17ij \tag{C.8}$$

$$\text{tableau de droite} \qquad ln(n_{ij}) = 6.57 - 2.07i - 1.32j - 0.88ij \tag{C.9}$$

Le tableau de gauche de la table 2 a un coefficient δ non nul. On sait que les lignes et les colonnes ne sont plus proportionnelles. l'interprétation des coefficients β et γ ne peut alors plus se faire directement. En effet, si A=0, ce qui correspond à i=0, l'équation 5 devient $ln(n_{ij}) = 6.67-1.97j$, alors que si A=1, elle devient $ln(n_{ij}) = (6.67-4.37)+(-1.97+4.17)j$ soit $log(n_{ij}) = 2.30 + 2.20j$. Pour A=0, le passage de B=0 à B=1 entraîne une multiplication des effectifs par $e^{-1.97} = 0.14$, donc une diminution, et pour A1, une multiplication par $e^{2.2} = 9$, soit une augmentation. La présence de cette interaction fait que l'effet de B dans une ligne est contraire de celui de l'autre ligne. C'est pour cela qu'on l'a appelée *interaction contrariante*. Par contre, si on examine l'effet de A pour B=0 et B=1, on n'a pas de contradiction avec le coefficient β. Dans le cas où B=0, alors $log(n_{ij}) = 6.67 - 4.37i$, et dans le cas où B=1, $log(n_{ij}) = 4.70 - 0.20i$. On obtient dans les deux cas A0 l'emportant sur A1.

Dans le tableau de droite, l'interaction n'est pas contrariante, dans la mesure ou A0 continue à l'emporter sur A1, et B0 sur B1. A la lecture des équations 8 et 9 ; nous retrouvons cela : l'interaction contrarie l'effet de B dès que $\beta + \delta$ est de signe contraire à δ, donc dès que δ est de signe contraire à β, et l'emporte sur lui en valeur absolue. C'est le cas de l'effet de B dans l'équation 8, où on a $\delta = 4.17$ de signe contraire à $\beta = -4.37$ et à $\gamma = -1.97$, mais δ l'emporte en valeur absolue seulement sur γ, ce qui produit un effet de B contraire selon les lignes. Dans l'équation 9, il n'y a pas ce problème, les 3 coefficients étant de même signe.

C.2.2 Les règles d'association

On reprend les définitions de la section 2.5.2, ainsi que les indices dont les formules sont données dans cette section. On extrait les règles dont le support dépasse un seuil donné (on ne garde pas les règles de support nul). Et on classe les règles extraites selon la valeur des autres indices. Plus cette valeur est élevée, plus la qualité de la règle est grande.

On dira que la règle A0→B0 est de meilleure qualité que la règle A0→B1 pour un indice donné "ind" si l'expression $\Delta(ind) = ind(A0 \rightarrow B0) - ind(A0 \rightarrow B1)$ est positive. Nous avons établi dans l'appendice de l'annexe B que les indices se divisent en 2 groupes. Pour le premier groupe, qui contient le support, la fréquence, la confiance et l'étonnement, cette expression est de même signe que $D = a - b$, donc que l'effet de B (en fait de B0 par rapport à B1), noté $\gamma = ln(b/a)$ et pour le second, qui contient la différence, l'intérêt, la satisfaction, la nouveauté, la conviction et l'implication[180], cette expression est de même signe que $P = ad - bc$, donc que l'interaction $\delta = ln(ad/bc)$.

[180] Pour l'implication, il faut toutefois que la valeur de $\frac{ad-bc}{N}$ soit assez grande en valeur absolue.

On procédera de la même façon que dans l'annexe B pour construire un jeu de "bonnes" règles. Nous rappelons ci-dessous comment les "bonnes" règles sont choisies. Quand les indices donnent des informations contradictoires, on peut choisir de rejeter les 2 règles correspondantes, ou bien de garder celles maximisant la valeur d'un indice particulier, adapté aux besoins spécifiques de la personne qui a collecté les données. On peut aussi imposer un seuil à la confiance, de façon arbitraire ou sur des bases statistiques [44], puis regarder les valeurs les plus élevées de certains autres indices [52]. Si le seuil de confiance dépasse 0.5, la seule des deux règles qui dépasse ce seuil est gardée. Si ce seuil est plus bas, les deux règles peuvent le dépasser, et dans ce cas les autres indices peuvent intervenir dans le choix de la meilleure. Nous allons voir si l'on retrouve les interactions dans le jeu de règles obtenu selon ces techniques.

Exemples sans interaction

Dans le premier tableau de la table 1, les 4 effectifs sont égaux. Les 8 règles obtenues ont donc toutes les mêmes indices qui sont indiqués dans le tableau 2. On ne peut pas avoir un jeu de règles qui les contient toutes[181], et on n'a aucune possibilité d'en privilégier une. Le résultat est donc qu'aucune règle n'est extraite de ces données.

Partie gauche		Partie droite	Frequ ence	Conf iance	Diffé rence	Inté rêt	Convic tion	Etonne ment	Nouv eauté	Satis faction	Impli cation
A=i	→	B=j	250	0.50	0	1	1	0	0	0	0.5
B=i	→	A=j	250	0.50	0	1	1	0	0	0	0.5

TAB. C.3 – Les indices des 4 règles (A=i)→(B=j), où i, j ∈ {0 ;1} et de leurs réciproques

Le second tableau de la table 1 contient deux lignes identiques, (pas d'effet principal pour A), alors que les deux colonnes ne sont pas identiques, mais restent proportionnelles, puisqu'il n'y a pas d'interaction entre A et B. Il y a un effet principal pour B, B1 étant plus fréquent que B0. On constate dans le tableau 3 que les indices du second groupe sont constants, du fait de l'absence d'interaction, et ne pouvant choisir entre les règles 5 et 6, pas plus qu'entre les règles 7 et 8, on les rejette toutes les quatre comme précédemment. Les indices du premier groupe nous permettent de faire un choix entre la règle 1 et la règle 2, en choisissant la règle 2, et entre la règle 3 et la règle 4, en choisissant la règle 3. Ce deuxième tableau nous fournit donc un ensemble de deux règles, qui sont la règle (A=0)→(B=1) et la règle (A=1)→(B=1). Ces deux règles ont tous leurs indices identiques. Elles ne sont pas contradictoires, mais l'information qu'elles apportent à elles deux n'a guère de valeur, car l'intervention de A ne change pas le rapport des chances entre avoir B=0 et B=1. On les supprime donc et le jeu de règles est vide.

Le troisième tableau de la table 1 contient cette fois deux effets principaux, celui de A, A=0 l'emportant sur A=1, et celui de B, B=1 l'emportant sur B=0. Il n'y a toujours pas d'interaction, donc les colonnes sont proportionnelles entre elles, ainsi que les lignes. On ne recalcule pas les indices du second groupe, car ils sont égaux aux même valeurs, vu l'absence d'interaction, les indices du premier groupe figurent dans le tableau 4. On peut choisir dans chacun des 4 couples de règles contradictoires une règle car tous les indices concordent à en privilégier une des deux. On garde ainsi les règles 2, 3, 5 et 8. Comme précédemment, on ne désire pas garder simultanément la règle 2 et la règle 3, qui apportent ensemble une information de peu de valeur. On peut choisir alors la règle 2, en prenant en compte les différences de valeurs d'indices, ou bien rejeter les 2

[181] On peut difficilement concevoir un jeu de règles contenant 2 règles de partie gauche identique, et de parties droites contraires.

No	Partie gauche		Partie droite	Frequ ence	Conf iance	Diffé rence	Inté rêt	Convic tion	Etonne ment	Nouv eauté	Satis faction	Impli cation
1	A0	→	B0	40	0.08	0	1	1	-5.25	0	0	0.50
2	A0	→	B1	460	0.92	0	1	1	0.46	0	0	0.50
3	A1	→	B1	460	0.92	0	1	1	0.46	0	0	0.50
4	A1	→	B0	40	0.08	0	1	1	-5.25	0	0	0.50
5	B0	→	A0	40	0.50	0	1	1	0	0	0	0.50
6	B0	→	A1	40	0.50	0	1	1	0	0	0	0.50
7	B1	→	A1	460	0.50	0	1	1	0	0	0	0.50
8	B1	→	A0	460	0.50	0	1	1	0	0	0	0.50

TAB. C.4 – Les indices des règles pour A et B sans interaction, et un effet principal de B

No	Partie gauche		Partie droite	Sup port	Confi ance	Etonn ement
1	A0	→	B0	60	0.1	-4.80
2	A0	→	B1	540	0.9	0.53
3	A1	→	B1	360	0.9	0.36
4	A1	→	B0	40	0.1	-3.2
5	B0	→	A0	60	0.6	0.03
6	B0	→	A1	40	0.4	-0.05
7	B1	→	A1	360	0.4	-0.45
8	B1	→	A0	540	0.6	0.30

TAB. C.5 – Les indices du premier groupe pour les règles avec A et B sans interaction, et deux effets principaux

règles, vu les valeurs proches de leurs indices. Le même problème se pose pour le choix entre les règles 5 et 8. On peut ainsi se retrouver avec une seule règle, la règle 2, deux règles, la règle 2 et sa réciproque la règle 8, ou aucune règle.

Pour conclure sur ces 3 exemples de non-interaction entre A et B, on voit que l'utilisation des seuls indices du groupe 1 (ceux de l'autre groupe sont constants) ne garantit pas l'extraction d'un jeu de règles d'association cohérent. On ne s'en aperçoit pas souvent pour 2 raisons principales. Les effets combinés des variables font qu'on a rarement égalité de tous les indices, ce qui permet de ne garder que quelques règles pour 2 variables, comme on l'a proposé pour le dernier tableau. De plus, on ne s'intéresse pas toujours aux modalités 0 des variables, ce qui fait disparaître les contradictions, car on examine alors une seule règle des 8 possibles. Notamment, dans les 2 derniers tableaux, la règle 3, A1→B1 sera extraite, compte tenu de sa valeur élevée de confiance, si on ne considère pas les indices du second groupe, alors que dans le meilleur des cas elle n'apporte aucune information utile.

Interaction contrariante ou non

On vérifie alors sur le tableau 6 comportant les valeurs des indices, que les règles privilégiées dans les couples (5, 6) et (7, 8) par les indices du premier groupe sont bien les règles 5 et 8, de partie droite A0, et que les règles privilégiées dans les couples (1, 2) et (3, 4) sont bien les règles 1 et 4, de partie droite B0, les indices du second groupe étant identiques. En reprenant

No	Partie gauche		Partie droite	Frequ ence	Conf iance	Diffé rence	Inté rêt	Convic tion	Etonne ment	Nouv eauté	Satis faction	Impli cation
1	A0	→	B0	720	0.8	0	1	1	0.68	0	0	0.5
2	A0	→	B1	180	0.2	0	1	1	-2.70	0	0	0.5
3	A1	→	B1	20	0.2	0	1	1	-0.30	0	0	0.5
4	A1	→	B0	80	0.8	0	1	1	0.08	0	0	0.5
5	B0	→	A0	720	0.9	0	1	1	0.71	0	0	0.5
6	B0	→	A1	80	0.1	0	1	1	-6.40	0	0	0.5
7	B1	→	A1	20	0.1	0	1	1	-1.60	0	0	0.5
8	B1	→	A0	180	0.9	0	1	1	0.18	0	0	0.5

TAB. C.6 – Les indices des règles sans interaction de la table 2

les conclusions précédentes, on peut alors aboutir à un jeu de règles vide, si on ne se fie qu'à la confiance, ou aux seules règles A0→B0 et sa réciproque B0→A0 si on prend en compte les autres indices du groupe 1.

No	Partie gauche		Partie droite	Frequ ence	Conf iance	Diffé rence	Inté rêt	Convic tion	Etonne ment	Nouv eauté	Satis faction	Impli cation
1	A0	→	B0	790	0.88	0.08	1.10	1.64	0.85	0.07	0.39	1.00
2	A0	→	B1	110	0.12	-0.08	0.61	0.91	-3.40	-0.07	-0.10	0.00
3	A1	→	B1	90	0.90	0.70	4.50	8.00	0.40	0.07	0.88	1.00
4	A1	→	B0	10	0.10	-0.70	0.13	0.22	-0.10	-0.07	-3.50	0.00
5	B0	→	A0	790	0.99	0.09	1.10	8.00	0.87	0.07	0.88	1.00
6	B0	→	A1	10	0.01	-0.09	0.13	0.91	-7.80	-0.07	-0.10	0.00
7	B1	→	A1	90	0.45	0.35	4.50	1.64	-0.20	0.07	0.39	1.00
8	B1	→	A0	110	0.55	-0.35	0.61	0.22	0.02	-0.07	-3.5	0.00

TAB. C.7 – les indices des règles avec interaction contrariante de la table 2

Le tableau de gauche de la table 2 a une interaction qui contrarie l'effet principal de B. En effet, dans la ligne de totaux, les effectifs correspondants à B0 et B1 n'ont pas changé, celui de B0 l'emportant toujours sur celui de B1. Par contre, dans la ligne correspondant à A=1, c'est l'inverse qui se produit maintenant, B1 l'emporte sur B0 . On ne peut plus parler d'effet principal de B pour le tableau de gauche, dans la mesure où pour A=1, on obtient un effet inverse de celui obtenu pour A=0. En examinant les indices correspondant dans le tableau 7, on voit que la règle 3 l'emporte maintenant sur la règle 4, alors que dans le couple de règles (1, 2), c'est toujours la règle 1 qui l'emporte, ses différences avec la règle 2 s'étant même accentuées[182] par rapport à celles du tableau 6 correspondant à l'absence d'interaction. Pour les lignes suivantes du tableau 7, on ne voit pas de différence importante, l'interaction n'ayant pas contrarié l'effet de A. En effet, que ce soit pour B=0 ou pour B=1, l'effectif correspondant à A0 l'emporte toujours sur celui correspondant à A1. Les indices du premier groupe privilégient donc comme auparavant les règles 5 et 8. Toutefois, dans la colonne correspondant à B=0, la différence entre A0 et A1 a été accentuée alors que dans la colonne correspondant à B=1, elle a été diminuée. Ce sont les indices du second groupe qui rendent compte de cet effet, et qui privilégient la règle 5 à la règle 8. Pour conclure sur cette interaction contrariante, on peut maintenant extraire un jeu de règles qui a vraiment du sens, et qui contient la règle 1, la règle 3 et la règle 5.

[182] La ligne de totaux étant constante, le basculement d'une ligne provoque un renforcement de l'autre.

No	Partie gauche		Partie droite	Frequ ence	Conf iance	Diffé rence	Inté rêt	Convic tion	Etonne ment	Nouv eauté	Satis faction	Impli cation
1	A0	→	B0	710	0.79	-0.01	0.99	0.95	0.65	-0.01	-0.06	0.15
2	A0	→	B1	190	0.21	0.01	1.06	1.01	-2.60	0.01	0.01	0.70
3	A1	→	B1	10	0.10	-0.10	0.50	0.89	-0.40	-0.01	-0.13	0.06
4	A1	→	B0	90	0.90	0.10	1.13	2.00	0.10	0.01	0.50	1.00
5	B0	→	A0	710	0.89	-0.01	0.99	0.89	0.69	-0.01	-0.13	0.06
6	B0	→	A1	90	0.11	0.01	1.13	1.01	-6.20	0.01	0.01	0.70
7	B1	→	A1	10	0.05	-0.05	0.50	0.95	-1.80	-0.01	-0.06	0.15
8	B1	→	A0	190	0.95	0.05	1.06	2.00	0.20	0.01	0.50	1.00

TAB. C.8 – Les indices des règles avec interaction non contrariante de la table 2

On constate donc que la modification du tableau sans interaction en un tableau avec interaction contrariante a permis de produire un jeu de règles plus informatif. On peut voir sur le tableau 8 que l'ajout d'interaction non contrariante permet également de faire un choix, grâce aux indices du second groupe entre les règles 1 et 4, et les règles 5 et 8. On obtient ainsi le jeu des deux règles 4 et 8.

Conclusion sur 2 variables

En cas de non interaction, les indices du groupe 1 permettent de faire apparaître les effets principaux, s'ils existent, par sélection de 4 règles parmi les 8 possibles, mais cette information est de peu d'intérêt car les jeux de règles contiennent alors des paires de règles avec même partie gauche et partie droite contraire. Il vaut mieux alors les éliminer. En cas d'interaction, le jeu de règles produit est beaucoup plus informatif, car on peut alors choisir avec les indices du groupe 2 les règles les plus appropriées. Si l'interaction contredit les deux effets principaux, on obtient quatre règles informatives, si elle en contredit un seul, on en obtient trois, et si elle n'en contredit aucun, on en obtient deux[183].

C.3 Trois variables

Partant du tableau 1 des valeurs des N sujets pour 2 variables A et B, on ajoute une troisième variable C. Par exemple, si on reprend l'illustration de ces définitions sur les étudiants, un étudiant aura une valeur de C=1 s'il redouble, et C=0 sinon. Dans le tableau 9 figurent les 8 cas possibles selon les différentes valeurs prises par ces trois variables, et dans la colonne des effectifs le nombre de sujets correspondant à chaque cas. En posant a'+a"=a, b'+b"=b, c'+c"=c et d'+d"=d, on se retrouve dans dans le cas du tableau de contingence du paragraphe 2.5.2, exprimant la relation entre A et B sans tenir compte de C.

[183] S'il n'y avait pas d'effet principal pour A par exemple, donc des totaux égaux à 500, l'interaction produirait des effets contraires sur les deux colonnes. On dirait alors qu'elle est également contrariante.

A	B	C	Effectifs
0	0	0	a'
0	0	1	a"
0	1	0	b'
0	1	1	b"
1	0	0	c'
1	0	1	c"
1	1	0	d'
1	1	1	d"
		total	N

TAB. C.9 – Les données sous forme de répartition des N sujets selon les 8 modalités de (A,B,C)

C.3.1 Le modèle log-linéaire

Si on note n_{ijk} le nombre de sujets pour lesquels on a $A = i$, $B = j$ et $C = k$, où $i, j, k \in \{0; 1\}$, et si aucun de ces effectifs n'est nul, on peut écrire l'équation suivante[184]

$$ln(n_{ijk}) = \alpha + \beta_1 i + \beta_2 j + \beta_3 k + \gamma_1 ij + \gamma_2 ik + \gamma_3 jk + \delta ijk \qquad (C.10)$$

avec $\alpha = ln(a')$, $\beta_1 = ln(c'/a')$, $\beta_2 = ln(b'/a')$, $\beta_3 = ln(a"/a')$, $\gamma_1 = ln(a'd'/b'c')$, $\gamma_2 = ln(a'c"/a"c')$, $\gamma_3 = ln(a'b"/a"b')$, $\delta = ln(a"b'c'd"/a'b"c"d')$. Comme dans l'équation 1, on a l'effet global, α, l'effet de chaque variable prise séparément, $\beta 1$ pour A, $\beta 2$ pour B, $\beta 3$ pour C, les interactions des variables prises deux à deux, γ_1 pour l'interaction A*B, γ_2 pour A*C, γ_3 pour B*C, mais on a en plus δ qui correspond à l'interaction A*B*C entre les trois variables. Pour interpréter l'interaction de trois variables exprimée par le coefficient δ, examinons le cas où elle est nulle, cela donne $a'd'/b'c' = a"d"/b"d"$, ce qui signifie que l'écart à la proportionnalité du tableau AxB pour C=0 est le même que celui du tableau AxB pour la valeur 1 de C[185].

Dans la table 10 figure un exemple d'interaction A*B*C nulle. On a pris pour cela les effectifs successifs du tableau 9, égaux respectivement à 180, 20, 600, 200, 200, 600, 20, 180, et on a reporté les 8 effectifs à leur place dans les deux tableaux de droite, le tableau de gauche étant la somme de ceux-ci pris deux à deux. On a indiqué également sous chaque tableau le modèle log-linéaire correspondant. On voit que l'interaction des deux tableaux de droite est la même, comme attendu.

Dans la table 11 les effectifs successifs sont 10, 50, 10, 800, 100, 130, 600, 300. Dans ce cas l'interaction A*B*C n'est pas nulle, mais de coefficient δ égal à -3.7. On voit que les interactions des deux tableaux de droite ont pour différence cette valeur -3.7. Cela découle de l'égalité suivante :

$$\delta = ln(a"b'c'd"/a'b"c"d') = ln(a"d"/b"c") - ln(a'd'/b'c')$$

dans laquelle le membre de gauche exprime l'interaction des 3 variables A,B et C et le membre de droite est la différence des coefficients d'interaction des modèles des 2 variables A et B pour C=1 et pour C=0.

[184] Comme dans le modèle à deux variables, on choisit une écriture qui n'est pas celle utilisée habituellement pour tester ce modèle. Signalons toutefois que les modèles ont été testés en utilisant la procédure *catmod* avec l'instruction *loglin* de SAS, et que toutes les interactions non nulles de plus haut niveau sont très significativement non nulles (la probabilité d'avoir une valeur si importante par hasard est inférieure à 0.01).

[185] cela signifie également l'égalité des écarts à la proportion pour les tableaux AxC pour B=0 et B=1, et pour les tableaux BxC pour A=0 et A=1

AxB sans tenir compte de C			
AxB	B=0	B=1	total
A=0	200	800	1000
A=1	800	200	1000
total	1000	1000	2000
$ln(n_{ij}) = 5.3 + 1.4i + 1.4j - 2.8ij$			

AxB sachant C=0			
AxB	B=0	B=1	total
A=0	180	600	780
A=1	200	20	220
total	380	620	1000
$ln(n_{ij}) = 5.2 + 0.1i + 1.2j - 3.5ij$			

AxB sachant C=1			
AxB	B=0	B=1	total
A=0	20	200	220
A=1	600	180	780
total	620	380	1000
$ln(n_{ij}) = 3 + 3.4i + 2.3j - 3.5ij$			

TAB. C.10 – Tableaux de contingence AxB selon C pour $ln(n_{ijk}) = 5.2 + 0.1i + 1.2j - 2.2k - 3.5ij + 3.3ik + 1.1jk$

AxB sans tenir compte de C			
AxB	B=0	B=1	total
A=0	60	810	870
A=1	230	900	1130
total	290	1710	2000
$ln(n_{ij}) = 4.1 + 1.3i + 2.6j - 1.2ij$			

AxB sachant C=0			
AxB	B=0	B=1	total
A=0	10	10	20
A=1	100	600	700
total	110	620	720
$ln(n_{ij}) = 2.3 + 2.3i + 1.8ij$			

AxB sachant C=1			
AxB	B=0	B=1	total
A=0	50	800	850
A=1	130	300	430
total	180	1400	1280
$ln(n_{ij}) = 3.9 + 1.0i + 2.8j - 1.9ij$			

TAB. C.11 – Tableaux AxB selon C pour $ln(n_{ijk}) = 2.3 + 2.3i + 1.6k + 1.8ij - 1.3ik + 2.8jk - 3.7ijk$

Pour ce qui est de l'interprétation, on procède de façon descendante, comme pour le modèle à deux variables. Quand l'interaction δ est nulle, les trois interactions entre 2 variables peuvent être interprétées. Dans la table 10, elles sont égales à -3.5, 3.3, 1.1. Ce qui signifie qu'une fois C fixé, l'interaction entre A et B est de -3.5, et qu'une fois B fixé l'interaction entre A et C est 3.3, et ainsi de suite pour la dernière.

Quand l'interaction des trois variables n'est pas nulle, on ne peut se prononcer sur les interactions entre deux variables au simple vu de l'équation du modèle. Par exemple, dans la table 11, le coefficient correspondant à l'interaction A*B dans le modèle général à trois variables est $\gamma_1 = 1.8$. On ne peut toutefois pas en déduire une interaction positive entre A et B. En effet, le coefficient d'interaction entre A et B pour C=0 est positif (1.8), alors qu'il est négatif pour C=1 (-1.9).

Si on regarde maintenant le tableau de gauche de ces deux tables, où figure la liaison entre les variables A et B sans tenir compte de C, on ne voit pas apparaître de relation simple entre le coefficient de l'interaction de ce tableau, et ceux des interactions A*B avec C fixé à 0 ou 1. On peut vérifier ceci en exprimant $ad - bc = (a' + a")(d' + d") - (b' + b")(c' + c")$ sous la forme $(a'd' - b'c') + (a"d" - b"c") + (a'd" + a"d' - b'c" - b"c')$. Ce qui montre, par exemple, que si $a'd' < b'c'$ et $a"d" < b"c"$, c'est à dire que les 2 coefficients d'interaction entre A et B à C fixé sont négatifs, on peut avoir $ad > bc$, donc le coefficient d'interaction entre A et B sans tenir compte de C positif dès que $a'd" + a"d' - b'c" - b"c'$ est assez grand. Ce résultat contre-intuitif par lequel une liaison entre deux variables peut se trouver contredite par l'apparition d'une troisième variable est appelé le paradoxe de Simpson[186], comme examiné en détail en annexe B.

Pour conclure sur le modèle loglinéaire à trois variables binaires, sa formulation permet d'écrire simplement les relations de proportionnalité à trois dimensions. Toutefois, si l'interprétation de ces relations dans le cas d'une interaction non nulle entre trois variables est difficile, la plus grande simplicité en cas d'interaction A*B*C nulle n'empêche pas que l'apparition d'une troisième variable peut contredire la relation trouvée sur deux variables.

[186] Par exemple, pour la table 10, l'interaction A*B*C est nulle, l'interaction B*C est de 1.1, pour A=0 comme pour A=1, alors qu'elle est de -1.0 sans tenir compte de A

C.3.2 Les règles d'association

Les règles étant formées uniquement d'une partie gauche et d'une partie droite, l'apparition de la variable C se fait à gauche ou a droite. On s'intéresse ici à la transformation d'une règle où ne figurent que A et B en une règle ou figurent les trois variables. Les règles d'associations extraites par les algorithmes d'extraction les plus courants [13] comportent des conjonctions de propriétés en partie gauche ou droite. Nous nous limitons à ceux ayant une conjonction en partie gauche. Nous en obtenons ainsi 16, en dédoublant les 8 règles (voir celles-ci table 4) par l'ajout à gauche de C=0 ou C=1. Le problème que nous nous posons est alors le suivant : Quand une règle $(A = i) \rightarrow (B = j)$ l'emporte sur la règle contraire $(A = i) \rightarrow (B = 1 - j)$, peut-on avoir pour C=0 ou C=1 l'inverse ? La réponse est oui. Illustrons cela par un exemple avant de l'établir de façon plus générale.

Avec les effectifs suivants dans le tableau 9 : 970 20 250 250 250 30 30 200, nous obtenons les règles du tableau 12.

No	Partie gauche		Partie droite	Frequence	Confiance	Différence	Intérêt	Conviction	Etonnement	Nouveauté	Satisfaction	Implication
1	A0 C0	→	B0	970	0.80	0.16	1.3	1.8	0.57	0.10	0.44	1.00
2	A0 C0	→	B1	250	0.20	-0.16	0.6	0.8	-0.99	-0.10	-0.25	0.00
3	A0 C1	→	B0	20	0.07	-0.56	0.1	0.4	-0.18	-0.08	-1.54	0.00
4	A0 C1	→	B1	250	0.93	0.56	2.5	8.6	0.32	0.08	0.88	1.00
5	A0	→	B0	990	0.66	0.03	1.0	1.1	0.39	0.02	0.08	0.97
6	A0	→	B1	500	0.34	-0.03	0.9	1.0	-0.67	-0.02	-0.05	0.08
7	A1 C0	→	B0	250	0.89	0.26	1.4	3.4	0.17	0.04	0.71	1.00
8	A1 C0	→	B1	30	0.11	-0.26	0.3	0.7	-0.30	-0.04	-0.41	0.00
9	A1 C1	→	B0	30	0.13	-0.50	0.2	0.4	-0.13	-0.06	-1.38	0.00
10	A1 C1	→	B1	200	0.87	0.50	2.4	4.9	0.23	0.06	0.79	1.00
11	A1	→	B0	280	0.55	-0.09	0.9	0.8	0.04	-0.02	-0.24	0.00
12	A1	→	B1	230	0.45	0.09	1.2	1.2	-0.07	0.02	0.14	0.99
13	B0 C0	→	A0	970	0.80	0.05	1.1	1.2	0.48	0.03	0.20	1.00
14	B0 C0	→	A1	250	0.20	-0.05	0.8	0.9	-1.41	-0.03	-0.07	0.02
15	B0 C1	→	A0	20	0.40	-0.35	0.5	0.4	-0.01	-0.01	-1.35	0.00
16	B0 C1	→	A1	30	0.60	0.35	2.4	1.9	0.02	0.01	0.46	1.00
17	B0	→	A0	990	0.78	0.03	1.0	1.2	0.48	0.02	0.14	0.99
18	B0	→	A1	280	0.22	-0.03	0.9	1.0	-1.39	-0.02	-0.05	0.08
19	B1 C0	→	A0	250	0.89	0.15	1.2	2.4	0.15	0.02	0.58	1.00
20	B1 C0	→	A1	30	0.11	-0.15	0.4	0.8	-0.43	-0.02	-0.20	0.00
21	B1 C1	→	A0	250	0.56	-0.19	0.7	0.6	0.03	-0.04	-0.74	0.00
22	B1 C1	→	A1	200	0.44	0.19	1.7	1.3	-0.10	0.04	0.25	1.00
23	B1	→	A0	500	0.68	-0.06	0.9	0.8	0.18	-0.02	-0.24	0.00
24	B1	→	A1	230	0.32	0.06	1.2	1.1	-0.53	0.02	0.08	0.97

TAB. C.12 – Règles avec une partie contenant une modalité de A et une modalité de B, et éventuellement en partie gauche une modalité de C

Nous convenons d'abord de choisir parmi les 2 règles contradictoires successives toutes les règles ayant une plus grande valeur pour les indices des 2 groupes. Ainsi on obtient, dans l'ensemble des 6 premières règles les 3 règles A0 C0→B0, A0 C1→B1, A0→B0. Si on imagine que A0 est "être de sexe féminin", B0 est "échouer à l'examen", et C0 est "ne pas être redoublant", on voit que ces règles se traduisent respectivement ainsi : "les non redoublantes échouent à l'examen", "les redoublantes réussissent l'examen", "les filles échouent à l'examen". La troisième

règle contredit la deuxième règle. On préfère en général l'éliminer afin de garder un jeu de règles cohérent sans perdre trop d'information.

En procédant ainsi, on voit qu'il reste la 7 et la 10 (impossible de choisir entre la 11 et la 12, les indices ne s'accordant pas), la 13 et la 16 (la 17 est éliminée pour éviter la contradiction avec la 13 et la 15), la 19 (impossible de choisir entre la 21 et la 22, la 23 et la 24, les indices étant discordants). Soit un jeu de 7 règles.

Essayons de retrouver mathématiquement comment cela a pu se produire. Pour comparer les deux règles A0 C0→B0 et A0 C0→B1, au lieu de calculer pour chaque indice "ind" la différence Δ(ind)=ind(A0 C0→B0)-ind(A0 C0→B1), nous nous contentons, comme dans la partie précédente, de calculer les 2 différences D' et P' sur le tableau de contingence correspondant, qui est le tableau 13. On trouve $D' = a' - b'$ et $P' = a'(d' + b" + d") - b'(c' + a" + c") = (a'd' - b'c') + (a'b" + a'd" - a"b' - b'c")$

AxB	B0	B1	total
A0 C0 A0 C1, A1 C0, A1 C1	a' c'+a"+c"	b' d'+b"+d"	a'+b' c'+a"+c"+d'+b"+d"
total	a'+c'+a"+c"	b'+d'+b"+d"	N=a'+b'+c'+d'+a"+b"+c"+d"

TAB. C.13 – Tableau de contingence pour la règle A0 C0→ B0

Pour le tableau permettant de comparer la règle A0 C1→ B0 à la règle A0 C1→ B1, il suffit de remplacer les symboles primes par des symboles secondes et inversement. On a donc $D" = a" - b"$, et $P" = a"(d" + b' + d') - b"(c" + a' + c') = (a"d" - b"c") + (a"b' + a"d' - a'b" - b"c')$

on constate alors que

$D = a - b = (a + a") - (b + b") = (a' - b') + (a" - b") = D' + D"$

$P = ad - bc = (a'd' - b'c') + (a"d" - b"c") + (a'd" + a"d' - b'c" - b"c') = P' + P"$.

Il est tout à fait possible que D positif se décompose en un nombre positif et un nombre négatif, ainsi que pour P, comme on vient de le voir dans l'exemple, ce qui montre comment la contradiction est possible.

A l'opposé, on peut avoir un accord total, que D' et D" soient positifs, et que P' et P" le soient également. Par exemple, la règle 1, A0 C0→B0 et la 7, A1 C0→B0 s'accordent avec la règle C0→B0. Dans ce cas, les règles 1 et 7 n'apportent pas plus d'information que la règle C0→B0, et on convient de ne garder que cette dernière règle. De la même façon, la 3 et la 9 sont remplacées par la règle C1→B1 et la 13 et la 19 par C0→A0. Avec la 15, cela donne un jeu de 4 regles.

C.4 Conclusion

En nous aidant du formalisme du modèle log-linéaire, nous avons essayé de comprendre les relations entre plusieurs variables binaires et leurs conséquences sur le jeu de règles d'association qui en découle. Nous avons examiné le modèle à deux variables, puis à trois variables, et le choix des règles d'association d'après les indices. Dans le modèle à deux variables, nous avons remarqué la piètre qualité du jeu de règles trouvé en absence d'interaction, ou en cas d'interaction non "contrariante". Ce que nous avons retrouvé sous forme de redondance dans le modèle à trois variables, le jeu de règles {A0 C0→B0, A1 C0→B0, C0→B0} devant être remplacé par la règle C0→B0 . Dans ce dernier modèle sont également apparues des contradictions dues à une règle générale A0→B0 obtenue par agrégation de 2 règles partielles contradictoires A0 C0→B0 et A0

C1→B1. Ces contradictions sont apparues dans un jeu de règles sélectionnées à l'aide de leurs valeurs pour certains indices de qualité couramment utilisés. Quels que soient les indices utilisés pour construire le jeu de règles d'association, il convient de contrôler que les interactions n'ont pas produit certaines de ces contradictions et/ou ces redondances avant de le donner à interpréter aux experts du domaine dont sont issues les données.

Bibliographie

[1] Abdi, H., Introduction au traitement statistique des données expérimentales Presses Universitaires de Grenoble, 1987.

[2] R. Agrawal, R. Srikant, H, "Fast algorithms for mining association rules in large databases", Research Report RJ 9839, IBM Almaden Research Center, San Jose, California, June 1994.

[3] Agrawal, H. Mannila, R. Srikant, H. Toivonen and A. I. Verkamo, "Fast Discovery of Association Rules, in Advances in Knowledge Discovery and Data Mining", U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy editors, AAAI Press / MIT Press, Menlo Park, California, 1996, p. 307-328.

[4] Alker, H.R., traduit par Paradeise, C. "Introduction à la sociologie mathématique" Canada. Larousse Université. 1973. Chapitre 6 : Corrélation et causalité.

[5] Anderson C.J. Applied categorical data lecture notes. University of Illinois, Urbana-Champaign. 2002

[6] Aracil J., Introduction à la Dynamique des Systèmes, traduction de Ossandon M. Presses Universitairs de Lyon. Lyon 1984.

[7] Ardilly P., Les techniques de sondage Editions technip, Paris 1994

[8] Armatte M., "Robert Gibrat et la loi de l'effet proportionnel", Math. Inf. Sci. Hum., 33ème année, n°129, 1995, p5-35

[9] Azé J. et Kodratoff Y., "Évaluation de la résistance au bruit de quelques mesures d'extraction de règles d'association", Extraction de connaissances et apprentissage, Volume 1, 2002, EGC'2002, p. 143-154.

[10] Baillargeon, La régression linéaire, Edition des trois Sources, Quebec, 2000

[11] Barbut M., Monjardet B., Ordre et classification, Tome 2, Paris, Hachette, 1970

[12] Bastide Y., Data mining : algorithmes par niveau. techniques d'implantation et applications, Thèse d'informatique, Université Blaise Pascal, Clermont-Ferrand, 2000.

[13] Bastide R Y., Taouil R., Pasquier N., Stumme G., Lakhal L., "Pascal : un algorithme d'extraction des motifs fréquents" , *Technique et science informatiques*, 21(1), 2002, p. 65-75.

[14] Batola L."Statistiques et économétrie". Masson, Paris. 1983

[15] Batanero C., Godino J.D., Navarro-Pelayo V., "The use of implicative and correspondence analysis of assessing Pupils combinatorial reasoning", Actes du colloque de l'ARDM (UFM de Caen), 1995.

[16] Bavaud François, Modèles et données : une introduction à la Statistique uni-, bi- et trivariée. Paris ; Montréal (Qc) : L'Harmattan, 1998

[17] Belohlavek R. "Fuzzy Galois connections" Mathematical logic quaterly, 45, p. 497-504, 1999.

[18] Ben Naceur-Mourali, Christophe Gonzales Une unification des algorithmes d'inférence de Pearl et de Jensen revue d'intelligence artificielle, RSTI série RIA, Vol 18, no 2/2004 Lavoisier, Paris, 2004, p. 229-260

[19] Benzécri J.-P. & F. & coll., Pratique de l'analyse des données, Tome 5, Paris, Dunod, 1970.

[20] Bertier P., Bourroche J.M., "analyse des données multidimensionnelles", Presses Universitaires de France, 1975.

[21] Bertail P. "Le bootstrap : une revue de la littérature", document de travail INSEE n°9201, 1991

[22] Besse P., Stabilité de l'Analyse en composantes principales par Ré-échantillonage, Approximation par la théorie des perturbation. Document N°05-89 du laboratoire de Statistique et Probabilités de l'université Paul Sabatier de Toulouse 1989

[23] Birkhoff, G. "Lattice theory", American Mathematical Society colloquium publications volume 25. New York, 1948.

[24] Blancheteau M., Magnan A., Psychologie expérimentale et psychologie du développement Hommage à César Florès. L'Harmattan, Paris, 1994

[25] Botta M., Boulicaut J.-F., Masson C., Meo R. : A Comparison between Query Languages for the Extraction of Association Rules. DaWaK 2002 p. 1-10

[26] Boudon R. L'analyse mathématique des faits sociaux. Plon 1967. Paris

[27] Boudjlida N. "Bases de données et systèmes d'informations. Le modèle relationnel : langages, systèmes et méthodes", Dunod, Paris. 1999

[28] Boulicaut J-F., A. Bykowski, and C. Rigotti. Approximation of frequency queries by means of free-sets. In : Proceedings of the Fourth European Conference on Principles and Practice of Knowledge Discovery in Databases PKDD'00, Lyon (F), September 13-16, 2000.Springer-Verlag LNAI 1910, pp. 75-85.

[29] Boulicaut J.-F., "From KDD scenario description to data mining qualitative benchmarks", Réunion de travail sur les scénarios prototypiques de l'Action Spécifique Discovery Challenge. http ://users.info.unicaen.fr/ bruno/asdisco

[30] Brandes U., Gaertler M., and Wagner D. , Experiments on Graph Clustering Algorithms. Proc. 11th Europ. Symp. Algorithms (ESA '03), Springer LNCS.(2003)

[31] Breiman L. Bagging predictors. Machine Learning 24 p. 123-140 1996

[32] Breslau L., Cao P., Fan L., Phillips G., Shenker S., « Web Caching and Zipf-like Distributions : Evidence and Implications », Proceedings of IEEE Infocom, NewYork, Mars 1999, p. 126-134.

[33] Brin S., Motwani R., Ullman J.D., Tsur S. Dynamic itemset counting and implication rules for market basket data. Proceedings ACM SIGMOD p255-264, 1997

[34] Brijs T., Vanhoof K., Wets G. "Reducing Redundancy in Characteristic Rule Discovery by Using IP-Techniques"

[35] Buntine W.L., Variational extensions to EM and multinomial PCA. In 13th European Conference on Machine Learning (ECML'02), Helsinki, Finland, 2002.

[36] Cadot M., Cuxac P., François C., (2006) Règles d'association avec une prémisse composée : mesure du gain d'information. EGC 2006 : p. 599-600

[37] Cuxac P., Cadot M., François C., (2005) Analyse comparative de classifications : apport des règles d'association floues. *EGC 2005* : p. 519-530

[38] Cadot, M., Maj, J.-B., Ziadé T., (2005) Association Rules and Statistics, dans *Encyclopedia of Data Warehousing and Mining*, Edité par John Wang, Montclair State University, USA, p. 94-98

[39] Cadot M. , (2005) "A Simulation Technique for extracting Robust Association Rules", CSDA 2005 (Chypre)

[40] Cadot, M. and di Martino, J. (2004) A Data Cleaning Solution by Perl Scripts for the KDD Cup 2003 Task 2. , revue SIGKDD Explorations. 2003. vol. 5. n° 2. pp.154-155

[41] Cadot M., Napoli A., (2003), "Règles d'association et interaction entre variables binaires", *SFC2003, Neuchatel, 10-12 septembre 2003* - Proceedings pp 87-90, 2003.

[42] Cadot M., Napoli A., 2003, *Règles d'association et "Paradoxe de Simpson*, rapport interne Loria, octobre 2003

[43] Cadot M., Napoli A., 2003, "Une optimisation de l'extraction d'un jeu de règles s'appuyant sur les caractéristiques statistiques des données", *RSTI-RIA-ECA*-16/2003 pp. 631-656

[44] Cadot M., Napoli A., et Nahama-Fourguette, V., (2003), "Comparaison de deux techniques d'extraction automatique de règles dans les bases de données. Illustration sur des données issues d'un questionnaire sur les peurs." LORIA A03-R-052.

[45] Cadot M., Napoli A., 2003, "Une optimisation de l'extraction d'un jeu de règles s'appuyant sur les caractéristiques statistiques des données", Rapport de Recherche LORIA numéro A02-R-162, Nancy, décembre 2002

[46] Cadot M., "Mathématiques et Psychologie", Atelier du colloque de la commission inter Irem, Mathématiques, sciences économiques et sociales, Dijon 1998. p. 49-55.

[47] Cadot, M., Vugdalic, A., Statistiques pour les psychologues, cours polycopié de Deug1 de psychologie de l'Université de Reims. 1997. Reims

[48] Cadot, M.. *Modélisation mathématique des cohortes scolaires* , Mémoire de DEA d'Analyses économique et politique, Université de Dijon, 1995

[49] Carrez C. "Des structures aux bases de données" Bordas, Paris 1990

[50] Groupe Chadule, "Initiation aux méthodes statistiques en géographie" Masson, Paris 1974

[51] Chen Q. "Mining Exceptions and Quantitative Association Rules in Olap Data Cube" Thèse, Simon Fraser University, 1999

[52] Cherfi H. et Toussaint Y., "Adéquation d'indices statistiques à l'interprétation de règles d'association", Actes des 6èmes Journées internationales d'Analyse statistique des Données Textuelles, *JADT'02*, Saint-Malo, IRISA-INRIA Vol. 1, p. 233-244, 2002.

[53] Clerc M., L'optimisation par essaims particulaires, versions paramétriques et adaptatives Lavoisier. Paris, 2005

[54] Cochran W.G. et Snedecor G.W., Méthodes Statistiques, traduit par une association de coordination technique agricole, Paris. 1966

[55] Coombs C.H., Dawes R.M., Tversky A. Psychologie mathématique. Tome 1 : Modèles et processus de décision traduit par J.-P et R. Poitou, Presses Universitaires de France, Paris, 1973.

[56] Cornuéjols A., Miclet L., Apprentissage artificiel, concepts et algorithmes. Eyrolles, Paris., 2002

[57] Cristofor L., Simovici D. (2002), "Generating an informative cover for association rules", ICDM 2002, Proceedings pp. 597-600. Japan, December 2002.

[58] Crucianu Michel, Jean-Pierre Asselin de Beauville , Romuald Boné - Méthodes factorielles pour l'analyse des données : méthodes linéaires et extensions non-linéaires. Hermès - Lavoisier, 2004

[59] Dagnelie P.,*Théorie et méthode statistiques*, Grenoble, Presses Agronomiques de Gembloux, 1970.

[60] Dagnelie P., Principes d'expérimentation : planification des expériences et analyse de leurs résultats. Grenoble, Presses Agronomiques de Gembloux, Edition électronique : http ://www.dagnelie.be 2003.

[61] Davey B.A., Priestley H.A., "Introduction to Lattices and Order", Cambridge University Press, 1990.

[62] Dekang L., "An information theoretic definition of similarity". Proceedings of ICML'98, p. 296-304, 1998

[63] Demeuse M., Les échelles de mesure : Thurstone, Likert, Guttman et le modèle de Rasch. Publications du service de pédagogie de l'Université de Liège, Note technique D/2000/8794/2. Liège, 2000

[64] Diatta J. (2003), "Génération de la base de Guigues-Duquenne-Luxenburger pour les règles d'association par une approche utilisant les mesures de similarité multivoies", CAP2003 Laval, Proceedings pp. 281-297

[65] Dickes, P., Tournois, J., Flieller, A., Kop, J.-L., "La psychométrie : théories et méthodes de la mesure en psychologie" Presses universitaires de France. Paris, 1994

[66] Diday E., "Symbolic Data Analysis end the SODAS Project : Purpose, History, Perspectives". dans : Analysis of symbolic data, H.-H Bock et E. Diday editors, Springer-Velrlag, Berlin, 2000. P. 1-23

[67] Dor E., "Econométrie" Pearson Education, Collection synthex. Paris, 2004

[68] Droesbeke J.-J., Fichet B., Tassi P., éditeurs. Analyse statistique des durées de vie. Journées d'Etude en Statistiques de l'Association pour la Statistique et ses Utilisations, Edition de l'Université de Bruxelles, Ellipses,.1988

[69] Droesbeke J.-J., Fichet B., Tassi P., éditeurs. Séries chronologiques : théorie et pratique des modèles ARIMA Journées d'Etude en Statistiques de l'Association pour la Statistique et ses Utilisations, Edition de l'Université de Bruxelles, Ellipses,.1988

[70] Droesbeke J.-J., Fine J., éditeurs. 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,.1996

[71] Dubois D., Prade H., *Théorie des possibilités*, Paris, Masson, 1988.

[72] Dubois D., Prade H., What are fuzzy rules and how to use them. Fuzzy Sets and Systems, n°84, pp 169-185.

[73] Edgington, E.S., Randomization tests. Marcel Decker, New York, USA Statistics : Textbooks and Monographs 1995

[74] Efron, B. Jolivet E. Hourdan R. traduction de Yahi N. et Saporta G., Le bootstrap et ses applications, discrimination et régression. CISIA Saint-Mandé, 1995

[75] El-Bèze M., J.-M. Torres-Moreno, F. Béchet, Peut-on rendre automatiquement à César ce qui lui appartient ? Application au jeu du Chirand-Mitterrac. TALN'2005 / DÉfi Fouille de Textes, Dourdan, 10 juin 2005,

[76] Fabris C.C., A.A. Freitas. Discovery of surprising patterns by detecting occurences of Simpson's paradox. Research and developpement in intelligent systems XVI (Proc ES99. The 19th SGES Int. Conf. of Knowledge-based systems and applied artificial intelligence) p148-160. Springer-Verlag, 1999.

[77] Faure R., Heurgon E., "Structure ordonnées et algèbres de Boole", Gauthier-Villars, Paris, 1971

[78] Fayyad U.M., C.C., Piatetsky-Shapiro G.,Smyth P., Uthurusamy R., *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press, 1996.

[79] Ferré L. "Systèmes d'Information Logiques : un paradigme logico-contextuel pour interroger, naviguer et apprendre", Thèse, Rennes 1, 2002

[80] Ferré S., Jouve B. Partionnement d'une classe de graphes orientés. Mathématiques et Sciences Humaines, N° 158, Été 2002

[81] Fidelis M.V., Lopes H.S., Freitas A.A., Discovering comprehensible classification rules with a genetic algorithm. Proc. Congress on evolutionary computation (CEC-2000) pp805-810. La Jolla, CA, USA, 2000.

[82] Fisher, R.A. The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics, 7 : 179-188 (1936)

[83] Le Floc'h A., Fisette C., Missaoui R., Valtchev P., Godin R., "JEN : un algorithme efficace de construction de générateurs pour l'identification des règles d'association", Journées de statistiques juin 2003, Lyon

[84] Foucart T., L'analyse des données, Presses Universitaires de Rennes, collection Didact Sciences, Rennes 1997

[85] Francisci D., Brisson L., Collard M., "A scalar evolutionnary approach to rule extraction", Rapport de Recherche ISRN I3S/RR-2003-12-FR, 2003

[86] Freitas A.A. On rule interestingness measures. Knowledge-Based Systems journal 12 (5-6), 309-315. Oct. 1999.

[87] Freitas A.A., Understanding the Crucial Role of Attribute Interaction in Data Mining. Artificial Intelligence Review 16(3), Nov. 2001, pp. 177-199.

[88] Freund Y. et Shapire R.E. A decision theoretic generalization of on-line learning and an application to boosting. In Proceedings of the second European Conference on Computational Learning Theory, P. 23-37. Springer-Verlag 1995

[89] Frome, P.M., Eccles, J.F., "Parents' Influence on Children's Achievement-Related Perceptions". Journal of Personality and social Psychology 1998, Vol 74, No 2,435-452

[90] Dazy F., Le Barzic J.-F., "L'analyse des données évolutives, méthodes et applications". Editions Technip, Paris. 1996

[91] Fu Y., "Discovery of multiple-Level Rules from large Databases, Thèse, Simon Fraser University, 1996

[92] Glymour C., Madigan D., Pregibon D., Smyth P., Statistical theme and lessons for Data Mining. Data Mining and Knowledge Discovery 1 (1) p11-28. 1997

[93] Godement R., "Cours d'algèbre", Hermann, Paris, 1966

[94] Godin R., Mineau G., Missaoui R., Mili H., "Méthodes de classification conceptuelle basées sur les treillis de Galois et applications", Revue d'Intelligence Artificielle, 1995, 9(2), p.105-137

[95] Godin R., Missaoui R., "An incremental Concept Formation Approach for learning from databases", Theoretical Computer Science, Volume 133, p. 387-419, 1994

[96] Good P., Permutation tests, A practical Guide. Resampling Methods for testing hypotheses, Springer, New York, USA Springer Series in statictics 2000

[97] Gonzales C., Les réseaux bayésiens, Revue d'intelligence artificielle, RSTI série RIA, Vol 18, no 2/2004 Lavoisier, Paris, 2004

[98] Govaert G., "Analyse des données", Lavoisier, Hermès Sciences. 2003

[99] Gras R., Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques , Thèse soutenue à l'Université de Rennes I, 1979.

[100] Gras R., Bailleul M., "La fouille dans les données par la méthode d'analyse implicative statistique", Journées du 23 et 24 juin 2000 organisées par l'IUFM de Caen et l'ARDM.

[101] Gras R., Méthodes d'analyses statistiques multidimensionnelles en didactique des mathématiques. Actes du colloque ARDM de Caen 27 - 29 janvier 1995 - publié par l'ARDM

[102] Gras R. et collaborateurs, *L'implication statistique, une nouvelle méthode exploratoire de données*, La pensée sauvage, Grenoble, 1996.

[103] Gras R., P. Kuntz, R. Couturier, F. Guillet, "Une version entropique de l'intensité d'implication pour les corpus volumineux", Extraction des connaissances et apprentissage, 2001, Volume 1, Numéro 1-2, p.69-80.

[104] Guermeur Y., Paugam-Moisy H., Théorie de l'apprentissage de Vapnik et SVM, Support Vector Machines. READ, Vol. 3, N. 1, 17-38. (1999).

[105] Guigues J.L. et Duquenne V. (1986) "Familles minimales d'implications informatives résultant d'un tableau de données binaires", Math. Sci. Hum. n°95, pp. 5-18

[106] Guillaume S. (2000) "Traitement des données volumineuses, mesures et algorithmes d'extraction de règles d'association et règles ordinales", Thèse en informatique, Nantes, déc. 2000.

[107] Guillet F. (2004) "Mesure de qualité des connaissances en ECD", Cours donné lors des journées EGC 2004, Clermont-Ferrand, 20 janvier 2004.

[108] Guiraud, P., Problèmes et méthodes de la statistique linguistique, Paris, Presses Universitaires de France, 1960.

[109] Guyon X., Statistiques et économétrie, du modèle linéaire aux modèles non linéaires. Ellipses, Paris. 2001

[110] Guyon Isabelle, André Elisseeff. "An introduction to variable and feature selection" Special issue on variable and feature selection, I. Guyon, A. Elisseeff, editors. Journal of Machine Learning Research, Volume 3 (Mar). Pages 1157-1182 2003

[111] Haining R., Spatial data analysis in the social and environmental science. Cambridge University Press. 1990

[112] Hambleton R.K. et Swaminathan H., Item Response Theory : Principles and applications. Kluwer. Dordrecht. 1986

[113] Han J. and Kamber M., Data Mining : Concepts and Techniques, Morgan Kaufmann Publishers, San Francisco, 2001.

[114] Han J., J. Pei, Y. Yin et R. Mao, Mining Frequent Patterns without Candidate Generation : A Frequent-Pattern Tree Approach. Data Mining and Knowledge Discovery, 8, pp. 53-87, 2004,

[115] Harman Harry H. Modern Factor Analysis.University of Chicago Press, Chicago, 1974

[116] Hastie T., Tibshirani R., Friedman J., The Elements of Statistical Learning Data Mining, Inference and Prediction. Springer Series in Statistics, Canada. 2002

[117] Hérault J., Jutten C. et Ans B., Détection de grandeurs primitives dans un message composite par une architecture de calcul neuromimétique en apprentissage non supervisé. Actes du Xème colloque GRETSI , Nice, France, pages 1017-1022, Mai, 1985

[118] Herman J., "Analyse de données qualitatives", tome 2 : "traitement d'enquêtes, modèles multivariés", collection méthodes +programmes, Masson, Paris, 1990

[119] Hilderman R.J., Hamilton H.J., 1999 Knowledge Discovery and Interestingness Measures : A Survey

[120] Hoc J.-M., L'analyse planifiée des données en psychologie. PUF Paris 1983.

[121] Hoel P., Statistiques mathématiques, Colin Paris. 1991.

[122] Hofmann Thomas, "Learning the Similarity of Documents : an information-geometric approach to document retrieval and categorization", in Advances in Neural Information Processing Systems 12, pp-914-920, MIT Press, 2000.

[123] Holte R.C., Acker L.E., Porter B.W., Concept Learning and the problem of small disjuncts. Proc. Int. Joint Conf. AI (IJCai-89) p. 813-818

[124] Holzinger Karl J. and Harry H. Harman. Factor Analysis. A Synthesis of Factorial Methods. University of Chicago Press, Chicago, 1941.

[125] Hotelling H., Analysis of a complex of statistical variables into principal components. The Journal of educational psychology, vol. 24, p. 417-441 et 498-520. 1933

[126] Howel D.C., Statistical Methods for Psychology, Duxbury, A Division of International Thomson Publishing Inc., 1997

[127] Hoyle R.H. Structural equation modeling. Concepts, Issues and Applications. Sage, London, 1995

[128] Hyvärinen A., Survey on Independent Component Analysis. Neural Computing Surveys 2 :94–128, 1999

[129] Jakulin A., "Attribute Interactions in Machine Learning" Master's thesis University of Ljubljana, Slovenija 2003

[130] Jensen D., Induction with Randomization Testing :Decision-Oriented Analysis of Large Data Sets. these mai 1992. Saint Louis, Missouri.

[131] Jensen D., Multiples comparisons in induction algorithms. Kluwer Academic Publishers, 1998 Boston p1-33

[132] Jensen D., Neville J., Linkage and autocorrelation Cause Feature Selection Bias in relational learning, Proceedings of the 19th international conference on machine learning ICML2002 Morgan-Kaufmann p259-266

[133] Jensen F.V., An introduction to Bayesian Networks University College London. 1995

[134] Johnston, J., Econometric Methods, MacGraw-Hill NY, USA, 1972

[135] Joreskog K.G et Sorbom D. Lisrel VI, user's guide, 3ème édition, Mooresville, IN : Scientific Software. 1984

[136] Jutten C., Herault J., Une solution neuromimétique au problème de séparation de sources Traitement du Signal [Trait. Signal], Vol. 5, N° 6-NS, p. 389-403. 1988

[137] Kalos, M.H., Whitlock P.A., Monte Carlo Methods. John Wiley & Sons. NY, USA, 1986, Vol. I

[138] Kamber M. et Shingal R., Evaluating the Interestingness of characteristic rules. Proc. second Int. Conf. Knowledge Discovery and Data Mining, p263-266 AAAI, 1996

[139] Kerre, Etienne E., "A comparative study of the behavior of some popular fuzzy implication operators on the generalized modus ponens", pp 281-295, in Zadeh, L.A., Kacprzyk J. "Fuzzy Logic for the management of uncertainty", John Wiley & sons, New York, 1992

[140] Kira K., Rendell L.A., A pratical approach to feature selection. Machine learning Proceedings of the international Conference ICML'92 p. 249-256, Morgan Kaufmann.1992

[141] Kline P., An easy guide to factor analysis. Londres. Routledge. 1994

[142] Kodratoff Y., "Rating the Interest of Rules Induced from Data and within texts ", *12th IEEE -International Conference on Database and Expert Systems Aplications-Dexa 2001*, Munich, sept 2001.

[143] Kononenko Igor, Marko Robnik-Sikonja, Uros Pompe, ReliefF for estimation and discretization of attributes in classification, regression and ILP problems. In A. Ramsay (ed.) : Artificial Intelligence : Methodology, Systems, Applications : Proceddings of AIMSA'96, pp.31-40, IOS Press, 1996

[144] Kohonen T., Self-Organizing Maps. Springer Series in Information Sciences, Vol. 30, Springer, Berlin, Heidelberg, New York, 1995,

[145] Igor Kononenko, Edvard Simec, Marko Robnik-Sikonja. Overcoming the myopia of inductive learning algorithms with ReliefF. Applied Intelligence, 7 :39-55 1997,

[146] Kruskal J.B., Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis Psychometrika, 29, 1964, p 1-27

[147] Kullback S., Leibler R.A., On Information and Sufficiency, Annals of Mathematical Statistics, 22, p79-86, 1951

[148] Kuntz P., Guillet F., Lehn R., Briand H. "A User-Driven Process for Mining Association Rules". PKDD 2000. Proceedings pp. 483-489

[149] Larher A., Implication statistique et applications à l'analyse de démarches de preuve mathématiques, Thèse de Mathématiques et application, Université de RennesI, Rennes, 1991.

[150] Larose D.T., Des données à la connaissance, une introduction au data mining. Traduction et adaptation de Vallaud T., Vuibert, Paris. 2005

[151] Laveaux D., Grégoire J., "Introduction aux théories des tests en psychologie et sciences de l'éducation", Bruxelles, 2002, De Boeck.

[152] Lavrac N., Flach P.A., Zupan B., "Rule Evaluation Measures : A Unifying View" ILP-99, Proceedings pp. 174-185.

[153] Le Floc'h A., Fisette C., Missaoui R., Valtchev P., Godin R., "JEN : un algorithme efficace de construction de générateurs pour l'identification des règles d'association". Journées de statistiques, juin 2003, Lyon

[154] LeCun Y., Modèles connexionnistes de l'apprentissage (connectionist learning models). PhD thesis, Université P. et M. Curie (Paris 6), June 1987.

[155] Lee D.D.and Seung H. S., Learning the parts of objects by nonnegative matrix factorization, Nature, 401, 788791. (1999)

[156] Lefébure R., Venturi G., Data Mining, Gestion de la relation client, Personnalisation de sites web Eyrolles, Paris, 2001

[157] Legendre P., Legendre L., Numerical Ecology Elsevier, Amsterdam, The Netherlands Developments in Environmental Modelling. 1998

[158] Lelu A., "Local Component Analysis : a neural model for information retrieval" - International Joint Conference on Neural Networks - Washington, pp.II.43-48, IEEE, 1989

[159] Lelu A., A. Georgel. "Neural models for orthogonal and oblique factor analyses : Towards dynamic data analysis of large sets of highly multidimensional objects" - Actes d'ICNN90 (Paris), pp.829-832, Kluwer, Dordrecht, 1990

[160] Lelu A., Martine Cadot, Sylvain Aubin. Coopération multiniveau d'approches nonsupervisées et supervisées pour la détection des ruptures thématiques dans les discours présidentiels français. Semaine du Document Numérique 2006 / 2ème DÉfi Fouille de Textes. Fribourg, Suisse. 21-22 septembre 2006,

[161] Lelu, A. ; Cuxac, P ;. Cadot, M. Document stream clustering : experimental an incremental algorithm and. AR-based tools for highlighting dynamic trends. COLLNET 2006 Nancy. 10-12 mai 2006

[162] Lenca P., Meyer P., Picouet P., Vaillant B., Lallich S., "Critères d'évaluation des mesures de qualité en ECD, JS 2003, Proceedings pp. 647-650, Lyon, 2003.

[163] Léon A., J. Cambon, M. Lumbroso, F. Winnykamen. "Manuel de psychopédagogie expérimentale", Paris 1977, Presses Universitaires de France

[164] Leonard M., cours, Université de Genève, 1994

[165] Leray P., Francois O., Réseaux bayésiens pour la classification, Méthodologie et illustration dans le cadre du diagnostic médical, Revue d'intelligence artificielle, RSTI série RIA, Vol 18, no 2/2004, Lavoisier, Paris, 2004, p. 168-193

[166] Lerman I.C., "Rôle de l'inférence statistique dans une approche de l'analyse classificatoire des données", Méthodes d'analyses statistiques multidimensionnelles en didactique des mathématiques. IUFM de Caen, 1995

[167] Lindley D. V. et Novick M.R., The role of exchangeability in Inference. Annals of Statistics 9, p45-48. 1981

[168] Liu B., Hsu W., Ma Y., Integrating classification and Association rule mining. Proc. 4th Int. Conf. on Knowledge Discovery an Data Mining (KDD-98) p80-86 AAAI Press 1998

[169] Liu B., Hsu W., Ma Y., Mining association rules with multiple supports. Proc. 5th ACM SIGKDD Int. Conf. on Knowledge Discovery an Data Mining (KDD-99) p125-134 1999

[170] Luxenburger M., "Implications partielles dans un contexte", Mathématiques, Informatique et Sciences humaines, n°113, pp. 35-55 1991

[171] Major J et Mangano J., Selection among rules induced from a hurricane database. The Journal of Intelligent Information Systems, 4, 1995, p39-52

[172] Mandelbrot, B. "An information theory of the statistical structure of language" W.E. Jackson (ed.), Communication Theory, Acad. Press. 1953. pp. 503-512

[173] Manly B.F.J., Randomization, Bootstrap and Monte Carlo methods in Biology. Chapman & Hall/CRC, Boca Raton, Florida, USA. Texts in Statistical Science, 1997

[174] Masson, A.M., Cadot, M., Ansseau, M., Perfectionnisme : effets du sexe et de l'échec. Revue l'Encéphale, Paris, vol. XXIX, p125-135. 2003

[175] Masson, A.M., Cadot M., Pereira A.M., Depreeuw E., Ansseau, M., Version francophone du TASTE (test for ability to study and evaluation). L'Encéphale, Paris, 2001 vol27. p.527-538

[176] Maurin J., "Simuation déterministe du hasard". Masson, Paris, France 1975

[177] Mephu Nguifo E. et Njiwoua P., "Feature Extraction, Construction and Selection : A Data Mining Perspective", vol 453, Chapter Using Lattice-based Framework as a Tool for Feature Extraction. Kluwer Academic Publishers, Boston, 1998.

[178] Meyn S.P., Tweedie R.L., "Markov Chains and Stochastic Stability". Springer-Verlag, London, UK. 1993

[179] Mielikainen T. and Mannila H., "The Pattern Ordering Problem". Knowledge Discovery in Databases : PKDD 2003, 7th European Conference on Principles and Practice of Knowledge Discovery in Databases, Cavtat-Dubrovnik, Croatia, September 22-26, 2003, Proceedings. Lecture Notes in Computer Science 2838, Springer, 2003. p. 327-338

[180] Mooney C.Z. et Duval R., Bootstrapping, : A Nonparametric Approach to Statistical Inference. Sage Publications, Series : Quantitative Applications in the Social Sciences. London. No 95. 1993

[181] Morineau, A., Nakache, J.-P., Krzyzanowski, C., *Le modèle log-linéaire et ses applications*, Cisia-Ceresta, Paris 1996.

[182] Morineau, A. (éd.) *Aide-mémoire statistique*. Saint-Mandé : CISIA-CERESTA, 1995

[183] Morin H., *Théorie de l'échantillonnage*. Les presses de l'université Laval, 1993

[184] Nakache, J.-P., Confais J., Statistique explicative appliquée : analyse discriminante, modele logistique, segmentation. Editions Technip, Paris, France 2003

[185] Njiwoua Njamen G.P., "Contribution à l'apprentissage symbolique automatique par l'usage du treillis de Galois". these Université d'Artois, Lens 2000

[186] Nolt J., Rohatyn D., Varzi A., *Logic*. Schaum's outline series, McGraw-Hill, 1998.

[187] Oates T. et Jensen D., "Large Datasets Lead to Overlay Complex Models : an Explanation and a Solution", 4th International Conference on Knowledge Discovery and Data Mining, sept 1998, Proceedings p.294-298.

[188] Oja E.A., "A simplified neuron model as a principal components analyzer", Journal of Mathematical. Biology, Vol. 15, pp. 267-273, 1982

[189] Oja, E., Ogawa, H., and Wangviwattana, J., Learning in nonlinear constrained Hebbian networks. In T. Kohonen, K. Mäkisara, O. Simula, and J. Kangas (Eds.), Artificial Neural Networks. Amsterdam : Elsevier, pp. 385 390 (1991)

[190] Padmanbhan B., Thuzhilin A., "Small is beautiful : Discovering the Minimal Set of Unexpected Patterns".

[191] Pasquier N., "Data Mining : Algorithmes d'Extraction et de Réduction des Règles d'Association dans les Bases de Données" Thèse, Université de Clermont-Ferrand II, 2000

[192] Pearl J., "Causality models, reasoning, and inference", Cambridge University Press, 2000, 267 - 279.

[193] Pearson K., On lines and planes of closest fit to systems of points in space. Phil. Mag, n° 2 (6ème série), p. 559-572. 1901

[194] Petit J.-M., "Découverte d'implications antre gènes à partir des données de biopuces". Séminaire IRISA, Rennes, janvier 2003.

[195] Piatetsky-Shapiro G., "Discovery, Analysis, and presentation of strong rules", Knowledge Discovery in Databases, Piatetsky-Shapiro G., Frawley W.J. (eds). AAAI/MIT Press, 1991, p. 229-248.

[196] James S. Press, The role of Bayesian and frequentist multivariate modeling in statistical Data Mining, dans "Statistical Data Mining and Knowledge Discovery", H. Bozdogan, Chapman & Hall/CRC, Boca Raton, US, 2004

[197] Popper K.R., "The logic of scientific discovery" (1935) London. Hutchinson. 1972

[198] Rakotomalala R., La distribution théorique des séquences. Technical Report, laboratoire Eric, Université Lumière, Lyon 2, 1995.

[199] Rasiowa H., "Toward fuzzy logic", pp 5-25, in Zadeh, L.A., Kacprzyk J. "Fuzzy Logic for the management of uncertainty", John Wiley & sons, New York, 1992

[200] Rasiowa, H., Cat Ho, N., "LT-fuzzy logic", pp 121-139, in Zadeh, L.A., Kacprzyk J. "Fuzzy Logic for the management of uncertainty", John Wiley & sons, New York, 1992

[201] Reuchlin, M., "Précis de statistique". Paris. Presses Universitaires de France. 1976

[202] Reuchlin, M., Bacher, F., "Les différences individuelles dans le développement cognitif de l'enfant". Paris. Presses Universitaires de France. 1989 Appendice méthodologique p. 231-293

[203] Rosenblatt D., A. Lelu, A. Georgel, "Learning in a single pass : A neural model for instantaneous principal component analysis and linear regression". Actes de la First IEE Conference on Neural Computing, pp.252-256, IEE, Londres, 1989

[204] Rosenblatt, Frank, The Perceptron : A Probabilistic Model for Information Storage and Organization in the Brain, Cornell Aeronautical Laboratory, Psychological Review, v65, No. 6, pp. 386-408. 1958

[205] Rouanet H, Le Roux B., "Analyse des données multidimensionnelles", Dunod, Paris, 1983

[206] Rouanet H, Bernard J.M., Le Roux B., Statistiques en sciences hmaines : l'analyse inductive des données. Dunod, Paris, 1990

[207] Roussel P., Durrieu F., Campoy E., El Akremi A., Méthodes d'Equations Structurelles : Recherche et Applications en Gestion. Economica. Paris 2002

[208] Rousseau R. La statistique descriptive et ses applications en éducation et en psychologie. Les presses de l'Université de Laval, Quebec, 1971

[209] Rubinstein R.Y., Simulation and Monte Carlo Method. Wiley-Interscience, NY, US. 1981

[210] Saporta G., Probabilités, analyse des données et statistique,. Technip, Paris, 1990.

[211] Savasere A., Omiecinski E., Navathe S., An efficient algorithm for mining association rules in larges databases. VLDB95, p432-444. Morgan Kaufmann, 1995

[212] Schwartz , Daniel, Méthodes statistiques à l'usage des médecins et des biologistes Paris , Flammarion, 1991

[213] Schmacker R.E., Lomax R.G., A beginner's guide to structural equation modeling, Lawrence Erlbaum Associates Publishers, 1996, Mahwah, New Jersey

[214] Siegel S., Castellan N.J. Jr., Nonparametric statistics for the behavorial sciences. McGraw-Hill 1988

[215] Sikonja M.R., Theorical and empirical analysis of ReliefF and RReliefF. Machine learning, 2002, p36-63

[216] Silberschatz, A., Tuzhilin, A., On Subjective Measures of Interestingness in Knowledge Discovery. KDD 1995 : 275-281

[217] Simpson, E. H., "The interpretation of interaction in contingency tables", Journal of the Royal Statistical Society, Series B, 13, 238-241, 1951.

[218] Smyth P. et Goodman R.M., Rule Induction using information theory, dans Piatetsky-Shapiro G. et Frawley W.J. editors, Knowledge Discovery in Databases, AAAI/MIT Press, p 229-238, 1991

[219] Spearman, C. E., General intelligence objectively determined and measured. American Journal of Psychology, 5, 201-293.1904

[220] Sprent P., "Pratique des statistiques non paramétriques", édition de 1987 traduite par Ley J.P., Collection "techniques et pratiques" de l'INRA

[221] Srikant R., Agrawal R., "Fast Algorithms for Mining Association Rules", 20th VLDB Conf., Sept. 1994

[222] Stumme G., Taouil R., Bastide Y., Pasquier N., Lakhal L., Computing Iceberg Concept Lattices with Titanic. Data & knowledge Engineering, vol 42, n°2, 189-222. 2002

[223] Stumme G., Taouil R., Bastide Y., Pasquier N., Lakhal L., "Intelligent Structuring and Reducing Association with Formal Concept Analysis", BDA 1999, Bordeaux

[224] Suzuki E., Kodratoff Y., Discovery of Surprising Exception Rules Based on Intensity of Implication. *Second European Symposium on Principles of Data Mining and Knowledge Discovery*. Springer-Verlag, London, UK. Source Lecture Notes In Computer Science. 1998, p. 10-18.

[225] Taouil R., Algorithmique du treillis des fermés : application à l'analyse formelle de concepts et aux bases de données. Thèse d'informatique, Université Blaise Pascal, Clermont-Ferrand, 2001.

[226] Teytaud O. et Lallich S., "Bornes uniformes en extraction de règles", Conférence d'Apprentissage, CAP'2001, Grenoble, p. 133-148.

[227] Toinoven H., Klementtinen M., Ronkainen P., Hatönen K., Mannila H., "Pruning and Grouping Discovered Association Rules", ECML'95

[228] Toinoven H., Sampling large databases for association rules. 22nd VLDB Conf.,p.134-145, Morgan Kaufmann,1996

[229] Tournois J., Dickes P., Pratique de l'échelonnement multidimensionnel, de l'observation à l'interprétation. De Boeck, Université de Bruxelles 1993

[230] Urban Hjorth J.S., Computer Intensive statistical methods : Validation Model Selection and Boostrap. Chapman & Hall/CRC, London, UK, 1994

[231] Wang W., "Predictive Modeling Based on Classification and Pattern maching Methods", Thèse, Simon Fraser University, 1999

[232] Wang Z., "Collaborative Filtering Using Error-Tolerant Facsicles", These de Simon Fraser University, 2001

[233] Whalen, T. et Schott B., "Presumption, prejudice and regularity in fuzzy material implication", pp 265-280, in Zadeh, L.A., Kacprzyk J., "Fuzzy Logic for the management of uncertainty", John Wiley & sons, New York, 1992

[234] Whittaker, J., *Graphical models in applied multivariate Statistics*, John Wiley, 1990.

[235] Wille R., "Restructuring lattice theory : an approach based on hierarchies of concepts", in I. Rival(ed.), Ordered sets, Dordrecht-Boston, Reidel, 1982

[236] Winer B.J., Brown D.R., Michels K.M.() "Statistical principles in experimental design" (third edition) 1991

[237] Wolpe, J. et Lang, P.J. (1964). A fear survey schedule for use in behaviour therapy. Behaviour Research and Therapy. 2, 27-30.

[238] Wolpe, J., La pratique de la thérapie comportementale. Paris, Masson. 1976

[239] Yule, G.U., . "Notes on the theory of association of attributes in statistics", Biometrika 2, 121-134. 1903

[240] Zadeh, L.A., "Fuzzy Sets", Information & control., 8, 338-353. 1965

[241] Zadeh, L.A., Kacprzyk J. "Fuzzy Logic for the management of uncertainty", John Wiley & sons, New York, 1992

[242] Zaki M.J., Parthasarathy S., "New algorithms for fast Discovery of Association Rules", Technical Report 651, University of Rochester, New York, Juil. 1997

[243] Zaki M.J., Hsiao C.-J., "An efficient Algorithm for Closed itemset Mining", 0-Porc. SIAM Int. Conf. Data Mining, Arlington, VA, p457-473

[244] Zaki M.J., Gouda K., Fast vertical mining using diffsets. Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, D.C. 2003. p. 326-335

[245] Zhu H., "On-Line Analytical Mining of Association Rules", Thèse, Simon Fraser University, 1998

[246] Zighed D.A., Rakotomalala R., Graphes d'induction, Apprentissage et Data Mining Hermes Sciences Publications, Paris 2000