Genese et developpement de l'analyse statistique implicative: retrospective historique

Gênese e desenvolvimento da análise estatística implicativa: retrospetiva histórica

Genesis and development of statistical implicative analysis: historical retrospective

RÉGIS GRAS [1]

Resume

Cet article présente tout d'abord, l'origine de la situation fondamentale dans laquelle la nécessité d'organiser des comportements de réponse d'élèves à un test de didactique des mathématiques est apparue, en respectant la complexité a priori d'exercices. Cela a conduit à la création d'un indice d'implication entre items de réponses, pour évaluer des règles comme : « si a alors généralement b ». puis à des représentations du préordre partiel obtenu entre les réponses. La théorie s'est ensuite développée, sous la poussée des applications variées rencontrées, par extension de la nature des variables de comportement à des variables non binaires. Enfin, une relation topologique duale a été établie entre les sujets et les variables.

Mots-clés: *Taxonomie, Rapport non-linéaire tout/partie, Dialectique, Système dynamique, Stabilité structurelle, Propriété émergente, Règle, Métarègle, Intensité d'implication, Graphe implicatif, Hiérarchie cohésitive, Variables binaire, Modale, Numérique, Intervalle, Floue, Vectorielle, Supplémentaire, Structure topologique duale*

Resumo

Este artigo apresenta, em primeiro lugar, a origem da situação fundamental na qual surge a necessidade de organizar comportamentos de respostas de alunos em relação a um teste de didática da Matemática, respeitando a complexidade a priori dos exercícios. Este fato levou à criação de um índice de implicação entre itens de respostas, para avaliar regras como: "se a, então geralmente b", depois à representações da pré-ordem parcial obtida entre as respostas. A teoria desenvolveu-se depois, a partir das aplicações variadas encontradas, por extensão da natureza das variáveis de comportamento a variáveis não binarias. Enfim, uma relação topológica dual foi estabelecida entre os sujeitos e as variáveis.

Palavras-chave: *Taxonomia, Relação não linear parte/tudo, Dialética, Sistema dinâmico, Estabilidade estrutural, Propriedade emergente, Regra, meta-regra, Intensidade de implicação, Grafo implicativo, Hierarquia Coesitiva, Variáveis binaria, Modal, Numérica, Intervalo, Fuzzy, Vetorial, Suplementar, Estrutura topológica dual.*

[1] Ecole Polytechnique de l'Université de Nantes, Equipe Connaissance et Décision, Laboratoire d'Informatique de Nantes-Atlantique (LINA), UMR 6241, E-mail : regisgra@club-internet.fr, http://math.unipa.it/~grim/homegras_03.htm

Introduction

L'A.S.I. est le dernier plat (ultime ?) sorti de la marmite dans laquelle j'ai, au cours de mes 62 ans de service dans l'Education Nationale, ajouté, mélangé un nombre important d'ingrédients en une sorte de melting-pot jubilatoire et passionné.
Je me propose, avec une telle composition de la marmite, de vous raconter l'histoire de l'Analyse Statistique Implicative, sa raison d'être, son fondement épistémologique, ses développements, sans rentrer dans les détails techniques, les formules mathématiques, bref en ne faisant appel qu'à votre bon sens et votre intuition[2].

Problématique d'ordre psycho-didactique

Au cours des années 70, dans le cadre des Instituts de Recherche sur l'Enseignement des Mathématiques, j'ai fréquenté une nouvelle fois des classes du secondaire, particulièrement du 1er cycle où j'y ai conduit et évalué une expérience nationale tout en participant à la formation continue et aux recherches des enseignants de ces classes. J'ai donc été le témoin des difficultés d'apprentissage des élèves et quelquefois l'acteur des tentatives de remédiation aux obstacles qui s'opposent à l'assimilation des notions enseignées. Ces obstacles n'étaient pas toujours rencontrés par les élèves mais certains étaient récurrents et relativement partagés. Leur nature relevait de la didactique mais aussi souvent de l'épistémologie en s'opposant à l'acquisition de leurs connaissances[3]. Je pouvais observer ces problèmes d'apprentissage aussi bien directement par l'attitude ou l'expression orale des élèves mais également à travers des questionnaires ou des travaux écrits. Faisant l'hypothèse que les attitudes ou les comportements de réponse étaient globalement identifiables, je disposais de données constituées de traces laissées par les élèves. A l'occasion de résolution d'exercices de mathématiques ou de problèmes, une certaine hiérarchie de difficultés segmentait l'ensemble des élèves interrogés. Plus la difficulté s'accroissait, plus le nombre de réussites diminuait ce qui peut sembler une tautologie : on peut en effet s'attendre à ce que tout élève qui réussit une épreuve jugée difficile, dans un contexte qui serait comparable, réussirait a fortiori ce qui était facile. Ce qui pourrait le contester, ce sont les incohérences par rapport à cet attendu. L'idée que la

[2] « … il n'y a de connaissance vraie que par l'intuition, c'est-à-dire par un acte singulier de l'intelligence pure et attentive, et par la déduction, qui lie entre elles les évidences », M.Foucault, « Les mots et les choses », p. 103
[3] « *C'est en termes d'obstacles qu'il faut poser le problème de la connaissance scientifique* » écrit G.Bachelard dans « La formation de l'esprit scientifique »

difficulté a priori soit définissable objectivement par ma propre pratique ne tenait plus en tant que prédicteur, même si elle était le plus souvent respectée. Ainsi, c'est la relation stable et relativement prévisible entre réussites et échecs, entre comportements de réponse qui m'intéressait plutôt que la réussite ou l'échec à un item donné. Ce qui rejoint l'opinion de H. Poincaré qui dit « *que les mathématiciens n'étudient pas les objets mais les relations entre les objets* ».

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

Relativement à un questionnaire constitué d'items spécifiant chacun des objectifs cognitifs de la taxonomie, en théorie on aurait pu attendre des réussites organisées linéairement selon la complexité a priori. Ce qui n'a pas été observé. A l'ordre total présumé s'est substitué un **préordre partiel** (comme sont définis les stades différentiels de développement de l'enfant chez Piaget). Ce qui signifie que des élèves pouvaient dans certains cas et pour quelques-uns d'entre eux, réussir à un item *a* jugé difficile tout en échouant à un item *b* jugé plus facile. Et ceci sans remettre en cause l'affirmation que «<u>généralement</u> la réussite à *a* s'accompagne de la réussite à *b* » et sans que sa réciproque ne soit nécessairement vraie. Mon intérêt va alors porter sur ce type de relation non symétrique, tenter de pondérer la qualité de son caractère approximatif et d'organiser si possible l'ensemble des couples de variables-items en jeu de ce préordre partiel.

Quels **outils statistiques** étaient alors à ma disposition pour qualifier et quantifier cette relation non symétrique entre deux variables ? Un **test paramétrique** non symétrique ? mais pour réfuter quelle hypothèse ? que les élèves qui ont répondu à *a* ont aussi répondu à <u>b</u> ? que faire de la réfutation ? la ranger sagement ? établir une liste de cas réfutés ou

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

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

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

Atlan[5]. L'objectif est, si possible, de la conserver et la qualifier en mettant en évidence la faible probabilité de la dérogation à la règle sur des bases statistiques. Belle illustration de l'adage : « **l'exception confirme la règle** ».

Une mesure statistique de la quasi-implication. Représentations associées

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

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

Laquelle vous étonnerait ? A laquelle accorderiez-vous la meilleure qualité prédictive ? Dans le premier cas, la règle est stricte mais la confiance en elle est fragile. Dans le second cas, c'est le contraire, l'étonnement est plus grand en dépit de la moindre valeur de la

[5] « …[en accord avec Jung] *si la fréquence des coïncidences n'excède pas de façon significative la probabilité qu'on peut leur calculer en les attribuant au seul hasard à l'exclusion de relations causales cachées, nous n'avons certes aucune raison de supposer l'existence de telles relations*. », Atlan H., (1986) A tort et à raison. Intercritique de la science et du mythe, Paris : Seuil, p.160

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

fréquence conditionnelle. Ce paradoxe relatif à l'acceptabilité de la règle est soluble dans la subjectivité. La statistique ASI va en restaurer une composante objective.

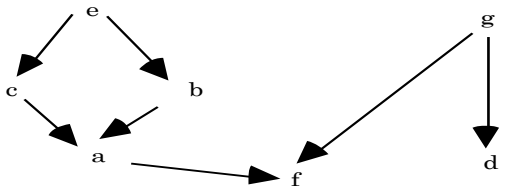
Certes, l'intensité d'implication est d'autant plus proche de 1 que la qualité pressentie de l'implication est grande. La relation qu'elle établit entre variables s'exprime fréquemment en termes de causalité. Mais cette *explicabilité causale* d'une variable par une ou plusieurs autres (J.-M. Levy Leblond parle de « **champ causal** » dans « Aux contraires ») que l'intensité d'implication évalue, n'est en rien **déterministe.** D'ailleurs, elle n'est pas **transitive** comme nous l'avons formalisé par l'étude des **règles d'exception** où les règles a =>b et b=>c ne s'accompagnent pas de la règle a =>c. Elle ne relève pas non plus d'un **déterminisme probabiliste** comme elle est souvent interprétée à tort : si 0.95 est une intensité d'implication de *a* sur *b*, cela ne signifie pas que *b* se réalise avec la probabilité 0.95 si *a* est réalisée.

Ainsi, partis de recherche de règles strictes, respectant la logique platonicienne, j'ai transigé et cherché des quasi-règles, ne respectant plus la logique formelle en raison de ses contre-exemples. Cette démarche illustre -nous y reviendrons- le mode de pensée que l'on dit **dialectique**[7] car il accepte les contradictions et les intègre pragmatiquement tout en enrichissant la connaissance.

Supposons donc effectué le calcul de l'intensité d'implication pour chaque paire de variables de la situation expérimentale. On ne conserve pour chaque paire que celle relative au couple conduisant à la plus grande intensité d'implication. Que faire du tableau de toutes ces valeurs ? Comment en dégager les lignes de force qui les structureraient comme le fait un plan factoriel ? Disposant d'un ensemble de relations binaires valuées, j'ai fait le choix de le représenter par un **graphe orienté pondéré et sans cycle,** image plus facilement appréhendable par l'utilisateur expert du domaine, par exemple, la didactique, la psychologie, la sociologie, la médecine, etc.. En général, il ne se réduit pas à un chemin linéaire puisque à une même « cause » peuvent être associés plusieurs « effets » et un « effet » peut être le fruit de plusieurs « causes ». Le problème de représentation d**'ontologies** est alors compatible avec ce cadre (travaux en collaboration avec **Jérôme David**). Les graphes suivants illustrent ces deux situations.

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

Figura1



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

Figura 2

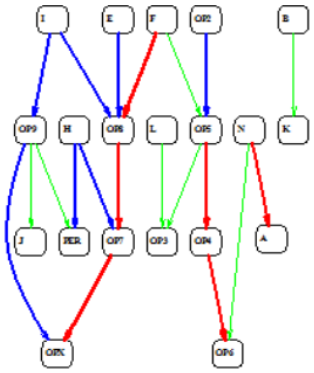
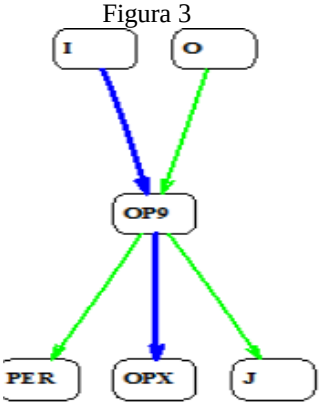


Figura 3



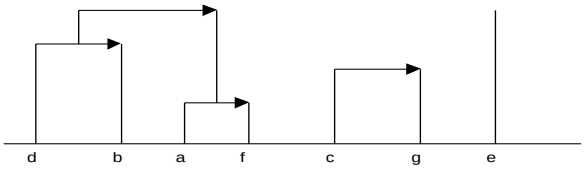
L'expert analyse et interprète alors les différents chemins en termes conceptuels en donnant du sens dans son domaine d'expertise à des **chemins** du graphe, à des **réseaux** comme dit Marc Bailleul, à un cône ascendants-descendants d'une variable au rôle d**'attracteur** qu'elle joue (Fig. 3), mais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

En psychologie du développement selon Piaget, la notion **d'abstraction réfléchissante** décrit le passage d'un niveau de conceptualisation à un niveau supérieur, chacun des niveaux étant constitué de règles portant sur des objets, puis d'opérations sur ces objets, puis sur des opérations sur ces opérations, etc... (par ex. schèmes, procédures, conception, méthode, ..). On retrouve, d'ailleurs, dans une théorie mathématique ces mêmes élargissements lorsque l'on passe d'un théorème à un corollaire ou, par exemple, dans l'étude des fonctions à celle, en analyse fonctionnelle, de fonction de fonctions. D'où l'idée de construire un second plan de relations implicatives, celui de **règles de règles**

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

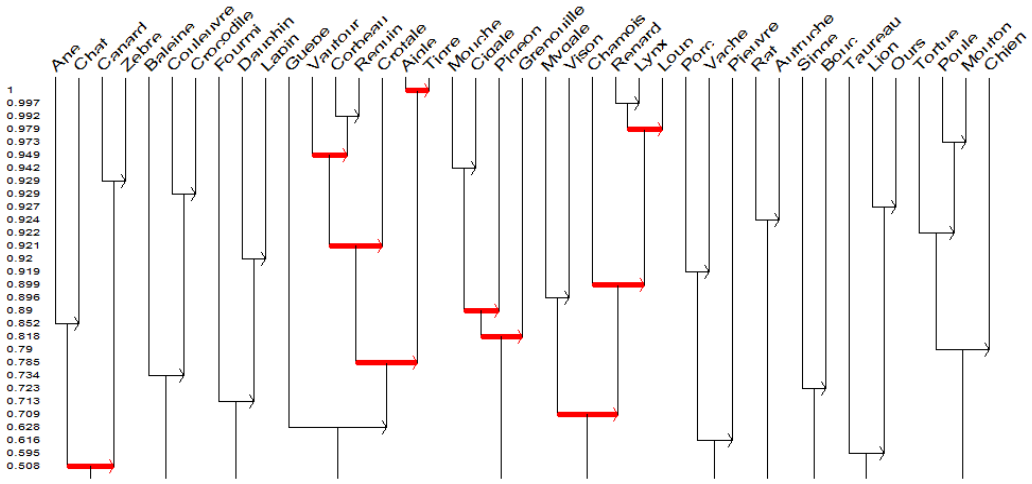
selon une **hiérarchie dite cohésitive** en raison de l'indice de **cohésion** qui permet d'engendrer des classes orientées de règles. A l'instar de certaines méthodes d'analyse de données, l'intérêt de passer, non linéairement, du niveau relationnel à celui de hiérarchie m'est apparu et nous avons étendu notre recherche de règles à celui de règles de règles ou méta-règles ou règles généralisées.

Figura 4. Graphe de la hiérarchie orientée



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

Figura 5. Graphe de la hiérarchie orientée



Un détour par la philosophie des sciences

A travers ces deux types de représentations de règles simples ou généralisées, qui mettent au jour deux types de structures dans l'ensemble des variables, nous répondons à la philosophie structuraliste qui estime que le **« tout » est plus riche que la somme de ses**

« parties ». Je cite, à ce sujet, deux extraits du livre de L. Sève (ib. p. 58) « *…le **tout** ne se compose de rien d'autre que de ses **parties**, et pourtant il présente, en tant que tout, des propriétés n'appartenant à aucune de ses parties. Autrement dit, dans le passage **non additif**, **non linéaire** des parties au tout, il y a apparition de propriétés qui ne sont d'aucune manière précontenues dans les parties et ne peuvent donc s'expliquer par elles* » et de poursuivre *« Tout se passe donc comme si se produisait une génération spontanée de propriétés du tout…C'est le paradoxe de **l'émergence**».*

J'insiste sur cette propriété spécifique de l'ASI que la hiérarchie cohésitive satisfait de façon originale à travers l'extension des relations entre variables. C'est par un saut qualitatif, produit généralement par un effet de seuil dans la quantité (ex. : psychologie de groupe, vaporisation de l'eau…), que le tout, au prix d'une synthèse, prend son sens. Celui-ci s'extrait de la lecture véritablement dialectique du **rapport non-linéaire tout/partie**[9] (non-proportionnalité et non-additivité de la cause sur l'effet). Avec l'ASI, il y a alors **paradoxe** entre des contraires, lien et absence de lien, car le tout, constitué (on devrait dire *organisé*) de parties en un **système dynamique**, possède des propriétés que ne possèdent pas les parties et qui sont généralement de **niveau supérieur**. De la même façon, et métaphoriquement, en linguistique, la signification d'une phrase ne se fait pas par l'analyse de chacun de ses mots mais par l'interaction de ceux-ci[10]. La logique qui sous-tend ce rapport tout/parties est **dialectique** (et non pas dichotomique) car elle concilie interactivement des contradictions : règle et non-règle, **instabilité d'un système dynamique et stabilité structurelle**. Elle se fonde en règle sur l'inexistence importante du contre-exemple et en métarègles sur la négation de l'entropie, du désordre. En cela, la logique dialectique s'oppose à la logique stricte (du mathématicien) sans, bien sûr, être un sophisme. La fécondité et l'originalité de l'ASI tiennent à ce caractère, particulièrement dans l'analyse hiérarchique manifestement non-linéaire où le tout fonde son sens, non par addition des propriétés de ses parties (sous-classes) mais par la synthèse

[9].. comme le montrent les équations différentielles non-linéaires de l'indice fondamental par rapport aux paramètres cardinaux des observations, contrairement à ce qui est observable avec d'autres indices concurrents..

« *La société n'est pas constituée d'individus, mais exprime la somme des relations, des rapports où ces individus se situent les uns par rapport aux autres* » (K. Marx, « Manuscrits de 1857-1858). Une rue n'est pas la somme des maisons qui y figurent. De même une ville n'est pas somme de ses rues, etc.

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

des interactions inférentielles. C'est par la notion de **niveau significatif** que nous pouvons mettre en évidence le phénomène de **propriété émergente**. En ce sens, l'A.S.I. apparaît comme une sorte d'avatar de l'apprentissage non linéaire des connaissances, apprentissage fait de ruptures et de reconstruction dialectique (cf. l'épistémologie de G. Bachelard ou de Lev Vygotsky). A l'opposé de l'A.S.I., le rapport tout-parties serait **linéaire** dans le cas d'emboîtements de classes comme en clustering fondée sur la similitude jusqu'à son nœud terminal.

Après ce pas de côté**, remontons le temps**. A l'occasion d'un dialogue avec Jacques Philippé, en 1990, je me suis aperçu que ma problématique coïncidait en partie avec la sienne. D'où mon intégration proposée par Henri Briand d'être associé au groupe de recherche de l'IRESTE, nouvellement Polytech'Nantes.

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

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

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

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

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

Ce logiciel est opérationnel à travers le monde puisque 26 pays, gréco-latins pour la plupart, le possèdent et l'utilisent[11]. Il y joue, pour la recherche, le double rôle de **révélateur et d'analyseur**. Si bien que des questions sur les possibilités et les limites de l'ASI se font jour à travers des questions d'ordre général mais aussi spécifiques du fait des traditions et des cultures différentes. Outre le lien épistolaire, elles se renforcent au fil de nos rencontres internationales sur l'ASI, de ASI 1 à ASI 6 en 2012 (et prochainement ASI 7 au Brésil). Ce sont ces questions qui poussent la théorie et son outil vers des développements continus comme nous allons le voir. Pour illustrer ceci, je citerai Anne Lauvergeon dans « La femme qui résiste » (Plon, 2012) : « … *lorsque l'on produit, on finit également par concevoir* ».

Extension de la méthode à d'autres variables

C'est donc à travers les différentes situations rencontrées que la limitation aux variables binaires, ayant servi à donner un sens ensembliste à l'implication, est apparue contraignante. Celui qui a tiré le premier est Marc Bailleul qui a voulu rechercher des relations de type préférences entre des assertions d'enseignants. Des **variables modales** (« un peu », « beaucoup », « pas du tout », …), devaient être traitées. Il a proposé un

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

premier indice satisfaisant à la mesure d'expression du type : « **si pour *a* la modalité « *un peu* » est choisie alors généralement pour *b* une modalité supérieure ou égale est choisie »**. Dans sa thèse, Marc Bailleul a obtenu d'excellents résultats, dont certains imprévisibles, relativement à 4 conceptions de l'enseignement des mathématiques sous le regard des enseignants.

Afin de procéder, relativement à la même problématique, par extension de l'intensité d'implication entre variables booléennes, avec J.-B. Lagrange, nous avons alors défini une nouvelle mesure portant sur des **variables numériques** et modales permettant d'attribuer une valeur à des énoncés comme : « **si $a = \alpha$ alors généralement $b \geq \beta$** ». Jean-Claude Régnier, de son côté, a ramené la problématique de la recherche de relations entre préférences à celle de **variables de rangs** qui a fourni une autre extension de l'A.S.I...

De là, suite à une question d'E. Diday se plaçant dans le cadre des **variables symboliques** et par une extension des variables numériques, nous avons affecté des valeurs à des expressions « si ***a* prend ses valeurs dans l'intervalle I_a alors généralement *b* prend ses valeurs dans I_b** ». L'idée maîtresse a consisté à partitionner l'étendue des valeurs prises par chaque variable en sous-intervalles maximisant leur variance interclasse. L'intérêt de cette nouvelle catégorie de variables, **dites intervalles**, se manifestait dans l'enseignement, pour comparer hiérarchiquement des performances dans des disciplines différentes et dans la recherche de transfert de compétences spécifiques vers d'autres compétences.

On voit alors que de ces **variables-intervalles** traitées en collaboration avec E. Diday et Pascale Kuntz, nous parvenons naturellement à **des variables floues** comme nous les rencontrons dans l'étude de relations du type : « si la tension *a* d'un sujet est plutôt élevée alors généralement son rythme cardiaque *b* l'est aussi ». On pressent les applications pratiques qui peuvent en découler dans des problèmes de construction ou de détections de pannes. La collaboration de Fabrice Guillet, Maurice Bernardet, Raphaël Couturier, Filippo Spagnolo et moi-même m'a permis de modéliser la notion de variable floue dans le cadre des variables-intervalles et d'en faire une présentation dans un congrès sur la logique floue.

Un problème est survenu un jour dans le traitement implicatif d'un ensemble trop large de variables au point de rendre illisibles les graphes obtenus. Avec Raphaël Couturier, Fabrice Guillet et Robin Gras, nous avons défini une relation d'équivalence entre les

variables, basée sur leurs comportements implicatifs voisins, qui a permis de substituer à un paquet de variables un représentant leader de ce paquet. Cette **réduction** s'est avérée efficace dans de nombreuses autres situations, par exemple dans la thèse de Laurence Ndong.

Un petit arrêt pour parler des collaborations dans l'équipe COD, toujours appuyées sur des choix épistémologiques en réponse à des attendus sémantiques. L'explosion d'une multitude de questions applicatives ou théoriques, toujours non symétriques, a conduit à des travaux communs entrepris, depuis 1990, avec certains membres de l'équipe du COD actuel : je citerai par exemple, avec Pascale Kuntz sur **les hiérarchies de règles** (ah ! la démonstration de l'ultramétricité de la hiérarchie ! le modèle algébrique de la hiérarchie !), sur les **règles d'exception** avec Einoshin Suzuki, la **redondance de règles** avec Pascale, avec Julien Blanchard sur **l'analyse entropique**, les **variables temporelles**, **l'expression de gènes en bio-informatique** avec Gérard Ramstein, etc..

Par exemple, comment la nécessité de variables temporelles est-elle apparue ? Eh bien, pour rendre compte de l'évolution des relations entre variables économiques, sociales ou cognitives. En effet, peut-on expliquer dans l'enseignement, l'implication entre variables lorsque l'on procède à des interventions en cours d'année sur l'apprentissage ou encore en socio-psychologie par des interventions successives, par des entretiens ou par des stages (en collaboration avec D.Pasquier). Nous avons alors formalisé ces variables indexées par le temps en **variable vectorielle,** où une variable est modélisée par un vecteur paramétré par le temps. Puis, Julien Blanchard, en collaboration avec Fabrice Guillet et moi-même, a, de façon différente, défini des **variables séquentielles** modélisées par un processus de Poisson. Autant de nouveaux concepts et de nouveaux champs d'application nés de problématiques variées et d'extensions attendues.

Des extensions généralisantes de l'ASI ont vu le jour récemment :

- d'une part, à un ensemble continu de sujets, par exemple des couleurs, des opinions, muni d'une loi de distribution donnée, généralisation puisque jusqu'alors le modèle se limitait à des ensembles discrets et finis ;
- d'autre part, à l'ensemble des valeurs prises par les différents types de variables dans des espaces continus, par exemple des champs, munis de lois données.

J'insiste sur une remarque susceptible de satisfaire tout cartésien: lors de chaque extension, **nous nous sommes efforcés de prouver et nous y sommes parvenus, que la**

restriction au cas fondamental de variables binaires et d'espaces discrets était toujours satisfaite. L'emboîtement mathématique est original et rigoureux.

Rôle explicatif de variables supplémentaires

Une autre question m'a taraudé très tôt. Y a-t-il dans la population de sujets concernés par une étude, une structure interne qui expliquerait la structure des variables révélée par l'ASI ? Supposant une structure implicative de variables obtenue par un graphe ou une hiérarchie, est-il possible de désigner les sujets et les catégories de sujets plus ou moins responsables des éléments de ces structures ? Par exemple, si nous observons que dans l'ensemble de classes scolaires une certaine conception de notions géométriques entraîne certaines conduites de réponses, à quel type d'élèves peut-on plus spécifiquement l'attribuer ? Inversement quels sont les élèves qui y seraient réfractaires ? Avec les thèses de Harrisson Ratsimba-Rajohn et Marc Bailleul, nous avons formalisé et exploité, dans l'A.S.I., la notion de variable supplémentaire (ou **exogènes** par opposition aux variables **endogènes** analysées) et sa **contribution** (on parle aussi de **typicalité**) à des éléments de structure, des **réseaux** par exemple, sémantiquement expliquées. Mieux encore, nous avons pu définir de façon rétroactive une **structure topologique duale** sur l'ensemble des sujets. Ainsi, par exemple, une conception « emblématique de l'école républicaine » est renforcée dans un milieu scolaire favorisé où la distance inférée entre les élèves par un élément des structures de règles est la plus faible (travaux avec Dominique Lahanier-Reuter).

Mais puisque j'ai cité au fil des pages les apports et le rôle d'aiguillons de collaborateurs pour le développement de l'ASI, en particulier de l'équipe COD (Pascale Kuntz, Fabrice Guillet, Julien Blanchard, P. Peter et Gérard Ramstein), je dois mettre en exergue le rôle critique, poil à gratter mais constructif que joue Jean-Claude Régnier tant à l'égard de la théorie et ses applications qu'à l'égard de la diffusion, la popularisation de l'ASI puisqu'il a pris en charge la présidence des manifestations internationales ASI. Combien de lapsus, d'erreurs, de maladresses, d'excès de précipitation, Pascale et Jean-Claude ont-ils redressés ! Je n'oublie pas non plus le dévouement fidèle et les compétences informatiques de Raphaël Couturier sans lequel l'ASI, privée de CHIC, ne serait qu'une construction théorique et peut-être spéculative à contempler. Il a comblé, avec CHIC, mon handicap en informatique comme Pascale a bien voulu le faire pour l'anglais. Je

crains d'ailleurs que cette dernière carence ait coûté une meilleure audience de l'ASI sur le plan international.

Cependant, d'autres équipes internationales participent également aux travaux sur le développement et les applications de l'ASI en utilisant le logiciel CHIC. Je cite, en particulier, étant donné leur régularité de participation :

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

Les spécificités de l'ASI et conclusion

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

- les **modèles successifs** de variables répondant à des contraintes épistémologiques explicites compatibles avec la sémantique des situations à modéliser ;
- la **non-symétrie** de la méthode;
- l'extension progressive de la nature des variables traitées tout en conservant les propriétés de plongement;
- les **capacités pédagogiques et ergonomiques** des représentations, en particulier pour l'examen des règles généralisées;

- **la dualité structurelle** des deux espaces en jeu : sujets et variables avec les notions de contribution et de typicalité aux structures;
- **l'extension du discret fini au continu** tant pour les variables que pour les sujets;
- **l'originalité du raisonnement dialectique** à la base de la définition des règles simples et généralisées;
- la **simplicité du modèle mathématique** sous-jacent lui assurant accessibilité plasticité et fécondité utiles pour répondre à des attentes applicatives dans de larges domaines.

D'aucuns diront que l'ASI, par l'amplitude de ses champs d'application et par l'homogénéité de ses propriétés analytiques et graphiques présente une **nature paradigmatique** originale. Doit-on le reconnaître ? Djamel Zighed, Directeur de l'Institut des Sciences de l'Homme de Lyon exprime cette idée de la façon suivante : « *l'ASI n'est pas une méthode mais un cadre théorique, large, dans lequel se traitent des problèmes modernes de l'extraction des connaissances à partir des données. C'est une théorie générale dans le domaine de la causalité parce qu'elle répond à des faiblesses d'autres théories, elle apporte un outillage formel et des méthodes pratiques de résolution de problèmes. Ses applications sont multiples... ».*

Pour terminer, je voudrais remercier tous les participants à ce 7ème colloque scientifique international sur l'Analyse Statistique Implicative organisé sous la direction scientifique et organisationnelle de Jean-Claude Régnier et de Saddo Ag Almouloud pour leurs contributions au développement du cadre théorique, méthodologique et applicatif au travers de la mise en œuvre des concepts et des outils qui le constituent, ainsi que par la confrontation aux défis théoriques traduits par des problématiques nouvelles.

Références

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

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

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

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

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

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

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

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