Analyse comparative d'indices d'implication discriminants fond´es sur une échelle de probabilité

Israël-César C. Lerman, Sylvie Guillaume

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Analyse comparative d'indices d'implication discriminants fondés sur une échelle de probabilité.

Israël-César LERMAN — Sylvie GUILLAUME

N° 7187

Février 2010

Thème BIO

ISSN 0249-6399 ISRN INRIA/RR--7187--FR+ENG

Analyse comparative d'indices d'implication discriminants fondés sur une échelle de probabilité.

Israël-César LERMAN* , Sylvie GUILLAUME†

Thème BIO — Systèmes biologiques
Équipes-Projets Symbiose

Rapport de recherche n° 7187 — Février 2010 — 85 pages

Résumé : Historiquement, l'élaboration d'une échelle de probabilité pour éprouver l'existence d'un lien entre *deux* attributs descriptifs a été établie dans l'optique des tests d'hypothèses statistiques. L'adaptation au problème de la comparaison mutuelle entre *plusieurs* attributs nécessite une normalisation préalable ; laquelle est indispensable pour que l'échelle de probabilité reste discriminante pour un nombre n d'observations augmentant de façon considérable (n pouvant atteindre plusieurs millions). C'est le cas de l'association *symétrique* traduisant un "degré d'équivalence" entre attributs qui s'est présenté en premier. Plus récemment, il s'est agi du cas de l'association *dissymétrique* traduisant un "degré d'implication" entre attributs, définissant ainsi ce qu'on appelle une "règle d'association". Ce dernier cas sera étudié de façon plus accentuée ici. Différentes techniques de normalisation ont été proposées. La première est *contextuelle* par rapport à un ensemble potentiel de règles d'association. Elle conduit à l'*Intensité d'Implication Contextuelle* (IIC). La seconde raisonne par rapport à un échantillon dont la taille serait réduite à 100 et propose une *Valeur Test* notée VT100. Nous découvrirons différentes variantes pour une telle réduction. La troisième technique, conduisant à l'*Intensité d'Implication Entropique* (IIE), mélange un indice probabiliste non normalisé et un indice d'inclusion faisant appel à l'entropie de Shannon. L'objet de ce travail est l'analyse théorique et expérimentale de ces différentes approches par rapport à différents modèles de croissance du nombre n d'observations. Une vision nouvelle et des résultats originaux seront proposés. L'analyse comparative et expérimentale utilisera la base de données bien connues "Wages".

Mots-clés : Règles d'association, Vraisemblance du Lien, Intensité d'Implication, Indices normalisés, Comportements limites.

* INRIA Rennes - Bretagne Atlantique, and University of Rennes 1, France.Campus de Beaulieu, 35042 Rennes Cédex — lerman@irisa.fr

† Université d'Auvergne, Laboratoire d'Informatique, de Modélisation et d'Optimisation des Systèmes (LIMOS) guillaum@isima.fr

Centre de recherche INRIA Rennes – Bretagne Atlantique
IRISA, Campus universitaire de Beaulieu, 35042 Rennes Cedex (France)
Téléphone : +33 2 99 84 71 00 — Télécopie : +33 2 99 84 71 71

Analysis of Discriminant association rules based on a probability scale.

Abstract: Historically, a probability scale has been established in order to validate the dependency between the two components of a *single* couple of descriptive attributes. For the pairwise comparison of an attribute set, the adaptation of such a probability scale necessitates a preliminary normalization. The latter plays a very important part in preserving the discriminant property of the probability scale when the observation number becomes large. *Symmetrical* associations between attributes have been studied first (*Likelihood Linkage Analysis Classification method*). The quantified notion may be expressed by the intensity of an "equivalence degree" between attributes. More recently, *Disymmetrical* associations between attributes have been considered, leading to a very active research. In this case, the quantified notion may be expressed by the intensity of an "implication degree", giving "association rules" between attributes. Different normalization techniques have been proposed. The first of them has a *contextual* nature with respect to a relevant set of association rules. It leads to the notion of "Contextual Implication Intensity". In the second approach, the data are sumarized by means of 100 sized sample. A *Test Value* TV100 is then proposed. New interpretations and new alternatives of this type of approach are considered in our paper. The discussed last technique averages in a same index, the *Entropic Intensity of Implication*, a non normalized probability index and an inclusion index given by the notion of Shannon entropy. Theoretical and experimental analysis of these different approches are provided. Comparative behaviours of the different indices are studied when the observation number n increases according to specific experimental designs. This comparative analysis uses the well known data base "Wadges".

Key-words: Association Rules, Likelihood Linkage Analysis, Implication Intensity, Normalized Indices, Asymptotic behaviours.

1 Introduction générale et position du problème

Relativement à une base de données où on distingue un ensemble $\mathcal{A}$ d'attributs booléens observés sur un ensemble $\mathcal{O}$ d'entités (objets, individus, ...), le problème fondamental et bien connu en "Fouille des Données" ("Data Mining") est de pouvoir inférer un ensemble significatif et exploitable de règles d'association, on dit encore d'implication, entre attributs. Pour $(a, b) \in \mathcal{A} \times \mathcal{A}$, une règle de la forme $a \rightarrow b$ signifie que si a est à $VRAI$ sur un élément de $\mathcal{O}$, alors généralement - mais sans que cela soit un absolu - b est également à $VRAI$ sur cet élément de $\mathcal{O}$. Pour détecter de telles associations orientées, il importe de disposer d'un indice (on dit encore "coefficient" ou "mesure") statistique pertinent d'implication qui permette de dégager des "règles d'association" *intéressantes* ; c'est - à - dire, qui augmentent notre connaissance du réseau des tendances causales entre attributs de $\mathcal{A}$. Le souci d'une *bonne* mesure transparaît bien dans [29] où différents critères sont considérés pour organiser un ensemble de coefficients définissant des indices d'implication. D'autres études comparatives avec des facettes différentes sont également considérées dans [9, 11]. Dans les travaux de Agrawal et al. [1] qui ont lancé ces recherches dans le domaine de la "Fouille des Données", c'est un indice certes naturel, mais très simple, voire brut qui a été proposé. Relativement à l'évaluation du degré d'implication $a \rightarrow b$, il s'agit de la proportion conditionnelle (on dit encore "relative") des entités où b est à $VRAI$ dans l'ensemble des entités où a est à $VRAI$. Ainsi, ce coefficient, appelé "Confiance" peut s'écrire $card[\mathcal{O}(a) \cap \mathcal{O}(b)]/card[\mathcal{O}(a)]$ où pour un attribut booléen α on a noté $\mathcal{O}(\alpha)$ le sous ensemble de $\mathcal{O}$ où α est à $VRAI$. Ce coefficient est généralement accompagné de celui appelé "Support" qui représente la proportion d'éléments de $\mathcal{O}$ où la conjonction $a \wedge b$ des attributs a et b est à $VRAI$; soit : $card[\mathcal{O}(a) \cap \mathcal{O}(b)]/card[\mathcal{O}]$. Il est bien connu depuis les travaux de Agrawal et al. [1] que l'indice Confiance associé à celui Support permet d'obtenir dans le treillis des parties de l'ensemble $\mathcal{A}$ des attributs, un ensemble de règles d'association implicatives entre conjonctions d'attributs, appelés "itemsets" ou motifs. Il est également bien connu que les ensembles de règles ainsi obtenus restent le plus souvent difficilement exploitables, compte tenu notamment de leur importance numérique. D'autre part, on a pu mettre en évidence des effets non désirables de l'indice Confiance [9]. Cet indice n'est généralement pris en considération que si le Support est supérieur à un seuil donné, ce qui peut faire perdre certaines règles intéressantes pour lesquelles le Support est inférieur à ce seuil et que l'on appelle "pépites de connaissance". Posons ici $n = card(\mathcal{O})$, $n(a) = card[\mathcal{O}(a)]$, $p(a) = n(a)/n$, $n(b) = card[\mathcal{O}(b)]$, $p(b) = n(b)/n$, $n(a \wedge b) = card[\mathcal{O}(a) \cap \mathcal{O}(b)]$ et $p(a \wedge b) = n(a \wedge b)/n$ qui n'est autre que le Support. On constate que dès lors que les proportions $p(a \wedge b)$ et $p(a)$ sont fixées, la Confiance est invariable quelle que soit la taille de $card(\mathcal{O})$ de l'ensemble $\mathcal{O}$. Pour $n(a)$ fixé, la Confiance varie linéairement par rapport au nombre $n(a \wedge \bar{b})$ de contre-exemples ($\bar{b}$ est l'attribut opposé à b). On met en évidence dans [9] des situations où Confiance et Support ont des valeurs relativement grandes alors qu'on a la relation $p(a \wedge b) = p(a) \times p(b)$, caractéristique de l'indépendance entre les deux attributs a et b.

Cependant, compte tenu de son intelligibilité ainsi que de ses vertus algorithmiques, l'indice Confiance fait référence. Associé à Support, nous l'utilisons comme premier filtre pour sélectionner un ensemble de règles relativement im-

portant en taille, dans lequel un indice plus raffiné peut permettre d'obtenir un ensemble de règles *intéressantes* et de taille exploitable [15].

Le caractère non adapté de l'indice Confiance comme mesure de l'intérêt d'une règle d'association implicative a conduit à la proposition de nombreux coefficients. Des recherches importantes ont pu être effectuées pour l'analyse et la situation relative des différents coefficients [9, 11, 15, 29]. La mise à contribution de la notion d'indépendance statistique entre attributs intervient dans l'élaboration de nombreux coefficients. D'autre part, alors que l'indice Confiance a un caractère dissymétrique relativement à la mesure de l'implication $a \rightarrow b$ mentionnée ci-dessus, un certain nombre de coefficients ont un caractère symétrique [9, 11, 15, 26, 29].

Précisément, dans notre approche [15] c'est une hypothèse probabiliste d'*indépendance* que nous appelons aussi *hypothèse d'absence de liaison* entre attributs qui joue un rôle fondamental dans la conception d'un indice d'association qui se réfère à une échelle de probabilité. Nous appelons cet indice de *Vraisemblance du Lien* ou plus précisément, de *Vraisemblance du Lien Relationnel*. Dans l'article séminal [12], le problème équivaut à la comparaison symétrique entre deux attributs booléens a et b. Un indice "brut" de similarité est introduit sous la forme $n(a \wedge b) = card[\mathcal{O}(a) \cap \mathcal{O}(b)]$ (voir notations ci-dessus). On considère alors une hypothèse probabiliste d'absence de liaison $\mathcal{N}$ associant au couple (a, b) d'attributs un couple $(a^\star, b^\star)$ d'attributs aléatoires indépendants. Pour ce modèle aléatoire, $\mathcal{E}$ désignant l'espérance mathématique, $\mathcal{E}[card(\mathcal{O}(a^\star))] = n(a)$ et $\mathcal{E}[card(\mathcal{O}(b^\star))] = n(b)$. Nous mettons ici l'indice probabiliste de *vraisemblance du lien* sous la forme :

$$P_l^{\mathcal{N}}(a, b) = Pr^{\mathcal{N}}\{n(a^\star \wedge b^\star) < n(a \wedge b)\}, \tag{1}$$

où on a noté $n(a^\star \wedge b^\star) = card[\mathcal{O}(a^\star) \cap \mathcal{O}(b^\star)]$ (P comme Probabiliste et l comme localement restreint à la comparaison entre les deux attributs booléens a et b).

Pour une telle échelle de probabilité $P_l^{\mathcal{N}}(a, b)$ indique combien $n(a \wedge b)$ est invraisemblablement grand, eu égard à $n(a)$ et $n(b)$. Précisons ici que relativement à l'adoption de la forme (1) de l'indice $P_l^{\mathcal{N}}(a, b)$, le degré d'invraisemblance de la grandeur relative de $n(a \wedge b)$ est mesurée par $Pr^{\mathcal{N}}\{n(a^\star \wedge b^\star) \geq n(a \wedge b)\}$. Nous aurions pu - car il s'agit d'une convention dans la définition - mesurer ce degré d'invraisemblance par $Pr^{\mathcal{N}}\{n(a^\star \wedge b^\star) > n(a \wedge b)\}$. De toute façon, les deux valeurs diffèrent peu dans les conditions générales où n, $n(a)$, $n(b)$ et $n(a \wedge b)$ sont "suffisamment grands". En fait, il nous fallait surtout ici prendre une définition qui permette la comparaison la plus ajustée avec l'approche définie dans [23, 27].

Jusqu'à une valeur importante de $n = card(\mathcal{O})$, $P_l^{\mathcal{N}}(a, b)$ peut être calculé exactement. Cependant, pour n assez grand, l'approximation par la loi normale de la loi de $n(a^\star, b^\star)$ est excellente. Elle est de moyenne $n(a)n(b)/n$ et de variance dépendant de la manière dont on spécifie le modèle aléatoire $\mathcal{N}$. Ainsi, en introduisant l'indice $n(a \wedge b)$ centré et réduit par rapport à $\mathcal{N}$:

$$Q(a,b) = \frac{n(a \wedge b) - \mathcal{E}[n(a^\star \wedge b^\star)]}{\sqrt{var[n(a^\star \wedge b^\star)]}} \tag{2}$$

On a, pour n, $n(a)$ et $n(b)$ "suffisamment grands",

$$P_l(a,b) \simeq \Phi[Q(a,b)] \tag{3}$$

où Φ désigne la fonction de répartition de la loi normale centrée et réduite. Signalons qu'il peut nous arriver dans la suite, surtout dans un contexte où n n'est pas "assez grand" d'introduire pour le calcul de l'approximation dans le numérateur de $Q(a,b)$ un élément correctif dont la valeur absolue est 0.5. Compte tenu de l'expression (5) ci-dessous, il s'agit ici de $+0.5$. Cette technique est utilisée dans [27] et nous aurons encore une fois à comparer notre approche de la "Vraisemblance du Lien" avec celle des *p-values* [23, 27]. Signalons néanmoins que l'ordre de grandeur de cet élément correctif sur l'ensemble du coefficient Q est de $(1/4n)^{0.5}$ en supposant que $O[min(n(a), n(\bar{a}), n(b), n(\bar{b}))] = O(n)$.

L'approche classique dans les tests statistiques d'indépendance consiste à utiliser la probabilité complémentaire de $P_l(a,b)$, soit

$$\bar{P}_l(a,b) = 1 - P_l(a,b) = Pr^{\mathcal{N}}\{n(a^\star, b^\star) \geq n(a \wedge b)\} \tag{4}$$

pour décider du rejet ou non de l'hypothèse d'indépendance entre les attributs a et b. Ayant décidé un seuil α, le rejet est adopté dès lors que $\bar{P}_l(a,b)$ est inférieur ou égal à α. Indépendamment $\bar{P}_l(a,b)$ définit le seuil critique ; c'est-à-dire, celui à partir duquel on rejetterait l'hypothèse d'indépendance. Dans [22] on introduit la valeur test VT qui est définie comme "le nombre d'écarts types de la loi normale centrée et réduite qu'il faut dépasser pour couvrir cette probabilité". Cette probabilité est très précisément celle $P_l(a,b)$ que nous avons définie ; mais dans une toute autre optique. Dans la section 2 nous préciserons la situation de la démarche VT par rapport à celle VL de la *Vraisemblance du Lien*.

Sous l'appellation " *d'Implication*", Régis Gras [5] a proposé une adaptation dissymétrique de notre indice de la *Vraisemblance du Lien*. Au lieu d'évaluer le degré d'invraisemblance de la grandeur relative de $n(a \wedge b)$ dans le cadre du modèle aléatoire de l'hypothèse d'absence de liaison, il s'agit d'évaluer le degré d'invraisemblance de la petitesse de $n(a \wedge \bar{b})$. Cet indice représente en effet le nombre de contre-exemples de l'implication $a \rightarrow b$. La forme générale de l'indice est alors

$$\mathcal{I}(a \rightarrow b) = Pr^{\mathcal{N}}\{n(a^\star \wedge \bar{b}^\star) > n(a \wedge \bar{b})\} \tag{5}$$

Nous avons pu montrer dans [21] que le modèle aléatoire $\mathcal{N}$ le plus adapté conduisait pour l'indice aléatoire $n(a^\star \wedge \bar{b}^\star)$ à une loi de Poisson de paramètre $(n(a) \times n(\bar{b}))/n$. Ainsi la loi de probabilité de $n(a^\star \wedge \bar{b}^\star)$ peut être approchée de façon suffisamment précise par la loi normale de moyenne et de variance toutes les deux égales à $(n(a) \times n(\bar{b}))/n$ [4]. En introduisant l'indice brut $n(a \wedge \bar{b})$ centré et réduit par rapport à l'hypothèse d'absence de liaison à caractère Poissonnien (voir ci-après (16) pour la notation adoptée) on obtient :

$$Q_3(a,\bar{b}) = \frac{n(a \wedge \bar{b}) - (n(a) \times n(\bar{b}))/n}{\sqrt{n(a) \times n(\bar{b})/n}} \tag{6}$$

Davantage l'implication est prononcée, davantage $Q_3(a \wedge \bar{b})$ est petit (plus fortement négatif). Compte tenu de l'approximation par la loi normale, l'indice d'implication qui se réfère à une échelle de probabilité prend la forme :

$$\mathcal{I}(a \rightarrow b) \simeq 1 - \Phi[Q_3(a, \bar{b})] = \Phi[-Q_3(a, \bar{b})] \tag{7}$$

Ici encore, pour le calcul de l'approximation numérique, on peut introduire au niveau du numérateur de $Q_3(a, \bar{b})$ un coefficient correctif qui vaut 0.5. Cependant, on suppose qu'on se trouve dans le cas général où $min[n(a), n(\bar{a}), n(b), n(\bar{b})]$ est "assez grand" pour rendre négligeable l'influence de cette correction.

Considérons à présent l'optique des tests d'indépendance où l'hypothèse d'indépendance entre les attributs a et b est éprouvée contre l'hypothèse de l'implication $a \rightarrow b$. Dans ce cas, la $p-value$ se trouve définie par la probabilité complémentaire de (5) :

$$p = Pr^N\{n(a^\star \wedge \bar{b}^\star) \leq n(a \wedge \bar{b})\} \tag{8}$$

et la valeur test peut s'écrire :

$$VT = \Phi^{-1}(1 - p) \tag{9}$$

Les indices (3) ou (5) de vraisemblance du lien permettent de comparer toutes les paires d'attributs sur la même base conceptuelle en neutralisant l'influence de la distribution des tailles $\{n(a) \mid a \in \mathcal{A}\}$. Cependant, les valeurs de ces indices se trouvent - pour une même distribution des fréquences jointes de la forme $\{p(a \wedge b) = n(a \wedge b)/n \mid a \in \mathcal{A}, b \in \mathcal{A})\}$, très concentrées autour des valeurs 1 et 0, lorsque n est "suffisamment grand". Or on sait qu'en "Fouille des Données", le nombre d'entités peut être très grand, de l'ordre de plusieurs millions par exemple. Ainsi, la $p - value$ devient quasi-nulle dès lors que $n(a \wedge b) > n(a) \times n(b)/n$ [23, 27]. La raison de ce phénomène en est qu'un indice tel que (5) s'interprète comme celui de la confirmation d'un lien unique compte tenu de l'incertitude liée à la taille de l'échantillon. Alors que pour notre part, nous voulons utiliser l'*hypothèse d'absence de liaison* comme référence pour établir une échelle de probabilité pour la comparaison des liens mutuels. À cet effet nous avons proposé et justifié sur les plans expérimental et théorique [13, 3] la réduction "globale" des indices de la forme (2) ou (6) avant la référence à l'indice probabiliste via la fonction de répartition de la loi normale centrée et réduite. En considérant le cas de l'indice Q_3 (voir (6)) qui va plus particulièrement nous intéresser, cette réduction "globale" consiste d'abord à considérer la distribution empirique de Q_3 - sur un ensemble potentiel pour l'implication - $\mathcal{C}$ de couples (a, b) et à en calculer la moyenne et l'écart-type empiriques. On substituera alors à la distribution de Q_3, celle de l'indice que nous notons Q_3^g, obtenue en centrant avec cette moyenne et en réduisant avec cet écart-type. Ainsi, la distribution de Q_3^g est normalisée de moyenne nulle et de variance unité. Ce que nous pouvons appeler *"Intensité d'Implication Contextuelle"* (relativement à $\mathcal{C}$) donne alors pour le couple (a, b) d'attributs

$$\mathcal{I}_c(a \rightarrow b) = \Phi[Q_3^g(a, \bar{b})] \tag{10}$$

Ainsi, l'échelle de probabilité devient finement discriminante quel que soit le nombre n d'entités.

Dans le cadre de la proposition d'une valeur test [voir (8) et (9) ci-dessus] pour résoudre le problème d'une *p-value* quasiment égale à 0 pour n assez grand, A. Morineau et R. Rakotomalala [23, 27] proposent de se ramener *artificiellement* au cas du nombre d'observations $n = 100$. Ils considèrent en effet que cette valeur de la taille d'un échantillon est symbolique de la pratique de prise de décision dans les tests d'hypothèses. Ils précisent que le modèle consiste à répéter "un grand nombre de fois le tirage d'un échantillon de taille 100". Dans ces conditions, on associe à chaque échantillon la *p-value*. La moyenne des *p-values* conduit alors à une valeur test notée VT_{100} qui est parfaitement discriminante pour le choix d'un sous ensemble réduit de règles d'association implicatives. Nous analyserons cette démarche et sa mise en pratique dans les sections 3 et 4. Cette analyse nous permettra de proposer des solutions nouvelles dans l'esprit de VT_{100} (section 4). Nous montrerons en particulier que la solution proposée [23, 27] ne correspond en aucune façon au modèle de tirage annoncé. Les sections 2 et 3 sont consacrés au développement du cas de l'association symétrique entre attributs de $\mathcal{A}$; alors que dans les sections 4 et 5 on étudie le cas de l'association dissymétrique implicative entre attributs de $\mathcal{A}$.

La motivation première de ce travail est la comparaison de l'approche *Vraisemblance du Lien normalisée* (c'est-à-dire, supposant la réduction globale des similarités, qu'elles soient symétriques d'équivalence ou dissymétriques d'implication) avec l'approche VT_{100}. Cette comparaison se veut d'abord théorique ; mais aussi expérimentale par rapport à un modèle de croissance du nombre n d'observations. À cet égard, deux modèles de croissance notés M_1 et M_2 seront considérés ici. Pour M_1, relativement à un couple d'attributs booléens (a, b), les effectifs n, $n(a \wedge b)$, $n(a)$ et $n(b)$ associés à une situation réelle, sont multipliés par une constante k, k croissant. C'est ce modèle qui est considéré dans [23, 27]. Le second modèle M_2 nous a été suggéré par Y. Kodratoff [15]. Pour ce dernier modèle $n(a \wedge b)$, $n(a)$ et $n(b)$ sont invariables, seul n augmente à partir de sa valeur initiale considérée dans une situation réelle. Ce modèle se justifie de façon tout à fait pertinente compte tenu de la taille des bases de données actuelles et compte tenu du fait que pour un attribut booléen donné décrivant un aspect de la base, le nombre d'entités où cet attribut est à $VRAI$ est en général "très petit" par rapport à la taille de la base. D'autres modèles d'analyse ayant un caractère plus intrinsèque, du comportement des indices seront évoqués en section 7 réservée à la conclusion et aux perspectives de ce travail.

Dans la section 6 nous détaillerons les résultats de l'analyse expérimentale. Dans cette dernière on utilisera la base de données "Wages" [24]. On commencera par comparer les meilleures règles extraites par l'un ou l'autre des indices obtenus. On se focalisera ensuite sur la comparaison du comportement de l'indice de l'*Intensité d'Implication Contextuelle* et des différentes versions - conceptuellement bien distinctes - d'un indice de type VT_{100}. D'autre part aussi, nous considérerons un type d'indice proposé par R. Gras, dit de l'*Intensité Entropique d'Implication* qui mélange - à l'aide de la moyenne géométrique - un indice de vraisemblance du lien à caractère local (l'*Intensité d'Implication*) et un indice d'inclusion faisant appel à l'entropie de Shannon. On montrera notamment que l'influence du facteur probabiliste devient vite négligeable pour n augmentant et cela, aussi bien pour le modèle M_1 de croissance que celui

M_2. Dans l'étude expérimentale, relativement à la base de données "Wages", on se situera - dans la comparaison relative entre $\mathcal{O}(a)$ et $\mathcal{O}(b)$ -, par rapport à des situations spécifiques et bien différenciées relativement au paradigme de l'implication statistique [7].

2 Les approches VL et VT pour la similarité symétrique

Nous allons reprendre ici de façon plus précise et plus détaillée la partie concernée de l'introduction. La donnée est définie par la description d'un ensemble fini $\mathcal{O}$ d'entités par un ensemble fini $\mathcal{A}$ d'attributs booléens. Si on désigne par n le cardinal de $\mathcal{O}$ et par m celui de $\mathcal{A}$, on note

$$\mathcal{O} = \{o_i \mid 1 \leq i \leq n\} \tag{11}$$

et

$$\mathcal{A} = \{a^j \mid 1 \leq j \leq m\} \tag{12}$$

Un attribut a est représenté par le sous ensemble $\mathcal{O}(a)$ où l'attribut a est à $VRAI$. À un attribut a de $\mathcal{A}$ nous associons sa négation que nous notons $\bar{a}$. La représentation $\mathcal{O}(\bar{a})$ de l'attribut $\bar{a}$ est définie par le sous ensemble complémentaire de $\mathcal{O}(a)$ dans $\mathcal{O}$. Les cardinaux de $\mathcal{O}(a)$ et de $\mathcal{O}(\bar{a})$ sont respectivement notés $n(a)$ et $n(\bar{a})$. On a bien entendu $n(a) + n(\bar{a}) = n$.

Relativement à un couple (a, b) d'attributs booléens de $\mathcal{A} \times \mathcal{A}$, nous introduisons les conjonctions $a \wedge b$, $a \wedge \bar{b}$, $\bar{a} \wedge b$ et $\bar{a} \wedge \bar{b}$ qui sont respectivement représentés par $O(a \wedge b) = O(a) \cap O(b)$, $O(a \wedge \bar{b}) = O(a) \cap O(\bar{b})$, $O(\bar{a} \wedge b) = O(\bar{a}) \cap O(b)$ et $O(\bar{a} \wedge \bar{b}) = O(\bar{a}) \cap O(\bar{b})$. Les cardinaux de ces sous ensembles sont respectivement désignés par $n(a \wedge b)$, $n(a \wedge \bar{b})$, $n(\bar{a} \wedge b)$ et $n(\bar{a} \wedge \bar{b})$. Ces cardinaux prennent place dans la table de contingence croisant les deux attributs binaires $\{a, \bar{a}\}$ et $\{b, \bar{b}\}$. En rapportant les cardinaux à n, on définit les fréquences relatives ou proportions $p(a \wedge b)$, $p(a \wedge \bar{b})$, $p(\bar{a} \wedge b)$ et $p(\bar{a} \wedge \bar{b})$.

2.1 L'approche VL (*Vraisemblance du Lien*)

Nous avons déjà largement introduit cette approche [12, 18, 19, 20]. Relativement à deux attributs a et b de $\mathcal{A}$, on introduit un indice "brut" de similarité (on dit aussi d'association) $s(a, b)$ qui représente le nombre d'entités où les deux attributs sont à $VRAI$. Ainsi :

$$s(a, b) = n(a \wedge b) = card[\mathcal{O}(a) \cap \mathcal{O}(b)] \tag{13}$$

Pour évaluer la "grandeur" de $s(a, b)$, compte tenu des tailles $n(a)$ et $n(b)$, on introduit une hypothèse d'absence de liaison $\mathcal{N}$ qui associe au couple d'attributs observé (a, b), un couple $(a^\star, b^\star)$ d'attributs aléatoires indépendants, de telle sorte que les espérances mathématiques de $card[\mathcal{O}(a^\star)]$ et de $card[\mathcal{O}(b^\star)]$ soient respectivement égales à $n(a)$ et $n(b)$.

À cet égard, nous avons mis en évidence trois formes fondamentales de l'hypothèse d'absence de liaison que nous notons $\mathcal{N}_1, \mathcal{N}_2$ et $\mathcal{N}_3$ [21, 18]. Ces dernières se distinguent de par la manière dont on associe à un sous ensemble $\mathcal{O}(c)$ de $\mathcal{O}$, un sous ensemble aléatoire $\mathcal{O}(c^\star)$ d'un ensemble Ω correspondant à l'ensemble $\mathcal{O}$. À cet effet, désignons par $(P)(\Omega)$ l'ensemble des parties de Ω organisé en niveaux au moyen de la relation d'inclusion entre ensembles. Un niveau donné est composé de tous les sous-ensembles de Ω ayant la même cardinalité. Nous allons maintenant préciser chacun des modèles $\mathcal{N}_1$, $\mathcal{N}_2$ et $\mathcal{N}_3$.

Pour $\mathcal{N}_1$, $\Omega = \mathcal{O}$ et $\mathcal{O}(c^\star)$ est un élément aléatoire du niveau $n(c) = card(\mathcal{O}(c))$ de $(P)(\Omega)$, muni d'une distribution de probabilité uniforme.

Pour $\mathcal{N}_2$, $\Omega = \mathcal{O}$; cependant, le modèle aléatoire comprend deux pas. Le premier consiste dans le choix aléatoire d'un niveau de $\mathcal{P}(\mathcal{O})$. Le second, dans le choix aléatoire d'un élément de ce niveau, pourvu d'une distribution de probabilité uniforme. Plus précisément, le choix aléatoire du niveau suit une loi binomiale de paramètres n et $p(c) = n(c)/n$. Dans ces conditions, la probabilité de choix du k-ème niveau, $1 \leq k \leq n$, est donnée par la probabilité binomiale $C_n^k p(c)^k \times p(\bar{c})^{n-k}$, où $p(\bar{c}) = n(\bar{c})/n$.

$\mathcal{N}_3$ est défini par un modèle aléatoire à trois étapes. La première consiste à associer à $\mathcal{O}$ un ensemble aléatoire Ω. La seule exigence du modèle concerne la loi de probabilité du cardinal N de Ω. On suppose que N suit une loi de Poisson de paramètre $n = card(\mathcal{O})$. Les deux étapes suivantes sont celles du modèle $\mathcal{N}_2$. Plus précisément, pour $N = l$ et un ensemble Ω instancié par Ω_0 de taille l, $\mathcal{O}(c^\star)$ est un sous ensemble aléatoire de Ω_0. $\mathcal{O}(c^\star)$ n'est défini que pour $l \geq n(c)$ et dans ce cas nous définissons $\gamma = n(c)/l$. Dans ces conditions, la probabilité de choisir le niveau k de $\mathcal{P}(\Omega_0)$ est définie par la probabilité binomiale $C_n^k \gamma^k \times (1-\gamma)^{l-k}$. Et, pour un niveau donné, le choix aléatoire de $\mathcal{O}(c^\star)$ est fait de façon uniforme sur ce niveau.

Nous établissons dans [21, 18] que la distribution de l'indice brut aléatoire $S(a,b) = n(a^\star \wedge b^\star)$ est :
- hypergéométrique de paramètres $(n, n(a), n(b))$, pour le modèle $\mathcal{N}_1$;
- binomiale de paramètres $(n, p(a) \times p(b))$, pour le modèle $\mathcal{N}_2$;
- de Poisson de paramètres $(n, n \times p(a) \times p(b))$ pour le modèle $\mathcal{N}_3$.

À l'indice brut $s(a,b) = n(a \wedge b)$, nous associons l'indice statistiquement normalisé en centrant avec l'espérance mathématique et en réduisant avec l'écart-type, relativement au modèle aléatoire de l'hypothèse d'absence de liaison $\mathcal{N}$. Les modèles $\mathcal{N}_1$, $\mathcal{N}_2$ et $\mathcal{N}_3$ conduisent respectivement aux coefficients normalisés :

$$Q_1(a,b) = \frac{n(a \wedge b) - (n(a) \times n(b)/n)}{\sqrt{n(a) \times n(\bar{a}) \times n(b) \times n(\bar{b})/(n^2 \times (n-1))}} = \sqrt{n-1} \times \frac{p(a \wedge b) - p(a) \times p(b)}{\sqrt{p(a) \times p(\bar{a}) \times p(b) \times p(\bar{b})}} \tag{14}$$

$$Q_2(a,b) = \frac{n(a \wedge b) - (n(a) \times n(b)/n)}{\sqrt{(n(a) \times n(b)/n) \times [1 - (n(a) \times n(b)/n)]}} = \sqrt{n} \times \frac{p(a \wedge b) - p(a) \times p(b)}{\sqrt{p(a) \times p(b) \times [1 - p(a) \times p(b)]}} \tag{15}$$

$$Q_3(a,b) = \frac{n(a \wedge b) - (n(a) \times n(b)/n)}{\sqrt{n(a) \times n(b)/n}} = \sqrt{n} \times \frac{p(a \wedge b) - p(a) \times p(b)}{\sqrt{p(a) \times p(b)}} \quad (16)$$

Le coefficient $Q_1(a,b)$ est parfaitement symétrique dans le sens suivant :

$$Q_1(a,b) = Q_1(\bar{a},\bar{b}) \quad (17)$$

Le plus souvent si ce n'est le cas quasi général, les attributs booléens sont spécifiés de telle sorte que leurs fréquences relatives respectives soient inférieures à 0.5. Plus précisément, si pour un attribut booléen a, $\bar{a}$ représente l'attribut opposé, on suppose qu'on a :

$$p(a) < p(\bar{a})$$

Dans ces conditions, on a les inégalités :

$$Q_2(a,b) > Q_2(\bar{a},\bar{b}) \quad (18)$$

$$Q_3(a,b) > Q_3(\bar{a},\bar{b}) \quad (19)$$

où la dernière inégalité est plus prononcée que celle qui précède (le lecteur pourra aisément vérifier ce point). Ainsi, les deux coefficients les plus différenciés sont Q_1 et Q_3. D'ailleurs, la statistique du χ^2 associée au tableau de contingence 2×2 croisant $(a,\bar{a})$ avec $(b,\bar{b})$, peut être écrite sous l'une des deux formes suivantes :

$$\chi^2(\{(a,\bar{a}),(b,\bar{b})\}) = [Q_1(a,b)]^2 \;=\; [Q_3(a,b)]^2 + [Q_3(a,\bar{b})]^2 + [Q_3(\bar{a},b)]^2 + [Q_3(\bar{a},\bar{b})]^2 \quad (20)$$

L'indice $Q_3(a,b)$ est ce que nous appelons [18] la *contribution orientée* de la case (a,b) à la statistique du χ^2.

Conformément à ce que nous avons déjà écrit dans l'introduction, la forme générale de l'indice de "*vraisemblance du lien*" entre deux attributs booléens a et b de $\mathcal{A}$, est :

$$P_l^{\mathcal{N}} = Pr^{\mathcal{N}}\{S(a,b) \leq s(a,b)\} \quad (21)$$

où nous avons noté $s(a,b)$ pour $n(a \wedge b)$ et $S(a,b)$ pour $n(a^\star \wedge b^\star)$.

Comme nous venons de l'annoncer, nous retenons les deux formes $\mathcal{N}_1$ et $\mathcal{N}_3$ de l'hypothèse d'absence de liaison. Pour $\mathcal{N} = \mathcal{N}_1$, la probabilité (21) se calcule de façon exacte en utilisant la fonction de répartition de la loi hypergéométrique de paramètres $(n, n(a), n(b))$. Pour $\mathcal{N} = \mathcal{N}_3$, c'est la fonction de répartition de la loi de Poisson de paramètres $(n, n \times p(a) \times p(b))$ qu'on utilise. Qu'il s'agisse du cas hypergéométrique ou de celui de Poisson, moyennant d'une formule de récurrence sur la fonction logarithmique de la probabilité, le calcul exact des probabilités peut être obtenu pour des valeurs relativement élevées de n. Cependant, de façon plus pratique et suffisamment précise, on utilise le caractère

asymptotiquement normal de la loi de $S(a,b)$ quelle que soit la forme de l'hypothèse d'absence de liaison $\mathcal{N}_1$, $\mathcal{N}_2$ ou $\mathcal{N}_3$. Dans ces conditions, on obtient pour n "assez grand", une excellente approximation avec les relations :

$$P_l^{\mathcal{N}_1}(a,b) = \Phi[Q_1(a,b)] \tag{22}$$

et

$$P_l^{\mathcal{N}_3}(a,b) = \Phi[Q_3(a,b)] \tag{23}$$

où Φ désigne la fonction de répartition de la loi normale centrée et réduite.

2.2 L'approche VT ($ValeurTest$)

Cette approche s'est longtemps exprimée dans le cadre de l'association symétrique entre attributs booléens et relativement à une hypothèse d'indépendance où c'est le modèle aléatoire hypergéométrique $\mathcal{N}_1$ qui est utilisé [22, 23].

La philosophie de l'approche VL se situe délibérément dans l'optique de l'analyse des données où de toute façon l'hypothèse d'indépendance est à rejeter. Cependant, cette hypothèse a une importance cruciale pour déterminer une échelle de probabilité permettant la comparaison des liens entre attributs. Dans sa conception l'approche VT reste accrochée à la théorie des tests d'indépendance. Ainsi, un rôle essentiel est dévolu au seuil critique d'un test d'indépendance entre les deux attributs booléens. Reprenons dans ces conditions l'équation (4) dans le cadre d'une hypothèse d'absence de liaison faisant appel au modèle aléatoire $\mathcal{N}_1$.

$$\bar{P}(a,b) = 1 - P_l(a,b) = Pr^{\mathcal{N}_1}\{n(a^\star \wedge b^\star) > n(a \wedge b)\} \tag{24}$$

Comme nous l'exprimions déjà dans l'introduction, cette probabilité que nous notons ici p peut s'interpréter comme le le seuil critique de rejet de l'hypothèse nulle d'indépendance contre celle d'une association positive entre les deux attributs booléens a et b. Dans ces conditions, la valeur test VT pour (a,b) que nous pouvons noter $VT(a,b)$ est définie par "le nombre d'écarts types de la loi normale centrée et réduite qu'il faut dépasser" pour couvrir la probabilité complémentaire $1-p$ [23]. Dans ces conditions on a :

$$VT(a,b) = \Phi^{-1}(1-p) = \Phi^{-1}[P_l(a,b)] \tag{25}$$

Or l'approximation normale - en général excellente - de la loi de $S(a,b) = n(a^\star \wedge b^\star)$ sous l'hypothèse d'absence de liaison $\mathcal{N}_1$, donne

$$\Phi[Q_1(a,b)] \simeq P_l(a,b) = 1 - p \tag{26}$$

où $Q_1(a,b)$ est donné par (14). Dans ces conditions, on a la propriété importante suivante :

Proposition 1 *Au degré de précision près fourni par l'approximation normale de la loi de $S(a,b) = n(a^\star \wedge b^\star)$ sous l'hypothèse d'absence de liaison $\mathcal{N}_1$, on a pour le modèle hypergéométrique :*

$$VT(a,b) = Q_1(a,b) \tag{27}$$

De la même façon que nous l'avions déjà adoptée [21] c'est une forme Poissonnienne de l'hypothèse d'absence de liaison qui est considérée dans [27]. De façon conforme à l'argumentation ci-dessus, on obtient en tenant compte de (7) :

Proposition 2 *Au degré de précision près fourni par l'approximation normale de la loi de* $S(a,\bar{b}) = n(a^\star \wedge \bar{b}^\star)$ *sous l'hypothèse d'absence de liaison* $\mathcal{N}_3$*, on a pour le modèle Poissonnien :*

$$VT(a,\bar{b}) = -Q_3(a,\bar{b}) \tag{28}$$

3 VL normalisé et $VT100$

3.1 VL normalisé, *VLgrSymH et VLgrSymP*

Considérons d'abord le cas d'une comparaison symétrique entre attributs booléens en nous référant à l'une ou à l'autre des deux formes de l'hypothèse d'absence de liaison $\mathcal{N}_1$ ou $\mathcal{N}_3$. Dans la pratique c'est le modèle Poissonnien $\mathcal{N}_3$ qui sera retenu. En effet, toutes choses égales par ailleurs, ce dernier accentue la quantification de la ressemblance entre attributs rares.

Nous avons déjà présenté ci-dessus (sections 1 et 2) le principe et le calcul de l'indice de *vraisemblance du lien* entre deux attributs a et b de l'ensemble $\mathcal{A}$ des attributs booléens de description. Au niveau de l'introduction nous pouvons nous référer aux expressions (1)et (3) où la forme de l'hypothèse d'absence de liaison n'est pas spécifiée. Comme nous venons tout juste de le préciser nous nous référerons soit à l'hypothèse d'absence de liaison $\mathcal{N}_1$ à caractère hypergéométrique et de nature symétrique, soit à l'hypothèse d'absence de liaison $\mathcal{N}_3$ à caractère Poissonnien et de nature dissymétrique. La transposition au cas dissymétrique sera aisée à traiter.

Comme nous l'avons exprimé dans l'introduction l'indice (1) ou son calcul (3) tend pour n "assez grand" soit vers 1, soit vers 0. De façon plus précise, considérons un modèle de croissance des effectifs du tableau de contingence croisant $\{a,\bar{a}\}$ avec $\{b,\bar{b}\}$, pour lequel les proportions $p(a\wedge b)$, $p(a)$ et $p(b)$ sont invariantes. Il s'agit du modèle que nous avons noté M_1 dans l'introduction. Considérons les deux cas de figures suivants :

$$p(a\wedge b) > p(a)\times p(b) \tag{29}$$

et

$$p(a\wedge b) < p(a)\times p(b) \tag{30}$$

Dans le premier cas (29), $P_l^{\mathcal{N}}(a,b)$ tend vers 1 que $\mathcal{N} = \mathcal{N}_1$ ou que $\mathcal{N} = \mathcal{N}_3$. Plus précisément, en désignant par $r_1(a,b)$ [resp. $r_3(a,b)$] le coefficient invariant et indépendant de n, soit $Q_1(a,b)/\sqrt{n}$ [resp. $Q_3(a,b)/\sqrt{n}$], $P_l^{\mathcal{N}_1}(a,b)$ [resp. $P_3^{\mathcal{N}_1}(a,b)$] tend vers l'unité comme $\Phi[\sqrt{n}\times r_1(a,b)]$ [resp. $\Phi[\sqrt{n}\times r_3(a,b)]$]. En prenant à titre d'exemple la table de contingence de Ritschard [28] :

	a	$\bar{a}$	Total
b	48	72	120
$\bar{b}$	28	125	153
Total	76	197	273

Table 1

On a : $p(a \wedge b) = 48/273 = 0.176$, $p(a) = 76/273 = 0.278$, $p(\bar{a}) = 0.722$, $p(b) = 120/273 = 0.440$, $p(\bar{b}) = 0.560$. Dans ces conditions, $r_1(a,b) = 0.242$ et $Q_1(a,b) = 3.994$. D'autre part, $r_3(a,b) = 0.153$ et $Q_3(a,b) = 2.523$. Ainsi, $\Phi(Q_1(a,b)) \simeq 1$. et $\Phi(Q_3(a,b)) = 0.994$. En multipliant par 10 tous les effectifs du tableau de contingence les valeurs de $Q_1(a,b)$ et de $Q_3(a,b)$ deviennent respectivement égales à 12.63 et à 7.98. Les différentes bibliothèques de programmes donnent alors pour $P_l^{\mathcal{N}_3}(a,b)$ et *a fortiori* pour $P_l^{\mathcal{N}_1}(a,b)$ la valeur 1.

En vérité, comme nous l'avons déjà mentionné dans l'introduction, un indice tel que $P_l^{\mathcal{N}}(a,b)$ est conçu dans le cadre des tests statistiques d'indépendance, comme niveau de certitude quant à l'existence d'un lien, quelle que soit d'ailleurs la force de ce dernier. Ainsi, en "Fouille des Données" où le nombre d'observations est suffisamment grand on se retrouve - compte tenu aussi de la précision calcul dont on dispose - devant deux valeurs de $P_l^{\mathcal{N}}(a,b)$: 1, si $p(a \wedge b) > p(a) \times p(b)$ et 0, si $p(a \wedge b) < p(a) \times p(b)$. Alors que, comme nous l'exprimions dans l'introduction, notre optique est de nous servir de l'hypothèse d'absence de liaison comme référence pour établir une échelle de probabilité discriminante pour la comparaison des liens mutuels entre attributs. En fait, conceptuellement, un indice tel que $P_l^{\mathcal{N}}(a,b)$ a un caractère *local*. Pour cet indice la comparaison entre a et b est absolue. Elle n'a pas un caractère *relatif* dans le contexte d'un ensemble $\mathcal{A}$ d'attributs. Pour faire intervenir ce contexte, nous avons proposé de substituer à la distribution des indices

$$\{Q_\epsilon(a^j, a^k) \mid 1 \leq j < k \leq m\} \tag{31}$$

($\epsilon = 1$ ou 3, cf. (14) ou (16)), des indices centrés et réduits globalement par rapport à la moyenne et à l'écart-type de la distribution (31). On obtient la distribution des indices :

$$\{Q_{\epsilon g}(a^j, a^k) \mid 1 \leq j < k \leq m\} \tag{32}$$

($\epsilon = 1$ ou 3), où

$$Q_{\epsilon g}(a^j, a^k) = \frac{Q_\epsilon(a^j, a^k) - moy_e(Q_\epsilon)}{\sqrt{var_e(Q_\epsilon)}} \tag{33}$$

L'indice probabiliste de vraisemblance du lien *global* prend alors la forme ($\epsilon = 1$ ou 3) :

$$P_g^{\mathcal{N}_\epsilon}(a^j, a^k) = \Phi[Q_{\epsilon g}(a^j, a^k)] \tag{34}$$

où Φ est la fonction de répartition de la loi normale centrée et réduite, $1 \leq j < k \leq m$.

La distribution (32) étant de moyenne nulle et de variance unité, l'échelle

probabiliste définie par (34) devient fine et discriminante pour l'évaluation comparée des liens entre les différents attributs de $\mathcal{A}$.

Ce mode de réduction globale est apparu au milieu des années 70 et se trouve implanté dans la méthode de classification ascendante hiérarchique de l'*Analyse de la Vraisemblance des Liens Relationnels* (programme **CHAVLh**) [18, 19, 25, 14]. Ici l'hypothèse d'absence de liaison $\mathcal{N}_\epsilon$ a un caractère global. À la suite $(a^j \mid 1 \leq j \leq m)$ des attributs de $\mathcal{A}$, on associe une suite $(a^{j\star} \mid 1 \leq j \leq m)$ d'attributs aléatoires mutuellement indépendants. La justification théorique, notamment de la tendance asymptotiquement normale des variables aléatoires $Q_{\epsilon g}(a^{j\star}, a^{k\star})$, $1 \leq j \leq m$, est donnée dans [13, 3]

Donnons maintenant le sens des acronymes indiqués dans le titre. *VLgrSymH* indique : Vraisemblance du Lien, indice globalement réduit, comparaison Symétrique, modèle Hypergéométrique. Il correspond à l'indice (34) où $\epsilon = 1$. *VLgrSymP* indique : Vraisemblance du Lien, indice globalement réduit, comparaison Symétrique, modèle Poissonnien. Il correspond à l'indice (34) où $\epsilon = 0$.

3.2 $VT100$, ***VT100SymBarH***

$VT100$ a été présenté dans [23] dans le cadre de la comparaison symétrique entre attributs. D'autre part, c'est le modèle hypergéométrique qui est pris en compte dans la référence mentionnée. Nous allons ici rester dans ce cadre.

Nous avons déjà exprimé dans l'introduction le principe général de $VT100$. Relativement au test d'indépendance entre deux attributs booléens a et b, la méthode VT se focalise sur le seuil critique (la *p-value*) que nous avons pour notre part interprété sous la forme de l'expression (24). Comme nous l'avons mentionné, en cas d'association positive [cf. (29)] et pour n assez grand, le seuil critique devient "trop faible" pour être calculé et jouer le rôle d'une mesure discriminante. Dans ces conditions et pour n "grand", A. Morineau et R. Rakotomalala [23] proposent "de se ramener artificiellement au cas de nombre d'observations $n = 100$. Ils considèrent en effet que cette valeur de la taille d'un échantillon est typique de la pratique "classique" d'un test d'hypothèses. Le modèle aléatoire consiste à répéter "un grand nombre de fois le tirage d'un échantillon de taille 100". Les répétitions mutuellement indépendantes du tirage peuvent se conçevoir avec ou sans remise. À chaque tirage on associe la *p-value* du test d'indépendance qui ne concerne plus qu'un échantillon de taille 100. La moyenne des *p-values* va donner lieu à une valeur test notée $VT100$ qui sera tout à fait discriminante pour comparer les associations par paires entre attributs de $\mathcal{A}$.

En désignant par E un échantillon générique de taille 100 tiré de la base, on a à considérer une suite :

$$(E^{(1)}, E^{(2)}, ..., E^{(l)}, ..., E^{(L)}) \tag{35}$$

de L échantillons aléatoires indépendants (on suppose que le tirage est avec remise) de taille 100 chacun. Indépendamment du problème de la détermination de L, une telle procédure, comme d'ailleurs c'est admis dans [23], est très lourde.

Cependant, notre analyse théorique en section 4 de la démarche permettra à partir de justifications au titre de la statistique mathématique de proposer des variantes efficaces obtenues à partir de calculs simples.

La solution proposée par les auteurs consiste d'abord à substituer au tableau de contingence observé :

	a	$\bar{a}$	Total
b	$n(a \wedge b)$	$n(\bar{a} \wedge b)$	$n(b)$
$\bar{b}$	$n(a \wedge \bar{b})$	$n(\bar{a} \wedge \bar{b})$	$n(\bar{b})$
Total	$n(a)$	$n(\bar{a})$	n

Table 2

où n est "grand", celui ramené à 100 qui prend la forme suivante :

	a	$\bar{a}$	Total
b	$100 \times p(a \wedge b)$	$100 \times p(\bar{a} \wedge b)$	$100 \times p(b)$
$\bar{b}$	$100 \times p(a \wedge \bar{b})$	$100 \times p(\bar{a} \wedge \bar{b})$	$100 \times p(\bar{b})$
Total	$100 \times p(a)$	$100 \times p(\bar{a})$	100

Table 3

Cependant, dans ce dernier tableau où le total est ramené à 100, les contenus des cases ne sont plus des entiers, mais des décimaux. Considérons dans ces conditions le contenu de la case interne du précédent tableau concerné par la nature symétrique de l'association entre les attributs a et b, ainsi que le contenu des cases marginales qui l'encadrent. On considérera $\gamma = 100$, $p(a \wedge b)$, $\alpha = 100 \times p(a)$ et $\beta = 100 \times p(b)$. En s'appuyant sur un exemple les auteurs proposent d'encadrer le vecteur (α, β, γ) par les 8 vecteurs à composantes entières les plus proches de ce vecteur. Un principe de moyenne barycentrique permet de retrouver le vecteur (α, β, γ) à partir des 8 vecteurs. À chacun des 8 vecteurs à composantes entières, on associe la *p-value* et on détermine la valeur moyenne de ces *p-values* ; cette dernière moyenne étant pondérée conformément aux coefficients de la moyenne barycentrique mentionnée. Cette *p-value* moyenne donne alors lieu à $VT100$.

Nous allons à présent préciser formellement le calcul. À cette fin, en désignant par $[\xi]$ la partie entière d'un réel positif ξ, les 8 vecteurs encadrant le vecteur (α, β, γ) sont :

$$([\gamma], [\alpha], [\beta]) \ , \ ([\gamma], [\alpha], [\beta + 1]) \ , \ ([\gamma], [\alpha + 1], [\beta]) \ , \ ([\gamma], [\alpha + 1], [\beta + 1]) \ ,$$
$$([\gamma + 1], [\alpha], [\beta]) \ , \ ([\gamma + 1], [\alpha], [\beta + 1]) \ , \ ([\gamma + 1], [\alpha + 1], [\beta]) \ , \ ([\gamma + 1], [\alpha + 1], [\beta + 1])$$

Introduisons à présent les parties fractionnaires :

$$z = \gamma - [\gamma] \ , \ x = \alpha - [\alpha] \text{ et } y = \beta - [\beta]$$

Le développement de $[x + (1 - x)] \times [y + (1 - y)] \times [z + (1 - z)]$ donne lieu aux 8 coefficients de la moyenne barycentrique. On a :

$$(1-z)\times(1-x)\times(1-y)\times([\gamma],[\alpha],[\beta])+(1-z)\times(1-x)\times y\times([\gamma],[\alpha],[\beta+1])+$$
$$(1-z)\times x\times(1-y)\times([\gamma],[\alpha+1],[\beta])+(1-z)\times x\times y\times([\gamma],[\alpha+1],[\beta+1])+$$
$$z\times(1-x)\times(1-y)\times([\gamma+1],[\alpha],[\beta])+z\times(1-x)\times y\times([\gamma+1],[\alpha],[\beta+1])+$$
$$z\times x\times(1-y)\times([\gamma+1],[\alpha+1],[\beta])+z\times x\times y\times([\gamma+1],[\alpha+1],[\beta+1])=(\gamma,\alpha,\beta) \quad (36)$$

qui peut se mettre sous une forme condensée :

$$\sum_{(\epsilon,\eta,\zeta)\in\{0,1\}^3}[(1-\epsilon)+(-1)^{1+\epsilon}\times z]\times[(1-\eta)+(-1)^{1+\eta}\times x]\times[(1-\zeta)+(-1)^{1+\zeta}\times y]\times$$
$$([\gamma+\epsilon],[\alpha+\eta],[\beta+\zeta])=(\gamma,\alpha,\beta) \quad (37)$$

Ces équations sont dûes à la relation :

$$(1-(\xi-[\xi]))\times[\xi]+(\xi-[\xi])\times[\xi+1]=\xi \quad (38)$$

où ξ est un réel positif.

Désignons à présent par $p_{\epsilon\eta\zeta}$ le seuil critique (la *p-value*) associé au vecteur $([\gamma+\epsilon],[\alpha+\eta],[\beta+\zeta])$ où $(\epsilon,\eta,\zeta)\in\{0,1\}^3$, on considère la moyenne p_{cent} des 8 seuils critiques $p_{\epsilon\eta\zeta}$, pondérée par les coefficients du précédent développement ; soit :

$$\sum_{(\epsilon,\eta,\zeta)\in\{0,1\}^3}[(1-\epsilon)+(-1)^{1+\epsilon}\times z]\times[(1-\eta)+(-1)^{1+\eta}\times x]\times[(1-\zeta)+(-1)^{1+\zeta}\times y]\times p_{\epsilon\eta\zeta}$$
$$=p_{cent} \quad (39)$$

Dans ces conditions, $VT100$ est proposé sous la forme :

$$VT100=\Phi^{-1}(1-p_{cent}) \quad (40)$$

Exemple :

Considérons l'exemple du tableau de contingence suivant donné dans [23] :

	a	$\bar{a}$	Total
b	226		120
$\bar{b}$			
Total	346		2000

Table 4

Ramené à 100, ce tableau donne :

	a	$\bar{a}$	Total
b	11.30		28.40
$\bar{b}$			
Total	17.30		100

Table 5

Par rapport aux notations ci-dessus, on a :

$$\gamma = 11.30 \ , \ \alpha = 17.30 \text{ et } \beta = 28.40$$

D'autre part,

$$z = 0.30 \ , \ x = 0.30 \text{ et } y = 0.40$$

Les 8 vecteurs à composantes entières encadrant (γ, α, β) sont :

$$(11, 17, 28), (11, 17, 29), (11, 18, 28), (11, 18, 29), (12, 17, 28), (12, 17, 29), (12, 18, 28), (12, 18, 29).$$

Comme dans [23], nous allons nous référer au modèle hypergéométrique pour déterminer la suite des 8 valeurs de la *p-value*, respectivement associées à la suite des 8 vecteurs précédents. Nous obtenons :

$$p_{000} = 0.32 \times 10^{-4} \ , \ p_{001} = 0.72 \times 10^{-4} \ , \ p_{010} = 1.16 \times 10^{-4} \ , \ p_{011} = 1.6 \times 10^{-4} \ ,$$
$$p_{100} = 0.03 \times 10^{-4} \ , \ p_{101} = 0.03 \times 10^{-4} \ , \ p_{110} = 0.21 \times 10^{-4} \ , \ p_{111} = 0.28 \times 10^{-4}.$$

De sorte que

$$\begin{aligned} p_{cent} = 10^{-4} \times (&0.7 \times 0.7 \times 0.6 \times 0.32 + 0.7 \times 0.7 \times 0.4 \times 0.72 + \\ &0.7 \times 0.3 \times 0.6 \times 1.16 + 0.7 \times 0.3 \times 0.4 \times 1.6 + \\ &0.3 \times 0.7 \times 0.6 \times 0.03 + 0.3 \times 0.7 \times 0.4 \times 0.03 + \\ &0.3 \times 0.3 \times 0.6 \times 0.21 + 0.3 \times 0.3 \times 0.4 \times 0.28) \\ &= 10^{-4} \times 0.54 \end{aligned}$$

Il en résulte une valeur de 3.9 de $VT100$. Ainsi ici, la réduction à 100 conduit à une forte valeur.

Donnons ici le sens de l'acronyme *VT100SymBarH* : Valeur Test basée sur un échantillon de taille 100, comparaison Symétrique, approche Barycentrique, modèle Hypergéométrique. Cet indice est défini par (40).

4 Analyse de l'approche $VT100$, proposition de variantes simples et efficaces

4.1 Une réduction corrélative non linéaire et non paramétrique, *VT100SymCorH*

Nous gardons ici le caractère symétrique de l'association entre les deux attributs a et b. La technique adoptée dans [23, 27] est focalisée sur d'abord, l'évaluation d'une *p-value* sur un échantillon de taille 100 qui représenterait une sorte de *moyenne* de l'ensemble des échantillons de taille 100 dans la population formée par l'échantillon initial de taille importante. La valeur $VT100$ est alors

déterminée *a posteriori*.

La technique que nous proposons ici est focalisée sur la détermination directe de la valeur test qui n'est autre que l'indice $Q_1(a,b)$ (voir Proposition 1). Il en résultera pour $VT100$ un calcul plus direct et plus simple. Introduisons ici les fonctions indicatrices des sous ensembles $\mathcal{O}(a)$ et $\mathcal{O}(b)$ (voir sections 1 et 2) que nous notons également sans risque de confusion, a et b : $a(i) = 1$ (resp. 0) si et seulement si l'attribut a est à $VRAI$ (resp. $FAUX$) sur l'objet o_i, $1 \leq i \leq n$; de même, $b(i) = 1$ (resp. 0) si et seulement si l'attribut b est à $VRAI$ (resp. $FAUX$) sur l'objet o_i, $1 \leq i \leq n$. Reprenons avec des notations que nous avions eues, les paramètres de la table 2. On a en désignant par I l'ensemble des indices $\{1, 2, ..., n\}$:

$$\begin{aligned} s = n(a \wedge b) &= \sum_{i \in I} a(i) \times b(i) \\ u = n(a \wedge \bar{b}) &= \sum_{i \in I} a(i) \times [1 - b(i)] \\ v = n(\bar{a} \wedge b) &= \sum_{i \in I} [1 - a(i)] \times b(i) \\ t = n(\bar{a} \wedge \bar{b}) &= \sum_{i \in I} [1 - a(i)] \times [1 - b(i)] \end{aligned} \tag{41}$$

On a :

$$\begin{aligned} s + u = n(a) \ , \ s + v &= n(b) \\ v + t = n(\bar{a}) \ , \ u + t &= n(\bar{b}) \\ s + u + v + t &= n \end{aligned}$$

Inversement, la donnée du système :

$$\begin{aligned} s = n(a \wedge b) &= \sum_{i \in I} a(i) \times b(i) \\ n(a) &= \sum_{i \in I} a(i) \\ n(b) &= \sum_{i \in I} b(i) \end{aligned} \tag{42}$$

est équivalente au système (41). La construction de l'indice se fondera le plus directement sur ce dernier système.

L'hypothèse d'absence de liaison à caractère hypergéométrique peut prendre avec les expressions analytiques ci-dessus, une forme permutationnelle. À l'indice brut s on associe l'un des trois indices bruts aléatoires équivalents sur le plan de leurs distributions respectives :

$$S_{ab(\sigma)} = \sum_{i\in I} a(i) \times b[\sigma(i)]$$
$$S_{a(\sigma)b} = \sum_{i\in I} a[\sigma(i)] \times b(i)$$
$$S_{a(\sigma)b(\tau)} = \sum_{i\in I} a[\sigma(i)] \times b[\tau(i)] \tag{43}$$

où σ et τ sont deux permutations aléatoires indépendantes prises dans l'ensemble G_n des permutations sur I (il y en a $n!$), muni d'une probabilité uniforme. En désignant par S l'une ou l'autre des trois variables aléatoires équivalentes (on peut prendre la première pour des raisons de simplicité), on a en désignant par $\mathcal{E}$ l'espérance mathématique et par var la variance [18] :

$$\mathcal{E}(S) = n \times moy(a) \times moy(b)$$
$$var(S) = \frac{n^2}{n-1} \times var(a) \times var(b) \tag{44}$$

où

$$moy(a) = p(a) = n(a)/n \ , \ moy(b) = p(b) = n(b)/n \ ,$$
$$var(a) = p(a) \times p(\bar{a}) = n(a) \times n(\bar{a})/n^2 \ , \ var(b) = p(b) \times p(\bar{b}) = n(b) \times n(\bar{b})/n^2.$$

Ainsi, on retrouve exactement pour l'indice centré et réduit

$$Q_1(a,b) = \frac{s - \mathcal{E}(S)}{\sqrt{var(S)}} \tag{45}$$

où l'expression de $Q_1(a,b)$ est donnée en (14).

Le passage d'une table telle que Table 4 à celle telle que Table 5, peut être obtenu en adoptant une mesure de présence (resp. d'absence) d'un attribut booléen c sur un individu, égale à $\sqrt{100/n} \times c(i)$ (resp. 0). Pour ce codage, la valuation indicatrice de l'appartenance d'un élément i à $\mathcal{O}(c)$ (resp. $\mathcal{O}(\bar{c})$) est $\sqrt{100/n} \times c(i)$ (resp. $\sqrt{100/n} \times [1 - c(i)]$). Dans ces conditions, les paramètres s , u , v et t ci-dessus [voir (41)] deviennent s_r , u_r , v_r et t_r (r pour réduit) :

$$s_r = \sum_{i\in I} (\sqrt{100/n} \times a(i)) \times (\sqrt{100/n} \times b(i)) = \frac{100}{n} \times s$$
$$u_r = \sum_{i\in I} (\sqrt{100/n} \times a(i)) \times (\sqrt{100/n} \times (1 - b(i))) = \frac{100}{n} \times u$$
$$v_r = \sum_{i\in I} (\sqrt{100/n} \times (1 - a(i))) \times (\sqrt{100/n} \times b(i)) = \frac{100}{n} \times v$$
$$t_r = \sum_{i\in I} (\sqrt{100/n} \times (1 - a(i))) \times (\sqrt{100/n} \times (1 - b(i))) = \frac{100}{n} \times t \tag{46}$$

En notant $a'(i) = \sqrt{100/n} \times a(i)$ et $b'(i) = \sqrt{100/n} \times b(i)$, on a les relations :

$$\sum_{i\in I} a'(i) = s_r + u_r = 100 \times \frac{n(a)}{n}$$
$$\sum_{i\in I} b'(i) = s_r + v_r = 100 \times \frac{n(b)}{n}$$

D'autre part, il est aisé de voir que le système d'équation (46) est équivalent à celui :

$$\begin{aligned}\sum_{i\in I} a'(i) \times b'(i) = s_r = \frac{100}{n} \times s \\ \sum_{i\in I} a'(i) = \frac{100}{n} \times n(a) \\ \sum_{i\in I} b'(i) = \frac{100}{n} \times n(b)\end{aligned} \tag{47}$$

En partant de ces paramètres et en appliquant la démarche permutationnele ci- dessus, on aboutit exactement au même coefficient d'association (45) et cela parce que l'ensemble observé comprend n éléments.

Dans ces conditions, nous allons procéder à la construction d'un ensemble virtuel Ω formé de 100 macros individus :

$$\Omega = \{\omega_j \mid 1 \leq j \leq 100\}$$

sur lequel sont définis deux attributs α et β à valeurs dans l'intervalle $[0, 1]$, tels que :

$$\begin{aligned}\sum_{1\leq j\leq 100} \alpha(j) \times \beta(j) = s_r \\ \sum_{1\leq j\leq 100} \alpha(j) \times [1 - \beta(j)] = u_r \\ \sum_{1\leq j\leq 100} [1 - \alpha(j)] \times \beta(j) = v_r \\ \sum_{1\leq j\leq 100} [1 - \alpha(j)] \times [1 - \beta(j)] = t_r\end{aligned}$$

Ce système est équivalent au suivant :

$$\begin{aligned}\sum_{1\leq j\leq 100} \alpha(j) \times \beta(j) = s_r \\ \sum_{1\leq j\leq 100} \alpha(j) = s_r + u_r = \frac{100}{n} \times n(a) \\ \sum_{1\leq j\leq 100} \beta(j) = s_r + v_r = \frac{100}{n} \times n(b)\end{aligned} \tag{48}$$

C'est à partir de ce système que le coefficient d'association est obtenu conformément à la démarche permutationnelle ci-dessus [voir 43, (45)]. À l'indice brut s_r nous associons la variable aléatoire permutationnelle :

$$S_r = \sum_{1 \leq j \leq 100} \alpha(j) \times \beta[\sigma(j)] \tag{49}$$

où σ est un élément aléatoire pris dans l'ensemble G_{100} des permutations sur $(1, 2, ..., j, ..., 100)$, muni d'une probabilité uniforme. L'indice s_r centré et réduit se met sous la forme :

$$Q_1(\alpha, \beta) = \frac{s_r - \mathcal{E}(S_r)}{\sqrt{var(S_r)}} \tag{50}$$

On a pour l'espérance mathématique et la variance de S_r :

$$\begin{aligned} \mathcal{E}(S_r) &= 100 \times moy(\alpha) \times moy(\beta) \\ var(S_r) &= \frac{100^2}{99} \times var(\alpha) \times var(\beta) \end{aligned} \tag{51}$$

Compte tenu des deux dernières relations (48), on a :

$$\begin{aligned} moy(\alpha) &= \frac{n(a)}{n} = moy(a) \\ moy(\beta) &= \frac{n(b)}{n} = moy(b) \end{aligned}$$

Le modèle que nous allons considérer pour l'ensemble virtuel Ω est tel que les variances des variables α et β qui sont à valeurs dans l'intervalle $[0, 1]$ soient maximales ; c'est-à-dire, telles qu'on ait :

$$\sum_{1 \leq j \leq 100} \alpha(j)^2 \ \textit{maximum}$$

et

$$\sum_{1 \leq j \leq 100} \beta(j)^2 \ \textit{maximum} \tag{52}$$

Soit le cube $[0, 1]^m$ où m est un entier et soit $\vec{\gamma} = (\gamma_1, \gamma_2, ..., \gamma_j, ..., \gamma_m)$ un point du cube tel que $\sum_{1 \leq j \leq m} \gamma_j = m(c)$ où $m(c) \leq m$ est fixé. On a alors :

Lemma 1 *$\sum_{1 \leq j \leq m} \gamma_j^2$ est maximum si et seulement si toutes les composantes du vecteur $\vec{\gamma}$ sont égales à 0 ou à 1, à l'exception éventuelle d'une seule qui vaut $m(c) - [m(c)]$, où $[m(c)]$ désigne la partie entière de $m(c)$.*

Preuve

En effet, s'il y avait deux composantes γ_1 et γ_2 - pour fixer les idées - telles que $0 < \gamma_1 < 1$ et $0 < \gamma_2 < 1$, deux cas peuvent se présenter :

$$(i)\ \gamma_1 + \gamma_2 > 1$$
$$(ii)\ \gamma_1 + \gamma_2 < 1$$

Dans le premier cas, en remplaçant (γ_1, γ_2) par $(1, \gamma_1 + \gamma_2 - 1)$, la somme des composantes reste constante,

$$0 < \gamma_1 + \gamma_2 - 1 \text{ et } \gamma_1^2 + \gamma_2^2 < 1 + (\gamma_1 + \gamma_2 - 1)^2 = \gamma_1^2 + \gamma_2^2 + 2(1 - \gamma_1)(1 - \gamma_2)$$

Dans le second cas, en remplaçant (γ_1, γ_2) par $(0, \gamma_1 + \gamma_2)$, la somme des composantes reste constante,

$$0 < \gamma_1 + \gamma_2 < 1 \text{ et } \gamma_1^2 + \gamma_2^2 < (\gamma_1 + \gamma_2)^2.$$

Maintenant, si toutes les composantes sont égales à 0 ou à 1 à l'exception de γ_1 et γ_2 et que $\gamma_1 + \gamma_2$ est fractionnaire, reprenons les deux cas ci dessus (i) et (ii). Dans le premier cas il subsiste une dernière composante $\gamma_1 + \gamma_2 - 1 = m(c) - [m(c)]$. Dans le deuxième cas, il subsiste également une dernière composante $\gamma_1 + \gamma_2 = m(c) - [m(c)]$.

Comme le cube $[0, 1]^m$ est un compact, on a bien le résultat de maximalité. □

Theorem 1 *Si $n(a)$ et $n(b)$ sont des multiples entiers de $\frac{n}{100}$, alors, sous le modèle de variance maximale α et β, on a :*

$$Q_1(\alpha, \beta) = \sqrt{\frac{99}{n-1}} \times Q_1(a, b)$$

En effet, compte tenu du lemme précédent, on a :

$$\sum_{1 \le j \le 100} \alpha(j)^2 = \sum_{1 \le j \le 100} \alpha(j)$$

et

$$\sum_{1 \le j \le 100} \beta(j)^2 = \sum_{1 \le j \le 100} \beta(j)$$

et les variances respectives des variables α et β s'écrivent :

$$var(\alpha) = \frac{s_r + u_r}{100} \times \frac{t_r + v_r}{100}$$

$$var(\beta) = \frac{s_r + v_r}{100} \times \frac{t_r + u_r}{100}$$

Ainsi $var(\alpha) = var(a)$ et $var(\beta) = var(b)$. α et β étant deux booléens, on peut aisément voir que :

$$\frac{p(\alpha \wedge \beta) - p(\alpha) \times p(\beta)}{\sqrt{p(\alpha) \times p(\bar{\alpha}) \times p(\beta) \times p(\bar{\beta})}} = \frac{p(a \wedge b) - p(a) \times p(b)}{\sqrt{p(a) \times p(\bar{a}) \times p(b) \times p(\bar{b})}}$$

D'autre part, si

$$Q_1(a,b) = \sqrt{n-1} \times \frac{p(a \wedge b) - p(a) \times p(b)}{\sqrt{p(a) \times p(\bar{a}) \times p(b) \times p(\bar{b})}},$$

alors

$$Q_1(\alpha,\beta) = \sqrt{99} \times \frac{p(\alpha \wedge \beta) - p(\alpha) \times p(\beta)}{\sqrt{p(\alpha) \times p(\bar{\alpha}) \times p(\beta) \times p(\bar{\beta})}}$$

D'où le résultat. □

Nous avons justifié de façon exacte l'expression (50) de l'indice dans le cas d'un modèle virtuel d'un ensemble formé de 100 éléments pour lequel on a les équations (48), où $var(\alpha)$ et $var(\beta)$ sont maximales et où $n(a)$ et $n(b)$ sont des multiples entiers de $\frac{n}{100}$. Une telle justification reste très précise dans le cas où la dernière condition n'est pas satisfaite. En effet, les variances maximales proposées pour α et β restent très voisines si on remplaçait la seule composante fractionnaire (strictement comprise entre 0 et 1) du vecteur $\vec{\alpha}$ (resp. $\vec{\beta}$) par 0. Pour s'en rendre compte, désignons par $\vec{\gamma}$ un vecteur dont toutes les composantes sont booléennes sauf une seule qu'on notera ϵ telle que $0 < \epsilon < 1$. On notera $[\vec{\gamma}]$ le vecteur obtenu à partir de $\vec{\gamma}$ en remplaçant la composante ϵ par 0. Les variables associées à $\vec{\gamma}$ et à $[\vec{\gamma}]$ seront notées γ et $[\gamma]$. Pour une expression plus générale nous indiquerons ci-dessous par e le nombre commun de composantes des vecteurs $\vec{\gamma}$ et $[\vec{\gamma}]$.

Proposition 3 $|var(\gamma) - var([\gamma])| < \frac{1}{e}$

Commençons par noter respectivement, $e(\gamma)$ et $e([\gamma])$ la somme des composantes des vecteurs $\vec{\gamma}$ et $[\vec{\gamma}]$. Clairement, on a : $\epsilon = e(\gamma) - e([\gamma])$. On a dans ces conditions :

$$\begin{aligned} var(\gamma) &= \frac{1}{e} \times \{e([\gamma]) + \epsilon^2\} - \{\frac{1}{e} \times [e([\gamma]) + \epsilon]\}^2 \\ &= \frac{1}{e} \times e([\gamma]) - \frac{1}{e^2} \times e([\gamma])^2 + \frac{1}{e} \times \epsilon^2 - \frac{1}{e^2} \times [\epsilon^2 + 2e([\gamma])\epsilon] \\ &= var([\gamma]) + \frac{\epsilon}{e} \times \{\epsilon(1 - \frac{1}{e}) - 2 \times \frac{e([\gamma])}{e}\} \end{aligned} \quad (53)$$

Le complément de $var([\gamma])$ dans le dernier second membre peut se mettre sous la forme :

$$\frac{1}{e^2} \times \{(e-1)\epsilon^2 - 2e([\gamma])\epsilon\}$$

Ce trinôme du second degré en ϵ est décroissant pour $\epsilon \leq e([\gamma])/(e-1)$ et croissant pour $\epsilon > e([\gamma])/(e-1)$. Ainsi, la valeur minimale - pour $\epsilon = e([\gamma])/(e-1)$ - peut s'écrire $-\frac{1}{e-1} \times (\frac{e([\gamma])}{e})^2$. Comme $e([\gamma]) \leq e-1$; la valeur absolue de la dernière quantité est inférieure ou égale à $\frac{e-1}{e^2}$ qui est strictement inférieur à $\frac{1}{e}$.

Maintenant, la valeur maximale du trinôme du second degré a lieu pour

$\epsilon = 1$. Sa valeur maximale prend la forme $\frac{1}{e^2} \times \{e - 1 - 2e([\gamma])\}$. Comme $0 \leq e([\gamma]) \leq e - 1$, on obtient que la valeur absolue de cette valeur maximale est strictement inférieure à $\frac{1}{e}$. □

Exemple Considérons le passage de la table 4 concernant le croisement $\{a, \bar{a}\} \times \{b, \bar{b}\}$ à celle de la table 5 concernant le croisement que nous avons noté ici $\{\alpha, \bar{\alpha}\} \times \{\beta, \bar{\beta}\}$ et où $e = 100$. On a $e([\alpha]) = 17$ et $\epsilon = 0.30$. On obtient :

$$var(\alpha) = \frac{1}{100}(17 + 0.3^2) - [\frac{1}{100} \times (17 + 0.3)]^2 = 0.140971$$

$$var([\alpha]) = \frac{17}{100} \times \frac{83}{100} = 0.1411$$

La différence est 0.000129. Toutefois, la variance telle qu'elle est calculée dans l'indice adopte la même forme que dans le cas booléen, elle s'écrit :

$$\frac{17.3}{100} \times \frac{82.7}{100} = 0.143071$$

Dans ce cas la différence par rapport à la vraie valeur est : 0.0021.

Maintenant, en ce qui concerne la variable β, on a $e([\beta]) = 28$ et $\epsilon = 0.40$. On obtient :

$$var(\beta) = \frac{1}{100}(28 + 0.4^2) - [\frac{1}{100} \times (28 + 0.4)]^2 = 0.200944$$

$$var([\beta]) = \frac{28}{100} \times \frac{72}{100} = 0.2016$$

et donc :

$$var([\beta]) - var(\beta) = 0.000656$$

La variance telle qu'elle est proposée pour le calcul de l'indice est :

$$\frac{28.4}{100} \times \frac{71.6}{100} = 0.20344$$

Dans ce cas la différence par rapport à la vraie valeur est : 0.0024.

Dans les formes adoptées pour la variance de α et la variance de β ; qui correspondent d'ailleurs à d'excellentes approximations, on a, en se référant à l'expression (14) ci-dessus :

$$\frac{p(a \wedge b) - p(a) \times p(b)}{\sqrt{p(a) \times p(\bar{a}) \times p(b) \times p(\bar{b})}} = \frac{p(\alpha \wedge \beta) - p(\alpha) \times p(\beta)}{\sqrt{p(\alpha) \times p(\bar{\alpha}) \times p(\beta) \times p(\bar{\beta})}}$$

où p désigne un rapport relativement au nombre d'observations, qui est de n pour le premier membre et de $e = 100$ pour le second membre. Les paramètres du second membre s'obtiennent à partir des relations (48). La valeur commune du coefficient est égale à 0.37. En adoptant les expressions ci-dessus de $Q_1(a, b)$ et de $Q_1(\alpha, \beta)$, on obtient pour la table 4, $Q_1(a, b) = 16.74$ et pour la table 5, $Q_1(\alpha, \beta) = 3.73$. Cette valeur est quelque peu plus petite que celle 3.9 obtenue dans le cadre de la technique précédente (voir fin de la sous section 3.1).

Comme dans [23] nous nous sommes fondés sur le modèle hypergéométrique de l'hypothèse d'absence de liaison. Nous aurions pu également considérer le modèle Poissonnien de l'hypothèse d'absence de liaison. C'est précisément ce que nous considérerons dans la section suivante ; notamment pour des raisons de simplicité calcul et d'ailleurs, dans la pratique, il s'agira du modèle définitivement retenu.

Indiquons à présent le sens de l'acronyme *VT100SymCorH* : $VT100$ a déjà été précisé, *Sym* pour comparaison pour comparaison Symétrique, *Cor* pour approche corrélative et H pour modèle Hypergéométrique.

4.2 Une réduction par projection sur un ensemble aléatoire, $VT100SymProj$.

Dans les indices proposés pour $VT100$ (voir sous-sections 3.1 et 4.1 ci-dessus) il y a une certaine déconnexion entre la définition conceptuelle - d'ailleurs approximative - telle qu'elle est exprimée dans [23, 27]. Par rapport à la suite définie dans (35), du fait qu'on adopte la forme "Poissonnienne", l'indice $VT100$ s'écrit comme suit :

$$\frac{1}{L}\sum_{1\leq l\leq L}\frac{card[\mathcal{O}(a)\cap\mathcal{O}(b)\cap E^{(l)}]-\frac{card[\mathcal{O}(a)\cap E^{(l)}]\times card[\mathcal{O}(b)\cap E^{(l)}]}{card(E^{(l)})}}{\sqrt{\frac{card[\mathcal{O}(a)\cap E^{(l)}]\times card[\mathcal{O}(b)\cap E^{(l)}]}{card(E^{(l)})}}} \tag{54}$$

où $(E^{(1)}, E^{(2)}, ..., E^{(l)}, ..., E^{(L)})$ est une suite de L parties aléatoires indépendantes de l'ensemble $\mathcal{O}$, de même cardinal 100 et où L est "grand".

Une expression précise de l'indice voulu est la suivante :

$$\mathcal{E}\Big(\frac{card[\mathcal{O}(a)\cap\mathcal{O}(b)\cap E^{\star}]-\frac{card[\mathcal{O}(a)\cap E^{\star}]\times card[\mathcal{O}(b)\cap E^{\star}]}{card(E^{\star})}}{\sqrt{\frac{card[\mathcal{O}(a)\cap E^{\star}]\times card[\mathcal{O}(b)\cap E^{\star}]}{card(E^{\star})}}}\Big) \tag{55}$$

où $\mathcal{E}$ désigne l'espérance mathématique et où $E^{\star}$ est un sous ensemble aléatoire de taille 100 pris dans l'ensemble des parties de $\mathcal{O}$ de même cardinal 100, muni d'une probabilité uniformément répartie.

À défaut de calculer l'espérance mathématique du rapport, nous allons nous contenter d'en calculer une forme "approchée" à partir du calcul de $\mathcal{E}(card[\mathcal{O}(a)\cap\mathcal{O}(b)\cap E^{\star}])$ et de $\mathcal{E}(card[\mathcal{O}(a)\cap E^{\star}]\times card[\mathcal{O}(b)\cap E^{\star}])$. Ce calcul permet de donner une valeur exacte de l'espérance mathématique du numérateur. Appelons ici e le cardinal de $E^{\star}$.

Proposition 4

$$\mathcal{E}(card[\mathcal{O}(a)\cap\mathcal{O}(b)\cap E^{\star}]) = \frac{n(a\wedge b)\times e}{n} = e\times p(a\wedge b) \tag{56}$$

$$\mathcal{E}(card[\mathcal{O}(a)\cap E^{\star}]\times card[\mathcal{O}(b)\cap E^{\star}]) = \frac{e}{n}\times n(a\wedge b)+\frac{e(e-1)}{n(n-1)}\times\Big(n(a)\times n(b)-n(a\wedge b)\Big) \tag{57}$$

La première relation est classique et correspond à l'espérance mathématique d'un indice brut aléatoire. Établissons directement la seconde relation. À cette fin, introduisons les fonction indicatrices 1_A, 1_B et $1_{E^\star}$ des trois sous ensembles $\mathcal{O}(a)$, $\mathcal{O}(b)$ et $E^\star$. On peut écrire :

$$card[\mathcal{O}(a) \cap E^\star] = \sum_{x \in \mathcal{O}} 1_A(x) \times 1_{E^\star}(x) \tag{58}$$

$$card[\mathcal{O}(b) \cap E^\star] = \sum_{x \in \mathcal{O}} 1_B(x) \times 1_{E^\star}(x) \tag{59}$$

En désignant par $\mathcal{O}^{[2]}$ l'ensemble des couples d'objets à composantes distinctes et par $[x, y]$ un élément générique de cet ensemble, on a :

$$card[\mathcal{O}(a) \cap E^\star] \times card[\mathcal{O}(b) \cap E^\star] =$$
$$\sum_{x \in \mathcal{O}} 1_A(x) \times 1_B(x) \times 1_{E^\star}(x) + \sum_{[x,y] \in \mathcal{O}}^{[2]} 1_A(x) \times 1_B(y) \times 1_{E^\star}(x) \times 1_{E^\star}(y) \tag{60}$$

En tenant compte des relations :

$$\mathcal{E}[1_{E^\star}(x)] = \frac{e}{n}$$
$$\mathcal{E}[1_{E^\star}(x) \times 1_{E^\star}(y)] = \frac{e(e-1)}{n(n-1)}$$
$$\sum_{x \in \mathcal{O}} 1_A(x) \times 1_B(y) = n(a \wedge b)$$

et

$$\sum_{[x,y] \in \mathcal{O}^{[2]}} 1_A(x) \times 1_B(y) = [n(a) \times n(b) - n(a \wedge b)]$$

on obtient l'espérance mathématique de (60) :

$$\frac{e}{n} \times n(a \wedge b) + \frac{e(e-1)}{n(n-1)} \times [n(a) \times n(b) - n(a \wedge b)] \tag{61}$$

D'où le résultat annoncé.

Considérons l'indice aléatoire centré et réduit qui est sous le signe espérance dans (55) ; mais où $e = card(E^\star)$. L'espérance mathématique du numérateur se met sous la forme :

$$e \times \frac{n(a \wedge b)}{n} - \frac{e(e-1)}{n(n-1)} \times [n(a) \times n(b) - n(a \wedge b)]$$
$$= \frac{1}{n(n-1)} \times [(ne-1) \times n(a \wedge b) - (e-1) \times n(a) \times n(b)] \tag{62}$$

En remplaçant dans le dénominateur et sous le signe radical, l'élément aléatoire par son espérance mathématique, on obtient après calcul, une version nouvelle de VT_e :

$$\frac{\frac{1}{n(n-1)} \times [(ne-1) \times n(a \wedge b) - (e-1) \times n(a) \times n(b)]}{\sqrt{\frac{n(a \wedge b)}{n} + \frac{e(e-1)}{n(n-1)} \times [n(a) \times n(b) - n(a \wedge b)]}}$$
$$= \frac{\frac{1}{\sqrt{n(n-1)}} \times [(ne-1) \times n(a \wedge b) - (e-1) \times n(a) \times n(b)]}{\sqrt{(n-e) \times n(a \wedge b) + (e-1) \times n(a) \times n(b)}} \quad (63)$$

En considérant les effectifs de la table 4 ($n = 2000, n(a) = 346, n(b) = 568$ et $n(a \wedge b) = 226$) on obtient pour ce nouvel indice la valeur de 2.89. Cette valeur est notablement plus petite que celle 3.9 considérée ci-dessus ; est-ce parce que la conception mathématique correspond de façon beaucoup plus étroite à celle, exprimée intuitivement dans [23, 27]. On peut à titre indicatif signaler que l'indice calculé conformément à la sous section 4.1 ci-dessus, dans le cadre du modèle Poissonnien de l'hypothèse d'absence de liaison vaut 2.87.

5 Les approches VL et VT pour la similarité implicative ; le cas normalisé

5.1 Les approches VL et VT

Nous avons déjà en section 1 largement introduit ces approches dans le cas de la similarité probabiliste implicative et donc dissymétrique. Rappelons que par rapport à la similarité probabiliste symétrique et d'équivalence entre les deux attributs booléen a et b, au lieu de se poser la question "combien $n(a \wedge b)$ est invraisemblablement grand", on se pose la question "combien $n(a \wedge \bar{b})$ est invraisemblablement petit. L'idée de cette transposition est dûe à R. Gras [5]. Il s'agissait en quelque sorte de transposer ce qui était élaboré dans le cas symétrique [12, 17, 16]. L'hypothèse d'absence de liaison associe au couple d'attributs booléens (a, b), un couple d'attributs booléens aléatoire $(a^\star, b^\star)$ de telle sorte que les espérances mathématiques de $card[\mathcal{O}(a^\star)]$ et de $card[\mathcal{O}(b^\star)]$ soient respectivement égales à $n(a)$ et $n(b)$. L'analyse que nous avons effectuée dans [21] montre que c'est le modèle Poissonnien de l'hypothèse d'absence de liaison qui est pertinent pour l'évaluation de la similarité probabiliste. Nous l'avions déjà adopté dans le cas symétrique [17] car ce modèle met l'accent sur la rareté de la concomittance. Nous nous limiterons donc dans la suite à ce modèle qui a été validé aussi bien formellement que par l'expérience. Dans ces conditions, reprenons à partir de l'introduction la notion d"'Intensité d'Implication" qui correspond exactement à un indice de *Vraisemblance du Lien* sous la forme :

$$\mathcal{I}(a \rightarrow b) = Pr^{\mathcal{N}_3}\{n(a^\star \wedge \bar{b^\star}) > n(a \wedge \bar{b})\} \quad (64)$$

où $\mathcal{N}_3$ est l'hypothèse d'absence de liaison de Poisson définie dans [21].

$n(a^\star \wedge \bar{b^\star})$ suit une loi de Poisson de paramètre $n(a) \times n(\bar{b})/n$. Comme nous l'avons déjà exprimé, le calcul exact du second membre de (64) peut être obtenu jusqu'à des valeurs importantes de n. Cependant, toujours comme nous

l'avons exprimé dans l'introduction, l'excellente approximation normale de la loi de Poisson donne :

$$\mathcal{I}(a \rightarrow b) = 1 - \Phi[Q_3(a, \bar{b})] = \Phi[-Q_3(a, \bar{b})] \tag{65}$$

On reprend ici l'expression (7) et $Q_3(a, \bar{b})$) est donné par l'expression (6) qui correspond à l'indice $n(a \wedge \bar{b})$ centré et réduit.

Continuons à reprendre ce qui est exprimé dans l'introduction. p désignant le seuil critique du test de l'hypothèse d'indépendance contre celle de l'implication :

$$p = Pr^{\mathcal{N}_3}\{n(a^\star \wedge \bar{b^\star}) \leq n(a \wedge \bar{b})\} \tag{66}$$

la valeur test VT ("nombre d'écarts types de la loi normale centrée et réduite qu'il faut dépasser pour couvrir la probabilité $1 - p$") peut s'écrire $\Phi^{-1}(1 - p)$ et n'est autre que $Q_3(a, \bar{b})$ (voir Proposition 1 ci-dessus).

5.2 Le cas normalisé VL globalement réduit et $VT100$

5.2.1 VL globalement réduit : l'Intensité d'Implication Contextuelle, $VLgrImpP$

Le principe de l'élaboration de cet indice a été présenté en section 1. Reprenons ici l'ensemble $\mathcal{A}$ des attributs booléens [voir 12] :

$$\mathcal{A} = \{a^j \mid 1 \leq j \leq m\} \tag{67}$$

Nous allons distinguer dans l'ensemble $\mathcal{A} \times \mathcal{A}$ des couples d'attributs, un sous ensemble potentiel $\mathcal{C}$ de couples d'attributs (a, b) pour lesquels l'implication $a \rightarrow b$ "peut avoir un sens".

Une première condition qu'on peut exiger pour un couple d'attributs (a, b) entrant dans la composition de $\mathcal{C}$ est d'être tel que (voir (6)) :

$$n(a \wedge \bar{b}) < \frac{n(a) \times n(\bar{b})}{n} \tag{68}$$

Cette condition est d'ailleurs équivalente à :

$$n(a \wedge b) > \frac{n(a) \times n(b)}{n} \tag{69}$$

on le voit en écrivant $n(a \wedge \bar{b}) = n(a) - n(a \wedge b)$ et $n(\bar{b}) = n - n(b)$. La dernière relation (69) indique l'observation de la dépendance positive entre les deux attributs a et b.

Une deuxième condition qu'on peut exiger pour qu'un couple d'attributs (a, b) rentre dans la composition de $\mathcal{C}$, relativement à l'implication $a \rightarrow b$, est que :

$$n(a) < n(b) \tag{70}$$

En effet, dans le cas où l'implication $a \rightarrow b$ est observée exactement (sans contre exemple), l'ensemble $\mathcal{O}(a)$ des individus où a est à $VRAI$ est strictement

inclus dans $\mathcal{O}(b)$ individus où b est à $VRAI$. D'ailleurs, relativement à un couple (a, b) d'attributs pour lequel $n(a) < n(b)$, on a :

Proposition 5 *Pour $n(a) < n(b)$ et $n(a \wedge b) > \frac{n(a) \times n(b)}{n}$, on a :*

$$Q_3(a, \bar{b}) < Q_3(b, \bar{a}) \tag{71}$$

Preuve

Commençons par rappeler les expressions de $Q_3(a, \bar{b})$ et de $Q_3(b, \bar{a})$:

$$Q_3(a, \bar{b}) = \frac{n(a \wedge \bar{b}) - [n(a) \times n(\bar{b})/n]}{\sqrt{n(a) \times n(\bar{b})/n}} \tag{72}$$

et

$$Q_3(b, \bar{a}) = \frac{n(b \wedge \bar{a}) - [n(b) \times n(\bar{a})/n]}{\sqrt{n(b) \times n(\bar{a})/n}} \tag{73}$$

On peut aisément établir que les numérateurs de (72) et (73) sont identiques, les deux prenant la valeur commune :

$$\frac{1}{n} \times [n(a \wedge \bar{b}) \times n(\bar{a} \wedge b) - n(a \wedge b) \times n(\bar{a} \wedge \bar{b})] \tag{74}$$

Pour le voir, il suffit de décomposer par rapport à $n(a \wedge b)$, $n(a \wedge \bar{b})$, $n(\bar{a} \wedge b)$ et $n(\bar{a} \wedge \bar{b})$.

La condition $n(a \wedge b) > n(a) \times n(b)/n$ implique (voir ci-dessus) que le numérateur commun de $Q_3(a, \bar{b})$ et de $Q_3(b, \bar{a})$ est négatif. Enfin, la condition $n(a) < n(b)$ est équivalente à :

$$n(a) \times n(\bar{b}) < n(\bar{a}) \times n(b) \tag{75}$$

Elle implique par conséquent que :

$$Q_3(a, \bar{b}) < Q_3(b, \bar{a}) \tag{76}$$

Ainsi, $Q_3(a, \bar{b})$ est plus fortement négatif que $Q_3(b, \bar{a})$. Dans ces conditions (voir (7) ci-dessus), on a pour l'intensité d'implication de la *vraisemblance du lien* :

$$\mathcal{I}(a \rightarrow b) = \Phi[-Q_3(a, \bar{b})] > \mathcal{I}(b, a) = \Phi[-Q_3(b, \bar{a})] \tag{77}$$

□

La première version du sous ensemble potentiel $\mathcal{C}$ de couples d'attributs par rapport auquel la réduction globale peut être opérée est défini comme suit :

$$\mathcal{C}_0 = \left\{(a, b) \mid (a, b) \in \mathcal{A} \times \mathcal{A}, n(a \wedge b) > \frac{n(a) \times n(b)}{n} \text{ et } n(a) < n(b)\right\} \tag{78}$$

Definition 1 *La réduction globale est totale si elle est effectuée par rapport à $\mathcal{C}_0$.*

Considérons dans ces conditions la distribution de Q_3 sur $\mathcal{C}_0$:

$$\{Q_3(a,\bar{b}) \mid (a,b) \in \mathcal{C}_0\} \tag{79}$$

et désignons par $moy_0(Q_3)$ $var_0(Q_3)$ la moyenne et la variance de cette distribution ; nommément :

$$moy_0(Q_3) = \frac{1}{card(\mathcal{C}_0)} \sum_{(a,b)\in\mathcal{C}_0} Q_3(a,\bar{b}) \tag{80}$$

$$var_0(Q_3) = \frac{1}{card(\mathcal{C}_0)} \sum_{(a,b)\in\mathcal{C}_0} [Q_3(a,\bar{b}) - moy_0(Q_3)]^2 \tag{81}$$

Nous noterons l'indice globalement réduit sous la forme :

$$Q_3^{g0}(a,\bar{b}) = \frac{Q_3(a,\bar{b}) - moy_0(Q_3)}{\sqrt{var_0(Q_3)}} \tag{82}$$

Sa distribution

$$\{Q_3^{g0}(a,\bar{b}) \mid (a,b) \in \mathcal{C}_0\} \tag{83}$$

est de moyenne nulle et de variance unité. De sorte que la distribution des indices probabilistes de l'intensité d'implication

$$\{\mathcal{I}^0(a \to b) = \Phi(-Q_3^{g0}(a,\bar{b})) \mid (a,b) \in \mathcal{C}_0\} \tag{84}$$

est finement discriminante pour comparer entre elles les différentes implications $a \to b$ où $(a,b) \in \mathcal{C}_0$.

On peut admettre pour des raisons de significativité que les attributs booléens de $\mathcal{A}$ sont établis tels que :

$$max_{a\in\mathcal{A}}\Big(p(a) = \frac{n(a)}{n}\Big) \leq 0.5 \tag{85}$$

D'autre part, remarquons que l'indice "Confiance" $p(b \mid a) = n(a \wedge b)/n(a)$ représente la part de $\mathcal{O}(a)$ qui se trouve dans $\mathcal{O}(b)$ et le moins qu'on puisse demander pour l'implication $a \to b$, est que cette part soit strictement supérieure à 0.5. Dans ces conditions, on a nécessairement :

$$\frac{p(a \wedge b)}{p(a)} > 0.5 \geq p(b) \tag{86}$$

D'où la proposition :

Proposition 6 *La condition (86) étant remplie, si Confiance($a \to b$) est strictement supérieure à* 0.5 *; alors la dépendance entre les deux attributs booléens a et b est positive : $p(a \wedge b) > p(a) \times p(b)$.*

Dans ces conditions, il est naturel de considérer un ensemble potentiel que nous notons $\mathcal{C}_c$ où on remplace pour le filtrage la condition de dépendance positive par celle, d'un seuil (noté c) pour la Confiance strictement supérieur à 0.5. Ainsi on a :

$$\mathcal{C}_c = \Big\{(a,b) \in \mathcal{A}\times\mathcal{A}, n(a \wedge b) > c \times n(a) \text{ et } n(a) < n(b)\Big\} \tag{87}$$

Definition 2 *La réduction globale est partielle relativement à la Confiance si elle est effectuée par rapport à $\mathcal{C}_c$.*

Les expressions à produire sont absolument analogues à (80), (81), (82), (83) et (84), à cela près qu'il y a lieu de remplacer $\mathcal{C}_0$ par $\mathcal{C}_c$, on notera $Q_3^{gc}(a,\bar{b})$ le coefficient centré et réduit correspondant à (82) et $\mathcal{I}^{gc}(a \to b)$ l'intensité d'implication de la *vraisemblance du lien*.

Un dernier filtrage complémentaire qu'on peut considérer pour des raisons de significativité, concerne un seuil minimal pour le "Support" $p(a \wedge b)$. Dans ce cas, l'ensemble potentiel pour l'implication entre attributs de la forme $(a \to b)$ s'écrit :

$$\mathcal{C}_{cs} = \Big\{(a,b) \in \mathcal{A} \times \mathcal{A}, p(b \mid b) > c, p(a \wedge b) > s \text{ et } n(a) < n(b)\Big\} \qquad (88)$$

Definition 3 *La réduction globale est partielle relativement à la Confiance et au Support, si elle est effectuée par rapport à $\mathcal{C}_{cs}$.*

De la même façon que ci-dessus les expressions à produire sont analogues à (80), (81), (82), (83) et (84), à cela près qu'il y a lieu de remplacer $\mathcal{C}_0$ par $\mathcal{C}_{cs}$. Ici, relativement à un même couple d'attributs (a,b) de $\mathcal{C}_{cs}$, on notera $Q_3^{gcs}(a,\bar{b})$ le coefficient centré et réduit correspondant (82) et $\mathcal{I}^{gcs}(a \to b)$ l'intensité d'implication de la *vraisemblance du lien*.

Donnons à présent le sens de l'acronyme $VLgrImpP$: VL comme *Vraisemblance* du *Lien*, gr comme indice globalement réduit, Imp comme indice *Implicatif* et P comme modèle *Poissonnien*.

5.2.2 L'approche $VT100$ fondée sur la moyenne des $p-values$, $VT100BarP$

Le passage entre [23] et [27] correspond à la nécessité de considérer le modèle Poissonnien de l'hypothèse d'absence de liaison - lequel ayant un caractère dissymétrique - et à prendre comme indice brut le nombre de contre exemples noté $n(a \wedge \bar{b})$ [21]. La démarche a été largement détaillée en section 3.1 ci-dessus. Dans ces conditions, nous allons nous contenter de reprendre de [27] leur exemple illustratif défini par la table suivante :

	a	$\bar{a}$	Total
b	48	72	120
$\bar{b}$	28	125	153
Total	76	197	273

Table 6

La table ramenée à 100 prend approximativement la forme suivante :

	a	$\bar{a}$	Total
b	17.58	26.38	43.96
$\bar{b}$	10.25	45.79	56.04
Total	27.83	72.17	100

Table 7

Nous nous focalisons ici par rapport à cette table sur le vecteur associé à $(n(a \wedge \bar{b}), n(a), n(\bar{b}))$; soit : $(10.25, 27.83, 56.04)$. Ce vecteur correspond à celui (γ, α, β) de la sous section 3.1 ci-dessus ; de sorte que les 8 vecteurs à composantes entières qui l'encadrent sont :

$$(10,27,56),(10,27,57),(10,28,56),(10,28,57),(11,27,56),(11,27,57),(11,28,56),(11,28,57)$$

La suite des $p-values$ associée à la suite ordonnée des 8 configurations est :

$$0.117,\ \ 0.106,\ \ 0.0954,\ \ 0.0858,\ \ 0.1759,\ \ 0.1606,\ \ 0.1455,\ \ 0.1316$$

Donnons à titre d'exemple le calcul de la première. Il s'agit conformément au modèle Poissonnien $\mathcal{N}_3$ de la probabilité :

$$Pr_3^{\mathcal{N}}\{n(a^\star \wedge \bar{b}^\star) \leq (n(a \wedge \bar{b})\} \ ,$$

plus spécifiquement pour l'exemple de :

$$Pr_3^{\mathcal{N}}\{n(a^\star \wedge \bar{b}^\star) \leq 10 \ ,$$

où le paramètre de la loi de Poisson est $\lambda = 27 \times 56/100 = 15.12$.
La dernière probabilité se met sous la forme approchée :

$$\Phi[(10 + 0.5 - (27 \times 56/100))/\sqrt{(27 \times 56/100)}] = \Phi(-1.188)$$

qui vaut 0.117.

Le développement est de même nature que celui de la sous section 3.1. En posant $x = 56.04 - 56 = 0.04$, $y = 27.83 - 27 = 0.83$ et $z = 10.25 - 10 = 0.25$, le barycentre des 8 vecteurs ci-dessus par rapport au développement tel que celui de (39) redonne le vecteur $(10.25, 27.83, 56.04)$. On s'autorise alors à considérer cette même moyenne sur la suite associée des $p-values$; soit :

$$\begin{aligned} &0.75 \times 0.17 \times 0.96 \times 0.117 + 0.75 \times 0.17 \times 0.04 \times 0.106 + \\ &0.75 \times 0.83 \times 0.96 \times 0.0954 + 0.75 \times 0.83 \times 0.04 \times 0.0858 + \\ &0.25 \times 0.17 \times 0.96 \times 0.1759 + 0.25 \times 0.17 \times 0.04 \times 0.1606 + \\ &0.25 \times 0.83 \times 0.96 \times 0.1455 + 0.25 \times 0.83 \times 0.04 \times 0.1316 = \\ &0.1115 \end{aligned}$$

En utilisant l'approximation par la loi normale, la valeur test correspondante est $\Phi^{-1}(1 - 0.1115) \simeq 1.22$. Maintenant, en considérant directement la table 7 et en adoptant pour la valeur test la même forme analytique que s'il s'agissait d'un tableau d'entiers, on obtient :

$$-\frac{10.25 + 0.5 - (27.83 \times 56.04/100)}{\sqrt{(27.83 \times 56.04/100)}} = 1.227$$

qui représente une valeur très proche de la précédente.

Terminons par une remarque. L'élément correctif 0.5 a été utilisé pour chacune des configurations entières encadrant le profil $(10.25, 27.83, 56.04)$. De la même façon, il vient d'être utilisé sur ce profil même, reprenant en cela [27]. Mais cette fois, les valeurs du tableau de contingence sont des rationnels. Une valeur unitaire représente un macro individu. On peut dans ces conditions ne pas comprendre ici cet élément correctif, lequel est conçu lorsque chaque valeur unitaire représente un individu élémentaire et non un groupement d'ailleurs fractionnaire d'individus (dans l'exemple, - voir les tables 6 et 7 - où d'ailleurs n n'est pas très grand, chaque macro individu représente 2.73 individus élémentaires).

Il nous reste maintenant à préciser le sens de l'acronyme $VT100ImpBarP$: $VT100$ comme $\underline{V}aleur\ \underline{T}est$ se basant sur un échantillon de taille 100, Imp comme indice $\underline{Impl}icatif$, Bar comme approche $\underline{Bar}ycentrique$ et P comme modèle $\underline{P}oissonnien$

5.2.3 L'approche $VT100$ fondée sur une approche corrélative et ensembliste, $VT100ImpCorP$

.

Cette approche correspond à celle de la sous section 4.1 où c'est le modèle permutationnel (généralisation du modèle hypergéométrique) qui a été considéré dans le cadre d'une association symétrique évaluant le degré d'équivalence. Dans ce dernier cas, sur la base du tableau réduit à 100 (voir Table 7), l'évaluation porte sur la grandeur relative de l'indice brut $n(a \wedge \bar{b})$. Ici, l'association est dissymétrique et concerne le degré d'implication $a \rightarrow b$. Elle porte sur l'évaluation de la petitesse relative de $n(a \wedge \bar{b})$. D'autre part, c'est le modèle Poissonnien qui doit être retenu ici. Cependant, en travaillant par rapport au tableau réduit à 100 (voir Table 7), nous ne sommes plus dans le cas discret. Dans ces conditions, on peut reprendre la construction conceptuelle de la sous section 4.1 et associer aux deux attributs booléens a et b deux variables numériques $a^{'}$ et $b^{'}$, où :

$$\begin{aligned} a^{'}(i) &= \sqrt{\frac{100}{n}} \times a(i) \\ b^{'}(i) &= \sqrt{\frac{100}{n}} \times b(i) \end{aligned} \tag{89}$$

où $a(i) = 1$ (resp. 0) si l'attribut booléen a est à $VRAI$ (resp. $FAUX$) sur l'individu codé i. De même, $b(i) = 1$ (resp. 0) si l'attribut booléen b est à $VRAI$ (resp. $FAUX$) sur l'individu codé i. De sorte qu'on reprend les équations (46) et celles (47) par rapport à l'ensemble virtuel $\Omega = \{\omega_j \mid 1 \leq j \leq 100\}$ formé de 100 macro individus. Dans ces conditions, l'indice brut prend la forme :

$$n(\alpha \wedge \bar{\beta}) = \sum_{1 \leq j \leq 100} \alpha(j) \times [1 - \beta(j)] \tag{90}$$

l'évaluation de sa petitesse est définie par :

$$Pr^{\mathcal{N}_3}\{n(\alpha^{\star} \wedge \bar{\beta}^{\star}) \leq n(\alpha \wedge \bar{\beta})\} \tag{91}$$

où $\alpha^\star$ et $\beta^\star$ sont deux attributs aléatoires numériques indépendants, respectivement associés dans le cadre d'un modèle Poissonnien de l'hypothèse d'absence de liaison qui étend celui considéré dans le cas de la comparaison de deux attributs booléens. Or une telle extension a été établie dans [20] et, indépendamment, dans [8].

L'indice de la forme :

$$Pr^{\mathcal{N}_3}\{n(\alpha^\star \wedge \bar{\beta}^\star) > n(\alpha \wedge \bar{\beta})\} \tag{92}$$

Son calcul passe par l'approximation par la loi normale de la loi de $n(\alpha^\star \wedge \bar{\beta}^\star)$. Cette approximation se justifie pleinement compte tenu de la tendance "rapide" de la loi de $n(\alpha^\star \wedge \bar{\beta}^\star)$ vers la loi normale.

Reprenons les relations (46) et (47). Les relations de même type que (48) et (49) nous permettent d'écrire :

$$\begin{aligned} moy(\alpha) &= \frac{1}{100} \sum_{1 \leq j \leq 100} \alpha(j) = moy(a) \\ moy(\bar{\beta}) &= \frac{1}{100} \sum_{1 \leq j \leq 100} \bar{\beta}(j) = moy(\bar{b}) \end{aligned} \tag{93}$$

D'autre part, ayant $var(\bar{\beta}) = var(\beta)$, en adoptant les conditions (52), les valeurs considérées pour la variance sont :

$$\begin{aligned} var(\alpha) &= var(a) \\ var(\bar{\beta}) &= var(b) \end{aligned} \tag{94}$$

Le calcul de l'indice (92) ou celui associé de la $p - value$ qui prend la forme, déjà exprimée ci-dessus :

$$Pr^{\mathcal{N}_3}\{n(\alpha^\star \wedge \bar{\beta}^\star) \leq n(\alpha \wedge \bar{\beta})\} \tag{95}$$

passe par l'approximation par la loi normale de $n(\alpha^\star \wedge \bar{\beta}^\star)$. L'espérance mathématique et la variance de cette dernière variable aléatoire dans l'hypothèse d'absence de liaison de Poisson sont données, en tenant compte des relations établies ci-dessus, dans [20, 8] :

$$\begin{aligned} \mathcal{E}[n(\alpha^\star \wedge \bar{\beta}^\star)] &= 100 \times moy(a) \times moy(b) = 100 \times p(a) \times [1 - p(b)] \\ var[n(\alpha^\star \wedge \bar{\beta}^\star)] &= 100 \times [var(a) + moy(a)^2] \times [var(b) + moy(b)^2] \end{aligned} \tag{96}$$

Ainsi, l'indice $n(\alpha \wedge \bar{\beta})$ centré et réduit prend - calcul fait - la forme :

$$Q_3(\alpha \wedge \bar{\beta}) = \frac{n(\alpha \wedge \bar{\beta}) - 100 \times p(a) \times [1 - p(b)]}{\sqrt{100 \times p(a) \times [1 - p(b)]}} \tag{97}$$

Le calcul de la $p - value$ (95) en tenant compte de l'approximation par la loi normale et en usant de l'élément correctif 0.5 donne :

$$p = \Phi\Big(\frac{n(\alpha \wedge \bar{\beta}) + 0.5 - 100 \times p(a) \times [1 - p(b)]}{\sqrt{100 \times p(a) \times [1 - p(b)]}}\Big) \tag{98}$$

De sorte que :

$$\Phi^{-1}(1-p) = -\frac{n(\alpha \wedge \bar{\beta}) + 0.5 - 100 \times p(a) \times [1 - p(b)]}{\sqrt{100 \times p(a) \times [1 - p(b)]}} \tag{99}$$

dont la valeur 1.227 est très exactement la valeur test directement donnée ci-dessus (voir sous section 5.2.2) pour la table 7 et qui s'avère très proche de celle, définie par une opération barycentrique telle qu'elle est proposée dans [27]. Quant à l'intensité d'implication, sa valeur est donnée par $1 - p$. Elle vaut dans le cas de la table 7, 0.890.

Il nous reste maintenant à préciser le sens de l'acronyme $VT100ImpCorP$: $VT100$ comme $\underline{V}aleur\ \underline{T}est$ se basant sur un échantillon de taille 100, Imp comme indice $\underline{Imp}licatif$, Cor comme approche $\underline{Cor}relative$ et P comme modèle $\underline{P}oissonnien$.

5.2.4 L'approche $VT100$ fondée sur une réduction par projection sur un ensemble aléatoire de taille 100, $VT100ImpProj$

Il s'agit d'adapter dans le cas implicatif le développement obtenu en sous section 4.2 dans le cas symétrique. Le correspondant de l'indice (55) dans le cas implicatif donne :

$$\mathcal{E}\Big(\frac{card[\mathcal{O}(a) \cap \mathcal{O}(\bar{b}) \cap E^\star] - \frac{card[\mathcal{O}(a) \cap E^\star \times card[\mathcal{O}(\bar{b}) \cap E^\star]]}{100}}{\sqrt{\frac{card[\mathcal{O}(a) \cap E^\star \times card[\mathcal{O}(\bar{b}) \cap E^\star]]}{100}}}\Big) \tag{100}$$

Le niveau aléatoire de l'implication sur un ensemble aléatoire $E^\star$ de taille 100 est d'autant plus fort que l'indice aléatoire sous le signe $\mathcal{E}$ espérance est stochastiquement plus petit (plus négatif). L'application de la proposition 3 ci-dessus permet d'obtenir :

$$\mathcal{E}(card[\mathcal{O}(a) \cap \mathcal{O}(\bar{b}) \cap E^\star]) = \frac{n(a \wedge \bar{b}) \times e}{n} = e \times p(a \wedge \bar{b}) \tag{101}$$

$$\mathcal{E}\Big(card[\mathcal{O}(a) \cap E^\star] \times card[\mathcal{O}(\bar{b}) \cap E^\star]\Big) = \frac{e}{n} \times n(a \wedge \bar{b}) + \frac{e(e-1)}{n(n-1)} \times [n(a) \times n(\bar{b}) - n(a \wedge \bar{b})] \tag{102}$$

En adoptant les mêmes choix qui ont permis d'aboutir à l'expression (63) ci-dessus, on proposera comme valeur de l'espérance mathématique (100) :

$$\frac{\frac{1}{\sqrt{n(n-1)}} \times [(ne-1) \times n(a \wedge \bar{b}) - (e-1) \times n(a) \times n(\bar{b})]}{\sqrt{(n-e) \times n(a \wedge \bar{b}) + (e-1) \times n(a) \times n(\bar{b})}} \tag{103}$$

La valeur test est l'opposée de cette dernière valeur, où on peut si l'on veut remplacer au numérateur $n(a \wedge \bar{b})$ par $n(a \wedge \bar{b}) + 0.5$.

Dans le cas de la table 6, la valeur test $VT100$ dans cette troisième nouvelle

version (expression (103)) est égale à 1.320. Cette valeur test est sensiblement plus grande que celle donnée par l'une ou l'autre des deux autres techniques dont celle, proposée dans [27]. L'indice de "*vraisemblance du lien*" ou "*intensité d'implication*" vaut $\Phi(1.320) = 0.9066$.

Terminons cette sous section en précisant le sens de l'acronyme $VT100ImpProj$: $VT100$ comme $\underline{V}aleur\ \underline{T}est$ se basant sur un échantillon de taille 100, Imp comme indice $\underline{Imp}licatif$, et Proj comme par $\underline{Proj}ection$ sur un ensemble aléatoire.

5.3 Analyse comparée de $VLgr$ et $VT100$ par rapport à deux modèles de croissance du nombre n d'observations

5.3.1 Introduction

Les deux modèles $M1$ et $M2$ ont été présentés dans la dernière partie de l'introduction. Relativement à un couple d'attributs booléens (a, b) reprenons les paramètres déjà largement introduits (voir section 2) :

$$n\ ,\ n(a)\ ,\ n(\bar{a})\ ,\ n(b)\ ,\ n(\bar{b})\ ,\ n(a \wedge b)\ ,\ n(a \wedge \bar{b})\ ,\ n(\bar{a} \wedge b) \text{ et } n(\bar{a} \wedge \bar{b})$$

Ces paramètres prennent naturellement place dans le tableau de contingence 2×2 croisant $\{a, \bar{a}\}$ avec $\{b, \bar{b}\}$ (voir Table 2 ci-dessus).

Pour le modèle $M1$ les effectifs précédents sont tous multipliés par une même constante k. Ainsi, la suite des cardinaux précédents devient :

$$k{\times}n\ ,\ k{\times}n(a)\ ,\ k{\times}n(\bar{a})\ ,\ k{\times}n(b)\ ,\ k{\times}n(\bar{b})\ ,\ k{\times}n(a{\wedge}b)\ ,\ k{\times}n(a{\wedge}\bar{b})\ ,\ k{\times}n(\bar{a}{\wedge}b) \text{ et } k{\times}n(\bar{a}{\wedge}\bar{b}) \tag{104}$$

Pour le modèle $M2$ on accroît le nombre $n(\bar{a} \wedge \bar{b})$ d'éléments pour lesquels a et b sont à $FAUX$. Si x définit cet accroissement, la suite des cardinaux précédents devient :

$$n{+}x\ ,\ n(a)\ ,\ n(\bar{a}){+}x\ ,\ n(b)\ ,\ n(\bar{b}){+}x\ ,\ n(a{\wedge}b)\ ,\ n(a{\wedge}\bar{b})\ ,\ n(\bar{a}{\wedge}b) \text{ et } n(\bar{a}{\wedge}\bar{b}){+}x \tag{105}$$

Pour chacun des deux modèles $M1$ et $M2$ nous allons chercher à analyser les comportements des indices $VLgr$ (*Vraisemblance du Lien* après *réduction globale*) et $VT100$. À effet, nous allons considérer le cas implicatif défini par le nombre $n(a \wedge \bar{b})$ de contre exemples. D'autre part, c'est l'hypothèse d'absence de liaison de Poisson - qui précisément a un caractère dissymétrique - qui sera la référence. Ainsi, $VLgrImpP$ est ce que nous avons appelé "Intensité d'Implication Contextuelle" [voir (10), (33), (34), (82)].

L'ensemble des couples d'attributs qui est évalué et par rapport auquel la normalisation est effectuée a été noté $\mathcal{C}_{cs}$ et se trouve défini par (88). Ainsi, pour un couple d'attributs (a, b) faisant partie de $\mathcal{C}_{cs}$, l'indice concerné s'écrit avec des notations que l'on comprend (voir par exemple (82)), sous la forme :

$$Q_3^{gcs}(a,\bar{b}) = \frac{Q_3(a,\bar{b}) - moy_{cs}(Q_3)}{\sqrt{var_{cs}(Q_3)}} \tag{106}$$

Il intervient juste avant l'indice probabiliste via la fonction de répartition de la loi normale centrée et réduite.

La variante que nous prenons de $VT100$ est celle que nous avons appelée corrélative et ensembliste. L'indice est défini par l'expression (97) ci-dessus. Si ν désigne le nombre total d'observations du tableau de contingence obtenu à partir du tableau de contingence initial - croisant $\{a,\bar{a}\}$ avec $\{b,\bar{b}\}$ - par application du modèle $M1$ ou $M2$, le nouveau tableau pour lequel l'indice Q_3 est calculé, s'obtient en multipliant les contenus des différentes cases du tableau par la rapport $100/\nu$.

5.3.2 Comportement de $VLgr$ et $VT100$ par rapport au modèle $M1$

Comportement de $VLgr$

Proposition 7 *L'indice est invariant dans le contexte du modèle $M1$.*

Reprenons ici la famille de coefficients $Q_3(a,\bar{b})$ définie sur l'ensemble des couples d'attributs booléens appartenant à $\mathcal{C}_{cs}$ (voir (88)). Chaque coefficient est calculé sur la base de n observations, nous le noterons $Q_{3n}(a,\bar{b})$. Considérons une instanciation de la valeur de ν associée à un coefficient de proportionalité k (voir (104)). L'indice transformé que nous noterons $Q_{3\nu}(a,\bar{b})$ s'obtient - voir (16) pour la justification - au moyen de la relation :

$$Q_{3\nu}(a,\bar{b}) = \sqrt{\frac{\nu}{n}} \times Q_{3n}(a,\bar{b}) = \sqrt{k} \times Q_{3n}(a,\bar{b}) \tag{107}$$

De la sorte, en reprenant une expression telle que (106)) on a :

$$Q_{3\nu}^{gcs}(a,\bar{b}) = Q_{3n}^{gcs}(a,\bar{b}) \tag{108}$$

pour tout couple (a,b) d'attributs de $\mathcal{C}_{cs}$. □

Comportement de $VT100$

Proposition 8 *L'indice est invariant dans le contexte du modèle $M1$*

Ce résultat est trivial. En effet, le tableau de contingence réduit à 100 est invariant qu'il soit calculé sur la base de n observations ou sur la base de $\nu = k \times n$ observations et ce, en multipliant le contenu de chacune des cases du tableau de contingence 2×2, par k. □

5.3.3 Comportement de $VLgr$ et $VT100$ par rapport au modèle $M2$

Comportement de $VLgr$ Nous restons dans le contexte de l'ensemble $\mathcal{C}_{cs}$ des couples d'attributs (voir 88) défini à partir de la situation initiale où le nombre d'observations est n. Pour $(a,b) \in \mathcal{C}_{cs}$, l'indice $Q_3(a,\bar{b})$ obtenu après la transformation (105) sera noté $Q_{3x}(a,\bar{b})$; de sorte que la distribution initiale peut être notée :

$$\left\{Q_{30}(a,\bar{b}) \mid (a,b) \in \mathcal{C}_{cs}\right\} \tag{109}$$

Commençons par exprimer $Q_{3x}(a,\bar{b})$, on a :

$$Q_{3x}(a,\bar{b}) = \frac{n(a\wedge\bar{b}) - \frac{n(a)\times[n(\bar{b})+x]}{(n+x)}}{\sqrt{\frac{n(a)\times[n(\bar{b})+x]}{(n+x)}}} \tag{110}$$

Nous considérerons plus précisément le coefficient $-Q_{3x}(a,\bar{b})$ qui est - via l'approximation par la fonction de répartition de la loi normale centrée et réduite - une fonction croissante de l'indice probabiliste local d'implication de la "*vraisemblance du lien*" ("*intensité d'implication statistique*"). Une autre interprétation est l'indice VT (voir Proposition 2). Pour alléger nos notations posons : $\gamma = n(a\wedge\bar{b})$, $\alpha = n(a)$, $\beta = n(b)$ et $y = n(\bar{b}) + x$. $-Q_{3x}(a,\bar{b})$ se met sous la forme de la fonction :

$$\phi(y) = \frac{-\gamma + \frac{\alpha\times y}{\beta+y}}{\sqrt{\frac{\alpha\times y}{\beta+y}}} \tag{111}$$

Nous obtenons pour la dérivée :

$$\phi'(y) = \frac{1}{2} \times \frac{\alpha\times\beta}{\sqrt{\alpha\times y\times(\beta+y)}} \times [\frac{1}{\beta+y} + \frac{\gamma}{\alpha\times y}] \tag{112}$$

Ainsi, la fonction $-Q_{3x}(a,\bar{b})$ est strictement croissante par rapport à x.

Le calcul de la fonction dérivée seconde $\phi''(y)$ nous donne :

$$\begin{aligned}\phi''(y) = -\frac{1}{2} \times \frac{\alpha\beta}{(\alpha y)^{3/2}(\beta+y)^{3/2}} \times \Big(\frac{\gamma}{\alpha y} + \frac{1}{2}[1+\gamma\times(\frac{\beta+y}{\alpha y})] \times [1+3(\frac{\beta+y}{\alpha y})^{-1}]\Big) \\ = -\frac{1}{2}\alpha\beta(\frac{\sqrt{u}}{\alpha y})^3 \times \Big(\frac{\gamma}{\alpha y} + \frac{1}{2}(1+\frac{\gamma}{u})(1+3u)\Big)\end{aligned} \tag{113}$$

en ayant posé $u = \alpha y/(\beta+y)$.

Ainsi, $\phi''(y)$ est négatif ; et, l'accroissement relatif de l'indice qui est positif va en décroissant. Donnons ici quelques valeurs de $-Q_{3x}(a,\bar{b})$ dans le cas de l'exemple issu de la base de données "Wages" où : $n = 14743$, $n(a) = 4819$, $n(\bar{b}) = 3522$ et $n(a\wedge\bar{b}) = 225$.

x	$-Q_{3x}(a,\bar{b})$
0	27.712
1000	31.599
2000	34.687
10000	47.503
50000	60.234
100000	63.233

Table 8

Considérons à présent la transformation (105) appliquée à l'ensemble des couples d'attributs de l'ensemble $\mathcal{C}_{cs}$ défini par (88). On obtient la famille des indices :

$$\Big\{Q_{3x}(a',b') \mid (a',b') \in \mathcal{C}_{cs}\Big\} \tag{114}$$

Il est à remarquer que pour tout couple (a',b') d'attributs transformés, l'indice Confiance reste invariable ; alors que le Support - prenant la forme $n(a' \wedge b')/(n+x)$ - diminue.

L'indice $Q_{3x}(a,\bar{b})$ globalement centré et réduit par rapport à la distribution (114) peut être écrit sous une forme analogue à celle (106) :

$$Q_{3x}^{gcs}(a,\bar{b}) = \frac{Q_{3x}(a,\bar{b}) - moy_{cs}(Q_{3x})}{\sqrt{var_{cs}(Q_{3x})}} \tag{115}$$

où $moy_{cs}(Q_{3x})$ et $var_{cs}(Q_{3x})$ sont la moyenne et la variance de la distribution (114).

Dans nos simulations nous étudions pour certains couples particuliers d'attributs (a,b), la variation par rapport à x - croissant à partir de 0 - de l"'*Intensité d'Implication Contextuelle*" $\Phi[-Q_{3x}^{gcs}(a,\bar{b})]$. Si $-Q_{3x}(a,\bar{b})$ est croissant par rapport à x, sa version normalisée $-Q_{3x}^{gcs}(a,\bar{b})$ ne l'est pas nécessairement toujours.

Comportement de $VT100$ Pour être conforme au type de notations cidessus, nous indiquons par $-Q_{3x}^{100}$ la version adoptée de l'indice $VT100$ qui donne - via la fonction de répartition de la loi normale centrée et réduite - l'indice probabiliste de l'intensité d'implication (nous négligeons l'élément correctif 0.5 qui ne fait que perturber la construction conceptuelle).

En désignant par ν l'entier $n+x$, cet indice est calculé à partir d'un tableau de contingence dont les entrées sont définies à partir de la suite (105) à laquelle on applique le facteur multiplicatif $100/\nu$. On obtient :

$$\begin{aligned} -Q_{3x}^{100}(a,\bar{b}) &= -\frac{\frac{100}{\nu}n(a\wedge\bar{b} - \frac{100}{\nu^2}n(a)\times[n(\bar{b})+x]}{\sqrt{\frac{100}{\nu^2}n(a)\times[n(\bar{b})+x]}} \\ &= -10\times\frac{[n(a\wedge\bar{b}) - \frac{n(a)\times[n(\bar{b})+x]}{n+x}]}{\sqrt{n(a)\times[n(\bar{b})+x]}} \end{aligned} \tag{116}$$

Nous allons étudier le sens de variation de cet indice lorsque x varie (à partir de 0 en croissant). À cet effet, simplifions les notations et posons : $\gamma = n(a\wedge\bar{b})$, $\alpha = n(a)$, $\beta = n(b)$, $\bar{\beta} = n(\bar{b})$ et $y = n(\bar{b}) + x$.

Au coefficient mltiplicatif -10 près, l'indice (116) se met sous la forme :

$$\psi(y) = \frac{\gamma - \alpha\times\frac{y}{\beta+y}}{\sqrt{\alpha y}} \tag{117}$$

Nous obtenons pour la fonction dérivée :

$$\psi'(y) = [-\frac{\alpha\beta}{(\beta+y)^2}] - [\frac{\gamma}{2y} - \frac{\alpha}{2(\beta+y)}] \tag{118}$$

En réduisant au même dénominateur et en tenant compte du facteur multiplicatif négatif -10, le signe de la dérivée de la fonction initiale associée à $-Q_{3x}^{100}(a,\bar{b})$ est donnée par :

$$(\gamma-\alpha)y^2 + \beta(\alpha+2\gamma)y + \gamma\beta^2 \tag{119}$$

Il s'agit d'un trinôme du second degré en y. Son discriminant

$$\Delta = \beta^2 \times [(\alpha+2\gamma)^2 + 4(\alpha-\gamma)] \tag{120}$$

est strictement positif puisque $\gamma = n(a \wedge \bar{b}) < \alpha = n(a)$. Les deux racines se mettent sous la forme :

$$\begin{aligned} y' &= \frac{\beta(\alpha+2\gamma) + \sqrt{\Delta}}{2(\alpha-\gamma)} \\ y'' &= \frac{\beta(\alpha+2\gamma) - \sqrt{\Delta}}{2(\alpha-\gamma)} \end{aligned} \tag{121}$$

On voit immédiatement que y' est strictement positif alors que y'' est négatif ou nul. Le cas de nullité est exceptionnel et correspond à $n(a \wedge \bar{b}) = n(a)$. Ayant $\gamma < \alpha$, le trinôme est positif pour $y \leq y''$ et négatif pour $y > y'$. Ainsi, l'indice $-Q_{3x}^{100}(a,\bar{b})$ est croissant pour x croissant de 0 jusqu'à $y' - n(\bar{b})$ et décroissant pour x supérieur à $y' - n(b)$.

À titre d'illustration considérons l'exemple ci-dessus issu de la base de données "Wages". On obtient :

x	$-Q_{3x}^{100}(a,\bar{b})$
0	2.282
1000	2.518
2000	2.681
10000	3.020
15000	2.973
20000	2.887
30000	2.694
50000	2.367
100000	1.982

Table 9

Le calcul de la racine y' ci-dessus (voir 121) donne $y' = 12837.02$, il lui correspond une valeur $x' = y' - 3522 = 9315.02$. La valeur associée de l'indice est 3.02.

Un tel comportement de $-Q_{3x}^{100}(a,\bar{b})$, d'abord croissant puis décroissant peut surprendre.

5.4 Le cas de l'Intensité d'Implication Entropique

5.4.1 Introduction

Nous avons bien vu que dans le cas où on a $n(a \wedge \bar{b}) < n(a)n(\bar{b})/n$ l'"*Intensité d'Implication*" qui dérive de l'approche "*Vraisemblance du Lien*" tend "rapidement" vers 1 dès lors que n augmente (voir (7) dans l'Introduction, (64) et (65)). Plus précisément, compte tenu de la forme (65) du calcul de l'indice et dans le cadre du modèle $M1$ de croissance des effectifs du tableau croisant $\{a, \bar{a}\}$ avec $\{b, \bar{b}\}$ (voir (104)), pour une multiplication par k des effectifs, on a, en désignant par $-Q_3^k(a, \bar{b})$ l'indice concerné :

$$-Q_3^k(a, \bar{b}) = \sqrt{k} \times [-Q_3^1(a, \bar{b})] \tag{122}$$

(voir (16)).

Pour palier à cet inconvénient R. Gras et al. [6] ont proposé de remplacer l'indice probabiliste $\mathcal{I}(a \rightarrow b)$ (voir 64) par un indice qui représente la moyenne géométrique entre $\mathcal{I}(a \rightarrow b)$ et un "*Indice d'Inclusion*" fondé sur l'entropie de Shannon. Pour $0 < \xi < 1$, considérons l'entropie associée à la distribution binaire $(\xi, 1-\xi)$; soit $H(\xi, 1-\xi) = -\xi log_2 \xi - (1-\xi) log_2(1-\xi)$. Son instanciation pour chacune des deux distributions conditionnelles $\Big(p(b/a) = n(a \wedge b)/n(a), p(\bar{b}/a) = n(a \wedge \bar{b})/n(a)\Big)$ et $\Big(p(\bar{a}/\bar{b}) = n(\bar{a} \wedge \bar{b})/n(\bar{b}), p(a/\bar{b}) = n(a \wedge \bar{b})/n(\bar{b})\Big)$ donne respectivement les entropies qu'on notera $H(b/a)$ et $H(\bar{a}/\bar{b})$. La première est associée à l'expression formelle $(a \rightarrow b)$ (a implique b) et la seconde, à l'expression formelle de la contraposée $(\bar{b} \rightarrow \bar{a})$ ($\bar{b}$ implique $\bar{a}$). L'indice d'inclusion τ d'une règle prend la forme :

$$\tau_\omega(a \rightarrow b) = [(1 - H^\star(b/a)^\omega) \times (1 - H^\star(\bar{a}/\bar{b})^\omega)]^{1/2\omega} \tag{123}$$

où on considère $\omega > 1$ et où

$$\begin{aligned} H^\star(b/a) = H(b/a) \text{ si } p(\bar{b}/a) < \frac{1}{2} \\ = 1 \text{ sinon} \\ H^\star(\bar{a}/\bar{b}) = H(\bar{a}/\bar{b}) \text{ si } p(a/\bar{b}) < \frac{1}{2} \\ = 1 \text{ sinon} \end{aligned} \tag{124}$$

La valeur du paramètre ω correspond à un choix, celle conseillée que nous adopterons dans la suite est $\omega = 2$. Dans ces conditions, l'*Indice d'Implication Entropique* noté dans [10] IIE et que nous noterons ici $IIEG$ par référence à son introduction par Régis Gras, s'écrit :

$$IIEG(a \rightarrow b) = [\mathcal{I}_{p(b)}(a \rightarrow b) \times \tau_2(a \rightarrow b)]^{1/2} \tag{125}$$

où $\mathcal{I}_{p(b)}(a \rightarrow b)$ n'est autre que l'indice (64) calculé par rapport au modèle Poissonnien de l'hypothèse d'absence de liaison. L'indice inférieur $p(b)$ ci-dessus se réfère à la formalisation paramétrique introduite dans [10]. Le paramètre de référence du modèle aléatoire est l'indice "Confiance" $p(b/a)$. L'indice de "*Vraisemblance du Lien*" étant établi par rapport à l'hypothèse d'indépendance où

on a : $p(b/a) = p(b)$.

Une autre version de l'IIE correspond pour l'indice probabiliste d'implication à se situer par rapport à l'indétermination définie par : $n(a \wedge \bar{b}) = \frac{1}{2}n(a)$; c'est-à-dire, aussi, de façon équivalente, par : $n(a \wedge b) = \frac{1}{2}n(a)$; soit $p(b/a) = \frac{1}{2}$. Cette référence est d'autant plus naturelle que l'indice d'inclusion s'annule pour $p(b/a) > \frac{1}{2}$. Cette variante a été établie dans [2] et analysée dans [10]. Nous la noterons $IIEL$ par référence à [10](L comme Lallich et al.). Se référant au modèle binomial, l'indice probabiliste associé peut s'écrire :

$$\mathcal{I}_{0.5}(a \rightarrow b) = \sum_{k=n(a\wedge\bar{b})+1}^{n(a)} C_{n(a)}^{k}(0.5)^{n(a)} \tag{126}$$

Ainsi, la deuxième version de l'*Intensité d'Implication Entropique* s'exprime comme suit :

$$IIEL(a \rightarrow b) = [\mathcal{I}_{0.5}(a \rightarrow b) \times \tau_2(a \rightarrow b)]^{1/2} \tag{127}$$

Cet indice a bien été considéré dans [2]. Cependant, les auteurs ont préféré substituer à l'indice τ_2, celui défini par $\frac{1}{2}(1 + \tau_2)$. L'indice global obtenu que nous notons $IIEB$ (B comme Blanchard et al.) s'écrit :

$$IIEB(a \rightarrow b) = [\mathcal{I}_{0.5}(a \rightarrow b) \times \frac{1}{2}(1 + \tau_2)(a \rightarrow b)]^{1/2} \tag{128}$$

En utilisant l'approximation normale de la loi binomiale on obtient, avec une approximation très précise dans les cas usuels :

$$\mathcal{I}_{0.5}(a \rightarrow b) = \Phi\Big(\frac{-n(a \wedge \bar{b}) + \frac{n(a)}{2}}{\sqrt{n(a)/4}}\Big) \tag{129}$$

On propose dans [10] toute une famille d'indices IIE paramétrés par $\theta = p(b/a)$ où θ est un paramètre qui peut être fixé par l'utilisateur. Nous nous limiterons quant à nous dans notre analyse du comportement par rapport aux deux modèles de croissance $M1$ et $M2$, aux deux formes basiques (125) et (127). Les résultats pourront être étendus à toute la famille paramétrée.

5.4.2 Analyse du comportement de l'IIE par rapport à $M1$ et $M2$

Modèle $M1$ de croissance des effectifs Ce modèle où tous les cardinaux du tableau de contingence croisant $\{a, \bar{a}\}$ avec $\{b, \bar{b}\}$, sont multipliés par k, a été précisé par les relations (104).

Il est clair que les différentes fréquences relatives $p(b/a)$, $p(\bar{b}/a)$, $p(\bar{a}/\bar{b})$ et $p(a/\bar{b})$ sont invariables suite à une dilatation k telle qu'elle est définie par les relations (104). Ainsi, l'indice d'inclusion $\tau_2(a \rightarrow b)$ est constant. Considérons maintenant le facteur composant de IIE défini par l'indice d'implication probabiliste, soit $\mathcal{I}_{p(b)}(a \rightarrow b)$. Le cas d'intérêt est celui où $n(a \wedge \bar{b}) < \frac{n(a)\times n(\bar{b})}{n}$ qui s'exprime de façon équivalente par $p(a \wedge \bar{b}) < p(a) \times p(\bar{b})$ ou $p(a \wedge b) > p(a) \times p(b)$ ou aussi - en se référant à l'indice Confiance - $p(b/a) > p(b)$. Dans ce cas et pour k croissant, l'indice $\mathcal{I}_{p(b)}(a \rightarrow b)$ tend en croissant "très rapidement" vers l'unité.

En effet, nous avons bien vu qu'il pouvait être calculé avec une grande précision au moyen de la formule :

$$\mathcal{I}_{p(b)}(a \to b) = \Phi\{\sqrt{k} \times [-Q_3^1(a, \bar{b})]\} \tag{130}$$

(voir (122)), où nous avons noté $\mathcal{I}^k_{p(b)}(a \to b)$ la valeur de l'indice probabiliste pour une dilatation k.

À titre d'illustration considérons le cas de la table de contingence 6 où $n = 273$, $n(a \wedge \bar{b}) = 28$, $n(a) = 76$ et $n(\bar{b}) = 153$; de sorte que l'indice $-Q_3^1(a, \bar{b})$ s'écrit :

$$[-28 + (76 \times 153/273)]/\sqrt{(76 \times 153/273)} = 2.236$$

La valeur correspondante de $\mathcal{I}_{p(b)}(a \to b)$ est 0.9873. À partir de $k = 4$ - ce qui correspond à un effectif total de 1092 - la valeur de $\mathcal{I}_{p(b)}(a \to b)$ dépasse 0.999994 ; c'est-à-dire, est pratiquement égale à 1. Ainsi, l'indice $IIEG(a \to b)$ se réduit pratiquement à l'indice d'inclusion $[\tau_2(a \to b)]^{1/2}$ dès lors que le nombre n d'observations est "assez grand". Il en résulte que le caractère discriminant de $IIEG(a \to b)$ pour n "assez grand" est presque exclusivement dû à sa composante inclusive, sa composante définie par l'indice probabiliste de la "*vraisemblance du lien*" dite " *intensité d'implication*" n'intervenant pour ainsi dire pas.

Proposition 9 *Sous le modèle $M1$ de croissance, quel que soit le nombre initial n d'observations, dans le cas où $p(b/a) > p(b)$, l'*Indice d'Implication Entropique $IIEG(a \to b)$ *tend vers l'*Indice d'Inclusion $[\tau_2(a \to b)]^{1/2}$*, lequel restant invariant.* □

Nous avons pu nous rendre compte que cette tendance est "très rapide".

Le phénomène est en tout point analogue en ce qui concerne l'indice $IIEL(a \to b)$ où la référence pour l'indice probabiliste est l'indétermination $(n(a \wedge \bar{b}) = n(a)/2)$. Avec des notations que l'on comprend, il est clair que pour $n(a \wedge \bar{b}) < n(a)/2$, l'expression analogue à celle (130) devient :

$$\mathcal{I}_{p(b)}(a \to b) = \Phi\{\sqrt{k} \times [-\frac{n(a \wedge \bar{b}) - \frac{n(a)}{2}}{\sqrt{n(a)/4}}]\} \tag{131}$$

En reprenant l'exemple ci-dessus on a :

$$\mathcal{I}_{0.5}(a \to b) \simeq \Phi\Big(- \frac{28 - (76/2)}{\sqrt{76/4}} = \Phi(2.294) = 0.9891\Big)$$

D'autre part, pour $k = 4$ on obtient :

$$\mathcal{I}^4_{0.5}(a \to b) = \Phi(\sqrt{4} \times 2.294) = \Phi(4.588) > 0.999997$$

La condition $n(a \wedge \bar{b}) < n(a)/2$ étant équivalente à $p(b/a) > 0.5$, on peut énoncer :

Proposition 10 *Sous le modèle $M1$ de croissance, quel que soit le nombre initial n d'observations, dans le cas où $p(b/a) > 0.5$, les* indices d'implication entropiques $IIEL(a \rightarrow b)$ *et* $IIEB(a \rightarrow b)$ *tendent respectivement vers les indices d'inclusion* $[\tau_2(a \rightarrow b)]^{1/2}$ *et* $\left(\frac{1}{2} \times \{1 + [\tau_2(a \rightarrow b)]\}\right)^{1/2}$ *lequels restant invariants.* □

Modèle $M2$ de croissance des effectifs Ce modèle est défini par les relations (105). Nous allons commencer par analyser le comportement de $IIE(a \rightarrow b)$. Nous avons déjà largement étudié le comportement de $-Q_{3x}(a, \bar{b})$ (voir sous section 5.3.3). Cet indice est croissant par rapport à x ; d'abord fortement ; puis plus faiblement. De toute façon, $\mathcal{I}_{p(b)}(a \rightarrow b)$ tend "rapidement" vers 1. En reprenant l'exemple ci-dessus (Table 6), on a pour $x = 50$, $-Q_{350}(a, \bar{b}) = 2.86$; pour $x = 100$, $-Q_{3100}(a, \bar{b}) = 3.28$; pour $x = 200$, $-Q_{3200}(a, \bar{b}) = 3.81$; pour $x = 300$, $-Q_{3x}(a, \bar{b}) = 4.14$; ...$\Phi(4.14) = 0.999976$.

Considérons à présent l'indice d'inclusion $\tau_2(a \rightarrow b)$ [voir (123) et (124)]. Pour mieux voir, commençons par expliciter $H(b/a)$ et $H(\bar{a}/\bar{b})$:

$$\begin{aligned} H(b/a) &= -p(b/a)log_2[p(b/a)] - p(\bar{b}/a)log_2[p(\bar{b}/a)] \\ H(\bar{a}/\bar{b}) &= -p(\bar{a}/\bar{b})log_2[p(\bar{a}/\bar{b})] - p(a/\bar{b})log_2[p(a/\bar{b})] \end{aligned} \tag{132}$$

Vu que les éléments intervenant dans la définition de $H(b/a)$ et de $H^\star(b/a)$ ne font intervenir que $p(a \wedge b)$, $p(a)$ et $p(b)$, le facteur $[1 - H^\star(b/a)^2]$ est invariable. Examinons à présent $H(\bar{a}/\bar{b})$. $n(a \wedge \bar{b})$ est invariable (il s'agit simplement du complément de $n(a \wedge b)$ par rapport à $n(a)$). $n(\bar{b})$ étant remplacé par $n(\bar{b}) + x$, la condition $p(a/\bar{b}) < \frac{1}{2}$ devient moins restrictive puisqu'elle correspond à $n(a \wedge \bar{b})/[n(\bar{b}) + x] < \frac{1}{2}$. En désignant par b_x l'attribut b tel qu'il devient après la transformation $n(\bar{a} \wedge \bar{b}) \leftarrow n(\bar{a} \wedge \bar{b}) + x$, on montre aisément que $p(a/\bar{b}_x)$ diminue et tend vers 0 pour x "assez grand". D'autre part, la définition de $H(b_x/a)$ se faisant au niveau de $\mathcal{O}(a)$, on a $H(b_x/a) = H(b/a)$. En désignant par $IIE(a \rightarrow b_x)$ l' *Intensité d'Implication Entropique* résultant de l'accroissement x de $n(\bar{a} \wedge \bar{b})$, on a la proposition suivante :

Proposition 11 *Sous le modèle $M2$ de croissance $IIEG(a \rightarrow b_x)$ tend vers* $[1 - H(b/a)^2]^{1/8}$ *pour x assez grand.* □

Considérons à présent le cas des indices $IIEL(a \rightarrow b)$ et $IIEB(a \rightarrow b)$; on se place maintenant pour l'indice probabiliste d'implication, par rapport à l'indétermination (voir (128)). Il est clair que sous le modèle $M2$ de croissance $\mathcal{I}_{0.5}(a \rightarrow b)$ est invariable. D'autre part, le facteur multiplicatif $\tau_2(a \rightarrow b)$ en ce qui concerne $IIEL$ et $\left(\frac{1}{2} \times \{1 + [\tau_2(a \rightarrow b)]\}\right)^{1/2}$ en ce qui concerne $IIEB$ intervient de la même façon que pour celui $IIEG(a \rightarrow b)$. On peut ainsi énoncer la proposition :

Proposition 12 *Sous le modèle $M2$ de croissance $IIEL(a \rightarrow b_x)$ et $IIEB(a \rightarrow b_x)$ tendent respectivement vers* $[\mathcal{I}_{0.5}(a \rightarrow b) \times (1 - H(b/a)^2)^{1/4}]^{1/2}$ *et* $[\mathcal{I}_{0.5}(a \rightarrow b) \times \frac{1}{2}\left(1 + [(1 - H(b/a)^2)^{1/4}]\right)]^{1/2}$. □

6. Analyse Expérimentale dans le cadre de la base « Wages »

Dans cette section, nous exposons trois séries d'expériences qui ont été réalisées avec la base de données "*Wages*", données disponibles sur UCI KDD archive (Murphy et Aha, 1995). Nous commençons par décrire la base de données *Wages* avec laquelle nous effectuons nos comparaisons expérimentales des différents indices d'implication discriminants.

6.1. Description de la base

La base de données *Wages* se compose de *534* enregistrements décrits par *11* attributs dont *4* sont des attributs numériques : éducation (*nombre d'années d'étude*), expérience (*nombre d'années d'expérience professionnelle*), salaire (*dollars par heure*) et âge.

La description des attributs numériques est donnée dans le *tableau 6.1*, où pour chaque attribut, nous donnons sa valeur minimale, sa valeur maximale, sa moyenne et le nombre de valeurs distinctes.

Nom	Minimale	Maximale	Moyenne	Distinct
éducation	2	18	13,02	17
expérience	0	55	17,82	52
salaire	1	44,5	9,02	238
âge	18	64	36,83	47

TAB. 6.*1 : Description des attributs numériques.*

Les *7* attributs qualitatifs de cette base sont : région (*nord, sud*), sexe (*féminin, masculin*), syndiqué (*oui, non*), origine ethnique (*hispanique, blanche, autre*), emploi (*cadre, vendeur, employé, ouvrier, profession libérale, autre*), secteur (*fabrication, construction, autre*) et marié (*oui, non*).

La description des attributs qualitatifs est donnée dans le *tableau 6.2*, où pour chaque attribut, nous donnons les différentes valeurs prises par cet attribut, la sémantique associée à chacune de ces valeurs, le nombre d'individus vérifiant la valeur de l'attribut et le pourcentage d'individus correspondant pour l'attribut donné.

Nom	Valeur	Sémantique	Nombre	Pourcentage
région	0	nord	378	70,79
	1	sud	156	29,21
sexe	0	masculin	289	54,12
	1	féminin	245	45,88
syndiqué	0	non syndiqué	438	82,02
	1	syndiqué	96	17,98
origine ethnique	1	autre origine	67	12,55
	2	hispanique	27	5,06
	3	blanc	440	82,39
emploi	1	cadre	55	10,30
	2	vendeur	38	7,12
	3	employé	97	18,16
	4	ouvrier	83	15,54
	5	profession libérale	105	19,66
	6	autre profession	156	29,22
secteur	0	autre secteur	411	76,97
	1	fabrication	99	18,54
	2	construction	24	4,49
marié	0	non marié	184	34,46
	1	marié	350	65,54

TAB. 6.*2 : Description des attributs qualitatifs.*

Afin d'extraire les différentes règles, une étape de pré-traitement est indispensable, l'étape de discrétisation des attributs numériques.

Les *tableaux 6.3* à *6.6* restituent la discrétisation effectuée sur ces *4* attributs numériques par la méthode des quantiles ou effectifs égaux (*obtenir pour chaque intervalle des effectifs de même importance*). Le nombre d'intervalles retenu est *5* pour avoir des attributs du type " *attribut vérifiant les*

très faibles valeurs", " *attribut vérifiant les faibles valeurs*", " *attribut ayant des valeurs moyennes*", " *attribut ayant des valeurs au dessus de la moyenne*" et " *attribut vérifiant les fortes valeurs*".

Intervalle	Effectif
[*2 ; 11*]	*83*
12	*219*
[*13 ; 14*]	*93*
[*15 ; 16*]	*84*
[*17 ; 18*]	*55*

TAB. 6.*3 : Discrétisation de l'attribut "éducation".*

Intervalle	Effectif
[*0 ; 7*[	*104*
[*7 ; 13*[	*104*
[*13 ; 18*[	*106*
[*18 ; 29*[	*108*
[*29 ; 55*]	*112*

TAB. 6.*4 : Discrétisation de l'attribut "expérience".*

Intervalle	Effectif
[*1 ; 5*[	*107*
[*5 ; 6,67*[	*107*
[*6,67 ; 9*[	*100*
[*9 ; 12,5*[	*113*
[*12,5 ; 44,5*]	*107*

TAB. 6.*5 : Discrétisation de l'attribut "salaire".*

Intervalle	Effectif
[*18 ; 26*[	*91*
[*26 ; 32*[	*108*
[*32 ; 38*[	*121*
[*38 ; 48*[	*106*
[*48 ; 64*]	*108*

TAB. 6.*6 : Discrétisation de l'attribut "âge".*

Suite à cette discrétisation, nous parlerons des attributs "*éducation1*", "*éducation2*", "*éducation3*", "*éducation4*" et "*éducation5*" pour respectivement les attributs "*éducation* = [*2, 11*]", "*éducation = 12*", "*éducation* = [*13, 14*]", "*éducation* = [*15, 16*]" et "*éducation* = [*17, 18*]". Cette notation sera retenue pour les trois autres attributs numériques "*expérience*", "*salaire*" et "*âge*".

Après avoir effectué cette discrétisation pour les attributs numériques et après la transformation des attributs qualitatifs en attributs booléens par un codage disjonctif complet, nous obtenons une base de données composée de *40* attributs qualitatifs booléens et *534* individus.

Après avoir décrit la base de données et défini l'étape de pré-traitement à l'extraction des règles, nous exposons maintenant notre première série d'expériences : la comparaison des *20* meilleures règles.

6.2. Comparaison des *20* meilleures règles

L'objectif de cette première série d'expériences est de tenter de dégager des comportements communs entre les indices d'implication discriminants *VLgrImpP*, *VTeImpBarP*, *VTeImpCorP* et *VTeImpProj* en étudiant les *20* meilleures règles extraites par ces indices. Dans les expériences qui vont suivre la valeur de *e* considérée est *100*. Cependant, rien n'empêche de prendre une autre valeur de *e*.

Soit min_{sup} ($min_{sup} \in [0, 1]$) le support minimum défini par l'utilisateur et soit min_{conf} ($min_{conf} \in [0, 1]$) la confiance minimum définie par l'utilisateur.

Le protocole expérimental utilisé est le suivant :

1. Trouver l'ensemble Ccs des couples d'attributs (a,b) vérifiant les contraintes de support ($p(a \wedge b) \geq min_{sup}$), de confiance ($p(b/a) \geq min_{conf}$) et la contrainte suivante $n(a) \leq n(b)$,

2. Pour chacun de ces couples, calculer les quatre indices $Q_3(a,\bar{b})$ [cf. expression (6) et (16)], *VteImpBarP* [voir sous sections 5.2.2 et 3.2], *VTeImpCorP* [cf. expression (98)] et *VTeImpProj* [cf. expression (103)]. À partir de l'indice normalisé, on se référera à chaque fois à la fonction de répartition de la loi normale centrée et réduite pour obtenir l'indice probabiliste.

3. Calculer la moyenne empirique et l'écart-type empirique de $Q_3(a,\bar{b})$ pour tous les couples (*a,b*) de Ccs et ensuite normaliser l'indice $Q_3(a,\bar{b})$ avant de faire appel à la loi normale.

4. Pour chacun des indices, trier les règles extraites par ordre décroissant selon la valeur de l'indice et conserver les *20* meilleures règles.

L'algorithme permettant de réaliser cette première série d'expériences est le suivant :

Algorithme 1

Entrée : une table de données $O \times A$, un support minimum min_{sup} et une confiance minimum min_{conf}.

Sortie : les ensembles R_1, R_2, R_3 et R_4 des *20* meilleures règles pour chacun des *4* indices.

DEBUT

// **(0) Initialisation de la variable qui va récupérer les règles valides avec toutes les mesures des indices**

$R = \emptyset$;

// **(1) Calcul de l'ensemble *Ccs* des couples (*a,b*) vérifiant les contraintes**

$Ccs = \{(a,b) \in A \times A \,/\, p(a \wedge b) \geq min_{sup} \text{ et } p(b/a) \geq min_{conf} \text{ et } n(a) \leq n(b)\}$;

// **(2) Calcul d'une partie des indices**

Pour tout $(a,b) \in Ccs$ **faire**

Calcul du coefficient normalisé $Q_3(a,\bar{b})$;

Calcul des indices *VTeImpBarP*, *VTeImpCorP* et *VTeImpProj* ;

$R = [R ; a \;\; b \;\; Q_3(a,\bar{b}) \;\; VTeImpBarP \;\; VTeImpCorP \;\; VTeImpProj]$;

Fin Pour tout

// **(3) Calcul de** *VLgrImpP*

$moy_{cs}(Q_3) = \text{mean}(R[:, 3])$; // calcul moyenne empirique de $Q_3(a,\bar{b})$

$var_{cs}(Q_3) = \text{var}(R[:, 3])$; // calcul variance empirique de $Q_3(a,\bar{b})$

Pour tout $(a,b) \in Ccs$ **faire**

$$VLgrImpP = 1 - \phi\left(\frac{Q_3(a,\bar{b}) - moy_{cs}(Q_3)}{\sqrt{var_{cs}(Q_3)}}\right) \text{ ; // Calcul de l'indice } VLgrImpP$$

$R[(a,b), 3] = VLgrImpP$; // substitution de $Q_3(a,\bar{b})$ par *VLgrImpP* dans R

Fin Pour tout

// **(4) Restitution des *20* meilleures règles**

$R_1 = [\text{sortrows}(R, 3)](1{:}20, :)$; // *20* meilleures règles pour *VLgrImpP*

$R_2 = [\text{sortrows}(R, 4)](1{:}20, :)$; // *20* meilleures règles pour *VTeImpBarP*

$R_3 = [\text{sortrows}(R, 5)](1{:}20, :)$; // *20* meilleures règles pour *VTeImpCorP*

$R_4 = [\text{sortrows}(R, 6)](1{:}20, :)$; // *20* meilleures règles pour *VTeImpProj*

Retourner R_1, R_2, R_3 et R_4

FIN

L'extraction a été effectuée avec les contraintes suivantes : un support minimum min_{sup} de *0,01* et une confiance minimum min_{conf} de *0,80*.

Il s'avère que l'ensemble des *20* meilleures règles est **le même** pour ces quatre indices et est restitué dans le *tableau 6.5*. Pour chacune des règles $a \rightarrow b$ extraites, le *tableau 6.5* donne la prémisse *a*, la conclusion *b*, le support $p(a \wedge b)$, la confiance $p(b/a)$ et les valeurs des quatre indices étudiés.

Règles		Mesures des RA[1]		Indices étudiés			
Prémisse	Conclusion	support	confiance	*VLgrImpP*	*VTeImpBarP*	*VTeImpCorP*	*VTeImpProj*
âge5	expérience5	0,19	0,94	1	1	1	1
âge1	expérience1	0,15	0,88	1	1	1	1
autre profession	masculin	0,24	0,81	0,99	0,98	0,98	0,98
ouvrier	autre secteur	0,15	0,98	0,95	0,92	0,92	0,95
salaire1	non syndiqué	0,19	0,95	0,89	0,87	0,87	0,91
construction	autre profession	0,04	0,83	0,88	0,85	0,86	0,90
employé	autre secteur	0,16	0,91	0,83	0,83	0,84	0,88
construction	masculin	0,04	0,92	0,81	0,78	0,80	0,86
âge5	marié	0,16	0,81	0,80	0,83	0,83	0,86
féminin	non syndiqué	0,41	0,89	0,76	0,81	0,81	0,84
employé	non syndiqué	0,17	0,92	0,72	0,75	0,76	0,82
vendeur	non syndiqué	0,07	0,97	0,72	0,69	0,70	0,81
cadre	non syndiqué	0,10	0,95	0,71	0,71	0,72	0,81
salaire2	non syndiqué	0,18	0,91	0,69	0,74	0,74	0,80
profession libérale	autre secteur	0,17	0,87	0,68	0,74	0,75	0,80
féminin	autre secteur	0,38	0,83	0,68	0,77	0,77	0,80
cadre	autre secteur	0,09	0,89	0,64	0,68	0,69	0,77
éducation5	autre secteur	0,09	0,89	0,64	0,68	0,69	0,77
âge1	non syndiqué	0,15	0,90	0,62	0,69	0,69	0,76
expérience1	non syndiqué	0,17	0,89	0,61	0,69	0,69	0,76

TAB. 6.7 *: Ensemble des 20 meilleures règles extraites sur Wages.*

Afin de comparer les meilleures règles extraites par chacun des indices, nous transformons le *tableau 6.5* en remplaçant les valeurs de chacun des indices par le rang de la règle : le rang *1* indiquant que c'est la meilleure règle (*d'intensité la plus forte*) et le rang *20*, la plus mauvaise règle (*d'intensité la plus faible*) parmi l'ensemble R_i ($i \in \{1,..,4\}$) des *20* meilleures règles extraites. Le *tableau 6.8* restitue ces différents rangs pour chacun des indices.

[1]RA : règles d'association.

Règles		Indices étudiés			
Prémisse	**Conclusion**	***VLgrImpP***	***VTeImpBarP***	***VTeImpCorP***	***VTeImpProj***
âge5	expérience5	1	1	1	1
âge1	expérience1	2	2	2	2
autre profession	masculin	3	3	3	3
ouvrier	autre secteur	4	4	4	4
salaire1	non syndiqué	5	5	5	5
construction	autre profession	6	6	6	6
employé	autre secteur	7	8	7	7
construction	masculin	8	10	10	8
âge5	marié	9	7	8	9
féminin	non syndiqué	10	9	9	10
employé	non syndiqué	11	12	12	11
vendeur	non syndiqué	12	16	16	12
cadre	non syndiqué	13	15	15	13
salaire2	non syndiqué	14	14	14	14
profession libérale	autre secteur	15	13	13	15
féminin	autre secteur	16	11	11	16
cadre	autre secteur	17	19	19	17
éducation5	autre secteur	18	20	20	18
âge1	non syndiqué	19	18	18	19
expérience1	non syndiqué	20	17	17	20

TAB. 6.*8 : Rang des 20 meilleures règles pour chacun des 4 indices.*

Nous constatons que :

- les indices *VLgrImpP* et *VTeImpProj* classent les *20* meilleures règles exactement de la même façon,
- les indices *VTeImpBarP* et *VTeImpCorP* classent les règles de façon quasiment similaire puisqu'il y a juste une inversion entre la règle "*employé* → *autre secteur*" et "*âge5* → *marié*" qui arrivent en septième ou huitième position selon l'indice (*voir les cellules en grisé ou orangé du tableau 6.8*).

La différence vient entre ces deux groupes de mesures, c'est-à-dire le groupe G_1 = {*VLgrImpP, VTeImpProj*} et le groupe G_2 = {*VTeImpBarP, VTeImpCorP*}). Les *6* meilleures règles sont les mêmes pour ces deux groupes de mesures et ensuite les classements divergent. Cette divergence se visualise aisément grâce à la *figure 1* où nous avons retenu le classement d'une des deux mesures du groupe G_1 comme référence (*voir la première bissectrice de la figure 1*) et le classement de la mesure *VTeImpBarP* du groupe G_2 car elle a plus de divergence avec G_1 et que nous pouvons visualiser grâce à la courbe en pointillé de la *figure 1*.

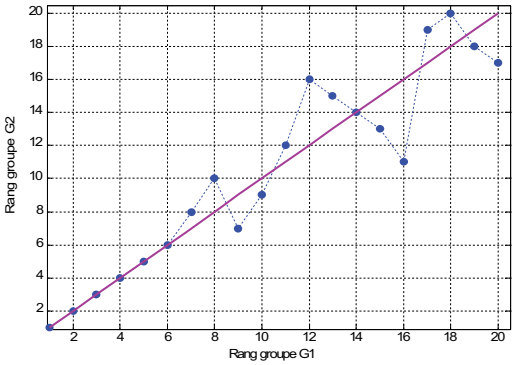

Fig. *1* : *Classement des règles pour les deux groupes de mesures.*

Les *6* règles sous la première bissectrice sont les règles qui ont été mieux classées par le deuxième groupe G_2 de mesures et les *7* règles au dessus de la première bissectrice sont celles qui ont été le mieux classées par le premier groupe G_1.

Relativement à l'évaluation d'une implication ($a \rightarrow b$) où a et b sont deux attributs booléens, il n'est pas étonnant que les comportements de *VTeImpCorP* et *VTeImpBarP* soient similaires. En effet dans les deux cas, partant du tableau des données $O \times \{a, b\}$, on procède à une réduction forcée à e objets. Pour *VTeImpCorP*, elle est directement établie à partir du tableau des données (*voir sous-sections 4.1 et 5.2.2*). Quant à *VTeImpBarP* la réduction se situe au niveau du tableau de contingence initial croisant $\{a, \bar{a}\}$ avec $\{b, \bar{b}\}$ et cela, en multipliant les contenus respectifs des cases par e/n et en encadrant le tableau des décimaux ainsi obtenu au moyen de *8* tableaux de contingence à effectifs entiers portant chacun sur e éléments (*voir sous-sections 3.2 et 5.2.2*).

Le principe de *VLgrImpP* est tout différent. Relativement à l'évaluation d'une implication $a \rightarrow b$, c'est un indice localement normalisé [cf. (6) ou (73)] qui est rapporté relativement à un ensemble potentiel de règles "*intéressantes*" pour l'implication [cf. (88)]. L'indice local pour un couple ($a \rightarrow b$) est alors empiriquement "*standardisé*" par rapport à la distribution de cet indice sur l'ensemble potentiel mentionné (*voir avant dernier alinéa précédant la sous-section 5.2.2*).

Le comportement de *VTeImpProj* est davantage comparable à celui de *VLgrImpP* qu'à celui des indices du groupe G_2 . La conception de l'indice *VTeImpProj* suppose également la normalisation intrinsèque d'un indice déjà constitué par rapport rapport à un modèle aléatoire (*voir sous-sections 4.2 et 5.2.4*). Bien que ne s'agissant pas d'une normalisation empirique, il y a dans la philosophie de la conception de *VTeImpProj*, un élément de ressemblance avec celle de *VLgrImpP*. Cependant, ces deux indices restent fondamentalement distincts.

L'évaluation comparée de ces différents indices (*VLgrImpP, VTeImpBarP, VTeImpCorP* et *VTeImpProj*) se fait conformément à une échelle de probabilité fournie par la fonction de répartition de la loi normale centrée et réduite.

Après avoir étudié les *20* meilleures règles extraites par les *4* indices d'implication discriminants, et qui a constitué la première série d'expériences, nous passons à la deuxième série d'expériences.

6.3. Étude de certaines situations caractéristiques

Cette seconde série d'expériences a pour objectif d'étudier le comportement limite des quatre indices *VLgrImpP*, *VTeImpBarP*, *VTeImpCorP* et *VTeImpProj* dans le contexte de données volumineuses et pour cela, nous étudions le comportement de ces indices dans certaines situations caractéristiques et selon les modèles de croissance M_1 et M_2 exposés précédemment.

Nous commençons par présenter ces situations caractéristiques, ensuite nous donnerons le protocole expérimental utilisé pour les deux modèles de croissance et nous terminerons par l'étude du comportement des *4* indices.

6.3.1 Situations caractéristiques

Nous savons qu'il existe différents états caractéristiques pour une règle $a \rightarrow b$ qui sont résumés dans la *figure 1b* et qui sont les suivants :

- l'**incompatibilité** : c'est le cas où il n'y a aucun individu qui vérifie à la fois a et b c'est-à-dire lorsque $n(a \wedge b)=0$ ou encore lorsque $p(b/a) = 0$,
- l'**indépendance** : c'est le cas où la réalisation de a n'augmente pas les chances d'apparition de b c'est-à-dire lorsque $n(a \wedge b)=\frac{n(a)n(b)}{n}$ ou encore lorsque $p(b/a) = p(b)$,
- l'**indétermination** ou l'**équilibre** : c'est le cas où lorsque a est réalisé, il y a autant de chances de voir apparaître b que non b, c'est-à-dire lorsque $n(a \wedge b)=\frac{n(a)}{2}$ ou encore lorsque $p(b/a)=\frac{1}{2}$,
- l'**implication logique** : c'est le cas où il n'y a pas de contre-exemple c'est-à-dire lorsque $n(a \wedge b)=n(a)$ ou encore lorsque $p(b/a) = 1$.

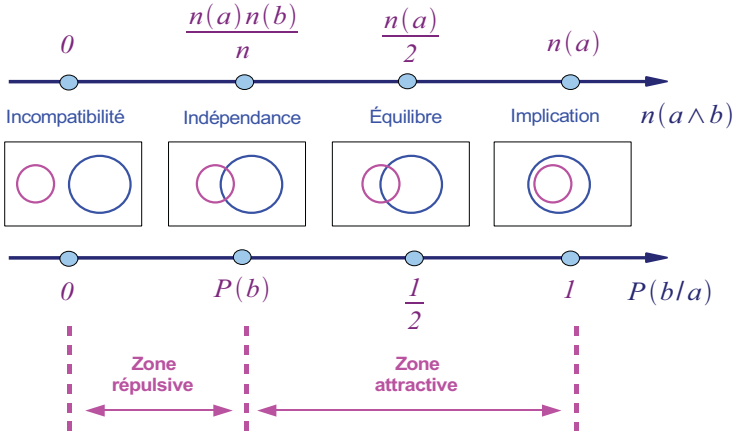

Fig. *1b* : États caractéristiques d'une règle.

La *figure 1b* a été reprise de la thèse de Sylvie Guillaume (Guillaume, 2000) à laquelle nous avons rajouté le point caractéristique d'indétermination qui a été abordé pour la première fois par Julien Blanchard dans (Blanchard et al. 2005).

Le cas de l'indétermination peut se trouver de part et d'autre de l'indépendance, position fonction de la taille $n(b)$ de la conclusion b par rapport à la taille n de l'ensemble d'apprentissage divisée par *2*. En effet, dans le cas de l'indétermination nous avons $n(a \wedge b)=\frac{n(a)}{2}$ et dans le cas de l'indépendance, nous avons $n(a \wedge b)=\frac{n(a)n(b)}{n}$. Ainsi, l'indétermination sera avant l'indépendance $\left(\frac{n(a)}{2}<\frac{n(a)n(b)}{n}\right)$ lorsque $n(b)>\frac{n}{2}$ et après $\left(\frac{n(a)}{2}>\frac{n(a)n(b)}{n}\right)$ lorsque $n(b)<\frac{n}{2}$.

Nous reprenons le vocabulaire défini dans (Guillaume, 2000) concernant les zones attractive et répulsive. Nous rappelons que la zone attractive est celle où lorsque la prémisse a est réalisée, les chances d'apparition de la conclusion b sont augmentées par rapport à l'hypothèse probabiliste d'indépendance ; c'est-à-dire lorsque $p(b/a) > p(b)$. La zone répulsive est donc celle où lorsque la prémisse a est réalisée, les chances d'apparition de la conclusion b sont diminuées c'est-à-dire lorsque $p(b/a) < p(b)$.

Après avoir défini les situations caractéristiques, nous donnons le protocole expérimental utilisé pour les modèles de croissance M_1 et M_2.

6.3.2 Protocole expérimental

Dans un premier temps, nous donnons une description formalisée du modèle de croissance M_1 (*voir* (*104*) *sous-section 5.3.1*) ainsi que l'algorithme utilisé pour tracer les différentes courbes permettant de révéler le comportement des indices dans certaines situations caractéristiques.

Modèle de croissance M_1

Ce premier modèle de croissance M_1 consiste à multiplier tous les effectifs $n(a)$, $n(b)$, $n(a \wedge b)$ et n par un coefficient multiplicateur k, k étant un entier non nul ($k \in \mathbb{N}^*$).

Soient les paramètres suivants :
- une règle $a \rightarrow b$ c'est-à-dire une situation caractéristique,
- un coefficient multiplicateur maximal $kMax$ ($kMax \in \mathbb{N}^*$),
- un pas d'évolution $pasK$ ($pasK \in \mathbb{N}^*$) pour le coefficient multiplicateur,
- un seuil minimum de support min_{sup} ($min_{sup} \in [0, 1]$) et
- un seuil minimum de confiance min_{conf} ($min_{conf} \in [0, 1]$).

Après avoir déterminé tous ces paramètres, nous utilisons le protocole suivant :

1. Rechercher l'ensemble Ccs [cf. (88)] des couples d'attributs (a,b) vérifiant les contraintes de support ($p(a \wedge b) \geq min_{sup}$) et de confiance ($p(b/a) \geq min_{conf}$).

2. Pour chaque coefficient multiplicateur k, calculer les quatre indices $VlgrImpP$ [expression de même type que (84) où l'indice supérieur 0 est remplacé par gcs], $VteImpBarP$ [cf. expression dans la sous-section 5.2.2], $VTeImpCorP$ [cf. expression (98)] et $VTeImpProj$ [cf. expression (103)]. Pour le calcul de $VLgrImpP$, il faut au préalable calculer l'indice $Q_3(a, \bar{b})$ pour chaque couple (a,b) de Ccs comme exposé dans la *section 6.2* (*comparaison des 20 meilleures règles*).

3. Tracer la courbe d'évolution $\mathsf{C_{O1}}$ des différents indices en fonction du coefficient multiplicateur k.

L'algorithme permettant de réaliser les courbes pour le modèle de croissance M_1 est le suivant :

Algorithme 2

Entrée : une table de données $\mathsf{O} \times \mathsf{A}$, une règle $a \rightarrow b$, un coefficient multiplicateur maximal $kMax$, un pas d'évolution $pasK$, un support minimum min_{sup} et une confiance minimum min_{conf}.

Sortie : une courbe $\mathsf{C_{O1}}$ d'évolution des indices en fonction du coefficient multiplicateur k.

DEBUT

// **(0) Initialisation de la variable qui va récupérer les valeurs des différents indices pour chaque coefficient multiplicateur *k***

$TabEvol = \emptyset$;

// **(1) Calcul de l'ensemble *Ccs* des couples (*a,b*) vérifiant les contraintes**

$Ccs = \{(a,b) \in \mathsf{A} \times \mathsf{A} / p(a \wedge b) \geq min_{sup} \text{ et } p(b/a) \geq min_{conf}\}$;

// **(2) Recherche de la valeur des indices pour chaque coefficient multiplicateur**

Pour $k = 1 : pasK : kMax$ **faire**

Calcul des indices $VTeImpBarP$, $VTeImpCorP$ et $VTeImpProj$;

Pour tout $(a,b) \in Ccs$ **faire**

Calcul du coefficient normalisé $Q_3(a, \bar{b})$;

Fin Pour tout

Calcul de l'indice $VLgrImpP$;

TabEvol = [TabEvol ; k $VLgrImpP$ $VTeImpBarP$ $VTeImpCorP$ $VTeImpProj$] ;

Fin pour

// **(3) Tracé de la courbe**

$\mathsf{C_{O1}}$ = TracéCourbe(TabEvol) ;

Retourner $\mathsf{C_{O1}}$

FIN

Les paramètres utilisés pour réaliser l'ensemble des courbes des *sections 6.3.3* à *6.3.9* sont les suivants : $kMax = 1\,000$, $pasK = 20$, $min_{sup} = 0$ et $min_{conf} = 0$.

Après avoir défini ce premier modèle de croissance, nous allons maintenant décrire formellement le deuxième modèle [voir (105), sous-section 5.3.1].

Modèle de croissance M_2

Ce deuxième modèle de croissance M_2 consiste à augmenter la taille n de l'ensemble d'apprentissage sans changer les effectifs $n(a)$, $n(b)$ et $n(a \wedge b)$ des ensembles $O(a)$, $O(b)$ et $O(a \wedge b)$. Cela revient donc à rajouter des individus ne vérifiant ni a et ni b.

Soient les paramètres suivants :
- une règle $a \rightarrow b$,
- une taille maximale pour l'ensemble d'apprentissage *nMax* ($nMax \in \mathbb{N}^*$),
- un pas d'évolution *pasN* ($pasN \in \mathbb{N}^*$) pour la taille de l'ensemble d'apprentissage,
- un seuil minimum de support min_{sup} ($min_{sup} \in [0, 1]$) et
- un seuil minimum de confiance min_{conf} ($min_{sup} \in [0, 1]$).

Après avoir choisi tous ces paramètres, nous utilisons le protocole suivant :

1. Rechercher l'ensemble Ccs des couples d'attributs (a,b) vérifiant les contraintes de support ($p(a \wedge b) \geq min_{sup}$) et de confiance ($p(b/a) \geq min_{conf}$).
2. Pour chaque valeur n de l'ensemble d'apprentissage, calculer les quatre indices *VLgrImpP*, *VTeImpBarP*, *VTeImpCorP* et *VTeImpProj*. Pour le calcul de *VLgrImpP*, il faut calculer au préalable l'indice $Q_3(a, \bar{b})$ pour chaque couple (a,b) de Ccs.

3. Tracer la courbe d'évolution C_{O2} des différents indices en fonction de la taille de l'ensemble d'apprentissage.

L'algorithme permettant de réaliser les courbes pour le modèle M_2 est le suivant :

Algorithme 3

Entrée : une table de données $O \times A$, une règle $a \rightarrow b$, une taille maximale pour l'ensemble d'apprentissage *nMax*, un pas d'évolution *pasN*, un support minimum min_{sup} et une confiance minimum min_{conf}.

Sortie : une courbe C_{O2} d'évolution des indices en fonction de la taille de l'ensemble d'apprentissage.

DEBUT

// **(0) Initialisations**

$TabEvol = \emptyset$; // initialisation de la variable qui va récupérer les valeurs des différents indices pour chaque taille n de l'ensemble d'apprentissage

$n_{min} = |O|$; // calcul de la taille initiale de l'ensemble d'apprentissage

// **(1) Calcul de l'ensemble *Ccs* des couples (*a,b*) vérifiant les contraintes**

$Ccs = \{(a,b) \in A \times A / p(a \wedge b) \geq min_{sup}$ et $p(b/a) \geq min_{conf}\}$;

// **(2) Recherche de la valeur des indices pour chaque taille de l'ensemble d'apprentissage**

Pour $n = n_{min}$: *pasN* : *nMax* **faire**

Calcul des indices *VTeImpBarP*, *VTeImpCorP* et *VTeImpProj* ;

Pour tout $(a,b) \in Ccs$ **faire**

Calcul du coefficient normalisé $Q_3(a, \bar{b})$;

Fin Pour tout

Calcul de l'indice *VLgrImpP* ;

TabEvol = [TabEvol ; *n VLgrImpP VTeImpBarP VTeImpCorP VTeImpProj*] ;

Fin pour

// **(3) Tracé de la courbe**

C_{O2} = TracéCourbe(TabEvol) ;

Retourner C_{O2}

FIN

Les paramètres utilisés pour réaliser l'ensemble des courbes des *sections 6.3.3* à *6.3.9* sont : $nMax = 6\,000$, $pasN = 20$, $min_{sup} = 0$ et $min_{conf} = 0$.

Après avoir exposé le protocole expérimental utilisé pour les deux modèles de croissance, nous étudions le comportement des *4* indices dans certaines situations caractéristiques.

Nous avons retenu huit règles issues de la base *Wages*, règles couvrant l'ensemble des situations exposées dans la *section 6.2.1* (*incompatibilité, indépendance, indétermination, implication logique, répulsion et attraction*). Pour chacune de ces huit règles, nous donnons le comportement des quatre indices pour les modèles M_1 et M_2.

6.3.3 Incompatibilité

La première règle étudiée et qui est la suivante "*éducation5* → *ouvrier*" illustre le cas de l'incompatibilité comme le révèle la *figure 2* qui donne la contingence des attributs "*éducation5*" et "*ouvrier*".

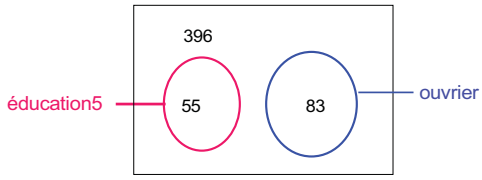

Fig. *2* : *Exemple de règle issue de Wages illustrant le cas de l'incompatibilité.*

L'évolution des quatre indices selon le modèle M_1 est donnée dans la *figure 3* et l'évolution selon le modèle M_2 peut être visualisée grâce à la *figure 4*.

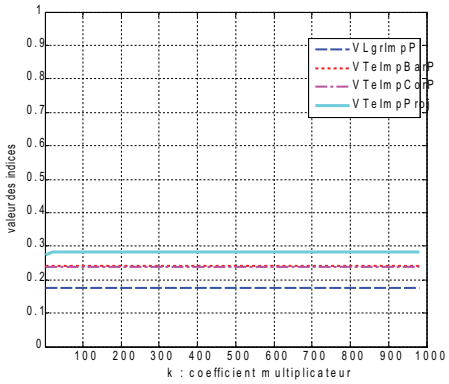

Fig. *3* : *Comparaison des indices selon le modèle M_1 pour la règle "éducation5 → ouvrier".*

Tout d'abord, nous vérifions une invariance des indices lorsque tous les effectifs sont multipliés par un coefficient multiplicateur *k* (*voir la figure 3*) comme cela a été démontré grâce aux (*Propositions 7 et 8*). Cependant, pour les très faibles valeurs de *k*, il y a une très légère croissance de la courbe *VTeImpProj*, croissance que nous retrouverons sur toutes les courbes suivantes, et cela pour toutes les situations étudiées. Nous observons un comportement similaire pour les indices *VTeImpBarP* et *VTeImpCorP* puisque les courbes sont quasiment confondues, ces deux mesures appartenant au groupe G_2 mis en évidence dans la première série d'expériences. Dans le cas de l'incompatibilité, c'est l'indice *VLgrImpP* qui est le plus sélectif (*valeurs les plus faibles*), ensuite les deux indices *VTeImpBarP* et *VTeImpCorP* du groupe G_2 et pour finir, l'indice *VTeImpProj*.

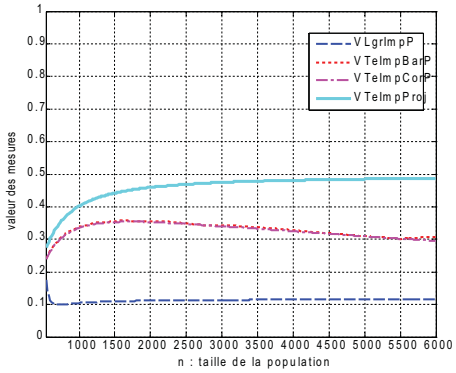

Fig. *4* : *Comparaison des indices selon le modèle M_2 pour la règle "éducation5 → ouvrier".*

Dans le cas de l'incompatibilité et pour le modèle M_2 (*voir la figure 4*), nous constatons un comportement similaire pour les indices *VTeImpBarP* et *VTeImpCorP* (*indices du groupe G_2*) puisque les courbes sont quasiment confondues comme pour le cas du modèle M_1 (*voir la figure 3)*. Ces courbes sont croissantes puis décroissantes comme cela a été démontré dans la sous-section 5.3.3. L'indice *VLgrImpP* décroît rapidement tout au début (*jusqu'à la valeur 734 pour la taille de l'ensemble d'apprentissage*) pour se stabiliser et tendre vers la valeur *0,1*. Quant à l'indice *VTeImpProj*, il croît pour tendre vers la valeur *0,5*.

6.3.4 Zone de répulsion

La deuxième règle "*syndiqué → féminin*" que l'on étudie (*la figure 5 permet de visualiser les différentes contingences*) correspond au cas où nous sommes dans la zone de répulsion c'est-à-dire que la réalisation de l'événement "*être syndiqué*" diminue les chances d'apparition de l'événement "*être de sexe féminin*". En effet, le nombre observé d'individus vérifiant à la fois la prémisse ("*être syndiqué*") et la conclusion ("*être de sexe féminin*") est de *28* alors que le nombre attendu dans l'hypothèse d'indépendance est de *44* (*c'est-à-dire 96 × 245 / 534*).

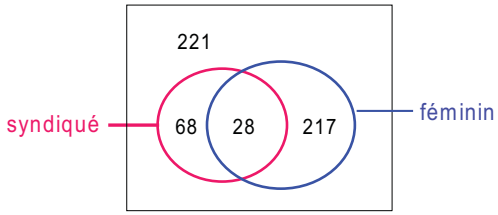

Fig. *5* : *Exemple de règle issue de Wages située dans la zone de répulsion.*

Les *figures 6* et *7* donnent respectivement l'évolution des quatre indices pour la règle "*syndiqué → féminin*" pour les modèles M_1 et M_2.

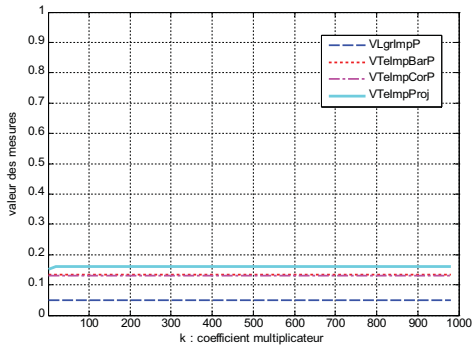

Fig. *6* : *Comparaison des indices selon le modèle M_1 pour la règle "syndiqué → féminin".*

Nous vérifions l'insensibilité des indices au coefficient multiplicateur *k*. De plus, nous observons, comme pour la *figure 3* (*règle "éducation5 → ouvrier", cas de l'incompatibilité, modèle* M_1), un comportement similaire pour les indices *VTeImpBarP* et *VTeImpCorP,* une plus grande sélectivité pour l'indice *VLgrImpP* et une moins grande sélectivité pour l'indice *VTeImpProj*.

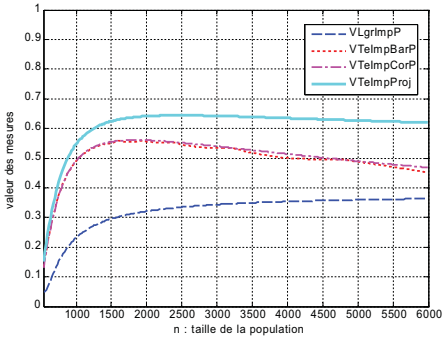

Fig. *7* : *Comparaison des indices selon le modèle* M_2 *pour la règle "syndiqué →féminin".*

On peut être surpris de constater que pour cette configuration de répulsion statistique entre *"syndiqué "* et "*féminin"* – mais où pourtant la conjonction *"syndiqué "* et "*féminin"* n'est pas vide – les valeurs des quatre indices sont inférieures au cas précédent de l'incompatibilité (*entre "éducation 5 " et* "*ouvrier"*). Cependant, l'incompatibilité porte sur les attributs beaucoup moins présents. Les produits des cardinaux dans ce dernier cas est 55×83=4565; alors que dans l'autre cas *"syndiqué →féminin"* il est de 96×245=23520. En se référant aux valeurs de l'indice $-Q_3(a,\bar{b})$, on obtient *-1,25* pour (*"éducation5 "→ "ouvrier"*) et *-2,23* pour (*"syndiqué "→ "féminin"*). Les implications précédentes sont en fait contre nature. Si on considérait les deux règles complémentaires (*"éducation5" → "non ouvrier"*) et (*"syndiqué "→ "non féminin"*), on obtient respectivement les valeurs : *2,92* et *2,42*.

Les courbes des indices *VTeImpProj*, *VTeImpBarP* et *VTeImpCorP* de la *figure 7* sont croissantes puis décroissantes contrairement à la courbe de l'indice *VLgrImpP* qui ne décroît pas mais tend vers une valeur proche de la valeur *0,4*. Nous observons toujours un comportement similaire pour les indices *VTeImpBarP* et *VTeImpCorP* et une même sélectivité pour les *4* indices (*c'est-à-dire de plus faibles valeurs pour VLgrImpP et de plus fortes valeurs pour VTeImpProj)*. Hormis les premières valeurs des indices correspondant à une taille de l'ensemble d'apprentissage faible, les valeurs des quatre indices sont plus élevées que dans le cas précédent, cas de l'incompatibilité. Cependant ici, le modèle d'évolution est bien différent. L'espérance du nombre de contre exemples dans l'hypothèse d'indépendance est croissante. Reprenons à cet égard le coefficient de la forme $-Q_{3x}(a,\bar{b})$ [cf. (110)] dont la réduction globale conduit à *VLgrImpP*($a \rightarrow \bar{b}$). Désignons par $\phi(x)$ [resp. $\psi(x)$] ce coefficient dans le cas de l'évaluation de (*"éducation5 "→ "ouvrier"*) [resp. (*"syndiqué "→ "féminin"*)], on peut vérifier en termes de dérivées que $\psi'(x)-\phi'(x)$ est positif. On comprendra dans ce cas qu'après l'opération de réduction globale, la courbe d'évolution associée à *VlgrImpP* (*"syndiqué "→ "féminin"*) (Fig. 7) devienne sensiblement plus élevée que celle associée à *VLgrImpP*(*"éducation5 "→ "ouvrier"*).

6.3.5 Indépendance

La troisième règle que nous étudions maintenant correspond au cas de l'indépendance. C'est la règle "*salaire3 → nord* " où nous pouvons visualiser la contingence des effectifs grâce à la *figure 8*.

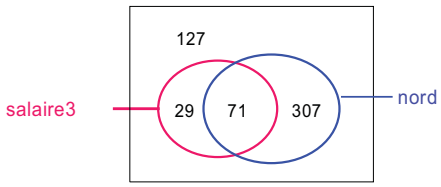

Fig. *8* : *Exemple de règle issue de Wages illustrant le cas de l'indépendance.*

Nous sommes dans le cas de l'indépendance car le nombre observé d'exemples (*c'est-à-dire 71*) est égal au nombre attendu d'exemples (*c'est-à-dire 100 × 378 / 534*).

Les *figures 9* et *10* donnent respectivement l'évolution des quatre indices pour la règle "*salaire3* → *nord*" pour les modèles M_1 et M_2.

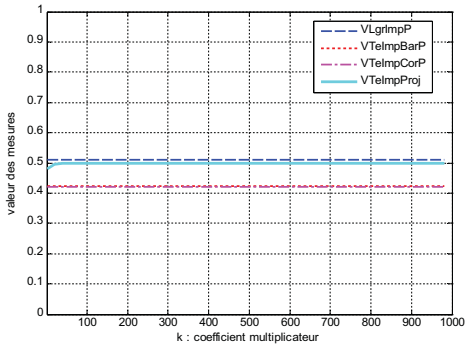

Fig. *9* : *Comparaison des indices selon le modèle M_1 pour la règle "salaire3 → nord"*.

Nous constatons toujours une invariance face au coefficient multiplicateur *k* et les valeurs des différents indices sont plus élevées que dans le cas de l'incompatibilité (*figure 3*) et de la répulsion (*figure 6*), ce qui est rassurant puisque cette règle est plus favorable que les deux cas précédents. Nous avons toujours une similitude de comportement pour les indices *VTeImpBarP* et *VTeImpCorP* mais la nouveauté est que l'indice *VLgrImpP* a les plus fortes égales valeurs alors que précédemment il avait les plus faibles valeurs. Toutefois, cette forte valeur commune est de *0,5* , ce qui semble traduire correctement l'indépendance.

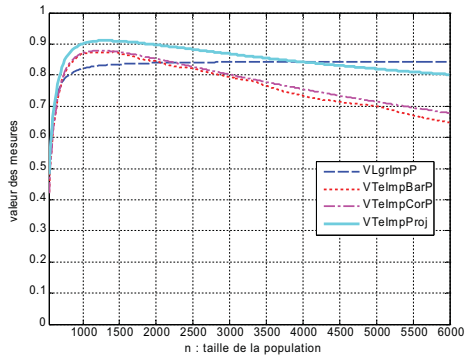

Fig. *10* : *Comparaison des indices selon le modèle M_2 pour la règle "salaire3 → nord"*.

Pour le modèle de croissance M_2, nous constatons toujours une croissance puis une décroissance pour les indices *VTeImpProj*, *VTeImpBarP* et *VTeImpCorP* contrairement à l'indice *VLgrImpP* qui ne décroît pas mais tend vers une valeur proche de la valeur *0,85*. Ce qui change par rapport aux courbes précédentes (*figures 4 et 7*), c'est que pour les fortes valeurs de *n* (*valeurs supérieures à 4 000*), les valeurs de l'indice *VLgrImpP* sont les valeurs les plus élevées, situation également observée pour le modèle de croissance M_1 de cette même règle.

6.3.6 Attraction proche de l'indépendance

La quatrième règle "*éducation3* → *non syndiqué*" est un cas proche de l'indépendance mais la règle se trouve dans la zone d'attraction puisque le nombre observé d'exemples (*c'est-à-dire 78*) est légèrement supérieur au nombre attendu (*c'est-à-dire* $76 = 93 \times 438 / 534$). La *figure 11* nous indique les différents effectifs de cette règle.

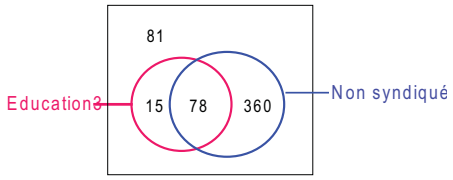

Fig. *11* : *Exemple de règle issue de Wages illustrant un cas proche de l'indépendance mais dans la zone d'attraction.*

Les *figures 12* et *13* donnent respectivement l'évolution des quatre indices pour la règle "*éducation3 → non syndiqué*" pour les modèles M_1 et M_2.

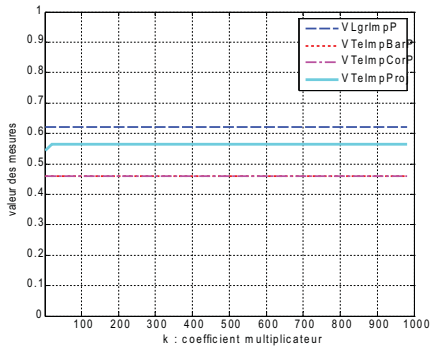

Fig. *12* : *Comparaison des indices selon le modèle M_1 pour la règle "éducation3 → non syndiqué".*

Même constat que pour les figures précédentes (*figures 3, 6 et 9*) : les courbes sont invariantes au coefficient multiplicateur *k*. Nous observons, comme pour le cas précédent de l'indépendance, que la courbe de *VLgrImpP* est celle qui a les plus fortes valeurs et non plus celle ayant les plus faibles valeurs (*cas d'incompatibilité et de répulsion*).

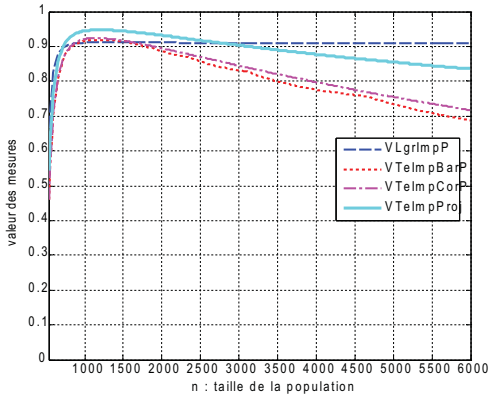

Fig. *13* : *Comparaison des indices selon le modèle M_2 pour la règle "éducation3 → non syndiqué".*

Nous obtenons des courbes similaires au cas précédent de l'indépendance avec la particularité que les valeurs de *VLgrImpP* deviennent plus élevées pour une valeur plus faible de la taille de l'ensemble d'apprentissage (*2 750 environ au lieu de 4 000*).

La cinquième règle "*construction → blanc*" étudiée correspond également à un cas où celle-ci est dans la zone d'attraction et très proche de l'indépendance puisque le nombre d'exemples observé est de *21* alors que le nombre attendu dans l'hypothèse d'indépendance est de *19* ($24 \times 440 / 534 \approx 20$). Nous avons une différence entre le nombre d'exemples observé et le nombre d'exemples attendu de *2*, comme pour le cas précédent, c'est-à-dire la règle "*éducation3 → non syndiqué*". La *figure 14* restitue la contingence des attributs "*construction*" et "*blanc*". Ce cas diffère de la situation précédente essentiellement par le support de la prémisse (*17,4% pour éducation3 contre 4,5% pour construction*) et la confiance (*83,9% pour la règle "éducation3 → non syndiqué" contre 87,5% pour la règle "construction → blanc"*) car le support de la conclusion est quasiment identique (*82% pour "non syndiqué" contre 82,4% pour "blanc"*). Nous

étudions un deuxième exemple concernant cette situation car nous obtenons des résultats quelque peu différents de ceux qui précèdent, à savoir la règle "*éducation3* → *non syndiqué*", et ceci pour le modèle de croissance M_2.

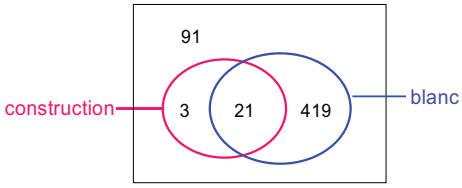

Fig. *14* : *Exemple de règle issue de Wages illustrant le cas d'une règle située dans la zone d'attraction mais proche de l'indépendance.*

Les *figures 15* et *16* donnent respectivement l'évolution des quatre indices pour la règle "*construction* → *blanc*" pour les modèles M_1 et M_2.

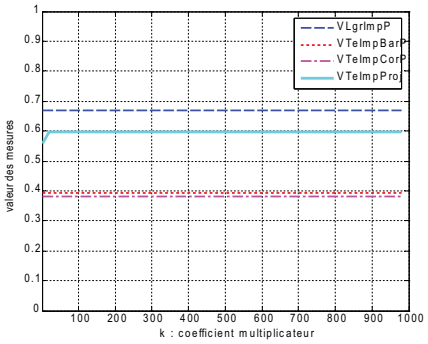

Fig. *15* : *Comparaison des indices selon le modèle M_1 pour la règle "construction → blanc".*

Pour le modèle de croissance M_1, nous obtenons des courbes similaires au cas précédent "*éducation3* → *non syndiqué*" avec la particularité que les valeurs des indices du groupe de mesures G_2 sont plus faibles.

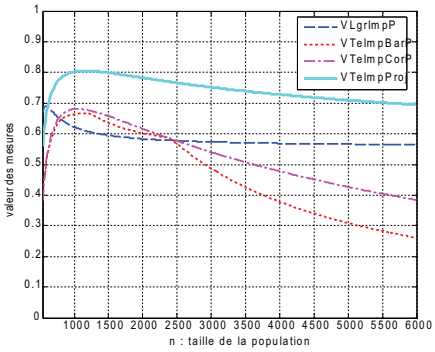

Fig. *16* : *Comparaison des indices selon le modèle M_2 pour la règle "construction → blanc".*

C'est la première figure où nous voyons une divergence sensible entre les indices *VTeImpBarP* et *VTeImpCorP* mais cependant ils évoluent de la même façon. Ce qui diffère des cas précédents étudiés est que la courbe de l'indice *VLgrImpP* décroît pour tendre vers une valeur proche de *0,67*. Ces courbes de la *figure 16* peuvent paraître surprenantes car nous avons des valeurs inférieures aux valeurs du cas précédent alors qu'à première vue, intuitivement, la présente configuration paraît plus implicative que la précédente (*Fig. 11*). Certes, la confiance est du même ordre dans les deux cas ; *0,875* ici et *0,84* pour le cas précédent. Dans les deux cas la taille de la conclusion est du même ordre ; *440* ici et *438* dans le cas précédent. Cependant, ce qui intervient est la taille de la prémisse ; *24* ici, alors que *93* pour le cas précédent. Ainsi, une implication engageant une prémisse de taille importante se trouve beaucoup plus ponctuée qu'une implication engageant une prémisse de faible taille.

6.3.7 Indétermination

La sixième règle "*éducation2* → *féminin*" correspond au cas de l'indétermination puisque nous avons quasiment autant d'exemples que de contre-exemples comme le montre la *figure 17* (*109 contre-exemples pour 110 exemples*).

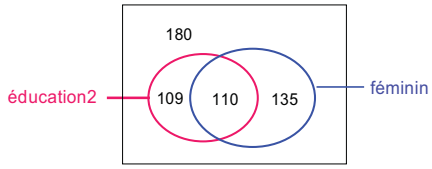

Fig. *17* : *Exemple de règle issue de Wages illustrant le cas de l'indétermination.*

Les *figures 18* et *19* donnent respectivement l'évolution des quatre indices pour la règle "*éducation2* → *féminin*" pour les modèles M_1 et M_2.

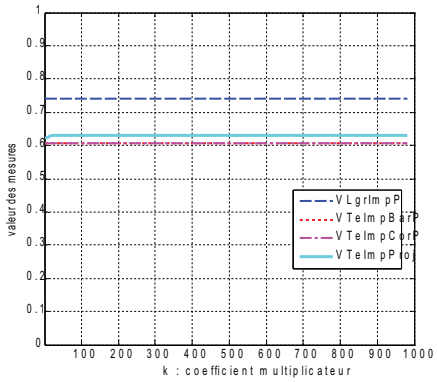

Fig. *18* : *Comparaison des indices selon le modèle M_1 pour la règle "éducation2 → féminin".*

Pour le modèle M_1, nous obtenons le même type de courbe que celles des *figures 9*, *12* et *15*.

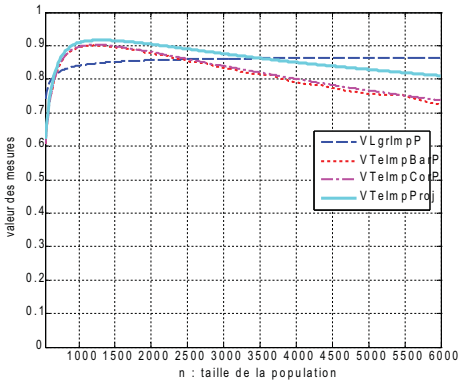

Fig. *19* : *Comparaison des indices selon le modèle M_2 pour la règle "éducation2 → féminin".*

Les indices ont le même comportement que pour les règles "*éducation3* → *non syndiqué*" (*attraction proche de l'indépendance*) et "*salaire3* → *nord*" (*cas de l'indépendance*).

6.3.8 Attraction

La septième règle "*féminin* → *non syndiqué*" est une règle située dans la zone d'attraction puisque le nombre observé d'exemples est égal à *217* et le nombre attendu est de *201* (*c'est-à-dire* $245 \times 438 / 534$). La *figure 20* restitue la contingence des attributs "*féminin*" et "*non syndiqué*".

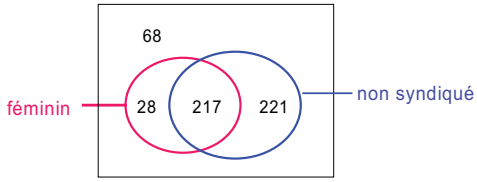

Fig. *20* : *Exemple de règle issue de Wages située dans la zone d'attraction.*

Les *figures 21* et *22* donnent respectivement l'évolution des quatre indices pour la règle "*féminin → non syndiqué*" pour les modèles M_1 et M_2.

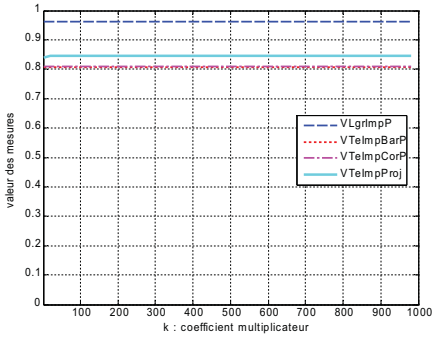

Fig. *21* : *Comparaison des indices selon le modèle M_1 pour la règle "féminin → non syndiqué"*.

Pour le modèle M_1, nous obtenons le même type de courbe que celles des *figures 9*, *12*, *15* et *18*.

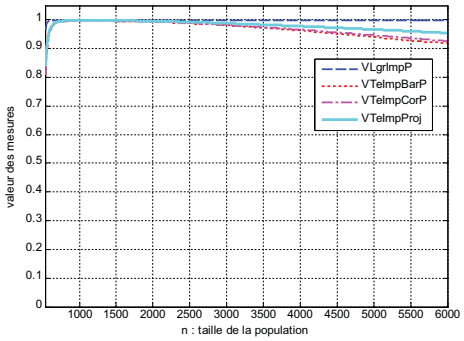

Fig. *22* : *Comparaison des indices selon le modèle M_2 pour la règle "féminin → non syndiqué"*.

Nous constatons des allures identiques aux cas des règles "*salaire3 → nord*" (*indépendance*), "*éducation3 → non syndiqué*" (*attraction proche indépendance*) et "*éducation2 → féminin*" (*indétermination*) avec comme particularité que l'indice *VLgrImpP* tend très rapidement vers la valeur *1*.

6.3.9 Implication

La dernière règle "*vendeur → non syndiqué*" correspond au cas de la quasi-implication puisqu'il n'y a qu'un contre-exemple comme l'illustre la *figure 23*. Le nombre observé est de *37* alors que le nombre attendu est égal à *31* (*c'est-à-dire* $38 \times 438 / 534$). Nous n'avons trouvé aucune règle dans la base *Wages* où l'ensemble des individus vérifiant la prémisse est inclus dans l'ensemble des individus vérifiant la conclusion.

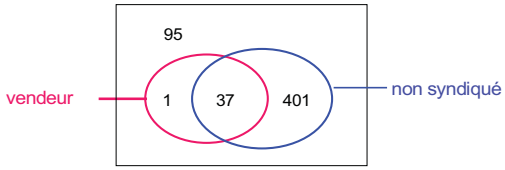

Fig. *23* : *Exemple de règle issue de Wages correspondant au cas de la quasi implication logique.*

Les *figures 24* et *25* donnent respectivement l'évolution des quatre indices pour la règle "*vendeur → non syndiqué*" pour les modèles M_1 et M_2.

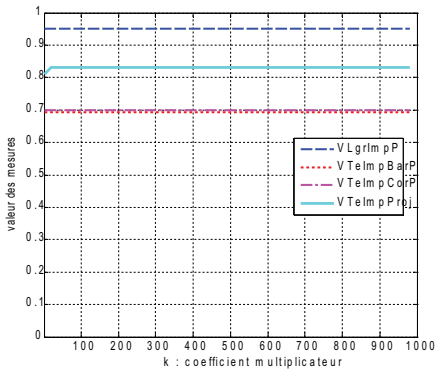

Fig. *24* : *Comparaison des indices selon le modèle M_1 pour la règle "vendeur → non syndiqué".*

Pour le modèle M_1, nous obtenons le même type de courbe que celles des *figures 9*, *12*, *15, 18* et *21*.

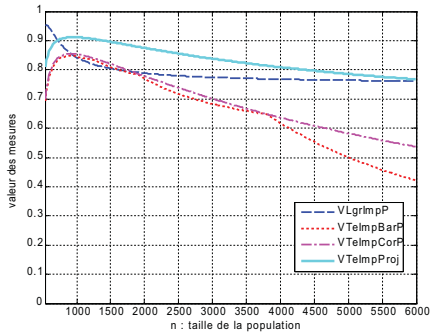

Fig. *25* : *Comparaison des indices selon le modèle M_2 pour la règle "vendeur → non syndiqué".*

Pour le modèle M_2, nous obtenons le même type de courbe que celle de la *figure 16*. Le phénomène qui a joué est – encore une fois – la faiblesse de la taille de la prémisse. Cependant, compte tenu d'une plus forte dépendance observée, les courbes ici (*Fig. 24 et Fig. 25*) sont plus élevées que celles correspondantes dans les *figures 15* et *16*.

6.4 Tendance comportementale de la fonction *VLgrImpP*

On peut remarquer que sous le modèle de croissance M_2, la fonction *VLgrImpP* est parmi les quatre fonctions *VLgrImpP*, *VTeImpBarP*, *VTeImpCorP* et *VTeImpProj,* la seule à avoir un comportement "*harmonieux*". Nous voulons dire qu'elle est globalement, soit continuellement croissante, soit continuellement décroissante ; alors que les trois autres fonctions ont un régime de variation non stable. Elles commencent par être croissantes, puis leurs allures changent et deviennent décroissantes. Cette instabilité du comportement est davantage marquée pour les indices du groupe G_2, (*VTeImpCorP et VTeImpBarP*). En effet, le comportement de *VTeImpProj* moins accidentée que pour les indices du groupe G_2 ; la décroissance de cette fonction reste modérée après son intervalle de croissance.

Le tableau suivant résume les caractéristiques des huit différentes situations étudiées relativement à la variation de la fonction *VLgrImpP*. Les cas d'abord de croissance puis de décroissance des indices

VTeImpCorP, VTeImpBarP et *VTeImpProj* se produisent de façon plus accentuée lorsque *VLgrImpP* est décroissant (*voir Fig. 16 et Fig. 25*). Pour chacune des situations ou règles ($a \rightarrow b$) on précise par le support $sup(a)$ de la prémisse, le support $sup(b)$ de la conclusion, le support $sup(a \rightarrow b)$ de la règle, la confiance $conf(a \rightarrow b)$ de la règle, le nombre *nbreObs* observé d'exemples, le nombre *nbreAtt* attendu d'exemples, la différence Δ entre le nombre observé d'exemples et le nombre attendu d'exemples et pour finir le type de courbe (*croissante ou décroissante*). Pour tous les supports du *tableau 6.9* nous indiquons le nombre d'individus ainsi que le pourcentage que ce nombre représente dans l'ensemble d'apprentissage.

Nous observons que les cas de décroissance interviennent lorsque le support $sup(a)$ de la prémisse est faible. Nous allons vérifier si c'est toujours le cas avec d'autres règles de la base *Wages*. Pour cela, nous avons recherché les règles ayant un support pour la prémisse inférieur à *10%* et une confiance supérieure à *80%*. Nous en avons trouvé deux : "*hispanique* → *autre secteur*" et "*vendeur* → *autre secteur*". Les *figures 26* et *28* donnent les contingences de ces deux règles et les *figures 27* et *29* les courbes d'évolution des indices.

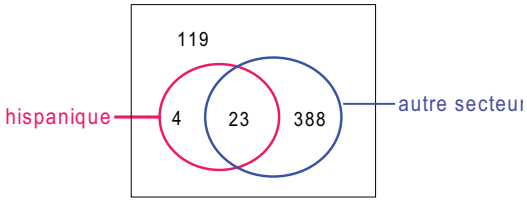

Fig. *26* : *Contingence de la règle "hispanique → autre secteur"*.

Le support $sup(a)$ de la règle "*hispanique* → *autre secteur*" est égal à *5%*, le support $sup(b)$ est égal à *77%* et la confiance est égale à *85%*.

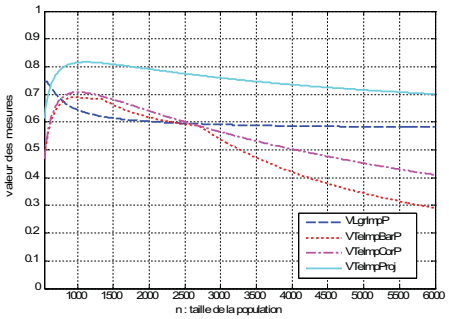

Fig. *27* : *Comparaison des indices selon le modèle M_2 pour la règle "hispanique → autre secteur"*.

Les deux figures suivantes étudient la règle "*vendeur* → *autre secteur*".

Règle	*sup(a)*	*sup(b)*	*sup(a→b)*	*conf(a→b)*	*nbreObs*	*nbreAtt*	Δ	Courbe
edu5 → ouv	*55 10,3%*	*83 15,5%*	*0 0%*	*0*	*0*	*8*	*-8*	décroissante
synd → fem	*96 18%*	*245 45,9%*	*28 5,2%*	*29,2%*	*28*	*44*	*-16*	croissante
sal3 → nord	*100 18,7%*	*378 70,8%*	*71 13,3%*	*71%*	*71*	*71*	*0*	croissante
edu3 → nonSynd	*93 17,4%*	*438 82%*	*78 14,6%*	*83,9%*	*78*	*76*	*2*	croissante
const → blanc	*24 4,5%*	*440 82,4%*	*21 3,9%*	*87,5%*	*21*	*19*	*2*	décroissante
edu2 → fem	*219 41%*	*245 45,9%*	*110 20,6%*	*50,2%*	*110*	*100*	*10*	croissante
fem → nonSynd	*245 45,9%*	*438 82%*	*217 40,6%*	*88,6%*	*217*	*201*	*16*	croissante
vend → nonSynd	*38 7,1%*	*438 82%*	*37 6,9%*	*97,4%*	*37*	*31*	*8*	décroissante

TAB. 6.9 : *Caractérisation des situations étudiées.*

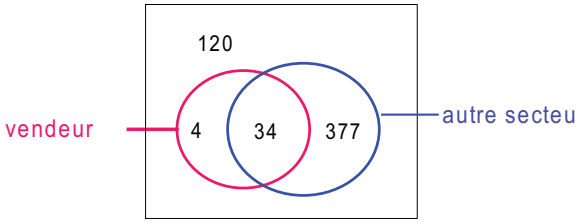

Fig. *28* : *Contingence de la règle "vendeur → autre secteur"*.

Le support *sup*(*a*) de la règle "*vendeur → autre secteur*" est égal à *7%*, le support *sup*(*b*) est égal à *77%* et la confiance est égale à *89%*.

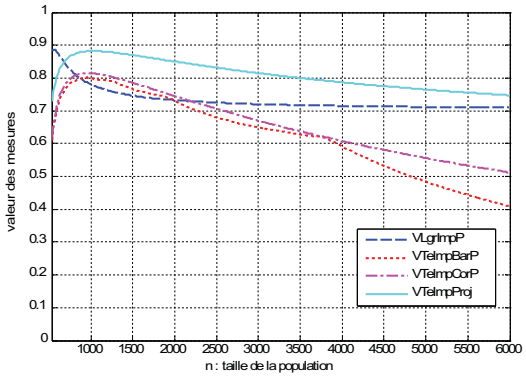

Fig. *29* : *Comparaison des indices selon le modèle M_2 pour la règle "vendeur → autre secteur"*.

Nous vérifions une décroissance pour la courbe de l'indice *VLgrImpP* dans le cas de la *figure 30.* Il en est d'ailleurs de même dans le cas de la *figure 32*. Cependant dans ce dernier cas la courbe est plus élevée ; ceci pouvant être dû à la consistance des effectifs en jeu.

Maintenant, nous souhaitons vérifier, quand le support de la prémisse est faible, si la taille de la conclusion peut avoir une influence et modifier la décroissance de la courbe. Nous avons donc sélectionné

une règle "*expérience3* → *âge3*" dont le support *sup*(*a*) de la prémisse est égal à *20%*, le support *sup*(*b*) de la conclusion est égal à *23%* et la confiance est égale à *72,6%*. La *figure 30* donne la contingence de la règle "*expérience3* → *âge3*" et la *figure 31* donne l'évolution des indices pour le modèle de croissance M_2.

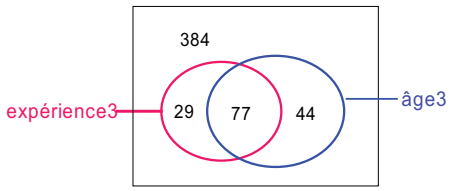

Fig. *30* : *Contingence de la règle "expérience3 → âge3".*

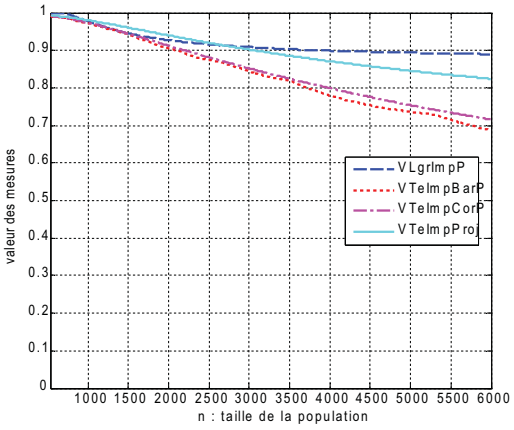

Fig. *31* : *Comparaison des indices selon le modèle M_2 pour la règle "expérience3 → âge3".*

6.5 Comparaison des huit différentes configurations pour un même indice

Nous regroupons maintenant sur un seul graphique toutes ces situations caractéristiques, et cela pour chaque mesure et pour chaque modèle de croissance afin d'avoir une vision globale du comportement de chaque indice. Les *figures 32* à *34* nous restituent les différentes courbes pour respectivement les indices *VLgrImpP*, *VTeImpBarP*, *VTeImpCorP* et *VTeImpProj* pour le modèle de croissance M_1 et les *figures 33* à *39* nous restituent les différentes courbes pour le modèle de croissance M_2. Afin de différencier chacune de ces courbes grâce à des petits symboles, nous avons du changer le pas d'évolution *pasK* et *pasN* et retenir la valeur *150* au lieu de *20*, ce qui explique les segments de droite présents pour les faibles valeurs de l'axe des abscisses (*surtout pour les courbes d'évolution pour le modèle de croissance M_2*).

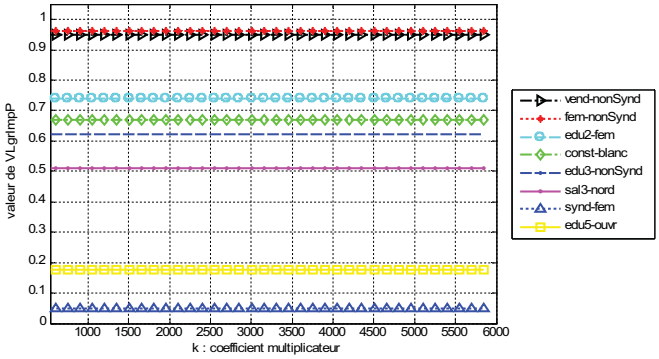

Fig. *32* : *Évolution de l'indice VLgrImpP pour les différentes situations et pour le modèle M_1.*

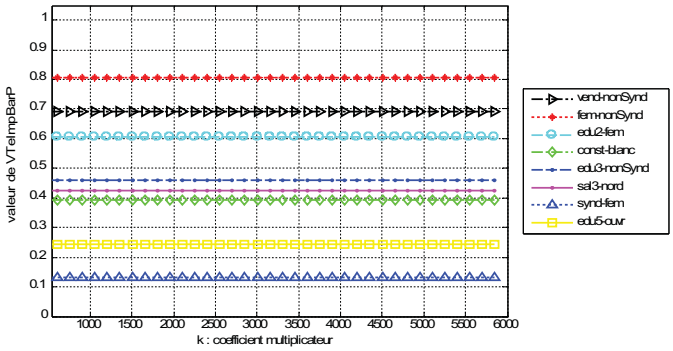

Fig. *33* : *Évolution de l'indice VTeImpBarP pour les différentes situations et pour le modèle M_1.*

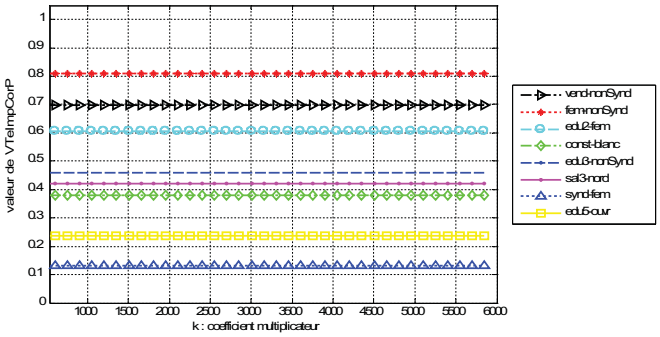

Fig. *34* : *Évolution de l'indice VTeImpCorP pour les différentes situations et pour le modèle M_1.*

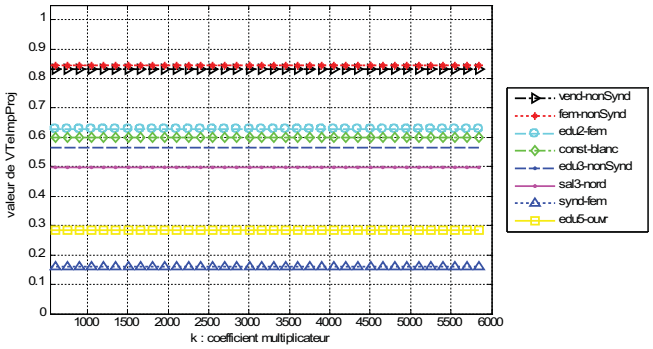

Fig. *35* : *Évolution de l'indice VTeImpProj pour les différentes situations et pour le modèle M_1.*

Après avoir étudié les courbes de croissance selon le modèle M_1, nous passons aux courbes selon le modèle M_2.

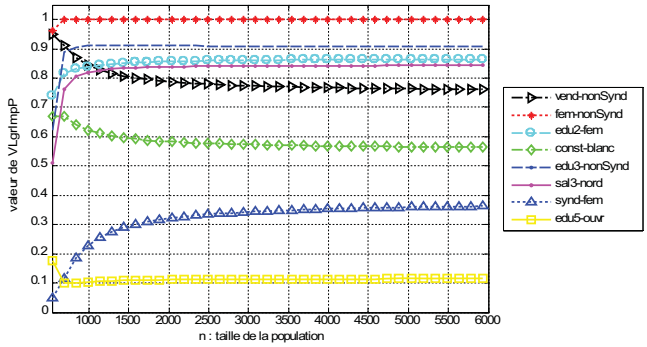

Fig. 36 : Évolution de l'indice VLgrImpP pour les différentes situations et pour le modèle M_2.

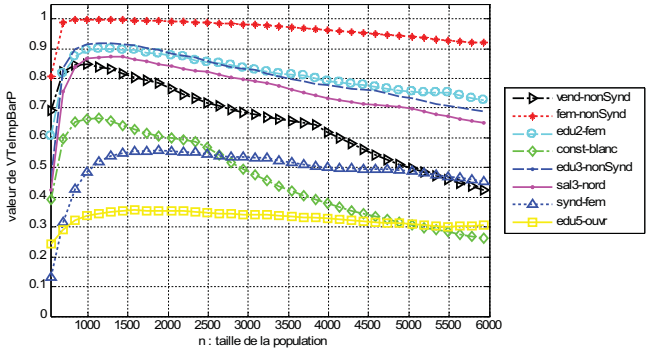

Fig. 37 : Évolution de l'indice VTeImpBarP pour les différentes situations et pour le modèle M_2.

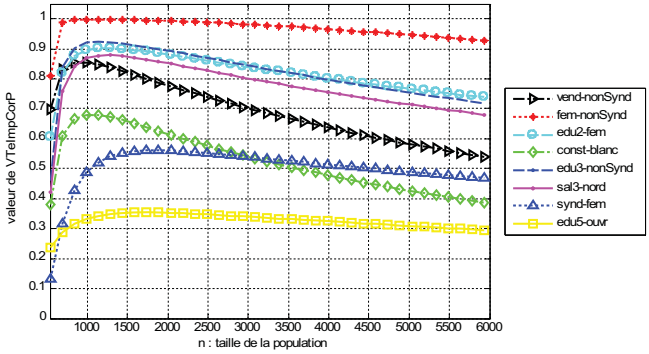

Fig. 38 : Évolution de l'indice VTeImpCorP pour les différentes situations et pour le modèle M_2.

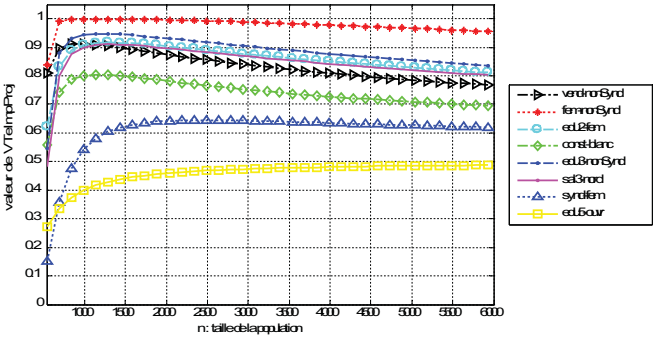

Fig. 39 : Évolution de l'indice VTeImpProj pour les différentes situations et pour le modèle M_2.

Comme établi ci-dessus (*Voir Propositions 7 et 8 sous-section 5.3.2*) les courbes sont représentées par des horizontales (*parallèles à l'axe des abscisses*). On constate que sur l'ensemble des huit règles ci-dessus, les deux ordres induits par les indices du groupe G_1, *VLgrImpP* et *VTeImpProj*, sont identiques. De même, les deux ordres induits par les indices du groupe G_2, *VTeImpCorP* et *VTeImpCorP*, sont identiques. Les règles *3*, *4* et *5* pour les indices du groupe G_1, deviennent respectivement de rangs *4*, *5* et *3* ; le reste demeurant invariant en termes de rangs.

Maintenant, nous allons effectuer le même travail que précédemment mais avec les indices *VLgrImpP*, *IIEG*, *IIEL* et *IIEB*. Tout d'abord, nous commençons par effectuer la comparaison des *20* meilleures règles extraites par ces différents indices.

6.6. Comparaison des *20* meilleures règles

L'objectif de cette expérience va être, comme dans le cas de la comparaison précédente (*voir section 6.2*), de dégager des comportements communs entre ces indices.

Le protocole expérimental est le même que celui utilisé dans la *section 6.2* sauf qu'au lieu de calculer la valeur des indices *VTeImpBarP*, *VTeImpCorP* et *VTeImpProj*, nous calculons la valeur des indices *IIEG*, *IIEL* et *IIEB*. De la même façon, l'extraction a été effectuée avec les mêmes contraintes : un support minimum min_{sup} de *0,01* et une confiance minimum min_{conf} de *0,80*.

L'ensemble des *20* meilleures règles est restitué dans le *tableau 6.7* et contrairement à l'étude précédente, nous n'avons pas le même ensemble pour chacun de ces indices. Pour chacune des règles $a \rightarrow b$ extraites, le *tableau 6.7* donne la prémisse a, la conclusion b, le support $p(a \wedge b)$, la confiance $p(b/a)$ et les valeurs des quatre indices étudiés. Afin de comparer d'une part les meilleures règles extraites par chacun des indices et d'autre part, de mettre en évidence des ensembles différents, nous transformons le *tableau 6.7* en remplaçant les valeurs de chacun des indices par le rang de la règle comme nous l'avons fait pour le *tableau 6.5*. Le *tableau 6.4* restitue ces différents rangs pour chacun des indices. Nous avons indiqué tous les rangs de toutes les règles même lorsque ceux-ci étaient supérieurs à *20*. Afin de repérer ces rangs supérieurs à *20*, nous les avons notés en italique et entre parenthèses.

Règles		Mesures des RA[2]		Indices étudiés			
Prémisse	**Conclusion**	**support**	**confiance**	***VLgrImpP***	***IIEG***	***IIEL***	***IIEB***
âge5	expérience5	0,19	0,94	1	0,98	0,98	0,99
âge1	expérience1	0,15	0,88	1	0,96	0,96	0,98
autre profession	masculin	0,24	0,81	0,99	0,88	0,88	0,94
ouvrier	autre secteur	0,15	0,98	0,95	0,99	0,99	1
salaire1	non syndiqué	0,19	0,95	0,89	0,98	0,98	0,99
construction	autre profession	0,04	0,83	0,88	0,93	0,93	0,97
employé	autre secteur	0,16	0,91	0,83	0,95	0,95	0,98
construction	masculin	0,04	0,92	0,81	0,97	0,98	0,99
âge5	marié	0,16	0,81	0,80	0,88	0,88	0,94
féminin	non syndiqué	0,41	0,89	0,76	0,80	0,81	0,91
employé	non syndiqué	0,17	0,92	0,72	0,95	0,95	0,98
vendeur	non syndiqué	0,07	0,97	0,72	0,99	1	1
cadre	non syndiqué	0,10	0,95	0,71	0,98	0,98	0,99
salaire2	non syndiqué	0,18	0,91	0,69	0,93	0,94	0,97
profession libérale	autre secteur	0,17	0,87	0,68	0,91	0,92	0,96
féminin	autre secteur	0,38	0,83	0,68	0,73	0,74	0,88
cadre	autre secteur	0,09	0,89	0,64	0,94	0,96	0,98
éducation5	autre secteur	0,09	0,89	0,64	0,94	0,96	0,98
âge1	non syndiqué	0,15	0,90	0,62	0,93	0,94	0,97
expérience1	non syndiqué	0,17	0,89	0,61	0,91	0,93	0,97
vendeur	autre secteur	0,06	0,89	0,57	0,94	0,96	0,98
éducation5	blanc	0,09	0,89	0,46	0,89	0,95	0,98
profession libérale	blanc	0,17	0,89	0,55	0,89	0,92	0,96
vendeur	blanc	0,06	0,89	0,43	0,88	0,96	0,98
construction	blanc	0,04	0,88	0,32	0,81	0,95	0,98
hispanique	autre secteur	0,04	0,85	0,39	0,85	0,94	0,97

TAB. *6.10 : Ensemble des meilleures règles extraites sur Wages.*

[2]RA : règles d'association.

Règles		Indices étudiés			
Prémisse	**Conclusion**	***VLgrImpP***	***IIEG***	***IIEL***	***IIEB***
âge5	expérience5	1	3	4	4
âge1	expérience1	2	7	9	9
autre profession	masculin	3	*(22)*	*(35)*	*(35)*
ouvrier	autre secteur	4	1	2	2
salaire1	non syndiqué	5	4	5	5
construction	autre profession	6	14	19	19
employé	autre secteur	7	8	13	13
construction	masculin	8	6	6	6
âge5	marié	9	*(21)*	*(33)*	*(33)*
féminin	non syndiqué	10	*(34)*	*(58)*	*(58)*
employé	non syndiqué	11	9	12	12
vendeur	non syndiqué	12	2	1	1
cadre	non syndiqué	13	5	3	3
salaire2	non syndiqué	14	13	17	17
profession libérale	autre secteur	15	17	*(22)*	*(22)*
féminin	autre secteur	16	*(43)*	*(60)*	*(60)*
cadre	autre secteur	17	10	10	10
éducation5	autre secteur	18	11	11	11
âge1	non syndiqué	19	15	16	16
expérience1	non syndiqué	20	16	20	20
vendeur	autre secteur	*(21)*	12	7	7
éducation5	blanc	*(29)*	18	15	15
profession libérale	blanc	*(24)*	19	*(21)*	*(21)*
vendeur	blanc	*(34)*	20	8	8
construction	blanc	*(46)*	*(31)*	14	14
hispanique	autre secteur	*(36)*	*(26)*	18	18

Tab. 6.11 : Rang des 20 meilleures règles extraites sur Wages.

Nous constatons que les indices *IIEL* et *IIEB* classent de la même façon les règles. Nous avons étendu cette étude et regardé le classement sur l'ensemble des *69* règles extraites et seul le classement de deux règles ont été inversé : les règles "*salaire5* → *blanc*" et "*hispanique* → *non syndiqué*" arrivant aux rangs *23* et *24* suivant l'indice considéré.

Comme ces deux indices classent les règles quasiment de la même façon, nous allons restreindre notre étude aux trois premiers indices : *VLgrImpP*, *IIEG* et *IIEL*.

Tout d'abord, nous allons comparer le comportement de l'indice *VLgrImpP* avec l'indice *IIEG*. La *figure 40* restitue le classement de l'indice *IIEG* par rapport à celui de *VLgrImpP* puisque le classement de *VLgrImpP* est matérialisé par la première bissectrice. Nous constatons une assez grande divergence dans le classement pour ces deux mesures en raison de la présence de nombreux pics, comme par exemple la règle classée en $3^{ème}$ position par *VLgrImpP* est en $22^{ème}$ position pour *IIEG*, ce qui fait un écart de classement égal à *19*. Les plus grands écarts de classement interviennent pour les règles classées par *VLgrImpP* aux rangs *3*, *9*, *10* et *16* avec des écarts de classement de respectivement *19*, *12*, *24* et *27*.

Nous allons effectuer ce même travail de comparaison mais avec les indices *VLgrImpP* et *IIEL*. La *figure 41* nous révèle le classement des règles obtenu grâce à *IIEL* par rapport à l'indice *VLgrImpP*

puisque, comme pour la *figure 40*, le classement de *VLgrImpP* est matérialisé par la première bissectrice. Nous constatons également une grande divergence dans ce classement puisque nous avons de nombreux pics, pics qui sont localisés au même endroit que précédemment, à savoir les rangs *3*, *9*, *10* et *16* mais avec des écarts plus importants puisque ceux-ci sont de respectivement *32*, *24*, *48* et *44*.

Nous constatons que l'allure générale des courbes des *figures 40* et *41* est quasiment identique puisque nous avons des pics et des creux aux mêmes endroits.

Comparons pour finir le classement des indices *IIEG* et *IIEL*, classement visible sur la *figure 42*. Cette fois-ci, la première bissectrice correspond au classement de l'indice *IIEG* et par conséquent, nous évaluons le classement de *IIEL* par rapport à *IIEG*. Nous constatons également une divergence de classement mais avec des écarts plus atténués puisque l'écart maximum est de *12* et intervient pour la dernière règle (*classée en 20eme position par IIEG*), l'écart immédiatement inférieur étant de *5*.

En conclusion, la plus grande divergence de classement s'opère entre l'indice *VLgrImpP* et les trois autres indices (*IIEG, IIEL et IIEB*) et nous avons un classement quasiment identique effectué par les indices *IIEL* et *IIEB*.

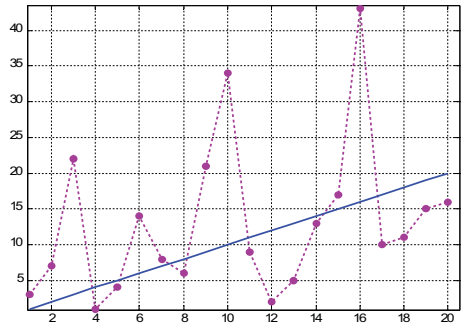

Fig. 40 : Classement des règles pour les indices VLgrImpP et IIEG.

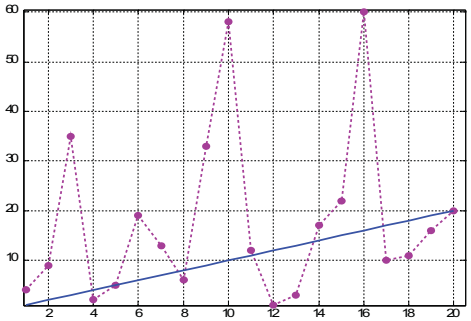

Fig. 41 : Classement des règles pour les indices VLgrImpP et IIEL.

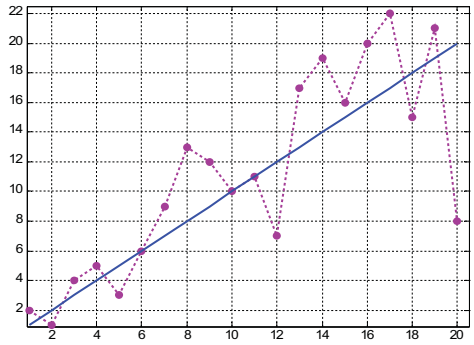

Fig. 42 : Classement des règles pour les indices IIEG et IIEL.

Après avoir réalisé cette étude comparative des classements, nous allons étudier le comportement de ces indices dans différentes situations caractéristiques, ce qui va constituer la deuxième série d'expériences.

6.7. Étude de certaines situations caractéristiques

Pour cette étude, nous reprenons les mêmes états caractéristiques que ceux qui ont été étudiés dans la *section 6.3* et selon les modèles de croissance M_1 et M_2.

6.7.1 Incompatibilité

La première règle étudiée "*éducation5* → *ouvrier*" illustre le cas de l'incompatibilité et nous rappelons la contingence de ces attributs grâce à la *figure 43*.

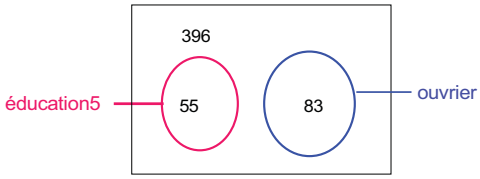

Fig. *43* : *Exemple de règle issue de Wages illustrant le cas de l'incompatibilité.*

L'évolution des quatre indices selon le modèle M_1 est donnée dans la *figure 44* et l'évolution selon le modèle M_2 peut être visualisée grâce à la *figure 45*.

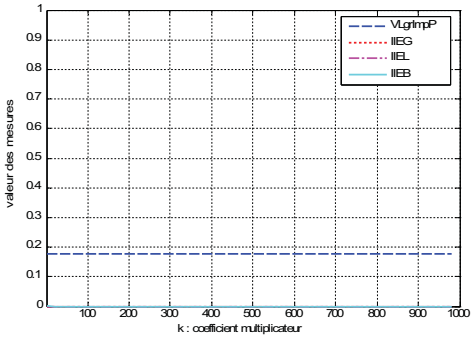

Fig. *44* : *Comparaison des indices selon le modèle M_1 pour la règle "éducation5 → ouvrier".*

Nous constatons une invariance des indices lorsque tous les effectifs sont multipliés par un coefficient multiplicateur *k* (*voir la figure 44*). Nous vérifions la *proposition 7* concernant l'indice *VLgrImpP* pour le modèle de croissance M_1. Quant aux trois autres indices, les *propositions 9* et *10* indiquent une invariance des indices dans le cas où $p(b/a) > p(b)$. Nous ne sommes pas dans ce cas de figure puisque $p(b/a) = 0$ et $p(b) = 0{,}155$ mais cependant nous constatons une invariance. Il est à noter que les courbes d'évolution des indices *IIEG*, *IIEL* et *IIEB* sont confondues, c'est pourquoi nous ne voyons que la courbe de l'indice *IIEB*

car celle-ci est représentée par un trait plein. Tous les indices rejettent cette règle où il y a incompatibilité entre les variables puisque les valeurs des indices sont très faibles.

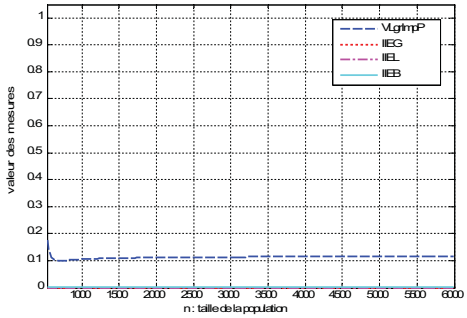

Fig. *45* : *Comparaison des indices selon le modèle M_2 pour la règle "éducation5 → ouvrier".*

Dans le cas de l'incompatibilité et pour le modèle de croissance M_2 (*voir la figure 45*), nous constatons une invariance des indices entropiques *IIEG*, *IIEL* et *IIEB* et nous notons que ces trois courbes sont quasiment confondues avec l'axe des abscisses. Pour l'indice *VLgrImpP*, la courbe est quasiment parallèle à l'axe des abscisses après une petite décroissance pour les faibles valeurs de n ($n < 735$). Ce dernier indice rejette également ce type de règle puisque les valeurs, pour n'importe quelle valeur de n, sont très faibles (*de l'ordre de 0,1*).

6.7.2 Zone de répulsion

La deuxième situation caractéristique étudiée est le cas d'une règle ("*syndiqué → féminin*") située dans la zone de répulsion dont nous rappelons la contingence grâce à la *figure 46*.

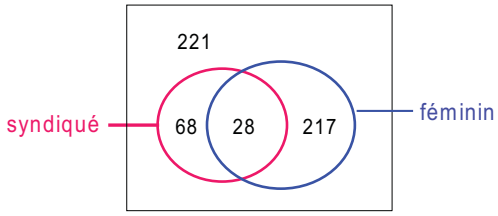

Fig. *46* : *Exemple de règle issue de Wages située dans la zone de répulsion.*

Les *figures 47* et *48* donnent respectivement l'évolution des quatre indices pour cette règle pour les modèles d'évolution M_1 et M_2.

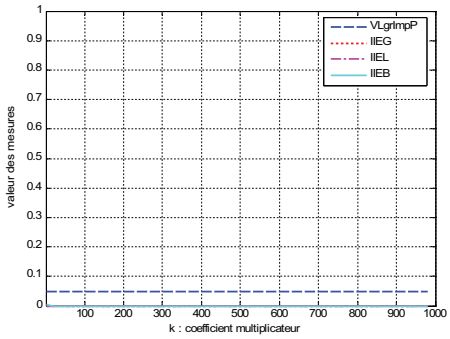

Fig. *47* : *Comparaison des indices selon le modèle M_1 pour la règle "syndiqué → féminin".*

Nous obtenons une situation très similaire au cas précédent de l'incompatibilité (*voir la figure 44*) c'est-à-dire des courbes pour les indices entropiques confondues avec l'axe des x et une courbe pour l'indice *VTgrImpP* invariante (*proposition 7*) avec une faible valeur (*de l'ordre de 0,05*), plus faible que celle du cas de l'incompatibilité (*de l'ordre de 0,18*). L'explication de l'obtention de valeurs inférieures a été donnée lors des commentaires accompagnant la *figure 7*, c'est-à-dire le cas de l'étude de la même règle

mais pour comparer le comportement des indices *VLgrImpP*, *VTeImpBarP*, *VTeImpCorP* et *VTeImpProj*. De nouveau, nous ne sommes pas dans le contexte d'application des *propositions 9* et *10* puisque $p(b/a) = 0{,}29$ et $p(b) = 0{,}46$ mais cependant nous constatons une invariance des indices entropiques selon le modèle de croissance M_1.

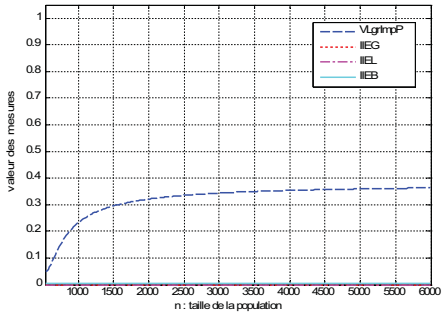

Fig. *48* : *Comparaison des indices selon le modèle M_2 pour la règle "syndiqué → féminin".*

Les courbes des indices entropiques pour le modèle de croissance M_2 (*figure 48*) sont confondues avec l'axe des abscisses, comme dans le cas de l'incompatibilité (*figure 45*), alors que la courbe de *VLgrImpP* est tout d'abord croissante pour tendre ensuite vers une valeur proche de *0,4*. Tous ces indices, par les faibles valeurs restituées, rejettent la règle "*syndiqué → féminin*" pour n'importe quelle valeur de la taille *n* de l'ensemble d'apprentissage.

6.7.3 Indépendance

La troisième étude correspond au cas de l'indépendance dont nous rappelons la contingence de la règle "*salaire3 → nord* " grâce à la *figure 49*.

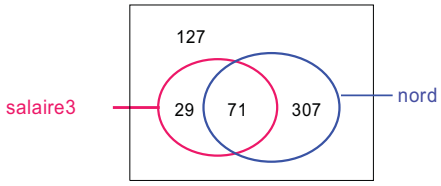

Fig. *49* : *Exemple de règle issue de Wages illustrant le cas de l'indépendance.*

Les *figures 50* et *51* donnent respectivement l'évolution des quatre indices pour la règle "*salaire3 → nord*" pour les modèles de croissance M_1 et M_2.

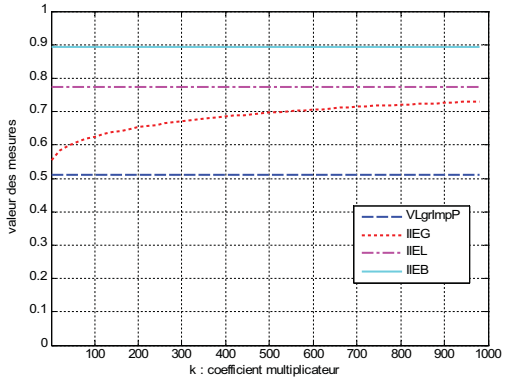

Fig. *50* : *Comparaison des indices selon le modèle M_1 pour la règle "salaire3 → nord".*

Nous constatons pour la première fois des comportements différents pour les indices entropiques. Seul l'indice *IIEG* n'est pas invariant en fonction du coefficient multiplicateur *k* mais nous ne sommes toujours pas dans le cadre d'application des *propositions 9* et *10* puisque $p(b/a) = p(b)$. La courbe de cet indice *IIEG* est croissante pour tendre vers une valeur proche de *0,73*. Nous constatons des valeurs relativement

élevées pour l'*IIEB* puisqu'elles sont de l'ordre de *0,9* puis viennent ensuite les valeurs de *IIEL* (*de l'ordre de 0,78*) et pour finir les valeurs de *IIEG*. L'indice le plus sélectif est *VLgrImpP* puisque ses valeurs sont proches de *0,52*, valeur proche de celle de l'indice de vraisemblance du lien dans le cas de l'indépendance (*égale à 0,50*). Dans ce cas de figure, l'indice *VLgrImpP* a le comportement le plus adapté.

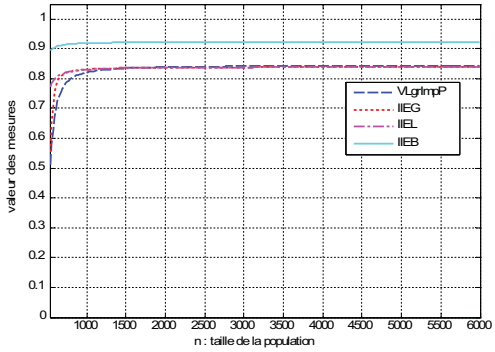

Fig. *51* : *Comparaison des indices selon le modèle M_2 pour la règle "salaire3 → nord"*.

Pour le modèle de croissance M_2, nous constatons des courbes croissantes pour ces quatre indices mais qui deviennent relativement vite constantes puisqu'elles tendent vers la valeur *0,84* pour les indices *VLgrImpP*, *IIEG* et *IIEL* pour des valeurs de taille d'ensemble d'apprentissage supérieures à approximativement *1000* et vers la valeur *0,92* pour l'indice *IIEB* pour des tailles d'ensemble d'apprentissage supérieures à *500*. Pour la première fois, l'indice *VLgrImpP* a un comportement quasiment identique aux indices entropiques et plus particulièrement aux indices *IIEG* et *IIEL*. Comme pour le cas précédent (*modèle de croissance M_1* , *figure 50*), c'est l'indice *IIEB* qui possèdent les plus fortes valeurs, ce qui n'était pas le cas pour l'étude de la situation d'incompatibilité et du cas d'une règle dans la zone de répulsion.

6.7.4 Attraction proche de l'indépendance

Nous étudions maintenant un cas proche de l'indépendance mais où la règle se trouve dans la zone d'attraction. Nous rappelons de nouveau la contingence de cette règle "*éducation3 → non syndiqué*" grâce à la *figure 52*.

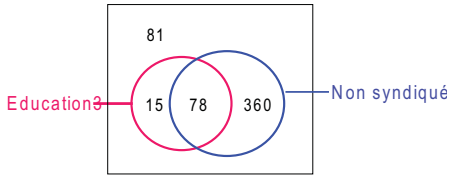

Fig. *52* : *Exemple de règle issue de Wages illustrant un cas proche de l'indépendance mais dans la zone d'attraction.*

Les *figures 53* et *54* donnent respectivement l'évolution des quatre indices pour la règle "*éducation3 → non syndiqué*" pour les modèles de croissance M_1 et M_2.

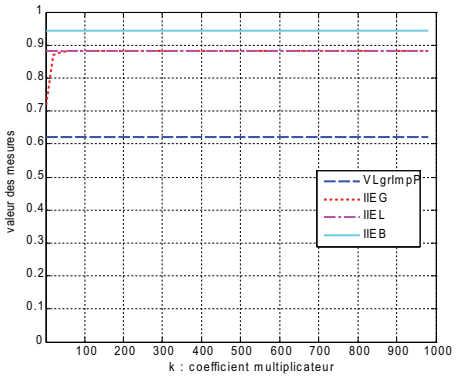

Fig. *53* : *Comparaison des indices selon le modèle M_1 pour la règle "éducation3 → non syndiqué"*.

Nous vérifions une invariance des indices comme cela a été prouvé grâce aux *propositions 7*, *9* et *10*. Cette fois-ci, nous sommes dans le cadre d'application des *propositions 9* et *10* puisque $p(b/a)$ est égale à *0,821* et $p(b)$ est égale à *0,820*, par conséquent nous avons $p(b/a) > p(b)$. Ce qui est surprenant, c'est cette forte croissance de la courbe de l'*IIEG* pour les faibles valeurs du coefficient multiplicateur *k*. Nous sommes dans le cas limite d'application de la *proposition 9* puisque $p(b/a) \approx p(b)$, ce qui peut expliquer cette petite croissance au départ. Les courbes des indices *IIEG* et *IIEL* sont très similaires pour cette situation caractéristique et comme dans le cas de l'indépendance, la mesure la plus sélective est *VLgrImpP* et la moins sélective est *IIEB*, ce qui n'est pas surprenant car ces deux cas expriment des situations relativement proches.

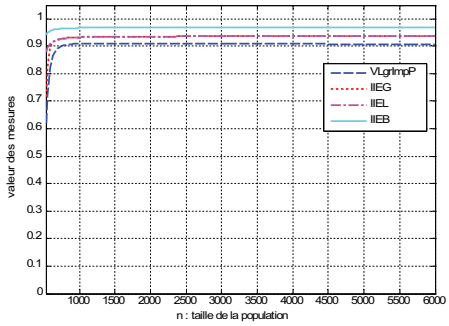

Fig. *54* : *Comparaison des indices selon le modèle M_2 pour la règle "éducation3 → non syndiqué"*.

Ces différentes courbes reflètent une situation assez proche du cas précédent (*indépendance, figure 51*) avec la particularité que la courbe de *VLgrImpP* n'est plus confondue avec celles des indices *IIEG* et *IIEL* mais est légèrement située au dessous.

Nous étudions maintenant un cas similaire au cas précédent (*zone d'attraction proche de l'indépendance*) mais avec une différence au niveau du support de la prémisse puisque celle-ci est plus faible (*4,5% contre 17,4%*) comme le restitue la *figure 55* qui donne la contingence de cette règle "*construction → blanc*". Nous avons vu précédemment, pour l'étude des indices *VLgrImpP*, *VTeImpBarP*, *VTeImpCorP* et *VTeImpProj*, que cette situation donnait des résultats différents et plus particulièrement pour l'indice *VLgrImpP*. C'est pourquoi, nous l'étudions de nouveau.

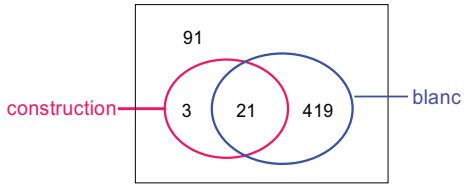

Fig. *55* : *Exemple de règle issue de Wages illustrant le cas d'une règle située dans la zone d'attraction mais proche de l'indépendance.*

Les *figures 38* et *57* donnent respectivement l'évolution des quatre indices pour la règle "*construction* → *blanc*" pour les modèles de croissance M_1 et M_2.

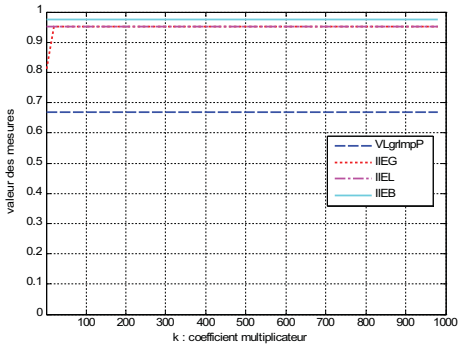

Fig. *56* : *Comparaison des indices selon le modèle M_1 pour la règle "construction → blanc".*

Pour le modèle de croissance M_1, nous obtenons des résultats proches du cas précédent (*voir la figure 53*), à savoir des courbes similaires pour les indices *IIEG* et *IIEL*, une plus grande sélectivité pour l'indice *VLgrImpP* et une moins grande sélectivité pour *IIEB*. Nous retrouvons également cette forte croissance de l'indice *IIEG* pour les faibles valeurs de *k*. Ce qui diffère avec le cas précédent, ce sont les valeurs prises par ces différents indices qui sont légèrement supérieures, ce qui n'est pas surprenant car la confiance de cette règle est légèrement supérieure à la confiance de la règle précédente puisque la confiance de la règle "*construction* → *blanc*" est de *87,5* % et la confiance de la règle précédente "*éducation3* → *non syndiqué*" est de *83,9* %.

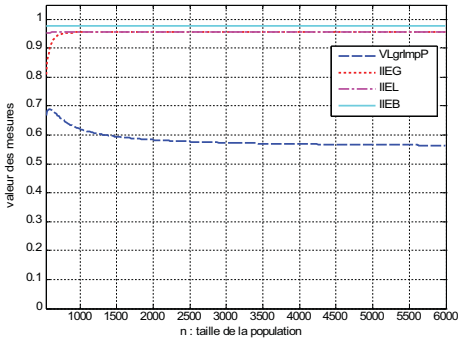

Fig. *57* : *Comparaison des indices selon le modèle M_2 pour la règle "construction → blanc".*

Ces courbes de la *figure 57* montrent une similitude de comportement des indices entropiques avec le cas précédent contrairement à l'indice *VLgrImpP* où la courbe de celui-ci décroît. Cette situation particulière pour l'indice *VLgrImpP* a déjà été discuté lors des commentaires de la *figure 16*.

6.7.5 Indétermination

Cette sixième règle "*éducation2* → *féminin*" va nous permettre d'étudier le cas de l'indétermination et la *figure 58* nous rappelle la contingence de cette règle.

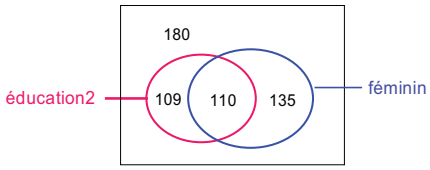

Fig. *58* : *Exemple de règle issue de Wages illustrant le cas de l'indétermination.*

Les *figures 59* et *60* donnent respectivement l'évolution des quatre indices pour la règle "*éducation2* → *féminin*" pour les modèles de croissance M_1 et M_2.

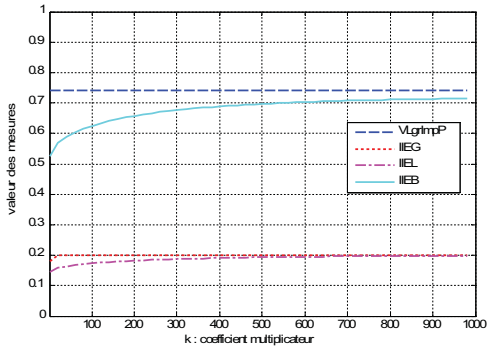

Fig. *59* : *Comparaison des indices selon le modèle M_1 pour la règle "éducation2 → féminin"*.

Pour le modèle M_1, nous obtenons une invariance uniquement pour l'indice *VLgrImpP* et pas pour les indices entropiques alors que nous sommes dans le cas d'application des *propositions 9* et *10* puisque la probabilité conditionnelle de cette règle est de *0,502* et la probabilité de la conclusion est de *0,459*.

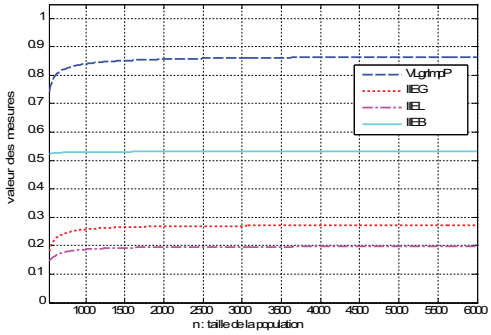

Fig. *60* : *Comparaison des indices selon le modèle M_2 pour la règle "éducation2 → féminin"*.

Pour le modèle de croissance M_2, nous obtenons des courbes très différentes de celles qui précèdent. C'est maintenant l'indice *VLgrImpP* qui a les plus fortes valeurs, les courbes des indices *IIEG* et *IIEL* ne sont plus étroitement liées et l'indice *IIEB* n'est celui qui a les plus fortes valeurs.

6.7.6 Attraction

La septième règle "*féminin → non syndiqué*" est une règle située dans la zone d'attraction et la *figure 61* nous rappelle la contingence des attributs "*féminin*" et "*non syndiqué*".

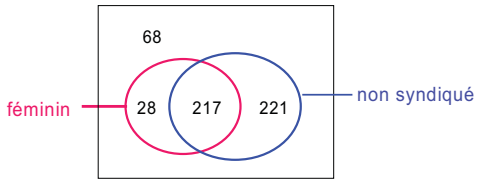

Fig. *61* : *Exemple de règle issue de Wages située dans la zone d'attraction.*

Les *figures 62* et *56* donnent respectivement l'évolution des quatre indices pour la règle "*féminin → non syndiqué*" pour les modèles de croissance M_1 et M_2.

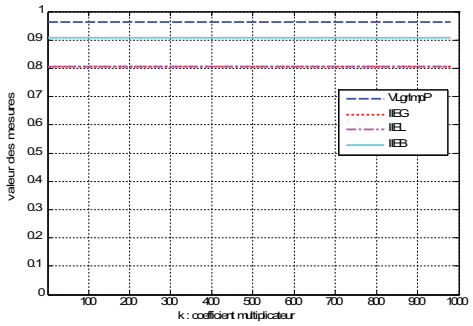

Fig. *62* : *Comparaison des indices selon le modèle M_1 pour la règle "féminin → non syndiqué"*.

Pour le modèle M_1, nous vérifions bien une invariance des indices en fonction du coefficient multiplicateur k. Nous retrouvons des courbes similaires pour les indices *IIEG* et *IIEL* mais cette fois-ci, c'est l'indice *VLgrImpP* qui possède les plus fortes valeurs et non plus l'indice *IIEB* comme pour les *figures 50*, *53* et *38*.

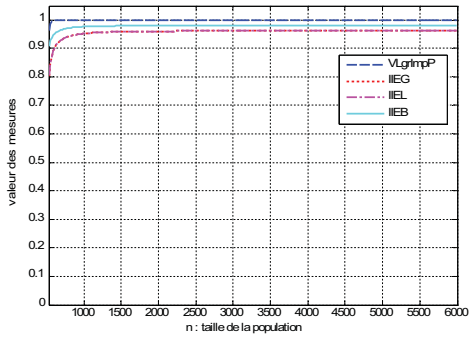

Fig. *63* : *Comparaison des indices selon le modèle M_2 pour la règle "féminin → non syndiqué"*.

Pour le modèle de croissance M_2, nous obtenons des valeurs élevées pour chacun des indices, ce qui n'est pas surprenant puisque nous sommes en présence d'une règle pertinente. Ces courbes ont des allures très similaires et c'est l'indice *VLgrImpP* qui attribue les plus fortes valeurs. Nous retrouvons des courbes confondues pour les indices *IIEG* et *IIEL*.

6.7.7 Implication

La dernière règle "*vendeur → non syndiqué*" correspond au cas de la quasi-implication et la *figure 37* nous rappelle la contingence de ces variables.

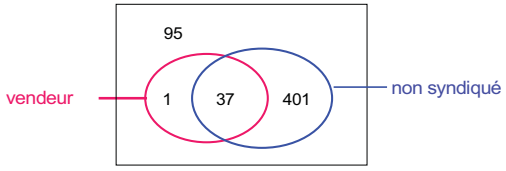

Fig. *64* : *Exemple de règle issue de Wages correspondant au cas de la quasi implication logique.*

Les *figures 36* et *35* donnent respectivement l'évolution des quatre indices pour la règle "*vendeur → non syndiqué*" pour les modèles de croissance M_1 et M_2.

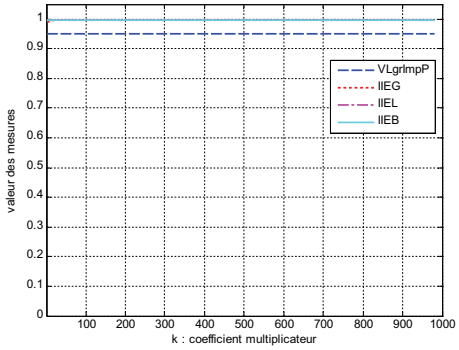

Fig. *65* : *Comparaison des indices selon le modèle M_1 pour la règle "vendeur → non syndiqué*".

Pour le modèle M_1, nous vérifions l'invariance des indices et nous observons des courbes confondues pour les indices entropiques. Les valeurs sont élevées pour chacun des indices et plus particulièrement pour les indices entropiques puisque ceux-ci ont des valeurs égales à *1*.

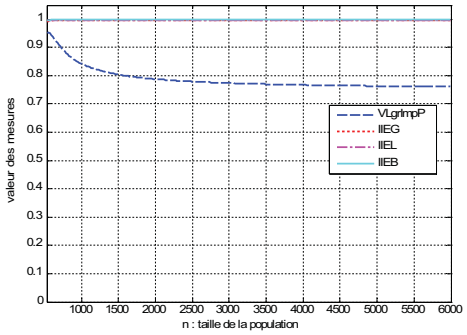

Fig. *66* : *Comparaison des indices selon le modèle M_2 pour la règle "vendeur → non syndiqué*".

Pour le modèle *M2*, nous obtenons des courbes confondues et invariantes pour les indices entropiques avec une valeur élevée pour toutes ces règles puisque celle-ci est égale à *1*. Pour l'indice *VLgrImpP*, nous constatons, comme pour la *figure 25*, une décroissance en raison du faible support de la prémisse de la règle. Cette décroissance a été discutée lors de l'étude précédente concernant le comportement comparatif des indices *VLgrImpP*, *VTeImpBarP*, *VTeImpCorP* et *VTeImpProj*.

7 Conclusion et Perspectives

Cette étude comporte clairement un versant d'analyse théorique et un versant d'analyse expérimentale qui sont intimement liés. L'objectif que nous nous sommes assignés a consisté à analyser de façon comparative le comportement différents indices d'implication comprenant une approche probabiliste et qui sont discriminants pour comparer entre elles différentes règles d'association issues d'une même base de données. Chacun de ces indices permet à sa façon de récolter, relativement à une même base de données, les "meilleures" ou plus exactement, les plus intenses. On a vu qu'il y avait deux types bien différenciés d'indices. Pour le premier, comprenant les indices $VLgrImpP$, $VTeImpProj$, $VTeImpCorP$ et $VTeImpBarP$, c'est l'échelle de probabilité de l'indice aléatoire $n(a^\star \wedge \bar{b}^\star)$ qui est géré par rapport à un contexte, ce qui permet d'en déduire une échelle de probabilité discriminante. Pour le deuxième type d'indices qui comprend $IIEG$, $IIEL$ et $IIEB$, il s'agit d'un indice d'inclusion pondéré multiplicativement par un indice probabiliste dont l'effet s'évanouit dès lors que le nombre d'individus dépasse un ordre de grandeur de 1000.

Deux modèles de croissance de la taille de la base de données $M1$ et $M2$ [voir (104) et (105)] ont été considérés pour étudier de façon relative les différents indices. Cette comparaison a pris comme appui différentes situations caractéristiques par rapport à la problématique de l'évaluation d'une implication de la forme $(a \rightarrow b)$ (voir sous section 6.3.1). C'est la base de données "Wages" [24] qui nous a servi de support à l'étude. Si le modèle $M1$ permet de constater de façon indiscutable comment chacun des indices ordonne les différentes situations caractéristiques, c'est le modèle $M2$ qui permet pour les différentes situations caractéristiques d'étudier le comportement, l'évolution comparée, des différents indices. Il ressort clairement que l'indice $VLgrImpP$ peut jouer un rôle de référence. C'est en effet d'une part dû à la clarté de sa conception et d'autre part, dû à la stabilité de son comportement lorsque la taille de la base augmente. Il est apparu expérimentalement, dans l'ensemble des indices du premier type, deux groupes notés G_1 et G_2 où G_1 comprend $VLgrImpP$ et $VTeImpProj$ et où G_2 comprend $VTeImpBarP$ et $VTeImpCorP$. Nous avons établi que ces deux derniers indices sont très semblables aussi bien relativement à leurs comportements respectifs que sur le plan conceptuel. À cet égard nous avons pu formaliser la nature de la conception de $VTeImpCorP$ en termes de statistique mathématique non paramétrique (voir sous-section 4.1 et 5.2.3). Dans ces conditions, il est légitime de s'interroger sur cette réduction forcée à un échantillon de taille e. Maintenant, relativement au problème de la sélection des r meilleures règles (nous avons pris $r = 20$), il a été remarquable de constater que les quatre différents indices relevant du premier type donnaient exactement les mêmes règles. De plus, il s'agissait exactement du même ordre pour les indices du groupe G_1 ; ainsi que du même ordre - à une transposition près - pour les indices du groupe G_2 ; les six meilleures règles étant classées de la même façon par chacun des deux ordres. Pour ce qui est des indices relevant du second type ($IIEG$, $IIEL$ et $IIEB$) il y a une différence sensible du comportement par rapport aux indices relevant du premier type. Plus précisément, en se situant par rapport à l'indice de référence $VLgrImpP$, il y a lieu de considérer les 26 meilleures règles sélectionnées par ce dernier indice pour couvrir les 20 meilleures règles de l'un ou l'autre des indices $IIEG$, $IIEL$ ou $IIEB$. Ces in-

dices ont un comportement assez similaire puisque les 20 meilleures règles pour $IIEG$ permettent de récolter 18 des meilleures règles pour $IIEL$ ou $IIEB$; les 6 meilleures règles étant les mêmes pour les trois indices. Par ailleurs, 15 ou 16 des meilleures règles sélectionnées par l'un ou l'autre de ces derniers indices se retrouvent parmi les 20 meilleures règles pour $VLgrImpP$. Dans ces conditions, la question se pose naturellement de savoir lequel des indices choisir ?

Nous avons vu que $VLgrImpP$ pouvait jouer un rôle de référence. Cependant, il y a derrière chacun des indices une approche spécifique qui a sa propre logique. On peut vouloir choisir l'un des indices ou se situer par rapport à un sous ensemble $\mathcal{IND}$ d'indices. En désignant par *indice* l'un des éléments de $\mathcal{IND}$ et par $\rho^{indice}(a \to b)$ le rang d'une règle $(a \to b)$ pour *indice*, on peut à partir d'un entier J ($J = 20$ ci-dessus), récolter un sous ensemble de règles de la forme $(a \to b)$ appartenant à $\mathcal{C}_{cs}$ telles que pour chacune, le minimum pour les différents indices de $\rho^{indice}(a \to b)$ est inférieur ou égal à J. L'ensemble des règles ainsi recueilli peut s'écrire :

$$\mathcal{R}_J^{\mathcal{IND}} = \Big\{(a \to b) \mid (a,b) \in \mathcal{C}_{cs} \text{ et } min\{\rho^{indice}(a \to b) \mid indice \in \mathcal{IND}\} \leq J\Big\} \tag{133}$$

On peut vouloir établir un classement (préordre total) de préférence sur l'ensemble des règles obtenu. À cet égard, on affectera à chaque règle retenue $(a \to b)$ la moyenne sur $\mathcal{IND}$ de $\rho^{indice}(a \to b)$.

L'évaluation de la qualité d'une règle par l'un ou l'autre des différents indices reste à des degrés divers intrinséque à une même base de données. C'est même une caractéristique mise en exergue pour ce qui concerne $VLgrImpP$. mais c'est également le cas pour les autres indices si on ne suppose pas nécessairement une stabilité au niveau de l'échantillonnage. Nous voulons dire que si une règle $(a \to b)$ est présente dans deux bases de données distinctes, son évaluation peut être différente selon qu'il s'agit de l'une ou de l'autre base. D'autre part, même si on suppose une stabilité du comportement statistique entre les deux bases, un indice de type VTe reste intimement lié à la taille de la base dans le cadre de laquelle il est calculé.

Revenons un instant - pour ce qui est du comportement de $VLgrImpP$ - au cas des deux configurations implicatives données par les figures 20 et 23. Le comportement de cet indice semble mieux répondre à notre intuition dans le premier cas que dans le second cas. On remarquera entre parenthèses que cette intuition semble davantage violée pour ce qui est des indices du groupe G_2. Nous avons attribué ce phénomène à la faiblesse de la taille de la prémisse dans la configuration donnée par la figure 23. En effet, lors d'une expérience menée - mais non rapportée ici - nous avons à partir de la configuration de la figure 23 augmenté de 10 en 10 le nombre d'exemples tout en préservant $n(b)$. Ainsi, la suite des valeurs considérées du couple $[n(a \wedge b), n(\bar{a} \wedge b)]$ est $[n(a \wedge b) + k \times 10, n(\bar{a} \wedge b) - k \times 10]$ pour $k \leq n(\bar{a} \wedge b)/10$. Déjà pour $k = 3$ où $[n(a \wedge b), n(\bar{a} \wedge b)] = (67, 371)$, la valeur limite de l'indice est de 0.92. On a constaté que pour $k = 6$ $[n(a \wedge b), n(\bar{a} \wedge b)] = (97, 341)$ où la valeur de la taille de la prémisse est bien inférieure à celle de la configuration de la figure 20, la valeur limite est sensiblement au dessus. Toutefois avant de se stabiliser,

la courbe est légèrement décroissante, surtout au tout début. Ainsi, pour cet indice ($VLgrImpP$), essentiellement contextuel, il y aura lieu d'étudier d'une part, le phénomène de croissance où la grandeur de $n(a \wedge \bar{b})$ semble jouer un rôle et d'autre part, l'importance de la valeur limite.

Nous avons ci-dessus surtout évoqué l'étude de l'influence de la taille de la prémisse. On peut également définir des modèles d'étude impliquant la taille de la conclusion. Signalons enfin deux modèles intéressants [29] où relativement au tableau de contigence 2×2 croisant $\{a, \bar{a}\}$ avec $\{b, \bar{b}\}$, on procède à la multiplication des contenus, soit des deux lignes $[n(a \wedge b), n(a \wedge \bar{b})]$ et $[n(\bar{a} \wedge b), n(\bar{a} \wedge \bar{b})]$ soit des deux colonnes $[n(a \wedge b), n(\bar{a} \wedge b)]$ et $[n(a \wedge \bar{b}), n(\bar{a} \wedge \bar{b})]$, respectivement par deux constantes k_1 et k_2.

Références

[1] R. AGRAWAL, T. IMIELINSKY, and A. SWAMI. Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD'93*, pages 207–216, 1993.

[2] J. BLANCHARD, F. GUILLET, H. BRIAND, and R. GRAS. Assessing rule with a probabilistic measure of deviation from equilbrium. In France École Nationale Supérieure des Télécommunications, Brest, editor, *11th International Symposium on Applied Stochastic Models and Data Analysis ASMDA*, pages 191–200, 2005.

[3] F. DAUDÉ. *Analyse et justification de la notion de ressemblance entre variables qualitatives dans l'optique de la classification hiérarchique par AVL*. PhD thesis, Université de Rennes 1, 1992.

[4] W. FELLER. *An introduction to probability theory and its applications*. John Wiley, 1964.

[5] R. GRAS. *Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques, Doctorat d'État*. PhD thesis, Université de Rennes 1, 1979.

[6] R. GRAS, P. KUNTZ, and H. BRIAND. Les fondements de l'analyse statistique implicative et quelques prolongements pour la fouille des données. *Mathématiques et Sciences Humaines*, (154-155) :9–29, 2001.

[7] S. GUILLAUME. *Traitement des données volumineuses. Mesures et algorithmes d'extraction des règles d'association et règles ordinales*. PhD thesis, Université de Nantes, 2000.

[8] J.-B. LAGRANGE. Analyse implicative d'un ensemble de variables numériques ; application au traitement d'un questionnaire à réponses modales ordonnées. *Revue de Statistique Appliquée*, (46, n°1) :71–93, 1998.

[9] S. LALLICH and O. TEYTAUD. Évaluation et validation de l'intérêt des règles d'association. *Mesures de Qualité pour la Fouille des Données, Cépaduès*, (RNTI-E-1) :193–218, 2004.

[10] S. LALLICH, B. VAILLANT, and P. LENCA. Parametrised measures for the evaluation of association rule interestingness. In France École Nationale Supérieure des Télécommunications, Brest, editor, *11th International Symposium on Applied Stochastic Models and Data Analysis ASMDA*, pages 220–229, 2005.

[11] P. LENCA, P. MEYER, B. PICOUET, and S. LALLICH. Évaluation et analyse multicritère des mesures de qualité des règles d'association. *Mesures de Qualité pour la Fouille des Données, Cépaduès*, (RNTI-E-1) :219–245, 2004.

[12] I-C. LERMAN. Sur l'analyse des données préalable à une classification automatique ; proposition d'une nouvelle mesure de similarité. *Mathématiques et Sciences Humaines*, (8) :5–15, 1970.

[13] I-C. LERMAN. Justification et validité statistique d'une échelle [0,1] de fréquence mathématique pour une structure de proximité sur un ensemble de variables observées. *Publications de l'Institut de Statistique des Universités de Paris*, (29) :27–57, 1984.

[14] I-C. LERMAN. Analyse de la vraisemblance des liens relationnels : une méthodologie d'analyse classificatoire des données. In Younès Bennani et Emmanuel Viennet, editor, *Apprentissage artificielet fouille de données, RNTI A3*, pages 93–126. Cépaduès, 2009.

[15] I-C. LERMAN and J. AZÉ. A new probabilistic measure of interestingness for association rules, based on the likelihood of the link. In F. Guillet and H.J. Hamilton, editors, *Quality measures in data mining, Studies in Computational Intelligence, vol. 43*, pages 207–236. Springer, 2007.

[16] I.C. LERMAN. Étude distributionnelle de statistiques de proximité entre structures finies de même type ; application à la classification automatique. *Cahiers du Bureau Universitaire de Recherche Opérationnelle*, (19) :1–52, 1973.

[17] I.C. LERMAN. Introduction à une méthode de classification automatique illustrée par la recherche d'une typologie des personnages enfants à travers la littérature enfantine. *Revue de Statistique Appliquée*, (XXI) :23–49, 1973.

[18] I.C. LERMAN. *Classification et analyse ordinale des données*. Dunod, 1981.

[19] I.C. LERMAN. Foundations of the likelihood linkage analysis (lla) classification method. *Applied Stochastic Models and Data Analysis*, (7) :379–397, march 1991.

[20] I.C. LERMAN. Conception et analyse de la forme limite d'une famille de coefficients statistiques d'association entre variables relationnelles. *Mathématiques, Informatique et Sciences humaines*, (118) :33–52, 1992.

[21] I.C. LERMAN, R. GRAS, and H. ROSTAM. Élaboration et évaluation d'un indice d'implication pour des données binaires i et ii. *Mathématique et Sciences Humaines*, (74-75) :5–35, 5–47, 1981.

[22] A. MORINEAU. Note sur la caractérisation statistique d'une classe et les valeurs-tests. *Bulletin technique du Centre de Statistique et d'Informatique Appliquée*, (2) :20–27, 1884.

[23] A. MORINEAU and R. RAKOTOMALALA. Critère vt100 de sélection des régles d'association. In Cépadués, editor, *Actes de Extraction et Gestion de Connaissances, EGC'2006*, pages 581–592, 2006.

[24] P.M. MURPHY and D.W. AHA. Uci repository of machine learning databases. Machine-readable collection, Dept of Information and Computer Science, University of California, Irvine, 1995.

[25] P. PETER, H. LEREDDE, and I. C. LERMAN. Notice du programme chavlh (classification hiérarchique par analyse de la vraisemblance des liens en cas de variables hétérogènes),. *Agence pour la Protection des Programmes*, (IDDN.FR.001.240016.000.S.P.2006.000.20700), Décembre 2005.

[26] G. PIATETSKY-SHAPIRO. Discovery, analysis, and presentation of strong rules. In *Knowledge Discovery in Databases*, pages 229–248. MIT Press, 1991.

[27] R. RAKOTOMALALA and A. MORINEAU. The tvpercent principle for the counterexamples statistic. In F. Guillet R. Gras, E. Suzuki and F. Spagnolo, editors, *Statistical Implicative Analysis*, pages 449–462. Springer, 2008.

[28] G. RITSCHARD. De l'usage de la statistique implicative dans les arbres de classification. In *Troisième Rencontre Internationale - Analyse Statistique Implicative*, pages 305–316, 2005.

[29] P-N. TAN, V. KUMAR, and J. SRIVASTAVA. Selecting the right interestingness measure for association patterns. In *Proceedings of the 8th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2002.

Centre de recherche INRIA Rennes – Bretagne Atlantique
IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)

Centre de recherche INRIA Futurs : Parc Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Grenoble – Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399