Chapitre 6 : Analyse statistique implicative entre variables vectorielles

Régis Gras*, Raphaël Couturier**

*Equipe Connaissances et Décision, Laboratoire d'Informatique de Nantes Atlantique
Ecole Polytechnique de l'Université de Nantes, UMR 6241
La Chantrerie BP 60601 44306 Nantes cedex
regisgra@club-internet.fr

** Institut Universitaire de Technologie de Belfort,
BP 527, rue E. Gros, 90016 Belfort *cedex*
raphael.couturier@iut-bm.univ.fcomte.fr

Résumé. Nous nous plaçons ici dans le cadre de la méthode d'analyse de données, l'analyse statistique implicative (A.S.I.). A l'instar de ce que nous avons fait pour passer des variables binaires aux variables numériques ou aux variables-intervalles, nous étendons le champ des traitements aux variables à valeurs vectorielles. Nous établissons un indice permettant de mesurer la qualité d'une règle entre variables vectorielles. Nous traitons des exemples portant l'un sur le baccalauréat, l'autre sur l'examen des critères de convergence des économies de l'Union Européenne.

1 Introduction

Les développements théoriques de l'analyse de données offrent des retombées enrichissantes pour l'Extraction de Connaissances. Ses développements et sa vitalité ne sont pas étrangères aux échanges induits. Par exemple, la construction d'indices permettant d'affecter une mesure non symétrique à des règles d'inférence partielle fournit des points d'application à l'extraction et à la représentation de règles d'association imprécises entre attributs binaires décrivant une population. Les démarches fondamentales se ramènent à la prise en compte d'une problématique commune aux deux domaines ; il s'agit de découvrir et de quantifier des règles non symétriques pour modéliser des relations du type *"si a, alors presque b"*. Qu'il s'agisse de réseaux bayésiens (Pearl, 1988), de treillis de Galois (Bernard et Poitrenaud, 1999) ou de fouille de règles (Agrawal et al, 1993) de très nombreuses mesures ont été proposées pour quantifier la pertinence de ces quasi-implications et optimiser leur extraction (par ex. (Hilderman et Hamilton, 1999) ou (Tan et Kumar, 2000)), Des travaux sur la qualité des règles d'association ont permis comparer leurs mesures selon des points de vue subjectifs et objectifs (Lenca et al, 2004). Cependant, à notre connaissance, ces travaux se limitent généralement à l'étude de mesures pour des règles entre attributs binaires ou conjonction de tels attributs.

Or les situations réelles, y compris celles pour lesquelles l'analyse implicative a créé son modèle statistique (la didactique des mathématiques), conduisent au traitement d'autres types

de variables. C'est ainsi que des extensions successives, pour répondre à des applications comme dans (Gras et al. 1996) nous ont conduits à intégrer dans la même théorie, des variables modales, numériques, floues, des variables sur intervalles, des variables-intervalles, des variables-rangs et à élaborer des outils de représentation graphique. Dans cet article, nous étudions, toujours dans le cadre de l'analyse implicative, le cas de variables à valeurs vectorielles.

2 Problématique

On veut comparer l'évolution temporelle d'un système de variables à d dimensions. Par exemple, sans pratiquer de test d'hypothèse mais sans l'exclure, on envisage :

une suite de *d* notes obtenues dans des disciplines, indépendantes a priori, examinée à des périodes régulières (trimestre) sur un ensemble E d'élèves soumis à un traitement expérimental. On cherche à savoir si, globalement, les performances à l'instant *t* impliquent celles obtenues à l'instant *t+1*, *t+2*, etc. et donc si le traitement expérimental est influent ;

des mesures (analyses médicales) de *d* paramètres, indépendants a priori, obtenues à des périodes régulières afin d'étudier l'effet d'un traitement sur une population *E* de patients. On cherche s'il existe des règles d'association entre les moments d'observation et donc s'il existe un effet du traitement ;

on fait l'observation à *d* instants (année, séquence, phase,…) de valeurs de paramètres dont on veut extraire d'éventuelles règles d'évolution d'un moment à un autre ; par exemple, l'évolution comparée de la fonction en entreprise occupée par des individus à *d* instants par rapport aux salaires respectivement octroyés à ces instants.

La représentation de ces suites est vectorielle. A une observation sur un sujet correspond le vecteur des observations de chacune de ses *d* composantes. Ces observations sont quantitatives, de nature binaire, modale ou intervalle. Une règle d'association entre deux vecteurs aura d'autant plus de sens que les composantes vectorielles présenteront une sémantique commune (par exemple : «progrès », « régression », etc.). Elle se substituera d'autant plus avantageusement à l'analyse implicative classique entre les composantes, modalités de chaque vecteur, que celles-ci présenteront des caractéristiques indépendantes dans leur essence même.

Mais les deux approches peuvent être complémentaires et informatives : l'analyse vectorielle est globale, l'analyse classique est ponctuelle. La première vise, en quelque sorte, à partir de profils individuels (la suite des composantes vectorielles) à dégager des règles d'association entre profils synthétiques. La seconde permet de dégager des règles d'association entre attributs binaires ou non, élémentaires ou composites (obtenu par la conjonction d'attributs élémentaires). Mais un attribut composite n'est pas un profil, il n'en est que la contraction. Ainsi, l'analyse vectorielle ne se ramène pas à l'analyse classique.

3 Cas de vecteurs à composantes binaires

3.1 Détermination d'un indice « vectoriel »

Les situations précédentes relèvent de variables numériques qu'il est possible de garder comme telles mais qu'il est aussi possible de binariser en ramenant chaque valeur à sa position par rapport à une norme : au-dessus, au-dessous. Mais d'autres situations peuvent se présenter sous une forme immédiatement binaire : par ex., une épreuve composée de d modalités où les observations seraient réussite-échec, présence-absence, etc. [1].

On cherche donc à comparer des vecteurs de type $\vec{a} = (a_1, a_2, \ldots, a_d)$, vecteur représentatif d'une **variable vectorielle** $\vec{a}$, et $\vec{b} = (b_1, b_2, \ldots, b_d)$, vecteur représentatif d'une variable vectorielle $\vec{b}$. On veut en extraire par exemple la règle $\vec{a} \Rightarrow \vec{b}$. Pour ce faire, on associe à chaque observation selon n sujets,

- d'une part , un vecteur à *d* dimensions, associé à la variable $\vec{a}$, de la somme sur les *n* sujets de chacune des *d* composantes, c'est-à-dire le vecteur-ligne : $\vec{A} = (\sum_{i=1}^{i=n} a_1(i), \sum_{i=1}^{i=n} a_2(i), \ldots, \sum_{i=1}^{i=n} a_d(i))$. $\vec{A}$ admet ainsi pour j^{ème} composante scalaire le nombre $\sum_{i=1}^{i=n} a_j(i)$. Ce nombre représente donc le nombre de fois (*cas binaire*) où apparaît la variable composante a_j (ou la somme de ses pondérations dans le *cas numérique*) dans l'ensemble de la population. Soit aussi $\vec{A}(i)$ le vecteur-ligne des *d* composantes du sujet *i* de la forme ($a_1(i)$, $a_2(i)$,… $a_j(i)$,… $a_d(i)$). $a_j(i)$ est la j^{ème} composante du vecteur $\vec{A}(i)$, c'est-à-dire la valeur prise par le sujet *i* selon la variable composante a_j. On fait de même avec la variable $\vec{b}$ afin d'obtenir le vecteur-ligne $\vec{\mathbf{B}}_i$;

- d'autre part, deux ensembles aléatoires de vecteurs-lignes, choisis au hasard et indépendamment l'un de l'autre, dont les sommes vectorielles-lignes $\vec{X}$ et $\vec{Y}$ respectives coïncident exactement et respectivement, composante par composante, avec celles de $\vec{A}$ et de $\vec{B}$. Par analogie avec le cas des variables *a* et *b* uniques (une seule composante), ces deux ensembles sont à comparer aux ensembles *X* et *Y* de sujets, de même cardinaux que ceux de *A* et *B* supports des variables respectives *a* et *b*. Mais, bien entendu, ce ne sont pas généralement, les mêmes sujets qui les ont constitués.

A ces vecteurs $\vec{A}$ et $\vec{B}$, on associe le vecteur $\vec{w}$ des contre-exemples observés aux implications successives selon les vecteurs-sujets $\vec{\mathbf{A}}(\mathbf{i})$ et $\vec{\mathbf{B}}(\mathbf{i})$, donc d'indice identique. Par exemple, un contre-exemple en *i* à $a_j \Rightarrow b_j$ apparaît lorsque $a_j(i) = 1$ alors que $b_j(i) = 0$, c'est-à-dire lorsque $a_j(i) \wedge \bar{b}_j(i) = 1$. Par suite, le vecteur des contre-exemples observés aura pour

[1] S'il y a nécessité d'étendre la méthodologie employée dans le cas d'une variable binaire, la généralisation au cas de variable numérique se fera comme dans le cas de la composante unique en utilisant la valeur corrigée des indices définis dans (Lagrange, 1998) ou (Régnier et Gras,,2005).

composantes scalaires les *j* scalaires de la forme $\sum_{i=1}^{i=n} a_j(i) \wedge \bar{b}_j(i)$ que nous notons, comme dans le cas de la variable unique $n_{a_j \wedge \bar{b}_j}$. Aux vecteurs aléatoires $\vec{X}$ et $\vec{Y}$, on associe de la même façon, le vecteur $\vec{W}$ des contre-exemples aléatoires de composantes $\sum_{i=1}^{i=n} x_j(i) \wedge \bar{y}_j(i)$ notés également $N_{a_j \wedge \bar{b}_j}$.

Or, $N_{a_j \wedge \bar{b}_j}$ suit, en tant que variable aléatoire relative à la règle présupposée $a_j \Rightarrow b_j$, et comme nous l'avons prouvé par ailleurs (Lerman et al1981) et Partie 1, la loi de Poisson de paramètre $\frac{n_{a_j}}{n}. n_{\bar{b}_j}$ ou bien la loi binomiale de paramètres n et $\frac{n_{a_j}}{n}. \frac{n_{\bar{b}_j}}{n}$. Alors, du fait de l'indépendance des variables composantes, la probabilité pour que les nombres de tous les contre-exemples aléatoires ne soient pas inférieurs aux nombres de contre-exemples respectivement observés est le produit des probabilités pour que cette propriété soit vérifiée sur chaque composante, à savoir :

$$\prod_{j=1}^{j=d} \left[1 - \Pr[N_{a_j \wedge \bar{b}_j} \leq n_{a_j \wedge \bar{b}_j}] \right]$$

On définit alors l'intensité d'implication de la règle $\vec{a} \Rightarrow \vec{b}$ par [2]

$$\varphi(\vec{a}, \vec{b}) = \left(\prod_{j=1}^{j=d} \left[1 - \Pr[N_{a_j \wedge \bar{b}_j} \leq n_{a_j \wedge \bar{b}_j}] \right] \right)^{\frac{1}{d}}$$

Par suite, compte tenu de la définition de l'indice classique de l'intensité d'implication d'une variable vers une autre : $\varphi(\vec{a}, \vec{b}) = \left[\prod_{j=1}^{j=d} \varphi(a_j, b_j) \right]^{\frac{1}{d}}$

Le logiciel C.H.I.C. (Couturier et Gras. 2005) et chapitre 11 permet le calcul des intensités d'implication dans le cas vectoriel. Nous en aurons deux applications plus loin.

3.2 Remarques

On peut comparer cette extension à la situation originelle où l'on a, pour une variable unique, un vecteur à une dimension. Dans le cas vectoriel présenté ici, une quasi-inclusion du vecteur représentatif de *a* dans un parallélépipède de dimension *d* définie par sa diagonale $\vec{b}$ se substitue à la quasi-inclusion de l'ensemble A dans l'ensemble B, supports respectifs de *a* et *b*.

[2] Une formule acceptable et alternative de celle-ci est . $1 - \prod_{j=1}^{j=d} \left[\Pr[N_{a_j \wedge \bar{b}_j} \leq n_{a_j \wedge \bar{b}_j}] \right]^{1/d}$

Toutes les formules où figure l'intensité d'implication $\varphi(\vec{a},\vec{b})$ peuvent être réécrites en remplaçant φ par ψ qui symbolise l'implication-inclusion sur des bases entropiques (Gras et al, 2001).

De même, ces formules peuvent être étendues au cas des variables numériques selon la transformation établie par (Lagrange, 1998), qui modifient l'écart-type de la loi de Poisson. On pourrait s'intéresser à l'événement où *k* composantes seulement parmi les *d* satisferaient l'inégalité entre la valeur aléatoires des contre-exemples et la valeur observée. Ce qui correspond à une exigence affaiblie quant à la règle $\vec{a} \Rightarrow \vec{b}$. Pour cela, il suffit d'envisager toutes les parties à *k* éléments parmi $\{(a_1,b_1),(a_2,b_2),....,(a_d,b_d)\}$, de faire la moyenne arithmétique des sommes des produits correspondants dans le crochet donnant $\varphi(\vec{a},\vec{b})$ et d'en prendre la puissance correspondante.

Par exemple, si *d=3* et *k=2*, on aurait :

$\varphi(\vec{a},\vec{b})=\{\frac{1}{3}[\varphi(a_1,b_1).\ \varphi(a_2,b_2).+\ \varphi(a_2,b_2).\ \varphi(a_3,b_3)+\ \varphi(a_3,b_3).\ \varphi(a_1,b_1)]\}^{1/2}$

Le respect de la sémantique quant au jugement associé à la mise en évidence d'une implication peut être est délicat. Il est loisible, pour certaines des variables composantes de la prémisse ou de la conclusion, de changer l'ordre défini a priori sur les valeurs (binaires ou non) que prennent ces composantes. Ce changement viserait à optimiser l'intensité d'implication $\varphi(\vec{a},\vec{b})$. Par exemple, on pourrait permuter les valeurs *0* et *1* dans le cas binaire ou, de façon générale, passer à la complémentation à *1* de chaque valeur.

3.3 Exemple 1

Une population est constituée de *85* Sujets, dont certains possèdent les mêmes caractéristiques que leur prototype. Ainsi on observe, dans le tableau 1, *20* Sujets répondant *0* en a_1 et de même selon b_1, *1* en a_2 et b_2

	$\vec{a}$		$\vec{b}$	
	a_1	a_2	b_1	b_2
20Sujets1	0	1	0	1
30Sujets2	0	0	1	0
30Sujets3	1	0	1	1
5Sujets4	1	1	0	0
Totaux	**35**	**25**	**60**	**50**

TAB. 1

On observe : $n_{a_1\wedge\bar{b}_1} = n_{a_2\wedge\bar{b}_2} = 5$. En utilisant le modèle de Poisson, on obtient en se limitant au calcul par l'intensité d'implication classique (et non pas « entropique ») :

$$\Pr[N_{a1\wedge\bar{b}1}\leq 5] = 0.05 = \Pr[N_{a2\wedge\bar{b}2}\leq 5] \text{ d'où } \varphi(\vec{a},\vec{b})=(0.95\text{x}0.95)^{1/2}= 0.95$$

4 Cas de vecteurs à composantes numériques

4.1 Indices vectoriels

Nous avons établi (Lagrange, 1998) un indice (Partie 1 Chap. 3) respectant l'approche statistique du cas binaire aux cas où :

1. les variables sont modales (degré de possession d'un attribut ou d'adéuation à celui-ci),
2. les variables sont numériques (nombre d'occurrences, valeurs de contingence)

L'indice adopté permet d'accorder une mesure à l'implication entre de telles variables. De plus, cet indice restreint au cas binaire coïncide avec l'indice mesurant traditionnellement la règle où prémisses et conclusions sont binaires.

Si a et b sont deux variables numériques, dont les valeurs observées respectivement $\{a_i\}_{i \in E}$ et $\{b_i\}_{i \in E}$ dans une population *E* de taille *n* et dont les moyennes et variances respectives sont m_a, m_b, s_a, s_b, alors l'indice de base est :

$$\tilde{q}(a,\bar{b}) = \frac{\sum_{i \in E} a_i \bar{b}_i - \frac{m_a m_{\bar{b}}}{n}}{\sqrt{\frac{(n^2 s_a^2 + m_a^2)((n^2 s_{\bar{b}}^2 + m_{\bar{b}}^2)}{n^3}}}$$

que nous avons aussi nommé, indice de propension.

et l'intensité de la règle $\vec{a} \Rightarrow \vec{b}$ est dans une approximation gaussienne :

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

expression dans laquelle $Q(a,\bar{b})$ est la valeur centrée réduite de $N_{a \wedge \bar{b}}$, pour l'une des lois retenues dans le modèle : binomiale ou de Poisson ou gaussienne.

Considérant maintenant des vecteurs de type $\vec{a} = (a_1, a_2, \ldots, a_d)$, vecteur représentatif d'une variable vectorielle $\vec{a}$, et $\vec{b} = (b_1, b_2, \ldots, b_d)$, vecteur représentatif d'une variable vectorielle $\vec{b}$ où $a_1, a_2, \ldots, a_d$ et $b_1, b_2, \ldots, b_d$ sont *2d* variables numériques. L'intensité d'implication de la règle $\vec{a} \Rightarrow \vec{b}$ est encore : $\varphi(\vec{a}, \vec{b}) = \left[\prod_{j=1}^{j=d} \varphi(a_j, b_j)\right]^{1/d}$

4.2 Exemple 2

Nous disposons des résultats en 1997[3] aux *3* baccalauréats (Scientifique S, Littéraire L et Economique et Social ES) dans les *26* académies de France, résultats exprimés en pourcentages d'admis définitifs. Ces résultats figurent dans le TAB 2 exprimés en ‰.

[3] Références extraites du QUID 99

Académie	S	L	ES	Académie	S	L	ES
Aix	760	762	714	Montpellier	777	742	734
Amiens	753	725	715	Nancy	752	765	795
Besançon	758	767	785	Nantes	781	815	807
Bordeaux	726	779	761	Nice	750	737	735
Caen	718	720	746	Orléans	787	779	817
Clermont	730	805	799	Paris	787	754	743
Corse	794	779	702	Poitiers	765	798	795
Créteil	712	729	680	Reims	700	762	765
Dijon	757	757	783	Rennes	779	810	803
Grenoble	789	795	821	Rouen	785	765	743
Lille	761	716	742	Strasbourg	783	799	822
Limoges	719	709	742	Toulouse	815	830	775
Lyon	788	770	798	Versailles	813	823	798

TAB. 2

Une première analyse à l'aide de C.H.I.C. des règles entre les trois variables numériques S, L et ES nous conduit à constater l'absence de relation à un seuil acceptable à partir duquel une certaine dépendance existerait (>0 ,50). Une nouvelle voie s'ouvre alors : peut-**on dire que l'ordre des succès définitifs dans l'une des séries « implique » l'ordre dans une** autre et dans laquelle ? Pour tenter de répondre à cette question, il nous suffit de subdiviser chaque intervalle de la valeur minimale à la valeur maximale donnée aux 26 académies par le tableau en sous-intervalles. Nous adoptons la technique des nuées dynamiques de E. Diday qui permet, en particulier, de maximiser la variance inter-classe des sous-intervalles créés.

On obtient, à l'aide du logiciel C.H.I.C., les partitions optimales suivantes pour des partitions en 3 sous-intervalles :

S1 de 700 à 730	L1 de 709 à 742	ES1 de 680 à 746
S2 de 750 à 765	L2 de 754 à 779	ES2 de 761 à 785
S3 de 777 à 815	L3 de 795 à 830	ES3 de 795 à 822

TAB. 3

Puis utilisant la version vectorielle de CHIC, on extrait les implications suivantes :

$\vec{S}$ -> $\vec{L}$ avec l'intensité 0.63339 $\vec{L}$ -> $\vec{S}$ avec l'intensité 0.65041

$\vec{S}$ -> E$\vec{S}$ avec l'intensité 0.55874 E$\vec{S}$ -> $\vec{S}$ avec l'intensité 0.56572

$\vec{L}$ -> E$\vec{S}$ avec l'intensité 0.88015 E$\vec{S}$ -> $\vec{L}$ avec l'intensité 0.91037

que l'on peut synthétiser ainsi : $E\vec{S} \overset{0.91}{\Rightarrow} \vec{L}$, $\vec{L} \overset{0.65}{\Rightarrow} S$ et $E\vec{S} \overset{0..57}{\Rightarrow} \vec{S}$ ou par FIG 1 :

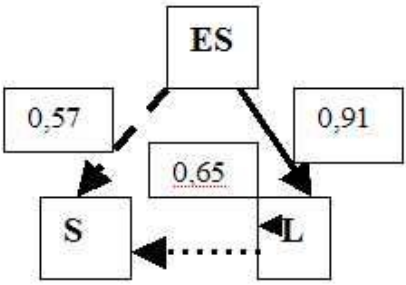

FIG. 1 – *Graphe implicatif des types de baccalauréats*

Ainsi, si les résultats au bac ES sont bons, alors ceux en L ont une propension à être meilleurs et, de façon moins décisive, ceux de S également. De plus, les résultats dans chaque académie en S ont tendance à être meilleurs, relativement à la tendance générale, à ceux de L et à en être prédicteurs dans l'ordre où ils apparaissent. Autrement dit, les académises où les résultats sont relativement faibles (resp. forts) en S, le sont aussi généralement en L et en ES.

On voit alors l'information complémentaire qu'apporte la méthode implicative par rapport au seul examen de la corrélation, voire de l'implication entre les colonnes. Elle dit dans quel sens et avec quelle intensité mesurable joue l'association entre les séries de baccalauréat.

4.3 Exemple 3

Nous disposons de « critères de convergence » entre États de l'Union européenne au 25-3-98, critères qui permettaient de comparer les économies des pays et d'en définir leur capacité à respecter la politique monétaire commune conditionnant l'ouverture à l'euro. Ces critères s'expriment selon plusieurs composantes : l'inflation, le déficit public (% du PIB), la dette publique (% du PIB) et le taux d'intérêt à long terme. Le tableau ci-dessous [4]donne les différentes valeurs observées sur 3 années : 1996, 1997 et 1998 sur les 15 pays de l'Union.

Nous avons encore affaire à des variables numériques : l'une pour 1996, une autre pour 1997, une troisième pour 1998. Les valeurs indiquées TAB. 4 ont été toutes ramenées à des nombres de l'intervalle [0,1] par division de chaque colonne par le maximum de la colonne. Seul le déficit a nécessité de translater les valeurs de telle façon que nous n'ayons que des nombres positifs, alors même que dans la plupart des pays, le déficit était toujours négatif.

[4] Tableau extrait du QUID 99, avec estimation des taux d'intérêt de la Grèce pour les trois années en question.

	96Infl	96défic	96dette	96Taux	97Infl	97défic	97dette	97Taux	98Infl	98défic	98dette	98Taux
Allemagne	0,22	0,41	0,48	0,62	0,35	0,23	0,5	0,7	0,38	0,1	0,52	0,75
Autriche	0,29	0,35	0,55	0,63	0,33	0,26	0,54	0,71	0,33	0,15	0,55	0,75
Belgique	0,27	0,43	1	0,65	0,29	0,33	1	0,73	0,29	0,31	1	0,76
Danemark	0,25	0,68	0,56	0,72	0,42	0,83	0,53	0,79	0,47	1	0,5	0,75
Espagne	0,4	0,29	0,55	0,87	0,46	0,25	0,56	0,8	0,49	0,18	0,57	0,84
Finlande	0,19	0,42	0,45	0,71	0,26	0,54	0,46	0,75	0,44	0,82	0,45	0,73
France	0,22	0,34	0,44	0,63	0,2	0,18	0,48	0,7	0,22	0	0,49	0,73
Grèce	1	0	0,88	1	1	0	0,89	1	1	0,18	0,91	1
Irlande	0,13	0,71	0,57	0,73	0,26	0,86	0,54	0,79	0,73	1	0,5	0,83
Italie	0,51	0,08	0,98	0,94	0,44	0,23	1	0,86	0,47	0,1	1	0,89
Luxembourg	0,19	1	0,05	0,63	0,26	1	0,06	0,7	0,36	1	0,06	0,75
Pays-Bas	0,15	0,52	0,61	0,62	0,4	0,46	0,59	0,7	0,51	0,33	0,59	0,73
Portugal	0,31	0,43	0,51	0,86	0,38	0,26	0,51	0,8	0,49	0,18	0,51	0,83
Royaume Uni	0,31	0,27	0,43	0,79	0,42	0,37	0,44	0,89	0,51	0,59	0,44	0,93
Suède	0,14	0,4	0,6	0,8	0,4	0,56	0,63	0,95	0,33	0,87	0,63	0,87

TAB. 4

Un premier traitement par C.H.I.C. des 12 variables numériques (4 pour chacune des 3 années) conduit au graphe implicatif suivant construit aux seuils .85 à .70:

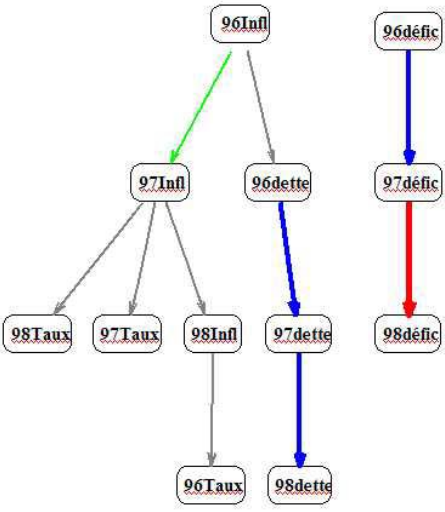

FIG. 2 – *Graphe implicatif des critères de convergence*

On observe une indépendance des 3 critères : inflation, dette et taux d'intérêt par rapport au critère déficit. Sur le plan économique, et cela paraît surprenant, le déficit n'aurait pas d'influence sur la dette et l'inflation. On constate, en revanche, la liaison attendue entre l'inflation et les taux d'intérêt, l'ensemble se présentant relativement peu lié à la dette.

On observe également qu'alors que déficit, inflation et dette croissent de 96 à 98, comme on pouvait l'attendre, les taux d'intérêt diminuent pendant cette période. Nous y reviendrons.

Arrêtant ici l'exploitation de ce graphe sur lequel bien d'autres considérations peuvent être énoncées, nous constatons que les natures des critères, bien qu'en partie liées, ne rendent pas bien compte du comportement global des économies de 96 à 98. D'où l'intérêt de considérer 3 variables vectorielles numériques : A code l'année 1996, présente les composantes A1, A2, A3 et A4 respectivement correspondant aux 4 critères, puis B code 1997, avec les composantes B1 à B4, et C code 1998 avec les composantes C1 à C4.

Le logiciel C.H.I.C. dégage les intensités d'implication suivantes :

$\varphi(A \Rightarrow B) = 0,71$; $\varphi(A \Rightarrow C)=0,70$; $\varphi(B \Rightarrow C)= 0,74$,

les intensités des relations réciproques étant plus faibles. Les trois variables s'enchaînent donc transitivement, au seuil 0,70, confirmant la croissance globale des valeurs des critères avec le temps, croissance freinée nécessairement par la décroissance des taux d'intérêt dont cette approche ne rend plus compte. Cette restriction illustre le propos tenu dans le dernier paragraphe du **§ 2**. La relation d'ordre dérivée de la sémantique des critères-composantes devrait être la même. Or pour les trois critères : inflation, déficit et dette, la sémantique de ces variables oriente l'évolution du critère vers la croissance, alors que pour le taux, dans les données, l'évolution est décroissante. Si nous complémentons, alors, à 1 chaque valeur du taux énoncé, nous obtenons des intensités d'implication supérieures aux précédentes, ce qui confirme le bien-fondé de la remarque au sujet des sémantiques d'ordre :

$\varphi(A \Rightarrow B) = 0,78$; $\varphi(A \Rightarrow C)=0,77$; $\varphi(B \Rightarrow C)= 0,79$,

5 Conclusion

Nous avons étendu la variété des variables prises en compte par la méthode d'analyse statistique implicative en conceptualisant la relation implicative entre deux variables vectorielles. De telles variables peuvent présenter des modalités binaires ou numériques représentées par les composantes vectorielles. Cette souplesse permet de dégager et quantifier une règle d'association dissymétrique entre deux vecteurs en exprimant par une mesure la dynamique inhérente aussi bien à l'ensemble des variables vectorielles qu'à celle de leurs modalités. Des exemples simples, traités à la main et des exemples réels traités par CHIC illustrent la construction théorique de l'extension de l'A.S.I. aux vecteurs.

Références

Agrawal, R., T. Imielinski. Et A. Swami (1993). Mining association rules between sets of items in large databases. In the1993 ACM SIGMOD international conference on management of data, ACM Press

Bernard J.-M., Poitrenaud S., (1999)L'analyse implicative bayesienne d'un questionnaire binaire : quasi- implications et treillis de Galois simplifié, *Mathématiques, Informatique et Sciences Humaines*, 147, 1999, 25-46

Couturier, R. et Gras R.,(2005) : CHIC : Traitement de données avec l'analyse implicative, *Extraction et Gestion des Connaissances, Volume I1, RNTI, Cepadues, Paris, p.679-684, , ISBN 2.85428.683.9*

Gras R. et al.(1996) *L'implication Statistique*, Grenoble, La Pensée Sauvage

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

Hilderman R.J., Hamilton H.J., (1999) Heuristics measures of efferestingness , *Proc. of the 3rd Eur. Conf. on Principles of Data Mining and Knowledge Discovery*, Lect. N. in Art. Int., 1999, p. 232-241.

Lagrange J.B. (1998) Analyse implicative d'un ensemble de variables numériques; application au traitement d'un questionnaire aux réponses modales ordonnées, *Revue de Statistique Appliquée XLVI-1*, Paris, I.H.P., 71-93

Lenca P., Meyer P., Vaillant B., Picouet P., Lallich S., (2004) Evaluation et analyse multi-critères des mesures de qualité des règles d'association, *Mesures de qualité pour la fouille de données, Collection RNTI-E-1, Cepaduès,219-246.*

Lerman, I.C., Gras R. et Rostam H., (1981) Elaboration et évaluation d'un indice d'implication pour données binaires , Mathématiques et Sc. Humaines, n°74, p. 5-35.

Pearl J., *Probabilistic Reasoning in intelligent systems*, San Mateo, CA, Morgan Kaufmann.

Régnier J.C. et R. Gras (2005). Statistique de rangs et analyse statistique implicative. *Revue de Statistique Appliquée*. 53 (1) : 5-38

Tan P., Kumar V., (2000) Interestingness measures for association patterns: a perspective, *Technical Report TR00-036, University of Minnesota, 2000.*

Summary

This text deals with statistical implicative analysis. As we have done for extending binary variables to numerical variables or interval-variables, we can extend the possible use of vectorial variables. We establish an index allowing us to measure the quality of a rule between vectorial variables. An exemple relative to first college degree and another concerning convergence criteria of European Union's economy are given.

.