UN ALGORITHME DE REGROUPEMENTS DE MODALITES DE VARIABLES DANS LE CADRE DE L'ANALYSE IMPLICATIVE DE DONNEES

D.LAHANIER-REUTER

RESUME

L'article développe un traitement statistique particulier dans le cadre de l'analyse implicative de données. Il expose un algorithme original de regroupements de modalités de variables qui permet de déchiffrer des lignes de force implicatives lorsque les modalités observées des variables étudiées sont en nombre élevé, par exemple lorsque l'une des variables est numérique.

MOTS -CLES

ANALYSE IMPLICATIVE ; ALGORITHME ; COMPARAISON DE DISTRIBUTIONS.

ABSTRACT

This paper develops a special statistical treatment consistent with the data implicative analysis theory. The presented algorithmic has been constructed for collecting together variable values. To exemplify this procedure we have chosen a didactical research.

KEY-WORDS

IMPLICATIVE ANALYSIS ; ALGORITHM ; DISTRIBUTIONS COMPARISON.

Equipe THEODILE

EA 1764

Université Charles-de-Gaulle Lille III, Domaine universitaire du « Pont de Bois »

BP 149

59653 Villeneuve d'Ascq cedex.

I. INTRODUCTION

Nous allons présenter ici un algorithme de regroupements de modalités observées de deux variables. Cet algorithme a pour but de mettre en évidence des lignes de force de type implicatif, au sens statistique du terme, dans le cas où les valeurs relevées des deux variables engagées sont soit numériques, soit en nombre élevé. En effet, la liste des implications statistiques entre différentes modalités des deux variables dans le cas évoqué est alors souvent difficile à interpréter.

Nous commencerons par développer l'exemple contextuel de recherche au cours duquel le problème de la gestion des implications statistiques entre données multiples est apparu de façon criante. Nous développerons ensuite l'une des solutions possibles à ce problème, en exposant la procédure algorithmique élaborée tout en la faisant fonctionner sur l'exemple contextuel choisi.

II. DESCRIPTION DE LA SITUATION EXPERIMENTALE ET INTERET DU RECOURS A L'ANALYSE IMPLICATIVE.

La situation de recherche contextuelle qui va servir de cadre initial au développement algorithmique est une situation rencontrée lors d'un travail de thèse sous la direction de Régis Gras.

Travaillant sur les conceptions locales et pragmatiques du hasard, nous étions parvenue à reconstruire trois conceptions différentes du hasard. Ces trois conceptions nous paraissaient relativement éloignées les unes des autres et nous les avions désignées respectivement sous les dénominations suivantes :

- le hasard de l'aléatoire ;
- le hasard de l'événement exceptionnel et improbable ;
- le hasard des deux possibles[1].

Nous cherchions alors à savoir si ces différentes conceptions étaient également mobilisées par des sujets d'âge et de parcours scolaires divers. Afin d'apporter des

[1] D. Lahanier- Reuter, *Etude de conceptions du hasard : approche épistémologique, didactique et expérimentale en milieu universitaire*, Thèse de doctorat, Université de Rennes I, 1998.

éléments de réponse à cette question, nous avons recueilli des productions écrites de cent deux sujets. Les sujets interrogés pouvaient être différenciés par leur âge et par leur niveau de scolarité. En effet, parmi ces cent deux sujets, on comptait dix huit élèves d'une classe de CM1/CM2, âgés de 9 à 10 ans, puis vingt quatre élèves d'une classe de terminale ES[2] , âgés de 17 à 20 ans, et enfin soixante étudiants de licence de Sciences de l'éducation, âgés de 20 à 45 ans. Chacun des sujets considéré se caractérise en conséquence par son appartenance à l'un des trois groupes : CM pour les élèves de CM1/CM2, TES pour les élèves de la classe de terminale, Etudiant pour les étudiants de licence.

En ce qui concerne la tâche d'écriture évoquée plus haut, il s'agissait de composer huit petits textes, à partir d'illustrations, avec pour seule contrainte d'utiliser le mot hasard.

Chacun des huit textes composés par un sujet de cette population a été classé - quand cela s'avérait possible - selon la conception du hasard que nous pensions y déchiffrer. Nous avons ainsi, pour la plupart des sujets de la population, pu leur faire correspondre un triplet, constitué du nombre de textes s'inscrivant dans la première conception, puis du nombre de ceux s'inscrivant dans la seconde et enfin de ceux relevant selon nous de la troisième et dernière conception reconstruite.

Le problème qui se posait alors était de savoir si, au cours de cette situation particulière, les sujets interrogés mobilisaient uniquement l'une de ces conceptions ou, au contraire, avaient plutôt tendance à diversifier leurs positions. Il était possible en effet d'obtenir aussi bien des triplets du type (8,0, 0) que du type (3,2,3). Le premier triplet indiquait alors pour nous un sujet que nous qualifions de « rigide », c'est-à-dire un sujet qui avait, au long des huit textes mobilisé une seule des trois conceptions reconstruites, dans ce cas celle du hasard aléatoire. En revanche, le second triplet (3,2,3) caractérisait, toujours selon nous, une certaine « souplesse » du sujet, puisqu'il rendait compte de la diversité des positions adoptées par le sujet au long de l'expérience.

[2] La classe de terminale ES est une option « économique et sociale ». Le programme de mathématiques de cette section est le plus riche en statistiques, comparativement à celui des autres sections.

Pour rendre compte de ce qu'il est possible de dénommer « rigidité » ou « souplesse » de ces sujets, nous avons élaboré alors une variable numérique qui permet de hiérarchiser les différents triplets obtenus : c'est l'entropie d'un triplet de somme constante égale à 8 (voir tableau 1). Une entropie faible traduit alors une rigidité des choix, tandis qu'une entropie élevée est plutôt à lire comme une souplesse des choix effectués par le sujet.

Tableau 1 : Entropie des différents triplets de somme 8.

Triplets[3]	8,0,0	7,1,0	6,2,0	5,3,0	4,4,0	6,1,1	5,2,1	4,3,1	4,2,2	3,3,2
Entropie	0.000	0.543	0.813	0.954	1.000	1.061	1.298	1.405	1.500	1.561

En définitive, en interrogeant le lien éventuel que l'on pouvait établir entre la qualité de la mobilité des choix du sujet (souplesse vs rigidité) et leur niveau scolaire, nous avons été amenée à définir des relations entre modalités observées d'une variable numérique - l'entropie - et celle d'une variable qualitative - le niveau scolaire - et à tenter d'en mesurer l'intensité. Le type d'analyse qui nous a alors semblé le plus pertinent est celui de l'analyse implicative de données et nous allons développer les raisons de ce choix.

III. RECOURS A L'ANALYSE IMPLICATIVE DE DONNEES.

Voici le tableau initial (tableau 2) de répartitions des 102 sujets selon l'entropie de leurs choix et leur situation scolaire.

Tableau 2 : Distributions de l'entropie selon les catégories de sujets

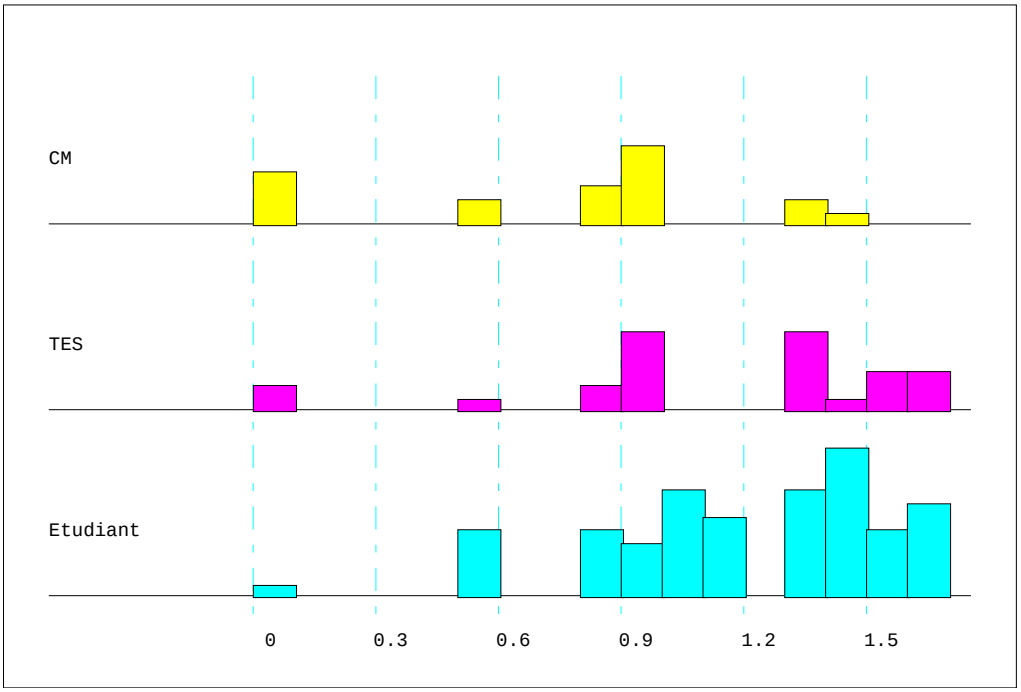
[3] Les triplets sont à lire aux permutations près.

Niveau scolaire Entropie	CM	TES	Etudiants	Totaux
0.000	4	2	1	7
0.543	2	1	5	8
0.813	3	2	5	10
0,.954	6	6	4	16
1.000	0	0	8	8
1.061	0	0	6	6
1.298	2	6	8	16
1.405	1	1	11	13
1.500	0	3	5	8
1.561	0	3	7	10
Totaux	18	24	60	102

Une simple lecture des données, s'attachant à l'absence de sujets, dans les catégories CM et TES, d'entropie 1,000 et 1,061, indique un déséquilibre presque symétrique des répartitions de l'entropie des élèves de CM et des étudiants. Ceci suggère un partage des élèves de CM en un groupe d'effectif relativement important et d'entropie faible et un groupe d'effectif peu important et d'entropie élevée, un partage équilibré des élèves de TES en deux groupes d'effectifs comparables d'entropie faible et d'entropie élevée et enfin une répartition des étudiants également en deux groupes, le premier d'entropie faible et d'effectif peu important, le second d'entropie élevée et rassemblant une large majorité de cette sous population (voir figure 1). Les élèves de CM seraient alors caractérisés par des choix rigides au contraire des étudiants dont les choix seraient plus souples.

Figure 1 : Distribution de l'entropie selon le statut scolaire des sujets[4]

[4] Histogramme réalisé à l'aide du logiciel ADSO : A. Dubus, ADSO 3, Trois-Monts, Trigone, CEDEP, 1995.



Evaluer l'influence de l'appartenance à un sous groupe de la population sur la distribution correspondante de l'entropie peut donner effectivement lieu à des traitements statistiques divers. Nous évoquerons ici uniquement l'analyse de la variance.

A. ANALYSE DE LA VARIANCE

L'analyse de la variance révélerait une absence significative d'homogénéité des variances inter-groupes et intra-groupe au seuil 1% (F = 8.899, s. à .01). Cependant, ce résultat ne peut être pris en compte, puisque, si les distributions peuvent effectivement être considérées comme normales[5], les variances des distributions de l'entropie selon le groupe de CM et celui des étudiants ne peuvent être considérées comme homogènes[6] au seuil de risque 5%. Ces calculs confirment certes les descriptions initiales. Mais, même si l'analyse de la variance avait été recevable, elle ne nous aurait permis d'établir que des comparaisons (les élèves de CM sont des sujets plus rigides dans leurs choix que ne les sont les étudiants) sans nous indiquer des caractérisations d'entropies attribuées aux catégories de sujets ni permettre une définition commune à ces sous populations de ce que représente - dans le contexte expérimental uniquement - une entropie

[5] Le test de Kolmogorov donne respectivement pour la distribution de l'entropie des CM : 0.172, des Tes 0.140 et pour celle des étudiants : 0.113.

élevée et une entropie faible. Il ne permet pas, par exemple, de définir ou d'approcher une valeur de l'entropie qui serait une valeur « seuil », séparant statistiquement les entropies observées sur le groupe des élèves de CM des entropies observées sur le groupe des étudiants.

B. CHOIX DE L'ANALYSE IMPLICATIVE DE DONNEES.

Par conséquent, l'outil statistique d'analyse de données que nous avons retenu est celui de l'implication statistique[7]. En effet, non seulement ce type d'analyse ne requiert aucune hypothèse quant à la distribution de la variable numérique, mais encore fournit une signification aux éléments descriptifs préétablis, singuliers à chacune des distributions étudiée séparément. Mais l'établissement de la liste des liens implicatifs significatifs est ici rapide, car elle est vide[8]. Et il est vrai que cette liste est, le plus souvent, peu exploitable, car elle est rarement lisible dès que le nombre de modalités observées de l'une ou l'autre des variables devient important.

Ces remarques nous ont poussée à tenter des regroupements de la variable entropie dans l'espoir de voir apparaître des liens statistiques implicatifs possédant une intensité intéressante. Deux possibilités sont envisageables : soit tester les regroupements pour l'ensemble des sous-populations, soit les tester sous-population par sous-population. Puisque le traitement algorithmique peut être transposé facilement d'un cas à l'autre, nous exposerons tout d'abord l'algorithme général, qui prend en charge le tableau initial dans sa complexité. Nous reviendrons ensuite sur l'autre possibilité de traitement.

IV. DESCRIPTION DE L'ALGORITHME ET MISE EN ŒUVRE SUR L'EXEMPLE TRAITE.

---

[6] F = 1.771, pour un couple (17,59).

[7] R. Gras, *L'implication statistique, nouvelle méthode exploratoire de données*, La Pensée Sauvage Editions, 1996.

[8] Dans d'autres cas, elle pourrait être illisible parce que trop longue. Nous avons rencontré ce cas en particulier lors de l'étude de croisement de variables numériques : D. Lahanier - Reuter, «Exemple d'une nouvelle méthode d'analyse de données : l'analyse implicative,», *Carrefours de l'éducation*, 2000.

A. FINALITE DE L'ALGORITHME.

Soit f et g deux variables. (Dans l'exemple traité, f est la variable « statut » et g la variable « entropie »).

On suppose que le nombre de modalités initiales observées de f et de g est au moins égal à 2.

On désigne par $a_1, a_2,\ldots a_i,\ldots$ les modalités de f. Leur nombre initial est de n_0.

On désigne par $b_1, b_2,\ldots b_j,\ldots$ les modalités de g. Leur nombre initial est de m_0.

La finalité de l'algorithme est de réduire au maximum le nombre des valeurs de g et de f tout en augmentant l'intensité des liens implicatifs significatifs entre valeurs de g et valeurs de f. Cet algorithme a ainsi pour but d'exhiber les regroupements de **taille maximale** des modalités observées de g et de f, pour lesquels il existe une implication statistique d'intensité maximale, du type : $\bigcup_{\substack{1\le i\le n_0\\ 0\le l\le n_0-i}} \{a_{i+l}\} \Rightarrow \bigcup_{\substack{1\le j\le m_0\\ 0\le k\le m_0-1}} \{b_{j+k}\}$, où les regroupements effectués ne se chevauchent pas.

Nous présenterons l'algorithme dans le cas où la réduction recherchée est celle des valeurs de g.

B. LISTE DES CONTRAINTES

Les contraintes qui vont guider la mise en place de cet algorithme sont les suivantes :

1. Seuls les regroupements de modalités connexes de l'entropie, en tant que variable numérique sont à envisager.
2. Seuls les regroupements n'ayant aucun élément commun pourront être réalisés simultanément.
3. Le gain ou la perte occasionnée par le nouveau regroupement sera déterminé en fonction de l'intensité des nouveaux liens implicatifs. En particulier, l'élaboration de l'algorithme s'appuie sur l'opération suivante : l'intérêt du regroupement de k (pour l compris entre 1 et m_0) modalités b_l *et* $b_{l+1}, \ldots b_{l+k-1}$ de g dépend de la comparaison de l'intensité des liens implicatifs associés respectivement aux implications $a_i \Rightarrow b_l$ et $a_i \Rightarrow b_{l+1}$....et $a_i \Rightarrow b_{l+k-1}$ à l'intensité de l'implication statistique $a_i \Rightarrow (b_l$ ou b_{l+1} ....ou $b_{l+k-1})$.

Ainsi, l'intérêt du regroupement de k modalités b_l, b_{l+1}, ...b_{l+k-1} de g est-il lié :

- pour toutes les modalités de f, soit pour toutes les valeurs de i comprises entre 1 et n_0, aux signes des indices[9] d'implication $q_{i,k,l}$ associés aux implications statistiques $a_i \Rightarrow (b_l$ ou b_{l+1} ...ou $b_{l+k-1})$. Un signe négatif traduit une intensité d'implication supérieure à 0.5.
- pour toutes les modalités de f, soit pour toutes les valeurs de i comprises entre 1 et n_0, à la comparaison de ces nouveaux indices d'implication $q_{i,k,l}$ aux indices d'implication $q_{i,l}$, $q_{i,l+1}$,...$q_{i,l+k-1}$ respectivement associés aux implications $a_i \Rightarrow b_l$, $a_i \Rightarrow b_{l+1}$, ...$a_i \Rightarrow b_{l+k-1}$. Toute augmentation de l'intensité correspond à une diminution de l'indice d'implication.

Par conséquent, l'expression $F_{i,k,l} = \left| q_{i,k,l} \right| (\inf(q_{i,l}, q_{i,\,l+1},...q_{i,l+k-1}) - q_{i,k,l})$ est positive si le regroupement proposé traduit un gain d'intensité implicative, négative dans le cas contraire.

En conséquence, le gain que représente le regroupement des k valeurs de g sur l'ensemble des modalités de f est lié au signe de l'expression :

$$F_{k,l} = \sup_{1 \le i \le n_0} \left| q_{i,j,l} \right| (\inf(q_{i,1}, q_{i,l+1}, \ldots q_{i,l+k-1}) - q_{i,k,l})$$

4.De même, le gain ou la perte occasionnée par le nouveau regroupement sera déterminé en fonction de la perte d'information que représente ce regroupement, c'est-à-dire, puisqu'il s'agit toujours de modalités numériques, en fonction de l'amplitude de l'intervalle que représente ce nouveau regroupement. Pour mesurer l'intérêt d'un regroupement de k valeurs connexes de g, on prendra en compte l'indicateur amplitude de l'intervalle [b_l, b_{l+k-1}] associé au regroupement $\{b_l, b_{l+1}, \ldots, b_{l+k-1}\}$, si les modalités b_j sont des valeurs discrètes. En revanche, si les

---

[9] Nous rappelons que l'indice q associé à l'implication a ⇒ b, est égal à $\frac{n_{a\wedge\bar{b}} - \frac{n_a n_{\bar{b}}}{n}}{\sqrt{\frac{n_a n_{\bar{b}}}{n}}}$ où n_a est l'effectif observé de la modalité a et où $n_{\bar{b}}$ est celui du complémentaire de la

b_j sont des intervalles de $\Re$, on prend en compte la différence entre les milieux des intervalles b_l et b_{l+k-1}.

L'indice d'intérêt final associé à un regroupement $b_l, b_{l+1},...b_{l+k-1}$ est donc :

$$I_{k,l} = \frac{F_{k,l}}{(b_{l+k-1} - b_l)}$$

Ainsi, $F_{k,l}$ étant donné, $I_{k,l}$ sera d'autant plus grand que l'amplitude sera faible.

C. DESCRIPTION DE L'ALGORITHME

Chaque étape de l'algorithme correspond à un essai de réduction d'une liste des valeurs de g, c'est-à-dire à un essai de réduction du nombre des lignes (ou des colonnes) d'un tableau croisé.

Chaque étape de l'algorithme se définit ainsi par la donnée d'un nombre **m** de valeurs distinctes de g : b_1, b_2, ...b_m. ($m \leq m_0$)

Sont associés à cette liste initiale, pour $1 \leq i \leq n_0$, $1 \leq j \leq m$:

- Les effectifs $E_{i,j}$ correspondants à ces différentes valeurs, au sein de la population ;
- Les indices des implications statistiques $q_{i,j}$ associés à chacune des implications statistiques $a_i \Rightarrow b_j$.

Pour toutes les valeurs successives de **k**, pour $2 \leq k \leq m$, l'intérêt (quant au gain en intensité implicative) de chacun des m-k+1 regroupements de k valeurs connexes de g parmi les m valeurs de g va être testé.

Pour $2 \leq k \leq m$, pour $1 \leq l \leq m-k+1$, chaque regroupement des m-k+1 valeurs de g : b_l, b_{l+1},...b_{l+k-1} sera noté $B_{k,l}$. Pour chacun de ces regroupements on calcule alors l'indicateur numérique $I_{k,l}$.

Si $I_{k,l}$ est négatif ou nul, le regroupement $B_{k,l}$ est déclaré non intéressant et n'est pas à considérer.

1) Si, à k fixé, les regroupements $B_{k,l}$ sont tous déclarés inintéressants, la valeur de k est incrémentée, tant que k est inférieur à m. Si k = m l'algorithme s'arrête.

---

modalité b. L'intensité φ d'implication est alors $\frac{1}{\sqrt{2\pi}} \int_q^{\infty} e^{-\frac{t^2}{2}} dt$.

2) Sinon, les indices $I_{k,l}$ positifs sont ordonnés. Les regroupements $B_{k,l}$ d'indices positifs sont effectués, dans l'ordre décroissant des $I_{k,l}$, à condition que ceci soit possible, c'est-à-dire :

- que le regroupement à effectuer soit compatible avec les regroupements déjà effectués ;
- que deux regroupements de même indice à effectuer puissent l'être simultanément.

Si aucun des regroupements intéressants n'est réalisable, la valeur de k est incrémentée tant que k est inférieur à m. Si k = m, l'algorithme s'arrête.

Sinon, les p regroupements $B_{k,l}$ de k valeurs intéressantes et réalisables sont désormais identifiés par les valeurs numériques β_l comme suit :

$$\beta_l = \frac{(b_l + b_{l+1} + \ldots . b_{l+k-1})}{k}$$

Les p valeurs numériques β_l viennent se substituer aux p x k valeurs de g ainsi regroupées. Le nombre de valeurs distinctes de g devient m-pk. Une nouvelle étape de l'algorithme est amorcée, tant que le nombre de valeurs distinctes de g reste supérieure à 1. Dans le cas contraire, l'algorithme s'arrête.

D. MISE EN ŒUVRE DE L'ALGORITHME SUR LE CAS CONSIDERE.

Dans l'exemple présenté, les variables f et g sont les variables « niveau scolaire » et « entropie ». Les modalités de f sont au nombre de trois, celles de g sont initialement au nombre de dix : $n_0 = 3$; $m_0 = 10$.

La première étape de l'algorithme suppose donc **m** = 10. La liste initiale des 10 valeurs de g, ainsi que, pour $1 \leq i \leq 3$ et $1 \leq j \leq 10$, les valeurs des effectifs $E_{i,j}$ et des indices $q_{i,j}$ se lit sur le tableau suivant :

Tableau 3 : fonctionnement de l'algorithme
Initialisation.

Niveau scolaire Entropie	CM	TES	Etudiants

$.b_1 = 0.000$	$E_{1,1} = 4$ $q_{1,1} = -0,675$	$E_{2,1} = 2$ $q_{2,1} = -0,075$	$E_{3,1} = 1$ $q_{3,1} = 0,417$
$.b_2 = 0.543$	$E_{1,2} = 2$ $q_{1,2} = -0,144$	$E_{2,2} = 1$ $q_{2,2} = 0,187$	$E_{3,2} = 5$ $q_{3,2} = -0,040$
$.b_3 = 0.813$	$E_{1,3} = 3$ $q_{1,3} = -0,307$	$E_{2,3} = 2$ $q_{2,3} = 0,076$	$E_{3,3} = 5$ $q_{3,3} = 0,120$
$.b_4 = 0.954$	$E_{1,4} = 6$ $q_{1,4} = -0.815$	$E_{2,4} = 6$ $q_{2,4} = -0.497$	$E_{3,4} = 4$ $q_{3,4} = 0.761$
$.b_5 = 1.000$	$E_{1,5} = 0$ $q_{1,5} = 0,347$	$E_{2,5} = 0$ $q_{2,5} = 0,400$	$E_{3,5} = 8$ $q_{3,5} = -0,443$
$.b_6 = 1.061$	$E_{1,6} = 0$ $q_{1,6} = 0,257$	$E_{2,6} = 0$ $q_{2,6} = 0,297$	$E_{3,6} = 6$ $q_{3,6} = -0,329$
$.b_7 = 1.298$	$E_{1,7} = 2$ $q_{1,7} = 0,211$	$E_{2,7} = 6$ $q_{2,7} = -0,497$	$E_{3,7} = 8$ $q_{3,7} = 0,198$
$.b_8 = 1.405$	$E_{1,8} = 1$ $q_{1,8} = 0,327$	$E_{2,8} = 1$ $q_{2,8} = 0,450$	$E_{3,8} = 11$ $q_{3,8} = -0,463$
$.b_9 = 1.500$	$E_{1,9} = 0$ $q_{1,9} = 0,347$	$E_{2,9} = 3$ $q_{2,9} = -0,238$	$E_{3,9} = 5$ $q_{3,9} = -0,040$
$.b_{10} = 1.561$	$E_{1,10} = 0$ $q_{1,10} = 0,438$	$E_{2,10} = 3$ $q_{2,10} = -0,139$	$E_{3,10} = 7$ $q_{3,10} = -0,152$

Pour la première étape, k vaudra tout d'abord 2. Nous allons donc présenter les 9 regroupements, notés $B_{2,l}$ pour $1 \leq l \leq 9$, des 10 valeurs initiales de l'entropie, deux à deux, ainsi que les indicateurs $F_{i,j,l}$, $F_{k,l}$ et $I_{2,l}$.

Tableau 4 : fonctionnement de l'algorithme : test de l'intérêt du regroupement des premières valeurs de l'entropie.

Niveau scolaire Entropie	CM	TES	Etudiants	$F_{k,l}$ $I_{k,l}$

$B_{2,1} = \{b_1\} \cup \{b_2\}$	$q_{1,2,1}$=-0.856 $F_{1,2,1}$=0.154	$q_{2,2,1}$=0.117 $F_{2,2,1}$=-0.022	$q_{3,2,1}$=0.394 $F_{3,2,1}$=-0.171	$F_{2,1}$=0.154 $I_{2,1}$=0.284
$B_{2,2} = \{b_2\} \cup \{b_3\}$	$q_{1,2,2}$=-0.474 $F_{1,2,2}$=0.079	$q_{2,2,2}$=0.276 $F_{2,2,2}$=-0.056	$q_{3,2,2}$=0.084 $F_{3,2,2}$=-0.010	$F_{2,2}$=0.079 $I_{2,2}$=0.293
$B_{2,3} = \{b_3\} \cup \{b_4\}$	$q_{1,2,3}$=-1.205 $F_{1,2,3}$=0.469	$q_{2,2,3}$=-0.445 $F_{2,2,3}$=-0.023	$q_{3,2,3}$=0.941 $F_{3,2,3}$=-0.773	$F_{2,3}$=0.469 $I_{2,3}$=3.326
$B_{2,4} = \{b_4\} \cup \{b_5\}$	$q_{1,2,4}$=-0.476 $F_{1,2,4}$=-0.162	$q_{2,2,4}$=-0.082 $F_{2,2,4}$=-0.034	$q_{3,2,4}$=0.313 $F_{3,2,4}$=-0.236	$F_{2,4}$=-0.034 $I_{2,4}$=-0.742
$B_{2,5} = \{b_5\} \cup \{b_6\}$	$q_{1,2,5}$=0.627 $F_{1,2,5}$=-0.232	$q_{2,2,5}$=0.724 $F_{2,2,5}$=-0.309	$q_{3,2,5}$=-0.801 $F_{3,2,5}$=0.287	$F_{2,5}$=0.287 $I_{2,5}$=4.705
$B_{2,6} = \{b_6\} \cup \{b_7\}$	$q_{1,2,6}$=0.501 $F_{1,2,6}$=-0.145	$q_{2,2,6}$=-0.190 $F_{2,2,6}$=-0.058	$q_{3,2,6}$=-0.154 $F_{3,2,6}$=-0.027	$F_{2,6}$=-0.027 $I_{2,6}$=-0.114
$B_{2,7} = \{b_7\} \cup \{b_8\}$	$q_{1,2,7}$=0.590 $F_{1,2,7}$=-0.223	$q_{2,2,7}$=-0.043 $F_{2,2,7}$=-0.019	$q_{3,2,7}$=-0.296 $F_{3,2,7}$=-0.050	$F_{2,7}$=-0.019 $I_{2,7}$=-0.181
$B_{2,8} = \{b_8\} \cup \{b_9\}$	$q_{1,2,8}$= 0.716 $F_{1,2,8}$=-0.279	$q_{2,2,8}$=0.216 $F_{2,2,8}$=-0.098	$q_{3,2,8}$=-0.528 $F_{3,2,8}$=0.034	$F_{2,8}$=0.034 $I_{2,8}$=0.361
$B_{2,9} = \{b_9\} \cup \{b_{10}\}$	$q_{1,2,9}$=0.825 $F_{1,2,9}$=-0.395	$q_{2,2,9}$=-0.397 $F_{2,2,9}$=0.063	$q_{3,2,9}$=-0.201 $F_{3,2,9}$=0.010	$F_{2,9}$=0.063 $I_{2,9}$=1.037

Par conséquent, six valeurs $I_{k,l}$ seulement sont positives : $I_{2,1}$, $I_{2,2}$, $I_{2,3}$, $I_{2,5}$, $I_{2,8}$, $I_{2,9}$. Ces valeurs ordonnées par ordre décroissant sont : $I_{2,5}$ (4,705), $I_{2,3}$ (3,326), $I_{2,9}$ (1,037), $I_{2,8}$(0,361), $I_{2,2}$(0,293) et $I_{2,1}$ (0,284). Par conséquent seulement quatre regroupements sont possibles à effectuer : $B_{2,5}$, $B_{2,3}$, $B_{2,9}$ et $B_{2,1}$.

Les nouvelles valeurs prises par g sont au nombre de 6 : b_1=(0.000+0.543)/2, b_2= (0.543+0.813)/2, b_3=(1.000+1.061)/2, b_4=1.298, b_5=1.405, b_6=(1.500+1.561)/2.

On continue, en reprenant k=2.

Soit

Tableau 5 : deuxième pas

Niveau scolaire Entropie	CM	TES	Etudiants

.b_1 = 0.2715	$E_{1,1}$ = 6 $q_{1,1}$ =-0.856	$E_{2,1}$ = 3 $q_{2,1}$= 0.117	$E_{3,1}$ = 6 $q_{3,1}$ = 0,395
.b_2 = 0.678	$E_{1,2}$ = 9 $q_{1,2}$ = -1,205	$E_{2,2}$ = 8 $q_{2,2}$ = -0,445	$E_{3,2}$ = 9 $q_{3,2}$= 0,941
.b_3 = 1.0305	$E_{1,3}$ = 0 $q_{1,3}$ = 0,627	$E_{2,3}$ = 0 $q_{2,3}$ = 0,724	$E_{3,3}$ = 14 $q_{3,3}$ = -0,801
.b_4 = 1.298	$E_{1,4}$ = 2 $q_{1,4}$ = 0,211	$E_{2,4}$ = 6 $q_{2,4}$ = -0.497	$E_{3,4}$ = 8 $q_{3,4}$ = 0,198
.b_5 = 1.405	$E_{1,5}$ = 1 $q_{1,5}$ = 0,327	$E_{2,5}$ = 1 $q_{2,5}$ = 0,450	$E_{3,5}$ = 11 $q_{3,5}$ = -0,463
.b_6 = 1.5305	$E_{1,6}$ = 0 $q_{1,6}$ = 0,825	$E_{2,6}$ = 6 $q_{2,6}$ = -0,397	$E_{3,6}$ = 12 $q_{3,6}$ = -0,201

Les regroupements deux à deux sont de nouveau testés :

Niveau scolaire Entropie	CM	TES	Etudiants	$F_{k,l}$ $I_{k,l}$
$B_{2,1} = \{b_1\} \cup \{b_2\}$	$q_{1,2,1}$=-2,367 $F_{1,2,1}$=2,750	$q_{2,2,1}$=-0,357 $F_{2,2,1}$=-0.031	$q_{3,2,1}$=1,522 $F_{3,2,1}$=-1,716	$F_{2,1}$=2,750 $I_{2,1}$=3,382
$B_{2,2} = \{b_2\} \cup \{b_3\}$	$q_{1,2,2}$=-0.488 $F_{1,2,2}$=-0,350	$q_{2,2,2}$=0.501 $F_{2,2,2}$=-0,474	$q_{3,2,2}$=-0.050 $F_{3,2,2}$=-0,037	$F_{2,2}$=-0,037 $I_{2,2}$=-0,053
$B_{2,3} = \{b_3\} \cup \{b_4\}$	$q_{1,2,3}$=0,924 $F_{1,2,3}$=-0,659	$q_{2,2,3}$=0,257 $F_{2,2,3}$=-0,194	$q_{3,2,3}$=-0,669 $F_{3,2,3}$=-0,089	$F_{2,3}$=-0,089 $I_{2,3}$=-0,165
$B_{2,4} = \{b_4\} \cup \{b_5\}$	$q_{1,2,4}$=0,590 $F_{1,2,4}$=-0,223	$q_{2,2,4}$=-0,043 $F_{2,2,4}$=-0,019	$q_{3,2,4}$=-0,296 $F_{3,2,4}$=-0,050	$F_{2,4}$=-0,019 $I_{2,4}$=-0,090
$B_{2,5} = \{b_5\} \cup \{b_6\}$	$q_{1,2,5}$=1,263 $F_{1,2,5}$=-1,183	$q_{2,2,5}$=0.072 $F_{2,2,5}$=-0,034	$q_{3,2,5}$=-0,737 $F_{3,2,5}$=0.202	$F_{2,5}$=0,202 $I_{2,5}$=0,804

Ainsi, deux regroupements vont être effectués à l'issue de cette étape : $B_{2,1}$ et $B_{2,5}$. On obtient enfin (voir tableaux 6 et 6bis), une partition des valeurs prises par l'entropie à deux classes. Sur chacun de ces intervalles, l'intensité q de

l'implication statistique entre la catégorie « CM » d'une part et « étudiant » d'autre part satisfait aux exigences émises (seuil de risque inférieur à .10).

Tableau 5 : Regroupements finaux des valeurs de l'entropie
Indices d'implications statistiques entre catégories de sujets et intervalles d'entropie.

Entropie	0,000 à 0,954	1,000 à 1,561
CM	**$E_{1,1}$= 15** **$q_{1,1}$ =-2,367**	$E_{1,2}$=3 $q_{1,2}$= 2,886
TES	$E_{2,1}$=11 $q_{2,1}$= -0,357	$E_{2,2}$=13 $q_{2,2}$= 0,436
Etudiant	$E_{3,1}$=15 $q_{3,1}$=1,522	**$E_{3,2}$=45** **q=-1,856**

Tableau 5 bis : Intensités des implications statistiques entre catégories de sujets et regroupements des valeurs de l'entropie.

Entropie	0,000 à 0,954	1,000 à 1,561
CM	**99%**	ns
TES	ns	ns
Etudiant	ns	**96%**

L'algorithme permet ainsi d'opposer deux groupes de sujets : les élèves de CM et les étudiants de licence. Les élèves de CM seraient caractérisés par une entropie faible, c'est-à-dire par des choix très rigides, au contraire des étudiants qui seraient quant à eux caractérisés par de choix plus diversifiés. De plus, ces opérations permettent de donner une définition - ponctuelle - de ce qu'est une entropie faible : c'est une entropie inférieure à 1, et de ce qu'est une entropie forte : une entropie supérieure ou égale à 1. En définitive, les élèves de CM seraient caractérisés par des choix recouvrant au maximum deux des trois

conceptions du hasard reconstruites, tandis que les étudiants seraient au contraire caractérisés par des choix englobant les trois mêmes conceptions (voir tableau 1). L'algorithme a été ensuite mené sur les valeurs de f , mais aucun regroupement ne s'avère intéressant.

V. CONCLUSION : AUTRE FONCTIONNEMENT POSSIBLE DE L'ALGORITHME

Il est aussi possible de rechercher, comme nous l'avons signalé au début de l'exposé opérationnel, de faire fonctionner cet algorithme, catégorie de sujet par catégorie de sujet, c'est-à-dire par valeur fixée de f. Dans ce cas, on recherche par exemple, quel regroupement de l'entropie conduit à des implications statistiques du type : « CM »$\Rightarrow$ « une entropie comprise entre a_1 et a_n », avec une intensité intéressante, puis quels sont ceux qui permettent d'écrire des implications statistiques du type : « TES » $\Rightarrow$ « entropie comprise entre a_i et a_m » etc.

Le traitement permet d'accéder ainsi à d'autres informations, qui risqueraient de passer inaperçues, dans le cas du traitement global. Mais rien n'assure que les regroupements intéressants retenus en dernière instance se recoupent. La mise en œuvre de l'algorithme sert moins dans ce cas à exhiber des oppositions entre groupes de sujets qu'à transcrire des lignes de force implicatives.

BIBLIOGRAPHIE

Gras R., 1996, *L'implication statistique, nouvelle méthode exploratoire de données*, La Pensée Sauvage Editions, Grenoble.

Gras R., Lahrer A.,1993, « L'implication statistique, une nouvelle méthode d'analyse des données », *Mathématiques, Informatique et Sciences Humaines* n°120, pp.5-31.

Hoel P. ,1991, *Statistique mathématique, tome II,* Armand Colin, Paris,.

Lahanier - Reuter D.,1998, *Etude de conceptions du hasard : approche épistémologique, didactique et expérimentale en milieu universitaire*, Thèse de doctorat, Université de Rennes I.

Lahanier - Reuter Dominique, 1999, *Conceptions du hasard et enseignement des probabilités et des statistiques*, P.U.F., Paris.

Lahanier - Reuter D., 2000, « Exemple d'une nouvelle méthode d'analyse de données : l'analyse implicative », *Carrefours de l'éducation.*

Lerman I.C., Gras R., Rostam H., 1981, « Elaboration et évaluation d'un indice d'implication pour les données binaires », *Mathématiques, Informatique et Sciences Humaines*, n°74, pp. 5-35.