Analyse informationnelle de banques de données en vue d'une exploitation rapide

par

Gérard COLSON
Albert CORHAY
Université de Liège - BELGIQUE

Abstract : L'analyse informationnelle est utilisée comme un outil descriptif des données chiffrées d'une banque quelconque de données. L'exploration progressive par programmes interactifs a pour objectif de mettre en évidence les faits saillants, les changements et les distances entre structures, et des degrés de dépendance.
Les mesures utilisées de la théorie mathématique de l'information sont l'entropie, l'espérance de gain en information et l'information mutuelle. L'ensemble des programmes ne se substitue pas à une analyse précise et détaillée des données, mais peut éventuellement la focaliser.

1. INTRODUCTION

Invités par des linguistes, nous avons à coeur de commencer cet exposé par une analyse de contenu du titre de cet article.

1.1. L'analyse informationnelle multivariée (A.I.M.) est une analyse de données catégorielles (fréquentielles) quantitatives positives s'appuyant sur des mesures issues de la théorie mathématique de l'information. Elle présente deux parties : l'analyse descriptive et la partie inférentielle.

En tant qu'outil descriptif des données, l'A.I.M. a déjà suscité de nombreux travaux dans les domaines les plus divers. A titre d'exemples, citons Mc DONALD (1952, taxonomie), VAN SOEST (1954, sociologie), ASHBY (1965, cybernétique), ATTNEAVE (1959, psychométrie), THEIL (1967, économie), THEIL (1969), LEV (1969), DUMAS (1976), COLSON (1973, 1977) CORHAY (1978, finance), DE JAEGER (1975), CHARNES, COOPER and LEARNER (1978, marketing), BROEKSTRA (1976, systèmes), etc.... Les travaux récents utilisant les mesures informationnelles additives et non additives visent essentiellement la classification ou la représentation d'observations (ex. SALLANTIN, 1977), SALLANTIN et VAN DER PYL (1978).

Les problèmes d'estimation statistique ont trouvé une intéressante généralisation de présentation dans la théorie mathématique de l'information (citons le précurseur KULLBACK, 1959).

Comme nous n'avons aucun objectif inférentiel, nous négligerons cet aspect encore plus important de l'A.I.M. L'A.I.M. que nous utilisons ici sera un pur outil descriptif basé sur trois types de mesures : l'entropie (H), l'espérance de gain d'information (I) et l'espérance d'information mutuelle ou transmise (I_d). Cette dernière mesure encore appelée distance de Kullback ou degré de dépendance informationnel est une généralisation non métrique de la distance du *chi-carré* ; elle apparaît à la fois comme une différence d'entropies et comme un cas particulier de l'espérance de gain d'information, réalisant ainsi un pont entre les deux types fondamentaux de mesures informationnelles et additives H et I.

1.2. Les seules banques de données traitables par l'A.I.M. sont constituées in fine de données quantitatives catégorielles positives. Observons qu'il est toujours possible de transformer un bloc de

données *irrégulier* – défini comme présentant des lacunes ou absence de données, et éventuellement des données négatives ou nulles – par extraction de "sous-blocs" réguliers, par arrangement de données ou par regroupement. Ces trois méthodes de régularisation ont évidemment des effets et des performances très différentes.

Des études déjà réalisées portent en finance sur des banques de bilans (DUMAS, 1976, COLSON, 1977, CORHAY, 1978) et en marketing sur la prévision d'une distribution de vente canaux/produits/marques (DE JAEGER, 1975).

Nous limiterons cette étude à un bloc de données irrégulier à trois dimensions (i, j, k) de maxima (I, J, K). L'analyse descriptive ne fournira que les trois mesures jointes à trois dimensions H(x,y,z), I(x', y', z' ; x, y, z) et I_d (x, y, z) pour les trois variables (x, y, z) à l'exclusion de tout schéma de liaison de mesures avec les interactions du second ordre ; par ailleurs, elle se limitera à des comparaisons et mesures sur des *tableaux* (à deux dimensions) et des *vecteurs* (à une dimension).

1.3. *Par exploitation rapide* de la banque, nous entendons

A. Détecter et localiser précisément des événements et faits saillants qui ressortent des données quantitatives.
B. mesurer des équilibres et des déséquilibres structurels.
C. comparer des structures au moyen de distances structurelles (I) et détecter des signes d'activité ou d'inactivité (voir A) avec le gain (I).
D. mesurer les degrés de dépendance des critères de classification (qui peuvent aussi servir à découvrir un suivi de politique – COLSON (1977).
E. visualiser certaines évolutions, certaines dissimilarités au travers des données via les diagrammes d'activité ou les triangles structurels (COLSON, 1977).

Par structure, nous entendons tout vecteur, tableau ou bloc constitués de parts relatives positives ou nulles qui somment sur un et prennent ainsi la forme d'une distribution de probabilités (p) conditionnelle, marginale ou jointe associée à un ensemble de données (x, y ou z, ou une combinaison).

1.4. Dans la section 2, nous ferons un rapide rappel des mesures informationnelles utilisées et de certaines de leurs interprétations. La section 3 présentera les choix et modes d'interrogation des données et nous terminerons par un schéma général d'exploration interactif dans la section 4.

2. RAPPEL DES MESURES INFORMATIONNELLES UTILISEES (Figure 1)

La figure 1.a. montre les formules de base utilisées dans les cas unidimensionnel – pour un *vecteur* x de distribution p –, bidimensionnel – pour un *tableau* (x, y) de distribution jointe (p_{ij}) –, tridimensionnel – pour un *bloc* (xyz) de distribution jointe (p_{ijk}) –. Pour l'entropie (formule (1)), on ne considère qu'un seul vecteur, tableau ou bloc. L'espérance de gain est une mesure de distance non métrique qui compare *deux* vecteurs, tableaux ou blocs, comme le montrent les formules (3) et (4). Sous sa forme première (3), I est associé à un graphe orienté et ne satisfait pas à la propriété de symétrie, c'est pourquoi on remplacera d'habitude I par la mesure symétrique IM (4). Le degré de dépendance apparaît comme une mesure de distance entre la structure jointe (p_{ij}) ou (p_{ijk}) observée et une structure hypothétique d'indépendance entre les critères (p_i, p_j) ou ($p_{i..}$ $p_{.j.}$ $p_{..k}$)), à la symétrie près (formules (2)).

La figure 1b montre que les mesures informationnelles ont une allure non linéaire et symétrique par

rapport à la distribution diffuse, ce qui crée une difficulté d'interprétation (1).

La figure 1c présente le schéma des liaisons existant entre les mesures bidimensionnelles. Par exemple, on a les relations suivantes :

$$H(x, y) = H(x) + H(y) - I_d(x, y) \quad (8)$$
$$I(x, y) = I(x) + I(y) - II_d(x, y) \quad (9)$$
$$H(x,y) = H(x) + H(y/x) = H(y) + H(x/y) \quad (10)$$
$$I(y, x) = I(x) + I(y/x) = I(y) + I(x/y) \quad (11)$$

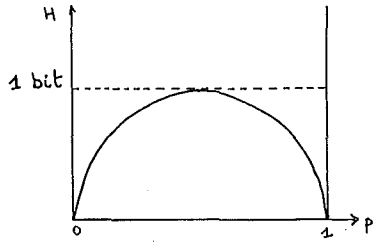
liant les mesures jointes aux mesures marginales et conditionnelles et au degré de dépendance I_d ou au degré de changement des degrés de dépendance II_d. Les formules correspondant à ces différentes mesures sont présentées à la figure 1d.

(1) Pour I, ceci est vrai pour la seule distribution de départ diffuse.

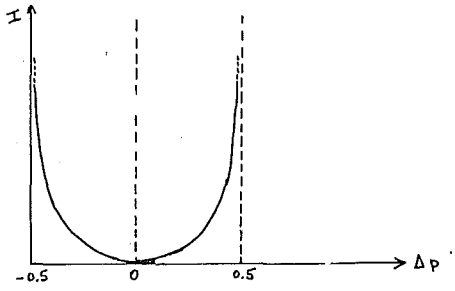
Figure 1a : Formules de base

	cas unidimensionnel	cas bidimensionnel	cas tridimensionnel
H	$H(x) = -\sum_{i=1}^{n} p_i \log p_i$ (1.1)	$H(x,y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij} \log p_{ij}$ (1.2)	$H(x,y,z) = -\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{l} p_{ijk} \log p_{ijk}$ (1.3)
I_d		$I_d(x,y) = \sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij} \log\left(\frac{p_{ij}}{p_{i.} \times p_{.j}}\right)$ (2.2)	$I_d(x,y,z) = \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{l} p_{ijk} \log\left(\frac{p_{ijk}}{p_{i..} \times p_{.j.} \times p_{..k}}\right)$ (2.3)
I_1	$I_1(x) = \sum_{i=1}^{n} q_i \log(q_i / p_i)$ (3.1)	$I_1(x,y) = \sum_{i=1}^{n}\sum_{j=1}^{m} q_{ij} \log(q_{ij} / p_{ij})$ (3.2)	$I_1(x,y,z) = \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{l} q_{ijk} \log(q_{ijk} / p_{ijk})$ (3.3)
I_2	$I_2(x) = \sum_{i=1}^{n} p_i \log(p_i / q_i)$	$I_2(x,y) = \sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij} \log(p_{ij} / q_{ij})$	$I_2(x,y,z) = \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{l} p_{ijk} \log(p_{ijk} / q_{ijk})$
I_M	$I_M(x) = \frac{1}{2}\left[I_1(x) + I_2(x)\right]$ (4.1)	$I_M(x,y) = \frac{1}{2}\left[I_1(x,y) + I_2(x,y)\right]$ (4.2)	$I_M(x,y,z) = \frac{1}{2}\left[I_1(x,y,z) + I_2(x,y,z)\right]$ (4.3)

Figure 1 b



$H(x) = -p \log_2 p - (1-p) \log_2 (1-p)$



$I_1(x,y) = q \log_2(q/p) + (1-q) \log_2((1-q)/(1-p))$

Figure 1c

$H(x,y)$ (1.2)	
$H(x)$	$H(y/x)$
$H(x/y)$	$H(y)$
	$I_d(x,y)$ (3.2)

$I_1(x,y)$ (3.2)	
$I_1(x)$	$I_1(y/x)$
$I_1(x/y)$	$I_1(y)$
	$II_{d_1}(x,y)$

Figure 1d

$H(x) = -\sum_{i=1}^{n} p_{i.} \log p_{i.}$

entropie marginale selon x

$H(y/x) = -\sum_{i=1}^{n} p_{i.} \sum_{j=1}^{m} p_j^{(i)} \log p_j^{(i)}$

entropie de y conditionnelle à x

$I_1(x) = \sum_{i=1}^{n} q_{i.} \log(q_{i.} / p_{i.})$

espérance de gain d'information marginale selon x

$I_1(y/x) = \sum_{i=1}^{n} q_{i.} \sum_{j=1}^{m} q_j^{(i)} \log(q_j^{(i)} / p_j^{(i)})$

espérance de gain d'information de y conditionnelle à x

$II_{d_1}(x,y) = \sum_{i=1}^{n} \sum_{j=1}^{m} q_{ij} \log \left[\left(\frac{q_{ij}}{q_{i.} \times q_{.j}} \right) \Big/ \left(\frac{p_{ij}}{p_{i.} \times p_{.j}} \right) \right]$

espérance de changement des informations mutuelles

3. CHOIX ET MODES D'INTERROGATION DES DONNEES

L'utilisateur du programme interactif dont le schéma d'exploration est annexé peut effectuer plusieurs choix successifs au niveau des types de mesures, des types d'analyse et du mode d'exploration.

3.1. *Les trois types de mesure* (H, I_d, I) peuvent être sélectionnées tout en sachant que la demande du schéma de liaisons fournira (H et I_d) pour les entropies et (I et II_d) pour les espérances de gain.

3.2. Il existe *deux types fondamentaux d'analyse* (si on exclut l'analyse particulière des degrés de dépendance, qu'on peut d'ailleurs inclure dans l'analyse des distances) qui sont
1. l'étude des dispersions ou encore des déséquilibres structurels où on utilise les entropies et
2. la comparaison des distances structurelles, employant I. En tenant compte du schéma de liaison bidimensionnel, la première étude peut être aussi vue comme une analyse d'incertitude, généralisation non métrique de l'analyse de variance. La comparaison des distances peut se faire de trois manières (S, A, L).

2.1. Par rapport à une même structure de référence (S).
2.2. De façon chaînée. I est alors calculé à la manière d'un indice chaîné. Si la chaîne se constitue par rapport à un axe temporel, on pourra parler d'une *analyse d'activité* au sens strict (A). La visualisation d'un tel indice par *diagramme d'activité* présentera l'allure d'une sorte d'encéphalogramme indiquant comme ce dernier des périodes d'activité plus ou moins intenses alternant avec des périodes de repos (cfr. DE JAEGER, 1975).
2.3. Librement (L), c'est-à-dire mesurer la distance d'une structure à n'importe quelle autre structure, y compris une structure hypothétique d'indépendance.

Enfin, on peut aussi effectuer des analyses composites, comme celle des bilans informationnels en finance (THEIL, 1969, COLSON, 1977, CORHAY, 1978).

3.3. *Le mode d'exploration* d'un bloc de données peut être une exploration générale ou une interrogation particulière ou focalisée s'adressant alors à un sous-bloc ou, et utilisant un des types de mesures ou un des types d'analyse. L'exploration peut enfin être interactive se focalisant éventuellement suite aux découvertes de faits saillants en procédant du général au particulier, ou elle peut être la simple recherche d'une vue synthétique des données en rapport avec le type de données.

Observons que le programme actuel est interactif et l'option générale de vue synthétique n'est pas explicitement incorporée. Par contre, l'exploration d'un bloc ou d'un sous-bloc peut se faire selon les trois angles de vue (trois axes d'exploration) possibles. La visualisation de données structurelles par triangles signalée plus haut n'est pas encore prévue.

4. SCHEMA GENERAL D'EXPLORATION DE LA BANQUE *(Figure 2)*

La figure 2 ci-après est l'organigramme simplifié du programme actuel. Les numéros renvoient aux explications suivantes :

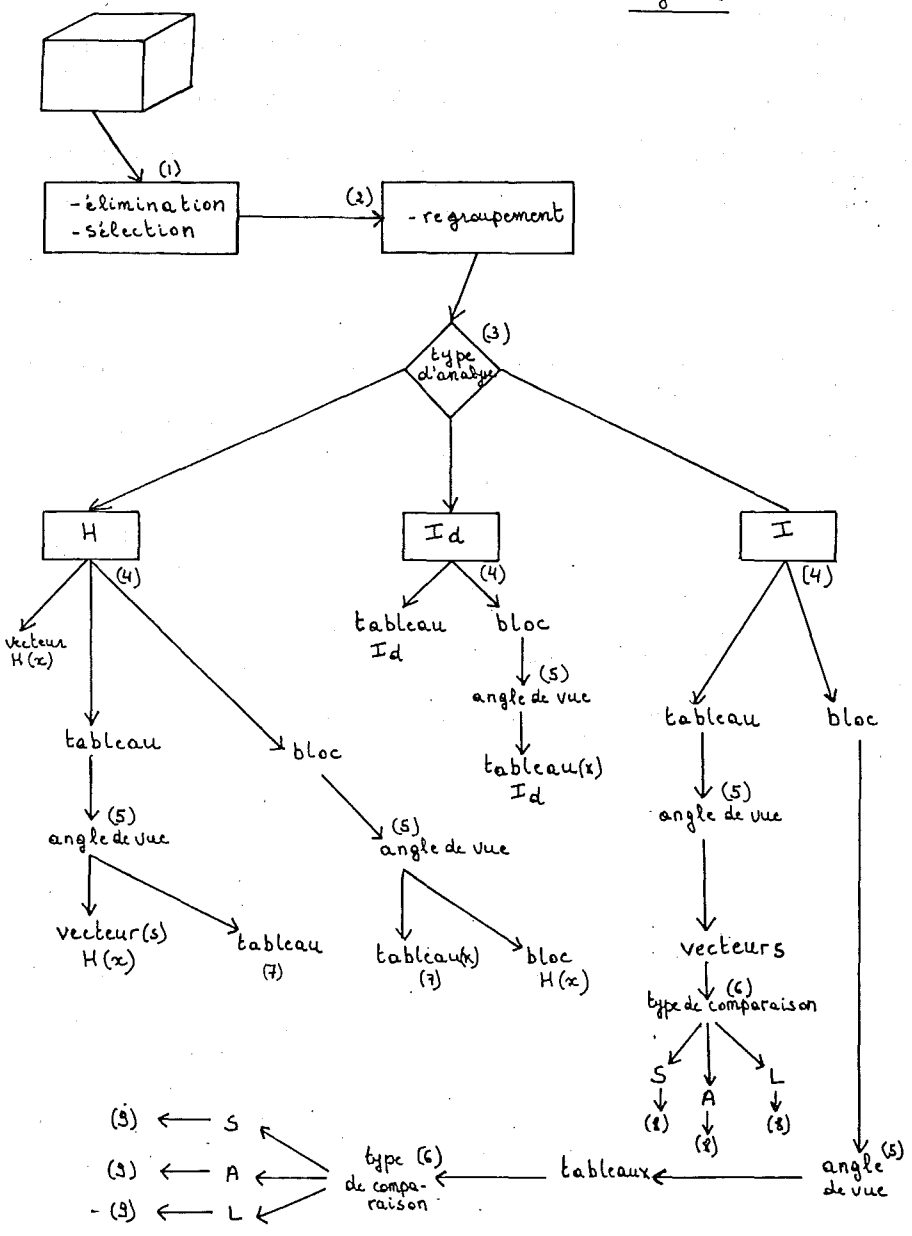
1. Procédure de constitution d'un sous-bloc régulier de données
   - soit par élimination de données du bloc de départ ou d'un sous-bloc constitué lors d'une étape précédente.
   - soit par constitution d'un nouveau sous-bloc.

(2) Procédure de regroupement de données.
(3) Type de mesure sélectionné H, I ou I_d.
(4) Le programme détermine la nature du sous-bloc traité : bloc, tableau ou vecteur. La poursuite de l'exploration dépend en effet du type de sous-bloc traité.
(5) Angle de vue de l'exploration, c'est-à-dire définition du ou des axe(s) sur lesquels va se faire l'exploration.
(6) Type de comparaison : par rapport à une structure de référence (S), de façon chaînée (A) ou libre (L).
(7) Les mesures effectuées sont toutes les entropies bidimensionnelles de la figure 1c.
(8) Les mesures effectuées sont toutes les espérances de gain d'information bidimensionnelles de la figure 1c.

REMARQUES

1. L'intervention de l'utilisateur se fait aux points (1), (2), (3), (5) et (6), sous forme de réponses aux questions posées par le programme.
2. Après avoir obtenu une ou plusieurs mesures informationnelles d'un sous-bloc, l'utilisateur peut
   - effectuer sur le même sous-bloc une autre mesure informationnelle, qu'elle soit de même type ou non (H, I ou I_d), qu'elle relève d'un autre angle de vue ou d'un autre type de comparaison (S, A ou L).
   - éliminer ou regrouper des données de ce sous-bloc et recommencer le processus d'exploration.
   - abandonner ce sous-bloc, reconstruire un nouveau sous-bloc à partir du bloc de départ et recommencer le processus d'exploration dès le début.
3. Avant d'effectuer une mesure informationnelle, le programme vérifie
   - si le sous-bloc traité (bloc, tableau ou vecteur) est régulier = sans donnée marquante, négative ou nulle.
   - si les vecteurs ou les tableaux qui sont comparés ont la (les) même(s) dimension(s).

Figure 2



REFERENCES

ASHBY, W.R., 1965, Measuring the internal informational exchange in a system, *Cybernetica,* 8, pp. 5-22.

ATTNEAVE, F., 1959, *Application of information theory to psychology,* Holt, Rinehart and Winston, New-York.

BROEKSTRA, G., 1976, Some comments on the application of informational measures to the processing of activity arrays, *International Journal of General Systems,* 3, pp. 43-51.

CHARNES, A., COOPER, W., LEARNER, D., 1978, Constrained information theoretic characterizations in consumer purchase behaviour, *J. Opl. Res. Soc.,* 29, 9, pp 833-842.

COLSON, G., 1973, Que peut apporter l'entropie à la science de gestion, *Rev. Belge des Sc. Comm.,* 1 et 2, pp. 43-71.

COLSON, G., 1977, Utilisation de l'analyse informationnelle pour orienter une analyse financière classique, *Colloque IRIA,* Analyse des données et informatique 7-9 sept., Versailles.

CORHAY, A., 1978, Utilisation de méthodes multivariées et de la théorie des graphes pour l'analyse des structures financières d'un échantillon de petites entreprises du secteur de la construction, *mémoire de licence,* EAA, Univ. de Liège.

DE JAEGER, J., 1975, Le monde actuel et potentiel des Pet Foods en Belgique, *Mémoire de licence,* EAA, Univ. de Liège.

DUMAS, M., 1976, Analyse informationnelle du secteur de l'électricité en Belgique, *Mémoire de licence,* EAA, Univ. de Liège.

KULLBACK, S., 1959, *Information theory and statistics,* New-York, Wiley.

LEV, B., 1969, An information theory analysis of budget variances, *The Accounting Review,* XLIV, 4.

Mc DONALD, D.K.C., 1952, Information theory and its application to taxonomy, *Journal of Applied Physics,* 23, pp. 529-531.

SALLANTIN, J., 1977, Approche commune des différents modèles en théorie de l'information, *colloque du CNRS,* groupe 22, "Développements récents de la théorie de l'information".

SALLANTIN, J. et VAN DER PYL, 1978, *Entropies, dimensions et représentation d'observations,* Séminaire questionnaires du groupe structure de l'information (C. PICARD), CNRS, 9 mars, Paris.

THEIL, H., 1969, On the use of information theory concepts in the analysis of financial statements, *Management Science,* 15, 9, pp. 459 à 480.

THEIL, H., 1967, *Economics and information theory,* Rand Mc Nally, Chicago, Illinois, Vol. 7 in : Studies in Mathematical and Managerial Economics.

VAN SOEST, J.L., 1954, A contribution of information theory to sociology, *Synthese,* 9, pp. 265-273.