LES CAHIERS DE L'ANALYSE DES DONNÉES

D. DOMENGÈS

Classification ascendante hiérarchique d'après un critère adapté aux tableaux de flux

Les cahiers de l'analyse des données, tome 7, n° 2 (1982), p. 169-172

<http://www.numdam.org/item?id=CAD_1982__7_2_169_0>

Les Cahiers de l'Analyse des Données
Vol. VII - 1982 - n° 2 - p. 169-172

CLASSIFICATION ASCENDANTE HIÉRARCHIQUE D'APRÈS UN CRITÈRE ADAPTÉ AUX TABLEAUX DE FLUX [C.A.H. FLUX]

par D. Domengès (1)

1 Les tableaux de flux

Nous dirons que ce sont des tableaux de correspondance I × J où les k(i,j) mesurent dans une unité homogène des mouvements de matière de biens ou de personnes ayant pour source i et pour but j. Un exemple classique est la matrice de Léontieff ou tableau des échanges industriels (T.E.I.). Dans ce cas on a :

I = un ensemble de produits ;

J = un ensemble de branches ;

k(i,j) = valeur (dans une monnaie quelconque) de la quantité du produit i absorbée au cours d'une année par la branche j.

Dans la pratique les branches peuvent être confondues avec les produits et par conséquent I = J.

Un autre exemple est celui des matrices d'échanges commerciaux internationaux : k(i,j) mesure alors les exportations du pays i vers vers le pays j.

Bien que dans de nombreux cas les ensembles I et J puissent être distincts, et que la description complète des flux requiert éventuellement un tableau ternaire (e.g. tableau des exportations par produits des pays de l'OCDE vers les pays de l'OPEP ; et tableau quaternaire même si l'on considère l'évolution temporelle), nous considérons ici exclusivement des flux carrés I × I. En règle générale, ces tableaux ne sont pas symétriques (k(i,i') ≠ k(i',i)).

2 Définition ou calcul des termes diagonaux

Une particularité des tableaux carrés de flux est que la définition des termes diagonaux k(i,i) est presque toujours discutable. Les sources statistiques omettent ordinairement ces termes : cependant les méthodes usuelles d'analyse statistique (analyse factorielle et C.A.H.) sont conçues pour l'analyse de tableaux complets.

Plusieurs voies mathématiques ont été explorées par B. Burtschy pour suppléer à l'absence de cette diagonale : reconstitution comme s'il s'agissait de données manquantes, calcul de distance par une formule ne comportant pas ces termes puis analyse du triple (cf. Burtschy Thèse et article à paraître).

(1) *Docteur 3° cycle.*

(*) *La présente note est extraite de la thèse de D. Domengès.*

D'un point de vue réel il semble que k(i,i) puisse être défini comme la limite de la somme des flux ayant pour source et pour but des sous-unités de i, à supposer que i soit décomposé en sous-unités aussi fines que possible. Bien que cette définition mette en jeu un processus de subdivision indéfini difficile à concevoir avec précision, elle devient concrètement applicable si, e.g., connaissant les échanges extérieurs entre pays, on se propose de calculer les échanges intérieurs à un bloc de pays ; il est d'autre part vraisemblable que les échanges, entre Belgique et Hollande sont comparables à ceux de deux régions de France telles que Lorraine et Champagne-Ardennes. De ce point de vue B. Burtschy a défini le terme diagonal k(i,i) d'une matrice d'échanges commerciaux internationaux par le volume du commerce intérieur au pays i ; et cela bien que selon ses propres critiques on recense avec le commerce intérieur des opérations locales qui ne sauraient être assimilées au commerce international. (Il conviendrait de reconnaître sur quels produits ou services portent les échanges internationaux : véhicules, navires, combustibles bruts ou raffinés etc. ; et recenser pour ces seuls postes ce qui est à la fois produit et consommé à l'intérieur d'un pays).

De cette discussion relative aux données, nous retiendrons que les termes diagonaux d'une matrice de flux ne sont jamais bien connus ; et qu'ils sont généralement surestimés. Cette dernière particularité aboutit facilement en analyse des correspondance à une suite de facteurs associés à des v. p. voisines de 1 et définis par un seul pays. Le traitement statistique sera d'autant plus sûr qu'il sera moins sensible aux termes diagonaux. De plus il importe de noter que si les k(i,i) sont incertains les k(c,c), mesurant les flux internes d'agrégats c définis à partir des unités de base i, le sont beaucoup moins. C'est selon ces principes, qu'en expérimentant sur les T.E.I. de la France nous avons conçu la méthode qui fait l'objet de la présente note.

3 *Taux d'association et critère d'agrégation*

Nous partons d'un tableau de flux k_{II} dont la diagonale est supposee connue. A ce tableau est associé un tableau symétrique $k_{S_{II}}$. La définition et le calcul des k(i,i') ou ks(i,i') sont étendus aux classes c et c' d'éléments i, suivant les notations et formules qui suivent :

k(i,i') = flux de i vers i'

kS(i,i') = k(i,i') + k(i',i) = flux mutuels entre i et i'

N.B. : on a donc kS(i,i) = 2 k(i,i).

kS(i) = $\Sigma\{kS(i,i') \mid i' \in I\}$ = total des flux ayant i pour source ou pour but ; (dans ce total, k(i,i) est compris deux fois, les flux internes ayant à la fois i pour source et pour but).

kS = total des termes du tableau carré des kS(i,i') (diagonale comprise).

Pour deux parties c, c' de I (c, c' $\subset$ I ; éventuellement c = c') on définit de même :

k(c,c') = $\Sigma\{k(i,i') \mid i \in c ; i' \in c'\}$ = total des flux ayant leur source dans c et leur but dans c'; (dans ce total les flux ayant leur source et leur but dans c $\cap$ c' sont comptés deux fois).

$$kS(c,c') = k(c,c') + k(c',c) \ ;$$

$$kS(c) = \Sigma\{kS(i) \mid i \in c\}.$$

Le taux d'association entre deux parties c et c' est alors :

$$ass(c,c') = (kS \ . \ kS(c,c')) / (kS(c) \ . kS(c')) \ .$$

Pratiquement cette définition destinée à construire une C.A.H. ne sera utilisée que pour des classes c et c' d'intersection vide. A la base, on notera que ass(i,i') = 1 correspond à l'absence de liens particuliers entre i et i' (cf. indépendance en probabilité) ; des valeurs supérieures ou inférieures à 1 caractérisant respectivement des échanges actifs ou faibles. Quant à ass(i,i) , bien que nous ne l'utilisions pas puisque l'agrégation ne se fait qu'entre éléments différents, il caractérise l'intensité des flux intérieurs à i relativement aux flux extérieurs ayant i pour source ou pour but.

Ceci posé c'est l'inverse du taux d'association, qui étant l'analogue d'une distance sera utilisé comme niveau d'agrégation . On notera :

$$niv(c,c') = 1/ass(c,c').$$

4 *Calcul du niveau d'agrégation et axiome de réductibilité*

Au sein d'un algorithme de C.A.H., les niveaux d'agrégation doivent être calculés de proche en proche ; de plus il est de beaucoup préférable que la hiérarchie construite ne présente pas d'inversion : i.e. que si i a été construit par agrégation de a et de b au niveau n , l'agrégation de c à c' ne se puisse faire qu'à un niveau n' supérieur (voire égal à n). Nous devons considérer de ce point de vue le taux d'association défini au § 3.

4.1 Calcul de ass(a ∪ b,c') : Soit a et b deux parties de I d'intersection vide (éventuellement chacune de ces parties peut-être réduite à un seul élément : a = {i}) et c' une autre partie :

$$a,b \subset I \ ; \ a \cup b = c \ ; \quad c' \subset I \ ; \quad \text{on a :}$$

$$\begin{aligned} ass(c,c') &= kS.kS(c,c') / (kS(c).kS(c')) \\ &= kS(kS(a,c') + kS(b,c')) / (kS(c).kS(c')) \\ &= (kS(a)/kS(c)) \, ((kS.kS(a,c')) / (kS(a).kS(c'))) \\ &\quad + (kS(b)/kS(c)) \, ((kS.kS(b,c')) / (kS(b).kS(c'))) \\ &= (kS(a) \ ass(a,c') + kS(b) \ ass(b,c')) / (kS(a) + kS(b)). \end{aligned}$$

Ainsi le taux ass(c,c') apparaît comme une moyenne entre les taux ass(a,c') et ass(b,c') pondérés respectivement par kS(a) et kS(b).

4.2 Axiome de la médiane : Cet axiome assure que dans la construction d'une hiérarchie on ne rencontre pas d'inversion ; il permet de plus,d'appliquer les algorithmes accélérés de C.A.H. ; agrégation des voisins réciproques, et graphes réductibles (cf. C. de Rham, *Cah.* Vol. V n° 2 ; M. Bruynooghe, *Cah.* Vol. III n° 1 ; et J. Juan, *Cah.* Vol. VII n° 2 ; et [*C.A.H. CHAINE RECIP.*] *ibid* § 2.1).

En bref (en remplaçant le terme de niveau d'agrégation par celui, plus familier de distance) nous dirons l'axiome assure que dans la construction d'une hiérarchie, l'agrégation de a à b ne peut créer une classe c qui soit le plus proche d'une autre classe c' que ne l'était chacune des deux classes a et b. Ici il est essentiel que la distance entre a et b qu'on a agrégés soit inférieure à celle

entre a et c' ou entre b et c' (sinon l'agrégation se serait faite entre ceux-ci, non entre ceux-là). La formule concerne donc un triangle a,b,c' dont ab est le plus petit côté ; on écrit (sans spécifier les cas de côtés égaux qui généralement en C.A.H. sont résolus par un choix arbitraire) :

$$niv(a,b) \leq \inf\{niv(a,c'), niv(b,c')\}$$
$$\Rightarrow \inf\{niv(a,c'), niv(b,c')\} \leq niv(a \cup b,c')$$

Dans le cas présent la démonstration est particulièrement simple; elle ne fait pas intervenir l'hypothèse qui précède le signe $\Rightarrow$, mais seulement le fait que $a \cap b = \emptyset$: en terme de taux d'association (inverses des niveaux d'agrégation) on a en effet toujours :

$$ass(a \cup b,c') \leq \sup\{ass(a,c'), ass(b,c')\},$$

parce que ass(a ∪b,c') est compris entre les deux taux ass(a,c') et ass(b,c') dont il est une moyenne.

5 *L'algorithme de classification*

Le calcul des distances rentre d'après la formule du § 4.2 dans le cadre de la procédure générale de l'algorithme de C.A.H. TI n° 4 § 2.1. Il suffit de poser :

```
DISTANCE(P1,L1,D1,P2,L2,D2,DIS12,PS,LS,DS,DIS1S,DIS2S):=
     (P1 + P2)/ ((P1/DIS1S) + (P2/DIS2S)) ;
```

Dans cette procédure, donnée ici sous sa forme générale, figurent des quantités L1, L2, LS que l'on n'a pas à définir ici ; les niveaux D1 et D2 des deux noeuds qu'on agrège, ainsi que DS du noeud S préexistant auquel on calcule la distance du noeud créé, sont des variables qui ont un sens mais n'entrent pas dans le calcul : seuls servent les poids P1, P2 des deux éléments agrégés, et leurs distances DIS1S, DIS2S à S. La formule écrite est celle du § 4.2 à ceci près que les niveaux (ou distances) sont comme on l'a dit au § 3 *(in fine)* les inverses des taux.

A la base on doit poser (en supposant pour simplifier l'écriture que le tableau DIS est écrit avec deux indices)

```
DIS[I,IP]:=1/ASS[I,IP] ;

P[I]= KS[I].
```

On remarquera que la construction faite ici ne tient aucun compte de la polarité des flux, autrement dit de la dissymétrie de la matrice initiale des k(i,i'). Mais il est possible de porter des polarités sur l'arbre des flux en calculant pour chaque noeud c = a ∪ b lequel est le plus grand des deux nombres k(a,b) (flux de l'aîné vers le benjamin) ou k(b,a) (flux du benjamin vers l'aîné) ; ou plus précisément en calculant le rapport k(a,b)/k(b,a).