REVUE FRANÇAISE D'AUTOMATIQUE, INFORMATIQUE, RECHERCHE OPÉRATIONNELLE. RECHERCHE OPÉRATIONNELLE

V. DEGOT

J. M. HUALDE

De l'utilisation de la notion de clique (sous-graphe complet symétrique) en matière de typologie de populations

Revue française d'automatique, informatique, recherche opérationnelle. Recherche opérationnelle, tome 9, n° V1 (1975), p. 5-18

<http://www.numdam.org/item?id=RO_1975__9_1_5_0>

R.A.I.R.O.
(9e année, janvier 1975, V-1, p. 5 à 18)

DE L'UTILISATION DE LA NOTION DE CLIQUE (SOUS-GRAPHE COMPLET SYMETRIQUE) EN MATIERE DE TYPOLOGIE DE POPULATIONS*

par V. DEGOT [1] et J. M. HUALDE [2]

Résumé. — *Les méthodes de typologie, qui consistent à rechercher des sous-ensembles homogènes à l'intérieur d'une population repérée par des variables quantitatives, sont d'un usage fréquent, en matière d'études de marchés par exemple. Les méthodes existantes, fondées sur l'agglomération progressive d'individus voisins, fournissent de manière simple et rapide des partitions de la population, mais on montre que le résultat peut être instable ou artificiel. Une nouvelle méthode est proposée qui procède de la recherche de cliques (sous-graphe complet symétrique). Cette méthode ne fournit pas nécessairement de partition de l'ensemble et produit simultanément plusieurs découpages possibles, ce qui n'est pas toujours un inconvénient. La rigueur de cette méthode est obtenue au prix d'un allongement du temps de calcul qui peut néanmoins rester acceptable avec l'algorithme nouveau qui est proposé.*

INTRODUCTION

Le présent article propose une nouvelle technique d'obtention des typologies. Celle-ci a fait appel à la notion de clique que suggère pour cet usage M. Bernard Roy dans son ouvrage *Algèbre moderne et théorie des graphes* (7), et elle met en œuvre un procédé de calcul qui semble plus satisfaisant, à divers titres, que ceux existant jusque là.

Si l'on se donne un ensemble d'objets, il est quelquefois possible d'étudier et de chiffrer certains critères — ou variables — concernant tous les objets de l'ensemble. Ces critères pourront alors permettre d'établir une relation entre les objets et il sera possible de dire, selon la relation, que les objets seront proches ou distants, ou bien semblables ou dissemblables, Dans ce contexte, le but de la typologie est de définir des sous-ensembles d'objets tels qu'à l'intérieur de ceux-ci, les objets soient proches ou semblables entre eux.

(1) Centre de Gestion Scientifique de l'E.N.S.M.P.
(2) Compagnie Internationale d'Informatique.
(*) Reçu le 15 juin 1973.

Partant de cette définition assez générale des typologies, nous allons, dans une première partie, proposer une classification de certaines méthodes de typologies existantes qui procèdent par calcul sur des graphes. Nous verrons que la qualité des résultats obtenus est plus liée aux principes qu'au détail des procédures utilisées. Ensuite, et comme une illustration des considérations abstraites de cette première partie, nous décrirons l'algorithme que nous avons élaboré pour obtenir des typologies.

I. UNE CLASSIFICATION DE QUELQUES TYPOLOGIES EXISTANTES

Après avoir énoncé la classification proposée, nous décrirons rapidement les méthodes concernées pour en souligner éventuellement les faiblesses.

I.1. Classification

Dans ce qui suit, nous poserons que les objets sont décrits comme des points dans un espace métrique. La relation entre deux objets utilisera donc la notion de distance topologique de deux points. Une relation possible entre deux objets consiste alors à se fixer un seuil et à poser comme semblables deux objets dont la distance est inférieure au seuil. Une autre relation consiste à classer par ordre croissant toutes les distances entre les objets pris deux à deux.

Dans ces deux cas, on peut résumer une partie de l'information initiale dans la construction d'un graphe symétrique :

– soit, ce graphe sera simplement constitué de sommets et d'arêtes et cela conduira au calcul sur les sommets ;

– soit, ce graphe sera constitué de sommets et d'arêtes indicées selon le classement indiqué plus haut, et ceci permettra le calcul sur les arêtes.

Par ailleurs, on peut opposer les méthodes qui calculent en utilisant des principes d'optimalité, et que nous nommerons récurrentes, à celles qui conduisent directement au résultat par un calcul séquentiel.

Le croisement de ces deux critères de classification conduit au tableau carré suivant (fig. 1) :

	Calcul sur les sommets	Calcul sur les arêtes
Méthodes non récurrentes	Typologie par accumulation	Typologie par concentration
Méthodes récurrentes	Méthode arborescente proposée par B. Roy	Méthode proposée dans cet article

Figure 1

Les méthodes non récurrentes sont évidemment d'un usage beaucoup plus simple, mais nous allons en montrer quelques limitations. Ensuite, nous rappellerons brièvement la méthode proposée par M. Bernard Roy.

I.2. Les méthodes non récurrentes

Nous avons vu plus haut que ces méthodes non récurrentes peuvent procéder de deux manières :

- soit en calculant sur les sommets : méthode par accumulation (6),
- soit en calculant sur les arêtes : méthode par concentration (6).

I.2.1. Méthode par accumulation

On part d'une liste de sommets classés d'une manière aléatoire et de la donnée d'un seuil. On regarde si le second sommet est à une distance du premier inférieure au seuil fixé, si oui, on le met dans la classe du premier sommet, sinon il constitue un second centre de classe. On continue ainsi pour l'ensemble des sommets ; on compare chacun d'eux, successivement, à chaque centre de classe et on l'ajoute à une classe déjà existante ou à la liste des centres de classes.

Il est évident que cette technique où l'on classe les points d'une manière aléatoire peut conduire à autant de résultats qu'il y a de classements différents de points. Et ceci n'est pas vrai uniquement si l'ensemble de la population est peu structuré par rapport aux variables envisagées (dans l'espace métrique de départ). Ce n'est que dans le cas le plus favorable que les individus choisis aléatoirement pour former des centres de classes sont près des centres qui existent réellement dans la population.

I.2.2. Méthode par concentration

La typologie par concentration procède à partir de la liste de toutes les distances des individus deux à deux. On considère les deux individus les plus proches, on les remplace par leur barycentre ([1]). On calcule les nouvelles distances ainsi induites ; puis, de nouveau, on sélectionne la distance la plus faible et on agglomère les deux points qu'elle concerne. La procédure d'arrêt dépend, soit d'un seuil de distance que l'on s'est fixé, soit du nombre de classes que l'on cherche à obtenir.

Cette technique présente, par rapport à la précédente, l'avantage de conduire à un résultat unique pour chaque population étudiée ; mais elle contient cependant des aspects contestables :

- le fait de remplacer des couples de points par leur barycentre peut conduire, lorsque cette opération a été répétée plusieurs fois, à regrouper dans

(1) On peut ou non pondérer les barycentres par le nombre de sommets qu'ils représentent.

la même classe des points assez distants ; ceci dans la mesure où il y a une agglomération de proche en proche avec déplacement des centres de gravité ;

— par ailleurs, si les distances les plus courtes, et donc prises en compte au début du processus, sont mal situées par rapport aux centres réels des classes, on peut être conduit à créer des « ponts » entre des classes réelles, si bien que le programme ne mettra pas celles-ci en évidence.

La faiblesse de ces deux méthodes provient donc du fait qu'elles opèrent d'une manière énumérative et que la définition des classes homogènes y est assez lâche :

— méthode par accumulation : peuvent appartenir à cette même classe les points situés dans une sphère ayant pour centre un centre de classe et pour rayon le seuil, donc la distance entre deux points semblables peut être deux fois le seuil ;

— méthode par concentration : la figure géométrique obtenue est complexe et dépend du classement des arêtes ; mais elle est telle que deux points de la même classe peuvent être à une distance supérieure au seuil.

I.3. Une méthode récurrente : calcul sur les sommets

Nous nous intéresserons parmi les méthodes récurrentes à celles qui font appel à une définition stricte des classes d'objets semblables avec la notion de cliques : une clique est un sous-ensemble des sommets dont le sous-graphe induit est complet (symétrique).

Les méthodes de typologie à l'aide de cliques procèdent en deux étapes : énumération des cliques maximales du graphe, ensuite, recherche de la couverture minimale du graphe à l'aide de cliques maximales.

I.3.1. Recherche des cliques maximales à partir des sommets

Dans cette technique, l'ordre arbitraire qui doit être introduit sur les sommets dans la technique de typologie par accumulation disparaît : tous les sommets interviennent avec le même rang dans une exploration arborescente de leurs combinaisons possibles, et où l'on progresse à l'aide d'un critère d'optimalité que nous allons décrire.

Nous ne ferons ici qu'exposer les principes de cette méthode afin d'illustrer les différences qu'elle présente avec celle que nous exposerons par la suite. Signalons qu'elle a fait l'objet de nombreux programmes et qu'elle se trouve développée, entre autres, dans les ouvrages (3), (5), (7).

Considérons un graphe $G = (X, U)$ et C_G l'ensemble des cliques de ce graphe.

Soit un sommet quelconque de X, par exemple x_1 dans une certaine numérotation. Il permet de séparer C_G en deux parties disjointes non vides :

– d'une part, les cliques qui contiennent x_1 et donc ne contiennent aucun des sommets x_i tels que $(x_1, x_i) \notin U$,

– d'autre part, les cliques qui ne contiennent pas x_1.

On peut ensuite séparer suivant le même principe l'une et l'autre parties de C_G ainsi définies :

– pour la première, on fera intervenir le premier sommet par ordre de numérotation de $\{ x_i/(x_1, x_i) \in U \} - x_1$,

– pour la seconde, on fera intervenir x_2.

On définit ainsi une procédure de séparation de C_G à laquelle se trouve associée une arborescence qui la caractérise :

– chaque sommet de l'arborecence représente une partie de C_G, celle-ci étant formée de cliques contenant tous les éléments d'un certain ensemble Ap et aucun de ceux d'un ensemble $\bar{A}p$,

– les arcs de l'arborescence traduisent la relation de filiation dans la procédure de séparation,

– tout sommet de l'arborescence tel que l'ensemble $X - Ap - \bar{A}p$ est non vide a deux suivants définis à partir du premier élément de cet ensemble, soit x_{ip}, par :

$$- \quad Ap \cup \{ x_{ip} \} \quad , \quad \bar{A}p \cup \{ x_j/(x_j, x_{ip}) \notin U \}$$

$$- \quad Ap \quad , \quad \bar{A}p \cup \{ x_{ip} \}$$

– tout sommet tel que l'ensemble $X - Ap - \bar{A}p$ est vide est un sommet terminal de l'arborescence. Ap est une clique. Qui plus est, tous ces sommets sont en correspondance biunivoque avec les éléments de R_G.

Pour obtenir uniquement des cliques maximales, considérons que, si A et $\bar{A}$ sont deux sous-ensembles disjoints du graphe et si S représente les cliques du graphe telles que : $A \subset S$ et $\bar{A} \cap S = \varnothing$, il ne peut exister de cliques maximales possédant cette propriété que si tout sommet

$$x \in B = \bar{A} - \{ x_i/\forall x_j \in A \quad , \quad (x_i, x_j) \in U \}$$

n'est pas relié à au moins un des sommets qui n'est ni dans B ni dans

$$\{ x_i/\forall x_j \in A \quad , \quad (x_i, x_j) \in U \}.$$

On appelle souvent ce principe : critère de maximalité (voir (3), (7)). (Pour l'exposé détaillé de l'algorithme permettant l'énumération des cliques maximales, nous renvoyons le lecteur à (7), chapitre VI, § A3*b*).

I.3.2. Recouvrement et ensembles stables extérieurement

L'énumération des cliques maximales n'est que la première étape de la typologie : il faut ensuite ne retenir parmi celles-ci que celles qui permettent de regrouper la totalité des points avec le minimum de cliques.

Pour calculer ces recouvrements, on utilise la notion d'ensembles stables extérieurement (e.s.e.). Considérons un graphe $B = (X, U)$, un e.s.e. A est tel que :

$$A = \{ x \in X / \forall y A, \exists x \in A, (x, y) \in U \}$$

Cette notion ne prend un sens pour le problème qui nous intéresse que si l'on considère un graphe bipartie $G = (X, Y, V)$ ou X est l'ensemble des points du graphe initial $B = (X, U)$, Y l'ensemble des cliques maximales et V la relation d'appartenance des points aux cliques (fig. 2).

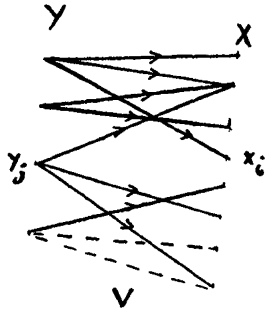

Figure 2

On voit que tout recouvrement de X avec des parties de Y correspond à un e.s.e. de G composé de sommets de Y formant le recouvrement. Réciproquement, tout e.s.e. de G ne comportant aucun sommet de X correspond à un recouvrement de X avec des parties de Y.

On introduit aussi la notion de couverture minimale (e.s.e.m.) qui correspond au fait d'obtenir une couverture de X avec le nombre minimum de parties de Y.

Dans la recherche de typologie par récurrence sur les arêtes que nous allons maintenant décrire, l'obtention du recouvrement se fait aussi de cette manière.

II. UNE AUTRE METHODE RECURRENTE : LA RECHERCHE DES CLIQUES A PARTIR DES ARETES

Nous avons précisé plus haut que l'ensemble U des arêtes pouvait être doté d'une relation de préordre total fondée sur les distances calculées dans l'espace de départ. On peut transformer ce préordre en ordre en ordonnant

arbitrairement les arêtes de mêmes longueurs. Si X est l'ensemble des sommets et U_n l'ensemble formé des n premières arêtes, on peut donc créer une famille de graphes ordonnés : $G_n(X, U_n)$.

Nous allons montrer que si l'on connaît l'ensemble C_n des cliques maximales du graphe G_n, on peut en déduire l'ensemble C_{n+1} des cliques maximales du graphe G_{n+1}.

(Une autre procédure, s'appuyant exactement sur le même principe, consiste à passer de l'étape $n + 1$ à l'étape n, c'est-à-dire à retirer les arêtes au lieu de les ajouter.)

II.1. Principes généraux de cette méthode

Considérons le graphe $G_n(X, U_n)$ et supposons connu sur ce graphe l'ensemble des cliques maximales CM_n.

On passe de $G_n(X, U_n)$ à $G_{n+1}(X, U_{n+1})$ en ajoutant l'arête $u_{n+1} \in U$. Cette arête relie deux sommets A et $B \in X$; la donnée de ces sommets permet de créer une partition de CM_n en trois classes :

- $CM_n(A)$, cliques maximales de G_n qui contiennent A,
- $CM_n(B)$, cliques maximales de G_n qui contiennent B,
- $CM_n(\overline{AB})$, cliques maximales de G_n qui ne contiennent ni A ni B.

On aura évidemment : $CM_n(\overline{AB}) \subset CM_{n+1}$.

Considérons maintenant uniquement $CM_n(A)$ et $CM_n(B)$. Soit a_i une clique de $CM_n(A)$ et b_j une clique de $CM_n(B)$: $a_i \cap b_j + \{ A, B \}$ est une clique de G_{n+1} (qui peut être réduite au couple de sommets $\{ A, B \}$ si $a_i \cap b_j \varnothing$).

Il est évident que cette clique n'est pas forcément maximale, comme la figure 3 permet de s'en convaincre :

$\alpha, \gamma \in CM_n(A)$
$\beta, \delta \in CM_n(B)$
$\alpha \cap \beta + \{ A, B \} \subset \delta \cap \gamma + \{ A, B \}$
$\quad - \quad \subset \gamma \cap \beta + \{ A, B \}$
$\quad - \quad \subset \alpha \cap \delta + \{ A, B \}$
$\alpha \cap \beta$
$\gamma \cap \delta$

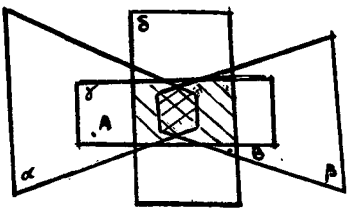

Figure 3

Le problème se ramène alors à celui de savoir, lors de la création de chaque nouvelle clique, d'une part si elle est maximale, c'est-à-dire si elle n'est pas incluse dans une clique déjà obtenue; d'autre part, si une clique déjà obtenue n'est pas incluse en elle. Nous allons voir comment ceci se trouve réalisé dans l'algorithme que nous allons décrire.

II.2. L'algorithme

Le schéma général de cet algorithme est formé de deux boucles emboîtées :

– la première correspond à l'ajout d'une arête et donc à la définition d'un graphe G_{n+1},

– la seconde, incluse dans la précédente, représente la comparaison itérative des cliques a_i de $CM_n(A) = \{ a_i \}$, $i \in [1, p]$ aux cliques b_j de

$$CM_n(B) = \{ b_j \} \quad , \quad j \in [1, q]$$

pour calculer CM_{n+1}.

C'est au niveau de cette seconde boucle que se situe la procédure permettant de décider si la clique $a_i \cap b_j + \{ A, B \}$ est maximale à ce stade du calcul et que nous allons décrire maintenant.

Appelons K l'étape où nous comparons a_i et b_j (on a $K \leqslant (i-1)q + j$ comme nous le verrons par la suite).

Soit CM^{K-1} l'ensemble des cliques retenues provisoirement comme maximales à l'étape précédente, nous allons constituer CM^K. Au début de l'étape K, on pose $CM^K = \varnothing$; on connaît par ailleurs certains éléments de CM_{n+1}, par exemple ceux de $CM_n(\overline{AB})$ ou d'autres comme il apparaîtra par la suite.

A partir de $\{ A, B \}$ et de a_i et b_j on procède comme suit :

1) Si $a_i \cap b_j = \varnothing$, on a évidemment $CM^K = CM^{K-1}$ et l'on passe à l'étape $K + 1$.

2) Si $a_i \cap b_j \neq \varnothing$ on passe en 2.1.

2.1) Si a_i et $b_j \subset a_i \cap b_j + \{ A, B \}$ on pose :

– $CM_{n+1} = CM_{n+1} + [a_i \cap b_j + \{ A, B \}]$,
– $CM_n(A) = CM_n(A) - a_i$,
– $CM_n(B) = CM_n(B) - b_j$,

et on passe à l'étape $K + 1$ (en effet a_i et $b_j \in CM_n$ donc $a_i \cup b_j \cup \{ A, B \}$ ne peut être incluse dans aucune autre clique de CM_{n+1}, donc est maximale. On peut donc éliminer a_i de $CM_n(A)$ et b_j de $CM_n(B)$).

Sinon, on passe en 2.2.

2.2) Si $a_i \subset a_i \cap b_j + \{ A, B \}$, on pose :

– $CM_{n+1} = CM_{n+1} + [a_i \cap b_j + \{ A, B \}$.
– $CM_n(A) = CM_n(A) - a_i$ et on passe à l'étape $K + 1$,

Sinon on passe en 2.3,

2.3) Si $b_j \subset a_i \cap b_j + \{ A, B \}$, on pose :

– $CM_{n+1} = CM_{n+1} + [a_i \cap b_j + \{ A, B \}]$
– $CM_n(B) = CM_n(B) - b_j$ et on passe à l'étape $K + 1$,

Sinon on passe en 2.4.

2.4) On a a_i et $b_j \not\subset a_i \cap b_j + \{A, B\}$:

2.4.1) S'il existe une ou plusieurs cliques $C_1 \ldots C_2$ de CM^{K-1} telles que $C_1, \ldots, C_l \subset a_i \cap b_j + \{A, B\}$ on pose :

– $CM^K = CM^{K-1} + [a_i \cap b_j + \{A, B\}] - [C_1, \ldots, C_l]$ et on passe à l'étape $K + 1$,

Sinon on passe en 2.4.2,

2.4.2) S'il existe au moins une clique C_l de CM^{K-1} telle que

$$a_i \cap b_j + \{A, B\} \subset C_l,$$

on pose :

– $CM^K = CM^{K-1}$ et on passe à l'étape $K + 1$ sinon on pose :

– $CM^K = CM^{K-1} + [a_i \cap b_j + \{A, B\}]$, et on passe à l'étape $K + 1$.

Lorsque toutes les cliques de $CM_n(A)$ auront été rapprochées de celles de $CM_n(B)$, on aura, d'une part, un ensemble CM_{n+1} et, d'autre part, un ensemble CM^K. On posera : $CM_{n+1} = CM_{n+1} + CM^K$ et CM_{n+1} représentera bien l'ensemble des cliques maximales du graphe $G_{n+1}(X, U_{n+1})$.

Nous allons maintenant voir comment on peut se fixer un nombre N d'arêtes que l'on prendra en compte et donc étudier la relation entre les procédures d'arrêt et l'utilisation du seuil variable.

II.3. La courbe reliant le nombre de types au seuil choisi

Si l'on considère les méthodes décrites précédemment, il est possible, pour un seuil donné, de définir des types dans une population étudiée. Pour celles de ces méthodes où la relation qui relie le seuil au nombre de types est univoque, on peut représenter cette fonction. Le graphe d'une telle fonction peut être de la forme de la courbe de la figure 4. Dans ce cas, la population paraît assez bien structurée :

– pour un seuil nul, on a évidemment autant de types que d'individus;

– comme la population est structurée, une croissance légère du seuil permet de regrouper un grand nombre d'individus (δ_1);

– on obtient ensuite un palier qui correspond au fait que les groupes sont assez nettement disjoints (de δ_1 à δ_2);

– puis on recommence à regrouper ces classes (après δ_2).

Parmi les méthodes que nous avons décrites, celles qui procèdent par récurrence sur les sommets opèrent avec un seuil fixe. La récurrence – ou l'énumération – sur les arêtes permet d'obtenir directement cette courbe par variation du seuil.

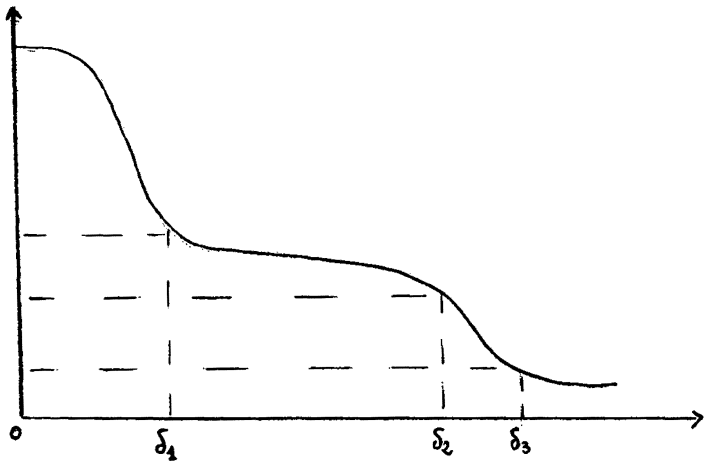

Figure 4

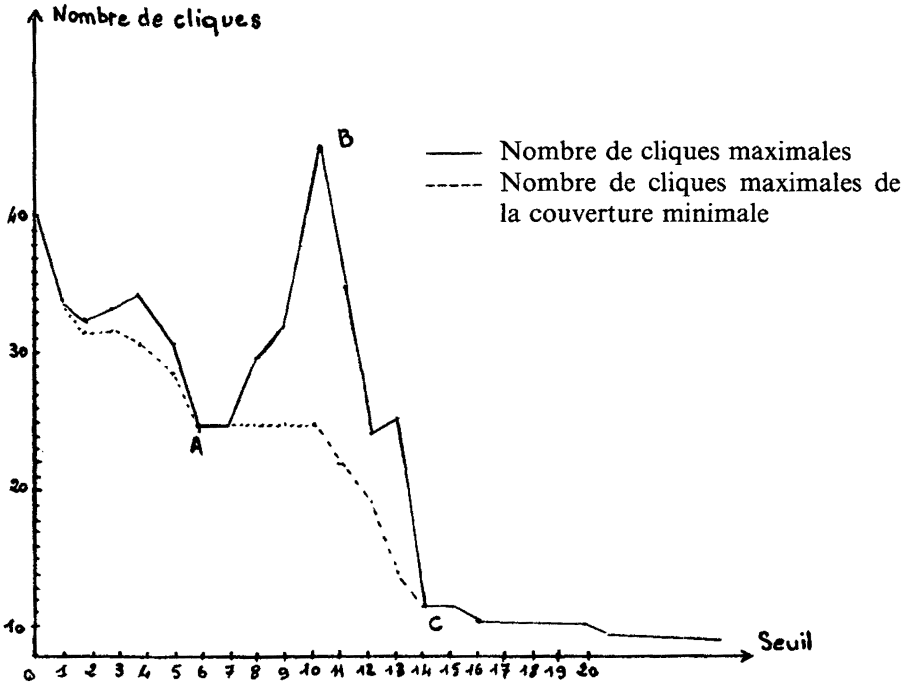

Figure 5

C'est le cas pour la méthode que nous venons d'exposer. Il est possible, chaque fois que l'on a augmenté de 10, 20, ou plus, le nombre d'arêtes de U_n de calculer la couverture minimale de G_n à l'aide de CM_n et donc de construire une courbe identique à celle présentée plus haut. La figure 5 est un exemple d'un tel calcul, mais sa signification n'est pas évidente.

Considérons par exemple les points A, B, C :

– de A vers B, l'ajout d'arêtes fait croître le nombre des cliques très rapidement sans modifier le nombre de cliques de la couverture minimale. Il s'agit évidemment de la création de « ponts » entre des cliques existant à l'étape A, selon le schéma suivant (fig. 6) :

– à l'étape i : cliques maximales : I, II, III, IV, a, b, c, d,

– à l'étape $i + 1$ (ajout de l'arête PQ) : cliques maximales : les précédentes plus PQC, PQD, PQE, PQF.

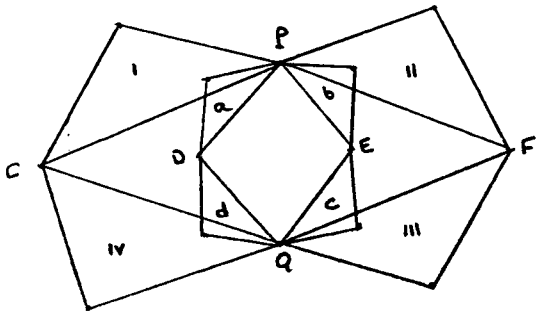

Figure 6

– de B vers C, l'ajout d'arêtes fait décroître le nombre de cliques maximales plus vite que le nombre de cliques de la couverture minimale. Il s'agit alors de l'agglomération rapide de cliques maximales selon le schéma suivant (fig. 7) :

– étape i : cliques maximales : 1, 2, 3, 4, couverture formée de deux cliques (par exemple 1,3),

– étape $i + 1$ (ajout de (P, Q)) : cliques maximales $PQCD$, $PQEF$, couverture formée de ces deux cliques.

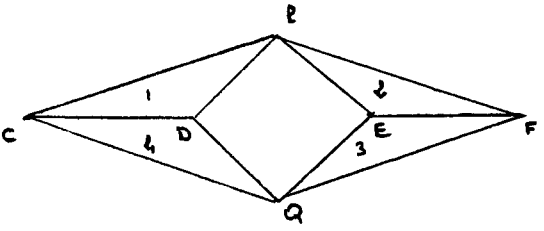

Figure 7

Pour le praticien, le point A apparaît comme celui qui, pour le seuil le plus grand (moins d'arêtes en tout) donne la structure la plus marquée (à l'intérieur d'une organisation donnée entre A et B).

L'utilité d'une telle courbe est essentielle dans la pratique car elle permet de savoir si la solution trouvée correspond à un état stable du système et ceci d'autant plus que la relation de distance utilisée qui agglomère plusieurs dimensions n'a pas en général de signification intrinsèque.

II.4. Quelques différences existant entre les méthodes récurrentes et les précédentes

Les typologies élaborées à partir des notions de cliques et d'e.s.e.m. procèdent en deux temps :

- énumération des cliques maximales du graphe,
- détermination des e.s.e.m.

Pour un graphe donné, il est évident que l'énumération des cliques maximales est unique, mais c'est par la détermination des e.s.e.m. que les résultats des méthodes utilisant les cliques vont différer des précédentes. Nous allons rapidement évoquer de quelles manières :

- les cliques retenues ne forment pas en général une partition du graphe,
- plusieurs ensembles différents de cliques peuvent être retenus comme couverture du graphe.

II.4.1. Les cliques de la couverture minimale ne forment pas une partition du graphe

En effet, pour obtenir une couverture qui soit une partition, il faudrait trouver un ensemble de cliques maximales qui soit lui-même une partition. Or, sauf exceptions, cela ne résulte pas de la définition des cliques maximales, car deux cliques maximales peuvent avoir plusieurs sommets en commun (fig. 8).

$ABCD$ clique maximale
$ARDE$ clique maximale
$ABCD \cap ACDE = ACD$
couverture minimale :
$ABCD$ et ADE

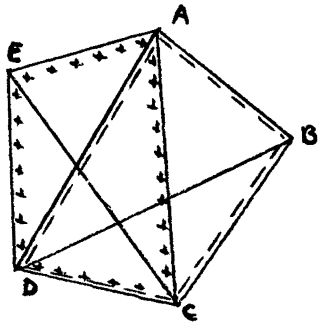

Figure 8

Dans le cas où l'on n'obtient pas de partitions à l'aide de cliques maximales, la couverture du graphe sera telle que les mêmes sommets y apparaîtront plusieurs fois, car cette couverture n'est faite que de cliques maximales. Notons qu'il peut être possible d'obtenir des partitions avec des cliques simples en introduisant des contraintes dans la recherche du recouvrement (voir (8)).

II.4.2. Pour un même graphe, on peut obtenir plusieurs couvertures minimales

Un schéma permet de se convaincre de cette propriété. Considérons trois cliques maximales sur un graphe de quatre sommets (fig. 9) :

- I et II forment une première couverture,
- I et III en constituent une seconde.

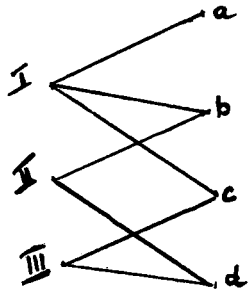

Figure 9

Cela résulte encore de ce que l'on n'obtient pas, en général, de partition à l'aide d'un ensemble de cliques maximales.

On peut, par des considérations annexes (couvertures correspondant à la dispersion minimum des points, par exemple), créer un ordre et donc opérer un choix sur ces divers recouvrements.

CONCLUSION

Les méthodes utilisant la notion de cliques conduisent à des résultats plus satisfaisants, sur le plan théorique, que celles existant auparavant, et ceci de deux manières :

- pour une population donnée, le résultat n'est plus instable,
- il n'y a pas création de classes artificielles.

Par ailleurs, sous la dernière forme décrite, ces méthodes fournissent des renseignements intéressants par l'intermédiaire de la courbe associant au seuil le nombre de classes obtenues.

Il faut cependant ajouter que les deux premières méthodes conduisent à des résultats beaucoup plus rapidement que celles utilisant les cliques, ce qui ne peut manquer d'être un avantage important pour traiter des populations nombreuses dans la pratique. Dans les méthodes itératives, le temps de calcul varie comme une puissance du nombre de points alors que dans les méthodes énumératives, la progression est plutôt d'un type linéaire.

Nous avons utilisé le modèle de recherche des cliques à partir des arêtes pour traiter le cas de la typologie de quarante magasins de grande surface connaissant leurs ventes de certains produits. Ces produits étaient d'ailleurs des représentants jugés typiques des classes d'une typologie effectuée sur certains articles vendus par tous les magasins. Une dizaine de produits avaient été retenus. La durée du temps de calcul sur un ordinateur IBM 360-40 était de l'ordre de quelques minutes pour obtenir la courbe de la figure 5.

Nous avons montré de plus comment l'introduction des typologies à l'aide de cliques posait deux questions :

– le fait de ne pas obtenir de partition lors de la recherche des cliques et du recouvrement peut-il constituer une gêne dans l'utilisation des résultats?

– comment tirer parti de ce qu'il y a généralement plusieurs couvertures, donc plusieurs systèmes de classes, pour un seuil donné?

Il semble qu'une méthode qui ne conduise plus à une partition et associe à un graphe plusieurs couvertures possibles, aille dans le sens d'une description plus précise de la population, c'est-à-dire moins arbitraire.

Enfin, en dehors de l'utilisation qui en est proposée pour la typologie, le nouvel algorithme décrit précédemment montre comment il peut être intéressant pour l'énumération de certains objets sur un graphe de substituer à la notion de recherche arborescente celle de construction progressive du graphe.

BIBLIOGRAPHIE

[1] C. BERGE, *Théorie des graphes et ses applications*, Dunod, 1958.

[2] M. BERRY et V. DEGOT, *Note sur les études de marché, critique des méthodes actuelles de segmentation et de typologie, vers une nouvelle méthode de typologie*. Publication École des Mines de Paris, 1970.

[3] G. DEMOUCRON, *Ensemble stable intérieurement d'un graphe*. Gestion, juillet-août 1968.

[4] J. L. GUIGOU, *Analyse économique et analyse multidimensionnelle*. Thèse complémentaire 14/11/72, Université de Dijon, Faculté des Sciences Économiques et de Gestion.

[5] J. C. HERTZ, *Quelques considérations sur les problèmes d'emploi du temps*, AFIRO n° 38, 1966.

[6] M. HUGUES, *Segmentation et typologie*, Bordas Management.

[7] B. ROY, *Algèbre moderne et théorie du graphe*. Dunod (particulièrement le chapitre VI, § A et B).

[8] B. ROY, *An algorithm for a general constrained set covering problem*, pp. 267-285 in Group Theory and computing, Académie Press, 1972.