APPLICATION A LA TYPOLOGIE

La typologie vise à constituer des groupes d'individus qui soient les plus similaires possibles, et de sorte que ces groupes soient aussi dissemblables que possible. Elle comporte cinq étapes :

- La collecte des données.
- Le calcul des proximités entre les éléments.
- La constitution des groupes en affectant chaque objet au groupe dont il est le plus proche.
- L'interprétation des résultats où chaque groupe est décrit à partir de ses principales caractéristiques.
- La validation des résultats pour déterminer la qualité de la solution obtenue.

En matière de typologie, les méthodes fondées sur l'agglomération progressive d'individus voisins, fournissent de manière simple et rapide des partitions de la population, mais le résultat peut être instable ou artificiel.

Dès lors une méthode qui procède par la recherche de cliques (sous graphe complet symétrique) a été proposée. Ainsi dans ce chapitre nous présentons un algorithme qui permet de déterminer l'ensemble des cliques maximales d'un graphe.

D'où dans ce qui suit nous considérons que les objets sont décrits comme des points dans un espace métrique. La relation entre deux objets utilise 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 construire 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 suivant :

	Calcul sur les sommets	Calcul sur les arêtes
Méthodes Non récurrentes	Typologie par accumulation	Typologie par concentration
Méthodes récurrentes	Méthode arborescente de B. Roy	Algorithme de recherche de clique maximale par les arêtes (1)

Fig. 1

(1) Méthode de V. Degot et J.M Hualde[7]

Les méthodes de typologie à l'aide de cliques opèrent en deux phases :

- Enumération des cliques maximales du graphe.
- Recherche de la couverture minimale du graphe à l'aide de cliques maximales.

Dans ce qui suit nous présentons uniquement l'algorithme de recherche de clique maximale par les arêtes.

4.1 La recherche des cliques à partir des arêtes

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

4.1.1 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 supplémentaire de U_{n+1} qui n'appartient pas à U_n. 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 à l'ensemble couple de sommets {A, B} si $a_i \cap b_j = \emptyset$).

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

$\alpha, \gamma \in CM_n$ (A)

$\beta, \delta \in CM_n$ (B)

$\alpha \cap \beta + \{A, B\} \subset \delta \cap \gamma + \{A, B\}$

$\alpha \cap \beta + \{A, B\} \subset \gamma \cap \beta + \{A, B\}$

$\alpha \cap \beta + \{A, B\} \subset \alpha \cap \delta + \{A, B\}$

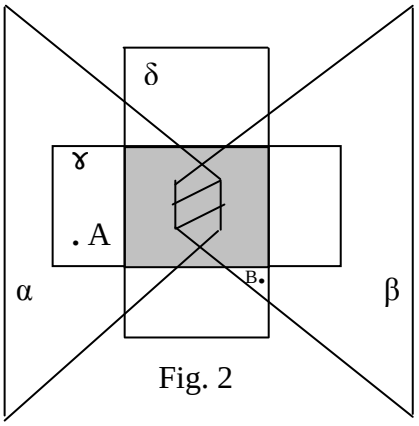

Fig. 2

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 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.

4.1.2 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 $CM_n(B) = \{b_j\}$, $j \in [1, q]$ 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. Soit CM^{K-1} l'ensemble des cliques retenues provisoirement maximales à l'étape précédente, nous allons constituer CM^K. Au début de l'étape K, on pose $CM^K = Ø$; on connaît par ailleurs certains éléments de CM_{n+1} par exemple ceux de CM_n ($\overline{AB}$).

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

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

2) Si $a_i \cap b_j \neq Ø$ 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, 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 $C_1,….,C_L$ de CM^{K-1} telles que $C_1,…..,C_L \subset a_i \cap b_j + \{A, B\}$ on pose :

- $CM^K = CM^{K-1} + [a_i \cap b_j + \{A, B\}] – [C_1,…, C_L]$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^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}).

4.1.3 La courbe reliant le nombre de cliques au seuil

Parmi les méthodes décrites précédemment, celles qui opèrent 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.

C'est le cas pour cette méthode. 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 de construire cette courbe. La figure 3 est un exemple d'un tel calcul.

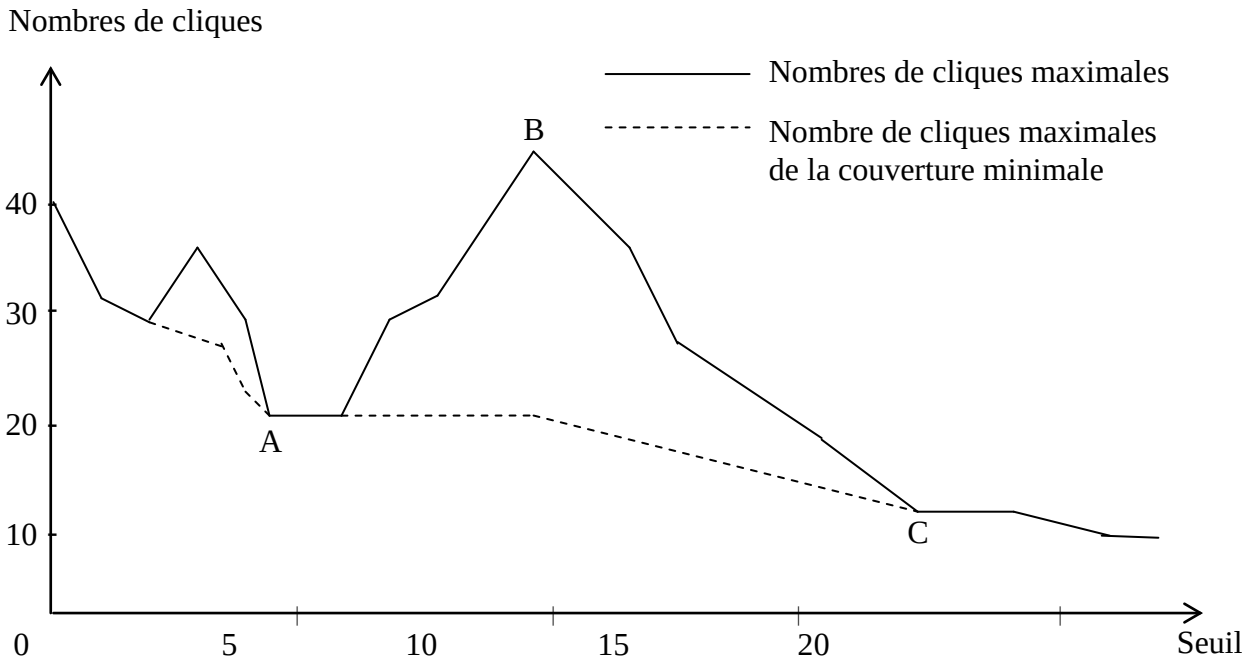

Figure. 3

De A vers B, l'ajout d'arêtes fait croître le nombre de 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 figure 4.

- A l'étape i, cliques maximales : I, II, III, IV, a, b, c, d.

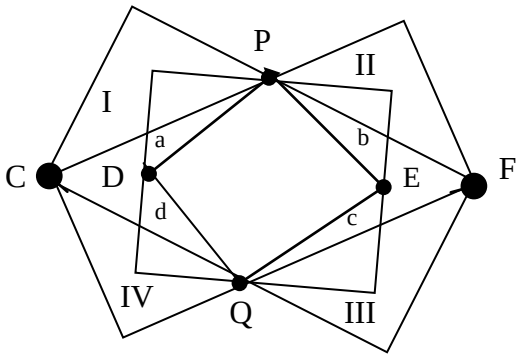

Fig. 4

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

- De B vers C, l'ajout d'arêtes fait décroître le nombre de cliques maximales. Il s'agit alors de l'agglomération rapide de cliques maximales selon le schéma suivant figure 5.
- Etape i, cliques maximales : 1, 2, 3, 4 couverture formée de deux cliques (par exemple 2,4 ou 1,3).

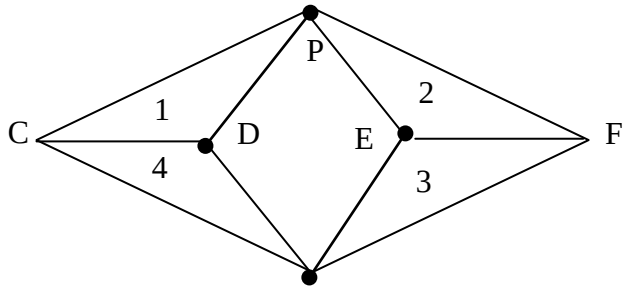

Fig.5

- Etape i+1 (ajout de (P, Q)), cliques maximales : PQCD, PQEF couverture formée de deux cliques.

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.

Remarque. Les typologies utilisant les cliques procèdent en deux étapes :

- Enumération des cliques maximales du graphe : pour un graphe donné l'énumération des cliques maximales est unique mais en général les cliques retenues ne forment pas une partition du graphe.
- Détermination de la couverture minimale : plusieurs ensembles de cliques peuvent être retenus comme couverture du graphe.

Exemple. Soit le graphe G de la figure 5.

- Ce graphe a 4 cliques maximales : 1, 2, 3, 4

- Pour sa couverture minimale on a le choix entre {1, 3} et {2, 4}. Nous allons maintenant décrire les principes généraux des méthodes par accumulation et par concentration.

4.2 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 ce n'est que dans le cas le plus favorable que les individus choisis aléatoirement pour former des centres de classe sont près des centres qui existent réellement dans la population.

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

On calcule les nouvelles distances 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 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. La faiblesse de ces méthodes provient donc du fait qu'elles opèrent d'une manière énumérative et que le concept de classe homogène y est mal défini.
- *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.

Par ailleurs nous présentons un autre algorithme sur les cliques : l'algorithme d'extraction d'une clique maximum dans le paragraphe suivant.

4.4 Algorithme d'extraction d'une clique maximum

La recherche se fait en deux fois. On recherche d'abord une clique maximale, puis on cherche s'il existe une clique plus grande. A chaque étape, on élimine les nœuds dont le degré n'est pas suffisant. Globalement, l'algorithme est le suivant :

Algorithme d'extraction d'une clique maximum[8]

On notera # E le nombre d'éléments d'un ensemble E.

#

Recherche d'une clique maximale C_m (maximum local)

(sous graphe complet tel que si on ajoute un nœud on perd la complétude)

#

trier l'ensemble des nœuds selon leur degré

C_m ⟵ {nœud de plus haut degré}

pour chaque nœud N, par degré décroissant

s'il existe une arête reliant N à tous les nœuds de C_m

C_m ⟵ $C_m \cup N$

#

Recherche d'une clique maximum C (maximum global)

#

Posons d_m le plus haut degré du graphe

C ⟵ C_m

Pour i de # C_m + 1 à d_m + 1

S'il existe une clique de taille i

C ⟵ une clique de taille i

Sinon

Retourner C

Exemple. Soit le graphe de la figure 6

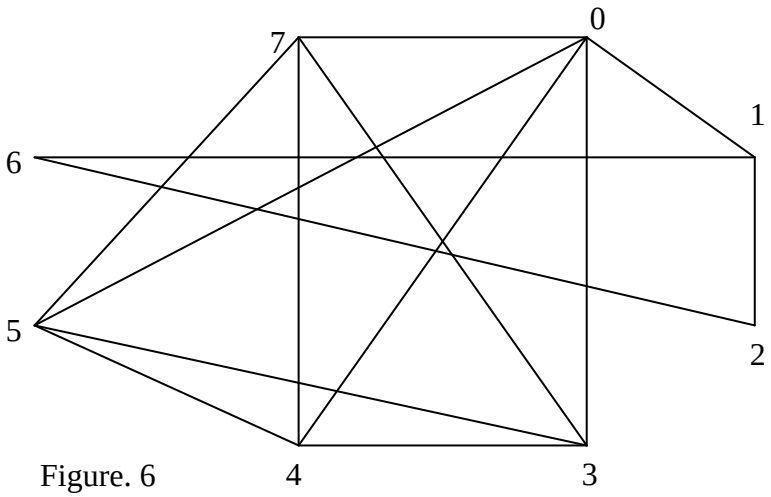

Figure. 6

Déterminons les degrés des sommets du graphe par ordre décroissant :

d (0) = 5

d (3) = d (4) = d (5) = d (7) = 4

d (1) = 3

d (2) = d (6) = 2

- Le sommet 0 a le plus haut degré d'où on a :

$C_m \longleftarrow \{0\}$

- On a d(3) = d(4) = d(5) = d(7) = 4 on choisit arbitrairement un sommet qui est lié au sommet 0, soit le sommet 3, d'où on a :

$C_m \longleftarrow C_m \cup \{3\}$

- Le sommet 4 est lié à tous les nœuds de $C_m = \{0, 3\}$ d'où on a

$C_m \longleftarrow C_m \cup \{4\}$

- Le sommet 5 est lié à tous les nœuds de $C_m = \{0, 3, 4\}$ d'où on a :

$C_m \longleftarrow C_m \cup \{5\}$

- Le sommet 7 est lié à tous les nœuds de $C_m = \{0, 3, 4, 5\}$ d'où on a :

$C_m \longleftarrow C_m \cup \{7\}$

- Le sommet 1 n'est pas lié à tous les sommets de C_m, de même que les sommets 2 et 6

D'où $C_m = \{0, 3, 4, 5, 7\}$ est une clique maximale, la recherche d'une clique maximale de ce graphe donne immédiatement la clique maximum.