Agrégation optimale sous contrainte de contiguïté :

aspects théoriques et mise en œuvre avec applications à des cas pratiques.

Marc CHRISTINE et Michel ISNARD (Insee, DCSRI)

1 Objectif de la procédure

La procédure décrite dans ce papier vise à permettre une agrégation d'une population comportant un nombre fini d'unités en groupes connexes dont le nombre a été fixé par l'utilisateur. Cette agrégation devra, si possible, vérifier des contraintes de taille (taille minimale et taille maximale identiques pour chaque groupe) et minimiser ou maximiser la variance intra-groupe d'une variable d'intérêt.

La procédure nécessite que l'on connaisse une *distance* (non nécessairement euclidienne) entre toutes les unités, un ensemble de *relations de contiguïté* entre unités qui permettra d'introduire le concept de *connexité* et, pour chaque unité, un *poids* qui intervient dans le calcul de la variance intra-groupe et une *taille* qui intervient dans les contraintes de taille.

Les différents termes ci-dessus seront définis dans la suite du papier.

Cet article reprend et enrichit une problématique déjà présentée aux JMS 2000[1]. Par rapport à ce précédent papier, cet article enrichit l'approche du fait de la possibilité de recourir à une distance non euclidienne [cf. §3] et il améliore les performances de l'algorithme mis en œuvre.

Toutes les procédures informatiques résultantes ont été écrites spécifiquement en langage SAS.

Les méthodes exposées seront mises en œuvre et illustrées sur des cas concrets. Un premier groupe d'exemples permettra de montrer comment construire une partition des communes au sein d'une région ou d'un département en un nombre fixé de classes connexes, en s'appuyant sur des critères de minimisation ou de maximisation de la variance intra-classe pour des variables quantitatives standards : ceci illustre le cas de la distance euclidienne.

Dans le second groupe d'exemples, on cherchera à appliquer ces techniques à des cas non euclidiens, par exemple la construction d'îlots de stabilité de l'emploi.

Le lecteur intéressé peut se procurer auprès de Michel Isnard (michel.isnard@insee.fr) la macro SAS utilisée et un guide d'utilisation.

2 Développements mathématiques

2.1 Cadre général

Considérons une population finie $\mathcal{P}$ de cardinal *N*, composée d'éléments appelés *unités statistiques*, indexées par un indice i. Sur chaque unité statistique, on suppose définis :

- Une variable d'intérêt x_i. Celle-ci doit être numérique.
- Un poids α_i (> 0).
- Une taille T_i (numérique, éventuellement entière).

[1] « Un algorithme de regroupement d'unités statistiques selon certains critères de similitude », Marc CHRISTINE et Michel ISNARD, VII[èmes] Journées de Méthodologie statistique, 4-5 décembre 2000.

Remarques sur la signification des variables.

- La variable x_i peut être de plusieurs natures différentes :

  - Ce peut être un *total* de variables élémentaires définies chacune sur des éléments de l'unité statistique i, considérée elle-même comme un ensemble.
    Exemple : le revenu total des ménages habitant la commune i, défini comme la somme des revenus de chaque ménage pris individuellement.

  - Ce peut être une *moyenne* de ces variables élémentaires définies chacune sur des éléments de l'unité statistique i.
    Exemple : le revenu moyen des ménages habitant la commune i, l'âge moyen des individus vivant dans la commune i.

  - Ce peut être un *ratio,* rapport de deux totaux de variables élémentaires définies chacune sur des éléments de l'unité statistique i.
    Exemples : le taux de chômage moyen des individus habitant la commune i, la densité, en nombre d'habitants au km², de cette commune.

  - Ce peut être une grandeur sans dimension.
    Exemple : la température en un point i du territoire à une date donnée.

- La variable T_i est une *mesure* de l'unité statistique i. Ce peut-être : le nombre d'éléments de l'unité statistique i, considérée elle-même comme un ensemble (nombre d'individus, nombre de ménages d'une commune, nombre de salariés d'un établissement), une longueur, une surface, un volume…

- Les variables α_i et T_i n'ont pas le même statut. Les variables α_i interviennent dans la définition du *centre de gravité et de l'inertie* de la population $\mathcal{P}$ relativement aux variables x_i (cf. infra), tandis que les variables T_i sont des *mesures* sur lesquelles on fera peser des contraintes. Néanmoins, il est possible de prendre $\alpha_i = T_i$. Cela prendra pleinement son sens lorsque l'unité statistique i est elle-même un ensemble et que T_i est son cardinal.

2.2 Centre de gravité, variance, inertie.

Il est d'usage de définir :

- le *centre de gravité* (pondéré) de la population $\mathcal{P}$ relativement aux variables d'intérêt x_i, qui sera la valeur : $\boxed{g = \dfrac{\sum_{i \in P} \alpha_i x_i}{\sum_{i \in P} \alpha_i}}$.

On a donc la relation : $(\sum_{i\in P}\alpha_i)g = \sum_{i\in P}\alpha_i x_i$, d'où : $\boxed{\sum_{i\in P}\alpha_i(x_i-g)=0}$.

- La *variance ou inertie*[2] de la population $\mathcal{P}$ relativement à ces mêmes variables :

$$\boxed{I=\frac{\sum_{i\in P}\alpha_i(x_i-g)^2}{\sum_{i\in P}\alpha_i}}.$$

2.3 Effet d'une partition de la population de référence

Supposons que la population $\mathcal{P}$ soit partitionnée en K *sous-populations* ou *groupes*[3], notées $\mathcal{P}_1$, $\mathcal{P}_2$, …, $\mathcal{P}_K$. Sur chaque groupe $\mathcal{P}_k$, on définit :

- le centre de gravité $g_k=\frac{\sum_{i\in P_k}\alpha_i x_i}{\sum_{i\in P_k}\alpha_i}$ [vérifiant la relation : $\sum_{i\in P_k}\alpha_i(x_i-g_k)=0$]
- et l'inertie $I_k=\frac{\sum_{i\in P_k}\alpha_i(x_i-g_k)^2}{\sum_{i\in P_k}\alpha_i}$.

Dans ce cas, l'inertie I relative à la population $\mathcal{P}$ sera qualifiée d'*inertie totale* pour la distinguer des inerties I_k.

On dispose alors de la classique équation d'**analyse de la variance** (ou théorème de HUYGHENS-KOENIG) :

Démonstration :

$$(\sum_{i\in P}\alpha_i)I=\sum_{i\in P}\alpha_i(x_i-g)^2$$

$$=\sum_{k=1}^{K}\sum_{i\in P_k}\alpha_i(x_i-g)^2$$

[2] Le terme « Inertie » sera préféré par la suite car il permet de traiter des cas plus généraux que celui du calcul de la décomposition de la variance d'une variable numérique définie sur les unités de la population $\mathcal{P}$.

[3] Il ne s'agit pas ici d'un groupe au sens opératoire du terme mais d'un sous-ensemble de $\mathcal{P}$.

$$= \sum_{k=1}^{K} \sum_{i \in P_k} \alpha_i (x_i - g_k + g_k - g)^2$$

$$= \sum_{k=1}^{K} \sum_{i \in P_k} \alpha_i [(x_i - g_k)^2 + 2(x_i - g_k)(g_k - g) + (g_k - g)^2]$$

$$= \sum_{k=1}^{K} \sum_{i \in P_k} \alpha_i (x_i - g_k)^2 + 2 \sum_{k=1}^{K} \sum_{i \in P_k} \alpha_i (x_i - g_k)(g_k - g) + \sum_{k=1}^{K} \sum_{i \in P_k} \alpha_i (g_k - g)^2$$

$$= \sum_{k=1}^{K} (\sum_{i \in P_k} \alpha_i) I_k + 2 \sum_{k=1}^{K} (g_k - g) \left[\underbrace{\sum_{i \in P_k} \alpha_i (x_i - g_k)}_{=0} \right] + \sum_{k=1}^{K} \left[(g_k - g)^2 (\sum_{i \in P_k} \alpha_i) \right]$$

$$= \sum_{k=1}^{K} \left[(\sum_{i \in P_k} \alpha_i)[I_k + (g_k - g)^2] \right].$$

D'où, en notant : $\omega = \sum_{i \in P} \alpha_i$ et $\omega_k = \sum_{i \in P_k} \alpha_i$:

$$I = \sum_{k=1}^{K} \left[\frac{\omega_k}{\omega} [I_k + (g_k - g)^2] \right].$$

Le terme $I^a = \sum_{k=1}^{K} \frac{\omega_k}{\omega} I_k$ constitue la variance (ou inertie) *intra-groupe* : il s'agit d'une moyenne pondérée des inerties propres à chacun des groupes P_k, les poids étant les ω_k.

Le terme $\sum_{k=1}^{K} \frac{\omega_k}{\omega} (g_k - g)^2$ représente la variance (ou inertie) *inter-groupes*, d'où la relation :

Inertie totale = Inertie intra-groupe + Inertie inter-groupes.

Ainsi, l'inertie totale étant une *constante de la population* (vis-à-vis du choix des variables x_i), la décomposition en les deux termes ci-dessus dépendra de la manière dont est construite la partition ; une partition pourra avoir une inertie intra-groupe plus élevée qu'une autre (donc une inertie inter-groupes *plus faible*) ou vice-versa.

Pour les cas limites, une inertie intra-groupe nulle signifie une *homogénéité parfaite* de chacun des groupes (au sein de chaque groupe, toutes les variables x_i prennent la même valeur), tandis qu'une

inertie inter-groupes nulle exprime que les centres de gravité de tous les groupes sont identiques. On peut alors parler de *similarité parfaite* des groupes, celle-ci étant mesurée par la distance du centre de gravité d'un groupe par rapport au centre de gravité de la population prise dans son ensemble.

Plus les groupes sont dissemblables, plus la variance inter-groupes est forte et vice-versa.

2.4 Passage à une partition de nombre de groupes inférieur.

On peut construire une partition *moins fine* qu'une partition donnée par agrégations successives des groupes qui la composent et réduire ainsi le nombre de groupes de la partition.

Voyons l'effet de l'agrégation de deux groupes sur la décomposition de l'inertie totale (qui reste constante) en étudiant, par exemple, la variation qui en résulte de l'inertie intra-groupe.

Supposons que l'on agrège les deux groupes $\mathcal{P}_{k_1}$ et $\mathcal{P}_{k_2}$, les autres restant inchangés. Notons $\mathcal{P}_{k_1k_2}$ le groupe résultant de cette agrégation, c'est-à-dire : $\mathcal{P}_{k_1k_2} = \mathcal{P}_{k_1} \cup \mathcal{P}_{k_2}$.

La variation d'inertie intra-groupe qui en résulte (elle ne met en jeu que les groupes $\mathcal{P}_{k_1}$ et $\mathcal{P}_{k_2}$) est :

$$\Delta I^a = \frac{\omega_{k_1} + \omega_{k_2}}{\omega} I_{k_1k_2} - \frac{\omega_{k_1}}{\omega} I_{k_1} - \frac{\omega_{k_2}}{\omega} I_{k_2},$$

où $I_{k_1k_2}$ est l'inertie intra-groupe du nouveau groupe $\mathcal{P}_{k_1k_2}$.

Or : $I_{k_1k_2} = \dfrac{\sum_{i \in P_{k_1} \cup P_{k_2}} \alpha_i (x_i - g_{k_1k_2})^2}{\sum_{i \in P_{k_1} \cup P_{k_2}} \alpha_i}$, où $g_{k_1k_2}$ est le centre de gravité du nouveau groupe $\mathcal{P}_{k_1k_2}$, soit :

$$g_{k_1k_2} = \frac{\sum_{i \in P_{k_1} \cup P_{k_2}} \alpha_i x_i}{\sum_{i \in P_{k_1} \cup P_{k_2}} \alpha_i} = \frac{\sum_{i \in P_{k_1}} \alpha_i x_i + \sum_{i \in P_{k_2}} \alpha_i x_i}{\sum_{i \in P_{k_1} \cup P_{k_2}} \alpha_i} = \frac{\left(\sum_{i \in P_{k_1}} \alpha_i\right) g_{k_1} + \left(\sum_{i \in P_{k_2}} \alpha_i\right) g_{k_2}}{\sum_{i \in P_{k_1} \cup P_{k_2}} \alpha_i} = \frac{\omega_{k_1} g_{k_1} + \omega_{k_2} g_{k_2}}{\omega_{k_1} + \omega_{k_2}}.$$

On constate que : $g_{k_1} - g_{k_1k_2} = g_{k_1} - \dfrac{\omega_{k_1} g_{k_1} + \omega_{k_2} g_{k_2}}{\omega_{k_1} + \omega_{k_2}} = \dfrac{\omega_{k_2} (g_{k_1} - g_{k_2})}{\omega_{k_1} + \omega_{k_2}}$.

Par suite :

$$I_{k_1k_2} = \frac{\sum_{i \in P_{k_1} \cup P_{k_2}} \alpha_i (x_i - g_{k_1k_2})^2}{\omega_{k_1} + \omega_{k_2}} = \frac{\sum_{i \in P_{k_1}} \alpha_i (x_i - g_{k_1k_2})^2 + \sum_{i \in P_{k_2}} \alpha_i (x_i - g_{k_1k_2})^2}{\omega_{k_1} + \omega_{k_2}}.$$

Mais :

$$\sum_{i\in P_{k_1}} \alpha_i (x_i - g_{k_1k_2})^2 = \sum_{i\in P_{k_1}} \alpha_i (x_i - g_{k_1} + g_{k_1} - g_{k_1k_2})^2$$

$$= \sum_{i\in P_{k_1}} \alpha_i (x_i - g_{k_1})^2 + 2\sum_{i\in P_{k_1}} \alpha_i (x_i - g_{k_1})(g_{k_1} - g_{k_1k_2}) + \sum_{i\in P_{k_1}} \alpha_i (g_{k_1} - g_{k_1k_2})^2$$

$$= \omega_{k_1} I_{k_1} + 2(g_{k_1} - g_{k_1k_2}) \underbrace{\sum_{i\in P_{k_1}} \alpha_i (x_i - g_{k_1})}_{=0} + (g_{k_1} - g_{k_1k_2})^2 \left(\sum_{i\in P_{k_1}} \alpha_i \right)$$

$$= \omega_{k_1} [I_{k_1} + (g_{k_1} - g_{k_1k_2})^2]$$

$$= \omega_{k_1} \left[I_{k_1} + \left(\frac{\omega_{k_2}(g_{k_1} - g_{k_2})}{\omega_{k_1} + \omega_{k_2}} \right)^2 \right].$$

On obtient, par symétrie, une formule analogue pour $\sum_{i\in P_{k_2}} \alpha_i (x_i - g_{k_1k_2})^2$.

Il en résulte :

$$I_{k_1k_2} = \frac{\omega_{k_1}}{\omega_{k_1} + \omega_{k_2}} \left[I_{k_1} + \left(\frac{\omega_{k_2}(g_{k_1} - g_{k_2})}{\omega_{k_1} + \omega_{k_2}} \right)^2 \right] + \frac{\omega_{k_2}}{\omega_{k_1} + \omega_{k_2}} \left[I_{k_2} + \left(\frac{\omega_{k_1}(g_{k_2} - g_{k_1})}{\omega_{k_1} + \omega_{k_2}} \right)^2 \right]$$

$$= \frac{\omega_{k_1} I_{k_1} + \omega_{k_2} I_{k_2}}{\omega_{k_1} + \omega_{k_2}} + (g_{k_1} - g_{k_2})^2 \frac{\omega_{k_1}\omega_{k_2}^2 + \omega_{k_2}\omega_{k_1}^2}{(\omega_{k_1} + \omega_{k_2})^3}$$

$$= \frac{\omega_{k_1} I_{k_1} + \omega_{k_2} I_{k_2}}{\omega_{k_1} + \omega_{k_2}} + (g_{k_1} - g_{k_2})^2 \frac{\omega_{k_1}\omega_{k_2}}{(\omega_{k_1} + \omega_{k_2})^2}.$$

Donc :

$$\Delta I^a = \frac{1}{\omega} [(\omega_{k_1} + \omega_{k_2}) I_{k_1k_2} - \omega_{k_1} I_{k_1} - \omega_{k_2} I_{k_2}],$$

soit :

$$\boxed{\Delta I^a = \frac{(g_{k_1} - g_{k_2})^2}{\omega} \frac{\omega_{k_1}\omega_{k_2}}{\omega_{k_1} + \omega_{k_2}}.} \quad \textbf{(1)}$$

Conséquence :

Par agrégation de deux groupes, **la variation d'inertie intra-groupe est toujours positive ou nulle, c'est-à-dire que l'inertie intra-groupe ne peut qu'augmenter dans une agrégation, et l'inertie inter-groupes diminuer.**

La question peut alors se poser, partant d'une partition donnée, de construire une nouvelle partition moins fine (et de nombre de groupes fixé), **minimisant l'inertie intra-groupe** (ou maximisant l'inertie inter-groupes). On mettra pour cela en œuvre un algorithme par agrégations successives **réalisant à chaque étape l'agrégation de deux groupes, de telle sorte que la variation d'inertie intra-groupe en résultant soit la plus faible possible.**

Un cas particulier est celui où la partition initiale est constituée des N singletons correspondant à chaque unité statistique prise isolément pour former un groupe. Il s'agit de la partition la plus fine possible qui constituera en général le point de départ d'un algorithme consistant à agréger progressivement les unités entre elles jusqu'à atteindre le nombre de groupes voulus. A chaque étape, seront choisis les deux groupes dont l'agrégation minimise l'inertie intra-groupe résultante (cf. § 4).

Un problème symétrique peut être la maximisation de l'inertie intra-groupe et l'algorithme correspondant sera une transposition du précédent en cherchant à assurer à chaque étape la variation d'inertie intra-groupe *la plus grande* possible.

Exemple : constituer une partition en groupes de tailles les plus semblables possible. Ici, on prendra $x_i = T_i$, $\alpha_i = 1$. L'inertie inter-groupes vis-à-vis de ces variables doit être la plus faible possible, donc l'inertie intra-groupe la plus grande.

2.5 Le problème de la contiguïté.

2.5.1 Position du problème

Le cœur et l'originalité de la problématique décrite ici, ainsi que l'une des principales complexités de la mise en œuvre de la procédure, résident dans le fait que l'on va se restreindre à chercher des *partitions connexes* de la population de référence $\mathcal{P}$.

A l'origine de cette problématique se trouve la question du partitionnement d'un territoire géographique composé d'unités statistiques qui peuvent être des communes ou des éléments de bâti ou des logements, caractérisés par un positionnement topologique, linéaire, ou le plus souvent spatial. Ce positionnement sera appréhendé par le concept de *contiguïté* : on construira les groupes par agrégations successives d'unités contiguës entre elles de façon qu'elles forment des sous-ensembles « d'un seul tenant ».

L'une des applications de cette idée est celle de la constitution d'unités primaires géographiques dans un plan de sondage à plusieurs degrés. Dans le cas où ces unités primaires correspondent à la zone d'action d'un enquêteur, qui doit se déplacer au cours des différentes enquêtes sur une portion de territoire relativement limitée, ce afin de réduire les coûts de déplacement, la connexité (jointe à des considérations de taille des unités) sera un moyen de répondre à ces impératifs pratiques.

Un autre intérêt réside dans la *cartographie* : construire et représenter des portions de territoire présentant des homogénéités fortes vis-à-vis de variables descriptives données permettra des interprétations plus faciles dès lors que ces parties de territoire seront connexes.

Les concepts utilisés ici peuvent être formalisés mathématiquement. On trouvera ce formalisme en annexe.

2.5.2 Application au cas d'une population finie.

Dans le cas d'une population de référence Φ de cardinal *N*, on peut exprimer la contiguïté entre éléments (ou unités statistiques) au moyen d'une *matrice de contiguïté*, de taille [*N, N*].

Une telle matrice C aura pour éléments $C_{i,j}$ définis par :

$C_{i,j}$ = 1 si les unités i et j sont contiguës

0 sinon.

La *symétrie* de la relation de contiguïté entraîne que cette matrice est *symétrique* et la *réflexivité* de la relation se traduit par le fait que la *matrice a des 1 sur la diagonale*.

Un sous-ensemble G, non vide, de la population de référence est assimilé à un vecteur ligne de taille [1, *N*] dont les composantes valent : G_i = 1 si l'unité i appartient à G, 0 sinon.

Enfin, une partition de Φ en K groupes sera assimilée à une matrice P de taille [K, *N*], dont les éléments sont : $P_{k,i}$ = 1 si l'unité i appartient au groupe Φ_k

0 sinon.

Une telle matrice possède un et un seul 1 par colonne (car chaque unité appartient à un et un seul groupe Φ_k) et au moins un 1 par ligne (car chaque groupe Φ_k est non vide).

2.5.3 Principaux résultats

Les principaux résultats sur la connexité ont été démontrés dans [1][4] :

Proposition 1 :

Soient deux groupes G et H inclus dans $\mathcal{P}$. Les deux propositions suivantes sont équivalentes :

1.	G et H sont contigus.
2.	(G C H') > 0.
	[H' est la transposée de H]

Proposition 2 :

Soit G un groupe inclus dans $\mathcal{P}$, comportant *q* éléments. Notons C^G la matrice de contiguïté restreinte aux *q* éléments de ce groupe (de taille [*q*, *q*]).

Alors :

$\forall\ n \in \mathbb{N}^*$: $\left(C_{i,j}^G\right)^{(n)} > 0 \iff$ il existe un chemin de longueur *n* entre les unités i et j.

On notera $\left(C_{i,j}^G\right)^{(n)}$ le terme d'indices (i, j) de la matrice $(C^G)^n$. La *stricte positivité d'une matrice* est définie par le fait que tous ses éléments sont > 0.

Enfin, pour $n \geq 2$, un chemin de longueur *n* entre les unités i et j sera une suite finie $\{i_1, i_2, ..., i_{n-1}\}$ d'éléments de G (distincts ou non des unités i ou j) telle que : $i \to i_1 \to i_2 \to ... \to i_{n-1} \to j$. La flèche représente ici la relation de contiguïté entre deux éléments.

Proposition 3 : **test de la connexité d'un groupe**

Avec les mêmes notations que dans la proposition 2, on cherche ici à ***déterminer si le groupe G constitué de q éléments est connexe ou pas***.

G est connexe si et seulement s'il existe un chemin entre deux quelconques de ses points constitué d'éléments de G. Il suffit évidemment que cette propriété soit vraie pour deux points *distincts* quelconques.

---

4 [1] Un algorithme de regroupement d'unités statistiques selon certains critères de similitude, Marc CHRISTINE, Michel ISNARD, JMS 2000, *in* Insee Méthodes n°100, Tome 2.

Si ce chemin est de longueur supérieure ou égale à *q*, le nombre de points intermédiaires sera d'au moins *q* - 1, soit *q* + 1 points en comptant les extrémités : le chemin passe donc au moins deux fois par un même point de G, c'est-à-dire constitue une *boucle*.

En éliminant l'ensemble des boucles, on peut obtenir un chemin de longueur inférieure ou égale à *q* - 1. Si sa longueur n'est pas égale à *q* - 1, on peut ensuite rajouter le nombre adéquat de boucles sur le premier élément pour obtenir un chemin de longueur strictement égale à *q* - 1.

On obtient donc :

$$\text{G connexe} \Leftrightarrow \forall i, j : C_{i,j}^{G(q-1)} > 0,$$

c'est-à-dire que tous les éléments de la matrice $(C^G)^{q-1}$ sont strictement positifs.

2.6 Le problème de la taille

La variable T_i sert, si l'utilisateur le désire, à mettre des contraintes sur les groupes à constituer dans une partition. En particulier, l'utilisateur peut spécifier que les tailles de tous les groupes doivent être comprises entre une taille minimale et une taille maximale. La taille d'un groupe G sera évidemment définie par : $T(G) = \sum_{i \in G} T_i$.

Ces contraintes sont *uniformes* (elles s'appliquent de manière identique à tous les groupes) et peuvent être relâchées par l'utilisateur en spécifiant une taille minimale nulle et une taille maximale égale à la taille totale de la population.

3 Généralisations à des cas non-euclidiens

3.1 Interprétation de l'inertie en termes de distances entre unités.

On reprend le contexte et les notations du § 2.

Considérons la quantité $D = \sum_{i,j \in P} \alpha_i \alpha_j (x_i - x_j)^2$. Elle peut s'écrire :

$$D = \sum_{i,j \in P} \alpha_i \alpha_j (x_i - g + g - x_j)^2$$

$$= \sum_{i,j \in P} \alpha_i \alpha_j [(x_i - g)^2 + 2\,(x_i - g)(x_j - g) + (x_j - g)^2]$$

$$= (\sum_{j \in P} \alpha_j) \sum_{i \in P} \alpha_i (x_i - g)^2 + 2 \sum_{i,j \in P} \alpha_i \alpha_j (x_i - g)(x_j - g) + (\sum_{i \in P} \alpha_i) \sum_{j \in P} \alpha_j (x_j - g)^2]$$

$$= \omega(\omega I) + 2 \left[\underbrace{\sum_{i \in P} \alpha_i (x_i - g)}_{=0} \right] \left[\underbrace{\sum_{j \in P} \alpha_j (x_j - g)}_{=0} \right] + \omega(\omega I) ,$$

d'où :

$$\boxed{D = 2\omega^2 I}, \text{ soit : } \boxed{I = \frac{1}{2\,\omega^2}\sum_{i,j\in P}\alpha_i\alpha_j(x_i - x_j)^2}.$$

De même, au sein d'un groupe $\mathcal{P}_k$: $I_k = \frac{1}{2\,\omega_k^2}\sum_{i,j\in P_k}\alpha_i\alpha_j(x_i - x_j)^2$.

La variance intra-groupe pourra alors s'écrire : $I^a = \sum_{k=1}^{K}\frac{\omega_k}{\omega}\frac{1}{2\,\omega_k^2}\left[\sum_{i,j\in P_k}\alpha_i\alpha_j(x_i - x_j)^2\right]$, soit :

$$\boxed{I^a = \frac{1}{2\,\omega}\sum_{k=1}^{K}\frac{1}{\omega_k}\left[\sum_{i,j\in P_k}\alpha_i\alpha_j(x_i - x_j)^2\right]}. \textbf{ (2)}$$

Dans cette expression, le terme $\left|x_i - x_j\right|$ peut s'interpréter comme une distance[5] entre les deux unités i et j : $d_{i,j} = \left|x_i - x_j\right|$.

3.2 Généralisation.

Il est possible de généraliser la notion d'inertie en s'affranchissant du contexte de calcul de la variance d'une variable d'intérêt x_i. Pour cela, on va introduire une pseudo-distance d entre les unités de la population de référence, non nécessairement euclidienne. Il s'agit plus précisément d'une fonction à valeurs positives ou nulles et non nécessairement nulle lorsqu'on on mesure la pseudo-distance d'une unité à elle-même.

En supposant toujours définis des poids α_i sur chaque unité i de la population, l'inertie intra-groupe, pour une partition de la population $\mathcal{P}$ en K groupes, sera donnée, par analogie formelle avec la formule **(2)**, par :

$$\boxed{I^a = \frac{1}{2\,\omega}\sum_{k=1}^{K}\frac{1}{\omega_k}\left[\sum_{i,j\in P_k}\alpha_i\alpha_j\, d_{i,j}^2\right]}. \textbf{ (3)}$$

.

[5] A strictement parler, on n'obtient de vraie distance qu'en la considérant définie sur l'ensemble-quotient de la population $\mathcal{P}$ par la relation d'équivalence : $i \sim j \Leftrightarrow \left|x_i - x_j\right| = 0.$ On convient ici que deux unités i et j ne sont pas discernables dès lors que $\left|x_i - x_j\right| = 0.$

Cette formule redonne l'expression de la variance intra-groupe pour les variables d'intérêt x_i, lorsque la distance considérée entre les deux unités i et j est : $d_{i,j} = \left| x_i - x_j \right|$.

Avec une telle expression, on ne disposera cependant plus des identités développées dans les paragraphes précédents. En particulier, on ne dispose plus de la notion de centre de gravité qui permettait d'écrire facilement la variation d'inertie résultant d'une agrégation de deux groupes. Cette variation ne peut être calculée qu'en appliquant la formule ci-dessus trois fois (une fois pour chaque groupe initial et une fois pour le groupe résultant) et en faisant le calcul direct de la variation d'inertie.

3.3 1ère application : construction de zones homogènes vis-à-vis des déplacements domicile-travail.

Sur un territoire géographique (département, région…), qui constituera la population de référence $\mathcal{P}$, on dispose d'unités statistiques qui sont les communes. On cherche à constituer une partition du territoire en bassins ou *îlots* d'emploi, considérés comme des sous-ensembles de communes astreints à des contraintes de taille et de connexité [cf. § 2.5], qui représenteraient des zones « de stabilité géographique » des migrations domicile-travail.

Dans un schéma de partition idéal, le territoire serait partitionné en zones connexes $\mathcal{Z}_k$ au sein desquelles tous les actifs y résidant y travailleraient également : aucun individu résidant dans une zone $\mathcal{Z}_k$ de la partition ne travaillerait à l'extérieur de cette zone.

Désignons par :

- $A_{i,j}$ = nombre d'actifs occupés habitant la commune i et travaillant dans la commune j.
- $A_{i,\bullet}$ = nombre d'actifs occupés habitant la commune i : $A_{i,\bullet} = \sum_{j \in Q} A_{i,j}$.

  Q désigne ici un univers de référence, sur-ensemble de $\mathcal{P}$. En effet, il se peut que des individus vivant dans la commune i travaillent à l'extérieur de la zone de référence $\mathcal{P}$.

  De ce fait, on *n'a pas* la relation : $\sum_{i \in P} A_{i,\bullet} = \sum_{i,j \in P} A_{i,j}$ *mais seulement :* $\sum_{i \in P} A_{i,\bullet} \geq \sum_{i,j \in P} A_{i,j}$.

Les $A_{i,j}$ représentent des *flux* et, plus généralement, l'exemple considéré peut s'intégrer dans un ensemble plus vaste de cas de figure où l'on définit des flux entre deux unités statistiques quelconques. On notera que ces flux ne sont pas nécessairement symétriques et qu'ils ne vérifient pas non plus la condition : $A_{i,i} = 0$.

Une zone « parfaite » $\mathcal{Z}$ serait alors un « îlot de stabilité » si et seulement si : $\dfrac{\sum_{i \in Z} \sum_{j \in Z} A_{i,j}}{\sum_{i \in Z} A_{i,\bullet}} = 1$ (1 représentant la valeur maximale du quotient figurant à gauche de cette égalité), c'est-à-dire lorsque la proportion des actifs occupés habitant et travaillant dans la zone représente 100% du nombre total d'actifs occupés vivant dans la zone (ou lorsque la totalité des flux émis à partir des unités statistiques de la zone sont dirigés vers des unités de la même zone).

Dans la pratique, on ne trouvera pas de zone parfaite en ce sens, même le territoire de référence ne l'est pas. On va alors définir l'inertie (intra-zone) de la zone $\mathcal{Z}$ par : $I(Z) = \dfrac{\sum_{i \in Z} \sum_{j \in Z} A_{i,j}}{\sum_{i \in Z} A_{i,\bullet}}$ et on va chercher à maximiser la somme des inerties intra-zone sur tous les groupes constitués, soit, si le territoire est partitionné en K groupes :

$$\boxed{J = \sum_{k=1}^{K} \left(\frac{\sum_{i \in Z_k} \sum_{j \in Z_k} A_{i,j}}{\sum_{i \in Z_k} A_{i,\bullet}} \right)}.$$

Intuitivement, cette fonction apparaît comme la somme, prise sur toutes les zones constituées, des % représentant la part des actifs résidant dans chaque zone qui travaillent dans la même zone.

Cette fonction à maximiser **s'interprète bien comme une inertie au sens où elle a été généralisée au § 3.2** (formule **(3)**). En effet, au facteur $\dfrac{1}{2\omega}$ près, elle peut s'écrire sous la forme :

$$\frac{1}{2\omega} J = I^a = \frac{1}{2\omega} \sum_{k=1}^{K} \frac{1}{\omega_k} \left[\sum_{i,j \in Z_k} \alpha_i \alpha_j \, d_{i,j}^2 \right],$$

en prenant pour paramètres :

$$\begin{cases} \alpha_i = A_{i,\bullet} \\ d_{i,j} = \sqrt{\dfrac{A_{i,j}}{A_{i,\bullet} A_{j,\bullet}}} \\ \omega_k = \sum_{i \in Z_k} \alpha_i = \sum_{i \in Z_k} A_{i,\bullet} \end{cases}.$$

La maximisation de J équivaut à celle de I^a, donc est justiciable de méthodes analogues à celles du § 2.4 pour déterminer une partition optimale du territoire $\mathcal{P}$ en termes d'inertie.

3.4 2ème application : le problème de la constitution des Zones d'Action Enquêteurs.

Un problème classique en statistique d'enquête est celui de l'affectation optimale des unités échantillonnées (en général des logements, des individus, plus rarement des établissements) entre différents enquêteurs.

Différents algorithmes ont déjà été proposés pour résoudre ce problème (voir notes internes Insee-UMS-ménages]. On va montrer ici qu'il constitue un cas d'application de la méthode développée dans cet article.

Considérons par exemple un échantillon de *N* logements, qui constituera la population de référence $\mathcal{P}$. Il s'agit de répartir les logements de cet échantillon entre K enquêteurs. L'affectation de l'échantillon entre les différents enquêteurs conduira à définir une partition de $\mathcal{P}$ en K *groupes* $\mathcal{P}_1$, $\mathcal{P}_2$, …, $\mathcal{P}_K$. Chacun de ces groupes sera affecté à un enquêteur, d'où le nom de *« Zones d'Action Enquêteurs »*[6].

Ce problème trouve son intérêt dans les situations où l'échantillon de logements n'est pas tiré dans des zones *affectées a priori* aux enquêteurs mais est sélectionné - peu importe comment - sur tout un territoire couvert par un réseau d'enquêteurs qui se partageront l'échantillon.

Les logements échantillonnés et les enquêteurs sont localisés dans l'espace. On peut résumer cette localisation par la distance d_{i,E_k} (≥ 0[7]) entre le logement i et l'enquêteur E_k. Ce peut être une distance euclidienne à vol d'oiseau, ou une pseudo-distance par la route ou un indicateur généralisé de coût : durée nécessaire pour que l'enquêteur E_k se rende au logement i, consommation d'essence, frais de déplacements engagés et / ou remboursés à l'enquêteur…).

L'optimalité recherchée.

Notons $\mathcal{P}_k$ l'ensemble des logements attribués à l'enquêteur E_k. On cherchera alors à minimiser un critère de type « coût total », soit : $\underset{P_1,P_2,\ldots,P_K}{Min} \sum_{k=1}^{K} \sum_{i\in P_k} d_{i,E_k}$.

S'il n'y avait pas d'autres contraintes, une solution consisterait à affecter chaque logement i à l'enquêteur dont il est le plus proche. Cela conduit à construire les groupes $\mathcal{P}_k$ comme suit :

Pour tout k ∈ {1, 2, ...K} : $\boxed{\mathcal{P}_k = \{i \in \mathcal{P};\ d_{i,E_k} = \underset{E_l}{Min}\ d_{i,E_l}\}}$.

[6] La difficulté supplémentaire, qui n'entre pas dans le champ de cet article, est que la population de référence $\mathcal{P}$ est, dans ce contexte, issue en général d'un échantillon *aléatoire*, susceptible de changer à chaque tirage d'échantillon, alors que l'ensemble des enquêteurs est fixe et que ceux-ci sont localisés de manière prédéterminée : les Zones d'Action Enquêteurs devront en fait réaliser une partition géographique *déterministe* du territoire couvert par l'ensemble potentiel de tous les échantillons.

[7] Par convention, les distances, quelle qu'en soit la signification concrète précise, sont comptées entre les communes de localisation du logement enquêté et de l'enquêteur. Par conséquent, la distance est conventionnellement égale à 0 si le logement enquêté se situe dans la commune de résidence de l'enquêteur auquel il a été attribué.

On peut supposer, pour simplifier, que les valeurs d_{i,E_k} sont toutes distinctes, ce qui entraîne que, pour chaque i dans $\mathcal{P}$, il y a un seul E_k qui réalise le minimum. Ainsi, il n'y a pas d'ambiguïté dans la définition des groupes $\mathcal{P}_k$ et ceux-ci sont disjoints[8].

Si ce n'était pas le cas, il faudrait modifier légèrement la définition des $\mathcal{P}_k$ en introduisant par exemple une contrainte d'ordre lexicographique :

$$\mathcal{P}_k = \{ i \in \mathcal{P};\ d_{i,E_k} = \underset{E_l}{Min}\ d_{i,E_l} \text{ et : } \forall\ m < k\ :\ d_{i,E_m} > d_{i,E_k} \}.$$

On obtient donc ainsi un optimum absolu au critère de distance. Mais des contraintes de « charge » doivent être prises en compte. Par ailleurs, certains enquêteurs seraient inemployés (cas où $\mathcal{P}_k = \varnothing$).

Les contraintes.

Chaque enquêteur E_k a une charge d'enquête (égale au nombre de logements qu'il peut ou doit interroger), notée C_k. Bien entendu, les charges doivent être a priori adaptées à la taille de l'échantillon. Les charges devant être supérieures ou égales à 1 pour tout enquêteur, ceci entraîne en particulier : $N \geq K$.

Souvent on raisonne en termes de bornes minimale et maximale : $C_{Min} \leq C_k \leq C_{Max}$.

Ceci entraîne :

$$K\, C_{Min} \leq N \leq K\, C_{Max}.$$

Si cette contrainte ne peut être satisfaite compte tenu des bornes relatives à chaque enquêteur, il faut soit modifier les limites de charge, soit modifier le nombre et la localisation du réseau d'enquêteurs pour que le problème ait une solution.

En tout état de cause, la solution optimale définie ci-dessus ne satisfait pas nécessairement ces contraintes et rend nécessaire une autre approche.

Une solution dans le cadre de la CAH contiguë.

On peut mettre en œuvre, pour résoudre le problème, les méthodes développées dans ce papier.

On suppose que les logements échantillonnés sont représentés comme des points d'un plan et localisés au moyen de leurs coordonnées : au logement i est associé le point M_i. L'objectif sera que les logements affectés à un enquêteur E_k soient les plus proches possibles les uns des autres et si possible pas trop éloignés de l'enquêteur, de manière à minimiser les coûts de déplacement.

On peut alors traiter ce problème dans le contexte de la contiguïté :

[8] Il ne forment cependant pas nécessairement une partition stricto sensu, certains d'entre eux pouvant être vides.

Pour un seuil de distance s donné, on définira la contiguïté entre deux logements i et j par :

$$\boxed{i \ R \ j \Leftrightarrow \exists \ E_k : d_{i,E_k} < s \ et \ d_{j,E_k} < s}.$$

[Mathématiquement : les points M_i et M_j appartiennent à un même cercle centré sur le point représentatif de l'un des enquêteurs].

D'autres définitions de la contiguïté seraient envisagables, par exemple :

$$\boxed{i \ R \ j \Leftrightarrow \begin{cases} \exists \ E_k : d_{i,E_k} < s \ et \ d_{j,E_k} < s \\ ou \\ M_i M_j < t \end{cases}},$$

où t est un autre seuil de distance.

On cherchera alors à *minimiser* une fonction du type de celle de la formule **(3)** :

$$I^a = \frac{1}{2\omega} \sum_{k=1}^{K} \frac{1}{\omega_k} \left[\sum_{i,j \in Z_k} \alpha_i \alpha_j \, d_{i,j}^2 \right],$$

en prenant pour paramètres :

$$\begin{cases} \alpha_i = 1 \\ d_{i,j} = \sqrt{M_i M_j} \\ \omega_k = \sum_{i \in P_k} \alpha_i \end{cases}.$$

Les variables de taille T_i seront égales à 1, les contraintes de charge pour chaque enquêteur s'exprimant alors en termes de contraintes de taille pour les groupes $\mathcal{P}_k$.

Les valeurs de s et de t doivent être adaptées pour que le problème ait une solution.

4 Description de l'algorithme

L'algorithme présenté ici comporte une variante visant à le rendre plus rapide (au prix éventuellement d'une certaine perte d'efficacité). Pour des raisons de clarté, cette variante n'est pas décrite dans le corps de l'algorithme, mais en fin de chapitre. Son embranchement est indiqué de la manière suivante : *variante 1*.

L'algorithme se déroule en deux phases :

- La première phase vise à effectuer une première agrégation de la population de référence en un nombre de groupes connexes choisi par l'utilisateur, respectant du mieux possible (cf. infra sur le relâchement des contraintes) les contraintes de taille fixées et maximisant ou minimisant la variance intra-groupe.

  Ou (*variante 1*)

- La seconde étape améliore la partition trouvée lors de la première étape en améliorant si nécessaire le respect des contraintes de taille et en optimisant la variance intra-groupe. Cette étape est rendue nécessaire parce que la méthode utilisée dans la première partie n'est pas optimale.

4.1 La première phase : une classification ascendante hiérarchique contiguë

La méthode utilisée ici est une méthode directement inspirée par la Classification Ascendante Hiérarchique (CAH) qui consiste, à une étape t de l'algorithme, à regrouper les deux groupes précédemment créés les plus « proches », c'est-à-dire minimisant l'augmentation de la variance intra-groupe liée à leur agrégation.

Par rapport à la CAH classique, cet algorithme a été modifié comme suit :

- L'agrégation ne peut avoir lieu qu'entre groupes contigus, puisque l'on veut obtenir une partition en groupes connexes ;
- Selon le choix de l'utilisateur, les groupes agrégés correspondent soit à une augmentation d'inertie maximale (si l'utilisateur souhaite maximiser l'inertie intra-groupe), soit minimale (dans le cas contraire).
- Le traitement des contraintes de taille sera décrit ci-dessous.

Cet algorithme récursif est initialisé avec la partition dont chaque groupe est un singleton correspondant à chaque unité de base et il s'arrête lorsque le nombre de groupes est égal à 1[9] ou lorsqu'à une étape donnée, il n'est pas possible d'agréger deux groupes quelconques sans que la taille du groupe résultant excède la taille maximale fixée par l'utilisateur[10]. On notera que, si au cours d'une itération, l'agrégation optimale (du point de vue de la variation d'inertie) conduit à constituer un groupe de taille supérieure à la taille maximale, cette agrégation n'a pas lieu et on recherche l'agrégation optimale *seulement sous contrainte d'éligibilité de la taille résultante*. Ainsi, on privilégie les contraintes de taille par rapport à l'optimalité de l'inertie.

[9] De ce fait, l'utilisateur dispose de toutes les partitions issues de la CAH, quel que soit le nombre de groupes constitué, et peut choisir celle correspondant au nombre de groupes souhaité.

[10] Cette contrainte de taille ne s'applique que si le nombre de groupes atteint à une étape donnée est supérieur ou égal au nombre fixé par l'utilisateur et que des agrégations ultérieures sont encore nécessaires pour parvenir à ce dernier nombre.

Dans le cas où la procédure ne peut se poursuivre à cause des contraintes de taille, un message est adressé à l'utilisateur et une taille maximale plus grande de 25% est fixée temporairement, puis l'algorithme se poursuit avec ce nouveau paramètre[11].

Si le nombre de composantes connexes de la population de référence est inférieur ou égal au nombre de groupes fixé par l'utilisateur, la première phase aboutit toujours à une agrégation en groupes connexes, mais ne respectant pas nécessairement les contraintes de taille fixées par l'utilisateur. Dans le cas où le nombre de composantes connexes est supérieur au nombre de groupes demandé par l'utilisateur, un message lui est adressé et aucune opération n'est conduite.

Le nombre de composantes connexes est calculé de la manière suivante : en partant d'une unité élémentaire, on construit l'ensemble des éléments qui sont reliés à cet élément initial par un chemin, quelle qu'en soit la longueur. Lorsque cette opération n'apporte plus d' « éléments nouveaux », l'ensemble des éléments ainsi reliés constitue une composante connexe. Il suffit de répéter cette étape jusqu'à ce que toutes les unités soient incluses dans une composante connexe pour avoir le nombre de composantes connexes (et leur composition).

Cette phase peut être relativement longue si le nombre d'unités de la population de référence est important.

4.1.1 La variante

- La *variante 1* consiste à ne pas mettre en œuvre la première phase. Il suffit que l'utilisateur fournisse une partition initiale en un nombre de groupes adéquat ne satisfaisant pas forcément les contraintes de taille. Le programme vérifie alors que cette agrégation ne comporte que des groupes connexes (dans le cas contraire, il s'arrête) et passe directement à la seconde phase. Si le nombre initial de groupes fourni par l'utilisateur est strictement supérieur au nombre de groupes final souhaité, une phase 1 réduite est mise en œuvre à partir de cette partition initiale.

4.2 La seconde phase : échanges ou transfert d'unités entre groupes

Sous l'ensemble des contraintes mentionnées ci-dessus (recours à un algorithme de type « CAH), contraintes de contiguïté et de taille), un optimum local est atteint.

Toutefois, comme nous l'avons indiqué ci-dessus, les contraintes initiales de taille fixées par l'utilisateur peuvent ne pas être respectées.

La seconde phase va donc chercher à réduire, en premier lieu, l'écart des tailles effectives des groupes constitués par rapport aux contraintes de taille, puis ensuite à optimiser l'intertie intra-groupe. Elle va procéder par transferts d'unités d'un groupe à un autre ou par échanges de deux unités appartenant à deux groupes différents.

Pour tester les contraintes de taille, une fonction additive, dite « critère », va être utilisée. Elle vaut, pour chaque groupe :

0 si la taille du groupe vérifie les contraintes de taille

ou la valeur absolue de la différence entre la taille du groupe et la borne de taille la plus proche, soit, pour un groupe G ne vérifiant pas les contraintes :

[11] A nouveau, ce relâchement de contraintes ne s'applique pas si la contrainte survient à une étape où l'on a déjà réduit le nombre de groupes obtenu par rapport au nombre souhaité.

$$CRIT(G) = \begin{cases} (taille(G) - T_{\max}) & si\ taille(G) > T_{\max} \\ ou & \\ (T_{\min} - taille(G)) & si\ taille(G) < T_{\min} \end{cases}$$

La méthode est relativement simple : pour tous les échanges ou transferts possibles d'unités, les influences sur les contraintes de taille et la variation de l'inertie sont calculées, puis on procède à l'échange qui diminue le plus le critère ou, si les contraintes de taille sont vérifiées, qui optimise le plus l'inertie intra-groupe.

Naturellement, un échange ou un transfert ne sont licites que si les groupes en résultant sont tous les deux connexes. Il convient donc de vérifier cette connexité avant de calculer les variations de critère ou de variance intra-groupe.

Cette seconde phase s'arrête lorsqu'il n'est plus possible d'améliorer le critère ou l'inertie intra-groupe. Chacune de ces itérations pouvant être longue, l'utilisateur peut en fixer un nombre maximal (plusieurs centaines).

5 Les résultats présentés

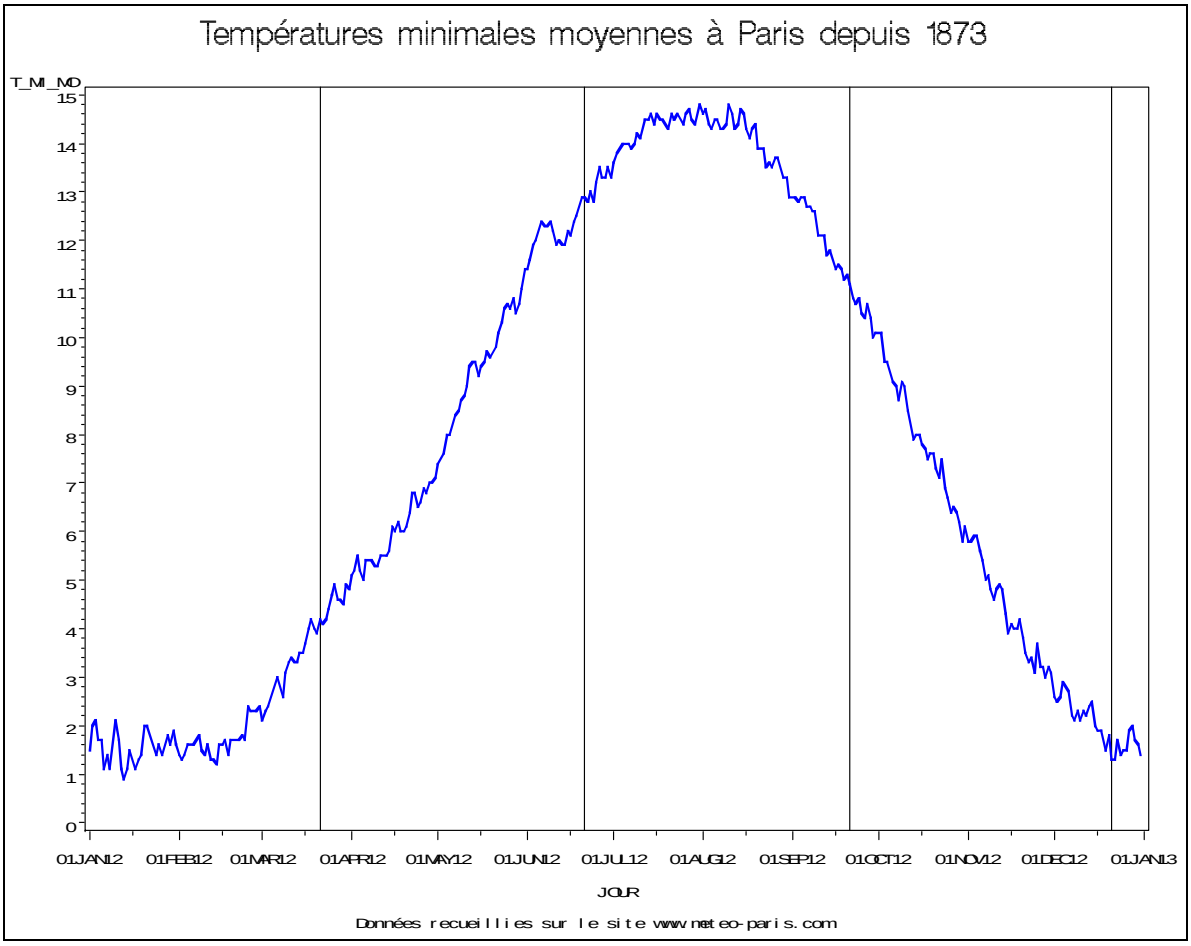
Ce chapitre reprend les différents résultats inclus dans la présentation orale et commente les différents points saillants de chacun.

5.1 Comment créer des saisons ?

Le premier exemple est un exemple simple. Il a le mérite de permettre la comparaison avec la valeur optimale qui n'est pas possible dans les autres cas.

Si les limites des saisons correspondent à des phénomènes astronomiques simples (équinoxe ou solstice), cet exemple se propose de reconstituer des saisons « optimales » selon un critère d'homogénéité de la température. On utilisera ici la température minimale moyenne depuis 1873 à Paris, trouvée sur le site www.meteo-paris.com. Dans ce cas, on a les contraintes suivantes :

- Une saison est un ensemble connexe de jours. Un jour est donc toujours contigu à son successeur et à son prédécesseur. Il s'agit donc d'un cas de contiguïté linéaire.
- Les saisons ont une longueur comprise entre 91 et 93 jours, la taille de chaque jour valant 1.
- Tous les jours sont équivalents et le poids de chacun vaut 1 aussi.
- On cherche à obtenir des saisons les plus homogènes en termes de température, donc à minimiser l'inertie intra-groupe.



Cet exemple a été choisi parce que le petit nombre d'agrégations possibles (il y en a 920) permet d'obtenir l'agrégation optimale. Le graphique ci-dessous donne l'agrégation optimale (en trait plein) et l'agrégation trouvé par la méthode (en pointillés, décalé vers le bas).

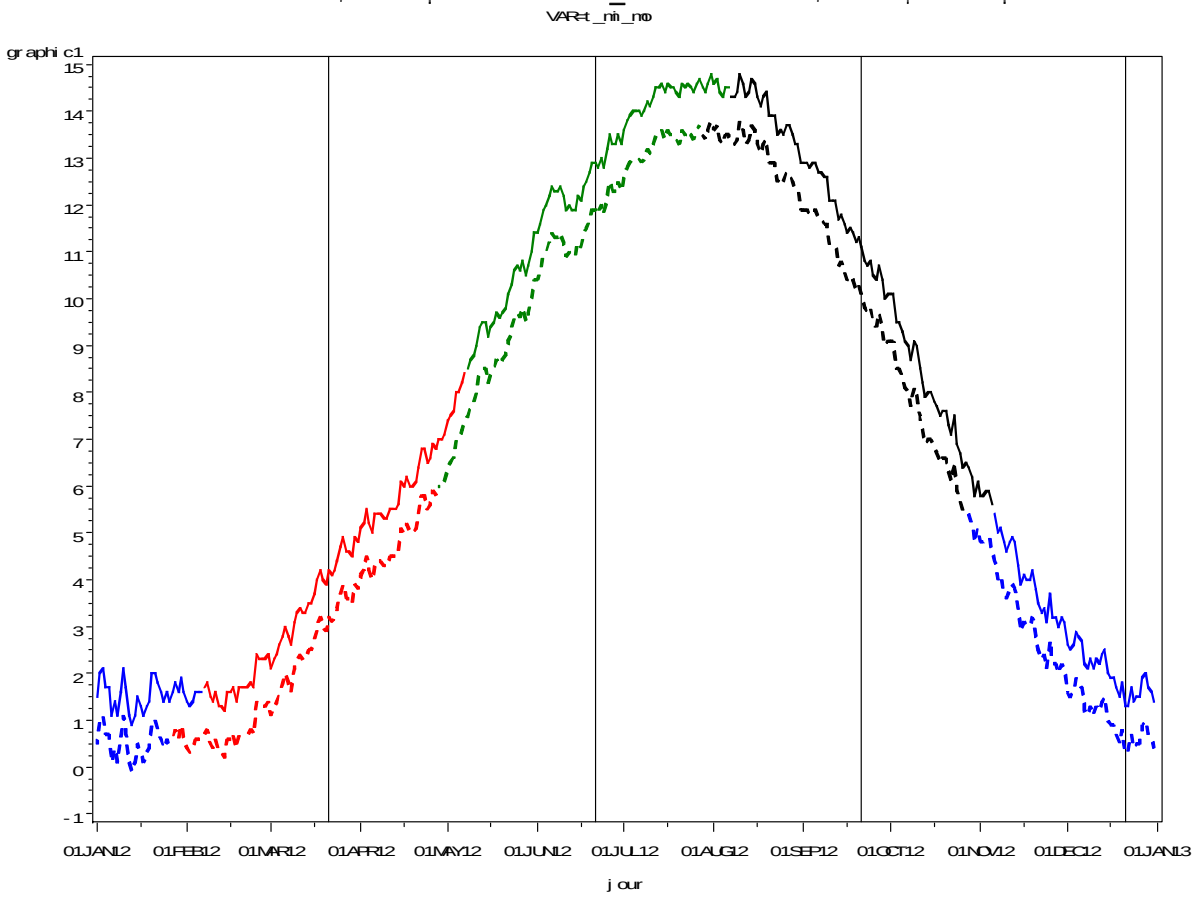
Les résultats montrent que l'agrégation obtenue par la méthode de CAH contiguë n'est pas optimale. En effet, l'inertie trouvée vaut 4,31 alors que l'inertie optimale vaut 4,28.

Les limites des saisons obtenues selon les deux méthodes sont différentes :

27 JAN, 27 AVR, 27 JUL et 26 OCT pour CAH_CONTIG

07 FEV, 08 MAY, 07 AUG et 06 NOV pour l'optimum

Variance intra : 4,3093 pour la CAH_CONTIG et 4,2839 pour optimum



5.2 Groupes fondés sur l'âge moyen par commune dans les Alpes-Maritimes

Le second exemple vise à agréger les 163 communes des Alpes-Maritimes en 5 groupes selon le critère de l'âge moyen de leurs habitants (l'âge moyen pour le département vaut 43,1 ans, données RP 2008). Le poids de chaque commune est égal à sa population statistique et sa taille est la part qu'elle représente par rapport à la population totale. Chaque groupe constitué doit avoir une taille comprise entre 15 et 35% de la population totale. La contiguïté entre communes est géographique.

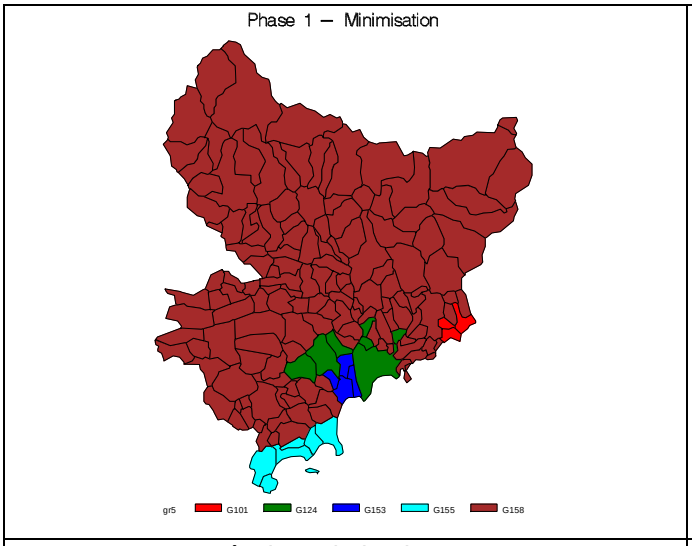
Cet exemple simple permet de donner une idée du nombre de partitions en 5 classes (connexes ou pas) de 163 unités statistiques. Ce nombre est de l'ordre de $7*10^{111}$. Le nombre de partitions en 4 classes est de l'ordre de $6*10^{96}$. Il est donc impossible d'explorer la totalité des solutions.

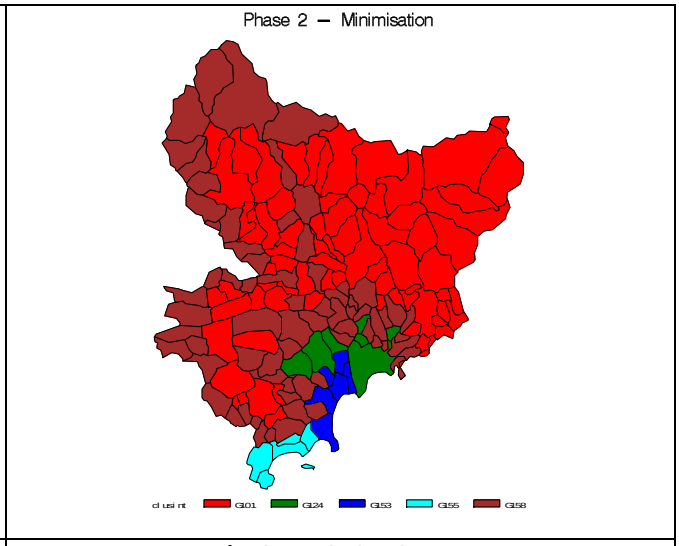
Cet exemple a été choisi car il va aussi nous permettre de comparer les résultats des deux phases de la procédure.

5.2.1 Minimisation de la variance intra-groupe

On cherche ici à construire les groupes les plus homogènes possible vis-à-vis de l'âge moyen.

Les graphiques ci-dessous montrent les résultats à la fin de la première et de la seconde phase.





Résultats de la phase 1

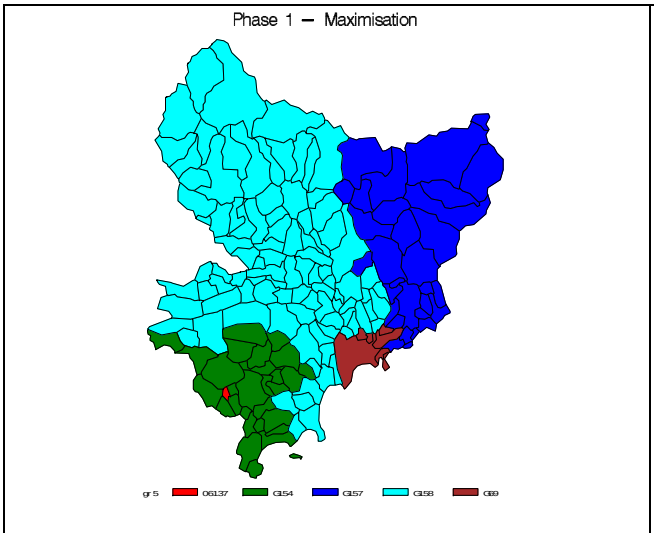
Groupes	Taille	Age moyen	Variance intra
G101	4,0%	45,4	0,002
G124	34,8%	43,1	0,013
G153	8,2%	43,2	0,126
G155	22,4%	45,6	0,309
G158	30,5%	41,0	2,752
		43,1	**3,202**

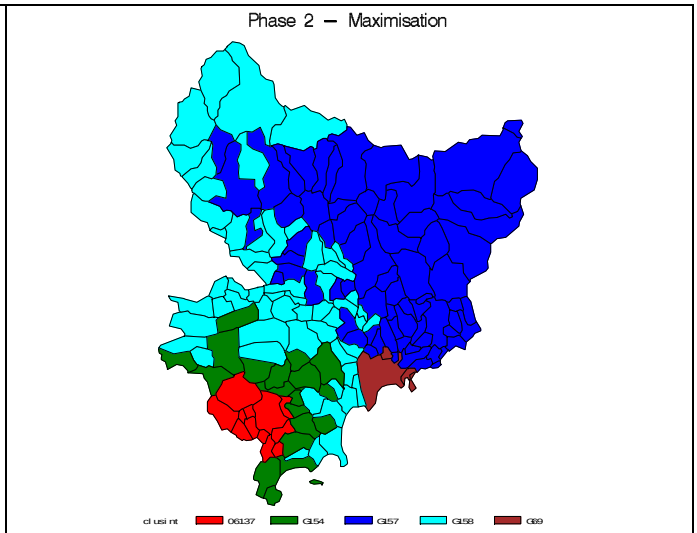
Résultats de la phase 2

Groupes	Taille	Age moyen	Variance intra
G101	16,1%	42,5	1,525
G124	34,8%	43,1	0,013
G153	16,7%	43,7	0,378
G155	15,3%	45,9	0,267
G158	17,1%	40,8	1,650
		43,1	**3,202**

5.2.2 Maximisation de la variance intra-groupe

Il s'agit des mêmes données, sauf qu'il est ici demandé une maximisation de la variance intra-groupe, c'est-à-dire une meilleure similarité entre les différents groupes, mais avec la plus forte hétérogénéité au sein de chaque groupe.





Résultats de la phase 1

Groupes	Taille	Age moyen	Variance intra
06137	0,1%	43,5	0
G154	26,2%	43,5	2,692
G157	8,2%	43,9	0,566
G158	22,0%	42,7	2,158
G69	30,0%	43,0	0,544
		43,1	**5,960**

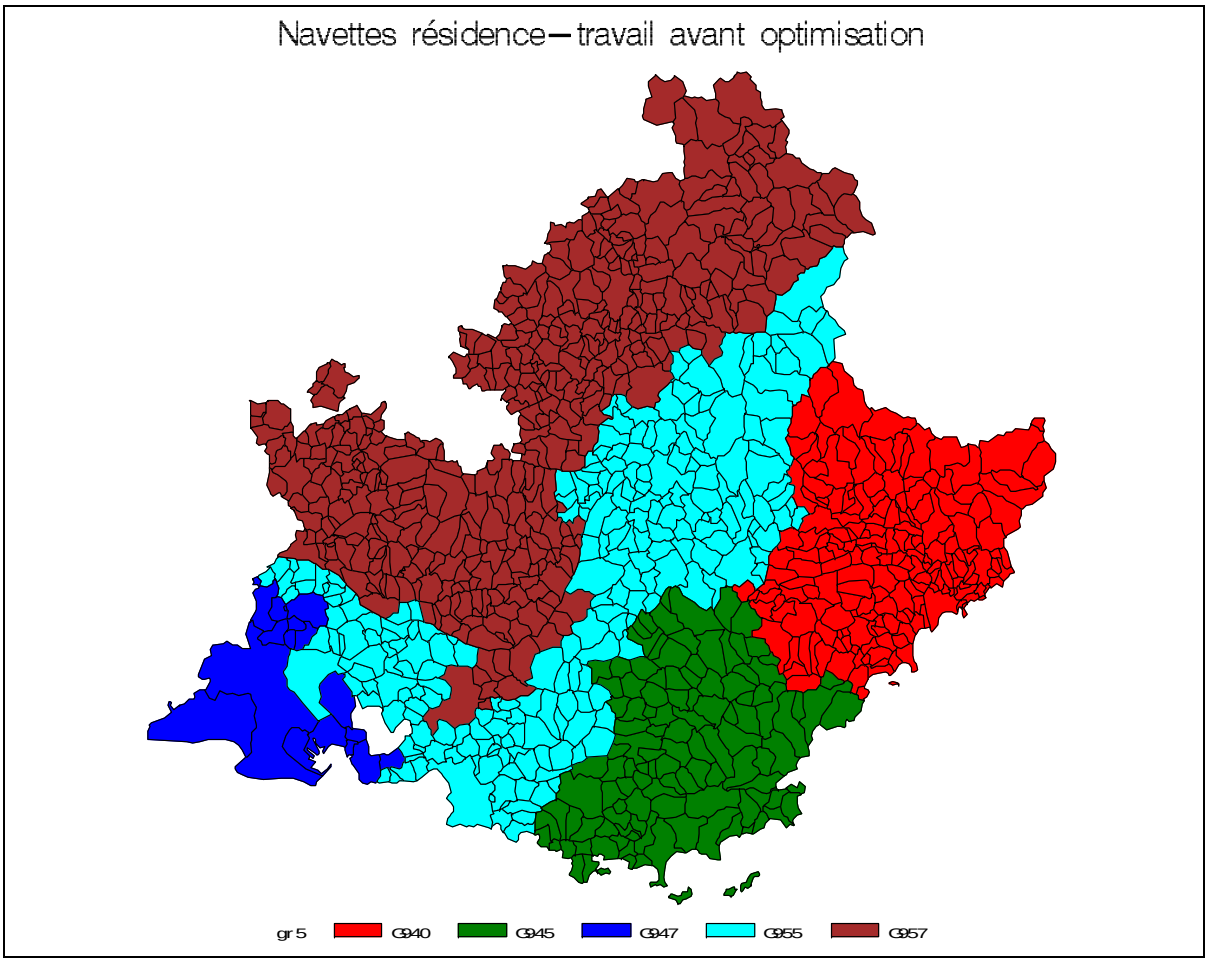
Résultats de la phase 2

Groupes	Taille	Age moyen	Variance intra
06137	9,20%	40,4	0,370
G154	16,5%	43,8	2,216
G157	15,0%	42,7	1,376
G158	26,2%	43,8	0,898
G69	33,1%	43,2	0,897
		43,1	**5,169**

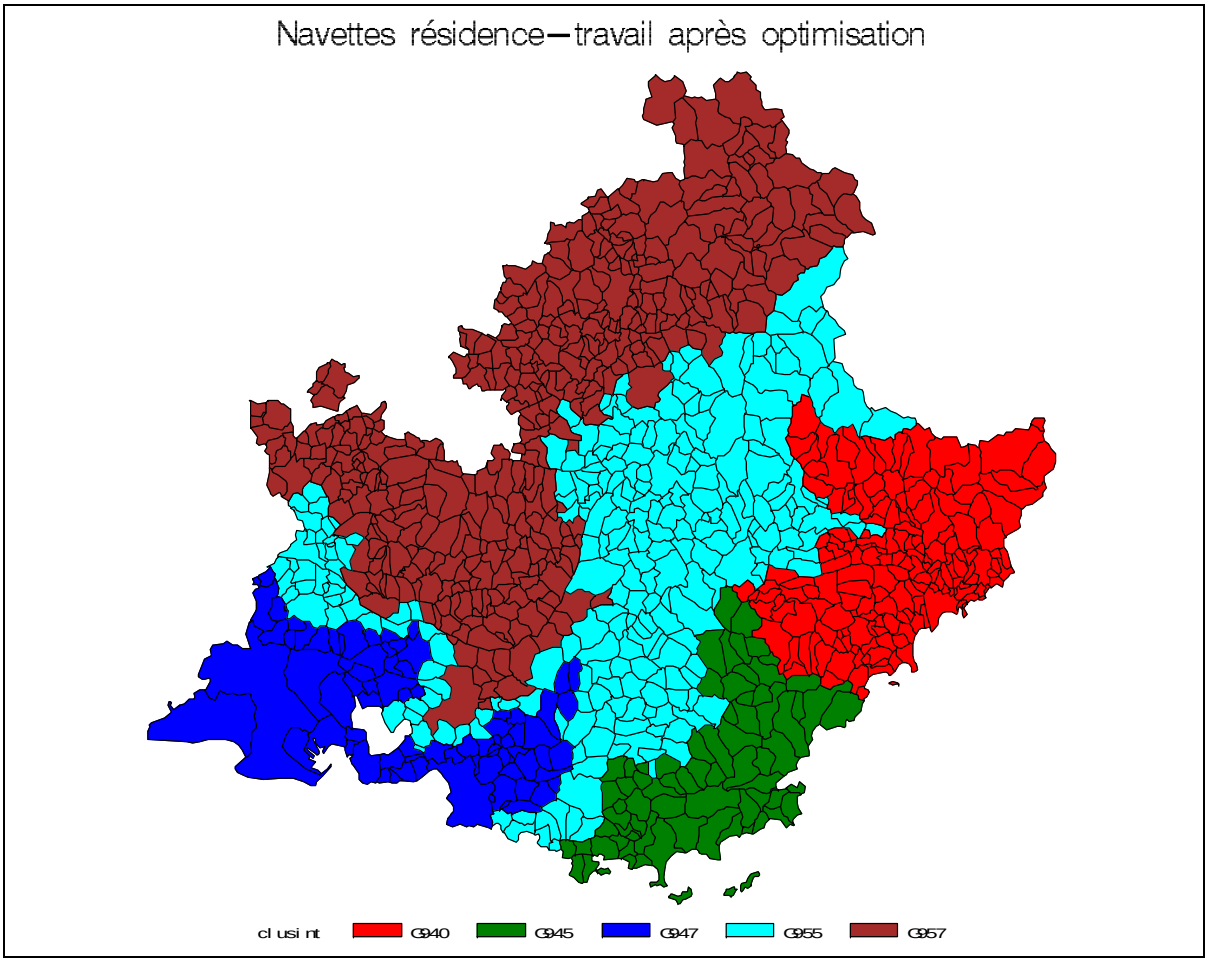
La commune 06137 est Spéracédès, commune de 1257 habitants en 2009.

5.3 Migrations professionnelles en région PACA

Le dernier exemple est un exemple plus complexe puisqu'il nécessite la mise en œuvre d'une distance non euclidienne telle que mentionnée ci-dessus [cf. § 3.3]. Les résultats de la première phase sont synthétisés dans la carte ci-dessous :



Les résultats de la seconde phase le sont dans la carte ci-dessous :



	Phase 1	Phase 2
% de personnes travaillant et résidant dans la même zone	91,4%	88,2%
Taille minimale des zones	5,7% (G947)	17,1% (G955)
Taille maximale des zones	28,1%(G940)	28,1%(G940)

Des exemples similaires pourraient être développés : construction de « bassins de vie » obtenus en comparant département de naissance et département de résidence.

6 Évolutions possibles

Actuellement, les limitations de taille mémoire empêchent de mettre en œuvre l'algorithme sur un nombre important d'unité statistiques, la limitation semblent être de l'ordre de 3000 unités statistiques. Les prochains travaux concerneront l'amélioration de l'algorithme sur ce point.

Par ailleurs, d'autres applications devraient être testées et leurs résultats analysés : en particulier, celles relatives à des distances non euclidiennes (données de flux) ainsi que la problématique de l'affectation d'un échantillon entre différents enquêteurs, en étudiant l'impact de différentes spécifications de la relation de contiguïté entre les unités échantillonnées.

7 ANNEXE

ENCADRÉ : sur les notions ensemblistes de contiguïté et de connexité.

Soit E un ensemble non vide muni d'une relation *R* réflexive et symétrique, dite *relation de contiguïté*.

Cette relation, définie sur E, peut s'étendre à deux *parties* de E.

- On dit que deux parties A et B de E, non vides, sont *contiguës* si et seulement si :

$$\exists\, x \in A, \exists\, y \in B : x\ R\ y\ .$$

  On constate sans problème que : $x\ R\ y \Leftrightarrow \{x\}$ et $\{y\}$ contiguës.

  Il y a donc cohérence entre les notions de contiguïté entre éléments et entre parties de E.

- On dit qu'une partie A de E, non vide, est *non connexe* si et seulement s'il existe deux parties non vides de E, C_1 et C_2, et non contiguës, telles que : $A = C_1 \cup C_2$.

  Une partie A sera dite *connexe* dans le cas contraire.

Par exemple, si E est une partie du plan euclidien usuel (dont les éléments sont des points), on notera $d(x, y)$ la distance euclidienne entre deux points *x* et *y* (à valeurs ≥ 0). Ces points peuvent figurer des logements, des maisons ou des immeubles dans un village, repérés sur un plan à grande échelle.

Soit $\delta > 0$. On peut prendre pour relation *R* celle définie sur E par :

$$\forall\ x \in E, \forall\ y \in E : x\ R\ y \Leftrightarrow d(x, y) < \delta\ .$$

La contiguïté entre les éléments s'interprète alors comme la proximité, à une distance moindre que δ.

Remarques.

- Dans la définition de la contiguïté, on notera ici que, du fait de la réflexivité de la relation *R*, les éléments *x* et *y* qui interviennent peuvent être égaux. De ce fait, **si deux parties A et B sont d'intersection non vide, elles sont contiguës,** puisque, si $x_0 \in A \cap B$, alors : $x_0\ R\ x_0$.

Inversement, **deux parties non contiguës sont nécessairement disjointes**. En revanche, deux parties peuvent être disjointes tout en étant contiguës.

- On notera qu'une partie non connexe contient au moins deux éléments, puisque les deux parties C_1 et C_2 qui figurent dans la définition sont non vides et non contiguës, donc disjointes. Par corollaire, ***un singleton est nécessairement connexe***. Par extension, l'ensemble vide est également connexe.

- La notion de non-connexité est plus facile à définir que sa négation (la connexité).

  Une partie A de E sera *connexe* si et seulement si l'une des conditions ci-dessous est satisfaite :

  Pour tout couple (C_1, C_2) de parties de E, non vides et non contiguës, on a : $A \neq C_1 \cup C_2$,

  ou :

  Pour tout couple (C_1, C_2) de parties non vides de E :

  ou bien elles sont contiguës, ou bien : $A \neq C_1 \cup C_2$,

  ou :

  Pour tout couple (C_1, C_2) de parties non vides de E :

  $A = C_1 \cup C_2 \Rightarrow C_1$ et C_2 sont contiguës.

Composante connexe

- Soit x un élément de E. On appelle *composante connexe* de x la réunion de toutes les parties connexes de E contenant x. Elle est toujours définie puisqu'il existe au moins une partie connexe contenant x : le singleton $\{x\}$.

  On la notera $C(x)$: $\boxed{C(x) = \bigcup_{\substack{x \in K \\ K \text{ connexe}}} K}$.

  On montre qu'il s'agit d'une partie connexe de E.

- Si l'on définit sur E la relation d'équivalence : $x \sim y \Leftrightarrow C(x) = C(y)$, on peut partitionner E en une réunion disjointe de parties connexes qui seront ses composantes connexes.
