UN ALGORITHME DE REGROUPEMENT D'UNITÉS STATISTIQUES SELON CERTAINS CRITÈRES DE SIMILITUDE

Marc CHRISTINE (*), *Michel ISNARD* (**)

(*) INSEE, Unité Méthodes Statistiques
(**) INSEE, Département de la Démographie

Construire une typologie d'une population de référence en constituant des classes homogènes vis-à-vis de certaines caractéristiques est une démarche traditionnelle en statistique empirique. Le problème se complique lorsque, en sus des critères qui permettent de construire la typologie, on rajoute des conditions sur le « positionnement » relatif des unités que l'on cherche à regrouper. Linéaire, ou le plus souvent spatial, ce positionnement sera appréhendé par le concept de *contiguïté.*

Cet article propose donc une méthode pour constituer de manière automatique une partition d'une population de référence formée d'unités de base, en agrégats astreints à différents types de contraintes : notamment critères de *similitude* (ou de dissemblance) vis-à-vis de certaines variables d'intérêt et critères de *contiguïté géographique.*

On décrira cette méthode sur le plan théorique et sur le plan algorithmique. Des programmes spécifiques ont été écrits pour la mettre en oeuvre. On étudiera enfin sur différents exemples, soit fictifs, soit réels, des applications de cette méthode.

Ce type de question a déjà été abordé dans la littérature. On se référera en particulier aux articles de L. LEBART et C. ROCHE parus en 1978 dans les Cahiers de l'Analyse des Données.

Néanmoins, par rapport aux approches utilisées dans ces papiers, notre perspective va plus loin :

- tout d'abord, l'approche est plus globale, et, en particulier, l'outil de classification ascendante hiérarchique (CAH) qui y est utilisé est, pour nous, un instrument ou une étape de la procédure, mais non le seul.
- ensuite, nous essayons de prendre en compte plusieurs types de critères pour constituer les classes : critères de similitude, critères de contiguïté et aussi critères de taille des agrégats formés.
- nous essayons de rechercher des solutions optimales (au moins localement), en mettant en oeuvre un algorithme d'amélioration d'une solution initiale.
- enfin, toutes les procédures informatiques résultantes ont été écrites spécifiquement en langage SAS.

1. Origine du problème.

Les échantillons des principales enquêtes nationales auprès des ménages réalisées par l'INSEE (à l'exception de l'enquête Emploi) reposent sur le concept d'*Echantillon-Maître*[1]. Il s'agit d'un échantillon à plusieurs degrés. Le 1 ^{er} degré est constitué *d'unités primaires*, qui sont, dans les zones entièrement rurales, des regroupements de communes, et, dans les zones urbaines, des Unités Urbaines (ou regroupement de petites Unités Urbaines)[2].

En préalable à la mise en place de l'échantillon-Maître, il convient donc de constituer ces unités primaires.

Le problème se pose principalement dans les zones rurales (dites « *strate de gestion 0* »), où il s'agit de former explicitement des regroupements de communes. En effet, pour les zones urbaines, quel que soit le degré d'urbanisation, les Unités Urbaines sont définies à partir de règles précises de densité de l'habitat et constituées à partir des données du dernier recensement de la population. On considère ces règles comme acquises et l'on ne cherche pas à construire des unités primaires urbaines selon d'autres critères.

La constitution des unités primaires en zone rurale doit obéir à plusieurs règles impératives :

- la ***connexité*** : ceci implique de ne regrouper que des communes contiguës entre elles.
- des ***seuils de taille minimale et/ou maximale (mesurée par l'effectif en nombre de logements de l'unité primaire).***

Ces deux premiers critères résultent essentiellement de considérations pratiques :

- les unités primaires correspondent en général 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.
- les contraintes minimales de taille s'expliquent par la nécessité d'avoir une réserve suffisante de logements pour réaliser toutes les enquêtes prévues pendant la période de vie de l'échantillon.

[1] On se réfèrera, pour plus de détails, aux différents articles présentés sur cette question dans la 1ère session des VII ^{èmes} JMS : « G. BOURDALLE, M. CHRISTINE, L. WILMS : Echantillons Maître et Emploi »

[2] La typologie des unités utilise le code TU99 : « tranche de taille d'unité urbaine ».

- la contrainte maximale se justifie à la fois par les raisons liées à la minimisation des déplacements des enquêteurs et par des raisons statistiques : dans un plan de sondage à deux degrés, on a en général intérêt à partir d'un grand nombre d'unités primaires, de tailles relativement homogènes et pas trop grandes, si l'on veut réduire la variance due au premier degré de tirage.

A ces deux premières règles s'en rajoute une troisième : **l'homogénéité (ou l'hétérogénéité) relativement à certaines variables d'intérêt** : c'est-à-dire que l'on cherche à construire des classes dont les unités constituantes soient assez « ressemblantes » (ou au contraire, très dissemblables), au vu de certaines caractéristiques socio-démographiques.

Ces conditions de similitude ou de dissemblance, que l'on va traduire en termes de variance, assez naturelles dans une démarche purement empirique, trouvent des justifications plus précises dans la théorie des sondages. Celles-ci seront éclairées, au moins sur certains cas de figure, dans les considérations du paragraphe suivant.

De même, pour l'échantillon Emploi, on a eu besoin de construire, en zone rurale, des regroupements de petites communes de façon à constituer des agrégats connexes, comprenant au moins 120 logements principaux. Ces regroupements devaient servir d'unités échantillonnées de 1 ^{ère} phase.

Dans ce cadre, l'homogénéité (ou la ressemblance) entre les communes vis-à-vis de différents critères socio-économiques semblait également pertinente pour présider au regroupement (quoique ce critère n'eût pas été introduit explicitement, notamment parce que la contiguïté géographique et la petite taille des unités concernées pouvaient l'entraîner naturellement).

Dans la pratique, la constitution des échantillons a été réalisée en cherchant à respecter au mieux les contraintes énoncées ci-dessus. Mais, en l'absence, au moment des travaux (réalisés au cours de l'année 2000), d'un algorithme permettant le traitement automatique de ces agrégations[3] - qui explique la genèse du présent travail -, la mise en oeuvre de ces dernières s'est faite de manière empirique.

[3] On notera toutefois que, pour la constitution de l'échantillon Emploi, on a mis en oeuvre un algorithme de regroupement des petites communes sur les seuls critères de taille minimale et de contiguïté. Cf. « la construction de l'échantillon de la future enquête Emploi en continu, Marc CHRISTINE, VII ^{èmes} JMS, Décembre 2000 ».

En particulier, le principe retenu pour l'Echantillon-Maître a consisté, pour la partie rurale de chaque région, à partir de partitions du territoire déjà existantes et vérifiant la condition de connexité, en l'occurrence, *les regroupements cantonaux*, et à modifier ceux-ci « à la main » pour satisfaire de manière impérative les limites de tailles et, seulement de manière approchée, les critères de similitude[4].

2. Lien avec la théorie des sondages.

2.1 Homogénéité dans les unités primaires et variance de l'estimateur

L'objet de ce paragraphe est d'apporter des éléments de justification théorique au choix du critère objectif (homogénéité ou hétérogénéité des classes constituées au vu de différents critères socio-démographiques) et à sa traduction mathématique dans la fonction à maximiser ou minimiser (variance intra dans une équation d'analyse de la variance), dans le cadre originel du travail (constitution d'unités primaires pour un plan de sondage à deux degrés tel que l'échantillon-Maître).

Dans un plan de sondage à deux degrés où les tailles des échantillons du 1er et du 2nd degrés sont fixes et dans lequel le second degré de tirage est aléatoire simple sans remise au sein de chaque unité primaire, l'estimateur de HORVITZ-THOMSON $\hat{\hat{T}}$ du total d'une variable d'intérêt Y définie sur l'ensemble de la population, a pour expression : $\hat{\hat{T}} = \sum_{j \in S_1} \frac{1}{\Pi_j^1} N_j \overline{y_j}$, avec les notations traditionnelles (rappelées en annexe A).

Comme détaillé en annexe A, sa variance se décompose donc en la somme de deux termes : un traduisant la variance liée au 1^{er} degré de tirage et l'autre relatif au second degré.

[4] Les interventions « manuelles » ont été confiées aux Directions Régionales de l'INSEE, avec l'assistance de différents outils cartographiques (cf. « utilisation de la cartographie en vue de la constitution d'échantillons », G. BOURDALLE, document présenté dans la session 1 des VII ^{èmes} JMS).

On cherche traditionnellement à réduire cette variance, ce qui peut conduire à agir sur chacun de ses deux termes. Malheureusement, ces derniers varient en général de façon antinomique, en raison de l'équation de l'analyse de la variance pour la décomposition de la variance totale de la variable, calculée sur la population, selon les unités primaires.

Néanmoins, on peut déterminer des cas où il est facile de voir l'incidence du mode de composition des unités primaires sur la variance totale.

C'est notamment le cas :

- lorsque le 1 ^{er} degré correspond à une stratification
- lorsque le 1 ^{er} degré est un tirage aléatoire simple.

Le lecteur intéressé se reportera, pour le détail des calculs, à l'annexe A.

On peut ainsi voir que, lorsqu'on constitue un échantillon stratifié, on a intérêt à avoir des unités primaires les plus homogènes possibles. Inversement, si l'on fait un tirage en grappes, les grappes doivent être le plus hétérogènes possibles.

De même, si le plan de sondage au 1er degré est aléatoire simple et que les unités primaires aient la même taille, les unités primaires doivent être le plus hétérogènes possibles.

D'autres cas intermédiaires peuvent se produire : si le tirage des unités primaires est lui-même stratifié, la stratification étant définie à partir des valeurs prises par le niveau moyen d'une ou de plusieurs variables d'intérêt, on a intérêt que les unités primaires soient relativement homogènes pour que la stratification ait un sens.

2.2 La constitution des unités primaires

Dans ces conditions, on comprend la stratégie qui peut présider à la constitution des unités primaires.

Si l'objectif était d'estimer un seul total d'une seule variable d'intérêt à partir d'un plan de sondage à 2 degrés, on s'arrangerait pour minimiser la variance totale de l'estimateur de ce total.

Dans la pratique, cependant, les échantillons (en particulier l'échantillon-Maître) ne sont pas construits pour estimer le total d'une seule variable d'intérêt, mais pour réaliser différentes enquêtes sur différents thèmes, d'où sont issues plusieurs variables d'intérêt.

Mais, dans ce cas, si l'on construit les unités primaires sur la base d'un critère de minimisation de la variance d'un estimateur du total d'un petit nombre de variables d'intérêt standard, on peut espérer qu'on satisfera, globalement, les conditions du problème pour une large gamme de variables d'intérêt décrivant le comportement des ménages, à condition que ces dernières soient bien corrélées avec les variables d'intérêt standards choisies.

Ceci explique que, dans la suite de ce papier, nous nous intéresserons à un critère de constitution des unités primaires qui s'écrira sous la forme suivante :

Minimiser **ou** **Maximiser**	**La variance intra d'une variable d'intérêt dans une décomposition de la variance totale de cette variable sur un ensemble d'unités primaires.**

Nous proposerons ici une méthode permettant de construire automatiquement les unités primaires par regroupement des unités de base, ici les communes, et cherchant à satisfaire trois types de critères :

- des critères de contiguïté géographique
- des seuils de taille (= effectif) maximaux et minimaux
- des critères d'homogénéité ou d'hétérogénéité « intra » relativement à une ou plusieurs variables d'intérêt, au sens défini ci-dessus.

Il faut noter que l'objectif de ce papier est de fournir un outil pour réaliser des classes de communes selon les critères ci-dessus et non de proposer une nouvelle typologie communale, même si la portée de la méthode peut aller au-delà de son objectif initial (cf. § 6.2 et 7).

Avant de décrire la méthode, il est nécessaire de faire des rappels sur les méthodes de partitionnement et d'agrégation pour voir comment elles peuvent s'appliquer.

3. Quelques rappels sur l'agrégation.

Considérons une population $\mathcal{P}$ composée d'unités élémentaires (ici des logements), en nombre N. Sur chacune des unités (*i*) est définie une variable d'intérêt y_i.

On définit :

- la moyenne empirique de la variable d'intérêt sur la population :
$$\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$
- la variance empirique[5] de la variable d'intérêt sur la population :
$$S^2 = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \overline{Y}\right)^2$$

Supposons que la population soit partitionnée en K unités primaires, définies provisoirement comme des regroupements de logements. Si l'on note :

- $\overline{Y_k}$ la moyenne empirique de la variable d'intérêt sur l'unité primaire numéro *k*.
- S_k^2 la variance empirique de la variable d'intérêt sur l'unité primaire numéro *k*, définie comme ci-dessus.
- N_k la taille (en nombre de logements) de l'unité primaire,

alors l'équation d'analyse de la variance pour la décomposition de la variance totale de Y sur $\mathcal{P}$ selon les K unités primaires va s'écrire classiquement :

$$S^2 = \sum_{k=1}^{K}\frac{N_k}{N}\left(S_k^2 + \left(\overline{Y_k} - \overline{Y}\right)^2\right).$$

La variance intra de cette décomposition vaut $\boxed{V_i = \frac{1}{N}\sum_{k=1}^{K} N_k S_k^2}$.

[5] Nous prenons, par rapport à l'annexe A relative au paragraphe précédent, une définition de la variance différant de la précédente par le nombre d'unités intervenant au dénominateur : la notation S_j^2 est utilisée ici au lieu de la notation S_j^{*2} de l'annexe A.

En **réalité**, les unités primaires vont être définies comme des regroupements *d'unités indivisibles*, les communes. Dans ce cas, l'équation d'analyse de la variance permet de décomposer à nouveau la variance S_k^2 sur l'unité primaire k en utilisant la décomposition de cette unité en communes. En notant, par analogie avec ce qui est fait plus haut, $\overline{Y_{jk}}$, S_{jk}^2 et N_{jk} les moyenne, variance et taille de la commune j de l'unité primaire k, on obtient alors :

$S_k^2 = \sum_{j \in C_k} \frac{N_{jk}}{N_k} \left(S_{jk}^2 + \left(\overline{Y_{jk}} - \overline{Y_k} \right)^2 \right)$, formule dans laquelle C_k représente l'ensemble de communes composant l'unité primaire k.

La variance intra qui nous intéresse dans la décomposition de la population en unités primaires a comme expression : $V_i = \frac{1}{N} \sum_{k=1}^{K} N_k \left\{ \sum_{j \in C_k} \frac{N_{jk}}{N_k} \left[S_{jk}^2 + \left(\overline{Y_{jk}} - \overline{Y_k} \right)^2 \right] \right\}$,

soit :

$$\boxed{V_i = \frac{1}{N} \sum_{k=1}^{K} \left\{ \sum_{j \in C_k} N_{jk} \left[S_{jk}^2 + \left(\overline{Y_{jk}} - \overline{Y_k} \right)^2 \right] \right\}}$$

Cette décomposition va nous permettre de traiter plusieurs problèmes lorsque l'on constitue des unités primaires par agrégation de communes ou lorsqu'on envisage de modifier la composition des unités primaires en échangeant des communes d'une unité primaire à l'autre. On notera que les degrés de liberté portent sur la composition de l'unité primaire en communes (dans l'objectif de minimiser la variance intra ci-dessus[6]) et non sur les moyennes ou les variances de la variable d'intérêt à l'intérieur d'une commune donnée, qui deviennent des paramètres exogènes de celle-ci. On notera avec intérêt que ces paramètres jouent le rôle de *statistiques exhaustives*, ce qui permet de n'utiliser l'information qu'au niveau communal et non au niveau individuel. Bien entendu, on ne modifie pas la composition des communes en logements et on s'interdit de définir les unités primaires à partir de fractions de communes.

[6] Ou de la maximiser si on vise à assurer l'hétérogénéité des unités primaires.

3.1 Problème 1 : Ajout d'une nouvelle commune à une classe existante

Supposons que nous rajoutions à une unité primaire donnée (composée de C_0 communes) une commune que nous appellerons *u*. On suppose ici que l'on est dans un processus de construction d'unités primaires et que la commune *u* n'est pas prélevée dans une autre unité primaire : ajouter la commune *u* ne modifie donc pas la composition et les paramètres des autres unités primaires.

Dans l'expression de Vi, seul va être modifié le terme correspondant à l'unité primaire en cours de modification, que l'on supposera indexée par 0.

Valeur initiale	$\frac{1}{N}\sum_{j\in C_0} N_{j0}\left[S_{j0}^2+\left(\overline{Y_{j0}}-\overline{Y_0}\right)^2\right]$
Valeur finale	$\frac{1}{N}\sum_{j\in C_0\cup\{u\}} N_{j0}\left[S_{j0}^2+\left(\overline{Y_{j0}}-\overline{Y_0}^{+}\right)^2\right]$

Dans les formules ci-dessus, $\overline{Y_0}$ (resp. $\overline{Y_0}^{+}$) représente la moyenne de la variable d'intérêt sur l'unité primaire avant (resp. après) l'agrégation.

La variation de variance intra due à cette agrégation par une unité « exogène » vaut alors :

$$\Delta V_i=\frac{1}{N}\left\{\sum_{j\in C_0} N_{j0}\left[\left(\overline{Y_{j0}}-\overline{Y_0}^{+}\right)^2-\left(\overline{Y_{j0}}-\overline{Y_0}\right)^2\right]+N_u\left[S_u{}^2+\left(\overline{Y_u}-\overline{Y_0}^{+}\right)^2\right]\right\}$$

Dans cette formule, l'indice u se rapporte aux caractéristiques (taille, moyenne et variance) de la commune qu'on agrège.

Après simplification, on obtient la formule suivante :

$$\Delta V_i=\frac{1}{N}\left[\frac{N_u N_0}{N_u+N_0}\left(\overline{Y_0}-\overline{Y_u}\right)^2+N_u S_u^2\right]$$

L'agrégation a donc toujours pour effet d'augmenter la variance intra (sauf pour le cas particulier limite où $\overline{Y_0} = \overline{Y_u}$ et $S_u^2 = 0$) toutes choses égales par ailleurs. On va voir dans le second problème que cela n'est plus le cas s'il y a échange entre unités primaires.

3.2 Problème 2 : Transfert d'une unité primaire à l'autre

Nous supposons, dans ce paragraphe, qu'il y a échange d'une commune entre deux unités primaires que l'on supposera être, pour simplifier, les unités 1 et 2.

Dans l'expression de la nouvelle variance intra, seuls vont être modifiés les termes relatifs aux unités primaires 1 et 2, les autres restant inchangés. On désignera par l'indice c la commune qui est transférée de l'unité primaire 1 à l'unité primaire 2, n_c, $\overline{Y_c}$, S_c^2 étant ses caractéristiques.

Notons que, dans cet exemple de transfert, aucune condition de connexité, ni de taille, n'intervient.

3.2.1 Calcul de la variation de la variance intra

Les termes concernant les unités primaires 1 et 2 qui apparaissent dans l'expression de la variance **initiale** sont :

$$\frac{1}{N}\left\{\sum_{j\in C_1} N_{j1}\left[S_{j1}^2 + \left(\overline{Y_{j1}} - \overline{Y}_1\right)^2\right] + \sum_{j\in C_2} N_{j2}\left[S_{j2}^2 + \left(\overline{Y_{j2}} - \overline{Y}_2\right)^2\right]\right\},$$

avec des notations analogues à celles des problèmes examinés ci-dessus.

Les termes qui apparaissent dans l'expression de la variance **finale** sont :

$$\frac{1}{N}\left\{\begin{array}{l} \displaystyle\sum_{j\in C_1-\{c\}} N_{j1}\left[S_{j1}^2 + \left(\overline{Y_{j1}} - \overline{Y}_1^+\right)^2\right] + \sum_{j\in C_2} N_{j2}\left[S_{j2}^2 + \left(\overline{Y_{j2}} - \overline{Y}_2^+\right)^2\right] \\ + n_c\left[S_c^2 + \left(\overline{Y_c} - \overline{Y_2^+}\right)^2\right] \end{array}\right\}$$

Dans ces dernières formules, les indices + correspondent aux moyennes de la variable d'intérêt calculées sur les nouvelles unités primaires après échange.

Après calculs et simplification, la variation de variance intra vaut :

$$\Delta V_i = \frac{n_c}{N}\left[\frac{N_2}{N_2 + n_c}\left(\overline{Y}_2 - \overline{Y}_c\right)^2 - \frac{N_1}{N_1 - n_c}\left(\overline{Y}_1 - \overline{Y}_c\right)^2\right]$$

Cette formule ne fait pas apparaître les variances intra relatives aux communes. Ceci s'explique par le caractère « conservatif » du processus : il n'y a pas d'échange de communes avec l'extérieur de l'ensemble constitué par les unités primaires 1 et 2. La formule ne fait intervenir que des caractéristiques avant le transfert, permettant ainsi le calcul sans qu'il soit nécessaire de recalculer les différentes caractéristiques des classes.

3.2.2 L'opportunité du transfert

On dira qu'un transfert est avantageux dans le cas où la variance intra diminue[7], c'est-à-dire si et seulement si :

$$\left|\frac{\overline{Y}_2 - \overline{Y}_c}{\overline{Y}_1 - \overline{Y}_c}\right| < \sqrt{\frac{N_1 (N_2 + n_c)}{N_2 (N_1 - n_c)}}$$

Cette formule sera réutilisée dans le paragraphe 5.2.

Plusieurs remarques peuvent être faites :

- on peut dire que le transfert est avantageux si et seulement si la moyenne (ou centre de gravité) de la commune qu'on transfère est plus proche (au sens d'une certaine distance) de celle de la 2ème UP que de la 1ère.
- le coefficient qui apparaît à droite de l'inégalité est toujours plus grand que 1.
- la formule pourrait être symétrisée en introduisant les tailles et moyennes des unités primaires, excluant la commune transférée.

[7] On inversera les sens si on vise à assurer l'hétérogénéité.

3.3 L'agrégation de deux unités primaires

Considérons deux unités primaires que nous noterons 1 et 2 et considérons l'unité primaire 0 qui est la réunion de ces deux unités primaires.

Là encore, dans le calcul de la variation de la variance intra, seuls interviennent les unités primaires 1 et 2, puisque les autres unités primaires ne sont pas touchées.

On aura donc, en repartant de la définition même de la variance intra :

$$\Delta V_i = \frac{1}{N}\left[(N_1 + N_2) S_0^2 - N_1 S_1^2 - N_2 S_2^2\right].$$

Après calculs et simplification, on retrouve la formule « classique » en CAH :

$$\boxed{\Delta V_i = \frac{1}{N}\frac{N_1 N_2}{N_1 + N_2}\left(\overline{Y_1} - \overline{Y_2}\right)^2}$$

On notera que cette formule est très similaire à celle qui apparaît dans le 1 ^{er} problème. Dans ce premier problème, apparaît un terme complémentaire représentant la variance intra d'une commune que l'on agrège : on « importe » en effet une commune de l'extérieur, qui vient avec sa propre variance, qui se rajoute donc à celle des autres communes présentes. Au contraire, dans le 3 ^{ème} problème, on se place dans un cadre conservatif : l'ensemble des communes est fermé, sans échange avec l'extérieur.

Remarques :

- La variation de la variance intra ne dépend pas de la variance à l'intérieur des communes.

- L'agrégation de deux classes se traduit toujours par une augmentation de la variance intra, sauf si les deux moyennes coïncident.

- Ce résultat permet une autre interprétation qui sera utilisée dans la procédure de CAH permettant l'initialisation de la constitution des unités primaires. Si l'on doit agréger ensemble deux classes de communes, il faut que l'augmentation de variance intra qui en résulte soit la plus faible possible. En d'autres termes, on aura intérêt à agréger les deux unités primaires réalisant le minimum de : $\Delta V_i = \frac{1}{N}\frac{N_\alpha N_\beta}{N_\alpha + N_\beta}\left(\overline{Y}_\alpha - \overline{Y}_\beta\right)^2$,

sur tous les couples (α, β) d'unités primaires. Ce critère est utilisé classiquement comme mesure de la distance entre classes dans la CAH.

4. Problèmes posés par la recherche de solutions.

Le problème posé revient donc à chercher une partition de la population en classes de communes vérifiant les contraintes suivantes :

- Ces classes de communes doivent être connexes.
- La partition doit minimiser[8] la variance intra classe.
- La taille de chacune des classes de communes doit être comprise entre une taille minimale et une taille maximale.
- Le nombre de classes doit être fixé. Cette contrainte a été ajoutée pour faciliter le déroulement de l'algorithme. L'utilisateur peut modifier ce nombre, s'il le juge utile.

Avant de parler des méthodes possibles pour trouver la (ou une) solution, il est nécessaire de revenir sur ces différentes contraintes et sur la manière de les combiner.

4.1 Contraintes incohérentes ? Contraintes incompatibles ?

Avoir un ensemble de contraintes pose nécessairement le problème de savoir s'il existe au moins une solution qui les vérifie. Il est possible de distinguer deux types d'impossibilité :

- On dira que l'ensemble des contraintes est **incohérent** s'il n'existe aucune solution qui respecte les deux dernières contraintes (seuils de taille et nombre de classes). Il suffit de vérifier pour cela que la taille de la population divisée par le nombre de classes se situe dans l'intervalle formé par la taille minimale et la taille maximale. Ainsi, le jeu de contraintes suivantes n'est pas cohérent : taille comprise entre 100 et 200 avec 5 classes

[8] Ou la maximiser si on vise à assurer l'hétérogénéité.

alors que la taille de l'ensemble des communes est de 10000. On voit ici qu'il est inutile de chercher une solution à ce problème.

- On dira que l'ensemble des contraintes est **incompatible** si l'ensemble des contraintes, *y compris celle de connexité*, ne peut pas être respecté.

Il est relativement simple par programme de détecter des contraintes incohérentes. En revanche, détecter un ensemble de contraintes incompatibles revient à montrer que les classes d'une solution d'un ensemble de contraintes cohérent ne sont pas toutes connexes. Dans la quasi-totalité des cas, la seule méthode envisageable consisterait à lister toutes ces solutions. Nous verrons plus loin que ce n'est pas faisable avec les moyens actuels, ce qui nous conduira à rechercher des solutions approchées.

4.2 Comment combiner ces contraintes ?

Devant un tel ensemble de contraintes, surtout quand elles risquent d'être incompatibles, plusieurs attitudes sont possibles :

- La première consiste à essayer d'en relâcher quelques unes et de les modifier de telle manière qu'elles puissent être atteintes. L'opérateur doit toujours pouvoir modifier ses contraintes initiales.
- Une seconde méthode, s'il apparaît que le problème n'est pas soluble, consisterait à chercher des solutions « proches » de la vérification des différentes contraintes. Par exemple, au lieu de fixer les seuils de taille des classes, on peut établir une fonction qui soit une distance (ou un écart) entre la taille réelle et les tailles théoriques visées. On cherchera alors à minimiser cet écart, la valeur du minimum n'étant nulle que si les contraintes de taille sont complètement satisfaites. Ce faisant, on voit que l'on disposerait alors de deux fonctions objectifs, qui ne peuvent être gérées simultanément que si on les pondère mutuellement. Il faudrait alors créer un nouveau critère comme combinaison linéaire des différents critères. Mais le problème qui se pose alors est l'influence des différentes contraintes dans la valeur du critère final.

La voie suivie dans ce papier est autre : elle consiste à privilégier une des contraintes, en sus de la contrainte de connexité des classes de communes, et d'essayer de vérifier au mieux l'autre. Ainsi, dans tout ce qui va suivre, le nombre de classes sera fixé, alors que la contrainte de taille pourra être relâchée.

4.3 Solution locale ? Solution globale ?

Si le jeu de contraintes est compatible, il y a une manière théorique relativement simple pour trouver une solution optimale. Il suffit de passer en revue toutes les solutions en termes de partition, de voir si elles sont compatibles avec les contraintes et de calculer, dans ce cas, la variance intra de cette partition. On conservera la solution qui minimise la variance intra. On parlera de solution **globale**.

Une autre méthode consiste à partir d'une solution compatible avec les contraintes et à l'améliorer (c'est-à-dire à chercher à minimiser ou à maximiser la variance intra) autant que possible, tout en vérifiant les contraintes. On parlera alors d'une solution **locale**.

La solution locale ne peut pas être meilleure que la solution globale et peut même en être éloignée. En pratique, ce qui permet de choisir l'une ou l'autre des voies est le nombre de solutions compatibles.

Or, ce nombre est très élevé dans le problème qui nous concerne.

L'annexe B fournit une relation permettant de calculer le nombre de manières de grouper n communes en p classes (sans tenir compte des autres contraintes). Ce nombre grandit très rapidement avec n et p et l'on peut donner un ordre d'idée de la valeur de ce nombre :

Communes	*Classes*	*Nombre de partitions*
11	2	1 023
11	5	246 730
15	5	2,1 E+8
23	5	9,6 E+13
53	5	1,8 E+34
53	15	1,1 E+50

Il est évident, au regard des nombres donnés ci-dessus, qu'une solution optimale globale ne peut pas être atteinte, même pour une répartition de 53 communes en 5 classes, sauf pour des exemples d'école qui serviront de tests aux méthodes proposées (cf. infra, § 6.1). Seule la voie des solutions locales nous est ouverte pour des données réelles.

5. La recherche d'une solution locale.

Comme indiqué plus haut, la recherche d'une solution locale fonctionne en deux temps :

- Recherche d'une solution initiale, vérifiant les contraintes.
- Amélioration de cette solution tout en vérifiant les contraintes.

Dans ce qui suit, seront présentés successivement la manière dont est obtenue la solution initiale, puis l'algorithme d'amélioration de cette solution.

5.1 La solution initiale

La solution initiale est une solution vérifiant les différentes contraintes mentionnées plus haut. Plusieurs méthodes sont possibles. On en distinguera plus particulièrement deux : une première appelée « agrégation successive » qui vise à essayer de bâtir des classes en vérifiant progressivement les contraintes, la seconde est basée sur une classification Ascendante Hiérarchique avec contraintes de contiguïté.

Parmi les autres méthodes possibles, on peut aussi partir de regroupements existants (comme les cantons qui sont des parties connexes[9]). Il faut noter que les procédures d'affectation aléatoire (comme celle utilisée dans la procédure FASTCLUS de SAS[10]) ne peuvent pas être utilisées, car les classes obtenues ne sont pas connexes.

5.1.1 La méthode « d'agrégation successive »

Elle se déroule de la manière suivante :

- Prendre la plus petite commune (par la taille) non encore agrégée.
- Prendre l'ensemble des communes contiguës à cette commune par ordre de taille croissante.
- Agréger chacune de ces communes tant que la taille minimale de la classe n'est pas atteinte. Une fois cette taille atteinte, repartir à l'étape 1 en éliminant les communes ainsi agrégées.

[9] Néanmoins, on peut perdre la connexité si l'on se restreint, à l'intérieur des cantons, aux seules communes rurales par exemple.

[10] Dont la méthodologie est décrite dans la documentation SAS.

Cette méthode permet bien de créer, dans la majorité des cas, des tailles de classes supérieures à la taille minimale figurant parmi les contraintes. Mais les contraintes du nombre de classes ou de taille maximale ne sont pas forcément respectées.

5.1.2 La Classification Ascendante Hiérarchique avec contiguïté

Une des manières classiques d'agréger les individus dans une population est la Classification Ascendante Hiérarchique (CAH). Toutefois, cette méthode ne garantit pas que les classes soient formées par agrégation d'unités contiguës. Il a donc été nécessaire de modifier l'algorithme classique pour tenir compte de cette contrainte.

5.1.2.1 L'algorithme de la CAH avec contiguïté

L'algorithme « classique » de la CAH est un algorithme fini qui consiste à agréger à chaque étape les deux classes constituées les plus proches. Au début de l'algorithme, il existe n classes constituées chacune d'un point. L'algorithme s'arrête après (n-1) itérations : la dernière classe constituée contient l'ensemble des communes. Traditionnellement, la distance entre deux classes quelconques est égale à l'augmentation de la variance intra qui aurait lieu si ces deux classes étaient agrégées. Ceci correspond bien au critère décrit dans le problème 3 ci-dessus.

La prise en compte de la contiguïté nécessite l'introduction d'une *matrice de contiguïté*, qui est un outil couramment utilisé en statistique spatiale[11]. Dans la pratique, cette matrice (dont l'élément générique vaut 1 si les deux unités de base, ici les communes, sont contiguës et 0 sinon) est construite à partir du fichier de contiguïté des communes, issu des données de l'IGN, qui donne, pour chaque commune, la liste de celles qui lui sont contiguës, en projection horizontale[12].

Cette matrice est *symétrique* et, pour des raisons de commodité d'écriture, nous supposerons également qu'elle possède des 1 sur la diagonale (ce qui correspond à la *réflexivité* de la notion de contiguïté, cf. annexe C).

La contiguïté des communes permet de définir une notion analogue pour les classes.

[11] Cf. Michel HANNOUN, « Un survol des méthodes de statistique spatiale », VII *èmes* JMS, 2 *ème* session.

[12] Ainsi, deux communes peuvent être contiguës sur le plan alors qu'elles sont séparées sur le terrain par un massif montagneux ou un obstacle naturel qui interdit un passage direct de l'une à l'autre par voie terrestre.

La prise en compte de la contrainte se fait alors de la manière suivante :

- Test de la contiguïté entre deux classes : le lecteur intéressé trouvera en annexe C la manière dont a été construit ce test de façon matricielle dans les programmes.
- Affectation d'une distance infinie (ou du moins supérieure aux différentes pertes de variance possible) si les classes concernées ne sont pas contiguës.

Ainsi, on n'agrège jamais deux classes non contiguës.

Cette modification de la procédure de CAH permet d'obtenir des classes connexes.

Une fois l'algorithme déroulé, il est possible de découper la population totale, si elle est connexe au départ, en K classes connexes, quel que soit le nombre de classes K (inférieur au nombre d'éléments). **Ce nombre de classes sera introduit comme paramètre de la macro SAS.** Il n'est pas possible d'en programmer automatiquement le calcul : il doit au contraire s'introduire comme contrainte en amont de la procédure, dont la valeur est laissée à l'initiative de l'opérateur et peut être modifiée par lui.

5.1.2.2 La contrainte de taille

Toutefois, les classes ainsi formées à l'issue de la CAH peuvent ne pas vérifier les contraintes de taille mentionnées plus haut. Une procédure dite d'*échange* va permettre de respecter ces contraintes, ou, à défaut, de s'en approcher le plus possible.

L'objectif ici est de minimiser une fonction des différentes tailles des classes, fonction positive sur l'ensemble des tailles possibles et prenant une valeur nulle à l'intérieur de l'intervalle $[T_{min}, T_{max}]$. La fonction utilisée est alors la suivante : $\sum_j f(t_j)$, où t_j représente la taille de la classe j. On peut prendre pour f la fonction suivante :

$$f(t) = \begin{cases} t - T_{max} \ si \ t > T_{max} \\ 0 \ si \ t \in [T_{min}, T_{max}] \\ T_{min} - t \ si \ t < T_{min} \end{cases}$$

D'autres spécifications pour f sont possibles.

La procédure d'échange va consister à transférer une commune de la classe G_1 à une classe G_2 dès que :

- Il y a une diminution du critère mentionné plus haut lors de l'échange.
- La classe G_1 reste connexe[13], même si la commune concernée lui a été enlevée.
- La classe G_2 reste connexe, après l'ajout de la commune considérée.

En pratique, l'avant-dernière condition revient à partir de l'ensemble des communes de G_1 non indispensables à sa connexité et la dernière à ne tester la validité d'un échange que pour les communes de cet ensemble qui sont au départ contiguës à G_2. Il conviendra donc d'examiner l'ensemble des communes de G_1 « éligibles » au sens précédent et de regarder s'il est possible d'échanger chaque commune de cet ensemble avec une des classes auxquelles elle est contiguë.

Bien entendu, cet échange produit en général une dégradation de la variance intra de la partition.

Après cette étape, on obtient donc une partition en K classes connexes, vérifiant, du mieux possible, les contraintes de taille.

Dans la programmation effective de l'algorithme, les différentes communes sont triées aléatoirement avant d'être traitées : on prend la première commune dans l'ordre aléatoire mentionné ci-dessus, on cherche à l'échanger avec toutes les classes auxquelles elle est contiguë. Si on ne peut pas l'échanger, on passe à la commune suivante. Le résultat, notamment en termes de variances intra, dépend cependant assez fortement de l'ordre de traitement des communes.

5.1.2.3 La rupture de connexité

Le déroulement de l'algorithme de recherche de la solution initiale est marqué par la volonté forte de ne pas rompre la connexité des classes à cause de la difficulté de reconstruire des classes connexes. C'est ce qui différencie la procédure par rapport à un échange « libre » entre classes (comme celui qui est mis en œuvre dans la procédure FASTCLUS) ; il est alors clair que cette contrainte de connexité conduit à un optimum en termes de variance intra qui est de moins bonne qualité qu'en l'absence de contrainte.

[13] La connexité d'une classe est testée suivant une procédure décrite en annexe B.

5.2 L'amélioration de la solution initiale

A partir de cette solution initiale, une procédure d'amélioration va être mise en œuvre. Elle est basée sur le même principe d'échange qui a été décrit dans le paragraphe précédent en ce qui concerne la taille.

Il s'agit, pour un point (= commune) donné, de tester s'il est *possible* et *pertinent* de le transférer d'une classe à l'autre. *Possible* fait ici référence au fait que les classes après le transfert doivent vérifier les contraintes de connexité et de taille. *Pertinent* fait référence au fait que la variance intra classe doit être plus petite après le transfert. On utilise pour cela la formule vue au paragraphe 3.2.2.

Plus précisément, l'algorithme se déroule de la manière suivante :

Etape 1a	Tri aléatoire des communes
Etape 1b	Prendre la première commune
Etape 2	Est-il possible et pertinent de transférer la commune en cours en testant l'agrégation sur toutes les classes contiguës ? Si **Oui**, la transférer, redéfinir la composition des classes et aller à l'étape 1b Si **Non**, passer à la commune suivante et passer à l'étape 2. Si c'était la dernière commune, passer à l'étape 3
Etape 3	Fin de l'algorithme

Comme on pouvait s'y attendre, l'ordre dans lequel on traite les communes ainsi que celui de la procédure d'échange (taille puis variance ou variance puis taille) influent assez fortement sur le résultat. Le programme informatique utilisé permet d'itérer cette étape et de choisir la meilleure solution possible à chaque étape.

6. Applications.

Les applications décrites dans ce paragraphe sont de trois natures différentes : un exemple purement d'école, différentes applications géographiques et le cas de la contiguïté linéaire.

Il est cependant difficile de montrer simultanément la configuration initiale, la dynamique de l'agrégation et les résultats finaux, même s'il pourrait être intéressant de confronter plusieurs variantes selon le type de contraintes que l'on impose ou le type de critères employés.

On se contentera donc ici de quelques schémas ou cartes, sans aller jusqu'à une cartographie très élaborée.

6.1 Exemple d'école.

Cet exemple a principalement pour but de comparer la solution globale, possible à identifier explicitement lorsqu'on travaille sur un petit nombre d'unités au départ, et différentes solutions locales. Celles-ci seront paramétrées pour mettre en évidence l'impact sur le résultat d'une modification, soit dans l'ordre d'examen des contraintes, soit de celui des unités, ou selon le mode de construction d'une solution initiale.

On examinera dans ce cadre un ensemble à 23 éléments à partitionner en 5 classes.

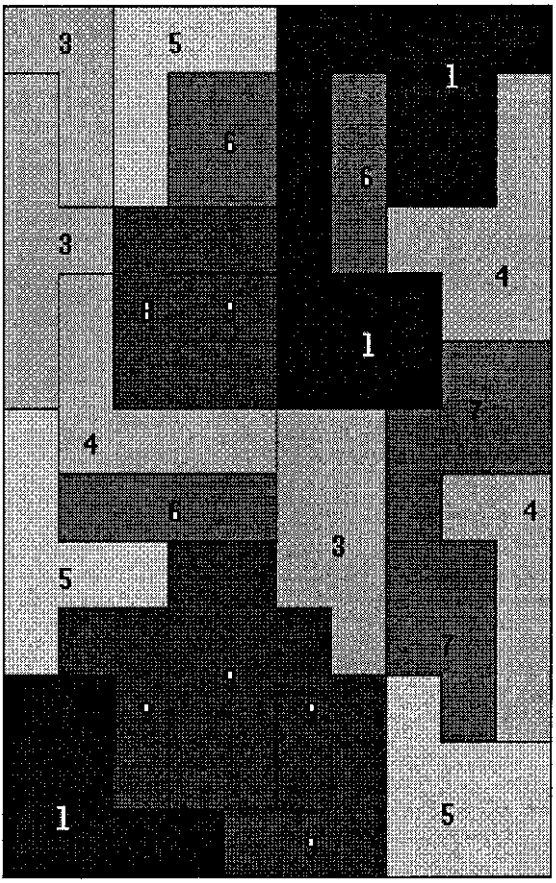
Cet exemple se veut surtout illustratif de la méthode. Sa portée est atténuée par le fait que, lorsqu'on travaille sur un faible nombre d'unités, il y a peu de solutions locales et l'on converge en général vers la solution globale.

6.1.1 Les données

La **carte A** décrit les données qui ont été utilisées pour cet exemple d'école. Comme indiqué ci-dessus, il y a 23 « communes » dont l'effectif est égal à la superficie de leur territoire. La variable d'intérêt (qui figure dans la carte à l'intérieur de chaque commune) a été générée aléatoirement. La répartition en 5 classes est la répartition faite automatiquement par la procédure GMAP de SAS.

Notons que, pour cet exemple, deux « communes » qui ont un angle en commun sont considérées comme contiguës.

CARTE A : situation initiale



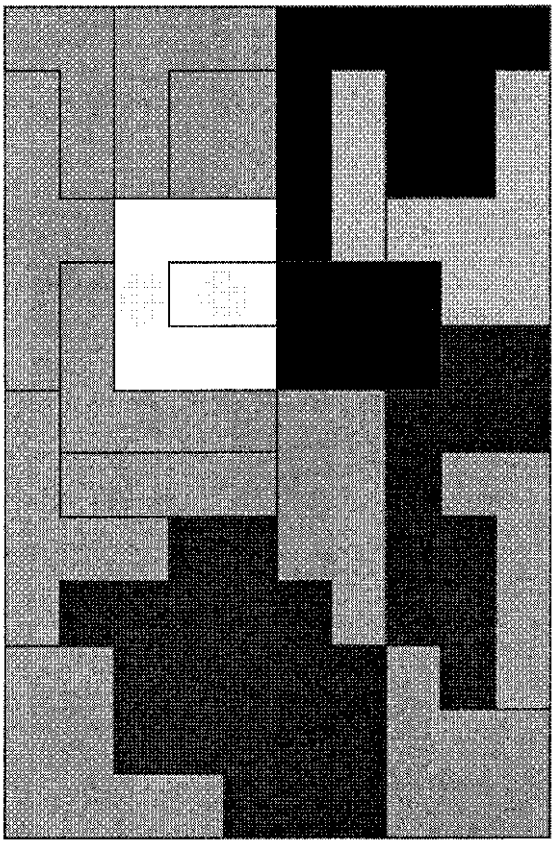
INTE 1.6 3.2 4.8 6.4 8

6.1.2 Après la CAH contiguë

La **carte B** représente le résultat de la CAH contiguë visant à partitionner l'ensemble des « communes » en 5 classes en cherchant à minimiser la variance intra.

La partition obtenue ne respecte pas les contraintes de taille fixée : la taille devait être comprise en 13 et 40.

CARTE B : la CAH contiguë



Résultats :

- 5 classes
- Moyenne et effectif des classes :

1	18
3,89	62
4,6	10
8,22	9
8,29	31

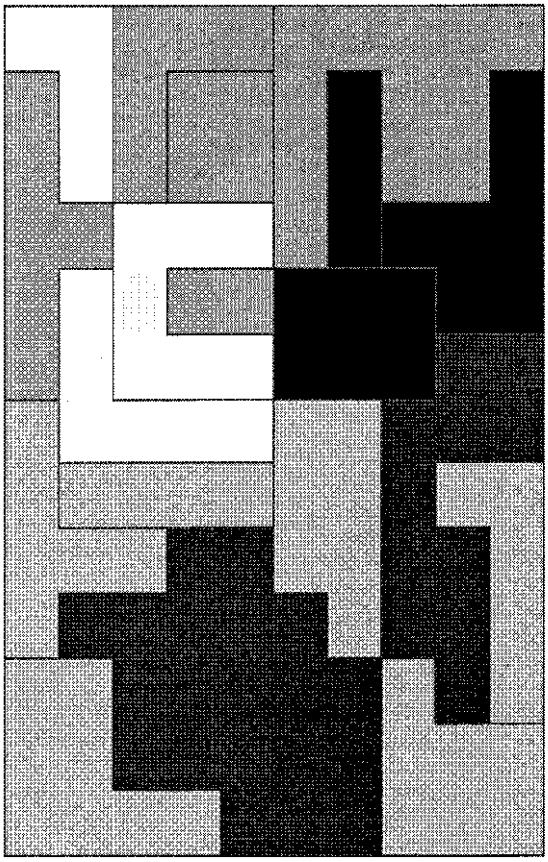
- V intra = 1,35

6.1.3 Deux échanges pour la taille

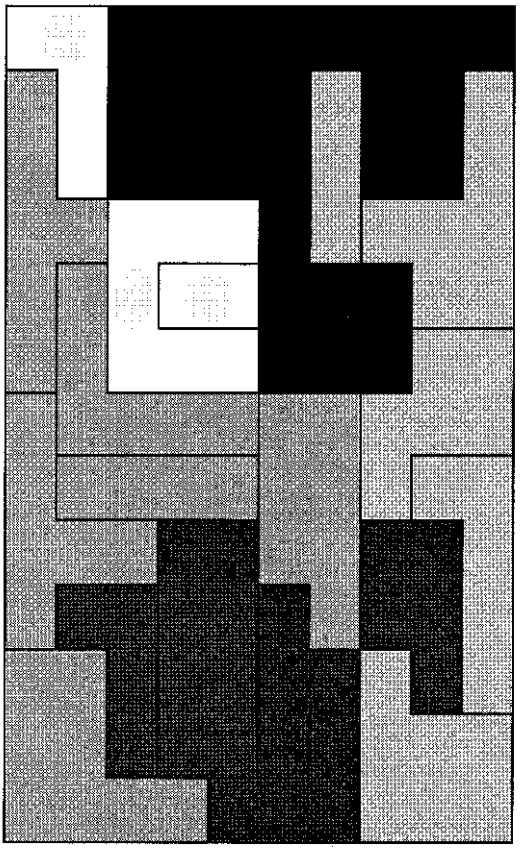
Les **cartes C et D** représentent deux résultats après les procédures d'échanges pour respecter les contraintes de taille. Ces deux résultats sont obtenus après différentes permutations aléatoires de l'ordre de traitement des « communes » comme indiqué plus haut.

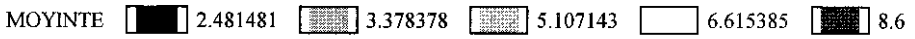
Le second résultat est meilleur que le premier (la variance intra vaut 2,65 au lieu de 3,42) et sera utilisé comme répartition initiale pour la troisième étape.

CARTE C : un 1er échange pour la taille



CARTE D : un échange de taille meilleur





Résultats (carte C) :

- 5 classes
- Moyenne et effectif des classes :

3,25	16
3,34	29
3,73	37
5,41	17
8,29	31

V intra = 3,42

Résultats (carte D) :

- 5 classes
- Moyenne et effectif des classes :

2,48	27
3,38	37
5,11	28
6,62	13
8,60	25

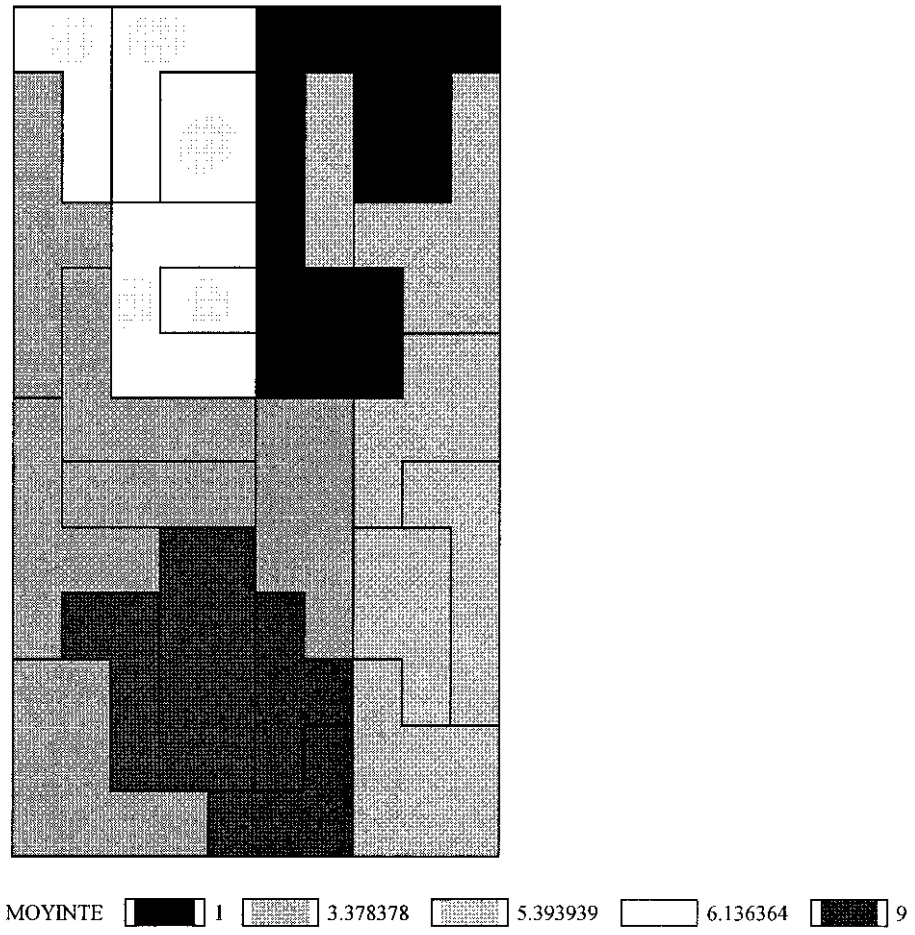
- V intra = 2,65

6.1.4 Le résultat définitif

La **carte notée E** représente le résultat définitif après les différents échanges minimisant la variance intra tout en vérifiant les contraintes de taille.

La variance intra est supérieure à celle obtenue après la CAH contiguë, mais présente une amélioration sensible par rapport celle obtenue après l'étape « taille ».

CARTE E : Minimisation de la variance



Résultats définitifs :

- 5 classes
- Moyenne et effectif des classes :

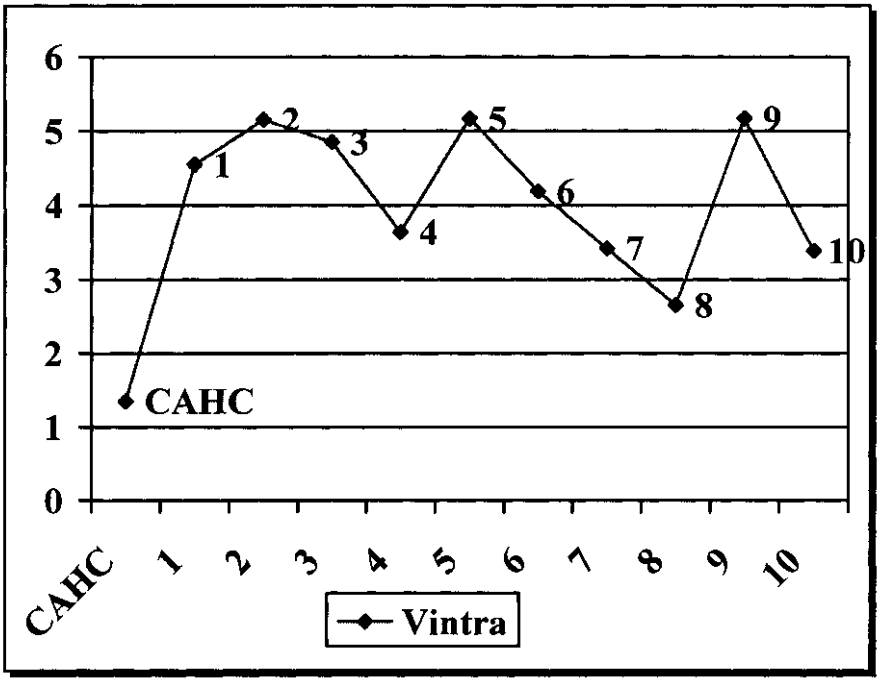
1	18
3,38	37
5,39	33
6,14	22
9	20

- V intra = 1,79

6.1.5 De l'influence de l'ordre des traitements des communes

La figure F représente la valeur de la variance intra après différents tris aléatoires de l'ordre de passage des « communes » dans l'algorithme d'échange pour la taille.

La macro SAS permet de choisir un nombre d'itérations de cette procédure d'échange et utilise la meilleure pour la suite de la procédure.



6.2 Applications spatiales.

Ces applications peuvent être réalisées dans le cadre décrit dans le § 1 : on cherche à construire des classes de communes, sur la base d'un critère de similitude, ici le revenu moyen déclaré à l'IRPP (sources 1996). La procédure a été testée sur le département des Alpes-Maritimes, tout d'abord globalement, puis en se restreignant aux communes rurales (de façon, notamment, à faire apparaître des non-connexités dans la population de base).

Plusieurs variantes ont été regardées : tout d'abord en fixant le nombre de classes, sans seuil de taille, puis en imposant un seuil de taille minimum, puis en prenant des seuils de taille analogues à ceux utilisés dans l'échantillon-Maître (entre 3600 et 7200 habitants), ce qui a imposé de mettre à part les communes de taille supérieure.

Une application en grandeur réelle a été construite pour la ***constitution de l'échantillon-Maître pour les extensions régionales d'enquête (EMEX)***[14]. Il s'agissait de constituer des groupes de communes (GRCOM) à l'intérieur de chacune des 253 UP rurales de l'EMEX.

L'algorithme décrit ici a été utilisé pour constituer, de manière automatique, des groupes de communes contiguës (au sens géographique) tout en privilégiant des associations de petites et grosses communes au sein des groupes : le critère retenu était donc la maximisation de l'hétérogénéité intra-GRCOM des tailles des communes. De plus, un seuil de taille a été imposé : 100 résidences principales au minimum dans chaque GRCOM.

La visualisation cartographique de quelques UP s'est révélée concluante, même si parfois quelques (rares) GRCOM ont tendance à "serpenter" dans l'UP.

Au total, 1880 GRCOM ont été créés sur la base de 3557 communes. Après la procédure, il est resté toutefois 12 GRCOM de moins de 100 résidences principales dont les 4 plus petits varient entre 39 et 63 résidences principales. Il s'agit en réalité de communes isolées au sein de l'UP, qui ont été laissées en l'état.

Notons enfin que la plupart des applications évoquées sont relatives à des regroupements de communes ; mais, bien entendu, la méthode peut s'appliquer pour réaliser des ***typologies infra-communales***, en regroupant par exemple des IRIS (pour les communes où ce découpage est réalisé). Cela nécessite néanmoins de disposer d'une matrice de contiguïté à ce niveau géographique.

[14] Effectuée par G. BOURDALLE et L. WILMS.

6.3 La « chaîne », ou contiguïté linéaire.

Bien que ce type d'application soit justiciable du même type d'approche théorique, elle a l'intérêt de montrer un autre cas d'utilisation de la méthode.

Supposons que, sur la population de référence, soit défini un *critère ordinal*. Par exemple, la taille en nombre de salariés permanents dans une population d'établissements, un indice de difficulté dans la description d'un chantier archéologique etc...

On peut alors classer la population de référence selon ce critère. Si l'on dispose d'une autre variable explicative, vis-à-vis de laquelle seront calculés les critères d'homogénéité (ou d'hétérogénéité) des classes, on peut chercher à partitionner de manière optimale la population étudiée, en un nombre de classes fixé, avec d'éventuels seuils sur les effectifs des classes et en *astreignant ceux-ci à respecter l'ordinalité*. Ainsi, chaque classe constituée doit pouvoir être décrite in fine comme l'ensemble des unités dont la variable ordinale correspondante prend une valeur délimitée entre deux bornes.

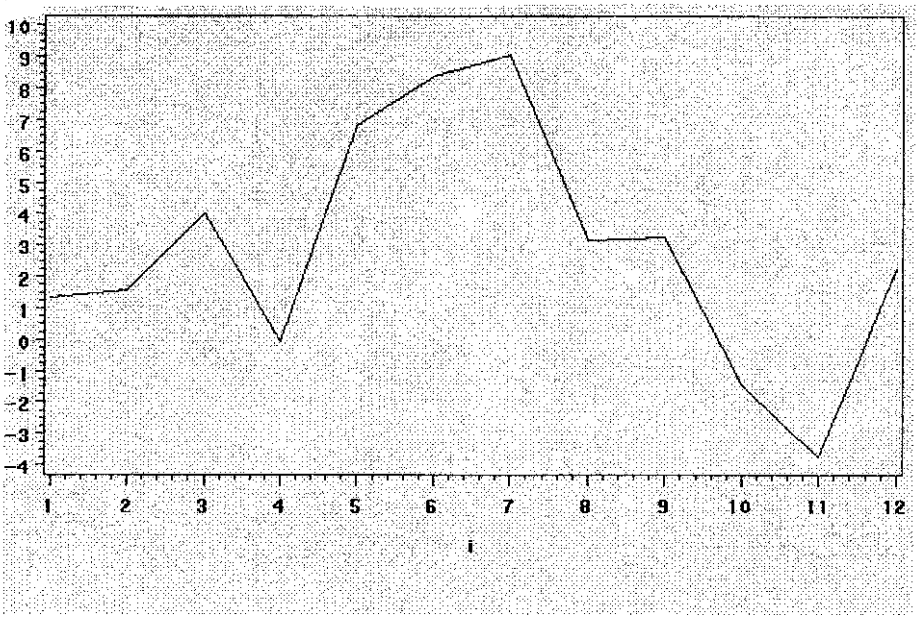
Il s'agit bien d'une application de la méthode générale : il faut, pour cela, définir la contiguïté à partir de l'ordinalité : *une unité (à l'exception des extrêmes) est contiguë à son prédécesseur et à son successeur exclusivement.*

Une autre application de ce cas de figure peut être à nouveau géographique : partitionner les communes littorales, par exemple, en K classes homogènes vis-à-vis du chiffre d'affaires moyen au m² généré par le tourisme, tout en respectant la linéarité de la côte (on aura ainsi des classes définies de la sorte : une classe composée des communes situées entre Menton et Nice, puis une autre, de celles situées entre Nice et Cannes etc...).

A titre d'illustration, on donne ci-dessous un exemple d'école de contiguïté linéaire, permettant de comparer l'effet de deux critères opposés, selon que l'on maximise ou que l'on minimise la variance intra.

6.3.1 Les données utilisées

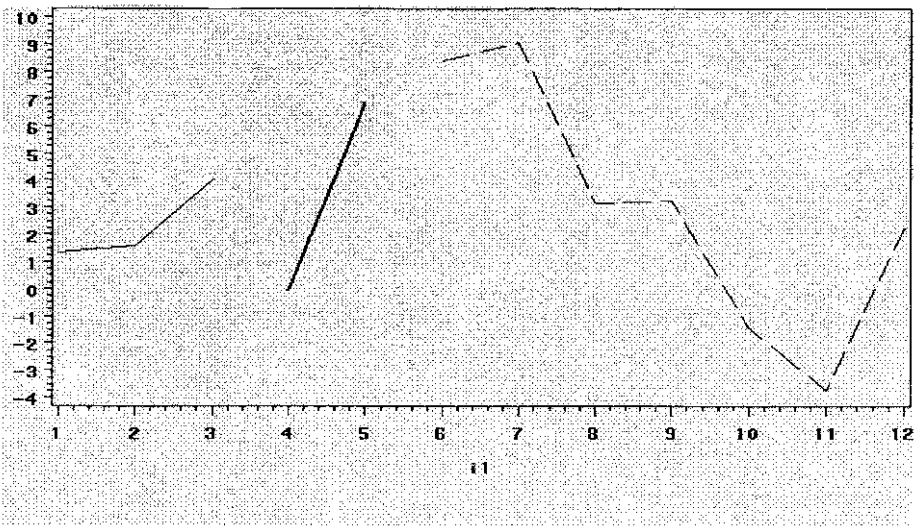
Les données utilisées sont les suivantes : il s'agit de 12 points, numérotés de 1 à 12. La variable d'intérêt est représentée par la courbe décrite dans la figure ci-dessous et chaque point a un effectif égal à 1. Comme indiqué ci-dessous, chaque point (n) est contigu aux points (n-1) et (n+1).



Deux types d'agrégation ont été appliqués à ces données : la première visait à agréger les points en 3 classes contiguës les plus *hétérogènes* possibles selon la variable d'intérêt choisie, la seconde à agréger ces mêmes points en 3 classes les plus *homogènes* possibles.

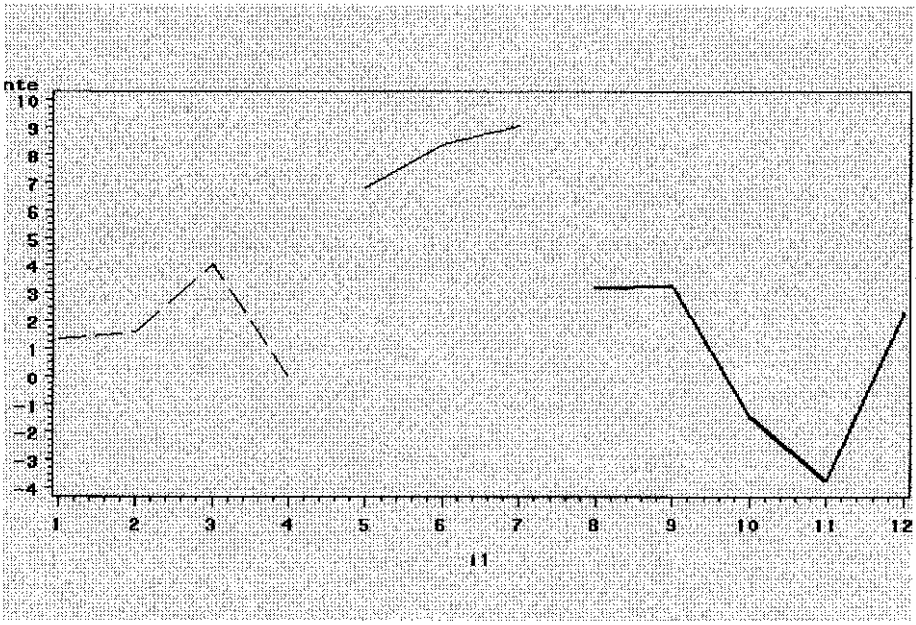
6.3.2 La maximisation de la variance intra

Il s'agit ici de créer des classes les plus hétérogènes possibles (il en résulte que les moyennes des classes sont les plus proches possibles). Le résultat est le suivant :



6.3.3 La minimisation de la variance intra

Il s'agit ici de créer des classes les plus homogènes possibles ; ces classes seront donc les plus différentes entre elles. Le résultat est donné dans la graphique ci-dessous :



6.4 Extensions de la notion de contiguïté.

Un certain nombre de zonages géographiques sont fondés sur des ***flux mettant en jeu deux communes*** (exemple : les flux domicile-travail qui président à la constitution des zones d'emploi et participent, parmi d'autres critères, à la définition des aires urbaines).

Des outils existent pour réaliser ce type de zonages[15].

[15] Par exemple , MIRABEL : Méthode informatisée de recherche et d'analyse de bassins par l'étude de liaisons.

L'algorithme développé dans ce papier permettrait de traiter également ces cas de figure au moyen d'une extension de la notion de contiguïté ; la « contiguïté », dans sa nouvelle acception, serait alors définie ainsi : deux communes sont contiguës si leurs territoires se touchent géographiquement ou si les flux migratoires entre elles satisfont une certaine condition, en termes absolus ou relatifs.

7. Conclusion.

7.1 Des extensions possibles de la méthode.

La méthode proposée est justiciable de plusieurs types d'extensions.

7.1.1 Cas de plusieurs variables d'intérêt.

On a traité jusqu'à présent essentiellement le cas d'une seule variable d'intérêt. L'extension à plusieurs variables peut se faire de différentes façons sans difficulté :

- prendre comme distance une norme euclidienne, dans l'espace des L variables d'intérêt (chaque unité de base est représentée par un vecteur de $\Re^L$). Pour éviter les problèmes d'unité de mesure ou d'incommensurabilité ou encore d'écarts très grands entre les dispersions des variables, on aura intérêt à normer chacune d'entre elles par leur variance empirique dans la population. On construit alors une *inertie* globale égale à la somme des variances des variables d'intérêt normées. On peut aussi, en fonction d'une hiérarchie subjective a priori, pondérer ces variables de manière adéquate dans l'expression de l'inertie.
- raisonner sur les variables principales issues d'une ACP des données initiales. C'est l'approche utilisée par L. LEBART dans son papier.

7.1.2 Homogénéité ou hétérogénéité.

Formellement, les problèmes sont identiques, comme on l'a vu : on maximisera une variance intra au lieu de la minimiser pour constituer des classes hétérogènes. Cette transposition doit bien entendu intervenir à toutes les étapes de la procédure (initialisation, procédure d'échange...).

Cette option a été implémentée dans les programmes informatiques.

7.1.3 Pondération des différentes contraintes.

On a vu que la combinaison des trois types de contraintes : nombre de classes, seuils de tailles des classes et connexité pouvaient être incompatibles ; la procédure doit pouvoir détecter ces impossibilités et l'opérateur doit alors intervenir pour modifier ses paramètres.

Néanmoins, on peut imaginer une extension consistant à relâcher certaines contraintes ou à modifier leur hiérarchie : à condition que l'on édicte a priori des seuils de relâchement, la procédure doit pouvoir générer automatiquement des solutions approchées ; plus généralement, on a vu que ce relâchement des contraintes peut s'interpréter comme une pondération entre les critères objectifs dans le processus de minimisation : miminiser une combinaison de la variance intra et d'une fonction d'écart par rapport aux seuils de taille ou aux nombres de classes, par exemple.

7.2 Apport de la méthode proposée.

Au total, la méthode proposée, par rapport aux outils traditionnels d'analyse des données, fournit un procédé systématique de regroupements d'unités statistiques sous un quadruple jeu de contraintes :

- **Contiguïté des unités regroupées** (qui entraîne la connexité des classes) : c'est la contrainte essentielle, celle qui est à l'origine de la mise en oeuvre de procédures spécifiques ;
- **Seuils d'effectifs** maximaux et minimaux pour les classes constituées ;
- **Nombre de classes** à constituer. Ces deux critères doivent évidemment être cohérents entre eux ;
- **Similitude (ou dissemblance) des unités regroupées**, traduite par l'objectif d'une minimisation (ou d'une maximisation) de la variance intra d'une variable d'intérêt, dans la décomposition de la variance totale selon la partition en classes.

Comme on l'a vu, l'optimalité absolue n'est en général pas accessible, mais la méthode permet de s'approcher d'optima relatifs.

L'intérêt réside aussi dans la *paramétrisation* de ces différentes contraintes : l'opérateur doit pouvoir (dans le respect de la cohérence rappelée ci-dessus) modifier les valeurs des contraintes et vérifier l'impact de ses modifications sur les résultats finaux. En revanche, s'il est clair que l'intervention manuelle est toujours nécessaire, ne serait-ce que pour la phase de validation et d'analyse des résultats

(impliquant évidemment un retour sur les valeurs des paramètres introduits), la méthode est entièrement automatisée.

Une *macro informatique* a été écrite, qui réalise de façon conviviale les opérations ci-dessus ainsi que les extensions évoquées dans le paragraphe précédent[16]. Le couplage avec des outils de cartographie, lorsque l'on se replace dans le contexte géographique du problème, permettra une meilleure visualisation des résultats (notamment, pour vérifier leur viabilité selon des critères non pris en compte : contraintes d'accessibilité en termes de voies de communication, par exemple, entre les différentes communes regroupées sur la base d'une contiguïté plane...).

7.3 Application au problème de l'échantillon-Maître.

L'origine du problème : constituer des unités primaires pour l'échantillon-Maître, a nécessité de mettre en relation théorie des sondages et analyse des données. Il convient de revenir en conclusion sur la portée de la méthode dans le cadre de ce problème originel.

La mise en oeuvre de cet outil permettra de générer automatiquement, en fonction des différents paramètres, variables d'intérêt ou critères introduits (minimisation / maximisation), des unités primaires fictives dans les zones rurales, qu'on pourra confronter aux unités primaires effectivement retenues dans l'échantillon-Maître 1999.

L'intérêt réside dans la possibilité de simuler plusieurs tirages d'enquêtes dans ces unités primaires et de comparer la précision des estimateurs des totaux de variables de référence avec celle que l'on obtiendrait dans l'échantillon-Maître et éventuellement avec les valeurs connues de ces totaux dans le recensement.

Ces travaux empiriques, à conduire au cours de la période à venir, permettront de mieux comprendre les comportements de différents plans de sondage, dans le cadre des tirages à deux degrés qui président à la logique de l'échantillon-Maître. Ils seront utiles pour la préparation ultérieure de la construction de nouvelles bases de sondage à partir des bases de logements construites dans le cadre du recensement rénové de la population.

[16] Cette macro peut être obtenue sur demande auprès des auteurs.

REFERENCES BIBLIOGRAPHIQUES.

C. ROCHE : « exemple de classification hiérarchique avec contraintes de contiguïté : le partage d'Aix-en-Provence en quartiers homogènes », les Cahiers de l'Analyse des Données, Vol III - 1978 - n°3.

L. LEBART : « Programme d'agrégation avec contraintes », les Cahiers de l'Analyse des Données, Vol III - 1978 - n°3.

Sas Institute, Documentation technique : « la procédure Fastclus ».

Annexe A
Liens avec la théorie des sondages

Considérons un plan de sondage à deux degrés où les tailles des échantillons du 1er et du 2nd sont fixes et dans lequel le second degré de tirage est aléatoire simple sans remise au sein de chaque unité primaire. Alors, l'estimateur de HORVITZ-THOMSON $\hat{\hat{T}}$ du total d'une variable d'intérêt Y définie sur l'ensemble de la population, a pour expression : $\hat{\hat{T}} = \sum_{j \in s_1} \frac{1}{\Pi_j^1} N_j \overline{y_j}$.

Sa variance a pour expression générale :

$$V\hat{\hat{T}} = \frac{1}{2} \sum_j \sum_{l \neq j} \left(\frac{T_j}{\Pi_j^1} - \frac{T_l}{\Pi_l^1} \right)^2 \left(\Pi_j^1 \Pi_l^1 - \Pi_{jl}^1 \right) + \sum_{j=1}^{K} \frac{N_j}{\Pi_j^1} \left(\frac{N_j}{n_j} - 1 \right) S_j^{*2}$$, avec

les notations suivantes :

Π_j^1 désigne la probabilité d'inclusion d'ordre 1 de l'unité primaire j au 1er degré.

Π_{jl}^1 désigne la probabilité d'inclusion d'ordre 2 des unités primaires j et l au 1er degré.

T_j représente le vrai total de la variable d'intérêt sur l'unité primaire j.

K représente le nombre total d'unités primaires constituées dans la population.

N_j est le nombre d'unités secondaires dans l'unité primaire j (=taille).

n_j est la taille de l'échantillon du second degré au sein de l'unité primaire j.

S_j^{*2} est la vraie variance de la variable d'intérêt au sein de l'unité primaire j, définie par $S_j^{*2} = \frac{1}{N_j - 1} \sum_{k=1}^{N_j} (Y_{jk} - \overline{Y}_j)^2$ où Y_{jk} est la valeur de la variable d'intérêt sur l'unité secondaire k appartenant à l'unité primaire j et $\overline{Y_j}$ la moyenne de cette variable d'intérêt sur cette même unité primaire.

Cette variance se décompose donc en la somme de deux termes : le premier traduit la variance liée au 1 er degré de tirage et le second est relatif au second degré.

Ces deux termes varient en général de façon antinomique.

Cette expression se simplifie dans un certain nombre de cas particuliers, définis notamment par la nature du tirage du 1er degré.

1. Le 1er degré est une stratification.

Dans ce cas, toutes les unités primaires sont tirées ($k=K$) et le premier terme de l'expression de la variance disparaît. La variance totale est entièrement commandée par le second terme (qui correspond à la variance due au 2nd degré). Cela conduit à rechercher des S_j^{*2} petits, c'est à dire **des unités primaires très homogènes**.

2. Le 1er degré est un sondage aléatoire simple

Dans ce cas, le terme de variance dû au 1er degré de tirage s'écrit : $K\frac{K-k}{k}\ V_i^T$,

dans laquelle V_i^T est la variance inter strates des totaux de la variable d'intérêt. Elle vaut : $V_i^T = \frac{1}{K-1}\sum_{j=1}^{K}\left(T_j - \overline{T}\right)^2$, avec $\overline{T} = \frac{1}{K}\sum_{j=1}^{K} T_j = \frac{T}{K}$, où T est le total de la variable d'intérêt sur l'ensemble de la population.

On voit que ce terme risque d'être important dans l'expression de la variance totale. Si tel est le cas, on aura intérêt à le réduire en faisant en sorte que les totaux par strate soient relativement voisins, c'est-à-dire que la variance inter soit faible, donc que les **unités primaires soient plutôt hétérogènes**.

Si toutes les unités primaires sont de même taille N_0, le plan de sondage équiprobable sans remise au 1er degré est alors équivalent au tirage avec probabilité proportionnelle à la taille et en constitue donc une approximation dans ce cas limite. On a alors les formules suivantes : $T_j = N_0\,\overline{Y_j}$ et $\overline{T} = N_0\,\overline{Y}$, et la taille de la population vaut $N = K\,N_0$.

Dans ce cas, la variance inter strate vaut : $V_i^T = \frac{N_0^2}{K-1}\sum_{j=1}^{K}\left(\overline{Y_j} - \overline{Y}\right)^2$.

On en déduit que la variance de l'estimateur du total vaut :

$$V\hat{\hat{T}} = K\left(\frac{K}{k}-1\right)\frac{N_0^2}{K-1}\sum_{j=1}^{K}\left(\overline{Y_j}-\overline{\overline{Y}}\right)^2 + \frac{N}{k}\sum_{j=1}^{K}\left(\frac{N_o}{n_j}-1\right)S_j^{*2}.$$

Or, l'équation d'analyse de la variance s'écrit ici :

$$VY = \frac{N_0}{N}\sum_{j=1}^{K}\left[\frac{N_0-1}{N_0}S_j^{*2} + (Y_j-\overline{\overline{Y}})^2\right] \approx \frac{N_0}{N}\sum_{j=1}^{K}\left[S_j^{*2} + (Y_j-\overline{\overline{Y}})^2\right].$$

On obtient donc la formule suivante :

$$V\hat{\hat{T}} \approx K\left(\frac{K}{k}-1\right)\frac{N_0^2}{K-1}\left[\frac{N}{N_0}VY - \sum_{j=1}^{K}S_j^{*2}\right] + \frac{N}{k}\sum_{j=1}^{K}\left(\left(\frac{N_0}{n_j}-1\right)S_j^{*2}\right)$$

On en déduit que :

$$\boxed{V\hat{T} \approx \frac{N^2}{k}VY + \frac{NN_0}{k}\sum_{j=1}^{K}S_j^{*2}\left(\frac{1}{n_j}-1\right)}$$

On constate que le terme qui fait intervenir les variances intra des unités primaires est doté d'un signe négatif, ce qui conduira à chercher à maximiser la variance intra pour minimiser la variance totale de l'estimateur, c'est-à-dire à constituer des **unités primaires hétérogènes**.

Annexe B
A propos du nombre de partitions de n communes en p classes

Pour bâtir une formule de récurrence, essayons de construire une partition d'un ensemble de $(n+1)$ éléments en $p+1$ classes à partir des partitions d'un ensemble de n éléments en p classes.

Pour cela, on considère un ensemble à n éléments et on s'intéresse à un n+1 ème élément, noté *e*. Deux cas disjoints sont possibles :

- Soit on l'affecte à l'une des classes existantes d'une partition de l'ensemble de n éléments en p+1 classes, soit p+1 choix de classes à multiplier par le nombre de partitions de n éléments en p+1 classes.
- Soit on crée une classe à lui tout seul, en la rajoutant aux p classes de la partition de l'ensemble de n éléments.

On a alors la formule de récurrence suivante :

$$f(n+1,p+1)=(p+1)\times f(n,p+1)+f(n,p)$$

Les conditions aux limites sont les suivantes :

- $f(n,1)=1$ car il n'y a qu'une manière possible de répartir n éléments en une classe.
- $f(n,n)=1$
- $f(n,p)=0 \quad dès\ que\ n<p$ car il est difficile de mettre n éléments dans un nombre plus grand de classes (celles-ci étant obligatoirement non vides) !!

Ainsi, on a, par exemple, le calcul suivant :

$f(3,2) = f(2,1) + 2 \times f(2,2) = 1 + 2 \times 1 = 3.$

On montre, au moyen de la formule, que, pour tout entier naturel n :

$f(n, 2) = 2^{n-1} - 1$

et que :

$$f(n, 3) = \frac{3^{n-1}}{2} + \frac{1}{2} - 2^{n-1}.$$

Le tableau suivant donne quelques valeurs de cette fonction.

Dans ce tableau, n est représenté en colonne et p en ligne. Ainsi, pour 23 éléments, il y a un peu plus de $9 * 10^{13}$ partitions en 5 classes.

Nombre de partitions d'un ensemble de n éléments (en ligne) en p classes (en colonne)

	1	2	3	4	5	6	7	8
1	1	0	0	0	0	0	0	0
2	1	1	0	0	0	0	0	0
3	1	3	1	0	0	0	0	0
4	1	7	6	1	0	0	0	0
5	1	15	25	10	1	0	0	0
6	1	31	90	65	15	1	0	0
7	1	63	301	350	140	21	1	0
8	1	127	966	1 701	1 050	266	28	1
9	1	255	3 025	7 770	6 951	2 646	462	36
10	1	511	9 330	34 105	42 525	22 827	5 880	750
11	1	1 023	28 501	145 750	246 730	179 487	63 987	11 880
12	1	2 047	86 526	611 501	1 379 400	1 323 652	627 396	159 027
13	1	4 095	261 625	2 532 530	7 508 501	9 321 312	5 715 424	1 899 612
14	1	8 191	788 970	1,0392E+07	4,0075E+07	6,3436E+07	4,9329E+07	2,0912E+07
15	1	16 383	2 375 101	4,2356E+07	2,1077E+08	4,2069E+08	4,0874E+08	2,1663E+08
16	1	32 767	7 141 686	1,7180E+08	1,0962E+09	2,7349E+09	3,2819E+09	2,1418E+09
17	1	65 535	2,1458E+07	6,9434E+08	5,6528E+09	1,7506E+10	2,5708E+10	2,0416E+10
18	1	131 071	6,4439E+07	2,7988E+09	2,8958E+10	1,1069E+11	1,9746E+11	1,8904E+11
19	1	262 143	1,9345E+08	1,1260E+10	1,4759E+11	6,9308E+11	1,4929E+12	1,7098E+12
20	1	524 287	5,8061E+08	4,5232E+10	7,4921E+11	4,3061E+12	1,1144E+13	1,5171E+13
21	1	1 048 575	1,7423E+09	1,8151E+11	3,7913E+12	2,6586E+13	8,2311E+13	1,3251E+14
22	1	2 097 151	5,2281E+09	7,2778E+11	1,9138E+13	1,6331E+14	6,0276E+14	1,1424E+15
23	1	4 194 303	1,5686E+10	2,9163E+12	9,6417E+13	9,9897E+14	4,3826E+15	9,7420E+15
24	1	8 388 607	4,7063E+10	1,1681E+13	4,8500E+14	6,0902E+15	3,1677E+16	8,2318E+16
25	1	1,6777E+07	1,4120E+11	4,6771E+13	2,4367E+15	3,7026E+16	2,2783E+17	6,9022E+17
26	1	3,3554E+07	4,2361E+11	1,8723E+14	1,2230E+16	2,2460E+17	1,6319E+18	5,7496E+18
27	1	6,7109E+07	1,2709E+12	7,4933E+14	6,1338E+16	1,3598E+18	1,1648E+19	4,7629E+19
28	1	1,3422E+08	3,8127E+12	2,9986E+15	3,0744E+17	8,2201E+18	8,2893E+19	3,9268E+20
29	1	2,6844E+08	1,1438E+13	1,1998E+16	1,5402E+18	4,9628E+19	5,8847E+20	3,2243E+21
30	1	5,3687E+08	3,4315E+13	4,8004E+16	7,7130E+18	2,9931E+20	4,1689E+21	2,6383E+22

(SUITE)

	9	10	11	12	13	14	15
1	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0
9	1	0	0	0	0	0	0
10	45	1	0	0	0	0	0
11	1 155	55	1	0	0	0	0
12	22 275	1 705	66	1	0	0	0
13	359 502	39 325	2 431	78	1	0	0
14	5 135 130	752 752	66 066	3 367	91	1	0
15	6,7128E+07	1,2663E+07	1 479 478	106 470	4 550	105	1
16	8,2078E+08	1,9375E+08	2,8937E+07	2 757 118	165 620	6 020	120
17	9,5288E+09	2,7583E+09	5,1206E+08	6,2022E+07	4 910 178	249 900	7 820
18	1,0618E+11	3,7112E+10	8,3910E+09	1,2563E+09	1,2585E+08	8 408 778	367 200
19	1,1446E+12	4,7730E+11	1,2941E+11	2,3467E+10	2,8924E+09	2,4358E+08	1,3917E+07
20	1,2011E+13	5,9176E+12	1,9008E+12	4,1102E+11	6,1069E+10	6,3025E+09	4,5233E+08
21	1,2327E+14	7,1187E+13	2,6827E+13	6,8330E+12	1,2049E+12	1,4930E+11	1,3087E+10
22	1,2420E+15	8,3514E+14	3,6628E+14	1,0882E+14	2,2497E+13	3,2952E+12	3,4562E+11
23	1,2320E+16	9,5934E+15	4,8643E+15	1,6722E+15	4,0128E+14	6,8629E+13	8,4794E+12
24	1,2062E+17	1,0825E+17	6,3100E+16	2,4930E+16	6,8888E+15	1,3621E+15	1,9582E+14
25	1,1679E+18	1,2032E+18	8,0236E+17	3,6226E+17	1,1449E+17	2,5958E+16	4,2994E+15
26	1,1202E+19	1,3200E+19	1,0029E+19	5,1495E+18	1,8506E+18	4,7790E+17	9,0449E+16
27	1,0656E+20	1,4320E+20	1,2352E+20	7,1823E+19	2,9207E+19	8,5411E+18	1,8346E+18
28	1,0067E+21	1,5385E+21	1,5019E+21	9,8540E+20	4,5151E+20	1,4878E+20	3,6061E+19
29	9,4530E+21	1,6392E+22	1,8060E+22	1,3327E+22	6,8551E+21	2,5345E+21	6,8969E+20
30	8,8301E+22	1,7337E+23	2,1505E+23	1,7798E+23	1,0244E+23	4,2338E+22	1,2880E+22

Annexe C : Le traitement de la contiguïté

1. Quelques notations

Dans la suite on notera :

1. $A_{ij}, i = 1..N, j = 1..N$ une matrice de contiguïté des N communes de la population, c'est à dire $A_{ij} = 0$ si les communes i et j ne sont pas contiguës et 1 sinon. Cette matrice est *symétrique* par hypothèse, ce qui correspond à la symétrie de la relation de contiguïté. On conviendra également qu'elle a des 1 sur la diagonale, ce qui correspond à la *réflexivité* de la relation de contiguïté (il s'agit d'une extension de la notion intuitive de contiguïté, qui sera utile par la suite). Partir d'une relation de contiguïté qui ne vérifierait cette dernière propriété pour aucun élément de la population conduirait à travailler ensuite sur la matrice A + I, au lieu de la matrice A.

2. Etant donné un sous-ensemble G, non vide, de la population de référence (qu'on baptisera *classe*), on notera G_j, le vecteur d'appartenance à la classe G : $G_j=0$ si j $\in$ G, et 1 sinon.

2. Lemme

Soient 2 classes G1 et G2. Les deux propositions suivantes sont équivalentes :

1. G1 et G2 sont contigus, c'est-à-dire il existe au moins un élément de l'une qui est contigu à un élément de l'autre. (F1)

2. $(G^1)'\ A\ G^2 > 0$ (F2).

2.1 Démonstration du lemme

Constatons d'abord que le réel $(G^1)'$ A G^2 est toujours positif ou nul, comme somme de produits de termes positifs ou nuls. La négation de (F2) est donc :
$(G^1)'$ A $G^2 = 0$.

Ce réel vaut $\sum_{i,j} G_i^1 A_{ij} G_j^2$, chacun des termes figurant dans cette expression étant aussi positif ou nul et la somme ayant lieu sur l'ensemble de la population. La négation de (F2) est donc équivalente à : $\forall$ i, j : $G_i^1 \, A_{ij} \, G_j^2 = 0$.

Cette expression est trivialement nulle pour les points i qui n'appartiennent pas à G1 et pour les points j qui n'appartiennent pas à G2. Pour tous les couples (i, j) appartenant à $G1 \times G2$, on a donc $A_{ij} = 0$. Par suite, aucun point de la première classe n'est contigu à un point de la seconde ; donc, les deux classes ne sont pas contiguës.

Il est clair que la réciproque est vraie, ce qui démontre le lemme.

2.2 Conséquences pratiques

Pour tester la contiguïté de deux classes, il suffit d'effectuer le produit scalaire décrit dans (F2).

3. Comment tester la connexité d'un ensemble de communes ?

Nous cherchons ici à ***déterminer si une classe G composée de k éléments est connexe ou pas***. Appelons, dans ce paragraphe, C la matrice de contiguïté réduite aux k éléments de la classe.

Nous allons montrer que :

La classe G est connexe si et seulement si C vérifie la propriété suivante : $C^{k-1} > 0$, c'est-à-dire que tous les éléments de la matrice C^{k-1} sont strictement positifs.

Nous allons fractionner la démonstration en deux parties : la première consiste à s'intéresser à la matrice C^{k-1}, puis à relier cette matrice à la connexité.

3.1 La matrice C^{k-1}

Considérons la matrice C^2.

Ses coefficients $C_{ij}^{(2)} = \sum_k C_{ik} C_{kj}$ ne sont différents de 0 que s'il existe au moins un élément k vérifiant $C_{ik} > 0$ et $C_{kj} > 0$, c'est-à-dire qu'il existe un élément k tel que $i \rightarrow k \rightarrow j$. La flèche représente ici une relation de contiguïté entre deux éléments. On dira que ce chemin est de longueur 2, que *k* soit distinct ou non de *i* et de *j*. On a donc l'équivalence suivante :

$C_{ij}^{(2)} > 0 \Leftrightarrow$ Il existe un chemin de longueur de 2 entre i et j.

Considérons le cas n.

Nous allons démontrer par récurrence que $C_{ij}^{(n)}$ est strictement positif si et seulement s'il existe un chemin de longueur n entre i et j.

On a la formule suivante $C_{ij}^{(n)} = \sum_k C_{ik}^{(n-1)} C_{kj}$.

Donc : $C_{ij}^{(n)}$ est strictement positif si et seulement s'il existe k tel que : $C_{ik}^{(n-1)} > 0$ et $C_{kj} > 0$, c'est-à-dire si et seulement s'il existe un chemin de longueur n-1 entre i et k et un chemin de longueur 1 entre k et j, pour un certain k, c'est-à-dire si et seulement s'il existe un chemin de longueur n entre i et j.

On a donc la propriété suivante :

$\forall$ n $\in$ N* : $C_{ij}^{(n)} > 0 \Leftrightarrow$ il existe un chemin de longueur n entre i et j.

3.2 La connexité de G

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

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

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

On a donc la propriété suivante :

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

CQFD.