Méthodes d'agrégation d'unités statistiques sous contrainte de contiguïté

Marc CHRISTINE
Insee, Direction de la Méthodologie
et de la coordination statistique et internationale

121212

Ce travail a été réalisé avec Michel ISNARD (Insee)

Plan de la présentation

- 1. Introduction
- 2. Deux exemples introductifs
- 3. Le cadre général
- 4. Description rapide de la méthode
- 5. Applications à des exemples simples
- 6. Application à une distance non euclidienne
- 7. Une nouvelle piste à explorer
- 8. Conclusion

1. Que veut-on faire ?

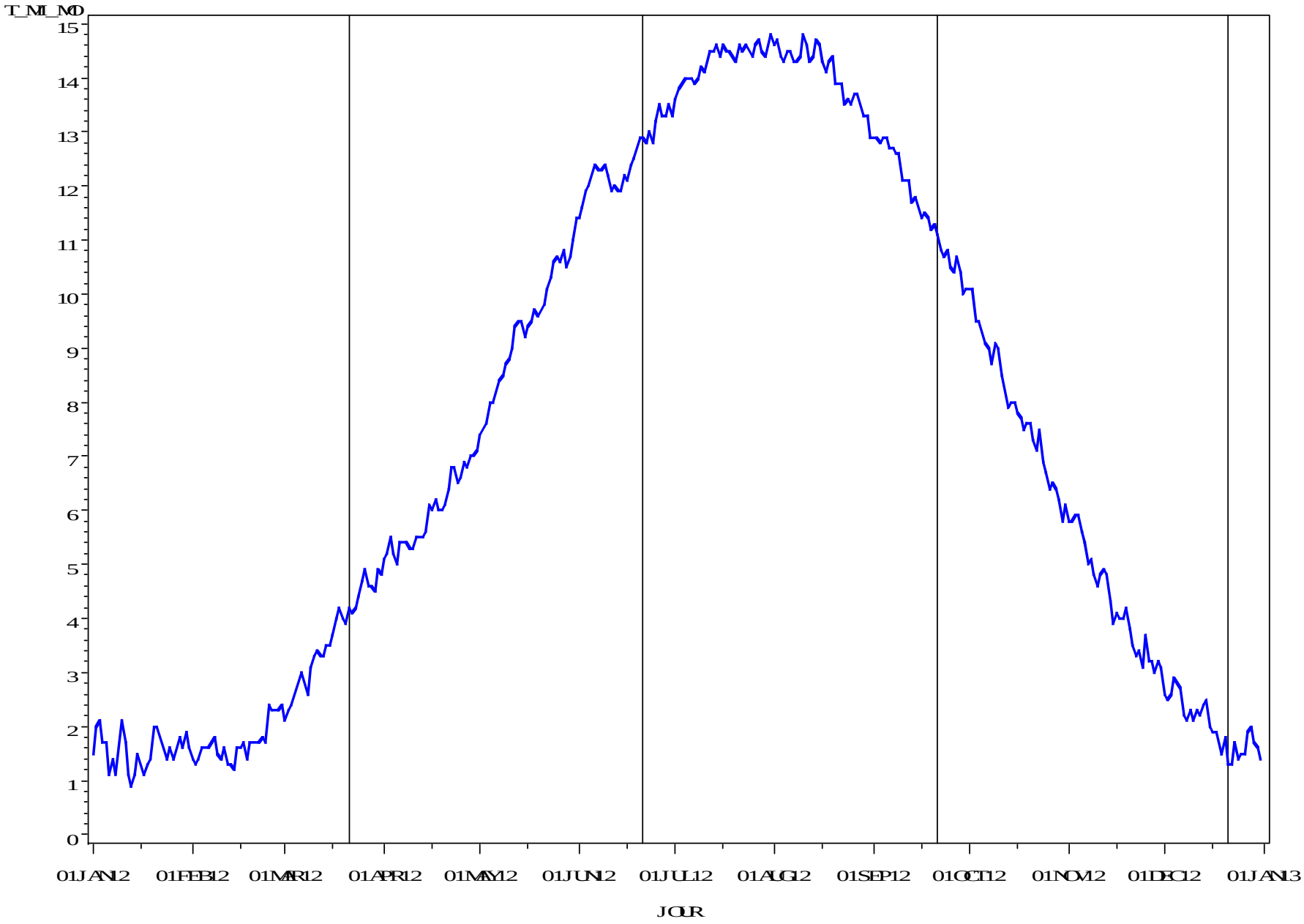
- A partir
  - d'une population d'unités statistiques *pondérées*
  - entre lesquelles a été définie une *distance* fondée sur des caractéristiques des unités (euclidienne ou pas) …
  - … et une relation de *contiguïté*

- Partager la population en K (nombre fixé par l'utilisateur) classes *connexes*, vérifiant des contraintes uniformes de *taille* (également fixées par l'utilisateur) et minimisant ou maximisant *l'inertie intra-classe*.

2. Deux exemples introductifs

- Premier exemple

Températures minimales moyennes à Paris depuis 1873



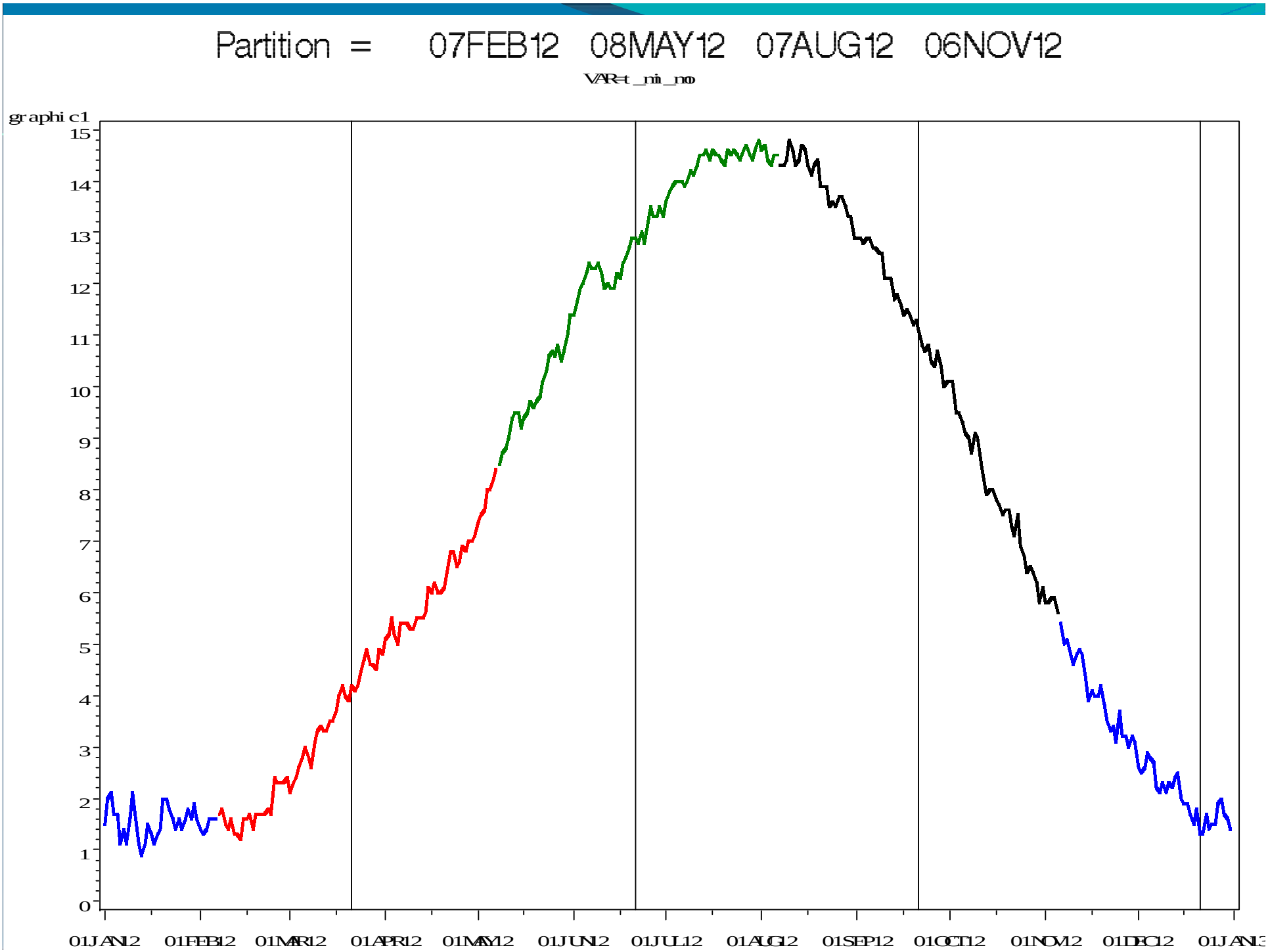
Données recueillies sur le site www.meteo-paris.com

Comment recréer des saisons ?

- Trouver toutes les partitions en plages de longueur comprises en 91 et 93 jours
  - 1er janvier et 31 décembre se touchant
- Calculer la variance intra-classe de chacune de ces partitions (pour la variable « température »)
- Trouver la partition ayant la plus petite variance intra-classe
  - Saisons les plus homogènes possibles.

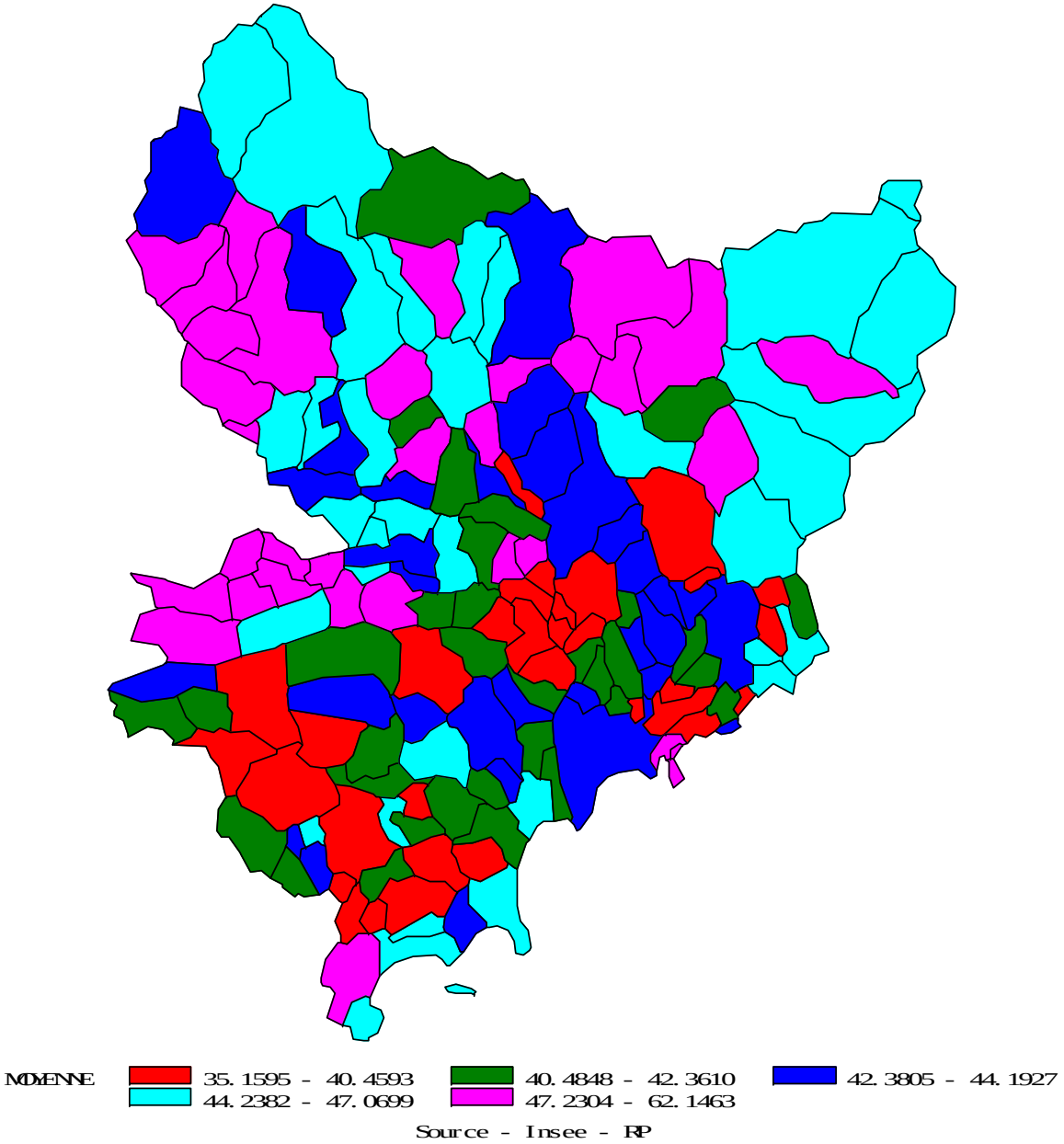
Comment recréer des saisons ?

- 920 partitions possibles


Partition = 07FEB12 08MAY12 07AUG12 06NOV12
VAR=t_mi_no
graphic1
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
01JAN12
01FEB12
01MAR12
01APR12
01MAY12
01JUN12
01JUL12
01AUG12
01SEP12
01OCT12
01NOV12
01DEC12
01JAN13

Deuxième exemple

Age moyen dans les communes des Alpes Maritimes



Nombre de partitions à tester

- 10^{111} partitions en 5 classes des 163 communes des Alpes Maritimes
  - Et … $0{,}000\ldots\varepsilon$ % partitions connexes
- Nécessité de trouver une autre méthode…
- … et de fournir un algorithme et un programme permettant d'atteindre des solutions.

3. Le cadre général

- Pour chaque unité *i* de la population de référence $\mathcal{P}$:

- Une variable d'intérêt numérique ou vectorielle : x_i
- Un poids : $\alpha_i > 0$
- Une taille numérique : T_i

- Centre de gravité de la population :

$$g = \frac{\sum_{i \in P} \alpha_i x_i}{\sum_{i \in P} \alpha_i}$$

- Variance ou *inertie* de la population :

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

- Equation d'***analyse de la variance*** pour un partitionnement de $\mathcal{P}$ en K classes :

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

$$\text{Avec : } \omega = \sum_{i \in P} \alpha_i \text{ et } \omega_k = \sum_{i \in P_k} \alpha_i$$

- On cherchera à *minimiser* ou *maximiser* l'inertie intra-classe :

$$I^a = \sum_{k=1}^{K} \frac{\omega_k}{\omega} I_k$$

- … qui s'écrit aussi :

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

- … et se généralise sous la forme :

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

- … faisant apparaître une distance $d_{i,j}$ entre les unités *i* et *j*.

- Agrégation de deux classes (***cas euclidien***) :
  - La variation d'inertie intra-classe résultante est :

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

  - Elle est positive ou nulle.

  - Si l'on agrège une unité à une classe, la variation devient :

$$\Delta I^a = \frac{(g_k - x_0)^2}{\omega} \frac{\omega_k \alpha_0}{\omega_k + \alpha_0}$$

- Cette variation d'inertie intra-classe peut s'interpréter :
  - Comme une distance entre deux classes…
  - … ou comme une distance entre une unité et une classe,
  - … calculées à partir des centres de gravité de ces classes et des poids afférents.

Le problème de la connexité

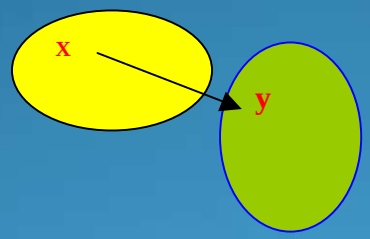
- **Sur un ensemble E, on définit une relation de *contiguïté* R, réflexive et symétrique.**

- **On la traduit par la matrice de contiguïté (réflexive et symétrique) :**

$$C_{i,j} = \begin{cases} 1 \text{ si les unités } i \text{ et } j \text{ sont contiguës} \\ 0 \text{ à défaut} \end{cases}$$

- Définition de la contiguïté entre parties :

  Deux parties A et B de E seront dites *contiguës* si et seulement si

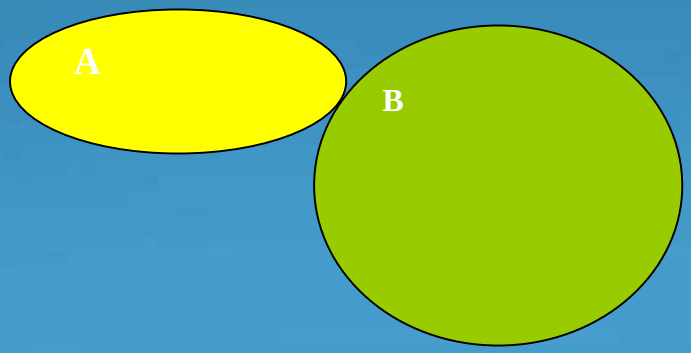


$$\exists\, x \in A, \exists\, y \in B : x\ R\ y$$

- Une partie A de E sera dite *connexe* si et seulement s'il n'existe pas de décomposition de A en deux parties *non contiguës* sous la forme :

$$A = C_1 \cup C_2$$

- La réunion de deux parties connexes et contiguës est connexe.



4. Description rapide de la méthode

- Une phase d'agrégation proche d'une classification ascendante hiérarchique.
- Une seconde phase d'échange entre les classes créées précédemment afin de respecter les critères de taille et d'optimiser l'inertie intra-classe.

Les contraintes initiales

- Le programme doit fournir des classes connexes respectant au mieux les contraintes données par l'utilisateur.

- L'utilisateur désirant utiliser cette méthode doit :
  - Fournir une liste d'unités statistiques pondérées
  - … et une distance entre deux quelconques de ces unités (construite à partir d'une variable d'intérêt)
  - Indiquer les unités contiguës
  - Indiquer le nombre de classes qu'il souhaite créer et les contraintes de taille (uniformes) qu'il souhaite voir respecter
  - Indiquer s'il souhaite maximiser ou minimiser l'inertie (=variance) intra-classe.

La première phase

- Bâtie sur le même algorithme qu'une CAH.
- A chaque étape, on choisit les classes à agréger : ***celles dont l'agrégation minimise ou maximise la variation d'inertie intra-classe.***
- **Cela équivaut à agréger soit les classes les plus proches, soit les plus éloignées, au sens de la distance entre classes définie ci-dessus.**

- Mais on limite l'agrégation aux classes *contiguës*, c'est-à-dire aux classes dont au moins une unité est contiguë à une unité de l'autre classe.

- **Les classes formées :**
  - **sont connexes,**
  - **mais ne respectent pas forcément les contraintes de taille.**

La seconde phase

- A partir des classes créées lors de la première phase,
  - On améliore le respect des contraintes de taille (et d'optimisation de l'inertie)…
  - … en échangeant des unités de classe à classe ou en transférant des unités d'une classe à une autre.
  - On privilégie d'abord le critère de taille…
  - …, puis, celui-ci rempli, le critère d'inertie.

- Il faut vérifier qu'un échange ou un transfert d'unités ne détruit pas la connexité des classes en résultant.

⇒ mise en œuvre de ***tests de connexité :***

- Pour une classe G :

$$Pour\ \ n > 0 : \left(C_{i,j}^{G}\right)^{(n)} > 0 \Leftrightarrow$$

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 $(C^G)^n$.

- Pour une classe G de q éléments :

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

- La procédure s'arrête quand plus aucun échange ni transfert n'améliore le respect des contraintes de taille ni l'optimisation du critère d'inertie.

- L'expérience montre que les contraintes de taille sont respectées, dès lors qu'elles ne sont pas trop « sévères ».

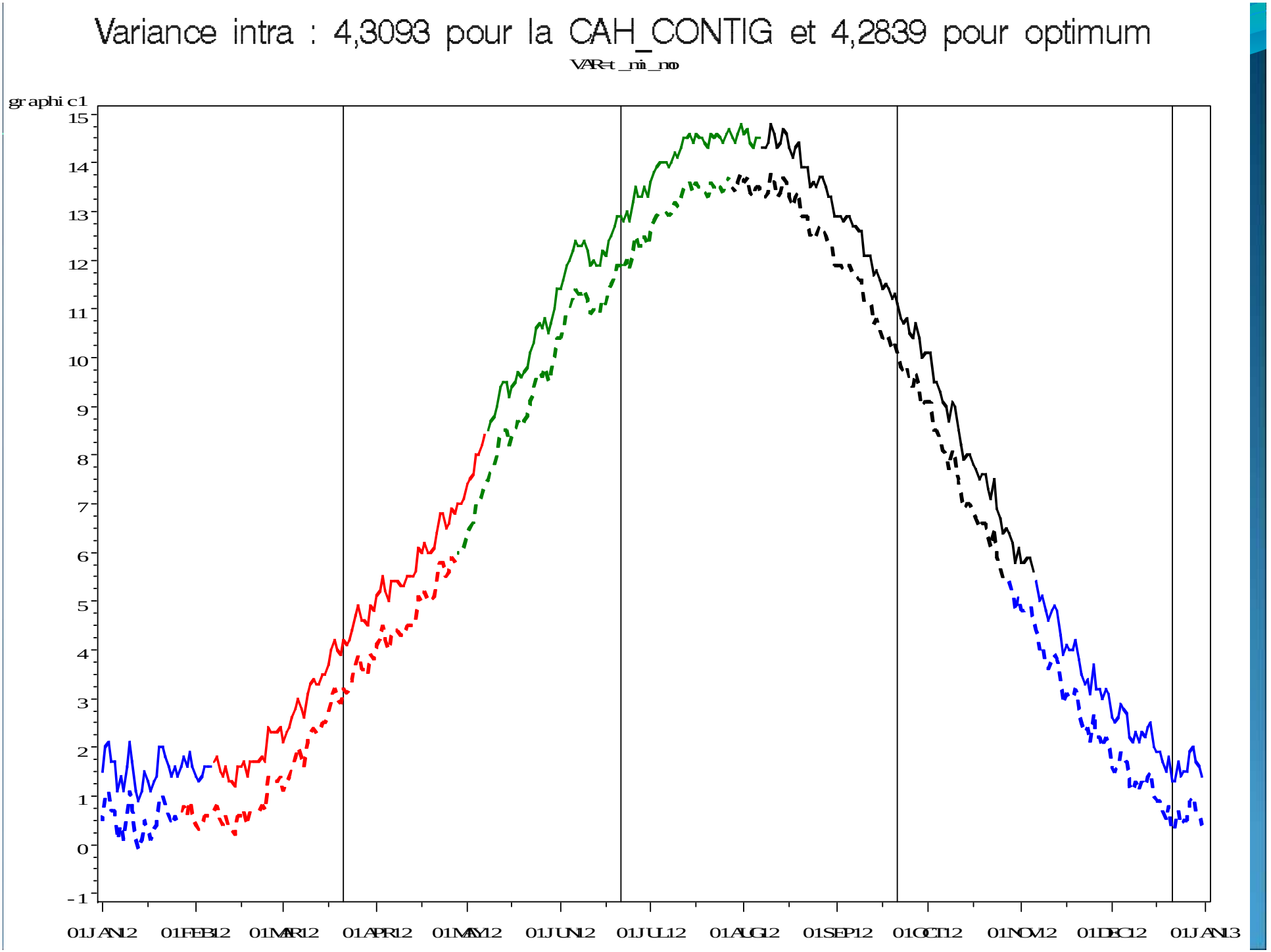
La méthode en résumé

- Tous les classes créées sont des classes connexes.
- Les contraintes de taille sont très fréquemment vérifiées.
- On n'est pas assuré de l'optimalité de l'inertie.

5. Retour sur les saisons

- Maximisation de l'inertie intra-saison (= variance de la température) avec des saisons de longueur comprise en 91 et 93 jours.
- Minimisation de l'inertie intra-saison avec les mêmes contraintes.
- La contiguïté est temporelle.

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



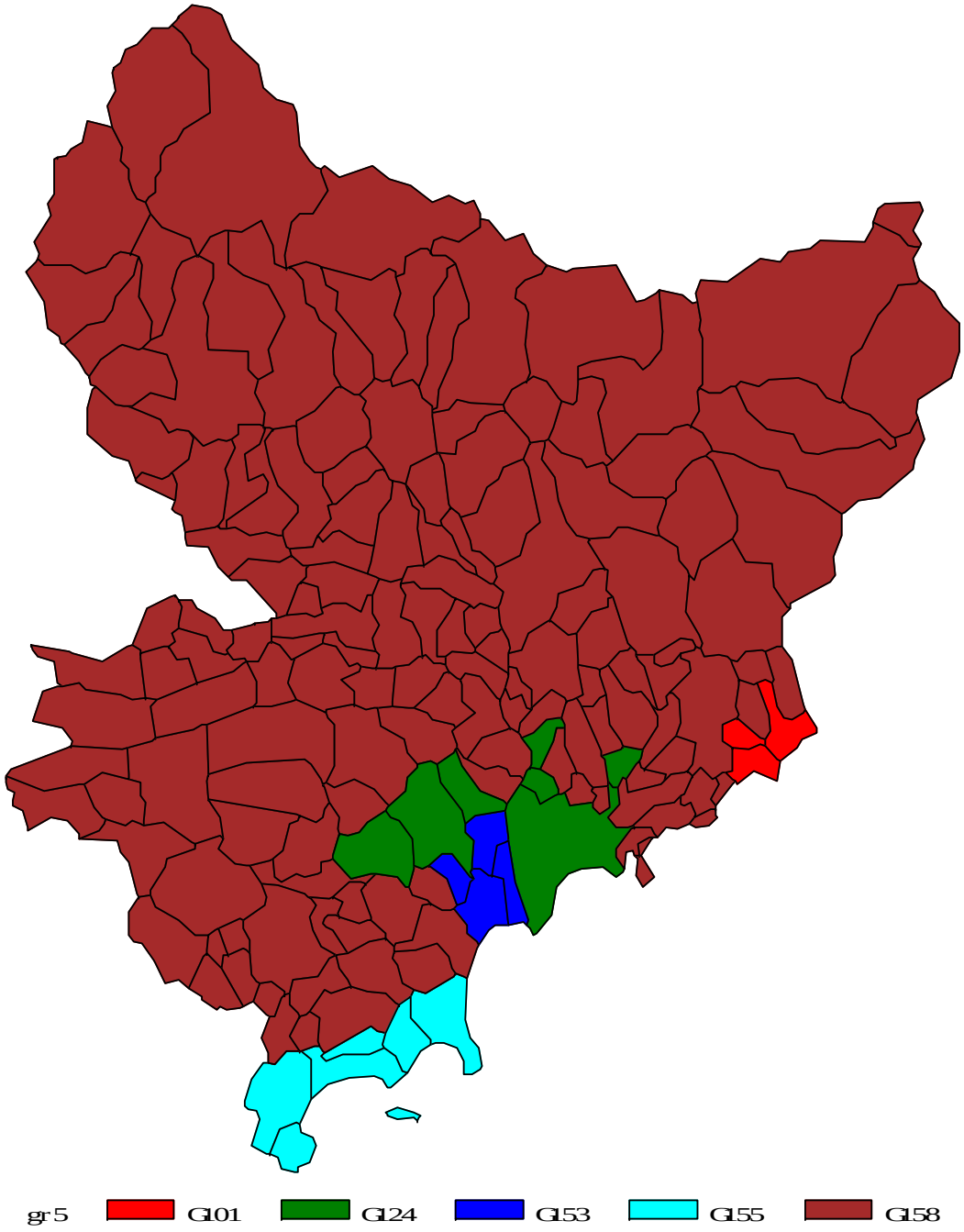
Retour sur les saisons

- Limites 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
- Variances intra-classe
  - 4,3093 pour la CAH CONTIG
  - 4,2839 pour l'optimum
- NON OPTIMAL … (mais écart faible)

Un exemple sur les Alpes-Maritimes

- Variable : Age moyen par commune (RP 2008)
  - Minimisation de la variance intra-classe : reclassement en 5 classes connexes les plus homogènes possibles,
  - Contraintes de taille : chaque classe doit comprendre entre 15 et 35% de la population.

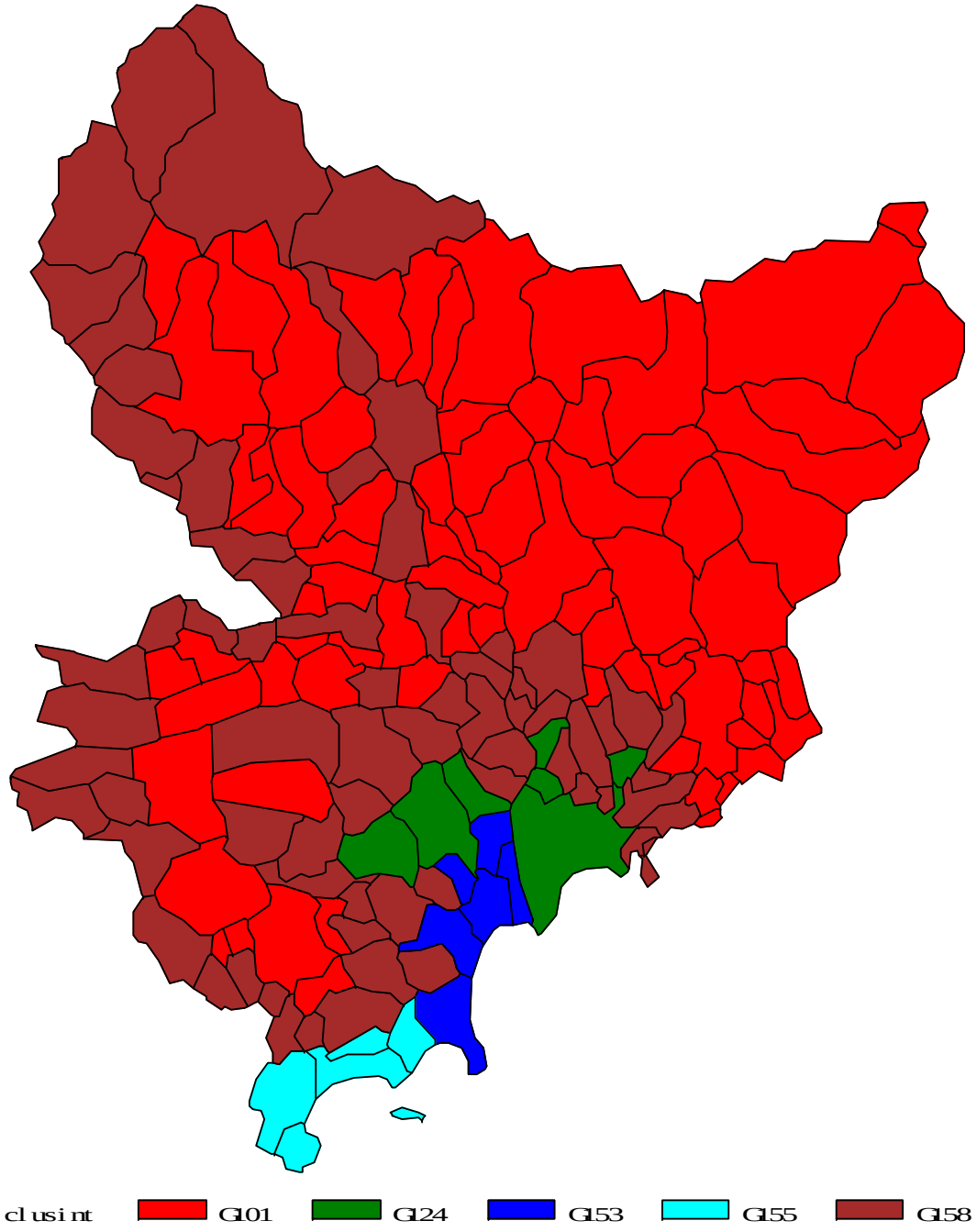
Phase 1 – Minimisation



Résultats de la première phase

classes	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

Phase 2 – Minimisation



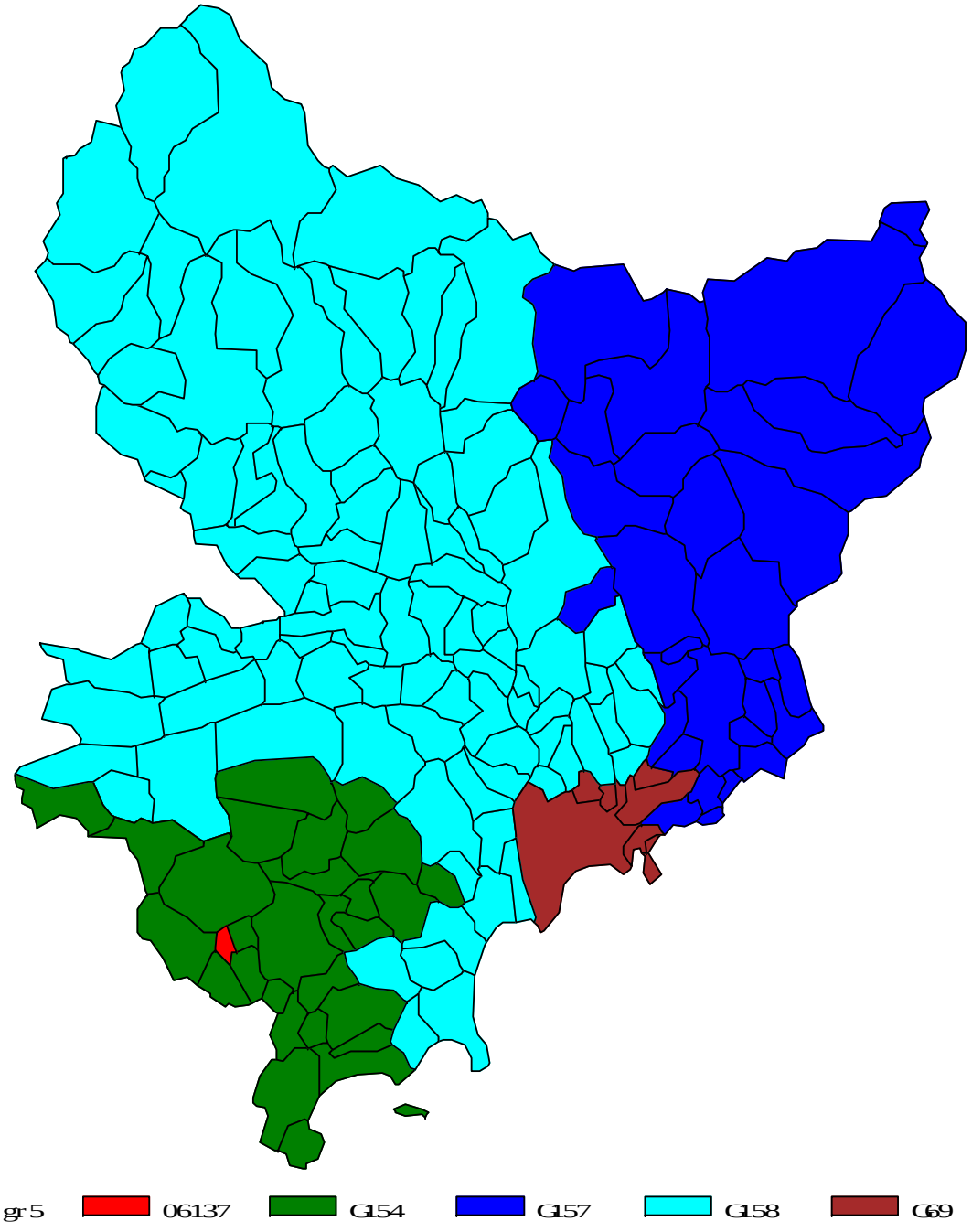
Résultats de la seconde phase

classes	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	1,65
		43,1	**3,830**

Un second exemple sur les Alpes-Maritimes

- Variable : Age moyen par commune (RP 2008)
  - **Maximisation** de la variance intra-classe : reclassement en 5 classes connexes les plus proches possibles de la population totale,
  - Contraintes de taille : chaque classe doit comprendre entre 15 et 35% de la population.
  - Contiguïté : ***géographique***.
- Inertie totale = 6,088

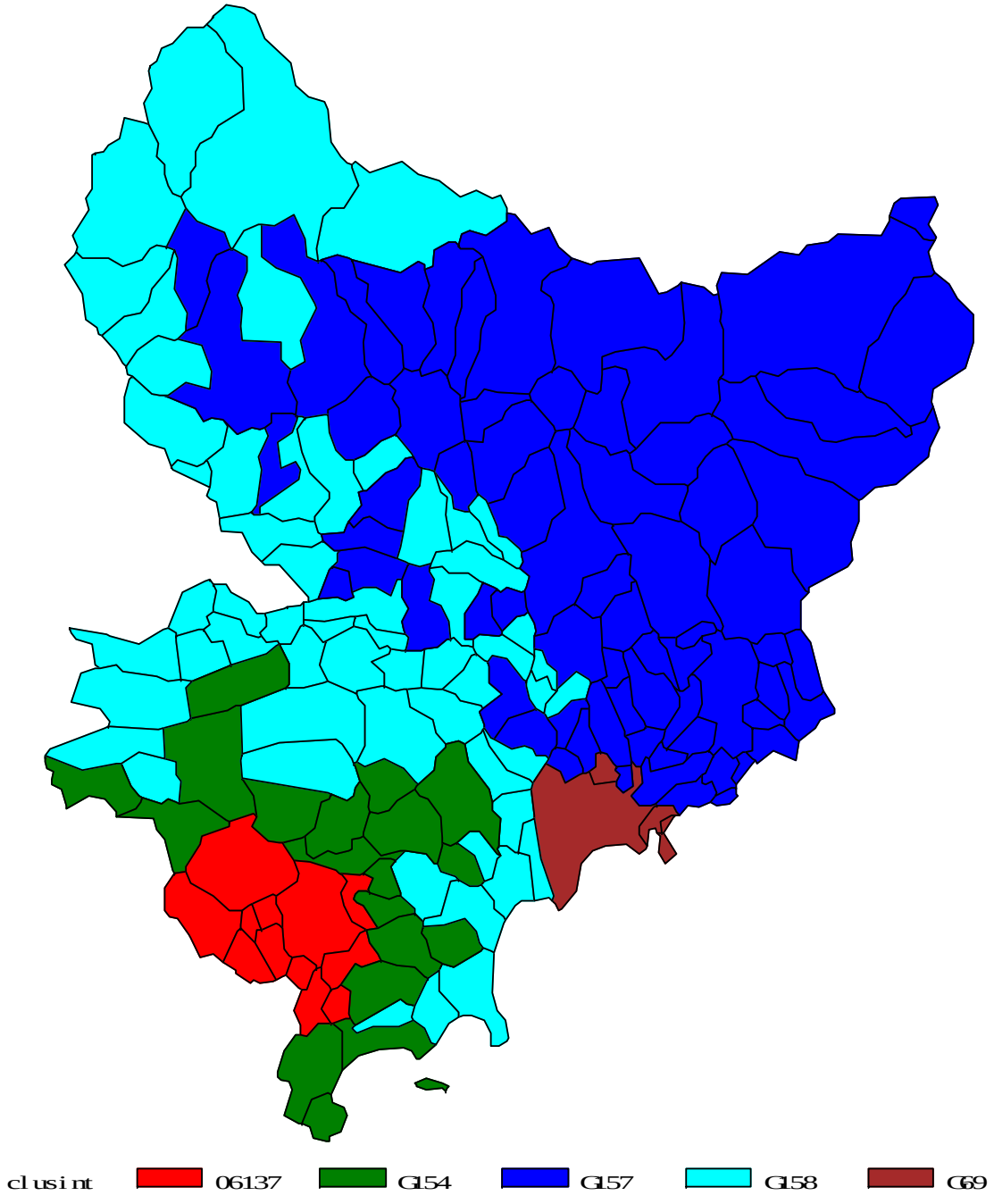
Phase 1 – Maximisation



Résultats de la première phase

classes	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**

Phase 2 – Maximisation



Résultats de la seconde phase

classes	Taille	Age moyen	Variance intra
06137	9,2%	40,4	0,370
G154	16,5%	43,8	2,216
G157	15,0%	42,7	1,376
G158	26,2%	43,8	0,897
G69	33,1%	43,2	0,309
		43,1	**5,168**

Extensions possibles

- Classes *les plus hétérogènes possibles*
  - Inertie intra-classe à ***maximiser*** et non plus à minimiser.
- Extensions à plusieurs variables.
- Extensions à des distances non euclidiennes.

6. Une extension au cas non euclidien

- On a vu que l'inertie intra-classe peut s'écrire :

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

- Ce qui peut se réécrire :

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

Une nouvelle interprétation

- La formule (2) permet d'étendre la méthode à des distances non euclidiennes entre unités.
- Mais elle ne permet plus d'utiliser la notion habituelle de centres de gravité de chaque classe.

Exemple : Navettes domicile-travail en PACA

- On cherche à créer des classes connexes de communes qui maximisent en un certain sens la proportion de personnes résidant et travaillant dans la même zone.

- La contiguïté est ***géographique***.

- La part des actifs résidant dans une zone Z et y travaillant s'écrit :

$$I(Z) = \frac{\sum_{i \in Z} \sum_{j \in Z} A_{i,j}}{\sum_{i \in Z} A_{i,\bullet}}$$

  - Où A_{ij} représente la population résidant dans la commune *i* et travaillant dans la commune *j*.

- Pour une partition du territoire en K zones, on cherchera à maximiser :

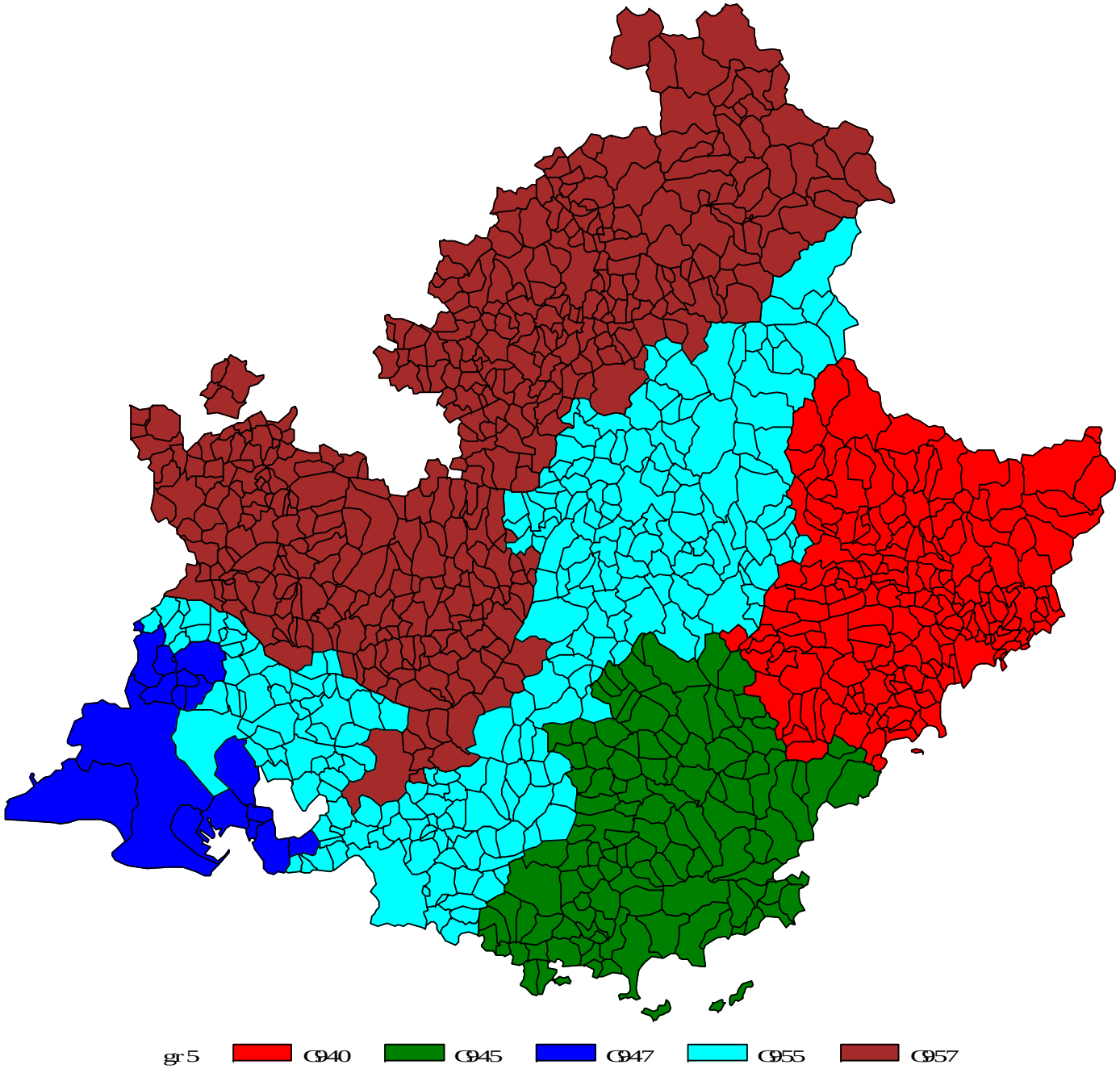
$$J = \sum_{k=1}^{K} \left(\frac{\sum_{i \in Z_k} \sum_{j \in Z_k} A_{i,j}}{\sum_{i \in Z_k} A_{i,\bullet}} \right)$$

- Interprétation : obtenir des « îlots d'emploi stables ».

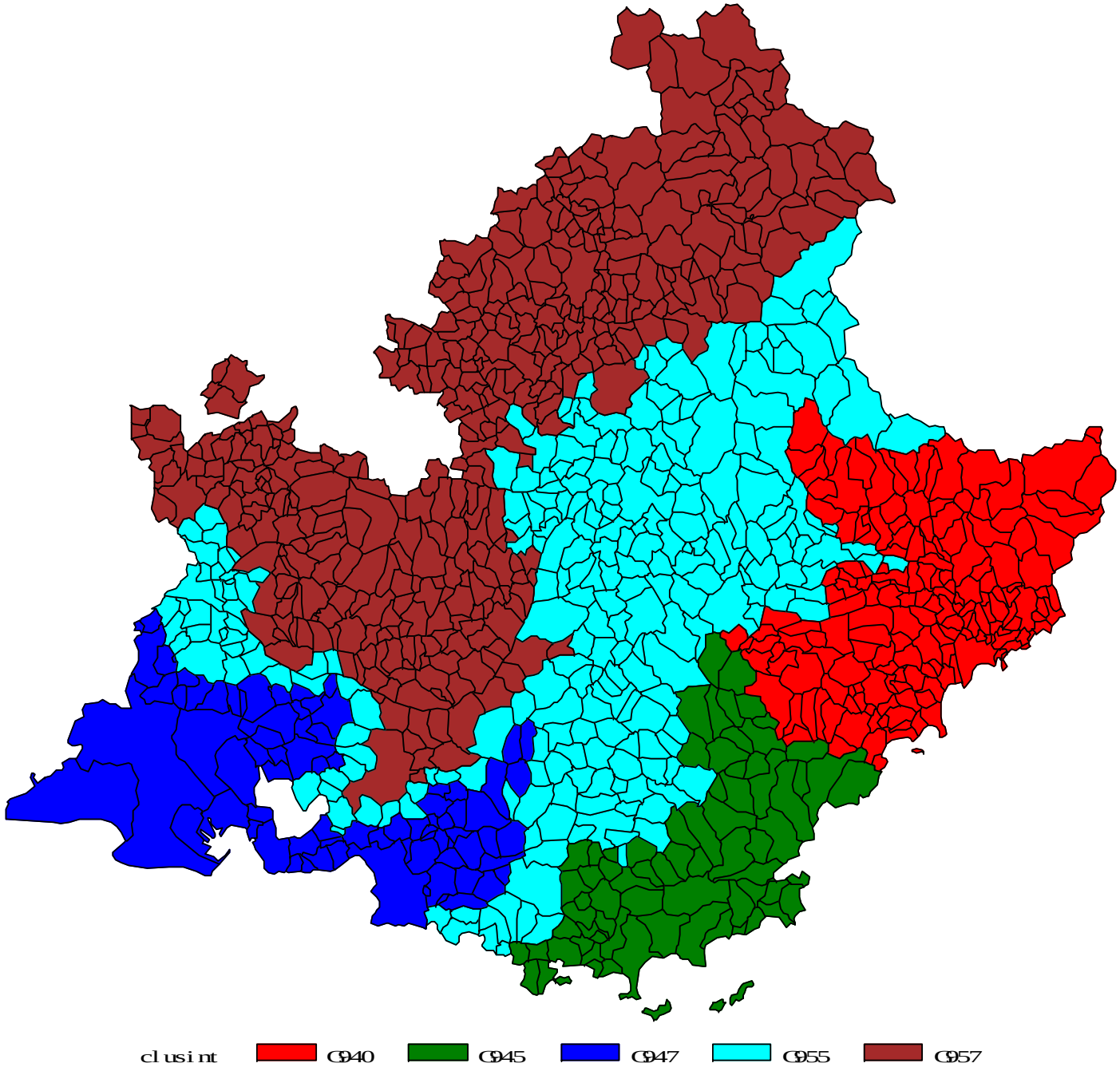
- On interprète ce critère comme une inertie intra-classe en prenant :

$$\begin{cases} \alpha_i = A_{i,\bullet} \\ d_{i,j} = \sqrt{\dfrac{A_{i,j}}{A_{i,\bullet} A_{j,\bullet}}} \end{cases}$$

Navettes résidence – travail avant optimisation



Navettes résidence–travail après optimisation



Les résultats

	Phase 1	Phase 2
%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)

2ème exemple : affectation des enquêteurs à un échantillon

- Problème : affecter un échantillon de *n* Unités statistiques (logements, établissements …)…
- … entre K enquêteurs.
- Enquêteurs et logements sont localisés dans l'espace.
- Distance unité *i* /enquêteur *k* : d_{i,E_k}
- On cherche à minimiser une fonction de coût.

- Idée naturelle :

$$\underset{P_1, P_2, \ldots, P_K}{Min} \sum_{k=1}^{K} \sum_{i \in P_k} d_{i, E_k}$$

- Solution sans contrainte :

$$P_k = \left\{ i \in P; d_{i, E_k} = \underset{E_l}{Min}\ d_{i, E_l} \right\}$$

- Mais il y a des contraintes :
- Chaque enquêteur a une charge *minimale* et *maximale* d'enquêtes à réaliser :

$$C_{Min} \leq C_k \leq C_{Max}$$

=> on peut essayer de résoudre le problème par une méthode de type CAH contiguë.

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

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

  - Ou (avec un autre seuil de distance *t*) :

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

- On cherchera alors à *minimiser* une fonction du type de celle de la formule (**2**) :

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

- ... en prenant pour paramètres :

$$\begin{cases} \alpha_i = T_i = 1 \\ d_{i,j} = \sqrt{\|i-j\|} \\ \omega_k = \sum_{i\in P_k}\alpha_i \end{cases}$$

7. Une nouvelle méthode à explorer

- Idée : utiliser une méthode inspirée des ***nuées dynamiques***.
- Rappel de la méthode :
  - On tire des points de manière aléatoire (nous nous limiterons à un noyau par classe).
  - On affecte chaque Unité statistique à la classe dont elle est la plus proche au sens d'une distance à déterminer.
  - On crée ainsi de nouvelles classes.
  - On calcule les noyaux de ces classes.
  - On itère la procédure en repartant de ces nouveaux noyaux.

- Convergence (dans le cas d'une distance euclidienne):
  - On s'arrête quand les classes restent stables.
  - On obtient un optimum local de l'inertie (ici un minimum).

- Il faut adapter la procédure :
  - Intégrer les contraintes de taille et de contiguïté.
  - Définir le noyau de chaque classe :
    => c'est l'unité U_{j_0} qui satisfait la condition :

$$\underset{j}{Min} \sum_{i \neq j} \left(\frac{\alpha_i}{\sum_{k \neq j} \alpha_k} \right) d^2(U_i, U_j)$$

  - Définir une distance entre un point et une classe (voir plus loin).

- Principe général de la méthode :
  - A une étape de la procédure, on va chercher à agréger des unités à des classes déjà constituées.
  - Pour tenir compte de la contiguïté, les unités candidates pour être agrégées à une classe seront astreintes à appartenir à ***la couronne de la classe*** = ensemble des unités contiguës à l'une au moins des unités de la classe mais n'appartenant pas à la classe.
  - On affecte alors une unité à une classe selon un critère de distance : les couples unités-classes sont triés selon un critère de distance et on réalise l'affectation de l'unité à la classe correspondant à la distance minimale.

- On a le choix entre deux distances possibles entre une unité et une classe :
  - Variation d'inertie au moment de l'agrégation (la somme des variations étant égale à l'inertie intra-classe de la classe finale en cas de distance euclidienne) :

$$I^a_{G\cup\{i_0\}} - I^a_G - I^a_{\{i_0\}}$$

  - Distance entre l'unité et le « noyau de la classe » défini ci-dessus
    - C'est la distance au sens de la distance initiale entre deux unités.

- On recalcule, pour la classe qui a été impactée par l'agrégation :
  - la nouvelle couronne
  - les nouvelles distances des unités de la couronne à la nouvelle classe
  - … ce qui nécessite le calcul du nouveau noyau (***pour la seconde définition de la distance*** classe-unité).
- On supprime l'unité qui vient d'être affectée de la liste des candidats à l'agrégation à d'autres classes.
- On itère.
- La procédure s'arrête quand on a affecté toutes les unités.

- Une fois cette phase d'affectation terminée, on peut l'itérer plusieurs fois en réinitialisant à partir de nouveaux noyaux.
- Selon la distance classe-unité et le type de la distance entre unités, il est possible que l'itération ne minimise pas l'inertie intra-classe.
- Une procédure d'échanges comme dans le cas de la CAH permet d'améliorer la solution obtenue.

- L'initialisation se fait par un tirage ***aléatoire*** d'unités…
- … en nombre égal à celui des classes à constituer.

- Pour tenir compte des <u>contraintes de taille</u> :
  - Lorsqu'à une étape de la procédure, les classes constituées ont une taille inférieure à T_{Min} , on privilégie l'agrégation d'unités permettant d'atteindre ou dépasser ce seuil, indépendamment des critères de distance.
  - Les classes posant problème sont traitées par ordre croissant de leur taille, jusqu'à ce que toutes atteignent le seuil T_{Min}.

  - Il est plus difficile de prendre en compte le critère de taille maximum.

- Avantages de la méthode :
  - On peut raisonner avec n'importe quelle distance.
  - On peut relancer plusieurs fois la procédure avec une initialisation aléatoire : on peut comparer deux solutions…
  - … alors que la CAH est *déterministe* et qu'on ne sait pas si son unique résultat est loin ou pas d'un optimum global

- Une question : peut-il être opportun d'initialiser à partir d'une configuration déterministe ***bien choisie*** ?

8. Conclusion

- L'algorithme de CAH contiguë a été écrit en SAS et une macro SAS est disponible sur demande (quel que soit le type de distance).

- Il y actuellement des limitations sur le nombre d'unités statistiques à traiter (charges en temps de calcul et mémoire)

  => des améliorations du programme sont à l'étude, ainsi qu'une réécriture en R.

- D'autres applications doivent être testées et leurs résultats analysés et comparés :
  - Application à des données de flux (déplacements domicile –travail, comparaison lieu de naissance / lieu de résidence..) : ***comparaison avec les zones d'emploi.***
  - Comparaison de la méthode par CAH contiguë et de celle par nuées dynamiques.
  - Affectation d'un échantillon entre différents enquêteurs minimisant un critère de temps ou de distance de déplacement : ***simulation sur données réelles.***

Merci de votre attention

Marc.christine@insee.fr
Michel.isnard@insee.fr