

Analyse des données avancée

Chapitre 14 : Les k-means et la consolidation



Jaouad Madkour

Faculté d'Économie et de Gestion de Tanger · 27 août 2026

L'algorithme des centres mobiles

Choisir le nombre de classes

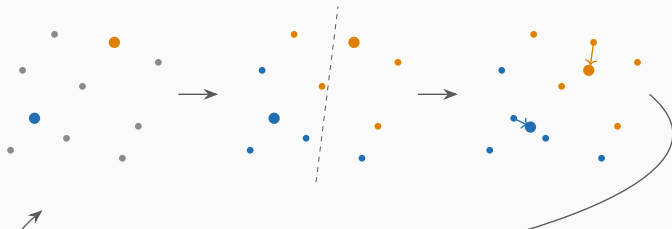
La consolidation

Ce qu'il faut retenir

L'algorithme des centres mobiles

Les trois étapes

1. Choisir k centres au hasard
2. Affecter chaque point au plus proche
3. Recalculer les centres



Trois étapes, répétées jusqu'à ce que plus rien ne bouge. Sur nos 300 clients, six à vingt itérations suffisent.

Chaque tour *diminue* l'inertie intraclasse : l'algorithme converge donc toujours. **Mais il converge vers un optimum local**, qui dépend des centres de départ — d'où la règle : relancer plusieurs fois, ou partir des centres d'une CAH.

Le critère

$$\min I_{\text{intraclasse}} = \frac{1}{n} \sum_{k=1}^K \sum_{i \in C_k} d^2(i, G_k)$$

Chaque itération le fait décroître : l'algorithme **converge toujours**.

Mais il converge vers un optimum local

Le nombre de partitions possibles de 300 clients en 4 classes dépasse 10^{178} . **Aucun algorithme ne les explore toutes.**

Les k -means descendent depuis leur point de départ jusqu'au premier creux rencontré. Un autre départ peut mener à un autre creux — et il n'existe aucun moyen de savoir si l'on tient le meilleur.

Quatre départs aléatoires, $k = 4$

Départ	Inertie interclasse	Itérations
1	76,51 %	22
7	76,54 %	8
42	76,54 %	16
2026	76,54 %	6

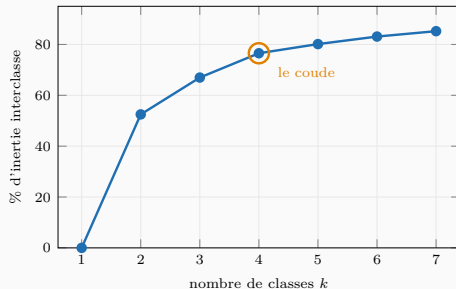
Trois fois le même optimum : bon signe

Quand des départs très différents aboutissent à la même partition, c'est que la structure est **robuste**.

Si les quatre lignes avaient donné 60, 71, 65 et 74 %, il aurait fallu conclure que le nuage n'a pas de groupes nets et **le dire dans le rapport**.

Choisir le nombre de classes

La courbe de l'inertie



Le même coude que l'éboulis, et la même logique.

De 3 à 4 classes, on gagne 9,6 points. De 4 à 5, seulement 3,6.

Au-delà, chaque classe supplémentaire ne fait plus que découper du bruit. La courbe monte encore, mais elle monterait jusqu'à 100 % pour $k = n$ — une classe par client.

Aucun critère ne prouve que 4 est le bon nombre. La CAH le suggérerait, le coude le confirme, et l'interprétation tranche.

Ce que chacune sait faire

	CAH	k -means
Nombre de classes	trouvé	à fournir
Réaffectation	impossible	à chaque tour
Calcul	$O(n^3)$	rapide
Résultat	unique	dépend du départ

Les défauts de l'une sont les qualités de l'autre

La CAH trouve le nombre de classes mais ne se corrige jamais. Les k -means se corrigent sans cesse mais ne trouvent pas le nombre.

La solution évidente : les enchaîner.

La consolidation

En trois temps

1. **ACP** pour obtenir des coordonnées propres.
2. **CAH de Ward** pour déterminer le nombre de classes et une partition initiale.
3. ***k*-means partant des centres de la CAH** pour corriger les affectations regrettables.

Le résultat sur nos clients

Trois itérations. **Vingt-six clients sur trois cents changent de classe.**

L'inertie interclasse passe de **74,93 % à 76,50 %** — un gain de **1,57 point**.

Le tableau de croisement

CAH \ consolidée	C1	C2	C3	C4	total
Classe 1	61	0	0	0	61
Classe 2	17	61	3	1	82
Classe 3	0	0	59	0	59
Classe 4	0	0	5	93	98
total	78	61	67	94	300

La consolidation : le meilleur des deux méthodes.

On part des **centres de la CAH** — qui ont fourni le nombre de classes — et on laisse tourner les k -means.

Trois itérations. 26 clients sur 300 changent de classe.

L'inertie interclasse passe de **74,93 %** à **76,50 %**.

Peu de mouvement, et c'est la bonne nouvelle : la partition ne dépendait pas de l'algorithme. Si la moitié des clients avaient changé de classe, il aurait fallu conclure qu'il n'y a pas de segments dans ces données.

Peu de mouvement est une bonne nouvelle

Vingt-six clients déplacés sur trois cents, c'est 8,7 %. **La partition ne dépendait donc presque pas de l'algorithme.**

Si la moitié des clients avaient changé de classe, la conclusion aurait été tout autre : il aurait fallu écrire que les données ne contiennent pas de segments stables, et renoncer à la segmentation.

La consolidation est un test, autant qu'une amélioration

Elle améliore la partition d'un point et demi — c'est peu.

Mais elle **vérifie** que deux logiques différentes — hiérarchique et itérative — retrouvent les mêmes groupes. **C'est le seul contrôle disponible en l'absence de test statistique.**

Ce qu'il faut retenir

Six points

1. Les k -means alternent **affectation** et **recalcul des centres**, jusqu'à stabilité.
2. Ils minimisent l'inertie **intraclasse** et convergent toujours — vers un **optimum local**.
3. Il faut donc **relancer plusieurs fois** : ici quatre départs donnent **76,5 %**.
4. Le **coude** de la courbe d'inertie confirme $k = 4$.
5. **CAH puis k -means** : l'une donne le nombre, l'autre corrige les affectations.
6. Chez nous, la consolidation déplace **26 clients** et gagne **1,57 point** — la partition est stable.

La suite

Quatre classes, et quatre nombres pour les décrire. Ce n'est pas encore une segmentation.

Une segmentation, c'est quatre portraits qu'un directeur commercial peut lire — et le chapitre 15 les construit.