Analyse de regroupements

Analyse multidimensionnelle appliquée

Léo Belzile

HEC Montréal

Analyse de regroupements

Objectif: regrouper des observations de telle sorte que

- les observations d'un même groupe soient le plus semblables possible,
- les groupes soient le plus différent possible les uns des autres.

Chaque observation se voit assigner une étiquette de groupe.

On procède ensuite à une analyse descriptive, segment par segment.

Exemples

- Programmes de fidélisation et résolution d'entités
- Segmentation de la clientèle de transport en commun et élaboration de forfaits
- Démarchage d'organismes de charité
- Segmentation de quartiers de Los Angeles et de New York selon leur vote
- Profils des électeurs albertains
- Positionnement de joueurs lors de match de la NBA

Illustration

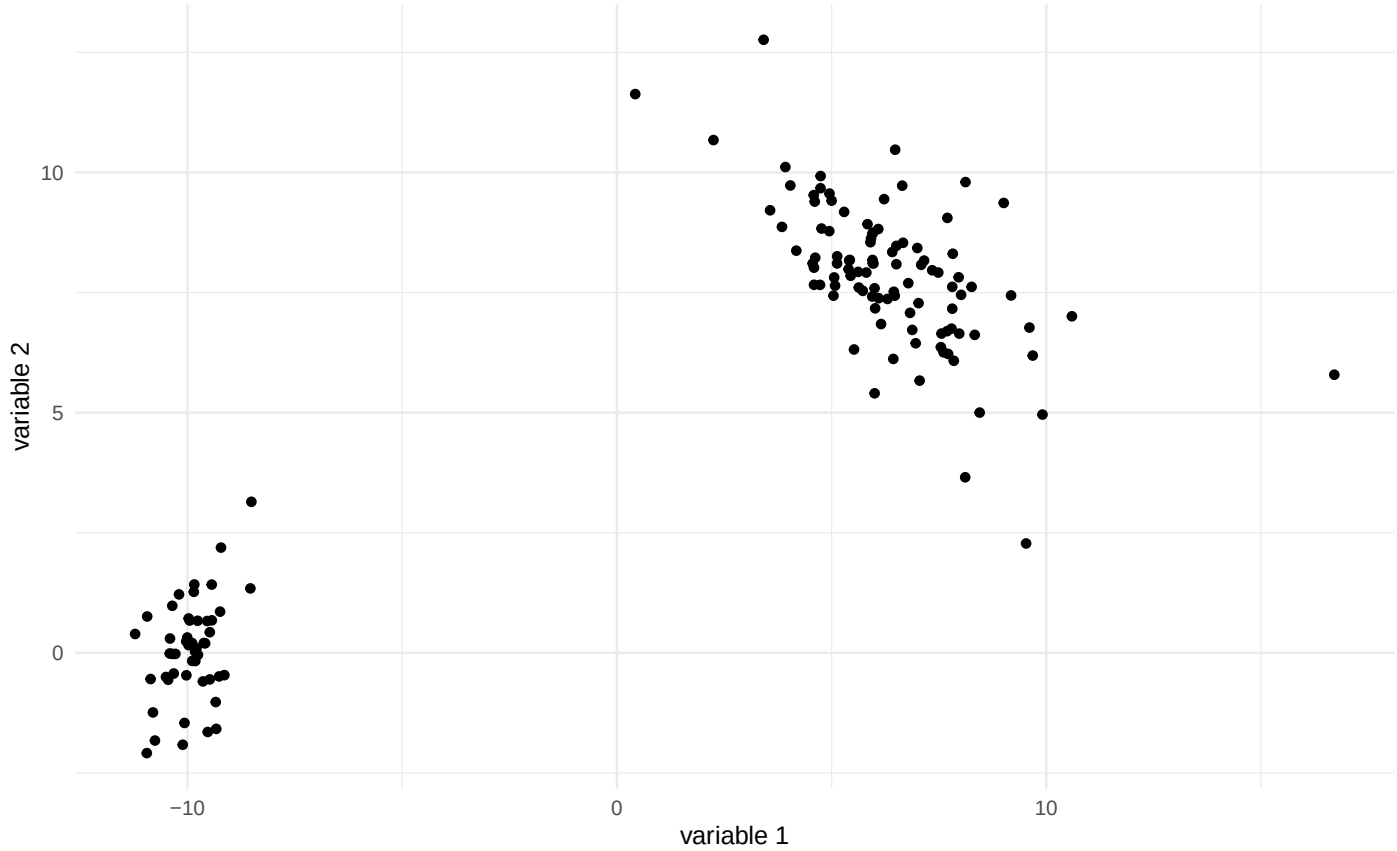
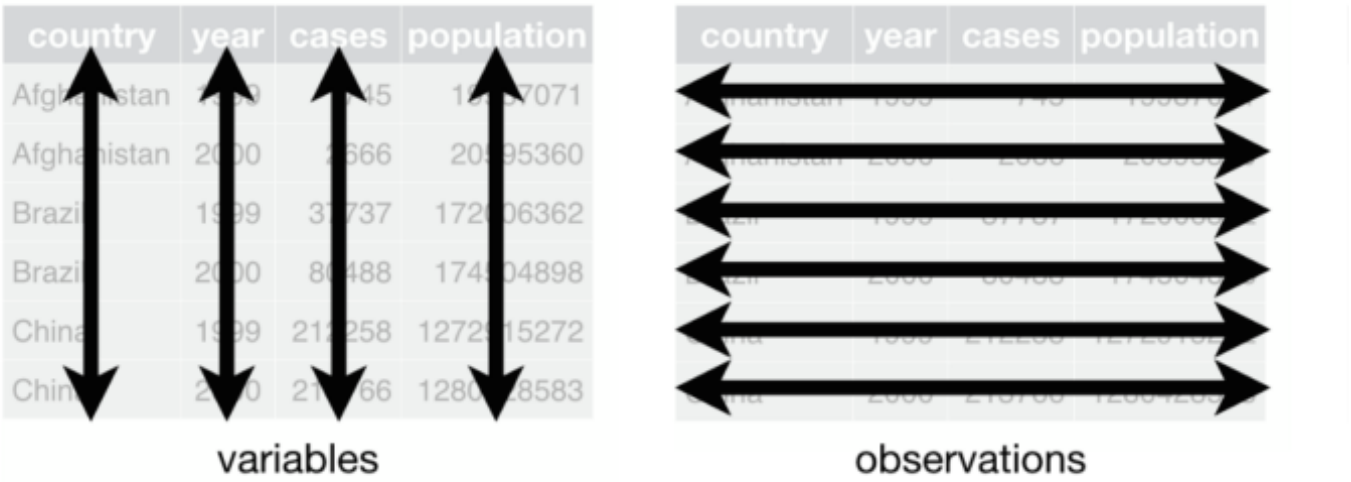


Figure 1: Données simulées avec deux regroupements hypothétiques.

Analogie avec analyse factorielle

En analyse factorielle, on combine des variables similaires.

Pour l'analyse de regroupements, on regroupe des observations.



Ce sont des méthodes dites d'apprentissage non-supervisé: l'objectif est de déduire la structure présente dans un ensemble de points X sans étiquette préalable (contrairement à la classification).

Structure de la base de données

Quelles variables $X_1, \ldots, X_p$ sont d'intérêt?

- Choisir des variables pertinentes pour faire ressortir les différences
- Créer de nouvelles variables explicatives

Pour les données longitudinales, on va typiquement aggréger les bases de données marketing par identifiant client.

Exemple avec transport en commun

La carte Opus enregistre

- les temps de passage
- le type de déplacement (REM, métro, bus)
- le nombre de passages
- les abonnements
- le profil client (études, rabais pour personnes âgées)

Quelles variables créer ou conserver?

- Nombre de passages mensuels
- Abonnement mensuel ou annuel (oui/non)
- Type de déplacement (soir, jour)
- Nombre d'allers-retours hebdomadaires en heure de pointe
- Variabilité de la fréquentation

À votre tour

Vous avez toutes les données transactionnelles associées à des comptes d'épicerie avec un compte de fidélisation.

Quelles variables pourriez-vous créer à partir des données aggrégées pour créer des segments?

Diviser pour régner

Souvent, il existe une division naturelle des données.

Les jeunes avec des abonnements de transport publics l'utilisent principalement pour aller à l'école.

On peut faire la segmentation séparément pour ces sous-groupes.

Choix des variables

Recommandations: choisir les variables pertinentes qui font ressortir les effets voulus.

- inclure de nombreuses variables similaires dilue les différences.
- transformez les variables pour diminuer la corrélation.

Typiquement, ne pas utiliser les variables sociodémographiques (âge, revenu, sexe, etc.)

- on compare plutôt leur répartition au sein des regroupements.

Exemple - dons à un organisme de charité

La base de données `dons` contient 19353 observations pour 16 variables.

- Trois grandes catégories: personnes qui n'ont pas donné, dons uniques, dons multiples.
- valeur des dons (total, min, max) et des promesses, nombre de dons, fréquence, ancienneté, etc.

Une rapide exploration des données révèle que près de 61% des employé(e)s n'ont pas donné à l'organisme.

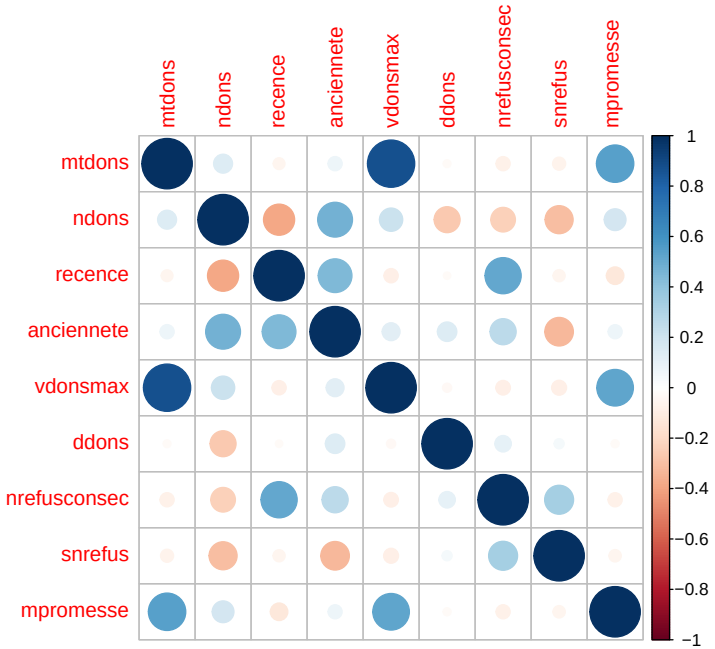
Une poignée de dons sont très élevés, mais la plupart des montants tourne autour de 5$, 10$, 20$, etc.

Modification des données

On se concentre sur les personnes qui ont donné.

```
donsmult <- hecmulti::dons |>
  filter(ndons > 1L) |>
  mutate(mtdons = vdons/ndons,
         snrefus = nrefus/anciennete*mean(anciennete),
         mpromesse = case_when(
           npromesse > 0 ~ vpromesse/npromesse,
           TRUE ~ 0)) |>
  select(!c(
    vradiations, # valeurs manquantes
    nindecis, vdons, ddonsmax,
    ddonsmin, vdonsmin, npromesse,
    vpromesse, nrefus, nradiations)) |>
  relocate(mtdons)
```

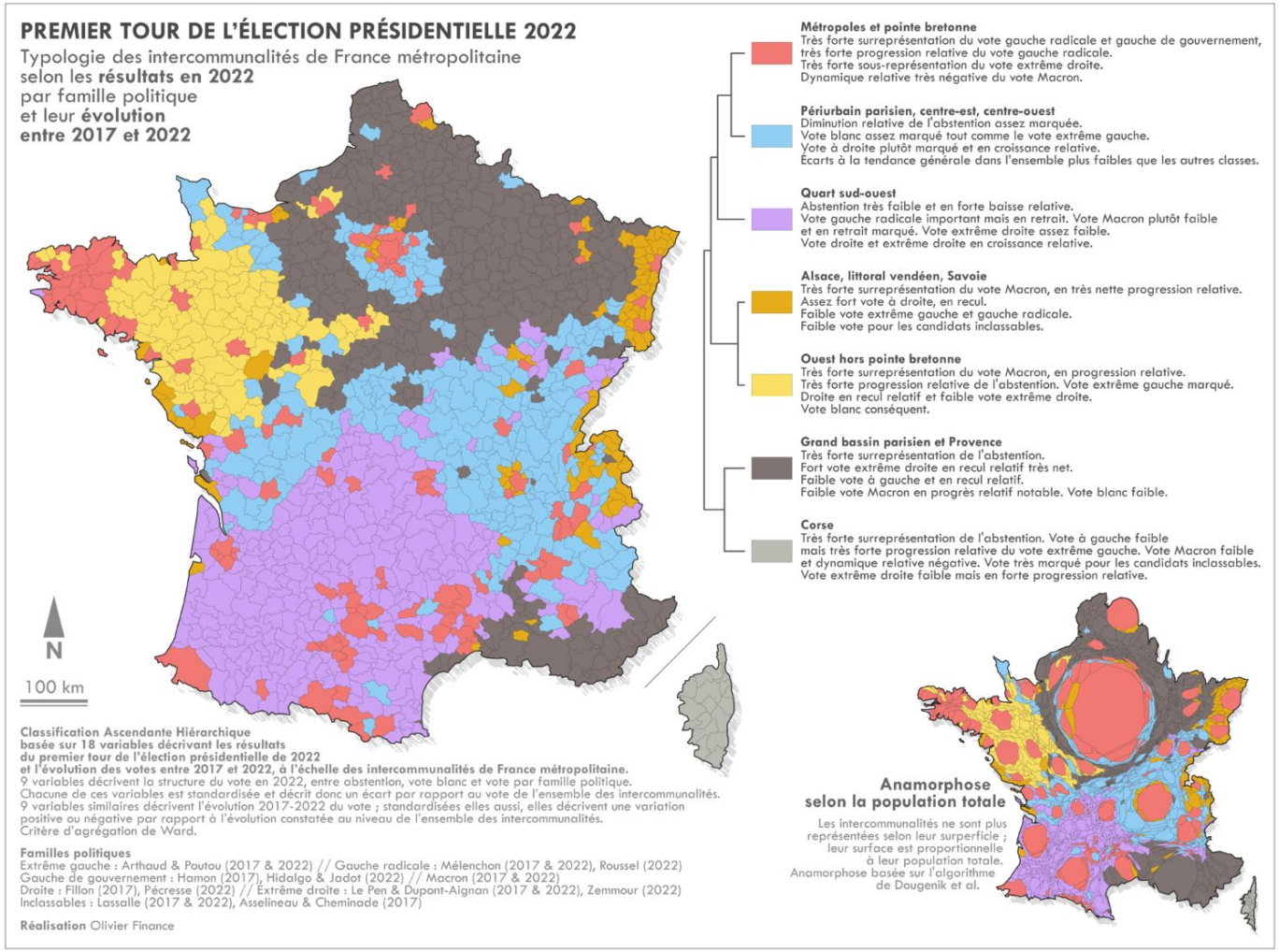
Corrélation pour nouvelles données



Étapes d'une analyse de regroupements

1. Choisir les variables pertinentes à l'analyse. Cette étape peut nécessiter de créer, transformer de nouvelles variables ou d'aggréger les données.
2. Décider quel méthode sera utilisée pour la segmentation.
3. Choisir les hyperparamètres de l'algorithme (nombre de regroupements, rayon, etc.) et la mesure de dissimilarité.
4. Étiqueter les observations.
5. Obtenir le prototype de chaque groupe.
6. Valider la qualité de la segmentation.
7. Interpréter les regroupements obtenus à partir des prototypes.

Exemple: typologie des votants en France



Mesures de dissimilarité et de similarité

Comment mesurer si deux observations appartiennent à un même regroupement et sont similaires?

Une mesure de dissimilarité sert à quantifier la proximité de deux objets à partir de leurs coordoonnées.

Plus la dissimilarité est élevée, moins les observations sont semblables (plus éloignées).

Mesures de dissimilarité

Quelques propriétés des mesures de dissimilarité:

1. positivité: la distance entre deux observations est nulle si et seulement si on a les mêmes caractéristiques pour toutes les variables explicatives et strictement positive sinon.
2. la dissimilarité est la même peu importe l'ordre des observations (symmétrie)

Toute distance est une mesure de dissimilarité.

Distance euclidienne

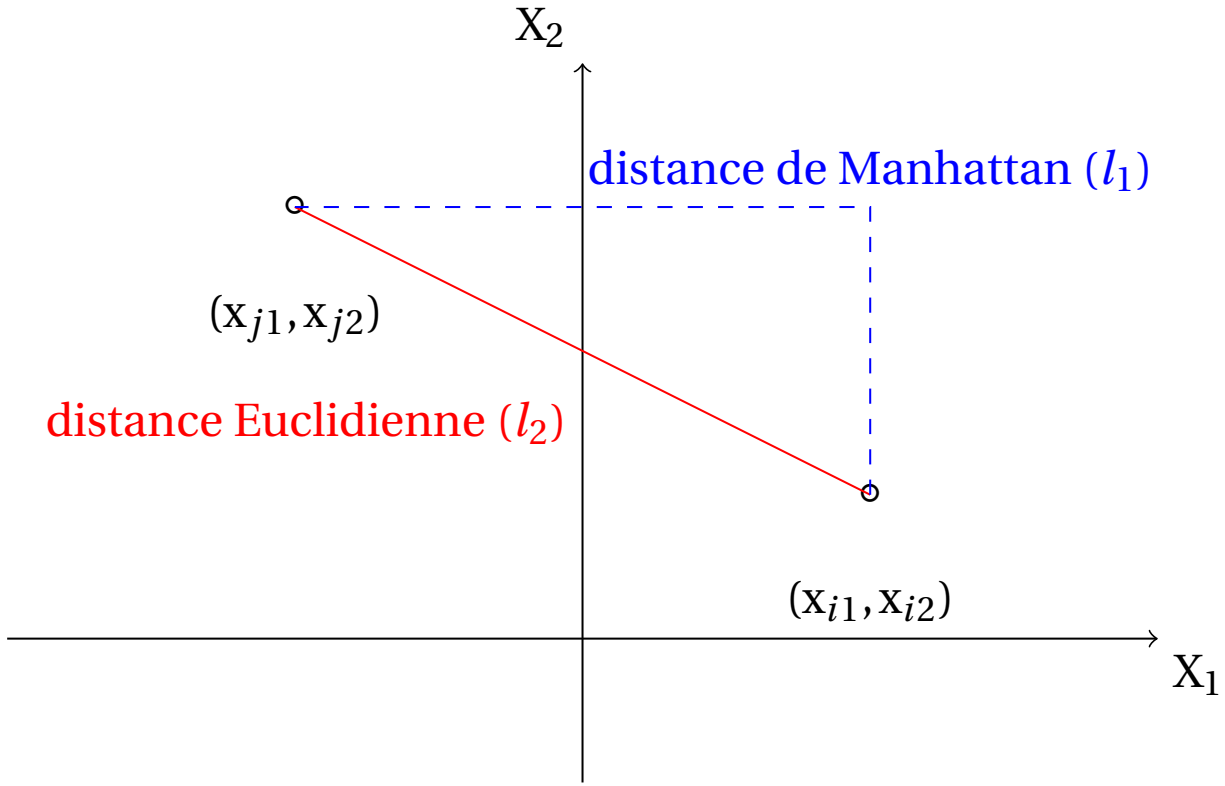
La mesure de dissimilarité la plus utilisée en pratique est la distance euclidienne.

La distance entre les vecteurs ligne X_i et X_j (deux lignes de la base de données) est

$$d(\mathsf{X}_i, \mathsf{X}_j; l_2) = \left\{(X_{i1} - X_{j1})^2 + \cdots + (X_{ip} - X_{jp})^2\right\}^{1/2}.$$

C'est tout simplement la longueur du segment qui relie deux points dans l'espace p dimensionnel.

Autres mesures de dissimilarité



- La distance de Manhattan est la somme des valeurs absolues entre chaque composante, $|X_{i1} - X_{j1}| + \cdots + |X_{ip} - X_{jp}|$.
- La distance l_∞, soit le maximum des différences entre les coordonnées des vecteurs d'observations i et j, $\max_{k=1}^{p} |X_{ik} - X_{jk}|$

Autres mesures de dissimilarité

Pour les données catégorielles nominales, on peut assigner une dissimilarité de 0 si les variables ont la même modalité et 1 sinon.

Pour le cas de variables mixtes, la distance de Gower permet de traiter les valeurs manquantes et standardise automatiquement.

dissimilarités dans R

Avec notre base de données `donsmult`, le stockage des distances prend environ 750MB!

```
# Distance euclidienne, de Manhattan, de Gower
d1 <- dist(donsmult, method = "euclidean")
d2 <- dist(donsmult, method = "minkowski", p = 1)
d3 <- cluster::daisy(donsmult, metric = "gower")
# Voir aussi ?flexclust::dist2
```

Les objets de class `dist` ne stockent que la matrice triangulaire inférieure (puisque la distance est symmétrique).

Au plus fort la poche

Le poids accordé à une variable explicative dépend de son étendu et de sa variabilité.

- Plus la variable est grande, plus elle aura un impact dans le calcul des distances
- Problème de standardisation (résultats différents selon les unités de mesure)

Standardisation

Généralement, on standardise les données avant l'analyse de regroupements.

- Soustraire la moyenne et diviser par l'écart-type empiriques (fonction `scale` dans R)
- ou utiliser des mesures robustes: soustraire la médiane et diviser par l'écart absolu à la médiane (mad)

Notez qu'il est illogique de standardiser les variables catégorielles (déclarer obligatoirement les variables binaires en facteurs et traiter à part!)

Standardisation avec R

```r
# Standardisation usuelle
# (soustraire la moyenne, diviser par écart-type)
donsmult_std <- scale(donsmult)
# Extraire moyenne et écart-type
dm_moy <- attr(donsmult_std, "scaled:center")
dm_std <- attr(donsmult_std, "scaled:scale")
# Standardisation robuste
donsmult_std_rob <- apply(
  donsmult,
  MARGIN = 2,
  FUN = function(x){(x - median(x))/mad(x)})
```

Être aux abonnés absents

Attention aux valeurs manquantes, rarement supportées par les algorithmes d'analyse de regroupements.

Quelques solutions

- ignorer les variables explicatives avec beaucoup de valeurs manquantes.
- faire une segmentation manuelle si les valeurs manquantes déterminent des regroupements (ex: temps entre dons valide uniquement pour dons multiples).
- imputer les données manquantes (voir chapitre sur les données manquantes)

Algorithmes pour la segmentation

L'analyse de regroupements cherche à créer une division de n observations de p variables en regroupements.

1. méthodes basées sur les centroïdes et les médoïdes (k-moyennes, k-médoides)
2. mélanges de modèles
3. méthodes basées sur la connectivité (regroupements hiérarchiques agglomératifs et divisifs)
4. méthodes basées sur la densité (pas couvertes)

Notation pour la complexité

- La notation $\mathrm{O}(\cdot)$ nous renseigne sur l'ordre du nombre d'opérations nécessaires pour le calcul. Par exemple, additionner une colonne de chiffres nécessite $\mathrm{O}(n)$ flops
- De même, le stockage d'une matrice de distance, qui contient $n(n-1)/2$ entrées distinctes, est $\mathrm{O}(n^2)$.

Plus le chiffre (ou la puissance) est élevé, plus le calcul ou le stockage est coûteux.

Critères pour sélection

- Complexité: plus un algorithme a une complexité élevée (coût de calcul et quantité de stockage), moins il sera susceptible d'être applicable à des mégadonnées.
- Choix des hyperparamètres: plusieurs paramètres (nombre de groupes, rayon, choix de la dissimilarité, etc.) à spécifier selon les méthodes.

Méthodes basées sur les centroïdes et les médoïdes

On assigne chaque observation à un de K regroupements, représentés par un prototype, disons μ_k pour le regroupement k.

- le nombre K est fixé apriori

On cherche à assigner les observations aux groupes de manière à minimiser la distance entre les observations et les prototypes.

K moyennes

Probablement la méthode de regroupement la plus populaire en raison de son faible coût (linéaire en n et p).

La fonction objective considère la distance totale entre les observations et les prototypes.

$$\min_{\mu_1,\ldots,\mu_K} \sum_{i=1}^{n} \min_{c_i \in \{1,\ldots,K\}} d(\mathsf{X}_i, \mu_{c_i}) \tag{1}$$

distance entre obs. i et son prototype μ_j

L'allocation optimale de n observations à K groupes est un problème NP complet: on cherchera plutôt une solution approximative au problème d'optimisation.

Initialisation des K-moyennes

Initialisation: on sélectionne préalablement

- un nombre K de regroupements et
- les coordonnées de départ pour les prototypes.

Algorithme EM

L'algorithme de type EM itère entre deux étapes:

1. Assignation (étape E): calculer la distance entre chaque observation et les prototypes; assigner chaque observation au prototype le plus près.
2. Mise à jour (étape M): calculer les coordonnées optimales des prototypes de chaque groupe (avec la distance Euclidienne, c'est le barycentre des observations du groupe).

L'algorithme termine après un nombre prédéfini d'itérations ou lorsque l'assignation ne change plus (solution locale).

Visionner l'animation en ligne ou ce site.

Forces et faiblesses

Quelques forces ($+$) et faiblesses ($-$) de l'algorithme des K moyennes

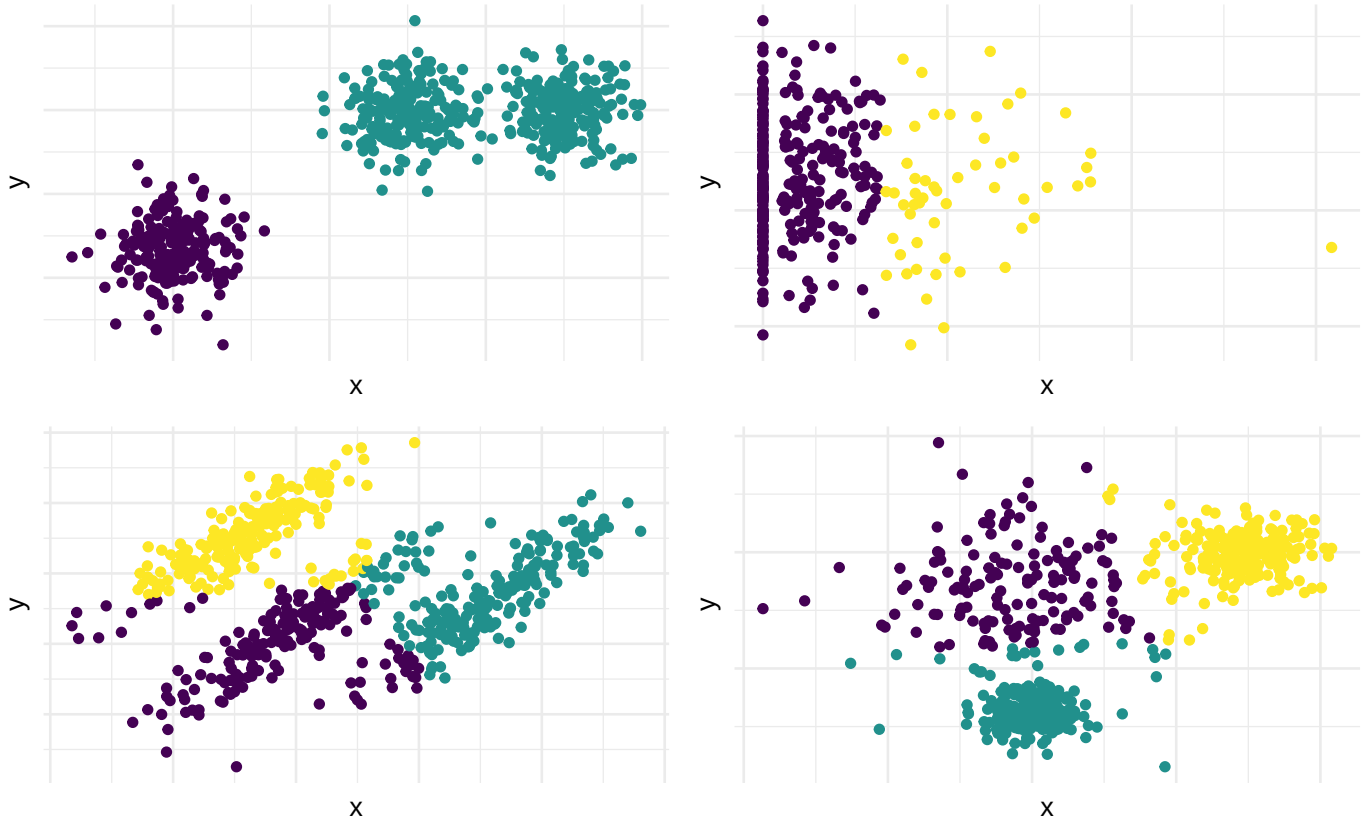
- ($+$) Complexité linéaire dans la dimension et dans le nombre de variables.
- ($+$) L'algorithme converge rapidement vers une solution locale (garantie théorique).
- ($-$) Regroupements globulaires d'apparence sphérique (distance Euclidienne).
- ($+$) Pour les prédictions, on peut assigner les nouvelles observations au barycentre le plus près.

Performance

- (—) Chaque observation est assignée à un seul des K regroupements (partition rigide).
- (—) Valeurs aberrantes pas étiquetées à part (manque de robustesse pour moyenne).
- (—) Sensible aux valeurs initiales des prototypes.
- (—) Les prototypes ne correspondent pas à des observations du groupe.

Performance des K-moyennes

Illustration de segmentations problématiques avec K-moyennes



Séparation linéaire de l'espace

Avec la distance euclidienne, la partition de l'espace est linéaire.

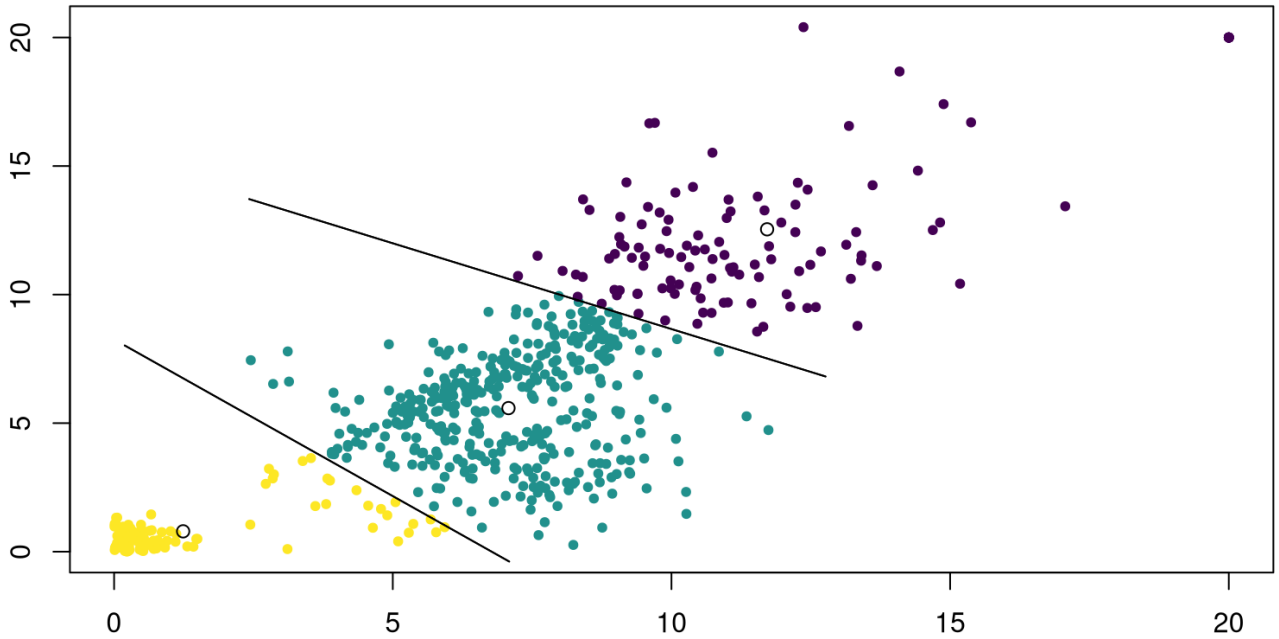


Figure 2: Partitions de Voronoï pour les regroupements avec séparateur linéaire.

Hyperparamètres avec K-moyennes

1. le choix de la mesure de distance
2. les valeurs initiales des prototypes
3. le nombre de groupes K

Distance

Avec la distance euclidienne l_2, les prototypes correspondent avec le barycentre (moyenne variable par variable) des observations du regroupement.

Avec la distance de Manhattan l_1, les prototypes correspondent avec la médiane variable par variable (K-médianes).

Autrement, optimisation nécessaire dans l'étape M de l'algorithme.

Initialisation

- Choisir aléatoirement K observations dans la base de données.
- Répéter plusieurs fois
- Prendre la meilleure segmentation du lot (celle avec la valeur optimale de la fonction objective).

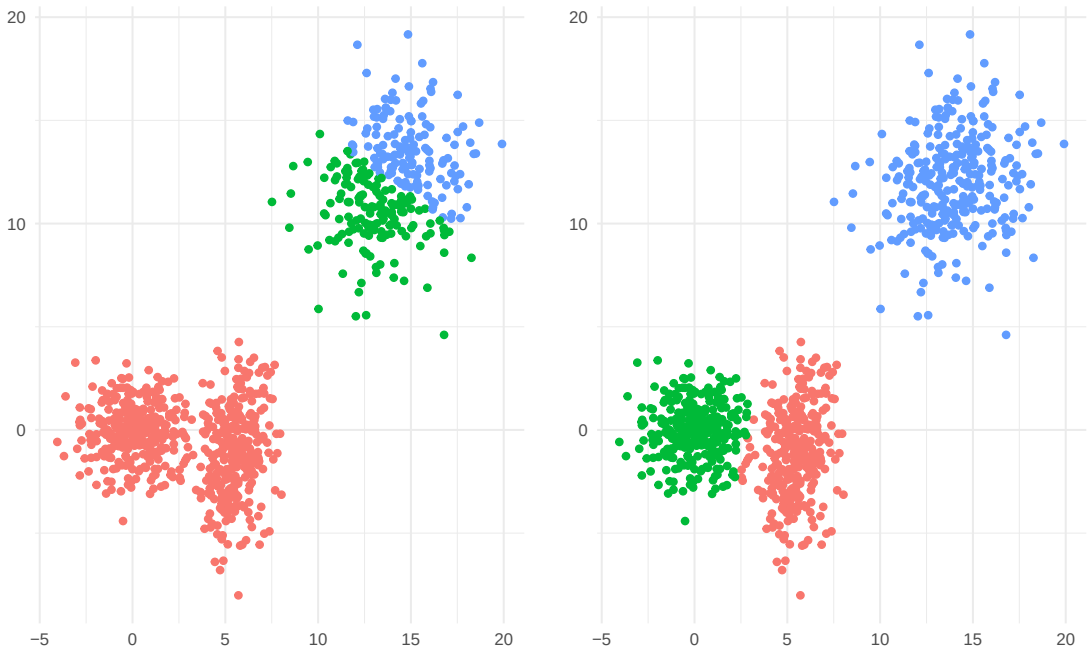


Figure 3: Regroupements pour $K = 3$ groupes avec une mauvaise initialisation principale (gauche) et une bonne initialisation (droite).

K-moyennes dans R

```r
set.seed(60602)
kmoy5 <- kmeans( # distance euclidienne
    x = donsmult_std, # données
    centers = 5L, # nb groupes
    nstart = 10, # nb initialisation aléatoire
    iter.max = 25) # nb étapes maximum dans optimisation

kmoy5$cluster # étiquettes
kmoy5$size # répartition
kmoy5$tot.withinss # fonction objective minimal
kmoy5$centers # barycentres (données standardisées)
```

Algorithme des K-moyennes^{++}

Choisir des observations comme valeurs initiales, mais avec échantillonnage préférentiel (points éloignés les uns des autres).

0. sélectionner une observation au hasard pour μ_1

Pour $k = 2, \dots, K$

1. calcul de la distance carrée minimale entre l'observation X_i et les prototypes précédemment choisis,

$$p_i = \min\{d(\mathsf{X}_i, \mu_1; l_2)^2, \dots, d(\mathsf{X}_i, \mu_{k-1}; l_2)^2)\}$$

2. Choisir le prototype initial μ_k au hasard parmi les observations avec une probabilité de $p_i / \sum_j p_j$ pour l'observation X_i.

K-moyennes^{++} dans R

Utiliser le paquet `flexclust`

```
set.seed(60602)
kmoypp5 <- flexclust::kcca(
  x = donsmult_std,
  k = 5, # nb groupes
  family = flexclust::kccaFamily("kmeans"),
  control = list(initcent = "kmeanspp"))
# Vérifier répartition
kmed5@clusinfo
# Coordonnées des prototypes standardisés
kmed5@centers
# Étiquettes
kmed5@cluster
```

Choix du nombre de regroupements

Plusieurs critères généralement applicables

- silhouettes (`cluster::silhouette`)
- statistique d'écart (`cluster::clusGap`)

Critères plus spécifiques aux K-moyennes rattachés à la fonction objective

- graphique du R^2
- critère d'information bayésien "BIC"

Somme du carré des distances intra-groupes

La fonction objective de l'Equation 1 avec la distance euclidienne représente la somme du carré des distances (SCD)

$$\mathrm{SCD}_K = \mathrm{SCD}_{1,K} + \cdots + \mathrm{SCD}_{K,K};$$

où

$$\mathrm{SCD}_{k,K} = \sum_{i \in G_k} \|\mathrm{X}_i - \mu_k\|_2^2,$$

est la somme des distances euclidiennes au carré entre les observation du groupe G_k et leur barycentre μ_k.

Avec un seul groupe, la distance par rapport à la moyenne est $\mathrm{SCT} = \mathrm{SCD}_1$, la somme totale du carré de toutes les distances par rapport au barycentre global.

Somme du carré des distances

La valeur optimale de la somme du carré des distances mesure va mécaniquement* diminuer quand K augmente $\text{SCD}_1 > \text{SCD}_2 \cdots$.

En pratique, cela peut ne pas être le cas si le minimum local est sous-optimal.

Si la réduction de la somme du carré des distances est négligeable, on pourrait penser que l'ajout d'un groupe supplémentaire.

Critères R^2 et "BIC"

On peut mesurer le pourcentage de variance expliquée,

$$R^2_K = 1 - \frac{\text{SCD}_K}{\text{SCT}}.$$

On cherche un point d'inflexion (coude) à partir duquel l'amélioration est négligeable.

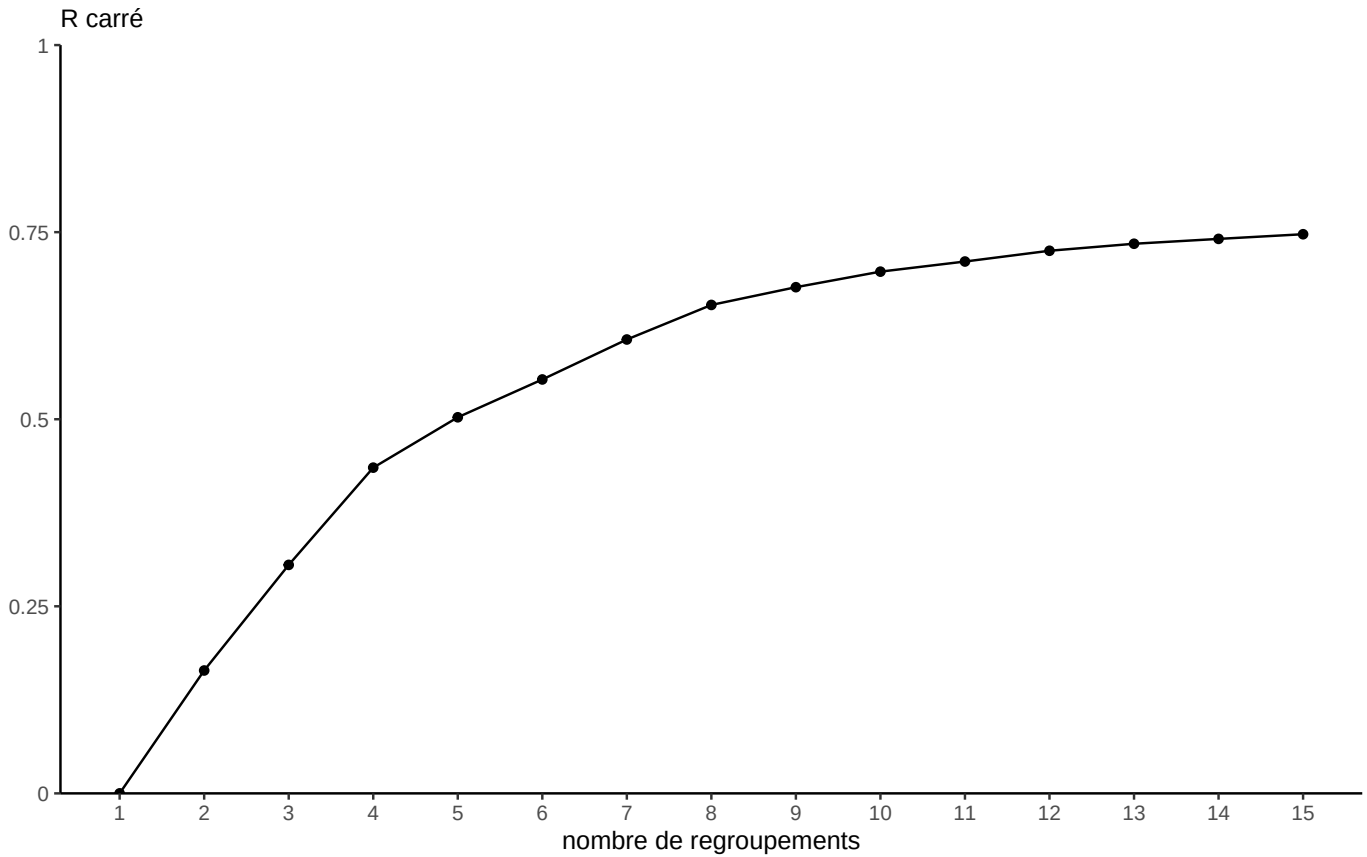
Puisque la somme du carré des distances diminue avec K, on peut considérer l'ajout d'une pénalité pour le nombre de paramètres estimés

$$\text{BIC} = \text{SCD}_K + \ln(n) K p$$

La plus petite valeur du "BIC" est préférable.

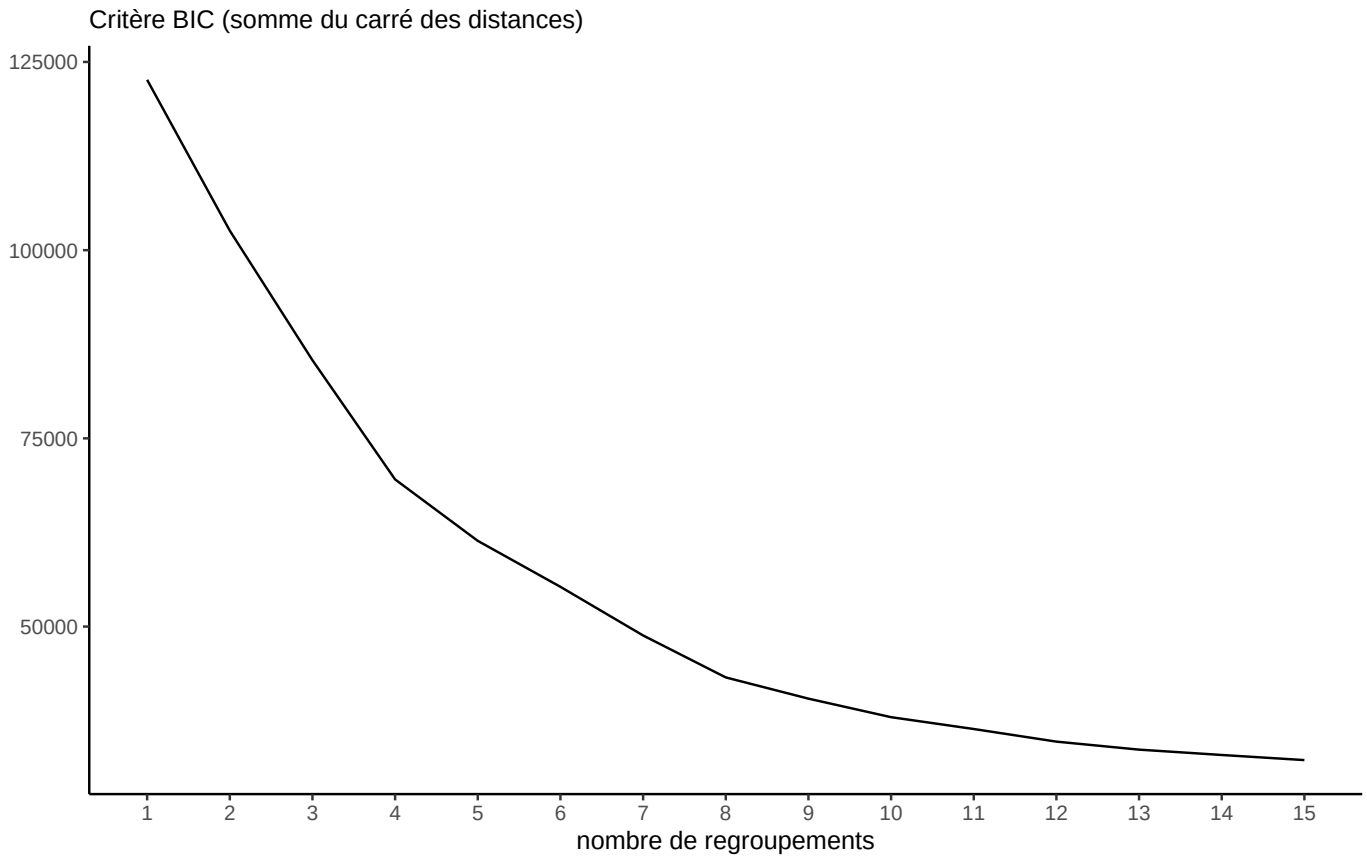
Graphique de coefficient de détermination R^2

Le pourcentage de variance expliqué augmente de manière plus ou moins constante jusqu'à 8 ou 9 composantes.



Critère BIC

Le critère suggère aussi un nombre élevé de regroupements, ici 15 (nombre maximum de 15).



Silhouettes

Pour chaque observation X_i, on calcule

- a_i, la moyenne des dissimilarités entre X_i et les observations de son regroupement
- b_i, le minimum parmi les $K-1$ dissimilarités moyennes entre X_i et les observations de chaque autre regroupement.

On calcule la silhouette

$$s_i = \frac{b_i - a_i}{\max\{a_i, b_i\}}$$

Il est possible que la silhouette s_i soit négative: cela indique généralement des observations mal regroupées.

De bons regroupements seront obtenus si la silhouette moyenne est élevée.

Graphique des silhouettes

Coûteux en calcul (nécessite matrice de dissimilarité), possible de faire avec un sous-échantillon aléatoire.

La segmentation de droite de la Figure 4 est supérieure parce que les regroupements sont plus homogènes et mieux équilibrés.

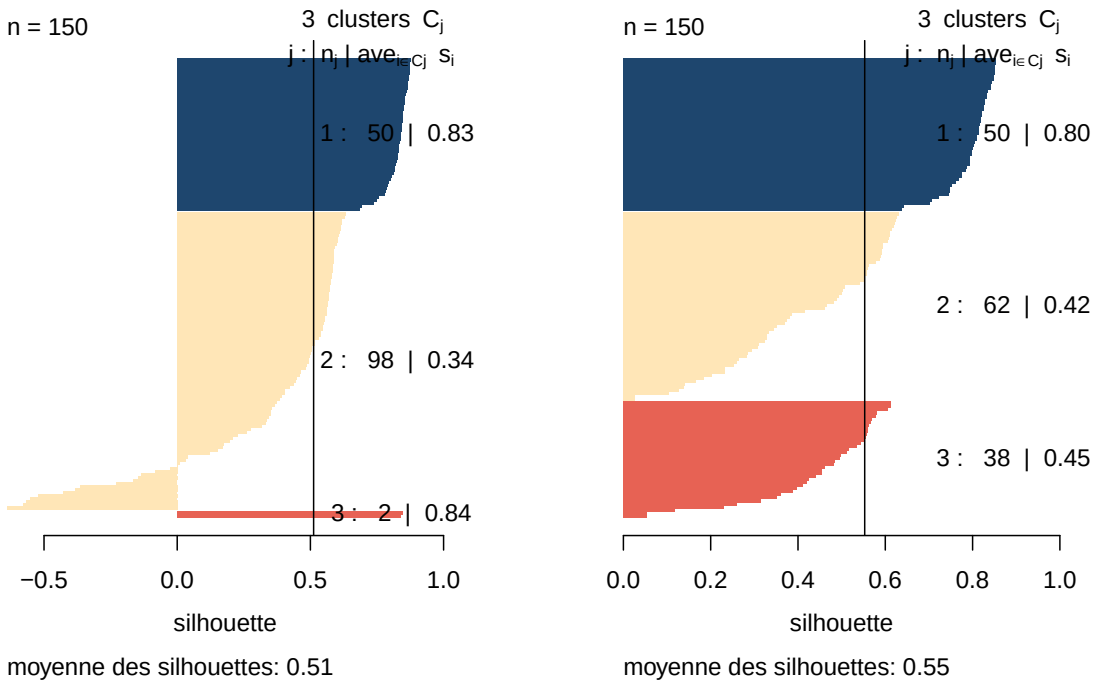
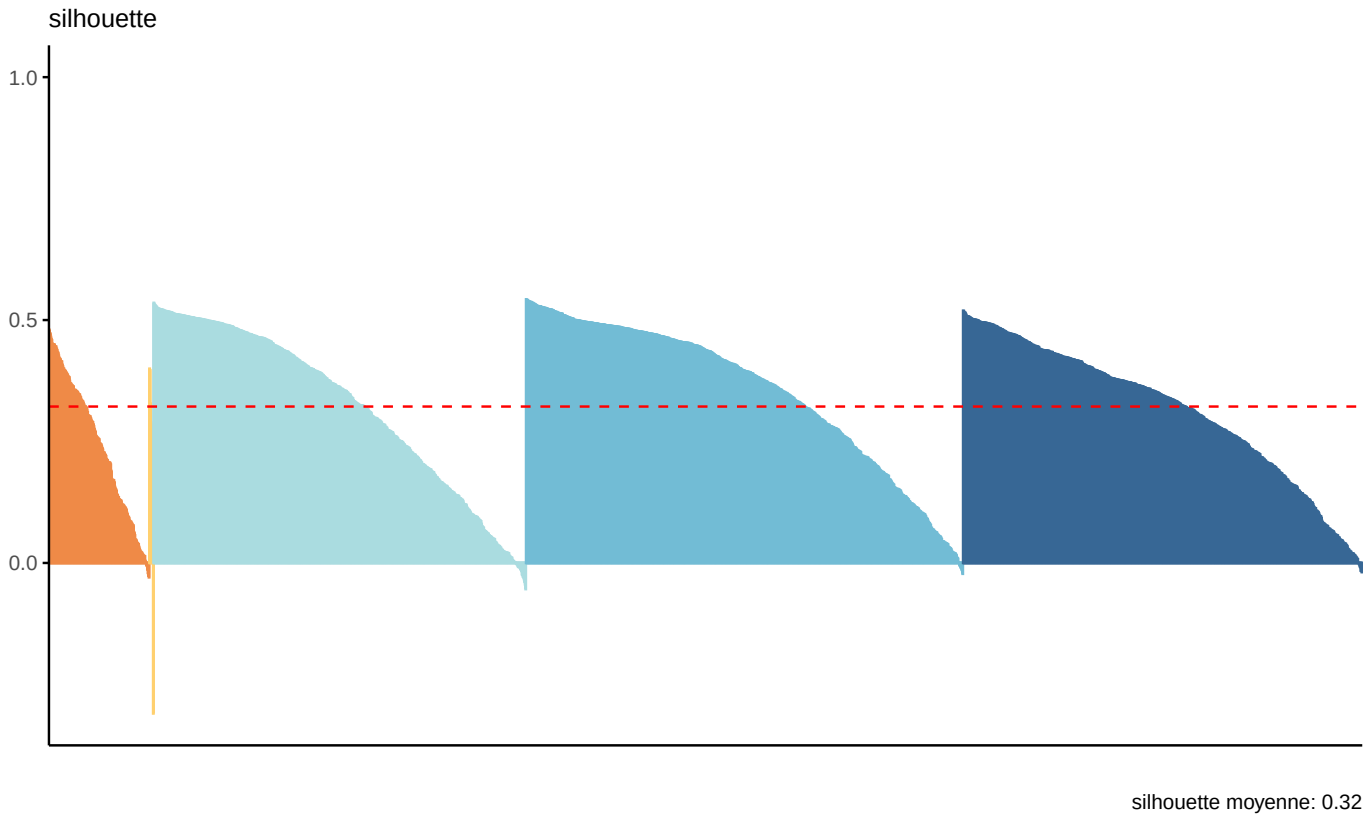


Figure 4: Profil des silhouettes pour deux regroupements d'un jeu de données.

Silhouettes avec segmentation des K-moyennes



Règle importante pour le choix de K

Utilisez votre jugement (et le gros bon sens).

Les segments doivent être interprétables.

Vérifiez que la taille des segments n'est pas fortement débalancée.

Statistiques descriptives par segment

Table 1: Moyenne des variables explicatives par segment (segmentation avec K-moyennes et cinq regroupements).

	1	2	3	4	5
décompte	993	64	3812	4496	4252
mtdons	13.92	445.49	24.98	15.32	12.11
ndons	2.98	11.38	13.71	4.00	4.63
recence	64.56	67.14	27.34	29.00	172.06
anciennete	219.46	255.45	252.77	83.59	247.85
vdonsmax	22.39	1069.30	61.19	22.53	19.23
ddons	7.49	1.92	1.65	1.60	1.87
nrefusconsec	1.82	0.52	0.47	0.62	3.15
snrefus	3.17	0.88	1.05	4.23	2.71
mpromesse	15.32	620.32	45.13	17.70	7.67

Interprétation des segments

Les regroupements obtenus sont interprétables:

- Groupe 1: Petits donateurs, faible nombre de dons. N'ont pas donné depuis longtemps. Refus fréquents et délai entre dons élevés
- Groupe 2: Grands donateurs fidèles: plus petit groupe. Ces personnes ont fait plusieurs dons, leur valeur maximale est élevée. N'ont pas donné récemment.
- Groupe 3: Petits fidèles. Dons plus élevés que la moyenne, nombre de dons élevés et récents.
- Groupe 4: Petits nouveaux. Moins d'ancienneté, dons récents et refus fréquents relativement à l'ancienneté.
- Groupe 5: Donateurs inactifs. Faible montant de dons, plutôt anciens, plusieurs refus.

Récapitulatif

- L'analyse de regroupement (clustering) est une méthode d'apprentissage non-supervisée
- Plusieurs choix de l'analyste (mesure de dissimilarité, algorithme, choix des hyperparamètres) impactent la segmentation et peuvent mener à des résultats très différents avec les mêmes données.
- L'algorithme des K-moyennes est le plus employé et son faible coût permet son utilisation avec des mégadonnées.