

Analyse de regroupements

Analyse multidimensionnelle appliquée

Léo Belzile

HEC Montréal

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.

Analogie avec analyse factorielle

En analyse factorielle, on combine des variables similaires.

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

country	year	cases	population
Afghanistan	2000	3366	20995360
Brazil	1999	37737	17206362
Brazil	2000	80488	17404898
China	1999	210258	127215272
China	2000	210266	128008583

variables

country	year	cases	population
Afghanistan	2000	3366	20995360
Brazil	1999	37737	17206362
Brazil	2000	80488	17404898
China	1999	210258	127215272
China	2000	210266	128008583

observations

Ce sont des méthodes dites d'apprentissage non-supervisé: l'objectif est de déduire les regroupements sans étiquette préalable (contrairement à la classification).

1. Choisir les variables pertinentes à l'analyse. Cette étape peut nécessiter de créer, transformer de nouvelles variables ou d'aggréger les données.
2. Choisir un algorithme et ses hyperparamètres (dissemblance, nombre de regroupements, etc.)
3. Regrouper et étiqueter les observations.
4. Valider la qualité de la segmentation.
5. Obtenir le prototype de chaque groupe et interpréter les regroupements à l'aide de ces derniers.

- 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

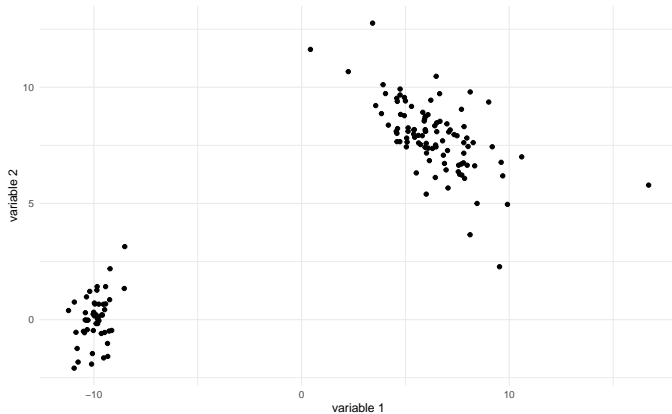


Figure 1: Données simulées avec deux regroupements hypothétiques.

Quelles variables X_1, \dots, 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 agréger les bases de données marketing par identifiant client.

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

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 agrégées pour créer des segments?

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.

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.

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

Nous avons plusieurs variables:

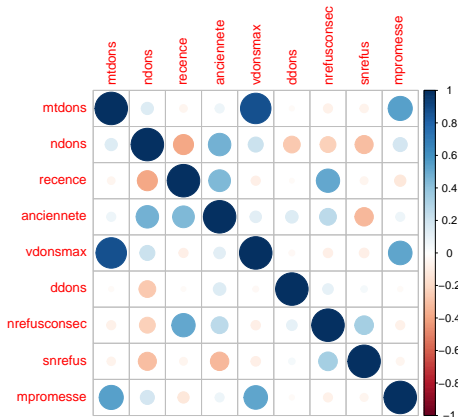
- la valeur des dons (total, min, max)
- la valeur des promesses,
- le délai entre les dons (moyen, min, max)
- le nombre de dons,
- la fréquence,
- l'ancienneté, etc.

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

Modification des données

```
1 donsult <- hecmulti::dons |>
2   filter(ndons > 1L) |> #seulement dons multiples
3   mutate(mtdons = vdons/ndons, # montant moyen
4          # nombre de dons selon l'anciennete
5          snrefus = nrefus/anciennete*mean(anciennete),
6          mpromesse = case_when(
7            npromesse > 0 ~ vpromesse/npromesse, #montant moy
8            TRUE ~ 0)) |> # zero si aucune promesse
9   select(!c(
10    vradiations, # trop de valeurs manquantes dans vradiation
11    nindecis, vdons, ddonsmax,
12    ddonsmin, vdonsmin, npromesse,
13    vpromesse, nrefus, nradiations)) |>
14   relocate(mtdons) # mettre mtdons dans la 1e colonne
```

Corrélation des nouvelles données



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 coordonnées.

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

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)

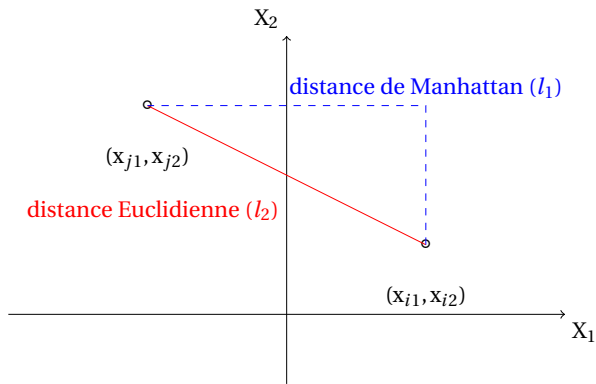
Toute distance est une mesure de dissimilarité.

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

$$d(X_i, X_j; l_2) = \{(X_{i1} - X_{j1})^2 + \dots + (X_{ip} - X_{jp})^2\}^{1/2}.$$

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



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

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.

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

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

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

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)

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!)

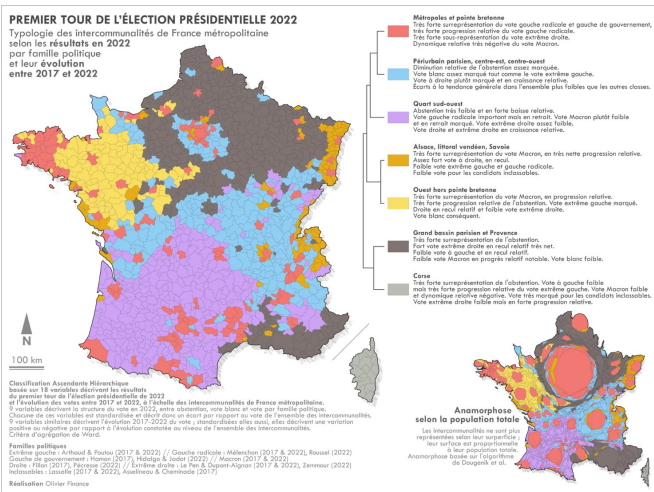
```
1 # Standardisation usuelle
2 # (soustraire la moyenne, diviser par écart-type)
3 donsmult_std <- scale(donsmult)
4 # Extraire moyenne et écart-type
5 dm_moy <- attr(donsmult_std, "scaled:center")
6 dm_std <- attr(donsmult_std, "scaled:scale")
7 # Standardisation robuste
8 donsmult_std_rob <- apply(
9   donsmult,
10  MARGIN = 2,
11  FUN = function(x){(x - median(x))/mad(x)}
```


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)

Exemple: typologie des votants en France



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 -médoïdes)
2. mélanges de modèles
3. méthodes basées sur la connectivité (regroupements hiérarchiques agglomératifs et divisifs)
4. méthodes basées sur la densité (pas couvertes)

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

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

- 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.

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 a priori

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

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.

$$\min_{\mu_1, \dots, \mu_K} \sum_{i=1}^n \min_{c_i \in \{1, \dots, K\}} d(X_i, \mu_{c_i}) \quad (1)$$

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: on sélectionne préalablement

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

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).

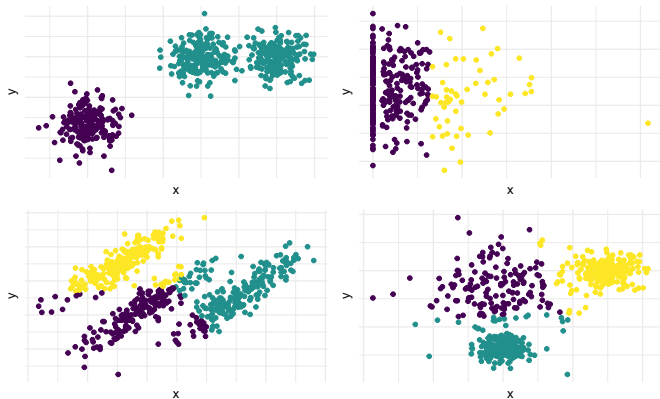
Visionner l'animation en ligne ou ce site.

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.

- (—) 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.

Illustration de segmentations problématiques avec K -moyennes



Séparation linéaire de l'espace

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

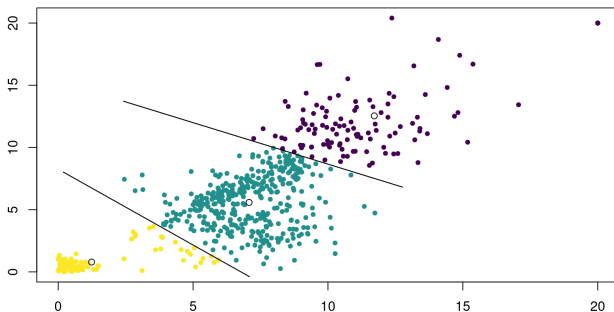


Figure 2: Partitions de Voronoï pour les regroupements avec séparateur linéaire.

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

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).

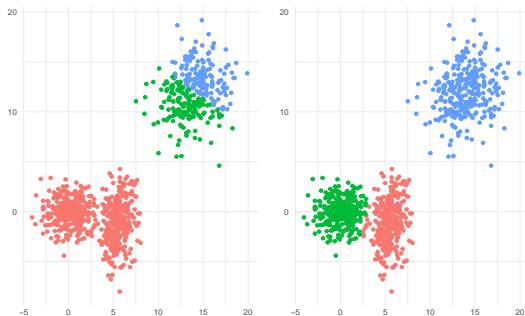


Figure 3: Regroupements pour $K = 3$ groupes avec une mauvaise initialisation principale (gauche) et une bonne initialisation (droite).


```
1 set.seed(60602)
2 kmoy5 <- kmeans( # distance euclidienne
3   x = donsmult_std, # données
4   centers = 5L, # nb groupes
5   nstart = 10, # nb initialisation aléatoire
6   iter.max = 25) # nb étapes maximum dans optimisation
7
8 kmoy5$cluster # étiquettes
9 kmoy5$size # répartition
10 kmoy5$tot.withinss # fonction objective minimal
11 kmoy5$centers # barycentres (données standardisées)
```

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

○ 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,

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

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 .

Utiliser le paquet flexclust

```
1 set.seed(60602)
2 kmoypp5 <- flexclust::kcca(
3   x = donsmult_std,
4   k = 5, # nb groupes
5   family = flexclust::kccaFamily("kmeans"),
6   control = list(initcent = "kmeanspp"))
7 # Vérifier répartition
8 kmed5@clusinfo
9 # Coordonnées des prototypes standardisés
10 kmed5@centers
11 # Étiquettes
12 kmed5@cluster
```

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"

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

$$\text{SCD}_K = \text{SCD}_{1,K} + \dots + \text{SCD}_{K,K};$$

où

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

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 $\text{SCT} = \text{SCD}_1$, la somme totale du carré de toutes les distances par rapport au barycentre global.

La valeur optimale de la somme du carré des distances mesure va mécaniquement* diminuer quand K augmente

$$SCD_1 > SCD_2 \dots$$

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.

On peut mesurer le pourcentage de variance expliquée,

$$R_K^2 = 1 - \frac{SCD_K}{SCT}.$$

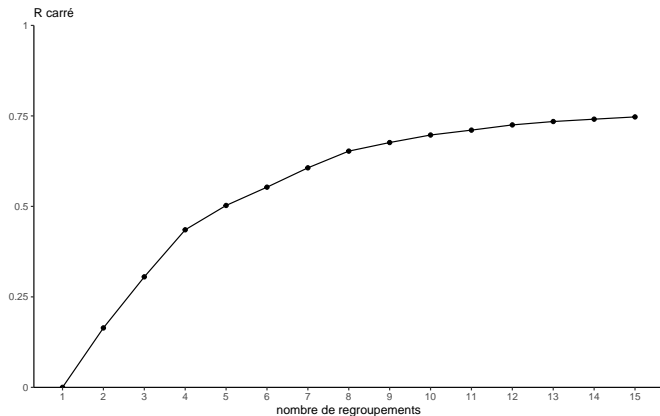
On cherche un point d'inflexion (coude) à partir duquel l'amélioration est négligeable.

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

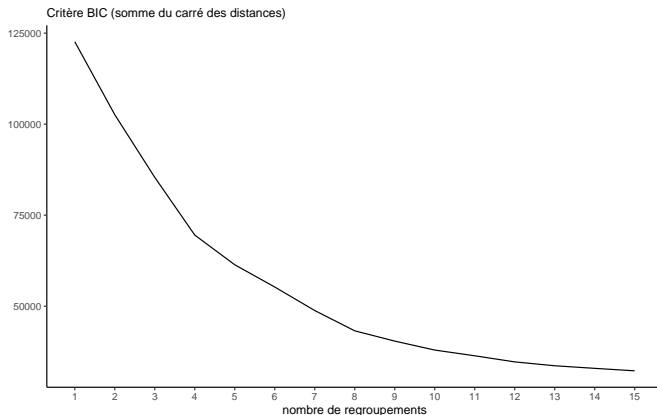
$$BIC = SCD_K + \ln(n)Kp$$

La plus petite valeur du "BIC" est préférable.

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



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



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.

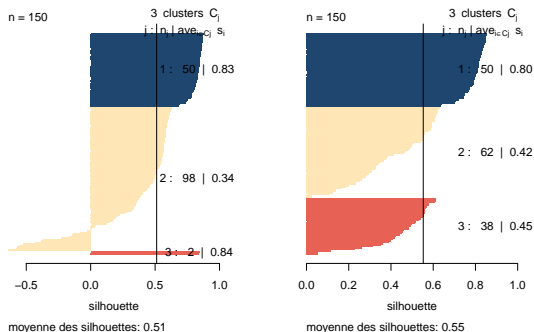
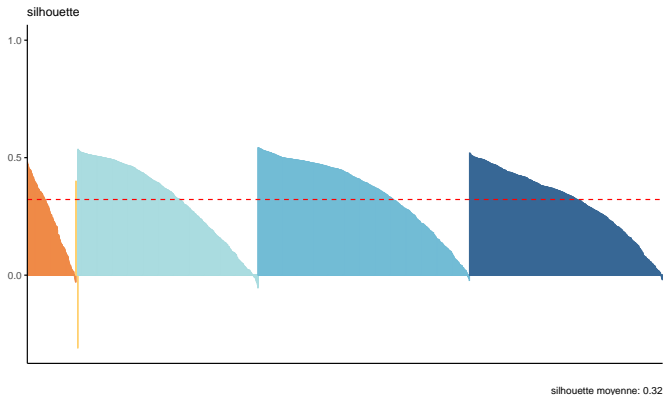


Figure 4: Profil des silhouettes pour deux regroupements d'un jeu de données.

Silhouettes avec segmentation des K -moyennes



Utilisez votre jugement (et le gros bon sens).

Les segments doivent être interprétables.

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

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
mtdots	13.92	445.49	24.98	15.32	12.11
ndots	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
vdotsmax	22.39	1069.30	61.19	22.53	19.23
ddots	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

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.

- 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.