Analyse de regroupements

Analyse multidimensionnelle appliquée

Léo Belzile

HEC Montréal

automne 2022

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.

Analyse **descriptive** des segments.

Illustration

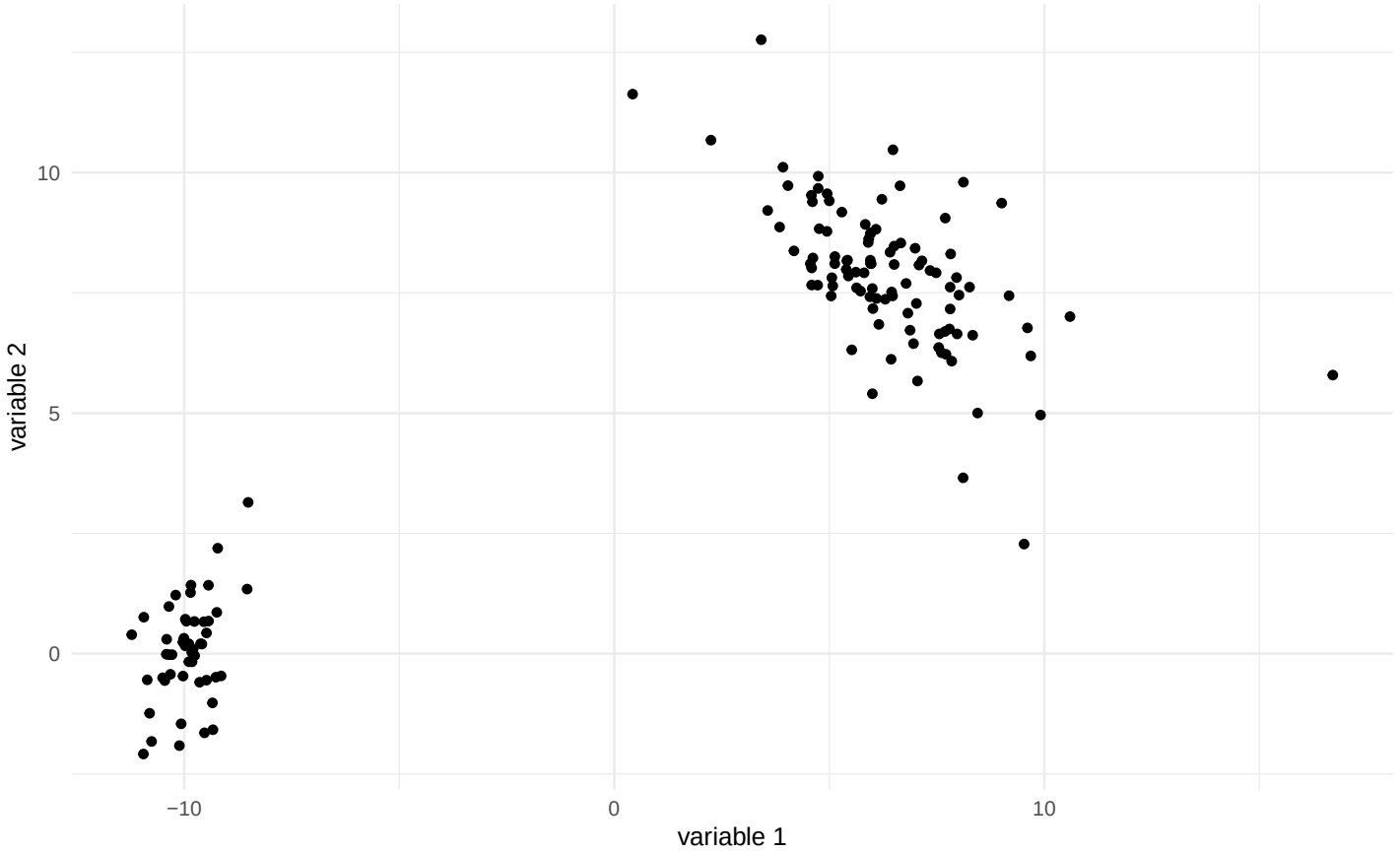
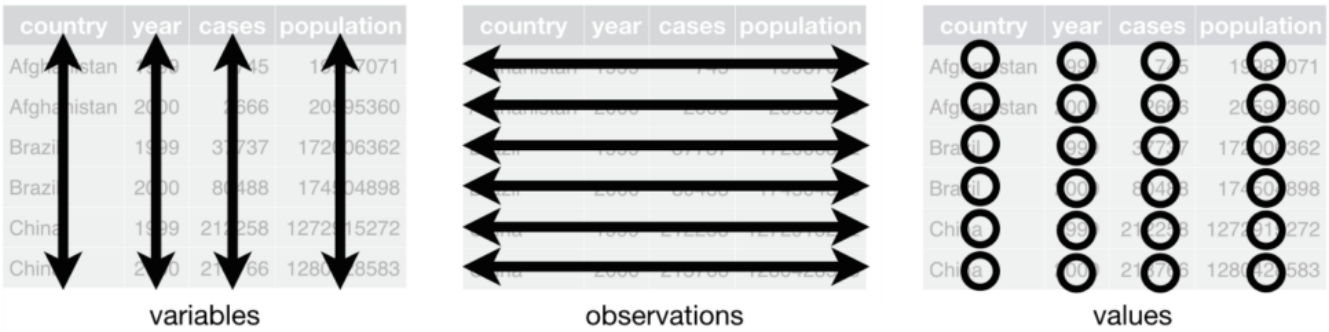


Figure 1: Données simulées avec deux regroupements hypothétiques.

Analogie avec analyse factorielle

En analyse factorielle, on combine des variables similaires (colonnes).

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



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).

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

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

Regrouper les bases de données marketing par identifiant client (aggrégation).

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

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.

À votre tour

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

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

Choix des variables

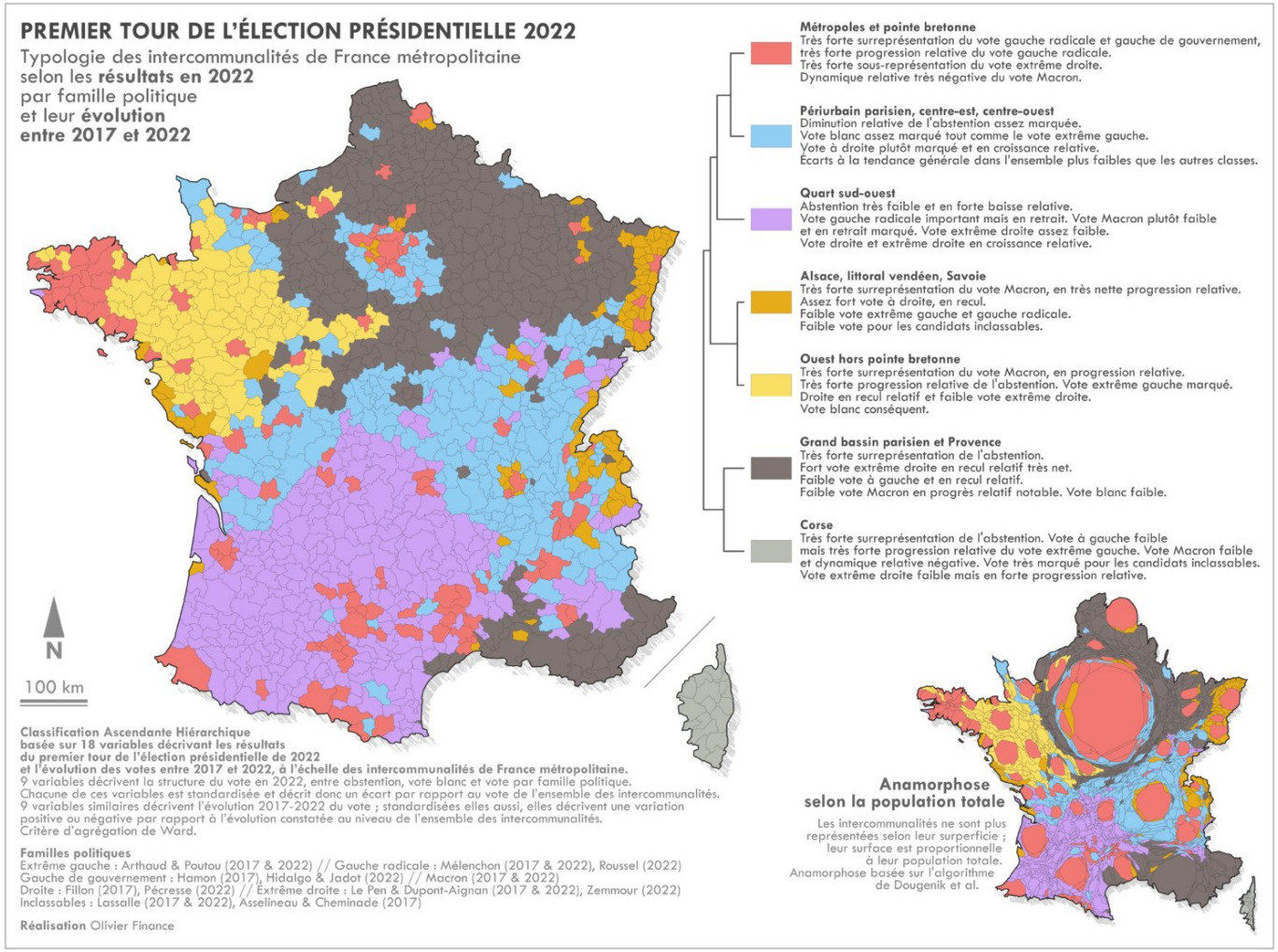
Recommandations: choisir les variables pertinentes qui font ressortir les effets voulus.

- inclure de nombreuses variables similaires dilue les différences.
- transformer 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: typologie des votants en France



Exemple - dons à un organisme de charité

La base de données `dons` contient 49730 observations pour 16 variables.

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

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.

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 et quelle mesure de dissemblance/similarité seront utilisées pour la segmentation.
3. Choisir les hyperparamètres de l'algorithme (nombre de regroupements, rayon, etc.)
4. Procéder à l'analyse de regroupements.
5. Calculer une mesure de qualité.
6. Assigner les étiquettes aux observations et calculer un prototype de groupe.
7. Interpréter les regroupements obtenus.