L’analyse de classification - hiérarchique et de nuées dynamiques

Cours Sol6210 Analyse quantitative avancée

© Claire Durand, 2024
Professeur titulaire,
Département de sociologie,
Université de Montréal

Deux grandes familles (Lebart, Morineau, Piron, 2000)

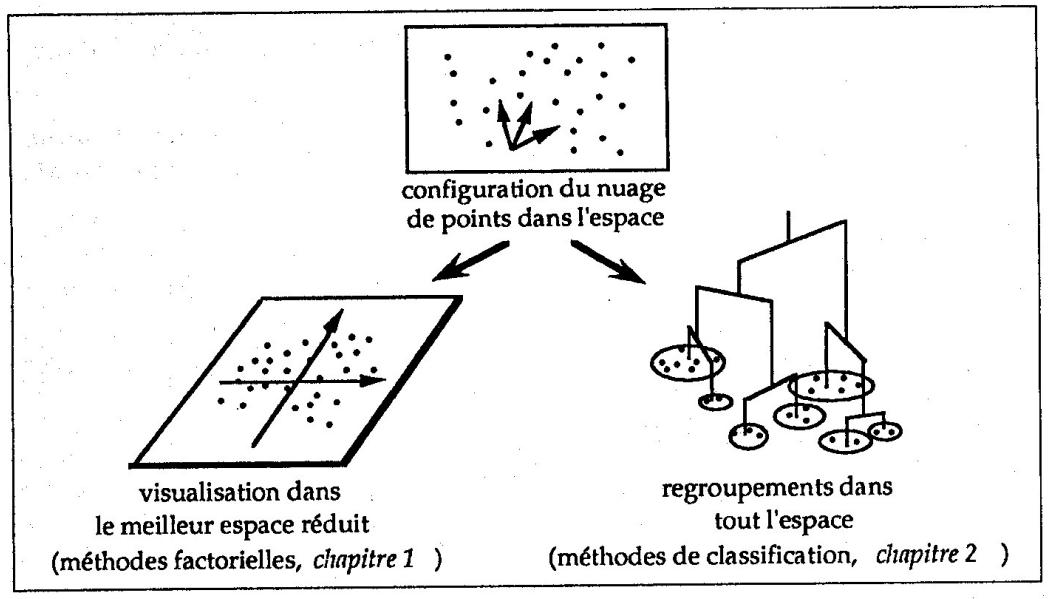

Figure 2
Les deux grandes familles de méthodes

Ces méthodes impliquent souvent de la même manière les individus (lignes) et les variables (colonnes). Les individus ne sont plus de simples intermédiaires utilisés pour calculer des moyennes ou des corrélations sur les variables, suivant le schéma de la statistique traditionnelle où ils ne sont que des réalisations d'épreuves indépendantes. La confrontation des espaces d'individus et de variables enrichira les interprétations.

Qu'est-ce que la classification?

- Un ensemble de méthodes
- Qui visent à regrouper les "cas" – qui peuvent être des individus, mais aussi des départements, des villes, des pays, etc....
- Selon la similitude de leurs réponses à un certain nombre de questions, d'indicateurs, de variables
- Ce qui permet d'identifier ou de valider, le cas échéant, des typologies.

Pourquoi l'analyse de classification...

- Pas de restrictions relativement au nombre de cas par variable
- Permet de sauvegarder une variable d'appartenance à une classe que l'on peut réutiliser dans d'autres analyses
- Tient compte des profils de réponse
- Permet de "visualiser" les profils de réponse
- Multiples mesures de distance disponibles
- Multiples algorithmes de regroupement
- Sert à construire/valider des typologies
- On peut rechercher des différences (et non seulement des similitudes)

Les étapes (selon Rapkin et Luke, 1993)

- 1. Identifier les “cas”: de quel type de cas s’agit-il? Est-ce qu’on les conserve tous?
- 2. Choisir, réduire, standardiser/pondérer les variables.
- 3. Décider de la mesure de distance entre les cas et entre les groupes de cas.
- 4. Choisir l’algorithme de regroupement
- 5. Déterminer le nombre de classes
- 6. Choisir le logiciel statistique et procéder à l’analyse

Les étapes (selon Rapkin et Luke, 1993)

Suite

7. Interpréter les profils de chaque classe
8. Vérifier la stabilité de la solution
9. Vérifier la validité de la solution
10. Présenter les résultats.

1. Identifier les cas

- Ils peuvent être
  a. Des individus
  b. Des professions
  c. Des catégories d'objet, villes, départements, universités, des textes, des discours, etc...
  d. Des variables...

2. Choisir, réduire, standardiser, pondérer les variables

Les variables sont métriques ou pseudo-métriques

- S'il y a une trop forte corrélation entre les variables, c'est équivalent à donner plus de poids à certaines informations, donc...
- Si nécessaire, faire une analyse factorielle avant l'analyse de classification pour réduire le nombre de variables et avoir des variables le plus indépendantes possible.
- Attention :
  a. Les mesures de distance sont liées à l'échelle de mesure
  b. Par contre, si on standardise, on perd l'information sur la moyenne et sur la variance d'origine.
  c. On peut ramener à la même échelle (voir diverses méthodes dans SPSS) sans standardiser.
- On pourrait choisir de donner des poids plus importants à certaines variables pour des raisons théoriques.

3. Décider de la mesure de distance

Elles ont chacune leurs avantages et inconvénients

- Entre autres :
  a. Euclidiennes: les distances plus fortes ont plus d'importance (parce qu'on les met au carré)
  b. "city-block" ou Manhattan: distance absolue
  c. Chebychev: distance maximum sur une variable
  d. Mahalanobis: basée sur la régression
  e. Lambda (pour variables de type 0,1)
  f. Corrélations (pour les variables).
- On peut utiliser n'importe quelle matrice de distance comme "entrée" pour l'analyse.

4. Choisir l'algorithme de regroupement

D'abord la structure

- Compact: chaque membre de la classe est plus similaire à tous les autres membres de sa classe qu'aux membres des autres classes.
- En chaîne: Chaque membre de la classe est plus similaire à un membre de sa classe qu'aux membres des autres classes.

4. Choisir l'algorithme de regroupement

Hiérarchique: valide pour un nombre plus restreint de cas (1000 + environ)

- Hiérarchique agglomératif :
  a. On joint ensemble les cas les plus semblables par étapes jusqu'à ce que tous les cas soient dans une seule classe
- Méthodes d'agglomération:
  a. Single linkage (voisin le plus proche) → en chaîne
  b. Complete linkage (voisin le plus éloigné) → en chaîne
  c. Moyenne entre les classes → compact
  d. Moyenne intra classe → compact
  e. Ward → compact
  f. Centroïde → compact
  g. Médiane → compact
- Hiérarchique divisif (peu disponible): On part d'une classe que l'on divise en écartant les cas les plus différents (comme les arbres de segmentation).

4. Choisir l'algorithme de regroupement

Itératif ou mixte

- Itératif (nuées dynamiques): On décide d'un point de départ au hasard pour chaque classe -- *il faut spécifier le nombre de classes au départ* --et on met dans chaque classe les points les plus proches du point de départ. On recalcule le centre de la classe et on refait l'agglomération jusqu'à la stabilité des classes.
  a. Les cas peuvent changer de classes en cours de processus.
  b. Méthodes multiples pour décider de la manière d'attribuer les cas à des classes.
  c. Problème: La classification peut dépendre du point de départ.
- Mixte: hiérarchique en deuxième étape sur classes (nombreuses: 100?) obtenues via nuées dynamiques.

5. Déterminer le nombre de classes

Les critères

- La soudaine augmentation de la distance entre les classes (test de l'éboulis ou du "coude de Cattell").
- Le nombre de cas par classe.
- La capacité de la solution à différencier entre les profils.
- L'homogénéité des profils internes des classes.
- L'ajout de variance expliquée par les classes.
- La stabilité des solutions.
- L'interprétation

6. Quel logiciel utiliser?

- SPSS ou STATA :
  a. Facile
  b. Multiples méthodes pour standardiser (SPSS)
  c. Multiples mesures de distance
  d. Multiples méthodes d'agglomération (Stata: Kmedian)
  e. SPSS a maintenant une two-step method qui permet de faire la classification avec des variables nominales et continues.
- SPAD ($), DTM-VIC, TRI-DEUX:
  a. Possibilité de faire l'analyse en combinaison avec l'analyse factorielle des correspondances; la distance est déterminée par le positionnement des cas sur les facteurs (axes).
  b. Représentation visuelle très intéressante avec SPAD.

7. Interpréter les profils

- Demande l'utilisation de procédures complémentaires:
  a. Anova, manova pour analyser les différences de moyennes et voir si les classes se distinguent sur toutes les variables.
  b. Graphiques représentant le processus d'agglomération (à utiliser quand il y a peu de cas: villes, pays, etc.).
  c. Graphiques de moyennes pour visualiser les profils.
  d. Dans SPAD: représentation visuelle des classes.
- Est-ce que les profils qu'on en dégage ont un sens dans le cadre de nos recherches, de la théorie, etc. Peut-on "mettre un nom" sur chaque type?

8. Vérifier la stabilité des classes

- Est-ce que la répartition des cas dans les classes change de façon sensible...
  a. Si j'utilise des paramètres de départ différents:
     i. Des mesures de distance différentes
     ii. Des méthodes d'agglomération différentes
  b. Je fais une première classification dans un sous-groupe et je regarde si ca tient lorsque j'utilise le reste des cas.
- Les analyses de prédiction (logistique, discriminante) pour estimer l'appartenance aux classes à partir des variables utilisées pour les produire donnent-elles des résultats similaires?
- Statistique de Rand.

9. Vérifier la validité des classes

Interne et externe

- Est-ce que chaque variable utilisée dans l'analyse a une relation significative avec la variable de classification? (Validité de construit)
- Est-ce que la classification est liée avec des variables externes auxquelles on pense qu'elle devrait être liée? (Validité critérielle - prédictive)

10. Présenter les résultats

Le visuel....

- Les analyses de classification ***exigent*** de recourir à des présentations visuelles pour permettre de bien comprendre ce qui distingue les classes.
  a. Graphiques en batons.
  b. Graphiques spécifiques (voir SPAD).

“Contraintes” de l’analyse de classification classique

- Les échelles de réponse doivent être de même ordre.
- **La nécessaire indépendance des variables.**
- Par rapport à d’autres procédures de classification
  a. Analyse discriminante
  b. Analyses de segmentation
  c. Régression logistique
  d. Analyses factorielles...