Analyse de données quantitatives (LPOLS1221)

Séance 5 : Analyses de classification

Le clustering : rappel théorique

Objectif : regrouper l'ensemble des unités d'analyse dans des groupes (plus ou moins) homogènes (où les observations soient le plus similaires possible), dans une logique de compromis entre la synthèse et une perte raisonnable d'informations. On vise donc :

- une forte similarité intra-groupe
- une faible similarité inter-groupe

Etapes :

1. **Choisir le critère de regroupement**. On choisit entre la distance (euclidienne) ou bien la ressemblance (coefficient de corrélation) entre les observations. Souvent, on opte pour le critère de la distance euclidienne :

$$d = \sqrt{\sum_{i=1}(a_{var_i} - b_{var_i})^2} \quad avec \begin{cases} a\{var_1; var_2; var_3; \dots; var_n\} \\ b\{var_1; var_2; var_3; \dots; var_n\} \end{cases}$$

   On utilise plutôt le carré de la distance euclidienne. SPSS fournit la matrice de distances euclidiennes séparant chaque deux observations, qui se présente sous la forme d'un tableau croisant. C'est à partir de ces distances que les regroupements peuvent être effectués, en commençant nécessairement par les unités les plus proches et en regroupant ensuite progressivement des unités différentes.
2. **Choisir un algorithme de regroupement des observations**. La **méthode de Ward** utilise la variance et propose de regrouper les classes *en minimisant la variance intra-classe, ce qui conduit à un accroissement de la variance inter-classe*. Pour cela, on calcule la distance euclidienne au carré des barycentres ou points d'équilibre (centroid) de classes (G) pondéré par le produit des poids (W) de classe divisé par la somme de ces poids. Au départ, toutes les classes contiennent donc une et une seule observation (singleton) et ont donc le même poids et un barycentre égal à l'observation elle-même. Le coefficient de Ward la plus faible entre deux classes déterminera la nouvelle agglomération. Lorsque deux classes sont rassemblées, les distances sont alors recalculées vis-à-vis des barycentres de classe actualisés et du poids de chaque classe.

   Coefficient de Ward pour deux classes A et B, de poids W et de barycentre G.

$$\Delta(A,B) = \frac{W_A W_B}{W_A + W_B} d^2(G_A, G_B)$$

   Il s'agit donc d'un algorithme itératif car dès que deux observations sont regroupées, il faut recalculer les distances de tous les points pour effectuer le prochain regroupement. Il y aura en tout et pour tout N-1 calcul pour arriver à l'agglomération totale des observations.
3. **Décider le nombre de clusters** (groupes) qui mieux représentent les données (décision à prendre du cas par cas, mais souvent entre 2 et 4/5). Pour cela, on analyse a) l'évolution des **coefficients de Ward** associées aux dernières étapes de regroupement et b) le **dendrogramme**, qui est la représentation graphique du

clustering hiérarchique (qui est très utile, mais nécessite généralement d'être redimensionné). Le dendrogramme illustre à partir de quelle étape les distances entre classes qui sont fusionnées deviennent importants (on s'intéresse donc à de longues distances entre deux lignes verticales successives).
L'exercice ici est de faire attention à faire certes, des regroupements (et donc de créer des variances intra-classe) et donc d'avoir un nombre « gérable » de clusters, mais également d'être vigilant à ne pas fusionner des individus trop différents.

4. **Générer une nouvelle variable** sur base de la solution de classification choisie. Cette nouvelle variable aura un nombre de modalités égal au nombre de clusters choisis. Ainsi, on attribue une modalité à chaque observation (unité d'analyse).
5. **Décrire les classes** sur base de tableaux croisés entre la nouvelle variable en lignes et chaque variable ayant servi à la classification (+ une ou plusieurs caractéristiques sociodémographiques) en colonnes. Ensuite, on résume le tout dans un tableau de synthèse (faisant attention aux p-valeurs des tests Khi-carré).

Pour la description des classes, il faut faire attention à trois aspects :
- **L'importance relative de la classe**
- **Les particularités spécifiques de la classe**
- **Ce qui distingue cette classe des autres.**

Les mêmes manipulations peuvent être reproduites, à la différence près que plutôt que d'inclure les variables d'origine pour la classification, on inclut les composantes qui ont été créées au préalable, à partir d'une analyse factorielle. L'avantage est que, dans ce cas, il s'agit de **variables directement standardisées**.

Manipulations SPSS

1. Analyse *(Analyze)*→ Classifier *(Classify)*→ Cluster hiérarchique *(Hierarchical cluster)*
2. Dans l'encadré « Variable(s) » (Variables), insérer la liste des variables numériques, dichotomiques ou ordinales sur lesquelles vous voulez effectuer la classification
3. Dans les onglets suivants, sélectionnez :
   - **Statistiques *(Statistics)* :** Le « planning des agglomérations » (Agglomeration schedule) indique les étapes du regroupement et peut être représentée graphiquement par un « icicle plot », tandis que la « matrice des distances » (Proximity matrix) vous permet de voir les distances entre les observations (donc une valeur plus petite signifie que les observations en questions sont plus proches l'une de l'autre) ; par contre, si l'on a beaucoup d'observations, afficher la matrice des distances risque d'alourdir significativement le output. Ensuite vous pouvez définir un nombre de classes (par exemple : entre 2 et 4).
   - **Tracé*(Plots)* :** sélectionner l'option « Dendrogramme »
   - **Méthode*(Method)* :** Sélectionner « Méthode de Ward », qui a l'avantage de créer des clusters d'effectifs plus ou moins similaires (autres options souvent utilisées sont notamment « centroid », « furthest neighbor » et « nearest neighbor »). Dans l'option distance, choisissez « carré de la distance euclidienne ». Dans le cas où les échelles des variables utilisées varient fortement (e.g. une variable est dans les miliers, tandis qu'une autre est dans les sous-unités), il peut s'avérer utile de standardiser les variables.
   - **Enregistrer *(Save)* :** Permet d'avoir la création d'une nouvelle variable avec les affectations de classes. Si vous sélectionnez plusieurs solutions dans statistiques, demander alors la même chose dans cet onglet (par exemple : entre 2 et 4).
   - **Cluster :** Observations

Applications

Utilisez la base de données EVS pour réaliser ces exercices.

1. Réalisez une classification des unités d'analyse (individus) **ayant participé à l'enquête en Belgique**, en fonction des facteurs qu'ils considèrent importants dans l'institution du mariage. Les variables suivantes seront utilisées :

Nom variable	Intitulé variable
V136	Important in marriage: faithfulness
V138	Important in marriage: same social background
V139	Important in marriage: shared religious beliefs
V145	Important in marriage: children
V146	Important in marriage: discuss problems

2. Réalisez une ACP sur ces cinq variables ci-dessus et ensuite effectuez à nouveau l'analyse de classification (toujours sur l'échantillon de Belgique), mais cette fois-ci, utilisant les résultats de l'ACP comme variables de départ.

3. Sélectionnez uniquement **les ressortissants belges, âgés de moins de 25 ans**. Puis, réalisez une classification des individus à partir des variables V268, V269, V270, V271, V272 et V273.

Nom variable	Intitulé variable
V268	Les immigrés prennent les travaux des locaux
V269	Les immigrés ébranlent la vie culturelle des pays d'accueil
V270	Les immigrés influencent à la hausse la criminalité
V271	Les immigrés sont une charge supplémentaire pour le système de sécurité sociale
V272	Les immigrés deviendront une menace pour la société
V273	Les immigrés conservent leurs propres habitudes/coutumes

4. Répondez aux questions suivantes :

   a. *Combien de classes proposez-vous de créer ? Quel(s) est/sont les éléments qui vous permettent de justifier ce choix du nombre de classes ?*

   b. *Quelle est la perte d'informations liée à ce regroupement ?*

   c. *Comment pouvez-vous caractériser les différentes classes ? Comment se distinguent-elles les unes des autres ?*
   Rem : pour la description des classes, utilisez les six variables mobilisées. Testez également la présence d'une relation entre la variable de classes et d'autres variables sociodémographiques que vous jugeriez pertinentes.