Séance 9 – Méthodes statistiques multivariées descriptives

1 Questions de cours

Question 1
La géographie mobilise fréquemment la statistique multivariée car elle s'intéresse à des phénomènes complexes,
qui ne peuvent être décrits par une seule variable. Les territoires, les populations ou les systèmes productifs
sont caractérisés simultanément par des variables sociales, économiques, démographiques, environnementales ou culturelles.
Les méthodes multivariées permettent de synthétiser cette complexité, de révéler des structures spatiales,
des gradients, des oppositions ou des proximités entre unités spatiales, tout en conservant l'essentiel de l'information.
Elles sont donc particulièrement adaptées à l'analyse territoriale.

Question 2
Les méthodes factorielles descriptives ont pour objectif principal l'exploration et la compréhension des données.
L'ACP vise à résumer l'information contenue dans un ensemble de variables quantitatives corrélées.
L'AFC analyse les relations entre les lignes et les colonnes d'un tableau de contingence.
L'ACM généralise l'AFC à plusieurs variables qualitatives.
L'AFM permet d'analyser conjointement plusieurs groupes de variables, en équilibrant leur contribution.
Ces méthodes n'ont pas un objectif prédictif mais interprétatif.

Question 3
Toutes les méthodes factorielles suivent une logique commune :
préparation des données, choix d'une métrique adaptée, calcul des axes factoriels, décomposition de l'inertie totale en valeurs propres,
projection des individus et/ou des variables dans des espaces de dimension réduite,
puis interprétation des axes à partir des contributions et des coordonnées.

Question 4
Les méthodes factorielles partagent des fondements mathématiques communs reposant sur l'algèbre linéaire et la géométrie euclidienne.
Elles cherchent toutes des axes orthogonaux maximisant l'inertie.
Elles ne sont donc pas indépendantes mathématiquement,
mais adaptées à des structures de données différentes.

Question 5
On distingue plusieurs types de tableaux :
le tableau individus-variables pour l'ACP,
le tableau de contingence pour l'AFC,
le tableau disjonctif complet pour l'ACM,
et les tableaux multiples structurés en groupes pour l'AFM.
Chaque structure impose une métrique spécifique et une méthode adaptée.

Question 6
Le choix de la méthode dépend avant tout du type de variables.
Les variables quantitatives appellent l'ACP.
Les variables qualitatives appellent l'AFC ou l'ACM.
Les données mixtes ou structurées en groupes appellent l'AFM.
L'AFM est souvent considérée comme la méthode la plus générale.

Question 7
La régression par les moindres carrés minimise les écarts verticaux entre les points observés et le modèle.
La régression orthogonale, utilisée implicitement en analyse factorielle,
minimise les distances perpendiculaires aux axes.
Cette dernière est plus cohérente avec une approche géométrique globale.

Question 8
Les valeurs propres mesurent la part d'inertie expliquée par chaque axe.
Plus une valeur propre est élevée, plus l'axe correspondant structure les données.
Elles permettent de hiérarchiser les axes et de décider du nombre de dimensions à conserver.

Question 9
La distance du khi-deux mesure l'écart entre un profil observé et un profil moyen, pondéré par les masses.
Elle est centrale en AFC et ACM car elle compare des distributions relatives
et neutralise les effets de taille des lignes et des colonnes.

Question 10
L'ACP repose sur des variables quantitatives et une distance euclidienne.
L'AFC analyse un tableau de contingence avec une distance du khi-deux.
L'ACM généralise l'AFC à plusieurs variables qualitatives via le tableau disjonctif complet.
Chaque méthode répond à une logique spécifique de données.

Question 11
Les profils lignes et colonnes expriment les distributions relatives internes aux lignes ou colonnes.
Ils permettent d'interpréter les écarts par rapport au profil moyen
et constituent la base de la distance du khi-deux.

Question 12
Les mappings représentent graphiquement les individus, variables ou modalités dans l'espace factoriel.
L'interprétation repose sur les proximités, les oppositions, la contribution aux axes
et la qualité de représentation ($\cos^2$).

Question 13
L'ACM est une généralisation de l'AFC.
Elle permet de traiter simultanément plusieurs variables qualitatives,
en analysant les associations entre leurs modalités
et les profils des individus.

Question 14
La classification vise à explorer la structure des données sans information préalable sur les classes.
Le classement ou l'analyse discriminante vise à affecter des individus à des groupes connus.
Les méthodes de regroupement reposent sur des distances et des critères d'agrégation.

Question 15
L'AFM est particulièrement adaptée aux données structurées en groupes de variables.
Elle permet de comparer les groupes, d'équilibrer leur influence
et d'analyser à la fois la structure globale et les structures partielles.

Question 16
Les méthodes factorielles s'inscrivent dans un cadre mathématique commun
mobilisant l'algèbre linéaire, la géométrie et la statistique multivariée.
Elles constituent un socle fondamental des analyses exploratoires modernes.

2 Mise en œuvre avec Python

Manipulation 1 – Analyse en composantes principales (ACP)

Manipulation 1.a
Les données sont importées à partir d'un fichier CSV à l'aide de la bibliothèque Pandas.
Cette étape permet de structurer les données sous forme de DataFrame
et de vérifier la nature des variables.

Manipulation 1.b
La colonne identifiant les villes est extraite afin de servir d'étiquette aux individus.
Les colonnes quantitatives restantes constituent la matrice d'analyse de l'ACP.

Manipulation 1.c
Les données sont centrées et réduites.
Cette étape est indispensable car les variables n'ont pas nécessairement la même unité ni la même variance.
La standardisation garantit une contribution équilibrée des variables.

Manipulation 1.d
Une ACP est calculée avec un nombre de facteurs égal au nombre de variables.
Cela permet d'analyser l'intégralité de l'inertie et d'identifier les axes dominants.

Manipulation 1.e
L'analyse des valeurs propres et de la variance expliquée montre que les premiers axes concentrent l'essentiel de l'information.
Le choix des deux premiers axes est justifié par leur inertie cumulée élevée.

Manipulation 1.f
Les résultats numériques de l'ACP (coordonnées, contributions, $\cos^2$)
sont organisés dans des DataFrames pour faciliter l'interprétation statistique.

Manipulation 1.g
Les coordonnées des individus sur les deux premiers axes sont projetées dans un plan factoriel.
Le mapping permet d'identifier des proximités géographiques ou climatiques entre les villes
et de donner une interprétation spatiale aux axes.