Université de Thessalie
Département d'Aménagement,
D'Urbanisme et Développement Régional

Enseignant : As. Pr. Marie-Noelle Duquenne

Ce cours est partie intégrante de l'U.E. 2.1. « Techniques, Méthodes et Outils ». L'objectif essentiel du cours est d'acquérir les connaissances fondamentales en matière d'analyse des données socio-économiques, démographiques et géographiques des unités territoriales. L'accent est effectivement mis sur l'étude des unités territoriales aux diverses échelles possibles (nationale, régionale et surtout locale) de façon à retranscrire de façon synthétique leur fonctionnement et spécificités (territorialité) et pouvoir dans un deuxième temps, procéder à leur représentation cartographique. Ce cours est en synergie totale avec l'ensemble des cours de l'U.E. 2.1. et plus spécialement les cours de cartographie thématique et de démographie spatiale.

Introduction :

Le recours aux méthodes d'analyse des données est aujourd'hui une pratique très courante dans de nombreuses disciplines scientifiques. Concernant plus précisément les phénomènes socio-économiques, les évolutions démographiques et plus généralement les dynamiques territoriales, ces méthodes contribuent à leur meilleur représentation et interprétation et fournissent bien souvent des informations qui peuvent être par la suite cartographiées.

Comme le soulignaient déjà en 1980 Bouroche et Saporta[1], «*contrairement à une idée très répandue, les méthodes d'analyse des données ont été élaborées depuis fort longtemps : H. Hotelling, dans les années 30, posait les fondements de l'analyse en composantes principales et de l'analyse canonique, en développant les travaux de C. Spearman et de K. Pearson qui dataient du début du siècle* ».

L'important développement des bases de données et l'essor toujours plus rapide des ordinateurs (capacité de stockage et de traitement) ont conduit à mettre au point non seulement de nombreuses méthodes pour synthétiser les informations volumineuses et repérer les grandes structures d'un vaste tableau de données quantitatives et qualitatives mais également des logiciels statistiques de plus en plus performants. C'est donc vers la fin des années 60

[1] Bouroche J.M., Saporta G., (1980), L'analyse des données, Edition PUF, Que sais-je, No 1854, 127 pages

que ces méthodes reposant sur une masse considérable de calculs, se sont fortement perfectionnées et furent largement vulgarisées.

La **Statistique classique** est axée sur l'étude d'un nombre restreint de caractères mesurés sur un nombre relativement limité d'individus[2]. La statistique classique reste néanmoins un outil indispensable dans la mesure ou il nous permet de mesurer les paramètres de tendance centrale et de dispersion de chacun des caractères étudiés, approche préliminaire bien utile pour comprendre la réalité des choses. Elle nous permet surtout d'étudier les relations entre deux caractères, c'est à dire de déterminer la nature et le sens de la relation logique – si elle existe – entre les deux caractères (voir annexe 1 sur la relation entre 2 caractères). C'est au travers de l'analyse statistique que se sont développées les méthodes d'estimation et de tests fondés sur des hypothèses probabilistes assez restrictives néanmoins.

Les **méthodes d'analyse des données** sont des méthodes d'analyse exploratoire qui permettent une étude globale des individus et variables grâce à des représentations graphiques suggestives. Or dans la réalité, le comportement des individus dépendent justement de nombreux caractères. Ces méthodes ont donc pour objectif premier de résumer des données volumineuses et de repérer les grandes structures d'un vaste tableau de données quantitatives et / ou qualitatives. Il s'agit par excellence d'une approche multidimensionnelle.

Notre cours a pour objectif de procéder à une présentation théorique de ces méthodes, de leur intérêt, de leurs biais et limites, ainsi qu'à une initiation à l'application concrète de ces divers outils, mettant en exergue les précautions nécessaires à leur utilisation. Il s'agit de familiariser les étudiants aux méthodes les plus courantes d'Analyse des Données Multidimensionnelles et multi variées. Les deux principaux types de méthodes de la statistique multidimensionnelle seront traités à l'aide du logiciel SPSS qui permet de traiter un important volume de données.

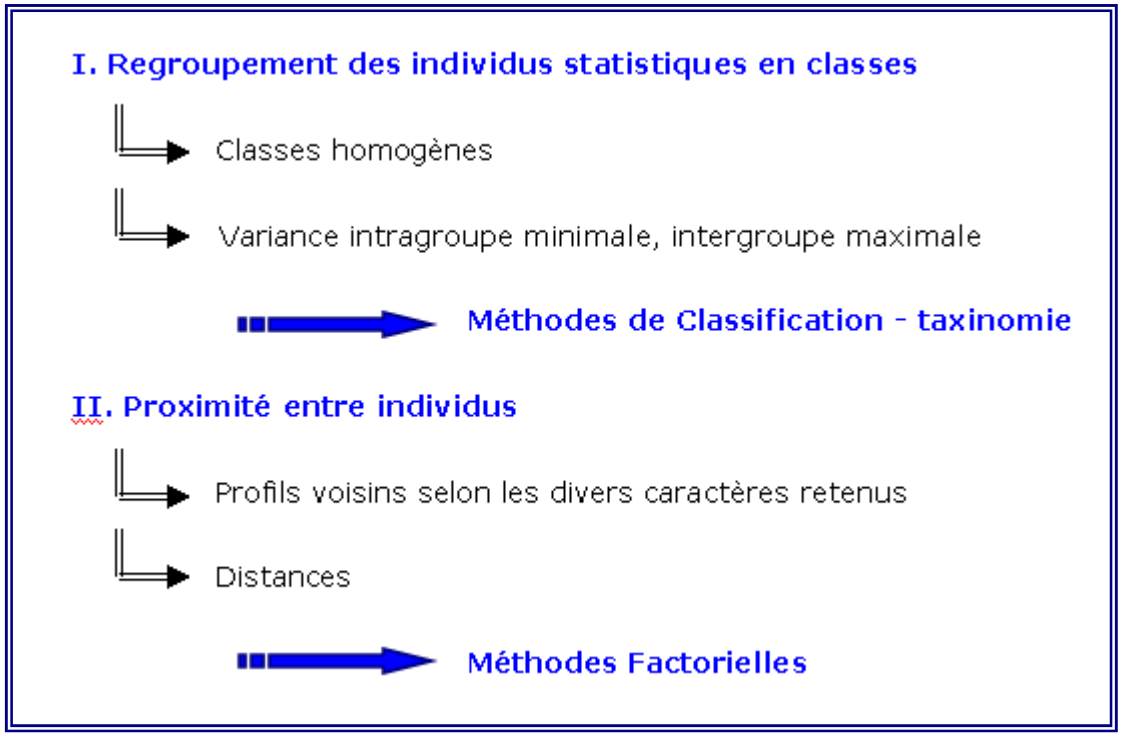
Il faut néanmoins souligner que ces méthodes reposent sur des calculs complexes d'algèbre linéaire. L'objectif de ce cours n'est pas de traiter en détail, les algorithmes qui permettent d'obtenir les résultats souhaités. Nous disposons désormais de nombreux logiciels qui prennent en charge cette vaste tache. Néanmoins, l'usage de ces logiciels comporte un risque, du fait de leur grande facilité d'emploi. En effet, en produisant rapidement des

[2] En fait de nos jours, le développement des logiciels statistiques nous permet de travailler sur un nombre très élevé d'individus. La notion d'individu est ici celle de la statistique. L'individu est un des éléments d'une population sur laquelle on observe un ensemble de caractères que l'on traduit par des variables. L'individu statistique est aussi bien une personne physique que morale, une unité territoriale, une entité administrative etc. Dans ce document, les individus étudiés porteront plus spécifiquement sur des unités territoriales et / ou collectivités territoriales telles les communes, les départements, les régions etc.

résultats, ces logiciels nous amènent à utiliser les méthodes en question de façon relativement mécanique, sans réfléchir suffisamment aux hypothèses de travail, au contexte général de l'étude et aux présupposés statistiques que chaque méthode sous-entend.

Comme le souligne très justement J.P. Georgin (2002), «*les logiciels spécialisés sont des boites noires : on entre les données à traiter, on sélectionne la méthode à utiliser et on obtient les résultats sans aucun regard sur les calculs intermédiaires. Aucune interactivité n'est donc possible…. Et il n'y a plus aucune interface entre la méthode et les résultats*»[3].

[3] Georgin, J.P. (2002) Analyse interactive des données (ACP, AFC) avec Excel 2000, Théorie et pratique, Société Francaise de Statistique, PUR, 267p.



➢ **les méthodes de classification** tentent de regrouper les individus en catégories – classes homogènes (cluster analysis etc), de façon à procéder à une typologie des individus étudiés. Ces opérations de découpage en classes d'une population sur laquelle on observe un ou plusieurs caractères (séries statistiques) sont basées sur le principe de la discrétisation à savoir rendre discontinue (s), une ou plusieurs séries mesurées sur une échelle continue de valeurs. Principalement, deux techniques seront abordées :

- la classification ascendante hiérarchique qui produit une suite de partitions en classes de plus en plus vastes è l'image des classifications bien connues des zoologistes en espèces, genres, familles, ordre etc.
- la classification non hiérarchique qui produit directement une partition en un nombre prédéterminé de classes.

➢ **les méthodes factorielles** tentent de déterminer les ressemblances ou les différences entre individus. Deux individus se ressemblent s'ils présentent un profil basé sur plusieurs caractères relativement voisin. Il s'agit donc de projeter le nuage de points sur un sous-espace, en perdant le moins d'information possible et de mettre en évidence les proximités entre individus. Cette projection sur un graphique permet de représenter graphiquement les proximités et éloignements entre individus. Trois techniques fondamentales seront abordées:

- l'analyse en composantes principales (plusieurs variables quantitatives),

- l'analyse des correspondances (deux variables qualitatives, représentées par un tableau de contingences)
- l'analyse des correspondances multiples (plus de deux variables qualitatives).

Les méthodes factorielles sont en fait utilisées de deux façons différentes :

- la première, la plus courante vise, pour une population statistique donnée, à réduire le nombre de variables initiales dont on dispose, en extrayant un nombre limité de variables composites (axes factoriels), tout en perdant le moins d'information initiale. Chaque facteur s'exprime sous la forme d'une combinaison linéaire des variables initiales. On admet donc que ces variables composites reflètent divers schémas de corrélation entre les variables initiales et que ces schémas de corrélation (limités en nombre) ne font perdre qu'un faible pourcentage de la variance totale du nuage.
- La deuxième vise à émettre des hypothèses relatives à des effets de causalité entre les diverses variables. Il s'agit de détecter la structure caractérisant les relations existant entre les variables initiales. Cette pratique est intéressante lorsque l'on souhaite procéder dans un second temps, à une analyse basée sur des modèles de régression linéaire.

Enfin, le cours se terminera par la présentation d'une méthode quelque peu spécifique de représentation et classification des données qui repose sur un traitement visuel des données, particulièrement efficace dans le cas de données géographiques, à savoir la **méthode Bertin**.

La logique de l'Analyse des données

Dans tous les cas, les méthodes appréhendées dans ce cours, ont pour objectif de conserver au mieux l'information contenue dans la ou les séries statistiques étudiées, tout en permettant une réduction du volume initial d'information de façon à obtenir la meilleure lisibilité possible. Ce principe de réduction de l'information et sa lisibilité est d'ailleurs primordial lorsque l'on souhaite procéder à un travail de cartographie des données.

La réduction du volume de données en quelques grandes dimensions doit cependant se faire avec une perte minimale d'information, ce qui est un compromis délicat qui exige que soient pris en compte, un certain nombre de paramètres :

- l'ordre de grandeur des phénomènes étudiés
- la forme des distributions
- leur dispersion
- l'existence éventuelle de cas particuliers, atypiques.

Une méthode d'analyse à la croisée de plusieurs sciences

L'analyse statistique unidimensionnelle et multidimensionnelle des données doit être considérée comme un outil contribuant largement à :

- l'analyse et la compréhension des phénomènes et comportements démographiques, sociaux, économiques etc, qui ne sont pas tous systématiquement quantitatifs, grâce à la production de ce que l'on pourrait qualifier de «méta-variables»
- les études prospectives qu'elles soient sectorielles ou territoriales,

Elle est également une étape préliminaire et incontournable de la cartographie et de la représentation visuelle des phénomènes et comportements.

Organisation et déroulement du cours

Chaque séance sera articulée de la façon suivante :

1ère Partie : (a) Rapide présentation théorique des principes et de la logique de la méthode envisagée ainsi que des éléments indispensables d'algorithme.

(b) Mise an application à l'aide de SPSS : exemples basés sur des données réelles, méthodes de lecture des résultats et interprétation

2ème Partie : Construction collective d'un dossier sur la base d'une thématique choisie par les étudiants : une fois le thème et l'objectif de recherche choisis, il s'agira de construire les variables appropriées puis finalement d'appliquer les méthodes.

Bibliographie de Base :

- Béguin M., Pumain D., (2003), La représentation des données géographiques, Statistique et cartographie, Armand Colin, Collection Cursus, 192 pages.
- Benzécri J.P. & F., (1984), Pratique de l'Analyse des Données, Dunod, 457 pages
- Bouroche J.M., (2002), L'analyse des données, PUF, Collection Que sais-je. No 1854, 8ème édition, 127 pages.
- Cibois P., (2000), L'analyse factorielle, 2000, PUF, Collection Que sais-je. N° 2095, 127 pages
- Dervin C., (1992), Comment interpréter les résultats d'une analyse factorielle des correspondances, Collection STAT-ITCF, 72 pages.
- Doise W., Clémence A., Lorenzi-Cioldi F., (1992), Représentations sociales et analyses de données, PUG, Grenoble, 264 pages.
- Dumolard P., Dubus N. Charleux L., (2005), Les statistiques en géographie, Edition Belin atouts Géographie, 240 pages.
- Fénelon J.P., (1999), Qu'est-ce que l'analyse de données?, Seisam, 311 p.
- Georgin J.P., (2002), Analyse interactive des données (ACP, AFC) avec Excel 2000. Théorie et pratique, Presses Universitaires de Rennes, Collection Didact Statistique, 266 pages.
- Groupe Chadule (1997), Initiation aux pratiques statistiques en géographie, Armand Colin, Collection U, 4ème édition, 203 pages
- Lebart L., Morineau A., Piron M., (2004), Statistique exploratoire multidimensionnelle, Dunod, 2ème édition, 439 pages.
- Sanders L., (1990), L'analyse des données appliquée à la géographie, Montpellier, Reclus, Alidade, 267 pages.
- Tomassone R., (1988), Comment interpréter les résultats d'une analyse factorielle discriminante, Collection STAT-ITCF, 56 pages