Université de Thessalie
Département d'Aménagement,
d' Urbanisme et de Développement Régional
Master Franco-hellénique
Population Développement Prospective (PODEPRO)

Semestre 2
U4: Techniques, méthodes et outils
Cours U4.1 : Analyse des données
Intervenant : Marie Noelle Duquenne
Durée : 25 heures

Objectif du cours :
L'important développement des bases de données a conduit à mettre au point de nombreuses méthodes pour synthétiser les informations volumineuses et repérer les grandes structures d'un vaste tableau de données quantitatives, mais aussi qualitatives. Le cours a donc 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.
Après un rapide rappel des principes essentiels de l'analyse statistique : théorie des probabilités et statistique descriptive, l'essentiel du cours a pour objectif 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 savoir :

- **les méthodes de classification**, qui tentent de regrouper les points (cluster analysis etc), de façon à procéder à une typologie des unités étudiées. Ces opérations de découpage en classes d'une ou plusieurs séries statistiques est basée 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
  - la classification non hiérarchique

- **les méthodes factorielles**, qui consistent à projeter le nuage de points sur un sous-espace, en perdant le moins d'information possible 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).

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 peut 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 du cours et méthode de travail

Chaque cours fera l'objet :

(i) d'une présentation théorique, concernant les principes, la logique et les algorithmes relatifs aux méthodes étudiées

Les étudiants intéressés par une présentation mathématique plus poussée, pourront se reporter aux documents qui seront fournis à chaque cours.

(ii) du traitement d'exemples concrets, basés sur des données réelles, l'accent étant mis sur les méthodes de lecture et interprétation des résultats

A la fin de chaque cours, un document de T.P. (Travaux Pratiques) sera distribué aux étudiants qui seront donc appelés à appliquer par eux-mêmes, les méthodes présentées auparavant. Cela signifie qu'un travail personnel assez conséquent est exigé de la part des étudiants. Ces derniers devront s'organiser sous forme de tandem.

Déroulement du cours

1		2h	Introduction Rappel de la théorie des probabilités, principes et lois. Principes de la réduction de l'information : - Variables et indices - Ordre de grandeur - Forme des distributions - Dispersion et concentration - Irrégularités des séries
2		2h	Les processus de discréditation La relation entre caractères (corrélation, etc) Les tests statistiques La notion de distance
3		2h	Les méthodes de classification hiérarchiques et non hiérarchiques à partir de l'étude de certains exemples classiques - Classification hiérarchique : fournir un ensemble de partitions plus ou moins fines obtenues par regroupements successifs de parties - Classification non hiérarchique : produire une partition en un nombre k fixé de classes
4		3h	Présentation et analyse des exemples de classification hiérarchiques et non hiérarchiques qu'auront étudié et préparé les groupes d'étudiants
5		3h	L'Analyse en Composantes Principales (ACP)
6		3h	Présentation et analyse de deux exemples d'ACP
7		3h	L'Analyse des Correspondances – Tableaux de contingence
8		3h	L'Analyse des Correspondances Multiples
9		3h	Présentation et analyse des exemples d'Analyse des Correspondances qu'auront étudié et préparé les groupes d'étudiants
10		2h	La Méthode Bertin

Si nécessaire, les 4 dernières séances pourront être portées de 2 à 3heures

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.
- 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

1ère Séance : Théorie des Probabilités

Si la notion de probabilité est assez ancienne, elle remonte en fait à plus de trois siècles et demi, ce n'est que dans les années 1930 que Kolmogorov formalise les fondements du calcul des probabilités et en en fait une construction axiomatique cohérente.

1. Notions de probabilités

Il existe deux manières essentielles de définir une probabilité : (a) *probabilités inductives ou expérimentales* et (b) *probabilités déductives ou théoriques*.

(a) **Probabilité expérimentale ou inductive**: la probabilité est déduite de toute la population concernée par expérimentation.

> Par exemple, si l'on observe le nombre de naissances dans un pays durant trois années et que l'on constate que parmi les 10.000 naissances, 5.150 sont des garçons et 4.850 des filles, on en déduit que P[garçon] = 0.515 (51,5%). Cette probabilité a été évaluée ***à posteriori*** (ex-post).

(b) **Probabilité théorique ou déductive**: cette probabilité est connue grâce à l'étude du phénomène sous-jacent sans expérimentation. Il s'agit donc d'une connaissance ***a priori*** (ex-ante) par opposition à la définition précédente.

> Par exemple, dans le cas classique du dé parfait, on peut dire, sans avoir à jeter un dé, que P["obtenir un 4"] = 1/6.

Comme il n'est pas toujours possible de déterminer des probabilités *a priori*, on est souvent amené à réaliser des expériences. Il faut donc pouvoir passer de la première à la deuxième solution. Ce passage est supposé possible en terme de limite (*i.e.* avec une population dont la taille tend vers la taille de la population réelle).

C'est bien parce qu'il est relativement peu courant de connaître à priori les probabilités de réalisation d'évènements que la notion de probabilité est liée à la notion d'expérience ou de mesure. De plus, on peut toujours affecter au résultat de l'expérience, une valeur numérique, ce qui nous permettra ainsi de définir une variable dite aléatoire, correspondant précisément à la mesure numérique de l'expérience.

> Par exemple, dans le cas du jeté d'une pièce de monnaie, on peut toujours décider que Pile = 1 et Face = 0. Le résultat de l'expérience ne peut être que 1 ou 0.

Cependant la variable a un caractère **aléatoire** dans la mesure ou la répétition de l'expérience telle que lancer la pièce de le même façon, plusieurs fois de suite (reproduction à l'identique) ne donne pas toujours le même résultat. Le résultat est aléatoire car il est incertain : on ne peut savoir avec certitude quel sera le résultat de l'expérience, c'est que l'on appelle communément le **hasard**. On entend par hasard, l'ensemble des causes qui font que le résultat n'est pas prévisible et provoquent finalement une dispersion des résultats.

Le but finalement de la théorie des probabilités n'est pas de tenter de décrire selon un processus déterministe les causes de la variabilité des résultats mais d'en prendre acte et de **fournir un cadre alternatif de quantification des résultats d'un très grand nombre d'expériences, en donnant à chaque issue possible de l'expérimentation une mesure, sa probabilité**. Cette dernière va dépendre de la façon dont se déroule l'expérimentation laquelle conditionne finalement le nombre total d'issues qui peuvent découler de l'expérimentation.

> Par exemple, dans une urne, nous avons 3 billets de banque: 1 billet de 5€, 1 billet de 10€ et 1 billet de 20€.. Nous décidions d'en tirer deux au hasard. Le nombre total d'issues de l'expérimentation dépend du mode de tirage.
> **Tirage avec remise** : 9 issues possibles

(5,5), (5,10), (5,20), (10,5), (10,10), (10,20), (20,5), (20,10), (20,20)
Tirage sans remise : 6 issues possibles
(5,10), (5,20), (10,5), (10,20), (20,5), (20,10)
La probabilité d'obtenir un et un seul billet de 20€ est égale à 4/9 dans le 1er cas et à 4/6 dans le second cas

La différence de résultat constaté dans l'exemple précédent est due au fait que le nombre de combinaisons possibles découlant de l'expérimentation diverge. Cela nous amène en toute logique à nous pencher sur la notion d'analyse combinatoire.

2. Principales règles de l'analyse combinatoire

2.1. Factorielle

Si une action peut être obtenue de n1 façons différentes, puis suivant cette action, de n2 façons différentes indépendantes des précédentes, puis de n3 façons différentes etc…, alors, le nombre de possibilités correspondant à l'ensemble de ces actions est égal à: $n! = \prod_1^k n_i$

On appelle **factorielle n** et l'on note n! le nombre : $n! = \prod_1^n i$ (4 ! = 4.3.2.1)

2.2. Arrangements de p objets parmi n

Nombre de possibilités de ranger p objets choisis parmi n et l'ordre a une signification :

$$A_n^p = \frac{n!}{(n-p)!} = n(n-1)...(n-p+1)$$

Il y a 6 arrangements possibles de 2 symboles parmi 3 différents A, B, C
(A,B) , (A,C) , (B,A) , (B,C) , (C, A) , (C,B)
Les arrangements (A,B) et (B,A) sont différents car l'ordre a une signification.

2.3. Permutations de n objets

Arrangement de n objets parmi n en tenant compte de l'ordre : $P_n = A_n^n = n!$

Ainsi, il y a 6 = 3! permutations possibles de 3 symboles différents A, B, C:
(A,B,C) , (A,C,B) , (B,A,C) , (B,C,A) , (C,A,B) , (C,B,A)

2.4. Combinaisons de p parmi n objets

On ne tient pas compte de l'ordre des objets dans le rangement : $C_n^p = \frac{n!}{p!(n-p)!}$

La notation anglo-saxonne pour les combinaisons est un peu différente : $\binom{p}{n}$

Il y a 3 combinaisons possibles de 2 symboles parmi 3 différents A, B, C :
(A,B) , (A,C) , (B,C)
La combinaison (A,B) est équivalente à la combinaison (B,A) car l'ordre n'a pas de signification.

Propriétés :

- $C_n^0 = C_n^n = 1$
- $C_n^p = C_n^{n-p}$
- $C_n^p = C_{n-1}^{p-1} + C_{n-1}^p$
- $\sum_{p=1}^{n} C_n^p = 2^n$

3. Epreuves et Evènements

Une **expérience** est dite **aléatoire** si ses résultats ne sont pas prévisibles avec certitude en fonction des conditions initiales.

On appelle **épreuve** la réalisation d'une expérience aléatoire.

On appelle **évènement** la propriété du système qui une fois l'épreuve effectuée est ou n'est pas réalisée.

Soit l'expérience aléatoire "lancer deux dés discernables" (et non pipés si l'on veut vraiment une expérience aléatoire) et l'évènement A = "obtenir un total de nombres supérieur à 10".
L'évènement A se réalise pour les épreuves (6,5), (5,6), (6,6).

Correspondance entre les opérateurs logiques et les ensembles.

Logique	Ensemble
évènement certain	espace entier Ω
état du système	élément $\omega \in \Omega$
évènement A	partie $\{A\} \subset \Omega$
évènement impossible	partie vide $\varnothing$
évènement contraire	partie complémentaire $\{\overline{A}\}$
l'évènement B entraîne l'évènement A	$\{B\} \subset \{A\}$
A et B	intersection $\{A\} \cap \{B\}$
évènements incompatibles	parties disjointes $\{A\} \cap \{B\} = \varnothing$
A ou B (ou non exclusif)	réunion $\{A\} \cup \{B\}$
A ou B exclusif	somme $\{A\} + \{B\} = (\{A\}) \cup \{B\}) - (\{A\} \cap \{B\})$

A partir de ces notions, on peut préciser le calcul de probabilités d'un évènement A:

Probabilité théorique: $P(A) = \dfrac{nombre\ de\ cas\ \ favorables}{nombre\ total\ \ de\ cas}$.

Probabilité expérimentale:. $P(A) = \dfrac{nombre\ d'epreuves\ qui\ realisent\ A}{nombre\ total\ d'epreuves}$

Cette 2ème approche est aussi appelée approche *fréquentiste. Elle ne* permet pas de donner une valeur ni même un sens à la probabilité d'un évènement non répétable du genre "neigera-t-il le 25 octobre 2990" ce qui limite de fait le champ d'application du calcul des probabilités.

Pour les fréquentistes, seules ont un sens les probabilités calculées à *posteriori* sur la base de la répétition d'un grand nombre d'évènements identiques; pour les subjectivistes, au contraire, la notion de probabilité *a priori*, évaluable en fonction d'un sentiment individuel d'incertitude, peut avoir un sens.

4. Espace probabilisé

4.1. Axiomatique de Kolmogorov

A chaque événement de l'ensemble Ω, on associe un nombre positif compris entre 0 et 1, sa probabilité.

Définition 1

On appelle probabilité sur (Ω, S) où Ω est l'ensemble des évènements et S une classe de parties de Ω), ou loi de probabilité, une application P *de S dans [0,1] telle que:*

- $P(\Omega) = 1$

Pour tout ensemble dénombrable d'évènements incompatibles $A_1, A_2, \ldots A_n$ *on a :*

- $P(\bigcup A_i) = \sum P(A_i)$

Définition 2

L' espace probabilisé est formé par le triplé (Ω,S ,P)

Une loi de probabilité n'est donc rien d'autre qu'une **mesure positive de masse totale 1**. On peut donc relier la théorie des probabilités à celle de la mesure.

4.2. Propriétés élémentaires

De l'axiomatique de Kolmogorov, on peut déduire les propriétés suivantes:

Propriété 1 : P(Ø) = 0

Propriété 2 : $P(\overline{A}) = 1 - P(A)$

Propriété 3 : $P(A) \leq P(B)\ si\ A \subset B$

Propriété 4 : $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Propriété 5 : $P(\cup A_i) \leq \sum_i P(A_i)$: Il n'y a stricte égalité que si tous les évènements A_i sont deux à deux incompatibles.

Propriété 6 : Continuité monotone séquentielle. Soit $A_1 \supset A_2 \supset A_3 \supset ... \supset A_n \supset \emptyset$

Si $\lim_{n\to\infty} A_n = \emptyset$ *alors* $\lim_{n\to\infty} P(A_n) = 0$

Propriété 7 : ***Théorème des probabilités totales:*** Soit $\Omega = \cup B_i$ un système complet d'évènements tel que $\{B_i\}$constitue une partition de Ω, alors :

$$\boxed{\forall A : P(A) = \sum_i P(A \cap B_i)}.$$

5. Probabilité conditionnelle - Théorème de Bayes

5.1. Théorème des probabilités composées

Soit deux évènements A et B réalisés respectivement n et m fois au cours de N épreuves. On a donc :

P(A) = n/N
P(B) = m/N.

Si de plus A et B sont réalisés simultanément k fois, on a $P(A \cap B) = k / N$.

Que peut-on déduire sur la probabilité de l'évènement B sachant que l'évènement A est réalisé?

Cette probabilité est appelée **probabilité conditionnelle de B sachant A** et se note P(B/A). Dans ce cas, P(B/A) = k/n

Par définition, on a : $\boxed{P(B/A) = \frac{P(A \cap B)}{P(A)} \; et \; P(A/B) = \frac{P(A \cap B)}{P(B)}}$

Conséquences
Deux évènements A et B sont dits indépendants si $P(A \cap B) = P(A).P(B)$ ou encore si P(B/A) = P(B) (l'information sur la réalisation de A n'apporte rien à l'évènement B) et de même P(A/B) = P(A).

5.2. Théorème de Bayes - Probabilités des causes

Soit un évènement A qui peut dépendre de N causes différentes, notées C_i et incompatibles deux à deux (on ne peut avoir deux causes réalisées simultanément). Etant donnée la réalisation de l'évènement A, quelle est la probabilité que ce soit l'évènement C_i qui en soit la cause ?

On cherche donc P(C_i/A)

Puisque toutes les causes sont incompatibles deux à deux et toutes les causes possibles à A sont supposées connues, on a alors :

- d'après le théorème des probabilités totales : $P(A) = \sum_i P(A \cap C_i)$

- puis en appliquant le théorème des probabilités conditionnelles, on a:

$$P(A \cap C_i) = P(C_i).P(A/C_i)$$

Alors $$\boxed{P(C_i / A) = \frac{P(C_i \cap A)}{P(A)} = \frac{P(C_i).P(A/C_i)}{\sum_i P(C_i).P(A/C_i)}}$$

Exemple: Deux machines M_1 et M_2 produisent respectivement 100 et 200 objets. M_1 produit 5% de pièces défectueuses et M_2 en produit 6%. Quelle est la probabilité pour qu'un objet défectueux ait été fabriqué par la machine M_1? En d'autres termes quelle est la probabilité que ce soit la machine 1 qui soit à l'origine de la pièce défectueuse?

L'évènement constaté A = «pièce défectueuse»

On cherche finalement $P(M_1/A)$

les causes sont les machines M_1 et M_2 :Compte tenu des productions de ces machines, on a : $P(M_1) = 1/3$ et $P(M_2) = 2/3$ (deux causes indépendantes et incompatibles).

Les probabilités conditionnelles sont donc $P(A/ M_1) = 0,05$ et $P(A/ M_2) = 0,06$