Formations de l'Urfist de Strasbourg

URFIST Strasbourg
Dir. Noël Thiboud
34, Boulevard de la Victoire
67070 Strasbourg
Tél : 03 68 85 08 00

UNIVERSITÉ DE STRASBOURG

Traitements de données d'enquêtes et initiation au logiciel SPAD

Jean-Paul Villette, Maître de Conférences

Traiter des données d'enquêtes . Produire un tableau de données .Organiser l'intelligibilité des données. Typologies : qu'est-ce que les individus d'un cluster ont en commun, qui les différencie des autres : les réponses clivantes.

Mercredi 2 novembre et lundi 7 novembre 2016, salle 312, 3ème étage de la Fac de Sciences Economiques et de Gestion, PEGE, 61 avenue de la Forêt Noire, Strasbourg
9h-12h et 13h30-16h30

Produire un tableau de données (numériques, qualitatives, textuelles) et premiers traitements

Penser, mesurer et représenter la nature et l'intensité des liens entre des variables

Typologies : la production de clusters intra-homogènes extra-hétérogènes. Les réponses clivantes.

Cas réels susceptibles d'être évoqués

COM

Jean-Paul VILLETTE :« ***les résultats obtenus grâce à l'analyse des données de l'enquête*** » communication au colloque« *Enquête sur un électorat en rupture ; le vote pour l'extrême-droite en Alsace* » Colloque du GREDA, 28-29 février 2008 , MISHA, Strasbourg.

Bernard AUBRY, Jean-Alain HERAUD, Jean-Paul VILLETTE « ***40 années d'évolutions de la structure socioprofessionnelle des cantons alsaciens »*** communication à la journée d'étude « *Les relations complexes : Habitat, Mobilité, Economie en Alsace* » de l'Association Prospective Rhénane (APR) 21 novembre 2008, Université de Strasbourg.

Laure PAIRET, Pascal POLITANSKY, Jean-Paul VILLETTE « ***Représentations et pratiques enseignantes à travers la question des représentations de la violence scolaire » communication à la journée d'étude*** : « Représentations et pratiques enseignantes » Université d'Artois. Décembre 2009

ACLN

« ***Gestion patrimoniale des réseaux d'eau potable et d'assainissement. Elaboration de politiques de renouvellement à partir des données d'inventaires départementaux : l'expérience du Bas-Rhin*** » Caty WEREY, Myriam CAMPARDON, Jean-Paul VILLETTE, Isabelle MELLAC-BECK in « Techniques Sciences Méthodes- La revue mensuelle des spécialistes de l'environnement . octobre 2009

Actes de conférences internationales à comité de lecture

CABASSUT R, VILLETTE J-P (2011) "*Exploratory data analysis of an european teacher training course on modelling*", in *Proceedings of 7ème Cerme* (Congress of European society for Research in Mathematics Education) Rseszow University.Poland

Actes de colloques nationaux à comité scientifique

Richard CABASSUT, Jean-Paul VILLETTE (2010) ***» Evaluation en formation de professeurs sur l'enseignement de la modélisation »***, in Actes du 37 eme colloque Copirelem. La Grande Motte

Mickael BENHAIM,., Jean-Alain HERAUD., Valérie MERINDOL., Jean-Paul VILLETTE : ***« La connectivité scientifique locale-globale des régions européennes : approche, mesures et incidences »***, Conférence Eurolio, St Etienne, 28-29 janvier 2012

FLEURY D,PEYTAVIN J-F, GODILLON S., SAINT-GERAND T, MEDJKANE M, PROPECK E., KAHN R. VILLETTE J-P " *l'approche territoriale de l'insécurité routière dans l'aménagement régional*' Association de Science régionale de Langue Française.2012

Participation à des contrats

Jean-Paul VILLETTE : « *Laboratoires et sources de financement de 2001 à 2007 : études typologiques* », Annexe1 au rapport final, pour l'ANR « *les financements de la recherche des laboratoires reconnus : évolutions récentes* » sous la direction du Professeur Patrick LLerena, BETA, Strasbourg, 2009

Jean-Paul VILLETTE *et* Laure PAIRET « *Enquête auprès de 73 entreprises. Régularités. Caractéristiques d'ensemble et réponses clivantes. »* Contribution au rapport du Fraunhofer - Institut für System und Innovationsforschung –ISI (Karlsruhe) à la GIZ (Deutsche Gesellschaft für Internationale Zusammenarbeit). Juillet 2012

René KAHN et Jean-Paul VILLETTE *« 690 550 déplacements, 477 IRIS, 2777 trajets, 5 motifs et 6 modèles de déplacements, 11 642 accidents survenus dans la LMCU (Lille Métropole Communauté Urbaine). Etudes typologiques ».* Contrat BETA- Laboratoire GEOSYSCOM-UMR IDEES 6266 du CNRS

Contrats

Jean-Paul VILLETTE *« 71 pôles de compétitivité .Etudes typologiques. Clusters de pôles et indicateurs clivants*. » contrat ADVENCIA-NEGOCIA / BETA. 2010.

Jean-Paul VILLETTE « *La différenciation des perceptions des besoins pour innover, formes d'innovations, capacités et coopérations pour innover. Le cas de 206 entreprises. Typologies d'entreprises et variables clivantes* ». contrat STRASBOURG-CONSEIL/ BETA. 2010.

Jean-Paul VILLETTE *« Typologies des 904 communes d'Alsace et indicateurs socio-économiques clivants*. **»** contrat BETA- APR (Association pour la Prospective Rhénane). Novembre 2010.

Jean-Paul VILLETTE *« 71 pôles de compétitivité .Etudes typologiques des évolutions récentes. Clusters de pôles et indicateurs clivants*. » contrat ADVENCIA-NEGOCIA / BETA. Janvier 2011.

Jean-Paul VILLETTE « *Ce que disent 1341 étudiants de l'Université de Limoges.180 questions, 200 000 réponses. Etudes typologiques »*». contrat STRASBOURG-CONSEIL/ BETA. Mai 2013.

Laure PAIRET, Pascal POLITANSKI, Jean-Paul VILLETTE « ***Enquête sur l'emploi transfontalier entre l'Alsace et le Bade-Wurtemberg. 1753 questionnaires. 80 questions.140 000 réponses. Etudes typologiques. Réponses clivantes »***». contrat STRASBOURG-CONSEIL/ BETA pour le Land Baden-Wüttemberg et le DFI (Deutsch-Französisches Institut). Janvier 2014

Jean-Paul VILLETTE « ***POI : les Parcours Orientation-Insertion de la région Midi-Pyrénées. 5276 individus. Etudes typologiques. »***». contrat STRASBOURG-CONSEIL/ BETA. Avril 2014.

Jean-Paul VILLETTE « ***Evaluation du dispositif « Nouvelle Chance pour l'Alternance de la région Aquitaine.61 306 individus. Etudes typologiques »***». contrat STRASBOURG-CONSEIL/ BETA. Avril 2014.

Jean-Paul VILLETTE « ***PDIP : les Parcours diplômants de la région Midi-Pyrénées. 3805 individus. Etudes typologiques. »***». contrat STRASBOURG-CONSEIL/ BETA. juin2014.

Jean-Paul VILLETTE « ***Ce que disent 2184 personnes du campus de la Doua à Lyon. Etudes typologiques. Réponses clivantes. 2184 questionnaires – 235 questions – 400 000 réponses. »***». Contrat STRASBOURG-CONSEIL/ BETA. juin2014.

Jean-Paul VILLETTE « ***PDIP : les Parcours diplômants de la région Midi-Pyrénées. 3805 individus. Etudes typologiques. »***». Contrat STRASBOURG-CONSEIL/ BETA. juin2014.

Laure PAIRET, Jean-Paul VILLETTE «*1223 cabinets d'Architectes, d'Avocats et d'Experts-Comptables. Etudes typologiques. Analyses de données textuelles*» contribution au rapport final « ***Etude relative au positionnement international des professions libérales françaises*** » de Strasbourg-Conseil au Ministère de l'Economie, de l'Industrie et du Numérique, Direction Générale des Entreprises, décembre 2015

Autres études :

Jean-Paul VILLETTE : rapport à Jean-Alain HERAUD, Doyen de la faculté des Sciences Economiques : » ***1733 étudiants de la Faculté de Sciences Economiques et de Gestion, année 2008-2009. Caractéristiques clivantes pour des indicateurs de résultats. Etudes typologiques. sans commentaires »*** juillet 2010.

Philippe BRETON, Laure PAIRET, Jean-Paul VILLETTE rapport au Président Alain BERETZ ***« Le fonctionnement des conseils vu par eux-mêmes . Synthèse et analyse des résultats du questionnaire rempli par les élus*** *(CA,CS,CEVU,CTP) en mai-juin 2010 ».* mission « Vie démocratique à l'UdS » septembre 2010.

Exercices

FSEG STRASBOURG
Faculté des sciences économiques et de gestion

UNIVERSITÉ DE STRASBOURG

Année universitaire 2015/2016
Master 1ère année
Mention Management des Projets et des Organisations
Semestre 1 – Session 1 / Contrôle terminal janvier 2016
Méthodes Quantitatives d'aide à la décision. Analyse des données
Jean-Paul Villette
Durée : 1h30 - Aucun document autorisé -Calculatrice autorisée
La note 20/20 correspondra à une partie du sujet. Les bonnes réponses sont très brèves.

1 – La provenance du dessert lors de la réception d'invités

Dans un questionnaire[1] on trouve :

Question: En général, lorsque vous recevez des invités, le dessert, ... ?

- Vous le faites vous-même :
- Vous l'achetez :
- Vous le faites en partie, avec des aides culinaires :
- Vos invités l'apportent :
- Ne se prononcent pas :

Est-ce que les réponses sont exhaustives ?exclusives? Significations concrètes

2-découpage d'une variable numériques en classes

cas réel. On a demandé à 105 passants de noter de 0 à 10 une affiche de publicité. On a les résultats :

note	%
0	0
1	5
2	1
3	4
4	7
5	12
6	13
7	15
8	24
9	11
10	8
	100

Proposez un découpage en classes

3 – tableau :

Dans un tableau de données provenant d'une enquête, que représente une ligne ? une colonne ? une cellule à l'intersection d'une ligne et d'une colonne ?

[1] Source site IFOP, année ?, Base: aux personnes déclarant terminer leur repas par une note sucrée,

4- Clusters

Revoici :

Source : « Capital « fév 94, « *dix quotidiens régionaux au banc d'essai* ». Les individus sont les quotidiens : **Ouest-France, Sud-Ouest,la Voix du Nord**… et les variables sont : *le nombre de pages, le pourcentage de pages régionale et locale, le % de pages nationale et internationale, le % de pages sportive, le % de pages économiques, le prix de vente, le taux de pénétration et le nombre de citations du maire.*

	nbre de pages	% pages régionales	% pages nat. et intern.	% pages sportive	% pages économique	prix de vente F	taux de pénétration[2]	citations du maire[3]
Ouest-France	42	33	18	12	2,3	4	40	4
Sud-Ouest	30	26	20	23	2,3	4	39	4
……..	……..	……..	……..	……..	……..	……..	……..	……..

Une typologie à 3 clusters résulte d'une Analyse en Composantes Principales suivie d'une Classification Hiérarchique .

Très brève interprétation du cluster

{ les DNA, La Voix du Nord, Ouest-France}	**Moyenne dans le cluster**	**Moyenne générale**
% pages économique	**4,1**	**2,2**
taux de pénétration	**42,7**	**35,6**
nbre de pages	**41,7**	**35,3**
% pages régionales	**28,0**	**27,4**
prix de vente F	4,1	4,1
% pages nat. et intern.	13,3	14,5
citations du maire	2,7	4,4
% pages sportive	12,0	15,6

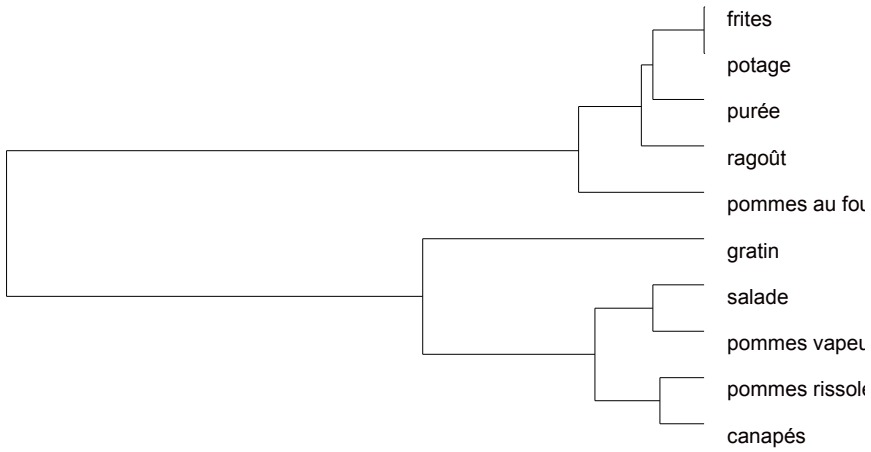
[2] *pourcentage de foyers qui lisent le quotidien dans la zone de diffusion*
[3] *nombre d'articles consacrés au maire de la ville siège du journal durant la période de l'enquête*

5- « pommes de terre : plats et variétés », une Classification Hiérarchique Ascendante

Il y a « 1 » lorsque la variété (en ligne) est recommandée pour le plat (en colonne).

	canapés	potage	salade	pommes vapeur	gratin	pommes au four	frites	pommes rissolées	ragoût	purée
Belle de Fontenay	1	0	1	1	0	0	0	1	0	0
BF15	1	0	0	1	1	0	0	1	0	0
Charlotte	1	0	1	1	0	0	0	1	0	0
Francine	0	0	1	1	1	0	0	0	0	0
Nicola	1	0	0	1	1	0	0	0	0	0
Pompadour	1	0	1	1	0	0	0	1	0	0
Ratte	1	0	1	1	0	0	0	1	0	0
Roseval	0	0	1	1	1	0	0	0	0	0
Rosine	0	0	0	0	1	0	0	0	0	0
Bintje	0	1	0	0	0	1	1	0	1	1
Estima	0	1	0	0	0	1	1	0	1	1
Manon	0	1	0	0	0	0	1	1	1	0
Monalisa	0	0	0	0	1	1	0	0	1	0
Samba	0	0	0	1	1	1	0	0	0	0

Classification hiérarchique directe



a- quelle est la règle statistique pour couper les branches

b- combien voyez-vous de groupes ? les décrire

c- quel est le plat le plus atypique ?

6–classification hiérarchique ascendante (CHA): à table !

Les arguments de proximité spatiale liée à des fréquences d'associations de modalités, rencontrés en Analyse Factorielles des Correspondances Multiples, se rencontrent aussi dans la conception de tableaux de bords par exemple : des contacteurs fréquemment utilisés conjointement (lave-glace et essuie-glace ..) doivent être proches. De même nous rangeons les fourchettes à côté des cuillères .Dans l'exercice suivant, j'emmène ma nièce au restaurant. Dans le tableau T, les **opérations** (manger un plat) identifiées par le nom du plat sont en ligne et les **outils** sont en colonne. On met 1 dans le tableau lorsque l'outil colonne est utilisé pour manger le plat-ligne, 0 sinon.

-

On considère le tableau

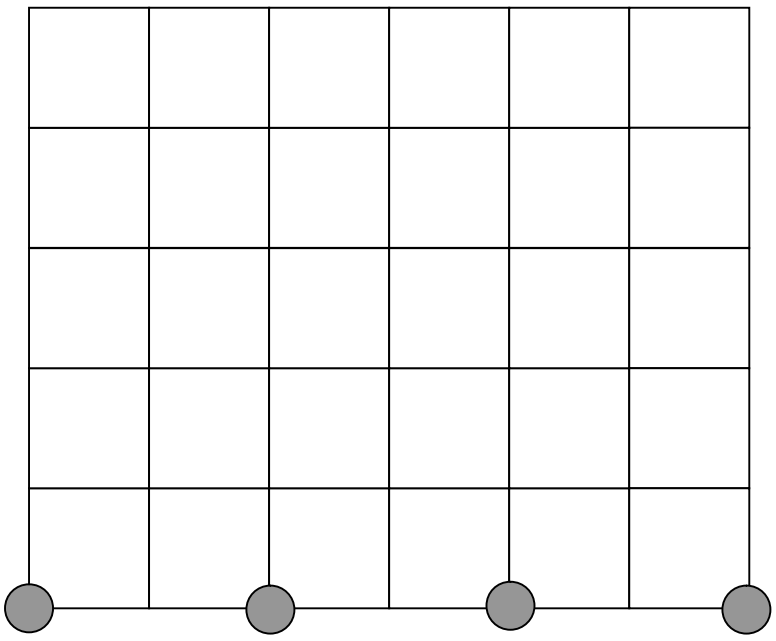
	fourchette	couteau	cuillère	doigts
frites	1	0	0	1
jambon	1	1	0	0
salade	1	0	0	0
Crème au chocolat	0	0	1	1

Voici un tableau de distances entre des « plats » mesurées par le nombre d' »outils » différents nécessaires.

distances	**{frites}**	**{jambon}**	**{salade}**	**{Crème au chocolat}**
{frites}	0	3	2	3
{jambon}		0	1	4
{salade}			0	3
{Crème au chocolat}				0

Calculer et représenter le *dendrogramme* de la Classification Hiérarchique Ascendante en utilisant la stratégie d'agrégation entre « paquets » dite « de la moyenne « :

$$d_{moy}(A,B)=\frac{\sum_{x\in A, y\in B} d(x,y)}{card(A).card(B)}$$



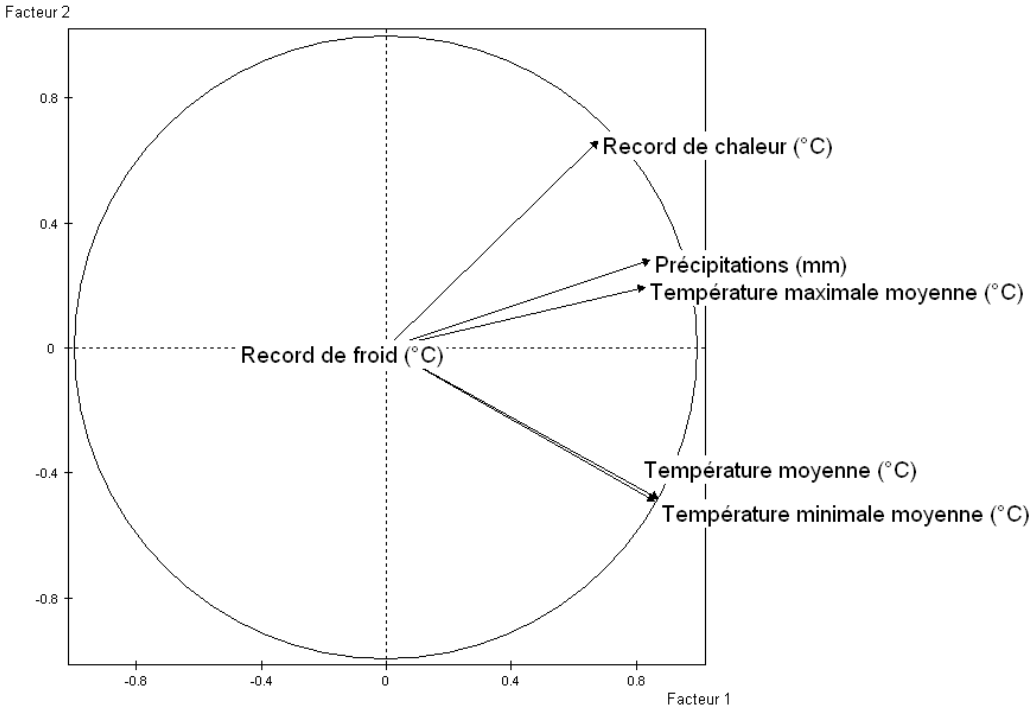
7– Données climatiques à Moscou :Analyse en Composantes Principales et Classification Hiérarchique Ascendante

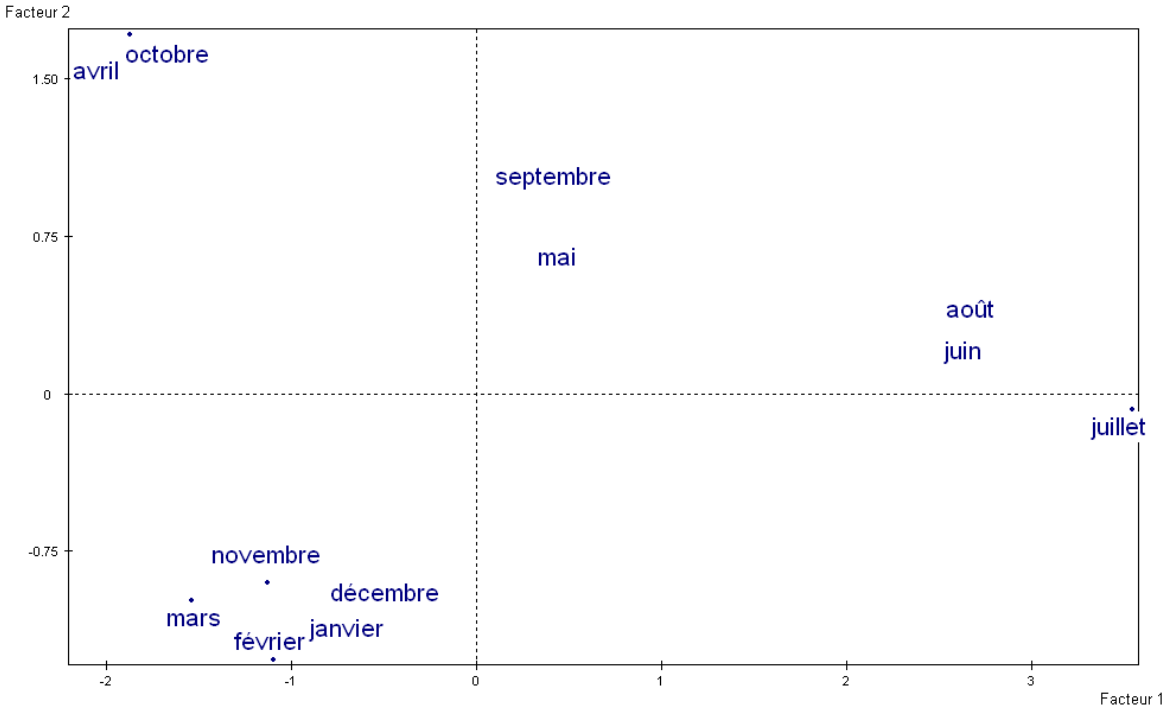
Voici le tableau des 12 mois (les individus) de l'année et de 6 variables numériques

Mois	Température minimale moyenne (°C)	Température moyenne (°C)	Température maximale moyenne (°C)	Record de froid (°C)	Record de chaleur (°C)	Précipitations (mm)
janvier	−9,1	−6,5	−4	−42,2	8,6	46
février	−9,7	−6,7	−3,7	−38,2	8,3	36
mars	−4,4	−1	2,6	−32,4	19,2	33
avril	2,2	6,7	11,3	−21	28,9	38
mai	7,7	13,2	18,6	−7,5	33,2	52
juin	12,1	17	22	−2,3	34,7	84
juillet	14,4	19,2	24,2	1,3	38,2	90
août	12,5	17	21,9	−1,2	37,3	80
septembre	7,4	11,3	15,7	−8,5	32,3	67
octobre	2,7	5,6	8,7	−16,1	24	66
novembre	−3,3	−1,2	0,9	−32,8	14,5	60
décembre	−7,6	−5,2	−3	−38,8	9,6	53

Analyse en Composantes Principales *(voir page suivante)*

que peut-on dire du moins d'avril**?**

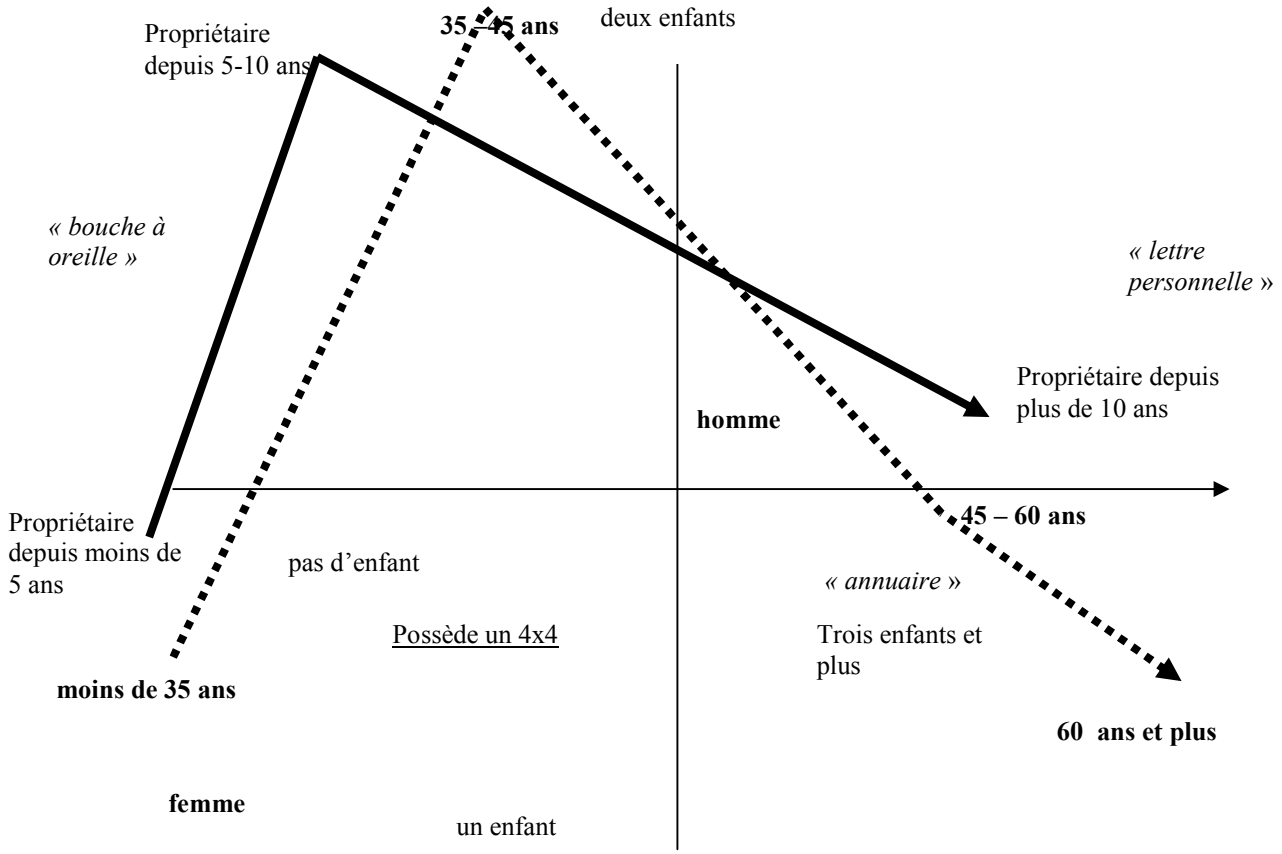

Facteur 2
0.8
0.4
0
-0.4
-0.8
Record de chaleur (°C)
Précipitations (mm)
Température maximale moyenne (°C)
Record de froid (°C)
Température moyenne (°C)
Température minimale moyenne (°C)
-0.8
-0.4
0
0.4
0.8
Facteur 1


Facteur 2
1.50
0.75
0
-0.75
avril
octobre
septembre
mai
août
juin
juillet
novembre
décembre
mars
janvier
février
-2
-1
0
1
2
3
Facteur 1

8– communication, une Analyse Factorielle des Correspondances Multiples

D'après un cas réel. Les individus sont des clients. Sont représentées ici quelques variables : **propriétaire** (*depuis moins de 5 ans, depuis 5-10 ans, depuis plus de 10 ans*), **enfants** (*pas d'enfant, un enfant, deux enfants, trois enfants et plus*), **l'âge** (*moins de 35 ans, de 35 à 45 ans, de 45 à 60 ans, 60 ans et plus*), **Le client a connu l'entreprise(**par le bouche à oreille, dans l'annuaire (amis, parents, voisins), par lettre personnelle), **possède un 4x4** (oui, non), et

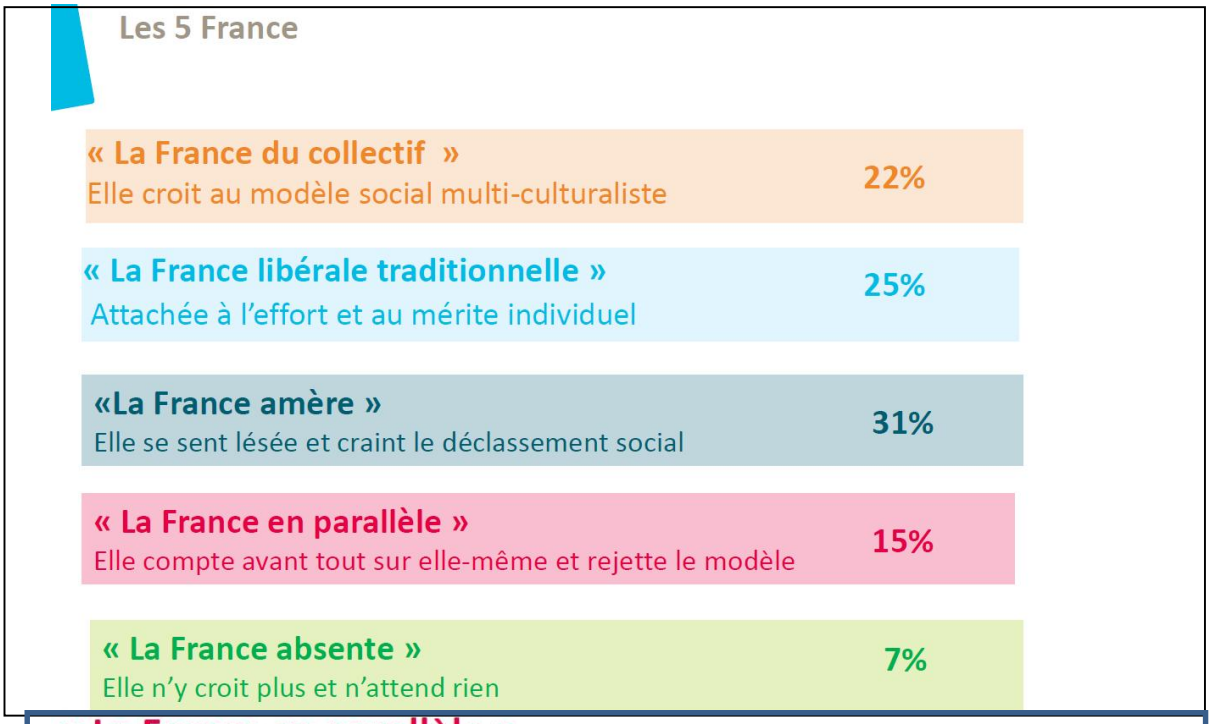
et **sexe de celui qui répond au questionnaire** (*homme, femme*),

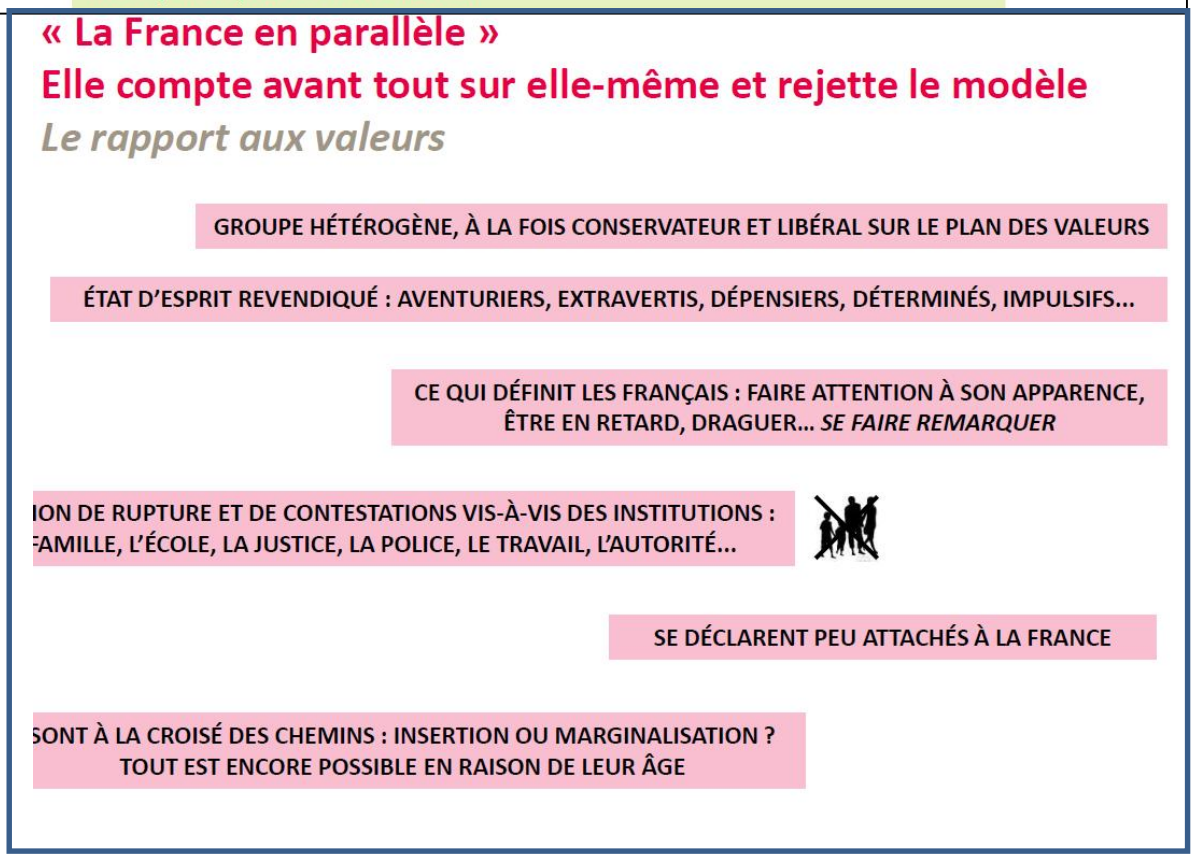


Que peut-on dire de « femme » ?

9 – question de méthodes

Ateliers CSA-Le Monde 3 dec 2013 « **Français, ce qui vous rassemble est-il plus fort que ce qui vous divise ?** »Extraits :





A- quelles méthodes statistiques ont-elles été utilisées pour obtenir ces cinq clusters ?
B***- Que faut-il comprendre, statistiquement, à :« se déclarent peu attachés à la France » » ?

Quelques points de repère : ACP [4]

- c'est une méthode de statistique exploratoire qui permet de déceler d'éventuelles régularités statistiques dans les co-relations des différentes variables quantitatives (numériques).
- Les liens entre les variables quantitatives sont pensés et mesurés en termes de **corrélations linéaires**. Un énoncé tel que : «les précipitations (la pluie) en mm et le nombre de jours d'orages sont positivement (linéairement) corrélés » doit se comprendre ainsi : les villes où les précipitations sont supérieures à la moyenne (58 mm) sont aussi les villes où le nombre de jours d'orages est supérieur au nombre moyen (4 jours). l'ACP est une méthode différentielle : elle met en évidence des écarts à la moyenne.

- Les variables quantitatives sont représentées par des vecteurs, des flèches, d'une couleur. Lorsque les variables sont centrées-réduites , la corrélation est alors le cosinus de l'angle formé par deux variables-vecteurs. **La corrélation se lit sur l'angle** :
    - Corrélation fortement positive <=> cosinus proche de 1 <=> angle aigu
    - Corrélation nulle, « indépendance » <=> cosinus proche de 0 <=> angle droit
    - Corrélation fortement négative <=> cosinus proche de –1 <=> angle obtus

- Les variables sont donc, géométriquement, des vecteurs flèches dans un espace de grande dimension. Ces vecteurs sont projetés sur un plan optimal en un certain sens : la figure géométrique initiale est la moins déformée possible, mais elle l'est .
- La qualité de la représentation d'une variable , ça se voit : dans ACP normée (ie avec centrage-réduction des variables) ce qui est le cas dès lors que l'on considère des corrélations et non pas des covariances, chaque flèche a la même longueur . Si sa représentation est petite c'est qu'elle est plutôt perpendiculaire au plan de projection, donc indépendante de ce plan.

- Une variable « supplémentaire-illustrative » est une variable qui est projetée sur un plan déjà constitué. On peut observer ses liens statistiques avec les autres variables mais elle n'a pas servi à structurer le nuage initial des individus, contrairement aux autres variables dites « actives ». La distinction variable active/supplémentaire-illustrative est d'un extrême importance méthodologique, même si , pratiquement les représentations peuvent être très semblables.

[4] deux livres utiles « *Analyse des Données* » Michel Volle , Economica et « *Statistique Exploratoire Multidimensionnelle* » Ludovic Lebart, Alain Morineau, M. Piron Dunod, 1995

Joindre cette page à votre copie, ne pas y écrire votre nom

Quelques points de repères en Analyses Factorielles des Correspondances Multiples

Une méthode pour penser les liens, en termes de sur et de sous représentations relatives des modalités de variables qualitatives , et pour les représenter, est l'Analyse Factorielle des Correspondances Multiples (1).

- c'est une méthode de statistique exploratoire qui permet de déceler d'éventuelles régularités statistiques dans les associations des modalités des différentes variables qualitatives.
- Les liens entre les variables qualitatives sont pensés et mesurés en termes de **sous et de sur représentations relatives**. Un énoncé tel que : chez les ménages OUVRIERS et CADRES SUPERIEURS les « assurances collectives accidents » sont sur-représentées, doit se comprendre ainsi : le pourcentage d '« assurances collectives accidents » dans ces catégories est plus élevé que le pourcentage moyen, celui de l'ensemble des ménages. l'AFCM est une méthode différentielle : elle met en évidence des écarts à la moyenne.
- Les variables qualitatives, sont représentées par leurs modalités, d'une même couleur. Lorsque la variable est **ordinale,** c'est le cas de la taille de l'agglomération, de l'âge du chef de ménage, les modalités successives sont reliées par une **ligne brisée orientée**.
- Les modalités sont, géométriquement, des points dans un espace de grande dimension. Ces points sont projetés sur un plan optimal en un certain sens : la figure géométrique initiale est la moins déformée possible, mais elle l'est .
- On obtient donc une représentation des modalités dans un plan dit « plan de projection » , et des règles de traductions, d'interprétations.
- Au centre du graphique se trouvent les pourcentages moyens. Les modalités de couleurs différentes situées du même côté, par rapport à ce centre sont mutuellement sur-représentées. Ainsi, dans le graphique « les comportements des français en matière d'assurance », en bas, chez les ménages habitant la « ville de Paris » , il y a une sur représentation des immeubles collectifs, des locataires, des employés…. Dans cette ville, les ménages ayant les caractéristiques précédentes sont relativement plus nombreux.
- Les modalités de couleurs différentes situées dans des directions opposées par rapport au centre du graphique sont mutuellement sous-représentées. Ainsi dans la « ville de Paris », il y a une sous-représentation des maisons individuelles, des AGRICULTEURS, des propriétaires….
- Des modalités de même couleur (d'une même variable) proches ont les mêmes caractéristiques c'est à dire les mêmes fréquences des modalités des autres variables
- Lorsque la ligne brisée orientée qui représente une variable ordinale se déploie bien dans le graphique, c'est le cas de la taille de l'agglomération, c'est qu 'elle est très significative , les différentes catégories : communes rurales, villes de moins de 20 000 habitants … sont, dans une certaine mesure, intra-homogènes extra-hétérogènes, lorsque l'on passe d'une classe de taille à une autre , il y a une variation sensible du pourcentage des modalités d'autres variables, notamment le pourcentage de locataires , de propriétaires, d'agriculteurs….
- Une variable supplémentaire-illustrative est une variable qui est projetée sur un plan déjà constitué. On peut observer ses liens statistiques avec les autres variables mais elle n'a pas servi à structurer le nuage initial des points-modalités, contrairement aux autres variables dites « actives ».La distinction variable active/supplémentaire-illustrative est d'une extrême importance méthodologique, même si , pratiquement les représentations peuvent être très semblables.

(1) un livre utile « *Analyses factorielles simples et multiples : objectifs, méthodes et interprétations* », Brigitte Escoffier et Jérôme Pages , Dunod 1998

FSEG STRASBOURG
Faculté des sciences économiques et de gestion

Joindre le sujet à votre copie, ne pas y écrire votre nom. Répondre sur le sujet ou sur la copie.

ITÉ DE STRASBOURG

Année universitaire 2014/2015
Master 1ère année mention Management des Projets et des Organisations
Semestre 1 – Session 1 / Contrôle terminal / Janvier 2015

« Méthodes Quantitatives d'aide à la décision : Analyse des Données»

Jean-Paul VILLETTE

Durée : 1 heure 30 . Tous documents interdits .Calculatrice autorisée

---

La note 20/20 correspondra à une partie du sujet. Les bonnes réponses sont brèves.

1 – tableau de données

Voici un tableau :

	N°Liste	Voix	N°Liste	Voix
Achenheim	Bigot	342	Richert	493
Adamswiller	Bigot	36	Richert	134
Albé	Bigot	92	Richert	117

Comment faudrait-il disposer les données pour pouvoir les traiter,

2– boissons chaudes, moments de la journée

Dans un questionnaire (*source site « IFOP »* Etude en 2000. échantillon de 1009 personnes). on trouve :

au petit-déjeuner, vous êtes plutôt …?
- ☐ thé
- ☐ café
- ☐ café décaféiné
- ☐ chocolat
- ☐ infusion
- ☐ rien de tout cela

- quelle est la variable ?
- est-ce que les modalités réponses sont exclusives ? exhaustives ? significations concrètes

3- questions

Dans un questionnaire[5] on trouve :

- Selon vous, quelles caractéristiques présentent les bouteilles Carola ? Elles sont :

☐ résistantes ☐ belles ☐ stables ☐ attirantes ☐ originales ☐ pratiques ☐ amusantes ☐

- pourquoi est-ce que les réponses ne sont pas exclusives ? que faut-il faire pour créer des questions à modalités exclusives (et exhaustives). Créer un petit exemple.

[5] « *Etude marché Carola* » Annelise DREVAL, mémoire de maitrise.2005- Université de Strasbourg

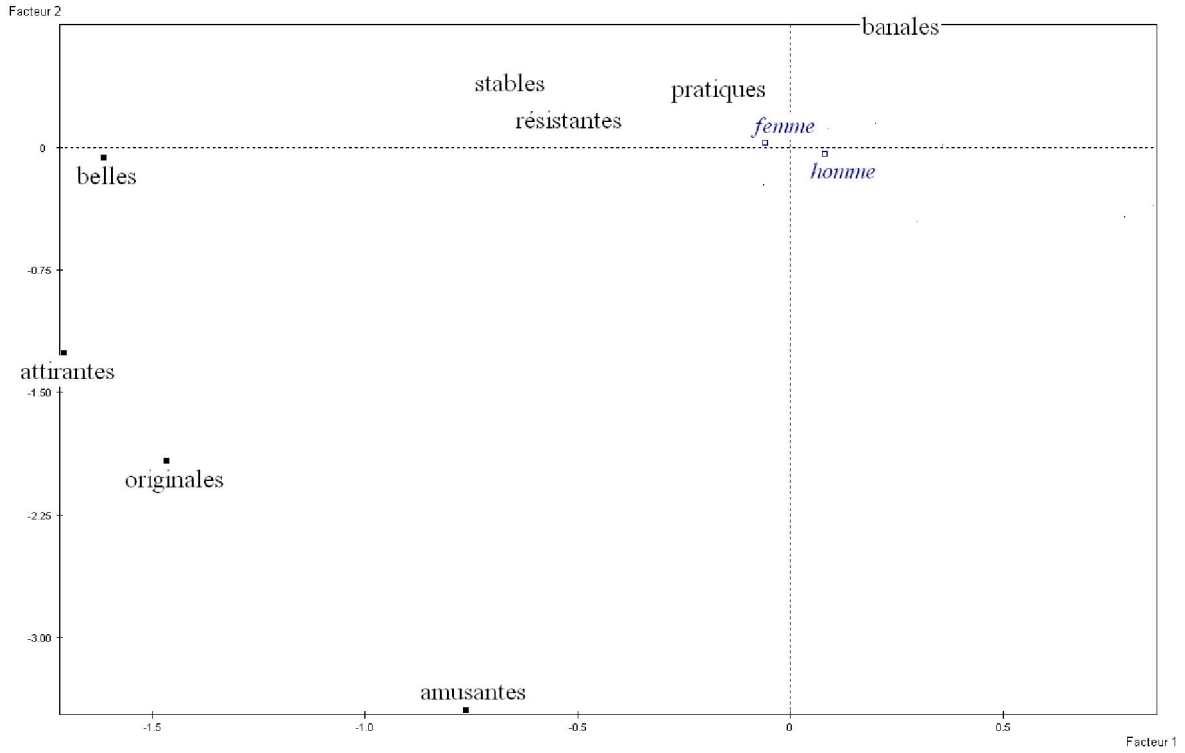
4: bouteilles, une Analyse Factorielle des Correspondances Multiples

Dans un questionnaire on trouve :

- Selon vous, quelles caractéristiques présentent les bouteilles Carola ? Elles sont :

☐ résistantes ☐ belles ☐ stables ☐ attirantes ☐ originales ☐ pratiques ☐ amusantes ☐

Ce sont les variables « actives ». la variable « *illustrative* » est *GENRE : homme, femme.*



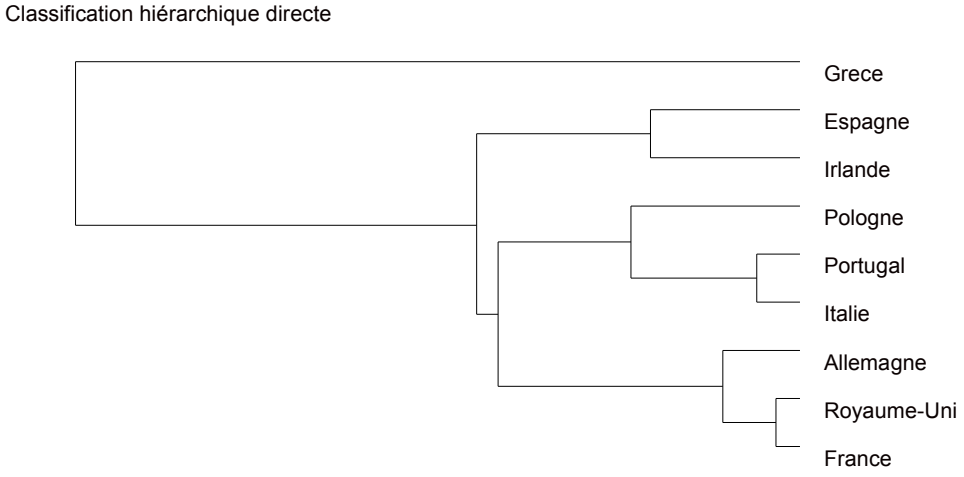
1- que peut-on dire de ceux qui trouvent les bouteilles « originales » ?

2- que peut-on dire du genre ?

5-Indicateurs économiques de neuf pays.Une Analyse en Composantes Principales suivie d'une Classification Hiérarchique Ascendante

	Taux chômage(%)	%croissance	%inflation	Dette/PIB	excedent/PIB	revenu moyenk€
France	8,6	0,5	2,4	86	-5,2	23,5
Allemagne	5,5	0,5	2,2	81	-1	21,5
Italie	10,3	-0,8	3,3	120	-3,9	18,1
Espagne	24,3	-0,3	2,7	68	-8,5	14,7
Royaume-Uni	8,1	-0,3	2,6	86	-8,3	20,6
Pologne	…	…	…	…	…	…
Grece	22,6	-4,7	1,2	165	-9,1	14
Portugal	15,2	-0,1	3,2	108	-4,2	10,6
Irlande	14,4	-1,1	2,6	108	-13,1	24

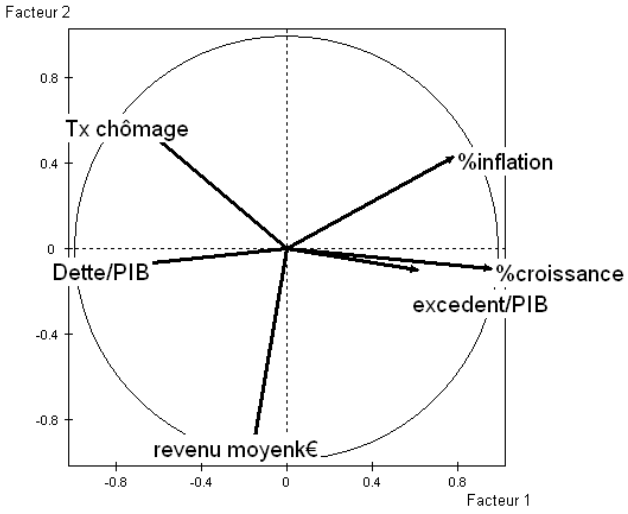
- *répondre sur la page suivante* : que peut-on dire de la Pologne (3 choses)
- Classification

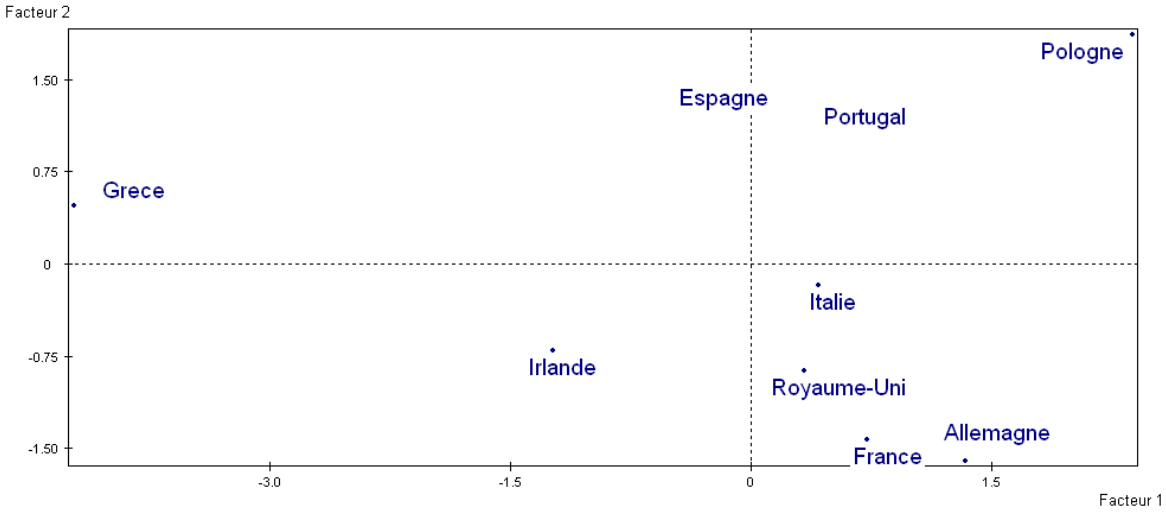


- Quelle est la règle statistique pour déterminer les groupes ?

- Quel est le pays le plus atypique ?

- On décide de considérer 4 groupes. Les colorier sur le graphe de la page suivante


Facteur 2
0.8
0.4
0
-0.4
-0.8
Tx chômage
%inflation
Dette/PIB
%croissance
excedent/PIB
revenu moyenk€
-0.8
-0.4
0
0.4
0.8
Facteur 1


Facteur 2
1.50
0.75
0
-0.75
-1.50
Pologne
Espagne
Portugal
Grece
Italie
Irlande
Royaume-Uni
Allemagne
France
-3.0
-1.5
0
1.5
Facteur 1

6– Eaux minérales :Analyses en Composantes Principales et Classification Hiérarchique Ascendante

Revoici le tableau de 13 eaux minérales (les individus) : **Talians, Courmayeur,…Volvic** et 6 variables numériques (la quantité de mg par litre (mg/l) de **Calcium** (Ca), **magnésium** (Mg), **potassium** (K), **sodium** (Na) , **carbonates**(HCO3) et **sulfates** (S04).

	Calcium	Magnésium	potassium	sodium	carbonates	sulfates
Talians	596	77	2	7	290	1530
……..	**……..**	**……..**	**……..**	**……..**	**……..**	**……..**

Très brève interprétation des trois clusters obtenus :

{ Talians, Courmayeur, Hepar, ContreX}	**Moyenne dans la classe**	**Moyenne générale**
sulfates	**1392**	**472**
Calcium	**539**	**224**
Magnésium	**85**	**36**
potassium	**3**	**2**
carbonates	**316**	**278**
sodium	8	13

{ Source Pyrénées, Perrier, Beckerich}	**Moyenne dans la classe**	**Moyenne générale**
sodium	**36**	**13**
carbonates	**321**	**278**
potassium	**3**	**2**
Magnésium	17	36
Calcium	96	224
sulfates	70	472

{ Vittel, Thonon, Evian, Valvert, Volvic, Cristaline}	**Moyenne dans la classe**	**Moyenne générale**
potassium	2	2
carbonates	230	278
sodium	4	13
Magnésium	14	36
sulfates	60	472
Calcium	78	224

7–classification hiérarchique ascendante (CHA): paroles d'enfants

.Source : « *Comment la parole vient aux enfants* » Bénédicte de Boysson-Bardies Ed. Odile Jacob 1996. Distribution des types de mots dans le vocabulaire d'enfants français, américains, suédois et japonais ayant un vocabulaire de moins de cinquante mots.

nombre de mots	français	américains	suédois	japonais
personnes	9	16	12	7
animaux	23	24	18	20
objets	44	51	44	29
verbes et adjectifs	24	11	25	26
onomatopées	2	5	1	15
expressions sociales	9	15	9	13

Une distance entre les vocabulaires des enfants des différents pays a été calculée

	{français}	{américains}	{suédois}	{japonais}
{français}	0	18	6	21
{américain}		0	19	30
{suédois}			0	22
{japonais}				0

Calculer et représenter approximativement le *dendrogramme* de la Classification Hiérarchique Ascendante en utilisant la stratégie d'agrégation entre « paquets » dite « de la moyenne « :

$$d_{moy}(A,B)=\frac{\sum_{x\in A, y\in B} d(x,y)}{card(A).card(B)}$$

Joindre cette page à votre copie, ne pas y écrire votre nom

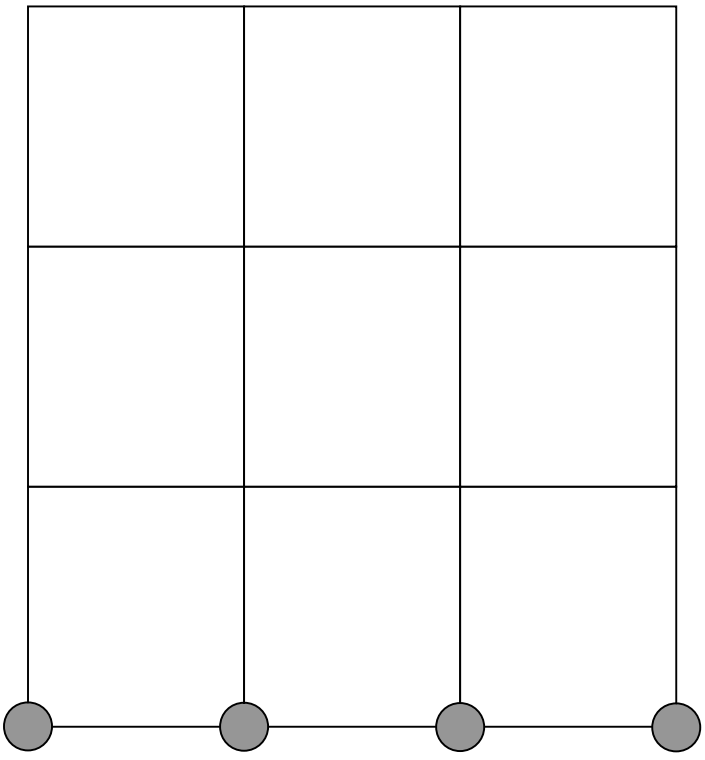
6 classification hiérarchique ascendante (CHA): paroles d'enfant

	{français}	{américain}	{suédois}	{japonais}
{français}	0	18	6	21
{américain}		0	19	30
{suédois}			0	22
{japonais}				0

Calculer et représenter le *dendrogramme*

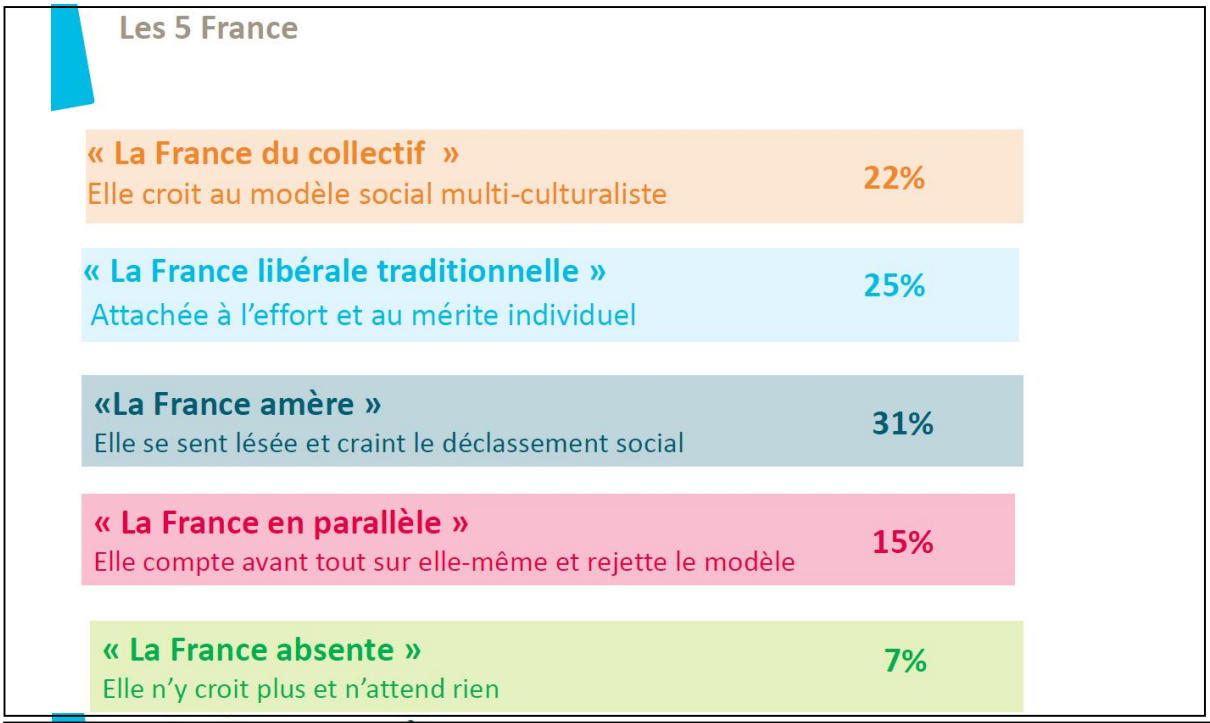
distance			

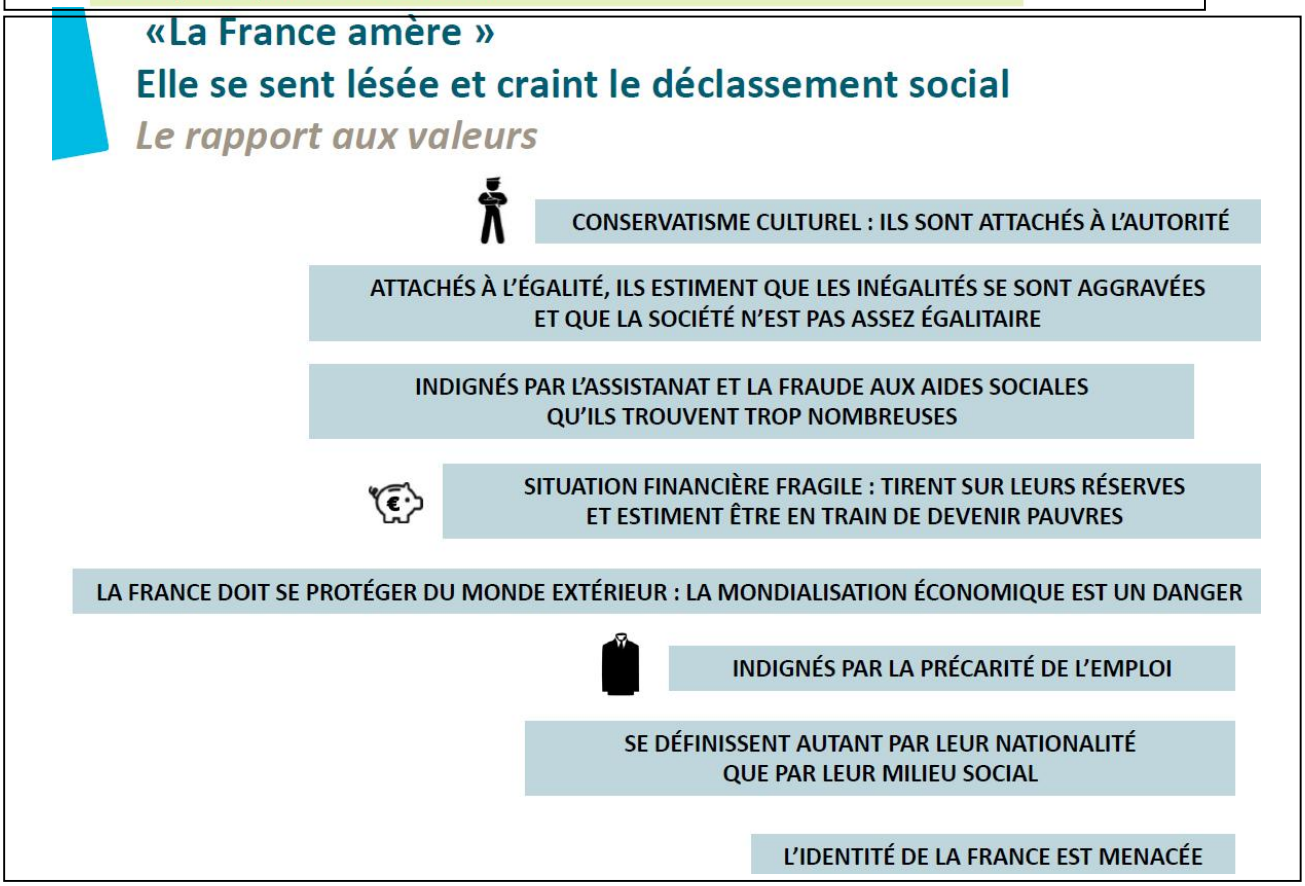
distance		



8 – question de méthodes

Ateliers CSA-Le Monde 3 dec 2013 « **Français, ce qui vous rassemble est-il plus fort que ce qui vous divise ?** »Extraits :



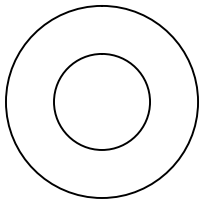


B- quelles méthodes statistiques ont-elles été utilisées pour obtenir ces cinq clusters ?

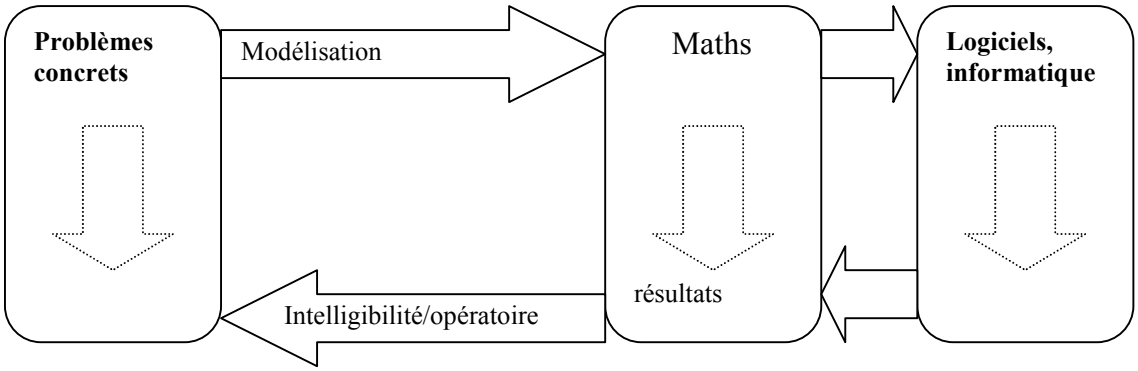
B***- Que faut-il comprendre, statistiquement, à :« L'IDENTITE DE LA FRANCE EST MENACEE » ?

Début 1 : modélisations

Terre, orange et bouts de ficelle



le schéma



« all models are wrong but some are usefull »

Paul WIGNER (physicien américain d'origine hongroise 1902-1994 , prix Nobel de Physique 1963) : « *the unreasonnable effectiveness of mathematics* »

James Clerk MAXWELL (physicien écossais 1831-1879) :« *les équations semblent plus intelligentes que nous* »

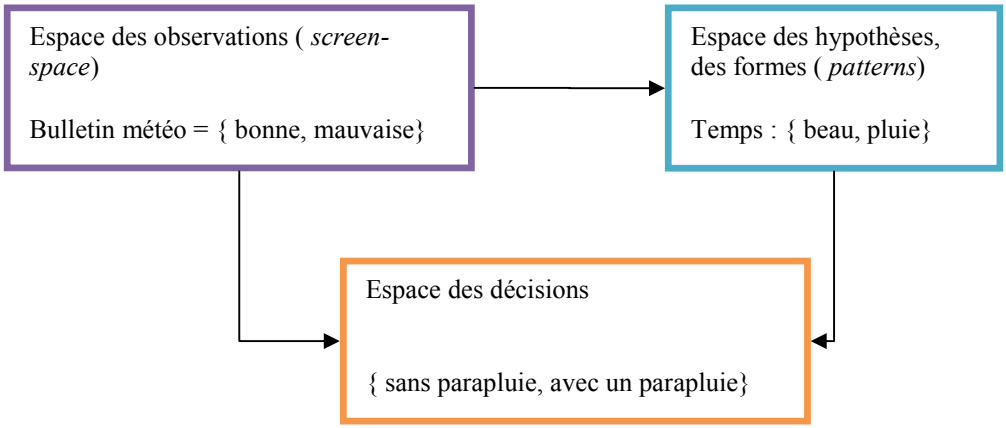
Exercice :
1-j'ai des pièces de 2 et de 5 €., 10 pièces au total, ce qui fait 38€. Combien de pièces de 2€, de 5 € ?
- essayer de résoudre par un raisonnement arithmétique
- modéliser

moralités :

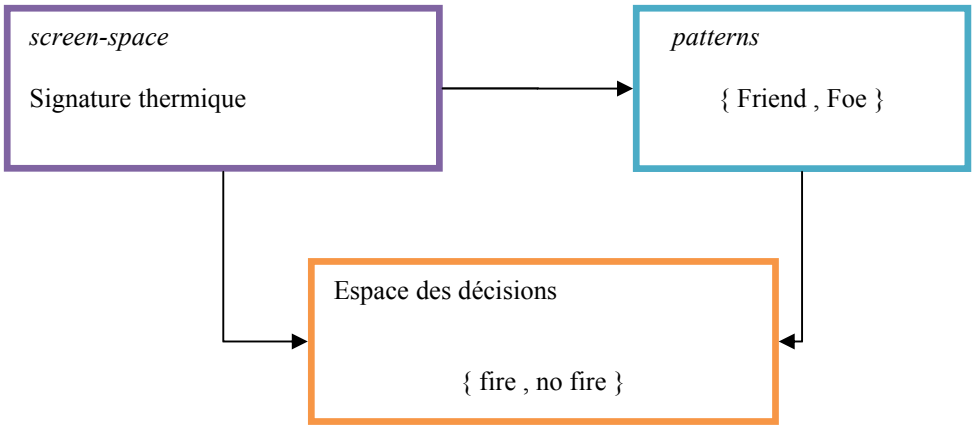
Début 2 : Outils Quantitatifs d'Aide à la Décision :

Reconnaissance des formes (*Pattern recognition*)

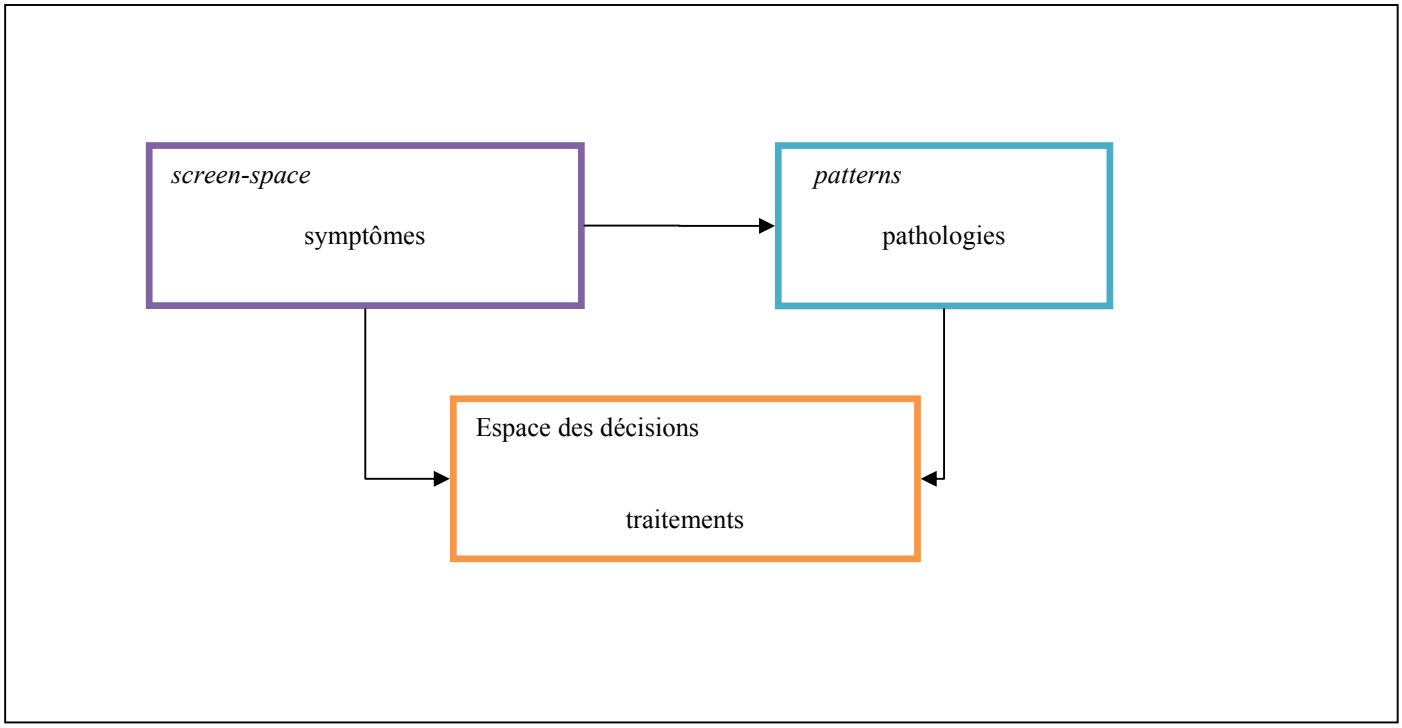
Il y lieu de distinguer trois espaces, et leurs rapports :



Exemple : FFI



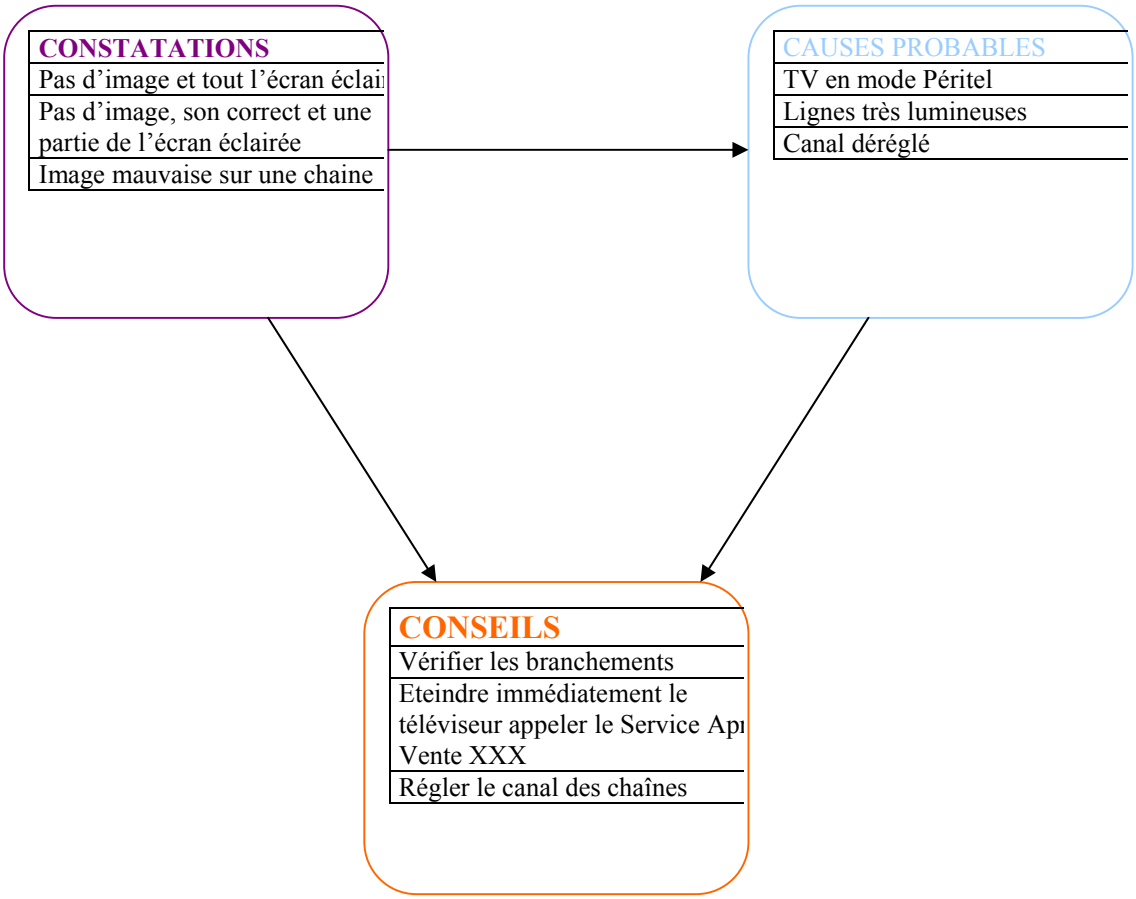
Autres exemples :



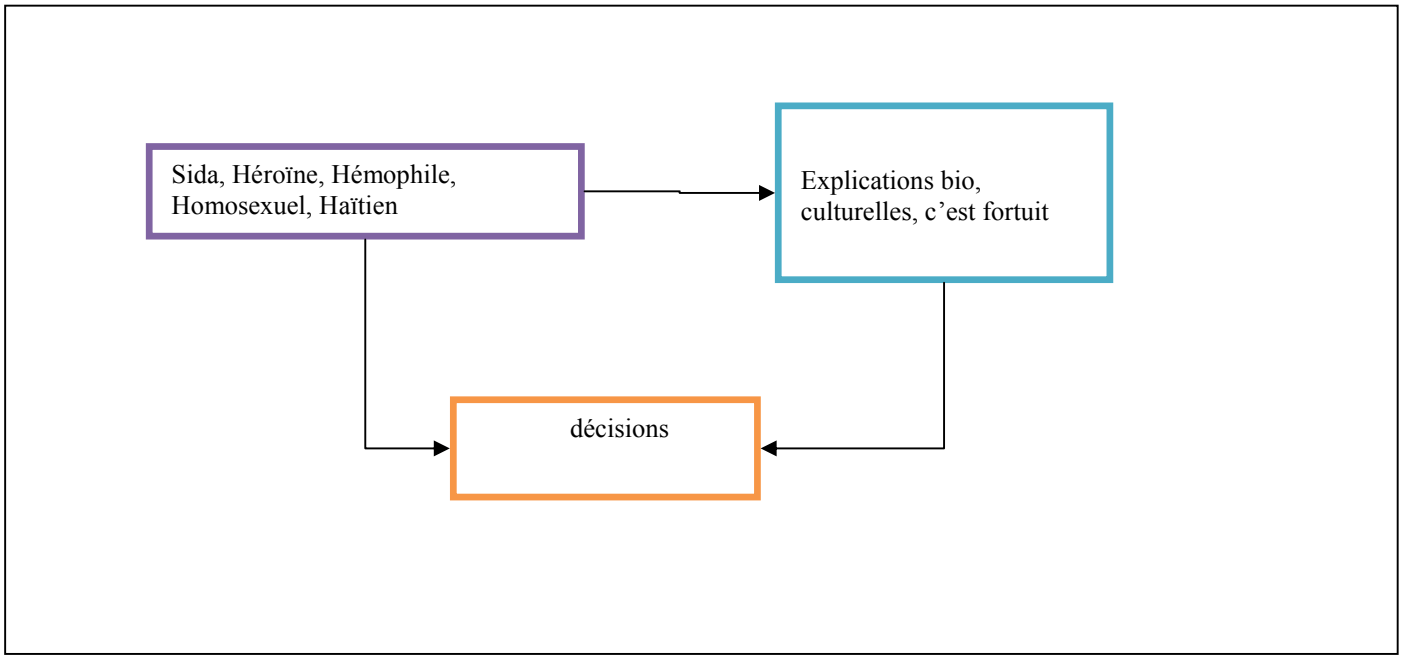
Notice S.A.V d'un hypermarché

Screen-space : constatations *patterns* : causes probables décisions : conseils

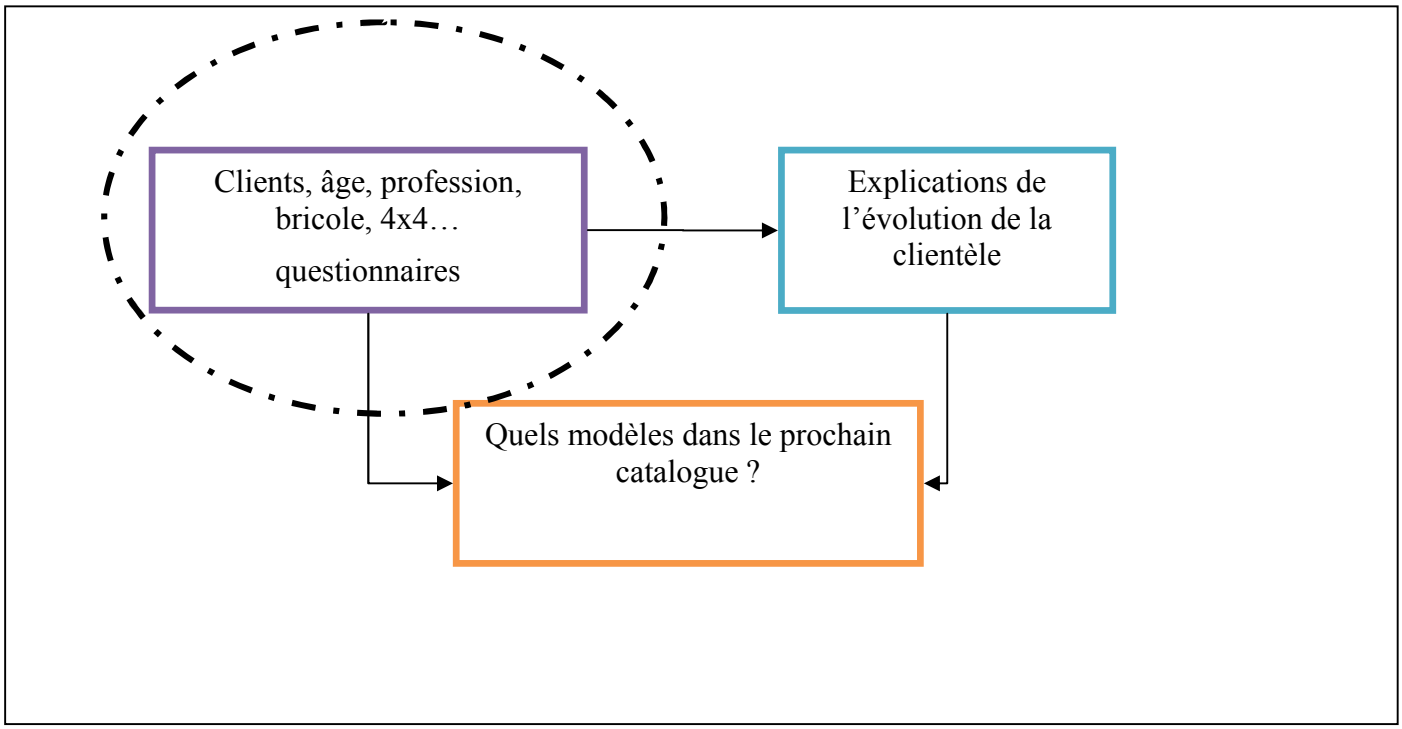
CONSTATATIONS	CAUSES PROBABLES	CONSEILS
Pas d'image et tout l'écran éclairé	TV en mode Péritel	Vérifier les branchements
Pas d'image, son correct et une partie de l'écran éclairée	Lignes très lumineuses	Eteindre immédiatement le téléviseur appeler le Service Après Vente XXX
Image mauvaise sur une chaine	Canal déréglé	Régler le canal des chaînes



Considérations méthodologiques : le temps, et la diversité des explications



nous verrons :



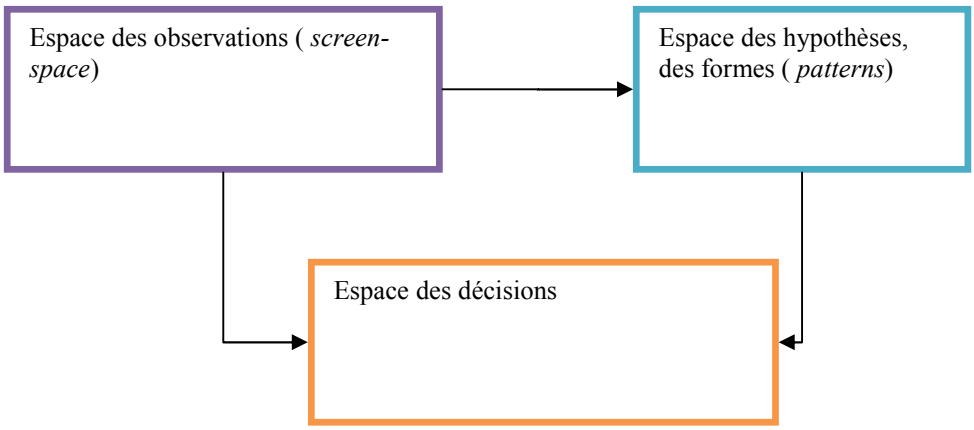
autres considérations méthodologiques : le botaniste et le fleuriste

Comprendre la différence de point de vue les rapports entre :

botaniste (statisticien) : description,
fleuristes : (économiste, gestionnaire..) : jugements de valeurs, d'utilité (la pertinence..), l'action .

Le temps qu'il fait	Point de vue/rôle	description	Interprétations/jugements de valeur	Décision
météorologue	« botaniste »	Il va pleuvoir		
jardinier	« fleuriste »		Bonne nouvelle	Je plante des carottes
vacancier	« fleuriste »		Mauvaise nouvelle	Je reste couché

Reconnaissance des formes (*Pattern recognition*)
Il y lieu de distinguer trois espaces, et leurs rapports :



Le rôle des uns, le rôle des autres et les confusions dans :

- « *des scientifiques s'inquiètent de la fonte de la banquise* ».
- dynamisme démographique,
- indicateur d'inégalité
- Sauver la planète
- catastrophe naturelle .

Die Philosophen haben die Welt nur verschieden interpretiert; es kömmt drauf an, sie zu verändern.
(*Les philosophes n'ont fait qu'interpréter le monde de diverses manières, il s'agit maintenant de le transformer*[6]

-

[6] Karl Marx, XI ème thèse sur Feuerbach, 1845.

P31

Évaluation de l'activité de rétrocession d'une PUI : Quelles sont les attentes des patients externes ?

Hehn M[1], Armand S[1], Grangeasse L[1], Talla M[1], Pailloud E[1], Villette JP[2]

[1] CH de Belfort-Montbéliard, Service Pharmacie,
14 rue de Mulhouse, 90 016 BELFORT CEDEX
[2] BETA, Faculté des Sciences Économiques ULP Strasbourg

INTRODUCTION

Dans le cadre d'une démarche d'amélioration continue de la qualité, l'équipe pharmaceutique a mené une enquête de satisfaction auprès des patients externes. Le parcours du patient de la consultation à la PUI a été évalué. L'objectif de ce sondage est de dresser un profil de leurs attentes. Ce travail constitue une première étape dans la mise en place d'indicateurs de qualité et participe à la démarche d'accréditation.

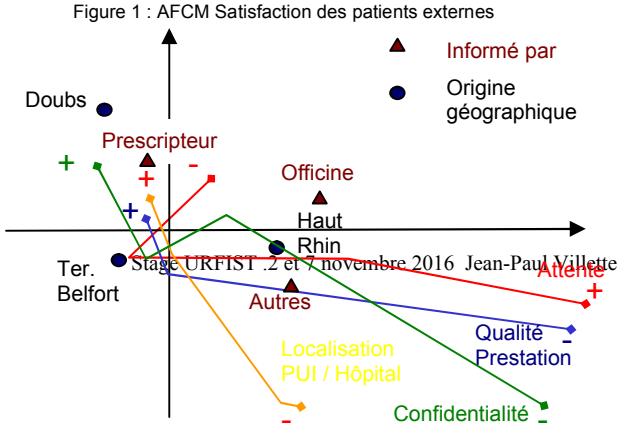
METHODOLOGIE

Les patients externes de la PUI ont répondu à un questionnaire composé de 6 items (répartis en 15 questions dont une ouverte) dont : - profil du patient - information sur le lieu de délivrance - orientation à travers l'hôpital - temps d'attente - confidentialité - appréciation globale de la prestation. Le sondage a duré quatre semaines. Un échantillon de 10 patients a permis de valider le questionnaire au préalable. Les différentes méthodes mises en œuvre sont notamment, l'Analyse Factorielle des Correspondances Multiples (AFCM) afin de déceler d'éventuelles typologies et la segmentation de la variable «appréciation globale de la prestation»[(1, 2)]

RESULTATS

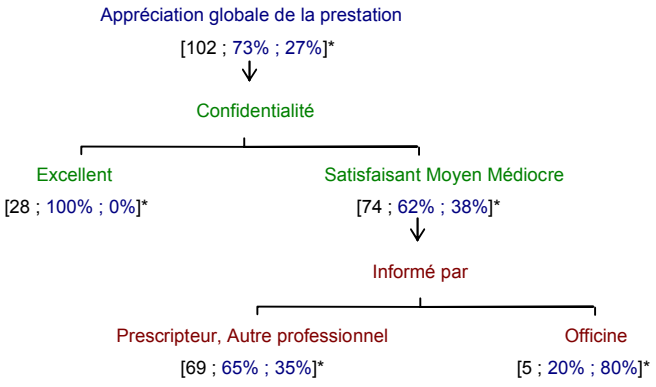
•102 questionnaires ont été complétés sur 298 dispensations effectuées sur la même période. 74.5% des patients sont originaires du Territoire de Belfort.
• L'appréciation globale de la prestation est jugée à 73% très satisfaisante et 27% partiellement satisfaisante.
• La figure 1 synthétise les résultats de l'AFCM. A l'origine, se trouvent les opinions exprimées par la moyenne des personnes interrogées. Ces dernières évaluent, de façon positive, notamment la qualité globale de l'accueil, le temps d'attente, la confidentialité, la localisation de la pharmacie dans l'hôpital. La qualité de la prestation est étroitement liée à la confidentialité et au temps d'attente.

Figure 1 : AFCM Satisfaction des patients externes



• Un arbre de segmentation (figure 2) de la variable appréciation globale de la prestation confirme et affine les résultats précédents.

Figure 2 : Arbre de segmentation de l'appréciation de la prestation



*[Effectif ; % très satisfait ; % partiellement satisfait]

• Ainsi, l'insatisfaction relative des patients sur la qualité de la prestation «s'explique» selon deux critères principaux : la confidentialité et l'information du patient sur le lieu de délivrance.

DISCUSSION

Vue la configuration de la PUI, les thèmes prioritaires du questionnaire abordent essentiellement les aspects organisationnels. Cependant, les commentaires (issus de la question ouverte) révèlent notamment l'importance de la relation de confiance établie entre le patient et l'équipe pharmaceutique.
Ces résultats confirment ceux d'une étude américaine[(3)] qui a démontré que la satisfaction perçue des patients repose davantage sur l'aspect relationnel et organisationnel que sur des critères techniques.

CONCLUSIONS

Le niveau de satisfaction globale est très élevé pour une file active de 300 patients par mois. A la suite de cette enquête, des mesures correctives ont été apportées. Une zone de confidentialité, séparée de la zone d'attente a été créée. Les praticiens hospitaliers ont bénéficié d'une information sur les médicaments à délivrance hospitalière. Cette étude sera renouvelée afin d'évaluer les progrès apportés et aborder de nouveaux thèmes d'amélioration.

Références

1. L Lebart, A Morineau, M Piron, *Statistique exploratoire multidimensionnelle*, édition Dunod, Paris 1995, 439p
2. G Celeux, JP Nakache, *Analyse discriminante sur variables qualitatives*, édition Polytechnica, 1994, 473p
3. LD Ried, F Wang, H Young, R Awiphan, *Patients satisfaction and their perception of the pharmacist*, JAPA, Nov-Dec 1999, 30(6):835-842

MARCHE DU PANSEMENT, DU DRAPAGE ET DE L'HABILLAGE STERILES POUR UNE MEILLEURE CONNAISSANCE DE L'OFFRE FOURNISSEUR

A.Ancédy[1], M.Hehn[1], L. Bertrand[1], JP. Villette[2], M.Talbert[1].
1.Centre hospitalier de Saint-Denis 93205 Saint-Denis - 2. BETA-UMR 7522 CNRS Université de Strasbourg

INTRODUCTION

Lors de la dernière consultation du groupement de commande RESAH-IdF-Dispositifs médicaux pour la fourniture de pansements et de drapage, le **critère « assistance technique et logistique »** contribuait à hauteur de **10%** à l'évaluation de l'offre fournisseur. Ce critère a été évalué à l'aide d'un « Questionnaire assistance Technique et Logistique » (QTL).
Ce travail a pour objectif d'identifier les diverses typologies de fournisseurs et de les croiser aux choix réalisés au terme de cette consultation.

MATERIEL & METHODES

➢ Extraction des données fournies dans les « Questionnaires assistance Technique et Logistique » (QTL).
➢ Analyse factorielle des correspondances multiples et classification hiérarchique ascendante (logiciel SPAD) réalisées à partir de 9 variables qualitatives (forme juridique, nombre d'employés, capital social, établissement pharmaceutique, délai de livraison, franco de port, développement durable, QTL complété, envoi dématérialisé) et 5 variables numériques associées (seuil de franco de port (€HT), frais de port (€HT), note au QTL, nombre de lots avec une offre et nombre de lots retenus).

RESULTATS

A. Typologies des fournisseurs soumissionnaires

➢ 58 fournisseurs ont proposé une offre.

Groupe 1
29 fournisseurs (50%)
ADDMEDICA, ADHESIA, APURA, BBRAUN, BD, BSN, CG Médical, COLLIN, DATASCOPE, DISTRIPHAC, EUROMEDIS, GENEVRIER, HYDREX, HYGEA, INNOSET, LCA, LEICA, MAURY, MEDTRONIC, MOLNLYCKE, MOLYPHARM, RAFFIN, PHAGOGENE, PHARMASET, SEPTODONT, SMITH&NEPHEW, STERIMA, SYLAMED, URGO

Groupe 2
14 fournisseurs (24%)
ABS-BOLTON, CAREFUSION, COLOPLAST, COVIDIEN, LCH, TETRA, TROIS M SANTE, BROTHIER, CLINICAL IHT, CONVATEC, HARTMANN, LOHMANN, THUASNE, VYGON

Groupe 3
7 fournisseurs (12%)
COOPER, GAMIDA, HEMODIA, INTERMED, ETHICON, GILBERT, HOLLISTER

Groupe 4
8 fournisseurs (14%)
ASTERIE, BIOMET, MAIL TUB, POURET, BAXTER, EIF, MEDISPORT, PETERS SURGICAL

B. Dans quels groupes sont les fournisseurs retenus?

➢ 40/58 **(69%)** fournisseurs ont été retenus.

	Groupe 1	Groupe 2	Groupe 3	Groupe 4
Effectif total	29	**14**	7	8
Fournisseurs retenus	19	**12**	5	4
% dans le groupe	66%	**86%**	71%	50%

➢ 4 groupes de fournisseurs sont mis en évidence.

Groupe 1

Modalités caractéristiques	Proportion groupe	Proportion effectif total
Franco de port : NON	83%	48%
Seuil franco de port : 100-400 € HT	72%	43%
Frais de port : 9-35 € HT	48%	24%

Groupe 2

Modalités caractéristiques	Proportion groupe	Proportion effectif total
Franco de port : OUI	93%	31%
Seuil franco de port : 0-80 € HT	100%	43%
Note QTL : 14.5-18/20	86%	31%

Groupe 3

Modalités caractéristiques	Proportion groupe	Proportion effectif total
QTL complété : NON	100%	22%
Envoi démat. : Non Renseigné (NR)	100%	26%
Dévelop. durable :NR	100%	33%

Groupe 4

Modalités caractéristiques	Proportion groupe	Proportion effectif total
Franco de port : NR	100%	14%
Seuil franco de port : NR	100%	14%
QTL complété : NON	75%	22%

➢ Les fournisseurs du groupe 2 présentent le meilleur profil logistique et technique, parmi ces derniers une grande majorité a été retenue (86%).

➢ L'effectif du groupe 1 est le plus important, cependant leurs prestations logistique et technique sont moins performantes.

➢ Les groupes 3 et 4 dont les effectifs sont restreints, se caractérisent par des données incomplètes ou absentes.

DISCUSSION - CONCLUSION

Il est dommage qu'un nombre important de fournisseurs retenus se situe dans le groupe 1, alors que ce groupe ne présente pas les meilleures performances par rapport aux caractéristiques étudiées.

Si l'on souhaite inciter les fournisseurs à mieux répondre dans ce domaine, il serait utile d'en augmenter la pondération dans les critères de choix. D'autre part, la reformulation du questionnaire devrait également permettre de réduire le taux de non réponse.

UE Organisation et gestion hospitalière – Politique des achats mars 2010

Début 3 : Analyse des données

Analyse des Données : synthèses de résultats

Les données ne sont pas données….coûts, processus de productions..

Synthèses/ typologies :
des groupes intra-homogènes extra-hétérogènes (*internal cohesion and external isolation of clusters*)
sélections/ évictions
ah ! le code de la route !

des groupes ? enjeux :

---

simplifier le sens est dans les groupes

simplifier : les arbres cachent les forêts . mettre les objets dans des sacs.

le sens est dans les groupes : match de football, phénomènes de mode, courants de pensée

distinctions et rapports :

Statistique Exploratoire (« Analyse des Données ») / Statistique Confirmatoire (« Econométrie »)

- Aussi : Statistique Exploratoire/ Statistique décisionnelle
- En anglais, « Analyse des Données » : *Data Exploratory Analysis (DEA), clustering, classification*

Le Big Data

Quelques mots. Tirer parti de beaucoup de liens faibles/ de quelques liens forts

Décès de tous âges	Année	Janvier	Février	Mars	Avril	Mai	Juin	Juillet	Août	Septembre	Octobre	Novembre	Décembre
1946	545 880	70 900	53 958	57 287	45 376	42 591	37 721	37 587	34 880	35 188	37 842	42 954	49 596
1947	538 157	60 453	56 891	56 442	45 121	42 605	37 894	38 364	36 763	35 768	40 488	41 361	46 007
1948	513 210	46 161	45 412	51 983	43 829	42 003	37 084	39 069	35 272	35 314	39 588	43 596	53 899
1949	573 598	87 861	58 592	52 772	44 154	41 896	39 141	40 042	37 372	36 267	40 534	47 049	47 918
1950	534 480	51 927	47 749	50 439	47 248	45 515	40 095	39 798	38 124	37 075	42 232	44 418	49 860
1951	565 829	63 048	58 340	58 959	49 690	46 496	40 359	40 199	37 670	37 288	42 997	42 370	48 413
1952	524 831	53 364	51 750	50 111	44 752	41 167	38 144	40 266	36 246	36 278	41 385	42 467	48 901
1953	556 983	69 112	73 023	54 100	43 573	42 651	37 468	36 807	36 685	35 488	40 885	43 243	43 948
1954	518 892	52 729	53 607	49 423	44 349	43 538	38 096	37 671	36 316	35 450	40 719	41 178	45 816
1955	526 322	52 017	46 623	56 974	45 364	41 749	38 477	38 211	37 969	35 745	41 772	43 595	47 826
1956	545 700	52 107	61 556	58 113	46 263	44 009	38 974	38 820	36 144	36 638	40 669	44 859	47 548
1957	532 107	56 610	45 707	43 656	41 437	40 495	39 094	39 774	36 455	36 454	45 258	52 329	54 838
1958	500 596	51 563	43 451	48 476	43 746	40 009	36 288	36 939	35 096	34 417	40 532	42 689	47 390
1959	509 114	47 483	46 355	48 627	48 180	43 832	37 602	38 584	35 222	34 954	39 524	42 899	45 852
1960	520 960	60 414	57 951	46 868	40 827	40 405	35 851	36 250	36 049	35 599	42 503	40 787	47 456
1961	500 289	50 816	41 813	43 080	41 093	40 232	39 362	37 511	36 889	35 808	40 365	44 631	48 689
1962	541 147	51 804	47 280	58 960	49 374	42 495	40 608	38 532	36 801	36 769	41 540	44 281	52 703
1963	557 852	58 163	57 997	63 038	47 911	43 129	38 910	40 210	37 276	38 310	41 888	41 254	49 766
1964	520 033	52 136	45 770	46 434	43 490	41 397	38 313	41 469	38 124	37 193	43 687	44 768	47 252
1965	543 696	50 205	49 851	64 174	44 094	43 481	41 080	39 191	38 705	39 029	42 809	43 727	47 350
1966	528 782	52 493	43 197	47 037	45 795	42 400	40 735	39 941	40 181	38 249	42 553	47 135	49 066
1967	543 033	52 771	49 087	49 256	45 422	44 657	41 305	42 527	40 224	40 241	42 453	44 877	50 213
1968	553 441	51 061	55 577	57 273	46 455	43 713	40 892	41 476	40 075	39 351	43 051	44 826	49 691
1969	573 335	50 881	46 746	50 615	47 426	44 932	42 277	44 000	41 160	40 915	44 077	45 581	74 725
1970	542 277	57 729	44 189	49 848	45 864	44 369	41 153	42 132	41 178	40 172	44 033	44 393	47 217
1971	554 151	55 079	46 097	52 813	45 523	44 005	42 310	45 569	40 991	40 482	44 621	45 538	51 123
1972	549 900	54 956	47 599	48 118	44 008	44 972	42 404	43 464	41 689	41 617	45 828	44 159	51 086
1973	558 782	64 762	46 668	48 974	45 607	44 328	41 801	42 236	43 518	40 570	46 284	45 080	48 954
1974	552 551	48 177	42 131	49 355	45 316	46 133	42 626	43 058	42 266	41 559	48 213	46 986	56 731
1975	560 353	55 805	44 628	49 246	47 957	46 400	43 697	44 704	45 287	42 164	46 640	44 745	49 080
1976	557 114	49 611	50 172	55 206	47 105	45 605	45 449	45 983	41 587	40 928	43 403	44 149	47 916
1977	536 221	49 306	43 491	48 498	46 991	44 659	41 209	42 834	41 370	41 290	43 629	43 932	49 012
1978	546 916	50 820	52 291	49 283	45 103	44 974	41 509	44 099	40 427	39 842	45 078	44 122	49 368
1979	541 805	50 317	43 825	48 683	46 123	46 375	41 870	44 242	41 213	41 059	45 289	45 464	47 345
1980	547 107	51 015	44 590	47 306	45 852	45 611	42 621	44 616	42 166	40 520	45 835	47 016	49 959
1981	554 823	55 289	50 153	50 199	44 564	46 040	42 658	43 935	41 958	40 219	45 618	45 182	49 008
1982	543 104	49 581	43 950	48 016	46 145	45 757	41 919	46 326	41 874	41 276	45 073	44 347	48 840
1983	559 655	50 526	49 778	51 319	46 372	45 008	43 395	49 570	42 000	41 446	44 655	45 808	49 778
1984	542 490	48 740	45 110	49 926	46 998	45 498	43 830	43 445	41 457	40 656	45 437	43 937	47 456
1985	552 496	57 357	47 112	51 271	45 653	45 269	41 673	43 230	41 105	40 213	43 963	47 412	48 238
1986	546 926	51 466	52 732	53 666	45 204	43 594	43 632	42 198	40 821	41 165	42 798	42 894	46 756
1987	527 466	51 229	43 613	45 600	43 337	43 786	42 176	42 516	41 422	39 308	43 554	43 478	47 447
1988	524 600	46 740	43 408	47 762	43 770	42 715	41 272	41 457	41 105	40 145	44 085	43 970	48 171
1989	529 283	48 428	43 397	45 384	42 154	43 099	41 112	42 710	40 663	40 376	45 098	43 021	53 841
1990	526 201	56 059	43 208	44 491	43 217	41 355	40 266	42 461	41 681	38 997	43 401	42 871	48 194
1991	524 685	47 211	46 308	45 463	42 788	42 962	40 314	41 871	41 130	39 195	44 502	44 288	48 653
1992	521 530	52 247	45 888	44 875	42 399	42 145	39 573	41 953	41 422	39 939	43 817	42 329	44 943
1993	532 263	49 231	43 185	49 688	43 686	42 331	40 678	41 063	40 987	40 571	43 938	44 017	52 888
1994	519 965	49 932	42 613	44 356	43 104	41 841	40 573	43 940	41 822	41 050	43 503	41 833	45 398
1995	531 618	48 788	40 937	46 454	45 670	43 965	41 742	43 819	42 078	40 873	43 462	43 346	50 484
1996	535 775	51 341	46 604	47 495	44 833	43 636	41 237	42 390	39 915	40 680	42 725	43 337	51 582
1997	530 319	59 227	44 788	45 218	43 954	42 644	39 325	42 105	42 519	39 394	42 266	43 100	45 779
1998	534 005	47 361	46 023	51 720	48 135	43 961	40 956	41 368	41 319	40 016	42 846	42 928	47 372
1999	537 661	51 436	50 019	50 049	43 339	42 664	40 123	41 799	41 012	39 754	43 651	43 308	50 507
2000	530 864	58 939	47 747	45 098	42 658	41 722	40 568	41 511	41 178	39 415	43 417	42 983	45 628
2001	531 073	49 529	42 513	46 133	43 594	44 212	41 846	43 230	42 518	40 973	43 132	44 583	48 810
2002	535 144	55 663	45 849	46 785	44 009	42 973	42 132	42 319	40 461	40 322	44 243	43 166	47 222
2003	552 339	50 920	44 667	47 177	44 479	42 928	42 604	43 760	56 550	41 137	43 786	43 954	50 377
2004	509 429	50 377	43 446	44 993	41 464	41 337	39 021	39 975	39 091	39 360	42 660	41 530	46 175
2005	527 533	48 186	50 324	53 401	43 365	42 433	40 712	39 780	38 870	38 973	42 551	41 991	46 947
2006	516 416	49 166	43 380	45 989	41 524	41 563	40 776	43 185	40 156	40 304	42 500	42 214	45 659
2007	521 016	49 225	44 229	45 401	42 342	41 112	39 837	41 885	40 044	39 721	44 448	44 549	48 223
2008	532 131	51 862	45 865	46 964	44 595	43 007	40 723	41 940	40 455	40 232	44 280	42 763	49 445
2009	538 116	58 938	46 677	46 407	43 550	42 818	40 796	41 384	41 066	40 033	44 885	43 238	48 324
2010	540 469	51 095	45 770	47 555	43 782	43 824	41 654	42 727	41 487	41 743	46 043	44 264	50 525
2011	534 795	52 088	45 390	46 906	43 075	43 325	41 391	41 824	41 861	41 188	44 935	44 274	48 538
2012	559 227	51 396	53 590	52 871	45 863	44 999	41 724	42 923	42 258	41 282	46 174	45 771	50 376
2013	558 408	55 134	50 603	53 241	46 036	44 410	42 324	44 145	41 622	41 986	45 063	44 452	49 392
2014	547 003	50 125	45 757	48 298	44 466	44 467	42 107	43 465	43 057	42 544	45 419	45 979	51 319

Source : Insee, statistiques de l'état civil

De 2005 à 2014 : 5 375 114

Un point de vue :

Confirmatory vs Exploratory Data Analysis

- **Confirmatory Analysis**
  - Inferential Statistics - Deductive Approach
    - Heavy reliance on probability models
    - Must accept untestable assumptions
    - Look for definite answers to specific questions
    - Emphasis on numerical calculations
    - Hypotheses determined at outset
    - Hypothesis tests and formal confidence interval estimation
  - Advantages
    - Provide precise information in the right circumstances
    - Well-established theory and methods
  - Disadvantages
    - Misleading impression of precision in less than ideal circumstances
    - Analysis driven by preconceived ideas
    - Difficult to notice unexpected results

- **Exploratory Analysis**
  - Descriptive Statistics - Inductive Approach
    - Look for flexible ways to examine data without preconceptions
    - Attempt to evaluate validity of assumptions
    - Heavy reliance on graphical displays
    - Let data suggest questions
    - Focus on indications and approximate error magnitudes
  - Advantages
    - Flexible ways to generate hypotheses
    - More realistic statements of accuracy
    - Does not require more than data can support
    - Promotes deeper understanding of processes
      - Statistical learning
  - Disadvantages
    - Usually does not provide definitive answers
    - Difficult to avoid optimistic bias produced by overfitting
    - Requires judgement and artistry - can't be cookbooked

www.geog.ucsb.edu/~joel/g210_w07/lecture_notes/lect01/oh07_01_2.html

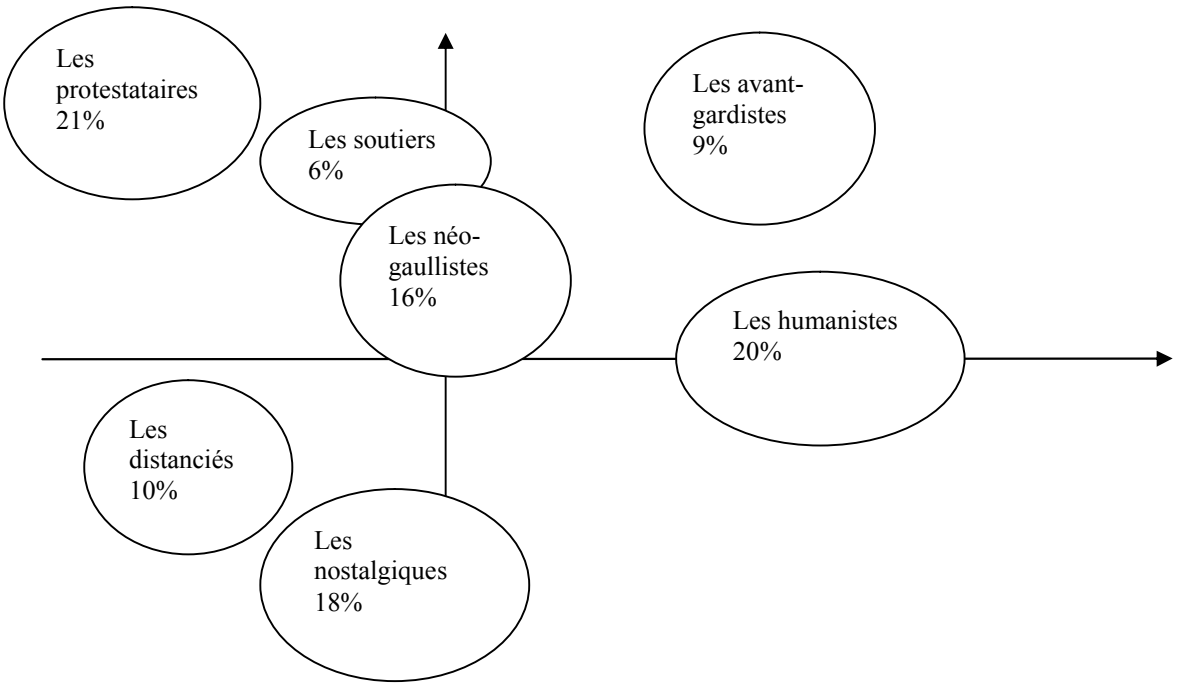
UC Santa Barbara dept of geograpphy

Exemple 1 : catalogue La Redoute, les huit « boutiques » femme :

1- classiques qualité
2- classiques paraître
3- traditionnelles
4- femmes actives petits prix
5- les bains
6- les modernes paraître
7- les modernes marques mode
8- les juniors

Exemple 2 : TNS-Sofres - Figaro Magazine[7]

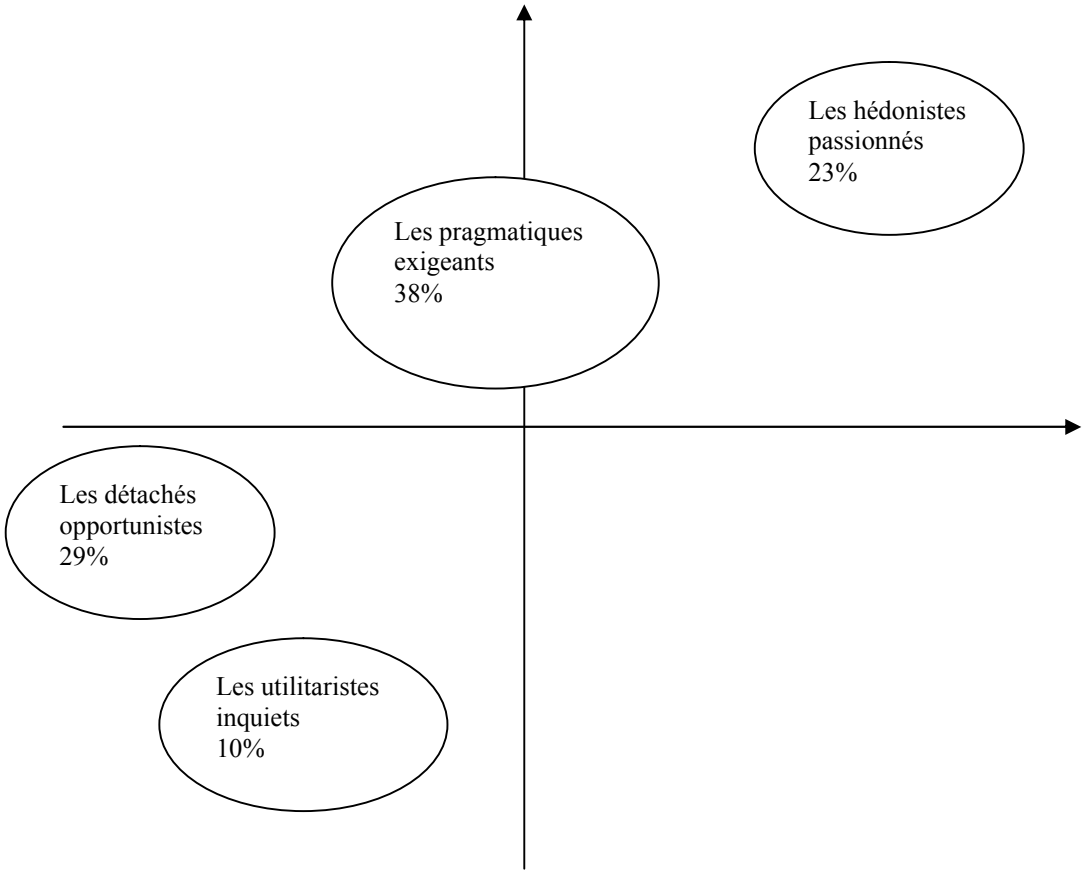
.....les sympathisants de droite se regroupent par affinités en sept familles distinctes mises en lumière par l'enquête TNS Sofres.....



[7] N° du samedi 20 novembre 2004

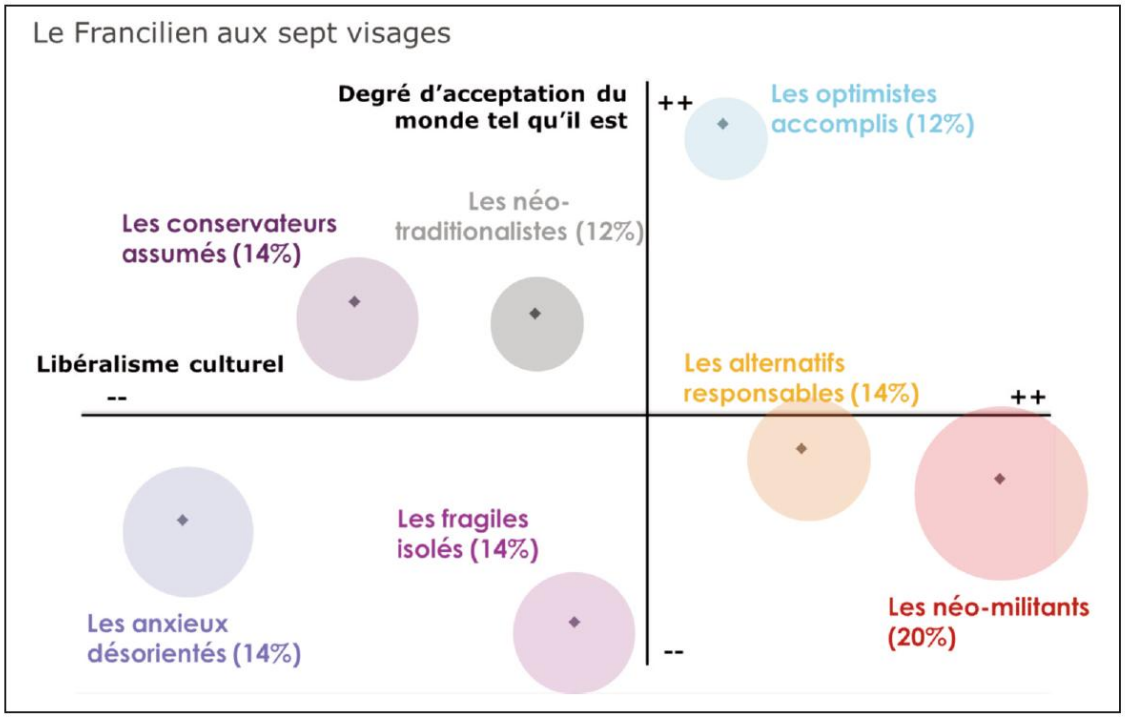
Exemple 3 : typologie des acheteurs, » le club des quatre »,

l'observatoire de l'automobile , groupe CETELEM, 2004



Exemple 4 : « Franciliens : un portrait qui trouble les lignes politiques classiques »

Guénaëlle Gault, TNS-SOFRES. NOTE n° 206 - Fondation Jean-Jaurès - 3 mars 2014

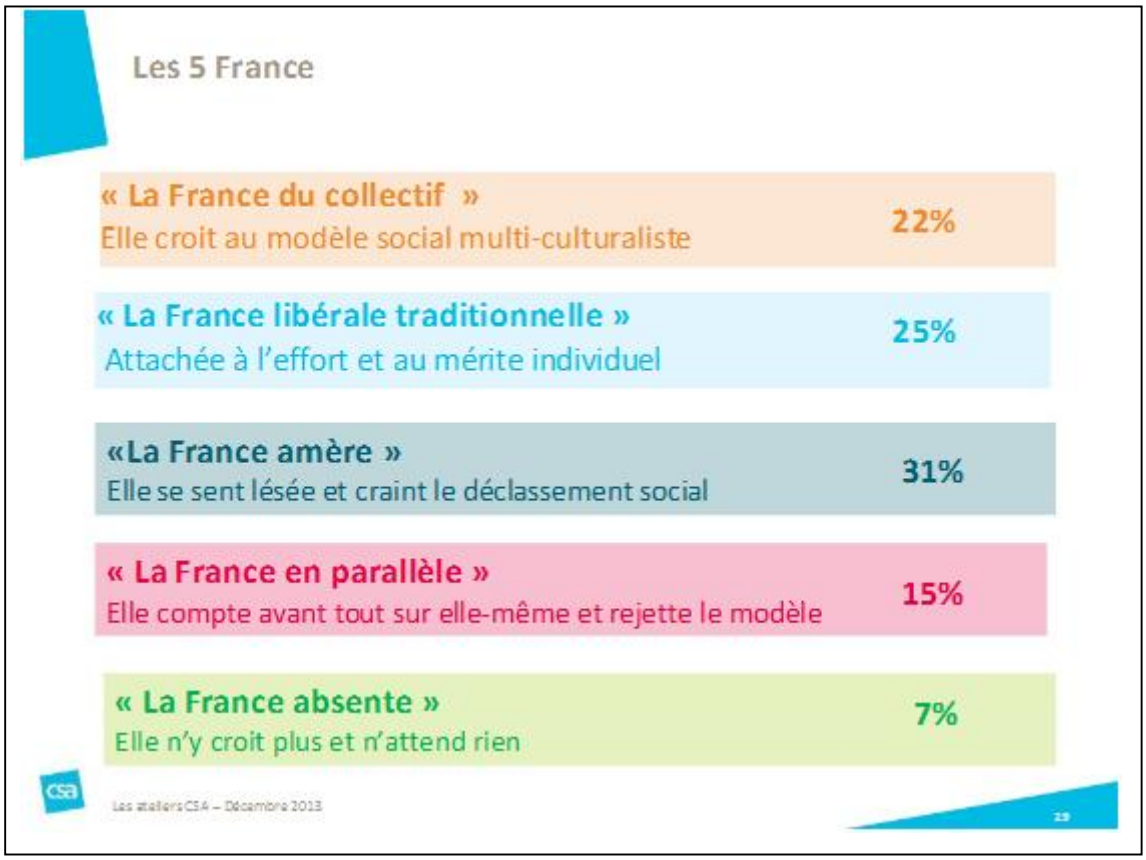


Les néo-militants : 20 % des Franciliens

	Ensemble des Franciliens	Néo-militants
Ce qui les distingue en termes de valeurs		
Il est normal qu'une femme puisse choisir d'avorter	89 %	98 %
L'homosexualité est une manière acceptable de vivre sa sexualité	81 %	98 %
Les couples homosexuels doivent avoir le droit d'adopter des enfants	51 %	92 %
La présence d'immigrés en France est une source d'enrichissement culturel	76 %	98 %
Favorables au droit de vote des étrangers résidant en France depuis plusieurs années aux élections municipales	63 %	97 %
D'accord avec l'idée selon laquelle pour établir la justice sociale il faut prendre aux riches pour donner aux pauvres	49 %	84 %
Il ne faut pas que les inégalités de revenus soient trop importantes car cela crée des tensions (*vs* il faut une assez grande inégalité de revenus en fonction des mérites)	49 %	71 %
La plupart de ceux qui bénéficient des aides sociales en ont vraiment besoin (*vs* n'en ont pas vraiment besoin)	47 %	75 %
Privilégient la solidarité collective (*vs* la responsabilité de chacun)	45 %	68 %
Il faut que l'Etat contrôle et réglemente plus étroitement les entreprises (*vs* leur fasse confiance et leur donne plus de libertés)	48 %	64 %
Il faut plus de libertés pour chacun (*vs* plus d'ordre et d'autorité)	40 %	72 %
Un monde meilleur serait un monde avec plus d'écologie	15 %	29 %
Ce qui les distingue en termes de représentation d'eux-mêmes		
Je vais arriver à m'en sortir	79 %	89 %
Se sent reconnu à sa juste valeur	77 %	89 %
Ce qui les distingue sur le plan sociopolitique		
Intérêt pour la politique	58 %	71 %
Vote F. Hollande au 1er tour 2012	24 %	49 %
Vote JL Mélenchon au 1er tour 2012	9 %	20 %
Sympathisants PS	25 %	46 %
Sympathisants EELV	5 %	12 %
Engagement associatif	29 %	38 %
Pratiques conso collaboratives (covoiturage / vente ou troc)	29 %/33 %	38 %/42 %
Ce qui les distingue en termes sociodémographiques		
Moins de 35 ans	31 %	38 %
Diplômés de l'enseignement supérieur	40 %	59 %
Cadres, professions intellectuelles	24 %	42 %
Salariés de l'Etat ou d'une collectivité locale	25 %	32 %
Revenus > 3 800 €	25 %	33 %
Sans religion	35 %	60 %
Ce qui les distingue sur le plan socio-territorial		
Centre-ville bourgeois	11 %	16 %
Paris branché	10 %	16 %
Paris populaire	7 %	12 %

Exemple 5 « Français, ce qui vous rassemble est-il plus fort que ce qui vous divise ? »

Les ateliers du CSA- Le Monde 2013



1-Individu

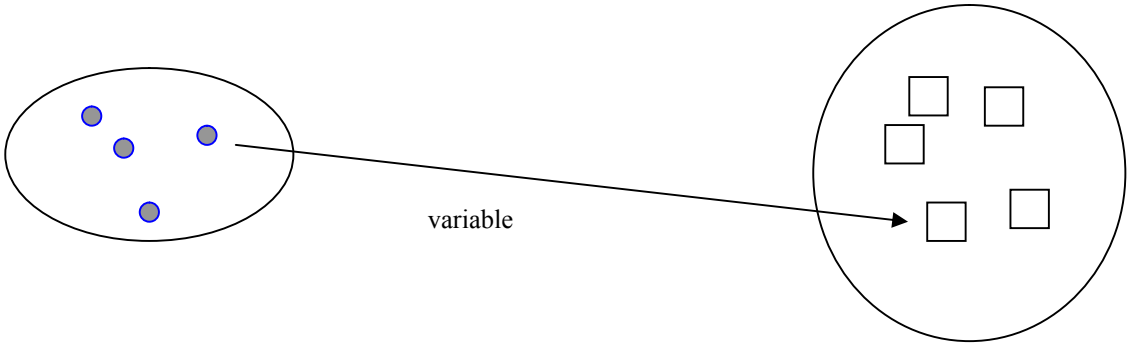
Il faut pouvoir distinguer conceptuellement et pratiquement les individus. Math : éléments d'un ensemble. Informatique : ligne d'un tableau (individu-ligne).

	GENRE	AGE	COULEUR	TABAC
Pierre	homme	23	claire	nonfumeur
Cécile	femme	34	claire	fumeur
Roxane	femme	21	claire	fumeur
Paul	homme	48	foncée	nonfumeur

2-Variable (caractère, indicateur, descripteur….)

une variable ne varie pas (enfin…) c'est une fonction :

I = { individus} ——————————— { modalités}



modalités : des chiffres : la variable est quantitative , numérique (SPAD : *variable continue C*)
modalités : des lettres : la variable est qualitative , texte (SPAD : *variable nominale N*)

enfin presque…
Excel : reconnaissance automatique, éventuellement erronée : le texte à gauche, le nombre à droite de la colonne

- idée de mesure
- fonction : protocoles d'opérations et de mesures
- informatique : colonne d'un tableau

	GENRE	AGE	COULEUR	TABAC
Pierre	homme	23	claire	nonfumeur
Cécile	femme	34	claire	fumeur
Roxane	femme	21	claire	fumeur
Paul	homme	48	foncée	nonfumeur

exemple , enquête

individu	ligne	**questionnaire**
variable	nom d'une colonne	**question**
modalités	cellule	**réponses**

Un questionnaire entier doit tenir sur une ligne. Ce n'est pas immédiat, il faut reconditionner les variables.

Cas réels :

capital privé national	capital privé étranger	forme_juridique	régime_exportation	taille_entreprise
100	0	SA	On_shore	moins_de_50_employés
100	0	SARL	On_shore	50_99_employés
100	0	SA	On_shore	100_employés_et_plus
100	0	SA	On_shore	100_employés_et_plus
33	67	SARL	Off_shore	50_99_employés
100	0	SARL	On_shore	50_99_employés
51	49	SA	On_shore	100_employés_et_plus
100	0	SARL	Off_shore	50_99_employés
100	0	SARL	On_shore	moins_de_50_employés
100	0	SARL	On_shore	moins_de_50_employés
100	0	SARL	On_shore	moins_de_50_employés
100	0	SA	On_shore	100_employés_et_plus
100	0	SARL	Off_shore	50_99_employés
34	66	SARL	Off_shore	100_employés_et_plus
51	49	SA	On_shore	100_employés_et_plus
100	0	SARL	On_shore	100_employés_et_plus

MOIS	TYPEJOUR	CAUSE	LUMINOSITE	INTEMPERIE	SU
Janvier	Lundi	Roule a gauche	Nuit éclairée	Aucune	Se
Janvier	Mercredi	Non respect du pieton en carrefour	Nuit éclairée	Aucune	Se
Janvier	Jeudi	Defaut de maitrise	Nuit éclairée	Aucune	Gra
Janvier	Jeudi	Non respect du pieton en carrefour	Plein jour	Aucune	Se
Janvier	Lundi	Non respect des feux tricolores	Nuit éclairée	Aucune	Se
Janvier	Mardi	Roule a gauche	Demi jour	Aucune	Se
Janvier	Jeudi	Non respect des feux tricolores	Plein jour	Aucune	Se
Janvier	Mercredi	Non respect des feux tricolores	Plein jour	Aucune	Se
Janvier	Jeudi	Non respect d'une balise	Plein jour	Aucune	Gra
Janvier	Samedi	Depassement dangereux	Nuit sans éclairage	Vent fort	

décision	âge	genre	section	décision très risquée	décision vouée à l'échec	verbatim lycéen
contourneur	19	fille	S	non	non	je me suis vraiment très très mal intégrée dans ma classe ainsi que mes camarades de même âge car des clans ont été formé entre doublants et triplants et j'ai été séparé de mes ami(s/es) qui se sont retrouvés dans une autre classe. Le changement de classe
contourneur	17	fille	S	oui	non	Une année bien trop courte avec beaucoup trop de lacunes
contourneur	17	garçon		non	non	
contourneur	18	fille		non	non	Il vaut mieux redoubler la Terminale plutôt que redoubler la première et en cas d'échec au bac, redoubler la Terminale
contourneur	17	garçon		non	non	mauvaise ambiance de classe
contourneur	20	garçon	ES	non	oui	le lycée m'obscurcit l'esprit
contourneur	17	fille	L	non	non	Je pense que ce questionnaire est éroné, car j'ai raté des mois d'école du à un problème de santé. Mis a part ce problème, j'ai toujours basé ma scolarité sur ce petit mot "quand-t-on veut on peut"

Enquêtes, sondages : choses diverses et ficelles du métier

- une variable ne varie pas : la question doit être la même (avoir le même sens !) pour chaque individu interrogé, c'est la réponse qui varie d'un individu à l'autre. Evidemment, on sait que ça n'est pas le cas.
- un poids et deux mesures
- nommer une variable : identifiants mnémotechniques, unités de mesures (VAk€, Pluiemm, Norages),
- jugements de valeur , INED : « *naissance illégitime* »
- Sexe ? Genre ? civilité
- enquête INED auprès des SDF : » *Comment vous êtes vous procuré votre repas hier soir ?* »
- questions : la violence *« tu préfères ton père ou ta mère ?* »,
- classer, c'est juger : catégorie, du grec katêgoria « accusation » , katêgorein « parler contre, accuser, blâmer…). (jugement catégorique)
- aux USA on distingue officiellement 63 catégories raciales. *Federal directive Race and Ethnic Standart for Federal Statistics and Administration Reporting – 1977*. des statistiques ethniques en France?
- *Selon l'étude de l'Université de Standford, 40% des Américains blancs sont d'accord avec au moins un adjectif négatif (violent, paresseux, râleur, irresponsable, fanfaron,…) pour qualifier les Noirs.*

- les questions : « le roi-de-France-est-il –chauve ? »

  « avez-vous déjà bénéficié d'une acupuncture ? », laïcité positive

* la technique du « fou ». exemples

* ajouter une réponse possible peut modifier l'ordre de préférence sur les autres
Voulez-vous partir maintenant ou reprendre un thé ? Voulez-vous partir maintenant ou reprendre un thé, ou de la cocaïne ?

- on a le droit de mentir, pas le droit de nuire. exemples-
- ce qu'on dit à l'oral, ce qu'on écrit : pourquoi les déclarations de sinistres aux assurances se font de plus en plus au téléphone.
- Les réponses n'engagent à rien
- Le coût des réponses n'est pas le même (une réponse gratifiante , consensuelle, conventionnelle etc est plus facile)…exemples

- Instituts de sondages, ce ne sont pas des instituts mais des entreprises privées . parties prenantes[8].
- « *un sondage, ça s'achète !* »

- « *push polls* « que l'on peut traduire par « sondage d'influence ». Primaires Parti Républicain USA 2000, Caroline du Sud , G W.Bush / MacCain « *Seriez-vous plus enclin ou moins enclin à voter John Mac Cain comme Président si vous saviez qu'il est le père illégitime d'un enfant noir ?* ».
- En 1948 les Instituts de sondage sont passés devant une commission du Sénat des USA. Ils avaient donné, à tort, les républicains vainqueurs.

- La question cruciale de l'anonymat. Garantie de l'Etat. Exemple fichier RMI.INSEE *« un accident de parcours dans l'anonymat se paye cher et pendant longtemps* ».

- Journée « Qualité dans les enquêtes » à la SFdS : « les questionnaires sont trop longs + trop complexes,+ **inadéquation questionneur/répondants »**

Les questions » scandaleuses » génèrent des sous et des sur-déclarations. Exemples.

<u>La qualité dans les enquêtes :</u>

- scène /obscène , ce que dit un serveur dans la salle de restaurant/cuisines
- on ne se comprend pas
- « quand un français dit 20h ça veut dire 20h30 » (un homme, français, à son épouse, allemande)
- questions très difficiles ; allume-cigare dans les automobiles, airbag
- les sondages : « qu'*est-ce qu'on dit au client ?* » dire la vérité au patient ? nous préférons une belle histoire. La gifle de Bayrou.
- il faut emballer le produit

<u>pourquoi c'est pas aussi bon que ça pourrait l'être ?</u>

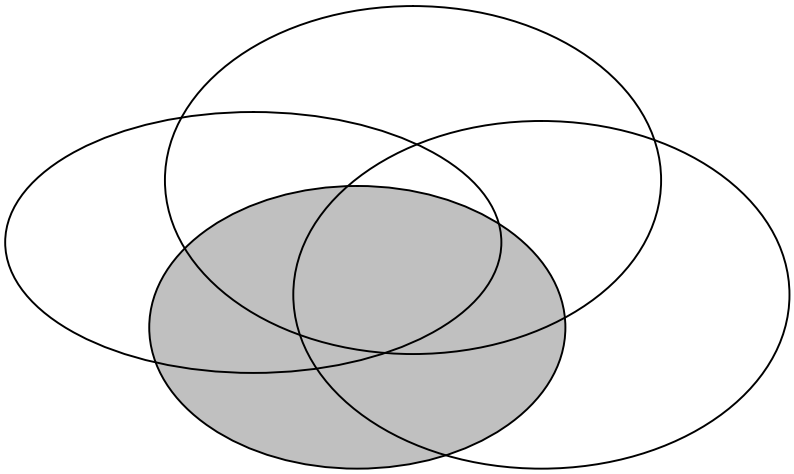
1. baisse de notre chiffre d'affaires (« *on dit oui à tout ce que demande le client »*)
2. *« garder la maîtrise »*
3. *« le client ne sait pas ce qu'il achète, il prend le moins cher »*

[8] « *Non, nos sondages ne sont pas trafiqués. OpinionWay ne dérange que les envieux* « Hugues Cazenave et Denis Pingaud, Président et Vice-Président d'Opinion Way in Le Monde du 31 juillet 2009 puis « *Si je me reconnaissais dans ce portrait des sondeurs, je choisirais un autre métier…* » Yannick Carriou, Directeur Général de TNS Sofres et « *Question de factures, pas de méthode, OpinionWay se victimise à bon compte* » Jean-Marc Lech, Coprésident d'IPSOS le 5 août 2009.

Dans la vie : 1-on ne mesure pas ce qu'on veut (indicateurs…) , 2-les données sont incomplètes (c'est subi : les missing data, ou voulu : les échantillons ⟶ inférences statistiques ..) 3-et le peu de données que l'on a sont « contaminées ».

It's not a bug , it's a feature !

1- La consommation de fuel lourd est un indicateur d'activité industrielle
- le nombre de navettes de la benne à ordures est un indicateur de fréquentation touristique…
- ménagères de moins de 50 ans

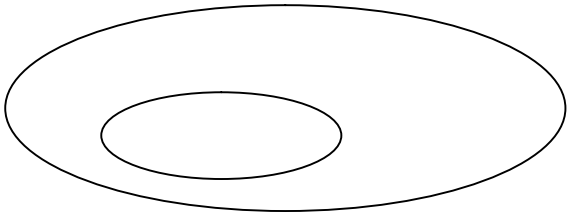


- comment trouver les » clients potentiels pour un fusil de chasse haut de gamme « ? (cas réel)
-

- **<u>Indicateur/marqueur très utile : l'âge</u>**

25 ans ? pas propriétaire de votre logement, contraintes d'emploi du temps faibles, financières fortes ,etc
45 ans ? propriétaire, contraintes d'emploi du temps fortes, financières fortes, etc
65 ans ? propriétaire , contraintes d'emploi du temps faibles, financières faibles, etc

2- échantillon / population (voir cahier tiré à part sur les sondages)



examen 2004 1 *– données

on peut trouver, dans le Oxford-Dictionary of statistics, à l'entrée « dirty data » :
dirty data *: Data with unusual, missing or incorrect values. Most real sets of data are dirty*.
Ce qui peut être traduit :données sales : données avec des valeurs étranges, manquantes ou erronées. La plupart des données sont sales.
Que comprenez-vous ?

Missing data :

Un exemple classique de tableau avec des données manquantes :

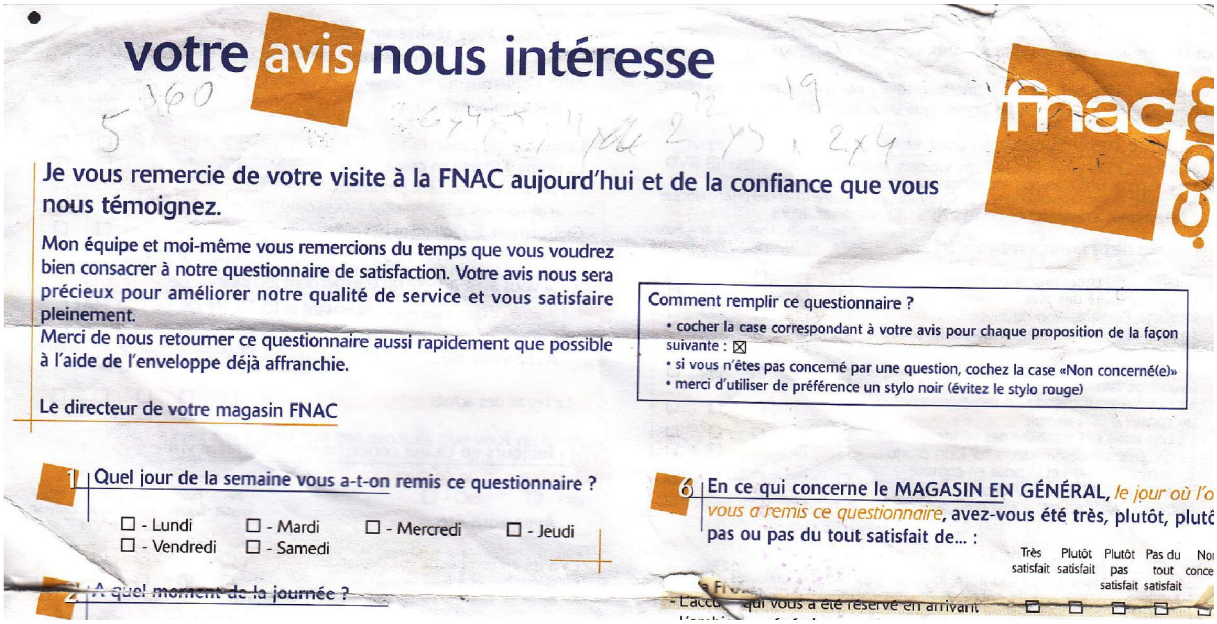
unit	Income	age	car
1		young	am
2	medium	medium	am
3		old	jap
4	low	young	jap
5	medium	young	am
6	high	old	am
7	low	young	jap
8	high	medium	am
9	high		am
10	low	young	am

Repris dans « *Fusion et greffes de Données* « G. Saporta , N. Fisher in La revue Modulad n°27, juin 2001

Pour un statisticien, **les non-réponses sont a priori des réponses comme les autres,** il y a mille interprétations possibles, qui ne sont pas de son ressort (oubli, lassitude, contestation de la question, interrogation, incompréhension, non, doute….).

En cas de donnée manquante, il faut laisser la case vide.

Des remarques :



votre avis nous intéresse

fnac

Je vous remercie de votre visite à la FNAC aujourd'hui et de la confiance que vous nous témoignez.

Mon équipe et moi-même vous remercions du temps que vous voudrez bien consacrer à notre questionnaire de satisfaction. Votre avis nous sera précieux pour améliorer notre qualité de service et vous satisfaire pleinement.
Merci de nous retourner ce questionnaire aussi rapidement que possible à l'aide de l'enveloppe déjà affranchie.

Le directeur de votre magasin FNAC

Comment remplir ce questionnaire ?
- cocher la case correspondant à votre avis pour chaque proposition de la façon suivante : ☒
- si vous n'êtes pas concerné par une question, cochez la case «Non concerné(e)»
- merci d'utiliser de préférence un stylo noir (évitez le stylo rouge)

1 Quel jour de la semaine vous a-t-on remis ce questionnaire ?

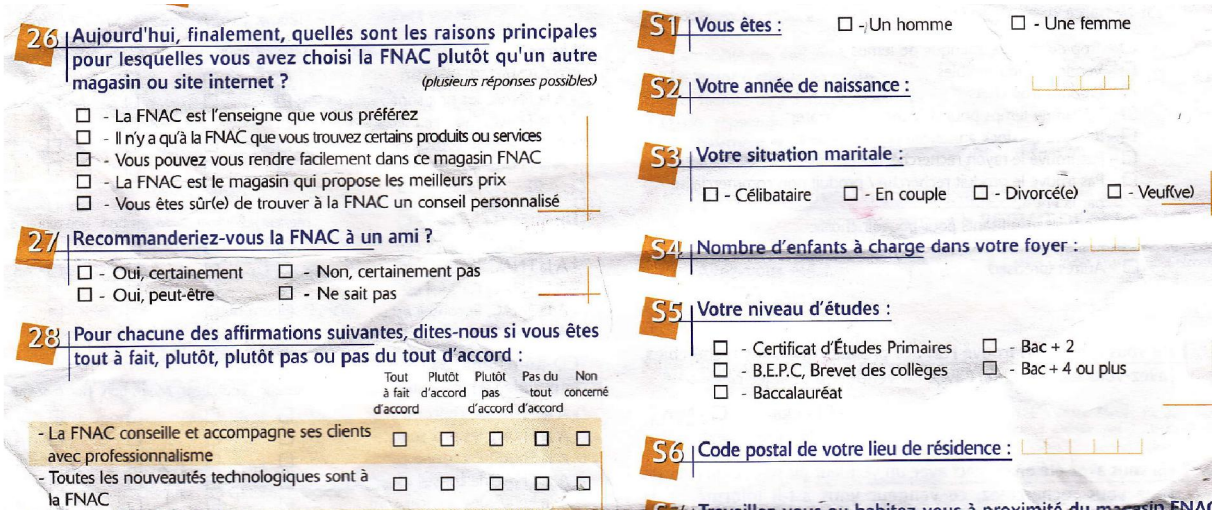
☐ - Lundi ☐ - Mardi ☐ - Mercredi ☐ - Jeudi
☐ - Vendredi ☐ - Samedi

2 A quel moment de la journée ?

6 En ce qui concerne le MAGASIN EN GÉNÉRAL, *le jour où l'o vous a remis ce questionnaire*, avez-vous été très, plutôt, plutô pas ou pas du tout satisfait de... :

Très satisfait | Plutôt satisfait | Plutôt pas satisfait | Pas du tout satisfait | Non conce

- L'accu qui vous a été réservé en arrivant



26 Aujourd'hui, finalement, quelles sont les raisons principales pour lesquelles vous avez choisi la FNAC plutôt qu'un autre magasin ou site internet ? *(plusieurs réponses possibles)*

☐ - La FNAC est l'enseigne que vous préférez
☐ - Il n'y a qu'à la FNAC que vous trouvez certains produits ou services
☐ - Vous pouvez vous rendre facilement dans ce magasin FNAC
☐ - La FNAC est le magasin qui propose les meilleurs prix
☐ - Vous êtes sûr(e) de trouver à la FNAC un conseil personnalisé

27 Recommanderiez-vous la FNAC à un ami ?

☐ - Oui, certainement ☐ - Non, certainement pas
☐ - Oui, peut-être ☐ - Ne sait pas

28 Pour chacune des affirmations suivantes, dites-nous si vous êtes tout à fait, plutôt, plutôt pas ou pas du tout d'accord :

	Tout à fait d'accord	Plutôt d'accord	Plutôt pas d'accord	Pas du tout d'accord	Non concerné
- La FNAC conseille et accompagne ses clients avec professionnalisme	☐	☐	☐	☐	☐
- Toutes les nouveautés technologiques sont à la FNAC	☐	☐	☐	☐	☐

S1 Vous êtes : ☐ - Un homme ☐ - Une femme

S2 Votre année de naissance :

S3 Votre situation maritale :

☐ - Célibataire ☐ - En couple ☐ - Divorcé(e) ☐ - Veuf(ve)

S4 Nombre d'enfants à charge dans votre foyer :

S5 Votre niveau d'études :

☐ - Certificat d'Études Primaires ☐ - Bac + 2
☐ - B.E.P.C, Brevet des collèges ☐ - Bac + 4 ou plus
☐ - Baccalauréat

S6 Code postal de votre lieu de résidence :

Les Clermontois majoritairement satisfaits

ualité du service public municipal

uête d'opinion menée du 22 au 27 juin 2011

▲ En 2011, décision a donc été prise de faire réaliser par un organisme spécialisé une enquête d'opinion sur la qualité du service public municipal.
En l'occurrence par l'Ifop, retenu à l'issue d'un appel à la concurrence, auteur de nombreuses enquêtes pour le compte de villes de toutes sensibilités politiques et chargé par la Ville de Clermont-Ferrand de mesurer la perception instantanée du niveau de qualité du service public municipal et du CCAS mais aussi celle de son évolution au cours des dernières années.
Et ce, dans treize domaines :

- Gestion de la ville et image de Clermont : utilisation des impôts locaux, travail accompli, modernisation de la ville.
- Environnement et cadre de vie : espaces verts, propreté, accessibilité.

Renseignements vous concernant

40 Quel âge avez-vous ?
Moins de 20 ans ☐ — 40 à 49 ans ☐
20 à 29 ans ☐ — 50 à 59 ans ☐
30 à 39 ans ☐ — 60 ans et plus ☐

41 Vous êtes ?... Un homme ☐ Une femme ☐

42 Dans votre foyer, qui exerce une activité professionnelle ?

	Oui	Non
• Vous-même	☐	☐
• Votre conjoint(e)	☐	☐

43 De combien de personnes se compose votre foyer, y compris vous-même ?
1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 et plus ☐

44 Avez-vous des enfants dans votre foyer ? Oui ☐ Non ☐
Si oui, merci de nous indiquer le nombre d'enfants appartenant à chacune des tranches d'âge suivantes :
3 ans et moins — De 4 à 5 ans inclus — De 6 à 10 ans inclus — De 11 à 14 ans inclus — De 15 à 17 ans inclus — 18 ans et plus

N° carte Auchan Waaoh 0 4 9

Remarques particulières concernant votre dernière visite dans ce magasin Auchan :

Vous pouvez également contacter votre magasin Auchan sur www.auchan.fr - en cliquant sur le lien | Contactez-nous | en bas de page.

Si vous souhaitez que le magasin apporte une réponse personnelle à vos remarques, merci d'indiquer vos coordonnées ci-dessous*.

Nom..... Adresse ..... E-mail:.....
Prénom..... Code Postal ..... Ville ..... Téléphone:.....

Voilà, c'est terminé ! Votre magasin Auchan vous remercie d'avoir bien voulu répondre à cette enquête.

*Vous disposez d'un droit d'accès, de modification, de rectification et de suppression des données vous concernant (loi «Informatique et Libertés» du 6 janvier 1978). Pour toute demande, adressez-vous à : SOFTCOMPUTING - 55 Quai de Grenelle - 75015 Paris

Soucieux de l'amélioration continue de nos prestations et afin de répondre au mieux à vos attentes. La SNCF vous demande votre avis sur l'information fournie en gare et la visibilité de nos agents sur le terrain.

Date :	/ Heure :	
	☐ Homme	☐ Femme
Âge	☐ 18 à 29 ans ☐ 30 à 44 ans ☐ 45 à 59 ans ☐ 60 à 74 ans ☐ 75 ans et plus	
CSP	☐ Artisan/commerçant ☐ Ouvrier ☐ Retraité ☐ Profession intermédiaire	☐ Autre ☐ Employé ☐ Cadre/chef d'entreprise ☐ Etudiant
Fréquence	☐ Occasionnel ☐ Fréquent ☐ Régulier	
Type de voyage	☐ Professionnel ☐ Loisir ☐ Les deux	
Train	☐ Transilien ☐ IC TER /Intercités ☐ TGV	

I - L'information :

Q1.Trouvez-vous les annonces sonores en gare, audibles ?
☐ Oui ☐ Non

Q2.Avez-vous déjà vécu une situation perturbée, en gare de Paris Est ?
☐ Oui ☐ Non => merci de répondre seulement de la question 9 à 12.

Q3.Quelle est votre dernière situation perturbée, en gare de Paris Est ?

Q4.Combien de temps avez-vous attendu avant d'avoir une information ?

Q5.Avez-vous été informé de la cause du retard ?
☐ Oui ☐ Non

Q6.Avez-vous été informé de la durée du retard ?
☐ Oui ☐ Non
Si oui, le retard réel correspondait-il à la durée du retard énoncé ?
☐ Oui ☐ Non

Q7.Quelles sont vos attentes en matière d'information, en situation perturbée ?

Marie-Claire juin 2015

8 Quelle est votre année de naissance ? 19... 62

9 Quelle est votre profession et, éventuellement, celle de votre conjoint, ou celle de vos parents si vous êtes étudiante ?

	LA VÔTRE 64	VOTRE CONJOINT 65
Agriculteur	1	1
Commerçant, artisan, chef d'entreprise	2	2
Cadre supérieur, profession libérale	3	3
Cadre moyen, technicien, contremaître	4	4
Employé	5	5
Ouvrier	6	6
Etudiant, à la recherche d'un premier emploi	7	7
Retraité, préretraité	8	8
Sans profession	9	9
Pas de conjoint		10

Je tiens une chronique pince-sans-rire sur le site OVIPAL (Observatoire de la vie politique en Alsace)
www.ovipal.com

Oh ! les chiffres, moi vous savez ...le point de vue d'un statisticien

Jean-Paul Villette, Maitre de conférences, Université de Strasbourg

1- lire l'étiquette d'un sondage

un exemple : sondage BVA pour Le Parisien-Aujourd'hui en France publié dimanche 16 mars 2014.Dans cette étude, réalisée les 13 et 14 mars auprès d'un échantillon de 997 personnes représentatif de la population française âgée de 18 ans et plus (méthode des quotas), recrutées par téléphone et interrogées par internet,

- **combien de personnes ont répondu ?**

Impossible de savoir : « auprès « est une expression poétique. Il faut des milliers d'appels pour obtenir que 1000 personnes décrochent et ne raccrochent pas aussitôt. L'application de filtres fait facilement tomber le nombre de répondants à quelques centaines. Exemple de filtre, pour l'Election présidentielle de 2012, un premier filtre était d'avoir voté en 2007 et le deuxième d'être certain d'aller voter en 2012.

Ici on ne sait rien.

- **pourquoi 997 ?**

C'est du marketing, ça fait sérieux sans doute. On peut obtenir des résultats utiles avec de très petits échantillons, c'est souvent le cas dans l'industrie, dans les sciences de la vie. Par exemple, des scientifiques ont trouvé que tous les individus peuvent être *classés* en *3* groupes distincts, 3 « entérotypes »,selon la nature des *bactéries* hébergées par le tube digestif**...** avec des échantillons de quelques dizaines d'individus (http://www2.cnrs.fr/presse/communique/2165.htm).

- échantillon représentatif ?

Ça ne veut rien dire. Représentatif (un synonyme est « proportionnel ») de quoi ? du genre ? alors la proportion hommes/femmes de l'échantillon serait celle de la population, de l'âge… ? Quelle est la « variable de contrôle » ?

Pour un statisticien, l'expression un « échantillon représentatif » , avec des guillemets, c'est à dire des pincettes, signifie qu'il permet, par des formules de calcul complexes d'avoir une estimation sans biais systématique de la variable mesurée. Il y un biais (une erreur, un écart entre la vraie valeur et la valeur estimée) mais il n'est pas systématique, c'est à dire qu'en gros la surestimation est aussi probable que la sous-estimation

- **des quotas ?**

On ne dit pas des quotas de quoi. Il y a une bonne idée et une bonne pratique. Qu'un individu ne soit pas tiré complètement au hasard, mais avec une probabilité conditionnelle à son genre, la taille de son agglomération, sa région …permet d'obtenir une estimation de bien meilleure qualité, en redressant les estimations par des coefficients basés sur la population. Voici la formule la plus simple avec une seule variable de contrôle :

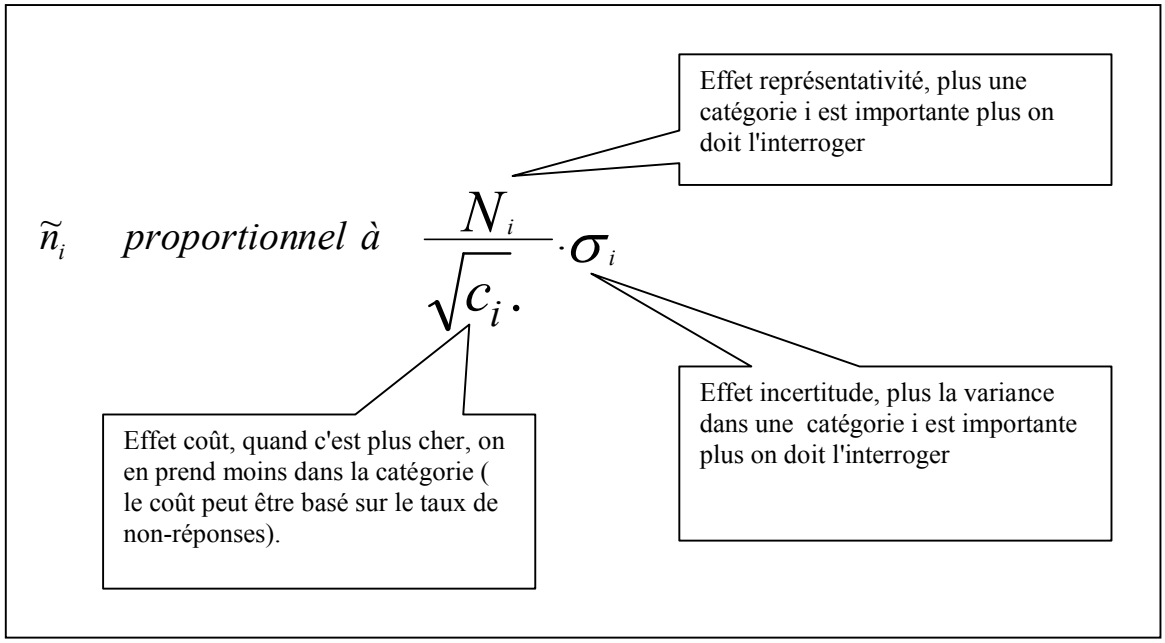
Une **variable de contrôle** définit des catégories i=1,…I. Dans la catégorie i, dans la population, il y a N_i individus et n_i individus dans l'échantillon. Un plan de sondage est (n_1,n_2,…,n_i,….n_I).La valeur de la variable X pour l'individu k de la catégorie i est $x_{i,k}$.

Quel que soit le plan de sondage, on peut calculer une estimation sans biais x* de x.

Estimation de la valeur moyenne dans la population: $\bar{x}^* = \frac{1}{N}\sum_{i=1}^{I} \frac{N_i}{n_i}\sum_{k=1}^{n_i} x_{i,k}$

Dans la pratique on ne respecte pas les quotas, c'est à dire à dire que les proportions dans l'échantillon de différentes variables de contrôle ne sont pas celles de la population parce que

1- ce n'est pas possible ou alors ce serait extrêmement couteux (on peut chercher longtemps l'agriculteur de 30 à 50 ans qui nous manque)
2- cela n'est pas nécessaire car les formules corrigent automatiquement les sur- ou sous-représentativités dans l'échantillon
3- surtout, cela n'est de toute façon pas souhaitable , on obtient de bien meilleures estimations en appliquant la « règle d'allocation de Neyman ».



alors un institut de sondage peut dire n'importe quoi ?

oui la liberté est totale. Il n'est même pas obligé d'interroger qui que ce soit.

mais il y existe une « commission des sondages «

oui, mais elle n'est, en gros, concernée que par les » sondages publics d'intention de vote susceptibles d'altérer la sincérité du scrutin ». Voir son site

http://www.commission-des-sondages.fr/

Un Institut de sondage qui dirait n'importe quoi, ça se saurait. Il perdrait toute crédibilité

Des enquêtes montrent que la confiance envers les Instituts de sondage est globalement faible Est-ce qu'un Institut qui fait un coup tordu risque d'en pâtir ? J'ai entendu cette question , posée à un responsable d'un grand institut de sondage à propos de l'IFOP, qui avait fait fort en 2012 (N. Sarkozy en tête dans les intentions de vote , sondage pour le Figaro). Je ne sais pas ce qu'il en est. J'ai compris de sa réponse que tous les instituts en pâtissent, pas spécialement le fautif.

A suivre

PS- pour ceux qu'intéresse l'obscure clarté des nombres, un site de statisticiens : http://www.penombre.org/

Oh ! les chiffres, moi vous savez ...le point de vue d'un statisticien

Jean-Paul Villette, Maitre de conférences, Université de Strasbourg

3- la différence entre un caillou et une intention de vote

- cible mouvante

On met un caillou sur une balance : 23 g. On peut le crier sur tous les toits, les 23 g n'en seront pas modifiés.

Les instituts des sondages ont évalué les intentions de vote pour les listes « Bleu-Marine » : 23%. En tête. Au moment de voter tout le monde connaissait ces chiffres. Alors il peut généralement se passer deux choses chez tous les électeurs, qu'ils aient l'intention de s'abstenir ou pas .

- l'effet « band-wagon » et l' »effet underdog »

L'effet ***bandwagon*** [9] , c'est la mobilisation en faveur du candidat placé en tête qui, dans notre exemple, aurait favorisé les listes Bleu-Marine. L'effet ***underdog*** , c'est au contraire la mobilisation pour les autres candidats. On peut identifier un effet ou l'autre dans différentes élections, notamment aux USA.

Les instituts sondages ont une communication sans faille nécessaire à leur combat pour nous convaincre de leur impartialité : les effets s'annulent .

- et en réalité ?

Je ne sais pas. Ceux qui volent au secours de la victoire sont généralement plus nombreux que les défenseurs d'une cause perdue.

- Et donc , pour les élections européennes de mai 2014 ?

Un résultat de 25%, c'est conforme aux prévisions. S'il y a eu des mouvements, ils se sont compensés. Un aspect statistique, si j'ose dire, de la question, c'est que les 27 millions d électeurs qui se sont abstenus n'ont pas manifesté d'objection à ce que les listes « Bleu-Marine » arrivent en tête. Cela ne les dérange pas. Cela les arrange ?

Jean-Paul Villette

PS- pour ceux qu'intéresse l'obscure clarté des nombres, un site de statisticiens : http://www.penombre.org/

[9] ***Band-wagon*** :*An activity that more and more people are becoming involved in.* ***under-dog*** *: a person, team, country,etc. that is thought to be in a weaker position than others and therefore not likely to be succesfull..*
(Oxford dictionary)

Concrètement : définir une variable quantitative-numérique

A chaque individu est associé un nombre

	âge
Pierre	23
Cécile	34
Roxane	21
Paul	48

Concrètement : définir une variable qualitative[10]

Les modalités doivent être exhaustives (à chaque individu correspond au moins une modalité, il faut avoir quelque chose à mettre dans la cellule) et exclusives (à chaque individu correspond au plus une modalité, faut pas avoir l'embarras du choix, une seule réponse possible).

Cela ne se fait pas tout seul. Il faut intervenir :

{France, Alsace, Autre}

	genre
Pierre	homme
Cécile	femme
Dominique	
Paul	homme

exhaustivité : créer une classe des inclassables : « autres », « divers », » nsp »…
exclusivité : l'illusion autoritariste : une seule réponse possible , plus subtil : « quel est le motif principal de votre déplacement . familial, professionnel …

Solution statistique :

Quelle langue parlez-vous ?

anglais : ☐
allemand : ☐
italien : ☐
autres : ☐

	anglais	allemand	italien	autre
Pierre	oui	oui	non	non

[10]« *variable catégorique* » en français du Québec, en anglais « *categorical variable* »

examen 2003, 2004

2 – question

Dans un questionnaire on trouve :

2. Quelles sont les raisons qui vous ont conduit à vous présenter à la Sous-Préfecture ?
- ☐ 1. Une réunion
- ☐ 2. dépôt de dossier ou délib
- ☐ 3. Rdv
- ☐ 4. Recherche d'information

Indiquez les réponses en cochant une ou plusieurs cases

Pourquoi est-ce que les réponses ne sont pas exclusives ? Que peut-on faire pour transformer cette variable en questions-variables à modalités-réponses exclusives ?

3- encore une question

sur le site www.expression-directe.com on peut trouver la question :

A propos de l'élargissement de l'Union européenne à dix nouveaux pays, quelle est parmi ces opinions celle qui se rapproche le plus de la vôtre ?
c'est une chance pour l'Europe tout entière
c'est un devoir historique et moral à l'égard de ces pays
c'est prématuré sans avoir fait au préalable la réforme des institutions
c'est une mauvaise chose pour la France
sans opinion

est-ce que les modalités réponses sont exclusives ? exhaustives ? significations concrètes.

3 – bonnes fêtes, mamans

Dans l'excellent livre de l'humoriste Alain Schriffres « *Ceux qui savent de quoi je parle comprendront ce que je veux dire* « . Laffont 1986, on peut lire, dans l'article « *bonnes fêtes , mamans* »

Pour le savoir, nous avons fait procéder à un sondage exclusif auprès d'un échantillon représentatif des préaux d'école… A la question, bouleversante de simplicité : « Pourquoi aimes-tu ta maman ? », quarante-huit citoyens de six à huit ans répondent de la façon suivante :

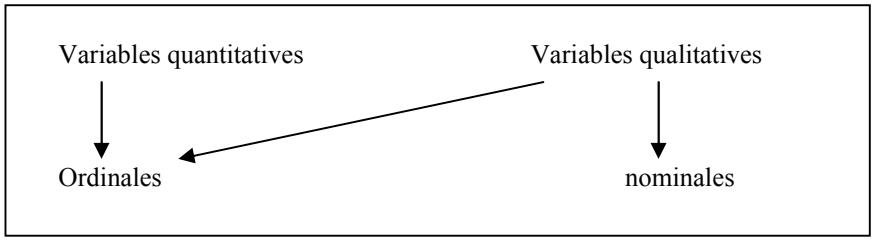
Parce qu'elle est gentille……..27
Parce qu'elle est belle……..12
Parce qu'elle est frisée……1
Parce qu'elle m'aime……6
Parce qu'elle travaille……4
Parce que c'est ma maman……2
Parce que c'est normal……..1
Parce qu'elle ne se met pas beaucoup de maquillage…1

2-a quelle est la « variable » ?
2-b les modalités-réponses sont-elles exhaustives ? Exclusives ? Significations concrètes.

Ah ! les variables ordinales !

On peut penser un ordre dans les modalités . Le botaniste-statisticien ne sait généralement pas, c'est au fleuriste-usager de décider. Le débat … *hot potato*

Petit, moyen, grand
Bronze, argent, or
Pas du tout, un peu , beaucoup
Homme, femme
(*« A Neufchâtel (Seine-Inférieure), on annonce le trépas par quinze coups de cloche pour un homme, douze pour une femme, et six pour un garçon et une fille. A Bully, à Esclavelles et à Bures, on tinte treize coups pour un homme, onze pour une femme et … sept pour les enfants. »*[11]
Extrême-Gauche, Gauche, Centre, Droite, Extrême-Droite
Lower-class, middle-class, Upper-Class
LLL, LLM,LLU,LML,…..MMM,……..,UUU
Les couleurs…
18-24 ans, 25-34 ans,….



<u>Ordinalité métier :</u>

Séminaire Société Française de Statistique :

Est-ce que cette variable est ordinale ?

policier
secteur public
retraité
secteur privé
artisan
étudiant

[11] Alain Corbin in « *Les cloches de la terre .Paysage sonore et culture sensible dans les campagnes au XIX ème siècle* ». Albin Michel

C’est du qualitatif ou du quantitatif ? points de vue

température ? sexe ? couleur ? salaire ? enfants ? tués ?

		Statisticien- qualitatif	botaniste quantitatif
usager-fleuriste	qualitatif	*	(i)
	quantitatif	(ii)	*

Ne pas perdre de vue, la distinction et les rapports entre :

- botaniste (statisticien) : description,
- fleuristes : (économiste, gestionnaire..) : jugements de valeurs, d’utilité (la pertinence..)

variable qualitative ⟶ variable(s) quantitative(s)

1- si la variable est ordinale pas du tout, un peu, beaucoup

échelle : 1 2 3

L'intervalle entre les barreaux peut varier :

Extrême-Gauche,	Gauche	Centre	Droite	Extrême-Droite

Piscine :

c'est facile	des amis ou des voisins m'ont aidé	j'ai fait appel à un artisan

« A Neufchâtel (Seine-Inférieure), on annonce le trépas par quinze coups de cloche pour un homme, douze pour une femme, et six pour un garçon et une fille. A Bully, à Esclavelles et à Bures, on tinte treize coups pour un homme, onze pour une femme et … sept pour les enfants. »[12]

homme	femme	enfant

2- si la variable a deux modalités { homme, femme}, { oui, non }, «{ pièce défectueuse, pièce non défectueuse }

	sexe(h)
Pierre	1
Cécile	0
Roxane	0

3-de manière générale I ⟶{ jaune, violet, rouge }

on substitue des variables quantitatives 0/1 les modalités deviennent des variables (aie !).

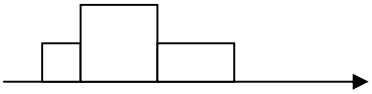
	couleur	**jaune**	**violet**	**rouge**
I1	jaune	1	0	0
I2	violet	0	1	0
I3	violet	0	1	0

Précautions d'emploi, abus dangereux…c'est fait automatiquement par le logiciel (Tableau Disjonctif complet voir page xx).

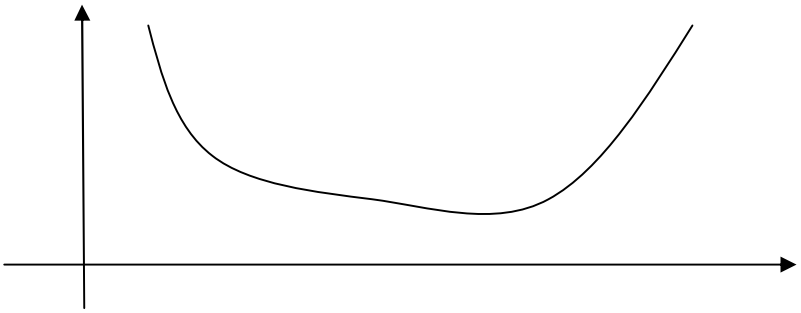
[12] Alain Corbin in « *Les cloches de la terre .Paysage sonore et culture sensible dans les campagnes au XIX ème siècle* ». Albin Michel

variable quantitative ⟶ variable qualitative ordinale

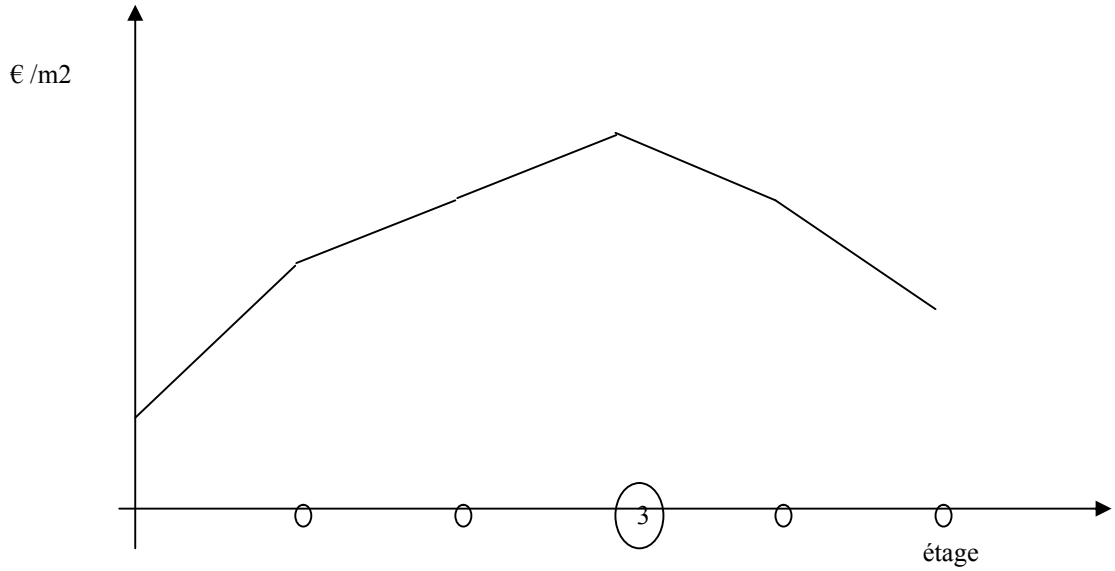
les enjeux…
découper en classes, en tranches , pour un diagramme différentiel



capturer des phénomènes de non-linéarité, non monotonie de jugement, exemples..

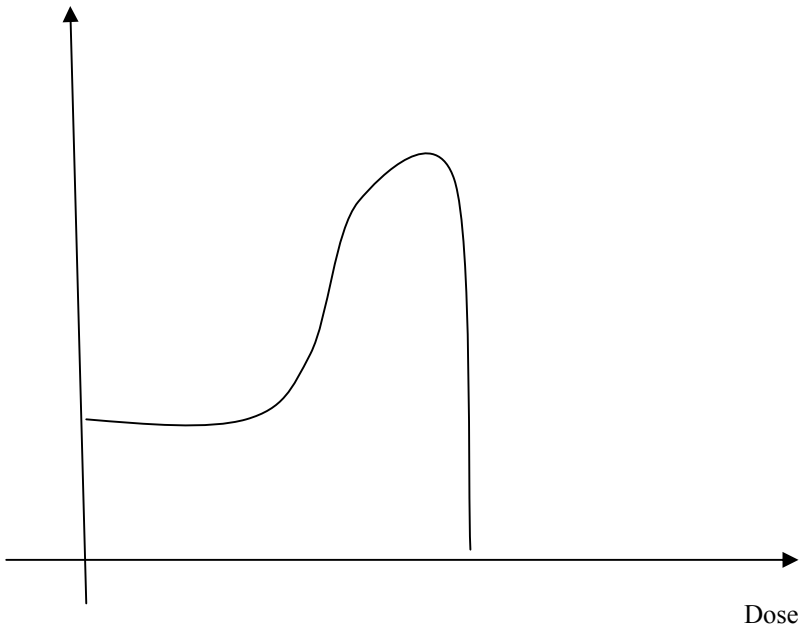


Prix du m2 dans un immeuble Haussmannien à Paris en fonction de l'étage

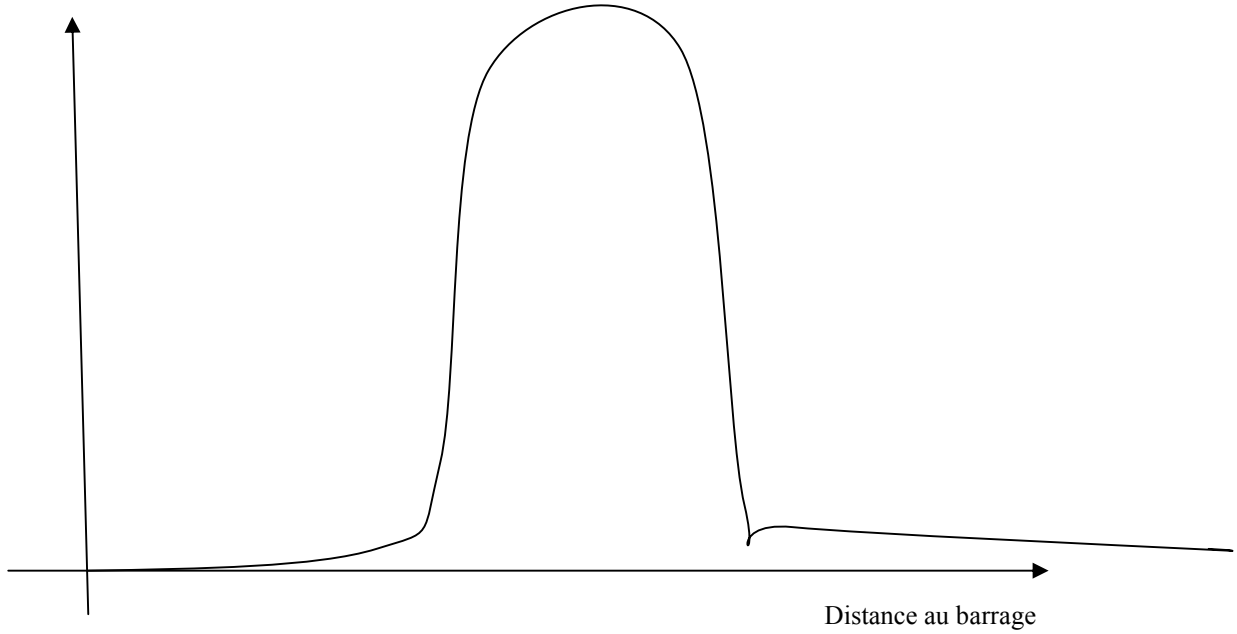


découpage statistique/métier

taux de guérison



Taux de personnes inquiètes



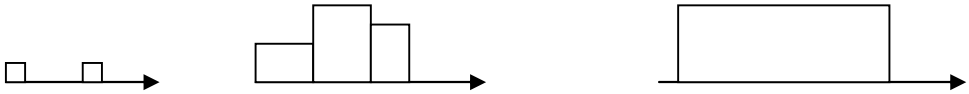
Pratiquement : gérer l'exhaustivité, l'exclusivité

FF/CA%	effectif	fréquence
1 2		
2 3		

Combien de classes ? la finesse du découpage.

i-**<u>découpage métier</u>** :Pertinence, usager classes d'ages pour Air France, pour un démographe..

ii- **<u>découpage du statisticien</u>** :statisticien , visibilité , découpage optimal ?



trop fin k ? trop grossier

ficelles : nombre de classes = √ nombre d'individus
nombre de classes = $1 + \text{Log}_2$(nombre d'individus)

Il ne faut pas créer de modalités trop ou trop peu fréquentées[13] . Un bon choix (en AFCM, dans les sondages,…etc…) <u>est de créer des classes de fréquences égales</u> (et par conséquence d'amplitudes inégales). Dans SPAD menu *outils>quantiles* .

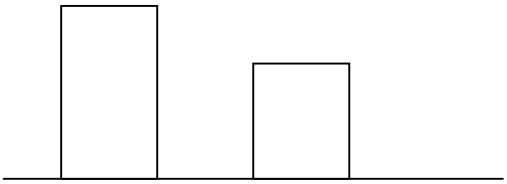
[13] Un plaisanterie de statisticien : « *la fréquence précède l'existence* » (Cf Jean-Paul Sartre, « *l'existence précède l'essence* » in « *l'Existentialisme est un humanisme* », 1946.

Traitement d'une variable qualitative : Classer / Compter / Comparer , représentation graphique

Remarque : toutes les variables sont qualitatives, par définition, ou après transformation.

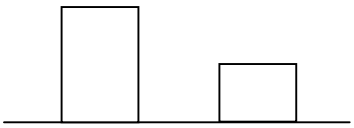
genre	effectif	fréquence %
homme	11356	63
femme	7257	37
	19613	100

Combien de chiffres derrière la virgule ? aucun[14]. La nuance ne doit pas faire perdre du vue la couleur.



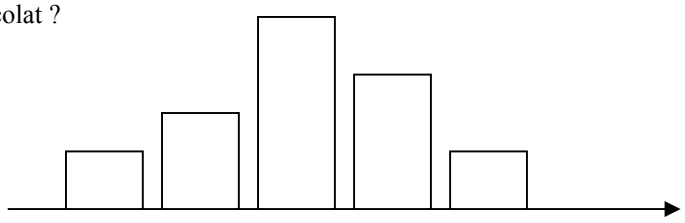
traits, formes, couleurs : pictogrammes

variable qualitative :

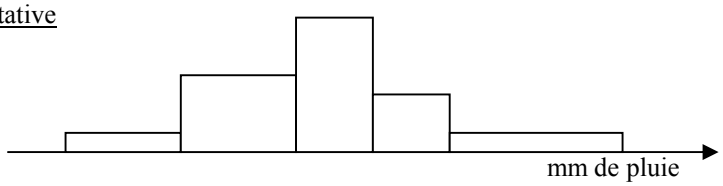


variable qualitative ordinale

vous aimez ce chocolat ?

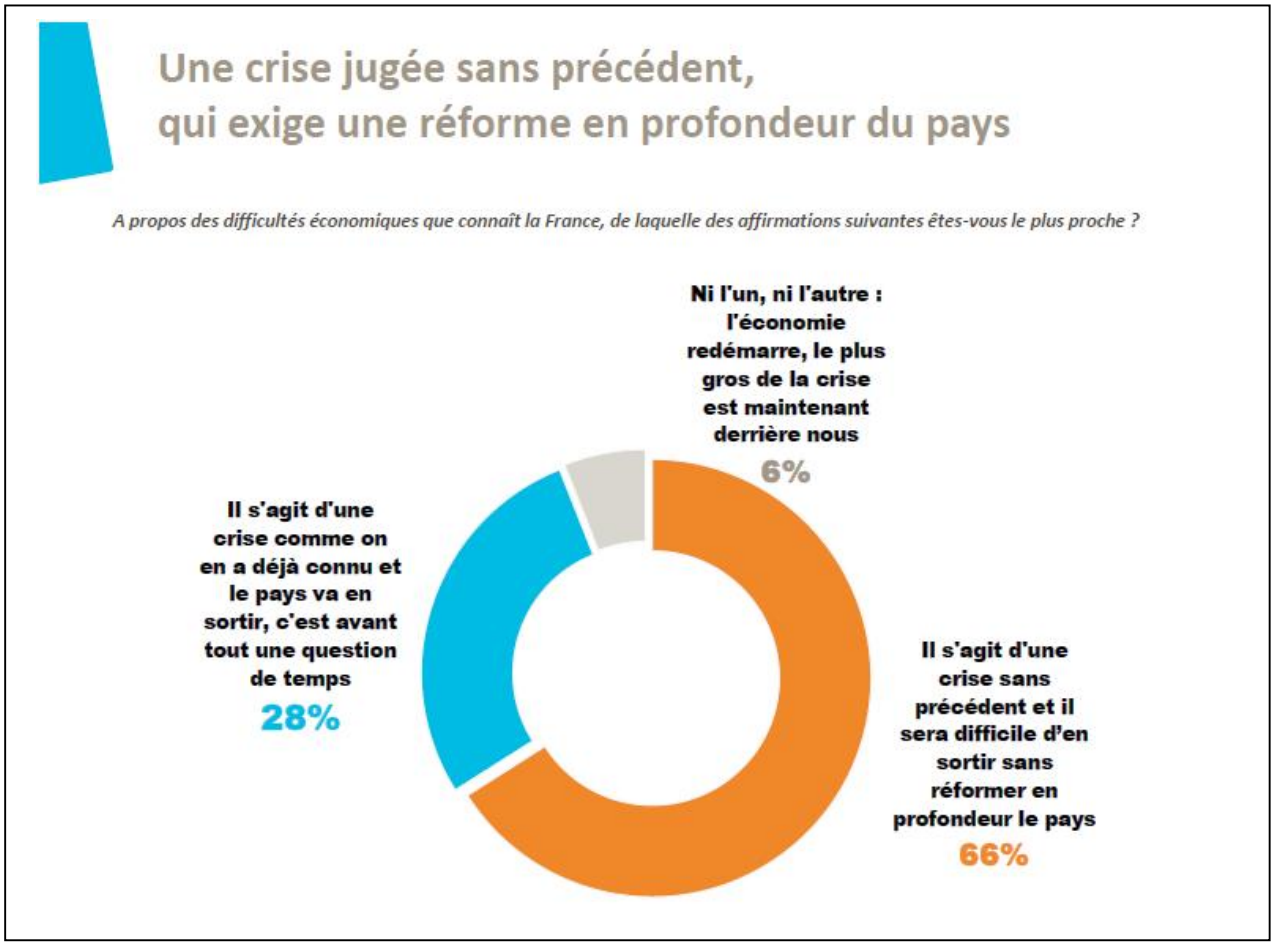


Variable quantitative

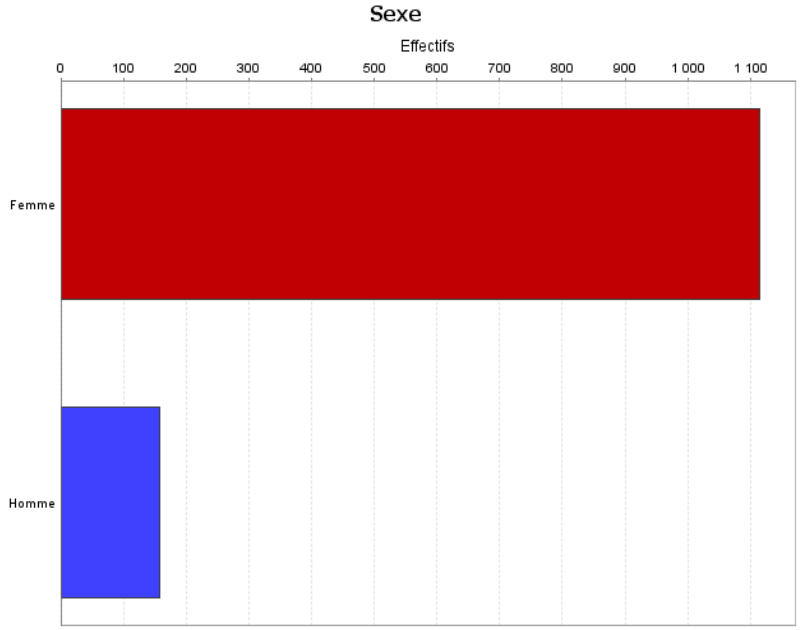


[14] Dans un musée d'Histoire Naturelle , le guide explique que le tyrannosaure est vieux de 70 millions et 6 ans. Une petite fille : « *vous êtes sûr ?* » . *« Pour être sûr, j'en suis sûr, lorsque j'ai pris mon travail ici, on m'a dit qu'il avait 70 millions d'années .Or ça, cétait il y a six ans ! ».*

3- Classer , Compter, Comparer , Repésentation graphique



Les diagrammes en anneaux ne facilitent pas les comparaisons mais permettent d'écrire les réponses.



des commentaires sur les « classer/compter/comparer » suivants :

sur le site www.expression-directe.com on pouvait trouver:

question du lundi 2 mai 2005

A propos de la campagne du référendum sur la Constitution européenne, avez-vous le sentiment que les médias en France sont …

… orientés en faveur du oui	73,1%
... orientés en faveur du non	8.4%
… ou se montrent équilibrés entre le oui et le non	15%
Sans opinion	3.4%

Nombre de votants : 13284

Sur le site Sofres- Taylor Young

Enquête réalisée le 7 octobre 2001 pour ABC News et le Washington Post auprès d'un échantillon de 506 personnes représentatif de la population américaine âgée de 18 ans et plus, Méthode des quotas (sexe, âge, profession du chef de ménage PCS) et stratification par département.

Question : approuvez-vous pu n'approuvez-vous pas la façon dont George W. Bush gère la riposte américaine face aux attaques terroristes du mois dernier à New-York et Washington ?

	%
Approuve	93
Désapprouve	5
Sans opinion	3

2- reconditionnement :

Dans une enquête on trouve la question :

buvez-vous de l'eau minérale ?

- sans occasion particulière
- au restaurant
- au travail
- pendant les repas de famille
- pendant une activité sportive
- autre

1- est-ce que les réponses sont exhaustives ? Exclusives ? Significations concrètes.
2- Que faire pour obtenir des variables à modalités exhaustives et exclusives ?

2 – variable qualitative ordinale
voici une variable qualitative ordinale

votre opinion sur ce film : vous avez aimé

□ pas du tout
□ un peu
□ beaucoup
□ une date dans l'histoire du cinéma !

proposez une transformation de cette variable en variable quantitative. Qu'est-ce qui est discutable ?

2 – question

Dans un questionnaire SNCF « *Votre satisfaction concernant les Salons Grand Voyageur* » on trouve :

Quel est le motif de votre voyage ?
O Domicile-Travail
O Domicile- études
O Déplacement privé ou loisir
O déplacement professionnel

Est-ce que les réponses sont exhaustives ? Pourquoi est-ce que les réponses ne sont pas exclusives ? Que peut-on faire pour transformer cette variable en questions-variables à modalités-réponses exclusives ? significations concrètes.

Objets centraux, en particulier Indicateurs de position centrale, et indicateurs de dispersions

Le problème :

711	239	739	63	993	909	77	618
925	660	737	827	512	334	980	318
592	392	516	380	431	213	618	973
9	363	391	19	846	653	477	504
845	655	132	313	34	906	397	558
773	149	775	486	34	263	794	463
568	543	133	446	733	35	70	389
493	858	640	196	336	978	260	732
706	591	42	924	17	182	711	694
941	613	257	885	583	923	501	122
499	293	435	232	324	439	885	542
210	598	213	176	163	216	327	252
73	374	636	261	201	470	464	13
617	606	312	259	378	828	944	52
607	833	562	84	634	509	175	830
39	142	29	26	585	771	645	279
43	306	20	745	897	166	922	365
76	625	912	819	462	701	841	73
75	980	214	676	602	797	562	212
849	901	11	344	777	893	692	309
739	746	576	773	474	381	732	98
330	298	219	817	996	823	435	810
964	661	32	358	754	115	523	840
69	869	248	25	387	859	709	951
145	587	244	497	445	561	774	331
718	69	251	395	488	881	657	530
685	791	373	445	929	894	973	3
557	917	744	903	585	786	873	58
418	457	403	716	394	362	901	41
30	371	161	432	870	894	789	99
419	110	803	302	317	777	624	705
687	808	371	874	722	932	936	41
924	367	291	524	546	335	821	118
933	640	451	377	400	205	184	476
580	46	11	357	276	247	883	20
435	846	119	908	601	784	931	663
238	977	56	527	359	12	47	96
552	68	513	908	442	734	295	382
329	539	459	561	955	386	145	291
63	368	643	206	265	275	978	818
461	878	986	945	441	187	220	809
655	152	920	321	984	193	395	39
648	697	462	216	707	266	951	455
965	803	682	973	567	552	394	976
238	286	571	532	119	987	442	442

De manière générale, à partir de n objets Θ_i i=1,…n et d'une mesure de la ressemblance/dissemblance de deux objets par une distance d, un objet central est la solution Θ^* du problème d'optimisation

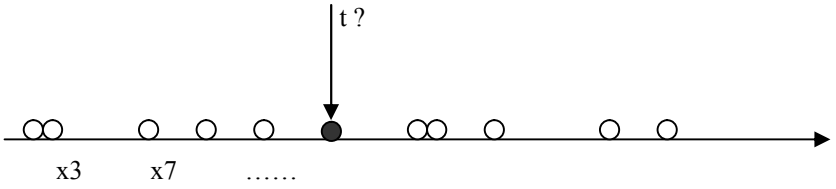
$$\min_{\Theta} V(\Theta)$$

avec

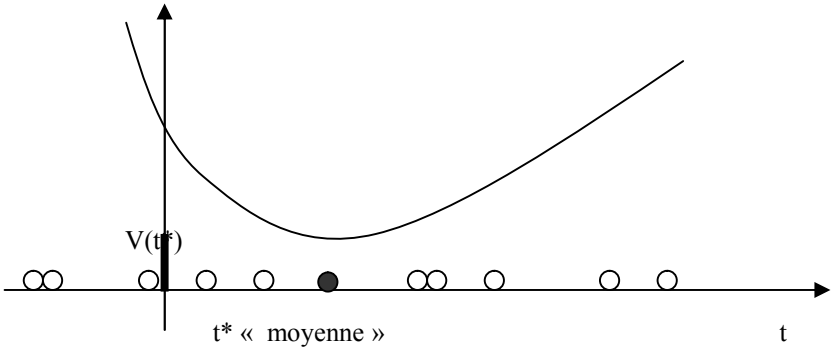
$$V(\Theta)=\frac{1}{n}\sum_{i=1}^{n} d(\Theta,\Theta_i)$$

la dispersion est mesurée par $V(\Theta^*)$

dans le cas de n observations numériques (données) $x_1, x_2,\ldots x_i,..x_n$ d'une variable X , on cherche le point • (la valeur de t) le plus proche possible des points blancs o (qui minimise V(t)).



$$V(t)=\frac{1}{n}\sum_{i=1}^{n} d(t,x_i)$$



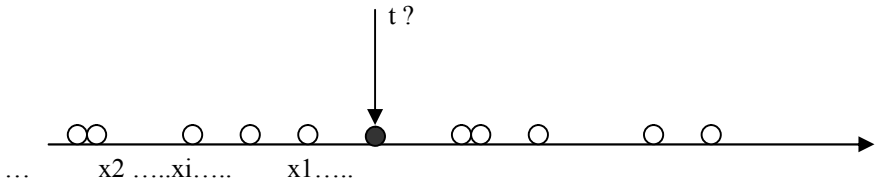
selon la distance

distance	**Objet central**	**dispersion**
d_1 de Manhattan	M_1 , médiane Me	$D=\frac{1}{n}\sum\lvert Me-x_i\rvert$
d_2 euclidienne	$\bar{x}=\frac{1}{n}\sum x_i$ M_2, moyenne	Variance, écart-type $V=\frac{1}{n}\sum\lvert\bar{x}-x_i\rvert^2$
d_q	Mq	Vq
d_M	mid-point, milieu $M=\frac{\max x_i+\min x_i}{2}$	amplitude Max xi – min xi
δ « hit or miss »	M_o, mode	% de valeurs non modales

La grande question c'est la **pertinence**, jamais évidente de tel ou tel indicateur, salaire moyen ? salaire médian ? ….**la signification concrète de ces indicateurs synthétiques**.
exemples …

Exercice : La production des indicateurs de position centrale et de dispersion

Un problème très général en « Analyse des Données » est le suivant : à partir d'un ensemble d'objets { Ob1, Ob2,…Obi,…Obn}, il s'agit de déterminer un objet Ob*, éventuellement fictif, c'est à dire n'appartenant pas à l'ensemble, qui soit en « moyenne » le plus « proche » possible des différents objets. Les objets sont, en pratique, des individus statistiques, des classements, des typologies selon différents critères… et dans le cas le plus simple, que nous allons considérer, une suite de nombres x1, x2,…xi,…xn résultant de n observations d'une variable X.



On considère les données xi i=1,…5 { -1 , 1 , 3 , 3 , 6 }

1- calculer la médiane Me, la moyenne M, le mode M0 et le « milieu » $M_\infty = \dfrac{\max_i x_i + \min_i x_i}{2}$

2- représenter graphiquement les différentes fonctions F : t ⟶ F(t) et déterminer les valeurs t1, t2, t∞ et t0 qui les minimisent.

$$F_1(t) = \sum_{i=1}^{5} |t - x_i| = |t-(-1)| + |t-1| + |t-3| + |t-3| + |t-6|$$

$$F_2(t) = \sum_{i=1}^{5} (t - x_i)^2$$

$$F_\infty(t) = \max_i (|t - x_i|)$$

$$F_0(t) = \sum_{i=1}^{5} \delta(t, x_i)$$

(la fonction δ fonctionne ainsi : δ(x,y)=0 si x=y et δ(x,y)=1 si x≠y)

3- remarques

exemples d'objets centraux

1- Irish Times 2005

D'après un récent sondage, le mâcheur de chewing-gum britannique type est une femme de moins de 24 ans qui vit dans le Nord de l'Angleterre et qui ne lit jamais un journal, hormis un tabloïd, The Sun..

2- Bruno Frappat, sur France-Culture , 2004

.. le lecteur type de « La Croix » est une femme, catholique pratiquante, de plus de 60 ans, habitant à la campagne…

3- *…L'électeur du non est un actif salarié de 35-54 ans.*

4-sur le site du Monde : *Pouvez-vous tracer le portrait-robot, s'il existe, du climatoseptique ?*
Les climatoseptiques se recrutent d'abord dans les tranches d'âge les plus âgées (plus de 65 ans). Il s'agit plus souvent d'hommes que de femmes, et on constate également une sur-représentation dans l'électorat de droite.

5- Objets : les classements d'étapes du tour de France

Objet central : le classement général

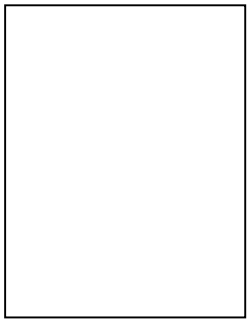
6- Objets : les notes en français, math, LVI, LVII … au bac

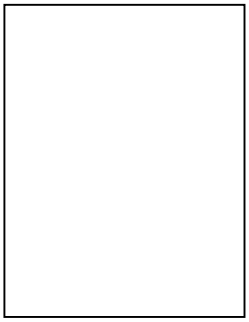
Objet central : la moyenne

7- objet central et dispersion

« *c'est l'histoire écossaise d'un anglais qui se noie dans un lac de 30 cm de profondeur en moyenne..* »

8- Rugby : mettre le ballon au « milieu » des joueurs, à l'engagement, au moment d'une mêlée.





examen 2002
1 – objet central

Parangon : ce mot a le sens général de modèle, par exemple parangon de vertu.
En Analyse des Données il désigne un objet central, représentatif d'un sous-ensemble d'objets. Une manière de produire un parangon à partir d'individus caractérisés par des variables qualitatives est de retenir, pour chaque variable , la modalité modale, c'est à dire la plus fréquente, par exemple :

	genre	**classe**	**couleur**
Pierre	garçon	moyens	jaune
Cécile	fille	moyens	jaune
Paul	garçon	petits	vert

Le parangon est un garçon de la classe des moyens habillé en jaune.

Calculer le parangon de ceux qui ont effectué un achat et un parangon de ceux qui ne l'ont pas fait dans le tableau ci-dessous

propriétaire	**statut**	**revenu**	**achat**
oui	marié	élevé	oui
oui	marié	moyen	oui
non	célibataire	élevé	oui
oui	marié	élevé	oui
non	célibataire	faible	non
oui	veuf	faible	non
non	célibataire	faible	non
non	célibataire	moyen	non

2 – ressemblances, distances

tableaux repris de « Le Data Mining » René Lefébure et Gilles Venturi Eyrolles 1999

On considère trois produits : Barre céréales, Crème dessert et Gâteau de riz, et leurs attributs (Y pour oui, N pour non)

	Barre céréales	**Crème dessert**	**Gâteau riz**
Chocolat	Y	N	Y
beurre	N	N	Y
Liquide	N	Y	N
Parfum mandarine	N	N	Y
Emballage métal	N	Y	Y
mini-dose	Y	Y	N
Sucre	Y	Y	Y
Riz	Y	N	Y
Edulcorant	N	N	Y
Colorant	N	N	Y

La distance entre deux produits est mesurée par le nombre d'attributs différents. Calculer
d(barre céréales, crème dessert) , d(barre céréales, gâteau de riz), d(crème dessert, gâteau de riz)
Quels sont les deux produits qui se ressemblent le plus ?

examen 2004

4**- cadeaux, des indicateurs de position centrale**

source « Economie et Statistiques » n° année . On a considéré la « valeur des cadeaux offerts à l'extérieur des ménages » . Voici quelques exemples :

	Médiane	moyenne
Vêtements, parures …	146 F	185 F
Livres, cassettes	80 F	109 F
Jouets et jeux	100 F	130 F
Etc,…etc		

Quelque soit le genre du cadeau, la valeur médiane est toujours inférieure à la valeur moyenne. Que comprendre ?

1 – objet central

on peut lire, dans le Monde-Campus cahier n°3 du 14 mai 1993

… l'enquête en cours à l'Institut Supérieur de Gestion sur la promotion 1992,… évalue le salaire médian (salaire le plus fréquemment cité) à …

que comprenez-vous ?

2 – il neige..

sur un site de données historiques de la météo concernant les Vosges, j'ai trouvé une phrase très pertinente :

il n'existe pas de hauteur de neige moyenne…

que comprenez-vous ?

– indicateurs de position centrale

- En France, de 2005 à 2010 le revenu moyen a augmenté en même temps que le revenu médian baissait. Que comprendre ?
- Cinq thermomètres indiquent, au même endroit et en même temps les températures : 20°, 20°, 21°, 20° et 24°. Calculer la température moyenne, modale, médiane. Quel est dans cet exemple l'indicateur le plus pertinent ?

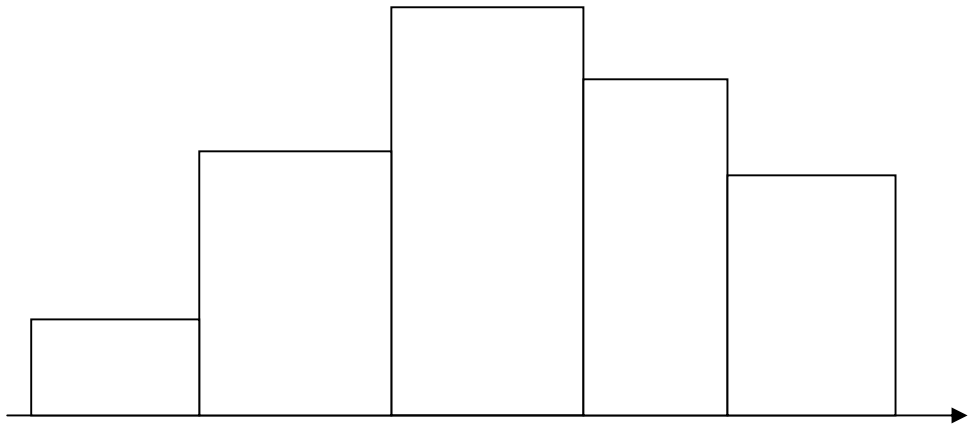
- d'après l'INSEE , « Le revenu moyen mensuel s'établit, en 2004, à 1 503 euros par personne, et le seuil de revenu qui sépare la population en deux est de 1 314 euros »

que peut-on comprendre ?

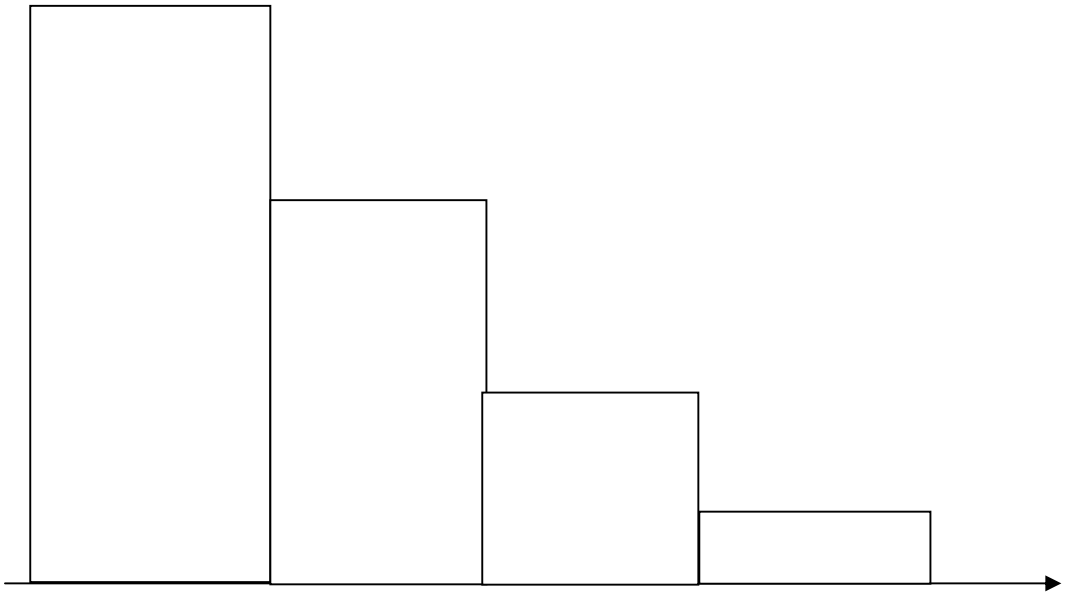
- la vie, la mort et les indicateurs statistiques

Vers 1660, Edmond HALLEY[15] établit des tables de mortalité pour des compagnies d'assurance. La durée de vie moyenne était 26 ans, un nouveau-né avait autant de chance de mourir avant 8 ans qu'après. Ceci parut curieux..Pouvez-vous expliquer ?

[15] Astronome anglais, une comète porte son nom.

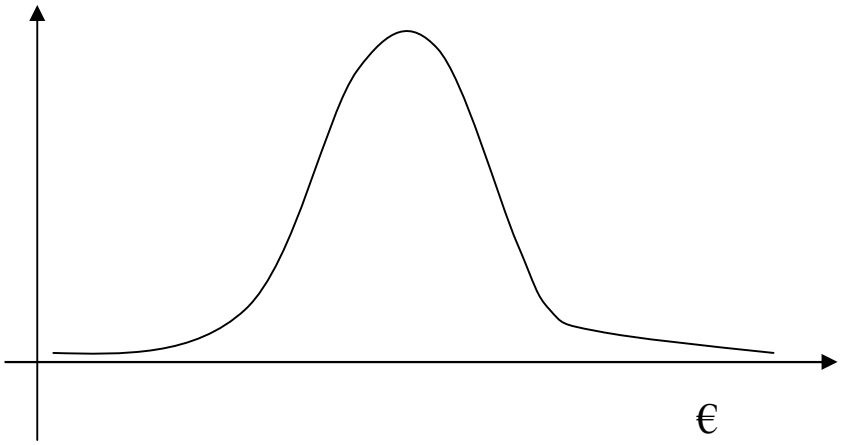


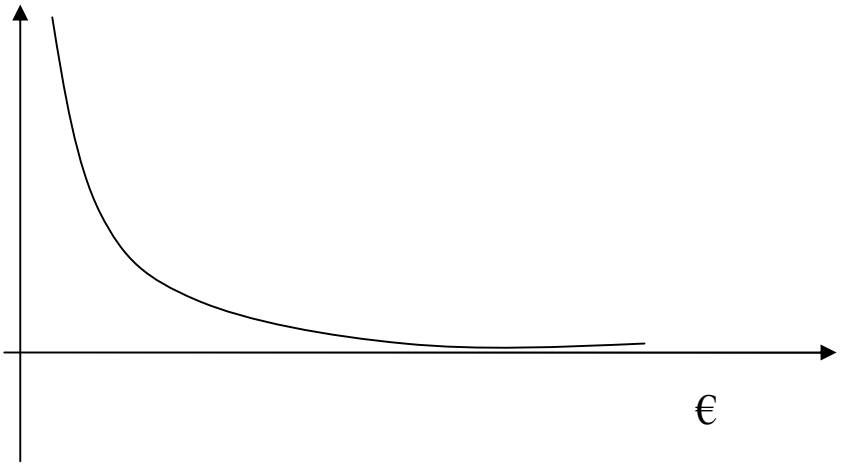
Ticket de caisse en €



Ticket de caisse en €

Exercice : quelle société ?





Données : X : -1 1 3 3 6

Médiane	Me=	dispersion :
Moyenne	M=	variance =
		Ecart-type =
Mode	Mo =	dispersion
Milieu	Mi =	amplitude =

Faire les calculs avec Excel

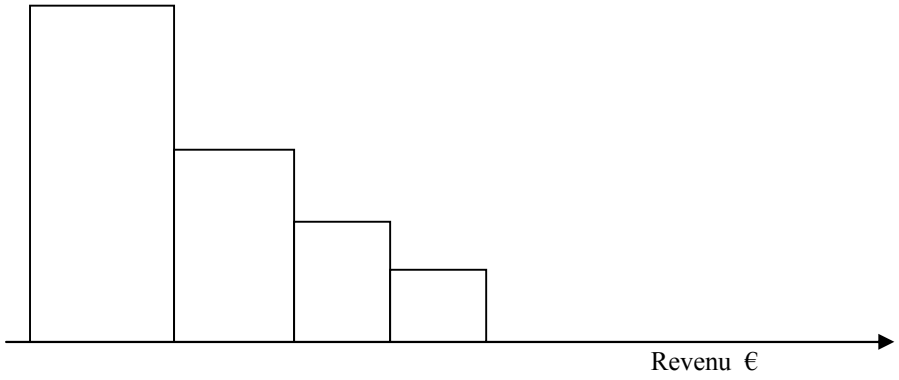
Pointures de chaussure 36,37,37….

Quel est l'indicateur le plus pertinent ?

fleuristes

		vendeur	anthropologue
	Moyenne		
	Médiane		
botaniste	Mode		
	milieu		

Revenu moyen ? revenu médian ?



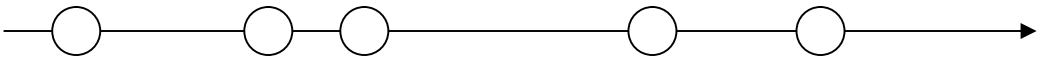
mesure dans les sciences sociales .

notes : 3 7 9 14 19

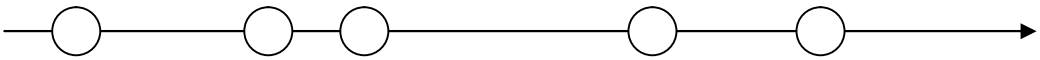
reçu si moyenne/médiane/mode/milieu supérieur(e) à 10 ?

où est-on incité à porter l'effort ?

note médiane



Milieu



Cas réels : temps de traitement de dossiers à la CAF, temps de retard des avions (AF).

Analyses factorielles des Correspondances Multiples. Penser, mesurer et représenter la nature et l'intensité des liens entre des variables qualitatives.

une variable : une couleur

Le tableau de données est de la forme

	SEXE	AGE	COULEUR
I1	homme	A3	foncée
I2	femme	A1	foncée
I3	homme	A3	foncée
I4	femme	A1	claire
I5	femme	A2	claire

à quoi correspond le Tableau Disjonctif Complet (TDC)

	GEN RE		AGE			COUL EUR	
	homme	femme	A1	A2	A3	claire	foncée
I1	1	0	0	0	1	0	1
I2	0	1	1	0	0	0	1
I3	1	0	0	0	1	0	1
I4	0	1	1	0	0	1	0
I5	0	1	0	1	0	1	0

A l'intersection d'un individu et d'une variable, ie les modalités de celle-ci, il y a exactement un « 1 » et des « 0 », ce qui résulte de l'exhaustivité et de l'exclusivité des modalités d'une même variable. Ex : l'individu n°3 et la variable AGE.

Croiser les variables deux à deux n'est pas raisonnable . Pratiquement, car un questionnaire ordinaire c'est facilement 70 questions, cela peut conduire à 100 variables (à cause des questions à réponses multiples), et donc à 100x99/2 = 5000 croisements deux à deux ! Théoriquement, car pour maîtriser les liens entre deux variables, il faut considérer leurs liens directs mais aussi tous les liens in- directs avec les autres variables.

Un classique

Voici un tableau très vraisemblable. On observe la température pendant 60 jours : elle est inférieure à 20°C, (« froid »), ou supérieure (« chaud ») l'autre variable est que le chauffage est sur « ON » ou sur « OFF », le thermostat est réglé à 18°C.

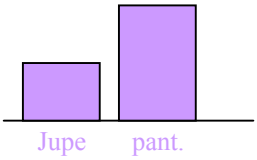
	froid	**chaud**
ON	16	1
OFF	2	41

Quand on chauffe, il fait froid, quand on ne chauffe pas, il fait chaud.

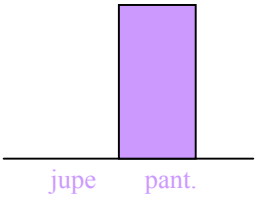
On peut s'attendre à ce que des croisements soient « intéressants » lorsque les modalités d'une variable ont des profils très différenciés sur l'autre variable, par exemple :

GENRE : homme, femme VETEMENT : jupe, pantalon

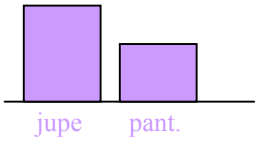
ensemble



homme

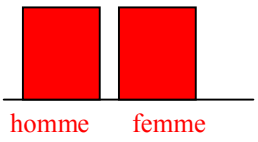


femme

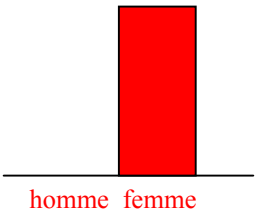


et réciproquement,

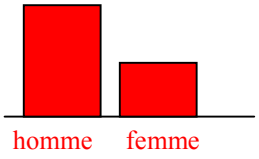
ensemble



jupe



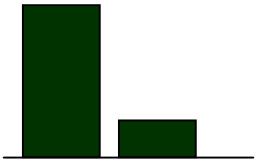
pantalon



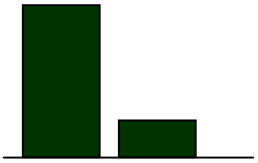
On peut aussi s'attendre à ce qu'au contraire des croisements soient « inintéressants » lorsque les modalités d'une variable ont des profils identiques sur l'autre variable, les variables sont alors dites « indépendantes » par exemple :

GENRE : homme, femme LIEU d'HABITATION : à la ville, à la campagne

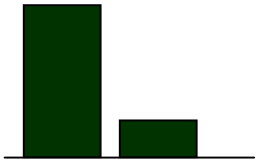
ensemble



homme

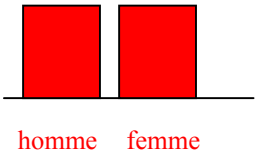


femme

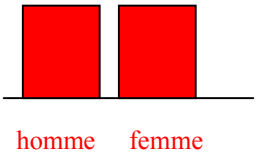


et réciproquement,

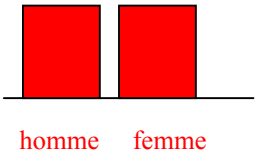
ensemble



ville



campagne



Il va s'agir de prévoir qu'un croisement est utile, ou pas.

La tableau de Burt [16]

C'est la juxtaposition des tableaux des croisements deux à deux des variables. D'un point de vue matriciel :

Burt = ^{t}TDC.TDC ^{t}TDC est la matrice transposée du Tableau Disjonctif Complet

Colorions :

	H	F	A1	A2	A3	cl	fo
I1	1	0	0	0	1	0	1
I2	0	1	1	0	0	0	1
I3	1	0	0	0	1	0	1
I4	0	1	1	0	0	1	0
I5	0	1	0	1	0	1	0

	I1	I2	I3	I4	I5
H	1	0	1	0	0
F	0	1	0	1	1
A1	0	1	0	1	0
A2	0	0	0	0	1
A3	1	0	1	0	0
cl	0	0	0	1	1
fo	1	1	1	0	0

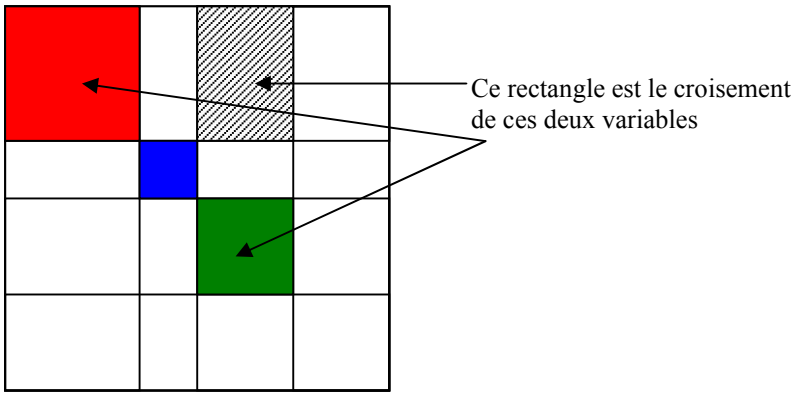
2	0	0	0	2	0	2
0	3	2	1	0	2	1
0	2	2	0	0	1	1
0	1	0	1	0	1	0
2	0	0	0	2	0	2
0	2	1	1	0	2	0
2	1	1	0	2	0	3

Un élément du tableau de Burt, c'est le nombre d'individus cumulant la modalité ligne avec la modalité colonne, il y a par exemple 2 hommes dans la cluster d'âge A3.

Sur la diagonale on retrouve les effectifs de chacune des modalités : 2 hommes, 3 femmes, 2 individus dans la cluster A1,…Les zéros dans les carrés le long de la diagonale résultent de l'exclusivité des modalités d'une même variable , alors qu'un zéro dans un rectangle, croisement de deux variables différentes résulte des données elles-mêmes : il n'y a pas d'homme A1 par exemple.

Un tableau de Burt est symétrique (Cf TD).

Le total de chaque carré ou rectangle est n, nombre total d'individus, 5 dans cet exemple.



Un tableau de Burt , on ne l'imprime pas, il est monstrueux : 50 variables à 6 modalités en moyenne (la variable Département en a 95), cela fait un tableau de 300x300 = 90 000 cases soit 8 000 000 000 de comparaisons deux à deux !

[16] **Sir Cyril Burt** 1883-1971 « *psychologue anglais, spécialiste de la statistique psychologique, il tenta d'adapter l'analyse factorielle aux tests psychologiques. Il a écrit notamment : Factors of the mind* (1940)… »in *Dictionnaire Robert des noms propres*. Par ailleurs célèbre pour avoir prouvé, en étudiant des jumeaux que l'intelligence s'expliquait à 93.346% par les gènes (« *the nature* ») contre 6.654% pour l'éducation (« *the nurture* »). Il s'est avéré après sa mort que les chiffres étaient faux, les lettres aussi, et qu'il avait inventé jusqu'au nom des co-auteurs de ses articles…

Stratégie de sélections/évictions (*pêche à la ligne*)

Quels sont les rectangles (croisements de variables) intéressants ?

On somme les contributions de toutes les cases d'un tableau croisant deux variables. Ce total est normalisé . On obtient le coefficient, le T^2 de Tchuprow [17] :
NL : nombre de lignes (nombre de modalités d'une variable)
NC : nombre de colonne (nombre de modalités de l'autre variable)

$$T^2 = \frac{1}{\sqrt{NL-1}} \cdot \frac{1}{\sqrt{NC-1}} \sum_{j=1}^{NC} \sum_{i=1}^{NL} c_{i,j} = \frac{1}{\sqrt{NL-1}} \cdot \frac{1}{\sqrt{NC-1}} \sum_{j=1}^{NC} \sum_{i=1}^{NL} \frac{(f_{i,j} - f_{i.} \cdot f_{.j})^2}{f_{i.} \cdot f_{.j}}$$

Il y a un rapport entre le T^2 de Tchuprow et le $\chi 2$ calculé d'un test d'indépendance de deux variables

$$\chi^2_{cal} = n.T^2 \sqrt{NL-1}.\sqrt{NC-1}$$

Par construction, $0 \leq T2 \leq 1$. Le T^2 est quelquefois noté T

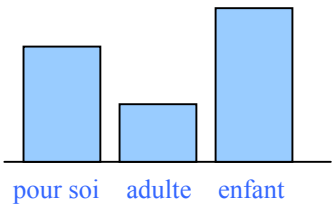
cas limites :

$T^2=0$ alors $c_{i,j}=0$ pour toutes les cases i,j. Les deux variables sont indépendantes. Les profils-lignes sont identiques entre eux, les profils-colonnes identiques entre eux. C'est le cas des variables GENRE et LIEU d'HABITATION : T2 (GENRE , LIEU d'HABITATION)=0 . La répartition homme/femme est la même à la ville et à la campagne. Il y a le même pourcentage d'hommes et de femmes à la ville et à la campagne.
$T^2=1$, alors il y a une exacte correspondance entre les lignes et les colonnes

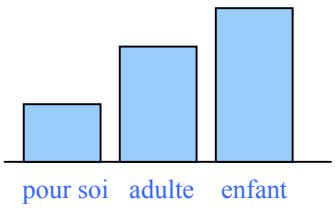
effectifs	maillot bleu	maillot blanc
Marseille	0	11
Strasbourg	11	0

Un T^2 élevé indique un croisement intéressant : il y a une forte différenciation des profils-lignes/profils-colonnes. Cette différenciation sera visible sur le représentation graphique des diagrammes différentiels .Cas réel : jouets en bois, variables GENRE et DESTINATAIRE (pour soi, pour un autre adulte, pour un enfant) . Le T^2 est élevé : il faut comprendre que la répartition des destinataires est très différente, pour les hommes et pour les femmes. J'imagine :

hommes



femmes



[17] **Alexander Alexandrovitch Чупров** 1874-1926,**Tchuprov**, Tchuprow, Tschuprow, Tschuprov, Chuprov, Tchouproff…

Stratégie d'agrégations/différenciations : Analyses Factorielles des Correspondances Multiples (*pêche au filet)*

Agrégations/différenciations des individus d'une part, des modalités d'autre part. Le tableau de départ est le Tableau Disjonctif Complet (le tableau de données).

* on a en vue de regrouper les individus qui se ressemblent, qui sont proches. La distance entre deux individus est mesurée par le nombre de modalités différentes (Cf exercice sur la distance orthographique)
Dans le tableau de la page * , d(I1,I2) = 1 + 1 + 0 = 2

*la distance entre deux modalités part de la distance quadratique entre leurs colonnes :
remarque : $0^2 = 0$, $1^2 = 1$, dans un TDC $x \in \{ 0, 1 \}$ et donc $x^2 = x$

$d(H,A1)=(1-0)^2+(0-1)^2+(1-0)^2+(0-1)^2+(0-0)^2 = 1+1+1+1+0 = 4$
de façon plus générale deux modalités A et B , a_i =1 si l'individu i a la modalité A, a_i=0 sinon. De même b_i.

A	B
a_1	b_1
a_2	b_2
….	….
a_i	b_i
…	…
a_n	b_n

$$d(A,B)=\sum_{i=1}^{n}(a_i-b_i)^2$$

$$d(A,B)=\sum_{i=1}^{n}(a_i^2+b_i^2-2a_i.b_i)=\sum_{i=1}^{n}a_i^2+\sum_{i=1}^{n}b_i^2-2\sum_{i=1}^{n}a_i.b_i=\sum_{i=1}^{n}a_i+\sum_{i=1}^{n}b_i-2\sum_{i=1}^{n}a_i.b_i$$

$d(A,B) = n_A + n_B - 2.n_{A\cap B}$

n_A : nombre d'individus qui ont la modalité A
n_B : nombre d'individus qui ont la modalité B
$n_{A\cap B}$: nombre d'individus qui ont les modalités A et B

<u>pour interpréter cette distance</u> :

On considère E, ensemble de tous les individus . Card (E)=n, le cardinal d'un ensemble c'est le nombre d'éléments. On note A le sous-ensemble de E des individus qui ont la modalité A, de même B.
Card(A)=n_A, card(B)=n_B.
On définit la « différence symétrique » de deux ensembles , puis son cardinal :

$$A\Delta B=(A\cup B)\cap(\overline{A\cap B}) \qquad \overline{A\cap B} \text{ le complémentaire de } A\cap B$$

$d(A,B)= card\,(A\,\Delta\,B) = n_A + n_B - 2.n_{A\cap B}$

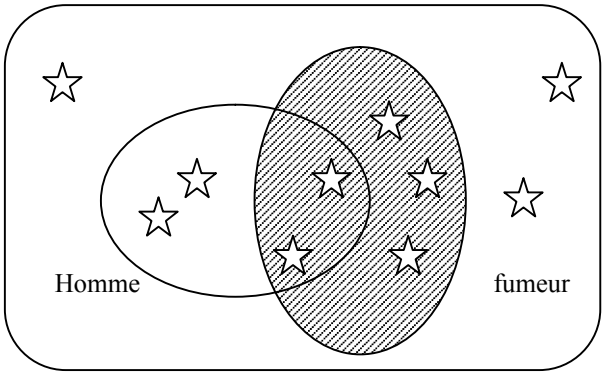
Lorsque les individus sont nombreux à avoir les deux modalités A et B, celles –ci sont rapprochées.

C'est une distance fondée sur la fréquence d'associations des modalités des variables.

Exercice : colorier, calculer, interpréter :

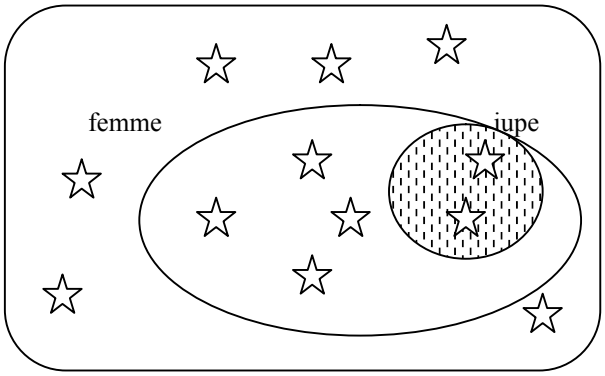
A	B	$a_i.b_i$	$(a_i-b_i)^2$
1	0		
1	0		
1	1		
0	1		
0	0		

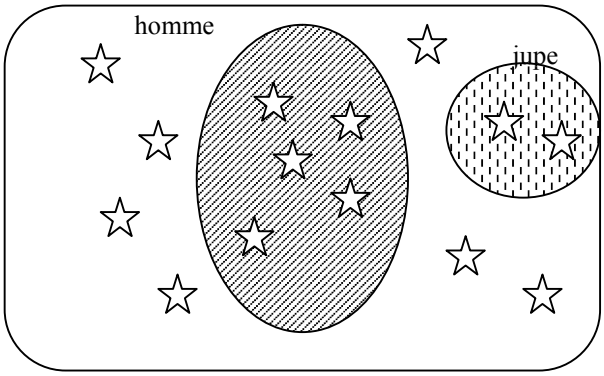
Illustrations : colorier AΔB et compter la distance



d(homme, fumeur) =

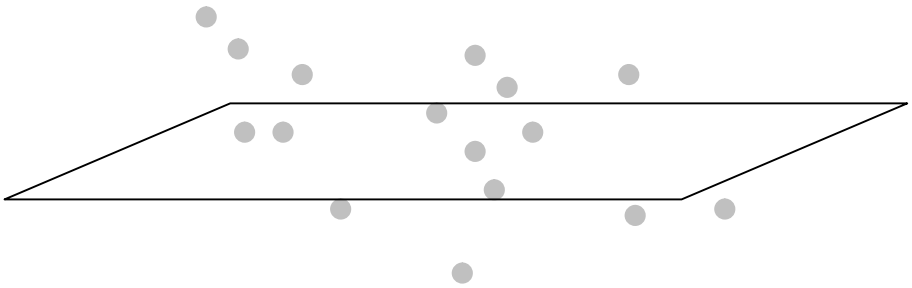
d(jupe,femme) =



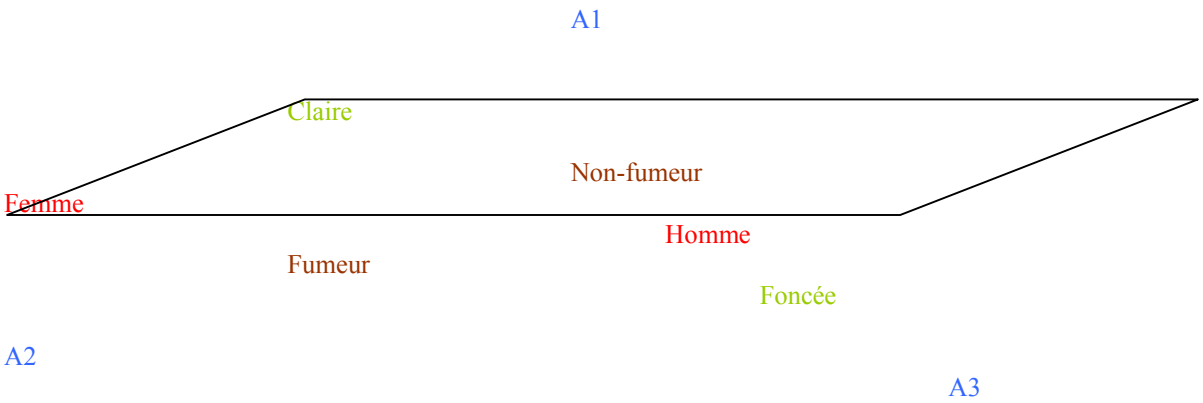


d(homme,jupe) =

Le nuage des points-individus dans l'espace des modalités

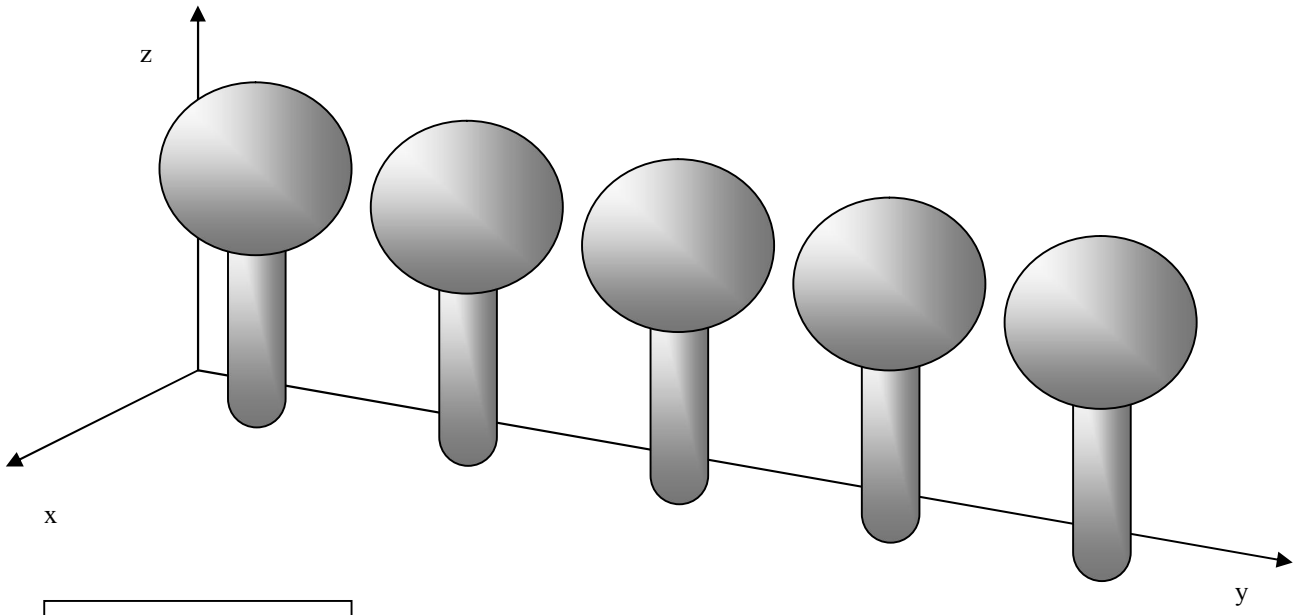


et le nuage des points-modalités dans l'espace des individus

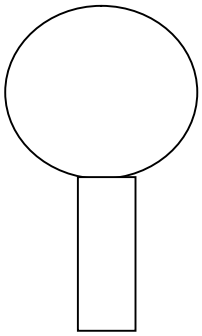


sont projetés sur le « meilleur plan »

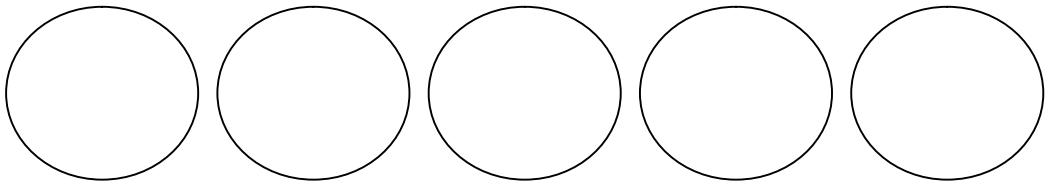
voir le cahier « mathématiques pour l'Analyse des Données »
les illusions d'optique, le meilleur « point de vue » / vecteurs et valeurs propres



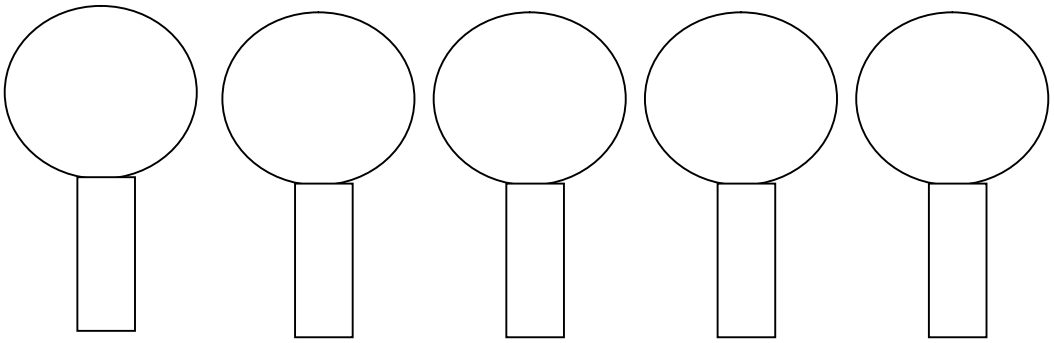
projection sur $P_{x,z}$



projection sur $P_{x,y}$

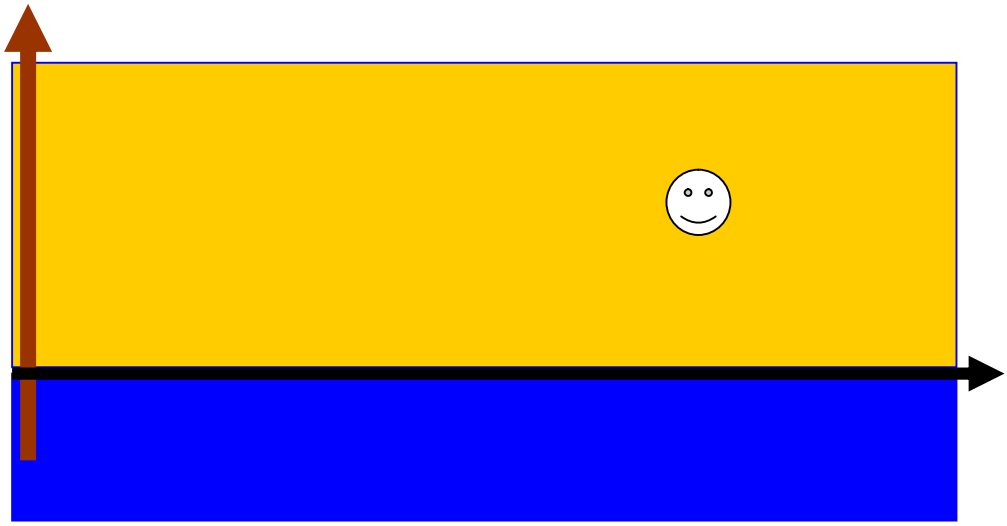


projection sur $P_{y,z}$

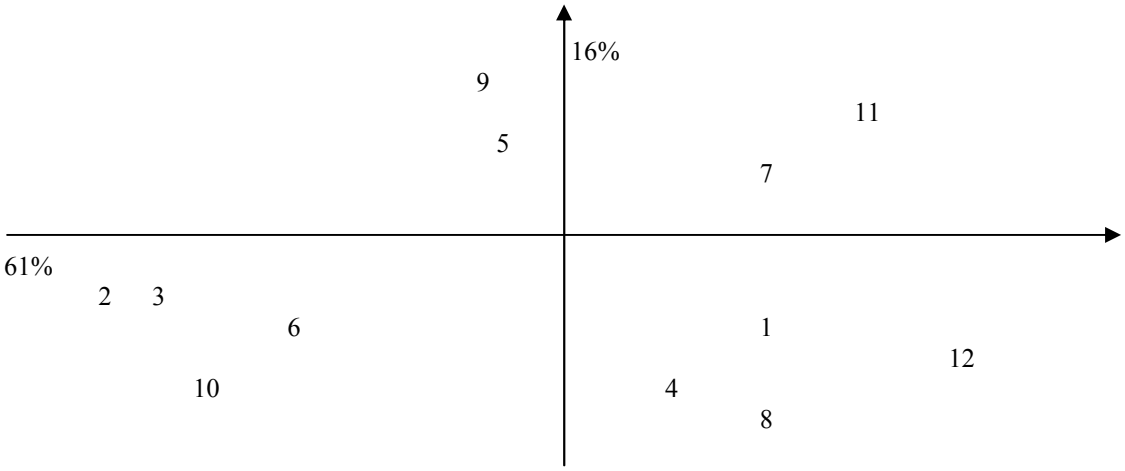


planisphère

rendez-vous à la plage



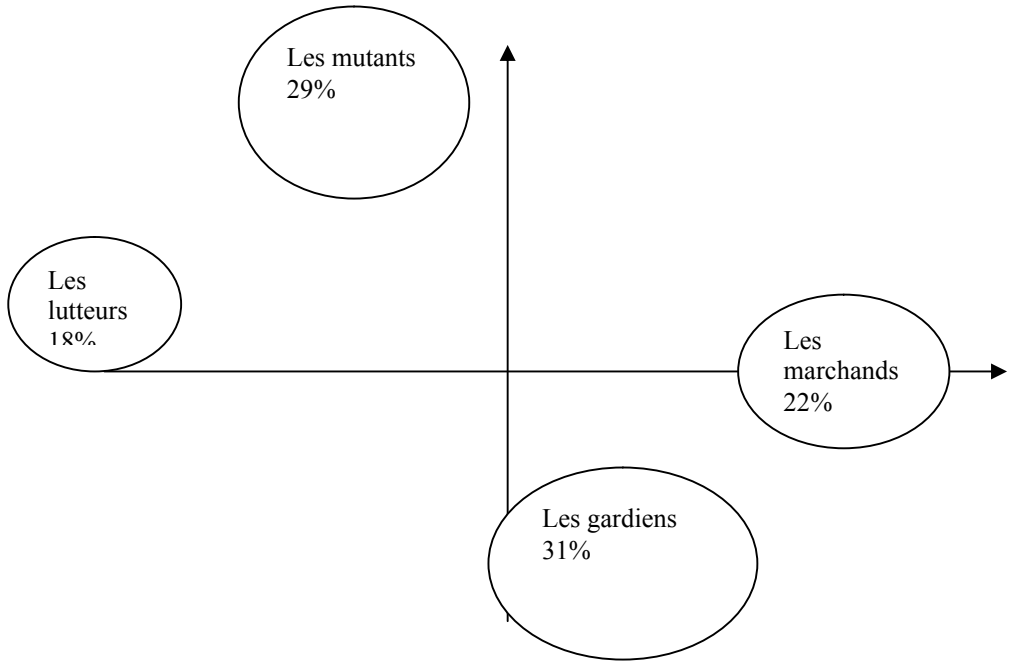
le graphique des individus



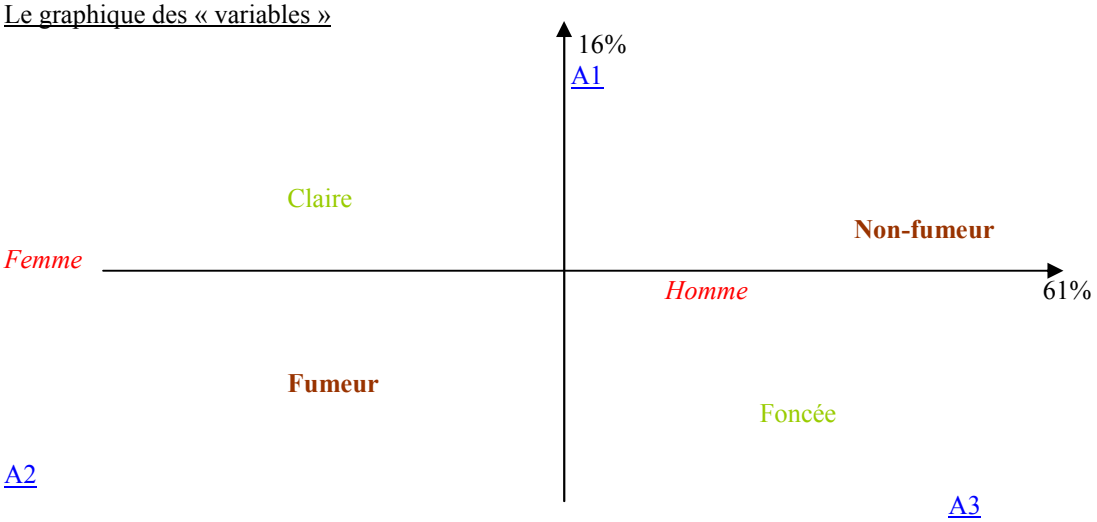
Sous réserve qu'ils soient bien représentés, des individus proches ont beaucoup de modalités en commun. On ne les représente généralement pas, raisons…..

Représentation de groupes , exemple Ipsos-Figaro Magazine

Exemple 3 : Ipsos- Figaro Magazine : 4 groupes, Ipsos mai 2000



Le graphique des « variables »



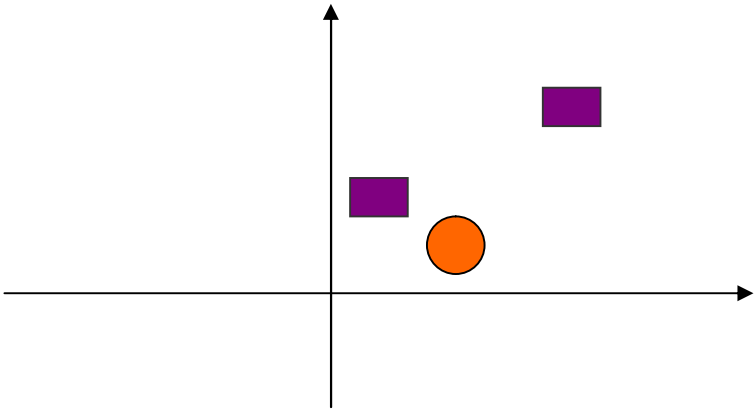
Arguments de fréquences relatives d'associations des modalités

- sur/sous représentations (relatives à la moyenne)
- analyse différentielle (on rend visible les écarts à la moyenne)
- position/ direction : analogie code de la route

position : *c'est ici !*	**direction : c'est par là !**
Bischheim	Paris Brest
modalités d'une même couleur	modalités de couleurs différentes
	angle aigu – même côté : sur-représentation relative et mutuelle *angle obtus -côtés opposés: sous-représentation relative et mutuelle*

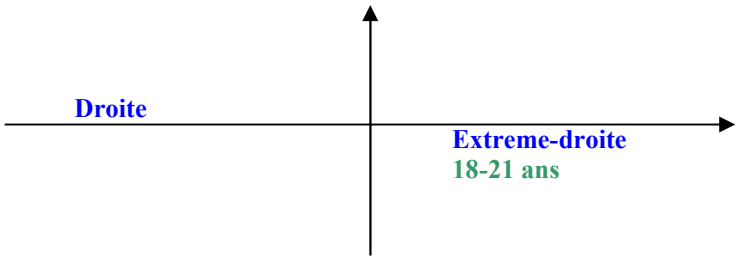
deux erreurs fréquentes

1-confusion position/direction



2-confusion relatif/absolu

Elections européennes 1984



	Droite	Ext-Droite	…	
18-21 ans	35	15	…	100
ensemble	40	10	…	100

Quelques points de repères en Analyses Factorielles des Correspondances Multiples

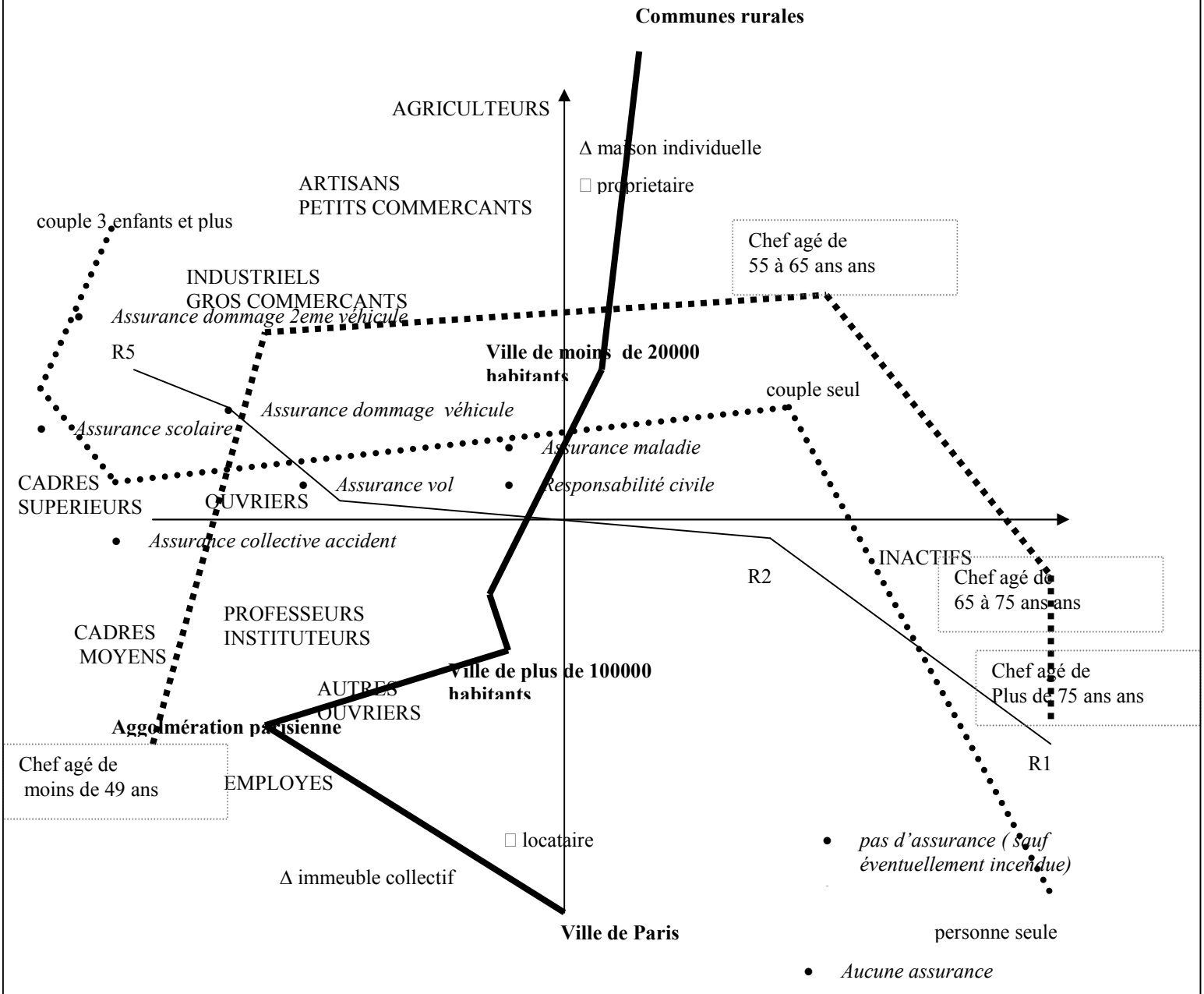
Une méthode pour penser les liens, en termes de sur et de sous représentations relatives des modalités de variables qualitatives , et pour les représenter, est l'Analyse Factorielle des Correspondances Multiples[18]

- c'est une méthode de statistique exploratoire qui permet de déceler d'éventuelles régularités statistiques dans les associations des modalités des différentes variables qualitatives.
- Les liens entre les variables qualitatives sont pensés et mesurés en termes de sous et de sur représentations relatives. Un énoncé tel que : chez les ménages OUVRIERS et CADRES SUPERIEURS les « assurances collectives accidents » sont sur-représentées, doit se comprendre ainsi : le pourcentage d '« assurances collectives accidents » dans ces catégories est plus élevé que le pourcentage moyen, celui de l'ensemble des ménages. l'AFCM est une méthode différentielle : elle met en évidence des écarts à la moyenne.
- Les variables qualitatives, sont représentées par leurs modalités, d'une même couleur. Lorsque la variable est ordinale, c'est le cas de la taille de l'agglomération, de l'âge du chef de ménage, les modalités successives sont reliées par une ligne brisée orientée.
- Les modalités sont, géométriquement, des points dans un espace de grande dimension. Ces points sont projetés sur un plan optimal en un certain sens : la figure géométrique initiale est la moins déformée possible, mais elle l'est .
- On obtient donc une représentation des modalités dans un plan dit « plan de projection » , et des règles de traductions, d'interprétations.
- Au centre du graphique se trouvent les pourcentages moyens. Les modalités de couleurs différentes situées du même côté, par rapport à ce centre sont mutuellement sur-représentées. Ainsi, dans le graphique « les comportements des français en matière d'assurance », en bas, chez les ménages habitant la « ville de Paris » , il y a une sur représentation des immeubles collectifs, des locataires, des employés…. Dans cette ville, les ménages ayant les caractéristiques précédentes sont relativement plus nombreux.
- Les modalités de couleurs différentes situées dans des directions opposées par rapport au centre du graphique sont mutuellement sous-représentées. Ainsi dans la « ville de Paris », il y a une sous-représentation des maisons individuelles, des AGRICULTEURS, des propriétaires….
- Des modalités de même couleur (d'une même variable) proches ont les mêmes caractéristiques c'est à dire les mêmes fréquences des modalités des autres variables.
- Lorsque la ligne brisée orientée qui représente une variable ordinale se déploie bien dans le graphique, c'est le cas de la taille de l'agglomération, c'est qu 'elle est très significative , les différentes catégories : communes rurales, villes de moins de 20 000 habitants … sont, dans une certaine mesure, intra-homogènes extra-hétérogènes, lorsque l'on passe d'une classe de taille à une autre , il y a une variation sensible du pourcentage des modalités d'autres variables, notamment le pourcentage de locataires , de propriétaires, d'agriculteurs….
- Une variable supplémentaire-illustrative est une variable qui est projetée sur un plan déjà constitué. On peut observer ses liens statistiques avec les autres variables mais elle n'a pas servi à structurer le nuage initial des points-modalités, contrairement aux autres variables dites « actives ».La distinction variable active/supplémentaire-illustrative est d'une extrême importance méthodologique, même si , pratiquement les représentations peuvent être très semblables.

[18] un livre utile « *Analyses factorielles simples et multiples : objectifs, méthodes et interprétations* », Brigitte Escoffier et Jérôme Pages , Dunod 1998

« Les comportements des français en matière d'assurance »

« La détention d'assurances dans l'espace des ménages représentés par leurs caractéristiques socio-démographiques » Daniel VERGER Economie et Statistique fev 85

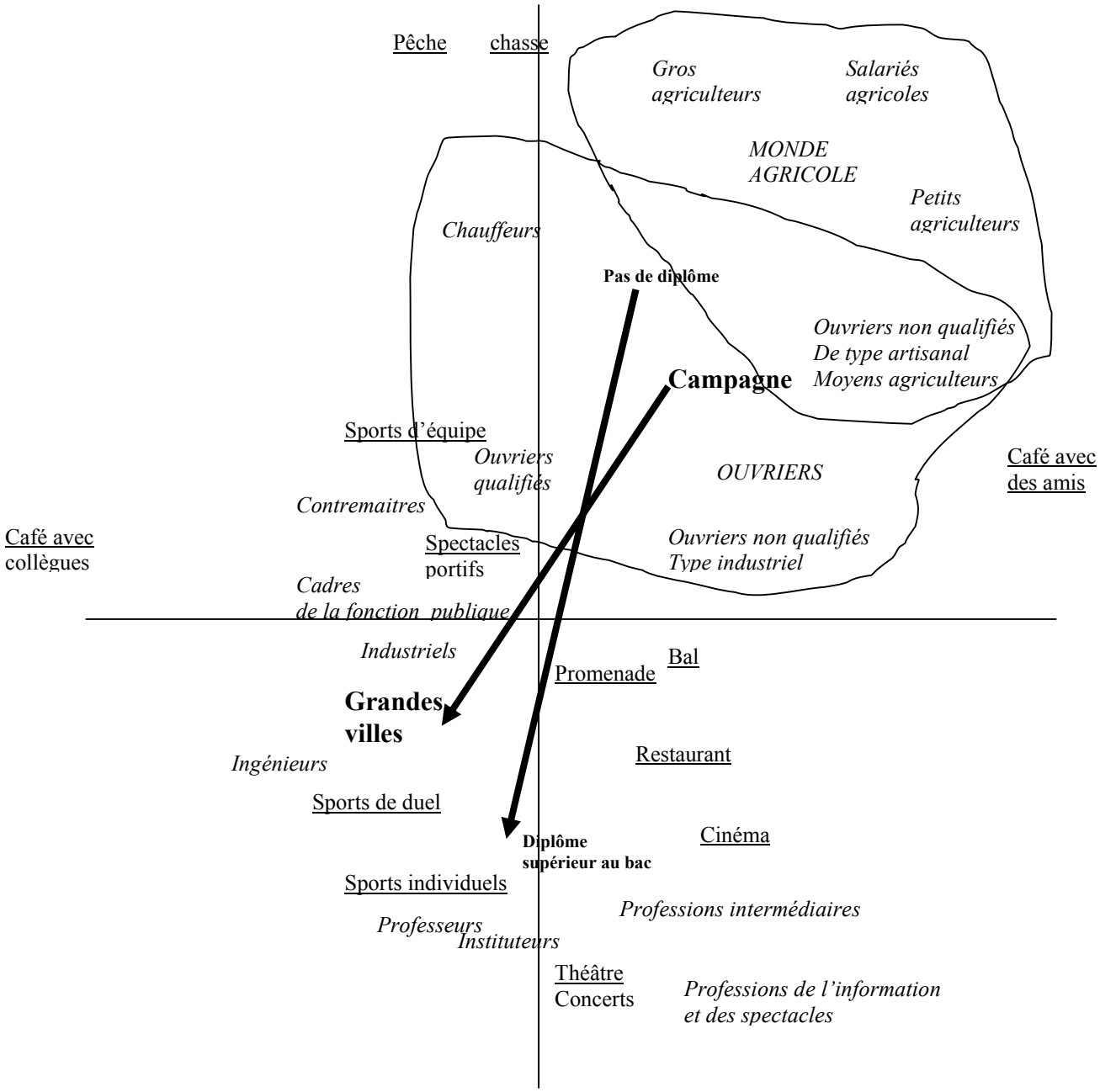


La méthode utilisée est celle de l'analyse factorielle des correspondances appliquée au tableau définissant pour chaque ménage (en ligne) ses caractéristiques socio-démographiques (en colonne).On peut ainsi dire que ces caractéristiques des ménages sont des variables actives et servent à déterminer les plans factoriels. Les variables relatives à la détention d'assurances ont été ensuite projetées sur ces plans factoriels par la méthode dite des « variables supplémentaires ».

La proximité d'un point « assurance » et d'un point « caractéristique du ménage » est susceptible de traduire une plus grande fréquence de l'assurance considérée parmi les ménages ayant cette caractéristique. Plus généralement la position des points « assurances » dans le plan factoriel permet de déceler les facteurs de disparité essentiels. Pour cela, il faut bien comprendre la signification des axes.

« Les sorties : une occasion de contacts «

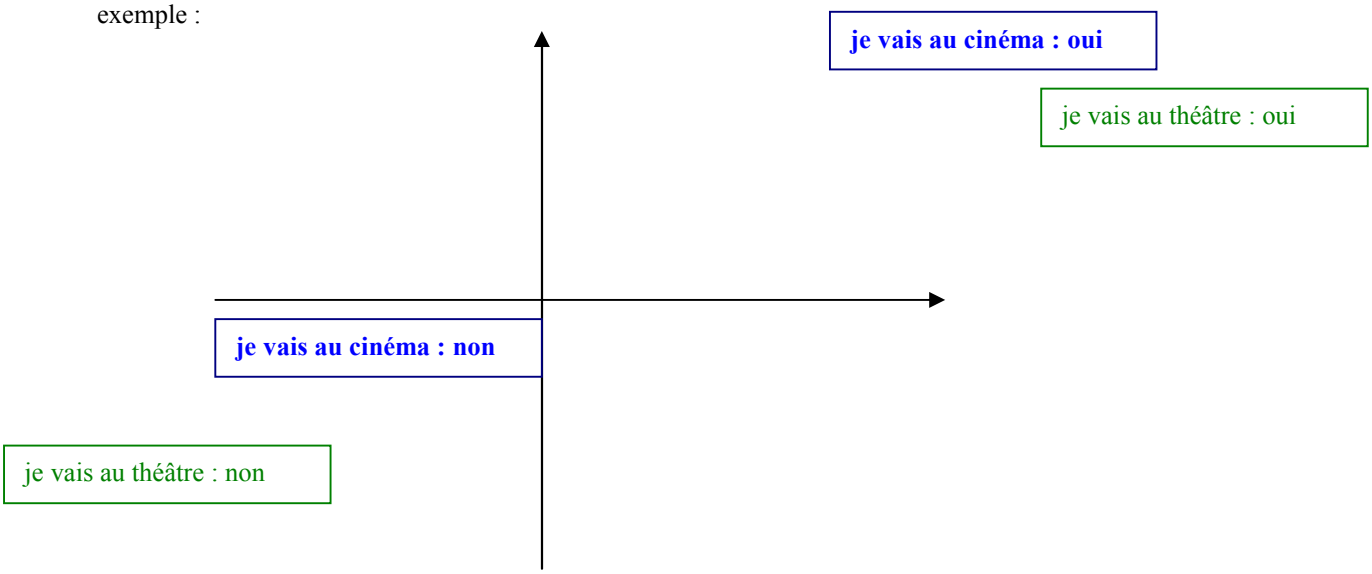
Olivier CHOQUET Economie et Statistique n°214 1988



la méthodologie mise en œuvre est celle de l'analyse factorielle des correspondances multiples. Variables actives (soulignées sur le graphique) :effectuer une sortie de type i avec quelqu'un d'extérieur à son ménage (loto exclu et café distingué selon la compagnie). Variables illustratives : diplôme, catégorie socio-professionnelle et taille de l'agglomération.

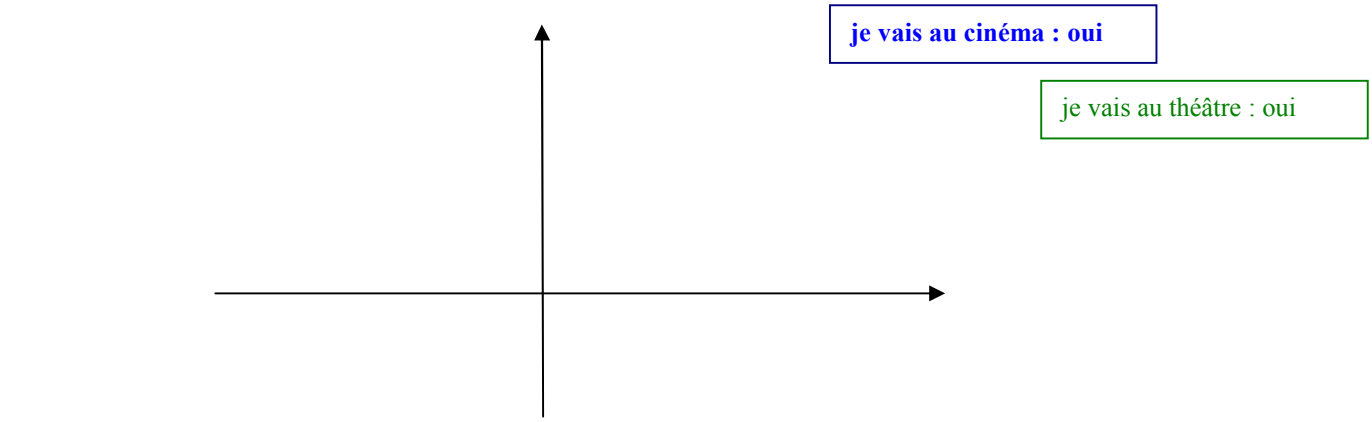
La représentation graphique dans le cas particulier de variables-questions à réponse oui/non

exemple :

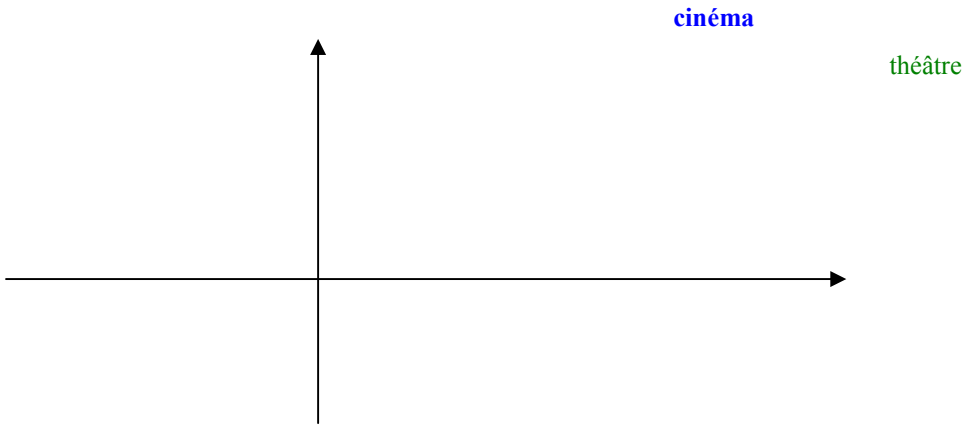


on efface la réponse "non" et on remplace la réponse "oui par la question (SPAD permet cette manipulation sur le graphique effacer/changer)

on obtient :



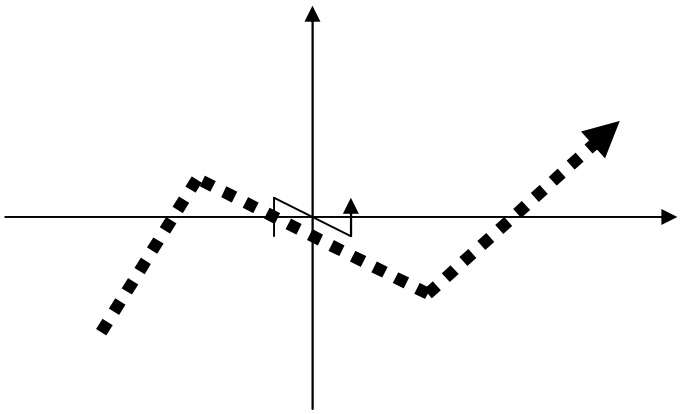
et finalement



aller au cinéma est plus fréquent pour ceux qui vont au théâtre, et réciproquement.

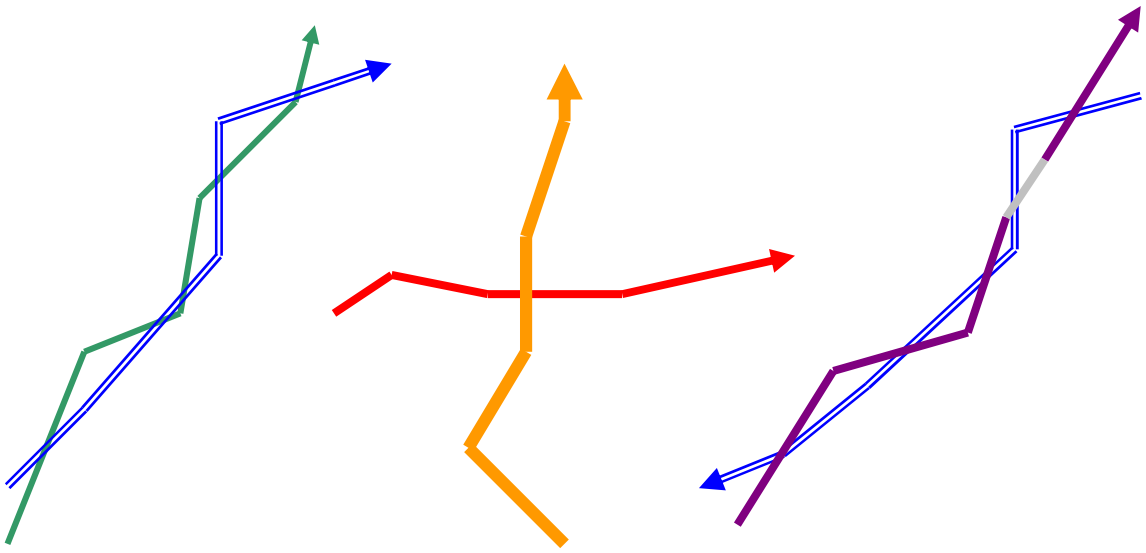
Le cas des variables qualitatives ordinales :

On relie, par une ligne brisée orientée les modalités successives d'une variable ordinale



Lorsque la ligne brisée orientée qui représente une variable ordinale se déploie bien dans le graphique , c'est qu'elle est très significative : lorsque l'on passe d'une modalité à la suivante il y a une variation sensible du pourcentage de modalités d'autres variables..

Voir les liens entre deux variables ordinales :



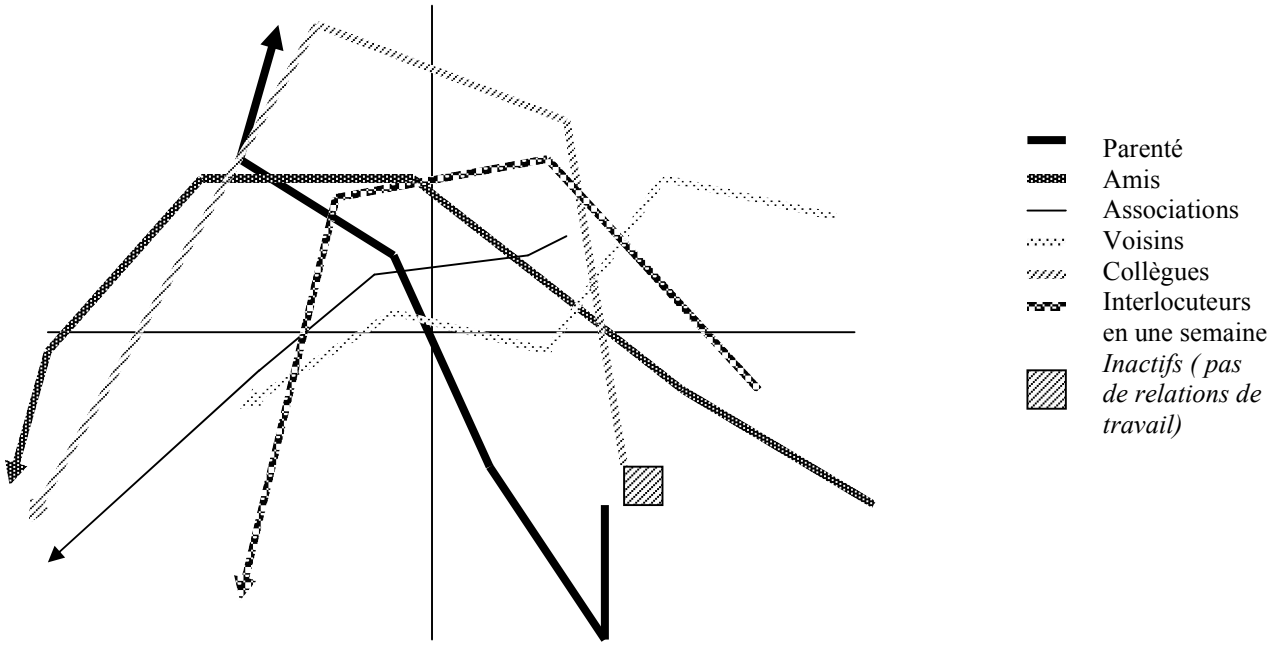
T2 >> Liens forts, même direction	T2 ≅ 0 indépendance	T2 >> lien fort, direction opposée

Interprétations : voir exemples

« La sociabilité « une Analyse Factorielle des Correspondances Multiples

d'après « La sociabilité, une pratique culturelle » François HERAN, Economie et Statistiques n°216 1988

cohérence de la sociabilité
l'espace des pratiques
Espace des variables actives dans l'analyse des correspondances multiples.
Intensité des principales relations.

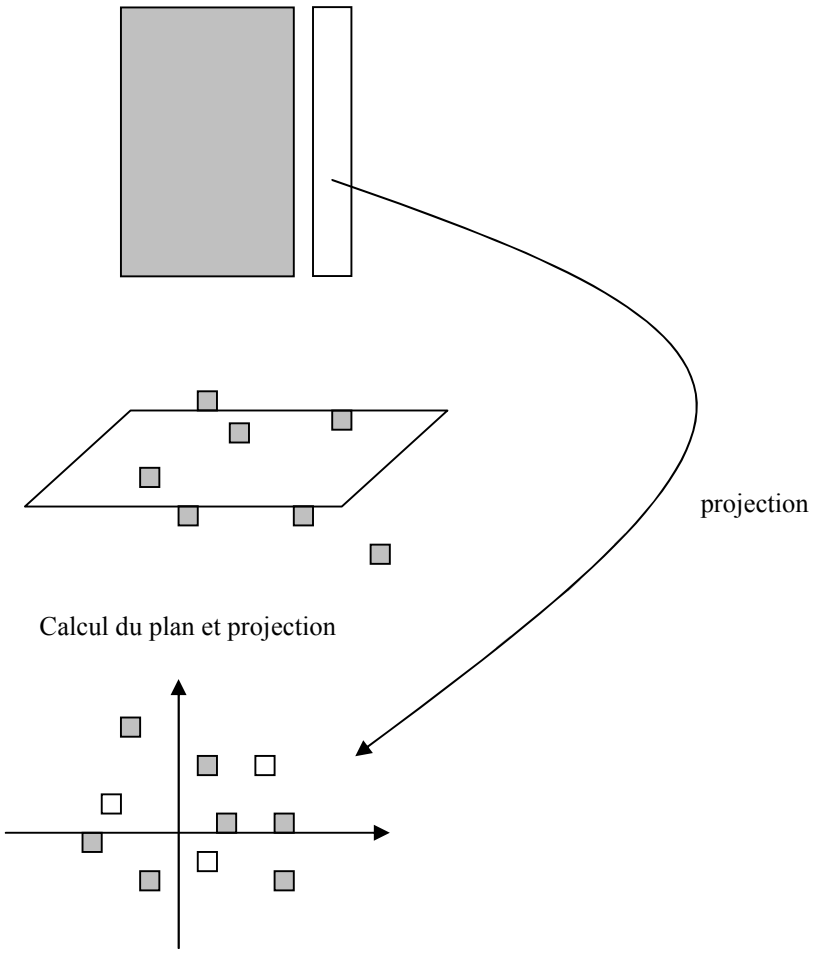


<u>On a retenu les indicateurs et les degrés d'intensité suivants</u> :

- nombre de membres de la parenté vus au moins une fois par mois (0,1 ou 2, 3 ou 4, 5 à 8, 9 à 12, plus de 12)
- nombre d'amis que l'on dit connaître personnellement (0, 1, 2, 3 ou 4, 5 ou plus)
- nombre d'adhésions aux associations (0, 1, 2, 3 ou plus)
- intensité des contacts avec le voisinage depuis un an (aucune conversation : simples conversations : visites : visites avec échanges de services : liens plus étroits)
- nombre de relations de travail que l'on revoit hors du lieu de travail (0 pour les inactifs : 0 pour les actifs : 1 à 3 : 4 à 9 : 10 ou plus)
- nombre de personnes différentes avec lesquelles on discute en une semaine (1 à 7 : 8 à 13 : 14 à 19 : 20 à 29 : 30 à 39 : 40 ou plus)

Objet (variable, individu…) actif / illustratif-supplémentaire

Une variable supplémentaire-illustrative est une variable qui est projetée sur un plan déjà constitué. On peut observer ses liens statistiques avec les autres variables mais elle n'a pas servi à structurer le nuage initial des points-modalités, contrairement aux autres variables dites « actives ».La distinction variable active/supplémentaire-illustrative est d'une extrême importance méthodologique, même si , pratiquement les représentations peuvent être très semblables.



<u>des raisons</u> :

- de fond : variables « lourdes », caractéristiques socio-économiques actives, puis les variables d'opinion en variables supplémentaires
- **variables « explicatives », « clivantes » en variables actives, et les variables « expliquées », « clivées » en variables supplémentaires**
- l'objet illustratif est fictif, et caricatural
- objet collectif : Alsace, Bade-Wurtemberg, Vénétie…. en variable active, puis France, Allemagne, Italie en variable illustrative

exemples, cas réels

5- « L'attachement régional « une Analyse Factorielle des Correspondances Multiples

source[19]. 715 individus habitant en Alsace.

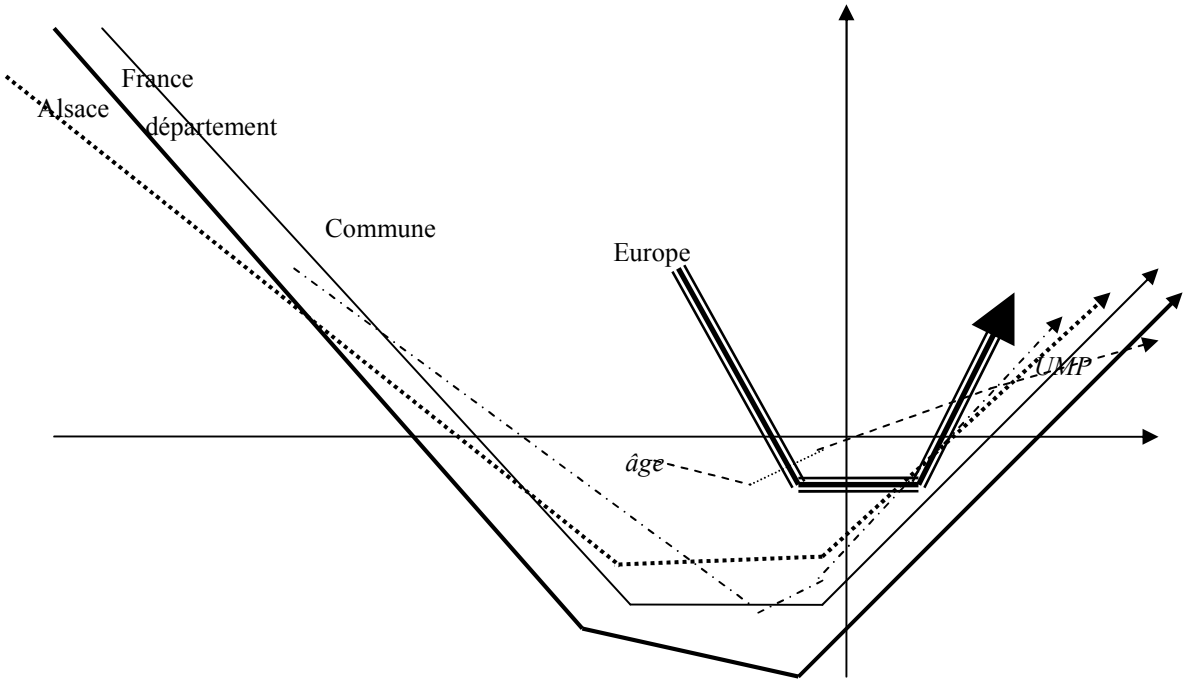
Variables actives :

Pouvez-vous me dire si vous êtes attaché, plutôt attaché, pas très attaché ou pas attaché du tout à :

- L'Europe
- La France
- L'Alsace
- Votre département
- La ville ou la commune où vous habitez

Variables illustratives :

Proximité partisane (après regroupements :- *Extrême-Droite, UMP, UDF, Gauche, Ecolo, Extrême-Gauche,…..*et **âge :***18-24 ans,25-34ans, 35-49ans,50-64 ans,65 ans et plus*



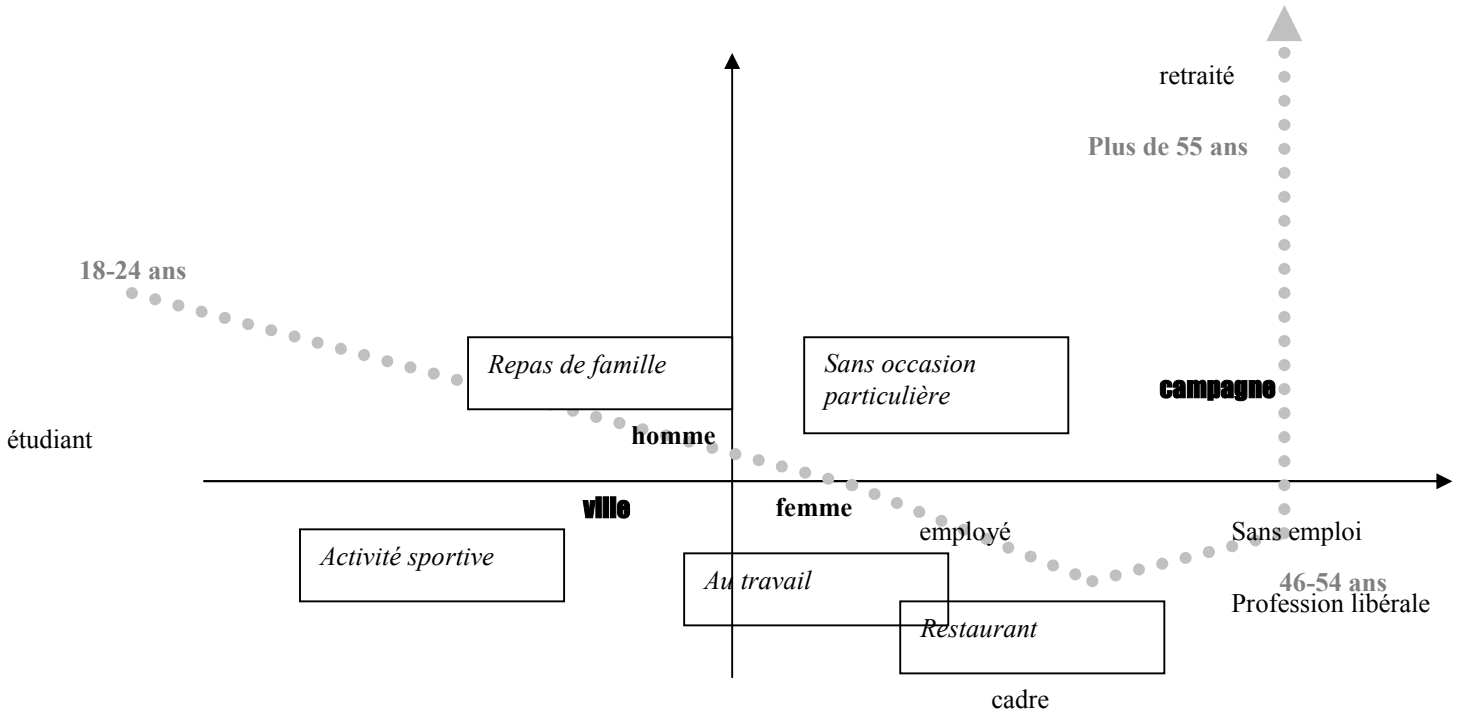
- *** que peut-on dire des liens entre les différents attachements (à l'Europe, La France,…) ?

[19] Source CDSP, Centre de Données Socio-Politiques, Sciences Po-Paris, 2004

6– Eaux minérales, une Analyse Factorielle des Correspondances Multiples

cas réel. On a interrogé 105 passants[20]
variables actives : GENRE : homme, femme), AGE : 18-24 ans,….45-54 ans, plus de 55 ans, CATEGORIE PROFESSIONNELLE : étudiant, cadre, employé, profession libérale, retraité, sans emploi , LIEU d'HABITATION : à la ville, à la campagne
variables illustratives: **buvez-vous de l'eau minérale** :

- *sans occasion particulière*
- *au restaurant*
- *au travail*
- *pendant les repas de famille*
- *pendant une activité sportive*
- *autre*



- que peux-t-on dire de ceux qui boivent de l'eau minérale au restaurant ?

[20] « Etude de marché CAROLA » Annelise DREVAL « mémoire de maîtrise » Université de Strasbourg I- 2005

Penser, mesurer et représenter la nature et l'intensité des liens entre des variables quantitatives-numériques.

Coefficient de corrélation linéaire

Une variable X : $x_1,x_2,\ldots x_i,\ldots x_n$ de moyenne MX et d'écart-type σ (SD *standart deviation*)

Valeur centrée-réduite (on s'affranchit de l'unité de mesure) : X' :

$$X' = \frac{X - MX}{\sigma} \quad \text{alors} \quad MX' = 0 \quad \text{et} \quad VX' = \sigma = 1$$

X' : $x'_1,x'_2,\ldots x'_i,\ldots x'_n$

le coefficient de corrélation linéaire ρ :

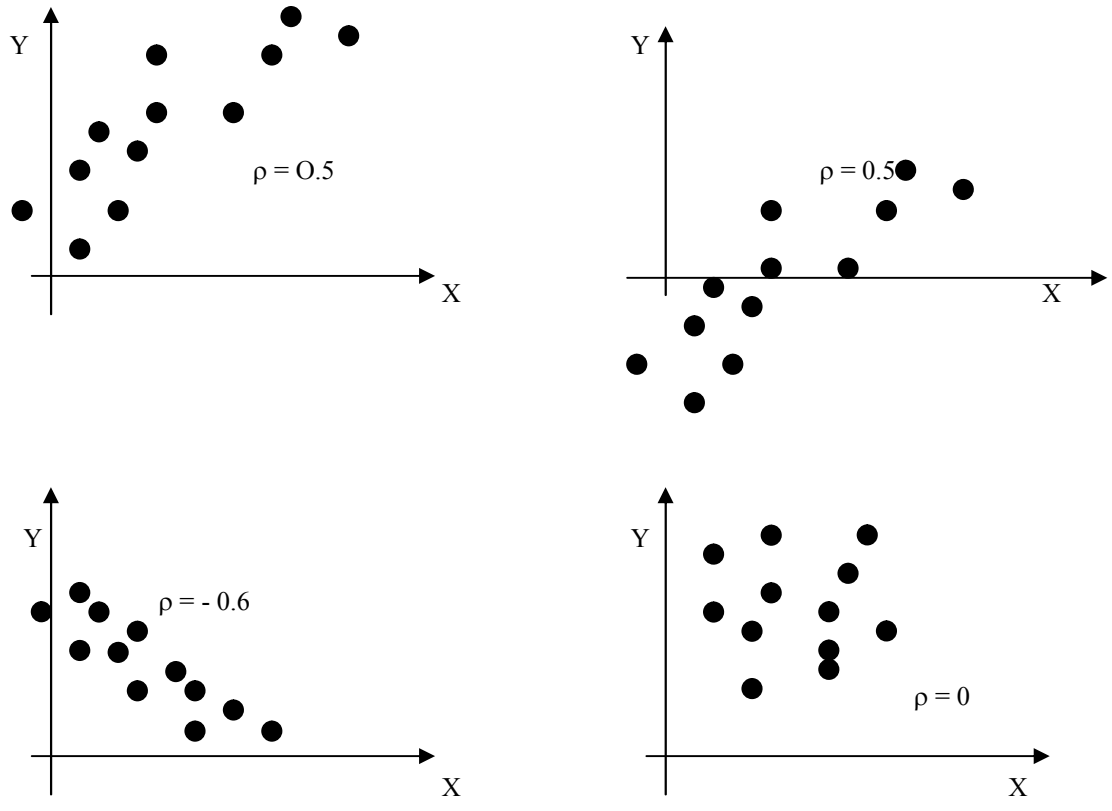
$$\rho(X,Y) = \rho(Y,X) = \frac{1}{n}\sum_i \frac{(x_i - MX)}{SDX} \cdot \frac{(y_i - MY)}{SDY} = \frac{1}{n}\sum_i x'_i \; y'_i$$

$-1 \leq \rho(X,Y) \leq 1$

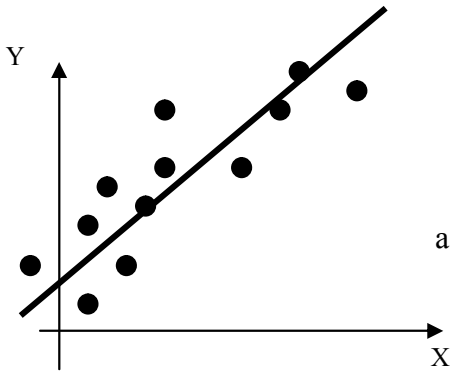
$\rho(X + a, Y + b) = \rho(X,Y)$

$\rho(\alpha.X + a, \beta.Y + b) = \rho(X,Y)$ si $\alpha > 0$ et $\beta > 0$

En fait, ρ mesure la corrélation linéaire des écarts (réduits) à la moyenne.



Droite des moindres carrés[21] / régression : réinventer la moyenne, la variance, le coefficient de corrélation linéaire.



$$y^* = a^* x + b^*$$

$$a^* = \frac{\sum_i (x_i - \bar{x}).(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \qquad b^* = \bar{y} - a^* \bar{x}$$

$$\bar{x} = \frac{\sum x_i}{n} \qquad \bar{y} = \frac{\sum y_i}{n}$$

$$y^* = \bar{y} + \rho . \frac{\sigma_Y}{\sigma_X} . (x - \bar{x})$$

* efficacité instrumentale , quant aux interprétations….

[21] « *Nouvelles méthodes pour la détermination des orbites des comètes* » , annexe « *Sur la méthode des moindres quarrés* »Adrien Marie LEGENDRE, 1805 */ regression* , mot introduit par Francis GALTON 1885

Une autre interprétation géométrique du coefficient de corrélation linéaire

Tableau centré-réduit : une variable-colonne X : $\sum x_i = 0$ MX=0 centrée

$$VX = \frac{1}{n}.\sum_{i=1}^{n}(x_i - 0)^2 = \frac{1}{n}.\sum_{i=1}^{n} x_i{}^2 = 1$$ réduite

et une autre variable Y $\sum y_i = 0$ $$VY = \frac{1}{n}.\sum_{i=1}^{n} y_i{}^2 = 1$$

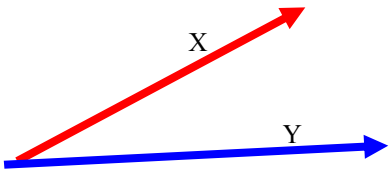
Tableau individus x variables : T

	…	X	…	Y	…	
1		x_1		y_1		
…		…		…		
…		…		…		
i		x_i		y_i		
…		…		…		
…		…		…		
…		…		…		
…		…		…		
…		…		…		
…		…		…		
n		x_n		x_n		

$$\rho(X,Y) = \frac{1}{n}\sum_i \frac{(x_i - MX)}{\sigma_X}.\frac{(y_i - MY)}{\sigma_Y} = \frac{1}{n}\sum_i x_i.y_i$$

car X et Y sont centrées-réduites

dans R^n



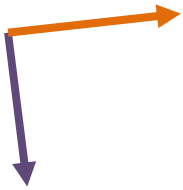
$$\cos(X,Y) = \frac{\langle X,Y\rangle}{\|X\|.\|Y\|} \qquad \|X\| = \sqrt{\sum_i x_i{}^2} = \sqrt{n} \qquad \|Y\| = \sqrt{\sum_i y_i{}^2} = \sqrt{n} \qquad \langle X,Y\rangle = \sum x_i.y_i$$

$$\cos(X,Y) = \frac{\sum_i x_i.y_i}{\sqrt{n}.\sqrt{n}} = \frac{1}{n}.\sum_i x_i.y_i = \rho(X,Y)\ !!!$$

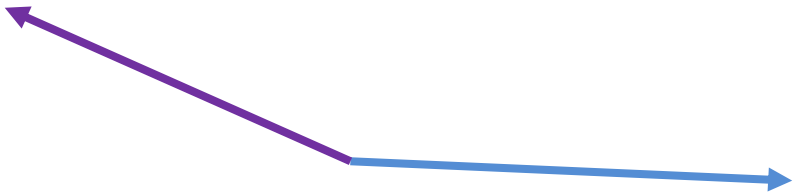
<u>variables- « flèches » , arguments de direction. Le lien entre deux variables se « voit » : c'est l'angle</u> .

angle aigu /même direction / cosinus proche de 1/ corrélation positive

Angle droit / directions ⊥ / cosinus nul / indépendance



angle obtus / directions opposées / cosinus proche de -1/ corrélation négative

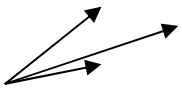


Ah ! les coefficients de corrélations …

- <u>la non-transitivité</u>

  $\rho(V_1,V_2)=0.71$ $\rho(V_2,V_3)=0.71$ $\rho(V_1,V_3)= ?$

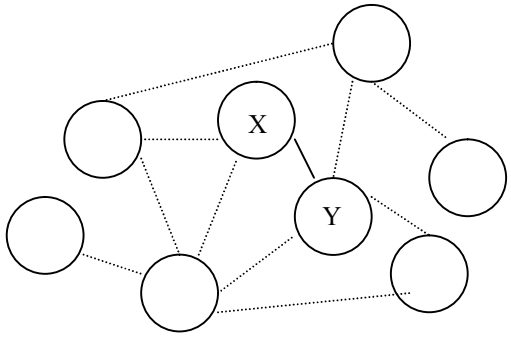
  $0 \leq \rho(V_1,V_3) \leq 1$



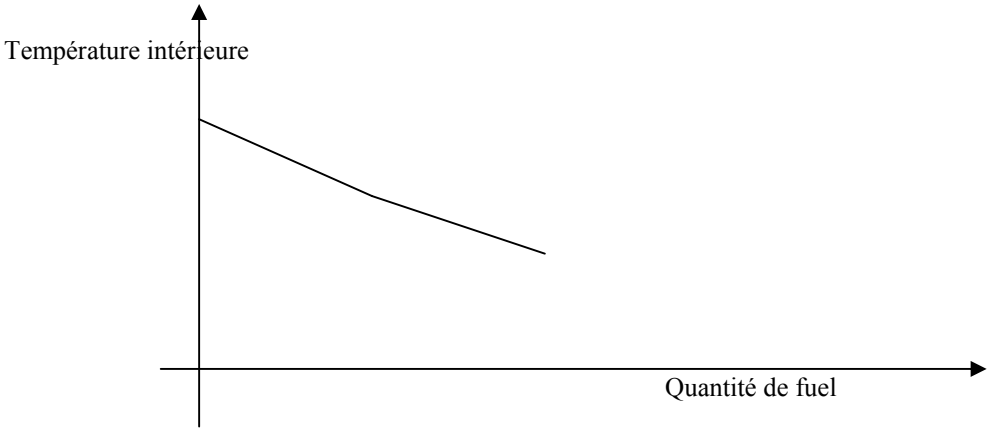
$\rho(V_1,V_2)=0.9$ $\rho(V_2,V_3)=0.9$ ……. $\rho(V_9,V_{10})=0.9$, $\rho(V_1,V_{10})= ?$

$-1 \leq \rho(V_1,V_{10}) \leq 1$

* effets directs, effets indirects, variables cachées

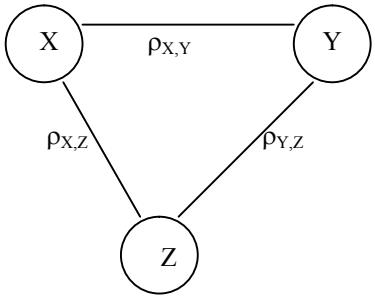


un classique :



Apparemment, plus on chauffe, plus il fait froid.

Variable cachée (Z) : température extérieure



On peut « neutraliser », par le calcul, l'effet de Z sur les rapports entre X et Y

$$\rho_Z(X,Y) = \frac{\rho_{X,Y} - \rho_{X,Z} \cdot \rho_{Y,Z}}{\sqrt{1-\rho^2_{X,Z}} \cdot \sqrt{1-\rho^2_{Y,Z}}}.$$

application numérique

X	Y	Z
Qté de fuel	**Température intérieure**	**température extérieure**
0	25	30
1	20	10
2	18	-5

les coefficients de corrélation:

	X	Y	Z
	Qté de fuel	**Température intérieure**	**température extérieure**
X	1	- 0,97072534	0,98718399
Y	- 0,97072534	1	0,98718399
Z	-0,9966159	0,98718399	1

attention aux décimales pour les calculs :

on a

ρ (X,Y)= ρ (Qté de fuel, Temp. intérieure)= - 0.97

apparemment, plus on chauffe, plus il fait froid.

puis

$$\rho_Z(X,Y) = \rho_{Temp.ext}(Qté\ fuel, Temp.\text{int}) = \frac{-0{,}9707-(-0{,}9966*0{,}9871)}{\sqrt{1-0{,}9966^2}.\sqrt{1-0{,}9871^2}.}$$

$$\rho_Z(X,Y) = \rho_{Temp.ext}(Qté\ fuel, Temp.\text{int}) = 0{,}99$$

Commentaires

l'intérêt de l'ACP : voir les liens entre toutes les variables…

Penser, mesurer et représenter la nature et l'intensité des liens entre des variables numériques. L'ACP : Analyse en Composantes Principales (normée)

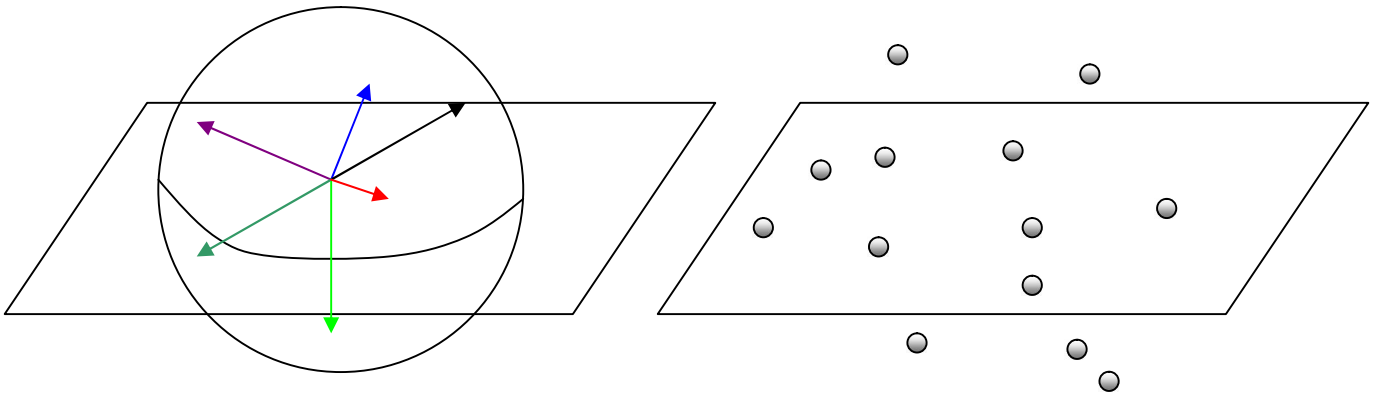
ACP normée : les variables du tableau sont centrées-réduites : on s'affranchit des unités de mesure.

Tableau individus x variables : T

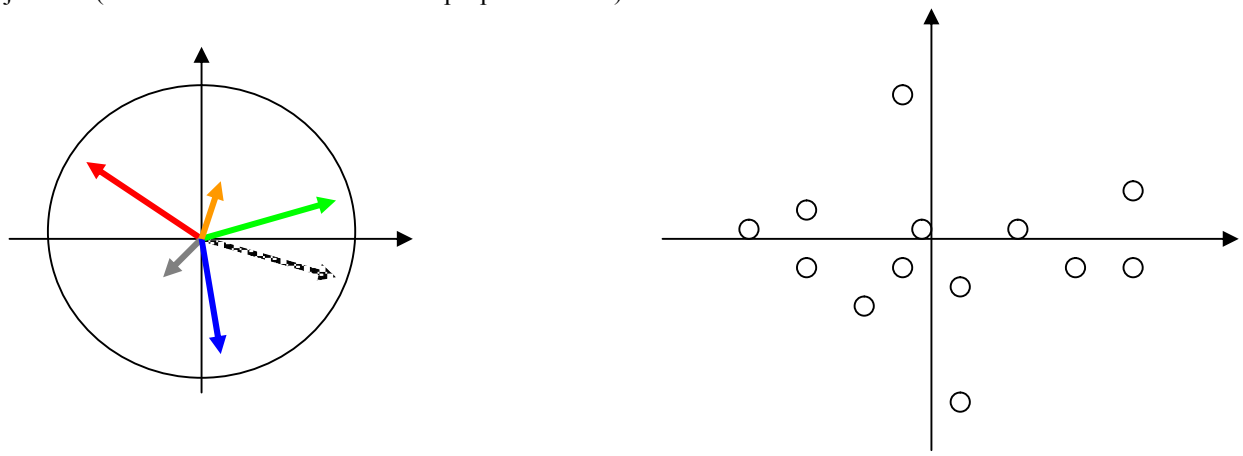
	V1	…..	…	Vj	…	Vm
1				$t_{1,j}$		
…				…		
…				…		
i	$t_{1,j}$			$t_{i,j}$		$t_{i,m}$
…				…		
…				…		
…				…		
…				…		
…				…		
…				…		
n				$t_{n,j}$		

La matrice des corrélations : $RO = \frac{1}{n}.{}^{t}T.T$

Les m variables-flèches dans R^n les n individus-points dans R^m



Projections (calculs des vecteurs et valeurs propres de RO)



Orientations / angles positions/distances

Quelques points de repère : ACP [22]

- c'est une méthode de statistique exploratoire qui permet de déceler d'éventuelles régularités statistiques dans les co-relations des différentes variables quantitatives (numériques).
- Les liens entre les variables quantitatives sont pensés et mesurés en termes de **corrélations linéaires**. Un énoncé tel que : «les précipitations (la pluie) en mm et le nombre de jours d'orages sont positivement (linéairement) corrélés » doit se comprendre ainsi : les villes où les précipitations sont supérieures à la moyenne (58 mm) sont aussi les villes où le nombre de jours d'orages est supérieur au nombre moyen (4 jours). l'ACP est une méthode différentielle : elle met en évidence des écarts à la moyenne.

- Les variables quantitatives sont représentées par des vecteurs, des flèches, d'une couleur. Lorsque les variables sont centrées-réduites , la corrélation est alors le cosinus de l'angle formé par deux variables-vecteurs. **La corrélation se lit sur l'angle** :
    - Corrélation fortement positive <=> cosinus proche de 1 <=> angle aigu
    - Corrélation nulle, « indépendance » <=> cosinus proche de 0 <=> angle droit
    - Corrélation fortement négative <=> cosinus proche de –1 <=> angle obtus

- Les variables sont donc, géométriquement, des vecteurs flèches dans un espace de grande dimension. Ces vecteurs sont projetés sur un plan optimal en un certain sens : la figure géométrique initiale est la moins déformée possible, mais elle l'est .
- La qualité de la représentation d'une variable , ça se voit : dans ACP normée (ie avec centrage-réduction des variables) ce qui est le cas dès lors que l'on considère des corrélations et non pas des covariances, chaque flèche a la même longueur . Si sa représentation est petite c'est qu'elle est plutôt perpendiculaire au plan de projection, donc indépendante de ce plan.

- Une variable « supplémentaire-illustrative » est une variable qui est projetée sur un plan déjà constitué. On peut observer ses liens statistiques avec les autres variables mais elle n'a pas servi à structurer le nuage initial des individus, contrairement aux autres variables dites « actives ». La distinction variable active/supplémentaire-illustrative est d'un extrême importance méthodologique, même si , pratiquement les représentations peuvent être très semblables.

[22] deux livres utiles « *Analyse des Données* » Michel Volle , Economica et « *Statistique Exploratoire Multidimensionnelle* » Ludovic Lebart, Alain Morineau, M. Piron Dunod, 1995

La pluie et le beau temps

Source « Le Monde » 1998. Les individus statistiques sont les villes où se sont déroulées les épreuves de la Coupe du Monde de Foot-ball
Les variables sont :
Temp$_b$: température minimale moyenne en °C.
Temp # : : température maximale moyenne en °C.
Mtemp : température moyenne en °C.
Pluiemm : précipitations en cumul moyen , en mm.
Norages : nombre moyen de jours d'orages.

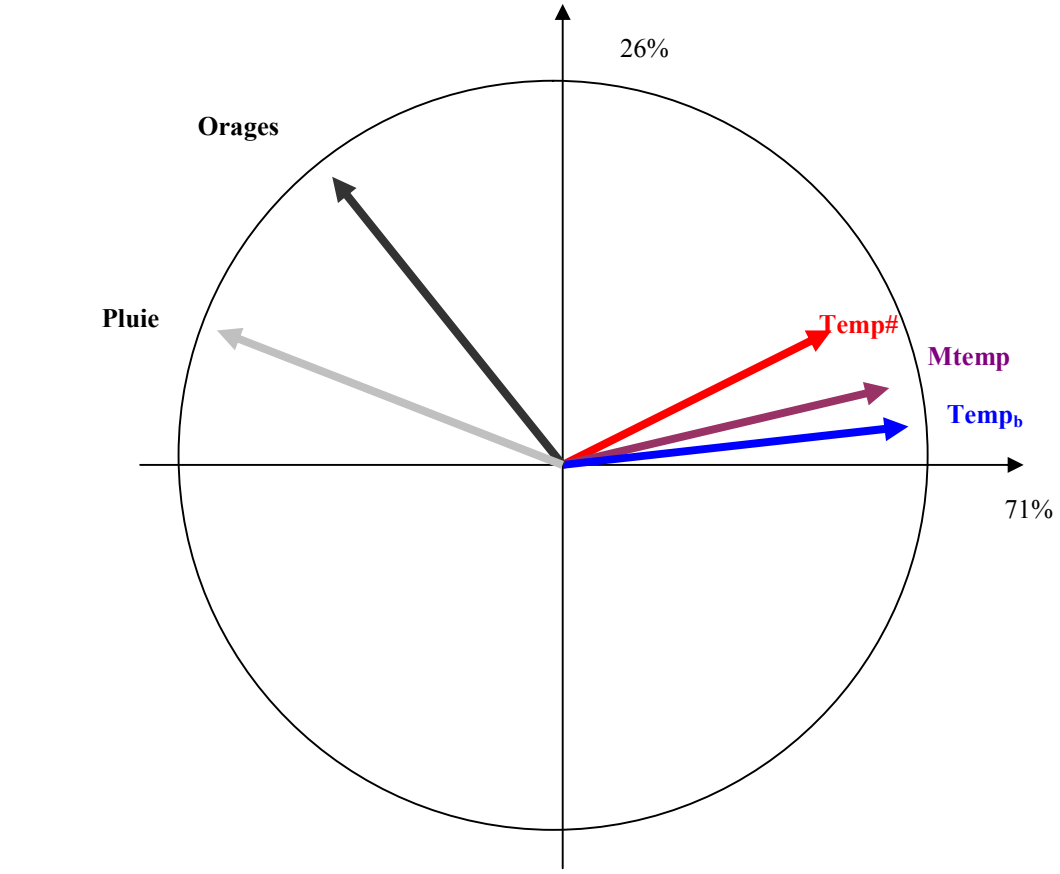
Pour chaque ville les variables sont des moyennes temporelles sur la période 10 juin-12 juillet de 1966 à 1995

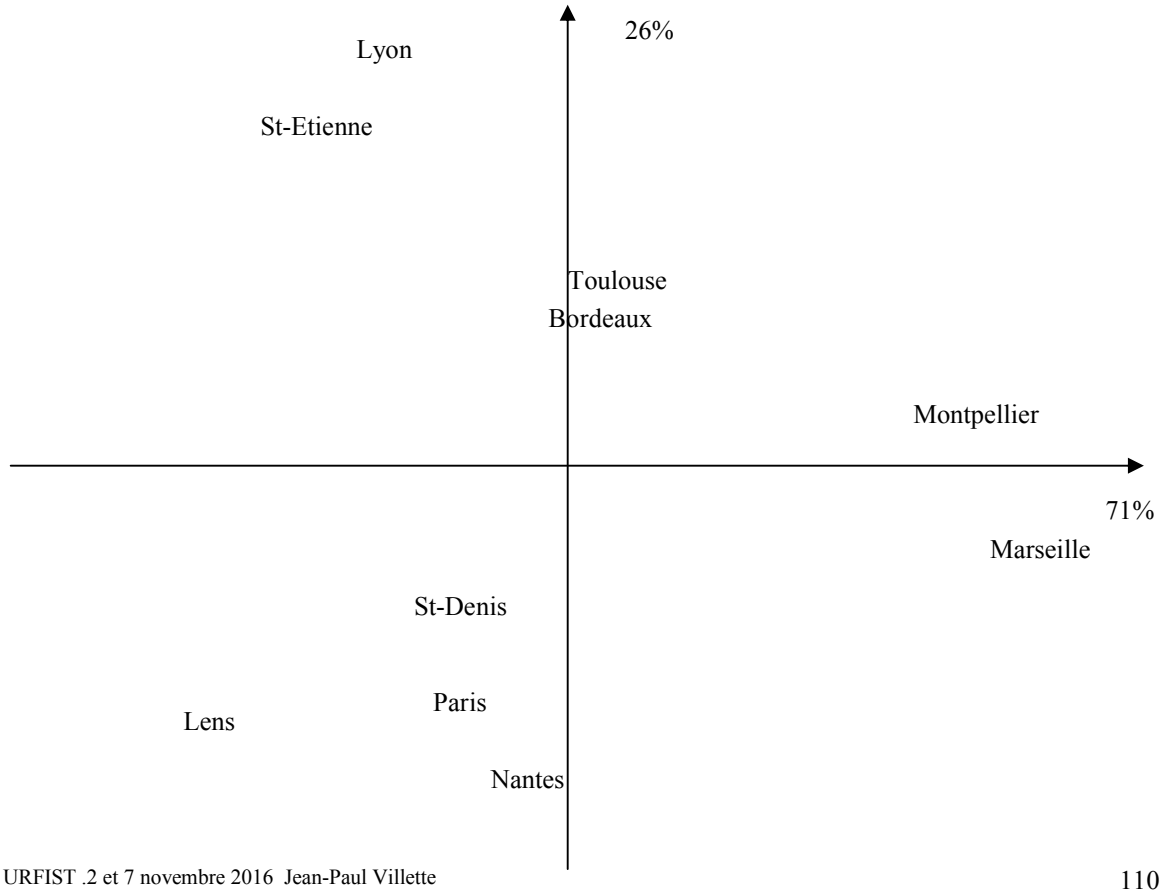
	Tempb	**Temp#**	**Mtemp**	**Pluiemm**	**Norages**
Bordeaux	14	25	19	62	5
Lens	12	21	16	73	4
Lyon	14	25	19	79	7
Marseille	17	27	22	24	2
Montpellier	16	27	21	33	4
Nantes	13	23	18	50	2
Paris	14	23	18	57	3
St-Denis	12	23	18	60	4
St-Etienne	12	24	18	82	7
Toulouse	14	25	20	67	5

Pour ces 10 villes :

	moyenne	écart-type
Tempb	14 °C	1.6 °C
Temp#	24 °C	1.8 °C
Mtemp	19 °C	1.7 °C
Pluiemm	59 mm	18 mm
Norages	4 jours	1.7 jours

La pluie et le beau temps : une Analyse en Composantes Principales





ANALYSE EN COMPOSANTES PRINCIPALES
STATISTIQUES SOMMAIRES DES VARIABLES CONTINUES
EFFECTIF TOTAL : 10 POIDS TOTAL : 10.00

NUM . IDEN - LIBELLE	EFFECTIF	POIDS	MOYENNE	ECART-TYPE	MINIMUM	MAXIMUM
1 . C2 - Tempb	10	10.00	13.80	1.60	12.00	17.00
2 . C3 - Temp=	10	10.00	24.30	1.79	21.00	27.00
3 . C4 - Mtemp	10	10.00	18.90	1.64	16.00	22.00
4 . C5 - Pluie	10	10.00	58.70	17.84	24.00	82.00
5 . C6 - Orages	10	10.00	4.30	1.68	2.00	7.00

MATRICE DES CORRELATIONS

	C2	C3	C4	C5	C6
C2	1.00				
C3	0.86	1.00			
C4	0.91	0.96	1.00		
C5	-0.79	-0.57	-0.70	1.00	
C6	-0.35	0.04	-0.17	0.77	1.00

MATRICE DES VALEURS-TESTS

	C2	C3	C4	C5	C6
C2	99.99				
C3	4.07	99.99			
C4	4.78	6.28	99.99		
C5	-3.36	-2.06	-2.73	99.99	
C6	-1.16	0.12	-0.55	3.24	99.99

VALEURS PROPRES
APERCU DE LA PRECISION DES CALCULS : TRACE AVANT DIAGONALISATION .. 5.0000
SOMME DES VALEURS PROPRES 5.0000
HISTOGRAMME DES 5 PREMIERES VALEURS PROPRES

NUMERO	VALEUR PROPRE	POURCENT.	POURCENT. CUMULE	
1	3.5535	71.07	71.07	**********
2	1.3047	26.09	97.16	*******
3	0.0933	1.87	99.03	***
4	0.0424	0.85	99.88	*
5	0.0061	0.12	100.00	*

RECHERCHE DE PALIERS ENTRE (DIFFERENCES SECONDES)

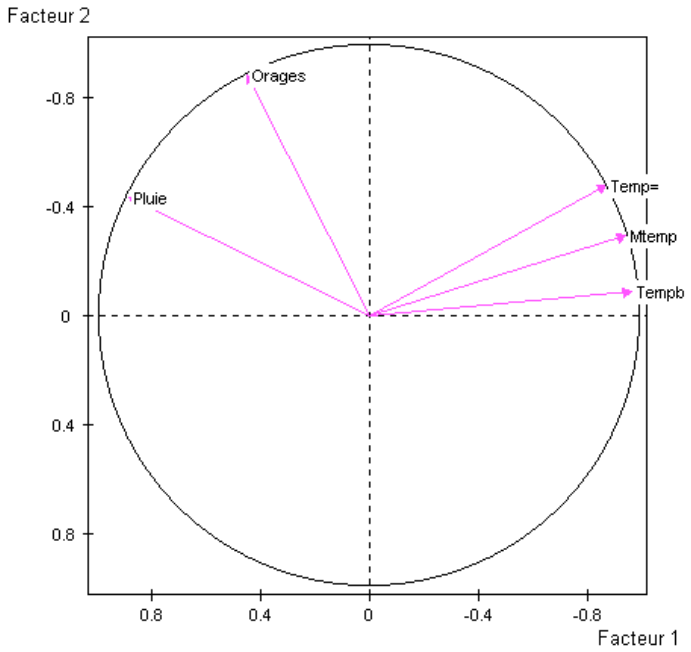
PALIER ENTRE	VALEUR DU PALIER	
1-- 2	1037.38	**********

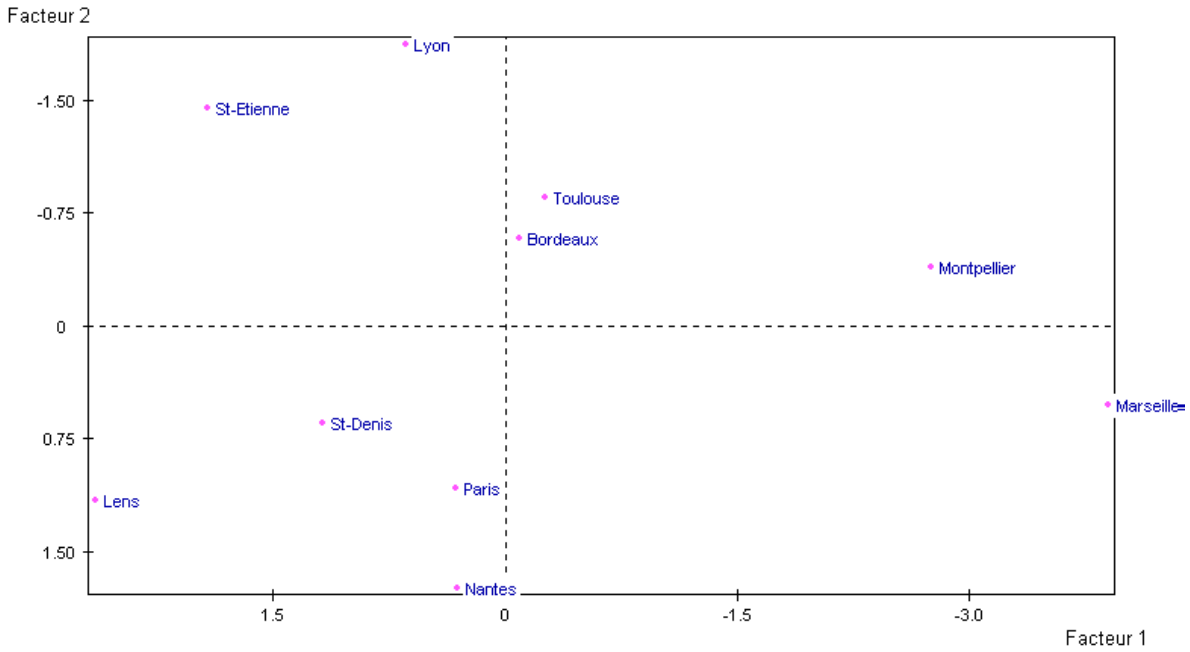
COORDONNEES DES VARIABLES SUR LES AXES 1 A 5
VARIABLES ACTIVES

VARIABLES	COORDONNEES					CORRELATIONS VARIABLE-FACTEUR					ANCIENS AXES UNITAIRES				
IDEN - LIBELLE COURT	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
C2 - Tempb	-0.96	-0.09	0.26	-0.02	0.00	-0.96	-0.09	0.26	-0.02	0.00	-0.51	-0.08	0.85	-0.09	
C3 - Temp=	-0.87	-0.48	-0.09	-0.03	0.06	-0.87	-0.48	-0.09	-0.03	0.06	-0.46	-0.42	-0.28	-0.16	
C4 - Mtemp	-0.94	-0.30	-0.09	0.13	-0.04	-0.94	-0.30	-0.09	0.13	-0.04	-0.50	-0.26	-0.29	0.63	-
C5 - Pluie	0.88	-0.43	0.10	0.13	0.03	0.88	-0.43	0.10	0.13	0.03	0.47	-0.38	0.34	0.63	
C6 - Orages	0.45	-0.89	0.00	-0.09	-0.03	0.45	-0.89	0.00	-0.09	-0.03	0.24	-0.78	0.00	-0.42	-

COORDONNEES, CONTRIBUTIONS ET COSINUS CARRES DES INDIVIDUS
AXES 1 A 5

INDIVIDUS			COORDONNEES					CONTRIBUTIONS					COSINUS CARRES				
IDENTIFICATEUR	P.REL	DISTO	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Bordeaux	10.00	0.38	-0.09	-0.59	0.04	-0.09	0.14	0.0	2.6	0.2	2.1	32.1	0.02	0.90	0.00	0.02	0.05
Lens	10.00	8.46	2.64	1.16	0.35	-0.14	-0.08	19.6	10.3	13.0	4.5	11.4	0.82	0.16	0.01	0.00	0.00
Lyon	10.00	4.06	0.64	-1.88	0.36	0.00	0.00	1.2	27.0	13.8	0.0	0.0	0.10	0.87	0.03	0.00	0.00
Marseille	10.00	15.51	-3.90	0.52	0.07	0.12	-0.05	42.8	2.1	0.6	3.6	3.5	0.98	0.02	0.00	0.00	0.00
Montpellier	10.00	7.91	-2.76	-0.39	-0.12	-0.39	-0.03	21.4	1.2	1.5	35.6	1.6	0.96	0.02	0.00	0.02	0.00
Nantes	10.00	3.20	0.31	1.74	-0.22	0.09	0.14	0.3	23.2	5.2	1.8	32.3	0.03	0.95	0.02	0.00	0.01
Paris	10.00	1.45	0.32	1.08	0.44	0.02	0.00	0.3	8.9	20.8	0.1	0.0	0.07	0.80	0.13	0.00	0.00
St-Denis	10.00	2.13	1.17	0.65	-0.57	0.00	-0.10	3.9	3.2	34.4	0.0	16.5	0.65	0.20	0.15	0.00	0.00
St-Etienne	10.00	5.89	1.92	-1.45	-0.31	-0.07	0.02	10.4	16.1	10.5	1.3	0.4	0.63	0.36	0.02	0.00	0.00
Toulouse	10.00	1.01	-0.26	-0.85	-0.04	0.46	-0.04	0.2	5.5	0.2	50.9	2.1	0.07	0.72	0.00	0.21	0.00


Facteur 2
Orages
Pluie
Temp=
Mtemp
Tempb
-0.8
-0.4
0
0.4
0.8
0.8
0.4
0
-0.4
-0.8
Facteur 1


Facteur 2
Lyon
St-Etienne
Toulouse
Bordeaux
Montpellier
Marseille
St-Denis
Paris
Lens
Nantes
-1.50
-0.75
0
0.75
1.50
1.5
0
-1.5
-3.0
Facteur 1

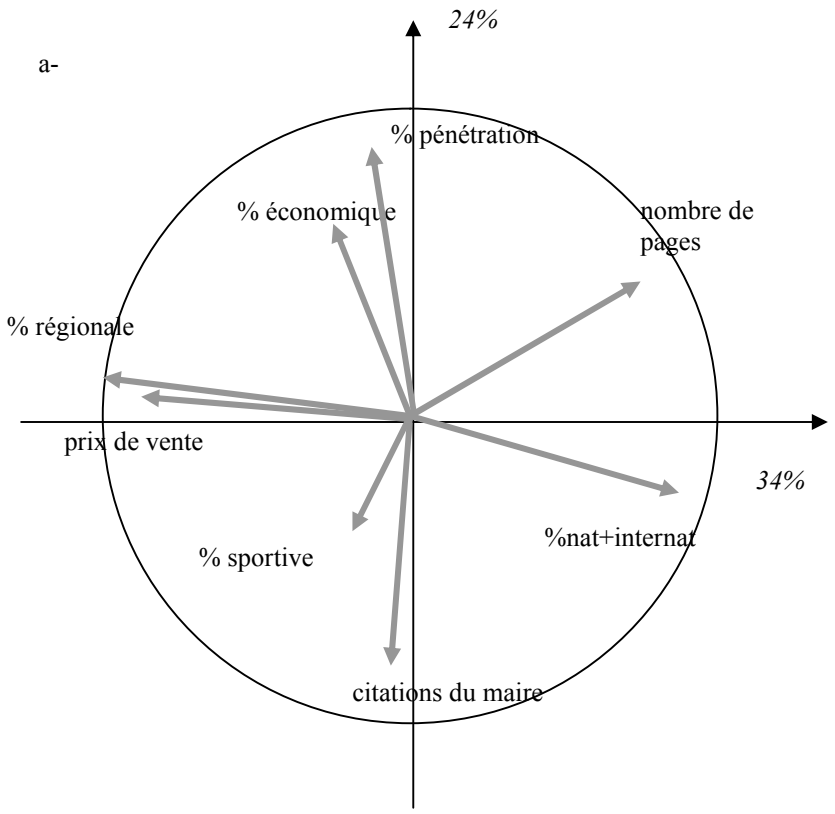
– quoditiens régionaux, ACP, Classification Hiérarchique Ascendante

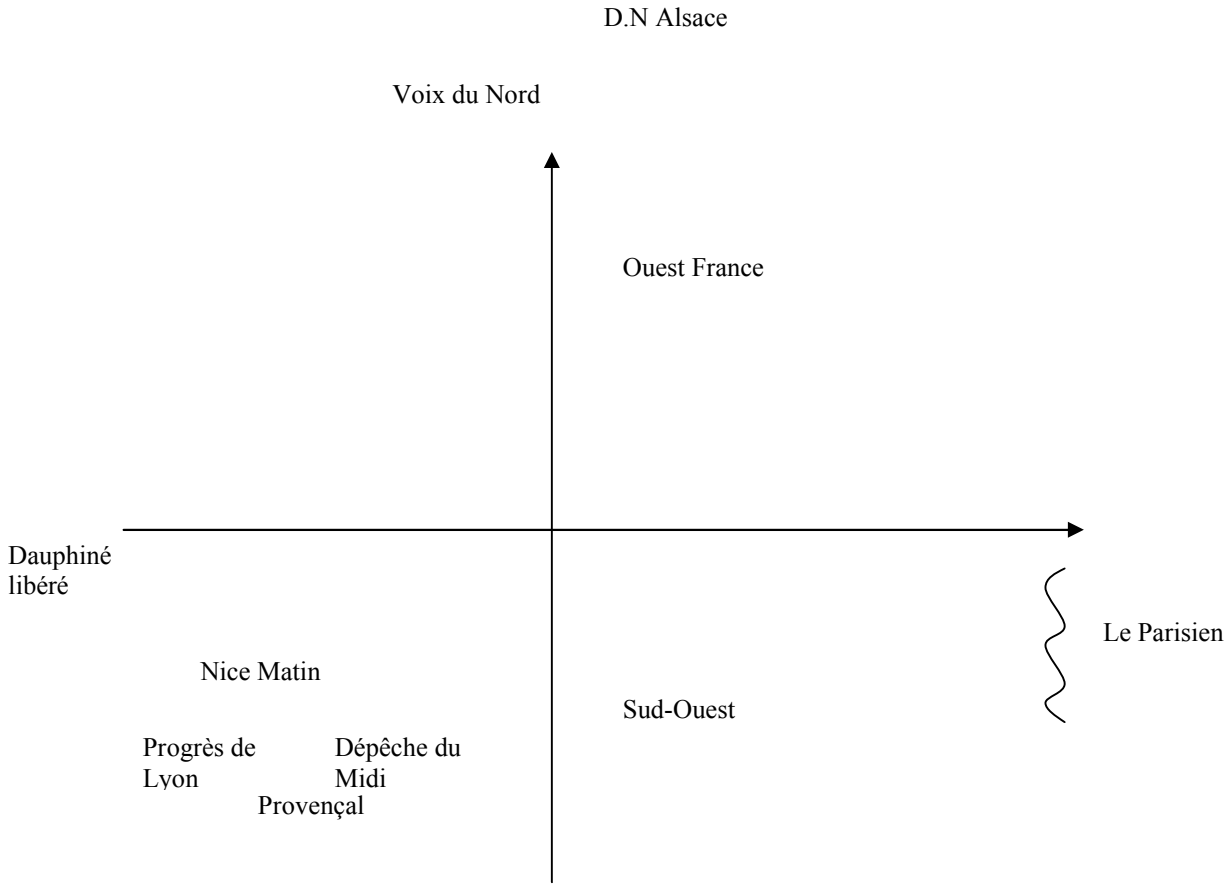
Source : « Capital « fév 94, « *dix quotidiens régionaux au banc d'essai* ». Les individus sont les quotidiens : **Ouest-France, Sud-Ouest,la Voix du Nord**… et les variables sont : *le nombre de pages, le pourcentage de pages régionale et locale, le % de pages nationale et internationale, le % de pages sportive, le % de pages économiques, le prix de vente, le taux de pénétration et le nombre de citations du maire.*

	nbre de pages	% pages régionales	% pages nat. et intern.	% pages sportive	% pages économique	prix de vente F	taux de pénétration [23]	citations du maire[24]
Ouest-France	42	33	18	12	2,3	4	40	4
Sud-Ouest	30	26	20	23	2,3	4	39	4
La voix du Nord	31	29	11	9	7,1	4	35	4
le Dauphiné Libéré	28	36	13	15	0,3	4,4	27	1
Le Progrès	30	28	12	16	1,6	4,5	23	5
Nice-Matin	33	32	10	20	2,1	4	40	7
La dépêche du Midi	30	28	12	18	2	4,2	28	5
les DNA	52	22	11	15	3	4,2	53	0
le Provençal	29	26	13	15	0,3	4,2	35	9
Le Parisien	48	14	25	13	1	3,5	ndisp.	5

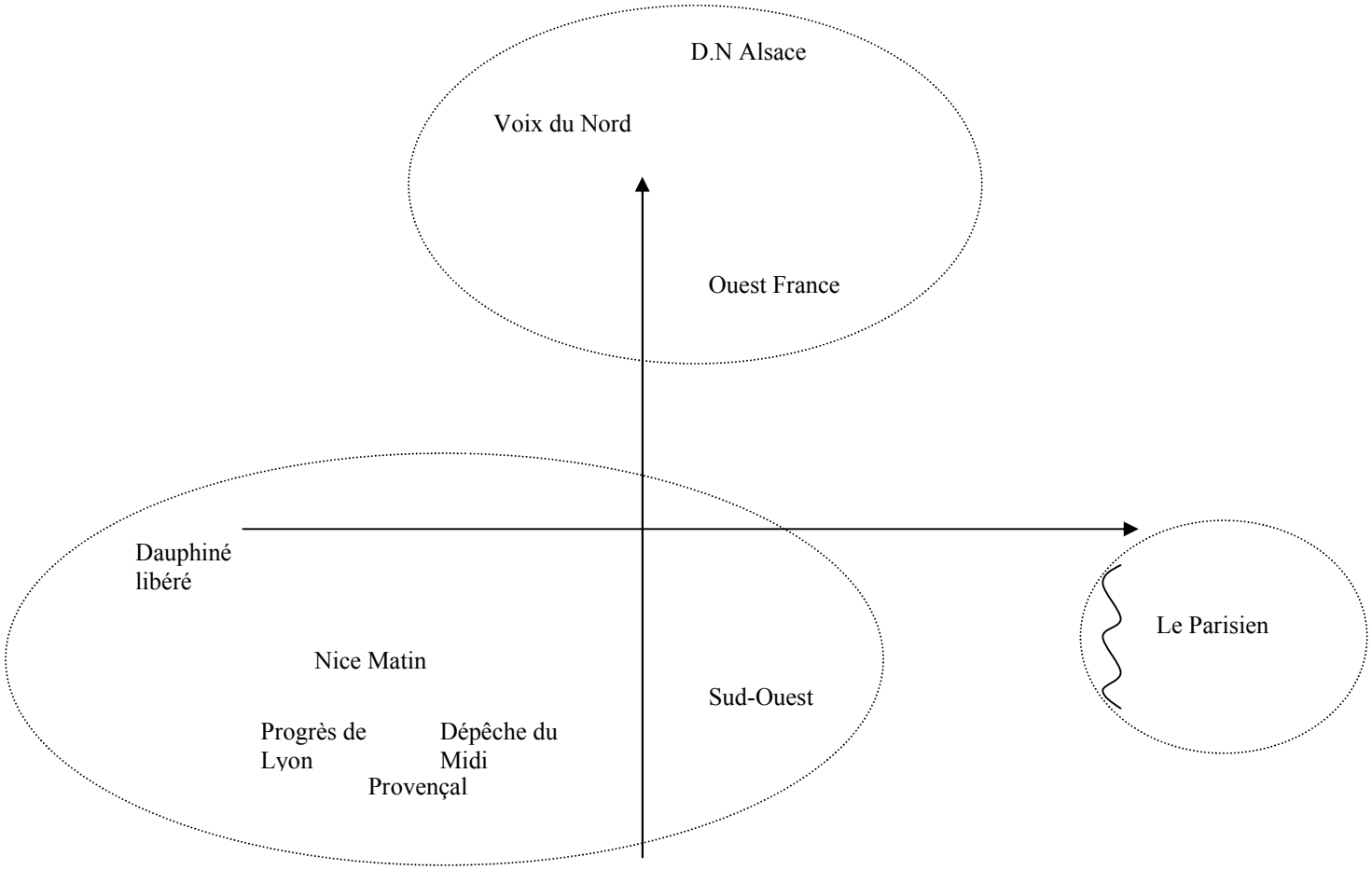
[23] *pourcentage de foyers qui lisent le quotidien dans la zone de diffusion*
[24] *nombre d'articles consacrés au maire de la ville siège du journal durant la période de l'enquête*

a-


24%
% pénétration
% économique
nombre de pages
% régionale
prix de vente
34%
%nat+internat
% sportive
citations du maire


D.N Alsace
Voix du Nord
Ouest France
Dauphiné libéré
Le Parisien
Nice Matin
Sud-Ouest
Progrès de Lyon
Dépêche du Midi
Provençal

janvier 05: une Classification Hiérarchique Ascendante (partie suivante du cours) permet de calculer trois groupes.



6– Données climatiques :Analyse en Composantes Principales et Classification Hiérarchique Ascendante

Voici le tableau [25]des 12 mois (les individus) de l'année et de 6 variables numériques

$Temp_b$: température minimale moyenne du mois, en °C
$Temp^{\#}$: température maximale moyenne du moins, en °C
Record min : température minimale absolue, en °C
Record max : température maximale absolue, en °C
Pluie mm : précipitations, quantité d'eau tombée de l'atmosphère, en mm
Soleil heures : durée d'ensoleillement, en heures

Strasbourg	**Tempb**	**Temp#**	**Record min**	**Record max**	**Pluie mm**	**Soleil heures**
janvier	-0,9	4	-16,8	15,7	25	51
février	-0,2	8	-9,9	20,4	42	100
mars	2,7	11,6	-6,8	24,1	33	134
avril	5,2	16,1	-3,8	27,8	45	178
mai	10,3	20,8	3,1	30,9	83	213
juin	12,9	23,6	5,8	33,1	80	230
juillet	13,7	24,3	6,7	33,9	68	214
août	14,2	25,7	4,8	36,3	51	249
septembre	10,6	21,4	3,2	30,8	55	182
octobre	7,2	15,1	-3,6	26,4	54	96
novembre	2,3	8,1	-9,9	18,2	67	62
décembre	0,7	5,5	-15,1	16,1	60	39

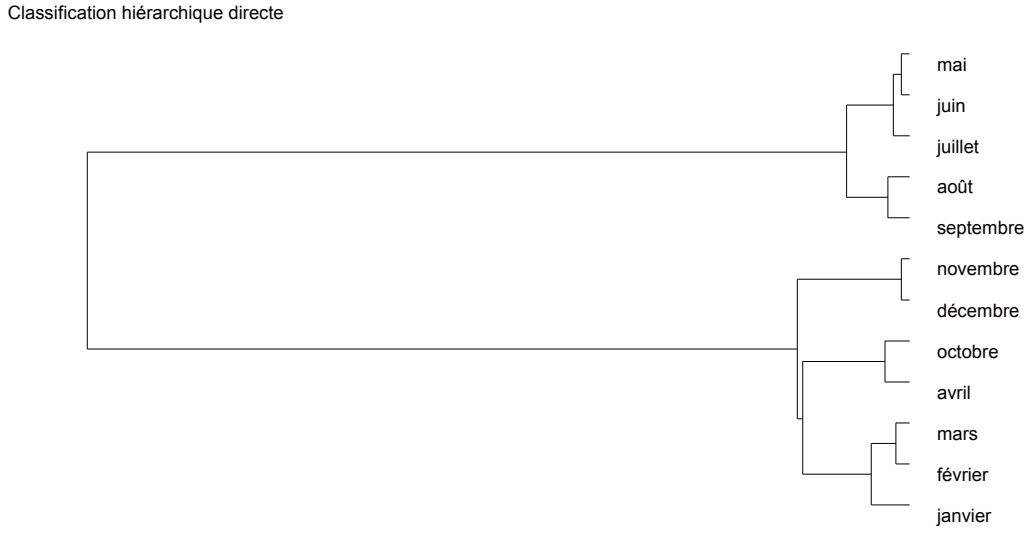
Analyse en Composantes Principales (voir page 5)

que peut-on dire du lien entre les variables « soleil » et « pluie » ?

que peut-on dire du moins de janvier ? (3 choses)

[25] Source : INSEE « *Tableaux de l'économie alsacienne* » 2002.données mensuelles de 1996 à 2000

Classification Hiérarchique Ascendante (CHA)



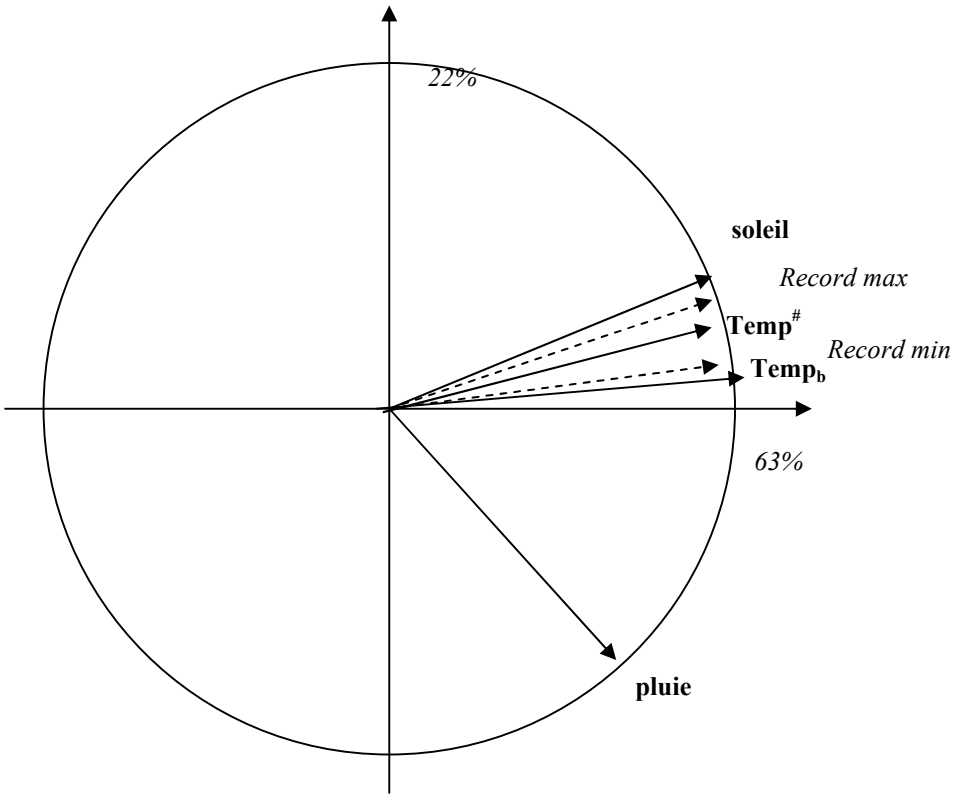
quelle est la règle statistique pour déterminer des groupes?

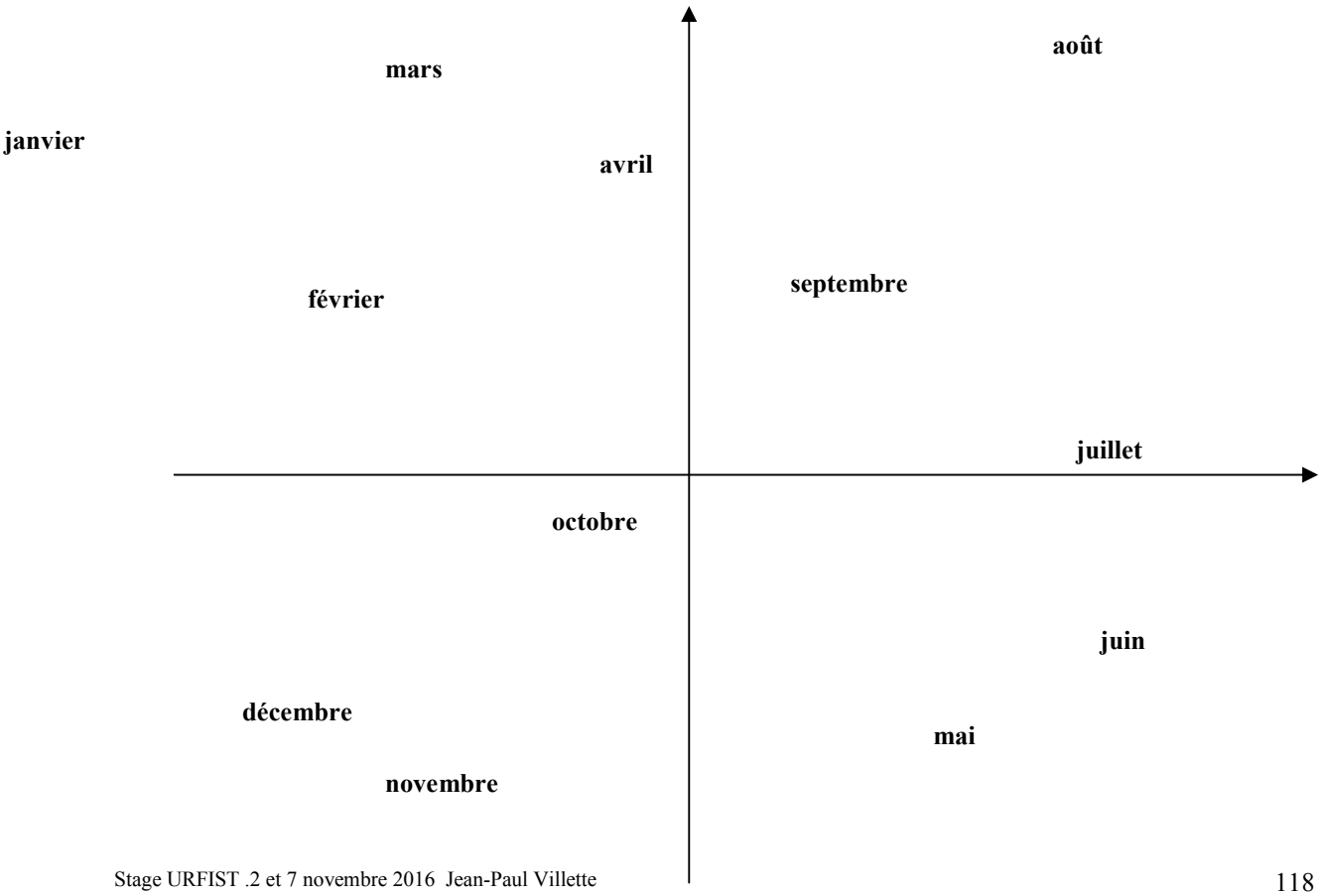
combien voyez-vous de groupes ? les décrire et les dessiner sur le graphe des individus de l'ACP qui se trouve à la page suivante.

Quel est le mois le plus proche du mois d'avril?

Quel est le mois le plus atypique?

vous pouvez joindre cette page à votre copie, ne pas y écrire votre nom





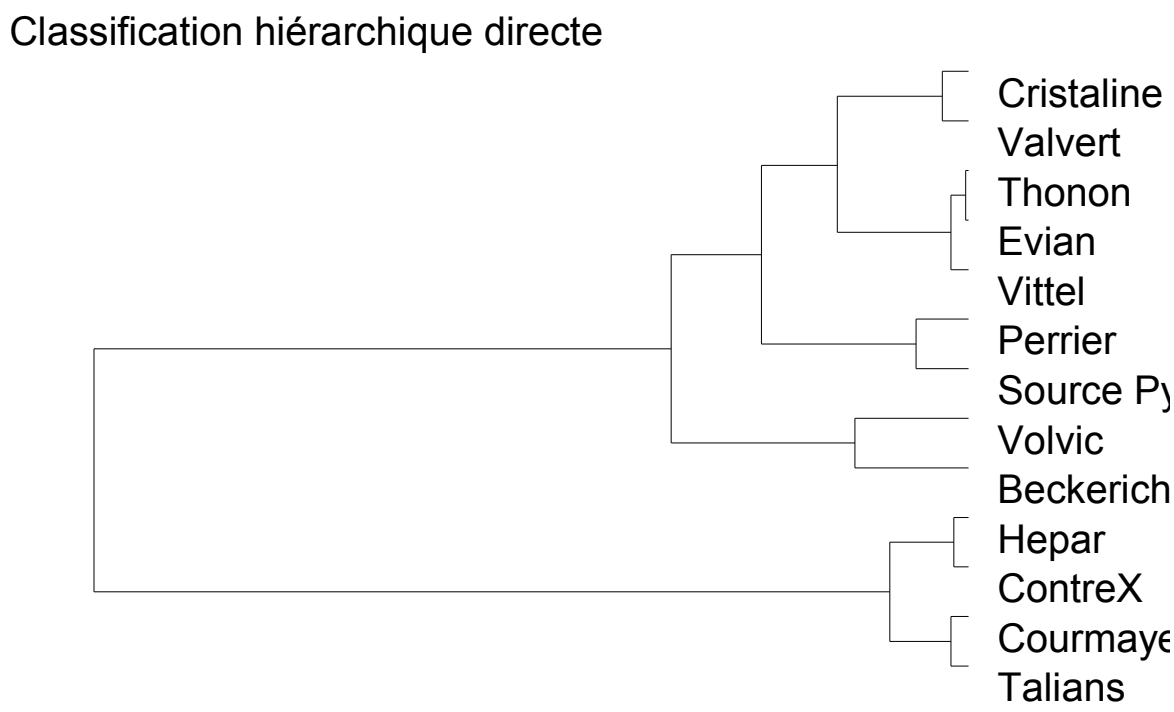
6– Eaux minérales :Analyses en Composantes Principales et Classification Hiérarchique Ascendante

Voici le tableau de 13 eaux minérales (les individus) : **Talians, Courmayeur,…Volvic** et 6 variables numériques (la quantité de mg par litre (mg/l) de **Calcium** (Ca), **magnésium** (Mg), **potassium** (K), **sodium** (Na) , **carbonates**(HCO3) et **sulfates** (S04).

	Calcium	Magnésium	potassium	sodium	carbonates	sulfates
Talians	596	77	2	7	290	1530
Courmayeur	517	67	2	1	168	1371
Hepar	555	110	4	14	403	1479
ContreX	486	84	3,2	9,1	403	1187
Vittel	202	36	2	3,8	402	306
Thonon	103,2	16,1	1,4	5,1	332	14
Evian	78	24	1	5	357	10
Valvert	67,6	2	0,2	1,9	204	18
Volvic	9,9	6,1	5,7	9,4	65,3	6,9
Source Pyrénées	49	12	1	35	186	17
Perrier	155	6,8	0	37	445	46,1
Beckerich	84	32	7	37	333	146
Cristaline	6,4	1,2	0,5	3	20	5

A- *répondre sur la page suivante* : que peut-on dire de l'eau « ContreX» ? (3 choses)

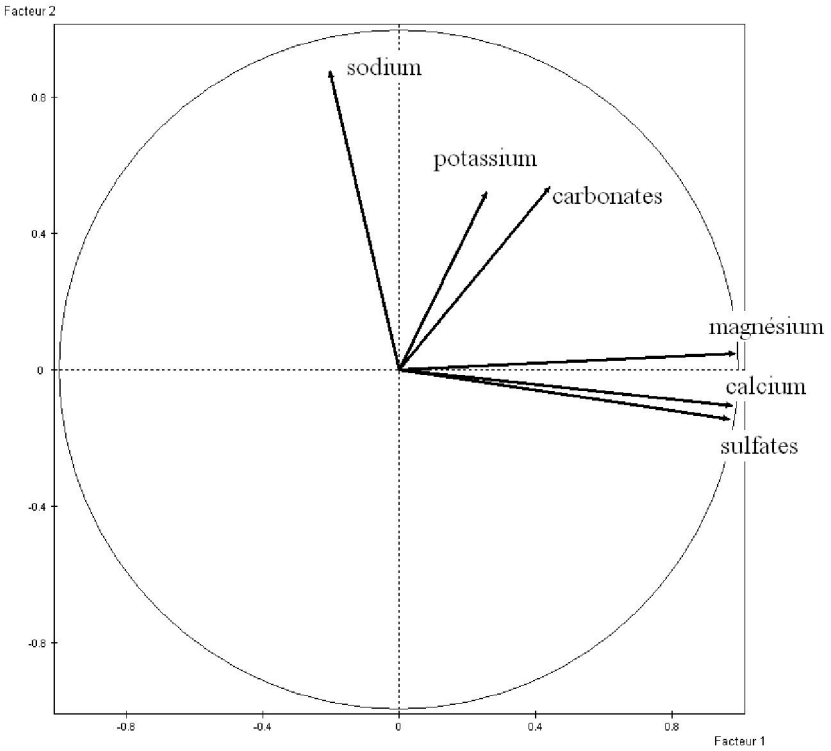
C- Classification Hiérarchique Ascendante (CHA)

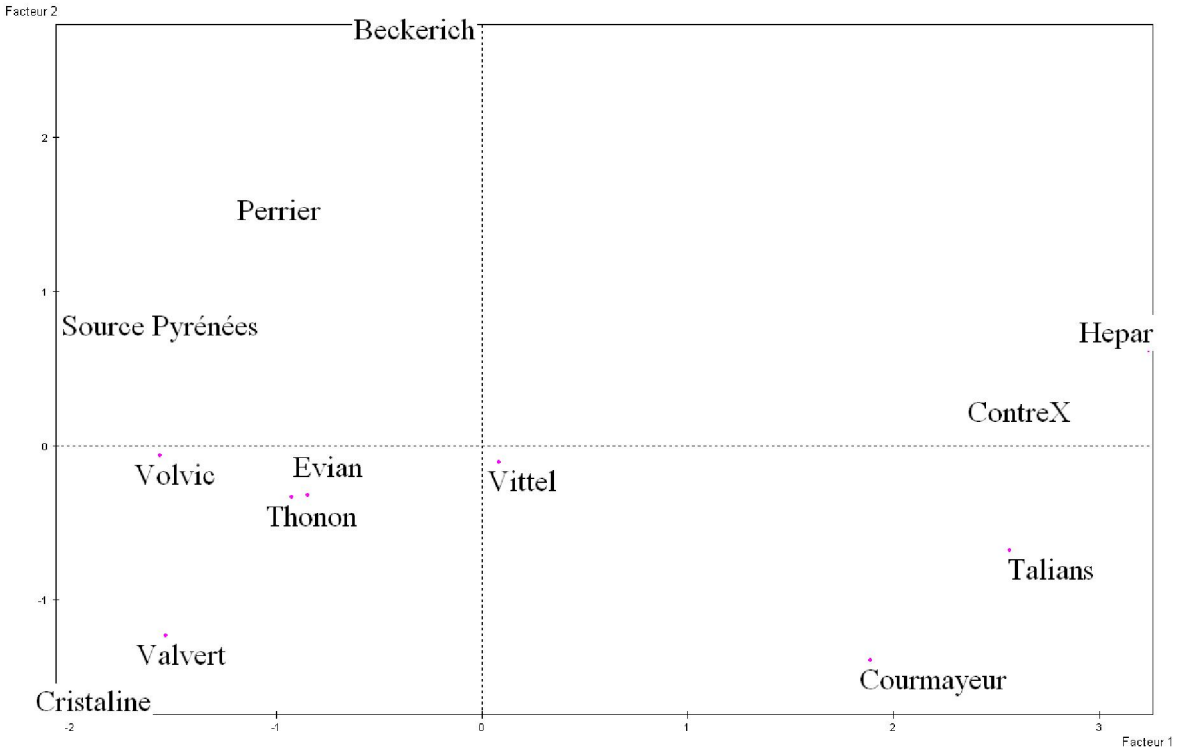


b- quelle est la règle statistique pour déterminer des groupes?
c- combien voyez-vous de groupes ? les représenter sur le graphe des individus de la page 4
d- quelles sont les deux eaux minérales les plus atypiques?

vous pouvez joindre cette page à votre copie, ne pas y écrire votre nom

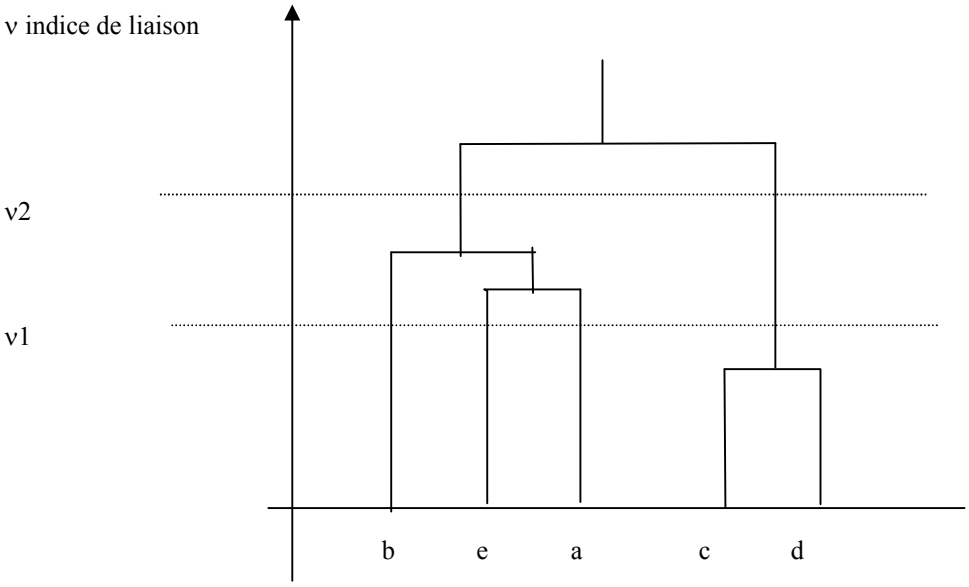
eaux minérales





Classification Hiérarchique Ascendante (CHA)

* à quoi ça ressemble, et comment « ça marche »
dendrogramme , du grec *dendron* arbre



ν_1 : { b } { e } { a } { c, d }

ν_2 : { b, e, a } { c, d }

- Il s'agit de répartir, automatiquement, les éléments (les objets) d'un ensemble en « paquets » (*clusters*, grappes) les plus intra-homogènes et extra-hétérogènes possibles.
- On constitue un paquet avec les deux objets les plus proches. Un processus d'agglomérations successives des éléments/paquets doit être défini par une « stratégie d'agglomération »

« rappels »

partition $\mathcal{P}$ d'un ensemble E

$\mathcal{P} = \{ E_i / i=1,\ldots k \}$ $\cup E_i = E$ $E_i \cap E_j = \varnothing$ si $i \neq j$

ex : les départements constituent une partition de la France

$\mathcal{D}$ = { Ain, Aisne, Allier,…} , la réunion des départements constituent la France entière, et deux départements sont disjoints.

rem : la distinction **classification/ classement**

Une CHA, c'est un processus d'agrégation, depuis les éléments jusqu'à l'ensemble lui-même. A un niveau croissant de l'indice ν correspond une partition de moins en moins fine

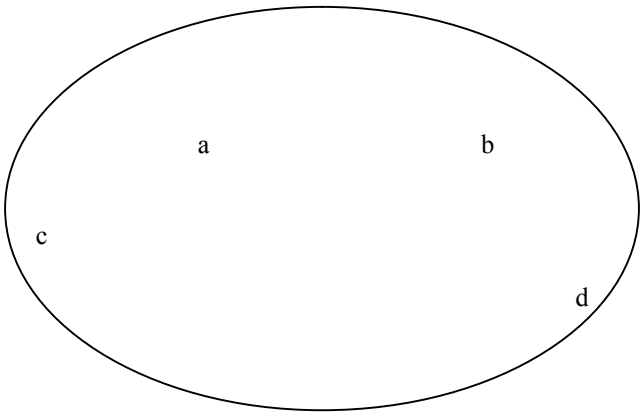
l'adjectif « hiérarchique » indique que les partitions sont « emboîtées »,. Par exemple les partitions { régions} et { départements} de la France sont emboîtées.

Définitions et procédures

Une hiérarchie $\mathcal{H}$ est un ensemble de partitions « emboîtées », par exemple les partitions « régions » et « départements » sont emboîtées.

$\mathcal{H} = \{ \mathcal{P}1, \mathcal{P}2, \mathcal{P}3 \}$

$\mathcal{P}1 = \{ \{a\}, \{b\}, \{c\}, \{d\} \}$
$\mathcal{P}2 = \{ \{a\}, \{b\}, \{c, d\} \}$
$\mathcal{P}3 = \{ a, b, c, d \}$



$\mathcal{H}$ est une hiérarchie ⇔

i- $\forall e \in E$, $\{e\} \in \mathcal{H}$. la partition la plus fine est un objet de la hiérarchie.

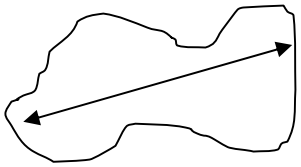
ii- $E \in \mathcal{H}$, la partition la moins fine est un objet de la hiérarchie.

iii- $A, B \in \mathcal{H}$, alors $A \cap B \in \{ A , B , \varnothing \}$ deux paquets sont sans élément communs ou bien l'un est contenu dans l'autre.

exemples :
« Alsace » ∩ »Bourgogne »=
« Alsace « ∩ »Bas-Rhin »=
« Alsace » ∩ »France » =

Hiérarchie indicée

Idée de grosseur , diamètre d'un ensemble



$\delta : \mathcal{H} \longrightarrow R^+$
$A \qquad \delta(A)$

Avec $\delta(e)= 0 \quad \delta(E)= 1$, et $A \subset B \Rightarrow \delta(A) \leq \delta(B)$

L'indice ν associé à une hiérarchie

- définition d'une distance (métrique)

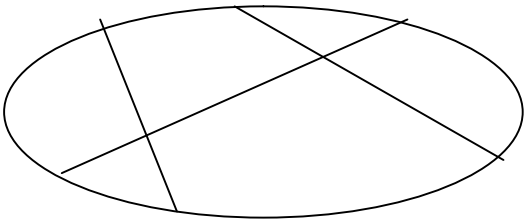
$d : ExE \longrightarrow R^+$
$(x, y) \qquad d(x,y)$

i- d(x,x)=0
ii- d(x,y)=d(y,x) symétrie
iii-d(x,z) ≤ d(x,y)+d(y,z) inégalité triangulaire

une **ultra-métrique** est une distance particulière

i- d(x,x)=0
ii- d(x,y)=d(y,x)
iii'-d(x,z) ≤ Max[d(x,y), d(y,z)] tous les triangles sont isocèles.

* ultramétrique associée à une partition



$E_1 \cup E_2 \cup E_3 \cup E_4 \cup E_4 \cup E_6 = E$ $\quad E_i \cap E_j = \varnothing$ si $i \neq j$

définition de u :

x, y $\in$ E

i- u(x,x)= 0

ii- u(x,y)= 1 si x et y appartiennent à des classes différentes

$\rho > 0$ ($0 < \rho < 1$) si $x \neq y$ mais dans la même classe

ex : les objets sont les villes de France, la partition est celle des Régions

u(Strasbourg,Paris)= 1
u(Strasbourg,Mulhouse)= 0.5
u(Lyon,Lyon)= 0

théorème
une ultramétrique est équivalente à une hiérarchie indicée (CHA)

- enjeux, conséquences pratiques

1 - $\mathcal{H}$, δ une hiérarchie indicée

On pose $u(x,y)=\delta(H_{x,y})$ $H_{x,y}$ étant le plus petit « paquet » contenant à la fois x et y.

u est bien une ultramétrique
$u(x,x)=\delta(\{x\})=0$ Cf i- dans la définition d'une hiérarchie
$u(y,x)=u(x,y)$
et on peut démontrer que
$u(x,z) \leq \text{Max}[u(x,y), u(y,z)]$

2 réciproquement

si u est une ultramétrique on définit l'indice Δ pour un sous-ensemble A de E

$$\Delta(A)=\frac{diamètre(A)}{diamètre(E)}$$

avec diamètre (A) = Max u(x,y) x, y ∈ E diamètre (A) = Max u(x,y) x, y ∈ E

il s'ensuit que $\Delta(\{x\}) = u(x,x)/\ \Delta(E)=0$

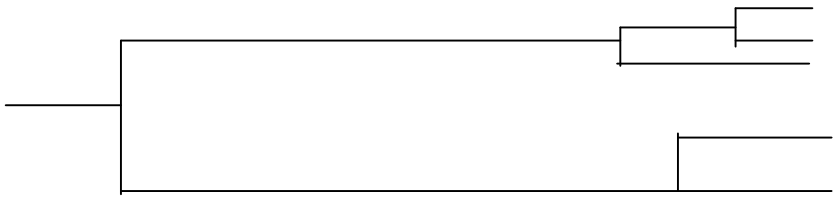
$\Delta(E) = \Delta(E)/\Delta(E)=1$ et on peut montrer que $A \subset B \Rightarrow \Delta(A) \leq \Delta(B)$

une grande question existentielle : mais au fait, est-ce qu'il existe des groupes , ou bien l'ensemble des objets n'est-il qu'un continuum ?

avec une CHA , on contourne la question de l'existence de groupes par une réponse paramétrée : ν indicateur de liaison, à quoi correspond un ordre d'apparition des agrégations

où regarder, où couper ?

une règle[26] : couper les branches les plus longues



résultats persistants, robustes quand on change la distance/ stratégie d'agrégation ?

quelle distance ? : le ppv génétique de la cigogne c'est le pélican brun, le ppv morphoanatomique c'est le héron cendré. Le champignon est le végétal le plus proche, génétiquement, de l'homme.

Dans la phrase bien connue est Analyse de Données Textuelles (ADT) : « *les poules du couvent couvent* », la distance orthographique entre « *couvent* » et « *couvent* » est nulle, la distance phonétique est faible, la distance sémantique est très forte.

* Stratégies d'agrégations : distance entre paquets

Quelques distances entre paquets :

$d_{max}(A,B)$= Max { d(x,y) / x∈ A , x∈ B }
$d_{min}(A,B)$= Min { d(x,y) / x∈ A , x∈ B }

$$d_{moy}(A,B)=\frac{\sum_{x\in A, y\in B} d(x,y)}{card(A).card(B)}$$

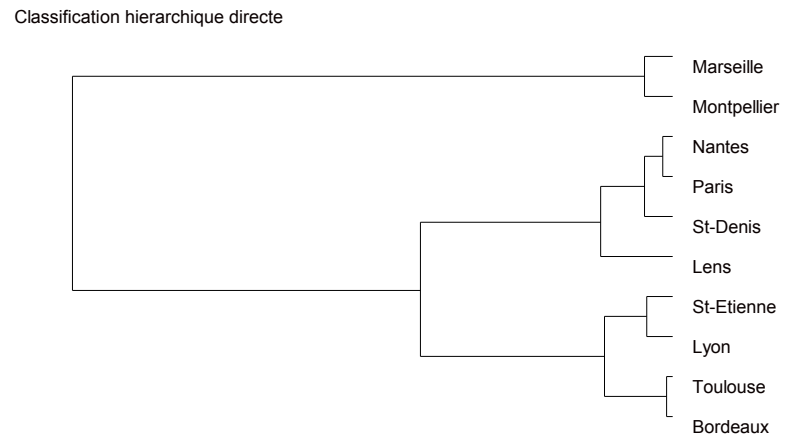
logiciel SPAD

Une CHA n'est possible qu'après une AFCM (Analyse Factorielle des Correspondances Multiples) ou une ACP (Analyse en Composantes Principales)

L'utilisateur n'a pas l'embarras du choix des distances et des stratégies d'agrégation. La distance est fondée sur les composantes des individus sur les premiers axes de projection. Très ingénieux et très pratique.

[26] Cf Michel VOLLE « Analyse des Données », page 291…

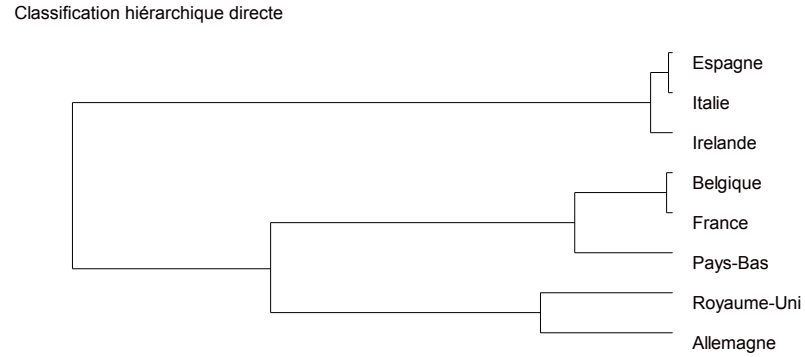
*données de la page *** de ce cahier :*



Examen sept 04

7– Sentiment d'appartenance religieuse, une Classification Hiérarchique Ascendante

On a considéré huit pays: Allemagne, Belgique, Espagne, France, Irlande, Italie, Pays-Bas et calculé une distance fondée sur des variables numériques (sentiment d'appartenance en % , catholique, protestant, autre religion, sans religion). Source « La Croix « nov 1998.



quelle est la règle statistique pour déterminer des groupes?

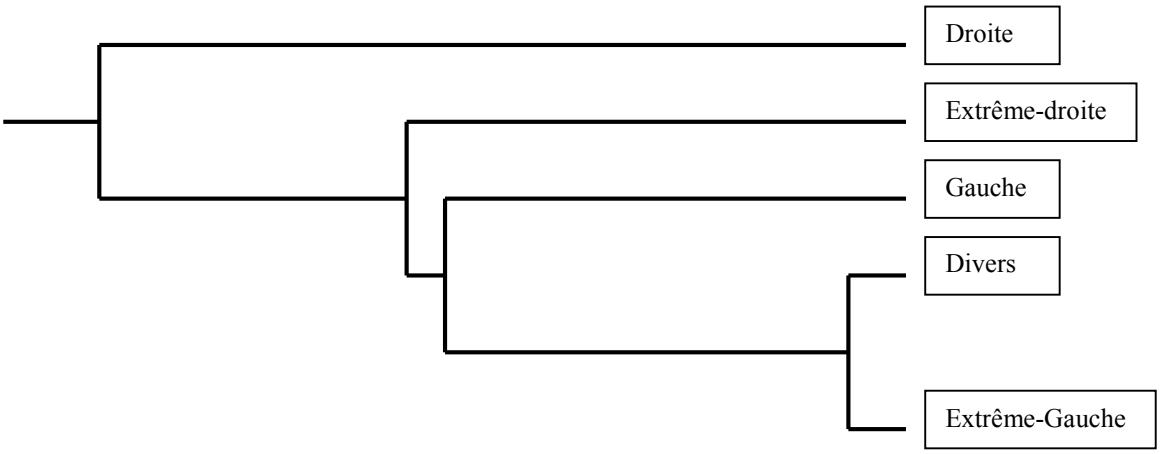
combien voyez-vous de groupes ? les décrire

4– pyramide des âges des électorats …, une Classification Hiérarchique Ascendante

source SOFRES mars 2004 . 1er tour des élections régionales. La distance est calculée à partir de la pyramide des âges des différents électorats. Pour simplifier je note:

Extrême-droite : FN / extrême-droite
Droite : UMP, UDF, divers droite
Gauche : PC, PS, verts, divers gauche
Extrême-gauche : LO, LCR, extrême-gauche
Divers : les autres

Voici le dendrogramme obtenu :



e- quelle est la règle statistique pour déterminer des groupes?

f- quels sont les 2 meilleurs groupes ? les décrire.

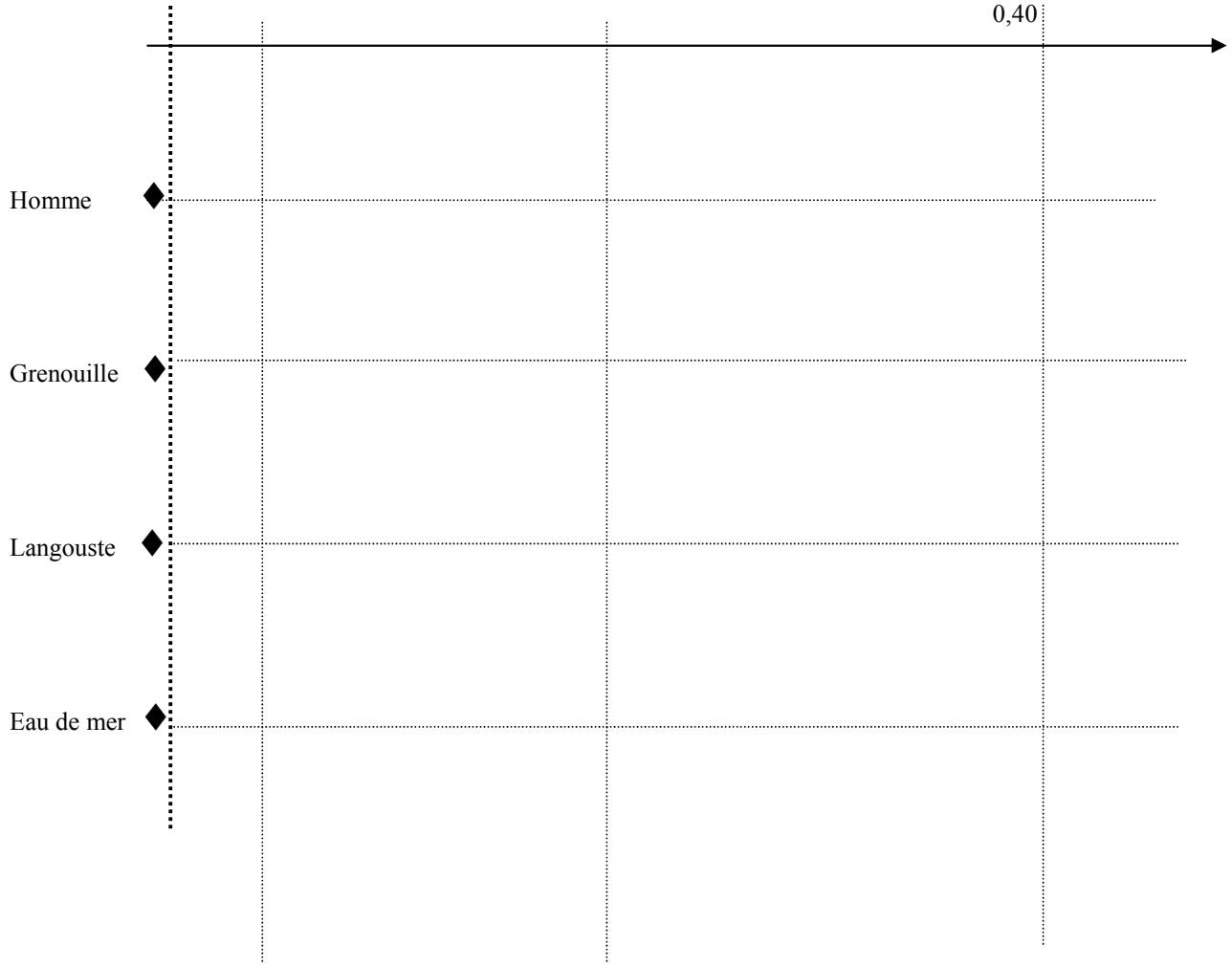
g- quel est l'électorat dont la pyramide des âges est la plus atypique ? pourquoi ?

Evolution. Stratégie d'agrégation : la moyenne

	homme	grenouille	langouste	eau de mer
homme	0	**0,04**	0,16	0,46
grenouille		0	0,14	0,56
langouste			0	0,20
eau de mer				0

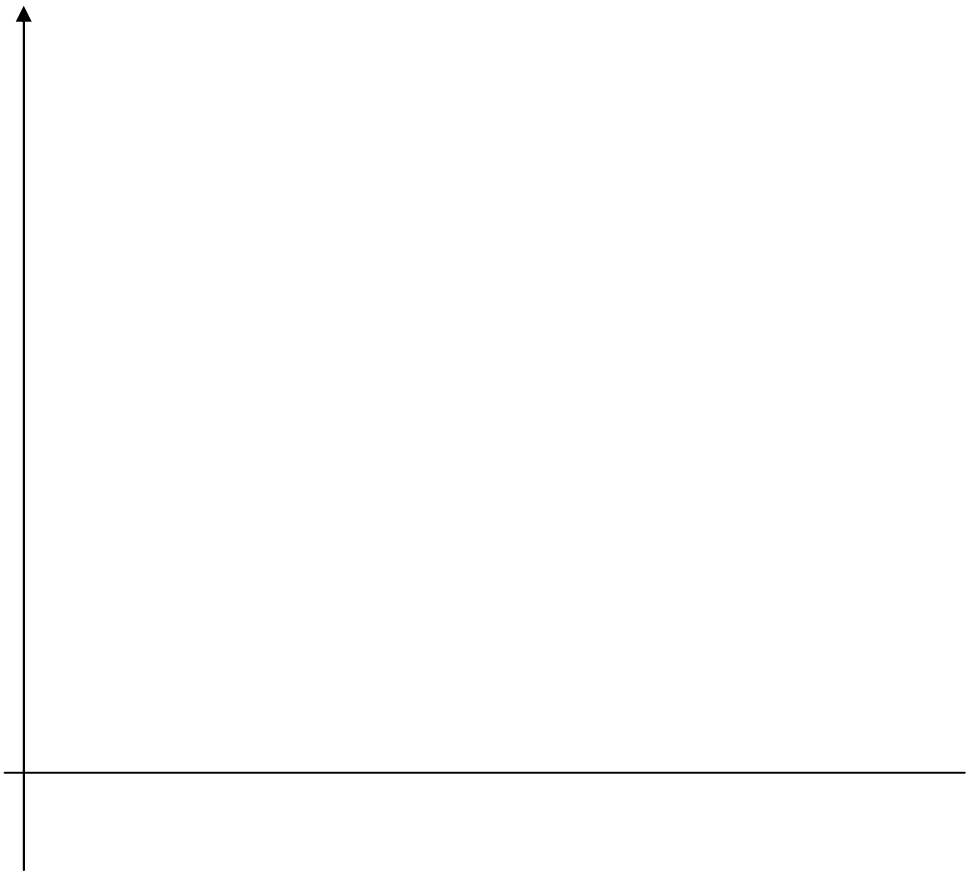
	{homme,grenouille}	langouste	eau de mer
{homme,grenouille}	0	**0,15**	0,51
langouste		0	0,20
eau de mer			0

	{homme,grenouille, langouste}	eau de mer
{homme,grenouille,langouste}	0	**0,40**
eau de mer		0

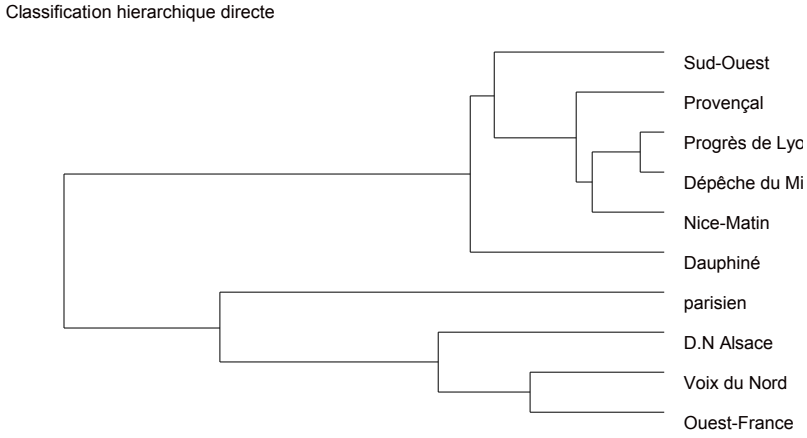


2 – corrélations, distances, Classification Hiérarchique Ascendante de variables

distance	**$temp_b$**	**Temp#**	**pluie**	**orages**
$temp_b$	0	0.2	3.6	2.7
Temp#		0	3.4	2.3
pluie			0	0.5
orages				0

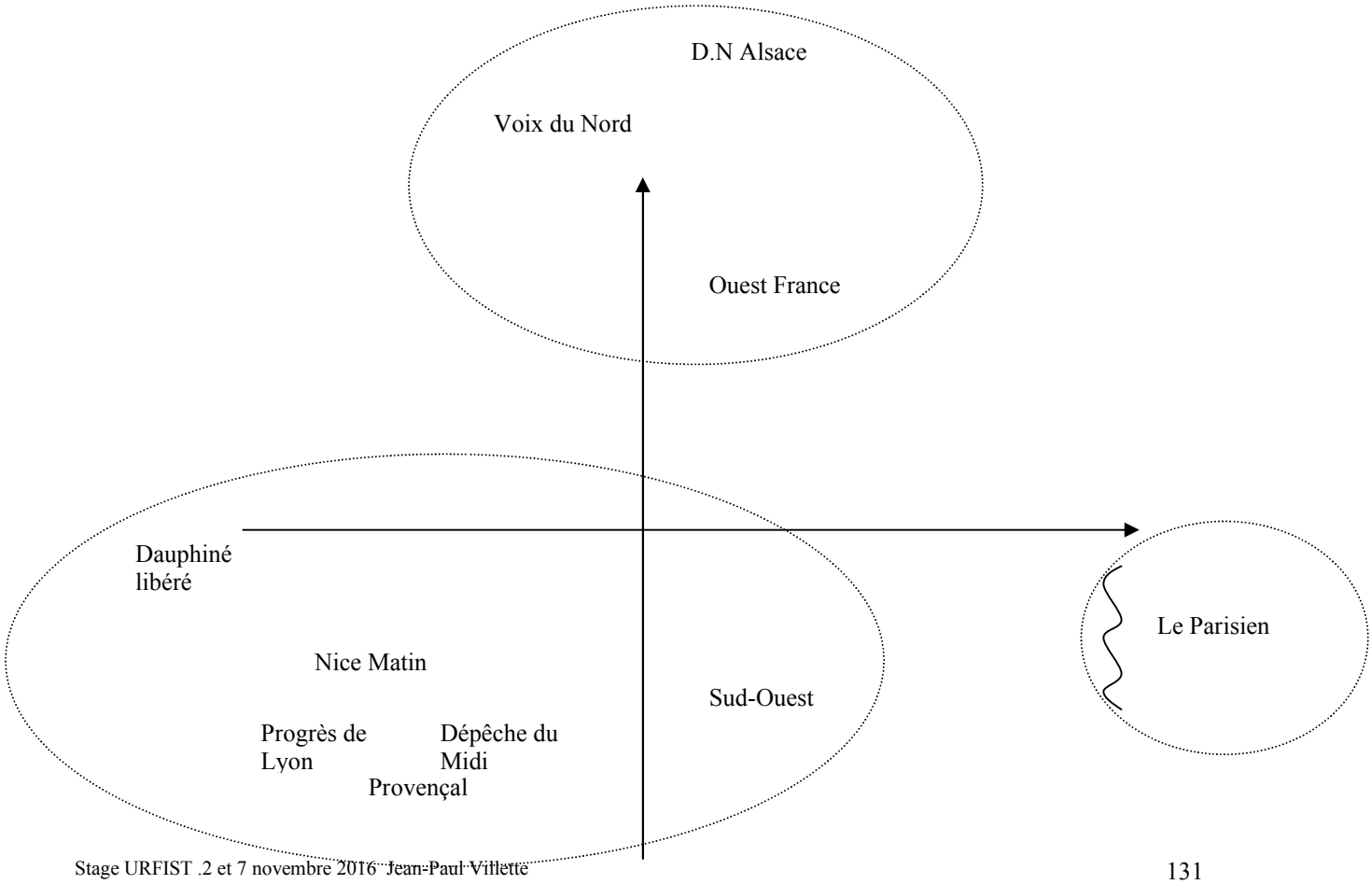


Examen 2003 :Voici le dendrogramme obtenu :



a-combien voyez-vous de groupes ?(il y a deux bonnes réponses possibles) les décrire. Quelle est la règle ?
b-quels sont les deux quotidiens les plus proches ?
c-quel est le quotidien le plus atypique ? pourquoi ?

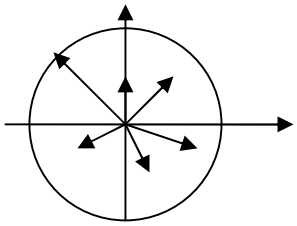
une typologie des individus sur le graphe des individus de l'ACP (page *), calculée par la CHA ci-dessus

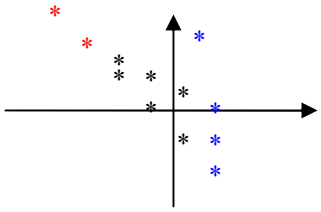


Une stratégie d'investigation :Analyse en Composantes Principale suivie d'une Classification Hiérarchique Ascendante

variables numériques, variables qualitatives *illustratives*	, variables numériques actives

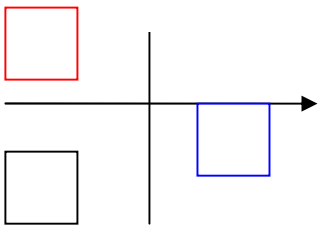
ACP



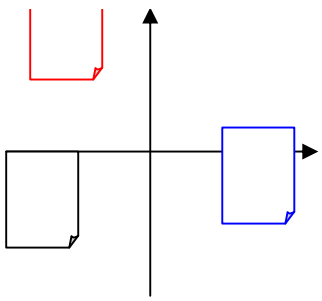


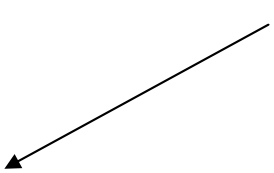
CHA, la production des clusters

Les variables numériques **clivantes**



*Les % des modalités **clivées** par les clusters, les moyennes des variables numériques illustratives éventuellement différenciées par les clusters*

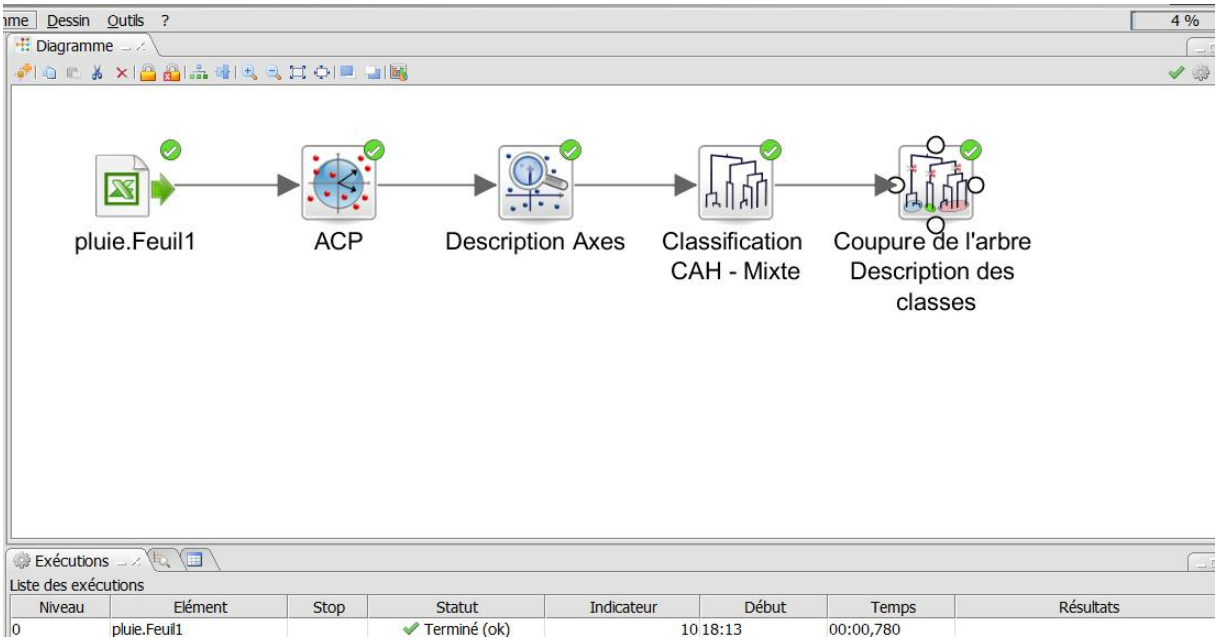




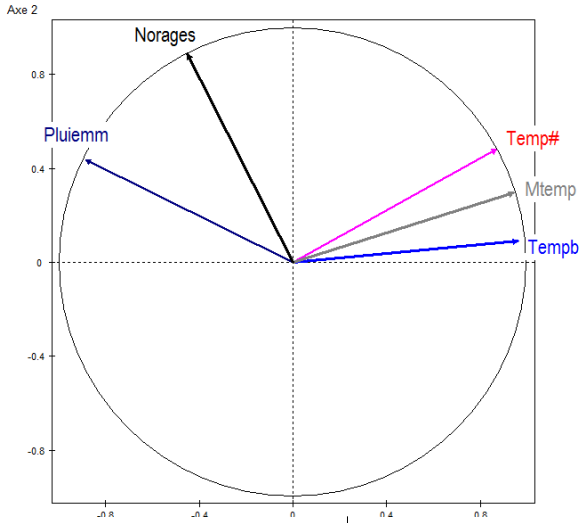
Entrer ce tableau dans Excel

	Tempb	Temp#	Mtemp	Pluiemm	Norages
Bordeaux	14	25	19	62	5
Lens	12	21	16	73	4
Lyon	14	25	19	79	7
Marseille	17	27	22	24	2
Montpellier	16	27	21	33	4
Nantes	13	23	18	50	2
Paris	14	23	18	57	3
St-Denis	12	23	18	60	4
St-Etienne	12	24	18	82	7
Toulouse	14	25	20	67	5

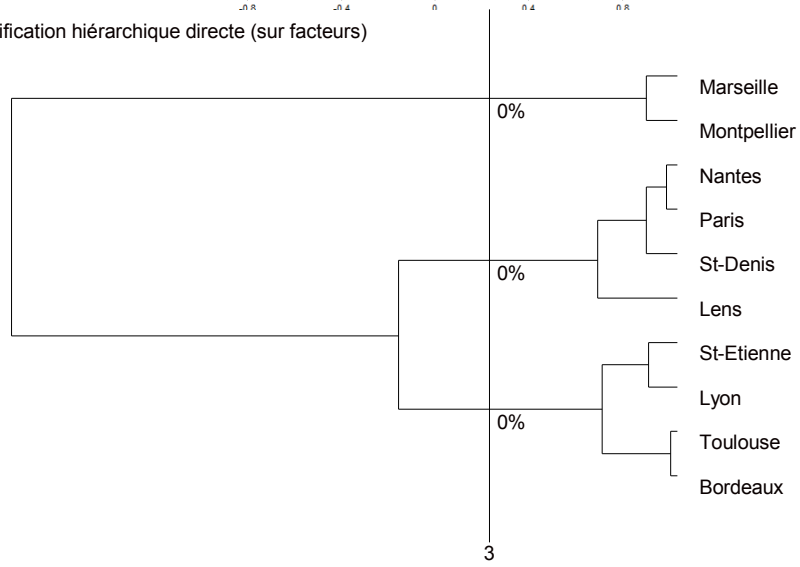
La procédure :

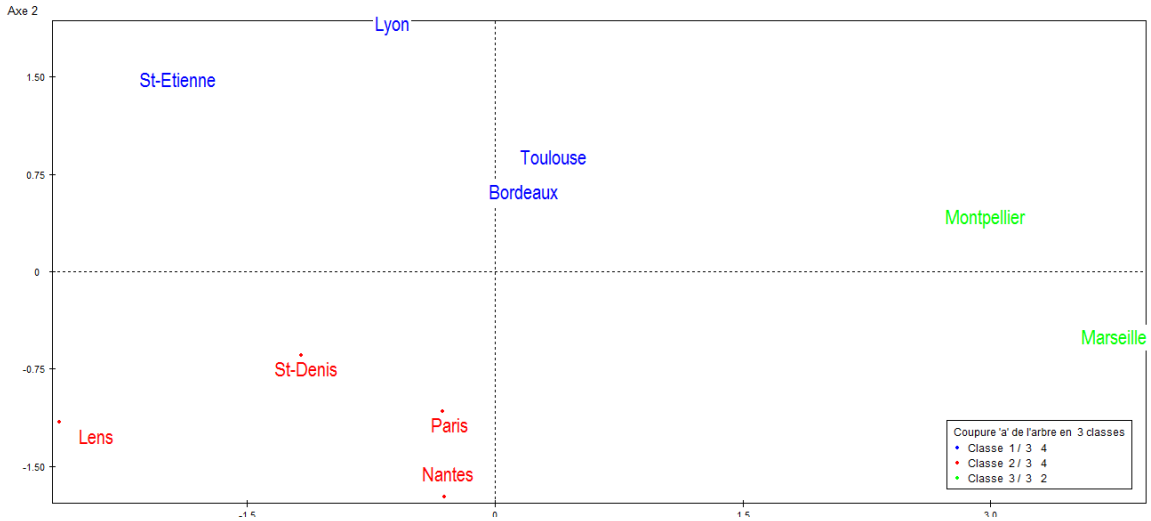


Tous les indicateurs en « actif »



Classification hiérarchique directe (sur facteurs)





Sorties du logiciel

Classe 1 / 3
Libellé de l'individu
Bordeaux
Lyon
St-Etienne
Toulouse

Classe 2 / 3
Libellé de l'individu
Lens
Nantes
Paris
St-Denis

Classe 3 / 3
Libellé de l'individu
Marseille
Montpellier

Classe 1 / 3 (Poids = 4.00 Effectif = 4)		
Variables caractéristiques	**Moyenne dans la classe**	**Moyenne générale**
Norages	**6**	**4**
Pluiemm	**73**	**59**
Temp#	**25**	**24**
Mtemp	**19**	**19**
Tempb	14	14

Classe 2 / 3 (Poids = 4.00 Effectif = 4)		
Variables caractéristiques	**Moyenne dans la classe**	**Moyenne générale**
Pluiemm	**60**	**59**
Norages	3	4
Tempb	13	14
Mtemp	18	19
Temp#	23	24

Classe 3 / 3 (Poids = 2.00 Effectif = 2)		
Variables caractéristiques	**Moyenne dans la classe**	**Moyenne générale**
Tempb	**17**	**14**
Mtemp	**22**	**19**
Temp#	**27**	**24**
Norages	3	4
Pluiemm	29	59

Une présentation des sorties :

Clusters et indicateurs les plus clivants :

{ Bordeaux, Lyon, Saint-Etienne, Toulouse}	Moyenne dans le cluster	Moyenne générale
Norages	**6**	**4**
Pluiemm	**73**	**59**
Temp#	**25**	**24**

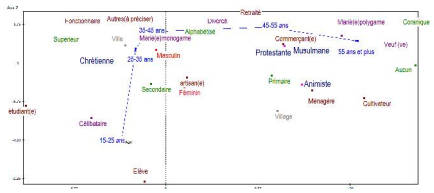
{Lens, Nantes, Paris, Saint- Denis }	Moyenne dans le cluster	Moyenne générale
Pluiemm	**60**	**59**
Norages	3	4
Tempb	13	14
Mtemp	18	19
Temp#	23	24

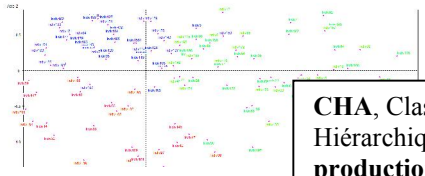
{ Marseille, Montpellie r}	Moyenne dans le cluster	Moyenne générale
Tempb	**17**	**14**
Mtemp	**22**	**19**
Temp#	**27**	**24**
Norages	3	4

une stratégie d'investigation : AFCM puis CHA

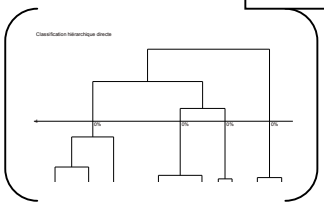
variables qualitatives (« actives » actives	*variables qualitatives »illustratives »*

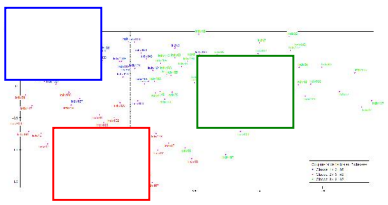
AFCM (Analyse Factorielle des Correspondances Multiples)





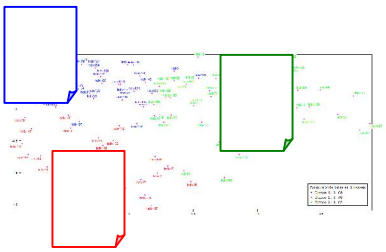
CHA, Classification Hiérarchique Ascendante, **la production des clusters**





Les % des modalités illustratives sont ensuite calculés par cluster, Des modalités sont éventuellement <u>clivées</u> par les clusters, lorsque leur % moyen dans un cluster est très différent du % dans l'ensemble des individus

Les modalités <u>clivantes</u>, elles sont sur- ou sous-représentées dans un cluster



Interpréter les axes ?bi-clustering.

Jean-Paul Villette 25 octobre 2011

En Statistique Exploratoire, une partie importante du travail consiste à reconditionner les données, éventuellement complexes (textes, images, copies d'écran de site web…) en un tableau avec des objets lignes, des objets colonnes, et, dans les cases, une mesure du couple (objet-ligne, objet-colonne). Le bi-clustering est alors un ensemble de procédures de traitement très efficaces.

Dans le cas le plus simple, les objet-lignes sont des individus, les objets-colonnes des variables numériques, et les mesures les valeurs prises par les variables sur les individus.

Les traitements en Statistique Exploratoire consiste à effectuer un bi-clustering, agrégation/différenciation des objets-lignes et , de manière duale, une agrégation/différenciation des objets-colonnes d'autre part.

La dynamo et le moteur électrique.

C'est la même machine, réversible. Avec du mouvement on produit de l'électricité, ou l'inverse, avec les deux. On peut produire de l'électricité avec un moteur électrique, mais la dynamo est plus efficace.

Je continue avec l'exemple de du tableau individus-variables numériques. Des traitements usuels sont les Analyses en Composantes Principales (ACP) et les ACP suivies d'une Classification Hiérarchique Ascendante sur les individus mesurés sur les premiers axes principaux.

1- bi-clustering orienté colonnes : la production d'indicateurs synthétiques

Les axes principaux d'une ACP sont les vecteurs propres de la matrice des corrélations (ou des covariances) et donc des combinaisons linéaires des variables-colonnes, avec des propriétés intéressantes. Ainsi l'indicateur de « vulnérabilité financière[27] »

$$V(E) = -0{,}45\frac{CPR + DLMT}{I + AVI} - 0{,}42\frac{CPR}{CPR + DLMT} - 0{,}42\frac{R + D}{DCT} - 0{,}48\frac{CPR}{CPR + DLMT + DCT - (R + D)} + 0{,}$$

réalise une synthèse, optimale en un certain sens, des variables fortement corrélées à l'axe.

Les individu-entreprises sont accessoires.

[27] Jean-Paul Villette, « *Contribution à l'étude de la vulnérabilité financière. Significations et mesures* » Thèse. Paris IX- 1985

2- **bi-clustering orienté lignes : la production de clusters d'objet-lignes, d'individus.**

On a en vue la production de clusters intra-homogènes et extra-hétérogènes[28]. Une méthode ingénieuse et très pratique, consiste à mesurer les individus sur les premières composantes de l'ACP normée sur les variables. Combien d'axes (synonymes de composantes) ? Il y a un algorithme dans le logiciel qui le détermine. L'idée est de conserver la structure des variables (les premiers axes) et d'éliminer le bruit de fond (les derniers).

L'aspect pratique est dans la procédure : il n'y a pas à faire les réglages de la Classification Hiérarchique, qui sont redoutables[29].Il est aussi dans des résultats qui sont probants, ou pas.

Sont produits des sous-ensembles d'individus (les **clusters**) et apparaissent des **variables clivantes** : celles qui ont des moyennes par cluster très différentes de la moyenne de l'ensemble des individus.

L'interprétation des axes n'est pas nécessaire **(ce sont les agrégations/différenciations des variables clivantes qu'il faut interpréter) ,** pas appropriée, et facilement contre-productive. Quand on part dans une interprétation : le premier axe oppose ceci à cela, le deuxième…et ainsi de suite, on génère implicitement 2, puis 4, puis 6 etc… clusters. Et si le nombre de clusters est en fait impair on se plante.

question existentielle : existe-t-il des clusters d'objets intra-homogènes extra-hétérogène ou bien s'agit-il d'un continuum d'objets ?

C'est la question que l'on peut nous poser et à laquelle

1- une Classification Hiérarchique Ascendante[30] ne répond pas ! une CHA donne les deux meilleurs clusters, les trois meilleurs…mais n'assure pas qu'ils sont bons.
2- répond un peu quand même. Il y a des niveaux de coupure statistiquement plus pertinents que d'autres (il faut couper les branches longues), on est alors renvoyé au choix de la mesure de cette longueur…

   Pratiquement, le logiciel SPAD, qui est bien fait, propose des niveaux de coupure de l'arbre statistiquement plus valides que d'autres. C'est à l'usager de retenir la typologie la plus utile.

---

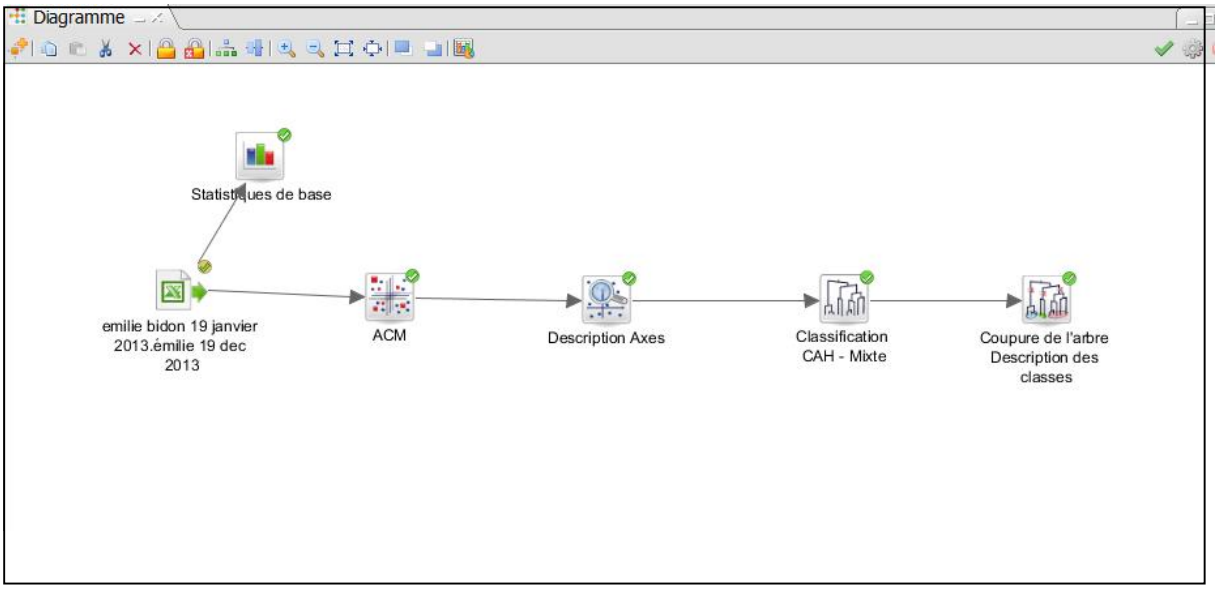
[28] *internal cohesion and external isolation of clusters*
[29] la CHA, c'était, rien que pour cette méthode, un cours de 20 heures en Master de Statistique, par Photis Nobelis.
[30] la classification est construite de manière *ascendante* (on part des objets que l'on agrège) mais est utilisée en *descendant* de l'ensemble à des sous-ensembles de plus en plus petits

Exercices avec SPAD

Entrer ce tableau dans Excel

	GENRE	AGE	COULEUR	TABAC
Pierre	homme	A3	claire	nonfumeur
Céline	femme	A2	claire	fumeur
Armelle	femme	A2	claire	fumeur
Paul	homme	A2	foncée	nonfumeur
Bernard	homme	A1	claire	fumeur
Jacques	homme	A2	claire	fumeur
Hervé	homme	A1	foncée	fumeur
Marc	homme	A3	foncée	fumeur
Nelly	femme	A1	claire	nonfumeur
Patricia	femme	A2	foncée	fumeur
Christophe	homme	A1	foncée	nonfumeur
Daniel	homme	A3	foncée	nonfumeur

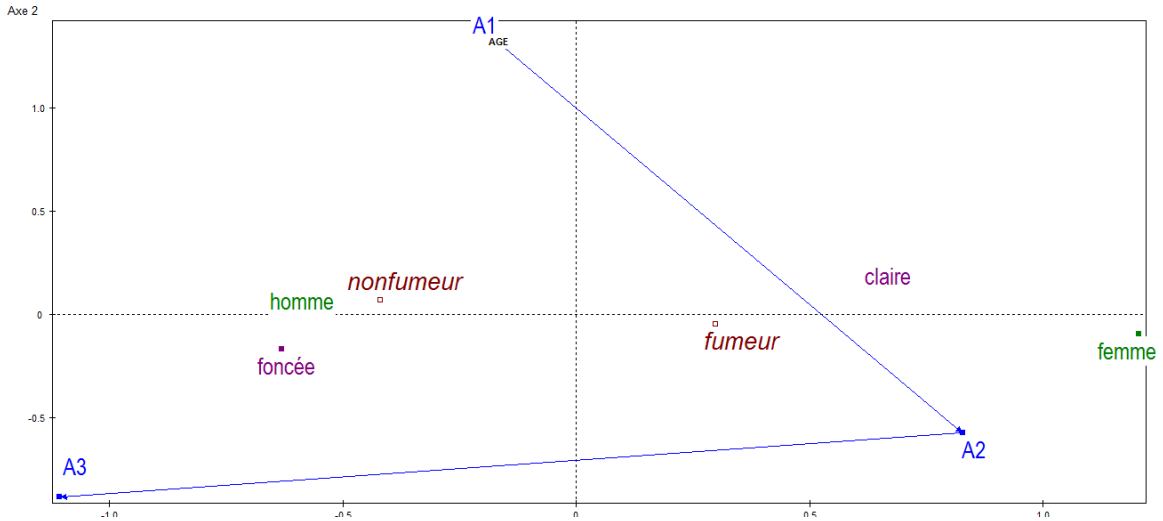
Dans SPAD

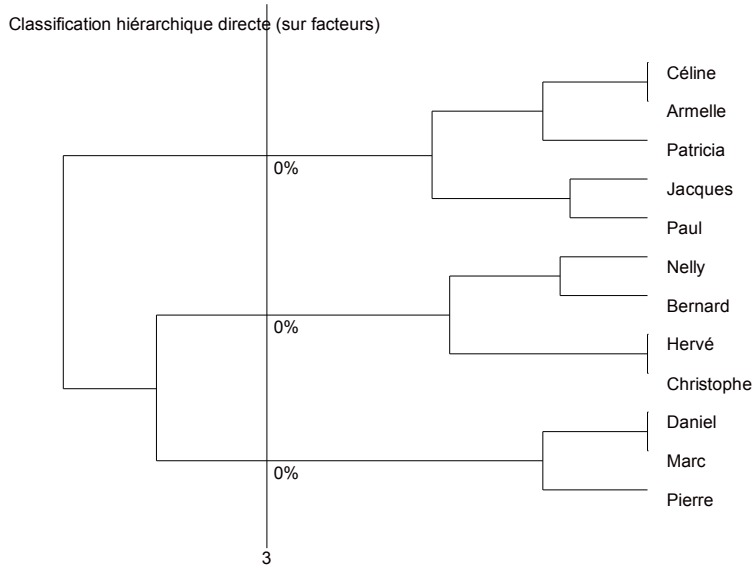


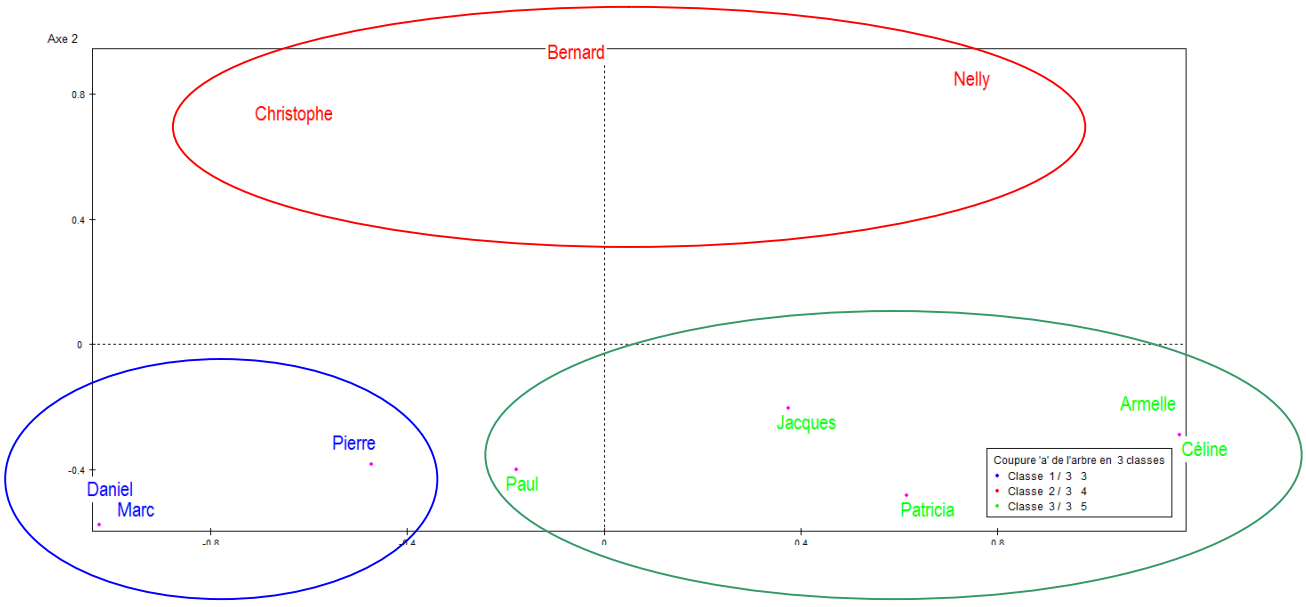
GENRE, AGE, COULEUR en variables nominales « active »
TABAC[31] en variable nominale « illustrative »

[31] Les variables et leurs modalités « illustratives » seront en *italique*

Graphe des modalités







Cluster « bleu », 3 individus (25%)	Modalités caractéristiques	% de la modalité dans la classe	% de la modalité dans l'échantillon	Poids
AGE[32]	**A3**	**100**	**25**	**3**
GENRE	**homme**	**100**	**67**	**8**
TABAC	***nonfumeur***	***67***	***42***	***5***
TABAC	*fumeur*	*33*	*58*	*7*
AGE	A1	0	33	4
GENRE	femme	0	33	4
AGE	A2	0	42	5

Cluster « rouge », 4 ind. (33%)	Modalités caractéristiques	% de la modalité dans la classe	% de la modalité dans l'échantillon	Poids
AGE	**A1**	**100**	**33**	**4**
COULEUR	**foncée**	**50**	**50**	**6**
COULEUR	**claire**	**50**	**50**	**6**
GENRE	**homme**	**75**	**67**	**8**
TABAC	***nonfumeur***	***50***	***42***	***5***
TABAC	*fumeur*	*50*	*58*	*7*
GENRE	femme	25	33	4
AGE	A3	0	25	3
AGE	A2	0	42	5

cluster « vert », 5 ind. 42%	Modalités caractéristiques	% de la modalité dans la classe	% de la modalité dans l'échantillon	Poids
AGE	**A2**	**100**	**42**	**5**
GENRE	**femme**	**60**	**33**	**4**
TABAC	***fumeur***	***80***	***58***	*7*
TABAC	*nonfumeur*	*20*	*42*	*5*
AGE	A3	0	25	3
GENRE	homme	40	67	8
AGE	A1	0	33	4

[32] Ecrire les modalités **sur-représentées** en gras,

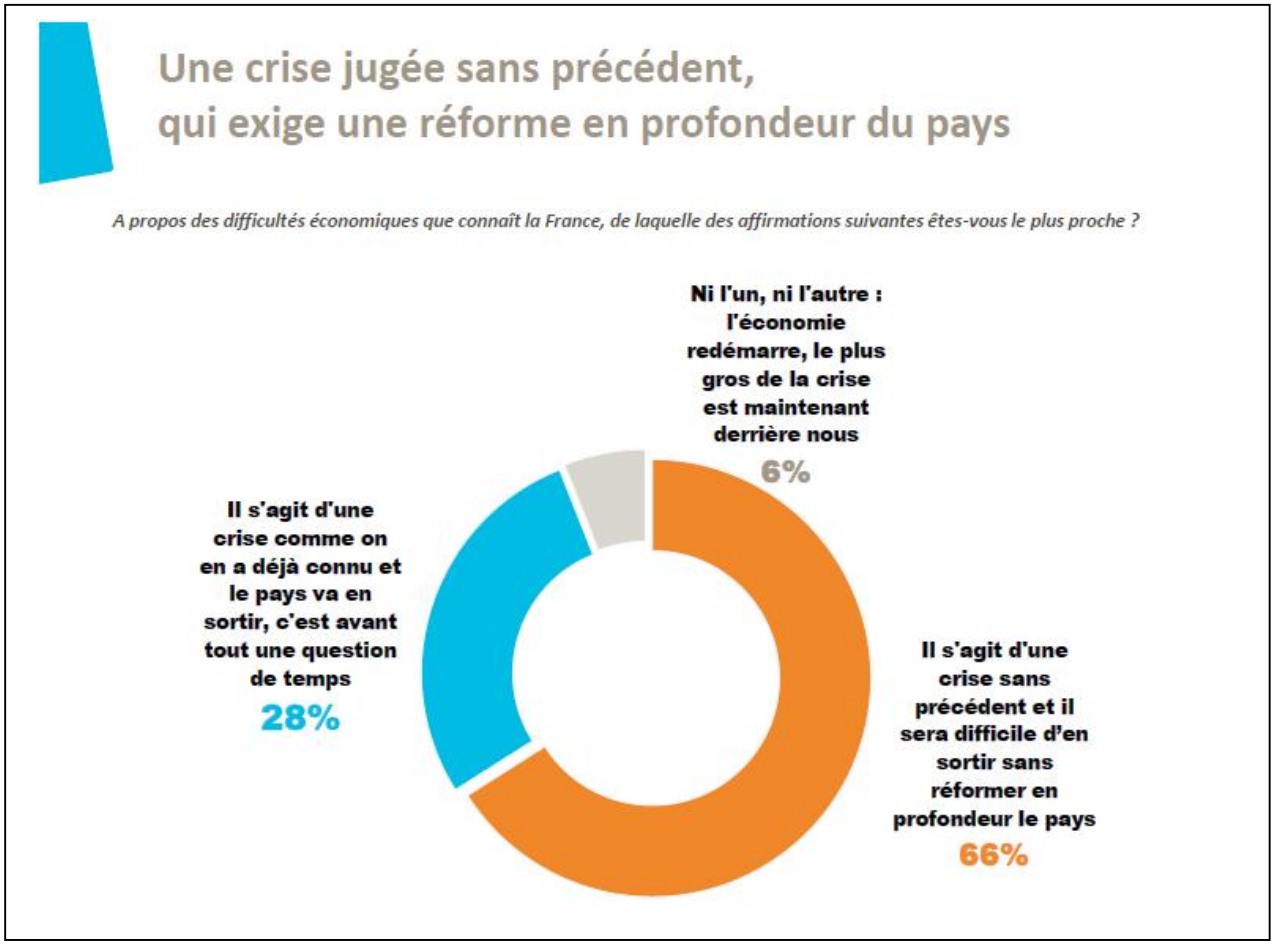
<u>Un exemple de présentation « commerciale » de cette procédure :</u>

Des extraits de

« Français, ce qui vous rassemble est-il plus fort que ce qui vous divise ? »

Les ateliers du CSA- Le Monde nov 2013

1- Classer , Compter, Comparer , Repésentation graphique



2 – AFCM et CHA

Les 5 France

« La France du collectif » — 22%
Elle croit au modèle social multi-culturaliste

« La France libérale traditionnelle » — 25%
Attachée à l'effort et au mérite individuel

«La France amère » — 31%
Elle se sent lésée et craint le déclassement social

« La France en parallèle » — 15%
Elle compte avant tout sur elle-même et rejette le modèle

« La France absente » — 7%
Elle n'y croit plus et n'attend rien

« La France libérale traditionnelle » 25%
Attachée à l'effort et au mérite individuel
Le portrait

- MAJORITAIREMENT DES HOMMES ÂGÉS DE PLUS DE 50 ANS
- RETRAITÉS DOTÉS D'UN CAPITAL CULTUREL PLUTÔT ÉLEVÉ
- NIVEAU DE VIE MOYEN SUPÉRIEUR À AISÉ
- PROPRIÉTAIRES DE LEUR LOGEMENT
- CATHOLIQUES : PRATIQUANTS RÉGULIERS OU OCCASIONNELS

POSTURE « LOYALTY » : VOTENT MASSIVEMENT À LA PRÉSIDENTIELLE ET EXPRIMENT UNE PRÉFÉRENCE PARTISANE POUR DES PARTIS DE GOUVERNEMENT (ESSENTIELLEMENT DE DROITE)

ESTIMENT QUE LES POLITIQUES ONT LES MOYENS D'AGIR MÊME S'ILS NE PEUVENT INFLUER QU'À LA MARGE SUR LE COURS DES CHOSES

•

Etude typologique des degrés d'attachement à la commune, au département, à la région, à la France,et à l'Europe déclarés par 715 habitants d'Alsace en 2004 à la SOFRES.

les données

715 répondants, habitant l'Alsace
fichier d'origine33 :

KIDEN	NUM	AGGLO	DEP	COM	ANNEE	Numero OIP	CATCOMM	nom de la r
Case n° 1	25	Category n° 3	Category n° 68	348	Category n° 4	Alsace	Category n° 2	reponse exa
Case n° 2	32	Category n° 1	Category n° 68	12	Category n° 4	Alsace	Category n° 1	reponse exa
Case n° 3	44	Category n° 4	Category n° 67	482	Category n° 4	Alsace	Category n° 5	reponse ine
Case n° 4	46	Category n° 4	Category n° 67	482	Category n° 4	Alsace	Category n° 5	reponse exa
Case n° 5	48	Category n° 1	Category n° 67	473	Category n° 4	Alsace	Category n° 1	reponse exa
Case n° 6	53	Category n° 4	Category n° 68	278	Category n° 4	Alsace	Category n° 3	reponse exa
Case n° 7	60	Category n° 2	Category n° 68	43	Category n° 4	Alsace	Category n° 2	reponse ine
Case n° 8	61	Category n° 3	Category n° 67	345	Category n° 4	Alsace	Category n° 2	reponse ine
Case n° 9	67	Category n° 4	Category n° 68	224	Category n° 4	Alsace	Category n° 5	reponse exa
Case n° 10	69	Category n° 3	Category n° 67	46	Category n° 4	Alsace	Category n° 3	reponse exa
Case n° 11	75	Category n° 1	Category n° 68	28	Category n° 4	Alsace	Category n° 1	reponse ine
Case n° 12	77	Category n° 2	Category n° 67	131	Category n° 4	Alsace	Category n° 2	reponse exa
……	……	……	……	……	……	……	……	……

715 répondants, habitant l'Alsace

[33] Source du fichier : CDSP, Centre de Données Socio-politiques, Sciences-Po Paris, 2004

typologies des degrés d'attachement à la commune, au département, à la région (Alsace), à la France et à l'Europe. Une Analyse Factorielle suivie d'une Classification hiérarchique ascendante .

variables « actives » susceptibles d'être clivantes dans la détermination de « clusters » d'individus : les questions concernant l'attachement. Des individus qui ont à peu près les mêmes réponses à ces questions seront agrégés dans un même cluster.
Variables « illustratives» susceptibles d'être clivées, différenciées par les clusters : toutes les autrès questions.
Tris à plat (715 individus)

attachement Europe	Effectif	
très attaché	146	20%
plutôt attaché	324	45%
pas très attaché	173	24%
pas attaché du tout	70	10%
NSP	2	0%

attachement France	Effectif	
très attaché	383	54%
plutôt attaché	264	37%
pas très attaché	46	6%
pas attaché du tout	22	3%

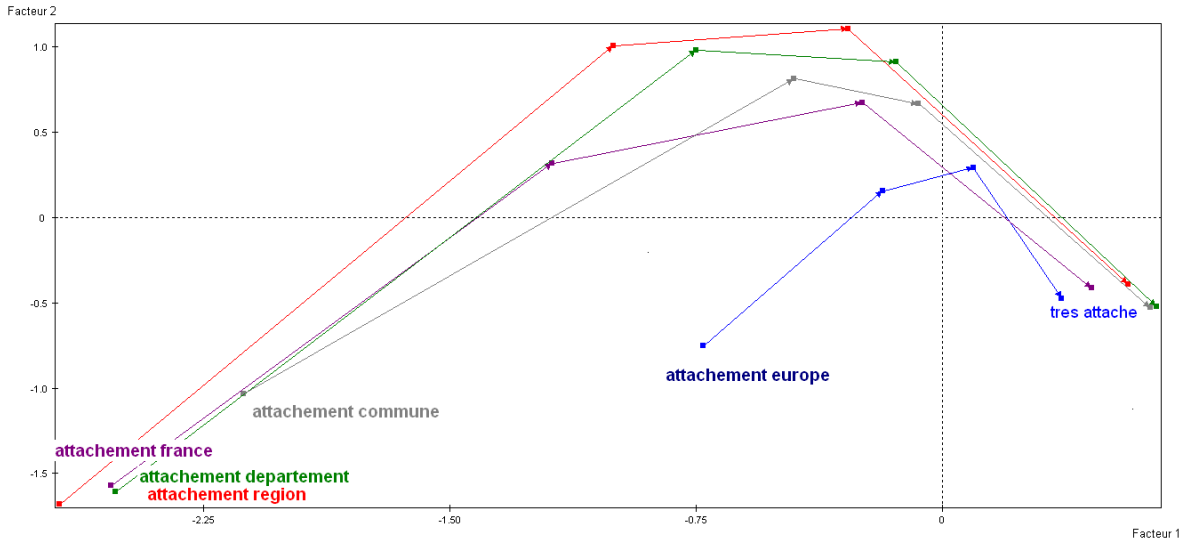
attachement région	Effectif	
très attaché	431	60%
plutôt attaché	174	24%
pas très attaché	60	8%
pas attaché du tout	50	7%

attachement département	Effectif	
très attaché	358	50%
plutôt attaché	220	31%
pas très attaché	80	11%
pas attaché du tout	57	8%

attachement commune	Effectif	
très attaché	316	44%
plutôt attaché	237	33%
pas très attaché	95	13%
pas attaché du tout	66	9%
NSP	1	0%

C'est l'attachement à la France qui est le plus fort (91% de plutôt ou très attaché)

AFCM (les variables sont ordinales) : un graphique d'une grande clarté



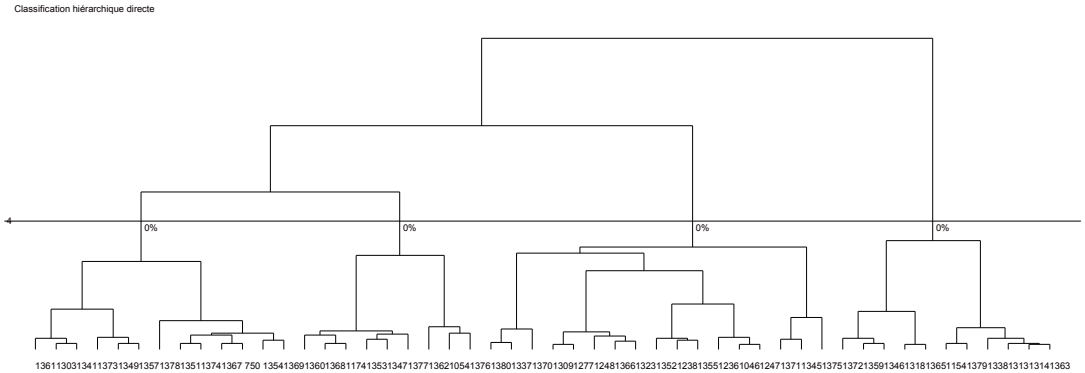
Le graphique est d'une grande clarté : plus le degré d'attachement à une collectivité est élevé, plus les degrés élevés aux autres collectivités sont fréquents. C'est nettement moins vrai pour l'attachement à l'Europe.

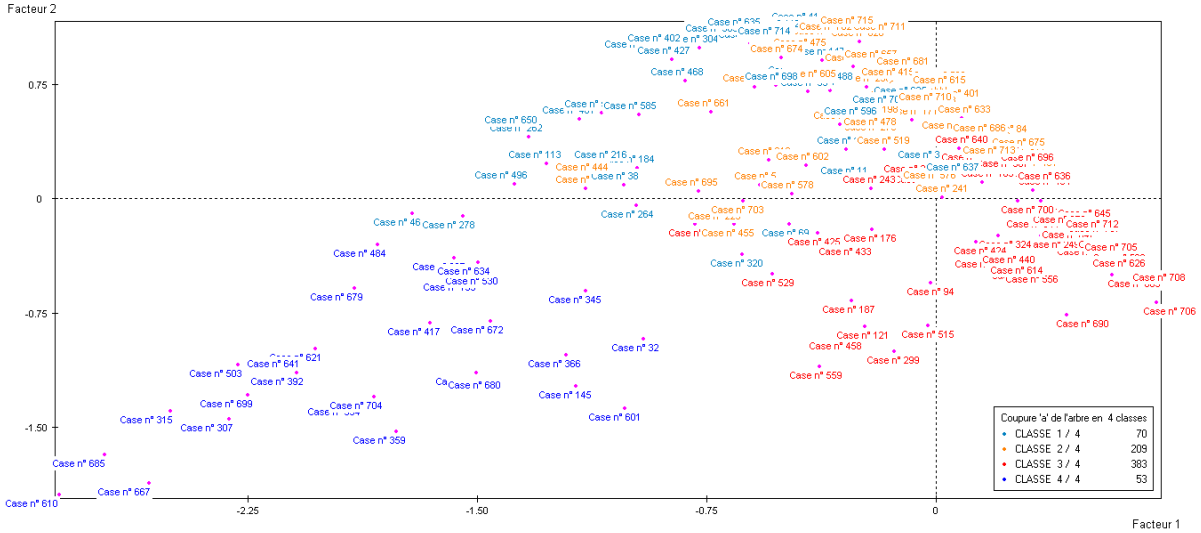
En particuliers, ceux qui sont « pas du tout « ou « pas très « attachés au département :

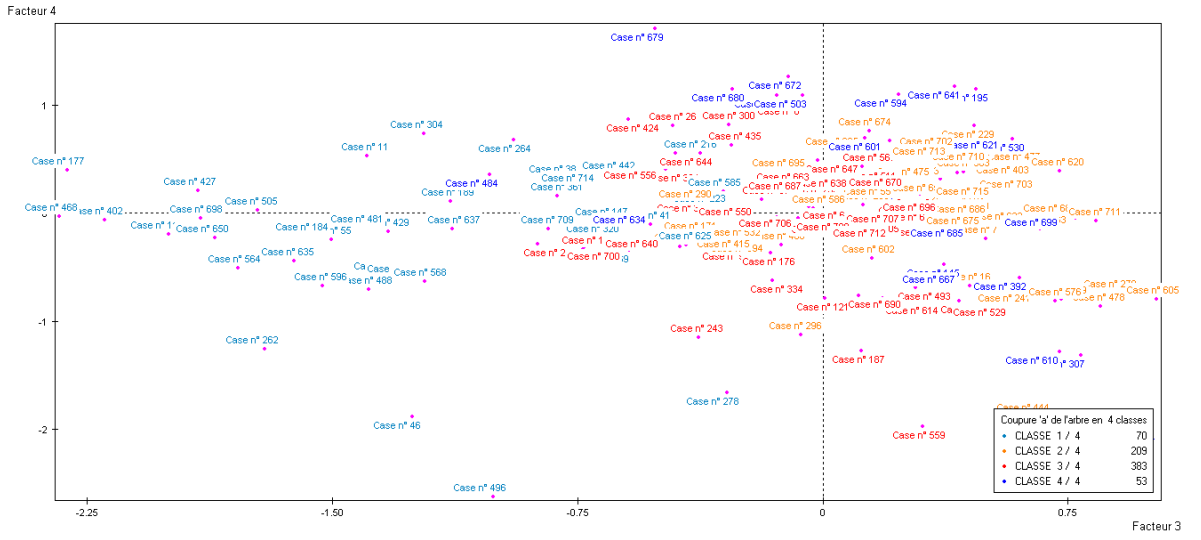
attachement région	Effectif	% / Total
très attaché	15	11
plutôt attaché	21	15
pas très attaché	52	38
pas attaché du tout	49	36
Total	137	100

ne sont pas attachés à la région non plus, très majoritairement (à 74%)

Spad suggère 4 « clusters » d'individus intra-homogènes extra-hétérogènes (des individus d'un même cluster ont à peu près les mêmes degrés d'attachements.







Quatre clusters d'individus : les « pas du tout », les « pas », les « plutôt » et les « très »

Les « pas du tout » 53 individus (7%)	Modalités caractéristiques	% dans le cluster	% dans l'échantillon	nombre de réponses
attachement département	pas attaché du tout	96	8	57
attachement région	pas attaché du tout	87	7	50
attachement commune	pas attaché du tout	74	9	66
attachement France	pas attaché du tout	25	3	22
attachement Europe	pas attaché du tout	30	10	70
attachement France	pas très attaché	19	6	46

Les « pas », 70 individus (10%)	Modalités caractéristiques	% dans le cluster	% dans l'échantillon	nombre de réponses
attachement département	pas très attaché	91	11	80
attachement région	pas très attaché	79	8	60
attachement commune	pas très attaché	46	13	95
attachement France	pas très attaché	24	6	46

Les « plutôt », 209 individus (29%)	Modalités caractéristiques	% dans le cluster	% dans l'échantillon	nombre de réponses
attachement département	plutot attaché	89	31	220
attachement région	plutot attaché	77	24	174
attachement France	plutot attaché	67	37	264
attachement commune	plutot attaché	60	33	237
attachement Europe	plutot attaché	57	45	324

Les « très », 383 individus (54%)	Modalités caractéristiques	% dans le cluster	% dans l'échantillon	nombre de réponses
attachement département	très attaché	91	50	358
attachement région	très attaché	98	60	431
attachement commune	très attaché	71	44	316
attachement France	très attaché	76	54	383
attachement Europe	très attaché	26	20	146

La question « attachement au département » est la plus clivante.

Caractéristiques et réponses sur-représentées dans les clusters :

les "pas du tout", 53 individus (7%)	Modalités caractéristiques	% de la modalité dans la classe	% de la modalité dans l'échantillon	nombre de réponses
attachement département	pas attaché du tout	96	8	57
attachement région	pas attaché du tout	87	7	50
attachement commune	pas attaché du tout	74	9	66
le sentiment de citoyenneté régionale	jamais	49	9	65
attachement France	pas attaché du tout	25	3	22
religion	*sans religion*	*40*	*10*	*69*
pratique religieuse	**Réponse manquante**	*40*	*10*	*71*
agréable a habiter	*s applique mal*	*30*	*8*	*54*
attachement Europe	pas attaché du tout	30	10	70
inscription sur les listes électorales	*non*	*30*	*12*	*84*
activité de la personne interrogée	*Etudiant*	*21*	*7*	*47*
profession de la personne interrogée	**Réponse manquante**	*21*	*7*	*47*
niveau diplôme	**Réponse manquante**	*21*	*7*	*47*
inégalités de développement dans la région	*oui beaucoup*	*38*	*19*	*135*
identification des politiques conduites par la gauche et la	*pas différentes du t*	*23*	*8*	*60*
le sentiment de citoyenneté régionale	*pas très souvent*	*28*	*12*	*89*
attachement France	*pas très attaché*	*19*	*6*	*46*
le sentiment de citoyenneté française	*jamais*	*11*	*3*	*18*
plutôt préservée de la pollution	*s applique mal*	*83*	*64*	*455*
profession du chef de famille	*Cadres sup*	*11*	*3*	*21*
indépendant, salarie	**Réponse manquante**	*23*	*10*	*73*
AGE5	*18-24 ANS*	*25*	*12*	*86*
optimisme avenir Europe	*TAF pessimiste*	*19*	*8*	*59*
image du projet du CR	*mauvaise direction*	*17*	*7*	*50*

les "pas", 70 individus (10%)	Modalités caractéristiques	% de la modalité dans la classe	% de la modalité dans l'échantillon	nombre de réponses
attachement département	pas très attaché	91	11	80
attachement région	pas très attaché	79	8	60
attachement commune	pas très attaché	46	13	95
attachement France	pas très attaché	24	6	46
plutôt dynamique	*s applique mal*	*27*	*10*	*70*
le sentiment de citoyenneté régionale	*jamais*	*26*	*9*	*65*
agréable a habiter	*s applique mal*	*21*	*8*	*54*
le sentiment de citoyenneté régionale	*pas très souvent*	*26*	*12*	*89*
niveau diplôme	*Diplôme sup*	*29*	*15*	*110*
profession de la personne interrogée	*Professeurs*	*10*	*3*	*21*
image du projet du CR	**Réponse manquante**	*34*	*21*	*147*
auto positionnement sur échelle gauche droite	*à gauche*	*29*	*17*	*122*
appréciation élargissement ue pour la France	*une bonne chose*	*67*	*53*	*381*

les "plutot", 209 individus (29%)	Modalités caractéristiques	% de la modalité dans la classe	% de la modalité dans l'échantillon	nombre de réponses
attachement département	plutot attaché	89	31	220
attachement région	plutot attaché	77	24	174
attachement France	plutot attaché	67	37	264
attachement commune	plutot attaché	60	33	237
le sentiment de citoyenneté régionale	assez souvent	41	26	185
attachement Europe	plutot attaché	57	45	324
auto positionnement sur echelle gauche droite	*au centre gauche*	*16*	*10*	*75*
le sentiment de citoyenneté francaise	*assez souvent*	*44*	*36*	*254*
le sentiment de citoyenneté régionale	*pas très souvent*	*18*	*12*	*89*
optimisme avenir Région	*plutôt optimiste*	*67*	*59*	*421*
AGE5	*25-34ANS*	*26*	*20*	*143*
la participation a des reunions concernant la région	*vous n avez pas le t*	*14*	*10*	*68*
activite chef de famille	*Temps complet*	*31*	*25*	*178*

les "très", 383 individus (54%)	Modalités caractéristiques	% de la modalité dans la classe	% de la modalité dans l'échantillon	nombre de réponses
attachement département	très attaché	91	50	358
attachement région	très attaché	98	60	431
attachement commune	très attaché	71	44	316
attachement France	très attaché	76	54	383
le sentiment de citoyenneté régionale	*très souvent*	*74*	*53*	*376*
AGE5	*65 ET PLUS*	*25*	*18*	*132*
image du projet du CR	*bonne direction*	*76*	*68*	*485*
agreable a habiter	*s applique bien*	*97*	*92*	*661*
auto positionnement sur echelle gauche droite	*a droite*	*21*	*15*	*110*
activite de la personne interrogee	*Retraite*	*28*	*22*	*158*
le sentiment de citoyenneté francaise	*très souvent*	*59*	*51*	*367*
attachement Europe	très attaché	26	20	146
activite chef de famille	*Retraite*	*13*	*9*	*67*
proximite partisane detaillee	*UDF*	*18*	*14*	*97*
elections régionales de mars 2004: changement president CR	*oui*	*23*	*18*	*132*
AGE5	*50-64AN*	*25*	*20*	*141*
satisfaction de l information sur activite du CR	*plutot bien*	*57*	*51*	*365*
niveau diplome	*Primaire*	*13*	*10*	*70*
religion	*catholique*	*75*	*70*	*499*
pratique religieuse	*une fois par mois*	*21*	*17*	*123*
satisfaction de l information sur activite du CR	*très bien*	*11*	*8*	*60*
le rôle à jouer dans la vie politique de la région	*oui tout a fait*	*9*	*7*	*48*
projet du CR pour la région	*oui*	*83*	*79*	*568*
proximite partisane detaillee	*UMP*	*17*	*14*	*103*
indépendant, salarie	*Employeur*	*4*	*3*	*18*

Post-it d'un statisticien benêt ;

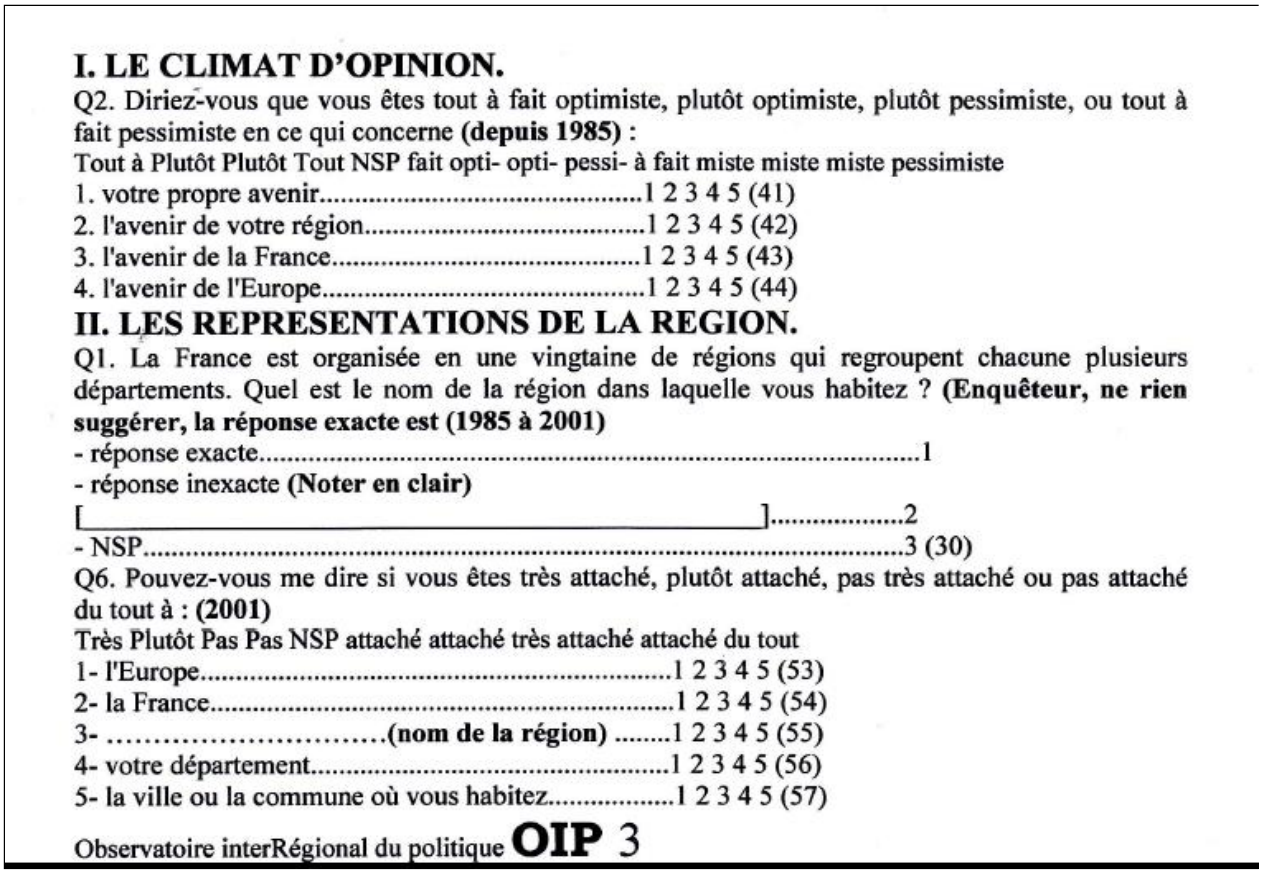
- des habitants d'Alsace « très attaché » ou « plutôt attaché « à la Région Alsace (Région $^{\#}$) et « pas très « ou » pas attachés du tout « au département (département $_{b}$) , il y en a très peu (5%).

	departement$^{\#}$	département$_{b}$	
région $^{\#}$	80%	5%	85%
région$_{b}$	1%	14%	15%
	81%	19%	100%

- des habitants région-philes, département-phobes, plutôt francophobes et europhiles , il n'y en extrêmement peu.

Annexe : Le verbatim des questions

Scan du guide d'entretien téléphonique



I. LE CLIMAT D'OPINION.

Q2. Diriez-vous que vous êtes tout à fait optimiste, plutôt optimiste, plutôt pessimiste, ou tout à fait pessimiste en ce qui concerne **(depuis 1985)** :

Tout à fait opti- / Plutôt opti- / Plutôt pessi- / Tout à fait pessi- / NSP

1. votre propre avenir.....1 2 3 4 5 (41)
2. l'avenir de votre région.....1 2 3 4 5 (42)
3. l'avenir de la France.....1 2 3 4 5 (43)
4. l'avenir de l'Europe.....1 2 3 4 5 (44)

II. LES REPRESENTATIONS DE LA REGION.

Q1. La France est organisée en une vingtaine de régions qui regroupent chacune plusieurs départements. Quel est le nom de la région dans laquelle vous habitez ? **(Enquêteur, ne rien suggérer, la réponse exacte est (1985 à 2001)**

- réponse exacte.....1
- réponse inexacte **(Noter en clair)**

[_____].....2

- NSP.....3 (30)

Q6. Pouvez-vous me dire si vous êtes très attaché, plutôt attaché, pas très attaché ou pas attaché du tout à : **(2001)**

Très attaché / Plutôt attaché / Pas très attaché / Pas attaché du tout / NSP

1- l'Europe.....1 2 3 4 5 (53)
2- la France.....1 2 3 4 5 (54)
3- .....**(nom de la région)** 1 2 3 4 5 (55)
4- votre département.....1 2 3 4 5 (56)
5- la ville ou la commune où vous habitez.....1 2 3 4 5 (57)

Observatoire interRégional du politique **OIP** 3

Remarque : 107 répondants sur 715 soit 15% ignorent qu'ils habitent dans la région Alsace, ce qui n'empêche pas d'y être très attachés, bien au contraire :

En ligne nom de la région
En colonne attachement région

Effectifs	très attaché	plutôt attaché	pas très attaché	pas attaché du tout	ENSEMBLE
réponse exacte	354	153	55	46	608
réponse inexacte	77	21	5	4	107
ENSEMBLE	431	174	60	50	715

Effectifs	très attaché	plutôt attaché	pas très attaché	pas attaché du tout	ENSEMBLE
réponse exacte	58%	25%	9%	8%	100%
réponse inexacte	72%	20%	5%	4%	100%
ENSEMBLE	60%	24%	8%	7%	100%