Analyse Factorielle des Correspondances AFC, Analyses des Correspondances Multiples ACM, analyses mixtes Analyse discriminante, classification

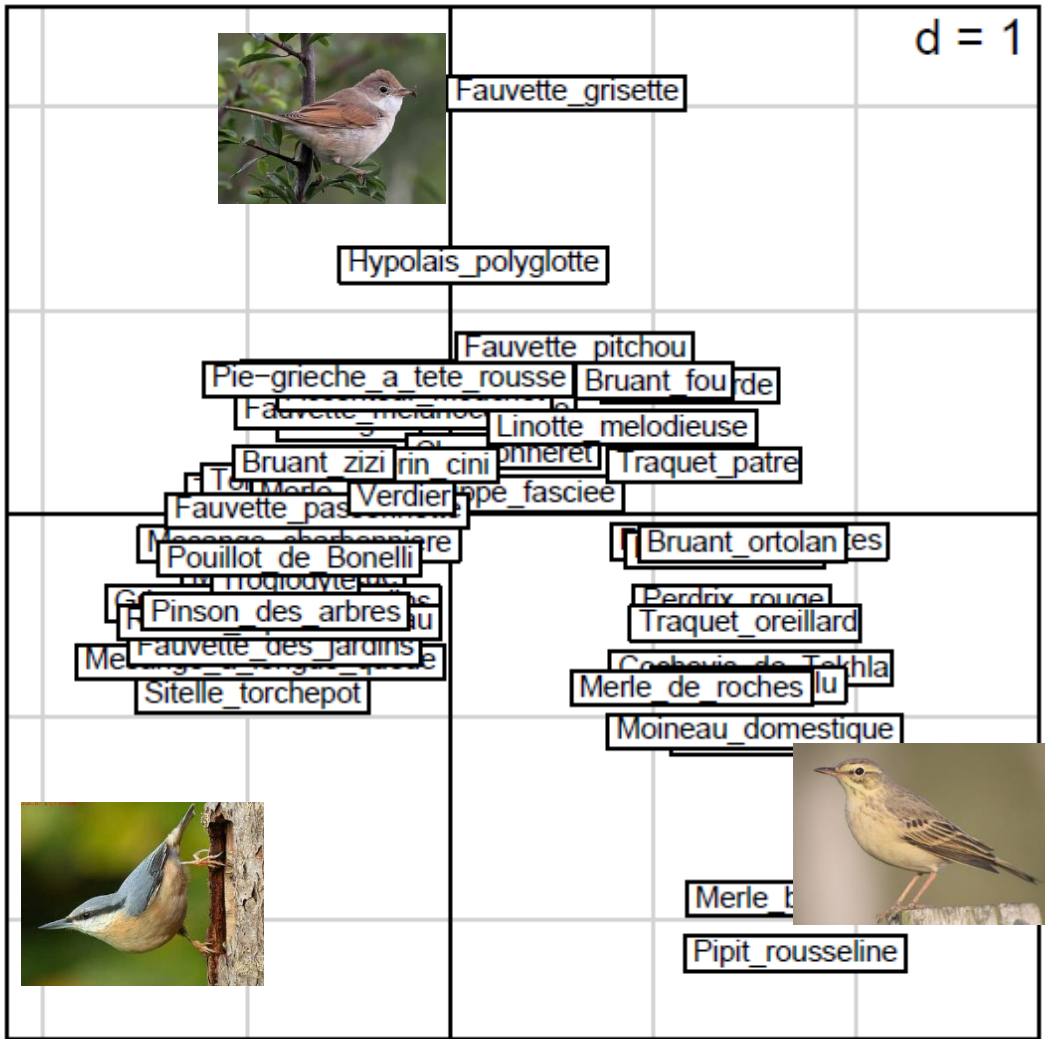
Emmanuelle Gilot-Fromont

Janvier 2024

VetAgro Sup

AFC

1. Les données
2. Questions
3. Ajustement des nuages
4. Interprétation
5. Exemples



VetAgro Sup

1. Les données

Table de contingence : effectifs obtenus en croisant deux variables qualitatives

	1		j		J
1	n_{11}				
i			n_{ij}		
I					n_{IJ}

I distributions en ligne

J distributions en colonne

Extension aux tableaux de données homogènes : notes, rangs…

VetAgro Sup

Exemple

	abattage	chronicité	guérison
arthrite	7	6	4
traumatisme	5	5	27
défaut d'aplomb	4	4	0
jarret droit	4	4	1
panaris	1	0	8
érosion des talons	1	4	13
clou de rue	0	4	5
ouverture ligne blanche	6	27	77
fourbure	0	2	20
fissure de muraille	0	0	3

242 taurillons ayant présenté une boiterie : tableau lésion / évolution :

10 distributions en lignes
3 distributions en colonnes

VetAgro Sup

Tableau des fréquences

- On considère souvent le tableau des fréquences :

$$f_{ij} = n_{ij}/n$$

- Elles permettent d'analyser les attractions / répulsions entre les modalités :

indépendance si	$f_{ij} = fi.\ f_{.j}$
attraction si	$f_{ij} > fi.\ f_{.j}$
répulsion si	$f_{ij} < fi.\ f_{.j}$

VetAgro Sup

Test d'indépendance

$$\chi_2 = \sum_{ij} \frac{(observé - théorique)^2}{théorique}$$

Théorique = $nf_{i.}\, f_{.j}$

Mesure l'écart entre le tableau observé et le tableau prédit sous l'hypothèse d'indépendance. Pour chaque case du tableau:

indépendance si $f_{ij} = f_{i.}\, f_{.j}$

attraction si $f_{ij} > f_{i.}\, f_{.j}$

répulsion si $f_{ij} < f_{i.}\, f_{.j}$

VetAgro Sup

Sous-exemple

	abattage	chronicité	guérison	
arthrite	7	6	4	17
traumatisme	5	5	27	37
	12	11	31	54

- Attraction arthrite / abattage

  $f_{ij} = 7/54 \quad > \quad fi.\ f_{.j} = 17/54 * 12/54 = 3{,}78/54$

- Répulsion chronicité / traumatisme

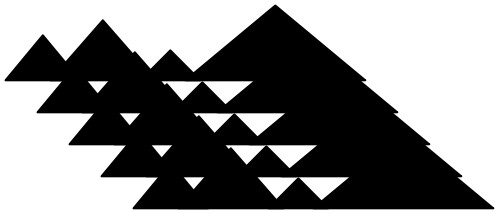
  $f_{ij} = 5/54 \quad < \quad fi.\ f_{.j} = 37/54 * 11/54 = 7{,}54/54$

VetAgro Sup

Profils-lignes et profils-colonnes

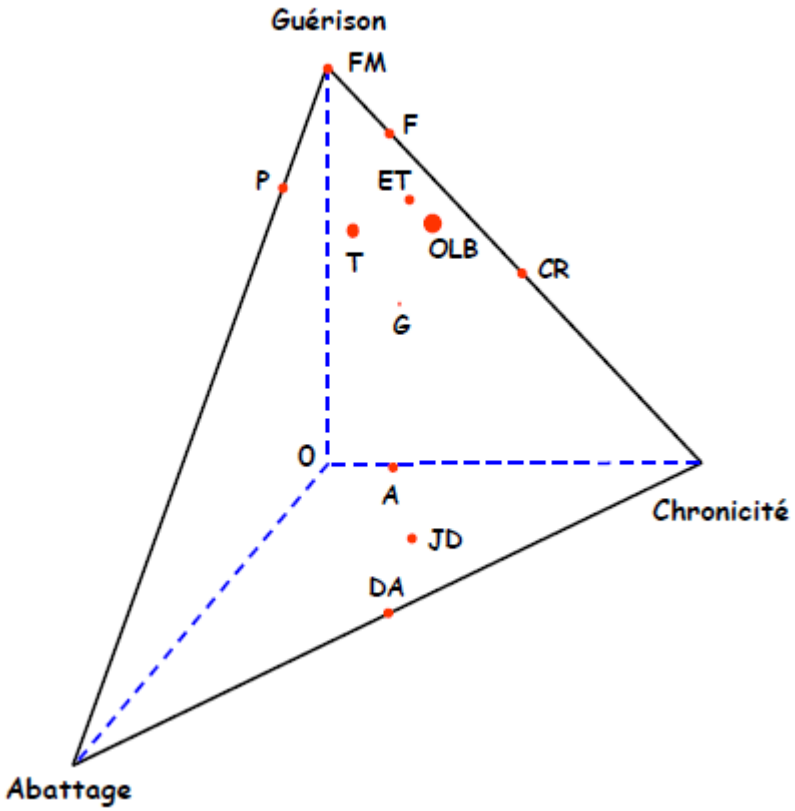
- Une ligne = une série de fréquences
  une colonne = une série de fréquences

- Indépendance = égalité entre les profils-lignes OU égalité entre les profils-colonnes



- Lignes et colonnes jouent un rôle symétrique
  La pondération n'est plus uniforme, elle vient des données

Interprétation géométrique



Une ligne i = une série de J valeurs numériques = un point dans l'espace à J dimensions, de poids $f_{i.}$

10 lignes dans l'espace à 3 dimensions (-1 = plan)

Ex: fissure de muraille : tous guéris

Nuage de profils-lignes; ressemblance = distance entre points-lignes

VetAgro Sup

Interprétation géométrique

- Une colonne j = une série de I valeurs numériques = un point dans l'espace à I dimensions, de poids $f_{.j}$

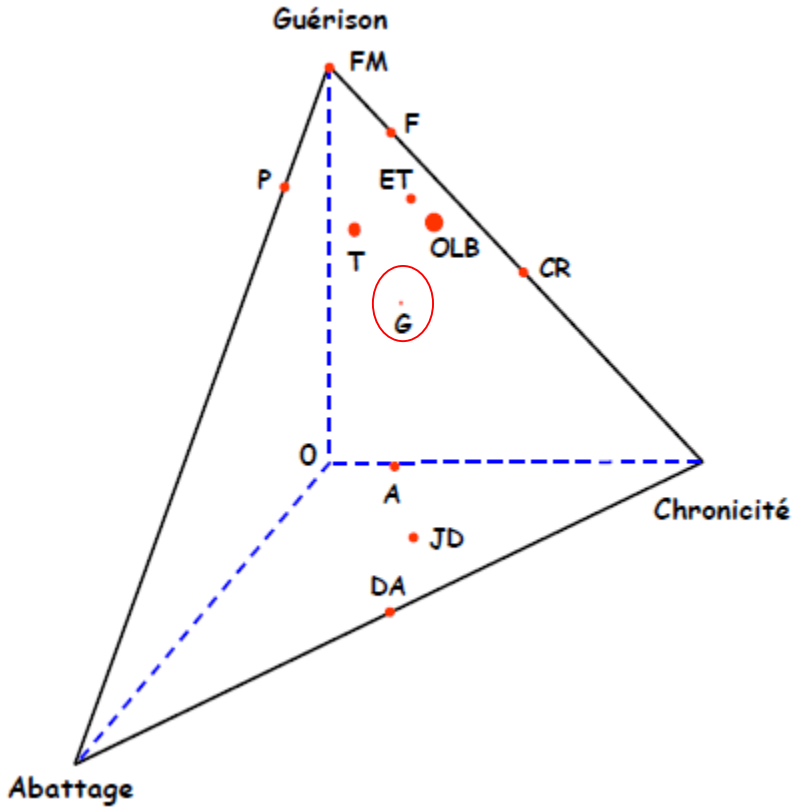
- Les deux nuages ne sont que deux représentations du même jeu de données : les analyses des lignes et des colonnes ne sont pas indépendantes

VetAgro Sup

2. Questions

- Ressemblance entre les profils-lignes : y a-t-il des ressemblances / différences entre lignes
  - y a-t-il des lésions qui ont le même profil d'évolution?
- Entre les profils colonnes : idem
  - y a-t-il des évolutions qui correspondent aux mêmes lésions?
- Correspondances :
  - à quelle lésion correspond quelle évolution?

Ressemblance



Distance entre profils-lignes =

- distance du chi-deux
- distance entre les points-lignes du nuage des lignes : propriétés euclidiennes

Barycentre du nuage G = profil moyen = somme des 10 profils-lignes

VetAgro Sup

3. Ajustement des nuages

Ajustement du nuage des lignes = chercher une suite privilégiée de directions dans le nuage de points telle que :

- la projection sur un axe factoriel conserve au mieux la forme du nuage = maximise l'inertie projetée du nuage
- chaque axe factoriel est orthogonal au(x) précédent(s)

VetAgro Sup

Ajustement du nuage des colonnes

Chercher une suite privilégiée de directions dans le nuage de points ayant les mêmes caractéristiques

Dualité:

- les inerties associées aux axes de mêmes rangs sont égales
- les facteurs de même rang sur les lignes et les colonnes sont liés par des relations de transition

Relations de transition

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_j \left(\frac{f_{ij}}{f_{i.}} G_s(j) \right)$$

$$G_s(j) = \frac{1}{\sqrt{\lambda_s}} \sum_i \left(\frac{f_{ij}}{f_{i.}} F_s(i) \right)$$

Fs(i) = projection de la ligne i sur l'axe de rang s du nuage des lignes

Gs(j) = projection de la colonne j sur l'axe de rang s du nuage des colonnes

VetAgro Sup

En pratique

- Transformation des données en fréquences
- Centrage
- Ajustement des nuages des profils-lignes : projection sur min(J-1),(I-1) axes d'inertie non nulle

4. Interprétation

- Analyse d'inertie

- Représentation graphique des projections des nuages

- Aides à l'interprétation

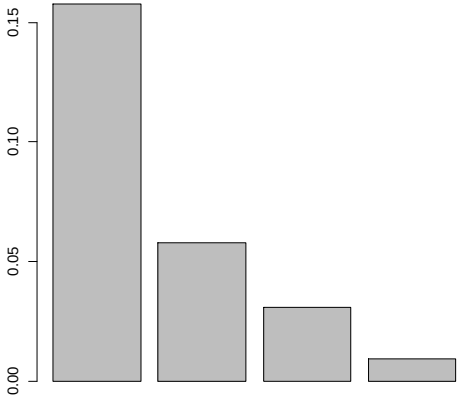
VetAgro Sup

Analyse d'inertie

Inertie totale du nuage de points :
chi2 / n = phi-deux de Pearson

Mesure l'intensité de la liaison entre les variables (chi2 = prise en compte de n = significativité)

Analyse du pourcentage d'inertie reprise par chaque axe, graphe des valeurs propres



VetAgro Sup

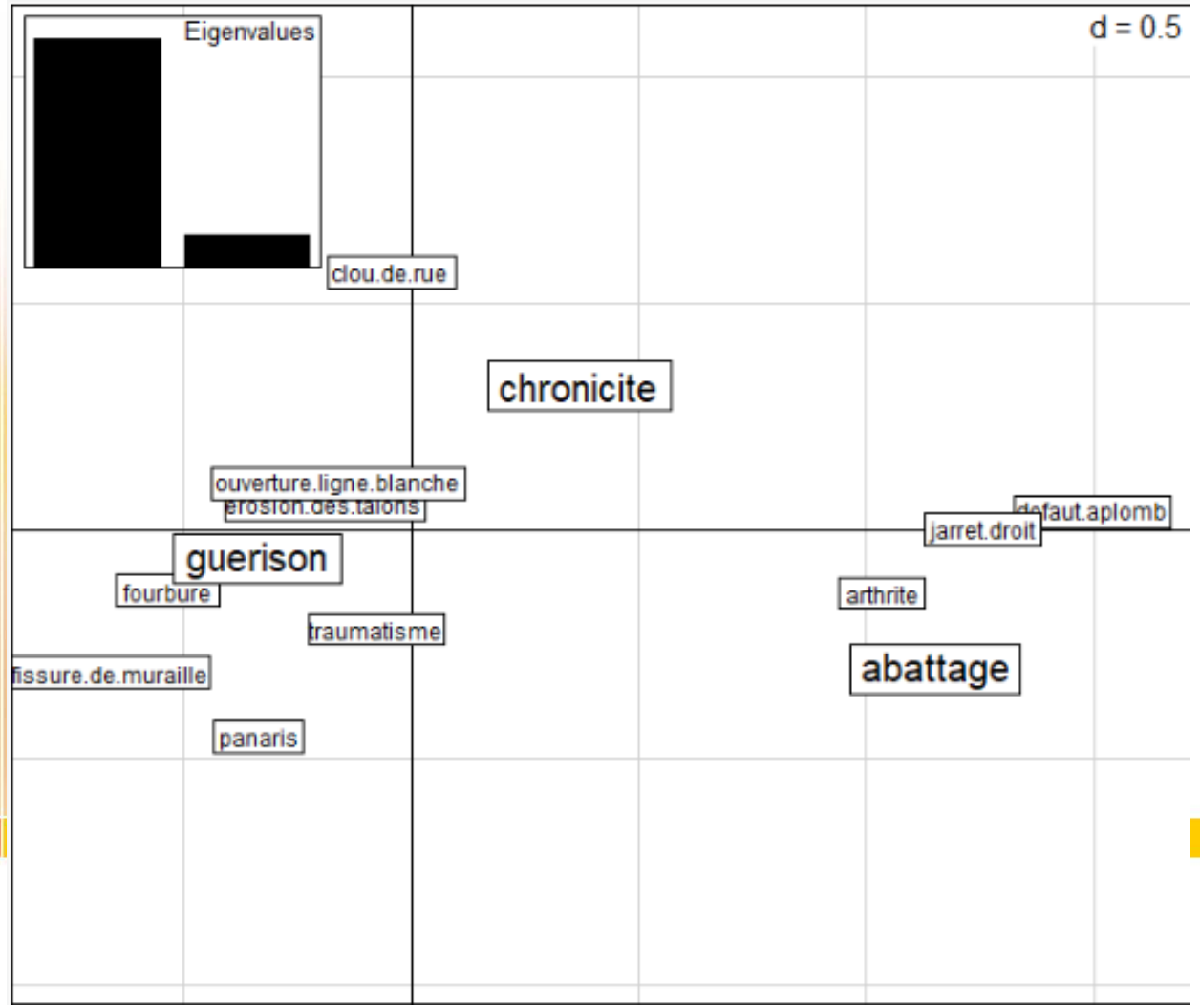
Représentation graphique

Représentation simultanée des projections des deux nuages sur les axes de même rang.

Les distances entre points-lignes s'interprètent comme des distances entre profils-lignes,
Idem pour les colonnes

A un coefficient près, le point représentant la ligne i est au barycentre des points-colonnes, et réciproquement

Représentation graphique



La projection maximise la corrélation canonique = écart à l'indépendance

VetAgro Sup

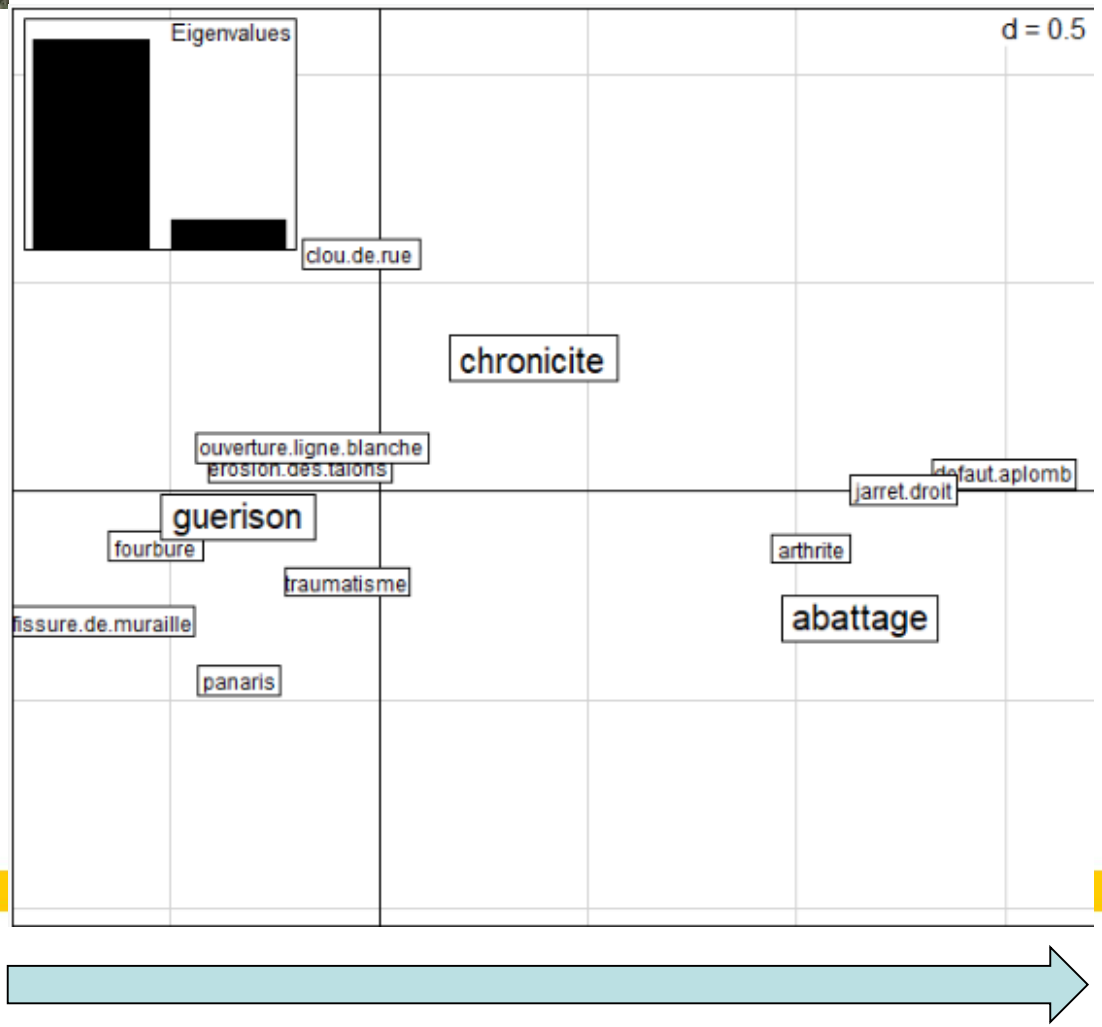
Aides à l'interprétation

Les contributions ne sont pas triviales car les poids des lignes et des colonnes ne sont plus uniformes

Mais souvent peu d'axes informatifs car peu de colonnes

VetAgro Sup

5. Exemples



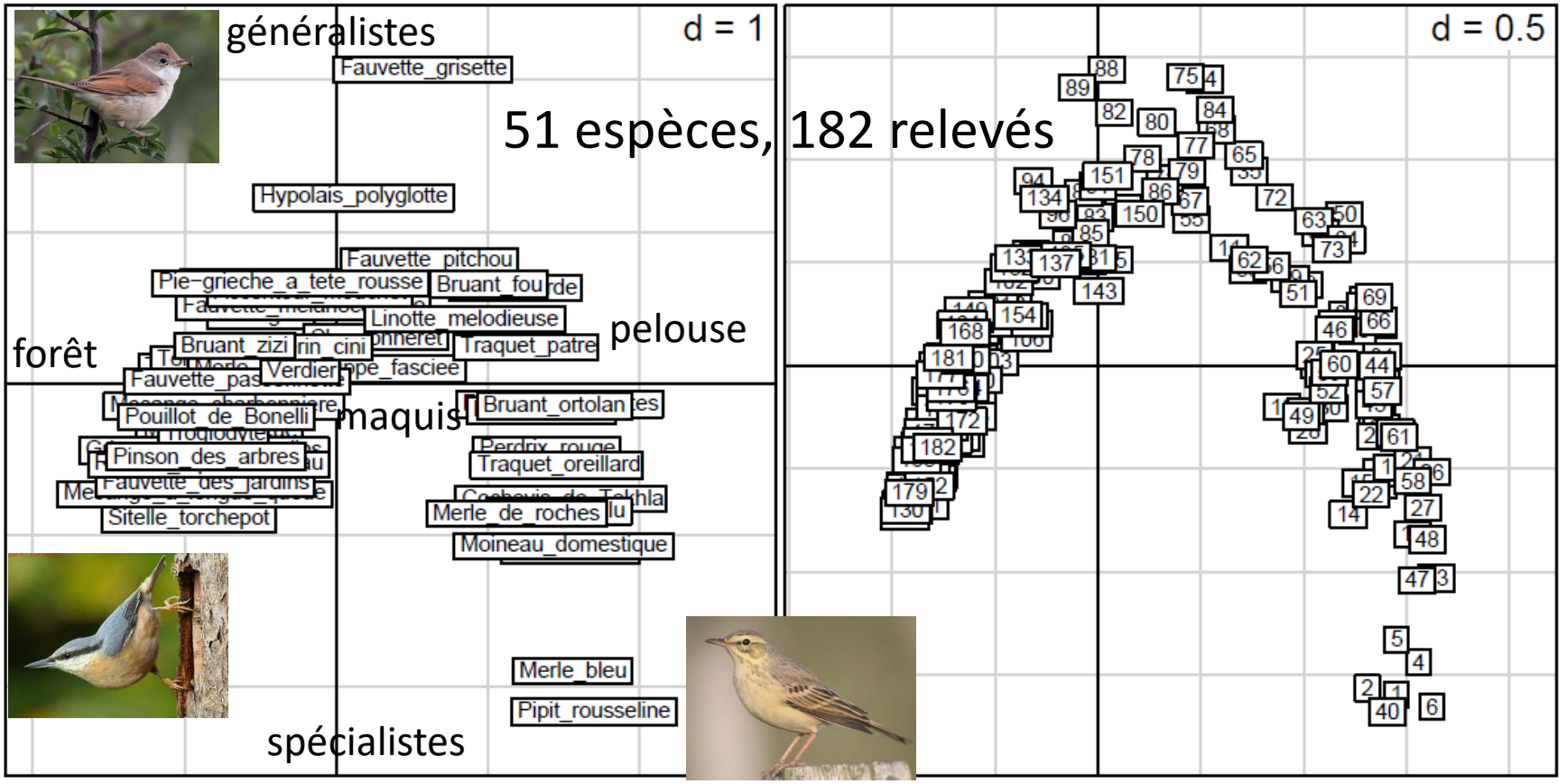
devenir de taurillons

Gradient de gravité

VetAgro Sup

Avifaune des Albères

Prodon et Lebreton 1981



Effet Guttman / fer à cheval : similarités entre les modalités du centre de l'axe 1

VetAgro Sup

Age et internet

Classe d'âge et activité principale sur internet

VetAgro Sup

ACM

1. Les données
2. Questions
3. Analyse
4. Interprétation
5. Exemples

VetAgro Sup

1. Les données

	1		j		J
1	x_{11}				
i			x_{ij}		
I					x_{IJ}

Variables qualitatives :
chaque variable comprend un petit nombre k de modalités sans signification numérique :

1 = bleu
2 = blanc
3 = rouge

VetAgro Sup

Choix des classes

- Nombre de classes :
  - si trop peu, on groupe des individus différents
  - si trop, effectifs faibles

- Choix des classes :
  - choisir des seuils ayant un sens
  - rechercher les cassures de la distribution
  - effectifs proches

- Si les classes sont bien choisies, la passage d'une variable quantitative en variable qualitative permet de détecter des effets non linéaires

Question	Réponses possibles	Poids (%)	Abréviation
Mode d'occupation	seul	48,30%	Seul
	colocataires	13,84%	Coloc
	en couple	13,05%	Couple
	avec les parents	23,50%	Parents
	non réponse	1,31%	NR1
Type d'habitation	cité universitaire	10,70%	Cité
	studio	28,20%	Studio
	appartement	30,29%	Appart
	chambre chez un particulier	5,22%	Chambre
	autre	19,84%	Autre
	non réponse	5,74%	NR2
Ancienneté hors du foyer familial	moins de 1 an	20,89%	< 1 an
	1 à 3 ans	24,80%	1-3 ans
	plus de 3 ans	28,72%	> 3 ans
	non applicable	24,80%	NA
	non réponse	0,78%	NR3
Eloignement de la fac	moins de 1 km	26,89%	< 1 km
	1 à 5 km	49,87%	1 à 5 km
	plus de 5 km	20,89%	> 5 km
	non réponse	2,35%	NR4
Superficie (m2)	moins de 10 m^2	9,14%	< 10
	10 à 20 m^2	17,75%	10 à 20
	20 à 30 m^2	24,80%	20 à 30
	plus de 30 m^2	39,16%	> 30
	non réponse	9,14%	NR5

Exemple

383 étudiants interrogés en 2001 sur leurs modalités de logement

Tableau disjonctif complet

	1		j		J
1					
i	0100		x_{ij}		001
I					

Une colonne par modalité

Une colonne = une indicatrice (présence /absence)

Une description complète des données est possible avec k-1 indicatrices par variable

VetAgro Sup

Tableau de Burt

	1	j	k	J
1				
j			TC(j,k)	
k		TC(j,k)		
J				

- Symétrique: J x J tables de contingences juxtaposées (cf AFC)
- Chaque TC représente une relation entre deux variables j et k
- Tableau équivalent à une matrice des corrélations (cf ACP)

2. Questions

- Sur les individus : cf ACP : typologie, ressemblances / différences, groupes

- Sur les variables :

  - liens entre variables (cf ACP); mais: représentation de ce lien : table de contingence (cf AFC)

  - résumer l'inertie à l'aide d'un nombre limité d'axes: information qualitative représentée par une valeur quantitative

Questions sur les modalités

- Une modalité = une colonne du tableau disjonctif complet : deux modalités se ressemblent d'autant plus qu'elles sont portées par les mêmes individus (TDC)

- Une modalité = une classe d'individus : deux modalités se ressemblent d'autant plus qu'elles sont associées aux mêmes modalités des autres variables (Burt)

3. Analyse

Sur le tableau disjonctif complet :

- Chaque individu est représenté par un ensemble de 0 et de 1 = point du nuage des individus dans l'espace des modalités
- Chaque modalité est représentée par les individus qui la portent (1) ou non (0)

=> On peut définir un nuage des individus et un nuage des modalités

Ajustement des nuages

- Recherche d'axes de projection qui conservent au maximum la forme des nuages

- Critère: maximisation du rapport de corrélation:

  η^2 = variance inter-modalités / variance totale

  Cf ANOVA

- Dualité: le nuage des individus et celui des modalités se projettent sur les mêmes axes

VetAgro Sup

Et les variables?

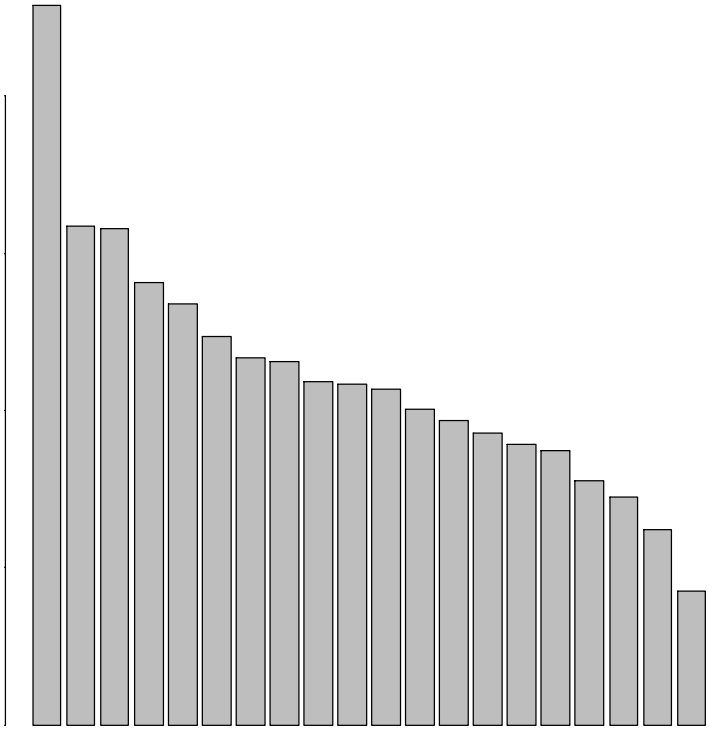
- Le tableau disjonctif complet considère les modalités comme autant de colonnes distinctes.
- Mais de par la redondance des indicatrices d'une même variable, le centre de gravité des modalités d'une même variable est au centre du nuage.
- Si au lieu d'analyser le tableau disjonctif complet on analyse le tableau de Burt = des groupes d'individus, on obtient le même résultat

4. Interprétation

- Analyse de l'inertie :

  Inertie totale = K/J - 1

- En ACM, les valeurs propres sont souvent nombreuses (nombre de dimensions du tableau = nombre total de modalités – nombre de variables) et peu discriminantes

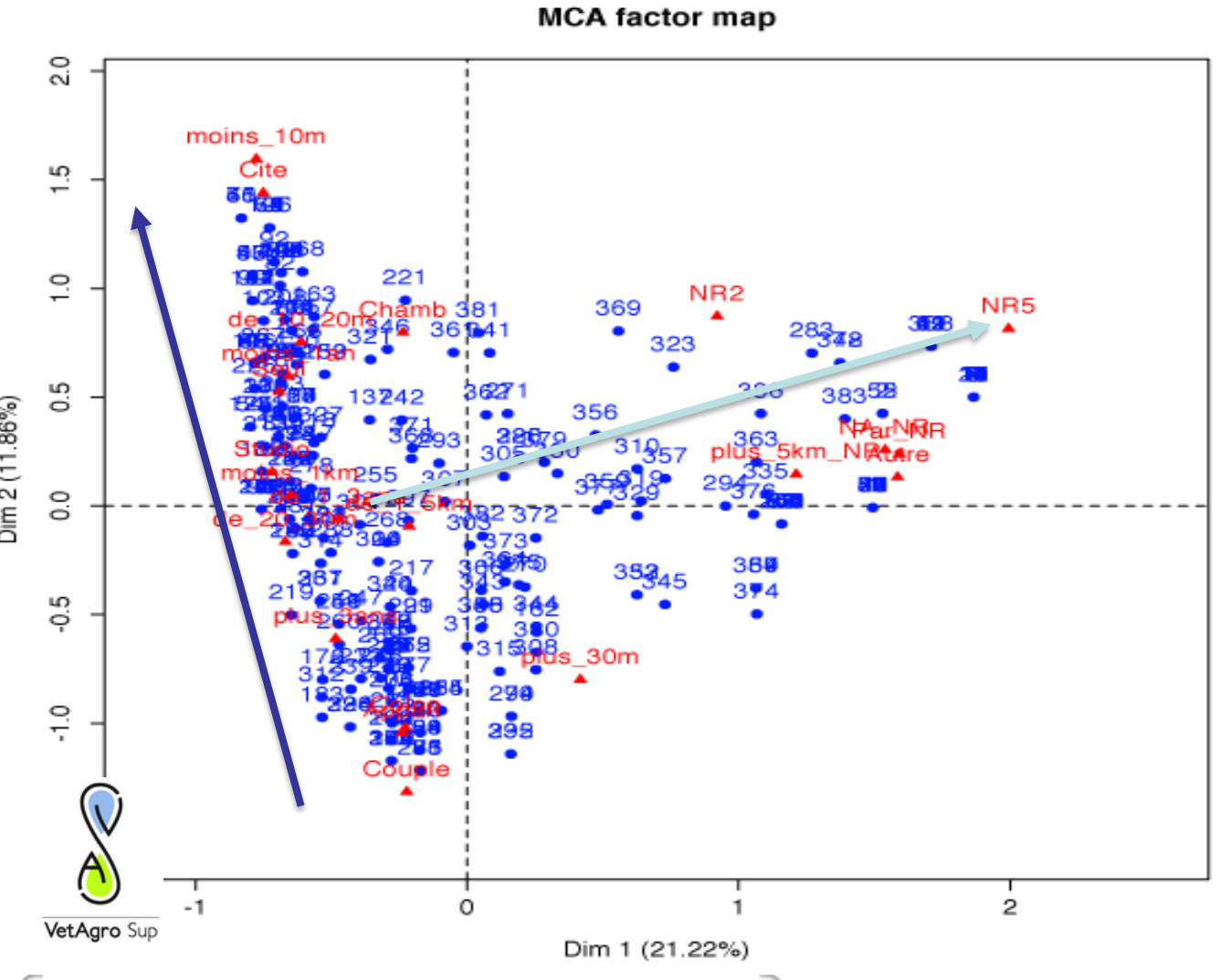


VetAgro Sup

Représentation graphique

- Représentation simultanée des points – modalités et éventuellement des points-individus (pas toujours intéressant)
- Chaque modalité est au barycentre des individus qu'elle représente
- La proximité entre modalités sur le plan factoriel représente leur association

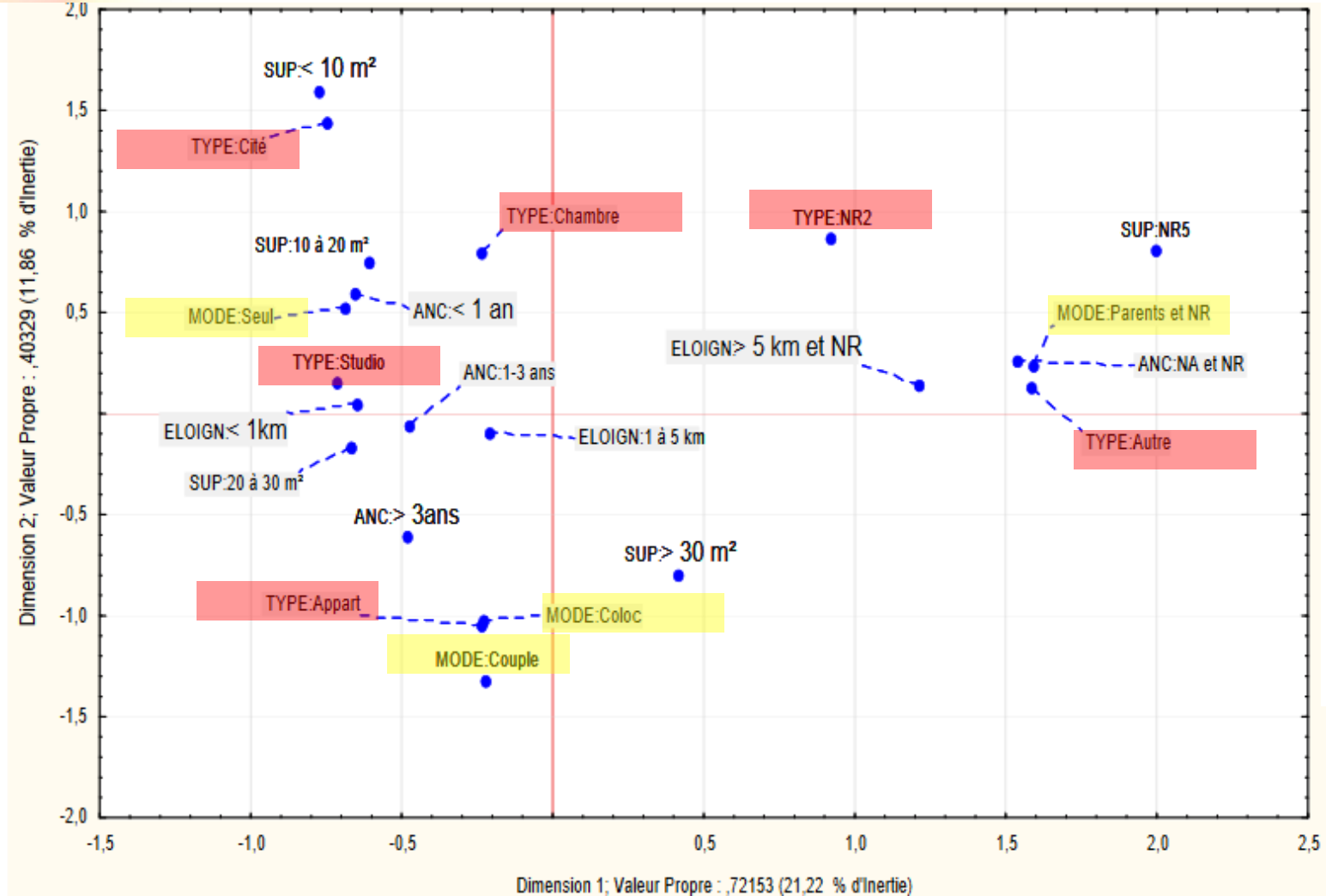
Exemple: logement étudiant



Points individus peu informatifs

Allure générale du nuage de points:
Axe 1: groupe à gauche, traine à droite
Axe 2: gradient du groupe de l'axe 1

Exemple



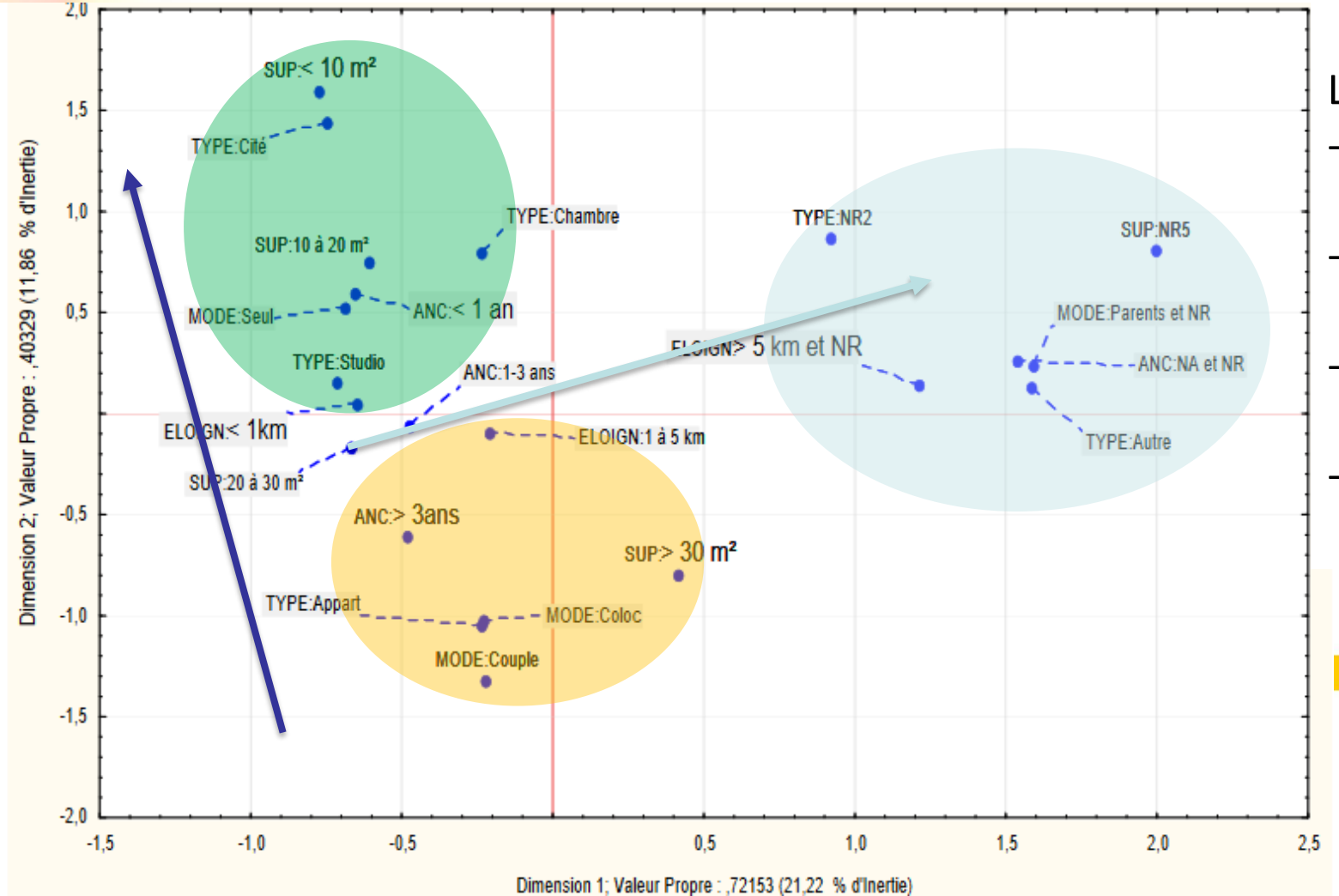
Le barycentre des modalités d'une même variable est au centre du nuage

Une modalité fréquente est proche du centre du nuage

Une modalité excentrée est une modalité peu représentée

Et les variables?

Associations entre variables = associations entre modalités



Logement étudiant :
- cité / moins de 10m

- non-réponses

- seul / moins de 1 an

- colocation / couple

Aide à l'interprétation

- Rapports de corrélation η^2 = contributions de chaque axe à la représentation de chaque modalité, sommés par variable

  N'interpréter que les variables dont les η^2 sont élevés pour un axe donné!

- Représentation des distributions de points par modalité

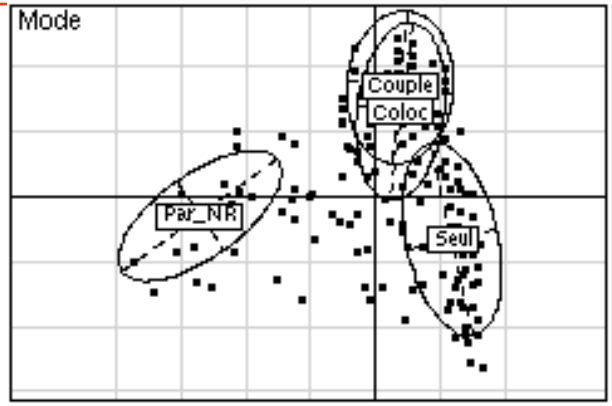
VetAgro Sup

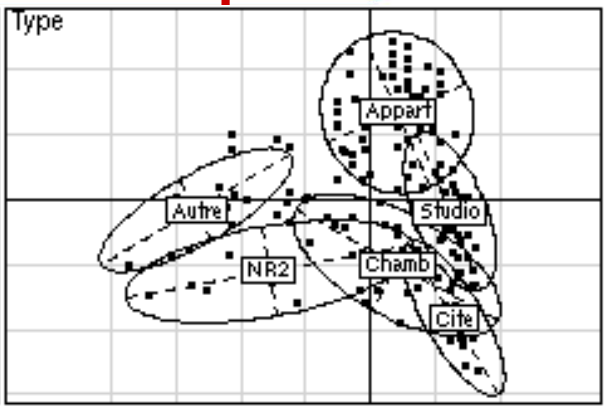
5. Exemples

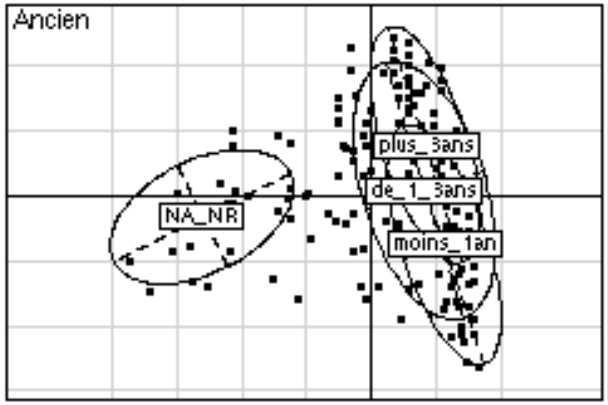
Logement étudiant :

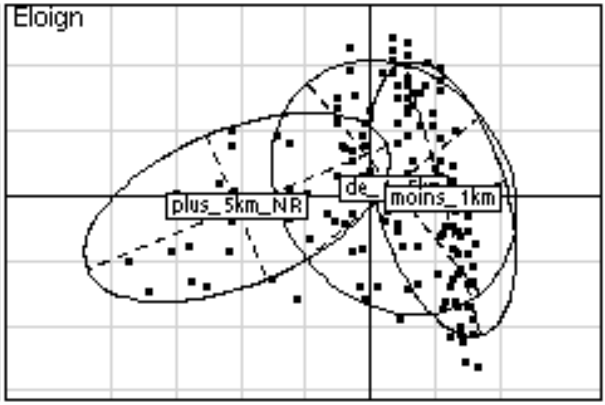
- axe 1: gradient répondants/non répondants + éloignement

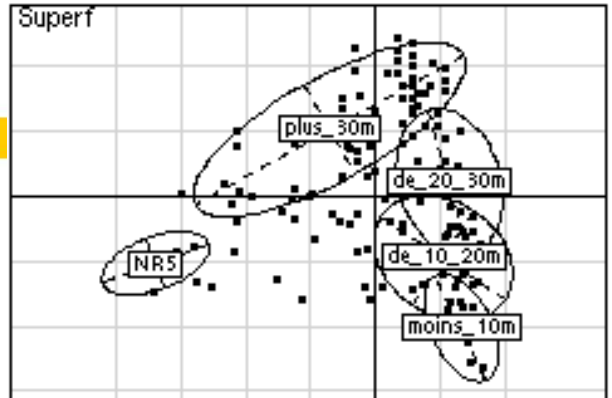
- axe 2: gradient ancienneté – taille du logement – nombre d'habitants







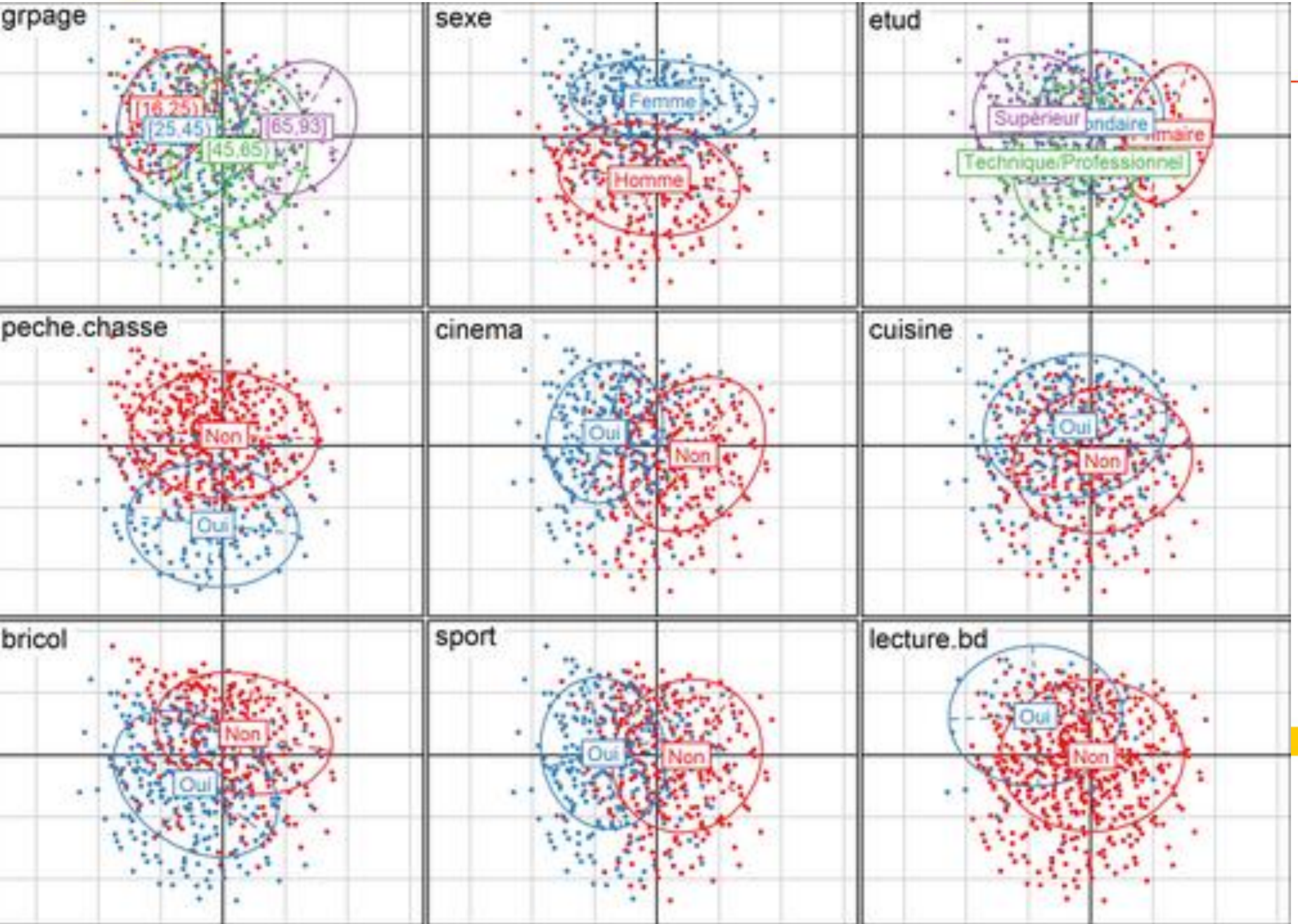




VetAgro Sup

Enquête loisirs selon age, sexe, niveau d'études atteint

Enquête loisirs



Analyses mixtes

1. Données, questions
2. Analyse
3. Interprétation

VetAgro Sup

1. données, questions

- Tableau d'individus décrit par un ensemble variables quantitatives + qualitatives

- Mêmes questions qu'en ACP et en ACM

Principe

- Hill et Smith 1976 -> méthode de Hill et Smith
- Généralisée sous la forme:

  Analyse Factorielle sur Données Mixtes AFDM

  Analyse canonique

2. Analyse

- ACP normée : la quantité maximisée est le coefficient de corrélation r

$$\sum_k r^2(k, F_s)$$

- ACM : quantité maximisée = rapport de corrélation η^2:

$$\sum_j \eta^2(j, F_s)$$

- En AFDM:

$$\sum_k r^2(k, F_s) + \sum_j \eta^2(j, F_s)$$

VetAgro Sup

En pratique

- Juxtaposition d'un tableau des données quantitatives centrées réduites et d'un tableau disjonctif complet dans lequel les 1 ont été remplacés par des $\sqrt{\lambda_s}$

- ACP centrée

VetAgro Sup

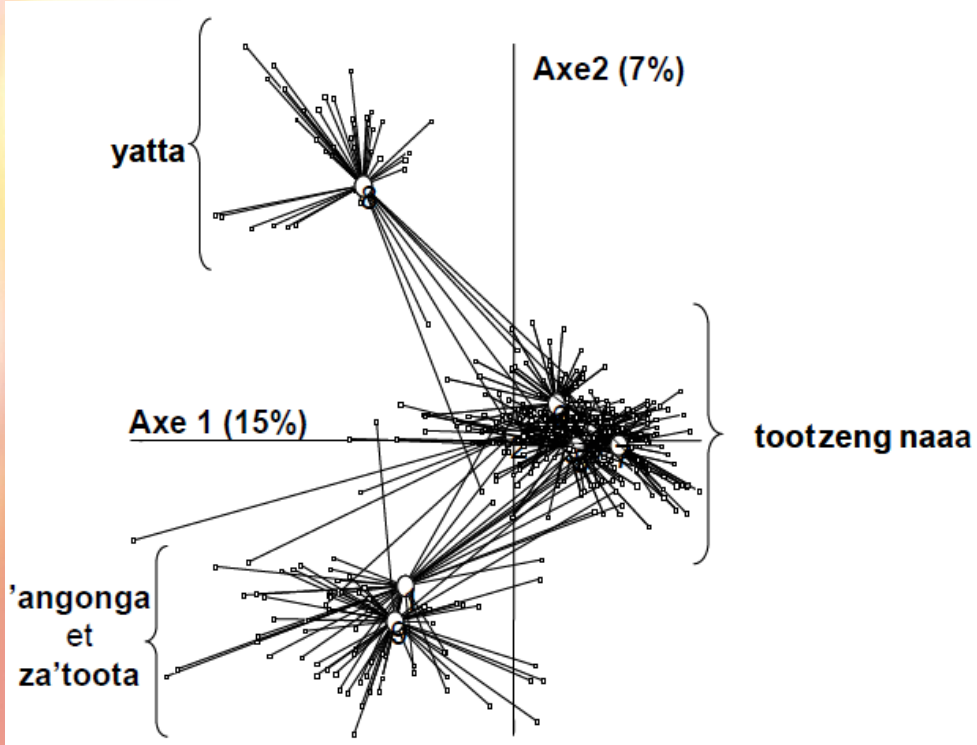
3. Interprétation

- Mêmes éléments d'interprétations que l'ACP et l'ACM : inertie, plans factoriels des lignes, des variables quantitatives et des modalités qualitatives, aides à l'interprétation

- La coordonnée de la variable quantitative k est r_k, celle de la variable quantitative vaut η^2

- Rapports de corrélations (au sens large, r^2 ou η^2) pour toutes les variables

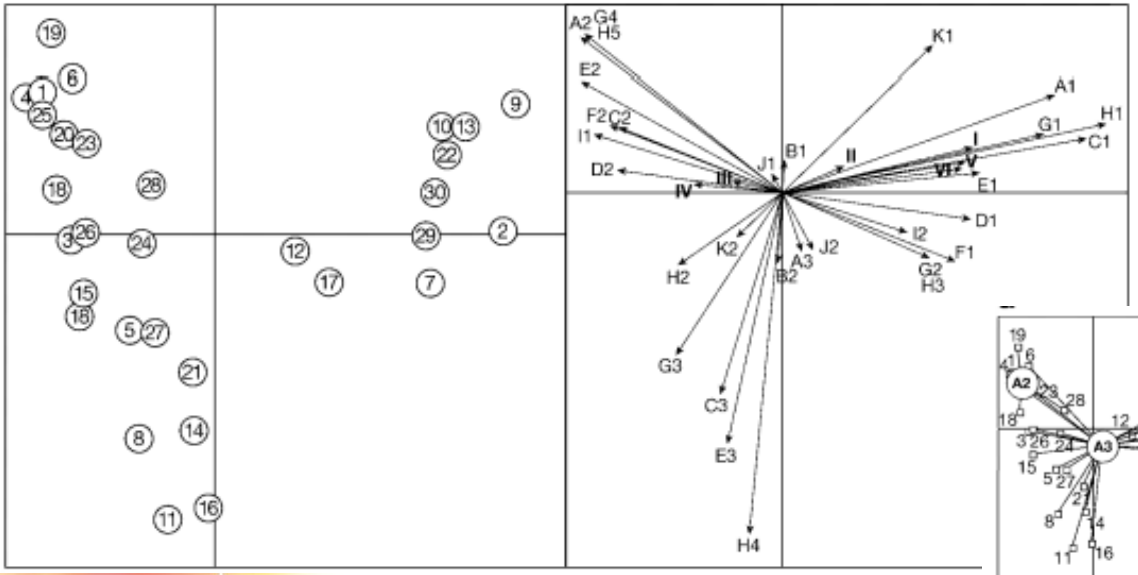
Exemple



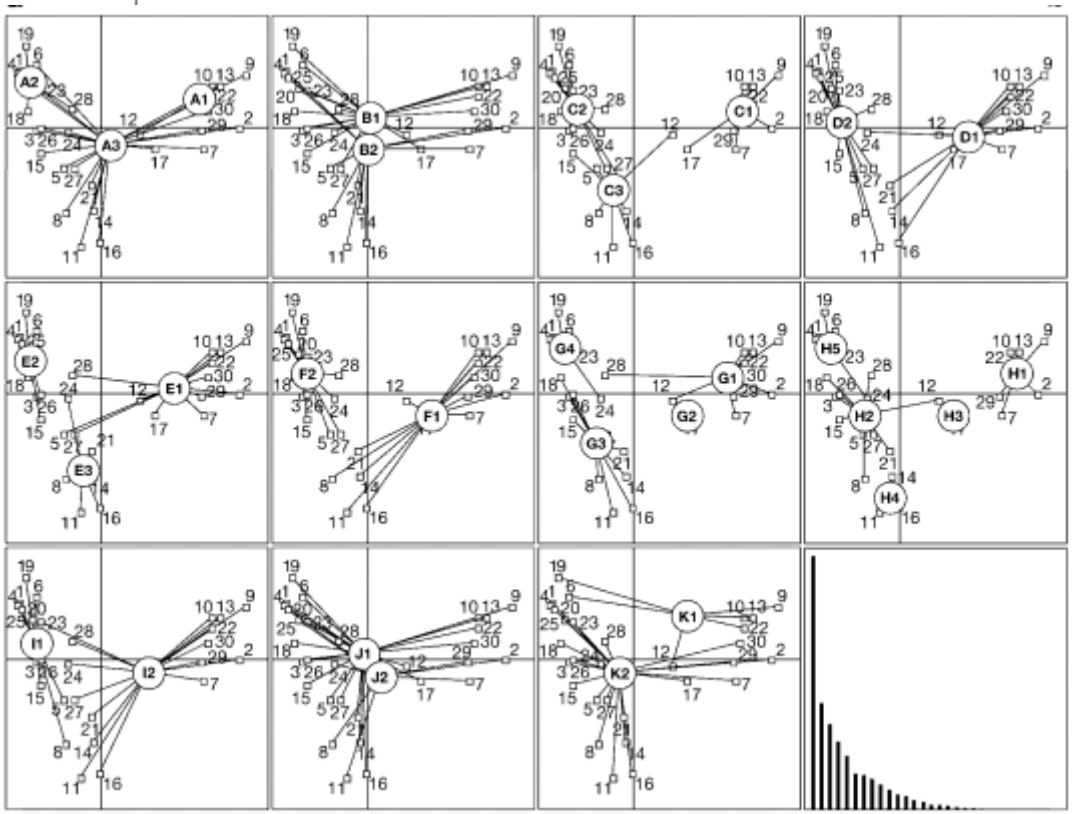
Plants de sorgho de 5 variétés, décrits par
- 14 microsatellites
- 4 caractères morphologiques (forme et compacité de la panicule, couleur des
grains, couverture de la glume) et
- des caractères phénologiques

Barnaud 2007

Exemple: les loricarinae



- 30 individus (1/genre)
- 17 variables qualitatives
- 11 variables quantitatives



Classification des genres et clé d'identification

Covain et Fisch-Muller 2007

VetAgro Sup

Analyse discriminante

Données, questions

Méthode

VetAgro Sup

1. données, questions

Un ensemble d'individus de plusieurs groupes décrits par les mêmes variables

Discrimination descriptive : chercher ce qui sépare des groupes prédéfinis

Discrimination prédictive : à quel groupe affecter un nouvel individu?

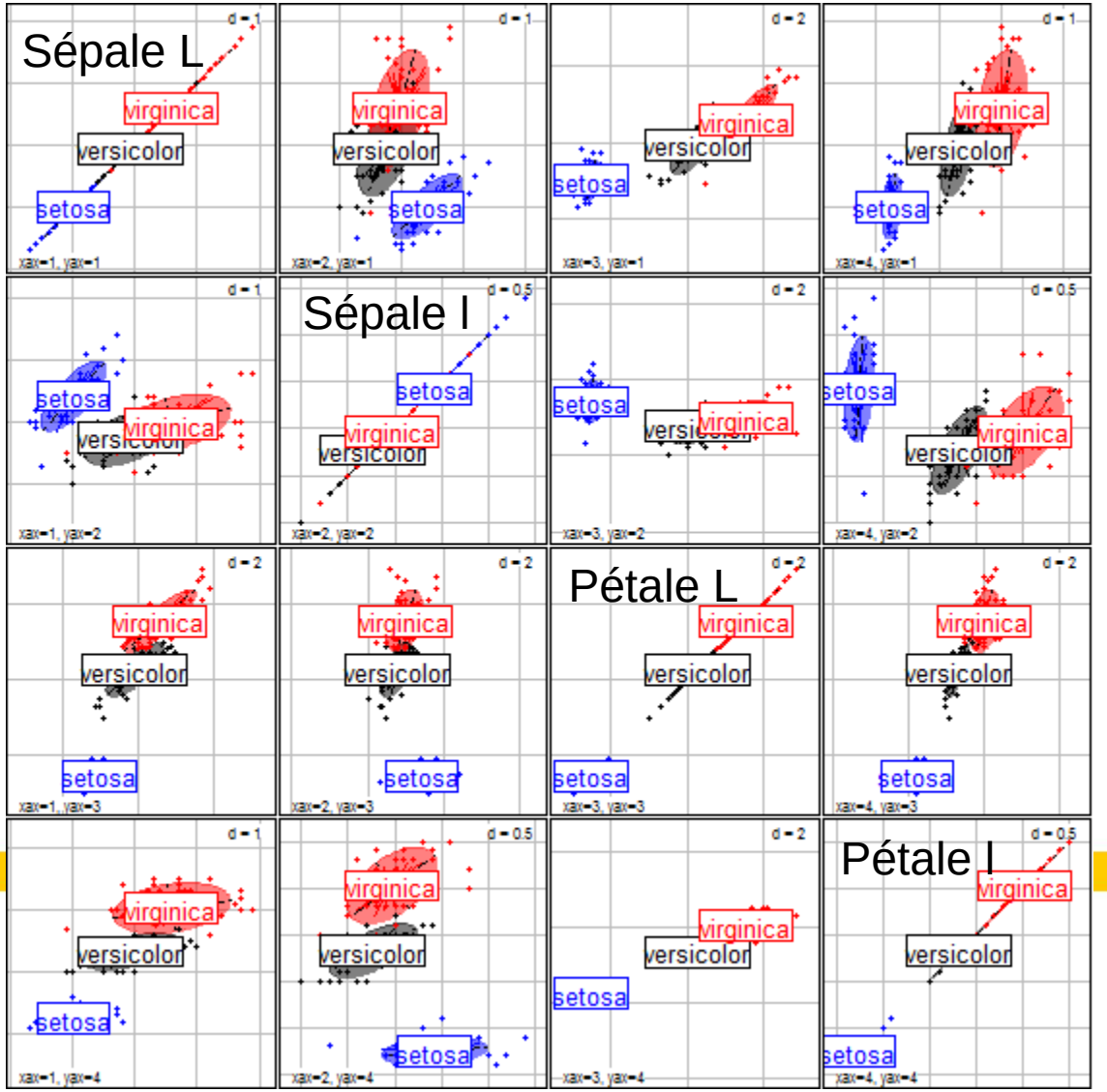
setosa

versicolor

virginica

Exemple

Les « iris de Fisher » (1937, données de Anderson): 4 variables, 3 espèces d'iris



VetAgro Sup

Fonction discriminante

Chaque variable discrimine (partiellement)

Principe: chercher une combinaison linéaire des variables de départ qui
- maximise la variance inter-groupes
- minimise la variance intra-groupe

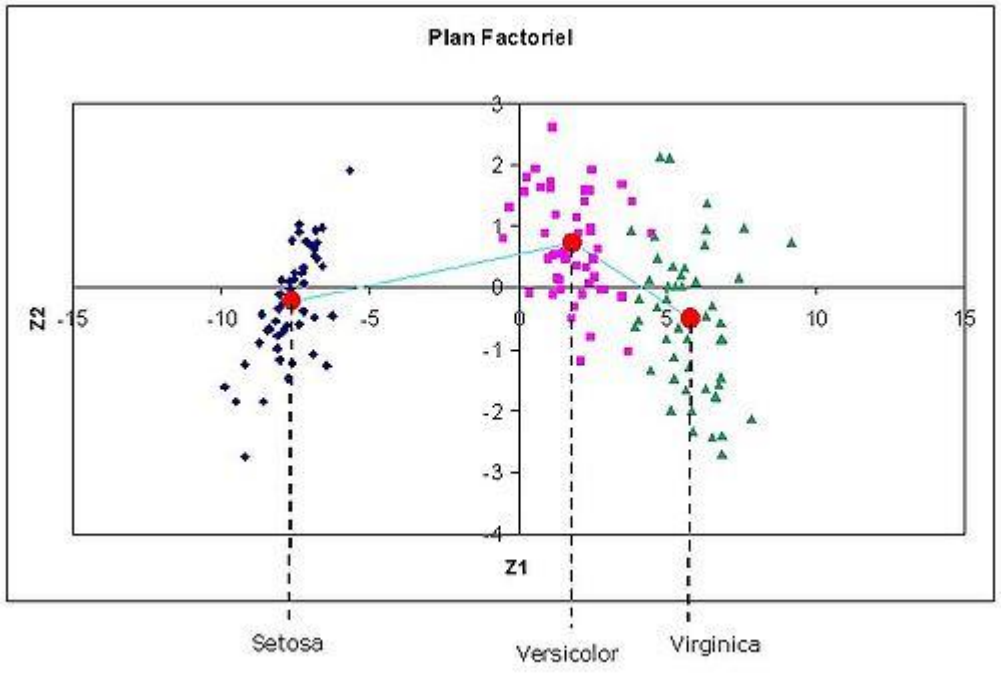
L'algorithme de l'analyse discriminante revient à une ACP des barycentres des groupes, chacun affectés de leur poids respectifs: ACP (ou AFC) pour avoir le tableau normé, puis analyse discriminante


setosa


versicolor


virginica

Exemple



F1 = discrimination maximale

F2 = deuxième axe possible

VetAgro Sup

Interprétation

- Les groupes sont-ils bien séparés? Différents tests possibles

- Qu'est-ce qui sépare? Discrimination descriptive: corrélation entre les premiers axes discriminants et les variables de départ

- [à quel groupe attribuer un nouvel individu? (discrimination prédictive)]

VetAgro Sup

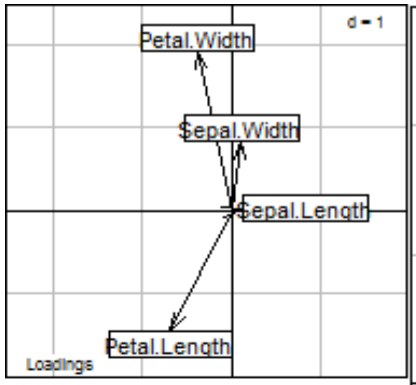
setosa

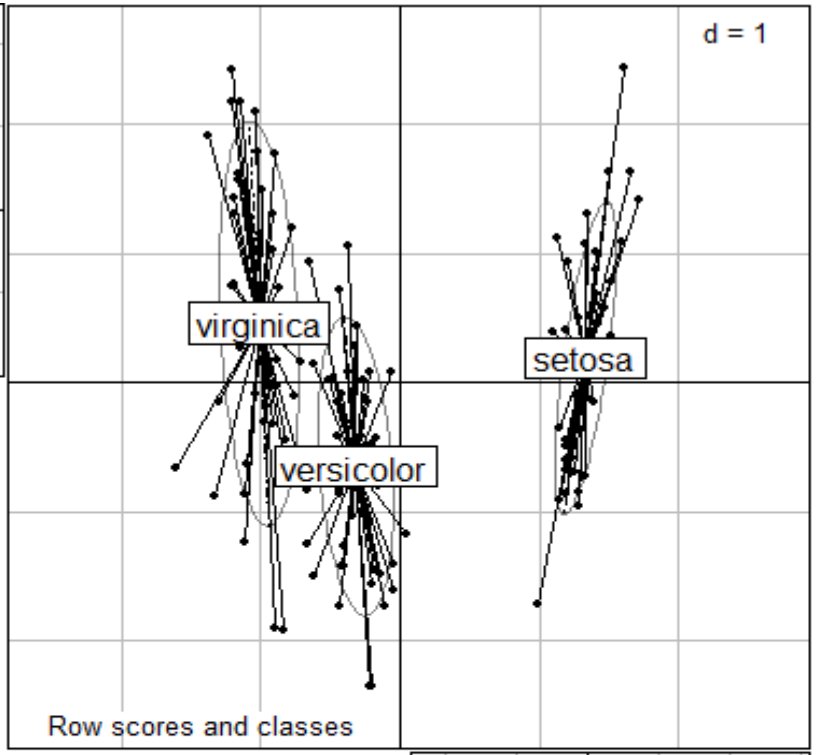
versicolor

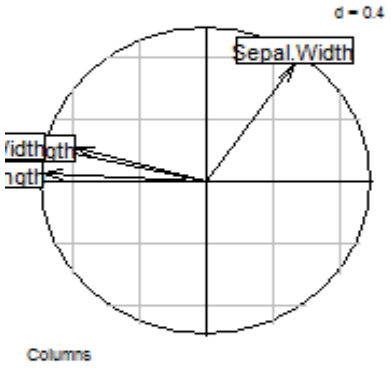
virginica

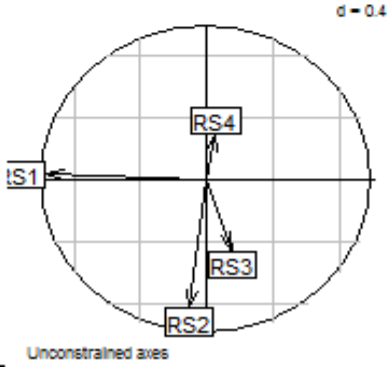
Exemple

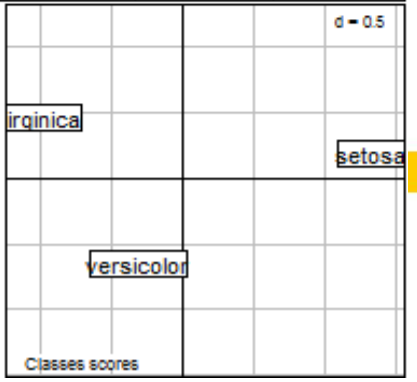
- Discrimination forte sur l'axe 1
- Axe 1 Setosa // autres, axe 2: versicolor/virginica
- Axe 1: pétales et longueur sépales, (=axe 1 de l'ACP des 4 variables); axe 2: largeur sépales (=axe 2 de l'ACP)







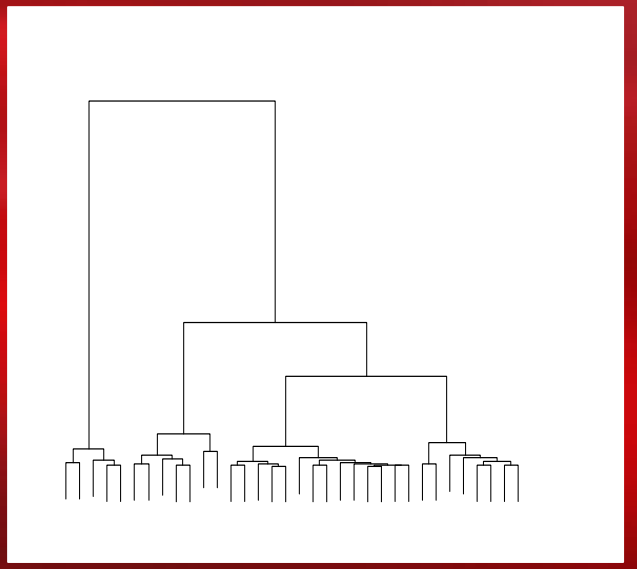




VetAgro Sup

Classification

1. Principe
2. Méthodes de mesure des distances
3. Méthodes de classification



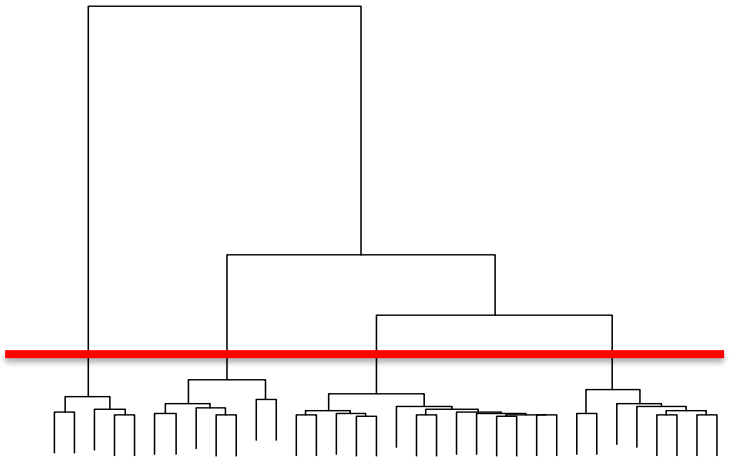
VetAgro Sup

1. Principe

Objectif: répartir les éléments d'un ensemble en groupes tels que:

- tous les éléments d'un groupe sont homogènes
- Les groupes sont au maximum différents entre eux

Hauteur des nœuds peut être proportionnelle à la différence entre les éléments



La troncature du dendrogramme produit une partition: ici 4 groupes

2. Méthodes de mesure des distances

Variables quantitatives: distance euclidienne
Y compris pour les distances entre points sur un plan factoriel d'ACM ou d'AFC

Variables qualitatives: indice de Jaccard

$$d = 1 - \frac{c}{p + q - c}$$

p/q = nombre de modalités présentes chez l'individu 1/2
c = nombre de modalités présentes chez les deux individus

VetAgro Sup

3. Méthodes de classification

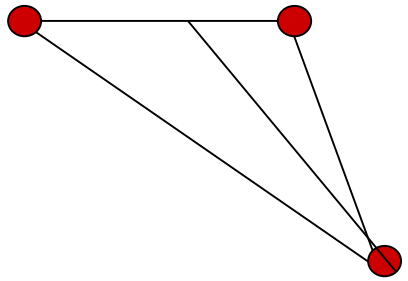
A partir de la matrice des distances des points deux à deux:

- classification ascendante hiérarchique: éléments groupés deux par deux à chaque étape
- classification descendante hiérarchique: un groupe coupé en deux à chaque étape
- partitionnement: une étape, une seule partition

Algorithmes ascendants

Regroupement des deux points les plus proches

Recalcul de la distance entre le nouveau groupe et chacun des autres points:



Lien complet lien moyen lien simple

Algorithme de Ward

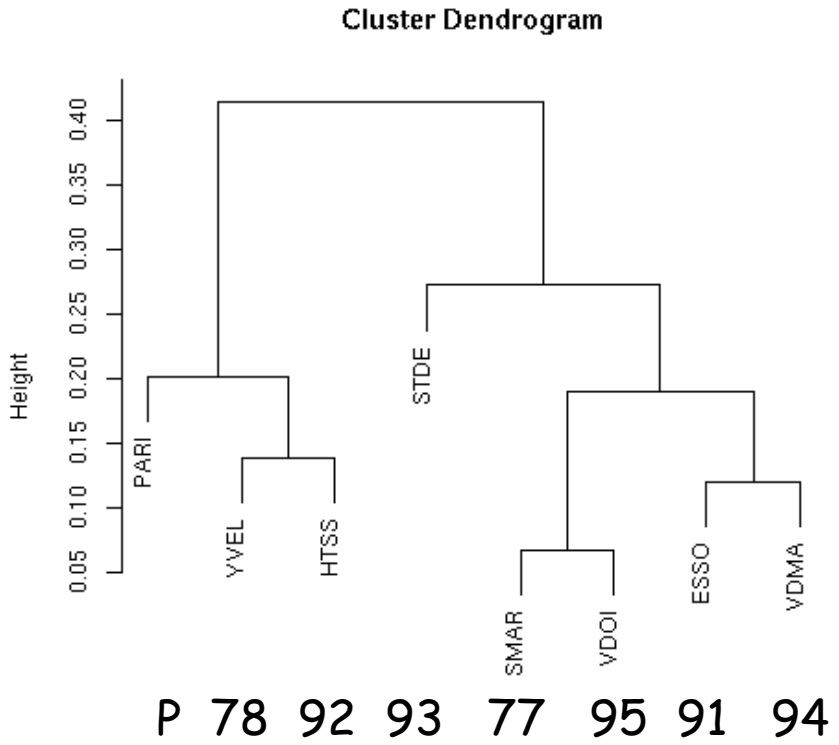
Algorithme ascendant dont le critère est que le regroupement doit faire augmenter l'inertie intra-groupe au minimum

Tendance à grouper des points proches et petits

Exemple

Départements de la région île-de-France : distribution des votes sur 8 listes au 1er tour des élections régionales de 2004

Distances = distances entre les distributions des départements



VetAgro Sup

Conclusion

Ce que font les analyses multivariées

- analyse de l'inertie d'un tableau / de la co-inertie de plusieurs tableaux

- analyse de la structure : groupes de lignes/colonnes, correspondances lignes/colonnes, partitionnement

- synthèse des informations

Ce qu'elles ne font pas

De l'inférence

Souvent complétées, par ex. par des modèles linéaires simples/multiples

VetAgro Sup