Cours 5: Exemples d'application
www.enseeiht.fr/~gergaud/teaching

Joseph Gergaud

INP ENSAT

30 novembre 2005

	PVP	AGR	CMI	...	Total
1872	18	0.5	0.1	...	100
1880	14.1	0.8	0.1	...	100
1890	13.6	0.7	0.7	...	100
1900	14.3	1.7	1.7	...	100
1903	10.3	1.5	0.4	...	100
1906	13.4	1.4	0.5	...	100
⋮	⋮	⋮	⋮		...
1971	12.8	2.8	7.1	...	100

- PVP: pouvoirs publics
- AGR: agriculture
- CMI: commerce et industrie
- TRA: travail
- LOG: logement et aménagement du territoire
- EDU: éducation
- ACS: action sociale
- ACO: anciens combatant;
- ...

Valeurs propres

1. Quelle est la quantité d'information en terme d'inertie dans le plan factoriel 1-2?

$$I_{1-2} = \lambda_1 + \lambda_2 = 4.9734 + 2.0499 = 7.0233$$

2. Quel pourcentage de l'inertie totale cela représente-t-il?
$\tau_2 = 7.0233/11 = 0.638$
3. Dans l'ACP la dernière valeur propre est nulle. Pourquoi? La somme des lignes de X est toujours de 100 $\Rightarrow rg(Z) \leq 10$

Interprétation des axes 1 et 2

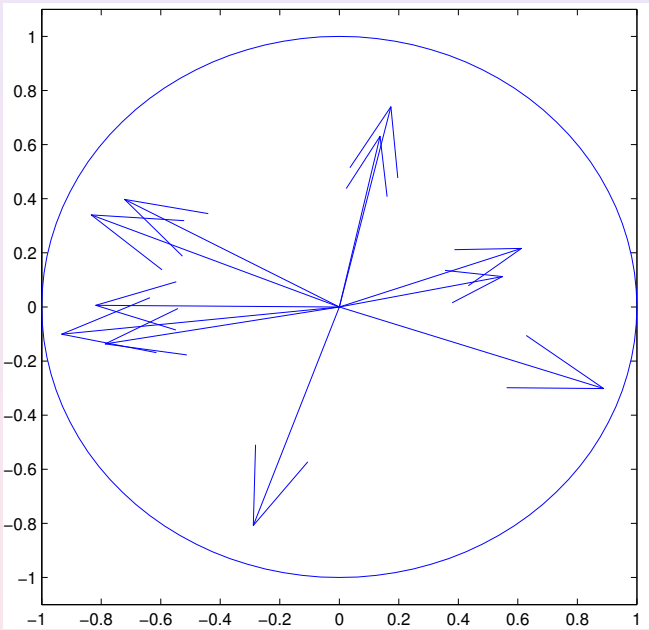


FIG.: Plan 1-2 des variables

- axe 1
  - AGR, CMI, LOG, EDU, ACS ont une corrélation avec l'axe 1 "proche" de -1
  - DEF, DET et DIV ont une corrélation avec l'axe 1 "proche" de $+1$
- axe 2
  - PVP et TRA ont une corrélation "proche" de $+1$ avec l'axe 2
  - ACO a une corrélation "proche" de -1 avec l'axe 2

Point bien représenté?

- $\cos^2(\theta_{22}) = 0.2384$, ce point n'est donc pas bien représenté sur l'axe 2.
- $\cos^2(\theta_{21}) + \cos^2(\theta_{22}) = 0.4511 + 0.2384 = 6895$. Ce point est bien représenté sur le plan 1-2.

$\cos^2\theta$

Remarque

Ce sont les $\cos^2$ *qui s'ajoutent. En effet la projection orthogonale du point* M_i *sur le plan 1-2 s'écrit*

$$\overrightarrow{OH}_i = (\overrightarrow{OM}_i|u)u = (\overrightarrow{OM}_i|u_{.1})u_{.1} + (\overrightarrow{OM}_i|u_{.2})u_{.2},,$$

avec $(\overrightarrow{OM}_i|u) = \cos(\theta)||\overrightarrow{OM}_i||$ *et*

$$(\overrightarrow{OM}_i|u_{.1}) = \cos(\theta_{i1})||\overrightarrow{OM}_i|| \quad (\overrightarrow{OH}_i|u_{.2}) = \cos(\theta_{i2})||\overrightarrow{OM}_i||$$

Par suite le théorème de Pythagore donne

$$\cos^2(\theta) = \cos^2(\theta_{i1}) + \cos^2(\theta_{i2}).$$

Interprétation

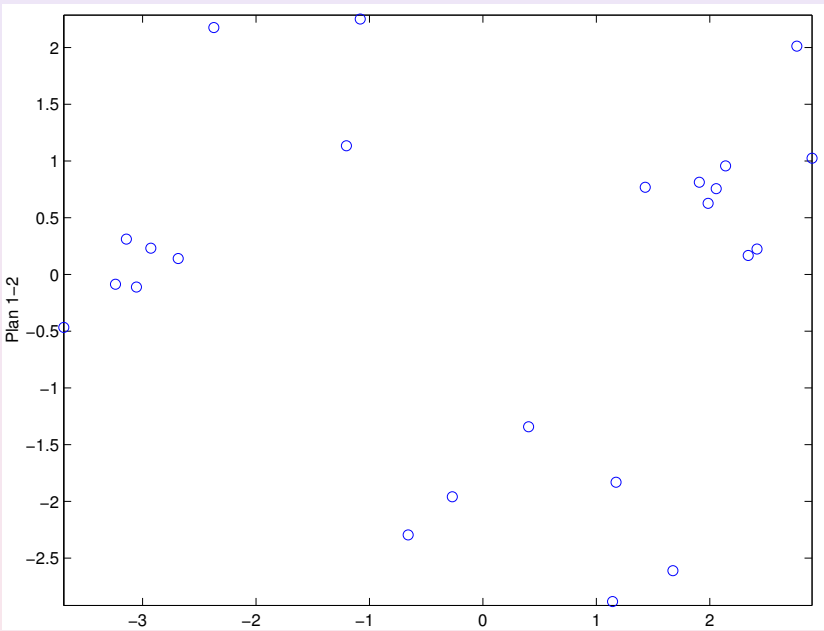


FIG.: Plan 1-2 des individus

- Aucun point n'est mal représenté
- 4 groupes d'années
  - Années de 1872 à 1920 (haut, droit): PVP, DEF et DET
  - Années de 1923 à 1938 (bas): ACO
  - Années 1947, 1950 et 1953: PVP et TRA
  - Années de 1956 à 1971 (gauche): AGR, CMI, LOG, EDU et ACS

- Donner les coordonnées des points M_i sur les axes 1 et 2 en fonction de Y et U.

$$\psi^Y = YU$$

- Démontrer que l'inertie de l'axe j est donné par la j*ème* valeur propre

$$I_j = \frac{1}{n}\sum_i (\psi_{ij}^Y)^2 = \frac{1}{n}(Yu_j/Yu_j) = (\frac{1}{n}{}^tYYu_j/u_j) = \lambda_j$$

- Donner les coordonnées des variables définies par les colonnes de Z sur l'axe j en fonction de U et Λ

$$\phi_{.j} = {}^tZv_{.j} = {}^tZ\frac{1}{\sqrt{\lambda_j}}Zu_{.j} = \frac{1}{\sqrt{\lambda_j}}{}^tZZu_{.j} = \frac{1}{\sqrt{\lambda_j}}\lambda_j u_{.j} = \sqrt{\lambda_j}u_{.j}$$

Budget-temps

- 27 "individus"=groupes d'individus identifiés par
  - caractère 1 = âge (1=-35, 2=+35, 3=+50)
  - caractère 3 = niveau d'étude (1=primaire, 2=secondaire, 3=supérieur)
  - caractère 4 = agglomération (1=rurale, 2=ville moyennes, 3=ville importante, 4=agglo parisienne)
- 16 variables:
  - Somm = Sommeil
  - Repo = Repos
  - Reps = Repas chez soi
  - Repr = Repas restaurant
  - Trav = travail rémunéré
  - ...

Inerties

- $\sum_{j=1}^{p} \lambda_j = p = 16 = \text{Inertie totale.}$
- Valeurs propres

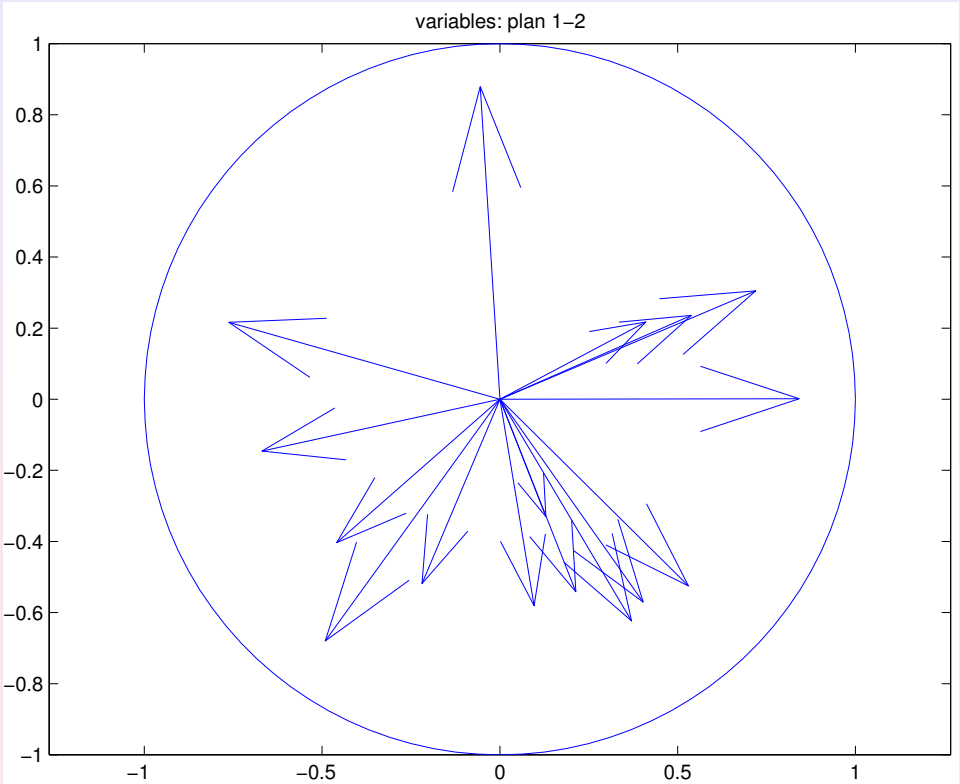
$$3.8711, 3.6605, 2.0066, 1.5143, 1.1266, \ldots 0.0198$$

- Pourcentage d'inertie

axes	% d'inertie	% d'inertie cumulée
1	24.20	24.20
2	22.88	47.07
3	12.54	59.61
4	9.47	69.68
5	7.04	76.12

- La sixième valeur propre est < 1 (critère de Kaiser).

Variables



- $r(\text{Ména},\text{Disq}) = 0.4966.$
- $r(\text{Trar},\text{Repr}) = -0.0101.$
- $r(\text{Repr},\text{Reps}) = -0.5273.$

FIG.: Plan 1-2

Interprétation des axes

- axe 1

+	-
Repas chez soi	Repas restaurant
Jardinage, bricolage	Loisirs estérieurs
Repos	Mémage
Fréquentation Média	Disque, cassette
	Lecture
	Trajet en voiture

  Interprétation: opposition entre les activités extérieures ou d'ouverture et les activités d'intérieures.

- axe 2

  Opposition entre le travail rémunéré et les activités de temps disponibles

Individus

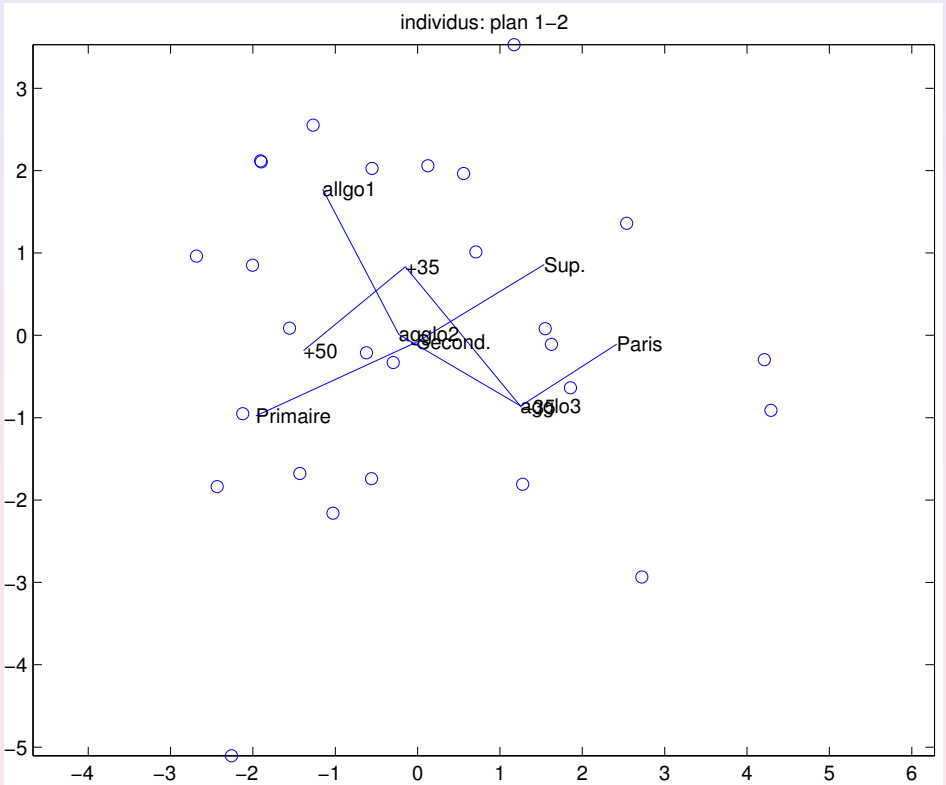


FIG.: Plan 1-2

- L'individu 1115, en bas et à gauche est éloigné des autres points Ceci suggère qu'il a un emploi du temps atypique par rapport aux critères des axes 1 et 2.
- L'individu 2123 est lui proche de l'origine. Il a un comportement moyen par rapport aux critères définis par des axes 1 et 2.

$\cos^\theta$ et contribution

- $$Ctr_{22} = 100\frac{(1/n)(\psi_{22}^Y)^2}{\lambda_2} = 100\frac{(1/27)(-5.1053)^2}{3.6602} = 27.27\%$$

- $$\cos^2(\theta_{15,1}) = \frac{(\psi_{ij}^Y)^2}{||\vec{OM}_i||^2}$$
$$= \frac{(-0.2970)^2}{(0.2970)^2 + (-0.3311)^2 + 0.7627^2 + \cdots} = 0.0106$$

Cet individu est donc très mal représenté sur l'axe 1.

Définition

- La variable âge à 3 modalité devient 3 individus supplémentaires:
  - age1=moyenne des données des individus de modalité 1;
  - age2=moyenne des données des individus de modalité 2;
  - age3=moyenne des données des individus de modalité 3;
- La variable niveau d'étude devient 3 individus supplémentaires:
  - niv1=moyenne des données des individus de modalité 1;
  - niv2=moyenne des données des individus de modalité 2;
  - niv3=moyenne des données des individus de modalité 3;
- La variable type d'agglomération devient 4 individus supplémentaires.

On projette ensuite c'est individus supplémentaires dans le plan 1-2; attention, il faut centrées et réduire ces données supplémentaires.

Nuage des individus

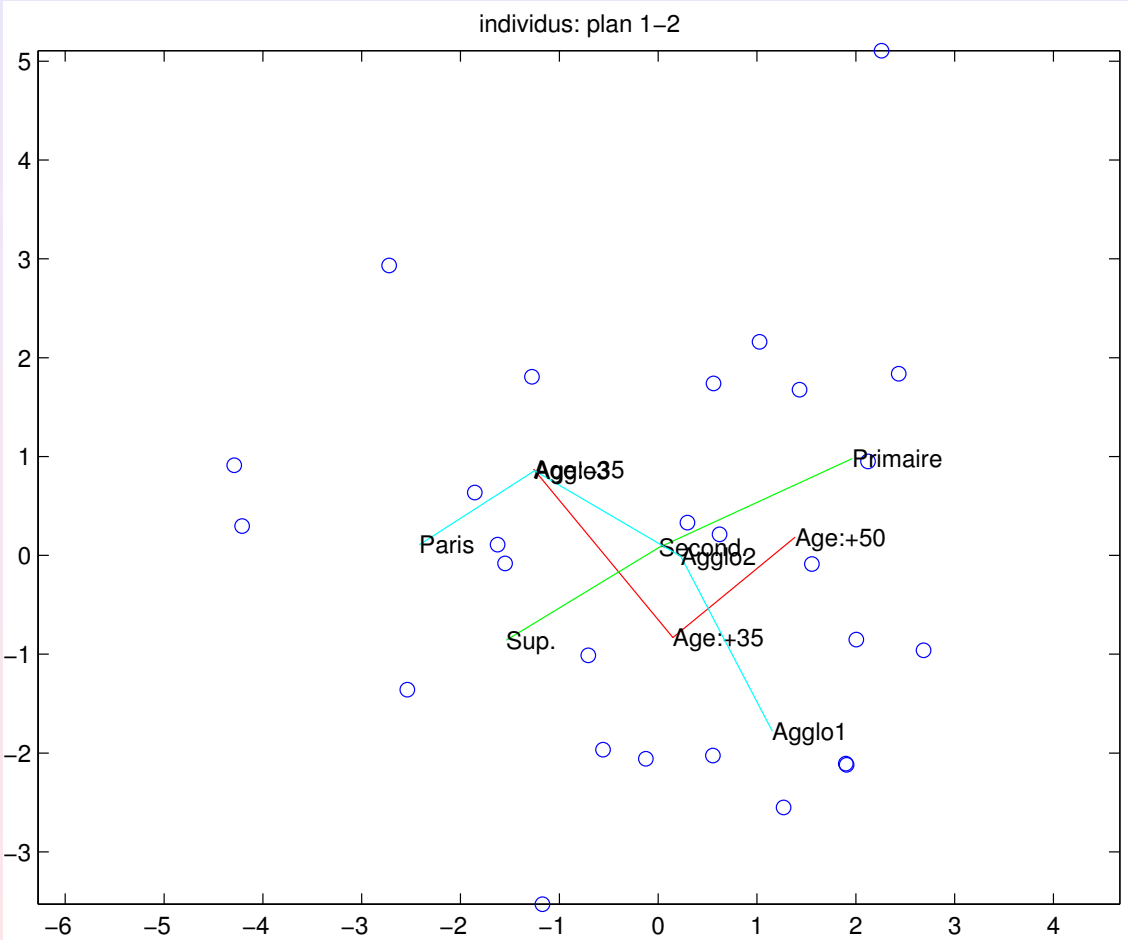


FIG.: Plan 1-2

Image du satellite SPOT région de Fabas

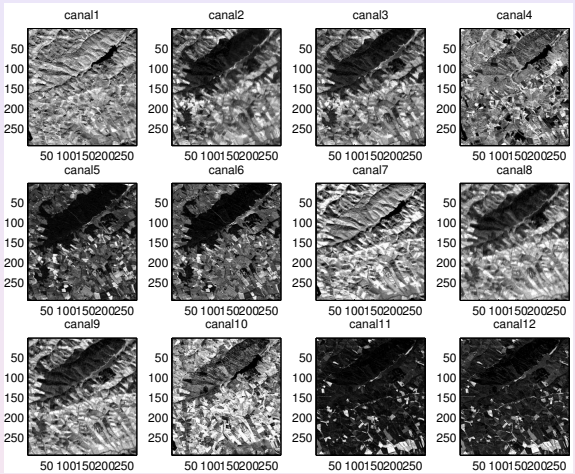


FIG.: 12 canaux initiaux

Remarque

Cette méthode est couramment utilisée dans les Systèmes d'Information Géographique (cours de J.P. Lacombe).

Données numériques

- 1 image $\longrightarrow$ 293 lignes$\times$282 colonnes de pixels
- 1 pixel $\longrightarrow$ un nombre (niveau) de 0 à 255

10 premières lignes et colonnes de la photo 1

150	82	80	163	74	109	107	85	98	58
116	97	109	174	87	109	113	85	78	54
89	82	87	127	74	87	113	85	98	77
100	112	109	117	117	117	125	74	98	89
150	173	153	141	104	91	166	117	127	123
161	188	175	206	80	109	208	138	167	113
155	203	182	226	67	98	208	170	196	87
161	203	190	242	60	87	202	181	196	67
183	120	146	157	74	98	190	138	137	164
194	143	146	106	74	80	250	159	157	152

Questions

1. la meilleure image noir et blanc;
2. la meilleure image couleur.

Données pour l'ACP: X

150	82	80	...
116	97	109	...
89	82	87	...
⋮	⋮	⋮	...

- un individu = un pixel
- une variable = une image
- x_{ij} = niveau du pixel i pour la j-ième longueur d'onde
- X est de dimension $(293 \times 282{,}12) = (82626{,}12)$

Individus

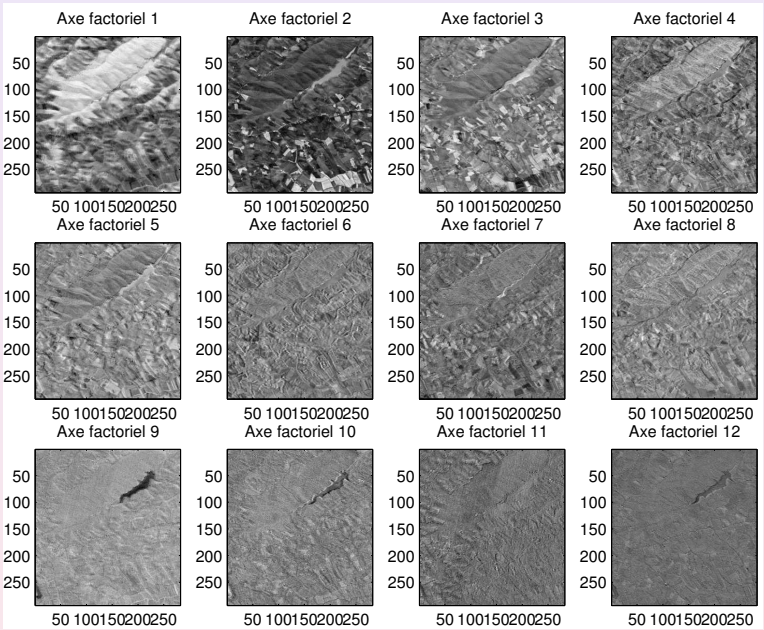


FIG.: 12 canaux ACP

- $\psi = ZU$
- Recodage: $\psi'_{.j} \in \{0, \ldots, 255\} \quad \rightarrow$ 1 image
- $\lambda =$ (5.5027,2.0452,1.7215,0.8298, ...,0.0233,0.0145)
- La meilleure image noire et blanc est $\psi'_{.1}$
- C'est celle de plus fort contraste
- Sur cette image on a 45.9% de l'information

Meilleure image couleur

- Un pixel d'une image couleur = (un niveau de rouge, un niveau de vert, un niveau de bleu)
- $\psi'_{.1}$ = canal rouge
- $\psi'_{.2}$ = canal vert
- $\psi'_{.3}$ = canal bleu
- On a sur cette image 77.25% de l'information

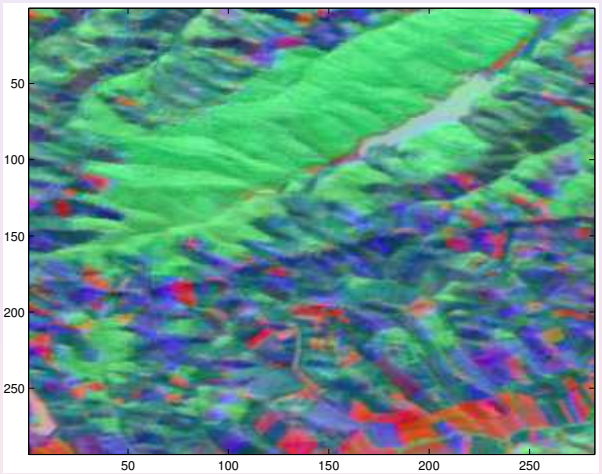


FIG.: Meilleure image couleur

Variables

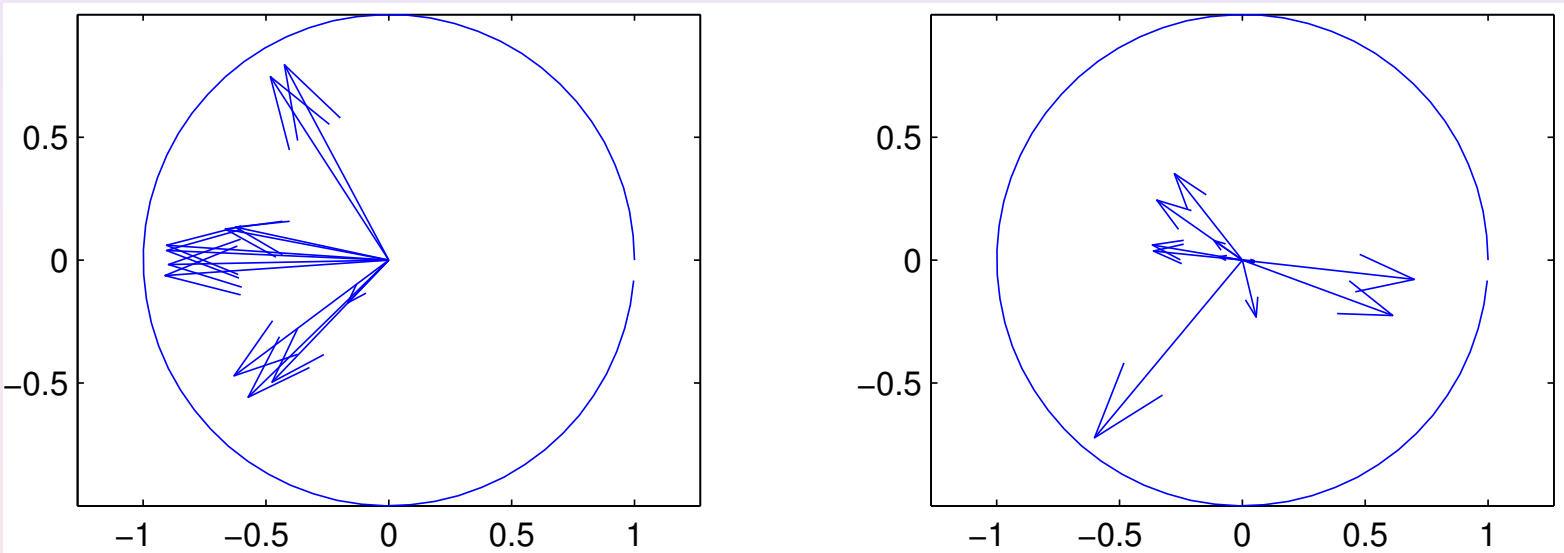


FIG.: Variables: plans 1-2 et 3-4

Remarques

- Images pseudo couleurs
- On cherchera ensuite les liens avec la végétation, la nature du terrain, ... (cf. cours de SIG de J.P. Lacombe)