2V481 – BM2

Cours 3

LICENCE SCIENCES ET TECHNOLOGIE

MENTION SCIENCE DE LA VIE – L2

LARSEN MARTIN
DEMEYRIER VIRGINIE
RYBARCZYK HERVÉ

Analyses multivariées

LICENCE SCIENCES ET TECHNOLOGIE

MENTION SCIENCE DE LA VIE – L2

LARSEN MARTIN
DEMEYRIER VIRGINIE
RYBARCZYK HERVÉ

Introduction

- En quoi cela consiste-t-il ?
  - En statistiques, les **analyses multivariées** ont pour caractéristique de s'intéresser à la distribution conjointe de plusieurs variables.
  - Les analyses multivariées sont très diverses selon l'objectif recherché, la nature des variables et la mise en œuvre formelle.
  - On peut identifier deux grandes familles :
    - celle des méthodes descriptives (visant à structurer et résumer l'information)
    - celle des méthodes explicatives visant à expliquer une ou des variables dites « dépendantes » (variables à expliquer) par un ensemble de variables dites « indépendantes » (variables explicatives).
  - Les méthodes appelées en français analyse des données ou « data mining » en sont un sous-ensemble.

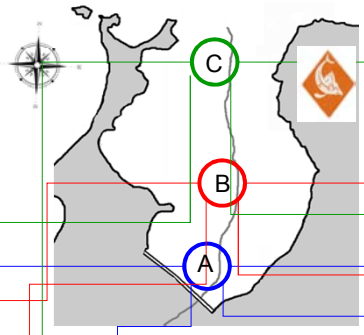
Introduction

- Objectif principal des analyses descriptives multivariées
  - Etudier ou décrire un ensemble de variables prises globalement.
  - Cette étude permet de synthétiser et de visualiser rapidement une grande quantité d'informations.
  - Elle repose sur un examen des interdépendances entre tous les variables.

Introduction

EXEMPLE DE TRAITEMENT DE DONNEES MULTIVARIEES

1 site
3 stations
nprélèvements/station



Faune
Macrofaune
Méiofaune

Micro-organismes
Abondance bactéries
Chlorophylle *a*

Données environnementales
Taille des grains
Sels nutritifs : (nitrite, nitrates, phosphates, silicates)

Données fonctionnelles
Production primaire
Respiration
Productivité

Introduction

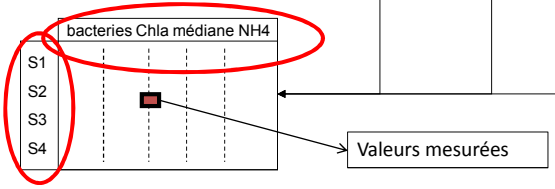
DONNEES « MESUREES » QUANTITATIVES

Micro-organismes
Abondance bactéries
Chlorophylle *a*

Données environnementales
Taille des grains
Sels nutritifs : (nitrite, nitrates, phosphates, silicates)

Données fonctionnelles
Production primaire
Respiration
Productivité

Un tableau de mesures site/variables



Analyses des composantes principales (ACP)

Analyse en composantes principales (ACP)

- Statistique exploratoire multidimensionnelle
- Ensemble de méthodes permettant de procéder à des transformations linéaires d'un grand nombre de variables intercorrélées de manière à obtenir un nombre relativement limité de composantes non corrélées.
- Cette approche facilite l'analyse en regroupant les données en des ensembles plus petits et en permettant d'éliminer les problèmes de multicolinéarité entre les variables.
- Pouvoir expliquer ou rendre compte de la variance observée (**inertie**) dans la masse de données initiales en se limitant à un nombre réduit de composantes, définies comme étant des transformations mathématiques pures et simples des variables initiales.

Analyse en composantes principales (ACP)

Variables aléatoires quantitatives en colonnes

	NH4	NOx	Si	PO4	BCR	GPP	PP	BB	MaB	MeB	Chla	Temp	median
SA1	148	44	125.8	15.7	0.00	13.8	0.5	16.5	15548	1849	28.96	8.30	64.70
SA2	139	8	85.0	20.2	2.55	7.6	0.2	19.1	11931	2863	30.57	13.70	87.37
SA3	105	3	66.7	4.4	6.51	14.1	0.6	29.1	20263	2329	23.43	16.10	161.17
SA4	109	2	84.0	6.6	3.56	9.5	0.4	33.2	18604	716	26.70	19.00	100.37
SA5	139	0	212.3	7.7	3.40	10.2	0.3	35.5	21473	2578	31.67	17.70	141.94
SA6	76	13	63.0	4.7	9.63	22.4	0.8	25.7	24107	1471	28.40	23.30	183.01
SA7	44	9	35.9	3.0	7.42	22.4	0.8	20.4	6731	1451	29.68	18.25	230.58
SA8	133	1	136.5	7.8	9.45	24.8	0.4	37.0	21418	1500	61.63	22.15	87.37
SA9	223	2	184.6	10.8	5.98	13.2	0.2	27.6	29290	1729	63.62	14.20	70.15
SA10	132	6	93.4	7.2	1.62	7.9	0.2	24.7	20458	1757	37.78	7.30	126.45
SA11	55	8	79.6	4	1.44	9.9	0.4	27.9	9854	1899	23.54	10.60	125.00
SB2	81	6	31.7	4.5	0.00	7.8	0.4	8.4	8570	1265	19.31	16.40	183.01
SB3	43	0	17.3	1.0	0.48	7.6	0.5	14.7	9158	1879	15.27	17.70	222.72
SB4	49	0	25.3	2.7	0.00	14.2	0.7	13.7	1642	685	19.74	19.10	238.71
SB5	114	0	88.0	9.0	3.03	12.9	0.6	18.9	7945	1256	23.36	19.60	180.91
SB6	31	13	41.2	3.8	3.55	17.2	0.9	24.3	4079	1163	18.11	23.10	191.67
SB7	25	16	32.3	2.9	5.33	20.8	1.2	27.0	2537	1211	17.89	22.30	235.97
SB8	24	4	28.8	2.7	0.96	26.6	0.7	27.9	792	1604	38.91	19.60	274.21
SB9	30	1	28.8	3.8	0.62	21.4	0.6	10.7	6680	1200	35.22	17.15	244.29
SB10	70	2	40.2	5.5	0.95	21.0	0.4	19.9	4192	1774	47.93	10.90	191.67
SB11	69	4	20.1	5.75	0.00	10.0	0.8	12.6	3697	905	12.92	14.60	189.46
SC2	17	8	2.0	2.0	0.00	6.1	0.7	4.2	2475	265	8.69	9.20	238.71
SC3	7	0	20.3	0.0	0.00	5.4	0.5	3.2	2515	274	12.00	15.40	238.71
SC4	6	0	4.0	0.0	0.00	3.6	0.3	10.2	1041	266	10.44	18.10	247.13
SC5	9	0	4.7	0.0	0.00	6.2	0.6	11.2	1831	274	9.59	14.80	258.82
SC6	9	0	8.0	0.5	1.18	5.0	0.6	7.5	1207	508	9.02	19.45	250.00
SC7	10	8	19.8	2.9	2.01	9.6	0.6	22.0	1583	564	15.91	22.30	252.90
SC8	20	5	9.9	2.4	0.00	10.7	0.7	12.2	1176	638	16.12	16.70	258.82
SC9	23	4	8.8	2.9	0.00	9.2	0.5	5.9	1234	671	18.82	15.30	258.82
SC10	17	3	3.9	1.7	0.00	4.3	0.2	6.1	800	253	18.04	11.90	267.94
SC11	0	4	8.5	0.9	0.00	5.8	0.5	7.9	78	218	11.61	14.00	244.29

Observations en lignes

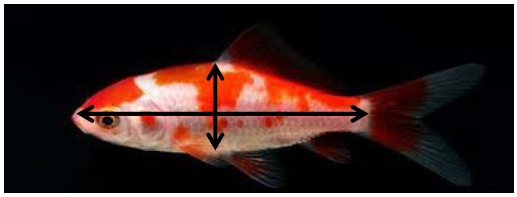
ATTENTION IL NE S'AGIT EN AUCUN CAS D'UN TABLEAU DE CONTINGENCE…

LES SOMMES PAR LIGNE NE VEULENT RIEN DIRE ICI …

- Comment sont-elles corrélées ?
- Peut-on regrouper ou ordonner les observations en fonction de ces variables ?

Analyse en composantes principales (ACP)

- Buts de l'ACP
  - Représenter en 2 ou 3 dimensions l'observation de plus de 3 variables
  - Réduire la dimension de manière pertinente : la réduction du nombre de variable fait perdre de l'information. Comment conserver l'information essentielle du jeu de données ?



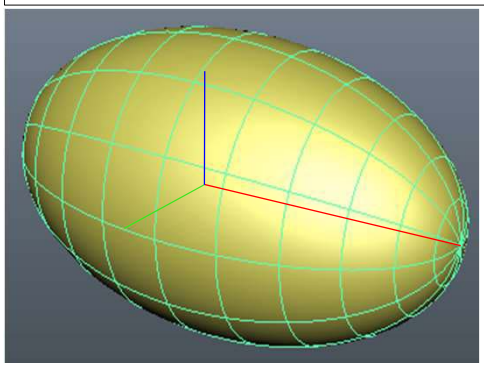
- Etude de la ressemblance entre individus
  => comparaison de profils
- Etude des liaisons entre variables
  => création d'un indicateur synthétique

Deux axes orthogonaux assimilables ici
à la Longueur (axe 1)
et à la Hauteur (axe 2)
qui décrivent un nuage de point

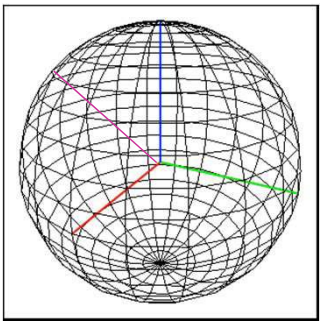
Analyse en composantes principales (ACP)

- Inertie, axes et plan

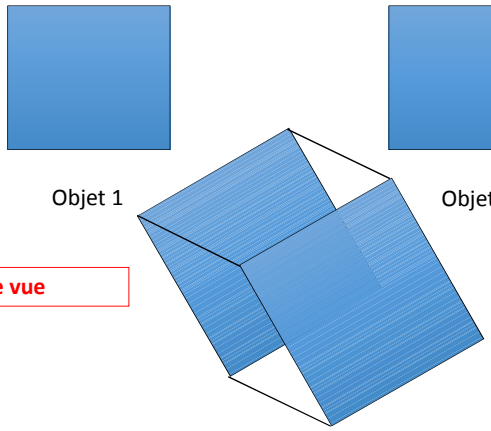
Un axe d'allongement donc une inertie non homogène qui « structure » des axes, donc des plans différents



Pas d'axe d'inertie privilégié
Chaque « direction » représente la même information



Analyse en composantes principales (ACP)

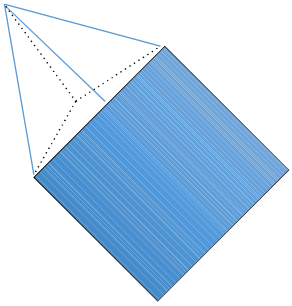


Objet 1

Objet 2

On CHANGE d'angle de vue

Objet 1

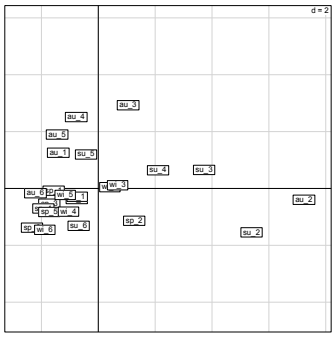


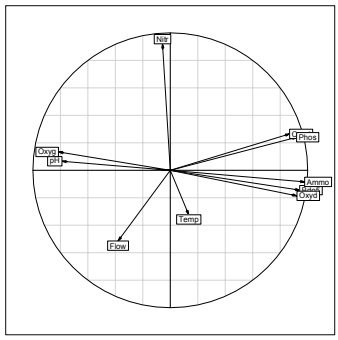
Objet 2

Analyse en composantes principales (ACP)

	Temp	Flow	pH	Cond	Oxyg	Bdo5	Oxyd	Ammo	Nitr	Phos
sp_1	10	41	8,5	295	110	2,3	1,4	0,12	3,4	0,11
sp_2	11	158	8,3	315	13	7,6	3,3	2,65	2,7	1,5
sp_3	11	198	8,5	290	113	3,3	1,5	0,4	4	0,1
sp_4	12	280	8,6	290	126	3,5	1,5	0,45	4	0,73
sp_5	13	322	8,5	285	117	3,6	1,6	0,48	4,6	0,94
sp_6	11	303	8,5	245	100	1,7	0,9	0,05	2,7	0,16
su_1	13	62	8,3	325	95	2,3	1,8	0,11	3	0,13
su_2	13	80	7,6	380	20	21	5,7	9,8	0,8	3,65
su_3	15	100	7,8	385	46	15	2,5	7,9	7,7	4,5
su_4	16	140	8	360	76	12	2,6	4,9	8,4	3,45
su_5	15	160	8,4	345	91	1,7	1,9	0,22	10	1,74
su_6	13	310	8,2	285	82	8,5	1,6	0,59	3,7	0,6
au_1	1	25	8,4	315	91	1,6	0,5	0,07	6,4	0,03
au_2	3	63	8	425	38	36	8	12,5	2,2	6,5
au_3	2	79	8,1	350	84	7,1	1,9	2,7	13,2	3,7
au_4	3	85	8,3	330	106	2	1,4	0,42	12	1,6
au_5	2	72	8,6	305	91	1,6	0,3	0,1	9,5	1,25
au_6	4	181	8,6	270	105	2,8	0,5	0,1	3,66	0,43
wi_1	3	118	8	325	100	1,6	1,2	0,17	1,8	0,19
wi_2	3	252	8,3	360	100	9,5	2,9	2,52	4,6	1,6
wi_3	3	316	8,3	370	100	8,7	2,8	2,8	4,8	2,86
wi_4	3	498	8,3	330	100	4,8	1,6	1,04	4,4	0,82
wi_5	2	390	8,2	330	100	1,7	1,2	0,56	5	0,6
wi_6	3	480	8,2	290	100	1,3	0,8	0,04	2,2	0,13

Tableaux de données
Variables Quantitatives





ACP

Les Observations

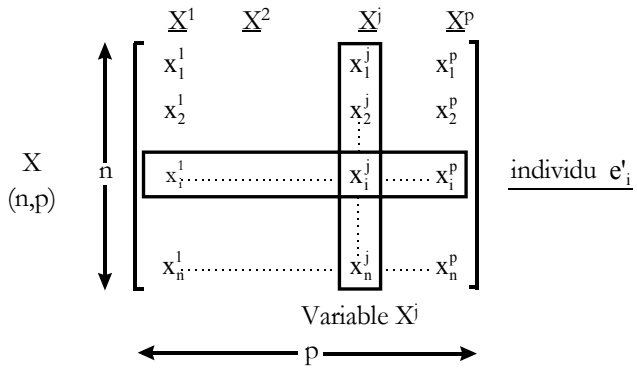
Les Variables

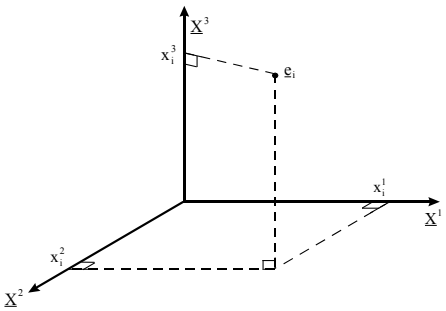
- Mise en place de l'ACP : 3 étapes
  - Centrer et réduire les données
  - Calculer les valeurs et vecteurs propres
  - Projeter les observations

Analyse en composantes principales (ACP)

- p variables quantitatives mesurées sur n individus

INDIVIDU = Élément de R^p
VARIABLE = Élément de R^n





Chaque axe : 1 variable

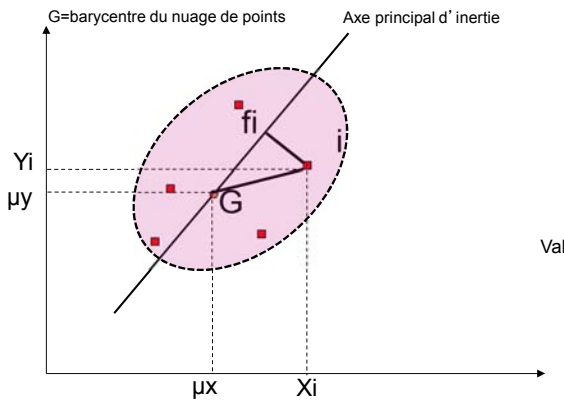
Généralisation à p dimensions d'un graphique cartésien

Analyse en composantes principales (ACP)

- Ajustement du nuage des individus

  ⇒ Recherche de la meilleure représentation axiale du nuage

- Principe général : la notion d'inertie (=variance en statistique)



Pythagore généralisé :

Mesure de la ressemblance par la distance

$(Gi)^2 = (Gfi)^2 + (i\ fi)^2$

$\Sigma(Gi)^2 = \Sigma(Gfi)^2 + \Sigma(i\ fi)^2$

Valeur propre — Inertie projetée — Inertie non-projetée (résiduelle)

Maximiser l'inertie projetée
Minimiser l'inertie résiduelle

Analyse en composantes principales (ACP)

- Principales étapes du calcul de l'ACP :
  - Calcul basé sur les règles du calcul matriciel

1. Centrer et réduire les données

M

	10	8,5	110
	11	8,3	13
	11	8,5	113

→ MC

-0,67	0,07	31,33
0,33	-0,13	-65,67
0,33	0,07	34,33

Matrice centrée par colonne

→ MCR

-1,15	0,58	0,55
0,58	-1,15	-1,15
0,58	0,58	0,60

Matrice centrée et réduite par colonne

moyenne	10,67	8,43	78,67
écart-type	0,58	0,12	56,89

2. Construire la matrice des corrélations entre les colonnes (transposée) de cette matrice

	Temp	Ph	Oxy
Temp	1		
Ph	-0,5	1	
Oxy	-0,47	0,99	1

Matrice des coef de corrélations

(3. Diagonaliser)

4. Avec ces nouvelles matrices, on recalcule des nouvelles coordonnées pour les lignes et les colonnes de notre tableau d'origine

Analyse en composantes principales (ACP)

	1	2	3
1	0,4548	0,8902	-0,0213
2	-0,6319	0,3057	-0,7121
3	-0,6274	0,3374	0,7001

Matrice des vecteurs propres

2,3525	0,64745	3,00E-15

Les trois valeurs propres

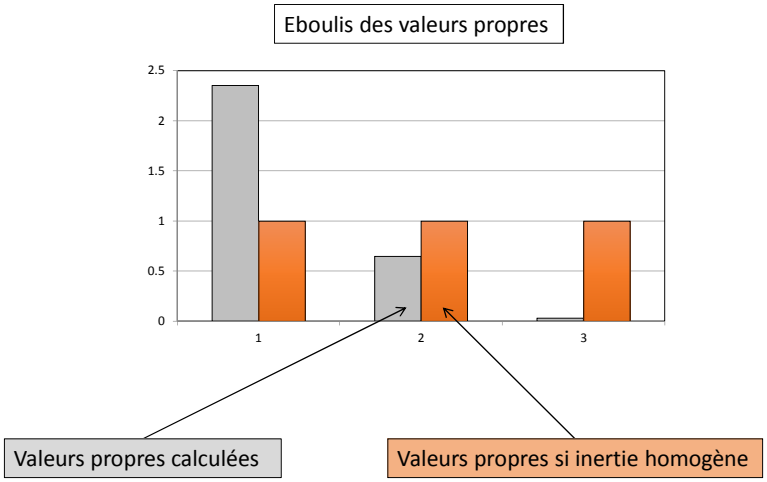
Les trois axes d'inerties sont associées aux valeurs propres
U1=2,3525 étant la première, la plus informative et détermine l'axe 1
U2=0,6474 étant la seconde et détermine l' axe 2 orthogonal au premier
U3=3,00 e-15 …. Ici insignifiant

La composante **c1** est le vecteur renfermant les coordonnées des projections des individus sur l'axe 1.

La composante **c2** est le vecteur renfermant les coordonnées des projections des individus sur l'axe 2.

Elles ne sont pas corrélées entre elles

Analyse en composantes principales (ACP)



Analyse en composantes principales (ACP)

- Nuage des individus
  - Pour obtenir les coordonnées des individus dans le plan factoriel, on pose chaque composante principale comme une combinaison linéaire des variables initiales

$$C^1 = U_1^1 X^1 + U_2^1 X^2 + U_3^1 X^3$$

ou plus généralement

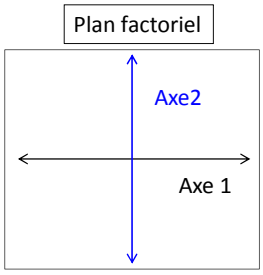
$$C^1 = U_1^1 X^1 + U_2^1 X^2 + ... + U_p^1 X^p$$

  - On se retrouve dans un environnement proche de la régression multiple sauf qu'ici les composantes synthétisent une ou plusieurs variables

Analyse en composantes principales (ACP)

- Représentation des individus
  - Les 2 premières composantes définissent un plan.
    ⇒Plan factoriel
  - Les axes factoriels 1 et 2 représentent le plus d'inertie du nuage de points

	1
1	0,4548
2	-0,6319
3	-0,6274

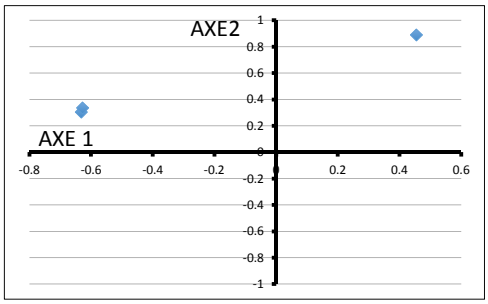


  - La première composante principale fournit les coordonnées des individus sur l'axe 1

Analyse en composantes principales (ACP)

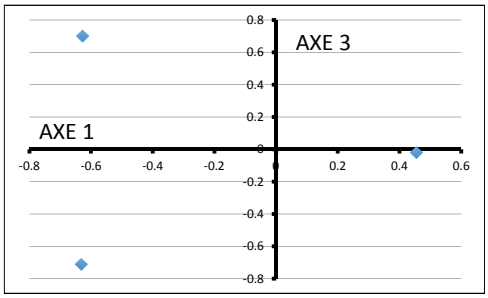
- Représentation des individus
  - Individus sur le 1er plan factoriel

	1	2
1	0,4548	0,8902
2	-0,6319	0,3057
3	-0,6274	0,3374



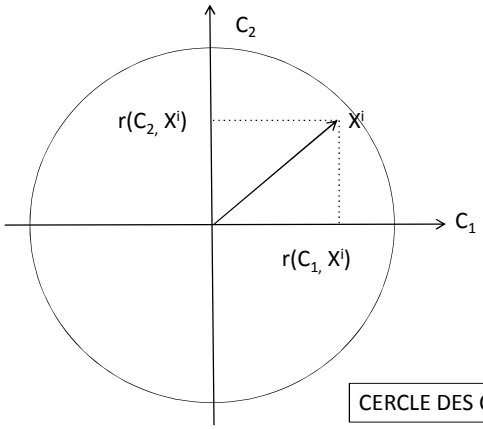
  - Individus sur le 2è plan factoriel

	1	3
1	0,4548	-0,0213
2	-0,6319	-0,7121
3	-0,6274	0,7001



Analyse en composantes principales (ACP)

- Représentation des variables
  - Les « proximités » entre les composantes principales et les variables initiales sont mesurées par les covariances et surtout **les corrélations**.



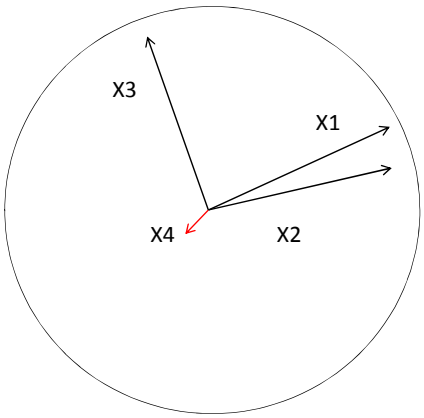
CERCLE DES CORRELATIONS

Analyse en composantes principales (ACP)

- Représentation des variables

X1 et X2 ont un r proche de 1
X1 et X3 ont un r=0

X4 n'appartient pas au plan (1,2)



Deux choses à regarder

1) La longueur des flèches
2) Les angles entre les flèches

Analyse en composantes principales (ACP)

- Exemple concret

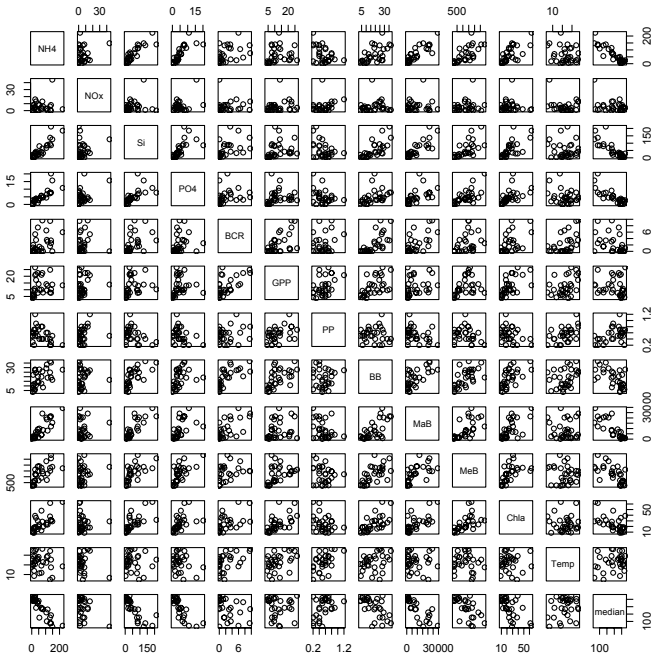
	NH4	NOx	Si	PO4	BCR	GPP	PP	BB	MaB	MeB	Chla	Temp	median
SA1	148	44	125.8	15.7	0.00	13.8	0.5	16.5	15548	1849	28.96	8.30	64.70
SA2	139	8	85.0	20.2	2.55	7.6	0.2	19.1	11931	2863	30.57	13.70	87.37
SA3	105	3	66.7	4.4	6.51	14.1	0.6	29.1	20263	2329	23.43	16.10	161.17
SA4	109	2	84.0	6.6	3.56	9.5	0.4	33.2	18604	716	26.70	19.00	100.37
SA5	139	0	212.3	7.7	3.40	10.2	0.3	35.5	21473	2578	31.67	17.70	141.94
SA6	76	13	63.0	4.7	9.63	22.4	0.8	25.7	24107	1471	28.40	23.30	183.01
SA7	44	9	35.9	3.0	7.42	22.4	0.8	20.4	6731	1451	29.68	18.25	230.58
SA8	133	1	136.5	7.8	9.45	24.8	0.4	37.0	21418	1500	61.63	22.15	87.37
SA9	223	2	184.6	10.8	5.98	13.2	0.2	27.6	29290	1729	63.62	14.20	70.15
SA10	132	6	93.4	7.2	1.62	7.9	0.2	24.7	20458	1757	37.78	7.30	126.45
SA11	55	8	79.6	4	1.44	9.9	0.4	27.9	9854	1899	23.54	10.60	125.00
SB2	81	6	31.7	4.5	0.00	7.8	0.4	8.4	8570	1265	19.31	16.40	183.01
SB3	43	0	17.3	1.0	0.48	7.6	0.5	14.7	9158	1879	15.27	17.70	222.72
SB4	49	0	25.3	2.7	0.00	14.2	0.7	13.7	1642	685	19.74	19.10	238.71
SB5	114	0	88.0	9.0	3.03	12.9	0.6	18.9	7945	1256	23.36	19.60	180.91
SB6	31	13	41.2	3.8	3.55	17.2	0.9	24.3	4079	1163	18.11	23.10	191.67
SB7	25	16	32.3	2.9	5.33	20.8	1.2	27.0	2537	1211	17.89	22.30	235.97
SB8	24	4	28.8	2.7	0.96	26.6	0.7	27.9	792	1604	38.91	19.60	274.21
SB9	30	1	28.8	3.8	0.62	21.4	0.6	10.7	6680	1200	35.22	17.15	244.29
SB10	70	2	40.2	5.5	0.95	21.0	0.4	19.9	4192	1774	47.93	10.90	191.67
SB11	69	4	20.1	5.75	0.00	10.0	0.8	12.6	3697	905	12.92	14.60	189.46
SC2	17	8	2.0	2.0	0.00	6.1	0.7	4.2	2475	265	8.69	9.20	238.71
SC3	7	0	20.3	0.0	0.00	5.4	0.5	3.2	2515	274	12.00	15.40	238.71
SC4	6	0	4.0	0.0	0.00	3.6	0.3	10.2	1041	266	10.44	18.10	247.13
SC5	9	0	4.7	0.0	0.00	6.2	0.6	11.2	1831	274	9.59	14.80	258.82
SC6	9	0	8.0	0.5	1.18	5.0	0.6	7.5	1207	508	9.02	19.45	250.00
SC7	10	8	19.8	2.9	2.01	9.6	0.6	22.0	1583	564	15.91	22.30	252.90
SC8	20	5	9.9	2.4	0.00	10.7	0.7	12.2	1176	638	16.12	16.70	258.82
SC9	23	4	8.8	2.9	0.00	9.2	0.5	5.9	1234	671	18.82	15.30	258.82
SC10	17	3	3.9	1.7	0.00	4.3	0.2	6.1	800	253	18.04	11.90	267.94
SC11	0	4	8.5	0.9	0.00	5.8	0.5	7.9	78	218	11.61	14.00	244.29

ACP sous R
library(ade4)
acp1=dudi.pca(donnees[,1:13], scale=T, center=T)

Analyse en composantes principales (ACP)

- Représentation des corrélations entre les variables

> pairs(donnees)



Analyse en composantes principales (ACP)

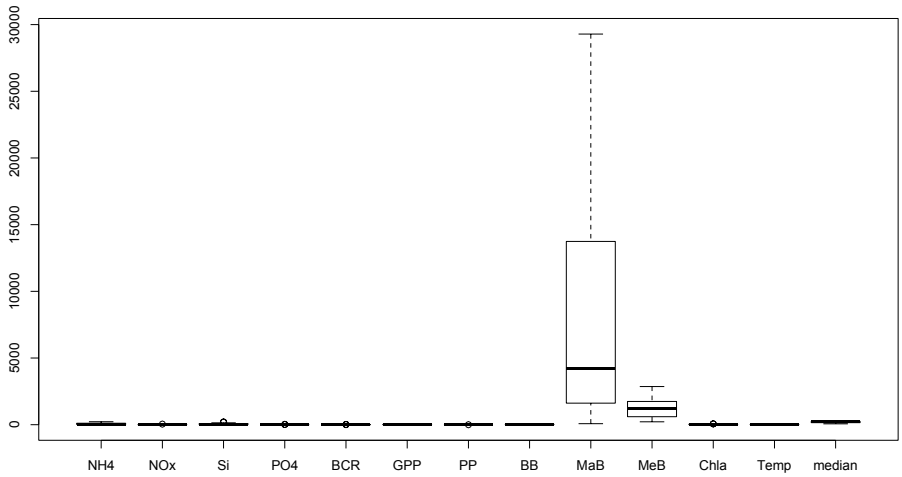
- Corrélations entre les variables

```
> cor(donnees)
             NH4          NOx           Si          PO4          BCR          GPP           PP           BB          MaB          MeB
NH4    1.0000000 0.190034615 0.89158493 0.8147979 0.45022704 0.1766309 -0.44951081 0.58992709 0.87141820 0.69613723
NOx    0.1900346 1.000000000 0.16718816 0.4360960 0.06138134 0.1965242  0.24267608 0.08102474 0.13213818 0.21240058
Si     0.8915849 0.167188156 1.00000000 0.6806899 0.49562210 0.2230915 -0.40270995 0.71352787 0.84521901 0.67745594
PO4    0.8147979 0.436096030 0.68068987 1.0000000 0.24575658 0.1420786 -0.35384805 0.39141097 0.57145991 0.68167401
BCR    0.4502270 0.061381337 0.49562210 0.2457566 1.00000000 0.5971436  0.18093662 0.71242348 0.65803137 0.41872362
GPP    0.1766309 0.196524187 0.22309148 0.1420786 0.59714356 1.0000000  0.44765011 0.55142139 0.22961079 0.38026791
PP    -0.4495108 0.242676081 -0.40270995 -0.3538481 0.18093662 0.4476501 1.00000000 -0.01443931 -0.36873293 -0.21989232
BB     0.5899271 0.081024741 0.71352787 0.3914110 0.71242348 0.5514214 -0.01443931 1.00000000 0.68212566 0.64037894
MaB    0.8714182 0.132138185 0.84521901 0.5714599 0.65803137 0.2296108 -0.36873293 0.68212566 1.00000000 0.64238970
MeB    0.6961372 0.212400576 0.67745594 0.6816740 0.41872362 0.3802679 -0.21989232 0.64037894 0.64238970 1.00000000
Chla   0.7183683 0.005312254 0.69731715 0.5234238 0.53290820 0.6072144 -0.35211821 0.63360776 0.66388813 0.57215458
Temp  -0.1950368 -0.208751001 -0.06314693 -0.2438825 0.48675818 0.4171393 0.50674144 0.31532499 -0.03933945 -0.09406076
median -0.9063454 -0.331994671 -0.83179224 -0.8153285 -0.42247367 -0.1092259 0.43531935 -0.60481962 -0.82658321 -0.65672849
            Chla          Temp       median
NH4    0.718368305 -0.19503684 -0.9063454
NOx    0.005312254 -0.20875100 -0.3319947
Si     0.697317152 -0.06314693 -0.8317922
PO4    0.523423831 -0.24388250 -0.8153285
BCR    0.532908205  0.48675818 -0.4224737
GPP    0.607214377  0.41713927 -0.1092259
PP    -0.352118214  0.50674144  0.4353194
BB     0.633607760  0.31532499 -0.6048196
MaB    0.663888126 -0.03933945 -0.8265832
MeB    0.572154580 -0.09406076 -0.6567285
Chla   1.000000000 -0.03325290 -0.6089094
Temp  -0.033252901  1.00000000  0.2377736
median -0.608909416  0.23777362  1.0000000
```

Analyse en composantes principales (ACP)

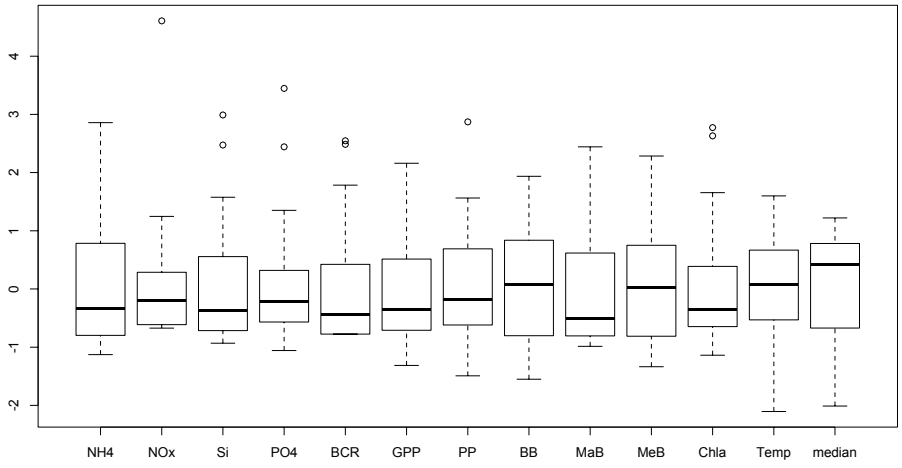
- Commande boxplot : visualiser l'hétérogénéité des données

```
> boxplot(donnees)
```



Analyse en composantes principales (ACP)

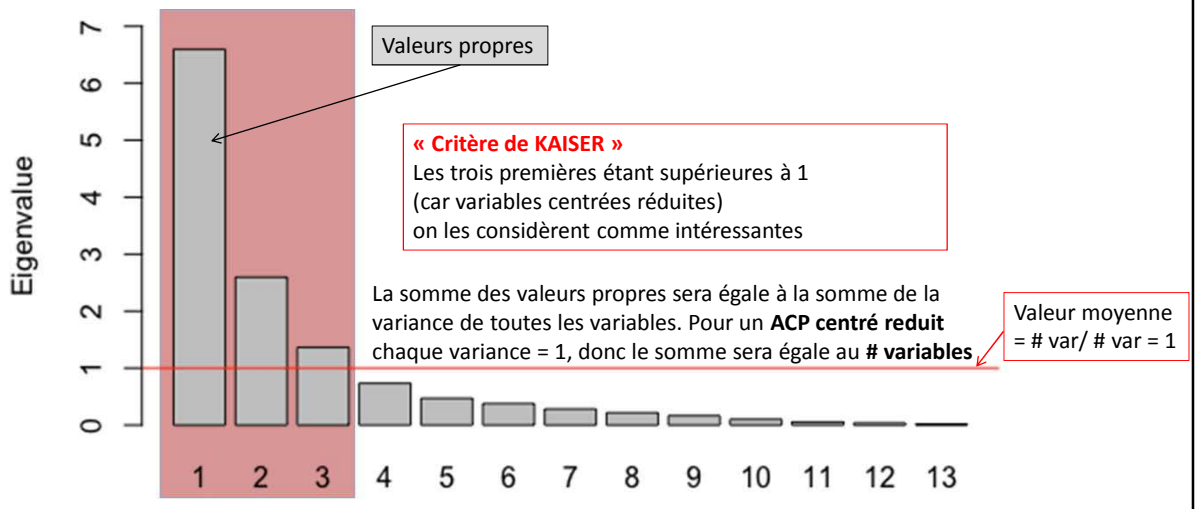
- Centrage et réduction : effacer l'hétérogénéité des données
  - donnees_CR <- apply(donnees,2,function(x){x-mean(x)}) # Centre
  - donnees_CR <- apply(donnees_CR,2,function(x){x/sd(x)}) # Reduction
  - boxplot(donnees_CR)



Analyse en composantes principales (ACP)

- Eboulis des valeurs propres

```
> barplot(acp1$eig, names.arg=c(1:length(acp1$eig)))
> abline(h=sum(acp1$eig/length(acp1$eig)),col="red")
```



Valeurs propres

« Critère de KAISER »
Les trois premières étant supérieures à 1
(car variables centrées réduites)
on les considèrent comme intéressantes

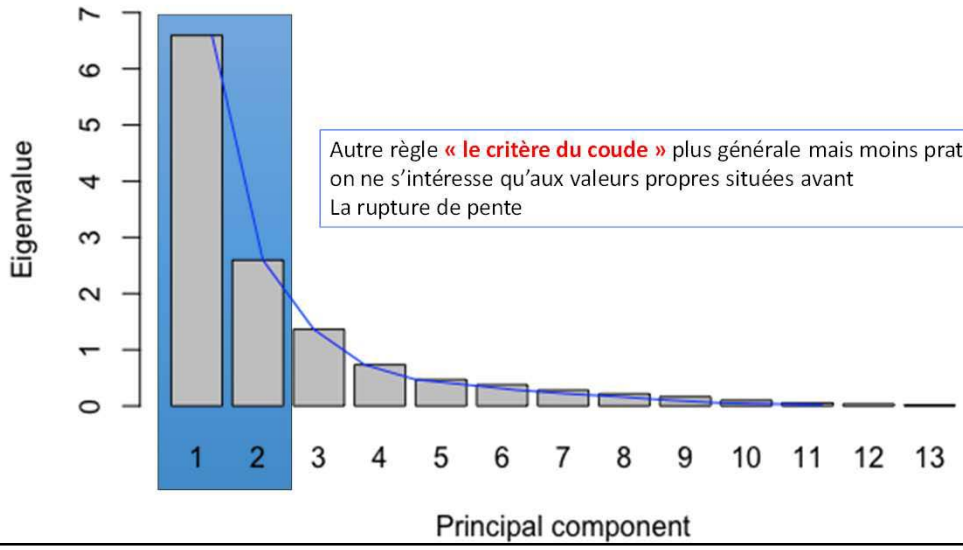
La somme des valeurs propres sera égale à la somme de la variance de toutes les variables. Pour un **ACP centré reduit** chaque variance = 1, donc le somme sera égale au **# variables**

Valeur moyenne
= # var/ # var = 1

Analyse en composantes principales (ACP)

- Eboulis des valeurs propres

```
> barplot(acp1$eig, names.arg=c(1:length(acp1$eig)))
> lines(acp1$eig, col="blue")
```



Autre règle **« le critère du coude »** plus générale mais moins pratique
on ne s'intéresse qu'aux valeurs propres situées avant
La rupture de pente

Analyse en composantes principales (ACP)

- Eboulis des valeurs propres

```
>pourc=round((acp1$eig/sum(acp1$eig))*100,2)
>pourc
>cumsum(pourc)
```

En ramenant à 100% de 13 (somme des valeurs propres
On obtient la part de l'inertie expliquée par les composantes
Ici

C1 = 50,72
C2= 20
C3=10,5

Ce qui donne pour le plan (1,2) plus de 70% de la variance de la Matrice d'origine expliquée.

PC	Explained variance	Cumulative exp. var
1	50.72	50.72
2	19.97	70.69
3	10.48	81.17
4	5.63	86.80
5	3.61	90.41
6	2.90	93.31
7	2.17	95.48
8	1.66	97.14
9	1.26	98.40
10	0.78	99.18
11	0.42	99.60
12	0.27	99.87
13	0.14	100.00

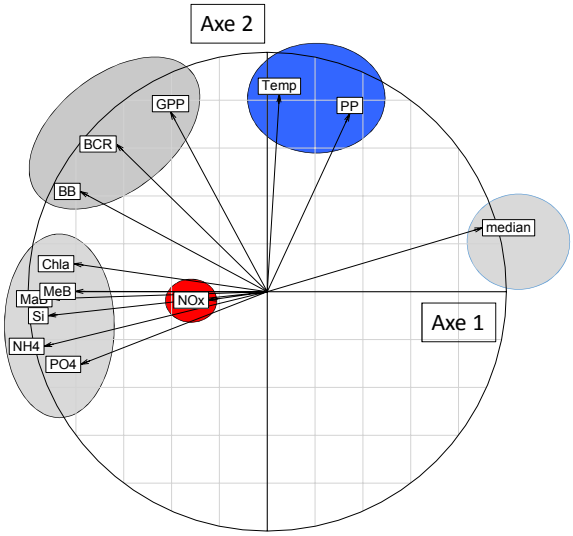
Analyse en composantes principales (ACP)

- Représentation des variables

> s.corcircle(acp1$co)
> acp1$co[,1:2]

R_i^2 = variance explique (entre PCi (Axe i) et une variable)
Quand dim(acp) = # variable (p), tt l'information d'une variable est distribue sur le p composant principal. Le somme de R_i^2 pour tt composant principal devrais donc expliques tt la variance (100%) de cette variable (cf. cours 2)

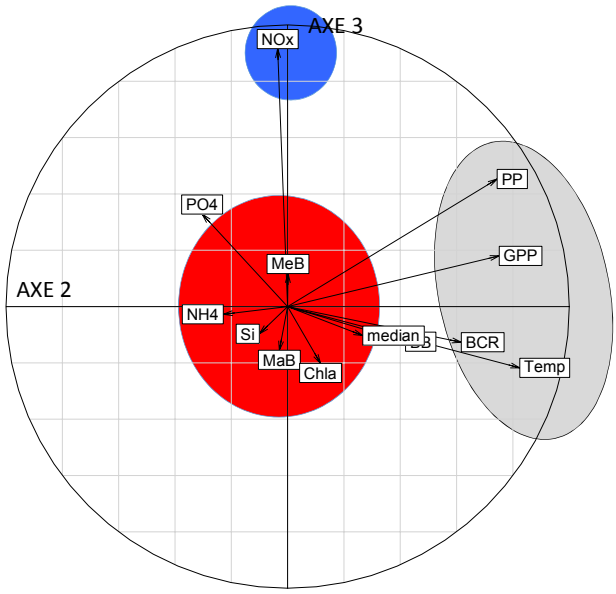
> sum(acp1$co[1,]^2) # =1



Analyse en composantes principales (ACP)

- Représentation des variables

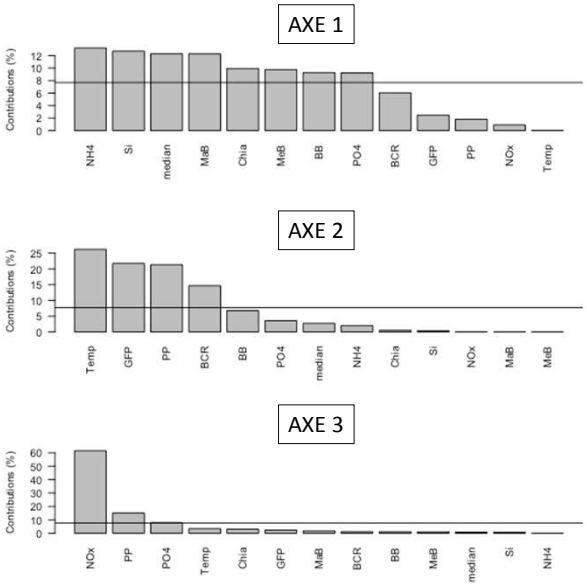
> s.corcircle(acp1$co, xax=2, yax=3)



Analyse en composantes principales (ACP)

- ctribution des variables aux axes

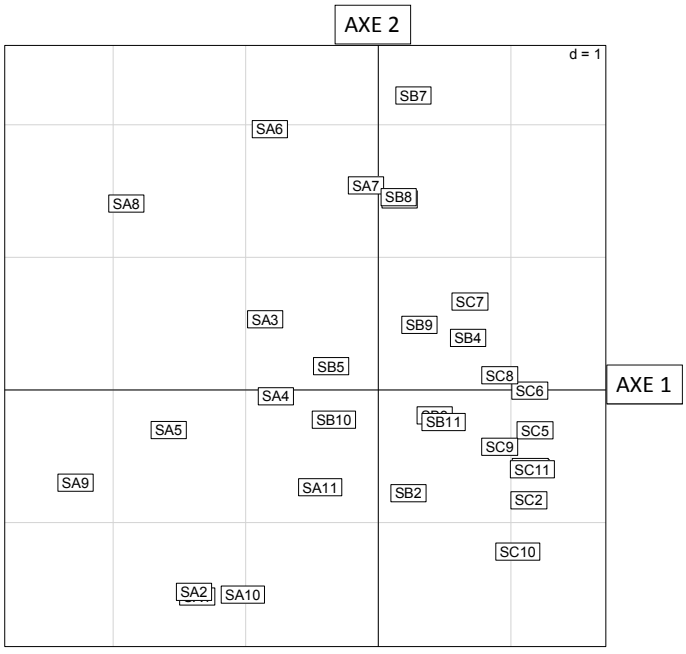
```
# For loop going through axis 1 to 3.
par(mfrow=c(3,1))
for (i in 1:3){
 # Calculate variable contribution
 var_ctrib <- acp1$co[,i]^2/sum(acp1$co[,i]^2)*100
 var_ctrib <- as.data.frame(var_ctrib)
 var_ctrib$var <- row.names(acp1$co)
 colnames(var_ctrib) <- c("Axis_ctrib","var")
 # Order descending
 var_ctrib <- var_ctrib[order(var_ctrib$Axis_ctrib,
          decreasing=TRUE),]
 var_ctrib <- as.data.frame(var_ctrib)
 # Plot data
 barplot(var_ctrib$Axis_ctrib,names.arg =
          var_ctrib$var,las=2, ylab="contributions (%)",
          main=paste("Axis ", i,sep=""))
 abline(h=100/length(acp1$eig))
}
```



Analyse en composantes principales (ACP)

- Représentation des individus

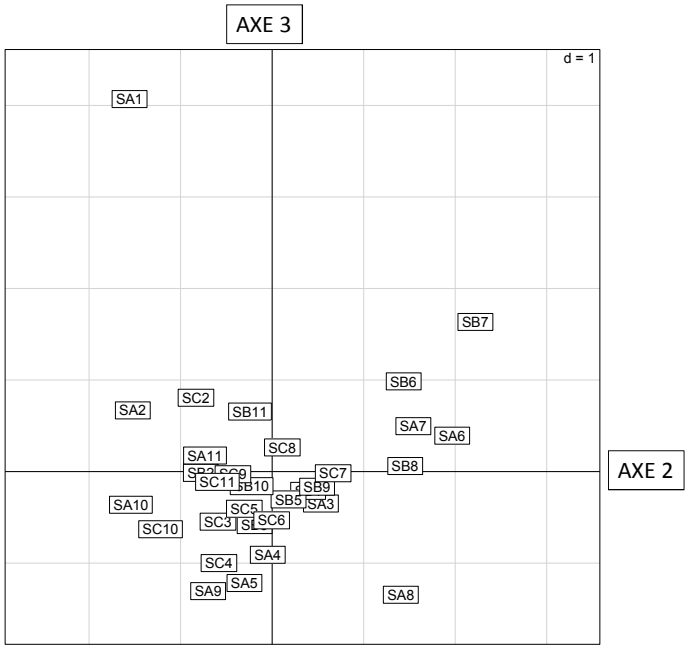
>s.label(acp1$li[,1:2])



Analyse en composantes principales (ACP)

- Représentation des individus

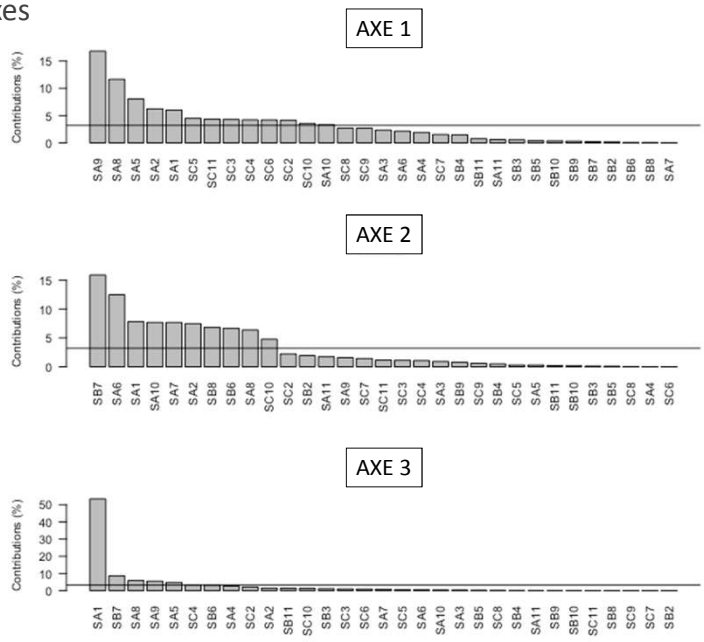
>s.label(acp1$li[,2:3])



Analyse en composantes principales (ACP)

- ctribution des individus aux axes

```
# For loop going through axis 1 to 3.
par(mfrow=c(3,1))
for (i in 1:3){
  # Calculate sample contribution
  obs_ctrib <- acp1$li[,i]^2/sum(acp1$li[,i]^2)*100
  obs_ctrib <- as.data.frame(obs_ctrib)
  obs_ctrib$var <- row.names(acp1$li)
  colnames(obs_ctrib) <- c("Axis_ctrib","obs")
  # Order descending
  obs_ctrib <- obs_ctrib[order(obs_ctrib$Axis_ctrib,
          decreasing=TRUE),]
  obs_ctrib <- as.data.frame(obs_ctrib)
  # Plot data
  barplot(obs_ctrib$Axis_ctrib,names.arg =
          obs_ctrib$obs, las=2, ylab="contributions
          (%)", main=paste("Axis ", i, sep=""))
  abline(h=100/length(acp1$li[,1]))
}
```



Analyse en composantes principales (ACP)

```
donnees$station <- as.factor(substr(row.names(donnees),1,2))
names(donnees)
[1] "NH4"   "NOx"   "Si"    "PO4"   "BCR"   "GPP"   "PP"    "BB"    "MaB"   "MeB"   "Chla"  "Temp"  "median" "station"
```

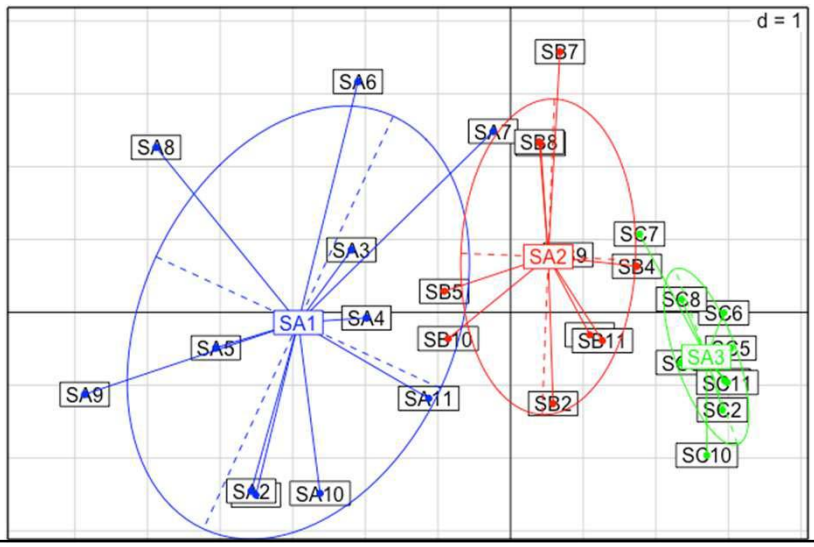
	NH4	NOx	Si	PO4	BCR	GPP	PP	BB	MaB	MeB	Chla	Temp	median	station
SA1	148	44	125.8	15.7	0.00	13.8	0.5	16.5	15548	1849	28.96	8.30	64.70	SA
SA2	139	8	85.0	20.2	2.55	7.6	0.2	19.1	11931	2863	30.57	13.70	87.37	SA
SA3	105	3	66.7	4.4	6.51	14.1	0.6	29.1	20263	2329	23.43	16.10	161.17	SA
SA4	109	2	84.0	6.6	3.56	9.5	0.4	33.2	18604	716	26.70	19.00	100.37	SA
SA5	139	0	212.3	7.7	3.40	10.2	0.3	35.5	21473	2578	31.67	17.70	141.94	SA
SA6	76	13	63.0	4.7	9.63	22.4	0.8	25.7	24107	1471	28.40	23.30	183.01	SA
SA7	44	9	35.9	3.0	7.42	22.4	0.8	20.4	6731	1451	29.68	18.25	230.58	SA
SA8	133	1	136.5	7.8	9.45	24.8	0.4	37.0	21418	1500	61.63	22.15	87.37	SA
SA9	223	2	184.6	10.8	5.98	13.2	0.2	27.6	29290	1729	63.62	14.20	70.15	SA
SA10	132	6	93.4	7.2	1.62	7.9	0.2	24.7	20458	1757	37.78	7.30	126.45	SA
SA11	55	8	79.6	4	1.44	9.9	0.4	27.9	9854	1899	23.54	10.60	125.00	SA
SB2	81	6	31.7	4.5	0.00	7.8	0.4	8.4	8570	1265	19.31	16.40	183.01	SB
SB3	43	0	17.3	1.0	0.48	7.6	0.5	14.7	9158	1879	15.27	17.70	222.72	SB
SB4	49	0	25.3	2.7	0.00	14.2	0.7	13.7	1642	685	19.74	19.10	238.71	SB
SB5	114	0	88.0	9.0	3.03	12.9	0.6	18.9	7945	1256	23.36	19.60	180.91	SB
SB6	31	13	41.2	3.8	3.55	17.2	0.9	24.3	4079	1163	18.11	23.10	191.67	SB
SB7	25	16	32.3	2.9	5.33	20.8	1.2	27.0	2537	1211	17.89	22.30	235.97	SB
SB8	24	4	28.8	2.7	0.96	26.6	0.7	27.9	792	1604	38.91	19.60	274.21	SB
SB9	30	1	28.8	3.8	0.62	21.4	0.6	10.7	6680	1200	35.22	17.15	244.29	SB
SB10	70	2	40.2	5.5	0.95	21.0	0.4	19.9	4192	1774	47.93	10.90	191.67	SB
SB11	69	4	20.1	5.75	0.00	10.0	0.8	12.6	3697	905	12.92	14.60	189.46	SB
SC2	17	8	2.0	2.0	0.00	6.1	0.7	4.2	2475	265	8.69	9.20	238.71	SC
SC3	7	0	20.3	0.0	0.00	5.4	0.5	3.2	2515	274	12.00	15.40	238.71	SC
SC4	6	0	4.0	0.0	0.00	3.6	0.3	10.2	1041	266	10.44	18.10	247.13	SC
SC5	9	0	4.7	0.0	0.00	6.2	0.6	11.2	1831	274	9.59	14.80	258.82	SC
SC6	9	0	8.0	0.5	1.18	5.0	0.6	7.5	1207	508	9.02	19.45	250.00	SC
SC7	10	8	19.8	2.9	2.01	9.6	0.6	22.0	1583	564	15.91	22.30	252.90	SC
SC8	20	5	9.9	2.4	0.00	10.7	0.7	12.2	1176	638	16.12	16.70	258.82	SC
SC9	23	4	8.8	2.9	0.00	9.2	0.5	5.9	1234	671	18.82	15.30	258.82	SC
SC10	17	3	3.9	1.7	0.00	4.3	0.2	6.1	800	253	18.04	11.90	267.94	SC
SC11	0	4	8.5	0.9	0.00	5.8	0.5	7.9	78	218	11.61	14.00	244.29	SC

Le nom des stations intervient ici comme variable de regroupement et prend la classe « factor »

Analyse en composantes principales (ACP)

ce qui permet d'utiliser une commande très utile « s.class »

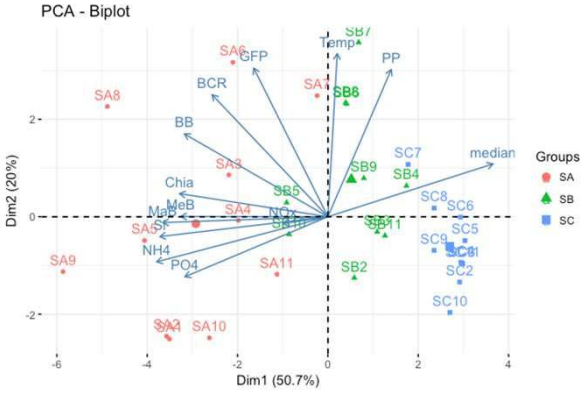
```
s.label(acp1$li)
s.class(acp1$li, fac=donnees$station, xax = 1, yax = 2, label = row.names(acp1$li), col=c("blue", "red", "green"),
add.plot = TRUE)
```



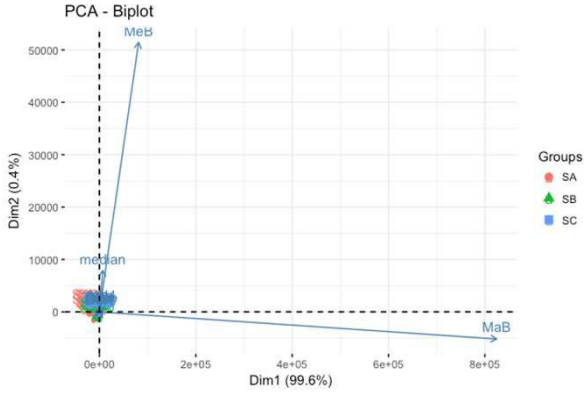
Analyse en composantes principales (ACP)

Après transformation on peut représenter les deux éléments (Variables et Individus) dans le même plan…
> biplot(acp1)
Ou
> fviz_pca_biplot(acp1) # du package « factoextra »

center=T, scale=T



center=F, scale=F

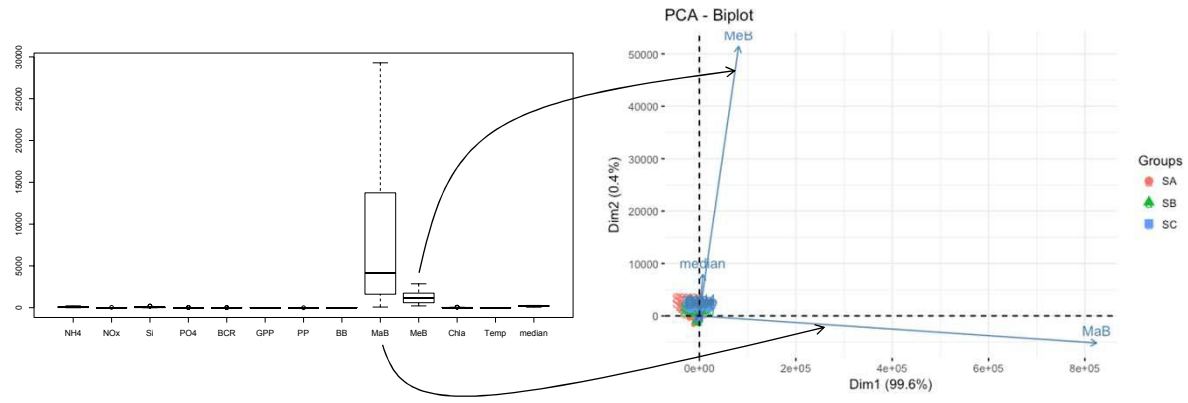


QUE REMARQUEZ VOUS sur ces deux résultats d'ACP issus du même tableau de données ?

Analyse en composantes principales (ACP)

Après transformation on peut représenter les deux éléments (Variables et Individus) dans le même plan…
> biplot(acp1)
Ou
> fviz_pca_biplot(acp1) # du package « factoextra »

center=F, scale=F



Non centre-réduit -> Les axes principaux sont les deux variables avec le plus de variance!

Adresses utiles pour aller plus loin

pbil.univ-lyon1.fr/R/enseignement.html

(site culte de l'université de Lyon, surtout pour les analyses multivariées sous R)

www.bio.umontreal.ca/legendre/

(le site du Maître des statistiques, plein de fonctions R à télécharger)

perso.univ-rennes1.fr/denis.poinsot/

(télécharger les deux doc statistiques, un classique et très pédagogique)