

Analyse de Données

CTD1 : Analyse en Composantes Principales
(ACP)

Double objectif en Analyse de Données

- 1) s'initier à l'analyse de données selon 2 approches :
 - 1) méthodes descriptives
 - 2) méthodes inférentielles
- 2) comprendre et utiliser les synergies entre le calcul scientifique et l'analyse de données (**SVD, valeurs propres**)

sujet du CTD1 : une méthode descriptive pour l'exploration de données **multivariées / multidimensionnelles**, l'ACP

- on veut décrire les données à travers leur **visualisation** et/ou le calcul de résumés numériques.
- ces résumés regroupent des **facteurs** qui **expliquent** la structure (parfois cachée) des données (clusters, ...)
- on parle d'**analyse factorielle** : la méthode clé, l'ACP, Analyse en Composantes Principales

Exemples de données manipulées

- Écologie : concentration du **polluant j** dans la **rivière i**
- Économie : valeur de l'**indicateur j** pour l'**année i**
- Génétique : expression du **gène j** pour le **patient i**
- Biologie : **mesure j** pour l'**animal i**
- Marketing : valeur de l'**indice de satisfaction j** pour la **marque i**
- Sociologie : **temps passé à l'activité j** par les individus de la **CSP i**
- imagerie : intensité lumineuse dans la **couleur j** du **pixel i**
- etc

tableau de données => X de taille (n, p), n individus (lignes), p variables (colonnes)

Un exemple de données en 12D

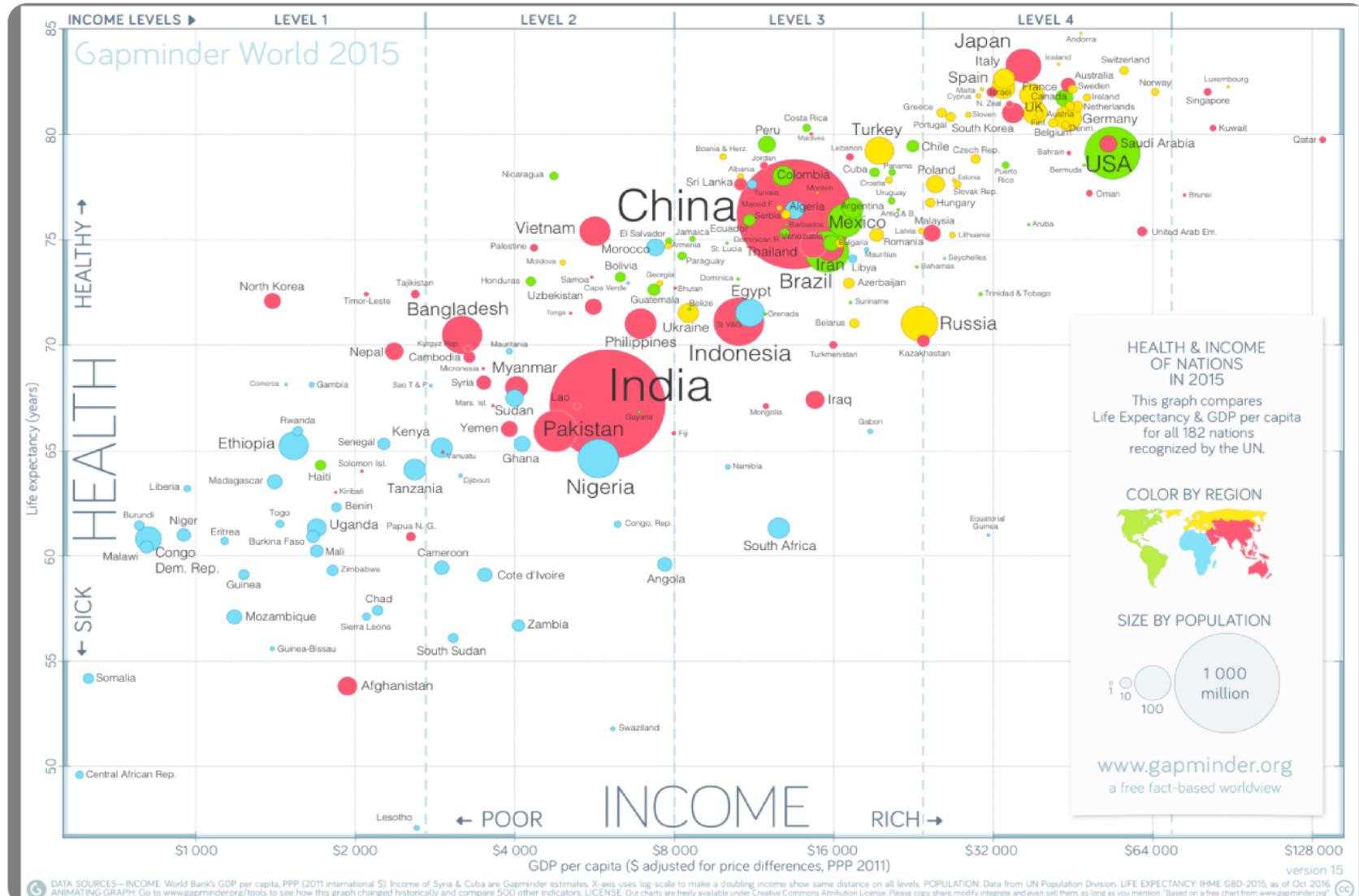
- 15 individus (lignes) : villes de France
- 14 variables (colonnes) :
 - 12 températures mensuelles moyennes (sur 30 ans)
 - 2 variables géométriques (latitude, longitude)

	Janv	Févr	Mars	Avri	Mai	Juin	juil	Août	Sept	Octo	Nove	Déce	Lati	Long
Bordeaux	5.6	6.6	10.3	12.8	15.8	19.3	20.9	21	18.6	13.8	9.1	6.2	44.5	-0.34
Brest	6.1	5.8	7.8	9.2	11.6	14.4	15.6	16	14.7	12	9	7	48.24	-4.29
Clermont	2.6	3.7	7.5	10.3	13.8	17.3	19.4	19.1	16.2	11.2	6.6	3.6	45.47	3.05
Grenoble	1.5	3.2	7.7	10.6	14.5	17.8	20.1	19.5	16.7	11.4	6.5	2.3	45.1	5.43
Lille	2.4	2.9	6	8.9	12.4	15.3	17.1	17.1	14.7	10.4	6.1	3.5	50.38	3.04
Lyon	2.1	3.3	7.7	10.9	14.9	18.5	20.7	20.1	16.9	11.4	6.7	3.1	45.45	4.51
Marseille	5.5	6.6	10	13	16.8	20.8	23.3	22.8	19.9	15	10.2	6.9	43.18	5.24
Montpellier	5.6	6.7	9.9	12.8	16.2	20.1	22.7	22.3	19.3	14.6	10	6.5	43.36	3.53
Nantes	5	5.3	8.4	10.8	13.9	17.2	18.8	18.6	16.4	12.2	8.2	5.5	47.13	-1.33
Nice	7.5	8.5	10.8	13.3	16.7	20.1	22.7	22.5	20.3	16	11.5	8.2	43.42	7.15
Paris	3.4	4.1	7.6	10.7	14.3	17.5	19.1	18.7	16	11.4	7.1	4.3	48.52	2.2
Rennes	4.8	5.3	7.9	10.1	13.1	16.2	17.9	17.8	15.7	11.6	7.8	5.4	48.05	-1.41
Strasbourg	0.4	1.5	5.6	9.8	14	17.2	19	18.3	15.1	9.5	4.9	1.3	48.35	7.45
Toulouse	4.7	5.6	9.2	11.6	14.9	18.7	20.9	20.9	18.3	13.3	8.6	5.5	43.36	1.26
Vichy	2.4	3.4	7.1	9.9	13.6	17.1	19.3	18.8	16	11	6.6	3.4	46.08	3.26

Problème : objectifs

- Représentation / Visualisation des données sous forme de graphiques simples
- Étude des individus
 - quand 2 individus se ressemblent au point de vue de l'ensemble des variables ?
 - peut-on faire un bilan des ressemblances ?
 - constructions de groupes d'individus
- Étude des variables
 - recherche de ressemblances entre les variables
 - on parle plutôt de liaisons
 - liaisons linéaires
 - coefficient de corrélation
 - matrice des corrélations
 - petit nombre d'indicateurs synthétiques pour résumer un grande nombres de variables

1) Visualisation de données



GapMinder

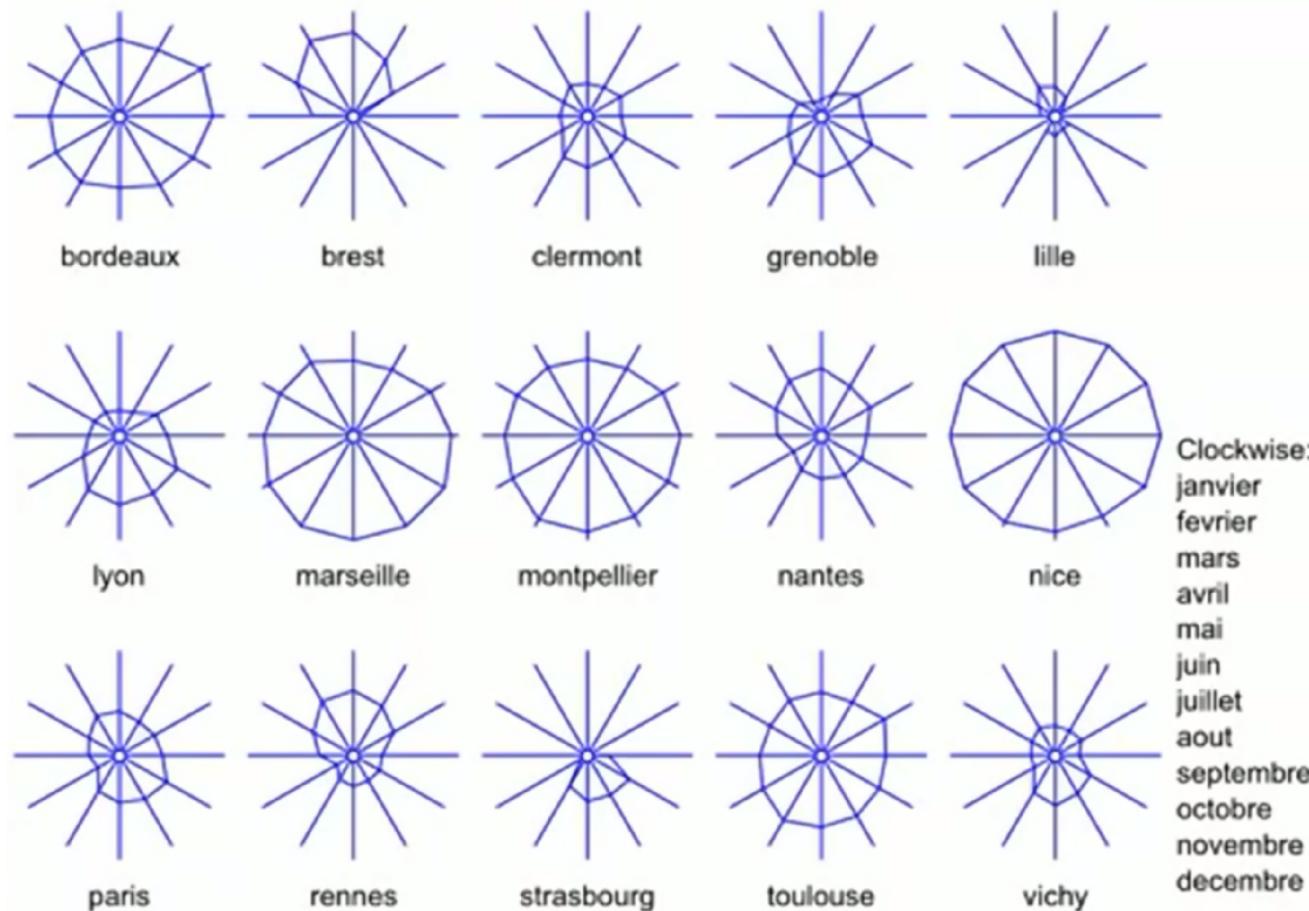
Un exemple de données en 12D

- 15 individus (lignes) : villes de France
- 14 variables (colonnes) :
 - 12 températures mensuelles moyennes (sur 30 ans)
 - 2 variables géométriques (latitude, longitude)

	Janv	Févr	Mars	Avri	Mai	Juin	juil	Août	Sept	Octo	Nove	Déce	Lati	Long
Bordeaux	5.6	6.6	10.3	12.8	15.8	19.3	20.9	21	18.6	13.8	9.1	6.2	44.5	-0.34
Brest	6.1	5.8	7.8	9.2	11.6	14.4	15.6	16	14.7	12	9	7	48.24	-4.29
Clermont	2.6	3.7	7.5	10.3	13.8	17.3	19.4	19.1	16.2	11.2	6.6	3.6	45.47	3.05
Grenoble	1.5	3.2	7.7	10.6	14.5	17.8	20.1	19.5	16.7	11.4	6.5	2.3	45.1	5.43
Lille	2.4	2.9	6	8.9	12.4	15.3	17.1	17.1	14.7	10.4	6.1	3.5	50.38	3.04
Lyon	2.1	3.3	7.7	10.9	14.9	18.5	20.7	20.1	16.9	11.4	6.7	3.1	45.45	4.51
Marseille	5.5	6.6	10	13	16.8	20.8	23.3	22.8	19.9	15	10.2	6.9	43.18	5.24
Montpellier	5.6	6.7	9.9	12.8	16.2	20.1	22.7	22.3	19.3	14.6	10	6.5	43.36	3.53
Nantes	5	5.3	8.4	10.8	13.9	17.2	18.8	18.6	16.4	12.2	8.2	5.5	47.13	-1.33
Nice	7.5	8.5	10.8	13.3	16.7	20.1	22.7	22.5	20.3	16	11.5	8.2	43.42	7.15
Paris	3.4	4.1	7.6	10.7	14.3	17.5	19.1	18.7	16	11.4	7.1	4.3	48.52	2.2
Rennes	4.8	5.3	7.9	10.1	13.1	16.2	17.9	17.8	15.7	11.6	7.8	5.4	48.05	-1.41
Strasbourg	0.4	1.5	5.6	9.8	14	17.2	19	18.3	15.1	9.5	4.9	1.3	48.35	7.45
Toulouse	4.7	5.6	9.2	11.6	14.9	18.7	20.9	20.9	18.3	13.3	8.6	5.5	43.36	1.26
Vichy	2.4	3.4	7.1	9.9	13.6	17.1	19.3	18.8	16	11	6.6	3.4	46.08	3.26

Représentation des données

- ... grâce à la corrélation entre mois successifs



Quid de variables non corrélées ?



Visualisation = projection en 2D



FIGURE: Quel animal ?

Visualisation = projection en 2D

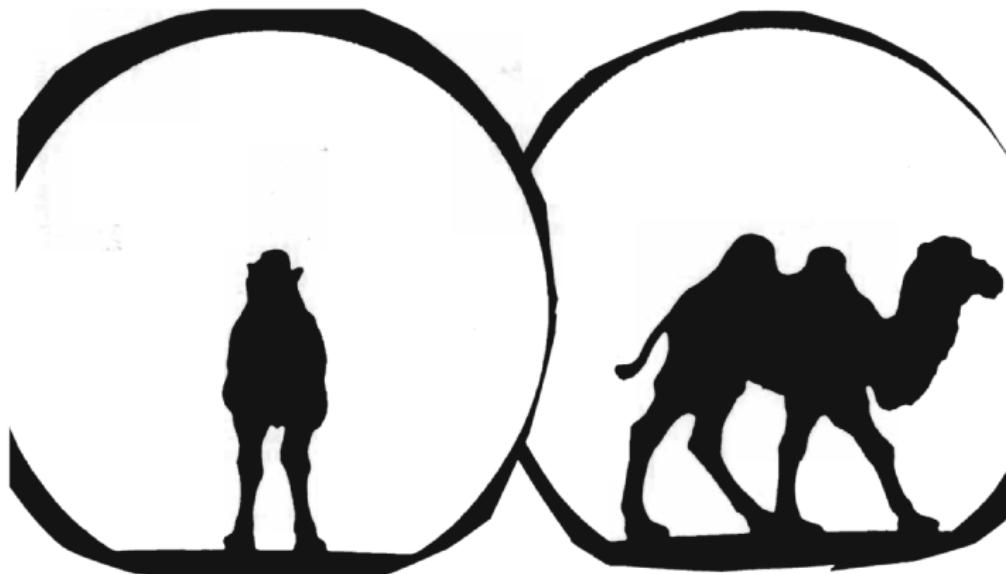
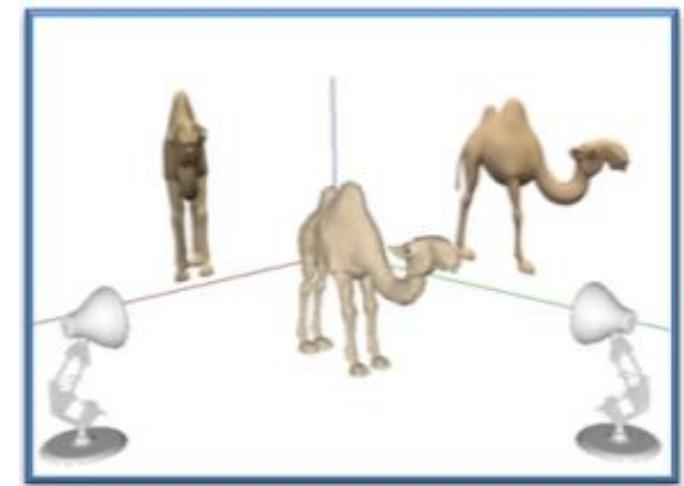
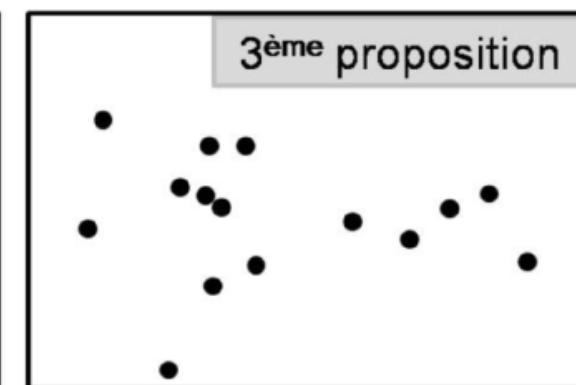
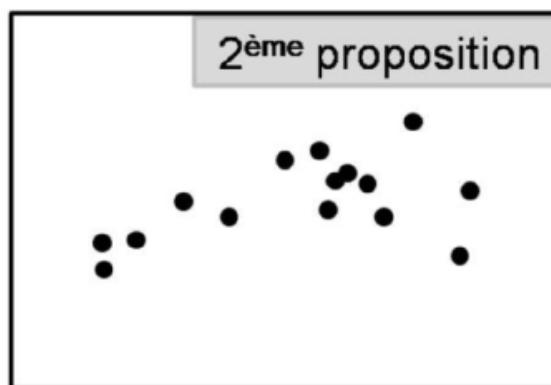
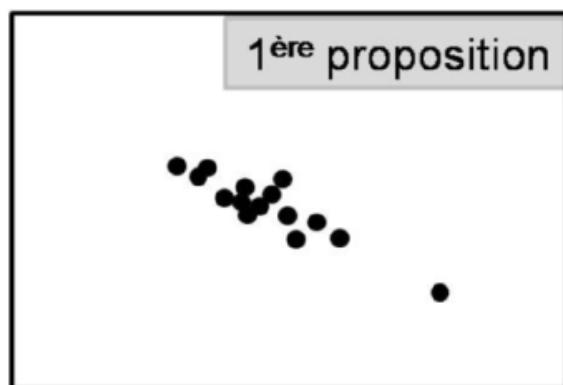
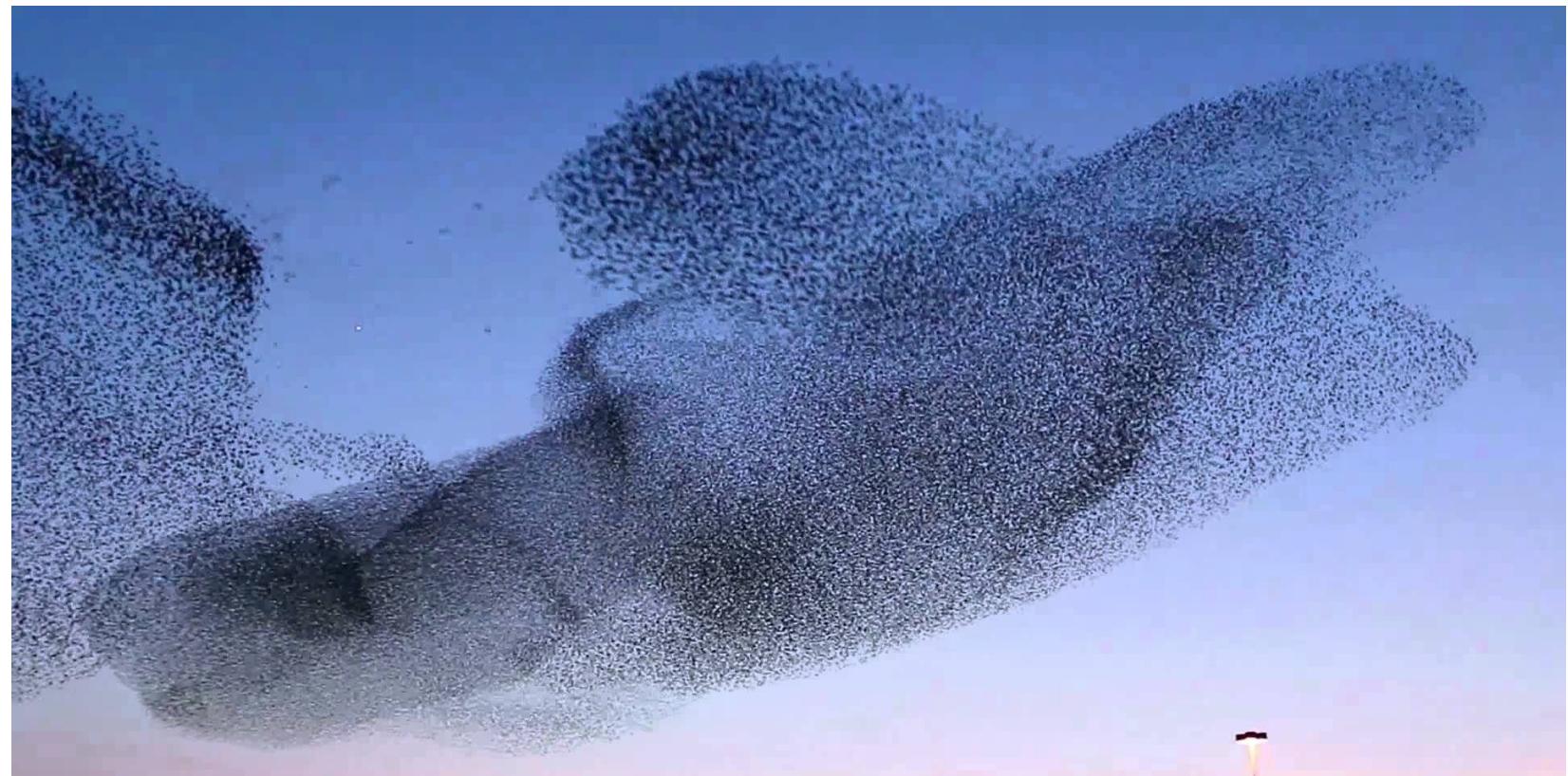


FIGURE: Quel animal ? (*illustration JP Fénelon*)

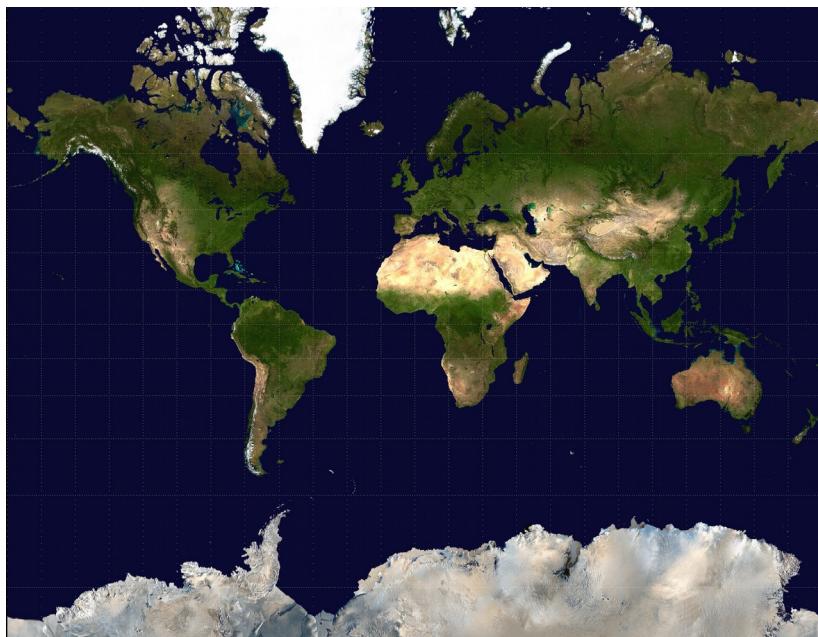


- exemple 3D matlab (separation.m)
- séparer les points revient à augmenter la dispersion

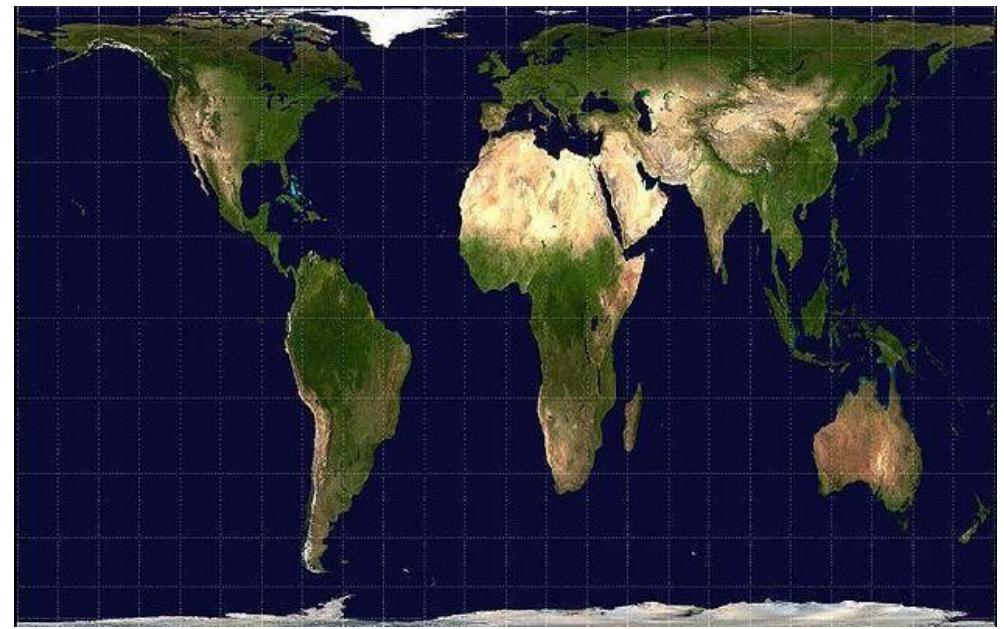
Visualisation = projection en 2D



projections 2D globe



projection de Mercator



projection Arno Peters

2) Réduction de dimension par ACP
(Analyse en Composantes Principales)

2.1) Principes de l'ACP

- méthode d'analyse factorielle sur des tableaux de données « individus - variables quantitatives »
- méthode descriptive, exploratoire
- n individus, p variables
- visualisation de données par des graphiques simples
- synthèse/résumé de grands tableaux (individus \times variables)
- interpréter les liaisons inter-variables, les similarités inter-individus
- choix d'effectuer l'analyse des individus
- réduction de données de p à $q << p$ dimensions ($q=2, q=3$)
- qualité ACP
 - restitue fidèlement le nuage de points
 - meilleure représentation de la diversité, de la variabilité
 - ne perturbe pas les distances entre individus

Objets manipulés

- tableau X [individus x variables] de dimension ($n \times p$)
 - espace des individus, distances entre individus (*figure*)
 - espace des variables, angles entre variables (*figure*)
-
- il faut centrer les données autour d'un individu moyen
 - il faut parfois réduire les données (unités différentes pour les variables)
-
- les individus se dispersent autour de l'origine/individu moyen selon des variances/covariances rangées dans une matrice :

la matrice de variance-covariance Sigma

2.2) Qu'est ce que la matrice de variance-covariance « empirique »

- rappels proba
- définition matricielle à partir du tableau/matrice des individus X
- que mesure-t-elle ?

un exercice : un point sur vos cerveaux...

On compte le nombre de cerveaux endormis en séances de TP à deux moments différents de la journée : 7h45-9h45 et 10h-12h.

	7h45-9h45	10h-12h
TP 1	6	2
TP 2	5	3
TP 3	6	1
TP 4	4	3
TP 5	4	1

1. Calculer la matrice de variance-covariance Σ . Y a-t-il une dépendance sur le taux de cerveaux endormis entre 7h45-9h45 et 10h-12h ? Expliquer la réponse.

- 1) approche « proba »
- 2) approche matricielle

Similarités entre individus

- délicat en présence de corrélations
 - quelle distance ? distance euclidienne ?
 - contre-exemple
- tenir compte des dépendances entre variables quand on veut mesurer la similarité inter-individus
 - distance qui utilise la matrice de variance-covariance empirique

Inertie du nuage de points

- dispersion des points autour de l'individu moyen
 - formulation avec la trace de sigma
- aussi appelée variance totale (généralisation de la variance à plusieurs dimensions)

qualité ACP

- restitue fidèlement le nuage de points
- meilleure représentation de la diversité, de la variabilité
- ne perturbe pas les distances entre individus

2.3) Analyse en q = 1 composante principale

- on cherche une droite sur laquelle on va projeter nos données (lien $X_c \leftrightarrow C_1$, schéma 2D)
- propriétés de la droite ? (*gif*)
- comment caractériser une droite ?
- formulation d'un problème sous contraintes faisant intervenir la matrice sigma

problème d'optimisation sous contrainte

- introduction du Lagrangien
- condition d'optimalité
- problème aux valeurs propres
- quelle couple propre choisir ?
- comment le calculer ?
- lien entre le vecteur principal et la composante principale

qualité de notre première composante principale

- variance de la première composante = plus grande valeur propre
- définition du contraste

2.4) extension à $q > 1$

- $q = 2$ plan, second vecteur propre (comment par rapport au premier ?)

3) Illustration de l'ACP / Interprétation des Composantes Principales

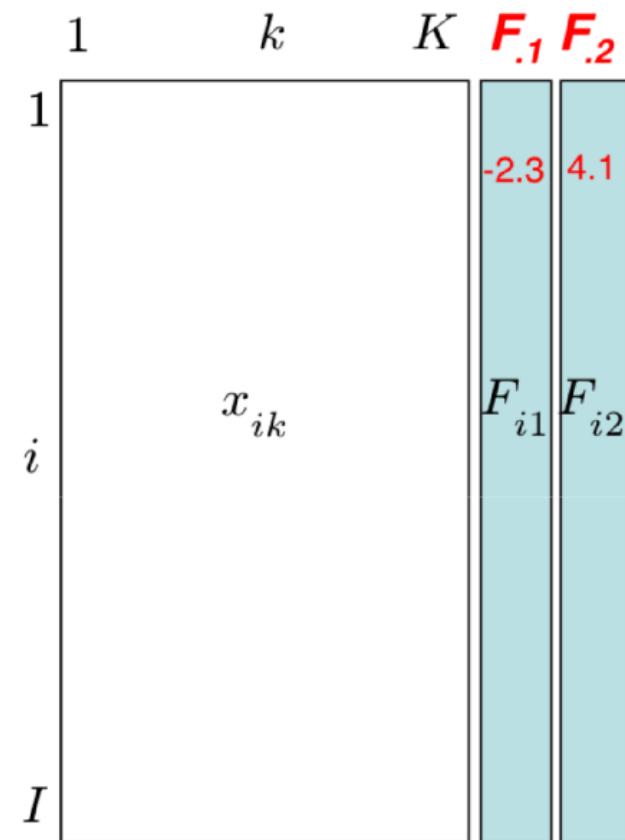
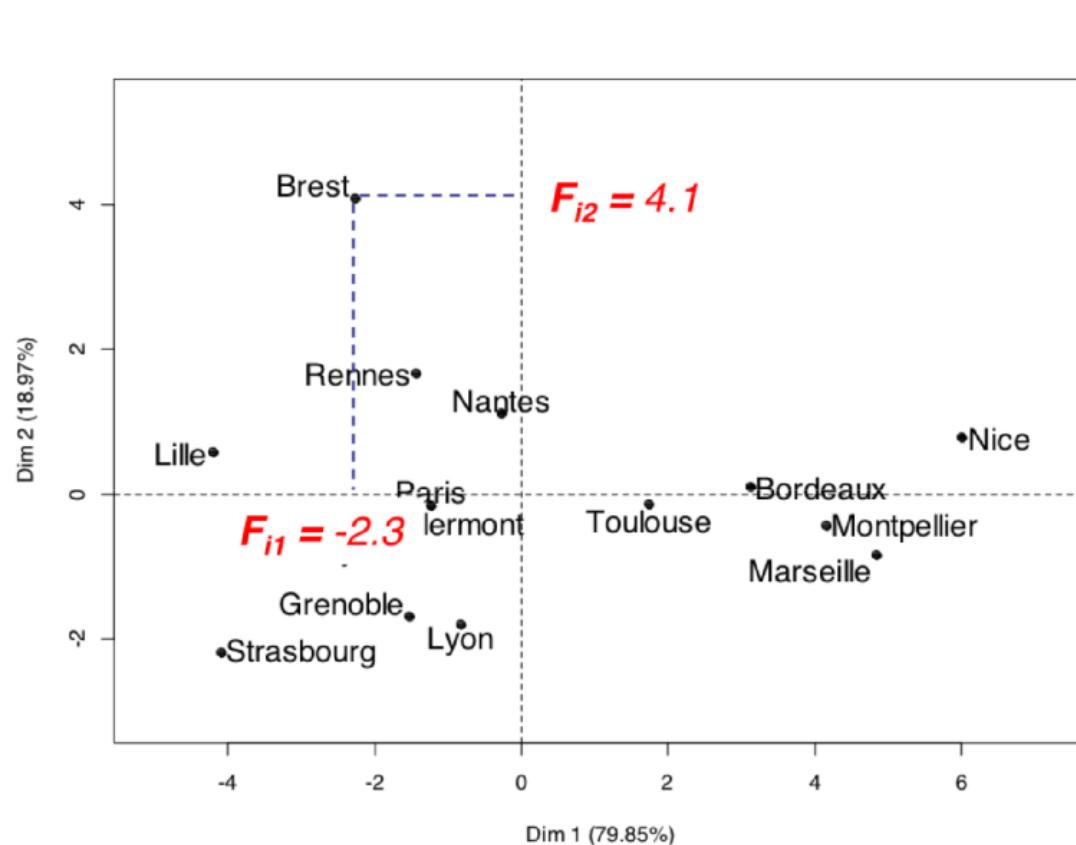
Les données brutes

	Janv	Févr	Mars	Avri	Mai	Juin	juil	Août	Sept	Octo	Nove	Déce	Lati	Long
Bordeaux	5.6	6.6	10.3	12.8	15.8	19.3	20.9	21	18.6	13.8	9.1	6.2	44.5	-0.34
Brest	6.1	5.8	7.8	9.2	11.6	14.4	15.6	16	14.7	12	9	7	48.24	-4.29
Clermont	2.6	3.7	7.5	10.3	13.8	17.3	19.4	19.1	16.2	11.2	6.6	3.6	45.47	3.05
Grenoble	1.5	3.2	7.7	10.6	14.5	17.8	20.1	19.5	16.7	11.4	6.5	2.3	45.1	5.43
Lille	2.4	2.9	6	8.9	12.4	15.3	17.1	17.1	14.7	10.4	6.1	3.5	50.38	3.04
Lyon	2.1	3.3	7.7	10.9	14.9	18.5	20.7	20.1	16.9	11.4	6.7	3.1	45.45	4.51
Marseille	5.5	6.6	10	13	16.8	20.8	23.3	22.8	19.9	15	10.2	6.9	43.18	5.24
Montpellier	5.6	6.7	9.9	12.8	16.2	20.1	22.7	22.3	19.3	14.6	10	6.5	43.36	3.53
Nantes	5	5.3	8.4	10.8	13.9	17.2	18.8	18.6	16.4	12.2	8.2	5.5	47.13	-1.33
Nice	7.5	8.5	10.8	13.3	16.7	20.1	22.7	22.5	20.3	16	11.5	8.2	43.42	7.15
Paris	3.4	4.1	7.6	10.7	14.3	17.5	19.1	18.7	16	11.4	7.1	4.3	48.52	2.2
Rennes	4.8	5.3	7.9	10.1	13.1	16.2	17.9	17.8	15.7	11.6	7.8	5.4	48.05	-1.41
Strasbourg	0.4	1.5	5.6	9.8	14	17.2	19	18.3	15.1	9.5	4.9	1.3	48.35	7.45
Toulouse	4.7	5.6	9.2	11.6	14.9	18.7	20.9	20.9	18.3	13.3	8.6	5.5	43.36	1.26
Vichy	2.4	3.4	7.1	9.9	13.6	17.1	19.3	18.8	16	11	6.6	3.4	46.08	3.26

Les données centrées réduites

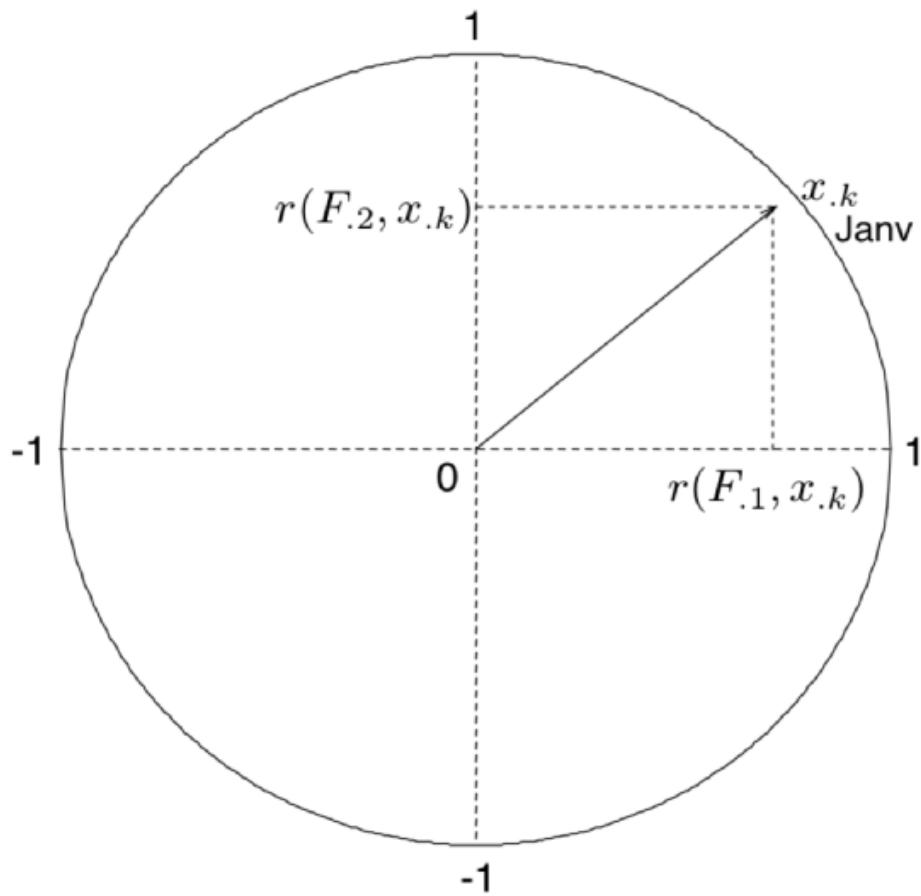
	Janv	Févr	Mars	Avri	Mai	Juin	juil	Août	Sept	Octo	Nove	Déce	Lat	Long
Bordeaux	0.84	0.98	1.40	1.33	0.94	0.85	0.52	0.74	0.90	0.84	0.67	0.72	44.5	-0.34
Brest	1.10	0.54	-0.29	-1.30	-1.95	-1.98	-2.06	-1.83	-1.28	-0.18	0.62	1.14	48.24	-4.29
Clermont	-0.71	-0.63	-0.50	-0.50	-0.44	-0.31	-0.21	-0.24	-0.44	-0.63	-0.76	-0.66	45.47	3.05
Grenoble	-1.28	-0.90	-0.36	-0.28	0.05	-0.02	0.13	-0.03	-0.16	-0.52	-0.82	-1.35	45.1	5.43
Lille	-0.81	-1.07	-1.51	-1.52	-1.40	-1.46	-1.33	-1.27	-1.28	-1.09	-1.05	-0.71	50.38	3.04
Lyon	-0.97	-0.85	-0.36	-0.06	0.32	0.38	0.42	0.27	-0.05	-0.52	-0.70	-0.92	45.45	4.51
Marseille	0.79	0.98	1.20	1.48	1.63	1.71	1.69	1.66	1.63	1.52	1.30	1.09	43.18	5.24
Montpellier	0.84	1.03	1.13	1.33	1.22	1.31	1.39	1.41	1.30	1.29	1.19	0.87	43.36	3.53
Nantes	0.53	0.26	0.11	-0.13	-0.37	-0.37	-0.50	-0.50	-0.33	-0.07	0.16	0.35	47.13	-1.33
Nice	1.82	2.03	1.74	1.70	1.56	1.31	1.39	1.51	1.86	2.08	2.05	1.77	43.42	7.15
Paris	-0.30	-0.41	-0.43	-0.20	-0.09	-0.19	-0.36	-0.45	-0.55	-0.52	-0.47	-0.29	48.52	2.2
Rennes	0.43	0.26	-0.23	-0.64	-0.92	-0.94	-0.94	-0.91	-0.72	-0.41	-0.07	0.29	48.05	-1.41
Strasbourg	-1.84	-1.85	-1.78	-0.86	-0.30	-0.37	-0.41	-0.65	-1.06	-1.60	-1.74	-1.87	48.35	7.45
Toulouse	0.37	0.42	0.65	0.45	0.32	0.50	0.52	0.69	0.74	0.55	0.39	0.35	43.36	1.26
Vichy	-0.81	-0.79	-0.77	-0.79	-0.57	-0.42	-0.26	-0.39	-0.55	-0.75	-0.76	-0.76	46.08	3.26

Projection des individus sur les 2 premières composantes principales

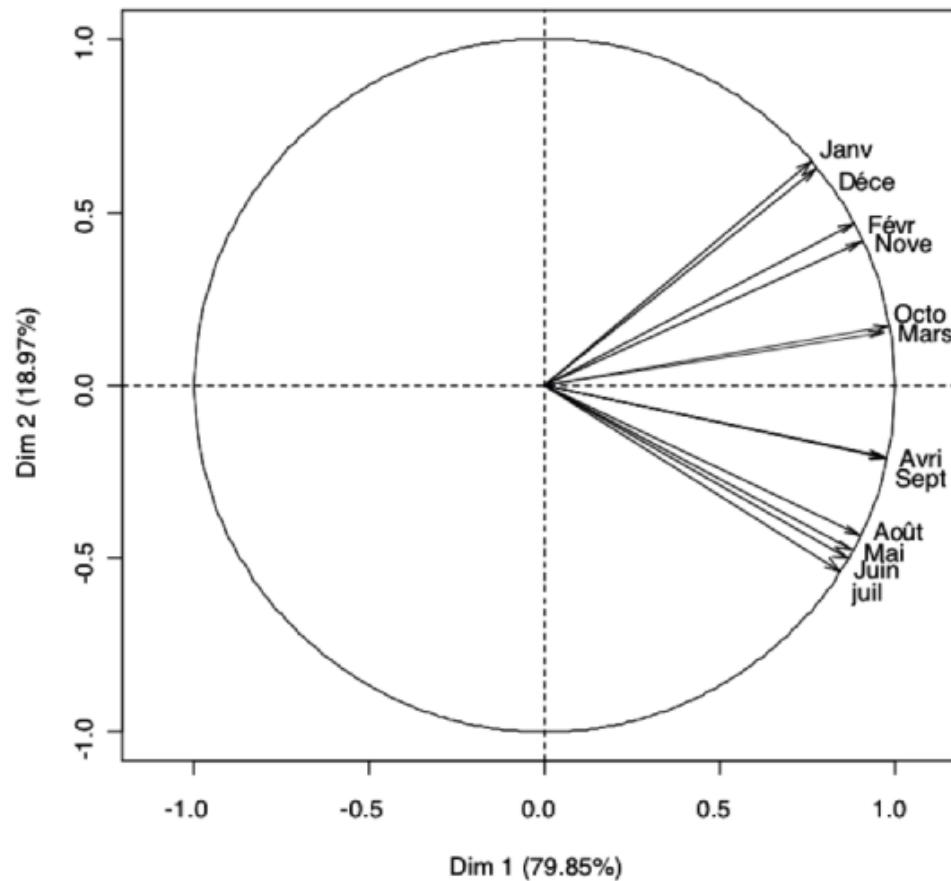


Quelle **interprétation** pour les axes de projection ?

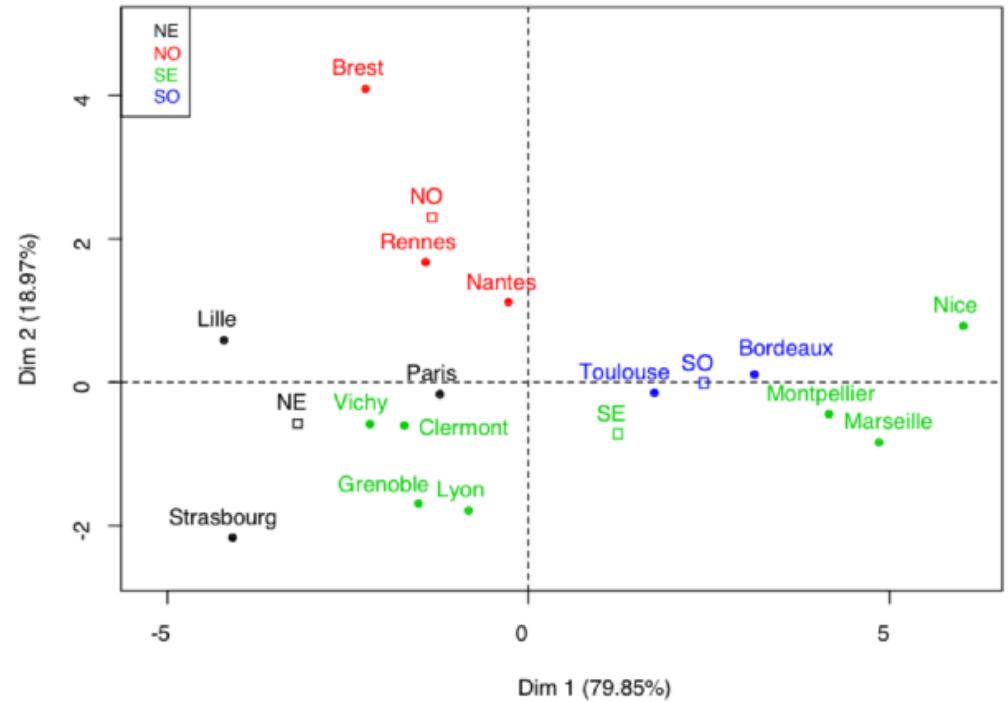
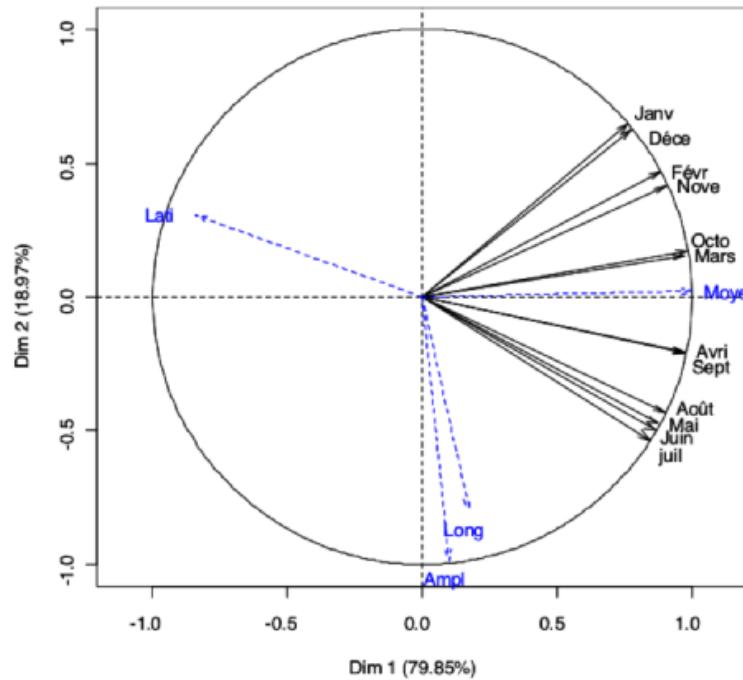
Cercle de corrélation



Projection des variables sur le cercle de corrélation



Information supplémentaire



Variables quantitatives

- Moyenne
- Amplitude
- Longitude
- Latitude

Modalités (NO, NE, SO, SE)

pour compléter : cours ACP François Husson (youtube)