Introduction au cours d'analyse de données

Jean-Marc Lasgouttes — Inria Paris
Jean-Marc.Lasgouttes@inria.fr

http://ana-donnees.lasgouttes.net/

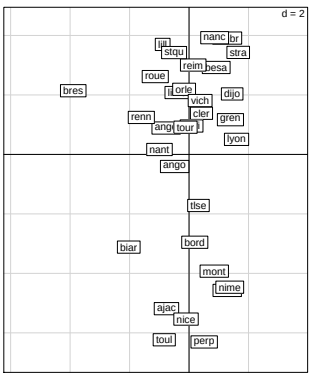
Exemple : la température en France

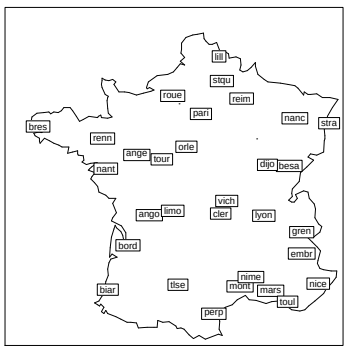
Contexte température moyenne mois par mois dans 25 villes de France.

Les données brutes sont difficiles a interpréter

	janv	fev	mars	avri	mai	juin	juil	aout	sept	oct	nov	dec
ajac	7.7	8.7	10.5	12.6	15.9	19.8	22.0	22.2	20.3	16.3	11.8	8.7
ange	4.2	4.9	7.9	10.4	13.6	17.0	18.7	18.4	16.1	11.7	7.6	4.9
ango	4.6	5.4	8.9	11.3	14.5	17.2	19.5	19.4	16.9	12.5	8.1	5.3
besa	1.1	2.2	6.4	9.7	13.6	16.9	18.7	18.3	15.5	10.4	5.7	2.0
biar	7.6	8.0	10.8	12.0	14.7	17.8	19.7	19.9	18.5	14.8	10.9	8.2
bord	5.6	6.6	10.3	12.8	15.8	19.3	20.9	21.0	18.6	13.8	9.1	6.2
bres	6.1	5.8	7.8	9.2	11.6	14.4	15.6	16.0	14.7	12.0	9.0	7.0
cler	2.6	3.7	7.5	10.3	13.8	17.3	19.4	19.1	16.2	11.2	6.6	3.6
dijo	1.3	2.6	6.9	10.4	14.3	17.7	19.6	19.0	15.9	10.5	5.7	2.1
embr	0.5	1.6	5.7	9.0	13.0	16.4	18.9	18.3	15.3	10.1	4.6	0.5
gren	1.5	3.2	7.7	10.6	14.5	17.8	20.1	19.5	16.7	11.4	6.5	2.3
lill	2.4	2.9	6.0	8.9	12.4	15.3	17.1	17.1	14.7	10.4	6.1	3.5
limo	3.1	3.9	7.4	9.9	13.3	16.8	18.4	17.8	15.3	10.7	6.7	3.8
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
mars	5.5	6.6	10.0	13.0	16.8	20.8	23.3	22.8	19.9	15.0	10.2	6.9
mont	5.6	6.7	9.9	12.8	16.2	20.1	22.7	22.3	19.3	14.6	10.0	6.5
nanc	0.8	1.6	5.5	9.2	13.3	16.5	18.3	17.7	14.7	9.4	5.2	1.8
nant	5.0	5.3	8.4	10.8	13.9	17.2	18.8	18.6	16.4	12.2	8.2	5.5
nice	7.5	8.5	10.8	13.3	16.7	20.1	22.7	22.5	20.3	16.0	11.5	8.2
nime	5.7	6.8	10.1	13.0	16.6	20.8	23.6	22.9	19.7	14.6	9.8	6.5
orle	2.7	3.6	6.9	9.8	13.4	16.6	18.4	18.2	15.6	10.9	6.6	3.6
pari	3.4	4.1	7.6	10.7	14.3	17.5	19.1	18.7	16.0	11.4	7.1	4.3
perp	7.5	8.4	11.3	13.9	17.1	21.1	23.8	23.3	20.5	15.9	11.5	8.6
reim	1.9	2.8	6.2	9.4	13.3	16.4	18.3	17.9	15.1	10.3	6.1	3.0
renn	4.8	5.3	7.9	10.1	13.1	16.2	17.9	17.8	15.7	11.6	7.8	5.4
roue	3.4	3.9	6.8	9.5	12.9	15.7	17.6	17.2	15.0	11.0	6.8	4.3
stqu	2.0	2.9	6.3	9.2	12.7	15.6	17.4	17.4	15.0	10.5	6.1	3.1
stra	0.4	1.5	5.6	9.8	14.0	17.2	19.0	18.3	15.1	9.5	4.9	1.3
toul	8.6	9.1	11.2	13.4	16.6	20.2	22.6	22.4	20.5	16.5	12.6	9.7
tlse	4.7	5.6	9.2	11.6	14.9	18.7	20.9	20.9	18.3	13.3	8.6	5.5
tour	3.5	4.4	7.7	10.6	13.9	17.4	19.1	18.7	16.2	11.7	7.2	4.3
vich	2.4	3.4	7.1	9.9	13.6	17.1	19.3	18.8	16.0	11.0	6.6	3.4

L'ACP propose une représentation en deux dimensions (à gauche)





Surprise La représentation ressemble à la carte de France. On peut l'interpréter.

Quelles données ?

Population groupe ou ensemble d'*individus* que l'on analyse.

Recensement étude de tous les individus d'une population donnée.

Sondage étude d'une partie seulement d'une population appelée *échantillon*.

Variables ensemble de caractéristiques d'une population.

- *quantitatives* : nombres sur lesquels les opérations usuelles (somme, moyenne,...) ont un sens ; elles peuvent être *discrètes* (ex : nombre d'éléments dans un ensemble) ou *continues* (ex : prix, taille) ;
- *qualitatives* : appartenance à une catégorie donnée ; elles peuvent être *nominales* (ex : sexe, CSP) ou *ordinales* quand les catégories sont ordonnées (ex : très résistant, assez résistant, peu résistant).

L'analyse de données

Statistiques descriptives synthétiser, structurer l'information contenue dans des données multidimensionnelles (n individus, p variables).

Deux groupes de méthodes

- *méthodes de classification* : réduire la taille de l'ensemble des individus en formant des groupes homogènes *(pas fait dans ce cours)* ;
- *méthodes factorielles* : réduire le nombre de variables en les résumant par un petit nombre de composantes synthétiques.

Deux types de méthodes factorielles

- variables numériques : *analyse en composantes principales* (ACP) ;
- variables qualitatives : *analyse factorielle des correspondances* (AFC) et *analyse des correspondance multiples* (ACM).

Ce que ce cours *n'est pas*

Un cours de statistique inférentielle il ne sera pas question ici d'estimation ou de régression ; nous utiliserons juste un peu de tests statistiques.

Un cours orienté « utilisateur » on cherche à la fois à savoir utiliser les méthodes d'analyse de données, et à comprendre les fondements mathématiques de ces méthodes.

Un cours « pratique » Les contraintes d'effectif et de matériel font que nous ne ferons qu'un peu de travaux pratiques à la fin.

Outils utilisés

Statistiques élémentaires on calcule des moyennes, variances, corrélations...

Matrices les tableaux de données sont vus comme des matrices : opérations élémentaires, vecteurs propres, valeurs propres...

Géométrie *(espaces métriques)* les données sont aussi vues comme des nuages de points en grande dimension : produits scalaires, normes, orthogonalité...

Statistiques inférentielles *(un peu)* on utilisera quelques tests statistiques.

Organisation du cours

Durée 12 semaines, le mardi en général de 13h15 à 16h30 pour les TDs et 16h45 à 18h15 pour le cours.

- Période 1 : analyse en composantes principales (ACP, ≈ 6 semaines)
- *IS sur l'ACP*
- Période 2 : analyse des variables qualitatives (AFC et ACM, ≈ 6 semaines)
- *DS sur ACP, AFC, ACM*

Type de cours pour chacune des deux périodes

- d'abord surtout du cours magistral
- quand les méthodes sont en place, surtout des analyses jeux de données en TD : une table de donnée, les sortie du logiciel et une série de questions pour guider l'analyse.
- quelques exercices de démonstrations mathématiques
- En fin de seconde période, deux séances de TP sur machine (analyse de données en R).

Notation (IS et DS)

- $4/5$ analyse d'un ou plusieurs jeux de données
- $1/5$ (petite) démonstration mathématique

Références

Ces références sont données à titre indicatif ; *aucun livre n'est demandé pour ce cours.*

Base du cours Gilbert Saporta, *Probabilités, analyse des données et statistique*, 3e édition, Technip, 2011.

Logiciel de traitement de données Les tables et graphiques présentés dans le cours et les analyses de données sont produites par le logiciel R (à l'aide du paquetage `ade4`). R est un logiciel libre (et donc gratuit) disponible pour Windows, macOS et Linux à l'adresse `http://www.r-project.org`.

Archives de ce cours cours, TD avec corrigé, données brutes sont disponibles à

`http://ana-donnees.lasgouttes.net/`