ACP
TP 2, par groupe de 3
Étude de cas sur données réelles environnementales

Dans cette séance, il s'agit d'acquérir un savoir-faire de l'ACP adapté aux données réelles et leurs interprétation. Dans ce TP, les différentes étapes sont :

— Faire une étude statistique préparatoire des données (moyenne, écart type, médiane, corrélation entre variables) : attention aux données manquantes ou égales à 0
— Mettre en œuvre une ACP avec analyse et interprétation des résultats (représentativité de l'ACP)
— Apporter des réponses aux questions posées sur cette étude

Le document à rendre

Le travail à rendre se fait sous la forme d'un rapport de type note de synthèse (pdf, word) mettant en évidence :

1. l'explication des choix réalisés pour votre analyse, les traitements sur les données,
2. la présentation et les explications des analyses faites sur ce type de données avec les codes source et graphes associés,
3. un fichier nommé à vos noms : `nom-trinome.pdf`.

Ressources :

— Vous disposez d'un exemple de programme R sur campus -> fichier temperat.R dans le dossier *cours*
— Utiliser la librairie FactoMineR :

```
if(!require(FactoMineR)) install.packages("FactoMineR")
library(FactoMineR)
```

→ documentation
→ un tuto FactoMineR sur ce site
— Vous pouvez me poser des questions par mail : thomas.galtier@emse.fr

1 Objet d'étude

Dans le cadre d'une étude d'impact environnemental industriel, on s'intéresse à l'émission de composés chimiques par des sites industriels de traitement de déchets par compostage localisés dans la Loire. L'émission de ces composés chimiques comporte des risques sanitaires potentiels pour les populations avoisinantes, et certaines conditions de fonctionnement peuvent favoriser l'émission de composés toxiques : entrants trop importants à différentes périodes de l'année, conditions de fermentation (anaérobie au lieu de dégradation aérobie), mauvaise gestion du site.

Afin d'identifier l'impact d'un site, il est nécessaire de comparer ses émissions à une référence. (C'est une méthode classique des statistiques explicatives. Pour pouvoir dire qu'une mesure est "bonne", "meilleure", "correcte", "pire" il faut préciser par rapport à quoi. Pour pouvoir dire qu'un phénomène à un impact, il faut le comparer à un cas où le phénomène n'est pas présent, sans quoi l'impact n'est tout simplement pas mesurable.) Pour cela on mesure la présence de composés avant et après installation du site, puis on les compare. La présence de composés avant installation est appelée bruit de fond. Des campagnes de mesure de ces composés ont été effectuées avant (dans le labels les 2 lettres BF) et après la mise en activité du site (lettre CA dans le labels) à différentes périodes de l'année (H pour hiver et E été).

On cherche donc à répondre aux questions suivantes :

- Existe-il une différence entre les observations avant et après la mise en activité du site ?
- Existe-il une différence entre les campagnes hiver / été ?
- Les localisation des points de mesure autour du site, montre-elle des regroupements de comportement (composés chimiques atmosphériques d'origine industrielle, automobile, milieu urbain, milieu rural...) ?

2 Les données : fichier TP2.xls sur campus

Deux campagnes de mesures (modalités BF2, BF3 de la variable qualitative *campagne*) ont été réalisées avant installation du site, et 4 après (modalités CA1, CA2, CA3, CA4 de la variable qualitative *campagne*). Ces campagnes de mesure effectuées ont été effectuées sur 33 sites (variable qualitative *localisation*), pour 14 familles de composés organiques volatiles. Les familles de composés chimiques volatiles considérées sont appelées : *6, B, T, E, X, 9_ane, 10_ane, 13_ane, 14_ane, 1_M_2_PA, BTM, FormicAcid, aceticacid, NonaDecanoicAc, Tot_OcNoDecana*. Pour chacune d'entre elles on dispose d'une variable quantitative à son nom, donnant la quantité de composants volatiles mesurée. On dispose également de variable quantitatives sur : l'environnement du site (variable *TYPE*), la saison (variable *SAISON*), Selon les sites, les 6 campagnes n'ont pas toutes été effectuées. Par conséquent on a 140 observations.

3 Exemple de marche à suivre

Vous devez choisir une stratégie de traitement de données pour essayer de répondre aux questions posées.

Étapes :

1. Statistiques de base : données manquantes ? aberrantes ?
   Gestion des données manquantes ou aberrante choisie ? Imputation ou suppression des observations contenant une données manquantes, des variables contenants beaucoup de données.
   Tableaux croisés, boxplot, analyses basiques selon les saison et l'environnement, et interprétations préliminaires.
2. Construire une matrice $\mathbb{X}$ R^p choisir si on normalise ?
3. ACP :
   (a) Identification du nombre de composantes. La réduction est-elle pertinente ?
   (b) Affichage des plans factoriels. Les individus sont-ils bien projetés ? Interprétation des plans factoriels. Groupes de points identifiables ?
   (c) Affichage et interprétations du cercle de corrélation.
   Pouvez vous donner un sens ou un nom aux composantes ?
4. Réponses aux questions dans un paragraphe de conclusion.