# Statistiques Exploratoires

TP Noté

A. L. N'Guessan, V. Roca & W. Heyse

Le 06 Octobre 2025

## Objectif du TP

Vous êtes chargé · e de réaliser une analyse exploratoire et une classification sur un jeu de données multidimensionnelles. Votre objectif est de comprendre les relations entre les variables, d'identifier des structures sous-jacentes et de proposer une classification des individus. Vous utiliserez pour cela des méthodes factorielles (ACP, AFC, ou ACM) et la Classification Ascendante Hiérarchique (CAH). Vous veillerez, dans chaque cas, à justifiez et à argumenter vos propos (comment vous contruisez vos raisonnements, comment vous arrivez à vos interprétation, ...)

## Données

Les caractéristiques de sols ont été mesurées sur des échantillons provenant de trois types de contours (Sommet, Pente et Creux) et à quatre profondeurs (0-10 cm, 10-30 cm, 30-60 cm, et 60-90 cm).

Un tableau de données avec 48 observations sur les 14 variables suivantes :

- **Contour** : un facteur avec 3 niveaux correspondant à l'endroit de prélèvement : `Depression` (Creux), `Slope` (Pente), `Top` (Sommet).
- **Depth** : un facteur avec 4 niveaux en cm correspondant à la profondeur de prélèvement : `0-10`, `10-30`, `30-60`, `60-90`.
- **pH** : pH du sol.
- **N** : azote total en %.
- **Dens** : densité apparente en g/cm$^3$.
- **P** : phosphore total en ppm.
- **Ca** : calcium en meq/100 g.
- **Mg** : magnésium en meq/100 g.
- **K** : potassium en meq/100 g.
- **Na** : sodium en meq/100 g.
- **Conduc** : conductivité.

Le *meq* est une unité de mesure du nombre d'ions présents (pour 100g ici).

Vous pouvez charger les données à l'aide du code : `readRDS(file = "Soils.RDS")`.

## Travail à réaliser

1. **Analyse descriptive initiale**
   - Décrivez brièvement le jeu de données (nature des variables, statistiques descriptives, distribution des variables, etc.).

- Identifiez d'éventuelles particularités dans les données (valeurs manquantes, valeurs aberrantes, etc.).

Selon les résultats de votre analyse descriptive, vous décidrez de vous orienter vers les méthodes factorielles (2.) et/ou vers la classification non supervisée (3.). Ici, toutes les méthodes peuvent être appliquées (modulo quelques modifications sur les données).

2. **Application d'une méthode factorielle adaptée**
   - Quelles variables et quels individus contribuent le plus aux axes principaux ?
   - Quelles associations ou oppositions observez-vous ?
   - Quelles structures latentes se trouvent dans les données ?
3. **Classification Ascendante Hiérarchique (CAH)**

   - Réalisez une CAH en utilisant les coordonnées des individus sur les axes factoriels comme entrée.
   - Proposez un découpage en classes (nombre à déterminer en fonction de votre analyse).
   - Visualisez les résultats avec un dendrogramme et interprétez les classes obtenues.
4. **Synthèse des résultats**
   - Proposez une interprétation globale des structures identifiées dans les données.
   - Quelles conclusions pourriez-vous en tirer selon le contexte des données ?

# Rendu attendu

- Un rapport PDF ou HTML synthétique contenant, les résultats de vos analyses appuyés sur des données ou des graphes pertinents à votre anlayse (inutile d'afficher des graphes pour remplir le rapport).
- Un script Rmd.

# Évaluation

- Qualité de l'analyse et justesse des interprétations (30%).
- Pertinence et rigueur dans l'utilisation des outils statistiques (30%).
- Clarté et organisation du rapport (20%).
- Qualité du code (20%).