Ecole supérieure d'agro-développement international

istom

Faire une ACP avec R

Promotion 115

Table des matières

Aide-mémoire : l'ACP normée **2**

Exercice 1 (guidé) : la fertilité des sols de quinze sites du Cameroun **4**

Exercice 2 (en autonomie) : typologie d'exploitations familiales au Burkina Faso **8**

Indications **10**

Corrections **11**

Correction de l'exercice 1 11

Ce TD comporte **deux exercices** à réaliser **avec R sous RStudio**, sur des données agronomiques réalistes :

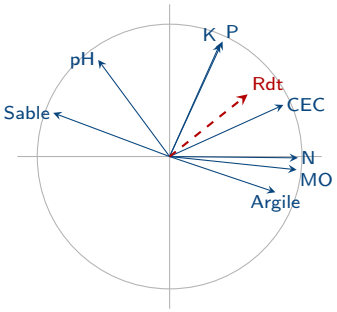
- **Exercice 1 (guidé)** : la fertilité des sols de quinze sites du Cameroun. Tout le code R est fourni ; à chaque étape, des encadrés **Lecture** expliquent les sorties et des lignes **✓ Contrôle** permettent de vérifier ses résultats ;
- **Exercice 2 (en autonomie)** : une typologie de 32 exploitations familiales au Burkina Faso. Aucun code n'est fourni : on choisit soi-même les fonctions et les réglages, et l'on justifie chaque choix.

Les jeux de données sont fournis dans les fichiers Excel `data_stat-0109.xlsx` et `data_stat-0110.xlsx`, à placer dans le dossier `data_raw` de votre projet RStudio. On utilise les packages `readxl` et `FactoMineR`.

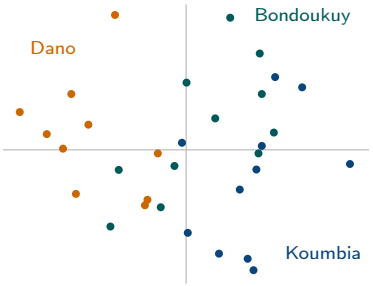
Ces exercices entraînent à mener une ACP comme on le fera en stage ou en mémoire :

- **les choix** : variables actives et supplémentaires, ACP normée, nombre d'axes à retenir ;
- **la lecture des sorties de R** : valeurs propres, cercle des corrélations, carte des individus, $\cos^2$, contributions, valeurs-test, description automatique des axes (`dimdesc`) ;
- **l'interprétation** : nommer les axes, contrôler la fiabilité de la lecture, relier les résultats à la question agronomique et rédiger une synthèse.

Un **aide-mémoire** (page 2) rassemble les formules, les commandes R correspondantes et les règles de lecture des graphiques. Les indications et les corrections sont en fin de document.



Exercice 1 (guidé)
fertilité des sols et rendement du maïs : le cercle des corrélations

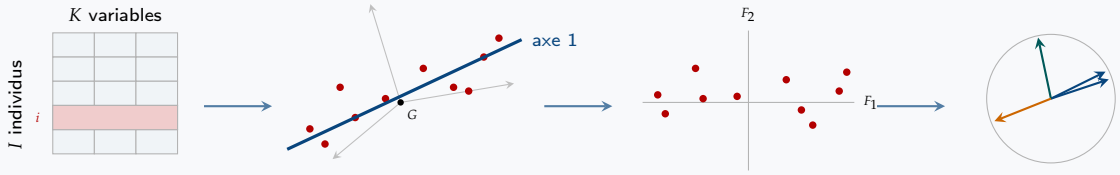


Exercice 2 (en autonomie)
typologie de 32 exploitations : la carte, colorée par village

Aide-mémoire : l'ACP normée

Ces deux pages rassemblent tout ce qui sert dans les quatre exercices : le principe de l'ACP, les formules « à la main » et leur équivalent sous R (`FactoMineR`), puis les règles de lecture des graphiques. Les notations sont celles du TD : I individus, K variables, z_{ik} la valeur centrée réduite de l'individu i pour la variable k.

Le principe : photographier un nuage de points sous son meilleur angle



1. le tableau
centré, réduit

2. le nuage des I points
axe 1 : direction d'étalement maximal

3. la carte (plan 1-2)
la « photographie » la plus fidèle

4. le cercle
les variables

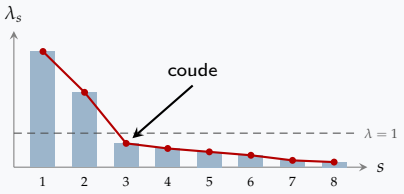
L'ACP cherche, dans l'espace à K dimensions, les directions orthogonales $U_1, U_2, \ldots$ le long desquelles le nuage est le plus étalé : la première porte la plus grande part d'inertie λ_1, la deuxième la plus grande part restante λ_2, etc. On lit ensuite les **individus** sur la carte et les **variables** sur le cercle, en faisant des allers-retours entre les deux.

Les étapes, à la main et sous R

Étape	À la main	Sous R (`res <- PCA(...)`)
Centrer, réduire	$z_{ik} = \dfrac{x_{ik} - \overline{X}_k}{\sigma_k}$ (σ_k : division par I)	`scale.unit = TRUE`
Distance au centre	$d^2(i,G) = \sum_k z_{ik}^2$ $\quad \mathcal{I}_{\text{tot}} = \frac{1}{I}\sum_i d^2(i,G) = K$	`res$ind$dist` (donne d)
Corrélations	$r_{kl} = \frac{1}{I}\sum_i z_{ik} z_{il}$ $\quad R = \frac{1}{I} Z^{\mathrm{T}} Z$	`cor(...)`
Valeurs propres	$\det(R - \lambda \mathrm{I}) = 0$ $\quad \sum_s \lambda_s = K$ $\quad \% = 100\,\lambda_s / K$	`res$eig`
Vecteurs propres	$R\,U_s = \lambda_s\,U_s$ $\quad \lVert U_s \rVert = 1$	`res$svd$V`
Coordonnées des individus	$F_s(i) = \sum_k u_{sk} z_{ik}$ $\quad \mathrm{Var}(F_s) = \lambda_s$	`res$ind$coord`
Qualité (individus)	$\cos^2_s(i) = \dfrac{F_s(i)^2}{d^2(i,G)}$ $\quad \sum_s \cos^2_s(i) = 1$	`res$ind$cos2`
Contribution (individus)	$\mathrm{contrib}_s(i) = 100\,\dfrac{F_s(i)^2}{I\,\lambda_s}$ $\quad$ moyenne $\dfrac{100}{I}$	`res$ind$contrib`
Coordonnées des variables	$r(X_k, F_s) = \sqrt{\lambda_s}\,u_{sk}$	`res$var$coord` (`$cor`)
Qualité (variables)	$\cos^2_s(X_k) = r^2(X_k, F_s)$ $\quad \sum_s r^2(X_k, F_s) = 1$	`res$var$cos2`
Contribution (variables)	$\mathrm{contrib}_s(X_k) = 100\,\dfrac{r^2(X_k, F_s)}{\lambda_s}$ $\quad$ moyenne $\dfrac{100}{K}$	`res$var$contrib`

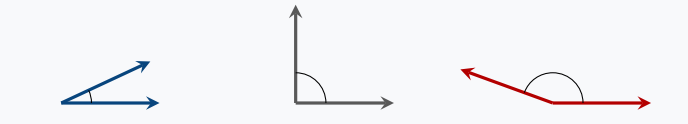
✓ **Contrôle** deux façons de retrouver la même valeur propre : $\lambda_s = \frac{1}{I}\sum_i F_s(i)^2$ (côté individus) $= \sum_k r^2(X_k, F_s)$ (côté variables).

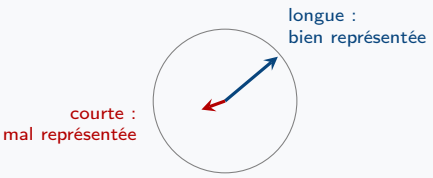
Combien d'axes ? Deux critères à croiser



- **Kaiser** : garder les axes tels que $\lambda_s \geq 1$ (ils résument au moins autant qu'une variable de départ).
- **Coude** : garder les axes situés avant la cassure de l'éboulis.
- Regarder aussi le **% d'inertie cumulé**, et ne garder que des axes que l'on sait **interpréter**.

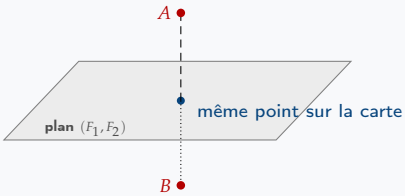
Lire le cercle des corrélations





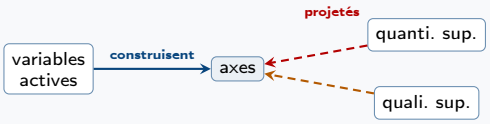
- La coordonnée d'une flèche sur un axe est la **corrélation** de la variable avec cet axe ; on nomme l'axe avec les variables qui y **contribuent** le plus (lues *axe par axe*).
- L'angle entre deux flèches approche leur corrélation ($r \approx \cos\theta$) **seulement si les deux flèches arrivent près du cercle**.

Lire la carte des individus



- Deux points proches sur la carte ne se ressemblent que s'ils sont **tous deux bien représentés** ($\cos^2$ sur le plan proche de 1) ; sinon, comme A et B, ils peuvent être éloignés.
- Un point situé dans la direction d'une flèche a une forte valeur de cette variable.
- Un individu de **contribution** très supérieure à $100/I$ « tire » l'axe : vérifier la donnée, refaire l'ACP en le mettant en supplémentaire.

Éléments supplémentaires



- Ils ne construisent pas les axes : on les place ensuite pour *interpréter* (ce que l'on cherche à expliquer, variables qualitatives).
- Une modalité est placée au barycentre de ses individus ; sa position sur un axe est significative si $|\text{v.test}| > 2$ (`res$quali.sup$v.test`, `dimdesc`).

Les pièges classiques. (1) Faire une ACP non normée sur des variables d'unités différentes : la variable de plus grande variance écrase les autres. (2) Interpréter la proximité de points ou l'angle de flèches mal représentés. (3) Croire qu'un axe classe du « meilleur » au « moins bon » : il oppose souvent deux profils. (4) S'inquiéter d'un axe inversé : son sens est arbitraire (un logiciel peut le retourner). (5) Nommer les axes avec les contributions *au plan* : elles mélangent les deux axes. (6) Conclure à une causalité : l'ACP décrit des associations.

Exercice 1 ACP guidée sous R : la fertilité des sols de quinze sites du Cameroun

Un projet d'appui à la filière maïs veut savoir *quels types de sols* on rencontre entre la zone forestière du Sud et les savanes de l'Adamaoua, et si ces types de sols expliquent les différences de rendement. Sur **15 sites**, on a analysé un échantillon composite de sol (horizon 0-20 cm). Le fichier `data_stat-0109.xlsx` contient, dans la feuille `donnees`, une ligne par site :

- **huit variables d'analyse de sol** : `pH` (pH eau), `MO` (matière organique, %), `N` (azote total, g/kg), `P` (phosphore assimilable, mg/kg), `K` (potassium échangeable, cmol$^+$/kg), `CEC` (capacité d'échange cationique, cmol$^+$/kg), `Argile` et `Sable` (%) ;
- `Rdt_mais` : rendement moyen du maïs sur le site (t/ha) ;
- `Zone` : zone agro-écologique (Forêt, Hauts plateaux, Transition, Savane).

La feuille `description` rappelle le sens et l'unité de chaque colonne (les données sont fictives).

☞ Récupérer les données

Scanner le QR code avec un téléphone, ou cliquer sur le lien `data_stat-0109.xlsx`, puis télécharger le fichier avec le bouton Download raw file de la page GitHub. L'enregistrer dans le dossier `data_raw` de votre projet RStudio.

https://github.com/ant0in3g/ant0in3g.github.io/blob/main/sTa7/td-R-lab/R-lab/data_raw/data_stat-0109.xlsx

Mode d'emploi

- Chaque question donne le code R à taper *dans un script* (*à enregistrer dans votre dossier OneDrive :* `STAT_Nom_Prenom`), pas dans la console : on garde ainsi une trace de tout ce que l'on fait. Pour exécuter une ligne, placer le curseur dessus et taper `Ctrl + Entrée`.
- Après chaque bloc de code : exécuter le code, comparer la sortie aux lignes ✓ **Contrôle**, lire les encadrés **Lecture**, puis répondre aux questions.
- Tout ce qui suit le signe # est un commentaire : R l'ignore.

1. Préparer la séance

a) Dans votre dossier de travail OneDrive, placer le fichier `data_stat-0109.xlsx` (encadré « Récupérer les données » ci-dessus) dans `data_raw`. Lancer RStudio à partir du fichier `_stat_Istom.Rproj`.

b) Installer les packages `readxl` et `FactoMineR` avec RStudio, puis les charger (*à chaque séance*) :

```
library(readxl)       # pour lire les fichiers Excel
library(FactoMineR)   # pour l'ACP
```

c) Importer les données :

```
sols <- as.data.frame(read_excel("data_raw/data_stat-0109.xlsx", sheet = "donnees"))
rownames(sols) <- sols$Site       # les sites deviennent les noms de lignes
sols$Site <- NULL                 # on supprime la colonne devenue inutile
sols$Zone <- as.factor(sols$Zone) # la zone est une variable qualitative (facteur)
str(sols)                         # structure du tableau
head(sols)                        # les 6 premieres lignes
```

✓ **Contrôle** `str` affiche `15 obs. of 10 variables` ; les neuf premières colonnes sont `num` (numériques) et `Zone` est un `Factor w/ 4 levels`.

Pourquoi faut-il mettre les noms des sites en noms de lignes, et non les laisser dans une colonne ? Quelle est maintenant la position (le numéro de colonne) de `Rdt_mais` ? de `Zone` ?

2. Regarder les données avant l'ACP

a) Afficher un résumé de chaque variable et les écarts-types :

```
summary(sols)
round(apply(sols[, 1:9], 2, sd), 2)   # ecart-type de chacune des colonnes 1 a 9
```

Lecture. `apply(tableau, 2, fonction)` applique la fonction à chaque colonne (le 2 désigne les colonnes) ; `round(..., 2)` arrondit à deux décimales. Attention : sd divise par $I-1$, alors que l'ACP utilise des variances « population » (division par I) ; la différence est ici faible.

Les variables ont-elles la même unité ? le même ordre de grandeur de dispersion ? En déduire qu'il faut faire une ACP *normée* (sur données centrées réduites). Que se passerait-il sinon ?

b) Afficher la matrice des corrélations des huit variables de sol :

```
round(cor(sols[, 1:8]), 2)
```

Repérer les quatre couples de variables les plus corrélés (en valeur absolue). Ces liaisons vous semblent-elles logiques pour un agronome ?

c) Regarder le rendement selon la zone :

```
boxplot(Rdt_mais ~ Zone, data = sols, ylab = "Rendement (t/ha)")
tapply(sols$Rdt_mais, sols$Zone, mean)   # rendement moyen par zone
```

Quelle zone a les meilleurs rendements ? Gardez ce résultat en tête : on cherchera à l'expliquer par les sols à la question 7.

3. Lancer l'ACP

a) On veut construire les axes avec les *huit variables de sol* seulement, et utiliser le rendement et la zone pour *interpréter*. Taper :

```
res <- PCA(sols, scale.unit = TRUE, quanti.sup = 9, quali.sup = 10,
           graph = FALSE)
```

Expliquer le rôle de chacun des arguments `scale.unit = TRUE`, `quanti.sup = 9`, `quali.sup = 10` et `graph = FALSE`.

b) Pourquoi est-il judicieux de mettre le rendement en variable *supplémentaire* plutôt qu'en variable active ?

c) **Que contient `res` ?** Taper `names(res)`. L'objet `res` est une liste qui range tous les résultats de l'ACP ; on y accède avec le signe `$`. Les notations sont celles du cours et des exercices faits à la main :

Dans R	Ce que c'est
`res$eig`	valeurs propres λ_s, % d'inertie, % cumulé
`res$var$coord` (ou `$cor`)	coordonnées des variables : $r(X_k, F_s)$
`res$var$cos2`, `res$var$contrib`	$\cos^2_s(X_k) = r^2(X_k, F_s)$ et contributions (%)
`res$ind$coord`	coordonnées des individus : $F_s(i)$
`res$ind$cos2`, `res$ind$contrib`	qualité de représentation et contributions (%)
`res$ind$dist`	distance au centre de gravité $d(i, G)$
`res$quanti.sup`, `res$quali.sup`	résultats des éléments supplémentaires

Dans ces tableaux, chaque ligne est un élément (variable ou site) et chaque colonne un axe (`Dim.1`, `Dim.2`, ...). Ainsi `res$ind$coord[, 1:2]` donne les coordonnées de tous les sites sur les axes 1 et 2, et `res$ind$coord["Meiganga", 1]` la coordonnée de Meiganga sur l'axe 1. Que donne `summary(res)` ?

4. Combien d'axes faut-il regarder ?

Rappels. En ACP normée, chacune des K variables actives a une variance égale à 1 : l'inertie totale vaut K. Le **critère de Kaiser** retient les axes tels que $\lambda_s \geq 1$, c'est-à-dire ceux qui résument au moins autant qu'une variable de départ (ici, avec $K = 8$, au moins $\frac{1}{8} = 12{,}5\,\%$ de l'inertie). Le **critère du coude** retient les axes situés *avant* la cassure de l'éboulis, là où les barres deviennent petites et à peu près de même hauteur.

a) Afficher les valeurs propres et l'éboulis :

```
round(res$eig, 2)
barplot(res$eig[, 2], names.arg = 1:nrow(res$eig),
        ylab = "% d'inertie", xlab = "axe")
```

Lecture. La première colonne de `res$eig` contient les valeurs propres, la deuxième le pourcentage d'inertie, la troisième le pourcentage cumulé.

✓ **Contrôle** $\lambda_1 \approx 4{,}61$.

Que vaut la somme des valeurs propres (`sum(res$eig[, 1])`) ? Pourquoi ?

b) Quel pourcentage d'inertie le plan $(1,2)$ restitue-t-il ?

c) Combien d'axes retenez-vous selon le critère de Kaiser ? selon le critère du coude ?

5. Le cercle des corrélations : interpréter les axes

a) Tracer le cercle et afficher les résultats des variables :

```
plot(res, choix = "var")          # cercle des correlations
round(res$var$cor[, 1:2], 2)      # coordonnees = correlations avec les axes
round(res$var$cos2[, 1:2], 2)     # qualite de representation
round(res$var$contrib[, 1:2], 1)  # contributions (en %)
```

À noter. Sur le cercle, la flèche en pointillés est celle du rendement, variable supplémentaire ; on l'utilisera à la question 7.

Tracer ensuite le cercle en coloriant les flèches selon leur qualité de représentation, puis selon leur contribution :

```
plot(res, choix = "var", habillage = "cos2")     # couleur = cos2 sur le plan
plot(res, choix = "var", habillage = "contrib")  # couleur = contribution au plan
```

Lecture. La couleur va du bleu (valeur faible) au rouge (valeur forte), selon l'échelle affichée à droite. Avec `"cos2"`, c'est la qualité de représentation sur le plan, $\cos^2_1 + \cos^2_2$. Avec `"contrib"`, c'est la contribution au *plan* $(1,2)$: la moyenne des contributions aux axes 1 et 2, pondérée par λ_1 et λ_2.

b) **Axe 1.** Quelles variables lui sont fortement corrélées (en valeur absolue, disons au-dessus de $0{,}8$) ? Dans quel sens ? Lesquelles y contribuent le plus (comparer à la contribution moyenne $100/8 = 12{,}5\,\%$) ? Proposer un nom pour cet axe.

c) **Axe 2.** Mêmes questions.

d) Sur le cercle colorié selon le $\cos^2$, quelle variable est la moins bien représentée sur le plan $(1,2)$? Calculer sa qualité de représentation $\cos^2_1 + \cos^2_2$, ou l'obtenir directement avec `round(rowSums(res$var$cos2[, 1:2]), 2)`. Sur quel axe se trouve le reste de son information ? (Regarder `round(res$var$cos2[, 3], 2)`.)

e) Sur le cercle colorié selon la contribution, lire l'échelle de couleurs : les contributions au plan sont-elles très différentes d'une variable à l'autre ? Comparer avec les contributions axe par axe affichées en a). Pourquoi vaut-il mieux lire les contributions *axe par axe* pour nommer les axes ?

f) **Faire le lien avec le calcul à la main.** Pour la variable `MO`, vérifier avec les valeurs affichées que :

- $\cos^2_1(\text{MO}) = r^2(\text{MO}, F_1)$;
- $\text{contrib}_1(\text{MO}) = 100 \times r^2(\text{MO}, F_1)/\lambda_1$;
- $r(\text{MO}, F_1) = \sqrt{\lambda_1}\, u_{1k}$, où u_{1k} est la coordonnée de `MO` dans le premier vecteur propre U_1, donnée par `res$svd$V[, 1]` (`MO` est la 2e variable).

g) Sur le cercle, les flèches de `P` et `K` sont presque confondues. Calculer le produit scalaire de leurs coordonnées sur le plan $(r(\text{P}, F_1)\, r(\text{K}, F_1) + r(\text{P}, F_2)\, r(\text{K}, F_2))$ et le comparer à leur corrélation $r(\text{P}, \text{K})$, calculée à la question 2 b). Pourquoi les deux valeurs sont-elles proches sans être égales ?

6. La carte des sites

a) Tracer la carte des sites coloriée selon la zone, et afficher les résultats des sites :

```
plot(res, choix = "ind", habillage = 10)    # couleur = zone (10e colonne)
round(res$ind$coord[, 1:2], 2)              # coordonnees F1, F2
round(res$ind$cos2[, 1:2], 2)               # qualite de representation
round(res$ind$contrib[, 1:2], 1)            # contributions (en %)
```

À noter. Sur la carte, les petits carrés dont le nom est écrit en italique sont les barycentres des zones (la position moyenne des sites de chaque zone).

b) Décrire la position des quatre zones sur la carte. En vous aidant de la question 5, caractériser les sols de chaque zone en une phrase.

c) Quels sont les deux sites qui contribuent le plus à l'axe 1 ? à l'axe 2 ? Comparer à la contribution moyenne $100/15 \approx 6{,}7\,\%$. Un site domine-t-il à lui seul un axe ? Tracer ensuite la carte coloriée selon la contribution au plan :

```
plot(res, choix = "ind", habillage = "contrib")  # couleur = contribution au plan
```

Les sites en rouge sont-ils bien ceux que vous avez repérés ? Où se trouvent-ils sur la carte ?

d) Afficher la qualité de représentation de chaque site sur le plan, tracer la carte coloriée selon le $\cos^2$, puis ne tracer que les sites bien représentés :

```
round(rowSums(res$ind$cos2[, 1:2]), 2)          # cos2 sur le plan (1,2)
plot(res, choix = "ind", habillage = "cos2")    # couleur = cos2 sur le plan
plot(res, choix = "ind", select = "cos2 0.6")   # sites avec cos2 > 0.6
```

Quels sites disparaissent de la carte ? Quel est le plus mal représenté ? Peut-on interpréter sa position près du centre ?

e) **Faire le lien avec le calcul à la main.** Pour Meiganga, vérifier que

$$\text{contrib}_1 = 100 \times \frac{F_1(i)^2}{I\,\lambda_1} \qquad \text{et} \qquad \cos^2_1 = \frac{F_1(i)^2}{d^2(i,G)},$$

avec `res$ind$coord["Meiganga", 1]`, `res$eig[1, 1]` et `res$ind$dist["Meiganga"]`.

f) **Retour aux données brutes.** Afficher les lignes de trois sites situés à trois coins différents de la carte (par exemple Mbouda, Meiganga, Ebolowa) avec

```
sols[c("Mbouda", "Meiganga", "Ebolowa"), ]
```

et vérifier que leurs valeurs sont cohérentes avec les noms donnés aux axes.

7. Les éléments supplémentaires : et le rendement ?

a) Afficher la corrélation du rendement avec les axes :

```
round(res$quanti.sup$cor[, 1:2], 2)
round(res$quanti.sup$cos2[, 1:2], 2)
```

Où se trouve la flèche du rendement sur le cercle ? Est-elle bien représentée ? Quels sols semblent les plus favorables au maïs ? Est-ce cohérent avec la boîte à moustaches de la question 2 ?

b) Afficher les coordonnées des barycentres des zones et leurs *valeurs-test* :

```
round(res$quali.sup$coord[, 1:2], 2)
round(res$quali.sup$v.test[, 1:2], 2)
```

Lecture. Une valeur-test supérieure à 2 en valeur absolue indique que les sites de la zone ont, sur cet axe, une position moyenne significativement différente de 0 (au seuil d'environ 5 %).

Quelles zones sont significativement d'un côté ou de l'autre de chaque axe ?

c) Obtenir la description automatique des axes :

```
dimdesc(res, axes = 1:2)
```

Lecture. Pour chaque axe, `$quanti` liste les variables significativement corrélées à l'axe (de la plus positive à la plus négative) ; `$quali` donne le R^2 de la zone, c'est-à-dire la part de la variabilité des coordonnées sur l'axe expliquée par la zone, avec sa p-valeur ; `$category` liste les zones significatives.

Le rendement est-il significativement lié à un axe ? Quelle part de la variabilité de l'axe 1 la zone explique-t-elle ? Que peut-on en conclure ?

8. Synthèse

Rédiger, pour le chef de projet, une synthèse de l'ACP : les deux grands gradients qui structurent les sols, les zones qui s'opposent, le lien avec le rendement et ce que cela suggère pour la fertilisation du maïs. Préciser les limites de l'étude.

Indication ▼ Correction ▼ [stat-0109]

Exercice 2 ACP en autonomie sous R : typologie d'exploitations familiales au Burkina Faso

Dans l'Ouest du Burkina Faso, une fédération de coopératives de producteurs veut adapter les services qu'elle propose à ses membres (conseil agricole, crédit de campagne, approvisionnement en intrants, commercialisation). Ses conseillers agricoles ont conduit une enquête auprès de **32 exploitations familiales** de trois villages : Koumbia (zone cotonnière), Bondoukuy et Dano. Avant de discuter des services avec les producteurs, ils veulent savoir *quels grands types d'exploitations* coexistent dans la zone et ce qui les distingue.

Le fichier `data_stat-0110.xlsx` (feuille `donnees`) contient une ligne par exploitation et les colonnes suivantes ; la feuille `description` précise les unités. Les données sont fictives, construites pour l'exercice.

Colonne	Signification
`Exploitation`	identifiant (E01 à E32)
`Actifs`	nombre d'actifs agricoles du ménage
`Surface`	surface cultivée (ha)
`Bovins, Petits_rum`	effectifs de bovins et de petits ruminants (têtes)
`Part_coton`	part de la surface cultivée en coton (%)
`Engrais`	dose moyenne d'engrais minéral (kg/ha)
`Rdt_cereales`	rendement moyen des céréales (kg/ha)
`Dist_marche`	distance au marché le plus proche (km)
`Revenu`	revenu agricole annuel du ménage (milliers de FCFA)
`Village`	Koumbia, Bondoukuy ou Dano
`Credit`	accès au crédit de campagne (oui/non)
`Traction`	équipement en traction animale (oui/non)

☞ Récupérer les données

Scanner le QR code avec un téléphone, ou cliquer sur le lien `data_stat-0110.xlsx`, puis télécharger le fichier avec le bouton Download raw file de la page GitHub. L'enregistrer dans le dossier `data_raw` de votre projet RStudio.

https://github.com/ant0in3g/ant0in3g.github.io/blob/main/sTa7/td-R-lab/R-lab/data_raw/data_stat-0110.xlsx

En autonomie

Cet exercice se fait sans code fourni : à vous de choisir les fonctions et les réglages, en vous appuyant sur l'exercice guidé précédent et sur le cours. **Chaque choix doit être justifié**, et chaque affirmation appuyée sur une sortie de R (valeur, graphique). Écrivez tout dans un script commenté.

Boîte à outils. `read_excel`, `as.factor`, `summary`, `cor`, `PCA`, `plot`, `dimdesc`, `tapply`, `boxplot`.

1. **Préparer les données.** Importer le fichier, mettre les identifiants en noms de lignes, déclarer correctement le type de chaque variable et faire un premier résumé du tableau.
2. **Choisir les variables.** Quelles variables mettez-vous en *actives* (celles qui définissent la ressemblance entre exploitations) ? Lesquelles en *supplémentaires* ? Justifier en particulier le statut du revenu. Faut-il réduire les variables ?
3. **Réaliser l'ACP et choisir le nombre d'axes à interpréter**, en croisant au moins deux critères.
4. **Interpréter les axes retenus** à l'aide du cercle des corrélations, des qualités de représentation, des contributions des variables et de la description automatique des axes. Donner un nom à chaque axe. Une variable est-elle liée à plusieurs axes à la fois ?
5. **Décrire la carte des exploitations.** Situer les trois villages, l'accès au crédit et l'équipement en traction animale : lesquelles de ces variables sont liées à quel axe ? Pour chaque grand type d'exploitation que vous identifiez, citer une ou deux exploitations représentatives et vérifier leur profil sur les données brutes.
6. **Le revenu.** Quels types d'exploitations ont les revenus les plus élevés ? les plus faibles ? La structure de l'exploitation suffit-elle à expliquer le revenu ?
7. **Contrôler la fiabilité de la lecture.** Repérer les exploitations mal représentées sur le plan et celles qui pèsent le plus dans la construction des axes. Ces dernières posent-elles problème ? Le vérifier.
8. **Note pour la fédération.** En une demi-page, à destination du conseil d'administration de la fédération, présenter la typologie obtenue et proposer, pour chaque type d'exploitation, des services adaptés à discuter avec les producteurs concernés. Indiquer les limites de l'analyse.

Indication ▼ [stat-0110]

Indications

Indication pour l'exercice 1 ▲

1. Si `read_excel` ne trouve pas le fichier, vérifier qu'il est bien dans le dossier `data_raw`, que RStudio a été lancé depuis le fichier `.Rproj` (`getwd()` affiche le dossier de travail) et que le chemin s'écrit avec une barre oblique `/`, jamais avec `\`. Les colonnes sont numérotées *après* la suppression de `Site`.
2. Comparer les écarts-types de `K` et de `Sable`. Les corrélations les plus fortes dépassent $0{,}8$ en valeur absolue.
3. Une variable supplémentaire est projetée *après* le calcul des axes : elle ne les influence pas.
4. L'inertie totale d'une ACP normée vaut le nombre de variables actives.
5. Pour nommer un axe, partir des variables de corrélation supérieure à $0{,}8$ en valeur absolue, en regardant leur signe. La qualité de représentation d'une variable sur le plan est la somme de ses $\cos^2$ sur les axes 1 et 2.
6. La contribution moyenne d'un site vaut $100/15$. `res$ind$dist` donne la distance $d(i, G)$, pas son carré.
7. Une variable qualitative supplémentaire se lit sur la carte des sites (barycentres des zones), pas sur le cercle.

Indication pour l'exercice 2 ▲

1. Les trois variables oui/non et le village doivent être des facteurs.
2. Une variable que l'on cherche à *expliquer* (un résultat économique, par exemple) se met plutôt en supplémentaire : on construit la typologie sans elle, puis on regarde où elle se place. Comparer les unités et les écarts-types des variables.
3. Comparer les valeurs propres à 1 et regarder l'éboulis.
4. Une flèche « en diagonale » est liée aux deux axes à la fois. Un axe ne classe pas forcément les exploitations de la meilleure à la moins bonne : il peut opposer deux façons de produire.
5. Pour lier une variable qualitative à un axe, lire la rubrique `$quali` de `dimdesc`, ou les valeurs-test (`v.test`) des modalités.
6. La fonction `tapply` calcule une moyenne par groupe, par exemple le revenu moyen selon l'accès au crédit.
7. Avec 32 exploitations, la contribution moyenne vaut $100/32 \approx 3{,}1\,\%$. L'argument `ind.sup` de `PCA` permet de refaire l'analyse sans quelques exploitations.

Corrections

Correction de l'exercice 1 ▲

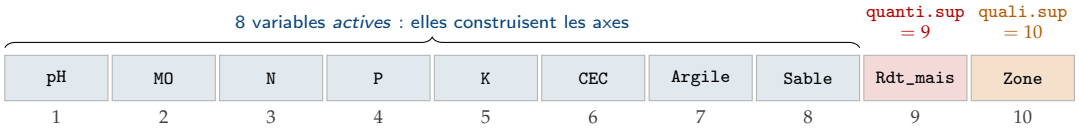
Comment lire cette correction

- Chaque réponse suit le même fil : **ce que R affiche** (tableau ou figure), **ce qu'on y lit**, puis **la conclusion agronomique**.
- Les valeurs sont celles de `FactoMineR`. Selon l'ordinateur, le signe d'un axe peut être inversé (toute la carte est alors retournée en miroir), sans conséquence sur l'interprétation.
- Les figures dessinées ici reprennent les sorties de R, avec les mêmes valeurs, pour les rendre plus lisibles.

1. Préparer la séance

c) `str(sols)` affiche `'data.frame': 15 obs. of 10 variables` : neuf colonnes `num` (de `pH` à `Rdt_mais`) et `Zone : Factor w/ 4 levels "Forêt","Hauts plateaux",...`

- Les noms des sites sont des *étiquettes*, pas une variable mesurée : en noms de lignes, ils servent à nommer les points de la carte, et `PCA` ne cherche pas à les traiter comme une donnée (une colonne de texte ferait d'ailleurs échouer la fonction).
- Après suppression de `Site`, les colonnes sont numérotées ainsi :



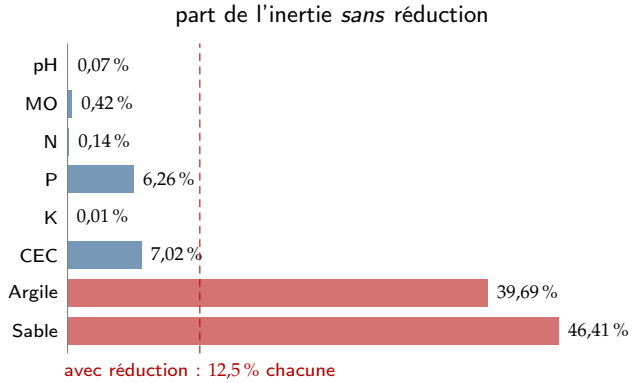
`Rdt_mais` est donc la colonne 9 et `Zone` la colonne 10 : ce sont ces numéros qu'on utilisera dans `PCA`.

2. Regarder les données avant l'ACP

a) **Unités et dispersions.** Les unités sont toutes différentes (pH sans unité, %, g/kg, mg/kg, cmol$^+$/kg) et les dispersions très inégales :

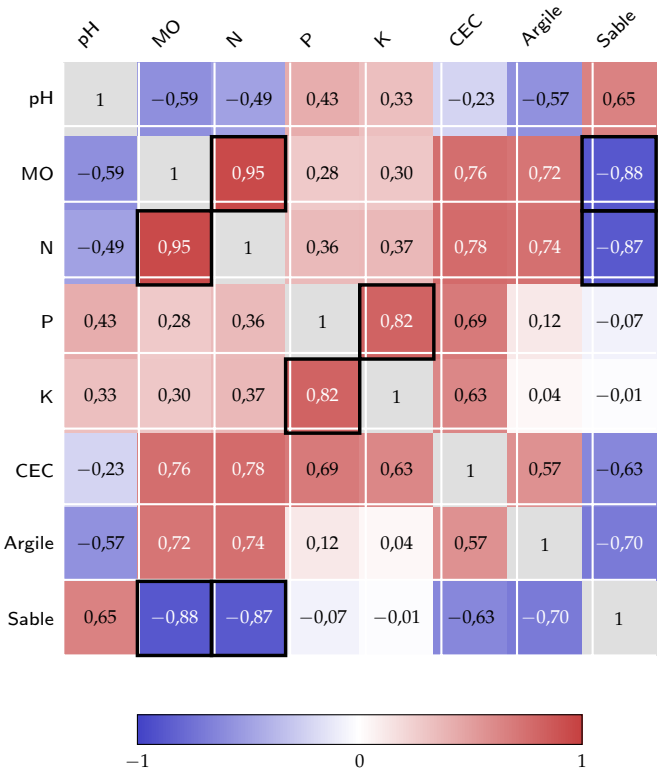
	pH	MO	N	P	K	CEC	Argile	Sable	Rdt_mais
Moyenne	5,67	3,25	1,75	11,75	0,42	11,93	30,6	47,0	2,55
Écart-type	0,47	1,14	0,66	4,41	0,16	4,67	11,10	12,01	0,68
Variance	0,22	1,29	0,43	19,43	0,03	21,79	123,26	144,14	

Sans réduction, l'inertie totale serait la somme des variances, $0{,}22 + 1{,}29 + \cdots + 144{,}14 \approx 310{,}6$ (avec les variances de `sd`, qui divisent par $I-1$; environ 290 avec la division par I de l'ACP, sans que les parts ne changent), et chaque variable y pèserait selon sa variance :



L'argile et le sable feraient à eux seuls 86 % de l'inertie : l'ACP ne verrait qu'eux, et le pH ou le potassium ne pèseraient presque rien, *simplement à cause de leur unité*. L'ACP **normée** donne à chaque variable une variance égale à 1, donc le même poids ($100/8 = 12{,}5$ % de l'inertie) : elle est indispensable ici.

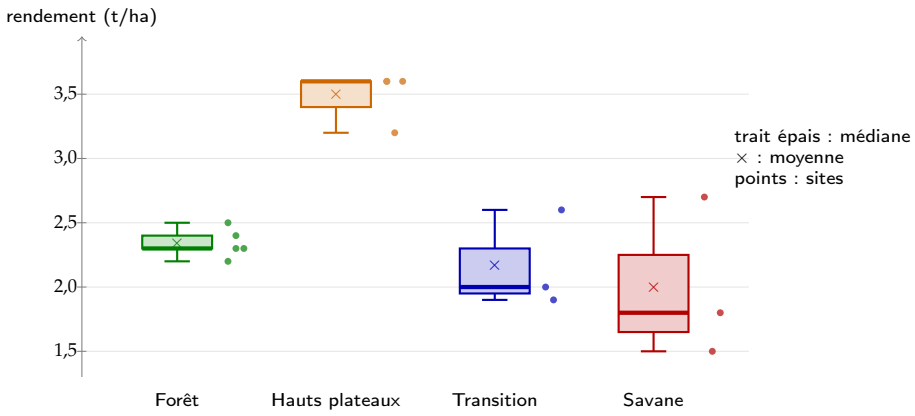
b) **Corrélations.** La matrice affichée par R se lit mieux en couleurs (rouge : liaison positive, bleu : liaison négative ; plus la couleur est foncée, plus la liaison est forte). Les quatre couples les plus corrélés sont encadrés :



Couple	r	Explication agronomique
MO – N	0,95	l'azote du sol est surtout organique : MO et N vont ensemble
MO – Sable	−0,88	les sols sableux retiennent mal la matière organique
N – Sable	−0,87	conséquence des deux lignes précédentes
P – K	0,82	P et K sont tous deux abondants dans les sols volcaniques

Ces liaisons sont logiques. Elles annoncent déjà deux « familles » de variables (matière organique et texture d'un côté, P et K de l'autre), que l'ACP va retrouver.

c) **Rendement selon la zone.** Rendements moyens : Hauts plateaux 3,50 t/ha, Forêt 2,34, Transition 2,17, Savane 2,00.



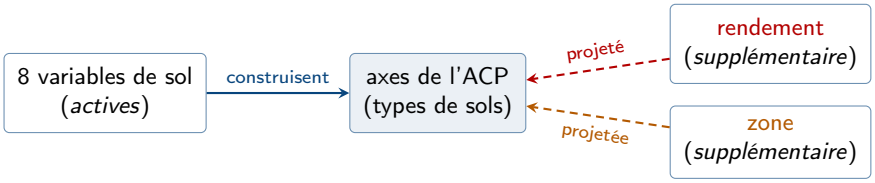
Les Hauts plateaux se détachent nettement (tous leurs sites sont au-dessus de 3 t/ha) ; les trois autres zones se chevauchent largement autour de 2 à 2,5 t/ha.

3. Lancer l'ACP

a) **Rôle des arguments.**

Argument	Rôle
`scale.unit = TRUE`	ACP *normée* : chaque variable est centrée et réduite (question 2 a)
`quanti.sup = 9`	la colonne 9 (`Rdt_mais`) est une variable quantitative *supplémentaire*
`quali.sup = 10`	la colonne 10 (`Zone`) est une variable qualitative *supplémentaire*
`graph = FALSE`	pas de graphique automatique : on les tracera nous-mêmes, un par un

b) **Pourquoi le rendement en supplémentaire ?** La question posée est « les types de sols expliquent-ils le rendement ? ». Il faut donc construire la typologie des sols *sans* le rendement, puis regarder où il se place :



En variable active, le rendement participerait à la définition des axes, et le raisonnement deviendrait circulaire (« les axes sont liés au rendement… parce qu'on les a construits avec lui »). Le rendement dépend en outre de bien d'autres facteurs que le sol (pluie, variété, conduite de la culture). Les éléments supplémentaires ne modifient pas les axes : ils sont placés ensuite, pour aider à les interpréter.

c) **Que contient `res` ?** `names(res)` affiche les sept éléments de la liste :

```
"eig"  "var"  "ind"  "svd"  "quanti.sup"  "quali.sup"  "call"
```

`summary(res)` affiche un résumé de toute l'ACP : le tableau des valeurs propres, puis, pour les premiers individus et les premières variables, les coordonnées (`Dim.1`, …), les contributions (`ctr`) et les $\cos^2$, et enfin les résultats des éléments supplémentaires.

4. Combien d'axes ?

a) **Valeurs propres.**

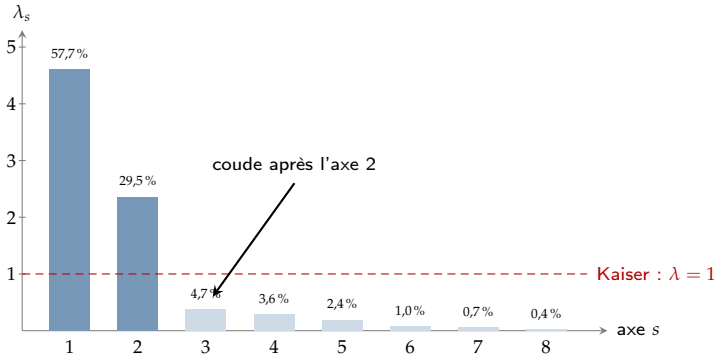
Axe	1	2	3	4	5	6	7	8
Valeur propre λ_s	**4,61**	**2,36**	0,38	0,29	0,19	0,08	0,06	0,03
% d'inertie	57,7	29,5	4,7	3,6	2,4	1,0	0,7	0,4
% cumulé	57,7	87,2	91,9	95,6	97,9	98,9	99,6	100

✓ **Contrôle** $\lambda_1 \approx 4{,}61$; le pourcentage est $100 \times \lambda_s / 8$, par exemple $100 \times 4{,}613/8 = 57{,}7\,\%$.

`sum(res$eig[, 1])` vaut **8** : en ACP normée, l'inertie totale est égale au nombre de variables actives, car chacune a une variance égale à 1.

b) Le plan $(1, 2)$ restitue $57{,}7 + 29{,}5 = \mathbf{87{,}2}\,\%$ de l'inertie : la carte sera fidèle pour la plupart des sites.

c) **Nombre d'axes.**



- *Kaiser* : $\lambda_1 = 4{,}61$ et $\lambda_2 = 2{,}36$ dépassent 1, $\lambda_3 = 0{,}38$ non : **deux axes**. (Chacun des deux premiers axes résume plus que deux variables de départ.)
- *Coude* : la chute est forte jusqu'à l'axe 2, puis les barres deviennent petites et presque égales : le coude est très net après le deuxième axe, même conclusion.

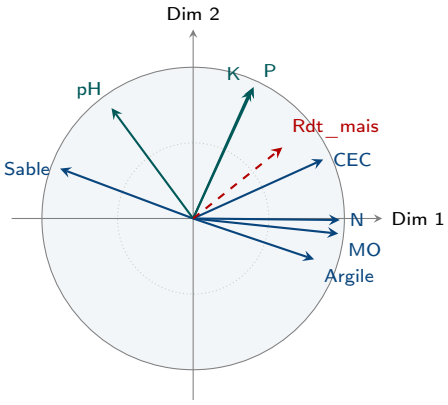
On interprète donc le plan $(1,2)$.

5. Le cercle des corrélations

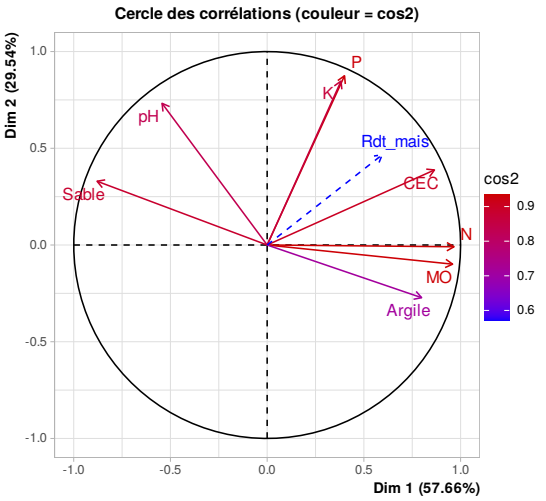
a) **Résultats des variables** (en gras : les contributions supérieures à la moyenne $100/8 = 12{,}5\,\%$) :

	Coordonnées		Qualité ($\cos^2$)			Contribution (%)	
Variable	$r(X_k, F_1)$	$r(X_k, F_2)$	axe 1	axe 2	plan	axe 1	axe 2
pH	−0,54	0,73	0,30	0,54	0,83	6,4	**22,7**
MO	0,96	−0,10	0,92	0,01	0,93	**19,9**	0,4
N	0,97	−0,01	0,93	0,00	0,93	**20,2**	0,0
P	0,40	0,87	0,16	0,77	0,93	3,5	**32,4**
K	0,38	0,85	0,15	0,72	0,86	3,2	**30,4**
CEC	0,86	0,39	0,75	0,15	0,90	**16,2**	6,4
Argile	0,80	−0,27	0,64	0,07	0,71	**13,8**	3,1
Sable	−0,88	0,33	0,77	0,11	0,88	**16,8**	4,6
Total			$4{,}61 = \lambda_1$	$2{,}36 = \lambda_2$		100	100
Rdt_mais (sup.)	0,59	0,47	0,35	0,22	0,57	—	—

✓ **Contrôle** chaque colonne de $\cos^2$ totalise la valeur propre de l'axe, et chaque colonne de contributions totalise $100\,\%$. Le rendement, supplémentaire, n'a pas de contribution : il n'a pas participé à la construction des axes.

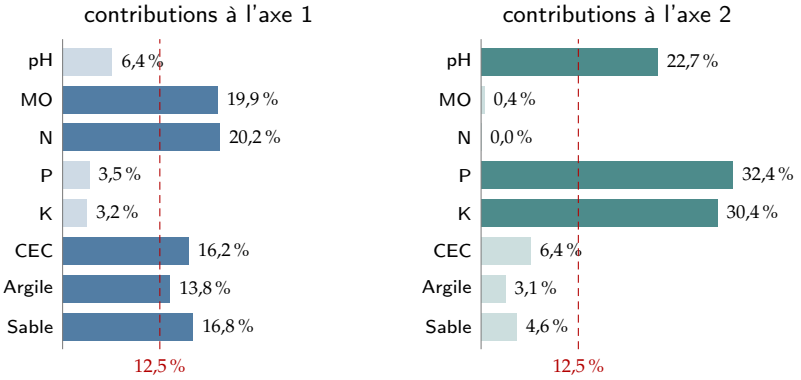


Cercle des corrélations (Dim 1 : 57,7 % ; Dim 2 : 29,5 %). En bleu : variables de l'axe 1 ; en vert : de l'axe 2.



Le même cercle sous R, couleur = $\cos^2$ (`habillage = "cos2"`).

Les contributions, axe par axe, se lisent d'un coup d'œil en barres :

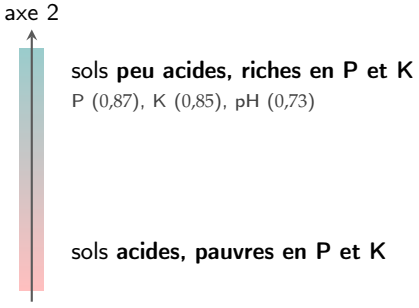


b) **Axe 1 (57,7 %).** Corrélations fortes et positives avec N (0,97), MO (0,96), CEC (0,86) et Argile (0,80) ; forte et négative avec Sable (−0,88). Ce sont aussi les cinq variables qui y contribuent plus que la moyenne (87 % de l'axe à elles cinq).



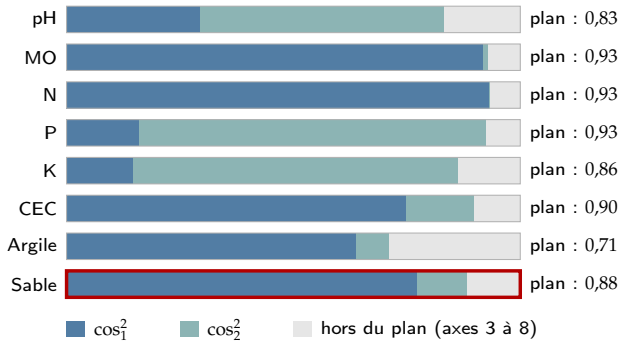
L'axe 1 oppose les sols *argileux, riches en matière organique et à forte CEC* (à droite) aux sols *sableux et pauvres* (à gauche) : c'est un axe « **réserve organique et texture** ».

c) **Axe 2 (29,5 %).** Corrélations fortes et positives avec P (0,87), K (0,85) et, dans une moindre mesure, pH (0,73) ; ces trois variables font $22{,}7 + 32{,}4 + 30{,}4 = 85{,}5\,\%$ de l'axe.



Il oppose, *à niveau organique donné*, des sols peu acides et riches en P et K (en haut) à des sols acides et pauvres en P et K (en bas) : c'est un axe « **acidité et richesse minérale** ». Les deux axes étant non corrélés, ce second gradient est indépendant du premier.

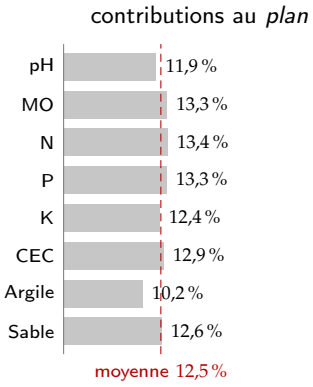
d) **Qualité de représentation.** Sur le cercle colorié selon le $\cos^2$ (figure de droite en a), la flèche de l'argile est la plus bleue :



L'argile est la moins bien représentée sur le plan : $\cos^2 = 0{,}64 + 0{,}07 = \mathbf{0{,}71}$; MO, N et P sont les mieux représentées (0,93). Le reste de l'information de l'argile est surtout sur l'axe 3 ($\cos^2_3 = 0{,}26$) : elle construit à elle seule 69 % de cet axe.

e) **Le piège des couleurs « contribution ».** Les couleurs paraissent contrastées (l'argile en bleu, MO et N en rouge), mais c'est parce que R étire l'échelle de couleurs entre la plus petite et la plus grande valeur. L'échelle affichée à droite ne va que de 10 à 13 % : les contributions au *plan* sont en réalité toutes proches de la moyenne 12,5 %. La contribution au plan est la moyenne des contributions aux deux axes, pondérée par les valeurs propres ; par exemple pour MO :

$$\frac{4{,}613 \times 19{,}93 + 2{,}363 \times 0{,}42}{4{,}613 + 2{,}363} = \frac{91{,}9 + 1{,}0}{6{,}976} \approx 13{,}3\,\%.$$



Pourtant, axe par axe (barres de la question a), les contributions sont très tranchées : MO et N font 40 % de l'axe 1 mais presque rien sur l'axe 2, P et K font 63 % de l'axe 2 mais presque rien sur l'axe 1. En mélangeant les deux axes, la contribution au plan efface ces différences : chaque variable compte « un peu » pour le plan. Pour nommer un axe, il faut savoir quelles variables ont construit *cet* axe : on lit les contributions **axe par axe**. La couleur selon la contribution est plus utile pour les individus (question 6).

f) **Lien avec le calcul à la main (MO).** R donne $r(\text{MO}, F_1) = 0{,}959$, $\lambda_1 = 4{,}613$ et, dans `res$svd$V[2, 1]`, $u_{12} = 0{,}446$:

$$\cos^2_1(\text{MO}) = r^2(\text{MO}, F_1) = 0{,}959^2 = 0{,}919 \qquad \text{(affiché : 0,92)}$$

$$\text{contrib}_1(\text{MO}) = 100 \times \frac{r^2(\text{MO}, F_1)}{\lambda_1} = 100 \times \frac{0{,}919}{4{,}613} = 19{,}9\,\% \qquad \text{(affiché : 19,9)}$$

$$r(\text{MO}, F_1) = \sqrt{\lambda_1}\, u_{12} = \sqrt{4{,}613} \times 0{,}446 = 2{,}148 \times 0{,}446 = 0{,}958 \qquad \text{(affiché : 0,96)}$$

Le logiciel applique exactement les formules vues à la main dans les exercices 1 et 2.

g) **P et K.** Produit scalaire des deux flèches sur le plan :

$$r(\mathrm{P},F_1)\,r(\mathrm{K},F_1)+r(\mathrm{P},F_2)\,r(\mathrm{K},F_2)=0{,}400\times0{,}382+0{,}875\times0{,}847=0{,}153+0{,}741=\mathbf{0{,}89},$$

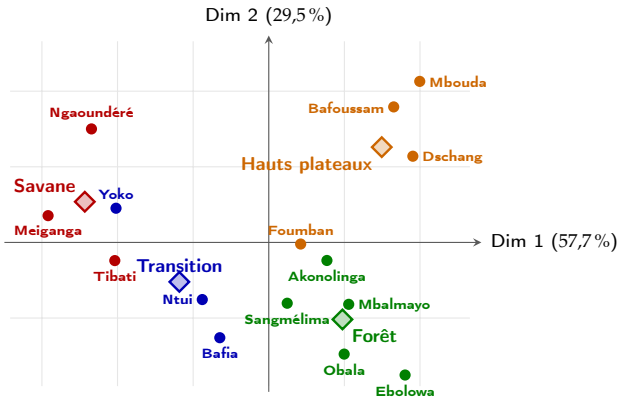
contre $r(\mathrm{P},\mathrm{K})=\mathbf{0{,}82}$. La corrélation est en fait la somme de ces produits sur *tous* les axes ; les axes suivants apportent de petites corrections, ici négatives :

Axe s	1	2	3	4	5	6	7	8
$r(\mathrm{P},F_s)\,r(\mathrm{K},F_s)$	0,153	0,741	−0,023	−0,016	−0,026	−0,014	0,003	0,001
somme cumulée	0,153	**0,894**	0,871	0,856	0,830	0,816	0,819	**0,820**

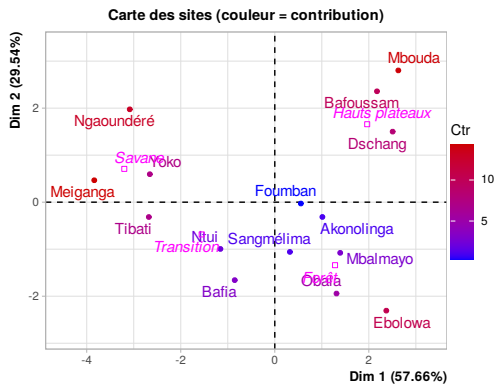
Les deux valeurs sont proches parce que les deux flèches sont longues ($\cos^2$ sur le plan 0,93 et 0,86), mais pas égales parce qu'une petite partie de l'information de P et de K se trouve sur les autres axes. Règle de lecture : deux flèches confondues et proches du cercle indiquent une forte corrélation positive, mais l'angle sur le plan n'en donne qu'une valeur approchée.

6. La carte des sites

a) **Carte des sites.**



Carte des sites, couleur = zone ; les losanges sont les barycentres des zones.



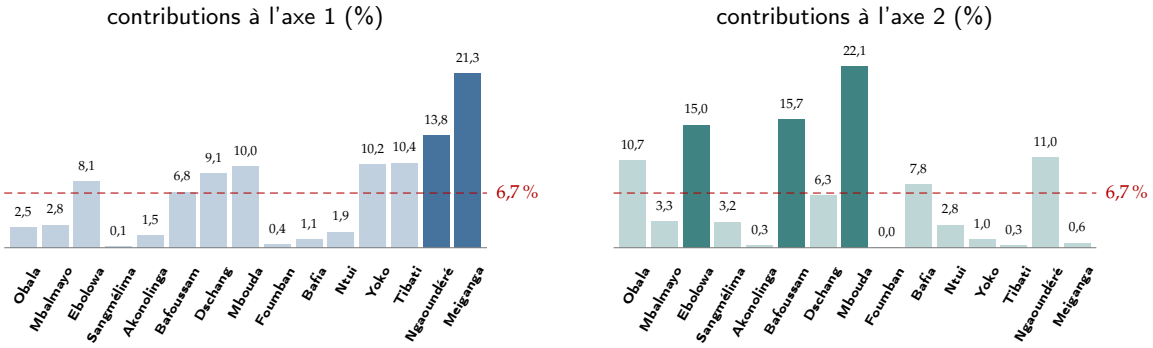
La carte sous R, couleur = contribution (`habillage = "contrib"`, voir c).

b) **Les quatre zones.** En combinant la position de chaque zone sur la carte et les noms des axes (question 5), puis en contrôlant sur les moyennes brutes par zone :

Zone	barycentre	pH	MO (%)	P (mg/kg)	K	Sable (%)	Rdt (t/ha)
Forêt	(1,28 ; −1,34)	**5,18**	**4,06**	9,48	0,36	38,8	2,34
Hauts plateaux	(1,97 ; 1,66)	5,80	**4,03**	**17,60**	**0,60**	40,5	**3,50**
Transition	(−1,56 ; −0,69)	5,87	2,60	8,57	0,30	51,7	2,17
Savane	(−3,20 ; 0,71)	6,10	1,53	10,93	0,42	**64,7**	2,00
Moyenne	(0 ; 0)	5,67	3,25	11,75	0,42	47,0	2,55

- **Forêt** (en bas à droite) : sols argileux et organiques, mais *acides* et pauvres en P et K (sols ferrallitiques désaturés).
- **Hauts plateaux** (en haut à droite) : organiques *et* riches en P et K, peu acides (sols volcaniques de l'Ouest).
- **Savane** (à gauche) : sableux et pauvres en matière organique, mais peu acides.
- **Transition** (à gauche, un peu en bas) : position intermédiaire, sols plutôt sableux et pauvres.

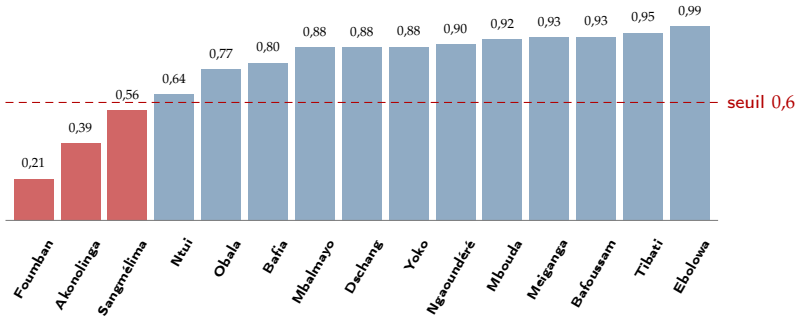
c) **Contributions des sites.** La contribution moyenne est $100/15 \approx 6{,}7\,\%$ (ligne pointillée) :



Axe 1 : Meiganga (21,3 %) et Ngaoundéré (13,8 %), soit deux à trois fois la moyenne : les sols les plus sableux et les plus pauvres. Axe 2 : Mbouda (22,1 %) et Bafoussam (15,7 %), suivis d'Ebolowa (15,0 %) à l'autre extrémité. **Aucun site ne domine à lui seul un axe** : les deux gradients sont portés par plusieurs sites.

Sur la carte coloriée selon la contribution au plan (figure de droite en a), les sites les plus rouges sont Meiganga (14,3 %), Mbouda (14,1 %), Ngaoundéré (12,8 %), Ebolowa (10,5 %) et Bafoussam (9,8 %) : ce sont bien ceux repérés ci-dessus. Ils sont tous *loin du centre*, aux extrémités du nuage ; les sites proches du centre (Foumban 0,3 %, Akonolinga, Sangmélima) sont bleus. La contribution dépend de la distance au centre, au carré.

d) **Qualité de représentation.** Les $\cos^2$ sur le plan, triés, avec le seuil 0,6 utilisé par `select = "cos2 0.6"` :



Les trois sites sous le seuil disparaissent de la carte : Foumban (0,21), Akonolinga (0,39) et Sangmélima (0,56). Tous sont près du centre de la carte. Foumban est le plus mal représenté : sa position près du centre ne doit *pas* être interprétée comme celle d'un « site moyen », car 79 % du carré de sa distance au centre se trouve sur les axes suivants (notamment l'axe 3, $\cos^2_3 = 0{,}22$).

e) **Lien avec le calcul à la main (Meiganga).** R donne $F_1 = -3{,}840$, $\lambda_1 = 4{,}613$ et $d(i,G) = 4{,}006$:

$$F_1^2 = (-3{,}840)^2 = 14{,}75, \qquad d^2(i,G) = 4{,}006^2 = 16{,}05,$$

$$\text{contrib}_1 = 100 \times \frac{F_1^2}{I\,\lambda_1} = 100 \times \frac{14{,}75}{15 \times 4{,}613} = 100 \times \frac{14{,}75}{69{,}2} \approx 21{,}3\,\% \qquad \text{(affiché : 21,3)}$$

$$\cos^2_1 = \frac{F_1^2}{d^2(i,G)} = \frac{14{,}75}{16{,}05} \approx 0{,}92 \qquad \text{(affiché : 0,92)}$$

On retrouve les valeurs de R.

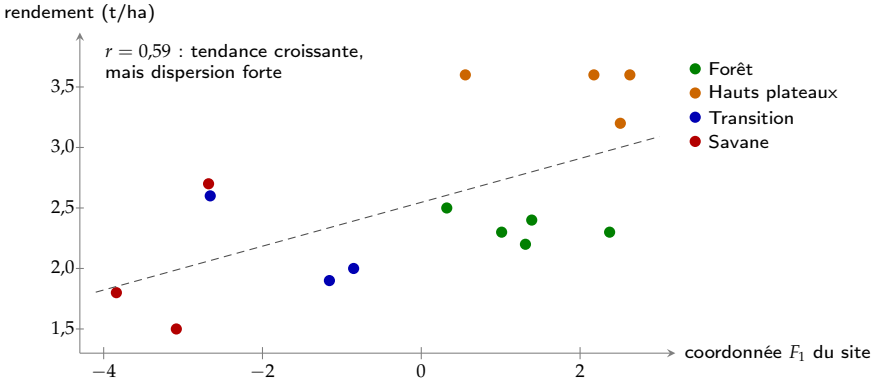
f) **Retour aux données brutes.** Les trois sites, comparés à la moyenne (**gras bleu** : nettement au-dessus, **gras rouge** : nettement au-dessous) :

Site	**position**	pH	MO	N	P	K	CEC	Argile	Sable
Mbouda	en haut à droite	**6,2**	**4,2**	**2,57**	**21,1**	**0,73**	**17,3**	44	40
Meiganga	à gauche	6,4	**0,8**	**0,73**	10,8	0,29	**5,8**	21	**65**
Ebolowa	en bas à droite	**4,9**	**4,5**	**2,50**	**7,3**	0,30	14,3	**46**	33
Moyenne		5,67	3,25	1,75	11,75	0,42	11,93	30,6	47,0

Mbouda est organique *et* très riche en P et K (maximums du tableau) ; Meiganga est très sableuse et pauvre (minimum de MO) ; Ebolowa est organique et argileuse (maximum de MO) mais acide et pauvre en P. Les données brutes confirment les noms donnés aux axes.

7. Les éléments supplémentaires

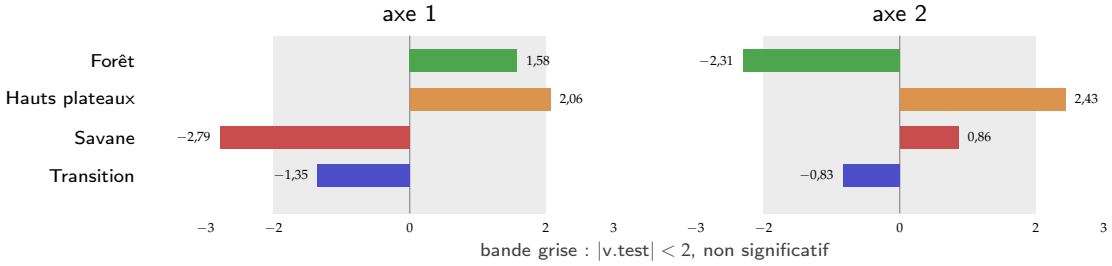
a) **Le rendement.** $r(\text{Rdt}, F_1) = 0{,}59$ et $r(\text{Rdt}, F_2) = 0{,}47$: la flèche du rendement pointe vers le *haut à droite*, du côté des Hauts plateaux. On le voit aussi en traçant le rendement de chaque site en fonction de sa coordonnée sur l'axe 1 :



Les meilleurs rendements s'observent sur les sols à la fois organiques *et* riches en P et K, peu acides (3,2 à 3,6 t/ha à Bafoussam, Dschang, Mbouda, Foumban). Les sols de forêt, pourtant organiques, plafonnent vers 2,2 à 2,5 t/ha (acidité, carence en P) ; les sols de savane et de transition (1,5 à 2,7 t/ha) sont pauvres. C'est cohérent avec la boîte à moustaches. La flèche est cependant assez loin du cercle ($\cos^2 = 0{,}35 + 0{,}22 = 0{,}57$) : le sol n'explique qu'une partie du rendement.

b) **Barycentres et valeurs-test.**

	Coordonnées		Valeur-test	
Zone	Dim 1	Dim 2	Dim 1	Dim 2
Forêt	1,28	−1,34	1,58	**−2,31**
Hauts plateaux	1,97	1,66	**2,06**	**2,43**
Savane	−3,20	0,71	**−2,79**	0,86
Transition	−1,56	−0,69	−1,36	−0,83



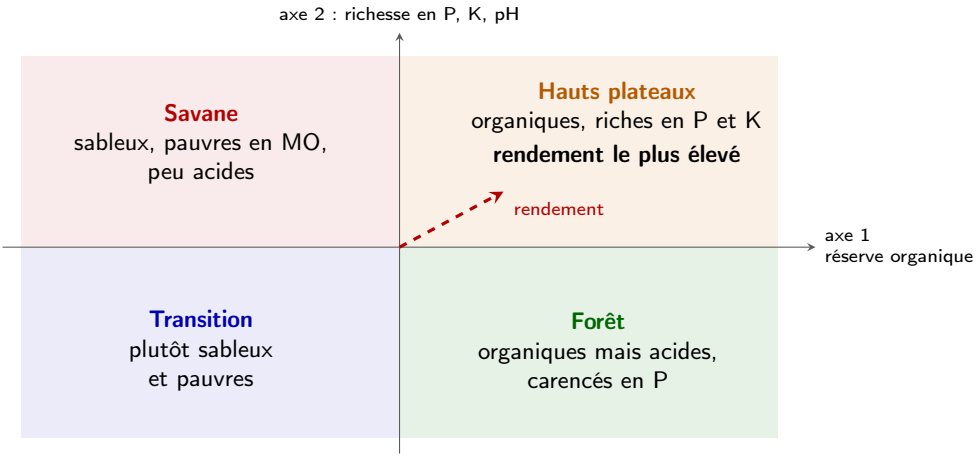
Les barres qui sortent de la bande grise ($|\text{v.test}| > 2$) sont significatives. Axe 1 : les Hauts plateaux sont significativement à droite, la Savane significativement à gauche. Axe 2 : les Hauts plateaux sont significativement en haut, la Forêt significativement en bas. La Transition n'est significative sur aucun axe (seulement trois sites, en position intermédiaire).

c) **Description automatique des axes.** `dimdesc` donne :

	Axe 1	Axe 2
`$quanti`	N (0,97), MO (0,96), CEC (0,86), Argile (0,80), Rdt_mais (0,59, $p = 0{,}02$) ; pH ($-0{,}54$), Sable ($-0{,}88$)	P (0,87), K (0,85), pH (0,73)
`$quali` (Zone)	$R^2 = 0{,}89$, $p \approx 10^{-5}$	$R^2 = 0{,}65$, $p \approx 0{,}008$
`$category`	Hauts plateaux ($+$), Savane ($-$)	Hauts plateaux ($+$), Forêt ($-$)

Le rendement n'est significativement lié qu'à l'axe 1 : sa corrélation avec l'axe 2 ($r = 0{,}47$) n'est pas significative avec seulement 15 sites ($p \approx 0{,}08$). La zone explique 89 % de la variabilité des coordonnées sur l'axe 1 : la typologie des sols colle très étroitement à la zonation agro-écologique.

8. Synthèse



Exemple de synthèse. Les sols des quinze sites s'organisent selon deux gradients qui résument 87 % de l'information. Le premier oppose les sols argileux et riches en matière organique de la forêt et des hauts plateaux aux sols sableux et pauvres des savanes de l'Adamaoua. Le second oppose les sols volcaniques des hauts plateaux, peu acides et riches en phosphore et potassium, aux sols acides et carencés en phosphore de la zone forestière. Le rendement du maïs suit surtout le premier gradient et il est maximal sur les hauts plateaux, qui cumulent les deux atouts. En zone forestière, la priorité serait la correction de l'acidité (chaulage) et l'apport de phosphore ; en savane, l'apport de matière organique et une fertilisation complète. Limites : quinze sites seulement, un seul prélèvement par site, des données fictives, et un rendement qui dépend aussi du climat et des pratiques ; l'ACP décrit des associations, elle ne prouve pas de causalité.

Ecole supérieure d'agro-développement international

istom



Étudiant(e) (Nom, Prénom) : .....

Promotion, groupe : .....

Email : .....