Chaque *bullet point* pourrait correspondre à une slide. Le diapo ici : <https://www.overleaf.com/5475982546jccbfgrtknhr#a17328>

Quelles conclusions tirer de nos études préliminaires ?

• **Variabilité sur la pluie ajoutée au sein de pays : l’information de la pluie aurait sinon été redondante avec celle du pays**

• **On crée 6 clusters qui pourront se substituer à l’information (pluie, température)** avec la méthode des k-means (meilleure que la CAH) : Inertia k-means: 3190.019 < 3617.478 et Inertia k-means by\_country: 526.5913 < 531.2891

⚠️ 10 pays ("Azerbaijan" "Brazil" "Egypt" "Iraq" "Lesotho" "Malawi" "Montenegro" "Romania" "Rwanda" "Zambia") sont attribués à deux clusters : on uniformise cela en les associant entièrement au cluster auquel ils ont été majoritairement associés

•On a un **dataframe data contenant 8 variables** : 4 catégorielles (type de culture, pays, année, cluster) et 4 continues (température, pesticides, pluie, rendement)

Par pays (101), par année (23, de 1990 à 2013 inclus, avec l’année 2003 manquante), par culture (10 au total), valeur de rendement [hg/ha], avec pour information les températures moyennes annuelles [°C], la quantité annuelle de pluie [mm], la quantité de pesticides utilisée [T]

13130 lignes (< 101\*23\*10 car 5.7 cultures par pays en moyenne)

• On a un **dataframe** **by\_country contenant 16 variables**: 3 catégorielles (pays, année, Cluster), 13 continues (10 rendements selon la culture), et {pluie, température, pesticides})

Par pays, par année, valeur de rendement [hg/ha] de chacune des cultures, avec pour information les températures moyennes annuelles [°C], la quantité annuelle de pluie [mm], la quantité de pesticides utilisée [T]

2250 lignes (< 101\*23 : il manque des années pour certains pays)

• Le **boxplot** **mean yield comparison for different crops**, issu de by\_country, et le **graphique yield(years)**, issu de data, montrent la variabilité des rendements.   
Le boxplot montre une variabilité en fonction des cultures mais aussi en fonction des pays pour chaque culture, notamment sur la pomme de terre. Le graphique yield(years) montre la légère croissance globale des rendements, d’environ 25% entre 1973 et 2013.

**Ensemble, ces deux graphes révèlent une nouvelle information**: **les cultures auraient pu être classées par rendement**, la pomme de terre ayant un très fort rendement, Manioc-Igname-Patates douces-Plantains un rendement moyen, et Blé-Maïs-Riz-Soja-Sorgho un faible rendement

• En raisonnant avec le **boxplot yield(cluster)**, on remarque qu’on a **homogénéisé** les variations de rendement, de telle sorte que : **pour chacun des groupes pluie-température**, **les rendements semblent globalement similaires : on peut s’attendre à un effet modéré des facteurs météorologique sur les rendements si l’on regarde toute culture confondue**

Seul le groupe 4 sort l’ordinaire, avec une moyenne de moyenne de température annuelle de 9°C et une moyenne de pluie annuelle de 748 mm/an, correspondant aux pays les plus froid (**montrer la carte associée**)

• Enfin, le **pairwiseplot résume très bien de nombreuses informations**, à savoir que**:**

* Les clusters (pluie-température) correspondent à des distributions bien définies
* **Les courbes de densité de température et pluie ne suivant aucune loi connue : nous ne pouvions donc pas utiliser de méthode de classification probabiliste comme les mixture-model, le k-means était donc le meilleur choix**
* Le Brésil, appartenant au cluster 1, utilise une quantité de pesticide bien plus grande que les autres pays
* **On n’observe aucune corrélation probante entre nos variables !!**

• Cette absence de corrélations se retrouve aussi au sein même de la classe de rendements : nous avons exécuté une **PCA sur les 10 colonnes de rendements de by\_country, correspondant chacune à une culture, pour observer de potentielles corrélations entre les rendements.** Malheureusement,2 dimensions n’expliquent même pas 50% de la variance : les critères du coude, de Kaiser et de la moyenne empirique indique qu’il faut en utiliser au moins 3.

La technique en composantes principales reproduit avec parcimonie la variation totale d'un grand nombre de variables (pour fixer les idées, dans les cas les plus courants : de 10 à 40) en un nombre sensiblement plus restreint de dimensions 🡪 ici, on va privilégier la réduction de variables, car aucune des dimensions n’est satisfaisante

• La **PCA** **sur les variables de data**, auxquelles on ajoute le rendement, **permet de se rendre compte des colinéarités existantes** : **la dimension 1 représente la température, et est plutôt fortement décorrélée de la variable de pesticides**. **Il n’y a en effet pas de raison évidente pour que ces deux variables soient associées.** La variable « pesticides » possède une direction similaire à celle des rendements, bien qu’elle soit mal représentée : **nous pouvons nous attendre à ce qu’elle ait un pouvoir explicatif sur la variable de rendements**. Notons néanmoins que cette PCA n’est pas vraiment pertinente, au sens où il est peu intéressant de réduire 4 variables sous deux dimensions : ces deux dimensions n'expliquent d’ailleurs à elles seules que 63% de la variance totale.

• **On va exclure certaines cultures : "Soybeans", "Cassava", "Sweet.potatoes", "Plantains.and.others", "Yams"**

**Raisons :**

* **Absentes de 45% des pays étudiés** (Igname et Plantain jusqu’à 80%)
* Pour gérer les NA, on ne **peut pas utiliser de méthodes inférentielles** (comme le bootstrap ou le jackknife), car le problème n’est pas un manque de données mais une inexistence des données : il serait absurde de créer artificiellement une valeur de rendement pour le manioc en France, étant donné qu’on ne cultive pas de manioc en France. Il faudrait donc soit avoir un jeu de données présentant beaucoup de données manquantes, soit un tableau by\_country très réduit, seule une poignéee de pays hébergeant en réalité toutes les cultures étudiées.
* La **PCA sur les 10 colonnes de rendements n’était pas concluante : effectuer une PCA sur les 5 cultures majoritaires donne des résultats bien plus significatifs !** Le rendement du maïs est très bien expliqué par la dimension 1, les rendements de riz et de sorgho sont corrélés d’une part, ceux du blé et des pommes de terre aussi dans une moindre mesure. **On s’attend** donc **à ce que les régressions sur les rendements du riz fonctionnent aussi pour celle du sorgho**.

**On s’intéressera donc aux pays cultivateurs de Maize, Potatoes, Rice, Wheat and Sorghum**

(df somecult pour les pays cultivant au moins l’une des 5, et somecultnona pour ceux cultivant les 5 à la fois)

**On crée un dataframe contenant 38 pays, correspondant aux pays ayant cultivé les 5 cultures ci-dessus durant les 23 années d’observations** (fullscnona), et un avec 85 pays, possédant au moins une des 5 cultures pendant 23 années (fullscdata) : il est ainsi créé un **plan d’expérience complet**

• On cherche alors à savoir :

(Faire un tableau récapitulatif des modèles qui n’ont pas fonctionné ? Variables entrées / R2 / p-valeurs du test de type II / Résidus linéaires / Loi normale / Homoscédasticité / Points aberrants – comme dans empreinte\_écologique)

# test de Shapiro ou de Kolmogorov-Smirnov pour l’hypothèse de normalité (Q-Q plot)

* **Au sein de pays ayant les mêmes conditions climatiques, la quantité de pesticide utilisée influence-t-elle le rendement des cultures, pour une culture donnée ?**

Sur fullscnona

1. Régression linéaire simple (pesticides) en filtrant les données pour ne garder qu’une seule culture à la fois (faire une reg lin sur maïs et une reg lin sur pommes de terre), pour deux cluster différents en terme de conditions climatiques (genre cluster 3 et 4) 🡺 **Boucle sur toutes les cultures et tous les clusters**

**La transformation log(Crop) ~log(pest) est celle qui rend le mieux le modèle homocédastique** - aucune transformation (racine carrée, réciproque, boxcox) ne fait sensiblement mieux

**Pourtant, même pour la meilleure régression (R^2 = 76%), correspondant aux pommes de terre du cluster 1(insérer graphe diagnostic), les erreurs ne sont pas indépendantes** – on observe une légère hyperbole, l’hypothèse d’homocédasticité n’est pas vraiment complétée. Néanmoins, le Q-Quantile-Quantile plot et un test de Shapiro nous indiquent qu’elles sont gaussiennes, et le retrait des observations 12,13 et 90 pourrait dans ce cas améliorer nos la complétion des hypothèses.

Très grande variabilité de résultats selon les cultures et les clusters (**insérer graphe r2**):

**Maïs, Pomme de terre et Sorgho plutôt bien expliqués par la régression simple** (> 50% avec une p\_value inférieure à 1‰) **dans les pays où il fait le plus chaud et où la pluie est abondante**, comme la Guyane, le Brésil ou la Papouasie-Nouvelle-Guinée. On remarque tout de même que cette **régression simple est insuffisante pour expliquer les rendements dans un cas général**, plus de la moitié des régressions effectuées n’atteignant même pas R^2 = 25%

**Réponse à la question posée : ça dépend des pays et des cultures, mais globalement non une régression linéaire simple ne suffit pas**

* **Comment expliquer les variations de rendement selon les variables disponibles ?**

Sur fullscdata

1. Régression linéaire multiple avec (pluie/température/pesticides)

**Modèle qui complète le mieux les hypothèses d’un modèle linéaire :**

Yield\_i = a.log(pest)\_i + b.log(temp)\_i + c.rain\_i

Ce modèle, effectué sur les cinq cultures confondues, dans les pays possédant 23 années observations, complète parfaitement les hypothèses d’homocédasticité, et d’indépendance des résidus. Quant à leur normalité, même si un test d’Anderson-Darling indique que l’hypothèse n’est pas respectée, le Q-Q plot représente une droite suffisamment correcte pour que nous nous en contentions**. Le modèle permet de rejeter H\_0 au seuil 1‰, et indique que l’effet des pesticides et de la température est 4 ordres de grandeur plus grand que celui de la pluie, bien que celle-ci reste statistiquement significative, et ne soit pas rejetée lors de la soumission du modèle au critère d’Akaike**. Nous soupçonnons la plus grande faiblesse prédictive de la pluie comme issue de sa très faible variabilité au cours des années, qui malgré notre introduction de quelques variations, fait d’elle une variable quasi-ordinale plutôt que continue.

Néanmoins, ce **modèle n’explique que 14% de la variabilité du logarithme du rendement, autant dire rien du tout** !

1. Régression linéaire générale avec année/Item/Cluster/pesticides

Cluster a l’avantage d’intégrer les paramètres pluie et température dans « Area », et d’en réduire les modalités

Du modèle complet log(yield) ~ Year \* Item \* Cluster \* log(pest), on arrive en retirant itérativement par l’analyse du test de type II à un modèle statistiquement similaire de la forme :

log(yield) ~ Year + Item + Cluster + log(pest) + Item:Cluster + Year:log(pest) + Item:log(pest) + Cluster:log(pest) + Item:Cluster:log(pest)

Avec un **R^2 de 76%** (p-value: < 2.2e-16) **et des hypothèses quasi complétées (seul le test d’Anderson-Darling annonce une non-normalité de la distribution des résidus)**

Néanmoins, **en retirant le paramètre de l’année et l’interaction triple**, **on conserve un R^2 quasi identique** (**75%**), de même pour les hypothèses

🡪 log(yield) ~ Item + Cluster + log(pest) + Item:Cluster + Item:log(pest) + Cluster:log(pest)

Les interactions peuvent se lire de la manière suivante : la relation entre le rendement et le type de culture dépend d’une part de la zone dans lequel cette culture pousse, d’autre part de la quantité de pesticides qui lui est appliquée. La relation entre le rendement et la quantité de pesticide dépend de la zone dans lequel

1. ANCOVA à deux facteurs : étude de l’effet du cluster et de la culture, après contrôle du volume de pesticides

<https://www.datanovia.com/en/lessons/ancova-in-r/>

* **Comment le pays et l’année affectent les rendements de chaque culture, en prenant en compte la température et la quantité de pluie tombée ?**

Sur fullscnona

1. **MANCOVA** - Analyse de la Covariance multivariée (« Multivariate analysis of covariance ») est une extension de l'Analyse de la Covariance (méthode [ANCOVA](https://fr.wikipedia.org/wiki/ANCOVA)) pour couvrir les cas où il y a plus d'une variable dépendante et où les variables dépendantes ne peuvent pas être simplement combinées

Conclusion :

Effet de la pluie non-conséquent, mais donnée globalement redondante avec celle du pays au vu de sa faible variabilité intra-pays, et donc redondante avec l’information du cluster

Effets principaux sur le rendement : type de culture (logique) et quantité de pesticides

Résidus n’ont jamais été normaux, tout notre raisonnement précédent s’appuie sur des bases non-fondées

Il nous manque certainement des variables : pluie, pédologie, etc. Mais en réalité, il est surtout certain que nous ne pouvons pas prédire des phénomènes non-linéaires avec des méthodes de régressions linéaires.

Pas de KNN car la prédiction est très mauvaise – il vaut mieux privilégier des méthodes d’arbre de décision, comme la *random forest* <https://www.kaggle.com/code/nishaanamin/crop-yield-prediction>

• Initialement, la valeur de la pluie était constante sur toutes les années d’observations, et était donc redondante avec l’information sur le pays : nous avons corrigé cela en important le dataset de la pluie, qui corrige légèrement le dataset proposé :

<https://data.worldbank.org/indicator/AG.LND.PRCP.MM?name_desc=false>

(ajoute de la variabilité dans les valeurs de pluie annuelles pour quelques années, pour quelques pays)

• [**https://www.kaggle.com/code/nishaanamin/crop-yield-prediction**](https://www.kaggle.com/code/nishaanamin/crop-yield-prediction) **pour de l’analyse basique de donnée**

Impossible de faire une présentation extensive : il faut se concentrer sur des choix subjectifs

* Analyse du ggpairs :

🡪 utiliser classification non-paramétrique : k-means (*Centroid models*, Hartigan and Wong algo) et CAH (*Connectivity models*)

K-means avec méthode du coude pour déterminer k

Que ça soit pour data ou pour by\_country, Iw\_kmeans < Iw\_HAC, on garde la clusterisation par kmeans avec k = 6

Problème : quelquefois, elle clusterise mal une dizaine de pays : pourquoi

🡪 patch en assignant le cluster majoritairement attribué à l’ensemble du pays

• On aurait pu aussi classer les cultures selon leur rendement : Sorgo-Soja-Blé-Riz-Maïs < Plantain-Manioc-Patates douces-Ignames < Pommes de terre

(voir graphique yield(year))

• PCA crops :

Rien n’est bien clair, 2 dimensions n’expliquent même pas 50% de la variance, les critères du coude, de Kaiser et de la moyenne empirique indique qu’il faut en utiliser au moins 3.

La technique en composantes principales reproduit avec parcimonie la variation totale d'un grand nombre de variables (pour fixer les idées, dans les cas les plus courants : de 10 à 40) en un nombre sensiblement plus restreint de dimensions 🡪 ici, on va privilégier la réduction de variables, car aucune des dimensions n’est satisfaisante

Avec 5 variables, deux dimensions suffisent à expliquer 74% de la variance, et on distingue clairement la ressemblance entre

• Information de l’année utile ? Effet de doublon avec les informations météorologiques annuelles ? 🡺 une variation similaire des températures et de la quantité de pluie tous pays confondu traduirait un effet « annuel »

Effet de l’année traduirait un manque de variables climatiques (évènement climatique extrême, gel, sécheresse, etc. – on n’a que la pluie totale et la température moyenne)

• Information du pays utile ? Effet de doublon avec les informations climatiques annuelles ? 🡺 une différence constante entre les pays des informations météorologiques annuelles traduiraient des différences entre pays de climat

Effet du pays traduirait peut-être un manque de données environnementales : pédologie, climat plus général, etc.

Choisir des questions précises, même si parcellaires

PCA : vp > 1 expliquent les variables / critère de Kaiser (moyenne empirique) et critère du coude pour vérifier

table.eig <- round(NAME$eig,2)

vec <- 1:12

plot(vec,table.eig[,2])

abline(h=mean(table.eig[,2]))

Se concentrer sur une seule année si le modèle ne donne rien : *Yi,j = (m + Pays\_i) + (b + gi)Pluie\_i,j + (d + ti)Temp\_i,j + Ei*

anc.INCGRCR=lm(Yij~Pays+Pluie+temp+pays\*pluie+pays\*temp,data=DATA)

summary(anc.INCGRCR)

Anova(anc.INCGRCR)

par(mfcol=c(2,2))

plot(anc.INCGRCR)

Ou moyenne des différences entre deux années ? voir « emprunte\_carbone » + tartinabilité du beurre

Faire un tableau récapitulatif des modèles qui n’ont pas fonctionné ? Variables entrées / R2 / p-valeurs du test de type II / Résidus linéaires / Loi normale / Homoscédasticité / Points aberrants

• **Comment expliquer les variations de rendement selon les variables disponibles ?**

**Y a-t-il une différence dans l’explication du rendement entre pommes de terre et maïs (+ de 90% des pays en cultivent), selon les pesticides, les variables climatiques, l’année ou le pays ?**

Tableau *yieldbycountry* :

• **Quelles cultures sont les mieux expliquées par les facteurs {pluie, température, pesticides}, tout pays et toute années confondues ? L’ajout de l’année ou du pays est-il pertinent pour gagner en information (faut-il pour cela exclure les variables où trop de données sont manquantes)** ?

• (En regroupant les pays selon les facteurs climatiques disponibles, peut-on expliquer le rendement de certaines cultures (commune à tous les groupes) à partir de la quantité de pesticide utilisée ?)

**« Au sein de pays ayant les mêmes conditions climatiques, la quantité de pesticide utilisée influence-t-elle le rendement des cultures »** ?

• Peut-on remplacer l’effet {pluie, température} par l’effet {année\*pays} ?

*•* Graphe d’interactions

PCA entre les différents rendements de cultures :

PCA sur tous les pays

PCA sur toutes les variables (à traduire en variable numériques donc)

🡺 permet de vérifier des collinéarités