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

Diapo

• (*À dire sur la page de plan*) À l'échelle mondiale, les pesticides sont la première cause de cancers [1] : ils causent maladies neurodégénératives ou cardiovasculaires [2], dégradent la biodiversité des sols [3], des eaux [4], et réduisent le potentiel de biocontrôle [5]. Pourtant, les producteurs participants au marché mondial y sont dépendants, car mis en concurrence ils doivent produire toujours plus. Cette prémice présuppose donc que les rendements sont dépendants de la quantité de pesticides utilisée. À travers notre analyse statistique, nous nous intéresserons d'une part à l'influence des pesticides sur le rendement, d'autre part à la multiplicité des variables pouvant expliquer les variations de rendement.

• *Conseiller au jury d'ouvrir notre code ? Sinon passer outre* - Les données de ce projet correspondent à une remise en forme des données disponibles à cette adresse https://www.kaggle.com/datasets/patelris/crop-yield-prediction-dataset, provenant initialement de la FAO (https: //www.fao.org/home/en/) et de https://data.worldbank.org/. Notre code est disponible ici : https://github.com/Dac-T/Agri_world_production

• Les rendements des 10 récoltes les plus consommées dans le monde sont disponibles pour 101 pays sur 23 années d'observation, de 1990 à 2013, 2003 exclue. À ces rendements [hg/ha] sont associés la quantité de pesticides utilisés [T], la quantité annuelle de pluie [mm/year] et la température moyenne [°C] de chaque pays pour chaque année. Notons qu'initialement, il était associé à chaque pays une unique valeur numérique de pluie annuelle, rendant cette information redondante. Nous avons corrigé cela en important de nouvelles données (https://data.worldbank.org/indicator/AG.LND.PRCP.MM?name_desc=false), ajoutant de la variabilité pour quelques années, pour quelques pays.

• Nos données sont disponibles sous deux formats. Le premier dataframe, *data,* contient 3 variables catégorielles et 4 variables continues : il donne pour chaque pays, chaque année et chaque type de culture le rendement associé, en complétant des valeurs de pluie, température et de quantité de pesticides utilisée. Il faut déjà retenir le fait que certains pays ont moins de 23 années d'observations.

• Le second, *by_country*, décline la variable « *Item* », qui représente le type de culture, en 10 colonnes, permettant d'associer à chaque culture son rendement particulier, dont la valeur dépend des variables catégorielles *pays* et *année*. Il nous faut noter l'existence de données manquantes sous ce format ; en effet, toutes les cultures ne sont pas présentes sur tous les territoires étudiés : il existe donc des lignes où les valeurs de rendement associées à une culture absente sont remplacées par NA.

• Afin d'améliorer la lisibilité de notre analyse, nous avons fait le choix de réduire le nombre de modalité de la variable « Area » : supposant que la paire (pluie-température) représente bien la localisation d'un pays, nous avons utilisé d'une part la méthode des kmeans, d'autre part une classification hiérarchique ascendante, que nous avons appliqué à nos deux dataframes. La méthode du coude nous a permis de déterminer que le nombre optimal de cluster était de 6, et la valeur de l'inertie intra-classe tirée de la méthode des kmeans étant inférieure à celle de la Classification Hiérarchique Ascendante, nous avons conservé la clusterisation issue de la première méthode.

• Le résultat de cette clusterisation est assez probant : les pays aux climats océanique ou continental sont représentés par le cluster 4 alors que les cluster 3 et 6 sont associés au climat aride. Cette cartographie révèle aussi un fort biais dans nos données, par l'absence de la Russie, des États-Unis et de la Chine, qui sont pourtant les plus gros pays producteurs à l'échelle mondiale.

• Ce premier graphique révèle la forte variabilité des rendements selon la culture : le rendement des pommes de terre est ainsi quasiment 8 fois plus grand que celui du soja ou du sorgho. Mais il permet aussi de constater la variabilité des rendements selon les pays : la pomme de terre, la patate douce et le manioc voient particulièrement leur rendement varier selon le pays dans lequel ils sont cultivés.

• L'évolution du rendement par culture révèle la légère croissance globale des rendements, d'environ 25% entre 1973 et 2013. On remarque enfin en combinant les informations issues des deux graphiques le fait que les cultures auraient pu être classées par rendement : la pomme de terre possède un très fort rendement, l'ensemble {Manioc-Igname-Patates douces-Plantains} un rendement moyen, et l'ensemble {Blé-Maïs-Riz-Soja-Sorgho} un faible rendement}

• En représentant les variations de rendement en fonction des cluster, toutes années et toutes cultures confondues, on remarque une certaine homogénéisation, de telle sorte que pour chacun des paires {pluie-température}, les rendements semblent globalement similaires : on peut s'attendre à un effet modéré des facteurs météorologique sur les rendements si l'on s'y intéresse toute culture confondue.

• Enfin, le pairwiseplot résume en son sein de nombreuses informations. Notons tout d'abord que les valeurs de pluie sont principalement comprises entre 0 et 2000 mm/an, la température moyenne tourne autour de 25°C, et les rendements principalement inférieurs à 10 T/ha.
[rain – temp] On peut aussi confirmer que les clusters, définis par la paire {pluie-température} correspondent à des distributions bien définies.
[rain – rain et temp - temp] On remarque ensuite que les courbes de densité de température et pluie ne suivant aucune loi connue : ne pouvant donc pas utiliser de méthode de classification probabiliste comme les mixture-model, le k-means était donc le meilleur choix.
[year-pest] Le Brésil, appartenant au cluster 1, en rouge, est le seul à utiliser une quantité de pesticide bien plus grande que les autres pays depuis les années 2000
[correlations] On n'observe aucune corrélation probante entre nos variables, notamment entre quantité de pesticides et rendement.

• 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 PCA sur les variables de notre dataframe *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 : cela ouvre un questionnement sur le pouvoir explicatif des pesticides 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, qui n'expliquent d'ailleurs à elles seules que 63% de la variance totale.

• Ainsi, au regard de notre analyse descriptive, nous faisons le choix d'exclure certaines cultures de notre analyse à venir : on retire le Soja, le Manioc, les Patates douces, le Plantain et l'Igname, qui, ne poussant que dans des climats chauds, sont absentes de 45 à 80% des pays étudiés. Cela introduisait d'ailleurs de nombreuses valeurs manquantes, problème qu'on ne pouvait pallier avec des méthodes inférentielles comme le bootstrap : ç'aurait en effet été absurde de créer artificiellement une valeur de rendement pour le manioc en France alors qu'on n'y cultive en pas.
Lorsque l'on conserve les 5 cultures majoritaires, la PCA donne des résultats bien plus probant ! 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 alors à ce que les régressions sur les rendements du riz fonctionnent aussi pour celle du sorgho.
Nous nous sommes donc substitués à la réduction de la dimension par une PCA la réduction de dimensions par le retrait de données.
Par conséquent, nous nous intéresserons aux pays cultivateurs de maïs, de pommes de terre, de riz, de blé et de sorgho.
On génère pour cela 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.

• De ces données, on se demande alors :
- 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 ?
- Comment expliquer les variations de rendement selon les variables disponibles ?
- Quels sont les effets du cluster et de la culture, après contrôle du volume de pesticides ?

• Question 1 :

À partir de notre dataframe contenant les 38 pays qui ont cultivé pendant les 23 années d'observations les 5 cultures dominantes, on crée une boucle sur chaque rendement associé à une culture, qui pour chaque cluster effectue une régression linéaire simple entre le logarithme du rendement et le logarithme du volume de pesticides utilisé. Nous nous sommes effectivement rendu compte que la transformation log~log était celle qui s'approchait le mieux de l'hypothèse d'homocédasticité, comparément à la racine carrée, à l'inverse, et au boxcox.

• Pourtant, même pour la meilleure régression (R^2 = 76%), celle qui correspond aux pommes de terre du cluster 1, les erreurs ne sont pas centrées, on observe en effet une légère hyperbole sur le graphe Res. v. fitted. De même, l'hypothèse d'homocédasticité n'est pas complétée d'après le test de Breusch-Pagan. Néanmoins, le 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.

• Rien qu'en s'intéressant au R^2, on se rend compte de l'extrême variabilité de la pertinence de la régression simple. Notons que tous les modèles présents sur ce graphe présentent une p-value au moins inférieure à 5%, nous permettant de rejeter l'hypothèse nulle correspondant à l'absence d'effet. On remarque que les rendements du maïs, de la pomme de terre et du sorgho sont plutôt bien expliqués par la régression simple (leur R^2 est supérieur à 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.
Toutefois, il est évident 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%.

• On ne peut donc pas affirmer qu'au sein de tout zone homogène en température et en quantité de pluie, pour une culture donnée, le volume de pesticide employé suffit à expliquer le rendement de cette culture.

• Question 2 :
On commence par chercher à construire une régression linéaire multiple. Après divers essais de fonctions de transformations et une minimisation du critère d'akaike, on en vient à ne pas considérer la pluie, pour écrire alors : Yield_i = a + b.log(pest)_i + c.log(temp)_i

Ce modèle, appliqué au rendement toutes cultures confondues de 85 pays complète parfaitement les hypothèses d'homocédasticité, et de structure 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 certes de rejeter H_0 au seuil de 1‰, mais il n'explique que 18% de la variabilité du logarithme du rendement, autant dire rien du tout ! Nous remarquons d'ailleurs que le coefficient *a* est 15 fois plus grand que *b*, associé à la température, et 120 fois plus grand que *c*, associé au volume de pesticides : il faut donc introduire d'autres variables dans notre modèle.

• Pour évaluer la pertinence de nos variables, nous avons effectué une MANCOVA sur les 5 valeurs de rendements attribuées à chaque culture, en s'intéressant aux effets de cluster et de l'année, en contrôlant ceux des pesticides, de la pluie et de la température. Les effets prédominants sont ceux des clusters, comparément aux années [6]. Notons que lorsque l'on s'intéresse aux analyses univariées, la pluie n'a des effets significatifs que sur les cultures de riz et de blé et la température n'a pas d'effet significatif sur les cultures de maïs. Notons qu'il est aussi rappelé l'hétérogénéité des variances et l'anormalité des résidus.

• Les résultats de la régression multiple et de la MANCOVA nous poussent alors à réaliser une régression linéaire générale sur les variables *Year* ; *Item* ; *Cluster* ; *Pesticides*. La variable cluster a en effet l'avantage d'intégrer les paramètres pluie et température dans « Area », et d'en réduire les modalités.
Nous partons du modèle complet avec toutes les interactions possibles, log(yield) ~ Year * Item * Cluster * log(pest), et en retirant des variables itérativement par l'analyse du test de type II, tout en contrôlant la similarité statistique au modèle initial, nous arrivons au modèle suivant :

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

Ce nouveau modèle minimise le critère d'information d'Akaike, passant de -6960 à -9073 : c'est d'ailleurs la valeur minimale que l'on a obtenue, tous modèles confondus. Même si le test d'Anderson-Darling annonce une non-normalité de la distribution des résidus, on peut au moins convenir du fait que l'absence de structure des résidus et l'homocédasticité sont des hypothèses complétées.

• On remarque alors que ce modèle nous permet d'expliquer 76% de la variabilité des rendements, mais surtout que les variables d'intérêts sont le type de culture, le cluster -donc la zone géographique et climatique, et la quantité de pesticides. Au contraire, la contribution de l'année est minimale.

• Nous générons alors un dernier modèle, en retirant le paramètre de l'année et l'interaction triple. Il devient log(yield) ~ Item + Cluster + log(pest) + Item:Cluster + Item:log(pest) + Cluster:log(pest), son AIC ne remonte qu'à -8707

Les graphiques de diagnostic ont une forme identique et le R^2 est à 75%. Nous conserverons ainsi ce nouveau modèle, qui a l'avantage d'être plus simple à interpréter. On peut en effet conclure que le logarithme du rendement dépend d'une part du type de culture, du cluster et du logarithme du volume de pesticide employé, mais aussi de leurs interactions mutuelles.

• Question 3
On a pu identifier précédemment les rôles majeurs du volume de pesticide, du cluster et du type de culture. Nous allons maintenant nous intéresser spécifiquement à ces deux dernières variables, en effectuant une ANCOVA à

deux facteurs, ce qui impose que l'on fixe l'usage de pesticides comme covariable (source de variation dont on retire les effets pour augmenter la puissance statistique de l'ANCOVA). Ainsi, malgré la significativité des interactions entre log(pest)*Item et log(pest)*Cluster, on les retire ici dans le cadre de l'ANCOVA.
Pour respecter l'hypothèse d'homogénéité des courbes de régression, nous sélectionnons 3 cultures compatibles, le maïs, la pomme de terre et le sorgho.
• Les graphes de diagnostic semblent convenir, notons toutefois qu'un test de Levene et un test d'Anderson-Darling révèlent que les variances des résidus ne sont pas homogènes et qu'ils ne sont pas normaux.
On arrive alors à expliquer 78% de l'effet sur le rendement, et après le contrôle des pesticides, il demeure une interaction statistiquement significative entre les clusters et les types de cultures ($p < 2.2 e\text{-}16$).

• On remarque sur le tableau du haut que l'effet du type de culture est statistiquement significatif sur tous les clusters, et de même sur le tableau du bas que l'effet du cluster est statistiquement significatif sur toutes les cultures.

*{*Tableau du haut *:*

```
full3cdata %>%
  group_by(Cluster) %>%
  rstatix::anova_test(log_yield ~ log(pest) + Item)
```

Tableau du bas *:*

```
full3cdata %>%
  group_by(Item) %>%
  rstatix::anova_test(log_yield ~ log(pest) + Cluster)}
```

On s'intéresse alors dans le détail à la différence des moyennes marginales du rendement, inter-cultures et inter-cluster.

• Une comparaison par paire avec correction de bonferroni entre les types de culture révèle des différences significatives entre les moyennes ajustées des trois cultures. Plus précisément, la seule comparaison non significative s'exprime entre le maïs et le sorgho dans le cluster 5.

• Une comparaison par paire avec correction de bonferroni révèle des similarités entre les moyennes ajustées des régions 2, 5 et 6, et entre les régions 1 et 3. Notons toutefois que le Sorgho présente des différences significatives entre les moyennes ajustées de tous les groupes comparés deux à deux, excepté les groupes 1 et 2. Autrement dit, toute chose égale par ailleurs, chaque cluster porte un effet statistiquement significatif distinct sur la culture du Sorgho, excepté les clusters 1 et 2.

On en retient ainsi le fait que le rendement, une fois la quantité de pesticides contrôlée, reste dépendant du type de culture, et de la région. Rappelons toutefois qu'encore une fois, les hypothèses sous-jacentes au modèle n'étaient pas respectées.

• À l'aide de la méthode des *kmeans*, nous avons pu créer avec succès 6 zones géographiques correspondant à des paires {pluie-température}, qui nous ont servie de modalités pour notre nouvelle variable catégorielle « cluster », bien plus pratique à analyser qu'une liste de 101 pays. Les analyses en composantes principales nous ont poussé à réduire notre jeu de donnée afin d'augmenter notre puissance statistique. Nos multiples régressions ont pu révéler les facteurs clé permettant d'expliquer les trois-quart de la variabilité du logarithme du rendement, à savoir la zone géographique, le type de culture, le logarithme de la quantité de pesticides utilisée, et toutes leurs interactions.

• Nous supposons que la faiblesse de l'effet associé à la pluie est issue de sa faible variabilité temporelle au sein d'un pays : malgré nos modifications, elle constituait plutôt une variable quasi-ordinale que continue.
• Il nous manque certainement d'autres informations, que ce soit en quantité de variables -nous aurions pu par exemple avoir des données pédologiques, ou bien en quantité de données. Nous déplorons en effet l'absence des plus grands producteurs mondiaux dans nos données, introduisant *de facto* un large biais sur les conclusions à tirer.
• Surtout, et c'est là la plus grande limite, nos résidus n'ont que très peu été normaux, nous n'avons jamais satisfait toutes les hypothèses que requièrent les modèles linéaires. C'est d'ailleurs pour cela que, malgré l'intérêt prometteur de nos régressions, nous n'avons pas performé de prédiction avec la méthode des *K-nearest-neighbors*. Ça a en effet déjà été effectué sur le même jeu de donnée, et la précision de la prédiction n'est que de 33%. Ce résultat fait sens : on ne peut pas prédire des phénomènes non-linéaires avec des méthodes simplistes qui supportent mal les grandes dimensions. Il faudrait à l'avenir songer à des méthodes plus élaborées, comme les forêts aléatoires ou le boosting.

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%) : car Cassava, sweet potatoes, yams, and plantains & others ne poussent que dans les climats chauds

- 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** !

2. 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)

Qui minimise le critère d'information d'Akaike, passant d'une valeur initiale de -6960 à -9073, et qui possède parmi tous nos modèles la valeur minimale

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, **pour un AIC qui ne retombe qu'à -8707**.

→ 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

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

https://en.wikipedia.org/wiki/Analysis_of_covariance
https://www.datanovia.com/en/lessons/ancova-in-r/

Pesticides : covariable – source de variation dont on retire les effets pour augmenter la puissance statistique de l'ANCOVA

Dérivée de la conclusion précédente, mais uniquement avec 3 cultures qui ont des courbes de régression visuellement homogènes (**ajouter graph**) – Maïs / Pomme de terre – Sorgho
Un test de Levene et un test d'Anderson-Darling révèlent que les variances ne sont pas homogènes et que les résidus ne sont pas normaux.
Malgré la significativité des interactions entre log(pest)*Item et log(pest)*Cluster, on les retire ici dans le cadre de l'ANCOVA.

➔ on arrive alors à expliquer **78%** de la variabilité du log du rendement
➔ L'effet de la culture est statistiquement significatif sur tous les groupes *(tibble in # Looking for the Items effect # Even with Bonferroni correction (p must be < 8e-3), everything is statistically significant)*

➔ **Graph lp_item** : une comparaison par paire avec correction de bonferroni entre les types de culture révèle des différences significatives entre les moyennes ajustées des trois cultures. Plus précisement, la seule comparaison non significative s'exprime entre le maïs et le sorgho dans le cluster 5.

➔ L'effet de la région (par Cluster) est statistiquement significatif sur toutes les cultures *(tibble in # Looking for the cluster's effect # Even with Bonferroni correction (p must be < 8e-3), everything is statistically significant)*
➔ **Graph lp_clust :** une comparaison par paire avec correction de bonferroni révèle des similarités entre les moyennes ajustées des régions 2, 5 et 6, et entre les régions 1 et 3. Notons toutefois que le Sorgho présente des différences significatives entre les moyennes ajustées de tous les groupes comparés deux à deux, excepté les groupes 1 et 2.

On en retient que le rendement, une fois la quantité de pesticides contrôlée, reste dépendant du type de culture, et de la région.
Rappelons toutefois qu'encore une fois, les hypothèses sous-jacentes au modèle n'étaient pas respectées.

- **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) pour couvrir les cas où il y a plus d'une variable dépendante et où les variables dépendantes ne peuvent pas être simplement combinées

Est-ce que le rendement des cultures (variable dépendante) diffèrent à cause des pays et de l'année (2 variables indépendantes) après avoir pris en compte la pluie et la température (2 covariables) ?

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
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.

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 (33% de précision) – 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 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