ACADEMIE DE MONTPELLIER

UNIVERSITE DE MONTPELLIER

MASTER

STATISTIQUE POUR LES SCIENCES DE LA VIE

MEMOIRE sur le stage

Analyse de données agronomiques d'une diversité de variétés de bananiers plantain.

- effectué du *13/02/2023* au *12/08/2023*
- à Montpellier
- sous la direction de Gaëlle Damour
- par Marina Pons
- soutenu le 12/09/2023
- devant la commission d'examen

J.N. BACRO, C. REYNES, R. SABATIER.

Remerciements

“

Je remercie tous les membres de l’unité GECO, pour leurs accueils, leurs bienveillances et les nombreuses conversations partagés ensemble. Je garderai un souvenir agréable de chacun d’entre eux. Ils m’ont accompagné sans le savoir dans une étape notable de mon parcours, ne l’oublions pas, qui est ce stage de fin de master.

Je remercie Sandrine AUZOUX d’avoir pris de son temps pour suivre l’évolution du projet tout en apportant sa joie de vivre et ses bons conseils.

Je remercie Pauline PUGEAUX de m’avoir facilité la compréhension du jeu de données et pour sa participation au projet.

Je remercie particulièrement Gaëlle DAMOUR, ma tutrice de stage, et Sylvain DEPIGNY, qui ont su m’accompagner et me conseiller tout au long de ce stage.

Je remercie Luc DE LAPEYRE DE BELLAIRE pour sa bienveillance, et une nouvelle fois, Gaëlle DAMOUR et Sylvain DEPIGNY de m’avoir permis de faire ce stage à leurs côtés.

“

Résumé

La collection de bananiers plantain du CARBAP au Cameroun est la plus grande du monde. En effet, elle compte 135 variétés différentes recouvrant tous les groupes botaniques existants. L'étude expérimentale a duré un peu moins de 5 ans et aura permis de collecter un grand nombre de données concernant des mesures phénotypiques de croissance et de rendement sur 5 cycles de croissance. Le jeu de données généré a été exploré pour approfondir les connaissances peu documentées sur la diversité des bananiers plantain. Pour cela une analyse descriptive a été effectuée, notamment à l'échelle variétale avec une visualisation des données sur une application web Shiny. Ensuite, une analyse des groupes botaniques, au travers des corrélations et d'arbres de classifications, a été effectuée. Une Analyse en Composantes Principales (ACP) a été réalisée dans le but de résumer la structure du jeu de données dans un espace réduit. À la suite de l'ACP, une Classification Hiérarchique sur Composantes Principales (HCPC) a été effectuée pour isoler de nouveaux groupes. Des variables explicatives du poids du régime ont été recherchées avec une régression linéaire simple et avec la méthode de forêt aléatoire.

Mots-clés :
CARBAP ; bananier ; plantain ; ACP ; HCPC ; forêt aléatoire ; corrélation ; Shiny ; Dashboard ; variété ; classification ; groupe morpho-taxonomique

Abstract

CARBAP's plantain collection in Cameroon is the biggest in the world. It is composed of 135 varieties representing all botanic groups. The experimental study lasted 5 years and many phenotypic variables describing growth and yield were measured throughout 5 cultural cycles. The dataset was explored to improve the knowledge of plantain diversity. A descriptive analysis of varieties was performed and visualized with a Shiny web application. Then, an analysis of the botanical groups, through correlations and classification trees, was performed. A Principal Component Analysis (PCA) were used to summarize the structure of the dataset in a reduced dimension. After the PCA, a hierarchical classification of the principal components (HCPC) was used to isolate the new groups. Predictive variables of bunch weight were sought with a simple linear regression and the random forest method (Random Forest).

KEYWORDS:
CARBAP; banana; plantain; ACP; HCPC; random forest; correlation; Shiny; Dashboard; variety; classification; morpho-taxonomic group;

Table des matières

1. Organisme d'accueil du stage 1

a. Présentation de l'organisme d'accueil : 1

b. Présentation de l'équipe encadrante : 1

2. Contexte de l'étude.....1

a. Objectif et missions du stage : 3

b. Présentation du jeu de données : 3

c. Gestion des données manquantes : 4

3. Analyse statistique – *effectué sur R Studio® version 2.4.3* 5

a. Plan d'analyse global : 5

b. Développement des étapes du plan d'analyse : 6

I. Analyse statistiques descriptives globales.....6

II. Analyse exploratoire.....14

III. Création de nouveaux groupes : HCPC.....19

IV. Prédiction du poids du régime.....21

V. Evolution temporelle 24

4. Application web SHINY 27

5. Discussions : 28

6. Conclusion de l'étude sur la collection de bananiers plantain et du stage 30

Bibliographie : 31

1. Organisme d'accueil du stage

a. Présentation de l'organisme d'accueil :

Ce stage prend vie grâce à la collaboration entre le master *« Epidémiologie, Données de santé et Biostatistiques pour les sciences de la vie »* de l'Université de Montpellier et le Centre de coopération Internationale en Recherche Agronomique pour le Développement (Cirad).

Cet organisme français, créé en 1984 par l'union de plusieurs instituts de recherches agricoles tropicales, est aujourd'hui présent sur tous les continents pour œuvrer dans la recherche, le développement et l'innovation agronomique en terrain tropical et subtropical. Son ambition actuelle est de participer à la création d'un monde plus durable tout en réussissant à nourrir le monde de façon saine et équilibrée, en particulier par l'accompagnement au développement de façon de cultiver basés sur des pratiques agroécologiques.

Il y a deux sites majeurs en France qui se situent à Montferrier-le-Lez. Un au sud, site de Lavalette, et l'autre au nord, site de Baillarguet. J'ai été, pour ma part, invitée à venir m'installer pour une durée de six mois sur le site de Baillarguet pour rejoindre l'équipe de l'unité GECO (Fonctionnement écologique et gestion durable des agrosystèmes bananiers et ananas, *https://ur-geco.cirad.fr*) du département en charge des performances des systèmes de production et de transformation tropicaux (« *Persyst* »). Plus exactement, j'ai rejoint l'équipe qui est chargée des recherches sur le bananier plantain. C'est ainsi que mon stage a débuté au côté de Gaëlle Damour, Sylvain Dépigny, Pauline Pugeaux, et Sandrine Auzoux.

b. Présentation de l'équipe encadrante :

- *Direction de l'unité de recherche :*

Luc DE LAPEYER DE BELLAIRE directeur de l'unité GECO au sein du département Persyst, Montpellier.

Philippe TIXIER directeur adjoint de l'unité GECO et responsable de l'axe scientifique 1 : *Dynamique des bioagresseurs et des communautés en conditions d'intensification écologique,* chercheur, écologie et modélisation, Laos.

Gaëlle DAMOUR directrice adjointe de l'unité GECO et responsable de l'axe scientifique 2 : *Diversité végétale et services écosystémiques dans les systèmes cultivés* ; chercheuse, agronomie et écologie fonctionnelle, Montpellier.

- *Encadrement du stage* :

Gaëlle DAMOUR tutrice du stage.

Sylvain DEPIGNY chercheur, agronomie, spécialiste du bananier plantain, Montpellier.

Pauline PUGEAUX ingénieure, agronomie, spécialiste du bananier plantain, Côte d'Ivoire.

Sandrine AUZOUX informaticienne, CIRAD unité AIDA, la Réunion.

2. Contexte de l'étude

L'étude nous emmène au Cameroun, entre les années 2013 et 2018, sur le terrain du Centre Africain de Recherches sur Bananiers et sur Plantains (CARBAP), où se trouve une collection regroupant plus d'une centaine de variétés différentes de bananiers plantain. Les plants de bananiers

plantain sont suivis en conditions de culture non limitantes. Cela signifie que les pratiques culturales appliquées ont permis de maximiser la sauvegarde des variétés et d'assurer l'expression de leur plein potentiel agronomique. Les bananiers plantain sont fertilisés et une lutte chimique et manuelle contre les principaux ravageurs des bananiers plantain (charançon, nématode, cercosporiose) est également appliquée. Pour lutter contre les charançons, il y a eu l'application de deux insecticides par an. Pour lutter contre les nématodes, il y a eu la mise en place d'une jachère avant l'implantation de la collection, puis l'application d'une dose de nématicide par an. Le bananier est également sensible à un champignon foliaire, la cercosporiose noire. Pour limiter sa propagation, un effeuillage manuel et un traitement chimique aérien ont été effectués régulièrement dans le but de minimiser les pertes de rendements. Un désherbage manuel et chimique ont été prodigués pour éviter la compétition entre les mauvaises herbes et les plants de bananiers plantain. Les bananiers plantain ont aussi été irrigués en saison sèche pour limiter les stress hydriques.

Mais avant tout, le bananier plantain qu'est-ce que c'est ? C'est une plante herbacée de la famille des *Musaceae (Lassoudière, 2010)*. Le bananier plantain provient du croisement de deux espèces sauvages *Musa acuminata* et *Musa balbisiana*. Les bananiers plantain font partie du groupe des bananes à cuire (triploïdes AAB). Ils sont parthénocarpiques et ils se multiplient donc uniquement par des rejets émergents du bulbe, au pied des pseudo-troncs *(Kwa et al, 2019)*. La banane plantain est le troisième aliment de base de la diète africaine en Afrique centrale et de l'Ouest. Elle peut être cuisinée de nombreuses manières, cuites, bouillies ou même frites et peut être considérée comme un superaliment puisqu'elle contient des fibres, des minéraux et de nombreuses vitamines *(Ekesa et al, 2015 ; Anyasi et al, 2018 ; Udomkun et al, 2021)*. Pour ces raisons, elle représente avant tout un enjeu de sécurité alimentaire et également un enjeu économique sur les marchés locaux et régionaux. Prévoyant la croissance de la consommation de bananes plantain à l'avenir, il est important de répondre à cette demande. Notamment dans les pays producteurs qui sont les principaux consommateurs, cela concerne majoritairement les pays « dans les zones intertropicales comme en Afrique de l'Ouest et du Centre » *(Kwa et al, 2019)*. L'objectif serait de ne pas avoir la même approche qu'avec la banane dessert où seule la variété *Cavendish* domine le marché actuellement. Dans l'idéal, il faudrait réussir à maintenir la culture d'une diversité variétale de bananiers plantain car les différentes caractéristiques variétales permettraient de répondre aussi bien aux différents usages attendus des bananiers plantain, qu'aux différentes problématiques de culture, les variétés, de bananiers plantain n'ayant pas nécessairement la même sensibilité aux différents ravageurs du plantain. Pour le Cirad, l'idée d'augmenter la production de bananes plantain pour alimenter les marchés, serait possible via l'intensification écologique de la culture de bananier plantain, c'est à dire en utilisant des pratiques agroécologiques plutôt que des pesticides. Le choix de variétés adaptées à des conditions de culture plus limitées et/ou résistantes à des ravageurs est une des pratiques agroécologiques à mettre en œuvre. C'est pour cela qu'il est nécessaire de connaître davantage la diversité des nombreuses variétés de bananiers plantain, dont celles à disposition au CARBAP, qui est la plus grande collection de variétés de bananiers plantain au monde.

Pour répondre à cette problématique, les chercheurs de l'unité GECO, ont décidé de lancer une vaste et ambitieuse étude qui vise à décrire un maximum de variétés de bananiers plantain. Nous voilà de retour au CARBAP au Cameroun, où l'étude débuta en 2013 et où Sylvain Dépigny, avec l'aide des partenaires du CARBAP et de Pauline Pugeaux, a acquis un grand nombre de données agronomique sur 135 variétés différentes de bananiers plantain. Il existe une description botanique des variétés avec des groupes morpho-taxonomiques, par type et par taille définie selon la dégénérescence de l'inflorescence et le nombre de feuilles émises *(Tezenas Du Montcel, 1983)*. Les variétés représentées par l'étude recouvrent tous les types et tailles existants. Les variétés se trouvant dans la collection ainsi que leur appartenance aux groupes sont répertoriées dans un ouvrage : le Musalogue *(Ibobondji et al, 2018)*. Cette description botanique ne permet cependant pas nécessairement de décrire le comportement agronomique des variétés *(Dépigny et Damour, 2022)*. L'étude aura eu une durée d'environ 5 ans ce qui représente plusieurs cycles de culture (période entre le démarrage du plant de bananier et la récolte de son régime) pour les bananiers plantain. Il faut savoir que les variétés n'ont

pas la même durée de cycle de culture dû à leurs caractéristiques différentes. Le premier cycle de culture est caractérisé par la plantation d'un premier plant de bananier plantain désolidarisé de son pied mère ce qui peut engendrer un stress supplémentaire *(M. Kwa et al, 2019)*. Contrairement aux cycles suivants où le rejet n'est pas privé de son pied mère. Les plants de bananier plantain sont aussi exposés à des problèmes climatiques et ils doivent faire face à des phytopathogènes et des bioagresseurs qui les rendent plus vulnérables à une chute précoce *(Lassoudière, 2010)*. Au sein de la collection, si un plant est perdu, il n'est plus suivi dans le cadre de l'étude. Les différentes mesures prisent sur le terrain représentent à la fois la croissance, le développement, mais aussi le rendement. Les fiches d'archives générées regroupant toutes les mesures ont été traduites en format numérique par Pauline Pugeaux dans le but d'exploiter ces données. Né alors le projet de stage, direction Montferrier-le-lez en 2023 dans le bureau de Gaëlle Damour !

a. Objectif et missions du stage :

L'objectif principal du stage était d'analyser les différences de valeurs de traits agronomiques entre les 135 variétés de bananiers plantain de la collection du CARBAP, ainsi que leurs évolutions sur 5 cycles de culture.

Les missions de ce stage étaient : 1. la prise en main du jeu de données et la familiarisation avec le cas d'étude, 2. la mise en forme et le nettoyage des données, 3. l'analyse des données et 4. la construction d'un outil de visualisation des caractéristiques variétales sous *Shiny*

L'analyse des données a porté sur deux axes de recherches. Le premier concernait la mise en évidence des différences observées entre variétés et leurs évolutions au cours des 5 cycles de culture. Le second portait sur la recherche de nouveaux groupes de variétés, sur la base de variables morphologiques et relatives au rendement, et révélant ainsi des similitudes de fonctionnement. L'idée était alors de rechercher des variables explicatives du rendement et/ou de la croissance au sein de chaque groupe.

b. Présentation du jeu de données :

Le jeu de données a été déposé sur le Dataverse CIRAD (https://dataverse.cirad.fr/), sous le DOI suivant : `doi:10.18167/DVN1/ZAFKNL`. Les données ne sont pour le moment pas rendues accessibles pour des raisons de confidentialités avant publication d'un article de recherche par l'équipe encadrante et auquel je serai associée.

Le jeu de données brut est stocké en version Excel, il a été construit, compilé et prénettoyé par Pauline Pugeaux. Il a une dimension de 202 colonnes et 675 lignes, correspondant à 135 variétés et 5 plants par variété. Les 675 plants ont été suivis sur 5 cycles. La période de plantation des différentes variétés s'étend du 30/05/2013 au 27/07/2013 et la période de dernière récolte (cycle 5) s'étend du 12/04/2016 au 09/01/2018. L'étude a donc duré 1685 jours soit 4 ans, 7 mois et 10 jours. Les données correspondantes ont été ajoutées les unes après les autres, en colonne. Les colonnes regroupent des variables relatives aux dates du cycle, des variables relatives à la morphologie des plants (pour la plupart prises de façon répétée dans le temps), des variables calculées et enfin des variables qualitatives issues du Musalogue *(Annexe 1)*.

Un premier diagnostic a été effectué et plusieurs éléments caractérisant le jeu de données brut sont à prendre en compte. Premièrement, les plants suivis au cours des différents cycles ne peuvent pas être considérés comme indépendants. Même si le bananier est coupé à la fin de chaque cycle, c'est son rejet, provenant du même bulbe, qui effectue le cycle suivant, le bananier se reproduisant de manière végétative. Il y a donc appariement entre les données prises au cours des 5 cycles sur un plant et ses rejets successeurs. Certaines variables sont mesurées plusieurs fois dans le temps au cours des cycles, semblables à des variables dites temporelles. Par exemple, la hauteur du pseudo-tronc a été mesurée pour les mois 1, 3, 6, 9, à la floraison et à la récolte. Le cycle 1 regroupe plus de variables que les

autres cycles. Aux cycles suivants la hauteur du pseudo-tronc n'est mesurée qu'à la floraison et à la récolte. Le cycle 2 est le cycle qui compte le moins de variables. En effet, dans le suivi d'une collection le premier cycle est utilisé pour une caractérisation plus précise de la croissance et du développement des variétés, d'ailleurs tous les rejets ont été conservés mais pas aux cycles suivants alors le cycle 2 s'en trouve perturbé. Les cycles suivants sont alors des cycles de routine et mesure les paramètres de rendement à la floraison et à la récolte qui intéresse plus particulièrement les agronomes. La durée de cycle est propre à chaque variété. Il existe un écart de plus d'un an et demi entre les variétés à la date de dernière récolte, rendant la compréhension du jeu de données un peu plus complexe. En effet, les mesures d'une même variable ont été prises aux mêmes stades physiologiques mais pas aux mêmes saisons. Il peut donc y avoir des biais d'interprétations dus à des conditions de culture, principalement des conditions météorologiques, qui ont pu être différentes entre variété. Ce biais s'intensifie au fur et à mesure des cycles, l'écart de temps se creusant toujours un peu plus. Pour finir, il y a un certain nombre de données manquantes liées au fait que les bananiers plantain peuvent présenter de forts taux de chutes dues aux conditions sanitaires et climatiques (vent, charançons). Une fois qu'un pied de bananier plantain ait chuté, il n'est plus suivi, à condition que le rejet ne puisse pas être récupéré non plus. En effet, il y a des plants qui ont chuté tardivement n'affectant pas le rejet et qui ont pu être suivie au cycle suivant quand même.

c. Gestion des données manquantes :

Une première évaluation a été faite sur les données manquantes du jeu de données brut avec la fonction *vis_miss* de la library *visdat* sur R Studio®. Les données manquantes s'élevaient à 21%. L'objectif a été de diminuer le nombre de données manquantes, voire de les remplacer, tout en conservant le maximum d'informations et de ne pas perdre en qualité. Cela ne concernait que des variables quantitatives. Il y avait une volonté de garder le jeu de données sous sa forme initiale pour pouvoir analyser chaque plant dans sa globalité, puisqu'un des objectifs d'analyse était de suivre leurs évolutions au cours des cycles. Tout d'abord un premier tri a été effectué sur le jeu de données : les bananiers plantain qui avaient des données manquantes dès le début du premier cycle ont été mis en évidence et retirés du jeu de données. Cela ne concernait que 10 plants : leurs croissances après plantation étaient soit inexistantes soit anormales *(Annexe 2)*. Des variables ont été enlevées : celles qui ont été jugées inutiles puisqu'elles n'apportaient aucune utilité descriptive mais plutôt de repère pour les agronomes. Les variables qui avaient des données manquantes supérieures à 50% ont aussi été supprimées. Le fait de perdre une observation signifie perdre 1/5 de l'information par variété. La volonté d'essayer d'imputer les données, même si cela n'était pas l'idéal, fut évoquée. Un travail de recherche bibliographie sur la gestion des données manquantes et sur différentes méthodes d'imputations a alors été effectué.

Dans un premier temps, il a fallu évaluer et déterminer la nature des données manquantes puisque cela peut avoir une incidence sur la qualité de l'imputation *(Misztal, 2013)*. La structure des données manquantes dans le jeu de données semble être à la fois monotone et sans structure. La structure monotone est définie si à partir d'une observation manquante toutes les observations suivantes manquent aussi pour un même individu. Il n'y a pas de structure quand la donnée manquante semble être aléatoire et qu'elle ne dépend ni de l'individu ni d'une autre variable. Il existe trois types de données manquantes : MCAR (Missing Completly At Random), MAR (Missing At random) et NMAR (Not Missing At Random). Les MCAR et MAR sont des données manquantes de façon aléatoires. Les MNAR ne sont pas aléatoires et sont dépendantes de la variable elle-même. Dans le cas du jeu de données, les données manquantes de type MNAR sont largement majoritaire dans le jeu de données et ce sont les plus délicates à traiter *(Little et al, 2002 ; Misztal, 2013 ; Imbert et al, 2020)*.

La méthode classique pour la gestion des données manquantes, appelée analyse des cas complets, consiste à supprimer les observations qui ont au moins une donnée manquante, mais elle devient obsolète si le jeu de données est de grande dimension et que le pourcentage de données manquantes

est élevé. *Graham (2009)* conseille de ne pas avoir recours à cette méthode si le groupe d'individus retiré représente plus de 5% de la population (*Imbert et al, 2020).* Dans notre cas l'application de cette méthode était impossible, il a donc fallu se tourner vers des méthodes d'imputations. Il existe l'imputation simple ou l'imputation multiple, avec plusieurs approches différentes comme celle par similarité ou par prédiction. Deux méthodes ont été testées sur le jeu de données, celle des k plus proches voisins (kNN) et celle de missForest.

La méthode des kNN est une approche basée sur la similarité par la distance entre individus. La distance est obtenue par un ensemble de variables entièrement observées. Les deux paramètres principaux de cette méthode sont le choix du nombre de voisins *k* et la distance *d*, c'est-à-dire les *k* individus ayant la distance *d* la plus faible. Le package R qui a été utilisé est VIM *(Kowarik et al, 2016)* avec les paramètres par défaut dont *k=5 (Sanjar et al, 2020).*

La méthode d'imputation par forêt aléatoire qui est basée sur l'algorithme Random Forest *(Breiman, 2001)* qui consiste à générer des arbres de classification pour prédire des variables et dans le cas de missForest pour prédire et remplacer les données manquantes. Le package R qui a été utilisé est missForest *(Stekhoven et al, 2012).*

Les valeurs remplacées n'étaient pas aberrantes mais l'information propre au jeu de données s'en trouve lourdement standardisée. La méthode kNN semblait être la plus adaptée dans ce cas puisqu'elle est basée sur la similarité entre 5 plus proches voisins et non pas sur la prédiction en fonction d'autres variables. Il peut y avoir des différences de comportement entre variétés. Il aurait été intéressant de comparer les méthodes pour aller plus loin. Cependant une première approche exploratoire permet d'évaluer cela ; elle consiste à comparer les données imputées avec les observations disponibles et de voir le comportement des données à travers des graphiques (*Imbert et al, 2020)*. Quand, dans la littérature les méthodes d'imputations sont confrontées entre elles, les erreurs d'imputations augmentent avec le pourcentage de données manquantes et la méthode d'imputation qui semble être une des plus performante est Amelia II qui est une méthode paramétrique d'imputation multiple par James Honaker *(Honaker et al, 2011).* Elle combine le maximum de vraisemblance avec une approche Bootstrap. Une transformation des données aurait été envisagée pour pouvoir approcher la loi normale mais une seconde hypothèse est à valider sur le fait que les données manquantes doivent être de nature MAR. Cette méthode ne pouvait pas être utilisée dans ce cas-là.

Finalement, les données manquantes étant MNAR et au vu du contexte de l'étude, il est fortement déconseillé d'imputer les données car il ne serait pas juste de qualifier les données manquantes autrement que données censurées. Les données censurées ne peuvent être remplacées. Il est statistiquement plus juste de n'utiliser que les données disponibles. Le premier cycle de culture sera donc analysé séparément tout en omettant les données censurées. Il sera le cycle de référence puisque c'est celui qui est le plus complet, il représente le mieux la collection. Il faut tout de même noter qu'il est différent des autres cycles puisqu'il n'a pas la même condition initiale à la plantation.

3. Analyse statistique – *effectué sur R Studio® version 2.4.3*

a. Plan d'analyse global :

I. Analyse statistiques descriptives globales
   a. Description de la population représentée au cycle 1
   b. Description des variétés
   c. Description des groupes botaniques : Taille et Type

II. Analyse exploratoire
   a. Corrélation entre variables en global

b. Corrélation entre variables par groupes botaniques
c. ACP globale

III. Création de nouveaux groupes : HCPC

IV. Prédiction du poids du régime
a. Modèle linéaire
b. Modèle linéaire multiple
c. Forêt aléatoire (*Random Forest*)

V. Evolution temporelle

b. Développement des étapes du plan d'analyse :

Les codes relatifs à cette analyse vont être déposés sur le GitHub®, une plateforme permettant le partage et la sauvegarde de scripts.

I. Analyse statistiques descriptives globales

Dans la première partie, le but est de faire une description détaillée des variétés de la collection pour pouvoir conserver l'information et décrire les groupes de Taille et Type issus du Musalogue pour pouvoir mieux les comprendre.

a. Description de la population représentée au cycle 1

Le jeu de données du cycle 1 n'est pas équirépété, le nombre de bananiers plantain peut varier de 1 à 5 par variété. Au total, il y en a 665. Les 135 variétés de la collection *(Annexe 3)* sont représentées. Le taux de données manquantes global du jeu de données du cycle 1 s'élève à 4,4%. La période de plantation du cycle 1 est du 30/05/2013 au 27/07/2013 et la période de récolte du 08/04/2014 au 17/02/2015.

Tableau 1 : Effectifs et proportions (%) de bananiers plantain selon les groupes Taille (Musalogue) au cycle 1.

Taille Musalogue	**Petit**	**Moyen**	**Géant**	**Total :**
Effectif	5	**393**	267	**665**
≈ %	*0,8*	***59,1***	*40,2*	***100***

La proportion de bananiers plantain appartenant au groupe Moyen est de 59,1%, il est le mieux représenté au cycle 1. La proportion du groupe Géant est de 40% et celle du groupe Petit représente moins de 1% des bananiers plantain du cycle 1 *(Tab. 1).*

Tableau 2 : Effectifs et proportions (%) de bananiers plantain selon les groupes Type (Musalogue) au cycle 1.

Type Musalogue	**Batard**	**Corne**	**Faux-corne**	**French**	**Total :**
Effectif	5	65	232	**363**	**665**
≈ %	*0,8*	*9,8*	*34,9*	***54,6***	***100***

La proportion de bananiers plantain appartenant au groupe French est de 54,6%, il est le mieux représenté. La proportion du groupe Faux-corne est de 34,9% et elle est de 9,8% pour le groupe Corne. Le groupe Batard représente moins de 1% des bananiers plantain du cycle 1 *(Tab. 2).*

La répartition de ces proportions illustre l'histoire des choix variétaux des producteurs et des consommateurs de bananiers plantain. Les bananiers plantain appartenant au groupe Moyen sont considérés comme moins sensibles aux stress biotique et abiotique, de plus les soins apportés à la plante et au régime sont relativement faciles à effectuer depuis le sol. Les bananiers plantain des groupes Géant et/ou French sont appréciés pour leurs régimes qui ont tendance à être plus lourds. Tandis que les bananiers plantain du groupe Faux-cornes sont appréciés pour leur goût.

Tableau 3 : Répartition des différents pays d'origines des variétés de bananier plantain représentés par la collection.

	Effectif	
Pays d'origine	***Variétés***	***Plants***
- Colombie	*1*	*5*
- Antilles françaises	*2*	*10*
- Côte d'Ivoire	*12*	*60*
- Ghana	*2*	*10*
- Nigeria	*32*	*153*
- Cameroun	*81*	*402*
- République centrafricaine	*1*	*5*
- Gabon	*1*	*5*
- Rwanda	*1*	*5*
Inconnue	*2*	*10*
Total	***135***	***665***

Environ 98% des variétés représentées par la collection sont originaires d'Afrique centrale et de l'Ouest et le reste provient d'Amérique centrale *(Tab. 3)*.

Tableau 4 : Sélection pratique d'une partie du lexique des variables qui se trouve en Annexe 1.

Variable	Description	Unité	Nature
Variete	Nom de la variété		Qualitative
Typemusalogue	Catégorie de Type de plantain d'après le Musalogue		Qualitative
Taillemusalogue	Catégorie de taille du plantain d'après le Musalogue		Qualitative
Cycle 1 / c1			
htfloc1	Hauteur du pseudo-tronc à la floraison	cm	Quantitative
htbulbfloc1	Hauteur du bulbe à la floraison	cm	Quantitative
c50floc1	Circonférence à 50 cm à la floraison	cm	Quantitative
nfefloc1	Nombre de feuilles émises à la floraison	/	Quantitative
nfvfloc1	Nombre de feuilles vivantes à la floraison	/	Quantitative
nrvfloc1	Nombre de rejets vivants à la floraison	/	Quantitative
ipfc1	Intervalle plantation - floraison	jours	Quantitative
nfvrecc1	Nombre de feuilles vivantes à la récolte	/	Quantitative
prc1	Poids du régime	kg	Quantitative
nmrc1	Nombre de mains du régime	/	Quantitative
ndrc1	Nombre de doigts du régime	/	Quantitative
ifcc1	Intervalle de temps entre la floraison et la récolte au cycle 1	jours	Quantitative
ipcc1	Intervalle de temps entre la plantation et la récolte au cycle 1	jours	Quantitative
pdrc1	Poids moyen des doigts du régime = (prc1 / ndrc1) * 1000	g	Quantitative

Le tableau 4 présente les variables les plus importantes du jeu de données et celles qui seront utilisées pour cette analyse.

Tableau 5 : Statistiques descriptives globales des variables principales au cycle 1. Les acronymes sont explicités au tableau 4.

	htfloc1	c50floc1	hbulbfloc1	nfefloc1	nfvfloc1	rdfvfloc1	nrvfloc1	prc1	nmrc1	ndrc1	ifcc1	ipc1
Unité :	*cm*	*cm*	*cm*	*/*	*/*	*/*	*/*	*kg*	*/*	*/*	*jours*	*jours*
Minimum	187	46,5	0	31	6	4	1	0	0	0	51	313
1er quartile	290	58,5	6	36	8	29	5	9,6	7	37	82	350
Médiane	315	62,6	7	38	9	30	8	12,5	7	68	89,5	363
Moyenne	323,47	66,84	8,43	40,05	9,03	31,91	8,08	13,26	7,45	74,18	89,8	393,08
2eme quartile	338,75	72	10	42	10	33	10	15,8	9	93	97	413
Maximum	504	103,5	20	63	19	56	32	39,5	21	257	125	628
IQR	48,75	13,5	4	6	2	4	5	6,2	2	56	15	63
sd	53,66	12,14	3,28	5,60	1,44	5,71	3,96	6,35	2,68	51,56	11,56	64,23
IC	4,22	0,95	0,31	0,44	0,11	0,45	0,31	0,5	0,21	4,04	0,93	5,11
CV (%)	17	18	39	14	16	18	49	48	36	70	13	16
NA's	*40*	*40*	*237*	*40*	*40*	*40*	*48*	*42*	*37*	*37*	*64*	*55*

Légendes : - IQR : écart inter-quartile *- IC : intervalle de confiance de la moyenne à 95%*
- sd : écart-type *- CV : coefficient de variation (sd/Moyenne) * 100*

Concernant les variables du tableau 5, la hauteur moyenne au cycle 1 d'un pseudo-tronc de bananier plantain (*htfloc1*), toutes variétés confondues, est de 323,5 ± 4,22 cm (*Moyenne ± IC ; Tab. 5)*. La circonférence à 50 centimètres du pseudo-tronc (*c50floc1*) est de 66,8 ± 0,95 cm de diamètre. Le nombre de feuilles émises (*nfefloc1*) moyen est de 40. Le nombre de rejets vivants (*nrvfloc1*) est de 7. Le poids du régime (*prc1*) moyen est de 13,3 ± 0,5 kg. Le nombre moyen de doigts du régime (*ndrc1*) est de 74 ± 4 et le nombre moyen de main du régime (*nmrc1*) est de 7. Le temps moyen, entre la plantation d'un bananier plantain et sa récolte (*ipc1*), est de 393 ± 5 jours au cycle 1.
Le coefficient de variation (CV) est utilisé pour représenter la dispersion des données pour des variables de rapport ou de ratio *(Moltusky, 2019)*. La variable ayant les données les plus dispersées autour de la moyenne est le *ndrc1* avec un CV de 70%. Ensuite, c'est le *nrvfloc1* et le *prc1* qui ont un CV de 49% et de 48%. La *hbulbfloc1* et le *nmrc1* ont un CV de 39% et 36%. La *c50floc1*, la *htfloc1*, le *rdfvfloc1*, le *nfvfloc1*, l'*ipc1*, l'*ifcc1* et le *nfefloc1* ont un CV entre 13 et 18%.

b. Description des variétés

Malgré le faible effectif de bananiers plantain par variété *(n ≤ 5)*, il était important de décrire les variétés de la collection avec les variables qui ont été prises. Un sommaire pour le cycle 1 sous forme de data table *(Annexe 4)* et 5 dossiers Excel, correspondant aux 5 cycles, ont été générés avec le logiciel R Studio pour y regrouper les statistiques descriptives par variété et triées par ordre alphabétique. Ces fichiers contiennent les informations d'une sélection de variables pour le data table et toutes les variables pour les fichiers Excel. Les statistiques descriptives classiques ont été calculées comme la moyenne, la médiane, le minimum, le maximum, le nombre de données manquantes, l'écart type, l'écart interquartile, les effectifs... Le travail accomplit sur Shiny, décrit en partie 4, utilise et valorise cette description statistique par une visualisation et une navigation facilitée au sein de ces données.

c. Description des groupes botaniques : Taille et Type

Il existe deux façons de classer les variétés de bananiers plantain et ses informations sont données par le Musalogue : par la Taille et le Type. Les différents groupes représentés par le jeu de données du cycle 1 sont décrits et explorés dans cette partie.

La classification par Taille est une classification donnée par le Musalogue. La taille d'un bananier plantain est ici définie comme un critère botanique basé sur le nombre de feuilles émises à la floraison. Le bananier plantain étant une monocotylédone érigée, plus le nombre de feuille est important et plus sa taille au sens commun tend à être élevée. Si le nombre de feuilles émises est inférieur à 32, la variété de bananier plantain est classée dans le groupe Petit, s'il est entre 32 et 38 elle est classée dans le groupe Moyen et s'il est supérieur à 38 elle est classée dans le groupe Géant.

Tableau 6 : Statistiques descriptives du nombre de feuilles émises à la floraison au cycle 1 (*nfefloc1*) par groupe de Taille (Musalogue).

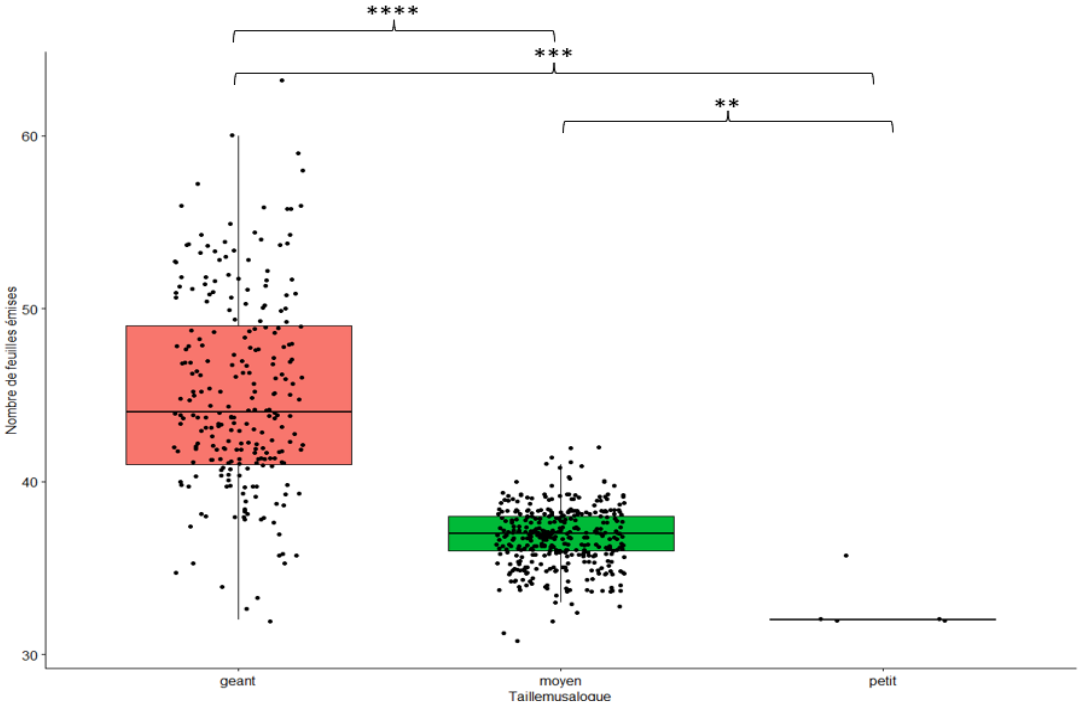
Statistiques nfefloc1 / Taille	*n*	Min.	Max.	Médiane	IQR	Moyenne	IC	Ecart-type
Géant	*247*	32	63	44	8	45,2	0,693	5,53
Moyen	*373*	31	42	37	2	36,7	0,174	1,71
Petit	*5*	32	36	32	0	32,8	2,22	1,79

- *n : effectif*
- *Min. : Minimum*
- *Max. : Maximum*
- *IQR : écart inter-quartile*
- *IC : intervalles de confiance à 95% de la moyenne*

Les bananiers plantain du groupe Géant ont une moyenne de nombre de feuilles émises (*nfefloc1)* d'environ 45 ± 1 *(Moyenne ± IC ; Tab. 6)* pour un minimum de 32 et un maximum de 63. Ceux du groupe Moyen ont une moyenne de *nfefloc1* d'environ 37 avec un minimum de 31 et un maximum de 42. Enfin, ceux du groupe Petit ont une médiane d'environ 32 pour un minimum de 32 et un maximum de 36.

Les statistiques descriptives de la variable *nfefloc1* par groupes de Taille issus du jeu de données de la collection ne sont pas totalement concordante avec la classification du Musalogue. En effet, dans le tableau 6, les minimums et maximums ne correspondent pas aux seuils définis par le Musalogue mais les valeurs moyennes (ou médianes pour le groupe Petit) se trouvent bien dans l'intervalle de nombre de feuilles émises à la floraison définissant les groupes.

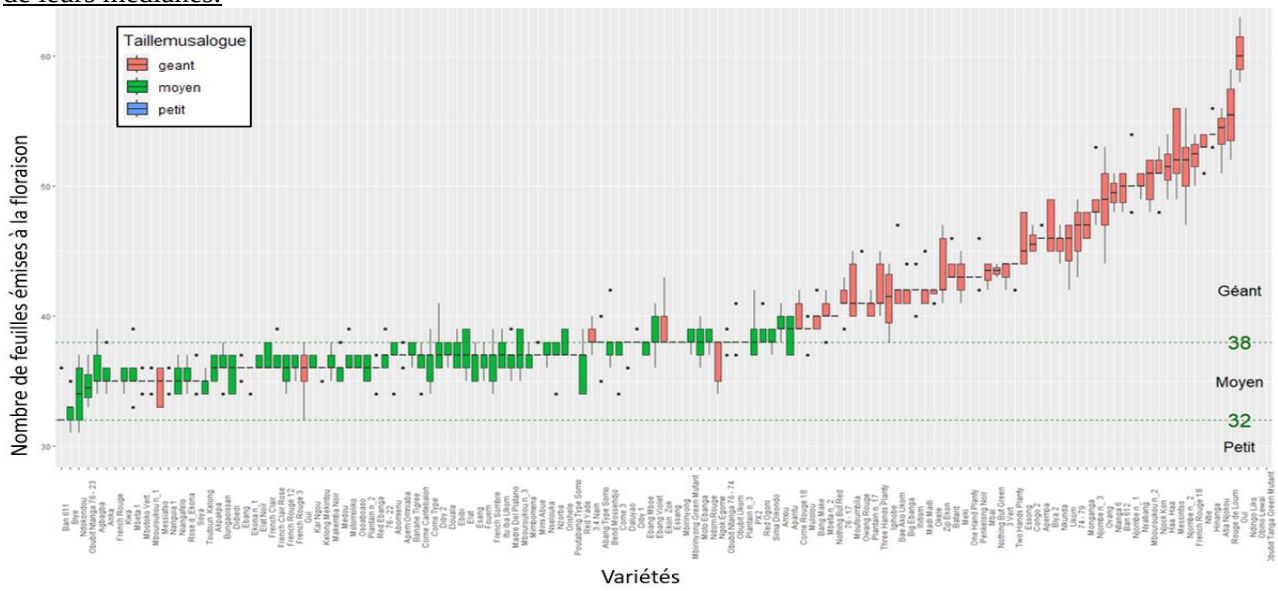
Figure 1 : Boxplots du *nfefloc1* selon les groupes Taille tel que renseigné dans le Musalogue (Taillemusalogue) avec résultat d'un test multiple de Wilcoxon.



*Résultats Wilcoxon (α = 5%) p-value adj. (Bonferroni) : ** $3e^{-3}$; ***$4{,}98e^{-4}$; ****$1{,}6e^{-79}$*

L'hypothèse selon laquelle la variable *nfefloc1* suit une loi normale est rejetée en global avec le test de Shapiro (α = 5% ; p-value = $1{,}71e^{-22}$) et par représentation graphique. Les tests non paramétriques ont été privilégiés. L'hypothèse selon laquelle il existe au moins un groupe Taille différent est validée avec le test non paramétrique de Kruskal-Wallis (α = 5% ; p-value ≈ 3.17e-80).
Un test multiple de Wilcoxon avec ajustement des p-valeurs avec la méthode de Bonferroni a été effectué par la suite pour tester les différences entre groupes deux à deux *(Fig. 1)*. Malgré le fait que les observations du jeu de données ne soient pas en concordance parfaite avec la classification du Musalogue, les résultats montrent que les trois groupes Taille sont significativement différents entre eux sur la base des valeurs observées. L'effectif du groupe Petit n'étant que de 5 bananiers plantains le résultat est à prendre avec précaution.

Figure 2 : Boxplots du *nfefloc1* selon les variétés de bananiers plantain triées par ordre croissant en fonction de leurs médianes.

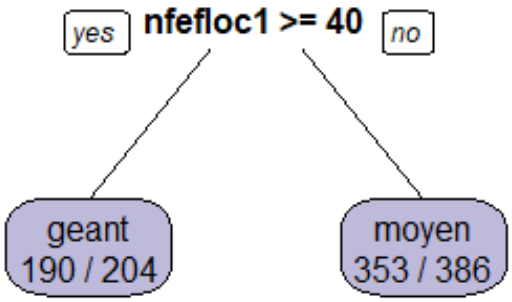


Le Figure 2 représente les boxplots de la variable *nfefloc1* par variétés classées par ordre croissant. Les trois groupes renseignés dans le Musalogue se devinent assez facilement, avec quelques erreurs. La variété « Ban 611 » de taille petite est la première variété à gauche.

Une variable a été créée temporairement sur les observations de la variable *nfefloc1* du jeu de données en fonction de la classification du Musalogue dans le but de confronter les deux évaluations par un test de concordance de Kappa de Cohen. Le résultat de la statistique de kappa s'élève à 0,76 (IC95% ± 0,10 ; p-value < 0,05), l'hypothèse que la concordance relève du hasard est rejetée. Le groupe Taille définie par le Musalogue concorde, d'après l'échelle établie par McHugh (2012), de façon modérée avec les mesures observées. Cela correspond à environ 87% de bonnes paires et 13% d'erreurs.

Un arbre de classification, avec la taille Musalogue en fonction du nombre de feuilles émises observé, a été généré sur R Studio® avec la *librarie rpart*. L'arbre de classification permet de trouver le meilleur arbre qui permet de partitionner de façon binaire des individus par rapport à *n* variables et permet d'évaluer le taux d'individus bien classés en fonction de groupes déjà existants *(Bertsimas et al, 2017)*. Cela complète l'exploration des groupes.

Figure 3 : Arbre de classification des groupes Taille (Musalogue) avec le nfefloc1.



L'arbre de classification indique un seuil de 40 feuilles, entre les groupes Moyen et Géant, et non pas 38. Le taux de bananiers plantain bien classé s'élève à 92%. Le groupe Petit a un effectif trop faible pour être comptabilisé dans l'arbre de classification. *(Fig. 3)*

Une nouvelle variable représentant les trois groupes de Taille a été codé de la façon suivante : *nfefloc1* ≥ 40 : Géant, *nfefloc1* ∈]32 ; 40] : Moyen et *nfefloc1* < 32 : Petit. La variable se nomme *new_Taille* et le choix de continuer les analyses sur la nouvelle variable *new_Taille* a été décidé par les chercheurs.

La classification botanique par Type dans le Musalogue est liée au stade de dégénérescence de l'inflorescence. Aucune mesure effectuée ne permet de décrire parfaitement cette dégénérescence. Néanmoins, une variable qui discrimine au mieux les groupes a tout de même été trouvée en explorant les données. Le nombre de doigts du régime (*ndrc1*) est la variable qui classe le mieux les groupes Type, le degré de dégénérescence de l'inflorescence se traduisant par des conformations de régime différent. Un doigt c'est une banane sur le régime.

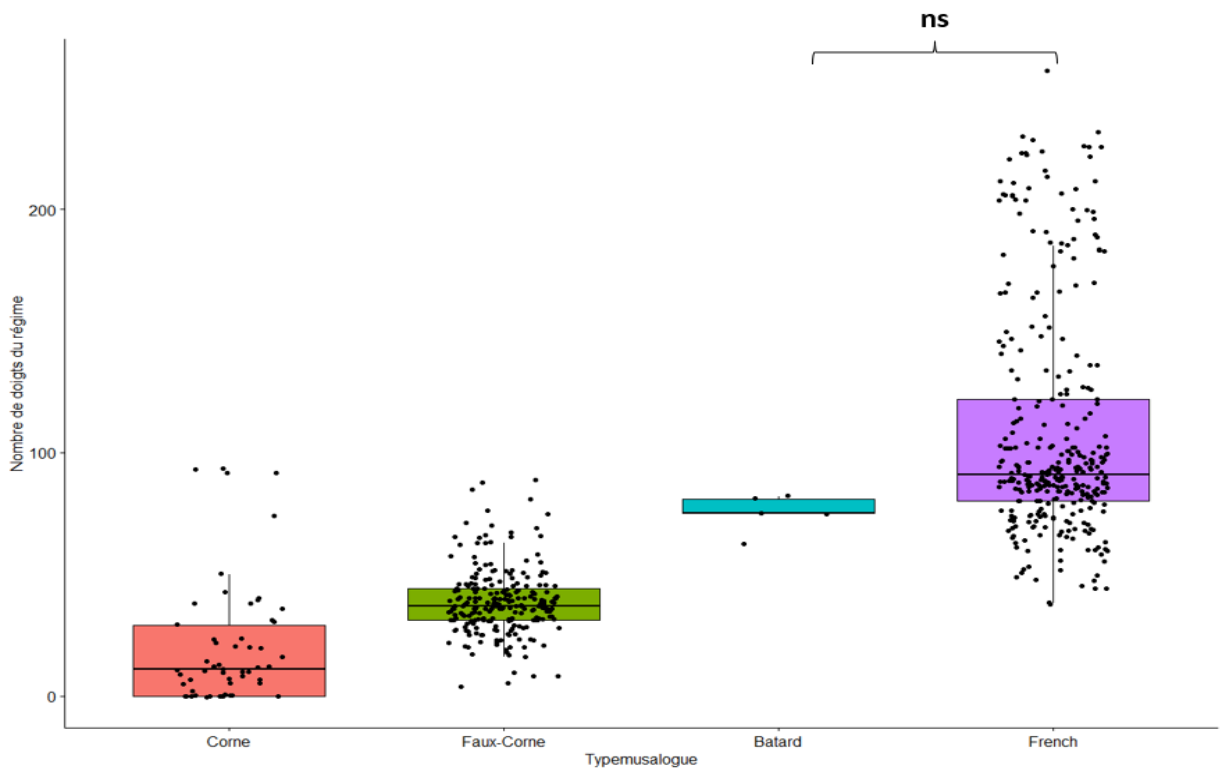
Tableau 7 : Statistiques descriptives du nombre de doigts du régime au cycle 1 (*ndrc1*) par groupe Type.

Statistiques ndrc1 / Type	*n*	Min.	Max.	Médiane	IQR	Moyenne	IC	Ecart-type
French	*333*	38	257	91	42	108	5,07	47
Batard	*5*	63	32	75	6	75,2	9,39	7,56
Faux-corne	*231*	4	89	37	13	38,8	1,79	13,8
Corne	*57*	0	94	11	29	20,1	6,74	25,4

- *n : effectif*
- *Min. : Minimum*
- *Max. : Maximum*
- *IQR : écart inter-quartile*
- *IC : intervalle de confiance à 95% de la moyenne*

Les bananiers plantain du groupe French ont une moyenne de nombre de doigts du régime (*ndrc1)* d'environ 108 ± 5 *(Moyenne ± IC ; Tab. 7)*. Ceux du groupe Faux-corne ont une moyenne de *ndrc1* d'environ 39 ± 2. Ceux du groupe Corne ont une moyenne de *ndrc1* d'environ 20 ± 7. Enfin, le groupe Batard a une médiane de *ndrc1* de 11.
Les statistiques descriptives de la variable *ndrc1* par groupes de Type issus du jeu de données de la collection montrent que ce sont les bananiers plantain du groupe French qui ont en moyenne le plus de doigts sur leurs régimes, ensuite les Batard, les Faux-corne et en dernier les Corne. Cela correspond aux différents stades de l'inflorescence décrits dans le Musalogue : un French à une inflorescence complète, les Batard se trouve entre les French et les Faux-corne, les Faux-corne ont une inflorescence partiellement complète et les Corne ont une inflorescence incomplète.

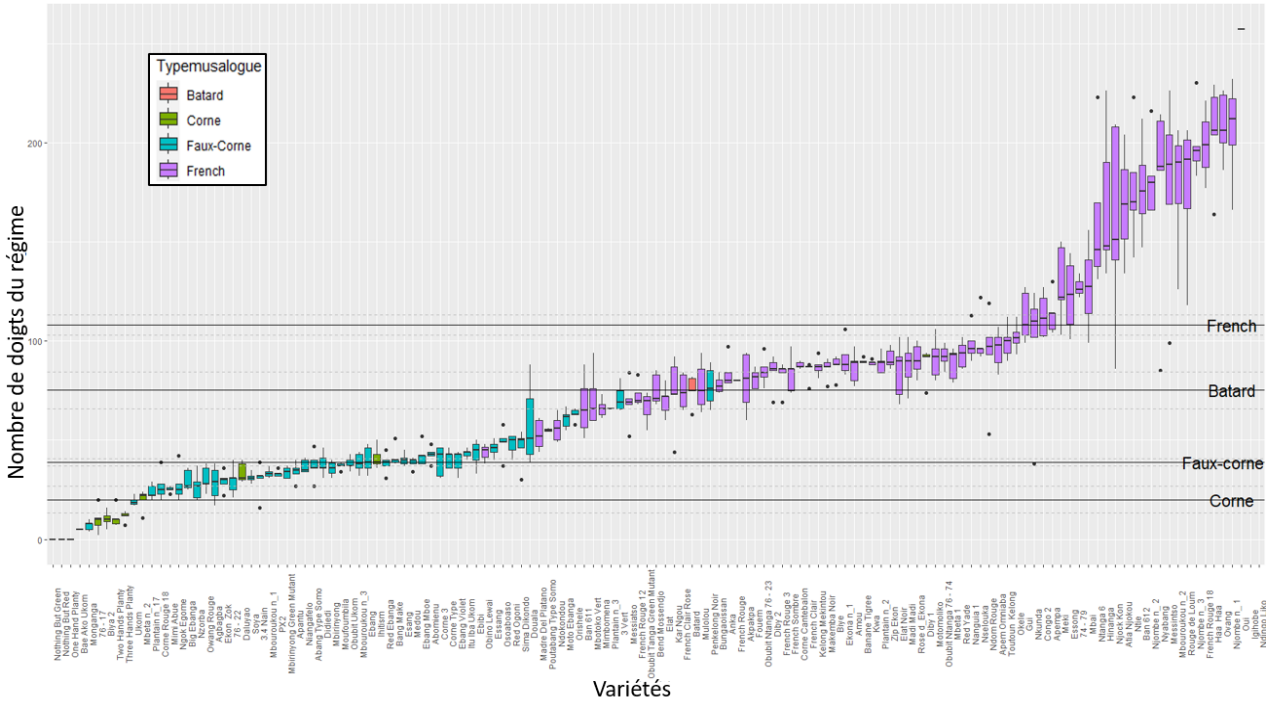
Figure 4 : Boxplots du *ndrc1* selon les groupes Type avec résultats d'un test multiple de Wilcoxon.



ns : p-valeur non significative

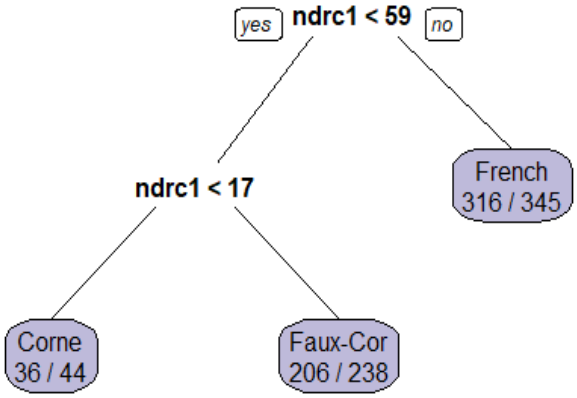
L'hypothèse selon laquelle la variable *ndrc1* suit une loi normale est rejetée en global avec le test de Shapiro (α = 5% ; p-value = $2{,}5e^{-20}$) et par représentation graphique. Les tests non paramétriques ont été privilégiés. L'hypothèse selon laquelle il existe au moins un groupe Type différent est validée avec le test non paramétrique de Kruskal-Wallis (α = 5% ; p-value ≈ $1{,}03e^{-93}$).
Un test multiple de Wilcoxon avec ajustement des p-valeurs avec la méthode de Bonferroni a été effectué par la suite pour tester les différences entre groupes deux à deux *(Fig. 4)*. Les résultats montrent que les quatre groupes Type sont significativement différents entre eux sur la base des valeurs observées, sauf entre le groupe French et Batard. L'effectif du groupe Batard n'étant que de 5 bananiers plantains le résultat est à prendre avec précaution. Le nombre de doigts du régime est ainsi différent pour chaque groupe Type.

Figure 5 : Boxplots du *ndrc1* selon les variétés de bananiers plantain triées par ordre croissant en fonction de leurs médianes.



Les différents groupes Type sont bien visibles, il y a cependant des variétés Corne qui se mélangent avec les variétés Faux-corne *(Fig. 5)*. Les seuils visibles sur la figure 5 représentent les moyennes par groupe et leurs intervalles de confiance. Il faut noter que le type Batard est un groupe qui se trouve entre les Faux-cornes et les Frenchs et que c'est le cas sur les figures 4 et 5 même s'il ne représente qu'une variété.

Figure 6 : Arbre de classification de ndrc1 en fonction des types.

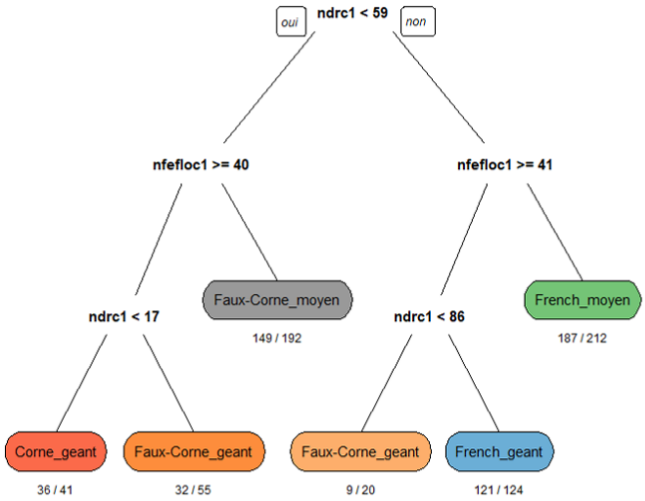


Un arbre de classification *(Fig. 6)* du nombre de doigts du régime en fonction des groupes Type a été généré. Les bananiers plantains du groupe Corne ont été classés à un seuil de *ndrc1* < 17, ceux appartenant au groupe Faux-corne ont été classés entre un *ndrc1* de 17 et 59 et ceux appartenant au groupe French ont été classés selon un seuil de *ndrc1* strictement supérieur à 59. Les pourcentages d'individus bien classés sont de 92% pour les French, 87% pour les Faux-corne et 82% pour les Corne. Le type Batard n'est pas assez bien représenté pour être classé par l'arbre. Cela fait une moyenne de 89% en tout. Cela confirme que le nombre de doigts du régime (*ndrc1*) pourrait refléter le type botanique des bananiers plantain et pourrait être un outil d'aide à la reconnaissance à la floraison.

Maintenant que les variables qui discriminent le mieux les groupes botaniques ont été mises en évidence, elles ont été regroupées dans un arbre de classification pour résumer et simplifier cette classification. Un arbre a été généré pour les groupes Type et *new_Taille* en fonction des variables *nfefloc1* et *ndrc1*.

Figure 7 :

(a) Arbre de classification des variables nfefloc1 et ndrc1 en fonction des groupes Type_Taille.



(b) : Résumé de l'arbre de classification des groupes Type_Taille.

Groupe :	Bien-classé	Effectif	nfefloc1	ndrc1
Corne_geant	87,8%	41	≥ 40	< 17
Faux-corne_moyen	77,6%	192	≤ 40	< 59
Faux-corne_geant	54,7%	75	≥ 40	[17 ; 86]
French_moyen	88,2%	212	≤ 41	< 59
French_geant	97,6%	123	≥ 40	> 86

Les bananiers plantain du groupe de French_géant sont les mieux classés *(97,6%, Fig. 7-b)* et les moins bien classés sont ceux appartenant au groupe de Faux-corne_géant *(54,7%, Fig. 7-b)*. Les groupes de Batard_géant, des Corne_moyen et de French_petit ne peuvent pas être inclus dans l'arbre. En moyenne les groupes ont un taux d'individus de bien classés de 81,2%. Les statistiques descriptives complètes sont en annexe 5 par groupe de *new_Taille* (a) et par groupe Type (b).

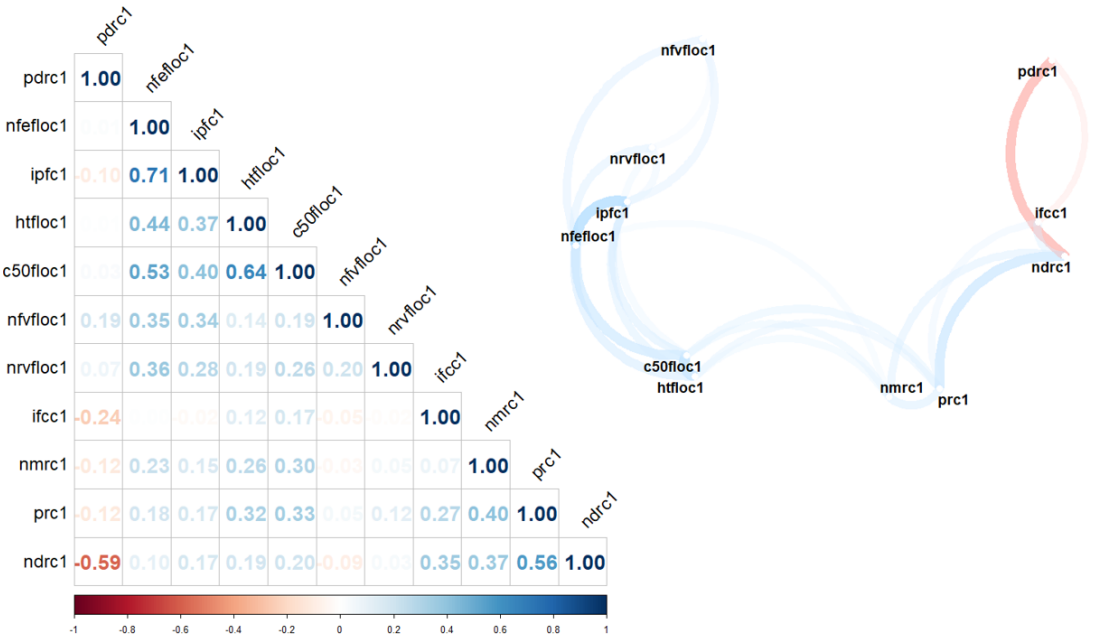
II. Analyse exploratoire

En seconde partie, le but était d'explorer la structure du jeu de données en observant les liens entre les différentes variables mesurées.

a. Corrélation entre variables en global

Pour calculer la matrice des corrélations entre variables, le test non paramétrique de Kendall est privilégié car les variables quantitatives du jeu de données du cycle 1 sont des variables ordinales *(Chen, 2002)*.

Figure 8 : Corrélations de Kendall d'une sélection de variables du cycle 1, avec deux représentations différentes.



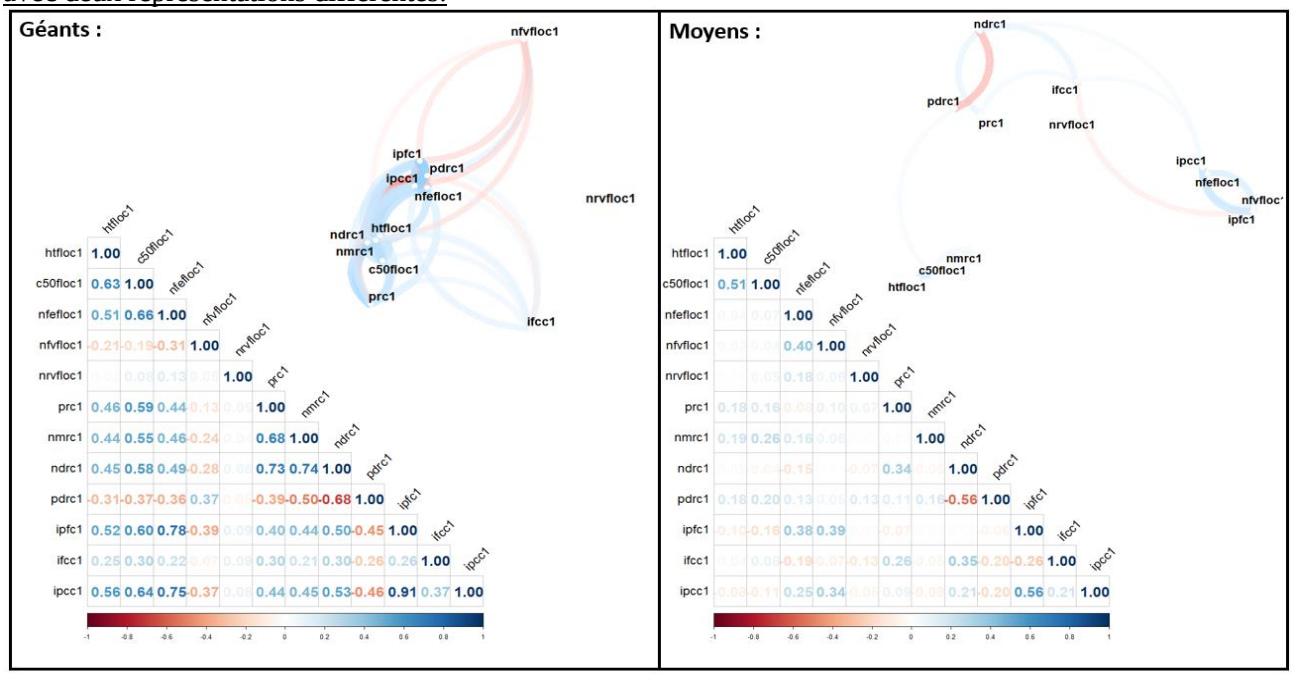
La figure 8 montre deux représentations des corrélations des variables du cycle 1. Il existe plusieurs corrélations positives et négatives entre les différentes variables mesurées au cycle 1. La corrélation la plus importante s'élève à 0,71 *(Fig. 8)* entre le nombre de feuilles émises (*nfefloc1*) et l'intervalle de temps entre la plantation et la floraison (*ipfc1*). En conditions de culture non limitantes, le rythme d'émission des feuilles est relativement régulier ; il est donc cohérent que le nombre de feuilles émises (*nfefloc1*) et l'intervalle entre la plantation et la floraison (ipfc1) soient corrélé. Le nombre de feuilles émises (*nfefloc1*) est également corrélé avec la hauteur du pseudo-tronc (*htfloc1* ; 0,53) et avec la circonférence à 50 centimètres du pseudo-tronc (*c50floc1* ; 0,64). L'intervalle de temps entre la plantation et la floraison (*ipfc1*) a tendance à avoir des corrélations positives plus fortes avec les variables relatives à la croissance de la plante qu'avec celles du régime. L'intervalle entre la floraison et la récolte (*ifcc1*) a tendance à avoir des corrélations positives plus forte avec les variables relatives au régime. Des corrélations existent entre les différents éléments du régime et son poids. Elle est plus forte pour le nombre de doigts que pour le nombre de mains mais cela s'explique par le fait que la variable nombre de mains ne varie pas beaucoup entre les différentes variétés ; en effet la variable nombre mains (*nmrc1*) à un coefficient de variation de 36% comparé à celle du nombre de doigts (*ndrc1*) qui s'élève à 70% *(Tab. 5)*. Le poids des doigts du régime (*pdrc1*) est négativement corrélé avec toutes les mesures en général mais il y a une faible corrélation (0,19) avec le nombre de feuilles vivantes. Sa plus forte corrélation est de -0,59 avec le nombre de doigts du régime, cela peut se traduire par le fait que plus il y a de doigts sur le régime (*ndrc1*) et moins son poids (*prc1*) a tendance à être élevé. Le nombre de rejets vivants (*nrvfloc1*) est la variable la plus corrélée avec le nombre de feuilles émise (*nfefloc1* ; 0,36).

Les corrélations sont cohérentes avec le comportement connu des bananiers dessert cultivés. Plus le pseudo-tronc d'un bananier plantain est grand, plus il est susceptible d'avoir une circonférence importante, de produire plus de feuilles et plus il est capable de produire un régime lourd.

b. Corrélation entre variables par groupe botaniques

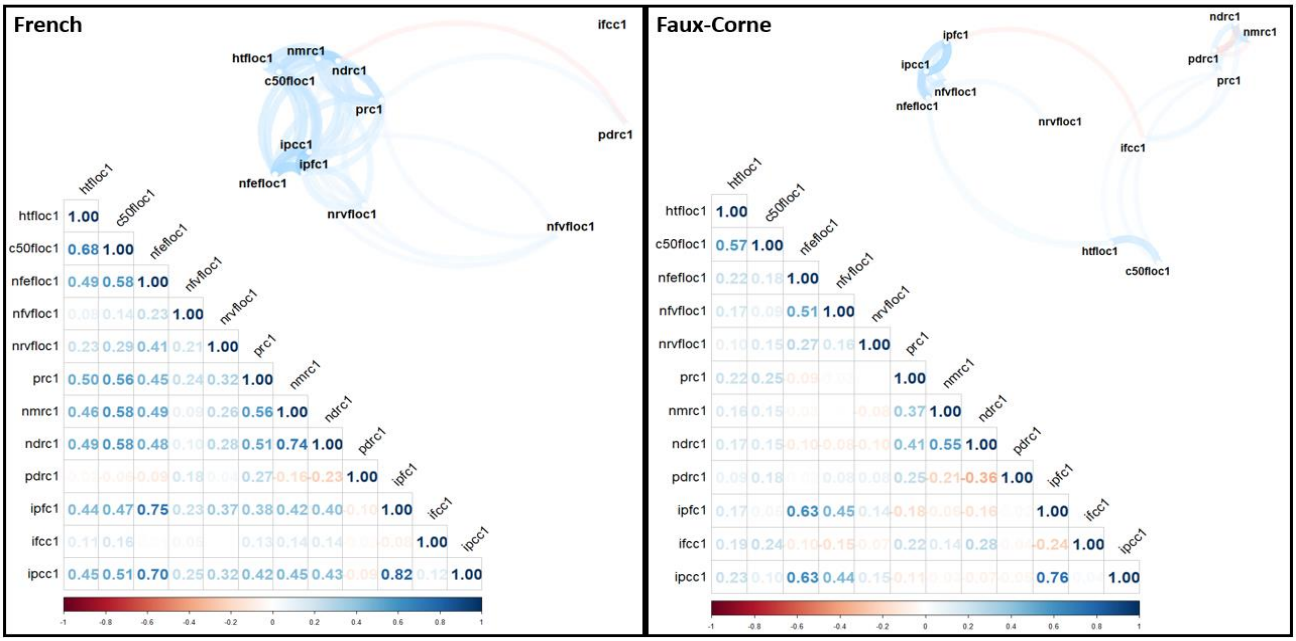
Le comportement des corrélations en global ne reflète pas nécessairement le comportement au sein des différents groupes séparément. Des corrélations différentes entre groupes pourraient suggérer des fonctionnements différents des bananiers plantain appartenant à différents groupes.

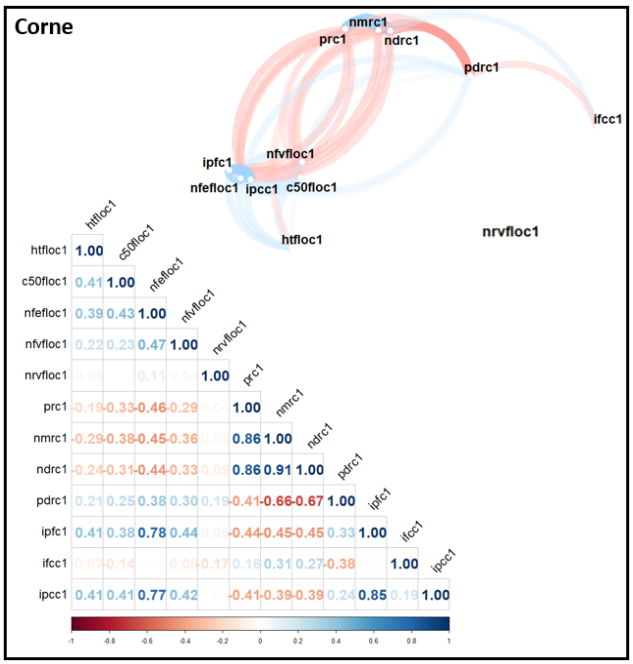
Figure 9 : Comparaison des corrélations de Kendall entre les groupes Taille (new_Taille) Géant et Moyen, avec deux représentations différentes.



La différence la plus notable en premier lieu est la différence d'intensité des corrélations entre les deux groupes *(Fig. 9)*. Il y a une tendance à avoir des corrélations plus fortes entre les variables chez le groupe Géant que chez le groupe Moyen. Pour le groupe Géant, il existe de fortes corrélations positives entre la plupart des variables (entre 0,40 et 0,91 en majorité) et des corrélations négatives et (-0,13 et -0,6) concernant le nombre de feuilles vivantes à la floraison (*nfvfloc1)* et le poids des doigts du régime (*pdrc1)* avec les autres variables. Pour le groupe Moyen, les corrélations sont plus faibles et sont plutôt inférieures à 0,40 sauf entre la *htfloc1* et la *c50floc1* (0,51), entre l'*ipfc1* et l'*ipcc1* (0,56) et entre le *nfefloc1* et le *rdvfloc1* (0,67). Le nombre de mains (*nmrc1*) n'est pas corrélé avec le *prc1* et le *ndrc1* mais un peu plus avec la *c50floc1* et la *htfloc1* (entre 0.20 et 0,26). Le *prc1* n'a pas de corrélations (négligeable : -0,12 à 0,18) avec les variables de croissance. Le nombre de feuilles émises (*nfefloc1*) est corrélé avec les intervalles de temps correspondant à la croissance de la plante (*ipfc1* et *ipc1*).

Figure 10 : Comparaisons des corrélations de Kendall entre les groupes Type : French, Faux-corne et Corne, avec deux représentations différentes.





Les corrélations par groupe de Type *(Fig. 10)* montrent encore que les différents groupes ont des tendances différentes. Pour le groupe French, les variables sont toutes corrélées positivement entre elles avec une assez forte intensité (> 0,30). Le groupe Faux-corne a des corrélations positives et négatives avec une intensité moins forte que le groupe French (< 0,30 pour la plupart). Les éléments du régime sont corrélés positivement entre eux ([0,37 ; 0,55]). La *htfloc1*, la *c50floc1* et le *nfefloc1* sont corrélées entre elles ([0,39 ; 0,43]). Le groupe Corne a globalement des corrélations plus intenses que le groupe Faux-corne. Les éléments du régime sont fortement corrélés entre eux ([0,86 ; 0.91]). Les variables de croissance (*htfloc1*, *c50floc1*, *nfefloc1*, *nfvfloc1*) sont corrélées entre elles ([0,22 ; 0,87]). Les éléments du régime sont corrélés négativement avec les variables de croissance.

Ce qui intéressant c'est que cela soutient l'hypothèse que les groupes botaniques ont bien des fonctionnements différents entre eux. Ce qui renforce l'hypothèse que la dégénérescence de l'inflorescence entraîne également des conséquences sur le fonctionnement interne du bananier. Cela pourrait permettre d'émettre une hypothèse : les bananiers plantain du groupe French sont non limités et répartissent leurs potentiels de matière végétale dans tous les éléments de la plante alors que les autres groupes font des compromis.

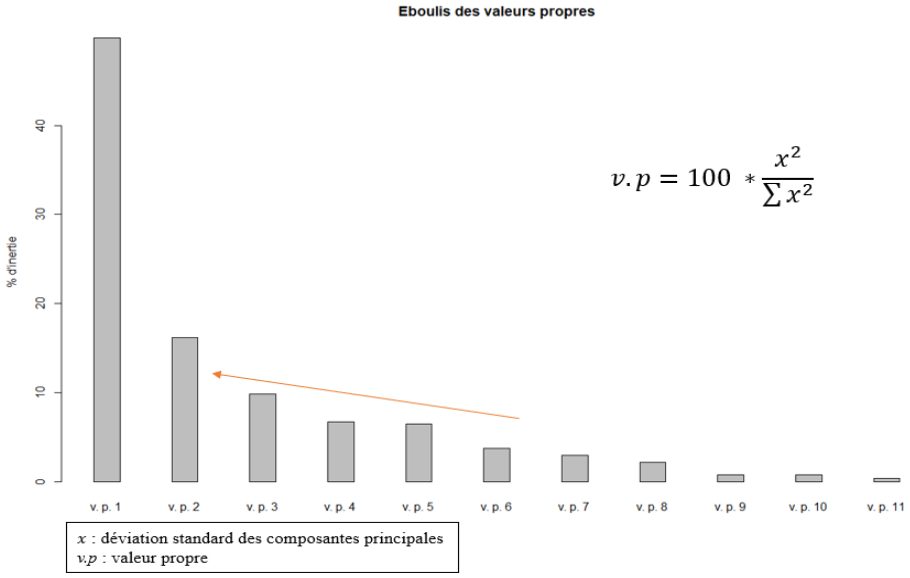
Ces différents comportements entre groupe botanique peuvent être observés aussi avec l'Analyse en Composantes Principales (ACP) faite à la suite.

c. ACP globale

L'Analyse en Composante Principale (ACP), est une méthode mise au point en 1901 par Pearson et par la suite reprise par Hotelling en 1933. Elle permet de générer la meilleure présentation possible des observations et des variables dans un même espace engendré par des combinaisons linéaires des variables de départ. Cette méthode ne supportant pas les données manquantes, elles sont omises du jeu de données. L'effectif est réduit à 588 dont 126 variétés. Il y a environ 53% de French, 39% de Faux-corne et moins de 1% de Batard (5 plants). Pour les groupes Taille, il y a environ 62% de Moyen, 37% de Géant et moins 1% de Petit (4 plants). Pour rendre comparable les variables qui n'ont pas la même unité, il faut les centrer et les réduire ce qui est fait par défaut avec la *librairie FactoMineR*. Avec les *librairies ade4* et *FactoMineR*, les nouvelles composantes sont calculées à partir des variables de départ et les résultats de l'ACP sont représentés.

Pour le choix du nombre de composantes en cas de données centrées réduites, il faut prendre l'éboulis des valeurs propres *(Fig.11)*.

Figure 11 :

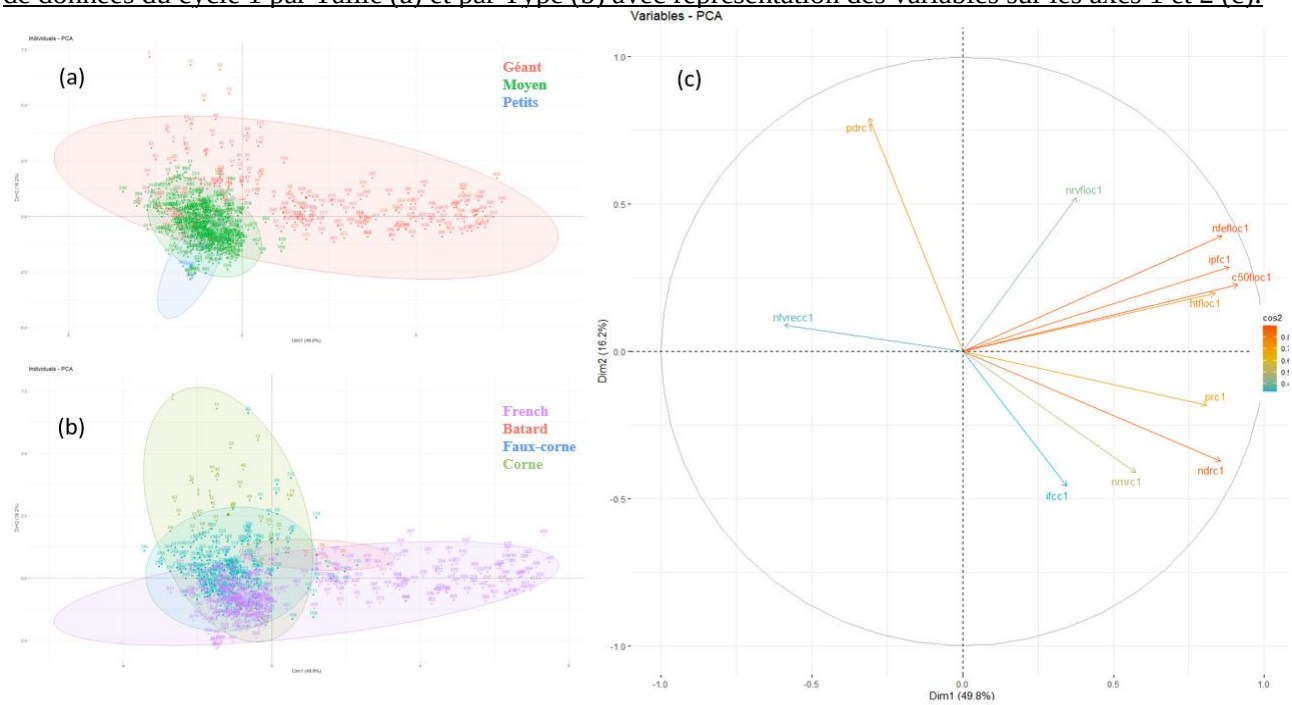


Il faut choisir le nombre de dimensions pour cela il existe deux critères : le critère de Kayser et celui de Saporta *(Saporta, 1990)*. Pour le critère de Kayser, il faut choisir les composantes principales qui ont des valeurs propres supérieures à 1. Dans notre cas, la composante 3 a une valeur propre égale à 1,08.

Pour le calcul du critère de Saporta : $1 + 2 * \sqrt{\frac{(nombre\ de\ variable - 1)}{(nombre\ d'observation - 1)}}$

Le résultat est égal à 1,26, ce qui correspond à la composante 2. Il est préférable de choisir deux dimensions, au vu des résultats des critères et de la Figure 11. L'éboulis des valeurs propres *(Fig. 11)* montre qu'en prenant 2 composantes, 66% de l'inertie est conservée.

Figure 12 : Répartition des individus sur les axes 1 et 2 lors de l'analyse en composantes principales sur le jeu de données du cycle 1 par Taille (a) et par Type (b) avec représentation des variables sur les axes 1 et 2 (c).



La répartition des individus montre qu'il existe des comportements différents entre les groupes Taille qui sont influencés par l'axe 1 *(Fig. 12.a)* et pour les groupes Type qui sont influencés par les deux axes *(Fig. 12.b)*. Les variables *c50floc1*, *nfefloc1* et *ndrc1* sont les variables qui ont la meilleure qualité de représentation dans l'ACP *(Fig. 12.c)*. L'*ifcc1* et *nfvrecc1* sont les variables qui sont le moins bien représentées. Les variables *nfefloc1*, *ipfc1*, *c50floc1* et *htfloc1* sont fortement corrélées entre elles. Les variables *prc1*, *ndrc1* et un peu *nmrc1* sont aussi corrélées entre elles. Ces deux groupes de variables constituent essentiellement l'axe 1. L'axe 1 explique 49,8% de la variabilité. L'axe 2 explique 16,2% de la variabilité, c'est la variable *pdrc1* qui l'explique principalement et l'*ifcc* mais elle est mal représentée.

Elle met en évidence une grande variabilité intergroupe chez les Géants et les Frenchs sur l'axe 1 et aussi avec le groupe des Cornes sur l'axe 2. Contrairement aux groupes Moyen et Faux-corne *(Fig. 12.a et b)*. La représentation des observations sur le plan 1-2 met en évidence un groupe homogène sur le plan 1-2 au sein de l'ensemble des bananiers plantain mais aussi des observations atypiques sur les deux axes. Il se peut que le groupe homogène cache la variabilité inexpliquée.

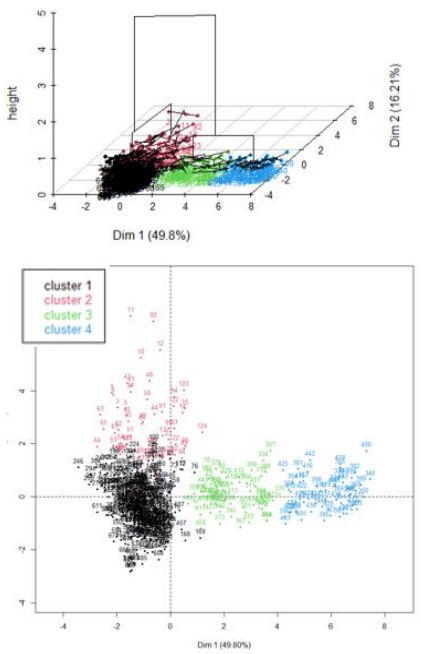
III. Création de nouveaux groupes : HCPC

En troisième partie, les chercheurs ont souhaité créer de nouveaux groupes de variétés sur la base du jeu de données dans le but de créer des groupes qui seraient basés sur des critères agronomiques et non botaniques comme l'est la classification morpho-taxonomique.

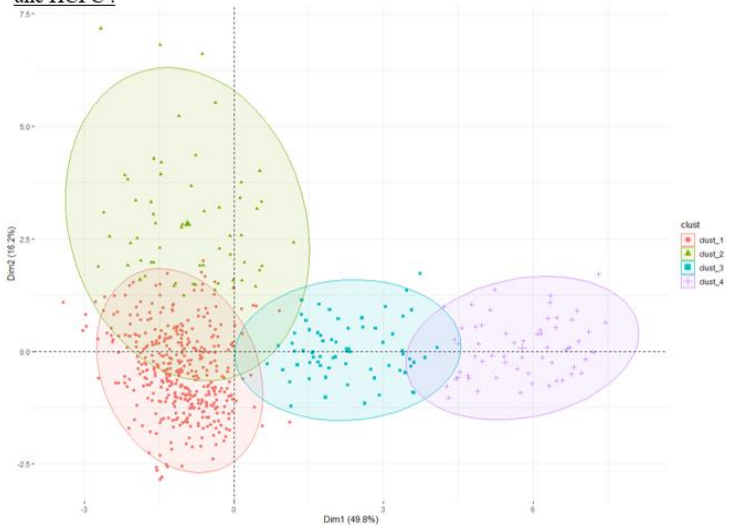
Pour cela, la méthode de classification Hiérarchique sur Composantes Principales (HCPC) est utilisée avec le critère de Ward. C'est un critère basé sur le théorème de Huygens et il décompose l'inertie totale entre les groupes et en son sein. Cela consiste à réaliser une classification hiérarchique sur les composantes principales de l'ACP effectuée précédemment. Les données sont partitionnées en sous-groupes et cela permet de regrouper les individus les plus semblables entre eux tout en maximisant la différence inter groupe *(Husson, 2010)*.

Figure 13 :

(a) Classification hiérarchique sur composantes principales (méthode Ward) :



(b) Projection des individus sur le plan 1/2 avec représentation des clusters après une HCPC :



L'HCPC a divisé les observations en quatre groupes par défaut *(Fig. 13)*. Afin de déterminer si les groupes sont significativement différents entre eux sur les axes, les coordonnées des observations correspondantes aux deux axes sont récupérées pour pouvoir faire une analyse de variance (ANOVA) sur chacune d'elles en fonction des nouveaux groupes. Les groupes ont des effectifs qui sont suffisamment importants *(Tab 9)* pour que le théorème central limite s'applique *(Motulsky, 2019)*. Cela fait deux modèles d'ANOVA à 1 facteur à tester. La normalité des résidus était validée mais pas l'homoscédasticité avec le test de Fligner et par représentation graphique. Le test ANOVA de Welch est utilisé quand l'homoscédasticité n'est pas validé *(Welch, 1951)*. Les deux tests sont significatifs, ce qui veut dire qu'il existe une différence entre les groupes. Le test post-hoc de Games Howell permet de mettre en évidence que tous les groupes sont significativement différents entre eux sauf pour les groupes 1 et 2 sur l'axe 1 et les groupes 3 et 4 sur l'axe 2.

Tableau 9 : Effectifs et proportions (%) de bananiers plantain dans les clusters issus de l'HCPC.

Cluster	**1**	**2**	**3**	**4**	**Total :**
Effectif	**404**	59	65	60	**588**
≈ %	***68,7***	*10,0*	*11,1*	*10,2*	***100***

Tableau 10 : Effectifs et proportions (%) des groupes Type et Taille *(new_Taille)* dans les clusters issus de l'HCPC.

Groupe / Cluster	**Géant**	**Moyen**	**Petit**	**French**	**Faux-Corne**	**Corne**	**Batard**
1	36 *(21,0)*	**360 *(97,0)***	8 (100)	200 *(63,9)*	**196 *(86,0)***	6 *(14,3)*	2 *(40)*
2	**48 *(22,8)***	11 *(2,5)*	0	0	23 *(10,1)*	**36 *(85,7)***	0
3	**60 *(28,8)***	5 *(0,5)*	0	**53 *(16,9)***	9 *(3,9)*	0	3 *(60)*
4	**60 *(27,4)***	0	0	**60 *(19,2)***	0	0	0
Total	**204 *(100)***	376 *(100)*	8 *(100)*	**313 *(100)***	228 *(100)*	42 *(100)*	5 *(100)*

Légendes : Effectifs *(Proportion %)* ; Groupes les moins bien représentés ; **Groupes les plus présents dans le cluster associé**

Le tableau 10 montre la répartition des individus appartenant aux différents groupes initiaux (Taille et Type) dans les nouveaux clusters de l'HCPC. Dans le cluster 1, il y a essentiellement des bananiers plantain du groupe Moyen et Faux-corne *(Tab 10)*. Dans le cluster 2, il y a majoritairement des bananiers plantain Corne et Géant. Dans le cluster 3, il y a majoritairement des bananiers plantain French et Géant. Le cluster 4 quant à lui ne compte que des bananiers plantain French et Géant. Cela permet de voir que les groupes issus de l'HCPC mélangent partiellement les groupes botaniques. Le cluster 1 présente des individus appartenant à un peu tous les groupes botaniques et il représente 68,7% des individus du jeu de données de l'ACP *(Tab. 9)*. Au contraire, les groupes 2, 3 et 4 excluent certains groupes botaniques.

Tableau 11 : Statistiques descriptives (moyenne et écart-type) des nouveaux clusters issus de l'HCPC sur les variables utilisées pour l'ACP, avec résultats de tests statistiques.

Variable[1] :	*c50floc1*		*nfefloc1*		*ipfc1*		*ndrc1*		*htfloc1*		*prc1*		*nmrc1*		*pdrc1*		*nfvrecc1*		*nrvfloc1*		*ifcc1*	
Groupe :	moy	sd	moy	sd	moy	sd	moy	sd	moy	sd	moy	sd	moy	sd	moy	sd	moy	sd	moy	sd	moy	sd
1	60,2	4,3	37	2,1	271	20,8	61,7	24,7	298,8	27,7	11,8	3,4	7,6	1,6	216,8	87,6	2,8	1,8	6,7	3,1	88,1	11,7
2	69,1	6	42,3	3,3	314,8	36,1	18,8	10,7	326,2	29,4	8,4	3,6	3,6	2,7	514,5	217,5	2,2	2	11,1	4,2	85	8,9
3	76,7	6,1	44,1	3,3	355,3	44,5	107,3	25,7	358,5	34,1	17,9	4,3	8,2	1,2	174,5	52,4	0,5	1	10,6	4,5	98,3	6,5
4	93,9	4,5	51,4	2,6	433,9	36,1	189,5	32,1	431,7	53,15	25,7	4,2	11,5	1,3	138,8	30,8	0,13	0,47	10,8	3,5	95,2	10,1
p-value : ANOVA de Welch **ou** Test Kruskal-Wallis	5.21e-85		1.73e-71		1.24e-61		< 2.2e-16		< 2.2e-16		< 2.2e-16		< 2.2e-16		< 2.2e-16		< 2.2e-16		< 2.2e-16		2.18e-21	
p-adjust[2] : Games Howell **ou** Wilcoxon	***		***		***		***		***		***		***		***		1-2		1-2 1-3 1-4		1-2 3-4	

[1] : classées par ordre d'importance (HCPC)
moy : moyenne
sd : écart-type de la moyenne

[2] : ajustement avec la méthode de Tukey et avec Bonferroni
**** : tous les groupes deux à deux sont significativement différents au seuil α 5%*

Le tableau 11 résume les moyennes et écart-types des variables utilisées dans l'ACP par cluster et regroupe les tests statistiques pour échantillon multiple appliqués par variable en fonction des nouveaux clusters. Les variables sont présentées par ordre d'importance des résultats d'HCPC. Pour chacune des variables, un premier test statistique a été effectué pour tester les différences entre cluster. Si les conditions le permettait une ANOVA de Welch était effectuée sinon un test non paramétrique de Kruskal-Wallis était utilisé. Si à l'issue du test la p-valeurs était significative au risque alpha 5% , alors un test multiple avec ajustement des p-valeurs était effectué, soit celui de Games-Howell dans le premier cas, soit celui de Wilcoxon pour tester les groupes 2 à 2. L'ANOVA de Welch est utilisée quand l'hypothèse d'homoscédasticité des résidus ne pouvait pas être validée dans le cadre d'une ANOVA. Les conditions de normalité des résidus et d'indépendance des résidus ont été évalués avec des tests statistiques et avec visualisation graphique.

Toutes les variables ont des différences significatives et cela dans chacun des clusters deux à deux sauf pour le *nfvrecc1*, le *nrvfloc1* et *l'ifcc1*. Pour le *nfvrecc1*, les clusters 1-2 sont différents entre eux mais pas les autres. Pour *nrvfloc1*, seul le cluster 1 est différent. Pour *l'ifcc1*, seul les couples de clusters 1-2 et 3-4 sont différents entre eux.

Les 4 clusters sont différents par leurs coordonnées spatiales de l'ACP et à travers les variables qui y ont été utilisées, surtout avec celles qui ont été importantes. Les variables *nfvrecc1*, *nrvfloc1* et *l'ifcc1* ne permettent pas de différencier significativement les 4 clusters entre eux. L'HCPC a partitionnée au mieux les observations projetées sur le plan 1-2 de l'ACP.

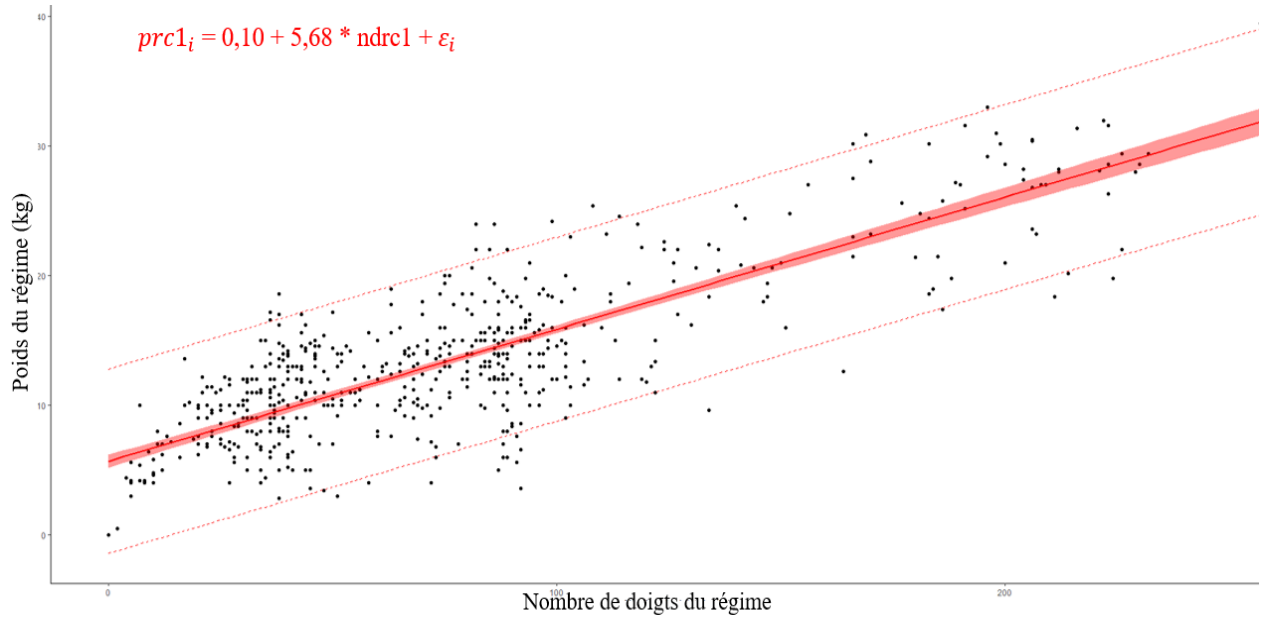
IV. Prédiction du poids du régime

En quatrième partie, l'objectif était d'essayer de prédire le poids du régime avec plusieurs méthodes.

a. Modèle linéaire

Comme vu précédemment, la variable du nombre de doigts est bien corrélée au poids du régime. Une régression linéaire est effectuée pour essayer de modéliser la variable quantitative continue "poids du régime" *(prc1)*.

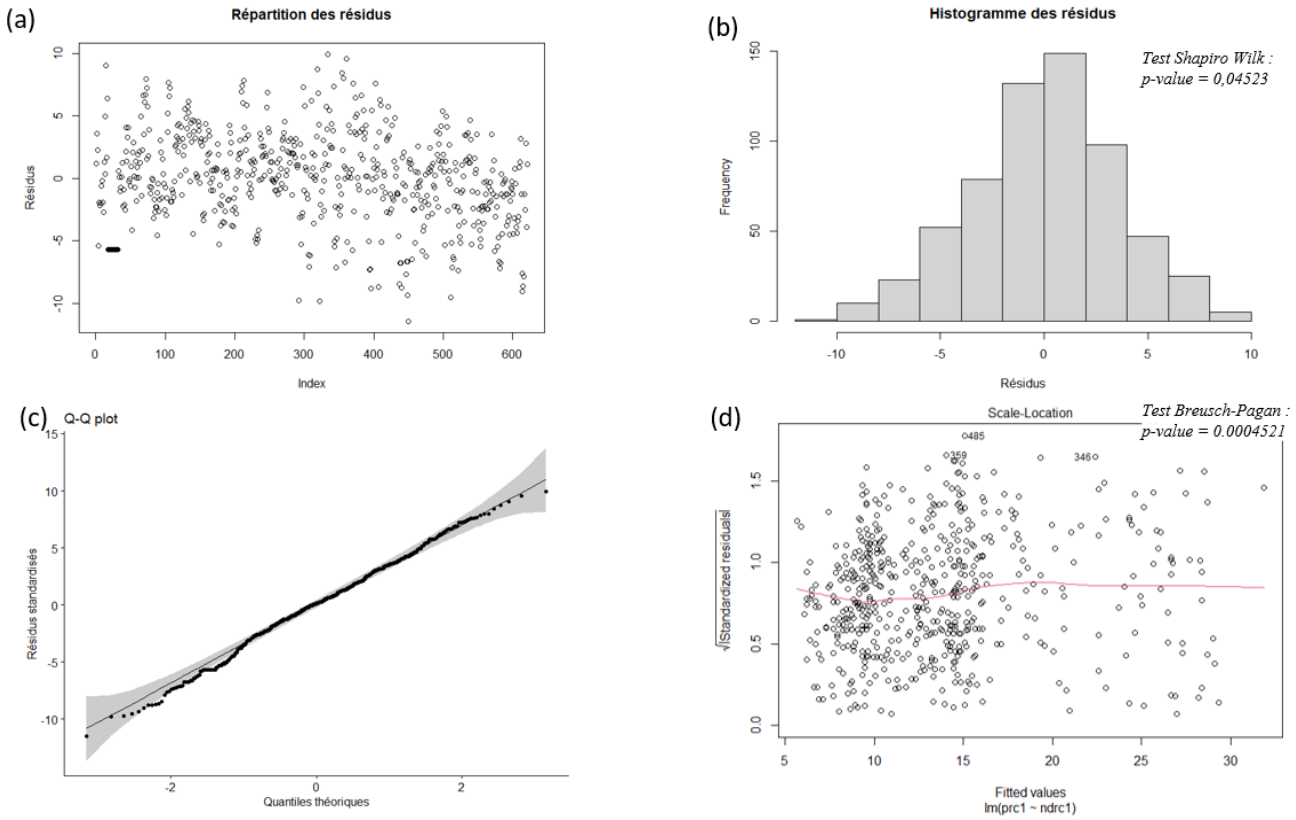
Figure 14 : Modèle linéaire, poids du régime en fonction du nombre de doigts du régime accompagné de sa bande de confiance à 95%, des bornes inférieure et supérieure des valeurs prédites.



ε_i : erreurs résiduels aléatoire
*ndrc1 : p-value <2e-16 ****

Pour valider le modèle *(Fig. 14)*, il faut analyser les résidus du modèle et valider trois hypothèses : les résidus sont indépendants, les résidus suivent une loi normale et la variance des résidus est homogène.

Figure 15 : Analyse des résidus pour la validation du modèle.



En ce qui concerne l'indépendance des résidus, sur le Figure 15.a), il semble y avoir une tendance dans les résidus mais elle reste correcte alors l'hypothèse d'indépendance peut être partiellement validée.

En ce qui concerne la normalité des résidus, l'histogramme des résidus *(Fig. 15.b)* se comporte comme une loi normale. Or, l'hypothèse de normalité est quand même rejetée par le test de Shapiro-Wilk avec une p-valeur de 0,045 (α = 5%). La p-valeur est proche du seuil de significativité de 0,05, de plus le test de Shapiro-Wilk a tendance à être sensible aux écarts de normalité surtout sur un grand échantillon, dans ce cas-là il est préférable de regarder le Q-Q plot *(Fig. 15.c)*. Il est correct, l'hypothèse de normalité des résidus est acceptée.
Ensuite, il faut vérifier que la variance des résidus est constante. Sur le Figure 15.d), la courbe de régression locale en rouge est globalement plate, cela veut dire que les résidus ont tendance à être répartis de façon homogène par rapport aux valeurs théoriques. Le test de Breusch-Pagan rejette cependant l'hypothèse d'homoscédasticité avec une p-valeur de 0,0045. Certains résidus se superposent, ce qui confirme un défaut d'indépendance.

Les hypothèses du modèle peuvent être partiellement validée. Le modèle a un R^2 de 0,68. Ce qui veut dire que 68% des observations sont expliquées par ce modèle. Le nombre de doigts du régime est significatif dans le modèle *(Fig. 14)*. Cependant, 32% de la variabilité reste inexpliquée par ce modèle. Il convient donc de rechercher des variables explicatives complémentaires. Le fait que le nombre de doigts du régime soit étroitement lié au poids du régime peut induire des erreurs *(Motulsky, 2019)*.

b. Modèle linéaire multiple

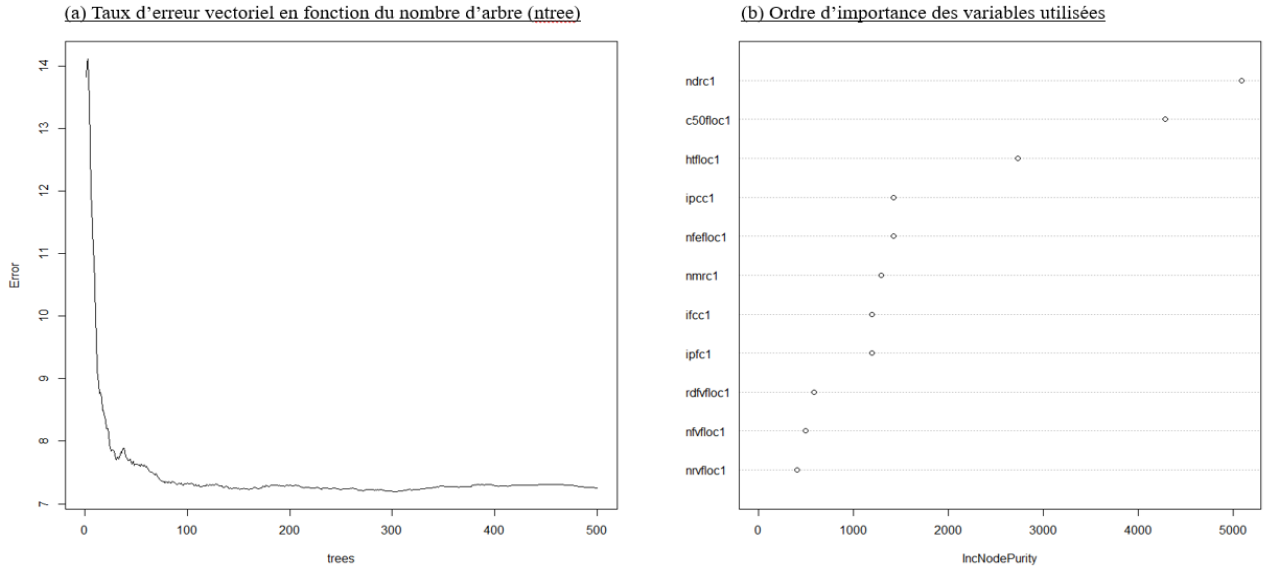
Il faut commencer par le modèle le plus complet (11 variables) avec leurs interactions et il faut supprimer les variables les moins significatives une par une. Cette méthode de sélection est appelée backward *(Benoudjit, 2002)*. Le modèle final ne respecte aucune des conditions d'application. L'hypothèse de normalité des résidus aurait pu être respectée mais pas l'hypothèse d'homoscédasticité, la variance des résidus augmentait. L'indépendance des résidus n'était pas respectée non plus. La prédiction du *prc1* avec un modèle linaire multiple n'a pas été poursuivie.

c. Forêt aléatoire *(Random Forest)*

L'arbre de régression a l'avantage de ne pas devoir valider des conditions initiales mais il a le désavantage de ne pas pouvoir être extrapolé hors du jeu de données d'apprentissage et ne supporte pas les données manquantes *(Breiman, 2001)*. Cet algorithme d'apprentissage automatique combine le résultat d'un ensemble d'arbre de décisions indépendants en un résultat final. Il peut résoudre des problèmes de régressions et de classifications. L'échantillon de données utilisé est divisé en sous-échantillon d'apprentissage et en sous-échantillon test (out-of-bag). La prédiction dans le cas d'une régression est faite par la moyenne d'arbres de décision individuels. Le sous-échantillon test subit une validation croisée dans le but de finaliser la prédiction.

Les dimensions du jeu de données pour Random Forest sont 588 individus et 12 variables. La variable *hbulbfloc1* a été retirée car elle contient 237 données manquantes *(Tab.5)*. Les paramètres mtry et mtree sont laissés par défaut (500 ; 3).

Figure 16 : Résultat de la régression avec Random Forest.



Ce modèle de régression permet d'obtenir un taux d'erreur sur la prédiction du poids du régime de 7,25% et le pourcentage de variabilité expliquée est de 79,36%. Ce qui est correct dans notre cas. Les variables classées par ordre d'importance *(Fig. 16)* sont le *ndrc1*, la *c50floc1* et la *htfloc1* ce qui est intéressant puisque le *ndrc1* est la variable la plus corrélée avec le *prc1* et qu'elle permet de classer les groupes Type de bananiers plantains. La *c50floc1*, la *htfloc1* et la *nefloc1* permettent d'avoir des informations sur la taille. Plus la taille est importante et plus les chances que le régime le soit aussi.

V. Evolution temporelle

Cette cinquième partie aspirait à suivre l'évolution du comportement des variables mesurées sur une même population de bananiers plantain sur 5 cycles de culture. Il existe au sein du jeu de données cinq variables indiquant la chute des bananiers plantain. Ces variables étaient appelées *« Codechute »* et les modalités étaient les suivantes ; 0 : pas de chute, 1 : chute avant la floraison, 2 : chute après la floraison et 3 : chute après la récolte. Un nouveau jeu de données est créé à partir du jeu de données initial : tous les plants de bananiers ayant chutés avant la floraison (code : 1) ont été supprimés, mais pas ceux qui ont chuté après la floraison (code : 2) et après la récolte (code : 3) puisque dans ces cas-là ils pouvaient être mesurés aux cycles suivants. Le rejet n'ayant pas forcément été impacté par la chute. Ce choix a été fait pour ne pas perdre trop d'informations et ne pas réduire à un tiers la population de bananiers plantain, les chutes étant fréquentes. Les jeux de données correspondant aux 5 cycles exempts des bananiers ayant chuté avant la floraison (code : 1), ont été compilés entre eux en fonction des variables qu'ils ont en commun. Le jeu de données final regroupe tous les bananiers plantain qui ont réussi à être suivis sur les 5 cycles.

La population totale est de 418 plants pour 117 variétés différentes. Le nombre de variétés et les effectifs par variété ont diminué.

Tableau 12 : Effectifs et proportions (%) de bananiers plantain selon les groupes Taille (de la variable *new_Taille*) dans le nouveau jeu de données.

Taille	**Petit**	**Moyen**	**Géant**	**Total :**
Effectif	9	**320**	85	**418**
≈ %	*2,2*	***76,6***	*20,3*	*100*

Tableau 13 : Effectifs et proportions (%) de bananiers plantain selon les groupes Type (Musalogue) dans le nouveau jeu de données.

Type Musalogue	Batard	Corne	Faux-Corne	French	Total :
Effectif	3	23	188	**204**	**418**
≈ %	*0,7*	*5,5*	*45,0*	***48,8***	*100*

Le nouveau jeu de données est majoritairement représenté par les bananiers plantain Moyen *(76,6%)*. Il y a 20,3% de Géant et 2,2% de Petit. Le groupe Type le plus présent est le French à 48,8%, ensuite le Faux-corne à 45% et le Corne à 5,5%. Le type Batard représente moins de 1% de l'échantillon *(Tab. 12 et 13)*.

La variable du poids du régime est prise comme exemple pour développer la démarche :

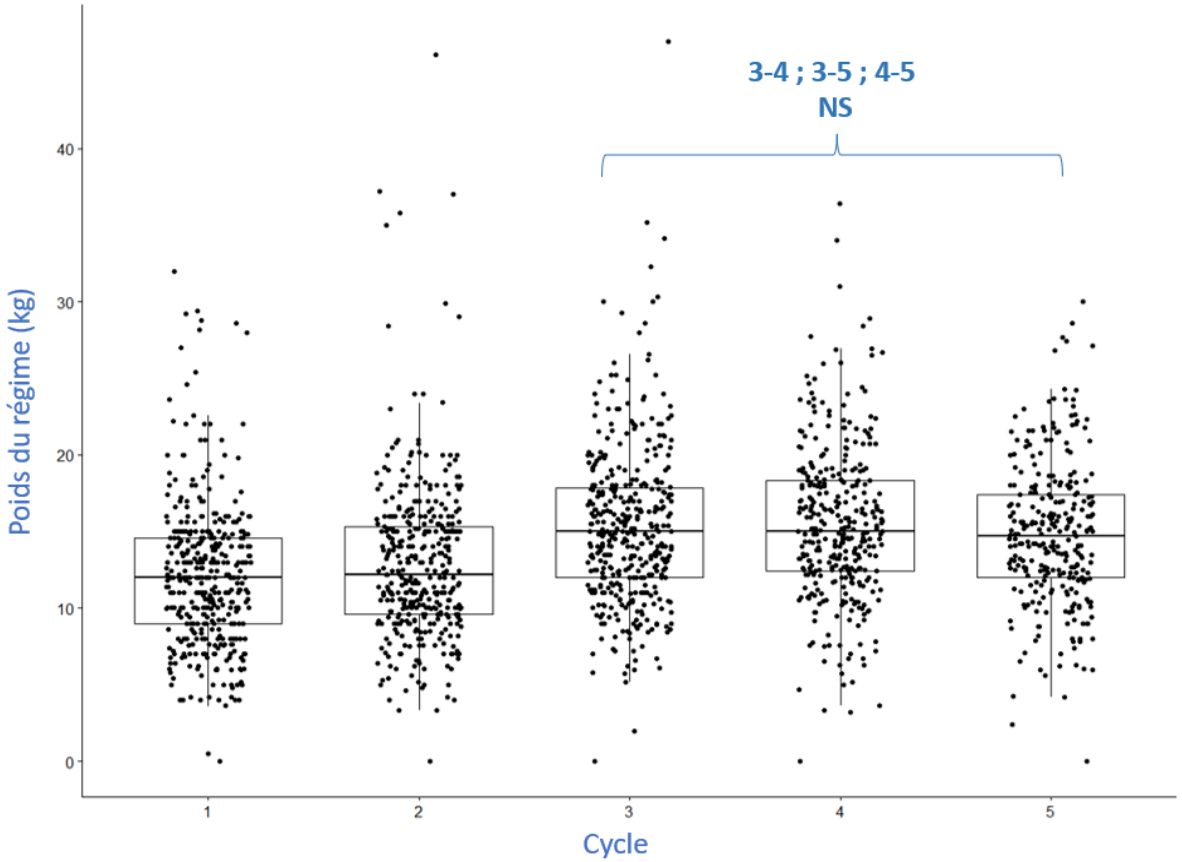
Tableau 14 : Statistiques descriptives du poids du régime pour chaque cycle.

Cycle	n	Min.	Max.	Médiane	IQR	Moyenne	Ecart-type	IC 95%
1	409	0	32	12	5,6	12,2	0,23	0,46
2	388	0	46,1	12,2	5,7	12,9	0,26	0,52
3	392	0	47	15	5,8	15,3	0,27	0,52
4	342	0	36,4	15	5,9	15,5	0,27	0,54
5	292	0	30	14,7	5,4	14,9	0,27	0,54

- *n : effectif*
- *Min. : Minimum*
- *Max. : Maximum*
- *IQR : écart inter-quartiles*
- *IC : intervalle de confiance à 95% de la moyenne*

La variable du poids du régime est particulière car les observations entre le cycle 1 et le 5 est divisée par 1,4 *(Tab. 14)*. Pour d'autres variables il y avait moins de données manquantes. Le jeu de données a toujours des données manquantes car, pour rappel, les bananiers plantain de ce nouveau jeu de données représentent ceux qui ont été suivis obligatoirement sur les cinq cycles consécutifs mais cela ne veut pas dire qu'ils n'ont pas chuté après la floraison ou la récolte (code : 2 et 3) générant des données manquantes au niveau des mesures prise à la récolte en particulier puisqu'un bananier qui chute avant d'avoir fini son régime ne pourra pas être mesuré au niveau du régime à la récolte mais bien au niveau de la hauteur du tronc à la floraison. Un bananier a plus de chance de chuté avant d'avoir fini son régime.

Figure 17 : Boxplots du poids du régime en fonction des cycles, avec représentation des résultats d'un test multiple de Wilcoxon pour données appariées.



NS : Non Significatif ; p-valeurs ajustées (Bonferroni) > 0,05 ; résultats test multiple de Wilcoxon (appariées)

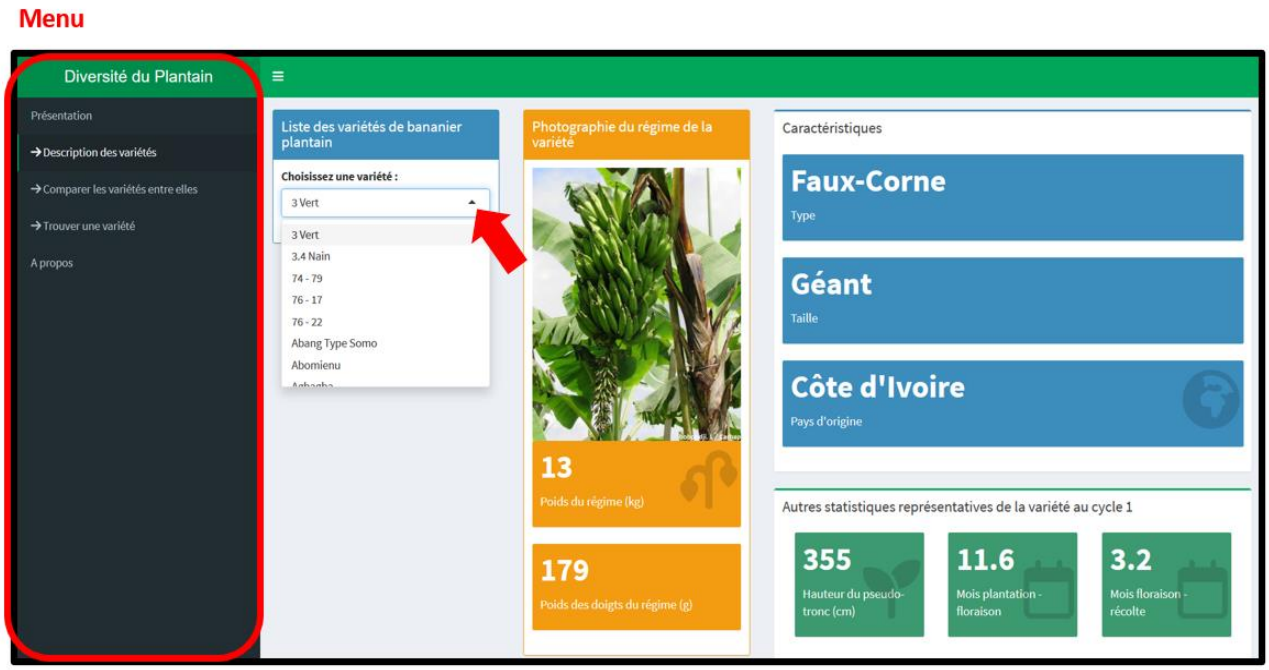
D'après le résultat du test de Kruskal-Wallis *(p value = 1.91e-35)*, il existe au moins un cycle différent. Le test de Kruskal-Wallis ne prend pas en compte l'appariement des données, or son homologue le test de Friedman ne gère pas les données manquantes il n'a pas pu être utilisé.
Le test multiple de Wilcoxon sur données apparié avec ajustement des p-valeurs avec la méthode de Bonferroni révèle que tous les cycles deux à deux sont significativement différents entre eux sauf pour ces couples de cycles : 3 et 4, 3 et 5, 4 et 5 *(Fig. 17)*. Pour les chercheurs ce résultat est cohérent car les trois derniers cycles correspondent au cycle de routine des bananiers. Le cycle 1 a des conditions initiales différentes.

Cette démarche a été faite sur plusieurs variables et quelques-unes ont été mises en annexe 6. Tous les cycles sont différents pour la hauteur du pseudo-tronc mais visuellement le cycle 1 se démarque. Pour la circonférence à 50 cm du pseudo-tronc ils sont tous différents sauf pour le couple 3 et 5. Pour le nombre de feuilles vivantes tous les couples sont différents mais visuellement les cycles 1 et 2 se démarquent *(Annexe 6.a)*. Pour les variables relatives au rendement *(Annexe 6.b)* elles ont visuellement la même tendance que le poids du régime, pour le nombre de doigts du régime les couple 2-4 et 4-5 ne sont pas significativement différents et pour le nombre de mains du régime se sont les couples 3-4 et 4-5. Les variables au cours des cycles ont globalement une grande variabilité et d'un point de vue global et agronomique les cycles sont plus ou moins stable entre eux sauf, comme attendu, le cycle 1 et le cycle 2 dans une moindre mesure.

4. Application web SHINY

Après avoir participé au nettoyage des données et avoir réalisé une analyse statistique descriptive et exploratoire, j'ai eu la possibilité au cours de mon stage, de participer à une formation sur le package Shiny de 2,5 jours. Cela m'a permis de créer une application de visualisation du jeu de données, en autonomie, par la suite.

Aperçu de l'onglet 2 de l'application (l'intégralité de l'interface de l'application a été mise en annexe 7) :



L'application se présente sous un format Dashboard composé de 5 onglets. Il y a une page de présentation *(Onglet 1, annexe 7)* qui permet d'informer sur le contexte du jeu de données et d'avoir quelques informations concernant l'étude. Un deuxième onglet *(Aperçu et Onglet 2 (a), annexe 7)* donne la possibilité de choisir une variété sur les 135 pour afficher sa fiche de présentation et le détail sur les variables correspondantes au pseudo-tronc, au régime, aux cycles et aux feuilles. Elles sont représentées par des boxplots et un tableau des statistiques descriptives détaillées *(Onglet 2 (b), annexe 7)*. Le troisième onglet permet de comparer les différentes caractéristiques des variétés sous forme de radar interactif *(Onglet 3, annexe 7)*. Le quatrième onglet permet de choisir des paramètres initiaux et une liste se met à jour en fonction des filtres précisés sur les paramètres *(Onglet 4, annexe 7)*. Le cinquième onglet *"A propos"* sert à afficher l'information liée à l'application *(Onglet 5, annexe 7)*.

Ce dashboard a été construit à but informatif et exploratoire avec un maximum de transparence sur le jeu de données. Il permet de profiter pleinement d'informations essentielles et détaillées qui s'y trouvent. Il peut être utilisé par les agronomes et les producteurs pour discuter de leurs choix de variétés.

5. Discussions :

Le travail mené apporte des résultats intéressants, dont il est néanmoins important de discuter la portée.

D'un point de vue expérimental, le fait que chacune des 135 variétés compte initialement 5 plants est considérable. Cette collection de variétés de bananiers plantain est la plus grande du monde, elle regroupe tous les groupes botaniques Taille et Type. Elle a été correctement conduite sur plusieurs années consécutives, contrôlant l'irrigation, les engrais, la lutte contre les bioagresseurs et champignons réduisant les biais associés. Un grand nombre de mesures agronomiques différentes ont été prises. Le jeu de données constitué est ainsi assez unique et d'une grande valeur. Cependant, ce faible nombre de réplicats par variétés ainsi que la structure spatiale de l'expérimentation amènent certaines limites qui seront discutées plus bas.

Le travail effectué lors des analyses a permis, dans un premier temps, de décrire les 135 variétés individuellement et de créer un outil de visualisation des données des variétés avec Shiny. Pour cela, des analyses descriptives ont été faites pour 135 variétés, ce qui représente une base des connaissances précieuses. Le jeu de données a permis de construire l'application web Shiny, qui a permis de comparer de manière synthétique et visuelle les variétés entre elles. C'est l'objectif principal de conduite d'une telle collection de bananiers plantain. Il a été constaté une grande hétérogénéité entre variétés et, dans une moindre mesure, au sein même d'une variété. Néanmoins, le contexte de l'étude et le jeu de données sont complexes et représentent des limites pour faire de l'analyse de données inférentielle. Les effectifs par variété sont faibles ($n \leq 5$) pour une analyse statistique et pour pouvoir comparer les variétés. Il aurait fallu pour cela des effectifs supérieurs à 30, ce qui n'était bien évidemment pas envisageable d'un point de vue expérimental sur une gamme aussi grande variétés. Il aurait été également souhaitable que les biais existants soient contrôlés ou minimisés en utilisant une configuration en bloc randomisé par exemple. En effet, les bananiers plantain ont été plantés de façon ordonnée par variété et par groupe botanique, ce qui empêche de minimiser les biais spatiaux. Notamment, les biais liés à la différence d'exposition au vent pour les chutes et des bioagresseurs, fréquent en bananeraie. Enfin, les bananiers se comportent différemment selon les conditions pédoclimatiques ce qui représente une limite à la généralisation des résultats.

Les analyses auront permis de mieux comprendre les groupes botaniques, notamment avec la description des variables les sous-tendant. Le nombre de feuille émise discrimine bien les groupes Taille *(Fig. 1 et 2)* et un arbre de classification a évalué le taux de bien classés à 92% *(Fig. 3)*. Le nombre de doigt du régime est une variable qui discrimine bien les groupes Type *(Fig. 4 et 5)* et un arbre de classification a évalué le taux de bien classés à 89% *(Fig. 6)*. Les deux variables ensemble permettent en moyenne de bien classer 81,2% des groupes Type et Taille en même temps *(Fig.7)*. Les corrélations *(Fig. 9 et 10)* permettent de voir qu'il y a des tendances différentes assez marquées entre groupes botaniques. L'ACP *(Fig. 12)* a permis de mieux comprendre la structure du jeu de données dans une dimension réduite. L'HCPC *(Fig. 13)* a isolé ses observations en 4 clusters. Ces nouveaux clusters sont différents entre eux par rapport aux variables qui ont été testées sur l'ACP et surtout avec celles qui étaient le mieux représentées par l'ACP *(Tab. 11)*. Les résultats laissent entrevoir des comportements différents entre groupes, mais difficiles à expliciter avec le jeu de données. Deux groupes (Petit et Batard) étaient mal représentés pour pouvoir les comparer aux autres, il aurait été possible de les supprimer mais le choix de garder l'information a été fait. Ce résultat va inciter les chercheurs à travailler sur d'autres données peut-être plus discriminantes pour mieux comprendre ces différences. Une tentative de construire de nouveaux groupes a montré une cohérence globale des groupes botaniques. Le travail mené sur la compréhension des groupes botaniques a été effectué car les chercheurs émettaient un doute sur leur intérêt pour traduire des différences de comportement agronomique. Ils ont pour objectif à l'avenir de créer des groupes sur des critères traduisant le

potentiel agronomique des bananiers est non sur une base morphologique, et ainsi plus proche du fonctionnement de la plante.

La recherche d'une prédiction du poids du régime a été abordée. Il peut être prédit avec le nombre de doigts du régime via un modèle linéaire simple qui explique 68% des observations *(Fig. 14 et 15)*. Le modèle n'est pas pertinent puisque la variable prédite (prc1 : poids du régime) et la variable explicative (ndrc1 : nombre de doigts du régime) sont intimement liées. Les hypothèses du modèle peuvent être validées mais avec prudence, le modèle n'est donc pas idéal mais il permet d'expliquer 68% de la variabilité du poids du régime avec le nombre de doigts. Il y a une tendance générale selon laquelle plus un régime a un grand nombre de doigts et moins le poids d'un doigt sera lourd. Les French ont tendance à avoir beaucoup de doigts mais avec un poids de doigt moins important contrairement aux Corne qui ont tendance à avoir peu de doigts mais plus lourd. Le modèle de régression multiple final n'a pas été poursuivi. Les résidus du modèle ne respectaient aucunes des hypothèses, il aurait été possible d'essayer de les contourner mais j'ai fait le choix de ne pas forcer le modèle au vu du contexte général et du fait que toutes les hypothèses devaient être corrigées. La régression avec les forêts aléatoires permet d'avoir un modèle plus complet et plus robuste avec un pourcentage de variabilité expliquée de 79,36% *(Fig. 16).* Cela a finalement permis d'avoir un modèle de régression qui prédit assez bien, avec 7,25% d'erreurs, et qui est propre à notre jeu de données. Les paramètres *mtry* et *mtree* auraient pu être optimisés. Quelques variations manuelles ont été testées mais le taux d'erreurs fluctuait peu. Cela aurait permis de rendre le modèle un peu plus précis mais dans notre cas ce n'est pas essentiel.

L'analyse exploratoire sur l'évolution des variables à travers les cycles a permis de mettre en évidence que le cycle 1 est différent au travers des variables analysées surtout pour les variables relatives à la croissance *(Annexe 6.a)*. Or, d'un point de vue agronomique il y a une certaine stabilité, pour la plupart des variables surtout celles relatives au régime, dans les 3 derniers cycles *(Fig. 17 et annexe 6.b)*. Il aurait été possible d'enlever tous les bananiers ayant chutés même après la floraison et la récolte et ainsi supprimer entièrement les données manquantes pour ne garder que les bananiers n'ayant jamais subi de chute. La perte d'information aurait été grande puisque cela concerne deux tiers de la population initiale. De plus, si un bananier chute après la récolte le poids de son régime pourra être mesuré quand même. Une chute tardive permet aussi de récupérer un rejet qui prendra la suite des cycles. Cette partie n'était qu'exploratoire. Elle a permis de voir l'évolution des variables au cours des cycles et de prendre du recul sur le déroulement de l'étude.

Enfin, J'ai suivi une formation accélérée sur le package Shiny de RStudio® et j'ai mis en pratique ces nouvelles connaissances grâce à l'élaboration de l'application web sur la diversité des plantains. La façon de réfléchir pour construire une application est différente d'un codage statistique classique. Notamment, par le fait que l'application est construite en deux sous-ensembles : l'interface utilisateur (ui) et le server. L'ui est utilisé pour construire l'interface de l'application et le server est utilisé pour le calcul des commandes. Cette nouvelle compétence me permettra à l'avenir de pouvoir présenter des résultats d'une analyse statistique et de valoriser un jeu de données via une interface ludique et épuré. Cela permet aussi de vulgariser l'information. Il y a deux formats possibles le Dashbord qui est une interface type tableau de bord et une page web où nous avons la liberté d'agencer la fenêtre à notre guise. Le Dashbord créé pour l'étude a permis de visualiser les données à l'échelle variétale et de les rendre accessibles avec une interface pratique et dynamique. L'application répond aux attentes des chercheurs, agronomes et producteurs, elle leur servira à comparer les variétés avec la plus grande transparence.

Pour aller plus loin, il aurait été possible d'essayer de faire une analyse spatiale des bananiers plantain sur la parcelle de culture du CARBAP au travers une modélisation, même si cela représente des limites étant donné les conditions de l'étude. Pour la recherche de variables explicatives du poids du régime, il aurait été possible d'utiliser une méthode qui permet de tester plusieurs hypothèses sur

sa prédiction et de faire le lien entre les variables choisies et leurs implications possibles. Cette méthode consiste en l'élaboration d'un modèle probabiliste d'équations structurelles (PSEM : Probabilistic Structural Equation Model). Pour finir, une partie des variables n'ont pas été utilisées donc il pourrait y avoir du potentiel avec les autres variables inexploitées présentes dans le jeu de données brut.

6. Conclusion de l'étude sur la collection de bananiers plantain et du stage

Les observations et les analyses ont permis de mieux comprendre le fonctionnement complexe et l'expression variée des bananiers plantain au Cameroun, la plus grande collection de bananiers plantain du monde. Les différentes variétés ont été décrites, les groupes botaniques ont été analysés, l'évolution des principales variables au cours des cycles a été étudiée et le poids du régime a été prédit. Le jeu de données et le script de l'analyse seront partagés sur un espace numérique. Le travail effectué au cours de ce stage a répondu à la demande de l'équipe de chercheurs tant dans la mise en forme du jeu de données, précieux pour eux, que dans les analyses, mais également dans un souci de partage et de valorisation des données. Le jeu de données n'a cependant pas permis de faire de statistique inférentielle au niveau variétale.

J'ai aimé le partage des connaissances qu'il y a eu dans l'équipe. D'un côté, j'ai apporté mon savoir et mon point de vue de biostatisticienne sur le jeu de données et de l'autre, j'ai reçu les connaissances et l'expertise des chercheurs dans leurs domaines. Cela a renforcé l'intérêt que je porte aux statistiques pour être au service des sciences et en collaboration avec leurs acteurs.

Ce stage m'aura permis d'effectuer tous les aspects de l'analyse de données. Du nettoyage jusqu'à la réalisation d'un outil de visualisation en passant par les différentes problématiques rencontrées en analyse de données.

Bibliographie :

I. Livre

Ibobondji, L. & Ngoh Newilah, G. & Tomekpe, K. & Ruas, M. & Chase, R. & Roux, N. (2018) *Musalogue : Catalogue de germoplasme de Musa. Plantains d'Afrique occidentale et centrale – Collection CARBAP*. Picq, C., Bioversity International, Montpellier (France), 297 p.

Kwa, M. & Temple, L. (2019). *Le bananier plantain : enjeux socio-économiques et techniques*. Quæ, 184 p.

Lassoudière, A. (2010). *L'histoire du bananier*. Quæ, 351 p.

Little, R.J.A. & Rubin, B. (2002). *Statistical Analysis with Missing Data*. John Wiley & Sons, 408 p.

Motulsky, H. (2019). *Biostatistique, une approche intuitive.* De Boeck Sup, 3eme ed, 608 p.

Saporta, G. (1990*). Probabilités, analyse de données et statistique*. Technip, Paris. 174 p.

II. Article

Anyasi, T.A. & Afam, I.O. & Jideani & Godwin, R.A. & Mchau. (2018). Phenolics and essential mineral profile of organic acid pretreated unripe banana flour.
Food Research International,104, 0963-9969.

Benoudjit, N. & Archambeau, C. & Lendasse, A. & Lee, J. & Verleysen, M. (2002). Width optimization of the Gaussian kernels in Radial Basis Function Networks.
ESANN, p. 425-432.

Bertsimas, D. & Dunn, J. (2017). Optimal classification trees.
Machine Learning, 106(7), 1039–1082.

Breiman, L. (2001). Random forests.
Machine Learning, 45(1):5–32.

Chen, P. & Popovich, P. (2002). Correlation : Parametric and Nonparametric Measures.
Sage University Papers Series on Quantitative Applications in the Social Sciences, no. 07-139, p 82.

Depigny, S. & Damour, G. (2022). Trait-based description of the agronomic and usage potential of a range of plantain varieties from Cameroon.
Experimental Agriculture, 58 : 17 p.

Ekesa, B. & Nabuuma, D. & Blomme, G. & Van den Bergh, I. (2015). Provitamin A carotenoid content of unripe and ripe banana cultivars for potential adoption in eastern Africa.
Journal of Food Composition and Analysis, 43, 0889-1575.

Graham, J. (2009). Missing data analysis: making it work in the real world.
Annual Review of Psychology, 60:549–576.

Honaker, J. & King, G. & Blackwell, M. (2011). Amelia II : A Program for Missing Data.
Journal of statistical software, 45 no 7.

Imbert, A. & Vialaneix. N. (2020). Décrire, prendre en compte, imputer et évaluer les valeurs manquantes dans les études statistiques : une revue des approches existantes.
Journal de la Société Française de Statistique, Vol. 159, No. 2, 1-55.

Kowarik, A. & Templ, M. (2016). Imputation with the R package VIM.
Journal of Statistical Software, 74(7):1–16.

McHugh, M. (2012). Interrater Reliability: The Kappa Statistic.
Biochemia Medica : Časopis Hrvatskoga Društva Medicinskih Biokemičara, 276–82.
Misztal, M. (2013). Some remarks on the data imputation “MissForest” Method.
Acta Universitatis Lodziensis, Folia Oeconomica, p : 169 – 179.
Sanjar, K. & Bekhzod, O. & Kim, J. & Paul, A. & Kim, J. (2020). Missing Data Imputation for Geolocation-based Price Prediction Using KNN–MCF Method" ISPRS
International Journal of Geo-Information 9, no. 4: 227.

Stekhoven, D. & Bühlmann, P. (2012). Missforest-non-parametric missing value imputation for mixed-type data.
Bioinformatics, 28(1):112–118.

Tezenas Du Montcel, H. & De Langhe, E. & Swennen, R. (1983). Essai de classification des bananiers plantains (AAB).
Fruits, vol. 38, n° 6, p. 461-474.

Udomkun, P. & Masso, C. & Swennen, R. & Innawong, B. & Kuate, A.F. & Alakonya, A. & Lienou, J. & Ibitoye, D.O. & Vanlauwe, B. (2021). Consumer Preferences and Socioeconomic Factors Decided on Plantain and Plantain-Based Products in the Central Region of Cameroon and Oyo State, Nigeria.
*Foods,*10, no. 8: 1955.

Welch, B. (1951) On the Comparison of Several Mean Values: An Alternative Approach.
Biometrika, 38, 330-336.

III. Littérature grise ou rapports internes

Husson, F. & Josse, J. & Pagès J. (2010). *Principal component methods - hierarchical clustering - partitional clustering: why would we need to choose for visualizing data?* Technical Report - Agrocampus Applied Mathematics Department, 17 p.
Unité Pédagogique de Mathématiques Appliquées 65 rue de Saint-Brieuc CS 84215 35 042 Rennes Cedex, France.

Table des matières ANNEXES

Annexe 1 : Lexique des variables du jeu de données BRUT……....p 1-5

Annexe 2 : Liste des plants de bananiers plantains avec absence de croissance (2) et croissance anormale (1) après plantation…….……p 6

Annexe 3 : Liste des 135 variétés représentées par la collection CARBAP (A : Z)……p 7

Annexe 4 : Capture d'écran d'une partie du data table généré avec R Studio…….p 8

Annexe 5 : Statistiques descriptives par groupes botanique Taille (a) et Type (b)…….p 9-10

Annexe 6 : Quelques variables en plus pour l'analyse sur 5 cycles de culture……..p 11

Annexe 7 : Application SHINY……p 12-15

Annexe 1 : Lexique des variables du jeu de données BRUT (5 pages)

Variable	**Description**	**Unité**	**Format**
Variete	Nom de la variété		Texte
Couleur	Couleur du pseudotronc		Texte
Typemusalogue	Type de plantain d'après le Musalogue		Texte
Taillemusalogue	Catégorie de taille du plantain d'après le Musalogue		Texte
Nanisme	Nanisme ou non		Texte
pays	Pays d'origine de la variété		Texte
longdoigts	Longueur du doigt médian	cm	Numérique
grade	Diamètre du doigt médian	cm	Numérique
ms	Taux de matière sèche du doigt médian		Numérique
fermete	Fermeté de la pulpe du doigt médian		Numérique
extraitsol	Extrait sec soluble de la pulpe à maturité		Numérique
n_plant	Numéro du plant, il y a 5 pieds par variété		Facteur
Dp	Date de plantation		Date
Cycle 1 / c1			
Dpfc1	Date de la première feuille émise au champ		Date
Df1c1	Date d'apparition de la première feuille ≥ 10 cm de large		Date
dmap1c1	Date 1 mois après plantation		Date
htmap1c1	Hauteur 1 mois après plantation	cm	Numérique
c50map1c1	Circonférence à 50 cm 1 mois après plantation	cm	Numérique
nfemap1c1	Nombre de feuilles émises 1 mois après plantation	/	Numérique
dmap3c1	Date 3 mois après plantation		Date
htmap3c1	Hauteur 3 mois après plantation	cm	Numérique
c50map3c1	Circonférence à 50 cm 3 mois après plantation	cm	Numérique
nfemap3c1	Nombre de feuilles émises 3 mois après plantation	/	Numérique
nfvmap3c1	Nombre de feuilles vivantes 3 mois après plantation	/	Numérique
rdfvmap3c1	Rang de la dernière feuille vivante 3 mois après plantation	/	Numérique
nrvmap3c1	Nombre de rejets vivants 3 mois après plantation	/	Numérique
dmap6c1	Date 6 mois après plantation		Date
htmap6c1	Hauteur 6 mois après plantation	cm	Numérique
c50map6c1	Circonférence à 50 cm 6 mois après plantation	cm	Numérique
nfemap6c1	Nombre de feuilles émises 6 mois après plantation	/	Numérique
nfvmap6c1	Nombre de feuilles vivantes 6 mois après plantation	/	Numérique
rdfvmap6c1	Rang de la dernière feuille vivante 6 mois après plantation	/	Numérique
nrvmap6c1	Nombre de rejets vivants 6 mois après plantation	/	Numérique
dmap9c1	Date 9 mois après plantation		Date
htmap9c1	Hauteur 9 mois après plantation	cm	Numérique
C50map9c1	Circonférence à 50 cm 9 mois après plantation	cm	Numérique
nfemap9c1	Nombre de feuilles émises 9 mois après plantation	/	Numérique
nfvmap9c1	Nombre de feuilles vivantes 9 mois après plantation	/	Numérique
rdfvmap9c1	Rang de la dernière feuille vivante 9 mois après plantation	/	Numérique
nrvmap9c1	Nombre de rejets vivants 9 mois après plantation	/	Numérique
sf1c1	Surface foliaire de la feuille 1	cm^2	Numérique
sf5c1	Surface foliaire de la feuille 5	cm^2	Numérique
sf10c1	Surface foliaire de la feuille 10	cm^2	Numérique
sf20c1	Surface foliaire de la feuille 20	cm^2	Numérique
sf30c1	Surface foliaire de la feuille 30	cm^2	Numérique

sf40c1	Surface foliaire de la feuille 40	cm^2	Numérique
sf(3)c1	Surface foliaire de l'antépénultième feuille avant la fleur	cm^2	Numérique
dfloc1	Date de la floraison		Date
htfloc1	Hauteur à la floraison	cm	Numérique
c50floc1	Circonférence à 50 cm à la floraison	cm	Numérique
hbulbfloc1	Hauteur du bulbe à la floraison	cm	Numérique
nfefloc1	Nombre de feuilles émises à la floraison	/	Numérique
nfvfloc1	Nombre de feuilles vivantes à la floraison	/	Numérique
rdfvfloc1	Rang de la dernière feuille vivante	/	Numérique
nrvfloc1	Nombre de rejets vivants à la floraison	/	Numérique
ipfc1	Intervalle plantation- floraison	jours	Numérique
drecc1	Date de récolte		Date
nfvrecc1	Nombre de feuilles vivantes à la récolte	/	Numérique
prc1	Poids du régime	kg	Numérique
nmrc1	Nombre de mains du régime	/	Numérique
ndrc1	Nombre de doigts du régime	/	Numérique
ifcc1	Intervalle de temps entre la floraison et la récolte au cycle 1	jours	Numérique
ipcc1	Intervalle de temps entre la plantation et la récolte au cycle 1	jours	Numérique
Codechutec1	Bananier n'ayant pas chuté = 0 ; chuté avant la floraison = 1 ; chuté entre la floraison et la récolte = 2 ; chuté à maturité = 3	/	Facteur
T_Joursipcc1	∑ températures (T) moyennes (0,4 * T_{max} + 0,6 * T_{min}) entre la plantation et la récolte	°C	Numérique
T_Joursipfc1	∑ températures (T) moyennes (0,4 * T_{max} + 0,6 * T_{min}) entre la plantation et la floraison	°C	Numérique
T_Joursifcc1	∑ températures (T) moyennes (0,4 * T_{max} + 0,6 * T_{min}) entre la floraison et la récolte	°C	Numérique
Cycle 2 / c2			
dfloc2	Date de floraison		Date
htfloc2	Hauteur du pseudo-tronc à la floraison	cm	Numérique
c50floc2	Circonférence à 50 cm à la floraison	cm	Numérique
hbulbfloc2	Hauteur du bulbe à la floraison	cm	Numérique
nfefloc2	Nombre de feuilles émises à la floraison	/	Numérique
nfvfloc2	Nombre de feuilles vivantes à la floraison	/	Numérique
drecc2	Date de récolte		Date
nfvrecc2	Nombre de feuilles vivantes à la récolte	/	Numérique
prc2	Poids du régime	kg	Numérique
nmrc2	Nombre de mains du régime	/	Numérique
ndrc2	Nombre de doigts du régime	/	Numérique
dobssucc2	Date observation du rejet successeur		Date
htsucc2	Hauteur du pseudotronc du rejet successeur	cm	Numérique
c50succ2	Circonférence à 50 cm du pseudotronc du rejet successeur	cm	Numérique
nfesucc2	Nombre de feuilles émises du rejet successeur	/	Numérique
nfvsucc2	Nombre de feuilles vivantes du rejet successeur	/	Numérique
ifcc2	Intervalle de temps entre la floraison et la récolte au cycle 2	jours	Numérique
T_Joursifcc2	∑ températures (T) moyennes (0,4 * T_{max} + 0,6 * T_{min}) entre la floraison et la récolte	°C	Numérique
Codechutec2	Bananier n'ayant pas chuté = 0 ; chuté avant la floraison = 1 ; chuté entre la floraison et la récolte = 2 ; chuté à maturité = 3		Facteur
Cycle 3 / c3			

DSc3	Date de sélection du successeur		Date
dF1c3	Date de la première feuille au cycle 3		Date
htF1c3	Hauteur du pseudotronc à la première feuille	cm	Numérique
c50F1c3	Circonférence à 50cm du pseudotronc à la première feuille	cm	Numérique
dobsifc3	Date d'observation de l'initiation florale		Date
htifc3	Hauteur de pseudotronc à l'initiation florale	cm	Numérique
c50ifc3	Circonférence à 50cm du pseudotronc à l'initiation florale	cm	Numérique
nfeifc3	Nombre de feuilles émises à l'initiation florale	/	Numérique
nfvifc3	Nombre de feuilles vivantes à l'initiation florale	/	Numérique
rdfvifc3	Rang de la dernière feuilles vivantes à l'initiation florale	/	Numérique
sf1c3	Surface foliaire de la feuille 1	cm^2	Numérique
sf(3)c3	Surface foliaire de la dernière feuille - 3	cm^2	Numérique
sf(1)c3	Surface foliaire de la dernière feuille - 1	cm^2	Numérique
sfetc3	Surface foliaire émise total	cm^2	Numérique
dfloc3	Date de floraison		Date
htfloc3	Hauteur du pseudo-tronc à la floraison	cm	Numérique
c50floc3	Circonférence à 50 cm à la floraison	cm	Numérique
htbulbfloc3	Hauteur du bulbe à la floraison	cm	Numérique
nfefloc3	Nombre de feuilles émises à la floraison	/	Numérique
nfvfloc3	Nombre de feuilles vivantes à la floraison	/	Numérique
drecc3	Date de récolte		Date
nfvrecc3	Nombre de feuilles vivantes à la récolte	/	Numérique
prc3	Poids du régime	kg	Numérique
nmrc3	Nombre de mains du régime	/	Numérique
ndrc3	Nombre de doigts du régime	/	Numérique
dobssucc3	Date observation du rejet successeur		Date
htsucc3	Hauteur du pseudotronc du rejet successeur	cm	Numérique
c50succ3	Circonférence à 50 cm du pseudotronc du rejet successeur	cm	Numérique
nfesucc3	Nombre de feuilles émises du rejet successeur	/	Numérique
nfvsucc3	Nombre de feuilles vivantes du rejet successeur	/	Numérique
ipfc3	Intervalle de temps entre la plantation et la floraison au cycle 3	jours	Numérique
ifcc3	Intervalle de temps entre la floraison et la récolte au cycle 3	jours	Numérique
ipcc3	Intervalle de temps entre la plantation et la récolte au cycle 3	jours	Numérique
Codechutec3	Bananier n'ayant pas chuté = 0 ; chuté avant la floraison = 1 ; chuté entre la floraison et la récolte = 2 ; chuté à maturité = 3		Facteur
T_Joursipfc3	∑ températures (T) moyennes (0,4 * T_{max} + 0,6 * T_{min}) entre la plantation et la floraison	°C	Numérique
T_Joursifcc3	∑ températures (T) moyennes (0,4 * T_{max} + 0,6 * T_{min}) entre la floraison et la récolte	°C	Numérique
T_Joursipcc3	∑ températures (T) moyennes (0,4 * T_{max} + 0,6 * T_{min}) entre la plantation et la récolte	°C	Numérique
Cycle 4 / c4			
DSc4	Date de sélection du successeur		Date
dF1c4	Date de la première feuille au cycle 3		Date
htF1c4	Hauteur du pseudotronc à la première feuille	cm	Numérique
c50F1c4	Circonférence à 50cm du pseudotronc à la première feuille	cm	Numérique
dobsifc4	Date d'observation de l'initiation florale		Date
htifc4	Hauteur de pseudotronc à l'initiation florale	cm	Numérique
c50ifc4	Circonférence à 50cm du pseudotronc à l'initiation florale	cm	Numérique

nfeifc4	Nombre de feuilles émises à l'initiation florale	/	Numérique
nfvifc4	Nombre de feuilles vivantes à l'initiation florale	/	Numérique
rdfvifc4	Rang de la dernière feuilles vivantes à l'initiation florale	/	Numérique
sf1c4	Surface foliaire de la feuille 1	cm^2	Numérique
sf(3)c4	Surface foliaire de la dernière feuille - 3	cm^2	Numérique
sf(1)c4	Surface foliaire de la dernière feuille - 1	cm^2	Numérique
sfetc4	Surface foliaire émise total	cm^2	Numérique
dfloc4	Date de floraison		Date
htfloc4	Hauteur du pseudo-tronc à la floraison	cm	Numérique
c50floc4	Circonférence à 50 cm à la floraison	cm	Numérique
nfefloc4	Nombre de feuilles émises à la floraison	/	Numérique
nfvfloc4	Nombre de feuilles vivantes à la floraison	/	Numérique
drecc4	Date de récolte		Date
nfvrecc4	Nombre de feuilles vivantes à la récolte	/	Numérique
prc4	Poids du régime	kg	Numérique
nmrc4	Nombre de mains du régime	/	Numérique
ndrc4	Nombre de doigts du régime	/	Numérique
dobssucc4	Date observation du rejet successeur		Date
htsucc4	Hauteur du pseudotronc du rejet successeur	cm	Numérique
c50succ4	Circonférence à 50 cm du pseudotronc du rejet successeur	cm	Numérique
nfesucc4	Nombre de feuilles émises du rejet successeur	/	Numérique
nfvsucc4	Nombre de feuilles vivantes du rejet successeur	/	Numérique
ipfc4	Intervalle de temps entre la plantation et la floraison au cycle 4	jours	Numérique
ifcc4	Intervalle de temps entre la floraison et la récolte au cycle 4	jours	Numérique
ipcc4	Intervalle de temps entre la plantation et la récolte au cycle 4	jours	Numérique
Codechutec4	Bananier n'ayant pas chuté = 0 ; chuté avant la floraison = 1 ; chuté entre la floraison et la récolte = 2 ; chuté à maturité = 3		Facteur
T_Joursipfc4	∑ températures moyennes (T) moyennes (0,4 * T_{max} + 0,6 * T_{min}) entre la plantation et la floraison	°C	Numérique
T_Joursifcc4	∑ températures (T) moyennes (0,4 * T_{max} + 0,6 * T_{min}) entre la floraison et la récolte	°C	Numérique
T_Joursipcc4	∑ températures (T) moyennes (0,4 * T_{max} + 0,6 * T_{min}) entre la plantation et la récolte	°C	Numérique
Cycle 5 / c5			
DSc5	Date de sélection du successeur		Date
dF1c5	Date de la première feuille au cycle 3		Date
htF1c5	Hauteur du pseudotronc à la première feuille	cm	Numérique
c50F1c5	Circonférence à 50cm du pseudotronc à la première feuille	cm	Numérique
dobsifc5	Date d'observation de l'initiation florale		Date
htifc5	Hauteur de pseudotronc à l'initiation florale	cm	Numérique
c50ifc5	Circonférence à 50cm du pseudotronc à l'initiation florale	cm	Numérique
nfeifc5	Nombre de feuilles émises à l'initiation florale	/	Numérique
nfvifc5	Nombre de feuilles vivantes à l'initiation florale	/	Numérique
rdfvifc5	Rang de la dernière feuilles vivantes à l'initiation florale	/	Numérique
sf1c5	Surface foliaire de la feuille 1	cm^2	Numérique
sf(3)c5	Surface foliaire de la dernière feuille - 3	cm^2	Numérique
sf(1)c5	Surface foliaire de la dernière feuille - 1	cm^2	Numérique
sfetc5	Surface foliaire émise total	cm^2	Numérique
dfloc5	Date de floraison		Date

htfloc5	Hauteur du pseudo-tronc à la floraison	cm	Numérique
c50floc5	Circonférence à 50 cm à la floraison	cm	Numérique
nfefloc5	Nombre de feuilles émises à la floraison	/	Numérique
nfvfloc5	Nombre de feuilles vivantes à la floraison	/	Numérique
drecc5	Date de récolte		Date
nfvrecc5	Nombre de feuilles vivantes à la récolte	/	Numérique
prc5	Poids du régime	kg	Numérique
nmrc5	Nombre de mains du régime	/	Numérique
ndrc5	Nombre de doigts du régime	/	Numérique
dobssucc5	Date observation du rejet successeur		Date
htsucc5	Hauteur du pseudotronc du rejet successeur	cm	Numérique
c50succ5	Circonférence à 50 cm du pseudotronc du rejet successeur	cm	Numérique
nfesucc5	Nombre de feuilles émises du rejet successeur	/	Numérique
nfvsucc5	Nombre de feuilles vivantes du rejet successeur	/	Numérique
ipfc5	Intervalle de temps entre la plantation et la floraison au cycle 5	jours	Numérique
ifcc5	Intervalle de temps entre la floraison et la récolte au cycle 5	jours	Numérique
ipcc5	Intervalle de temps entre la plantation et la récolte au cycle 5	jours	Numérique
Codechutec5	Bananier n'ayant pas chuté = 0 ; chuté avant la floraison = 1 ; chuté entre la floraison et la récolte = 2 ; chuté à maturité = 3		Facteur
T_Joursipfc5	Somme des températures (T) moyennes (0,4 * T_{max} + 0,6 * T_{min}) entre la plantation et la floraison	°C	Numérique
T_Joursifcc5	Somme des températures (T) moyennes (0,4 * T_{max} + 0,6 * T_{min}) entre la floraison et la récolte	°C	Numérique
T_Joursipcc5	Somme des températures (T) moyennes (0,4 * T_{max} + 0,6 * T_{min}) entre la plantation et la récolte	°C	Numérique
Variable en plus			
ipc1	Intervalle de temps entre la plantation et la récolte au cycle 1	jours	Numérique
icc2	Intervalle de temps entre la récolte du cycle 1 et la récolte du cycle 2	jours	Numérique
icc3	Intervalle de temps entre la récolte du cycle 2 et la récolte du cycle 3	jours	Numérique
icc4	Intervalle de temps entre la récolte du cycle 3 et la récolte du cycle 4	jours	Numérique
icc5	Intervalle de temps entre la récolte du cycle 4 et la récolte du cycle 5	jours	Numérique
if1f_1	Intervalle de temps entre la date d'observation de la première feuille et la floraison au cycle 1	jours	Numérique
if1f_3	Intervalle de temps entre la date d'observation de la première feuille et la floraison au cycle 3	jours	Numérique
if1f_4	Intervalle de temps entre la date d'observation de la première feuille et la floraison au cycle 4	jours	Numérique
if1f_5	Intervalle de temps entre la date d'observation de la première feuille et la floraison au cycle 5	jours	Numérique

Annexe 2 : Liste des plants de bananiers plantains avec absence de croissance (2) et croissance anormale (1) après plantation.

Numéro de ligne	***Variété***	***Numéro du plant***	***Code croissance***
125	*Ukom*	*5*	*2*
210	*Mbirinyong*	*5*	*2*
269	*Poutabang Type Sumo*	*3*	*1*
365	*Mbai*	*5*	*1*
426	*Okele*	*1*	*1*
531	*French Rouge*	*1*	*1*
533	*French Rouge*	*3*	*1*
534	*French Rouge*	*4*	*1*
535	*French Rouge*	*5*	*1*
635	*Obubit Ntanga 76 – 23*	*5*	*1*

Annexe 3 : Liste des 135 variétés représentées par la collection CARBAP (A : Z).

1 :50

3 Vert
3/4 Nain
74 - 79
76 - 17
76 - 22
Abang Type Somo
Abomienu
Agbagba
Akpakpa
Amou
Anka
Apantu
Apem Omniaba
Apempa
Atia Njokou
Bae Ako Ukom
Ban 611
Ban 612
Banane Tigree
Bang Make
Batard
Bend Mossendjo
Big Ebanga
Biya 2
Biye
Bungaoissan
Congo 2
Corne 3
Corne Cantebalon
Corne Rouge 18
Corne Type
Daluyao
Diby 1
Diby 2
Didiedi
Douala
Ebang
Ebang Mboe
Ebang Violet
Ebibi
Ekon Zok
Ekona n°1
Elat
Elat Noir
Esang
Essang
Essong
Fouem
French Clair
French Clair Rose

51 :100

French Clair Rose
French Rouge
French Rouge 12
French Rouge 18
French Rouge 3
French Sombre
Gui
Haa Haa
Hinanga
Igihobe
Ihitism
Itu Iba Ukom
Kar Ngou
Kelong Mekintou
Kwa
Madi Madi
Madre Del Platano
Makemba Noir
Mbai
Mbeta 1
Mbeta n°2
Mbirinyong
Mbirinyong Green Mutant
Mbotoko Vert
Mbouroukou n°1
Mbouroukou n°2
Mbouroukou n°3
Medou
Meki
Messiatso
Messintso
Mimbomena
Mimi Abue
Monganga
Moto Ebanga
Motomoliko
Moufoumbila
Mulolou
Nanguia 1
Ndingo Liko
Ndokondou
Ndom Rouge
Ngok Egome
Niangafelo
Njock Kon
Njombe n°1
Njombe n°2
Njombe n°3
Nkunda
Nothing But Green

101 :135

Nothing But Red
Nselouka
Ntanga 6
Ntie
Nyabang
Nzorba
Obino Lewaï
Obubit Ntanga 76 - 23
Obubit Ntanga 76 - 74
Obubit Tanga Green Mutant
Obubit Ukom
Okele
One Hand Planty
Orishele
Osoaboaso
Oui
Ovang
Owang Rouge
Penkelong Noir
Plantain n°17
Plantain n°2
Plantain n°3
Poutabang Type Somo
PX2
Red Ebanga
Red Ogoni
Red Yade
Rose d'Ekona
Rouge de Loum
Sima Dikondo
Soya
Three Hands Planty
Toutoun Kelong
Two Hands Planty
Ukom
Zip Ekon

Annexe 4 : Capture d'écran d'une partie du data table généré avec R Studio.

Group: Variete = 3 Vert
Dimensions: 5 x 12
Duplicates: 0

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	id [character]	1. 3 Vert_1 2. 3 Vert_2 3. 3 Vert_3 4. 3 Vert_4 5. 3 Vert_5	1 (20.0%) 1 (20.0%) 1 (20.0%) 1 (20.0%) 1 (20.0%)		5 (100.0%)	0 (0.0%)
3	Couleur [character]	1. Vert rouge	5 (100.0%)		5 (100.0%)	0 (0.0%)
4	Typemusalogue [character]	1. Faux-Corne	5 (100.0%)		5 (100.0%)	0 (0.0%)
5	Taillemusalogue [character]	1. geant	5 (100.0%)		5 (100.0%)	0 (0.0%)
6	Nanisme [character]	1. non	5 (100.0%)		5 (100.0%)	0 (0.0%)
7	pays [character]	1. Cote d_Ivoire	5 (100.0%)		5 (100.0%)	0 (0.0%)
8	htfloc1 [integer]	Mean (sd) : 357.2 (18.7) min ≤ med ≤ max: 340 ≤ 355 ≤ 385 IQR (CV) : 24 (0.1)	340 : 1 (20.0%) 341 : 1 (20.0%) 355 : 1 (20.0%) 365 : 1 (20.0%) 385 : 1 (20.0%)		5 (100.0%)	0 (0.0%)
9	c50floc1 [numeric]	Mean (sd) : 74.5 (4) min ≤ med ≤ max: 69 ≤ 74 ≤ 79.5 IQR (CV) : 4 (0.1)	69.00 : 1 (20.0%) 73.00 : 1 (20.0%) 74.00 : 1 (20.0%) 77.00 : 1 (20.0%) 79.50 : 1 (20.0%)		5 (100.0%)	0 (0.0%)
10	prc1 [numeric]	Mean (sd) : 13.2 (1.5) min ≤ med ≤ max: 10.6 ≤ 13.4 ≤ 14.6 IQR (CV) : 0.8 (0.1)	10.60 : 1 (20.0%) 13.20 : 1 (20.0%) 13.40 : 1 (20.0%) 14.00 : 1 (20.0%) 14.60 : 1 (20.0%)		5 (100.0%)	0 (0.0%)
11	nmrc1 [integer]	Min : 8 Mean : 8.4 Max : 9	8 : 3 (60.0%) 9 : 2 (40.0%)		5 (100.0%)	0 (0.0%)
12	ndrc1 [integer]	Mean (sd) : 71.2 (6.7) min ≤ med ≤ max: 65 ≤ 69 ≤ 81 IQR (CV) : 9 (0.1)	65 : 1 (20.0%) 66 : 1 (20.0%) 69 : 1 (20.0%) 75 : 1 (20.0%) 81 : 1 (20.0%)		5 (100.0%)	0 (0.0%)

Annexe 5 :

(a) Statistiques descriptives (arrondies à 0,1) par groupe de new Taille.

Géant	htfloc1	c50floc1	hbulbfloc1	nfefloc1	nfvfloc1	rdfvfloc1	nrvfloc1	prc1	nmrc1	ndrc1	ipfc1	ifcc1	ipcc1
Unité :	*cm*	*cm*	*cm*	*/*	*/*	*/*	*/*	*kg*	*/*	*/*	*jours*	*jours*	*jours*
Min.	187	51	5	40	6	4	3	0	0	0	263	51	333
1er quartile	325	69	7	42	9	32	7,8	9,4	6	28	315,2	84	398,5
Médiane	352	76	8	45	10	35	10	16	8	90	351,5	93	446
Moyenne	364	78,2	9,6	46	9,8	37,1	10,5	16,3	7,8	98,8	368,5	91	457,5
3eme quartile	409	89,9	12,8	49,5	11	42	1	22,2	11	166	416,8	99	516
Max.	504	103,5	20	63	19	56	32	39,5	13	257	593	122	628
NA	0	0	54	0	0	0	11	14	14	14	1	19	19

Moyen													
Min.	205	46,5	0	33	6	9	1	3	1	16	212	52	313
1er quartile	285	57,5	6	36	8	28	4	10	7	39	258	82	348
Médiane	300	60	7	37	9	29	6	12	7	62,5	268	88	356
Moyenne	299	60,3	7,4	36,7	8,6	28,9	6,8	12,1	7,51	62,5	268,6	88,8	357,6
3eme quartile	318	63	8	38	9	30	9	14,5	8	87	278	96	369
Max.	445	81	20	39	11	33	21	25,4	21	127	343	125	432
NA	*0*	*0*	*148*	*0*	*0*	*0*	*2*	*5*	*1*	*3*		*5*	*5*

Petit													
Min.	235	48	7	31	6	23	2	4	6	49	226	92	320
1er quartile	272	52	10	32	6	25	3	5,5	6	63,25	232	94	341
Médiane	280	58	10	32	7	26	4	7,5	6,5	80	241	109	357
Moyenne	284,1	56,5	11,4	31,8	7,3	25,6	5,3	8,2	6,5	75,9	245	106,8	351,8
3eme quartile	290	61	15	32	8	26	5	10,1	7	88	249	115	362
Max.	370	62	15	32	9	27	13	16	7	99	284	121	399
NA	*0*	*0*	*4*	*0*	*0*	*0*	*0*	*1*	*1*	*1*	*0*	*0*	*0*

(b) Statistiques descriptives par groupe de type.

French	htfloc1	c50floc1	hbulbfloc1	nfefloc1	nfvfloc1	rdfvfloc1	nrvfloc1	prc1	nmrc1	ndrc1	ipfc1	ifcc1	ipcc1
Unité :	*cm*	*cm*	*cm*	*/*	*/*	*/*	*/*	*kg*	*/*	*/*	*jours*	*jours*	*jours*
Min.	205	46,5	0	31	6	4	1	2,8	4	38	223	52	313
1er quartile	285	58	6	36	8	28	5	12	7	80	263	88	357
Médiane	315	62,5	8	38	6	30	8	14,8	7	91	288	95	376
Moyenne	333,7	69,4	9,2	40,9	6,8	33	8	16,1	7,3	108,4	323,9	94,2	413
3eme quartile	366	82	12	47	10	38	10	19,4	9	122	387	102	481
Max.	504	103	20	63	17	56	32	39,5	13	257	632	122	628
NA	*28*	*28*	*138*	*28*	*28*	*28*	*40*	*30*	*28*	*30*	*28*	*38*	*38*

Faux-corne													
Min.	187	47	4	34	6	9	2	3	2	4	212	51	313
1er quartile	299	59	6	37	8	29	5	8	7	31	262	79	348
Médiane	310	62	7	38	9	30	7	11	8	37	271	84	356
Moyenne	308,6	62,5	7,6	38,3	9	30,2	7,6	10,7	8,2	38,8	278	84,3	362,9
3eme quartile	325	64,5	10	39	10	31	10	13	8	44	289	89,3	369
Max.	385	81	15	48	19	41	19	22	21	89	379	125	481
NA	*1*	*1*	*78*	*1*	*1*	*1*	*1*	*4*	*1*	*1*	*1*	*4*	*4*

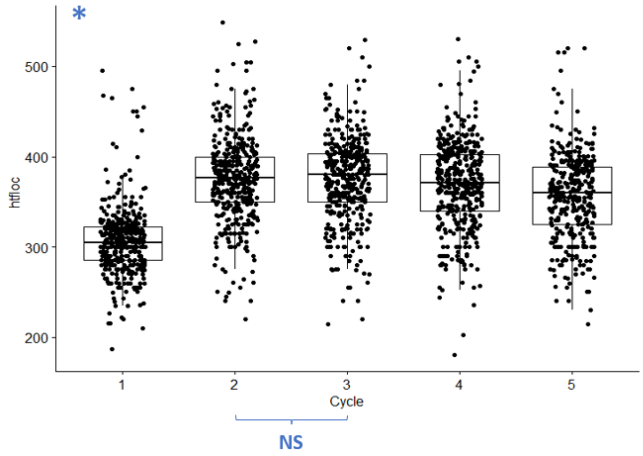
Corne													
Min.	235	52	5	36	8	29	3	0	0	0	243	63	328
1er quartile	307,8	67,4	7	40	9,3	31	7	0	0	0	278,2	82,5	369
Médiane	325	69	7	42	11	32	9	6,4	1	11	309,5	87	390
Moyenne	321	69	7,6	41,7	10,3	32,4	10,4	6,6	1,9	20,1	308,7	87,2	393,6
3eme quartile	339,2	73	8	44	11	33	13	10,4	3	29	333,8	93,5	425
Max.	387	77,5	11	49	13	42	24	18,6	9	94	390	113	488
NA	*11*	*11*	*20*	*11*	*11*	*11*	*7*	*8*	*8*	*8*	*11*	*22*	*22*

Batard													
Min.	350	70,5	7	43	12	32	6	15,6	8	63	315	68	385
1er quartile	355	73	7	43	12	32	7	19,4	8	75	321	70	392
Médiane	364	75,5	7	43	12	32	10	20	8	75	324	84	405
Moyenne	359,8	74,6	7,8	43,8	12,4	32,4	9,6	19,5	8,2	75	325,2	78,8	404
3eme quartile	365	76,5	7,8	44	12	33	12	20,6	8	75,2	329	84	413
Max.	365	77,3	10	46	14	33	13	22	9	81	337	88	425
NA	*0*	*0*	*1*	*0*	*0*	*0*	*0*	*0*	*0*	*82*	*0*	*0*	*0*

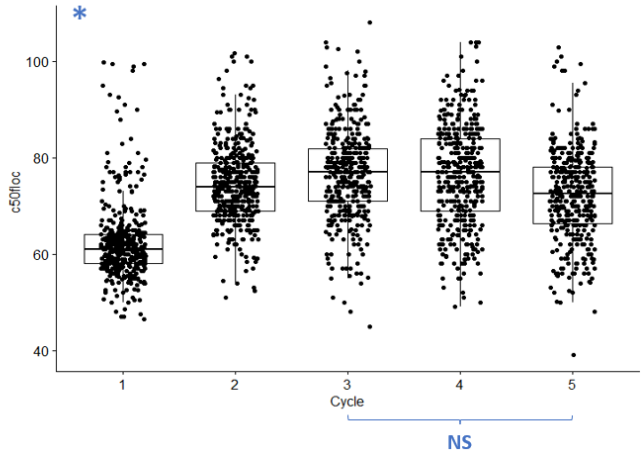
Annexe 6 : Quelques variables en plus pour l'analyse sur 5 cycles de culture.

(a) Variables relatives à la croissance

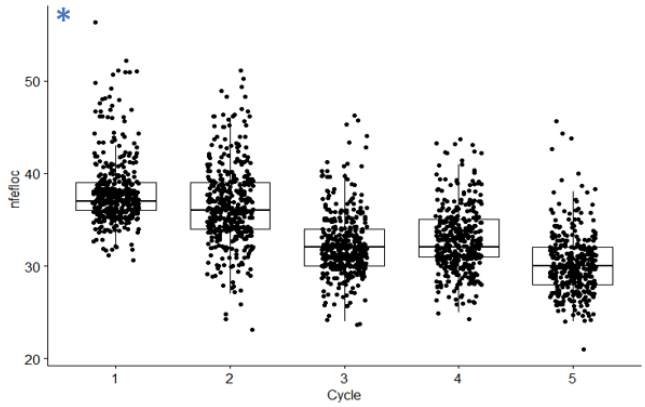
Hauteur du pseudo-tronc à la floraison (cm) :



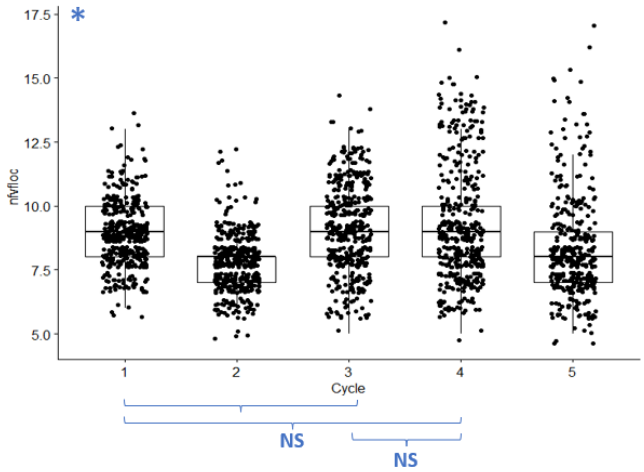
Circonférence à 50 cm du pseudo-tronc à la floraison (cm) :



Nombre de feuilles émises à la floraison :

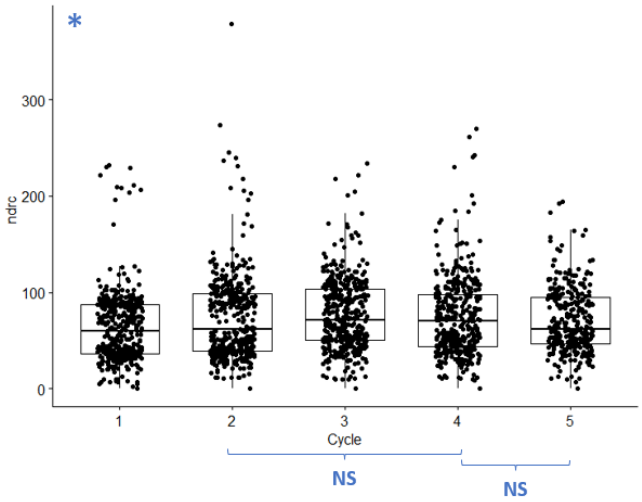


Nombre de feuilles vivantes à la floraison :

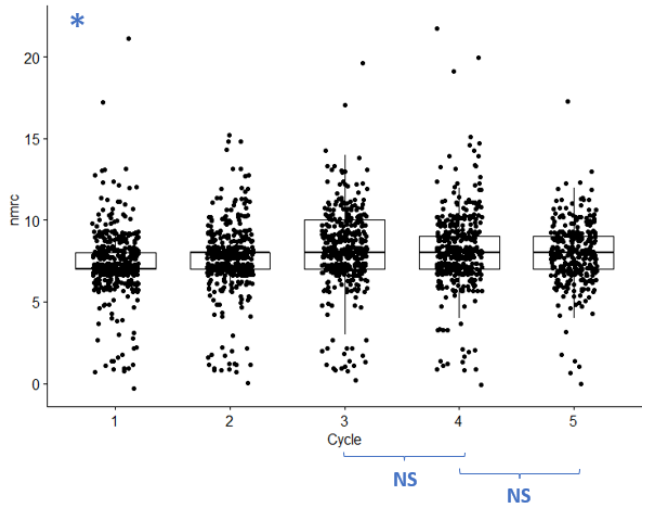


(b) Variables relatives au rendement

Nombre de doigts du régime :



Nombre de mains du régime :



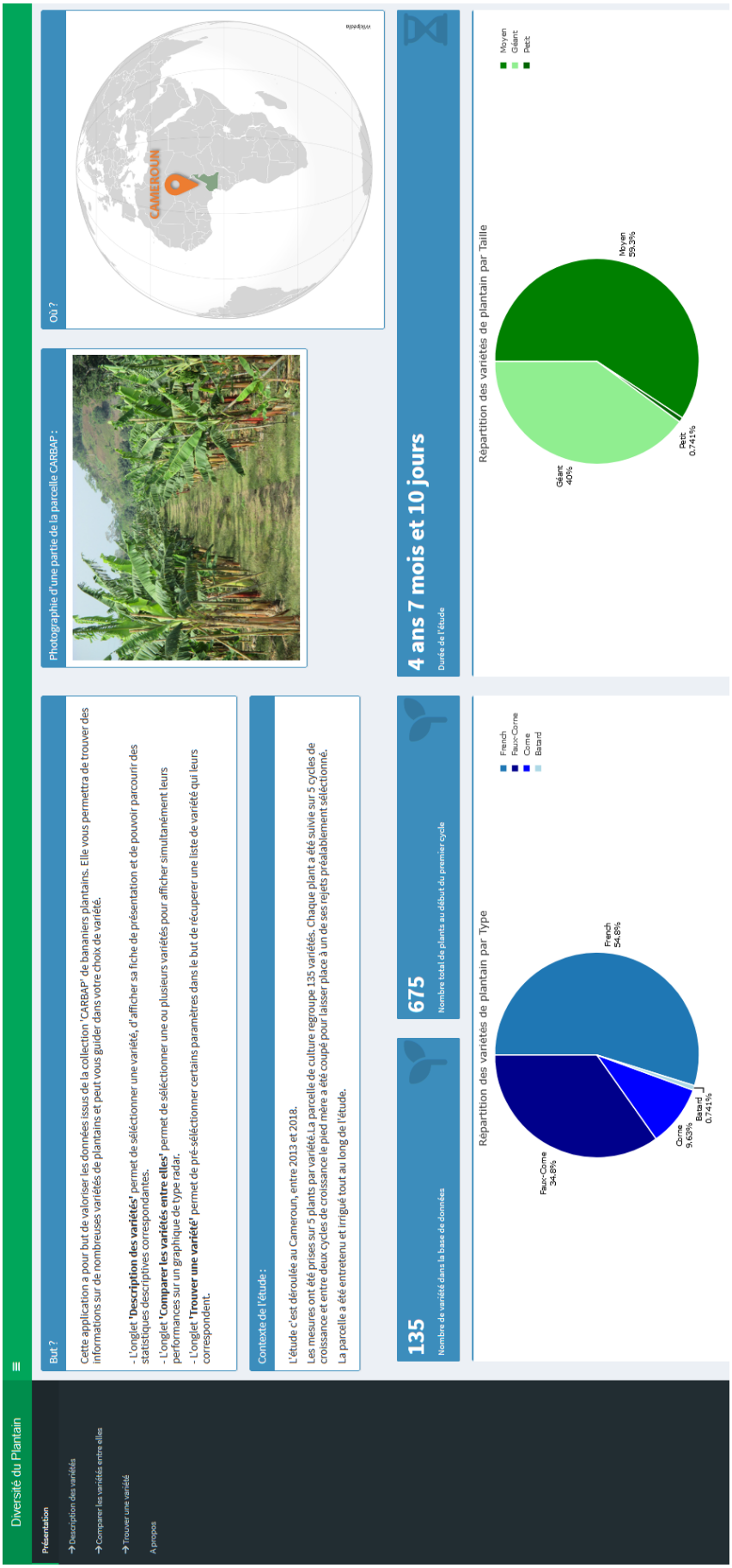
Légendes :

* : test de Kruskal-Wallis p-value < 0,05

NS : résultat du test multiple de Wilcoxon pour données appariées p-value > 0,05

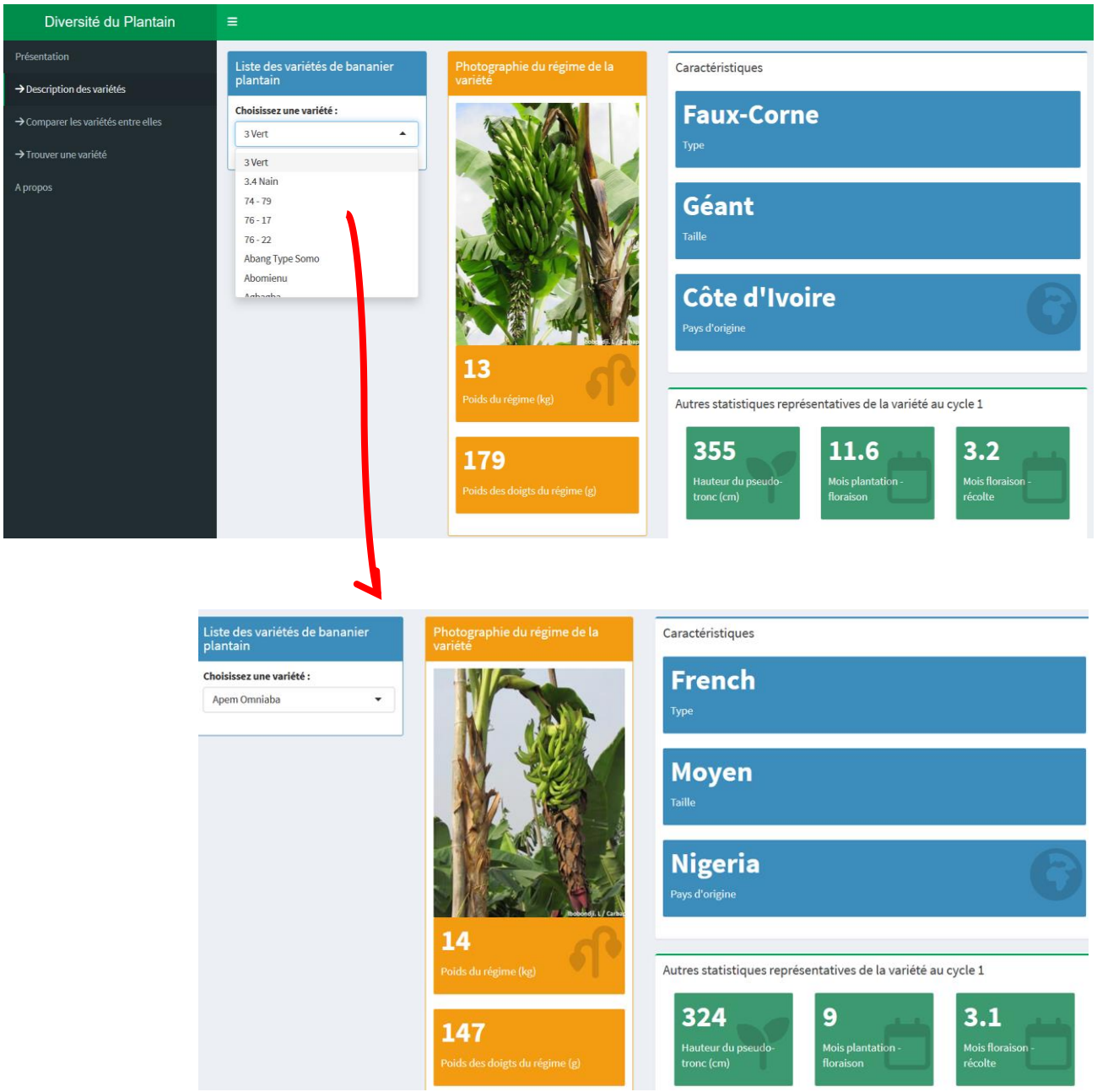
Annexe 7 : Application SHINY

1) Premier onglet « Présentation » :

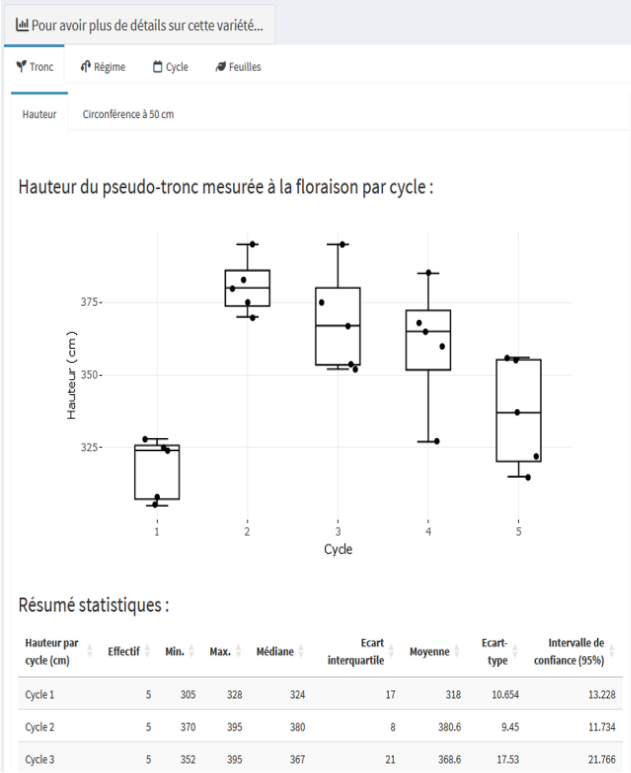


2) Deuxième onglet « Diversité des variétés » :

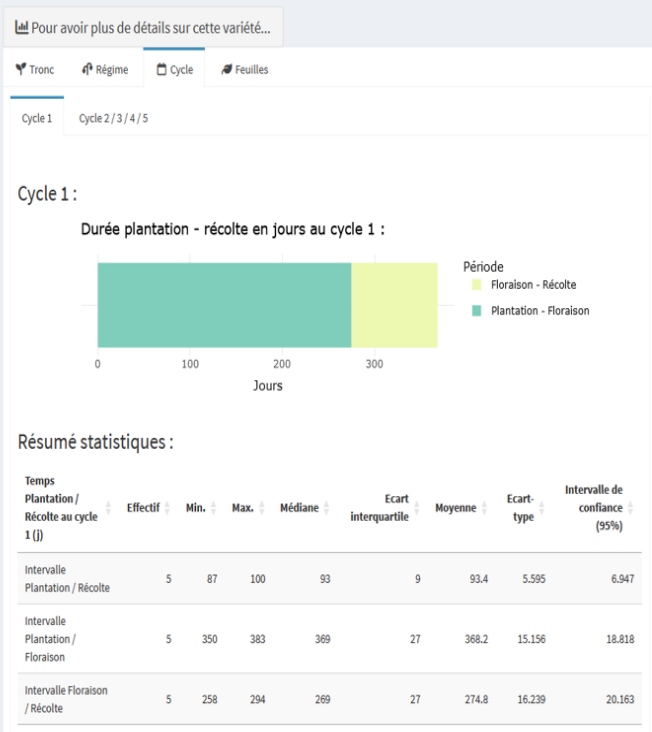
(a)



(b)

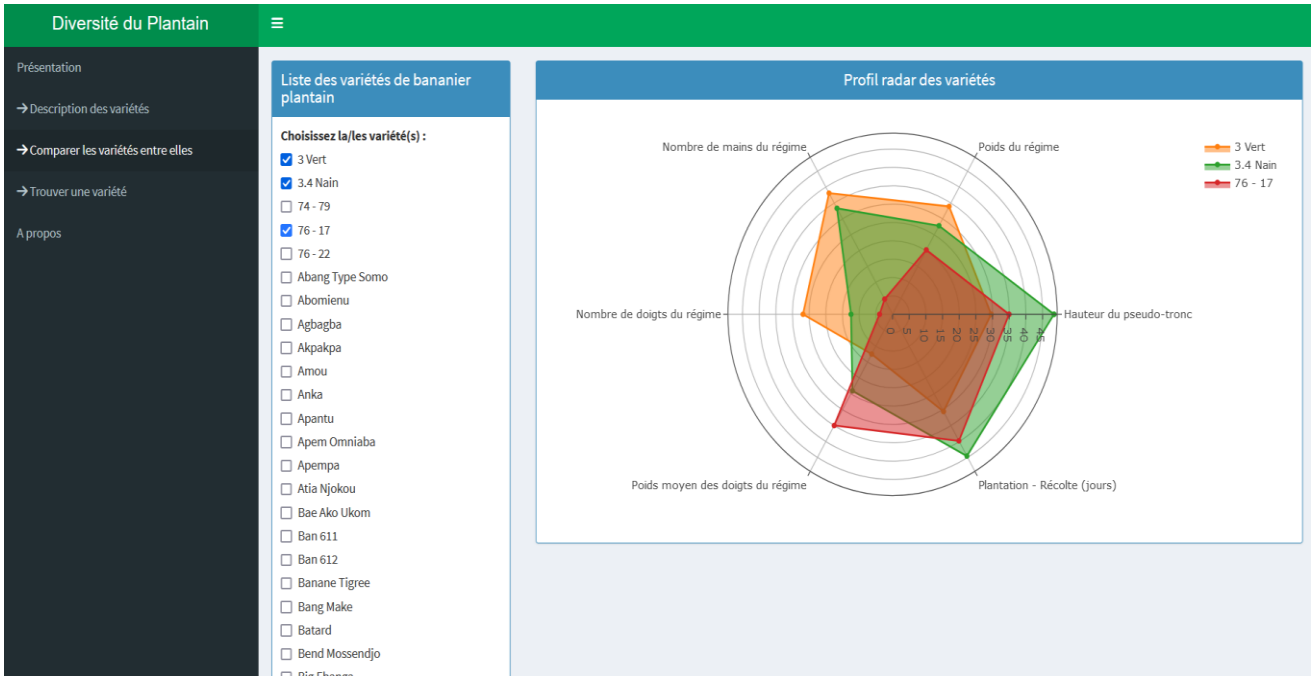


Hauteur par cycle (cm)	Effectif	Min.	Max.	Médiane	Ecart interquartile	Moyenne	Ecart-type	Intervalle de confiance (95%)
Cycle 1	5	305	328	324	17	318	10.654	13.228
Cycle 2	5	370	395	380	8	380.6	9.45	11.734
Cycle 3	5	352	395	367	21	368.6	17.53	21.766

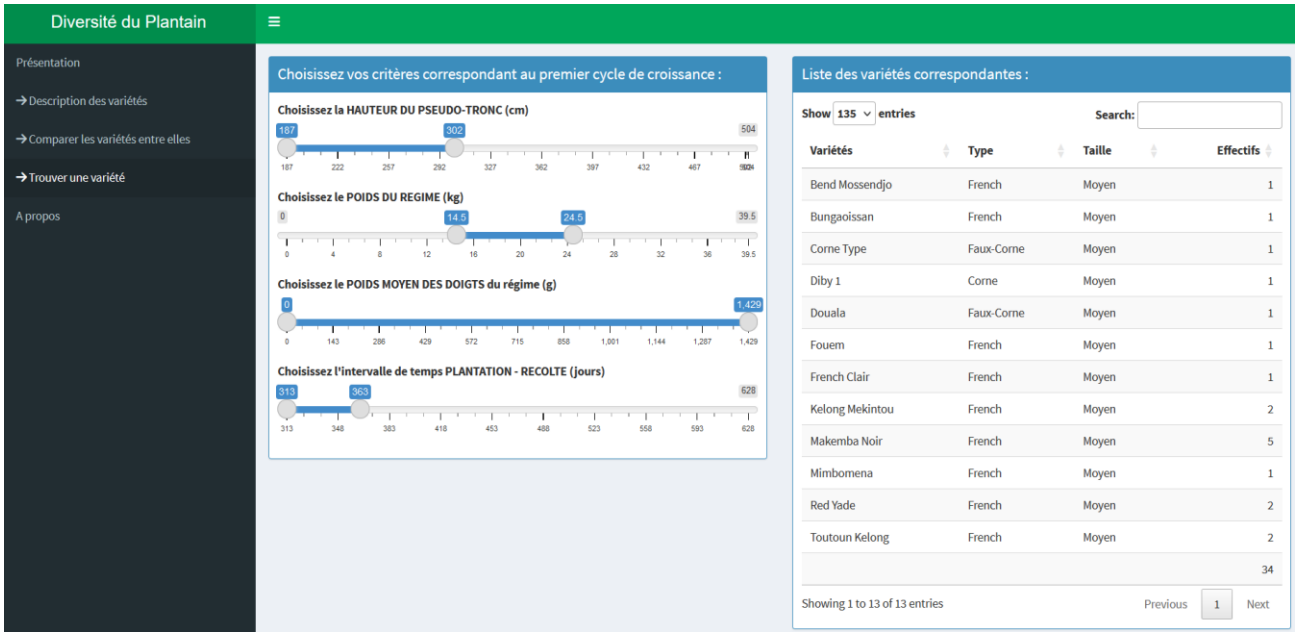
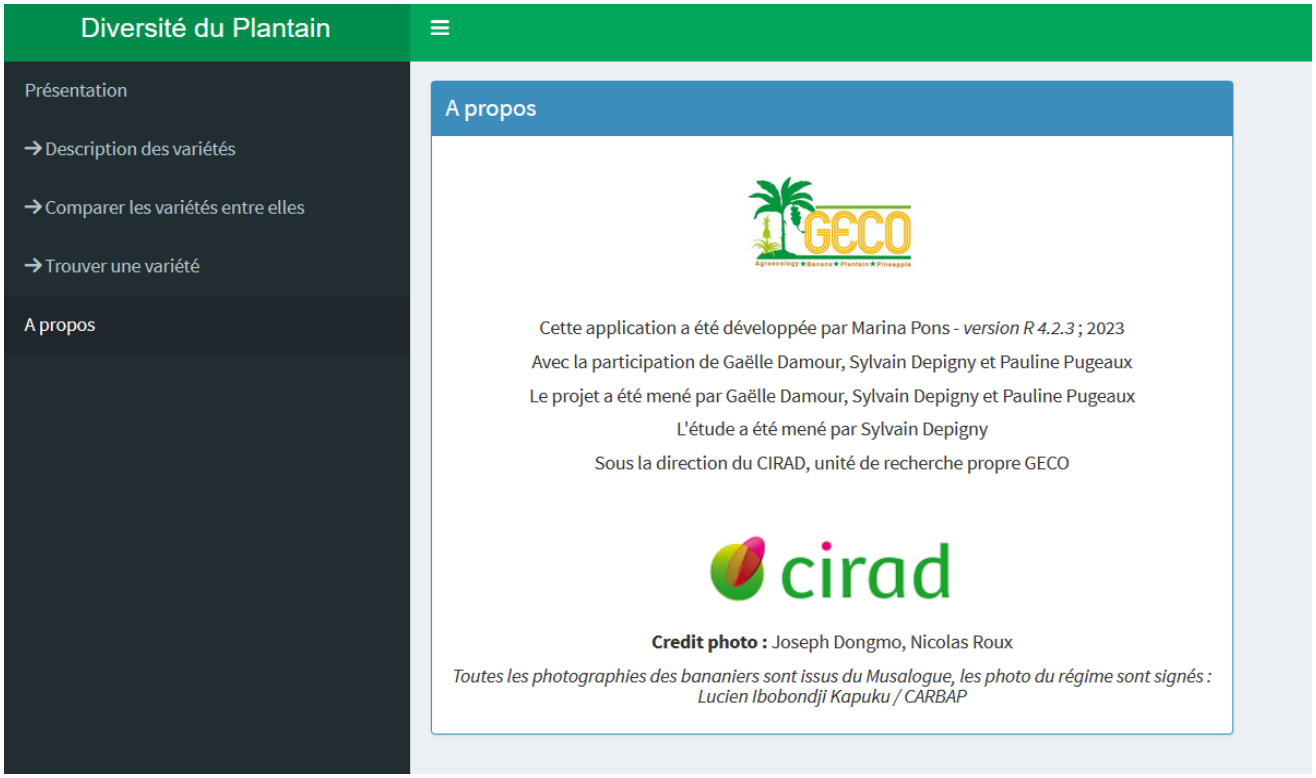


Temps Plantation / Récolte au cycle 1 (j)	Effectif	Min.	Max.	Médiane	Ecart interquartile	Moyenne	Ecart-type	Intervalle de confiance (95%)
Intervalle Plantation / Récolte	5	87	100	93	9	93.4	5.595	6.947
Intervalle Plantation / Floraison	5	350	383	369	27	368.2	15.156	18.818
Intervalle Floraison / Récolte	5	258	294	269	27	274.8	16.239	20.163

3) Troisième onglet « Comparer les variétés entre elles » :



4) Quatrième onglet « Trouver une variété » :



5) Cinquième onglet « A propos » :