Analyse des Données : Production et Qualité des Agrumes

Yibo Wang, Renwen Xu, Qiyao Zhou, Zhihan Zheng
Sorbonne Université
Encadrant : Yassin Mazroui

5 janvier 2025

Table des matières

1 Introduction

1.1 Contexte de l'Étude

Dans le contexte actuel de l'agriculture moderne, l'industrie des agrumes occupe une place centrale dans l'économie agricole. L'étude des facteurs déterminants de la production et de la qualité des fruits est essentielle pour :

1. **Améliorer les pratiques agricoles**, en favorisant une gestion optimale des intrants et des ressources.
2. **Optimiser les rendements**, en identifiant les leviers d'action les plus pertinents pour maximiser la productivité.

1.2 Source des Données

Les données utilisées dans cette étude proviennent d'un concours national de statistiques pour étudiants chinois et comprennent :

— **24 variables**, couvrant les caractéristiques des exploitations, les pratiques agricoles, les intrants et les résultats de production. Ces variables incluent :
  — • **Caractéristiques des exploitations :** type de verger (petit, moyen, intensif), superficie, espacement de plantation ;
  — • **Propriétés des arbres :** âge moyen des arbres, mode de plantation (haute ou basse tige) ;
  — • **Pratiques agricoles :** consommation d'engrais (chimiques et organiques), usage de pesticides, volume d'irrigation ;
  — • **Résultats de production :** rendement moyen par murendement moyen par mu [1] , proportion des fruits commercialisables.

— **289 exploitations agricoles**, représentatives de différentes régions, avec une diversité dans les pratiques agricoles et les types de vergers observés. Les données couvrent une période de 4 ans (2017-2020), permettant une analyse temporelle des évolutions.

Cette analyse poursuit deux objectifs principaux :

1. Identifier les facteurs influençant le rendement par mu à l'aide d'une approche multivariée.
2. Étudier les déterminants de la proportion de fruits commercialisables afin d'optimiser la qualité de la production.

1.3 Méthodologie

1.3.1 Variables d'Étude

Le jeu de données initial comprenait 24 variables, dont 20 ont été sélectionnées après le prétraitement pour l'analyse statistique.

1.3.2 Prétraitement des Données

L'échantillon initial comptait 289 observations collectées entre 2016 et 2020. Après l'application des critères d'exclusion, l'échantillon final se compose de :

— 289 observations valides
— 18 variables explicatives
— 2 variables dépendantes principales

2 Analyse Descriptive

2.1 Analyse du Rendement par Mu

2.1.1 Statistiques Descriptives

L'analyse de la distribution du rendement par mu a été réalisée à l'aide d'un histogramme de fréquence(Figure 1).

L'analyse statistique révèle une distribution bimodale avec une faible normalité, ainsi que la présence de valeurs aberrantes (rendements $<$ 2000 kg/mu). Ces éléments nécessitent un traitement préalable des données avant toute analyse approfondie.

Cette configuration impose deux interventions majeures pour la suite de l'analyse :

Premièrement, il est indispensable d'améliorer la normalité des données. La majorité des analyses statistiques paramétriques envisagées, telles que la régression linéaire et l'analyse de variance, reposent sur l'hypothèse de normalité. L'absence actuelle de normalité pourrait compromettre la validité de ces tests et entraîner des biais dans les conclusions. Une transformation des données sera donc nécessaire pour garantir le respect des conditions d'application.

1. Ici, 'mu' représente une unité de surface équivalente à 666,7 m^2.

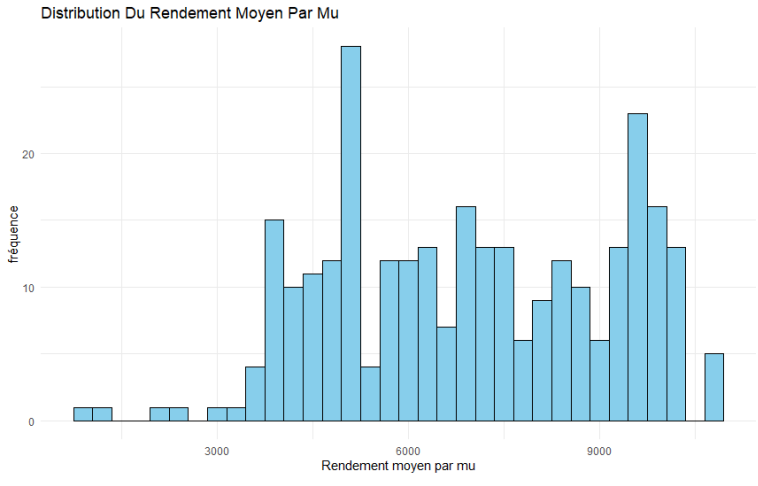

FIGURE 1 : Distribution du rendement moyen par Mu.

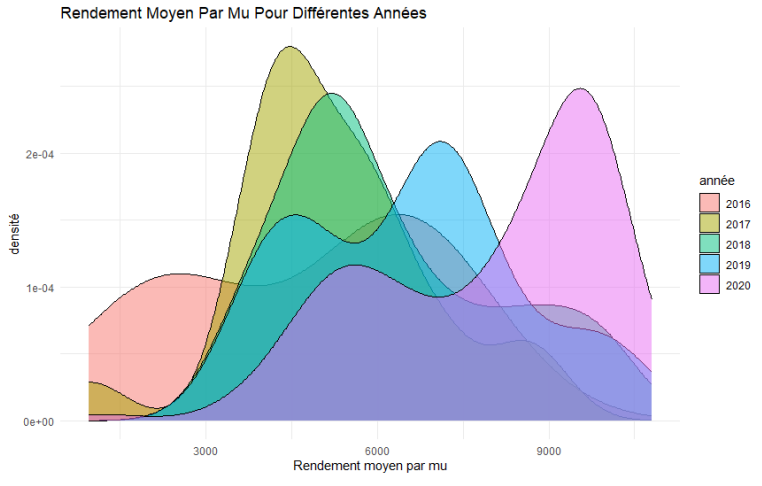

FIGURE 2 : Rendement moyen par Mu pour différentes années.

Deuxièmement, le traitement des valeurs aberrantes est essentiel. Les rendements inférieurs à 2000 kg/mu s'écartent significativement de la tendance générale observée. Ces valeurs extrêmes risquent de biaiser les estimations statistiques et d'occulter les relations réelles entre les variables d'intérêt. Leur exclusion permettra d'obtenir un jeu de données plus représentatif des conditions normales de production et d'assurer la fiabilité des analyses ultérieures.

2.1.2 Distribution des Rendements

L'analyse temporelle des rendements (Figure 2), basée sur les courbes de densité, permet de comparer les distributions des rendements pour chaque année d'observation. Cette analyse met en lumière plusieurs tendances significatives dans l'évolution de la production :

1. **Stabilité en 2017 et 2018 :** Les distributions de ces deux années présentent des pics de densité similaires, centrés autour de 4000 kg/mu. Cette constance reflète une relative stabilité des pratiques agricoles et des conditions de production durant cette période.
2. **Progression en 2019 et 2020 :** Les pics de densité se déplacent vers des rendements plus élevés, atteignant environ 6000 kg/mu en 2019 et 8000 kg/mu en 2020. Cette évolution peut être attribuée à des facteurs tels que :
   — L'amélioration des politiques agricoles mises en œuvre depuis 2018,
   — Le développement des techniques de production,
   — La maturation progressive des arbres fruitiers.
3. **Particularité de 2016 :** Les données de cette année, limitées à trois observations, présentent une distribution atypique, empêchant toute interprétation statistique fiable.

Par ailleurs, la normalité des distributions varie selon les années, ce qui conduit à deux décisions méthodologiques :
— L'exclusion des données de 2016, dont la faible représentativité risque de biaiser les résultats,
— L'amélioration de la normalité des distributions pour garantir l'application valide des tests statistiques paramétriques.

Cette analyse révèle une tendance positive dans les rendements sur la période étudiée, reflétant une amélioration continue des pratiques agricoles. Ces résultats constituent une base solide pour les analyses ultérieures des facteurs influençant la productivité.

2.2 Analyse de la Proportion de Fruits Commercialisables

L'analyse de la proportion des fruits commercialisables a été réalisée à l'aide d'un histogramme de fréquence (Figure 3), mettant en évidence une concentration marquée autour de 90%. Un examen détaillé révèle une asymétrie prononcée dans la répartition des observations. La majorité des exploitations affiche des taux compris entre 85% et 95%, avec un pic notable à 90%. En dehors de cet intervalle, la dispersion est limitée, et les observations inférieures à 80% sont quasiment inexistantes. Cette homogénéité reflète une certaine constance dans la qualité de production des exploitations étudiées.

Face à cette distribution, une classification binaire a été adoptée pour évaluer la qualité de production. Deux catégories ont ainsi été définies :
— **Qualité supérieure :** exploitations atteignant ou dépassant 90% de fruits commercialisables ;
— **Qualité standard :** exploitations n'atteignant pas ce seuil.

Ce choix, fondé sur la distribution observée, caractérisée par une médiane de 0.9020 et une moyenne de 0.8883, reflète une réalité empirique du secteur. Ces résultats trouvent leur fondement dans les observations mentionnées

dans les études scientifiques agricoles de référence [2], soulignant ainsi la pertinence de cette approche. Cette approche dichotomique présente plusieurs avantages méthodologique :

1. Elle permet une distinction claire et objective entre les exploitations.
2. Elle facilite l'identification des facteurs discriminants entre les deux catégories de qualité.
3. Elle constitue une base solide pour l'application de modèles de régression logistique, permettant une analyse approfondie des déterminants de la qualité.

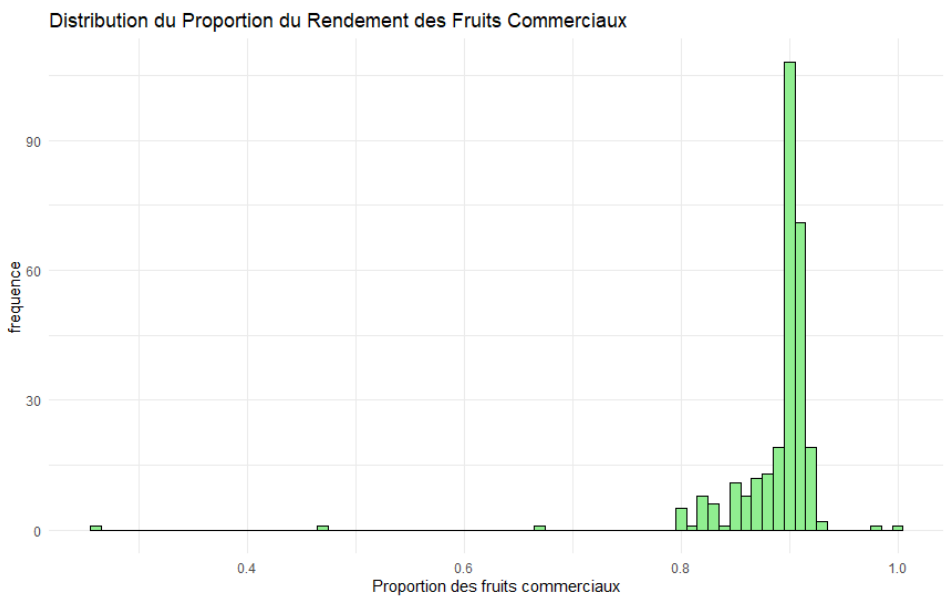

FIGURE 3 : Distribution du proportion du rendement des fruits commerciaux. En général, pour la grande majorité des exploitations agricoles, la production fruitière est de qualité, avec une proportion de fruits commerciaux qui se concentre autour de 90%.

Ces premières observations soulèvent des questions fondamentales : quelles pratiques agricoles et quelles caractéristiques des exploitations favorisent une proportion élevée de fruits commercialisables ? Ces interrogations guideront nos analyses statistiques approfondies dans les sections suivantes.

2.3 Analyse des Corrélations et Tests d'Hypothèses Complémentaires

L'analyse des relations entre les variables a révélé de fortes corrélations entre certaines d'entre elles. Pour quantifier ces relations, une matrice de corrélation symétrique a été calculée (Figure 4).

L'examen de cette matrice met en évidence plusieurs associations particulièrement élevées :

1. **Quantité moyenne de d'engrais organique utilisée par mu (variable 8) :** Corrélée très fortement avec :
   — Variable 6, la quantité moyenne d'engrais utilisée ($r = 0.98$),
   — Variable 4, le volume moyen d'irrigation par mu ($r = 0.98$),
   — Variable 2, la consommation moyenne de pesticides par mu ($r = 0.98$).
2. **Quantité moyenne d'engrais utilisée par mu (variable 6) :** Montre des corrélations élevées avec :
   — Variable 4, le volume moyen d'irrigation par mu ($r = 0.97$),
   — Variable 2, la consommation moyenne de pesticides par mu ($r = 0.98$).
3. **Volume moyen d'irrigation par mu(variable 4) et consommation moyenne de pesticides par mu(variable 2) :** Corrélés fortement entre eux ($r = 0.98$).

Ces corrélations élevées indiquent que ces variables capturent des aspects similaires de l'intensité des pratiques agricoles. Par conséquent, il est pertinent de les traiter comme un ensemble cohérent dans les analyses ultérieures.

Pour compléter notre analyse descriptive, des tests t de Student ont été réalisés afin d'évaluer l'effet de la technique de plantation en basse tige sur deux variables clés :

1. **Le rendement moyen par mu**,
2. **La proportion de fruits commerciaux**.

2. Hendre, P. D., S. A. Ranpise, and P. S. Pawar. "Effect of different irrigation and fertigation levels on fruit quality parameters in sweet orange (Citrus sinensis L. Osbeck) cv. Phule Mosambi." Journal of Pharmacognosy and Phytochemistry, vol. 9, no. 3, 2020, pp. 503–507. Web.

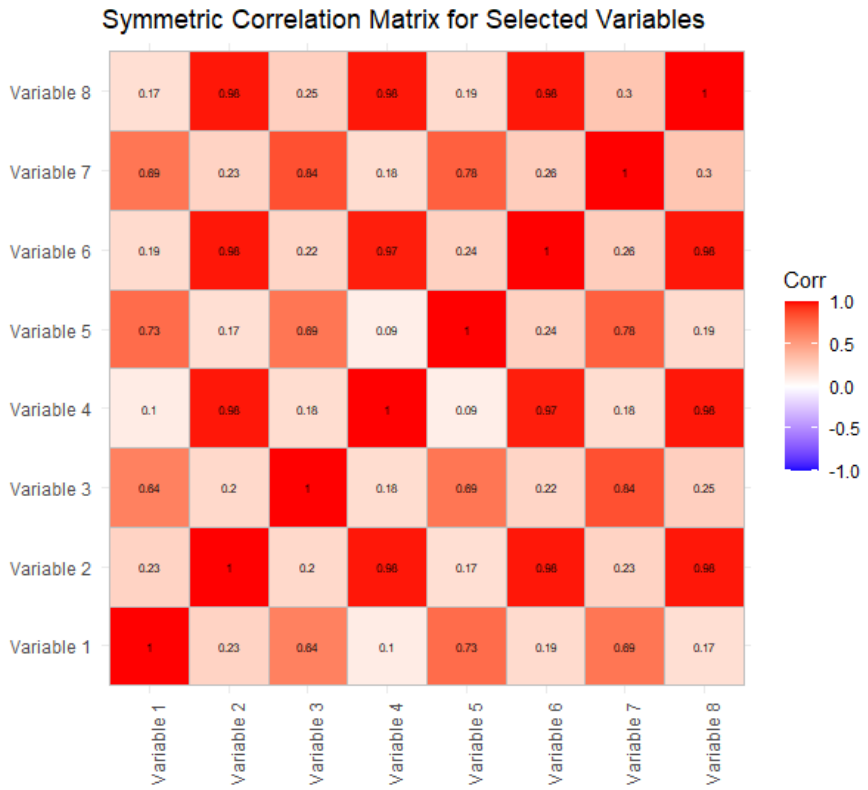

FIGURE 4 : Matrice de corrélation symétrique pour les variables sélectionnées. Les variables sont définies comme suit : **Variable 1** : quantité annuelle de consommation de pesticides (litres), **Variable 2** : consommation moyenne de pesticides par Mu, **Variable 3** : volume d'irrigation (mètres cubes), **Variable 4** : volume moyen d'irrigation par Mu, **Variable 5** : quantité d'engrais utilisée (kilogrammes), **Variable 6** : quantité moyenne d'engrais utilisée par Mu (kilogrammes), **Variable 7** : quantité d'engrais organique utilisée (kilogrammes), **Variable 8** : quantité moyenne d'engrais organique utilisée par Mu (kilogrammes).

Les résultats obtenus sont les suivants :

— Pour le rendement moyen par mu : p-value $= 0.7643$ (> 0.05),

— Pour la proportion de fruits commerciaux : p-value $= 0.9455$ (> 0.05).

Dans les deux cas, les p-values sont largement supérieures au seuil conventionnel de significativité (0.05), indiquant que la technique de plantation en basse tige n'a pas d'effet statistiquement significatif ni sur le rendement moyen par mu, ni sur la proportion de fruits commerciaux.

3 Amélioration de la Normalité des Données

Dans le cadre de notre analyse statistique, l'amélioration de la normalité des données constitue une étape essentielle. En effet, la plupart des tests statistiques paramétriques envisagés, tels que la régression linéaire et l'analyse de variance, reposent sur l'hypothèse fondamentale d'une distribution normale des données.

La transformation Box-Cox s'avère particulièrement appropriée pour cet objectif. Elle permet non seulement de réduire l'asymétrie, mais également d'améliorer l'homoscédasticité des données, tout en préservant les relations entre les variables, garantissant ainsi la validité des analyses statistiques.

3.1 Transformation Box-Cox

La transformation Box-Cox est définie mathématiquement par :

$$y(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{si } \lambda \neq 0 \\ \ln(y) & \text{si } \lambda = 0 \end{cases}$$

L'optimisation du paramètre λ a été réalisée en maximisant la log-vraisemblance de la transformation. Comme l'illustre la Figure 5, la valeur optimale obtenue est $\lambda = 0,666$. Cette valeur spécifique garantit une transformation efficace pour normaliser la distribution des données, selon la formule suivante :

$$\frac{y^{0,666} - 1}{0,666}$$

3.2 Évaluation de la Transformation

Pour évaluer l'efficacité de la transformation Box-Cox, nous avons comparé les diagrammes Quantile-Quantile (Q-Q plots) (Figure 6) des données avant et après transformation. Ces graphiques permettent de visualiser la correspondance entre les quantiles empiriques des données et les quantiles théoriques d'une distribution normale.

L'analyse des diagrammes Q-Q révèle une amélioration significative de la normalité des données après transformation. Dans le graphique de gauche, représentant les données brutes, une déviation importante par rapport à la ligne de référence est observée, en particulier aux extrémités. Cette déviation reflète une distribution initiale marquée par une forte asymétrie et la présence de valeurs extrêmes, confirmée par les écarts importants aux deux extrémités du graphique.

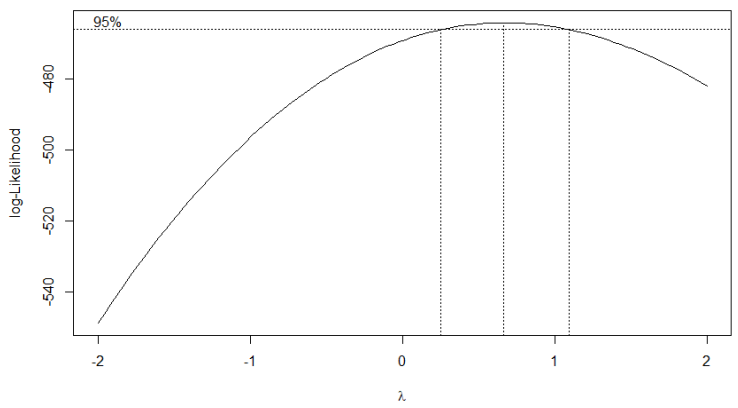

FIGURE 5 : Log-likelihood de la transformation Box-Cox. On observe que le maximum de log-likelihood est atteint autour de 0.6.

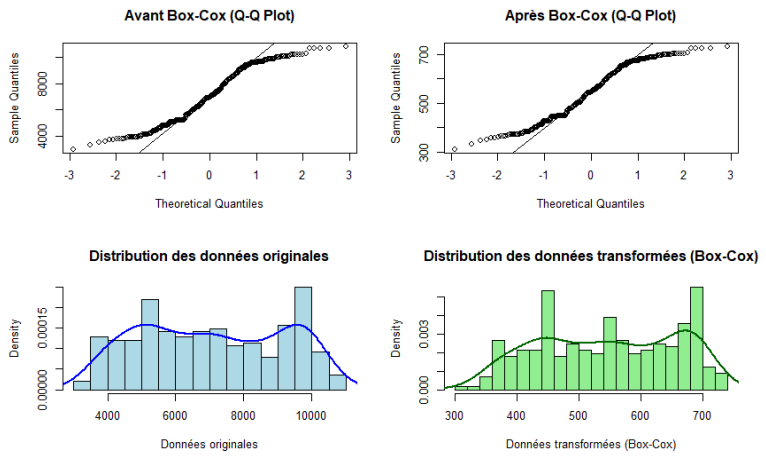

FIGURE 6 : Effet de la transformation Box-Cox. À gauche, les QQ-plots et la distribution avant la transformation ; à droite, après la transformation. On observe qu'après la transformation Box-Cox, les écarts sont nettement réduits, suggérant une amélioration de la normalité des données, bien que celles-ci ne suivent toujours pas une distribution normale parfaite.

Résultats du test de normalité de Shapiro-Wilk (brutes vs Box-Cox)

Type de données	**Valeur p**
Données brutes	8.508×10^{-9}
Données transformées (Box-Cox)	1.486×10^{-8}

Remarque : Malgré une légère augmentation des valeurs p après la transformation Box-Cox, elles restent très inférieures à 0,05. Cela confirme que les données ne suivent toujours pas une distribution normale.

En revanche, le graphique de droite, représentant les données transformées, montre un alignement nettement plus proche de la ligne de référence. Cette amélioration se traduit par plusieurs éléments notables :

1. **Réduction de l'asymétrie :** La distribution des données devient visiblement plus symétrique.
2. **Diminution de l'impact des valeurs extrêmes :** Une meilleure adéquation est observée aux extrémités du graphique.
3. **Homogénéité accrue :** Les points se répartissent de manière plus uniforme autour de la ligne de référence, suggérant une symétrie globale améliorée.

Bien que la transformation Box-Cox n'ait pas permis de générer une distribution parfaitement normale, elle a amélioré la normalité des données. Ces améliorations incluent une réduction de l'asymétrie, une atténuation de l'influence des valeurs extrêmes et une meilleure qualité globale du jeu de données. Ces ajustements contribuent à mieux respecter les hypothèses fondamentales des tests statistiques paramétriques, tels que la régression linéaire et l'analyse de variance, tout en offrant une base plus fiable pour les analyses ultérieures.

4 Analyse en Composantes Principales

4.1 Prétraitement des Données

La première étape de notre analyse en composantes principales (ACP) a consisté à préparer rigoureusement les données pour garantir la qualité de l'analyse. Les données ont été standardisées pour éliminer l'effet d'échelle

entre les variables. Les valeurs manquantes identifiées au cours de ce processus ont été imputées par les moyennes correspondantes, une méthode couramment utilisée pour préserver la structure globale des données et assurer la validité de l'analyse.

4.2 Sélection des Composantes

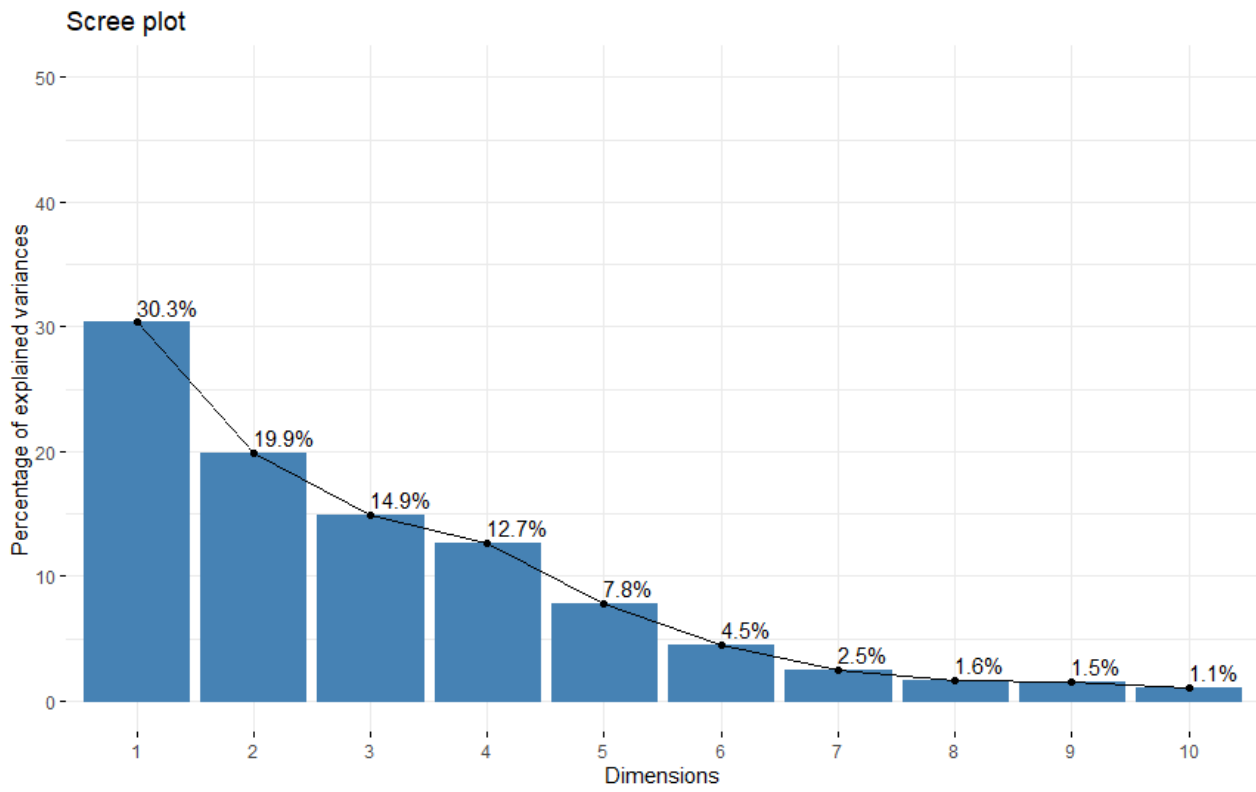

FIGURE 7 : Scree plot issu de l'Analyse en Composantes Principales (ACP). Le graphique montre que le premier composant principal (Dim 1) explique 30.3% de la variance totale, le deuxième (Dim 2) 19.9%, et le troisième (Dim 3) 14.9%. Les quatre premiers composants principaux (Dim 1 à Dim 4) cumulent environ 77.8% de la variance totale. Au-delà du cinquième composant (Dim 5), les contributions marginales des composants deviennent négligeables.

Pour déterminer le nombre optimal de composantes principales, deux critères complémentaires ont été appliqués :

1. **Critère de Kaiser :** Seules les composantes ayant des valeurs propres supérieures à 1 sont retenues. Ce seuil est illustré par une ligne horizontale rouge à $y = 1$.
2. **Proportion de variance cumulée expliquée :** Les composantes retenues doivent expliquer une proportion substantielle de la variance totale, généralement fixée à un seuil de 80%.

L'application conjointe de ces deux critères a conduit à sélectionner cinq composantes principales pour l'analyse.

4.3 Interprétation des Résultats

L'analyse des relations entre variables et composantes principales est rendue complexe par le grand nombre de variables impliquées. Bien que le cercle des corrélations (Figure 8) fournisse une visualisation utile, sa lisibilité est limitée par la densité d'informations. Une approche plus systématique a donc été adoptée en examinant la matrice des contributions des variables aux composantes principales (Voir Appendix Table 2).

Principales Contributions des Composantes :

1. **Première composante (30.3% de la variance totale) :** Liée aux intrants agricoles tels que les engrais, les pesticides et l'irrigation.
2. **Deuxième composante (19.9% de la variance totale) :** Représente une forte corrélation entre les intrants agricoles, indiquant une gestion intégrée des ressources.
3. **Troisième composante :** Associée à la structure physique des vergers, notamment la densité et l'espacement de plantation.
4. **Quatrième composante :** Reflète les aspects productifs et l'âge moyen des arbres.
5. **Cinquième composante :** Combine des éléments d'identification et des indicateurs de performance, tels que la proportion de fruits commercialisables.

Cette décomposition met en lumière la complexité des facteurs influençant la production d'agrumes. Elle souligne particulièrement le rôle clé de la gestion intégrée des intrants agricoles dans l'optimisation de la production.

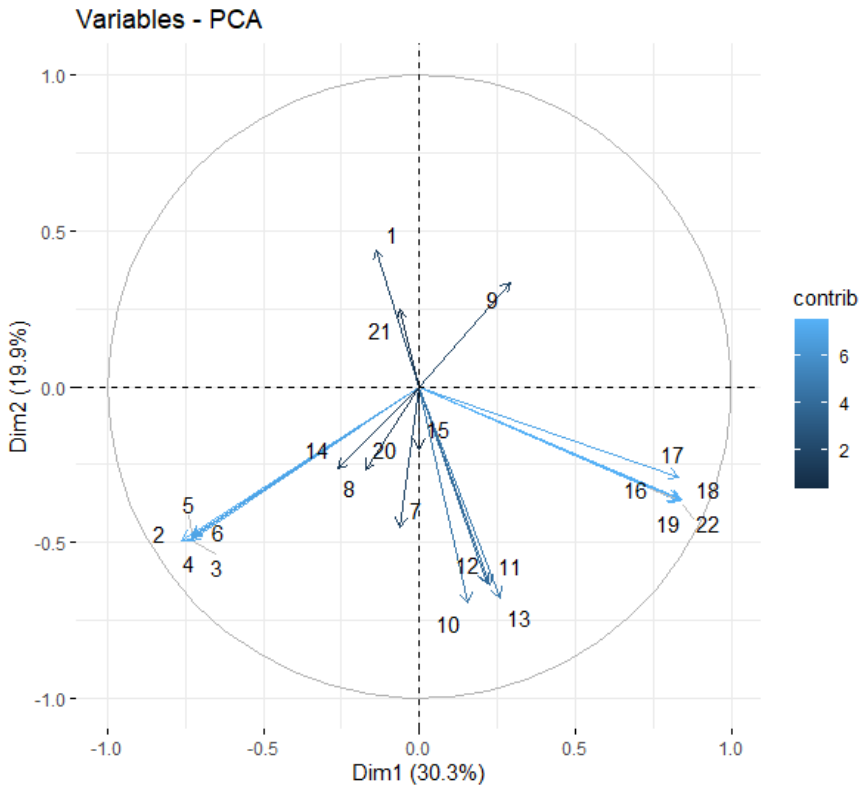

FIGURE 8 : Cercle de corrélation issues de l'Analyse en Composantes Principales (ACP). On observe deux groupes de variables ayant une contribution significative aux dimensions Dim1 et Dim2 : **(1)** Le groupe (2, 3, 4, 5, 6) correspond aux variables suivantes : quantité moyenne d'engrais utilisée par mu (kilogrammes), quantité de fertilisants utilisée, quantité moyenne d'engrais organique utilisée par mu (kilogrammes), volume moyen d'irrigation par mu, et consommation moyenne de pesticides par mu ; **(2)** Le groupe (16, 17, 18, 19, 22) correspond aux variables suivantes : quantité annuelle de consommation de pesticides (litres), quantité d'engrais organique utilisée (kilogrammes), volume d'irrigation (mètres cubes), quantité d'engrais utilisée (kilogrammes), et rendement total (kilogrammes).

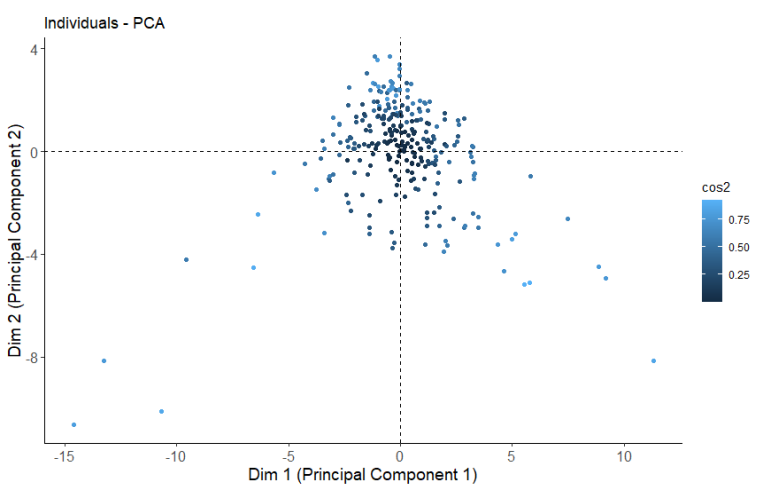

FIGURE 9 : Projection des individus sur les dimensions principales Dim1 et Dim2.

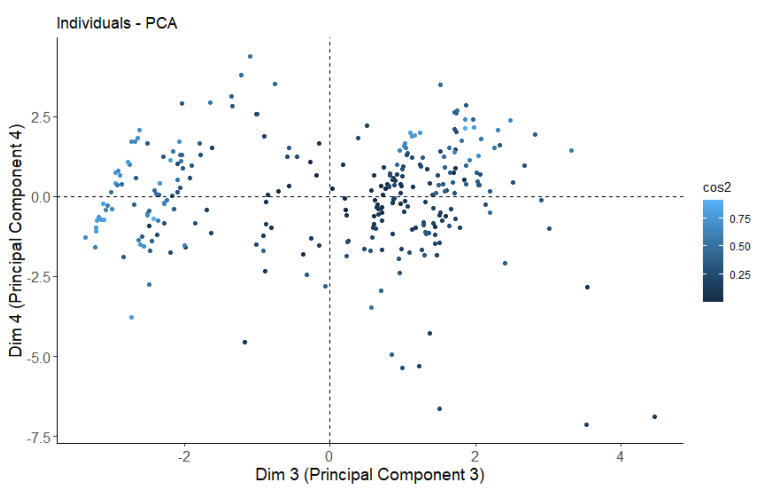

FIGURE 10 : Projection des individus sur les dimensions principales Dim3 et Dim4. Ces graphiques montrent la distribution des individus dans l'espace principal.

5 Modélisation par Régression Linéaire Multiple

5.1 Construction du Modèle

Pour identifier les facteurs influençant le rendement par mu, un modèle de régression linéaire multiple a été développé sur les données transformées par la méthode Box-Cox. La sélection des variables explicatives a été effectuée à l'aide d'une procédure stepwise, permettant de retenir systématiquement les variables les plus pertinentes pour le modèle.

5.2 Résultats du Modèle

Le modèle final retient six variables explicatives significatives :
— Le numéro d'échantillon ①,
— L'année ②,

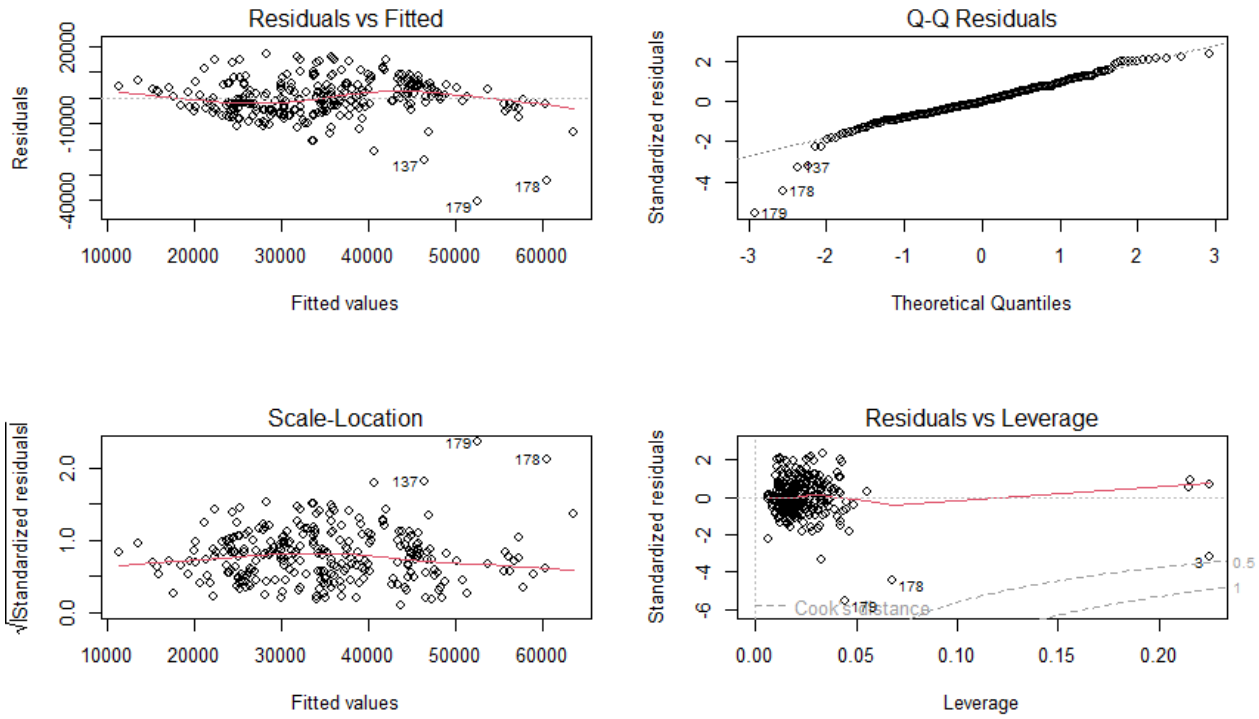

FIGURE 11 : Résultats de diagnostic pour le modèle de régression multiple. Les quatre graphiques montrent que le modèle s'adapte bien à la majorité des points, mais on observe quelques exceptions : **(1)** Dans les graphiques en haut à gauche, en bas à gauche et en haut à droite, les points 137, 178, et 179 apparaissent comme des valeurs aberrantes ; **(2)** Dans le graphique en bas à droite, les points 3, 178, et 179 s'approchent de la ligne de référence Cook's Distance à 0.5, indiquant une influence potentielle élevée sur le modèle.

— La superficie ③,
— L'âge moyen des arbres ④,
— L'espacement de plantation ⑤,
— La consommation annuelle de pesticides ⑥.

L'équation du modèle optimal s'écrit :

$$\frac{\text{rendement par mu}^{\lambda}-1}{\lambda} = -3,205 \times 10^6 + 121,5 \cdot ① + 1585 \cdot ② - 1,698 \cdot ③ + 5799 \cdot ④ - 3060 \cdot ⑤ + 8,609 \cdot ⑥$$

La validation du modèle a été effectuée à plusieurs niveaux :

1. **Comparaison de modèles :** L'analyse de variance (ANOVA) entre le modèle complet et le modèle simplifié donne une p-valeur de 0.9672 (> 0.05), confirmant que la simplification n'entraîne pas de perte significative de qualité prédictive.
2. **Analyse de la multicolinéarité :** Les facteurs d'inflation de la variance (VIF) sont tous inférieurs à 10, indiquant l'absence de multicolinéarité problématique entre les variables explicatives.
3. **Diagnostic des résidus :** L'examen des graphiques de diagnostic (Figure 11) montre :
   — Une distribution homogène des résidus autour de zéro,
   — Une absence de pattern systématique,
   — Une distribution approximativement normale des résidus,
   — Peu de points à effet levier important.

Ces résultats confirment que les hypothèses sous-jacentes de la régression linéaire multiple sont respectées et que le modèle est robuste.

TABLE 1 – Performances des modèles de régression multiple avant et après optimisation par Step-wise (R^2)

Ensemble de données	**Modèle complet (Full Data)**	**Modèle optimisé (Step-wise)**
Entraînement (Training Set)	0.7149572	0.7078819
Test (Test Set)	0.3306897	0.3575764

Remarque : Les résultats montrent que, bien que le modèle complet ait une légère meilleure capacité explicative sur l'ensemble d'entraînement ($R^2 = 0.7149572$), le modèle optimisé par Step-wise améliore la performance sur l'ensemble de test ($R^2 = 0.3575764$ contre $R^2 = 0.3306897$). Cela suggère que l'optimisation par Step-wise peut atténuer les problèmes de sur-ajustement, renforçant ainsi les capacités prédictives du modèle sur des données non vues.

5.3 Conclusions

Les analyses statistiques réalisées dans cette étude ont permis de dégager deux résultats majeurs :

1. **Impact des intrants agricoles :** Les variables liées aux intrants agricoles (engrais organique, engrais chimique, irrigation, pesticides) exercent des effets similaires sur le rendement. Cela se traduit par la sélection unique des pesticides comme variable représentative dans le modèle final.
2. **Configuration spatiale des vergers :** L'espacement de plantation influence significativement le rendement, tandis que la technique de plantation en basse tige n'a pas d'effet notable, comme le confirme son exclusion du modèle optimisé.

Ces résultats mettent en lumière l'importance d'une gestion intégrée des intrants agricoles et d'une configuration spatiale appropriée des vergers pour maximiser la production d'agrumes.

6 Analyse par Régression Logistique

6.1 Classification de la Qualité

Afin d'étudier les déterminants de la qualité de production, une variable binaire a été créée sur la base de la proportion de fruits commercialisables. En nous appuyant sur la distribution observée, un seuil de 90% a été défini pour distinguer deux catégories :

— **Production de haute qualité** : Proportion de fruits commercialisables $\geq 90\%$.
— **Production standard** : Proportion de fruits commercialisables $< 90\%$.

6.2 Modélisation

Un modèle de régression logistique a été développé pour prédire la qualité de production. Après une procédure de sélection *stepwise*, le modèle final a retenu les variables explicatives suivantes :

— **Numéro d'échantillon** : Variable d'identification,
— **Type de verger** : Catégorisé en petit, moyen ou intensif,
— **Densité de plantation** (plants/mu),
— **Quantité annuelle de consommation de pesticides** (litres),
— **Volume d'irrigation** (mètres cubes),
— **Utilisation de la technique de plantation en basse tige** (binaire : 0 pour non, 1 pour oui),
— **Volume moyen d'irrigation par mu**.

6.3 Résultats du Modèle

Le modèle logistique obtenu s'exprime par :

$$\begin{aligned}\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = {} & 13.72 - 0.0405 \cdot \text{numéro d'échantillon} - 0.8946 \cdot \text{type de verger moyen} \\ & - 2.5154 \cdot \text{type de verger petit} - 0.0336 \cdot \text{densité de plantation} \\ & + 0.0025 \cdot \text{quantité annuelle de pesticides} - 0.0839 \cdot \text{volume d'irrigation} \\ & + 1.6115 \cdot \text{technique basse tige} + 0.5625 \cdot \text{volume moyen d'irrigation par mu.}\end{aligned}$$

6.4 Validation et Performance du Modèle

La validation du modèle a été réalisée à plusieurs niveaux pour évaluer sa robustesse et sa capacité prédictive :

1. **Analyse de Déviance :** Une comparaison entre le modèle complet et le modèle optimisé montre une *p-valeur* de 0.9865, indiquant qu'il n'y a pas de perte significative de performance après simplification.
2. **Capacité Prédictive :** Sur le jeu d'entraînement, la précision est de 80.69% avec la matrice de confusion suivante :

	élevée	**faible**
élevée	104	26
faible	13	59

Sur le jeu de test, la précision est de 71.26% avec la matrice de confusion suivante :

	élevée	**faible**
élevée	38	18
faible	7	24

3. **Multicolinéarité :** L'analyse des facteurs d'inflation de variance (*VIF*) montre que toutes les valeurs sont inférieures à 3.1, confirmant l'absence de multicolinéarité problématique.
4. **AUC et ROC :** La courbe ROC du modèle optimisé présente une AUC de 0.82, indiquant une capacité prédictive satisfaisante pour différencier les deux catégories de qualité.

6.5 Conclusion

Le modèle de régression logistique final offre des performances solides pour prédire la qualité de production des agrumes, avec une précision globale satisfaisante et des variables explicatives pertinentes. Bien que les résultats sur le jeu de test soient légèrement inférieurs à ceux sur le jeu d'entraînement, le modèle reste robuste grâce à son AUC élevée et à l'absence de multicolinéarité. Ces résultats confirment la pertinence des variables sélectionnées et leur impact sur la qualité de production.

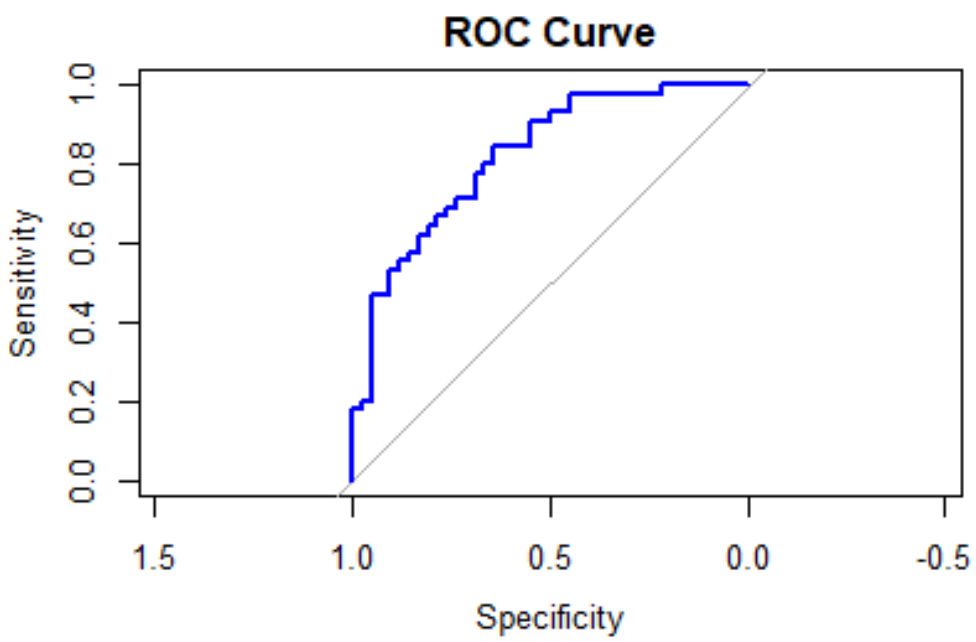

FIGURE 12 : Courbe ROC du modèle de régression logistique optimisé sur le jeu de test. La courbe montre une nette séparation par rapport à la diagonale aléatoire, indiquant une bonne capacité de discrimination. L'AUC est de 0,8201, reflétant une performance globale satisfaisante. L'accuracy sur le jeu de test est de 0,7126, avec une légère baisse par rapport à l'ensemble d'entraînement (0,8069), ce qui témoigne d'une bonne généralisation.

7 Conclusion Générale et Perspectives

7.1 Synthèse des Résultats

Cette étude statistique menée sur 289 exploitations d'agrumes a permis de mettre en lumière les facteurs déterminants du rendement et de la qualité :

- **Rendement :** Une augmentation continue entre 2017 et 2020 a été observée, attribuable à l'amélioration des pratiques agricoles.
- **Structure des données :** L'analyse en composantes principales a révélé que les pratiques liées aux intrants agricoles (engrais, irrigation, pesticides) sont fortement corrélées, confirmant une gestion intégrée par les exploitants.
- **Déterminants clés :** La régression linéaire multiple a souligné :
  - L'impact de l'âge des arbres.
  - L'importance de l'espacement de plantation.
  - La gestion des pesticides.
- **Qualité :** Le modèle de régression logistique a montré une précision prédictive élevée (71.26%), mettant en avant :
  - Le rôle crucial du type de verger.
  - La gestion de l'irrigation.
  - L'utilisation raisonnée des pesticides.

7.2 Implications Pratiques

À partir de ces résultats, plusieurs recommandations se dessinent :

— Encourager une **gestion intégrée des intrants agricoles** pour maximiser le rendement et la qualité.
— Porter une **attention accrue à l'espacement de plantation**, facteur clé influençant la productivité.
— Optimiser les pratiques d'**irrigation** et d'**utilisation des pesticides** pour garantir un impact positif sur la qualité et la durabilité.
— Considérer **l'âge des arbres** dans les stratégies de renouvellement pour maintenir des rendements stables à long terme.

7.3 Limites et Perspectives

Bien que cette étude offre des résultats significatifs, certaines limites doivent être prises en compte :
— **Période d'observation courte :** La période d'analyse (2017-2020) limite la portée temporelle des conclusions.
— **Données environnementales absentes :** Le manque d'informations sur les facteurs climatiques et les caractéristiques des sols réduit l'analyse globale.
— **Spécificité géographique :** Les exploitations étudiées étant localisées dans une seule région, cela peut limiter la généralisation des résultats.
Ces observations ouvrent plusieurs perspectives :
— Étendre l'étude sur une période plus longue pour analyser les **tendances temporelles**.
— Inclure des données climatiques et environnementales pour enrichir l'analyse.
— Comparer les résultats avec d'autres régions productrices pour évaluer les **différences structurelles**.
— Explorer davantage les **interactions complexes entre pratiques agricoles**, afin d'optimiser les stratégies intégrées.
Ainsi, cette étude jette les bases pour des pratiques agricoles plus efficaces et offre des pistes prometteuses pour des recherches futures sur la production d'agrumes.

A Table 2

TABLE 2 – Principales contributions pour chaque composante principale (PCA)

Composante Principale	**Principales Contributions**
Dim 1	quantité moyenne d'engrais utilisée par mu (kilogrammes), quantité de fertilisants utilisée, quantité moyenne d'engrais organique utilisée par mu (kilogrammes), volume moyen d'irrigation par mu, consommation moyenne de pesticides par mu
Dim 2	quantité annuelle de consommation de pesticides (litres), quantité d'engrais organique utilisée (kilogrammes), volume d'irrigation (mètres cubes), quantité d'engrais utilisée (kilogrammes), rendement total (kilogrammes)
Dim 3	densité de plantation (plants/mu), densité de plantation (plants par mu), espacement de plantation (mètres), utilisation de la technique de plantation en basse tige ? (0 non, 1 oui), rendement des fruits commerciaux (kilogrammes)
Dim 4	rendement moyen par mu, quantité annuelle de consommation de pesticides (litres), volume d'irrigation (mètres cubes), quantité d'engrais utilisée (kilogrammes), âge moyen des arbres (ans)
Dim 5	numéro d'échantillon, rendement moyen par mu, proportion du rendement des fruits commerciaux, âge moyen des arbres (ans), quantité d'engrais utilisée (kilogrammes)

```
# packages installation
required_packages <- c("ggplot2", "readxl", "ggcorrplot", "MASS", "ade4", "factoextra",
"ggrepel", "car", "pROC")
install_missing_packages <- function(packages) {
  missing_packages <- packages[!(packages %in% installed.packages()[, "Package"])]
  if (length(missing_packages) > 0) {
    install.packages(missing_packages)
  }
}
install_missing_packages(required_packages)
lapply(required_packages, library, character.only = TRUE)
# data loading
data <- read_excel("agrume1.xlsx")
summary(data)
##### 2. Analyses descripives #####
yield <- data$`rendement moyen par mu`
fruit_ratio <- data$`proportion du rendement des fruits commerciaux`
# 2.1.1 rendement moyen par mu distribution
ggplot(data, aes(x = yield)) +
  geom_histogram(binwidth = 300, fill = "skyblue", color = "black") +
  ggtitle("Distribution Du Rendement Moyen Par Mu") +
  xlab("Rendement moyen par mu") +
  ylab("fréquence") +
  theme_minimal()
# 2.1.2 rendement moyen par an
ggplot(data, aes(x = yield, fill = as.factor(année))) +
  geom_density(alpha = 0.5) +
  ggtitle("Rendement Moyen Par Mu Pour Différentes Années") +
  xlab("Rendement moyen par mu") +
  ylab("densité") +
  labs(fill = "année") +
  theme_minimal()
# Statistiques descriptives
summary_stats <- data.frame(
  Mean = mean(yield, na.rm = TRUE),
  Median = median(yield, na.rm = TRUE),
  SD = sd(yield, na.rm = TRUE),
  Quantiles = quantile(yield, probs = c(0.25, 0.5, 0.75), na.rm = TRUE),
  Variance = var(yield, na.rm = TRUE)
)
print(summary_stats)
# 2.2 proportion du rendement des fruits commerciaux
fruit_ratio <- data$`proportion du rendement des fruits commerciaux`
# distribution
ggplot(data, aes(x = fruit_ratio)) +
  geom_histogram(binwidth = 0.01, fill = "lightgreen", color = "black") +
  ggtitle("Distribution du Proportion du Rendement des Fruits Commerciaux") +
  xlab("Proportion des fruits commerciaux") +ylab("frequence") +theme_minimal()
# statistique descriptive
mean_value <- mean(fruit_ratio, na.rm = TRUE)
median_value <- median(fruit_ratio, na.rm = TRUE)
sd_value <- sd(fruit_ratio, na.rm = TRUE)
variance_value <- var(fruit_ratio, na.rm = TRUE)
quantiles <- quantile(fruit_ratio, probs = c(0.25, 0.5, 0.75), na.rm = TRUE)
range_values <- range(fruit_ratio, na.rm = TRUE)
fruit_ratio_stats <- data.frame(
  Statistic = c("Mean", "Median", "Standard Deviation", "Variance", "Q1", "Q2
(Median)", "Q3", "Min", "Max"),
  Value = c(mean_value, median_value,
sd_value, variance_value, quantiles[1],
quantiles[2], quantiles[3], range_values[1],
    range_values[2]
  )
)
print(fruit_ratio_stats)
# 2.3 Analyse des Corrélations
# retirer les variables qui n'ont pas de sens
filtered_data <- data[, !colnames(data) %in% c(
  "rendement total (kilogrammes)",
"superficie (mus)", "rendement des fruits commerciaux (kilogrammes)","rendement des
fruits de seconde qualité (kilogrammes)",
  "densité de plantation (plants par mu)"
)]
# Extraire les variables sous-jacentes
selected_columns <- c(
  "quantité annuelle de consommation de pesticides (litres)", "consommation
moyenne de pesticides par mu", "volume d'irrigation (mètres cubes)", "volume moyen
d'irrigation par mu", "quantité d'engrais utilisée (kilogrammes)", "quantité moyenne
d'engrais utilisée par mu (kilogrammes)", "quantité d'engrais organique utilisée
(kilogrammes)",
  "quantité moyenne d'engrais organique utilisée par mu (kilogrammes)"
)
filtered_data <- data[selected_columns]
# Renommer les colonnes pour des noms simplifiés
colnames(filtered_data) <- paste0("Variable ", seq_along(selected_columns))
# Calculer la matrice de corrélation
cor_matrix <- cor(filtered_data, use = "complete.obs")
# Carte thermique du coefficient de corrélation
ggcorrplot(cor_matrix, hc.order = FALSE,
  type = "full", lab = TRUE, lab_size = 2,
```

```
  colors = c("blue", "white", "red"),
  title = "Symmetric Correlation Matrix for Selected Variables",
  tl.cex = 9,  tl.srt = 90  )
# 2.4  Tests d'Hypothèses Complémentaires
# `utilisation de la technique de plantation en basse tige ?  (0 non, 1 oui)` -> `rendement moyen par mu`
ggplot(data, aes(x = as.factor(`utilisation de la technique de plantation en basse tige ? (0 non, 1 oui)`), y = yield)) +
  geom_boxplot(fill = "lightblue", color = "black") +
  ggtitle("Impact de la Technique de Plantation sur le Rendement Moyen Par Mu") +xlab("Utilisation de la technique (0 = non, 1 = oui)") +ylab("Rendement moyen par mu") +
  theme_minimal()
dev.off()
t_test_result <- t.test(
  yield ~ data$`utilisation de la technique de plantation en basse tige ?  (0 non, 1 oui)`,
  alternative = "two.sided")
print(t_test_result)
# `utilisation de la technique de plantation en basse tige ?  (0 non, 1 oui)` -> `proportion du rendement des fruits commerciaux`
t_test_result2 <- t.test(
  data$`proportion du rendement des fruits commerciaux` ~ data$`utilisation de la technique de plantation en basse tige ?  (0 non, 1 oui)`,  alternative = "two.sided")
print(t_test_result2)
##### 3. Amélioration de la Normalité des Données #####
# 3.1 Transformation Box-Cox
# Filtrer les données
data_filtered <- subset(data, `rendement moyen par mu` > 2000 & année != 2016)
y <- data_filtered$`rendement moyen par mu`
# Transformation de Box-Cox
boxcox_result <- boxcox(lm(y ~ 1), lambda = seq(-2, 2, by = 0.1))
# Obtenir le lambda optimal
lambda_opt <- boxcox_result$x[which.max(boxcox_result$y)]
y_transformed <- ((y^lambda_opt) - 1) / lambda_opt
# Configurer une disposition de 2 lignes et 2 colonnes pour les graphiques
par(mfrow = c(2, 2))
# Tracer le Q-Q plot des données originales
qqnorm(y, main = "Avant Box-Cox (Q-Q Plot)")
qqline(y)
# Tracer le Q-Q plot des données transformées par Box-Cox
qqnorm(y_transformed, main = "Après Box-Cox (Q-Q Plot)")
qqline(y_transformed)
# Tracer l'histogramme des données originales
hist(y, breaks = 20, col = "lightblue", border = "black", main = "Distribution des données originales", xlab = "Données originales", probability = TRUE)
lines(density(y), col = "blue", lwd = 2)  # Ajouter une courbe de densité
# Tracer l'histogramme des données transformées par Box-Cox #
hist(y_transformed, breaks = 20, col = "lightgreen", border = "black", main = "Distribution des données transformées (Box-Cox)", xlab = "Données transformées (Box-Cox)", probability = TRUE)
lines(density(y_transformed), col = "darkgreen", lwd = 2)  # Ajouter une courbe de densité
# Effectuer le test de Shapiro-Wilk pour vérifier la normalité #
shapiro_original <- shapiro.test(y)
# Test pour les données originales
shapiro_transformed <- shapiro.test(y_transformed)
# Test pour les données transformées
# Afficher les résultats des tests
cat("Résultat du test de Shapiro-Wilk pour les données originales:\n")
print(shapiro_original)
cat("\nRésultat du test de Shapiro-Wilk pour les données transformées:\n")
print(shapiro_transformed)
#### 4. Analyse en Composantes Principales ##### scaling
numeric_columns <- sapply(data_filtered, is.numeric)
X <- data_filtered[, numeric_columns]
X_scaled <- scale(X)
# la matrice de covariance, eigenvalues
# examiner s'il existe "NA" ou "infty"
any(is.na(X_scaled))  # NA
any(is.infinite(X_scaled)) # infty
# remplacer NA par la moyenne
X_scaled[is.na(X_scaled)] <- mean(X_scaled, na.rm = TRUE)
cov_matrix <- cov(X_scaled)
pca_results <- eigen(cov_matrix)
pca_results$values
# Scree plot
barplot(pca_results$values, main = "Scree Plot", xlab = "Principal Components", ylab = "Eigenvalues")
abline(h = mean(pca_results$values), col = "red", lty = 2)  # critère de Kaiser
# Proportion cumulative de la variance.
explained_variance <- pca_results$values / sum(pca_results$values)
cumulative_variance <- cumsum(explained_variance)
print(cumulative_variance)
ACP <- dudi.pca(X_scaled, scannf = FALSE, nf = 5)
# Scree Plot
fviz_eig(ACP, addlabels = TRUE, ylim = c(0, 50))
pca_var <- get_pca_var(ACP)
variable_numbers <- 1:nrow(pca_var$coord)
fviz_pca_var(ACP, col.var = "contrib", repel = FALSE, label = "none" ) +
```

```
geom_text_repel(aes(x = pca_var$coord[, 1], y = pca_var$coord[, 2], label =
variable_numbers ), size = 4,
box.padding = 0.3,    point.padding = 0.2,
segment.color = "gray" )
# Dim 1 2
fviz_pca_ind(ACP, col.ind = "cos2",    label = "none",        geom = "point"             ) +
theme_classic() + xlab("Dim 1 (Principal Component 1)") + ylab("Dim 2 (Principal
Component 2)") + theme(
      axis.line = element_line(color = "black"),
      axis.text = element_text(size = 12),
      axis.title = element_text(size = 14))
# Dim 3 4
fviz_pca_ind(ACP, axes = c(3, 4), col.ind = "cos2", label = "none", geom = "point"
) + theme_classic() +      xlab("Dim 3 (Principal Component 3)") + ylab("Dim 4 (Principal
Component 4)") + theme(
      axis.line = element_line(color = "black"),
      axis.text = element_text(size = 12),
      axis.title = element_text(size = 14)
   )
contributions <- pca_var$contrib[, 1:5]    # Contribution
print(contributions)
for (i in 1:5) { cat(paste("Composante principale", i, "main contribution：\n"))
  top_vars <- names(sort(contributions[, i], decreasing = TRUE)[1:5])
print(top_vars)
cat("\n")}
#### 5. Modélisation par Régression Linéaire Multiple ####
# Preprocess the data
data_regression<-data_filtered[, !names(data_filtered) %in% c("rendement des fruits
commerciaux (kilogrammes)", "rendement des fruits de seconde qualité
(kilogrammes)",
"rendement total (kilogrammes)", "densité de plantation (plants par mu)", "proportion
du rendement des fruits commerciaux")]
initial_model <- lm(`rendement moyen par mu` ~ ., data = data_regression)
# La transformée de Box-Cox
boxcox_results <- boxcox(initial_model, lambda = seq(-2, 2, 0.1), plotit = FALSE)
lambda_optimal<-boxcox_results$x[which.max(boxcox_results$y)]
cat("lambda optimal pour Box-Cox:", lambda_optimal, "\n")
if (lambda_optimal == 0) {
   data_regression$Y_transformed<- log(data_regression$`rendement moyen par mu`)
} else {
   data_regression$Y_transformed<-    (data_regression$`rendement moyen par
mu`^lambda_optimal - 1) / lambda_optimal
}
data_regression <- data_regression[, !names(data_regression) %in% c("rendement
moyen par mu")]
data_regression$année <- as.factor(data_regression$année)
data_regression$`type de verger` <- as.factor(data_regression$`type de verger`)
data_regression$année <- as.numeric(as.character(data_regression$année))
# Split the preprocessed data into training and testing sets
set.seed(42)    # Set a seed for reproducibility
train_indices <- sample(1:nrow(data_regression), size = 0.7 * nrow(data_regression))
train_data <- data_regression[train_indices, ]
test_data <- data_regression[-train_indices, ]
# Construire le modèle de régression linéaire multiple sur le training set
full_model_train <- lm(Y_transformed ~ ., data = train_data)
summary(full_model_train)
# Step-wise pour sélectionner le meilleur modèle sur le training set
best_model_train <- step(full_model_train, direction = "both")
summary(best_model_train)
# Utiliser ANOVA pour tester si best_model diffère significativement avec full_model
(training set)
anova(full_model_train, best_model_train)
# Examiner la colinéarité (training set)
vif(best_model_train)
# Visualiser les résidus (training set)
par(mfrow = c(2, 2))
plot(best_model_train)
# Calculate performance metrics for the training set
predicted_train <- predict(best_model_train, newdata = train_data)
# Calculate R-squared for the training set
tss_train <- sum((train_data$Y_transformed - mean(train_data$Y_transformed))^2)
ssr_train <- sum((train_data$Y_transformed - predicted_train)^2)
r_squared_train <- 1 - (ssr_train / tss_train)
cat("Training Set R-squared:", r_squared_train, "")
# Test the model on the testing set
predicted_test <- predict(best_model_train, newdata = test_data)
predicted_test[is.na(predicted_test)] <- mean(predicted_test, na.rm = TRUE)
# Calculate R-squared for the testing set
tss_test <- sum((test_data$Y_transformed - mean(test_data$Y_transformed))^2)
ssr_test <- sum((test_data$Y_transformed - predicted_test)^2)
r_squared_test <- 1 - (ssr_test / tss_test)
cat("Test Set R-squared:", r_squared_test, "
")
# Calculate R-squared for full_model on the training set
predicted_full_train <- predict(full_model_train, newdata = train_data)
tss_full_train <- sum((train_data$Y_transformed - mean(train_data$Y_transformed))^2)
ssr_full_train <- sum((train_data$Y_transformed - predicted_full_train)^2)
r_squared_full_train <- 1 - (ssr_full_train / tss_full_train)
cat("Training Set R-squared for Full Model:", r_squared_full_train, "
")
```

```
# Calculate R-squared for full_model on the testing set
predicted_full_test <- predict(full_model_train, newdata = test_data)
predicted_full_test[is.na(predicted_full_test)] <- mean(predicted_full_test, na.rm = TRUE)
tss_full_test <- sum((na.omit(test_data$Y_transformed) - mean(na.omit(test_data$Y_transformed)))^2)
ssr_full_test <- sum((na.omit(test_data$Y_transformed) - predicted_full_test)^2)
r_squared_full_test <- 1 - (ssr_full_test / tss_full_test)
cat("Test Set R-squared for Full Model:", r_squared_full_test, "
")
#### 6. Analyse par Régression Logistique ####
# Pre-processing the data
data$`qualité de production` <- ifelse(data$`proportion du rendement des fruits commerciaux` > 0.9, "élevée", "faible")
data$`qualité de production` <- as.factor(data$`qualité de production`)
summary(train_data$`qualité de production`)
data_logistic <- subset(data, select = -c(`proportion du rendement des fruits commerciaux`, `rendement moyen par mu`,
`rendement total (kilogrammes)`,
`rendement des fruits commerciaux (kilogrammes)`,
`rendement des fruits de seconde qualité (kilogrammes)`,
`densité de plantation (plants par mu)`))
# Handle missing values
data_logistic$`espacement de plantation (mètres)`[is.na(data_logistic$`espacement de plantation (mètres)`)] <- 3.5
data_logistic$`type de verger` <- as.factor(data_logistic$`type de verger`)
data_logistic$année <- as.numeric(as.character(data_logistic$année))
# Split the data into training and testing sets
set.seed(42)  # Set a seed for reproducibility
train_indices <- sample(1:nrow(data_logistic), size = 0.7 * nrow(data_logistic))
train_data <- data_logistic[train_indices, ]
test_data <- data_logistic[-train_indices, ]
# Build the logistic regression model on the training set
logistic_model <- glm(`qualité de production` ~ ., data = train_data, family = binomial)
# Perform step-wise variable selection
optimized_model <- step(logistic_model, direction = "both")
anova(optimized_model, logistic_model)
# Evaluate the optimized model on the training set
train_predicted_probs <- predict(optimized_model, newdata = train_data, type = "response")
train_predicted_classes <- ifelse(train_predicted_probs >= 0.5, "faible", "élevée")
# Confusion matrix and accuracy for the training set
train_confusion_matrix <- table(Predicted = train_predicted_classes, Actual = train_data$`qualité de production`)
train_accuracy <- mean(train_predicted_classes == train_data$`qualité de production`)
# Print results for training set
cat("\nConfusion Matrix for Training Set:\n")
print(train_confusion_matrix)
cat("Training Set Accuracy:", train_accuracy, "\n")
# Evaluate the optimized model on the test set
predicted_probs <- predict(optimized_model, newdata = test_data, type = "response")
predicted_classes <- ifelse(predicted_probs >= 0.5, "faible", "élevée")
# Confusion matrix and accuracy
confusion_matrix <- table(Predicted = predicted_classes, Actual = test_data$`qualité de production`)
accuracy <- mean(predicted_classes == test_data$`qualité de production`)
# Print results
print(confusion_matrix)
cat("Test Set Accuracy:", accuracy, "\n")
# Check multicollinearity in the optimized model
vif_values <- vif(optimized_model)
print(vif_values)
roc_curve <- roc(test_data$`qualité de production`, predicted_probs, levels = c("faible", "élevée"))
plot(roc_curve, main = "ROC Curve", col = "blue", lwd = 2)
cat("AUC:", auc(roc_curve), "\n")
coef(optimized_model)
```