# MÉMOIRE DE FIN DE FORMATION

MASTER 2 ECONOMIE : PARCOURS ECONOMETRICS, BIG DATA, STATISTICS

Calibration de modèles et visualisation de données pour explorer les avenirs plausibles de la zone des Niayes au Sénégal.

Crésus K. S. KOUNOUDJI

**08 Décembre 2021**

**Tuteurs de Stage :** Camille JAHEL & Etienne DELAY
**Enseignant référent :** Sébastien LAURENT

**Enseignement :** Aix-Marseille Université | Aix-Marseille School of Economics (AMSE)
**Accueil :** Cirad | Maison de la Télédétection, 500 Rue Jean François Breton, 34090 Montpellier

# ENGAGEMENT DE NON-PLAGIAT

Je soussigné, Kpessou Serges Crésus KOUNOUDJI
N° d'étudiant : 19024801

Déclare avoir pris connaissance de la charte relative à la lutte contre le plagiat de l'Université d'Aix-Marseille.

Je suis pleinement conscient que la copie intégrale sans citation ni référence de documents ou d'une partie de document publiés sous quelques formes que ce soit (publications internet, livres, rapports, etc...) est un plagiat et constitue une violation des droits d'auteur ainsi qu'une fraude caractérisée.

En conséquence, je m'engage à citer toutes les sources que j'ai utilisées pour produire et écrire ce document.

Fait le 27/20/2021 à Marseille.

Kpessou Serges Crésus KOUNOUDJI

AMU – AMSE 5-9 Boulevard Maurice Bourdet, CS 50498 13205 Marseille Cedex 1 France

# Remerciements

Je souhaite exprimer toute ma gratitude aux personnes qui ont contribué directement ou indirectement à la complétion de mon cursus et à la réalisation de ce rapport.

Notamment, à tout le personnel administratif et enseignant du Aix-Marseille School Of Economics (AMSE) que je représente par monsieur Sebastien LAURENT mon responsable de formation ; sans oublier mes camarades étudiants.

Mes remerciements s'adressent aussi à tout le personnel de la Maison de la Télé Détection à Montpelier où mon stage s'est déroulé. Aux différent(e)s stagiaires que j'ai pu y rencontrer ainsi qu'au personnel du CIRAD qui m'a accueilli, tout particulièrement Camille JAHEL et Etienne DELAY mes tuteurs de stage, je dis merci.

**Résumé**

Dans un contexte de changements globaux (démographiques, climatiques, etc.), les perspectives pour l'avenir de la zone des Niayes au Sénégal, face à l'avancée de l'urbanisation au détriment de l'agriculture, à la salinisation des nappes phréatiques et aux développements de zones minières et d'agro-industries, ne sont pas bonnes. Le projet Niayes 2040 mené par le Cirad vise, dans une démarche de prospective, à explorer les avenirs plausibles de la zone. Ce travail s'inscrit dans la suite de la première phase du projet qui a proposé un récit descriptif des scénarios plausibles pour la zone. Dans ce document, je développe mes démarches méthodologiques dans la calibration des modèles hydrologique et d'occupation du sol, utilisés pour représenter la dynamique de la zone des Niayes afin de prolonger les simulations dans les différents futurs hypothétiques et ainsi apporter une dimension empirique aux différents avenirs explorés. Dans un premier temps, j'ai effectué une analyse descriptive de la zone avec une typologie des exploitations et une analyse de la démographie interne à des fins de simulation. Ensuite, après une analyse de sensibilité pour réduire le nombre de paramètres à caler et pour optimiser les performances du modèle dans la calibration, j'ai procédé à la calibration et à la validation des modèles de simulation. Enfin, j'ai construit une représentation graphique de l'évolution de la piézométrie selon les différents scénarios.

*Mots clés* : Changements globaux, Niayes 2040, Cirad, Prospective, Scénarios plausibles, Simulation, Calibration, Validation des modèles, piézométrie...

**Abstract**

In a context of global changes (demographic, climatic, etc.), the outlook for the future of the Niayes zone in Senegal, given the advance of urbanization to the detriment of agriculture, the salinization of water tables, and the development of mining and agribusiness zones, is not promising. The Niayes 2040 project led by CIRAD aims to explore plausible futures for the area through a prospective approach. This work is a continuation of the first phase of the project, which proposed a descriptive narrative of plausible scenarios for the zone. In this document, I develop my methodological approaches to the calibration of the hydrological and land use models used to represent the dynamics of the Niayes area in order to extend the simulations into different hypothetical futures and thus bring an empirical dimension to the different futures explored. First, I conducted a descriptive analysis of the zone with a typology of farms and an analysis of the internal demography for simulation purposes. Then, after a sensitivity analysis to reduce the number of parameters to be calibrated to optimize the performance of the model in the calibration, I proceeded to the calibration and validation of the simulation models. Finally, I produce a graphical representation of the evolution of the piezometry according to the different scenarios.

*Keywords*: Global changes, Niayes 2040, Cirad, Foresight, Plausible scenarios, Simulation, Calibration, Model validation, piezometry...

# Sommaire

# Introduction

En Afrique subsaharienne, les économies nationales sont très largement dépendantes de l'agriculture. Au Sénégal en l'occurrence, l'agriculture représente d'après l'Agence Française de Développement [AFD21] 16% du produit intérieur brut (PIB) et fourni 70% des emplois. Ainsi, face au défi de la mondialisation et des changements globaux (climatiques et démographiques notamment), est-il légitime de vouloir anticiper l'avenir des territoires ruraux.

Ces questionnements sur le futur sont d'autant plus pertinents pour la zone des Niayes (une zone rurale du Sénégal qui s'étend sur la bande côtière entres les villes de Dakar et Saint-Louis) touchée de plein fouet par ces mutations (socio-économiques, climatiques, etc.). En effet, la zone des Niayes, qui fournit 60% des produits horticoles du Sénégal, est confrontée à une urbanisation croissante au détriment des terres agricoles, les activités minières et agro-industrielles, avec la nappe phréatique qui baisse de niveau et se salinise [al18a] [al19].

C'est dans ce contexte que le projet Niayes2040 conduit par le Centre de coopération internationale en recherche agronomique pour le développement (Cirad) et financé par l'AFD a vu le jour. L'objectif est, à travers les méthodes de prospectives, d'explorer et mesurer les avenirs plausibles de la zone des Niayes. Une première phase du projet, a permis à travers des activités participatives avec les experts [1] locaux suivant une méthodologie développée par Bourgeois R.[Rob12], d'identifier des scénarii, des narratifs plausibles avec les variables et les évènements pouvant faire penchant le future vers l'un ou l'autre de ces scénarii.

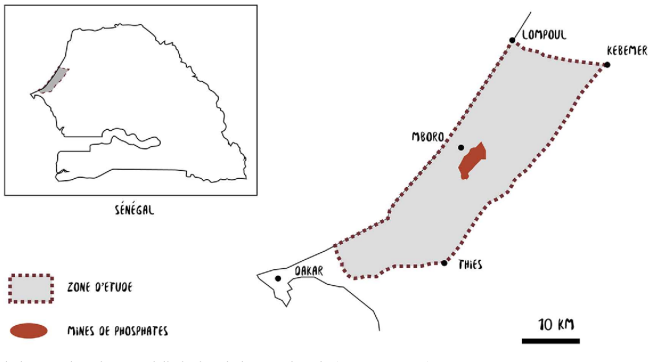


Figure 1 – carte de la zone des Niayes [al18a].

1. Les experts ici désignent des individus qui ne sont pas sélectionnés pour leurs fonctions ni pour représenter une organisation mais pour leurs connaissances et compétences personnelles [Rob12]

# 1 Cadre institutionnel

## 1.1 Présentation du Cirad

### Mission et Attributions du Cirad

Le Centre de coopération internationale en recherche agronomique pour le développement (Cirad) est un organisme français né de la fusion (en 1984) de plusieurs instituts techniques et de recherche structurés autour des filières tropicales [Cir21]. Pour reprendre les informations du site du Cirad, l'organisme français passe alors de l'expérimentation agricole dans un cadre d'économie coloniale à la production de connaissant et à l'appui technique aux nouveaux États indépendants, en agriculture, élevage et foresterie des pays tropicaux ainsi qu'aux partenariats de recherche pour le développement.

Avec la mondialisation et les changement globaux (climat, démographie, etc.) le Cirad, à travers ses productions de savoir scientifique, d'innovation et des formations, veut permettre d'atteindre les objectifs de développement durable. C'est dans cette dynamique qu'il s'est donné pour mission de : « contribuer à un monde plus durable et à la réalisation des objectifs de développement durable grâce à des systèmes agricoles et alimentaires qui nourrissent sainement les populations, qui rémunèrent décemment les productrices et les producteurs, résilients face aux changements globaux dont climatiques, tout en préservant la biodiversité et les ressources naturelles » [Cir21]. En pratique, le Cirad a pour mission de soutenir le développement agricole et rural des territoires d'outre-mer et de développer la coopération scientifique dans l'océan Indien et les Caraïbes.

Elle incarne cette mission dans sa stratégie, ses valeurs et l'orientation de ses recherches. En effet, le Cirad priorise dans sa stratégie la recherche « utile » sur six (06) thématiques prioritaire : la biodiversité, la transition agroécologique, les systèmes alimentaires durables, le changement climatique, les territoires comme levier de développement durable et inclusif, la santé (des plantes, des animaux et des écosystèmes). Cette mission est bien conduite d'ailleurs puisque le Cirad est aujourd'hui un référent scientifique et technique pour la plupart des filières agricoles tropicales. Avec son expertise, le Cirad qui est l'un des piliers de la capitale mondiale de la recherche agronomique, Montpellier, souligne le rôle pivot de l'agriculture et apporte une solution aux problématiques économiques, sociales, environnementales et sanitaires actuels. Cette expertise mise au service de tous, des producteurs aux politiques publiques.

### Structure organisationnelle du Cirad

Le Cirad engagé dans la mise en oeuvre d'une démarche scientifique et partenariale rigoureuse est présent sur tous les continents dans une cinquantaine de pays. Il mobilise les compétences de 1 650 salariés, dont 1 140 scientifiques, ainsi que d'un réseau mondial

d'environ 200 partenaires. Le Cirad est constitué de 29 unités de recherche (des unités mixtes de recherche - UMR, des unités propres de recherche - UPR, et une unité de services - US) réparties dans trois départements scientifiques : Systèmes biologiques (Bios), Performances des systèmes de production et de transformation tropicaux (Persyst) et Environnements et sociétés (ES) [Cir21].Le Cirad joue aussi un rôle de catalyseur de réseaux, en l'occurrence au sein des dispositifs de recherche et de formation en partenariat (dP), qu'il a créés en 2008 [Cir21].

## 1.2 Déroulement du stage

Au Cirad, le stage s'est déroulé sous la supervision de deux chercheurs du Cirad : Camille JAHEL, responsable de projet TETIS (Territoire Environnement télédictions Information Spatiale) et Etienne DELAY de l'UMR-SENS (Unité Mixte de Recherche - Savoirs, Environnement, Sociétés). La mission et le travail accompli durant le travail a été essentiellement de proposer et d'implémenter (automatiser sous logiciel R 4.0.5 en l'occurrence) des solutions statistiques et/ou des calculs/analyses numériques. Il a été aussi question de produire des sorties graphiques et des tableaux de valeurs liés aux différents calculs. Par ailleurs, le stage qui a eu lieu dans les locaux de la Maison de la télédétection à Montpellier s'est bien déroulé. Le cadre était idéal et l'environnement de travail bon. L'accueil chaleureux et la convivialité qui règne (avec le personnel de la maison de la télédétection, les chercheurs de différentes unités de recherches et leurs stagiaires etc.) favorisent un environnement de travail agréable. Mention spéciale à la cantine du Cirad, souvent excellente.

# 2 Cadre théorique de l'étude

## 2.1 Problématique

La pression démographique, l'avancement de l'urbanisation, la dégradation des ressources et les autres menacent qui pèsent sur la zone des Niayes ont conduit au déploiement de méthodes prospectives pour explorer et mesurer les avenirs plausibles de la zone. Pour y parvenir, les chercheurs du Cirad ont construit un modèle général de simulation pour reproduire la dynamique de la zone à partir des facteurs moteurs qui la caractérise.

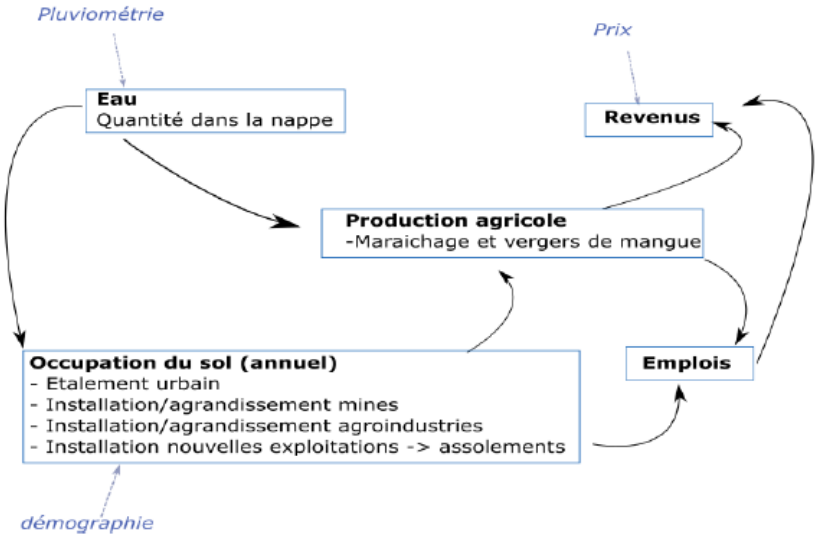


Figure 2 – structure du modèle générale [Rob21].

Il s'agira à partir de ce modèle de prédire les avenirs plausibles à travers ces grands agrégats en entrant des paramètres selon les scénarii. Mais avant toute prospective, le modèle est-il une bonne représentation des phénomènes réels étudiés ? Et même s'il reproduit bien les observations est-il adapté hors échantillon ? en d'autres termes, peut-on avoir confiance au modèle dans sa prédiction d'évènement qu'on ne connait pas ? etc. Les réponses aux questions précédentes sont indispensables pour transformer la première phase descriptive en mesure quantitative à travers le modèle de simulation. C'est le sujet de ce stage en l'occurrence sur les modèle Hydrologique (Eau nappe) et d'occupation du sol. Dans ce rapport nous allons décrire comment nous avons procédé pour y répondre et apporter des solutions quantitatives.

## 2.2 Intérêt de l'étude

En plus de sa contribution à la littérature, cette étude à une large portée économique, politique et sociale en générale. En effet, elle donne le moyen aux responsables politiques, aux différents acteurs sociaux concernés, de près ou de loin par l'avenir de cette zone, de visualiser le futur de la zone et de pouvoir agir à travers des points d'inflexion identifier en atelier, pour avoir le futur qui leur plairait.

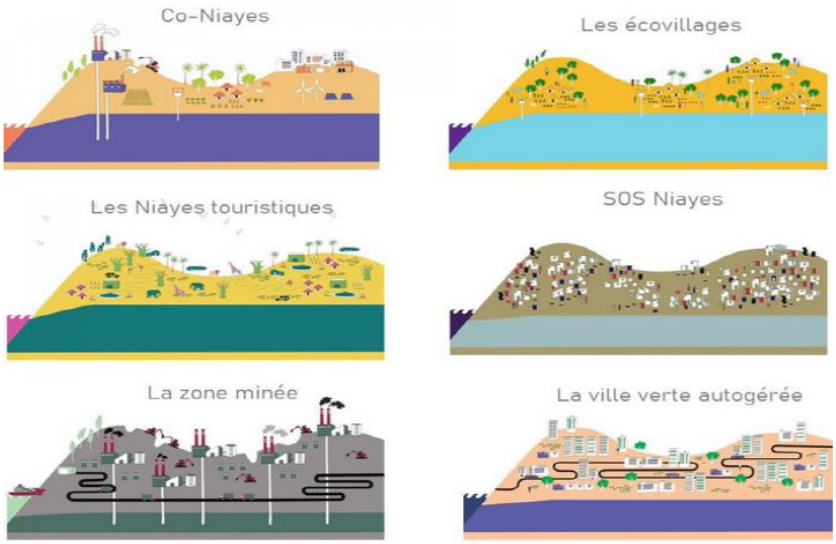


Figure 3 – illustration de différents scénario Niayes 2040 [Cir21].

## 2.3 Objectif de l'étude

L'objectif général ici c'est de calibrer les modèles hydrologiques et d'occupation du sol et d'irrigation et d'explorer graphiquement les scénarii.

Il s'agira en particulier de :

— **Faire une analyse descriptive des exploitations de la zone :** Pour dégager des profils d'exploitation afin de pourvoir les générer dans la simulation ; pour estimer la dynamique démographique à l'intérieur de la zone à des fins de projection démographique.

— **Faire des analyses de sensibilités :** Pour réduire le nombre de paramètre à calibrer et identifier les paramètres à estimer précisément.

— **Calibrer et valider les modèles de simulation :** Afin d'optimiser les performances des modèles dans la représentation des phénomènes réelles et assurer un bon pouvoir prédictif.

— **Fournir des représentations graphiques des simulations par scénarii**

## 2.4 Revue de littérature

### 2.4.1 La typologie des exploitations agricoles

Les méthodes de classification pour la production de typologies d'exploitation et/ou de ménage sont bien documentées dans la littérature. Citons quelques-unes des contributions sur ce sujet. Par exemple, Alvazez et al. [al18b] ont proposé une méthodologie pour générer et analyser des typologies d'exploitations agricoles. Dans ce papier, les auteurs ont construit, à partir des données d'exploitations agricoles de la zone orientale du Zambie, cinq (05) typologies différentes en utilisant l'Analyse des Composantes Principales (ACP) et la Classification Ascendante Hiérarchique (CAH). Chaque typologie est issue d'une méthode de classification et une sélection différente de variables. Après une analyse comparative des différents regroupements obtenus, les auteurs ont souligné l'effet déterminant de la méthode et des hypothèses de départ (qui conduisent le choix des variables de différenciation) sur la typologie obtenue. Ils préconisent un cadre d'analyse de la typologie moins subjective (dans le choix des variables par exemple) et basé sur les hypothèses des experts. Il s'agirait donc de choisir les variables discriminantes et une méthode en fonction des hypothèses formulés par les experts puis de consolider l'analyse sur la base de la connaissance des experts.

Dans le même ordre d'idée, Kuivanen et al. [al16] dans un papier publié dans le Journal of Rural Studies ont cherché à comparer les regroupements quantitatif (analyse statistique) et qualitatif (analyse qualitative, à partir à d'ateliers de groupe avec les acteurs locaux) d'exploitations agricoles du nord Ghana. On retient essentiellement de cette étude que l'analyse statistique réalisée avec ACP classe les ménages en six (06) groupes à partir de leur dotation en ressources (foncières) et de leurs objectifs de production tandis que l'analyse qualitative (participative) identifie elle cinq (05) groupes se distinguant par avec les caractéristiques structurelles des exploitations (taille de l'exploitation, revenu. . . ), le genre et l'Age ce qui donne plus de nuance dans le regroupement. Les auteurs concluent leurs analyses en combinant ces deux méthodes. En effet, d'après ce papier, l'analyse statistique a l'avantage de la reproductibilité et de l'objectivité mais elle est limitée dans son efficacité par les problèmes de collecte de données et les réalités locales. L'analyse qualitative quant à elle permet de contextualiser l'hétérogénéité des profils mais peut manquer de précision.

Dans la plupart des travaux consultés sur le sujet, l'ACP et/ou le CAH sont les méthodes les plus courantes pour constituer des typologies d'exploitation. Or ces méthodes ne permettent pas de prendre en compte des variables catégorielles ou mixtes. Kuentz-simonet et al. [al13] proposent une approche intéressante pour réaliser les typologies d'exploitations agricoles et pallier cette limite : l'algorithme ClustOfVar. Leur méthode se découpe en plusieurs étapes. Dans un premier temps, il s'agit de réaliser une réduction de dimension, de manière assez classique. Ensuite, un regroupement des variables est réalisé grâce à une CAH. On obtient alors des groupes de variables sur lesquels vont ensuite porter la classification. Ainsi, on réalise une Analyse factorielle des données mixtes (AFDM, qui permet d'utiliser aussi les

variables non numérique) sur chaque classe de variables pour obtenir une variable synthétique de chaque classe. Ces variables synthétiques constitueront les nouvelles variables dans le reste de l'analyse. La différence fondamentale avec la réduction classique de dimension réside dans la construction d'une variable synthétique qui ne sera construite qu'avec les variables d'une même classe. Ces variables synthétiques sont donc indépendantes par construction. Les variables synthétiques ainsi obtenues serviront de nouvelles variables pour la classification des observations

### 2.4.2 La calibration de modèle

Les modèles numériques ou de simulation, sont en général des représentations simplifiées d'un système réel complexe dans le but de les reproduire et/ou faire de la prospective. Ces modèles sont souvent paramétriques, c'est-à-dire contiennent des paramètres à estimer ou fixer, dont les valeurs déterminent l'état du système à travers les équations internes du modèle dont ils sont les paramètres d'entrée. Pour appliquer ce modèle, il faut être en mesure de contrôler l'incertitude dans les paramètres pour avoir une représentation satisfaisante du système selon des critères bien défini. Par exemple, le biais du modèle, c'est-à-dire à quel point les valeurs prédites par le modèle s'éloignent des valeurs observées, peut être un critère pour juger de la pertinence de la représentation. C'est là qu'intervient la notion de calibration. La calibration d'un modèle est le processus dans lequel on ajuste des paramètres d'un modèle de sorte à réduire l'incertitude et donc optimiser les performances du modèle [Mah18].

Il existe plusieurs méthodes de calibration. Selon le domaine d'étude l'approche peut différer (calage, optimisation, étalonnage, etc.), mais le principe reste le même. Une méthode de calibration assez classique, est la méthode dite de calage qui est une méthode graphique. Il s'agit, de tester différentes valeurs des paramètres jusqu'à obtenir une représentation graphique des simulations du modèle qui se rapprochent, se cale le plus possible au système réellement observé (la figure 15 illustre bien le principe). L'utilisateur peut aussi avoir comme critère de reproduire simplement les tendances dans les valeurs réelles observées. La figure 4 ci-dessous illustre la calibration d'un modèle paramétrique de pluie-débit en comparant les observations aux valeurs simulées. Par ailleurs, on peut aussi citer la méthode inverse qui est une méthode d'estimation statistique des paramètres et qui consiste à estimer la fonction $F(x)^{-1}$ inverse des observations $F(x) = y$. Le problème avec ces méthodes est qu'elles sont soit chronophage coûteuse soit non pertinente. En dehors du problème de dimensionalité pour les cas de nombreux paramètres (solutionner avec le choix limité de paramètres et de valeurs à partir des connaissances des expertes ou des tests de sensibilité par exemple) le coût que peut avoir un calage graphique est assez trivial.

En ce qui concerne la méthode d'estimation inverse, elle n'est pas toujours applicable. En effet, lorsqu'on ne dispose pas d'un modèle qu'on peut écrire comme fonction paramétrique « classique » (polynôme), [ce qui est notre cas puisque le modèle est un système d'équations dynamique avec des équations interdépendantes qui évoluent simultanément à chaque itéra-

tion dans le temps]„ on ne peut pas directement utiliser cette méthode. Bien qu'il existe des méthodes de construction de modèle de surface (on pense notamment aux méthodes développées par dans le projet GitLab Lagun lien dans les citations), nous nous sommes focalisés sur la méthode la plus utilisée dans la littérature : la méthode d'optimisation. Elle consiste à définir une fonction objective à optimiser (en l'occurrence à minimiser dans le cas d'une fonction de perte).

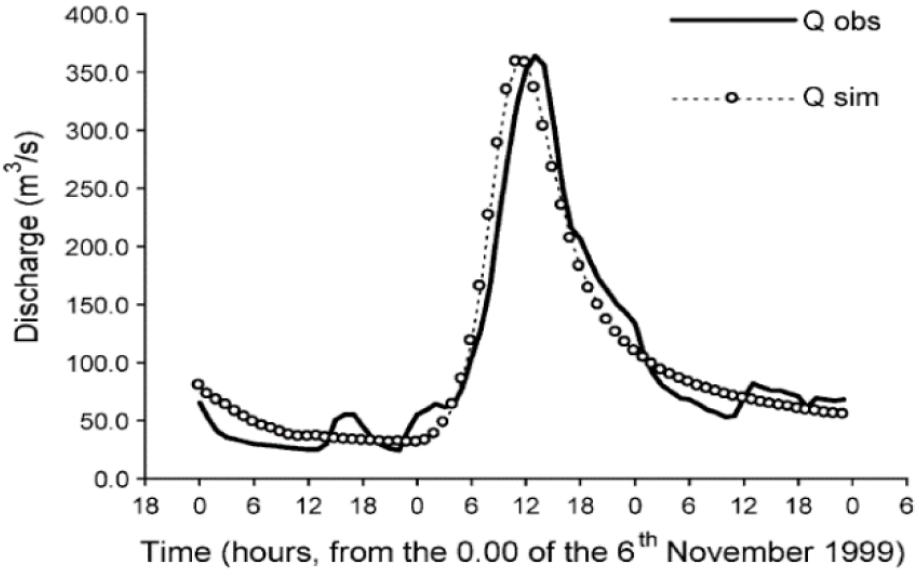


Figure 4 – Comparaisons des simulations d'un modèle aux valeurs réelles observées [Mon21].

### 2.4.3 L'analyse de sensibilité

Les modèles de simulations (en l'occurrence en hydrologie) sont généralement des modèles paramétriques. Ces paramètres peuvent être déterminés par des équations internes au modèle, être fixes, dépendre du temps et/ou de l'état du modèle [Mon21]. Les modèles qu'on étudie ici sont eux aussi paramétriques. Il faut donc optimiser les paramètres qu'on introduit de manière à trouver la combinaison de paramètres optimisant les performances du modèle. C'est le rôle de la calibration. Et pour ça, on calcule un indicateur pour chaque simulation issue de combinaison de valeurs du jeu de paramètres. On voit clairement qu'avec un grand nombre de valeurs possible et/ou un grand nombre de paramètres à calibrer, le nombre de combinaison explose. C'est l'équivalent de la malédiction de la dimensionalité. C'est là qu'intervient l'analyse de sensibilité. L'idée est de mesurer ou d'évaluer l'effet de l'incertitude dans le paramètre d'entrée sur l'incertitude de la sortie, pour réduire les paramètres à calibrer et donc les simulations à lancer. Ainsi, les paramètres dont l'effet est grand seront estimés finement avec la calibration tandis que les autres pourront être fixés ou estimés avec beaucoup moins de valeurs différentes sans risquer de perdre en précision.

La littérature sur l'analyse de sensibilité est abondante. Notamment saltelli [al06] qui développent la méthodologie théorique et empirique sur le sujet. Ici, nous allons nous con-

centrer sur l'une des méthodes que nous avons explorées : la méthode de Morris [Mor91] [Fra]. Elle est bien documentée, simple à comprendre et intuitive. Pour citer [Bui14], la méthode des effets élémentaires de Morris ou méthode de screening est une méthode d'analyse de sensibilité globale dite qualitative (classification des méthodes d'analyse de sensibilité cf. annexe A). Elle est dite qualitative parce que contrairement aux méthodes dites quantitatives (la méthode de Sobol, décomposition de la variance, etc.), elle ne permet pas de quantifier tous les types d'effet mais donne une approximation qui permet de classer les entrées par ordre d'importance [Cer17]. Même si c'est une sorte de One At Time (OAT) ''généralisée", elle présente les avantages des méthodes d'analyse globale. Elle permet ainsi d'explorer le domaine de définition des facteurs (contrairement aux méthodes d'analyse locale comme le OAT, qui font une analyse autour d'un point) et de tenir compte des non-linéarités et/ou interactions. Pour citer Ceria [Cer17], la méthode de Morris est souvent considérée comme une analyse préliminaire. En effet, elle nécessite un faible nombre de simulations par rapport aux autres méthodes et permet de déterminer les entrées influentes.

### 2.4.4 Le choix de la métrique de calibration

Dans la littérature sur l'évaluation de modèles, plusieurs types d'indicateurs permettent d'évaluer les similarités entre des données simulées et des données observées. D'après Liemohn et al. [al21], il existe plusieurs catégories de métriques en fonction de ce qu'on cherche à évaluer. On a entre autres : le biais, l'exactitude et la précision. Le Biais est souvent mesuré en prenant l'écart moyen entre les valeurs observées et les valeurs simulées. Cette mesure n'est pas très populaire parce qu'elle pose un problème : on peut avoir de grand écart entre les valeurs prédites par le modèle et les valeurs observées mais la moyenne sera nulle (les écarts négatifs et positifs se compensent dans la somme) ce qui pose un problème de lecture ou d'interprétabilité de la mesure. Pour régler ce problème on peut prendre la valeur absolue des écarts ; c'est l'Ecart Absolue Moyen (MAE : Mean Absolute Error). Cet indicateur mesure l'amplitude moyenne des erreurs du modèle quelle que soit la direction. L'inconvénient de cet indicateur est que la valeur absolue n'est pas toujours souhaitée mathématiquement. Chai et Draxler [R R14] donnent l'exemple de l'analyse de sensibilité que la valeur absolue rend impossible pour certains paramètres. Ce qui est problématique puisque l'analyse de sensibilité fait partie de notre démarche méthodologique de calibration et d'évaluation du modèle.

La mesure la plus repondue, notamment en hydrologie, juste après l'Erreur Quadratique Moyenne (MSE : Mean Squared Error) est sa Racine carrée le RMSE (Root- Mean Squared Error). Dans leurs papiers, Willmott et Matsuura [K M05] puis Willmott, Matsuura et Robeson [C W09] défendent l'idée que le RMSE, une mesure de l'erreur largement répandue dans la littérature sur l'environnement et le climat est ambiguë, inappropriée et souvent mal interprétée. En effet, le RMSE est fonction de trois (03) composantes de l'erreur (le biais, la variance de l'erreur, l'erreur résiduelle) au lieu d'une (le biais) ce qui peut entrainer un biais dans l'interprétation. Ensuite, toujours pour citer les auteurs Willmott et Matsuura [K M05],

la dépendance du RMSE à plusieurs composantes de l'erreur rends sa signification confuse. Enfin, en citant Paul Mielke, les auteurs fondent le problème d'interprétabilité de la mesure sur le fait que les statistiques basées sur la somme des carrés ne respecteraient pas l'inégalité triangulaire. Les auteurs préconisent fortement l'utilisation du MAE qui serait une mesure plus naturelle.

A tout cela, Chai et Draxler [R R14] déplorent que beaucoup de papiers citent les précédents pour justifier de la supériorité du MEA sur le RMSE et donc son utilisation. Même si certains problèmes soulevés sont fondés, Chai et Draxler [R R14] ne jugent pas pertinent la hiérarchisation des indicateurs. Le RMSE peut s'avérer être plus pertinent que le MEA dans certaines configurations et souvent la combinaison des métriques peut être nécessaire pour apprécier un modèle. Les auteurs ont montré que le RMSE n'était pas ambiguë contrairement à ce qui était affirmé, qu'il était approprié lorsque l'erreur est censée être normale et qu'il respecte l'inégalité triangulaire. Par ailleurs, il existe beaucoup d'autres indicateurs pour la plupart calculés à partir du MSE. En hydrologie en l'occurrence on a le coefficient d'efficacité de Nash-Sutcliffe (NSE) qui a les mêmes propriétés que le RMSE mais qui a l'avantage d'être normalisé donc de permettre de comparer des modèles complètement différents. Mais nous allons nous concentrer sur le RMSE puisque la littérature empirique en hydrologie donne une valeur seuil du RMSE à laquelle nous pouvons nous référer pour évaluer notre RMSE. Le NSE est évalué différemment du RMSE. Plus ce dernier est petit, plus le modèle est précis à l'inverse du NSE qui varie entre moins l'infini et 1 et qui doit être le plus proche de 1. Dans la littérature empirique hydrologique le RMSE doit-être inférieur ou égal 0.5, or dans la littérature sur le modèles pluie-débit un NSE inférieur ou égal à 0.5 témoigne d'un faible pouvoir de prédiction [Mon21].

**Coefficient d'efficacité de Nash-Sutcliffe ($NSE$):**

$$\boldsymbol{NSE = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}} \quad \text{avec} \quad \begin{cases} y_i = \text{les observations, } i = 1 \dots N \\ \hat{y}_i = \text{les simulations, } i = 1 \dots N \\ \bar{y} = \text{moyenne des observations} \end{cases}$$

Enfin, en calibration, les critères de performance de modèle sont vus comme des fonctions objectives ou fonctions de perte qui servent à mapper plusieurs variables sur une seule valeur qui représente intuitivement la qualité d'une caractéristique simulée [Mon21]. Dans cet ordre d'idée, le RMSE comme fonction objective de calibration de modèle se justifie pleinement avec ses propriétés statistiques (non biaisé et cohérent) et pratiques (pénalise les erreurs importantes). Le problème du RMSE c'est que les hypothèses sous-jacentes à ces propriétés statistiques, ne sont que rarement respectées en hydrologie par exemple. De plus en pratique, la pénalité, c'est-à-dire le poids accordé aux erreurs importantes par le RMSE, peut le rendre moins approprié pour servir de fonction objective par exemple dans la calibration de faible

débit dans un modèle pluie-débit. Cependant l'auteur souligne que les propriétés statistiques (cohérence et absence de biais, qu'on a d'ailleurs rarement avec les modèles hydrologiques) d'un indicateur ne sont pas nécessaires pour la calibration. En effet, ces propriétés permettent d'avoir les paramètres, qui donnent, dans l'ensemble, la meilleure simulation. Mais selon l'objectif (meilleure simulation locale par exemple), ces propriétés ne sont pas prioritaires. On préconise la créativité dans la construction de la métrique en fonction, entre autres, des objectifs d'applications et de conception [Mon21].

—

**La décomposition de l'erreur**

$$\text{MSE} = \text{Biais}^2 + \boldsymbol{\sigma_\epsilon^2} + \boldsymbol{\epsilon}$$

—

**Le Biais**

$$\text{Biais} = \frac{\boldsymbol{1}}{\boldsymbol{N}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} (\boldsymbol{y_i} - \boldsymbol{\hat{y}_i})$$

—

**L'écart absolu moyen**

$$\text{MAE} = \frac{\boldsymbol{1}}{\boldsymbol{N}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} |\boldsymbol{y_i} - \boldsymbol{\hat{y}_i}|$$

avec

- $y_i$ = les observations, $i = 1 \dots N$
- $\hat{y}_i$ = les simulations, $i = 1 \dots N$
- $\epsilon$ = erreur résiduelle
- $\sigma_\epsilon^2$ = variance de l'erreur résiduelle

—

**L'erreur quadratique moyen**

$$\text{MSE} = \frac{\boldsymbol{1}}{\boldsymbol{N}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} (\boldsymbol{y_i} - \boldsymbol{\hat{y}_i})^{\boldsymbol{2}}$$

—

**La racine de l'erreur quadratique moyen**

$$\text{RMSE} = \sqrt{\frac{\boldsymbol{1}}{\boldsymbol{N}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} (\boldsymbol{y_i} - \boldsymbol{\hat{y}_i})^{\boldsymbol{2}}}$$

A partir de la littérature existante, le RMSE nous semble un bon compromis entre biais et précision.

# 3 Cadre méthodologique de l'étude

## 3.1 Typologie des exploitations agricoles

Notre démarche méthodologique est fondée sur les préconisations de Alvarez et al. [al18b] et Kuivanen et Al. [al16]. D'abord il s'agit d'identifier dans la base de données les variables clés qui permettent de séparer statistiquement les exploitations. Ensuite, les partitionnements seront confrontés aux connaissances des experts notamment la concordance avec les enquêtes de terrains réalisées par les chercheurs. Nous avons mis en oeuvre différentes méthodes de partitionnement et avons ainsi obtenus différentes typologies d'exploitations pour la zone des Niayes. Nous avons ensuite présenté ces typologies aux chercheurs qui ont retenus celle qui était la plus en cohérence avec leurs connaissances du terrain. La typologie retenue était celle issue de la méthode ClustOfVar proposée par Kuentz-Simonet et al. [al13]. Nous avons donc choisi de la présenter dans cette partie. Les résultats des autres algorithmes sont mis en annexe. Le tableau suivant présente les algorithmes retenus pour obtenir une typologie des exploitations.

Table 1 – Algorithmes implémentés pour la typologie des exploitations

| **Algorithmes de réduction de dimensions** | **Algorithme de classification** |
|---|---|
| Classification de variables (Kmeans) et construction de variables synthétiques (AFDM) | Kmeans |
| Analyse Factorielle des Données Mixtes (AFDM) | CAH (Kmeans consolidation) |

## 3.2 La métrique : le cas particulier du modèle hydrologique

Dans notre cas, comme indiqué dans la sous-section 2.4.4 de la revue de littérature, sur le choix de l'indicateur, nous avons choisi comme fonction de perte un indicateur de performance de modèle largement utilisé dans la littérature, le RMSE.

**La calibration :**

$$\boldsymbol{\theta}_{\text{optimal}} = \text{argmin}\left(\|(\hat{\boldsymbol{y}} - \boldsymbol{y})\|\right) \quad \text{avec} \quad \begin{cases} \theta = \text{les paramètres du modèle} \\ \hat{y} = \text{les sorties de simulations du modèle} \\ y = \text{les valeurs de mesures observées} \end{cases}$$

$$\text{où } \|(\hat{\boldsymbol{y}} - \boldsymbol{y})\| = \text{RMSE}_{\text{Niayes}}$$

Le principal problème qui s'est posé à nous pour le calcul de notre RMSE, est le biais spatial du fait d'une sur-représentation de nos données observées dans une partie de la zone. En effet, les stations ne sont pas réparties de manière homogène spatialement, ce qui introduit un biais spatial non négligeable (figure 5), le RMSE final surreprésenterait les stations groupées au Nord- Ouest (entourées en rouge).

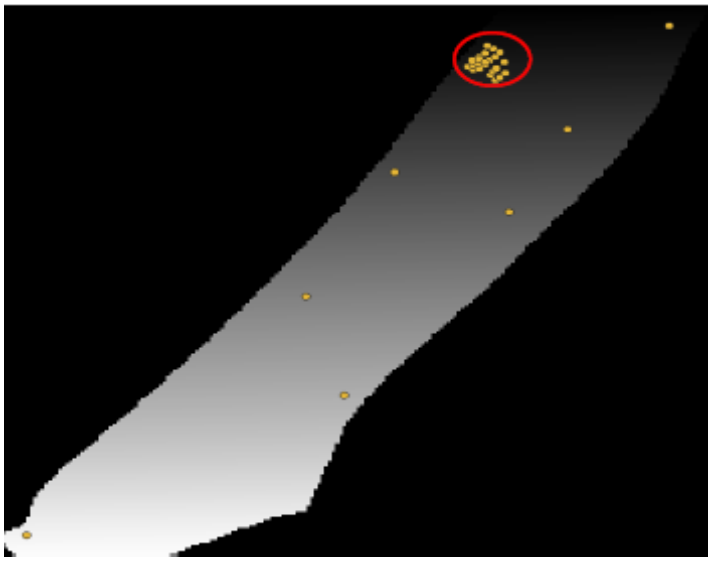

Figure 5 – Visualisation de la répartition des stations sur la zone d'étude.

Nos réflexions, nous ont mené à proposer plusieurs méthodes de calcul et/ou plusieurs indicateurs de performance différents pour au final retenir le calcul suivant :

**Le** $RMSE_{Niayes}$ :

$$\boldsymbol{RMSE_{Niayes}} = \frac{\sum_{i=1}^{n_1} \boldsymbol{RMSE_i} + \frac{1}{n_2} \sum_{j=1}^{n_2} \boldsymbol{RMSE_j}}{\boldsymbol{n_1 + 1}}$$

$$\text{où } \boldsymbol{RMSE} = \sqrt{\frac{\mathbf{1}}{\boldsymbol{N}} \sum_{k=1}^{N} (\boldsymbol{y_k} - \boldsymbol{\hat{y}_k})^2}$$

- $\hat{y}$ : les sorties de simulations du modèle
- $y$ : les valeurs de mesures observées
- $i = 1, \ldots, n_1$ les stations de la zone minière DGPZ
- $j = 1, \ldots, n_2$ les autres stations de la zone d'étude

## 3.3 Analyse de sensibilité

L'analyse s'est basée sur un plan d'expérience, le plan de Morris. Ce plan est constitué de plusieurs plan OAT avec les différents niveaux des facteurs. Morris [Mor91] recommande de discrétiser les domaines (intervalles de valeurs) des facteurs en cinq (05) niveaux de sorte que pour un intervalle de valeur $I$ on ait :

**Étape 0 : Discrétisation des valeurs des paramètres**

$$\boldsymbol{I = [a, b]}$$
$$\boldsymbol{\delta = \frac{b-a}{p-1}}$$
$$\boldsymbol{q_i = [a, a+\delta, a+2\delta, a+3\delta, b]}$$

avec
- $q_i$ : les valeurs d'entrée du paramètre $X_i$
- $\delta$ : le choc élémentaire
- $I$ : l'intervalle de valeurs du paramètre $X_i$
- $p$ : le nombre de niveaux (i.e valeurs de $q_i$)

N'ayant pas le même nombre de niveaux (nombre de valeurs dans l'ensemble des valeurs) pour nos paramètres, $\delta$ ici ne sera pas fixe. La méthode d'échantillonnage utilisée est celle de l'échantillonnage radiale itérée (plus performante que la méthode par les chemins [Fra]).

— **Application de la méthode de Morris**

— **Étape 1 : Choix du plan de départ $\boldsymbol{\theta_0}$**

$$\mathcal{F}(\boldsymbol{\theta_0} = \mathbf{x}_{1_0}, \mathbf{x}_{2_0}, \ldots, \mathbf{x}_{i_0}, \ldots, \mathbf{x}_{n_0}) = \text{RMSE}_0$$

— **Étape 2 : Choix aléatoire d'un $i^{ème}$ paramètre à choquer avec $\pm\delta$**

$$\mathcal{F}(\boldsymbol{\theta_1} = \mathbf{x}_{1_1}, \mathbf{x}_{2_1}, \ldots, \mathbf{x}_{i_1}, \ldots, \mathbf{x}_{n_0}) = \text{RMSE}_1 \quad \text{avec } \mathbf{x}_{i_1} = \mathbf{x}_{i_0} \pm \delta_i$$

— **Étape 3 : Calcul du premier effet élémentaire du $i^{ème}$ paramètre**

$$\text{EE}_{1_{x_i}} = (\text{RMSE}_1 - \text{RMSE}_0) \pm \delta_i$$

- $\delta_i$ : est le choc élémentaire du paramètre $x_i$
- $\theta_j$ : combinaison des $j^{ème}$ valeurs des paramètres d'entrée $x_i$ de l'ensemble $q_i$
- $\text{RMSE}_j$ : RMSE calculé pour $\theta_j$ comme paramètres d'entrée du modèle
- $\text{EE}_{jx_i}$ : $j^{ème}$ effet élémentaire du $i^{ème}$ paramètre d'entrée du modèle

Dans un premier temps, on choisit aléatoirement un plan de départ avec des valeurs $q_{i_0}$ (aléatoirement triées dans notre plan) pour les niveaux de chaque paramètre i. Ensuite on lance la simulation pour avoir la réponse correspondante à cette combinaison de facteurs (la réponse ici c'est le $RMSE_0$). La seconde étape consiste à choisir aléatoirement un paramètre parmi les $i$ paramètres, celui qu'on va choquer. C'est à dire aléatoirement augmenter ou diminuer d'un pas $\delta$ en gardant les autres paramètres à l'état initial. Après cela, on lance à nouveau la simulation avec la nouvelle combinaison ($RMS_1$). A l'étape 3, on calcule l'effet élémentaire $EE_{1_xi}$ pour le $1^{er}$ « choc » du paramètre $x_i$ . On répète toutes les étapes précédentes jusqu'à ce que le plan soit complet. Le plan est complet lorsque tous les facteurs ont été choqués une fois. Toutes les précédentes étapes sont répétées $R$ fois jusqu'à obtenir le nombre d'effets élémentaires voulu. On a donc un $R - échantillons$ pour chaque effet

éléments avec $N = R * (p + 1)$ expériences.

— **Les mesures de sensibilités de la méthode de Morris**

— **La moyenne des effets élémentaires absolus :**

$$\boldsymbol{\mu^* = \mathbb{E}(|EE_i|)}, \text{ c'est une mesure de l'importance de l'effet}$$

— **La moyenne des effets élémentaires :**

$$\boldsymbol{\mu = \mathbb{E}(EE_i)}, \text{ elle renseigne principalement sur le sens de l'effet}$$

— **L'écart type des effets élémentaires :**

$\boldsymbol{\sigma = \sigma(|EE_i|)}$, c'est une mesure des effets non linéaires et/ou des interactions.

$$\begin{cases} \text{EE}_i = \text{effet élémentaire du paramètre } x_i \\ \mu^* = \text{moyenne des effets élémentaires absolus} \\ \mu = \text{moyenne des effets élémentaires} \\ \sigma = \text{écart type des effets élémentaires} \end{cases}$$

— **Règles de Décision :**

Table 2 – Règles de décision de la méthode de Morris [Ioo]

| | $\sigma$ **Faible** | $\sigma$ **Élevé** |
|---|---|---|
| $\mu^*$ **Faible** | Facteur négligeable | Facteur influent, effet non monotone et/ou interactions |
| $\mu^*$ **Élevé** | Facteur influent, effet linéaire | Facteur influent, effet non linéaire et/ou interactions |

Le problème majeur que peut poser cette méthode c'est le coût en temps de calculs. En effet, pour des modèles complexes comme ceux qu'on étudie ici, et dont l'exécution peut prendre énormément de temps, créer plusieurs plans OAT et faire appel au modèle pour chaque combinaison de paramètres peut vite être très coûteux et contreproductif. Contreproductif puisque l'idée du test de sensibilité est de réduire la charge de calcul en triant les paramètres importants à bien explorer pour la suite des analyses. Pour cela, on a généralement recours aux « modèles de surface » (surrogate models). Ce sont des versions simplifiées des modèles initiaux qui ont vocation à reproduire (en déduisant à partir des données) les relations entre les paramètres d'entrée et les sorties dans une fonction. Ces modèles pourront ensuite être utilisés en lieu et place du vrai modèle pour les analyses d'exploration des paramètres etc.

Dans notre cas, à partir des connaissances des hydrologues, certaines valeurs ont préalablement été définies pour les paramètres et les simulations ont été réalisées pour des combinaisons de ces valeurs. Il s'agira donc de calculer l'indicateur de performance (RMSE) pour chaque simulation et d'utiliser ces valeurs déjà existantes comme plan de Morris et le RMSE comme réponse du modèle. C'est la raison pour laquelle le $\sigma$ n'est pas fixé pour nous ; il sera calculé directement dans le code et on n'aura pas non plus à discrétiser l'ensemble des valeurs des paramètres.

## 3.4 La validation ou vérification

La littérature sur les méthodes de validation en général (Leave-one-out, Validation croisée, etc.) et sur la validation de modèle calé en particulier est abondante. Le principe est souvent le même. Il s'agit de séparer les données en deux parties (ou plus) dont une servira à calibrer le modèle et l'autre à la vérification. Notre démarche méthodologique elle, est principalement fondée sur les stratégies de validation des modèles hydrologiques et de qualité de l'eau recommandée par Daggupati et al. [al15].

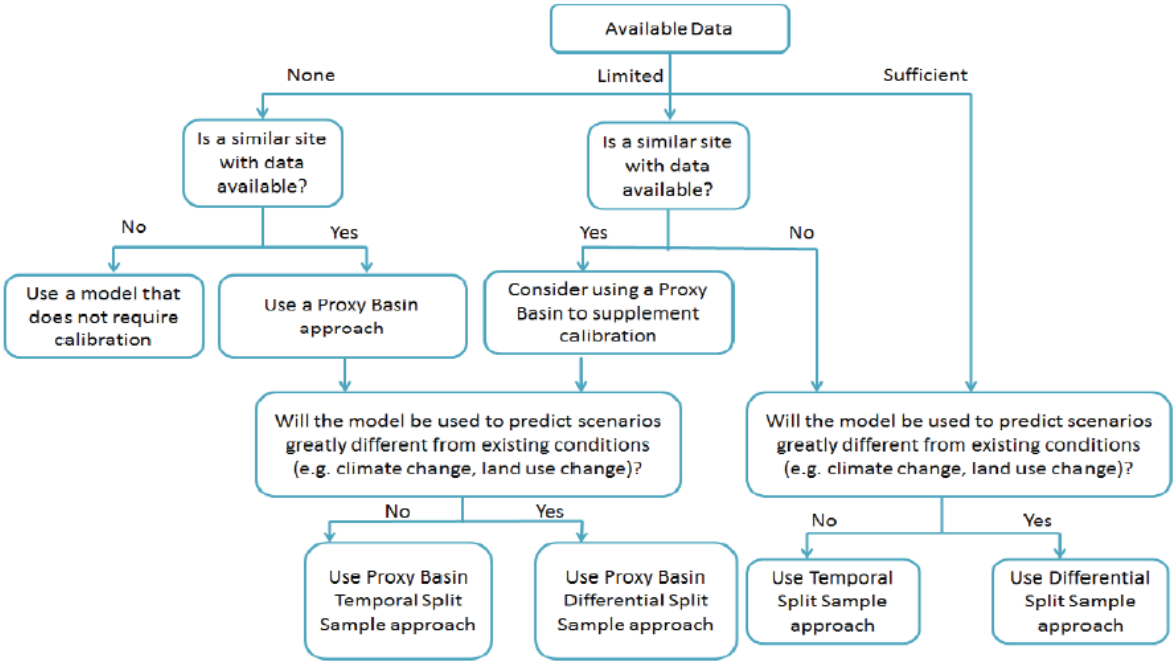


Figure 6 – Approches de répartition des données spatio-temporelles [al15].

La figure 6 ci-dessus, résume différents cas de figures et les méthodes de séparations de données adaptées pour la validation. Compte tenu de nos objectifs et des données dont nous disposons, nous avons utilisé la méthode de séparation temporelle (Temporal Split-Sampler). Le Temporal Split-Sampler, qui est la méthode la plus largement utilisée d'après les auteurs, consiste à subdiviser pour chaque position, les données de simulations et d'observation en deux parties. Une partie pour la calibration (pour nous 2008-2012) et une partie pour la vérification (2013-2016).

# 4 Résultats et interprétation

## 4.1 Analyse exploratoire des données

Toutes les analyses présentées dans cette section sont réalisées sous le logiciel R 4.0.5 à partir des données des bases PAPA et RGPHAE-2013. Les scripts sont consultables sur le dédié *GitHub repository* dont des captures sont en annexe **??**.

### 4.1.1 Présentation et sources des données

Dans ce chapitre, il s'agira principalement de décrire la zone d'étude et certaines de ses dynamiques à l'aide d'analyses statistiques à partir des données dont nous disposons. Ce travail de stage s'est appuyé sur l'analyse de deux jeux de données, l'un concernant les exploitations agricoles de la zone des Niayes, et l'autres des projections démographiques pour les différents arrondissements de cette zone [3].

Table 3 – Description des bases de données

| **Bases** | **Descriptions** | **Sources** |
|---|---|---|
| PAPA | La base PAPA est issue d'une enquête menée auprès des producteurs horticoles des Niayes entre 2015 et 2018. Elle est composée de 26 modules, couvrant des données socio-démographiques de l'exploitation, des informations sur les assolements, les itinéraires techniques, les investissements et les ventes. | Projet d'Appui aux Politiques Agricoles (PAPA) |
| RGPHAE 2013 | Projections démographiques sénégalaises entre 2013 et 2025, à l'échelle des arrondissements, réalisées à partir du Recensement Général de la Population et de l'Habitat, de l'Agriculture et de l'Élevage (RGPHAE) de 2013. | Agence nationale de la statistique et de la démographie (ANSD Sénégal) |

### 4.1.2 Analyses descriptives

La base de données utilisée dans cette sous-section, est extraite de la base PAPA. Elle est constituée d'individus statistiques (les lignes de la table) qui sont les producteurs représentés par un identifiant unique. En colonne de la table de données, nous avons les informations sur l'exploitation et le ménage (la taille de l'exploitation, les productions par culture, les méthodes d'irrigation, la taille du ménage etc.). Cette table de données contient 402 lignes et 119 colonnes.

Avant toute analyse statistique, les données ont été nettoyées. Notamment, une élimination raisonnée de variables de la base, non pertinentes pour nos analyses. Une élimination des variables redondantes, des colonnes qui ne varient pas, et des lignes sans valeur pour les données sur l'eau (variable d'intérêt pour notre travail). Ce qui a conduit à réduire notre base de données à 264 lignes et 27 colonnes. Dans cette analyse, nous nous sommes intéressés aux relations simples qu'on pourrait mettre en évidence entre certaines variables clés. La figure 7 présente les relations linéaires des variables deux à deux. On observe peu de relations linéaires claires, mise à part entre la taille du ménage et le nombre de personnes actives dans le ménage. Certains nuages de points prennent une forme qui suggère une relation linéaire : la surface cultivée et les intrants, ou les intrants et la taille de ménage.

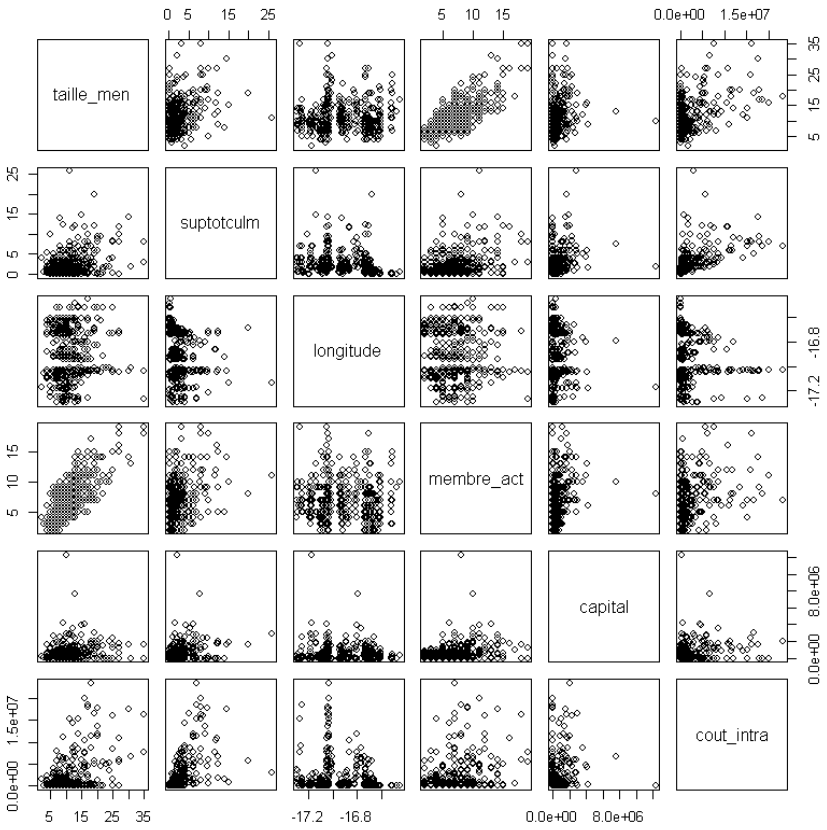


Figure 7 – nuage de points.

Pour tenter de faire ressortir des tendances, nous avons repris ces analyses considérant des groupes d'exploitation en fonction de leurs surfaces et localité. En effet, les Niayes étant positionnées le long de la bande côtière entre Dakar et Saint Louis, nous avons voulu tester l'hypothèse d'une évolution des exploitations en fonction de leur localisation (et notamment de leur distance à Dakar). Les figures 8 et 9 ci-dessous montrent, que la tranche de surface n'affecte pas la distribution de la taille des ménages (respectivement du capital), contrairement à la localité. La figure 9 montre en effet que les profils de taille de ménage sont différents en fonction de la commune. Pour le capital, le centrage reste plus ou moins le même, mais

de manière moins évidente.

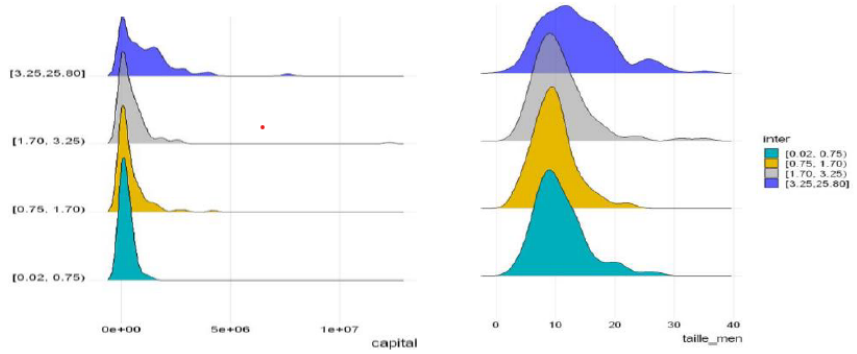


Figure 8 – Distribution de la taille des ménages et du capital par groupe de surface d'exploitation.

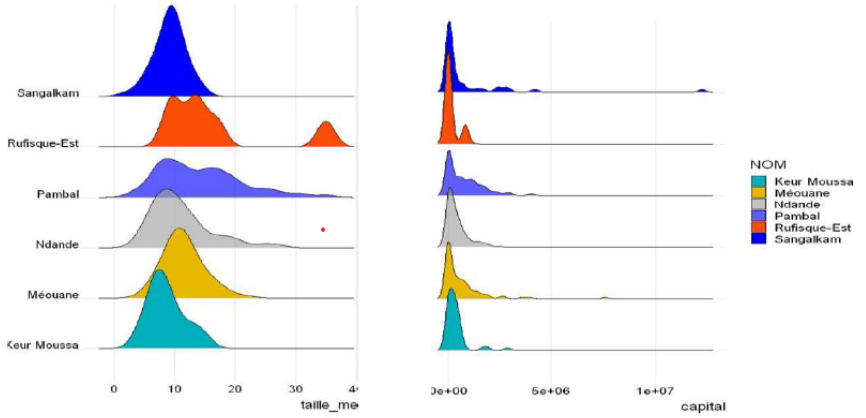


Figure 9 – Distribution de la taille des ménages et du capital par commune.

### 4.1.3 Statistiques démographiques

Dans les scénarii plausibles de la zone des Niayes, la démographie est l'un des premiers et principaux facteurs influençant les dynamiques du territoire. Il était donc important de pouvoir réaliser des projections démographiques couvrant la période entre 2006 et 2040. Pour cela, il a fallu estimer le taux de croissance démographique de la zone d'étude à partir des projections démographiques de la base RGPHAE-2013 (table de données chronologique de 2013 à 2025 pour chacune des 46 départements du Sénégal, leurs Arrondissements et communes). Cela permettra ensuite de refaire des projections démographiques jusqu'en 2040. Pour estimer le taux moyen à partir des données, nous avons calculé le taux de croissance suivant les deux hypothèses de croissance les plus communes. Pour calculer le taux de croissance r, on part des données disponibles et on en déduisant le taux de croissance r de la formule ci-dessous. D'abord, une hypothèse de croissance de la population avec la loi exponentielle. Ensuite, suivant une hypothèse de croissance géométrique. Enfin, nous avons effectué une projection à partir des données démographiques de la première année et retenu le taux estimé

qui correspondait le mieux aux projections de l'ANSD. Notamment le taux de croissance sous l'hypothèse de croissance géométrique. Les formules de calculs sont les suivantes.

— **Hypothèse de croissance exponentielle**

$$P_{t_n} = P_{t_0} \times e^{(r \times d)}$$

— **Hypothèse de croissance géométrique**

$$P_{t_n} = P_{t_0} \times (1 + r)^d$$

Où :

$P_{t_n}$ = est la population à la date, $t_n$

$P_{t_0}$ = est la population à la date initiale, $t_0$

$r$ = est le taux de croissance annuel moyen

$d$ = est la distance (en nombre d'années) entre la date initiale$t_0$ et la date $t_n$

Nous avons ainsi obtenu un taux de croissance de 0.0267, et pu réaliser des projections démographiques par arrondissement, qui ont été fournies à l'équipe de modélisation du Cirad pour être intégrées au modèle Niayes2040.

### 4.1.4 Typologie des exploitations agricoles

— **Transformation des données :**

Dans un premier temps, nous avons cherché à identifier d'éventuels individus très atypique dont il faudrait tenir compte pour avoir une analyse robuste. Disposant d'une base de données avec beaucoup de variables différentes, la visualisation du Positionnement multidimensionnel (MDS) des observations apparait comme une méthode pertinente de détection au moins graphique d'outliers. Le Multidimensional Scaling (MDS) est une méthode factorielle de réduction de dimension. Elle part d'une métrique (en l'occurrence une matrice de distance de la table de données) pour trouver dans un espace euclidien de moindre dimension (en général 2) qui admette cette métrique comme matrice de distance. Si la métrique d'entrée est une matrice de distance euclidienne de la table de données, le MDS revient à un ACP classique sur la table de données d'origine. Dans notre cas, disposant de données mixtes, on utilise une matrice de distance de Gower [Gow71] tirée de l'indice du même nom, qui mesure la similarité entre les individus en appliquant différentes méthodes selon le type de variables. En général, la distance euclidienne (ou de Manhattan) pour les variables quantitatives et la distance $\Phi_2$ (variante de la distance du $\chi^2$ qui est celle utilisée dans l'Analyse des Correspondance Multiple ACM) pour les variables qualitatives. Le résultat de cette opération est présenté dans la figure 10 ci-dessous. Sur la figure, aucune des observations ne se distingue franchement. On n'a pas de raison particulière de traiter des individus outliers dans la base.

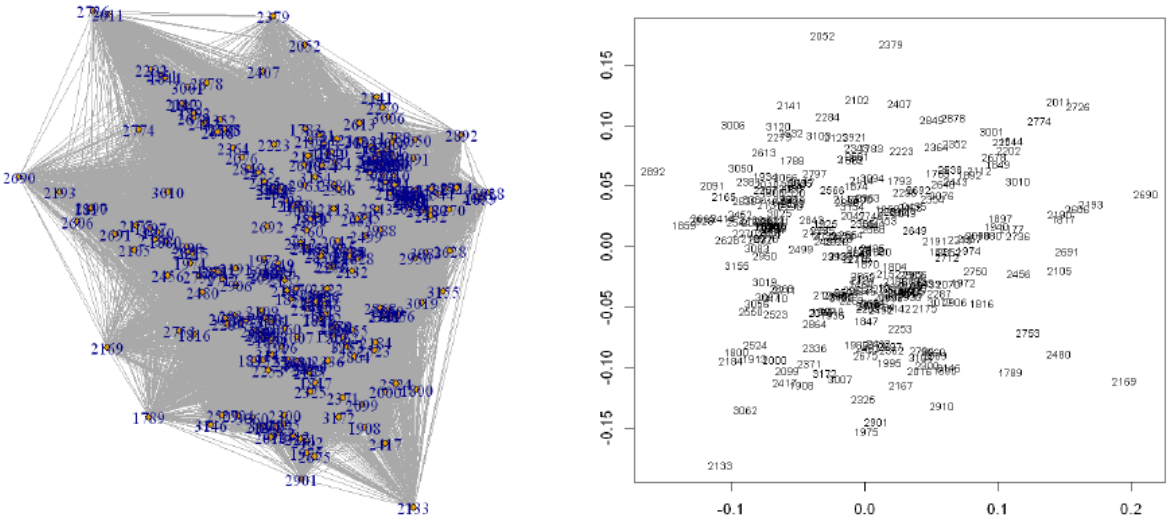

Figure 10 – Carte en 2D du Positionnement multidimensionnel (MDS) des observations.

Dans un second temps, nous avons normalisé les variables quantitatives. En effet, dans la plupart des algorithmes de classifications, les différences d'échelle reviennent à accorder plus ou moins de poids aux différentes variables. On voit clairement sur la figure 11 qui représente les boites à moustache des différentes variables que certaines variables dans le même plan ont leurs boîtes écrasées à cause de la différence d'échelle (figure à gauche). Nous avons donc centré et réduit les variables ce qui ajuste les échelles (figure à droite).

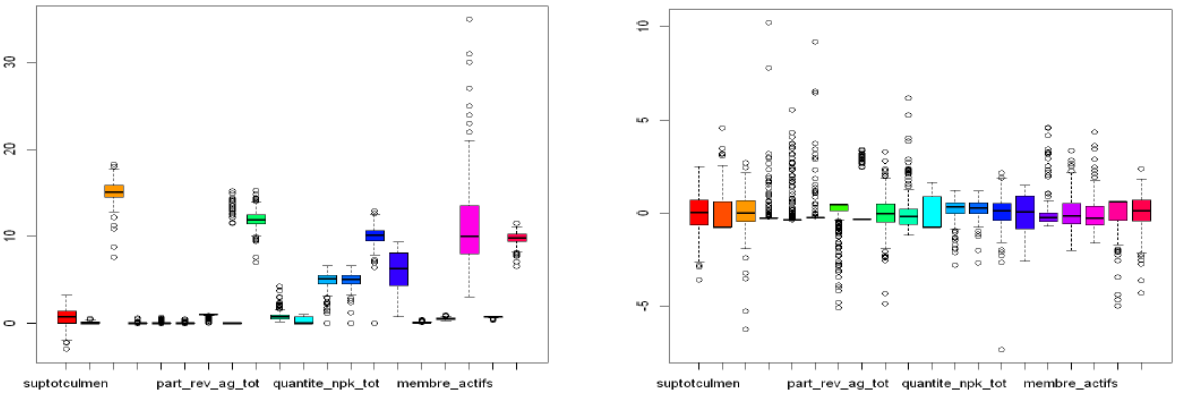


Figure 11 – Boites à moustache des différentes variables avant et après le centrage-réduction.

Il aurait cependant été imprudent de prendre la moyenne et l'écart-type pour mettre à l'échelle les données. En effet, en observant l'histogramme de différentes variables, nous nous sommes rendu compte que la grande majorité avait une distribution avec une lourde queue, ce qui suggère des valeurs aberrantes ou extrême (un exemple sur la figure 12 gauche). Nous avons donc log $-$*normalisé* ces variables (figure 11 droite) avant la mise à l'échelle. Il convient toutefois de souligner deux choses : d'abord ces transformations ne déforment pas le nuage de points, ensuite trois (03) variables avec de lourdes queues (dont la part de revenu *part_rev_transf_tot*, cf. annexe A) ne sont pas affectées par la log $-$*transformation*

et restent écrasées malgré la mise à l'échelle.

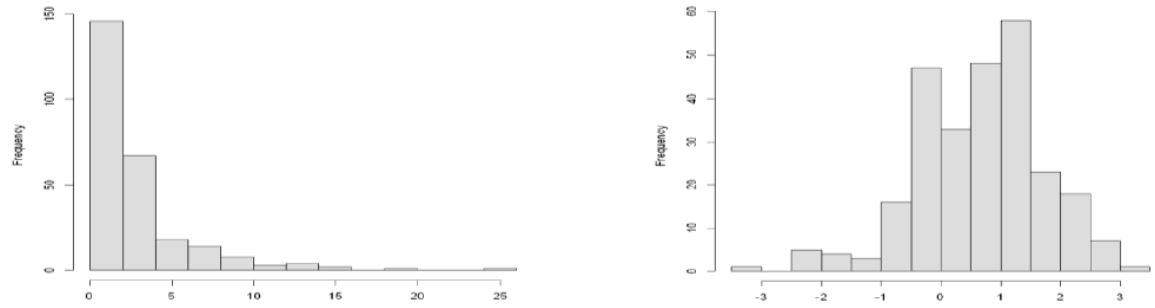


Figure 12 – Histogramme des valeurs de la superficie totale avant et après la log-normalisation.

Enfin, nous avons corrigé certaines variables qui apportent des informations intéressantes mais qui sont mécaniquement fortement corrélées à d'autres variables. Cette corrélation trompeuse peut biaiser nos analyses et/ou interprétations. Par exemple, la quantité totale d'urée en kilogramme (kg) est fortement corrélée à la superficie totale de l'exploitation en hectare (ha) ce qui est intuitif. Plus l'exploitation est grande, plus il y aura besoin d'urée. Mais ce qui peut être intéressant à voir, c'est si le fait d'utiliser plus l'urée que d'autres engrais peut caractériser des exploitations. La surface parasite cette information. Pour corriger cela, nous allons utiliser à la place la quantité d'urée par hectare (en divisant la quantité d'urée par la surface de l'exploitation). La démarche est la même pour les variables dans un cas similaire. La figure 13 montre la matrice de corrélation avant (à gauche) et après (droite) ces opérations.

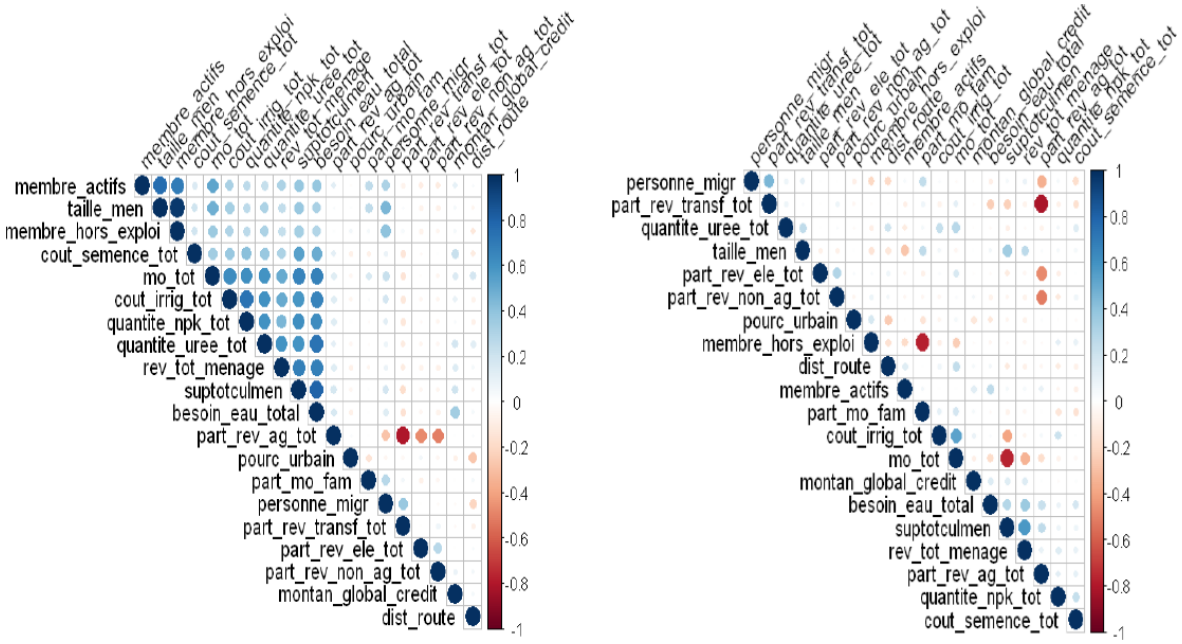


Figure 13 – Matrice de corrélation des variables quantitatives avant et après correction.

— **Classification des variables :**

La première étape de l'algorithme ClustOfVar est la classification des variables en vue de construire des variables synthétiques (gradients) pour représenter chaque groupe de variables. Pour ce faire, nous avons procédé à une CAH à partir des variables dont le dendrogramme est représenté sur la figure 14. Nous avons sélectionné ces variables en nous inspirant des variables habituellement utilisées pour les typologies d'exploitations. Ensuite en appliquant le critère de coude à la figure 16 représentant l'évolution des critères de classification issue du CAH, nous avons fixé le partitionnement idéal à cinq (05) classes.

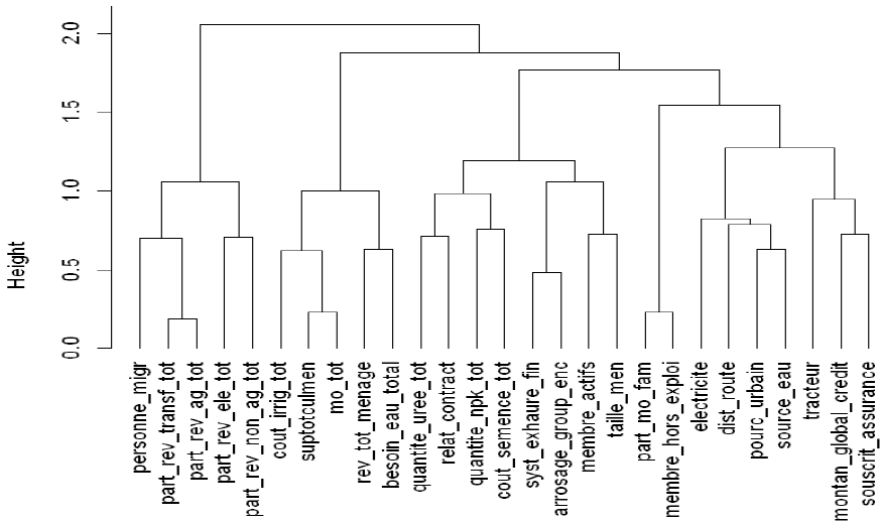


Figure 14 – Dendrogramme de la classification hiérarchique des variables.

.

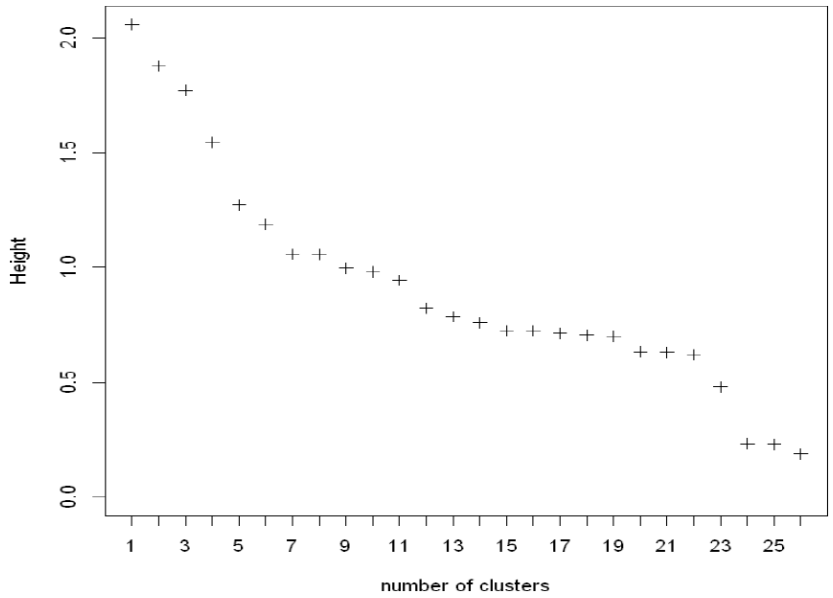


Figure 15 – Évolution du critère de classification des 27 variables.

Dès lors, nous avons utilisé le résultat du découpage du dendrogramme en cinq clusters comme partition initiale dans un algorithme de classification (le Kmeans). Les 27 variables

dont 20 quantitatives et 7 qualitatives, sont regroupées en 5 clusters comme suite :

Table 4 – Les capitaux (foncier, financier, matériel, travail)

| **Variables** | **Squared loading** | **Correlation** | **Coef** |
|---|---|---|---|
| suptotculmen | 0.832 | 0.91 | 0.57 |
| mo_tot | 0.736 | -0.86 | -0.54 |
| rev_tot_menage | 0.442 | 0.67 | 0.42 |
| cout_irrig_tot | 0.304 | -0.55 | -0.35 |
| tracteur = Non | 0.027 | NA | -0.02 |
| tracteur = Oui | | | 0.68 |
| besoin_eau_total | 0.194 | 0.44 | 0.28 |

Table 5 – La diversification des activités de l'exploitation

| **Variables** | **Squared loading** | **Correlation** | **Coef** |
|---|---|---|---|
| part_rev_ag_tot | 0.96 | -0.98 | -0.64 |
| part_rev_transf_tot | 0.63 | 0.80 | 0.52 |
| personne_migr | 0.28 | 0.53 | 0.35 |
| part_rev_non_ag_tot | 0.26 | 0.51 | 0.33 |
| part_rev_ele_tot | 0.21 | 0.46 | 0.3 |

Table 6 – L'insertion dans les réseaux

| **Variables** | **Squared loading** | **Correlation** | **Coef** |
|---|---|---|---|
| pourc_urbain | 0.533 | -0.73 | -0.54 |
| source_eau = ceane | 0.443 | NA | -1.85 |
| source_eau = puits_et_forage | | | 1.62 |
| source_eau = puits | | | -0.04 |
| source_eau = forage | | | 0.45 |
| source_eau = cours_eau | | | -1.16 |
| electricite = Non | 0.313 | NA | 0.34 |
| electricite = Oui | | | -0.5 |
| dist_route | 0.369 | 0.61 | 0.45 |
| montan_global_credit | 0.120 | 0.35 | 0.26 |
| souscrit_assurance = Non | 0.034 | NA | -0.01 |
| souscrit_assurance = Oui | | | 1.57 |

Table 7 – La famille

| **Variables** | **Squared loading** | **Correlation** | **Coef** |
|---|---|---|---|
| part_mo_fam | 0.89 | 0.94 | 0.71 |
| membre_hors_exploi | 0.89 | -0.94 | -0.71 |

Table 8 – Les itinéraires techniques

| **Variables** | **Squared loading** | **Corrélation** | **Coef** |
|---|---|---|---|
| arrosage_group_enc = Manuel | 0.576 | - | -0.09 |
| arrosage_group_enc = Mecanique | | | 0.36 |
| arrosage_group_enc = Mixte | | | -0.61 |
| arrosage_group_enc = Goutte_goutte | | | 2.05 |
| quantite_npk_tot | 0.278 | -0.53 | -0.37 |
| taille_men | 0.275 | -0.52 | -0.36 |
| relat_contract = Non | 0.273 | - | -0.04 |
| relat_contract = Oui | | | 3.37 |
| quantite_uree_tot | 0.266 | -0.52 | -0.36 |
| cout_semence_tot | 0.236 | -0.49 | -0.34 |
| membre_actifs | 0.114 | 0.34 | 0.23 |
| quantite_uree_tot | 0.266 | -0.52 | -0.36 |
| cout_semence_tot | 0.236 | -0.49 | -0.34 |
| syst_exhaure_fin = Manuel | 0.072 | - | 0.22 |
| syst_exhaure_fin = Manuel_et_Mecanique | | | -0.4 |
| syst_exhaure_fin = Mecanique | | | 0.04 |

La colonne Squared loading indique la corrélation entre les variables de la classe et la variable synthétique de cette classe pour les variables qualitatives. Dans le cas de variables quantitatives, la corrélation est donnée par le carré du coefficient de corrélation (colonne Corrélation). La colonne Coef donne les coefficients associés à chaque variable dans la construction de la variable synthétique. On peut lire ces valeurs comme on lit la contribution ou le cos2 pour l'ACP. Le regroupement ainsi obtenu permet d'avoir une première interprétation des groupes de variables et donc de nommer les variables synthétiques construites dans ces groupes. Par exemple, la variable synthétique du cluster 2 rendra compte des principales sources de revenus des ménages. Les ménages avec un revenu principalement agricole s'opposent aux ménages avec un revenu fortement basé sur les transferts (ou sur l'élevage, les revenus non-agricole et avec les membres du ménage ayant migré).

— **Identification des individus caractéristiques de chaque classe :**

Les variables synthétiques obtenues dans la deuxième étape en réalisant une AFDM sur chaque cluster, nous ont servi dans l'algorithme des K-means pour classifier les observations

à partir de notre nouvelle table de données constituée de cinq (05) variables quantitatives (les variables synthétiques de chaque cluster et des 213 observations de départ). Le choix du nombre optimal de clusters pour la typologie des observations s'est fait à partir de l'évolution du critère de classification dans la figure 16 ci-dessous.

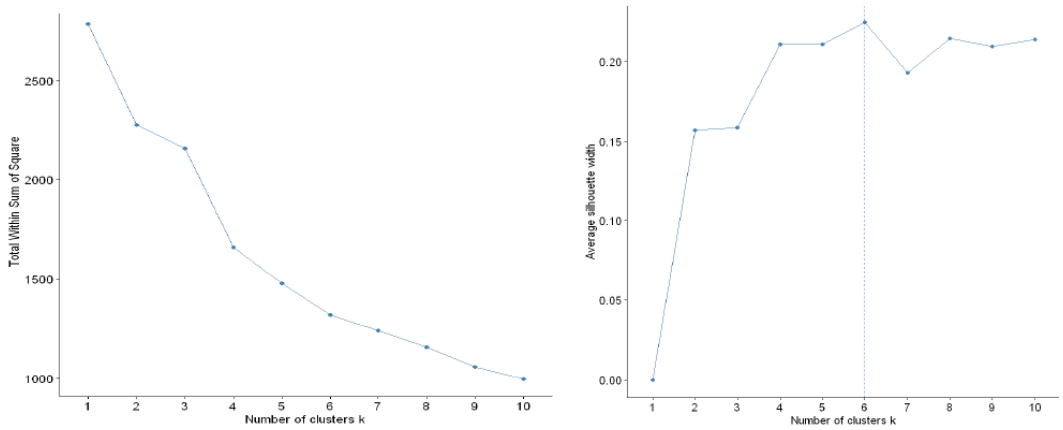


Figure 16 – Évolution du critère de classification des observations.

La méthode du coude (à gauche) suggère un partitionnement en 6 classes maximum et 4 minimum. Les silhouettes moyennes (à droite), qui représentent à quel point les variables ressemblent à leurs groupes et diffèrent des autres, suggèrent un découpage en 6 classes. Mais on voit que les silhouettes ne sont pas très élevées (autour de 0,3) et qu'un découpage en 4 ou 5 classes serait plus ou moins équivalents. Compte tenu de nos objectifs d'obtenir à la fin un nombre réduit de types d'exploitations, nous avons choisi le découpage en 4 classes. La figure 17 ci-dessous montre une projection dans un plan factoriel de la classification obtenue. Dans ce plan, on a un peu moins de 50% (environ 47,2%) de l'inertie totale.

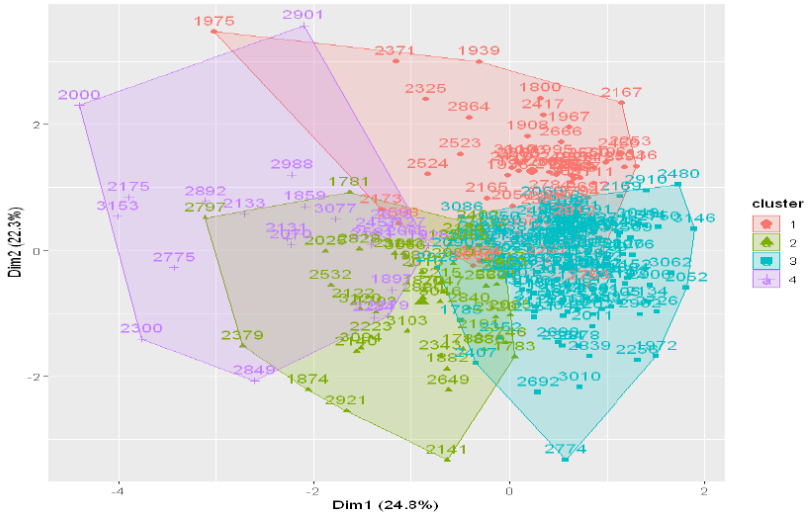


Figure 17 – Représentation du partitionnement des observations obtenues par les K-means.

L'objectif principal de cette typologie était de pourvoir générer des exploitations agricoles « types » dans les simulations à partir de la description des classes (cf. annexe A). Nous avons donc identifié des profils types (individus parangons) en prenant les individus les plus proches (distance euclidienne) du centre de gravité de la classe (pour plus de diversité dans les caractéristiques, nous avons pris les quatre (04) plus proches).

Table 9 – Les individus types dans le cluster

| **Parangons** | **Cluster** | **Distance au centroïde** |
|---|---|---|
| 3093 | 1 | 0.5 |
| 2115 | 1 | 0.7 |
| 2265 | 1 | 1 |
| 2818 | 1 | 1 |
| 2880 | 2 | 0.4 |
| 3046 | 2 | 0.8 |
| 2047 | 2 | 1.2 |
| 2840 | 2 | 1.2 |
| 1840 | 3 | 0.8 |
| 1858 | 3 | 0.9 |
| 2465 | 3 | 0.9 |
| 2831 | 3 | 1 |
| 1859 | 4 | 1.6 |
| 2070 | 4 | 1.6 |
| 3077 | 4 | 1.8 |
| 2123 | 4 | 2.0 |

## 4.2 Calibration et vérification

Cette partie présente le travail réalisé pour calibrer et vérifier les sorties de deux modules du modèle Niayes 2040 : le module de dynamiques d'occupation des sols (étalement des espaces urbains et irrigués), et le module de dynamiques de la nappe phréatique. Le module de dynamiques d'occupation du sol permet de simuler l'étalement urbain et l'espace irrigué à partir d'une carte initiale d'occupation du sol de 2006. Il fournit en sortie des cartes de type raster d'étalement des espaces urbains et irrigués à différentes dates. Le module de dynamiques hydrologique permet de simuler l'évolution du niveau de la nappe phréatique (piézométrie) à partir de l'année 2008. Il fournit en sortie des rasters de séries temporelles de piézométrie. Ces deux modules partent de données en entrées fixes et des paramètres à fixer (qui sont des variables dont la valeur est choisie en début de chaque simulation), pour reproduire les processus et fournir des sorties de simulation. Notre travail consiste à :

— Proposer un indicateur pour évaluer les sorties de simulation.

— Utiliser cet indicateur pour conduire des tests de sensibilité qui permettent d'évaluer

l'effet de chacun des paramètres sur les sorties du modèle et sélectionner ceux ayant le plus d'importance.

— Calibrer les modèles : il s'agit de choisir la combinaison de paramètres qui conduisent à la meilleure sortie

— Vérifier/valider les modèles : le modèle est fait tourner avec ce jeu de paramètres calibrés sur une période qui n'a pas été utilisée pour la calibration. Les sorties sont comparées avec des données observées non utilisées pour la calibration, et le RMSE final est calculé.

### 4.2.1 Présentation et sources des données

Les données utilisées pour la calibration et la vérification sont les données sorties des simulations du modèle Niayes2040 et des données d'observation (tableau 10).

Table 10 – Description des données de calibration

| **Type de donnée** | **Descriptif** | **Résolution** | **Source** |
|---|---|---|---|
| | **Données d'observation** | | |
| Utilisation et occupation du sol | Raster d'occupation du sol pour les années 2006, 2014 et 2018 | Grille de 1*1 km | Jolivot, 2021. |
| Stations d'observations de la profondeur de la nappe (Piézométrie) | Série de valeurs piézométriques issues de stations géolocalisées, mensuelles ou annuelles entre 2006 et 2020. | 24 Stations | Direction de la Gestion et de la Planification des Ressources en eau (DGPRE) et Industrie Minière la GCO |
| | **Données de simulation** | | |
| Utilisation et occupation du sol | Raster d'occupation du sol simulé pour les années 2006, 2014 et 2018 | Grille de 1*1 km | Simulation du module dynamique d'occupation du sol |
| Piézométrie simulée | Série de valeurs piézométriques simulées avec un pas de temps décadaire | 24 points correspondants aux 24 stations | Simulation du module hydrologique |

Le module de dynamique de la nappe Quaternaire contient donc 5 paramètres dont il est possible d'ajuster les valeurs : $larésolution(Res)$, l'ajustement de $laperméabilité(varperm)$, l'ajustement de la réserve utile du sol ($\Delta_S$), le nombre d'itérations

pour $lesécoulementslatéraux(nbiterations)$, et $laconsommationurbaine(Consourb)$. Le tableau 11 suivant donne les valeurs testées pour ces différents paramètres.

Table 11 – Valeurs des paramètres

| **Paramètres** | **Intervalle de valeurs** | **Pas** |
|---|---|---|
| resolution | [200, 1000] | 100 |
| varperm | [-0.001, 0.001] | 0.0005 |
| $\Delta S$ | [-60, 60] | 30 |
| nbiterations | [3, 50] | 20 |
| Consourb | 0.03 et 0.06 | - |

Pour le module occupation du sol, 8 paramètres ont été testés menant à 56 000 simulations (tableau 12 ci-dessous).

Table 12 – Valeurs des paramètres testées

| **Paramètres** | **Intervalle de valeurs** | **Pas** |
|---|---|---|
| $Coef_{voisin_irr}$ | [0, 9] | 3 |
| $Coef_{route}$ | [0, 9] | 3 |
| $Coef_{pente_faible}$ | [0, 9] | 3 |
| RatioListe | 0.2 ; 0.4 ; 0.6 | - |
| $Coef_{urb_voisin}$ | [0, 9] | 3 |
| $Coef_{urb_route}$ | [0, 9] | 3 |
| popHaIrrig | 9 ; 12 ; 15.0 | - |
| popHaUrb | 60 ; 65 ; 70.0 | - |

### 4.2.2 Analyse de sensibilité

— **Modèle Hydrologique (Hydro):**

D'après le graphique à gauche sur la figure 18 représentant les écart-types des effets élémentaires en fonction des moyennes des effets élémentaires absolues, les variables Résolution et Varperm sont influentes, avec des effets non-linéaires et/ou d'interactions (le plus influent étant de peu la résolution). Les autres sont négligeables. Après avoir fixé un Varperm optimal en regardant la valeurs du Varperm qui optimisent le RMSE, on a retiré ce paramètre de l'analyse. Pour la deuxième analyse (graphique à droite), un nouveau paramètre a été testé : la quantité d'eau moyenne par hectare utilisée pour l'irrigation (Moyirrigculture). Cette dernière analyse indique que la quantité d'eau moyenne pour l'irrigation (Moyirrigculture) et la consommation urbaine (Consurb) sont négligeables. Par ailleurs, la conclusion est la même que précédemment pour la résolution. Elle est influente, avec des effets non-linéaires et/ou interactions. L'effet du Varstock est toute relative mais il est bien distinct des effets nuls ou

quasi-nulle des variables Consurb et Moyirrigculture. Une analyse plus poussée (quantitative) serait une bonne piste pour mieux saisir l'intensité et/ou la nature de l'influence de la Varstock. La même analyse pour le Varperm a été menée pour le Varstock afin de fixer des valeurs de ce paramètre qui optimisent le RMSE.

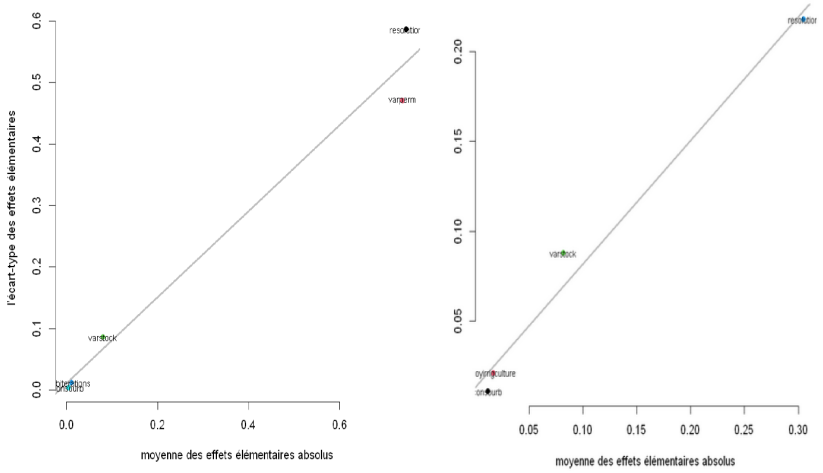


Figure 18 – Visualisation du plan de Morris du modèle hydrologique.

Table 13 – Rang de Morris modèle hydrologique - $1^{ère}$ série de simulations

| | Moyenne ($\mu_{EE}$) | Ecart type ($\sigma_{EE}$) | Rang Morris |
|---|---|---|---|
| resolution | 0.747 | 0.587 | 1 |
| varperm | 0.738 | 0.471 | 2 |
| varstock | 0.080 | 0.087 | 3 |
| nbiterations | 0.010 | 0.012 | 5 |
| Consourb | 0.005 | 0.004 | 6 |

Table 14 – Rang de Morris modèle hydrologique - $2^{ème}$ série de simulations

| | Moyenne ($\mu_{EE}$) | Ecart type ($\sigma_{EE}$) | Rang Morris |
|---|---|---|---|
| resolution | 0.305 | 0.0218 | 1 |
| varstock | 0.083 | 0.088 | 2 |
| moyirrigculture | 0.017 | 0.021 | 3 |
| Consourb | 0.012 | 0.011 | 4 |

Les tableaux 13 et 14 ci-dessus donnent l'ordre d'importance des variables pour les deux analyses de sensibilité du modèle hydrologique. L'ordre est déterminé par la somme des deux mesures de sensibilité ; plus elle est élevée plus le paramètre est important donc à estimer précisément. Les résultats d'analyse du modèle occupation du sol sont présentés en annexe A.

— **Modèle Occupation du Sol et Irrigation (OCS):**

La figure 19 ci-dessous présente le plan de Morris des variables testées dans le modèle occupation du sol et irrigation. L'analyse de ce graphique suggère que la variable voisUrbCoef est très influente et a un effet non linéaire et/ou monotone. Il y a ensuite la variable expCoeff qui elle aussi est très influente avec un effet linéaire. Les variables routesCoef, penteInfCoeff, distrouteCoeff et ratioliste sont plus ou moins dans la même région du plan avec des effets non négligeable linéaire. Les autres sont négligeables avec des effets quasi-nulles. Le tableau des rangs de Morris permet de consolider ces observations en fournissant l'ordre d'importance des paramètres.

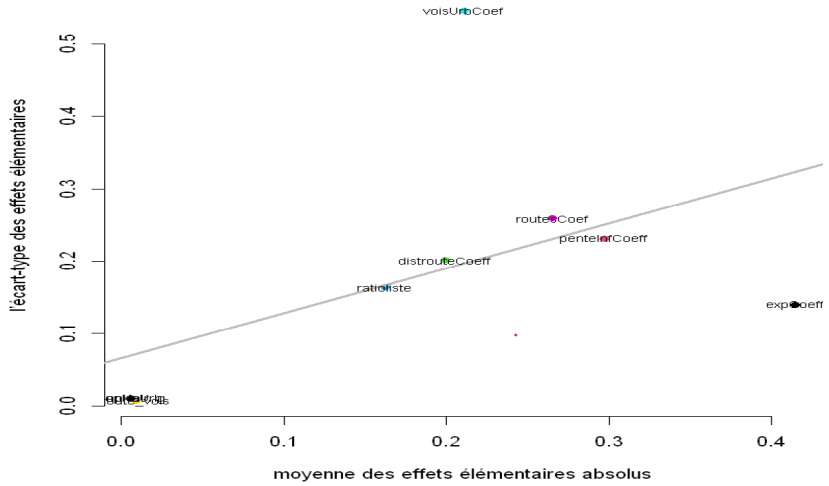


Figure 19 – Visualisation du plan de Morris du modèle occupation du sol et irrigation.

Table 15 – Rang de Morris modèle occupation du sol et irrigation

| Variable | Moyenne des effets | Écart type des effets | Rang Morris |
|---|---|---|---|
| voisUrbCoef | 0.211 | 0.546 | 1 |
| expCoeff | 0.415 | 0.140 | 2 |
| penteInfCoeff | 0.297 | 0.231 | 3 |
| routesCoef | 0.265 | 0.259 | 4 |
| distrouteCoeff | 0.199 | 0.202 | 5 |
| ratioliste | 0.162 | 0.164 | 6 |
| popHaIrrig | 0.007 | 0.010 | 7 |
| popHaUrb | 0.006 | 0.010 | 8 |
| route_vois | 0.009 | 0.007 | 9 |

### 4.2.3 La calibration de modèle

La calibration que nous avons effectuée est une combinaison des méthodes de calage graphique et d'optimisation. En effet, pour le module de dynamiques d'occupation du sol,

notre calibration a consisté à identifier les combinaisons de paramètres qui minimisent le RMSE. Tandis que pour le modèle Hydro, en plus de la méthode précédente, les chercheurs ont confirmé graphiquement les simulations retenues en comparant les tendances des tracés de ces simulations aux valeurs tracées des observations. La figure 20 ci-dessous présente les graphiques que nous avons générés à partir des simulations pour permettre aux chercheurs d'évaluer visuellement le comportement de leur modèle.

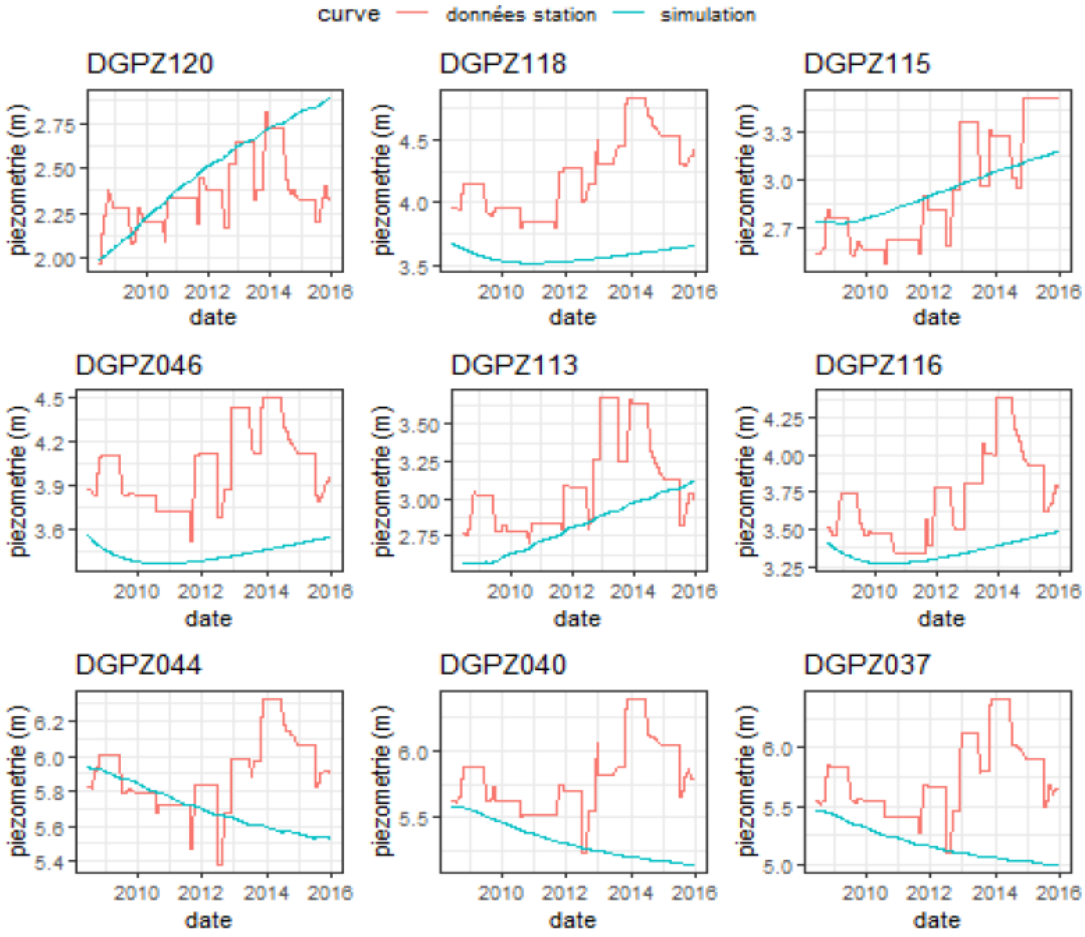


Figure 20 – Visualisation du calage des simulations du modèle aux valeurs réelles observées.

De cette façon, la calibration revient à ''entraîner" le modèle à ressembler à des données qu'on a observées. Cela pose le problème d'applicabilité du modèle pour de la prospective par exemple, puisqu'il risque d'être mauvais pour simuler des données hors de cet échantillon d'entrainement ; sur les réalisations futures en l'occurrence. Ce qui serait incompatible avec l'objectif de ce projet qui est de faire de la prospective sur les avenirs plausibles des Niayes avec ce modèle. D'où l'étape de validation ou de vérification du modèle.

### 4.2.4 La validation ou vérification

Pour chacune des stations, les données ont été divisées en deux parties et la calibration a été faite avec les RMSE calculés sur la première partie. On a ensuite calculé les RMSE pour la simulation avec les meilleurs paramètres ainsi obtenu. Le Modèle est validé puisque le RMSE est resté bon (en dessous du seuil empirique) sur l'échantillon de vérification avec les paramètres de calibration. Nos modèles sont calés.

Table 16 – Validation du modèle Hydro

| Calibration Hydro : 2008-2012 | | | | | |
|---|---|---|---|---|---|
| Meilleure simulation | consourb | moyirrigculture | varstock | resolution | rmse_agg_dgpz |
| Simulation 4 | 0.03 | 40.0 | 0.0 | 300.0 | 0.24 |
| Vérification Hydro : 2013-2016 | | | | | |
| Meilleure simulation | consourb | moyirrigculture | varstock | resolution | rmse_agg_dgpz |
| Simulation 4 | 0.03 | 40.0 | 0.0 | 300.0 | 0.39 |

# Conclusion

L'objectif générale de notre travail était de préparer des données d'entrée pour faire tourner un modèle des dynamiques hydrique et d'occupation du sol de la zone des Niayes au Sénégal, puis de calibrer et de vérifier les sorties de ces modèles. Ainsi, nous avons grâce à l'approche ClustOfVar, identifié quatre (04) profils d'exploitation dans la zone. Ces profils, et notamment les caractéristiques de classe des individus parangons, ont ensuite été utilisés par l'équipe de recherche du projet Niayes2040 pour reproduire les caractéristiques des exploitations agricoles dans leur modèle. Nous avons ensuite estimé les taux de croissance démographique de la zone. Ces taux ont été utilisés comme paramètre d'entrée du modèle Niayes2040 pour simuler la dynamique d'occupation du sol. Il a ensuite fallu calibrer et vérifier ce modèle. Pour cela, nous avons proposé et calculé plusieurs indicateurs de calibration, sur lesquels nous nous sommes basés pour réaliser des tests de sensibilités. Ces tests nous ont permis d'identifier les paramètres à calibrer de manière fine. Puis, nous avons réalisé la calibration sur une partie des données puis validé le modèle sur une autre. Nous avons enfin produit des graphiques pour permettre de visualiser les sorties du modèle. Le travail que nous avons présenté dans ce document a appuyé l'équipe de recherche du projet Niayes 2040 dans la simulation des différents scénarios pour le futur des Niayes.

# References

[AFD21] AFD. « L'AFD et le Sénégal : favoriser une croissance inclusive et protéger l'environnement ». In: (2021). URL: https://www.afd.fr/fr/page-region-pays/senegal (visited on 2021).

[al06] A. Saltelli et al. « Sensitivity analysis in pratice : A guide to assessing scientific model ». In: (2006). URL: https://academic.oup.com/jrsssa/article/168/2/466/7084302?login=false (visited on 2021).

[al18a] C. Camara et al. « Rapport des ateliers de coconstruction de scénarios prospectifs pour la zone sud des Niayes ». In: (2018). URL: https://doi.org/10.18167/agritrop/00433 (visited on 2021).

[al19] Clémentine Camara et al. « Quel avenir pour l'espace agro-sylvo-pastoral de la zone sud des niayes à l'horizon 2040 ? » In: (2019). URL: https://dumas.ccsd.cnrs.fr/dumas-03866273/document (visited on 2021).

[al16] K.S. Kuivanenet al. « A comparison of statistical and participatory clustering of smallholder farming systems – A case study in Northern Ghana ». In: (2016). URL: https://www.sciencedirect.com/science/article/abs/pii/S0743016716300493 (visited on 2021).

[al21] M. Lihemohn et al. « RMSE is not enough: Guidelines to robust data-model comparisons for magnetospheric physics ». In: (2021). URL: https://www.sciencedirect.com/science/article/pii/S1364682621000857 (visited on 2021).

[al15] Prasad Daggupati et al. « A Recommended Calibration and Validation Strategy for Hydrologic and Water Quality Models ». In: (2015). URL: https://www.sciencedirect.com/science/article/abs/pii/S0010465510005321 (visited on 2021).

[al18b] Stéphanie Alvarez et al. « Capturing farm diversity with hypothesisbased typologies: An innovative methodological framework for farming system typology development ». In: (2018). URL: https://doi.org/10.1371/journal.pone.0194757 (visited on 08/18/2023).

[al13] V. Kuentz-Simonet et al. « Une approche par classification de variables pour la typologie d'observations : le cas d'une enquête agriculture et environnement ». In: (2013). URL: https://hal.science/hal-00876254/file/bx2013-pub00039083.pdf (visited on 2021).

[Bui14] Samuel Buis. « Analyse de sensibilité des modèles de simulation ». In: (Oct. 26, 2014). URL: https://www.eccorev.fr/IMG/pdf/Sensibilite_SBuis.pdf (visited on 2021).

[C W09] K. Matsuura et S. Robeson C. Willmott. « Ambiguities inherent in sums-of-squares-based error statistics ». In: (2009). URL: https://www.sciencedirect.com/science/article/abs/pii/S1352231008009564 (visited on 2021).

[Cer17] Paul Ceria. « Les plans de Morris ». In: (2017). URL: http://paulceria.com/page/Incertitude/afm_virtuel_morris.php (visited on 2021).

[Cir21] Cirad. « Explorer et quantifier les futurs plausibles de la région des Niayes au Sénégal - Niayes 2040 ». In: (2021). URL: https://www.cirad.fr/dans-le-monde/cirad-dans-le-monde/projets/projet-niayes-2040 (visited on 2021).

[Fra] Jessica Cariboni et Andrea Saltelli Francesca Campolongo. « An effective screening design for sensitivity analysis of large models ». In: (). URL: https://www.sciencedirect.com/science/article/abs/pii/S1364815206002805 (visited on 2021).

[Gow71] J. C. Gower. « A General Coefficient of Similarity and Some of Its Properties ». In: (1971). URL: https://members.cbio.mines-paristech.fr/~jvert/svn/bibli/local/Gower1971general.pdf (visited on 2021).

[Ioo] Bertrand Iooss. « Analyses d'incertitudes et de sensibilité de modèles complexes - Applications dans des problèmes d'ingénierie ». In: (). URL: https://www.math.univ-toulouse.fr/~baehr/meteo_SMAI/Pres/Pres_Iooss.pdf (visited on 2021).

[K M05] C. Willmott et K. Matsuura. « Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance ». In: (2005). URL: https://www.sciencedirect.com/science/article/abs/pii/S1352231008009564 (visited on 2021).

[Mah18] S. Mahévas. « Introduction à la calibration de modèles complexes ». In: (2018). URL: https://reseau-mexico.fr/sites/mexicoD8/files/PresentationsEC2018LaRochelle/mexicoEC2018-expose_IntroductionCalibration.pdf (visited on 2021).

[Mon21] Alberto Montanari. « Model calibration and validation ». In: (2021). URL: https://www.albertomontanari.it/node/87 (visited on 2021).

[Mor91] Max D. Morris. « Factorial Sampling Plans for Preliminary Computational Experiments ». In: (1991). URL: https://www.tandfonline.com/doi/abs/10.1080/00401706.1991.10484804 (visited on 2021).

[R R14] T. Chai et R. R. Draxler. « Root mean square error (RMSE) or mean absolute error (MAE) ? – Arguments against avoiding RMSE in the literature ». In: (2014). URL: https://gmd.copernicus.org/articles/7/1247/2014/ (visited on 2021).

[Rob12] Bourgeois Robin. « Guide Méthodologique de Prospective Territoriale Avec Application À Une Prospective Du Monde Agricole et Rural À Mayotte (Document de travail non publié) ». In: (2012). (Visited on 2021).

[Rob21] Bourgeois Robin. « Niayes 2040: Comité de pilotage, (Document de travail non publié) ». In: (2021). (Visited on 2021).

# Glossaire

**A** | **C** | **D** | **M** | **N** | **O** | **P** | **S** | **U**

**A**

**Agriculture** L'agriculture est un processus par lequel les êtres humains aménagent leurs écosystèmes et contrôlent le cycle biologique d'espèces domestiquées, dans le but de produire des aliments et d'autres ressources utiles à leurs sociétés..

**Agro-industrie** L'agro-industrie est un terme regroupant l'ensemble des industries ayant un lien direct avec l'agriculture..

**C**

**Calibration** La calibration d'un modèle empirique ou semi-empirique consiste à ajuster l'ensemble des valeurs des paramètres de façon à avoir le meilleur ajustement possible entre le comportement du modèle et celui de l'objet d'étude. Pour quantifier globalement la qualité de cet ajustement, on utilise le plus souvent le critère dit "des moindres carrés"..

**Changements globaux** Les Changements globaux change désignent le changement des conditions climatiques dans l'atmosphère terrestre liées aux activités humaines..

**D**

**Dynamique de la zone des Niayes** Dynamiques territoriales des Niayes: divergences économiques entre industrie extractive minière et agriculture (littoral Nord du Sénégal)..

**M**

**Modèle de simulation** Utilisation d'une représentation, forcément simplifiée d'un objet, d'un phénomène construite dans le but d'en faciliter l'étude, d'en mieux comprendre le comportement, d'en prédire les propriétés, d'en prévoir l'évolution etc. pour prédire les propriétés, prévoir le comportement de l'objet modélisé..

**Modèles hydrologique** Un modèle hydrologique, ou modèle pluie-débit, est un outil numérique de représentation de la relation pluie-débit à l'échelle d'un bassin versant..

**N**

**Nappes phréatiques** Un aquifère est un sol ou une roche réservoir originellement poreuse ou fissurée, contenant une nappe d'eau souterraine et suffisamment perméable pour que l'eau puisse y circuler librement. Ainsi, la nappe phréatique est une nappe d'eau que l'on rencontre à faible profondeur. Elle alimente traditionnellement les puits et les sources en eau potable. C'est la nappe la plus exposée à la pollution en provenance de la surface..

**Niayes 2040** Le projet explore les futurs plausibles qui s'offrent à la région des Niayes au Sénégal et les différentes actions qui pourraient être réalisées pour tendre vers un futur collectivement négocié..

**O**

**Occupation du sol** L'occupation du sol est pour la FAO [1998] « la couverture (bio-)physique de la surface des terres émergées » et donc le type d'usage (ou de non-usage) fait des terres par l'Homme..

**P**

**Piézométrie** Le niveau, la cote ou la surface piézométrique est l'altitude ou la profondeur de la limite entre la nappe phréatique et la zone vadose dans une formation aquifère. Ce niveau est mesuré à l'aide d'un piézomètre..

**Prospective** Ensemble de recherches concernant l'évolution future des sociétés et permettant de dégager des éléments de prévision..

**S**

**Salinisation des nappes phréatiques** La salinisation est le processus par lequel la concentration de sels et de minéraux dans les eaux souterraines augmente, détériorant ses paramètres de qualité. Ce processus est considéré comme un type de pollution du sol et de l'eau qui affecte davantage les aquifères côtiers..

**Scénarios plausibles** décrit une approche pour imaginer des histoires, des scénarios « affections → affects humains1 → idées et actions → affections », sur tout sujet, scénarios qui se sont passés ou auraient pu se passer jusqu'à une situation sociale existante ou ayant existé ou qui aurait pu exister, ou scénarios qui pourraient advenir. Dans ces scénarios, la temporalité n'est qu'ordinale (avant, pendant, après) et non cardinale (ni durée, ni datation), sauf à raccrocher des événements avérés à ces scénarios [Andre Moulin, 2021]..

**U**

**Urbanisation** C'est donc un processus de développement des villes et de concentration des populations dans celles-ci..

# A Annexe

## Scripts R

Lien vers mes Scripts R sur le Git repository dédié.

## Prétratitement des données

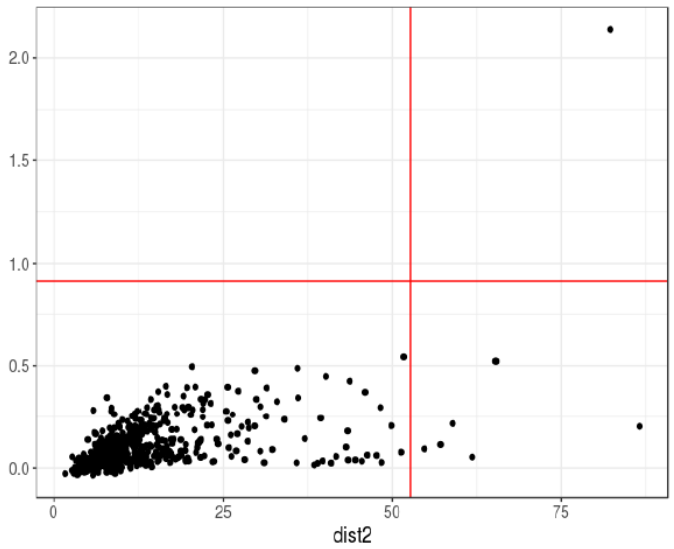


Figure 21 – illustration de la détection graphique d'outliers dans un plan factoriel [source].

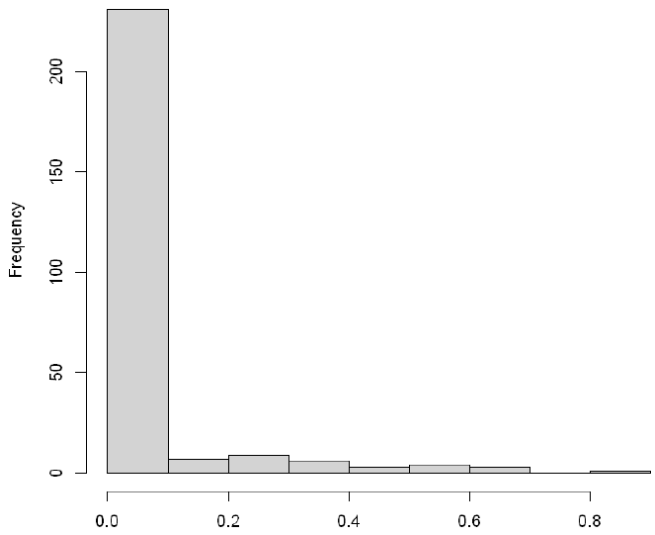


Figure 22 – histogramme de fréquence de la part de revenu de transfert dans la part du revenu total avant et après la log-normalisation.

## Méthodes de calcul des métriques et d'analyse de sensibilité

Le Chapitre 4.2 est consacré à la phase d'évaluation et d'optimisation des modèles hydrologique et d'occupation du sol. Dans ce cadre, il est important de déterminer quelles métriques utiliser pour évaluer les sorties des modèles. Il s'agit ici de présenter les réflexions que nous avons menées dans le choix de nos indicateurs. Dans le tableau ci-dessous, on décrit quelques-uns des autres calculs proposés.

Table 17 – Les indicateurs explorés.

| **Indicateur** | **Méthode de calcul** | **Intérêt** |
|---|---|---|
| RMSE | Calcul de la moyenne des RMSE des stations de la Zone minière DGPZ avant de calculer la moyenne de la Zone d'étude en utilisant cette première moyenne pour représenter toute la zone. | Annuler le problème de spatialité en résumant la proximité et le nombre de stations en zone minière par l'unique valeur de la moyenne de la zone. |
| RMSLE | Le calcul est le même que pour le RMSE en ce qui concerne la moyenne sur la zone d'étude. Mais à la place du RMSE, on utilise le RMSLE où la formule est celle du RMSE mais à la place des valeurs prédites et observées, on fait une log-transformation. | Les mêmes propriétés que précédemment ; à la différence qu'on profite des propriétés du log qui lisse les valeurs aberrantes tout en pénalisant les erreurs importantes. Avec les propriétés du Log, on ne risque pas d'avoir d'explosion de la valeur de l'indicateur dû aux échelles de valeurs ; On peut aussi lire le résultat comme une erreur relative. |
| MMAD | Il s'agira de prendre plutôt la médiane de l'écart médian à la médiane des RMSE. | L'indicateur ne sera pas sensible aux valeurs aberrantes et le problème de spatialité ne devrait pas se poser. |

## Algorithmes de classification

Le tableau suivant présente tous les algorithmes implémentées pour obtenir des typologies d'exploitation.

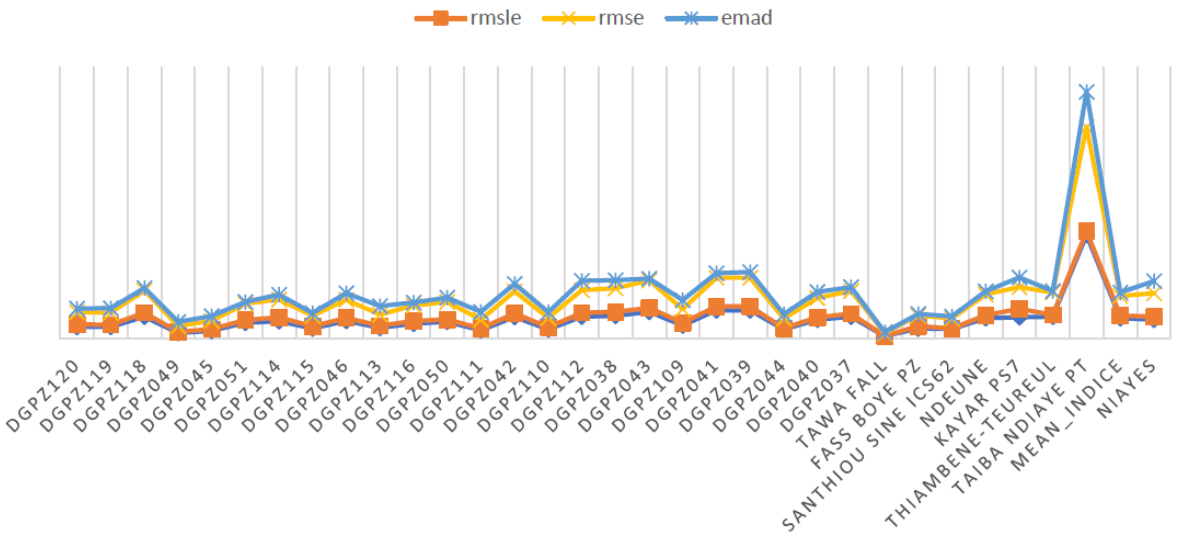


Figure 23 – Quelques indicateurs calculés par station.

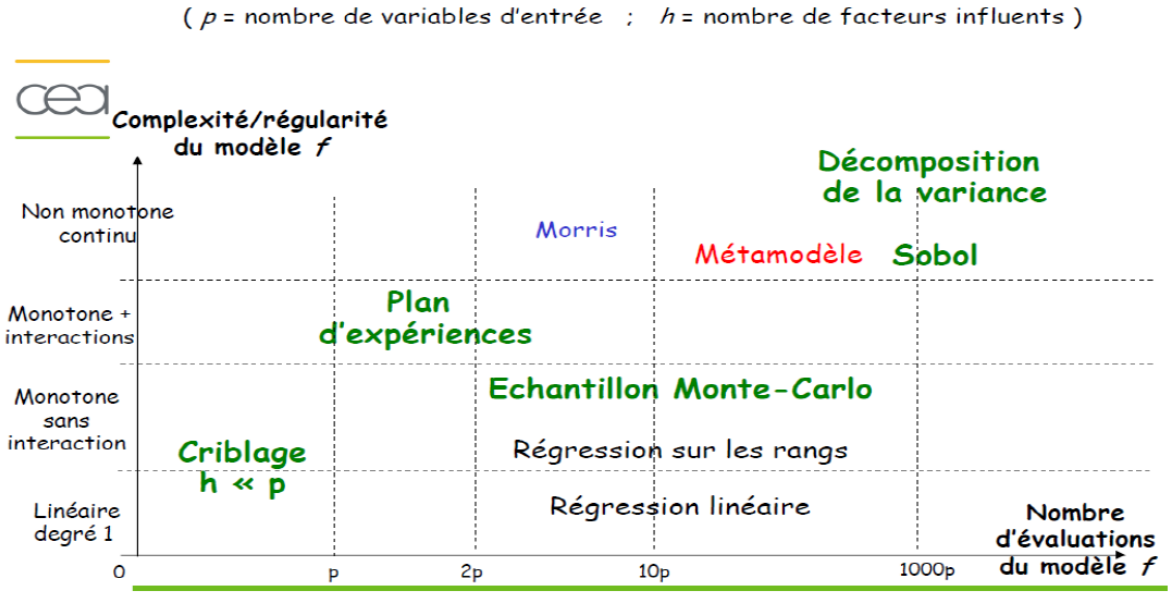


Figure 24 – Classification des méthodes d'analyse de sensibilité.

## Caractérisation des classes

Les tableaux suivant présentent respectivement, les effectifs dans les classes, la caractérisation des classes par les variables qualitatives avec les proportions sur les modalités et la caractérisation des classes par les variables quantitatives avec les moyennes à l'intérieur des groupes.

Table 18 – Algorithmes testés pour les typologies d'exploitation

| **Algorithmes de réduction de dimensions** | **Algorithme de classification** |
|---|---|
| Classification de variables (Kmeans) et construction de variables synthétiques (AFDM) | Kmeans |
| Analyse Factorielle des Données Mixtes (AFDM) | CAH (Kmeans consolidation) |
| Non-metric Multidimensional Scaling (MDS) : Sammon's non-linear mapping | Kmeans |

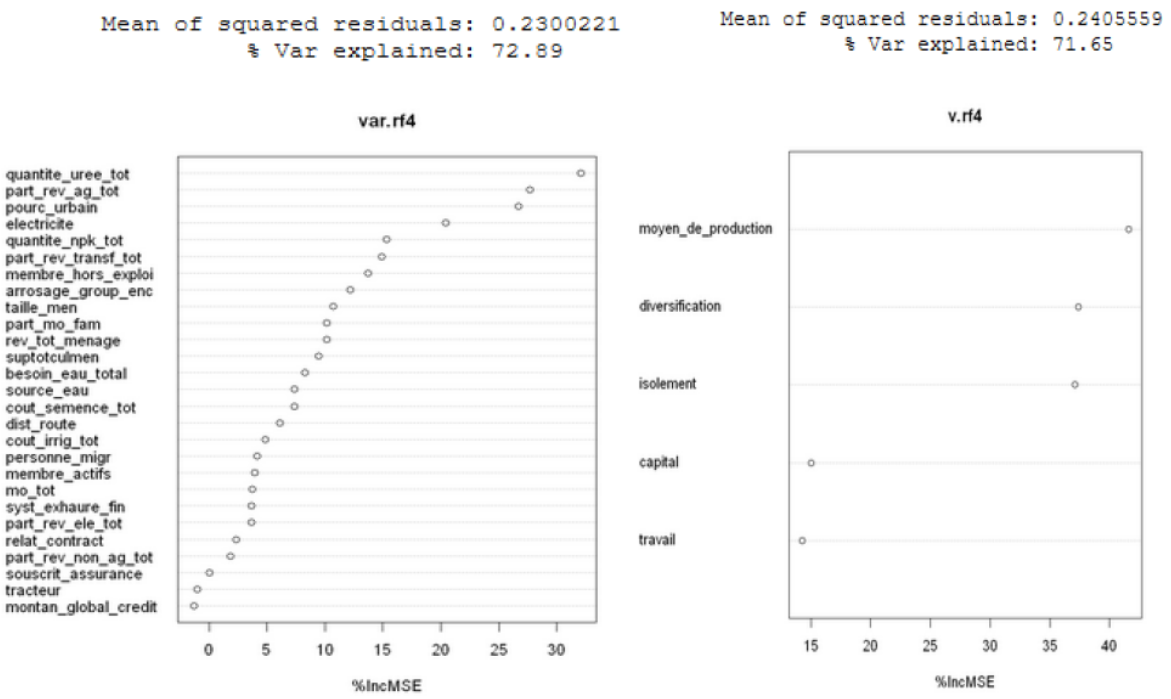


Figure 25 – Perte d'inertie liée au retrait de chaque variable de la classification obtenue avec un RandomForest.

Table 19 – Caractérisation des classes : effectifs.

| | **Effectifs** | **Proportion (en %)** |
|---|---|---|
| Cluster 1 | 55 | 20.8 |
| Cluster 2 | 48 | 18.2 |
| Cluster 3 | 137 | 51.9 |
| Cluster 4 | 24 | 9.1 |

## Paramètres optimaux

Table 20 – Caractérisation des classes : les variables qualitatives

| | | **Cluster 1** | **Cluster 2** | **Cluster 3** | **Cluster 4** |
|---|---|---|---|---|---|
| tracteur | Non | 100% | 100% | 96% | 100% |
| | Oui | 0% | 0% | 4% | 0% |
| relat_contract | Non | 95% | 100% | 100% | 100% |
| | Oui | 5% | 0% | 0% | 0% |
| source eau | Ceane | 13% | 0% | 0% | 8% |
| | Cours_eau | 4% | 0% | 2% | 0% |
| | Forage | 14% | 13% | 20% | 4% |
| | Puits | 69% | 85% | 74% | 88% |
| | Puits_et_forage | 0% | 2% | 4% | 0% |
| syst_exhaure_fin | Manuel | 18% | 10% | 16% | 29% |
| | Manuel_et_Mécanique | 5% | 19% | 21% | 4% |
| | Mécanique | 76% | 71% | 63% | 67% |
| souscrit_assurance | Non | 100% | 100% | 99% | 99% |
| | Oui | 0% | 0% | 1% | 1% |
| electricite | Non | 15% | 71% | 74% | 58% |
| | Oui | 85% | 29% | 26% | 42% |
| arrosage_group_enc | Goutte_goutte | 1% | 6% | 0% | 0% |
| | Manuel | 33% | 33% | 34% | 38% |
| | Mécanique | 53% | 40% | 25% | 58% |
| | Mixte | 5% | 21% | 41% | 4% |

Table 21 – Caractérisation des classes : les variables quantitatives.

| **Variables** | **Cluster 1** | **Cluster 2** | **Cluster 3** | **Cluster 4** | **% d'inertie** |
|---|---|---|---|---|---|
| suptotculmen | 2.87 | 1.04 | 4.26 | 1.41 | 15.48 |
| personne_migr | 0.38 | 0.54 | 1.21 | 2.12 | 13.31 |
| rev_tot_menage | 5997345.00 | 2492788.96 | 8280951.00 | 2915067.71 | 7.32 |
| part_rev_ele_tot | 0.00 | 0.00 | 0.01 | 0.08 | 11.21 |
| part_rev_transf_tot | 0.01 | 0.01 | 0.02 | 0.38 | 62.60 |
| part_rev_non_ag_tot | 0.01 | 0.00 | 0.01 | 0.08 | 14.24 |
| part_rev_ag_tot | 0.98 | 0.99 | 0.96 | 0.46 | 77.89 |
| montant_global_credit | 309D9.09 | 32208.33 | 140824.82 | 62500.00 | 1.75 |
| cout_irrig_tot | 374025.09 | 355004.17 | 831359.85 | 185395.83 | 6.22 |
| mo_tot | 1.85 | 2.44 | 2.96 | 2.04 | 6.08 |
| part_mo_fam | 0.13 | 0.72 | 0.24 | 0.41 | 23.54 |
| quantite_npk_tot | 308.84 | 164.25 | 764.92 | 200.00 | 11.39 |
| quantite_uree_tot | 283.24 | 220.38 | 785.17 | 180.50 | 8.05 |
| cout_semence_tot | 86724.33 | 36003.76 | 164434.00 | 49039.63 | 9.92 |
| dist_route | 881.40 | 4307.48 | 1954.21 | 2178.23 | 13.22 |
| pourc_urbain | 0.11 | 0.02 | 0.03 | 0.05 | 20.50 |
| membre_actifs | 5.82 | 6.52 | 7.78 | 5.96 | 7.29 |
| taille_men | 8.53 | 10.19 | 13.39 | 11.00 | 13.90 |
| membre_hors_exploit | 8.31 | 8.38 | 12.47 | 10.00 | 16.82 |
| besoin_eau_total | 56979.33 | 22707.33 | 118012.88 | 12928.16 | 11.12 |

Table 22 – Paramètres optimaux déterminer pour Les variables Varperm et Varstock pour différentes stations.

| **Stations** | **Varperm Optimal** | **Varstock Optimal** |
|---|---|---|
| DGPZ037 | 0 | -60 |
| DGPZ038 | 0.001 | -60 |
| DGPZ039 | -0.0005 | -60 |
| DG P7040 | 0 | 30 |
| DGPZ041 | -0.0005 | -60 |
| DGPZ042 | -0.001 | -60 |
| DGPZ043 | 0 | -6 |
| DG PZ044 | 1 | 60 |
| DGPZ045 | 0 | 50 |
| DGPZ046 | -1 | 0 |
| DGPZ049 | -1 | -30 |
| DGPZ050 | 0.0005 | -60 |
| DGPZ051 | -1 | -30 |
| DGPZ109 | 0.0005 | -60 |
| DGPZ110 | 0.0005 | 60 |
| DGPZ111 | -0.001 | -30 |
| DGPZ112 | -0.001 | -60 |
| DGPZ113 | 0 | -60 |
| DGPZ114 | -0.0005 | -60 |
| DGPZ115 | 0 | -60 |
| DGPZ116 | -0.001 | -30 |
| DGPZ118 | -0.0005 | -30 |
| DGP7119 | -0.001 | -60 |
| DGPZ120 | -0.001 | -60 |
| FASS BOYE Pz | 0.0005 | 30 |
| KAYAR PS7 | 1 | -60 |
| NDEUNE | -0.001 | 60 |
| SANTHIOU SINE ICS62 | 0.0005 | -30 |
| TAWA FALL | -0.001 | 60 |