Université Paris 8 | Vincennes - Saint-Denis

Laboratoire LIASD, Ecole doctorale Cognition, Langage, Interaction - ED 224

Université de Nouakchott,Faculté des Sciences et Techniques

Unité de recherche en Calcul Scientifique, Informatique et DataScience (CSIDS)

Estimation de la qualité des cultures de riz par des systèmes intelligents

par Yahjeb Bouha Khatraty

Date de soutenance : 12 Décembre 2023

Membres du jury :

Ahmedou Haouba, Professeur des universités à l'Université de Nouakchott (Mauritanie) ; Président
Khalid Benabdeslem, Professeur associé - HDR à l'université de Lyon1 (France) ; Rapporteur
Imed-Riadh Farah, Professeur à l'Université de la Manouba (Tunisie) ; Rapporteur
Myriam Lamolle, Professeur à l'Université Paris 8 (France) ; Examinatrice
Sadok ben Yahia, Professeur à l'Université de Technologie de Tallinn (Estonie) ; Examinateur
Nédra Mellouli, Professeur associé à l'Université Paris 8 (France) ; Directrice de thèse
Mohamedade Farouk Nanne, Professeur - HDR à l'Université de Nouakchott ; Directeur de thèse
Mamadou Tourad DIALLO, Docteur à l'Université de Nouakchott ; Encadrant

Remerciements

Je tiens à exprimer ma profonde gratitude envers Madame Nedra Mellouli, Monsieur Mohamedade Farouk Nanne et Monsieur Tourad Diallo, mes directeurs de thèse et mon encadrant, pour leur soutien inestimable, leur expertise et leurs conseils éclairés qui ont été essentiels dans la réalisation de cette thèse. Leur dévouement inlassable à l'excellence académique, leur disponibilité constante et leur confiance en mes capacités ont été une source d'inspiration et de croissance personnelle tout au long de ce parcours de recherche. Leur mentorat bienveillant a façonné ma vision de la recherche et m'a encouragé(e) à repousser les limites de mes connaissances.

Je tiens également à exprimer ma profonde reconnaissance envers mes parents, Mohamed El-Mamoun et Mallassa, pour leur amour inconditionnel, leur soutien indéfectible et leurs sacrifices qui ont rendu cette thèse possible. Leur encouragement constant, leur confiance en mes capacités et leur présence réconfortante ont été des sources d'inspiration et de force tout au long de mon parcours académique. Leur soutien inébranlable a été le fondement de ma réussite.

Enfin, je souhaite exprimer ma gratitude à mes frères et sœurs pour leur soutien inconditionnel.

Résumé

La gestion agricole en Mauritanie est aujourd'hui confrontée à des enjeux cruciaux, notamment la sécheresse et la rareté de l'eau. Pour répondre à ces défis pressants, cette thèse présente une solution innovante qui apporte des contributions significatives dans ce domaine en plein essor. Au cœur de notre contribution réside une amélioration substantielle de l'algorithme DBSCAN, dédié au regroupement de séries temporelles multivariées. Cette amélioration se matérialise par l'intégration d'un epsilon adaptable et l'utilisation ingénieuse du Deep Auto Encoder pour réduire avec précision les dimensions des données. Cette combinaison offre une meilleure compréhension des tendances agricoles, permettant ainsi une gestion plus éclairée. Mais notre innovation ne s'arrête pas là. Nous avons également conçu une architecture de modèle de prédiction, spécialement calibrée pour les séries temporelles multivariées. Basé sur un réseau de neurones LSTM, ce modèle se révèle exceptionnellement performant pour anticiper les rendements agricoles, ouvrant ainsi la voie à une gestion plus efficiente des ressources agricoles. L'application concrète de notre travail se manifeste à travers la mise en place d'un système de gestion en temps réel des rizières. Ce système intègre des données provenant de capteurs Meraki et des données satellitaires, fournissant ainsi une source d'informations en continu pour une gestion proactive des cultures. De plus, notre système est respectueux de l'environnement, puisqu'il fonctionne grâce à l'énergie renouvelable, renforçant ainsi sa durabilité.

En regardant vers l'avenir, nous envisageons d'adopter la technologie LoRa pour rendre notre modèle totalement indépendant de la couverture du réseau GSM, élargissant ainsi en-

core davantage les horizons de notre solution.

Mots clés : Agriculture intelligente, Prédiction par LSTM, Clustering par DBSCAN, Reduction des données, Deep clustering

Abstract

Agricultural management in Mauritania is currently facing critical challenges, notably drought and water scarcity. In response to these pressing issues, this thesis introduces an innovative solution that brings significant contributions to this rapidly evolving field. At the core of our contribution lies a substantial enhancement of the DBSCAN algorithm, dedicated to clustering multivariate time series data. This improvement is materialized by the integration of an adaptive epsilon and the ingenious utilization of the Deep Auto Encoder to precisely reduce data dimensions. This combination provides a better understanding of agricultural trends, thus enabling more informed management. However, our innovation doesn't stop there. We have also designed an architecture for a prediction model, specifically calibrated for multivariate time series. Based on a LSTM neural network, this model proves exceptionally effective in forecasting agricultural yields, thereby paving the way for more efficient resource management. The practical application of our work is evident through the implementation of a real-time rice field management system. This system integrates data from Meraki sensors and satellite data, offering a continuous source of information for proactive crop management. Furthermore, our system is environmentally friendly as it operates on renewable energy, reinforcing its sustainability. Looking to the future, we plan to adopt LoRa technology to make our model completely independent of GSM network coverage, further expanding the horizons of our solution.

Table des matières

I Introduction et Fondements théoriques **1**

1 Digital Twins DT pour l'agriculture de précision **3**

1.1 Agriculture de précision 3

1.2 Concept des jumeaux numériques 4

1.3 Approches existantes de DT et leurs limitations 7

1.4 Formalisation du problème 13

1.5 Nouvelle architecture proposée 15

2 Données et métriques d'évaluation utilisées **18**

2.1 Données utilisées 18

2.2 Metriques d'evaluation 25

2.3 Mesures de simularité 32

II Analyse et clustering des séries temporelles **37**

3 Clustering des séries temporelles **39**

3.1 État de l'art sur le clustering des séries temporelles 39

3.2 Modèles de clustering spatio-temporel 46

3.2.1 DBSCAN 47

3.2.2 MDST-DBSCAN 51

3.2.3 Choix de ϵ adaptatif pour densité hétérogène . . . 56
3.2.4 Techniques de réduction de dimension . . . 57
3.2.5 Révision des clusters . . . 64
3.3 Évaluation comparative entre différents modèles . . . 66

4 Prédiction de la qualité et de la quantité du rendement du riz 85
4.1 Utilisation des modèles statistiques pour la prédiction . . . 85
4.2 Utilisation des modèles d'apprentissage automatique et profond pour la prediction . . . 86
4.3 Architecture du modèle utilisé . . . 96
4.4 Application des modèles . . . 100
4.4.1 Application des modéles sur des données de ROSSO . . . 100
4.4.2 Mesure de la fiabilité des données IoT de ROSSO via données satellitaires 103

III Prédiction et optimisation du rendement des rizières en Mauritanie 106

5 Interface de monitoring administrateur 108
5.1 Installation du système de monitoring . . . 108
5.1.1 Matériels utilisés . . . 109
5.2 Interface Meraki . . . 112
5.2.1 Données collectées . . . 112
5.2.2 Services fournis par Meraki . . . 113

6 Interface de monitoring utilisateur 118
6.1 Exploitation des données IoT . . . 119
6.1.1 Analyse des données provenant de dispositifs connectés à l'Internet des objets (IoT). . . . 119
6.2 Intégration des données IoT et des données satellitaires . . . 124
6.2.1 Clustering des données IoT et satellitaires . . . 126

6.2.2 Prediction du NDVI 134

Conclusion Générale **139**

Bibliographie **140**

Liste des tableaux

3.1 Récapitulatif des combinaisons de paramètres de UMAP et des scores de silhouette . . . 72
3.2 DBSCAN avec UMAP et t-SNE sur les données ECG . . . 75
3.3 Spiral - Aggregation - Flame . . . 80
3.4 Compound-ECG . . . 80
3.5 Données du Covid . . . 80
3.6 DBSCAN(Minpts=10) avec PCA, LDA et DAE sur les données ECG . . . 81
3.7 Révision du regroupement (données mal regroupées) . . . 83
3.8 Temps de calcul en secondes sans/avec Joblibspark . . . 84

4.1 Performances des modèles de prédiction . . . 102

5.1 Caractéristiques du matériel utilisé . . . 112

Table des figures

1.1 Architecture du DT du systeme de surveillance du champs agricole 17

2.1 Variation des différentes variables 20
2.2 Exemple des donnees ECG 22
2.3 Les données de Covid 19 23
2.4 le champ de l'étude à Tekane, en Mauritanie 23
2.5 Données satelittaires de la rizére à Tekane 24
2.6 Données synthétiques 25

3.1 Approche Innovante de Regroupement pour les Séries Temporelles Multivariées 44
3.2 Taxonomie des Méthodes de Regroupement Spatio-Temporel et Profond . . . 47
3.3 Schéma Descriptif de DBSCAN 48
3.4 Loss vs Epoque 63
3.5 Variation des hyperparamètres sur le jeu de données ECG 63
3.6 Variation des hyperparamètres sur le jeu de données Covid 64
3.7 Architecture de DAE pour les données ECG 64
3.8 Architecture de DAE pour les données Covid 65
3.9 dbscan sur ECG 68
3.10 DBSCAN sur des données synthétiques 70
3.11 DBSCAN avec ϵ adaptatif des hypothèses H_1, H_2 et H_3 sur les données de Compound 81

3.12 DBSCAN avec ϵ adaptatif de l'hypothèse H_1 sur d'autres ensembles de données. 81
3.13 MDST-DBSCAN classique 82
3.14 D-MDST DBSCAN 83

4.1 Architecture du modèle LSTM 97
4.2 Architecture du modèle GRU 98
4.3 Décomposition des données de l'ONM et des données satellites 105
4.4 Matrice de corrélation des variables 105

5.1 Installation du système de monitoring 111
5.2 Interface de Meraki 113
5.3 Données du capteur MT10 114
5.4 Données du capteur MT14-1 (valeurs maximales) 115
5.5 Données du capteur MT14-2 (valeurs minimales) 116
5.6 graphe economie d'energie via MT14 117

6.1 Matrice de corrélation des indices satellitaires avec les variables collectées par notre système. 126
6.2 Clustering des données satellitaires et IoT 1 130
6.3 Clustering des données satellitaires et IoT 2 131
6.4 Ensemble d'images du mois de mai et juin 2023 135
6.5 Prédiction de l'indice de végétation NDVI 136

Liste des abréviations

Abbreviation	Label
ADF	Augmented Dickey-Fuller
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DAE	Deep AutoEncoder
DT	Digital Twins
DTSC	Deep Temporal Series Clustering
ECG	Electrocardiogramme
IA	Intelligence Artificielle
IoT	Internet des objets
LDA	Linear Discriminant Analysis
LSTM	Long Short-Term Memory
LSH	Locality-Sensitive Hashing
NDMI	Normalized Difference Moisture Index
NDSI	Normalised Difference Snow Index
NDVI	Normalized Difference Vegetation Index
NDWI	Normalized Difference Water Index
ONM	Office National de la Météorologie
PCA	Principal Component Analysis
t-SNE	t-Distributed Stochastic Neighbor Embedding
TVOC	Total Volatile Organic Compounds
TIC	Technologies de l'information et de la communication
UMAP	Uniform Manifold Approximation and Projection

Première partie

Introduction et Fondements théoriques

Chapitre 1

Digital Twins DT pour l'agriculture de précision

Dans ce chapitre introductif, nous plongerons dans les concepts fondamentaux de l'agriculture de précision et des jumeaux numériques, tout en explorant les approches existantes et en formalisant le problème qui guidera notre recherche. En outre, nous présenterons notre nouvelle architecture proposée pour l'optimisation de ces domaines.

1.1 Agriculture de précision

L'agriculture de précision représente une révolution majeure dans le secteur agricole grâce à l'intégration de technologies numériques avancées. Son objectif principal est d'optimiser la gestion des exploitations agricoles en utilisant des solutions innovantes basées sur les données et l'analyse, visant à accroître l'efficacité, la productivité et la durabilité tout en réduisant l'impact environnemental. Face aux défis croissants de la demande alimentaire mondiale, la diminution des ressources naturelles et les pressions environnementales, l'agriculture de précision se présente comme une solution prometteuse en exploitant les technologies de l'information et de la communication (TIC) pour améliorer les pratiques agricoles Mizik (2023).

Cette approche intègre de nombreuses tendances émergentes et avancées de recherche. L'utilisation de l'Internet des objets (IoT) et des capteurs connectés joue un rôle crucial en permettant aux agriculteurs de surveiller en temps réel divers paramètres tels que l'humidité

du sol, la température, l'humidité de l'air et la qualité de l'air, ce qui permet de prendre des décisions éclairées sur la gestion des cultures Fathy and Ali (2023), Hundal et al. (2023), Saranya et al. (2023), Senapaty et al. (2023).

Parallèlement, l'analyse des données massives (Big Data) est un pilier essentiel de l'agriculture de précision. Elle permet de comprendre les tendances et les relations complexes entre les facteurs environnementaux et les rendements des cultures, favorisant ainsi l'optimisation des pratiques agricoles Melgar-García et al. (2022), Osinga et al. (2022), San Emeterio de la Parte et al. (2023). De plus, l'utilisation croissante de l'intelligence artificielle (IA) joue un rôle clé en résolvant des problèmes complexes et en automatisant certaines tâches, telles que la prédiction des rendements des cultures, l'identification des maladies des plantes, la recommandation de stratégies de gestion des cultures et l'optimisation des ressources Foster et al. (2023), Javaid et al. (2023), Sharma et al. (2023).

En outre, l'agriculture de précision a permis le développement de systèmes de pulvérisation intelligents qui utilisent des technologies de reconnaissance d'image pour identifier et pulvériser de manière ciblée les mauvaises herbes et les ravageurs, réduisant ainsi la quantité d'intrants utilisés et contribuant à une agriculture plus durable Nasir et al. (2023).

En explorant ces avancées, nous comprenons comment l'agriculture de précision redéfinit la manière dont nous cultivons nos terres et prenons soin de nos cultures. Pour aller plus loin dans cette approche, nous examinerons maintenant le concept passionnant des jumeaux numériques et son rôle clé dans l'agriculture de précision.

1.2 Concept des jumeaux numériques

Le concept des jumeaux numériques ou Digital Twins (DT) est une approche novatrice qui gagne en popularité dans le domaine de l'agriculture, notamment en raison des avancées technologiques dans le domaine de l'informatique et de la modélisation. Les jumeaux numériques sont des représentations virtuelles d'objets agricoles réels, créées à partir de données en temps réel ou de données historiques, et gérées à l'aide d'outils informatiques avancés

Liu et al. (2023).

Dans le contexte de l'agriculture, les jumeaux numériques sont utilisés pour modéliser divers éléments, tels que des cultures, des animaux, des équipements agricoles, des sols, des conditions météorologiques, etc. Ils permettent de recueillir, d'analyser et de visualiser les données agricoles de manière plus efficace et précise, ce qui peut aider les agriculteurs à prendre des décisions éclairées pour optimiser leurs opérations agricoles Cesco et al. (2023), Purcell et al. (2023).

Les jumeaux numériques peuvent être créés à l'aide de techniques de modélisation 3D, de simulation et d'apprentissage automatique. Ces modèles virtuels sont capables de représenter de manière réaliste la géométrie, la structure et le comportement des objets agricoles, en intégrant diverses données, telles que des données géospatiales, des données météorologiques, des données de capteurs agricoles, des données historiques sur les cultures, etc Chen et al. (2023), Neethirajan and Kemp (2021).

L'utilisation des jumeaux numériques dans l'agriculture offre de nombreux avantages. Tout d'abord, ils permettent de simuler et d'optimiser des scénarios agricoles avant leur mise en œuvre sur le terrain, ce qui peut aider à minimiser les risques et les coûts associés à l'expérimentation réelle. En utilisant des modèles virtuels, les agriculteurs peuvent tester différentes stratégies de gestion, optimiser l'utilisation des ressources, prédire les rendements des cultures et évaluer les impacts potentiels des changements environnementaux Menon et al. (2023).

De plus, les jumeaux numériques offrent la possibilité de surveiller et de contrôler les conditions agricoles en temps réel. En intégrant des capteurs et des systèmes de surveillance dans les modèles virtuels, les agriculteurs peuvent obtenir des informations en temps réel sur les conditions météorologiques, les niveaux d'humidité, les taux de croissance des cultures, etc. Cela leur permet de prendre des décisions éclairées pour optimiser la gestion des cultures, notamment en ajustant les niveaux d'irrigation, les pratiques de fertilisation, etc Atalay (2023), Cesco et al. (2023), Slob et al. (2023) .

Par ailleurs, les jumeaux numériques peuvent également contribuer à optimiser la planifi-

cation et la logistique agricole. En intégrant des modèles virtuels dans les systèmes de gestion agricole, les agriculteurs peuvent optimiser les itinéraires de transport, les calendriers de plantation, les stratégies de récolte, etc. Cela peut permettre de réduire les coûts de transport, de minimiser les pertes de récolte et d'optimiser l'utilisation des ressources, ce qui peut avoir un impact positif sur la rentabilité globale de l'exploitation agricole et aussi d'un point de vue écologique. Melesse et al. (2023), Singh et al. (2023).

Les jumeaux numériques peuvent également être utilisés pour faciliter la communication et la collaboration entre les différents acteurs de la chaîne d'approvisionnement agricole. En partageant les modèles virtuels avec les partenaires de la chaîne d'approvisionnement, tels que les fournisseurs, les distributeurs, les transformateurs, etc., il est possible d'améliorer la coordination des activités, la traçabilité des produits et la transparence tout au long de la chaîne d'approvisionnement Melesse et al. (2023).

Enfin, les jumeaux numériques peuvent contribuer à améliorer la durabilité et la gestion environnementale dans l'agriculture. En permettant une gestion plus précise et ciblée des ressources agricoles, telle que l'eau, les fertilisants et les pesticides Cesco et al. (2023), Silva et al. (2023), les jumeaux numériques peuvent contribuer à réduire la consommation excessive de ressources peu disponibles et minimiser les impacts environnementaux négatifs. De plus, en utilisant des modèles virtuels pour prédire les impacts potentiels des changements climatiques et d'autres facteurs environnementaux sur les cultures, les agriculteurs peuvent prendre des mesures proactives pour s'adapter et atténuer les risques Maksimović (2023).

En somme, les jumeaux numériques sont une approche innovante qui offre de nombreuses opportunités pour améliorer la gestion agricole en utilisant des modèles virtuels basés sur des données en temps réel ou historiques. Ils permettent d'optimiser les opérations agricoles, de prendre des décisions éclairées, de faciliter la communication et la collaboration, et de favoriser la durabilité environnementale. Dans les sections suivantes de ce chapitre, nous explorerons plus en détail les différentes composantes et les applications spécifiques des jumeaux numériques dans le contexte de l'agriculture, ainsi que les défis et les perspectives futures de cette approche prometteuse.

1.3 Approches existantes de DT et leurs limitations

Selon les recherches de Singh et al. (2022), bien que les jumeaux numériques soient largement utilisés dans d'autres secteurs commerciaux, leur implémentation dans l'agriculture demeure encore limitée. La plupart des propositions pour les jumeaux numériques en agriculture se trouvent encore au stade conceptuel ou au stade de prototype, comme l'a mentionné Pylianidis et al. (2021). Malgré les avantages potentiels qu'ils offrent, tels que la simulation de scénarios, la prédiction des performances des cultures, l'optimisation des ressources et la prise de décision en temps réel, les jumeaux numériques ne sont pas encore largement utilisés dans le secteur agricole. Cela peut être dû à plusieurs défis, notamment la complexité des données agricoles, les coûts associés à la mise en place de systèmes de jumeaux numériques et les préoccupations liées à la confidentialité et à la sécurité des données agricoles. Néanmoins, malgré ces limitations, l'intérêt pour les jumeaux numériques en agriculture continue de croître et de nombreuses recherches et expérimentations sont en cours pour développer leur utilisation dans le secteur agricole et permettre ainsi une meilleure gestion des cultures et une optimisation des rendements agricoles.

L'étude de Moghadam et al. (2020) a apporté une contribution significative à l'utilisation des jumeaux numériques en agriculture, en se concentrant spécifiquement sur le suivi de la croissance des plantes. Dans cette recherche, un dispositif innovant appelé AgScan3D+ a été développé, utilisant la technologie Lidar 3D et des caméras pour surveiller la croissance des arbres et la production de fruits. Grâce à ce dispositif, les chercheurs ont été en mesure de créer un jumeau virtuel pour 15 000 arbres, permettant ainsi une modélisation détaillée de certaines caractéristiques clés de la canopée.

En se basant sur la densité et la distribution des feuilles, le jumeau virtuel développé dans cette étude a permis de modéliser des paramètres importants tels que la densité des feuilles et la capacité de pénétration de la lumière dans la canopée. Cette modélisation précise a permis aux chercheurs de mieux comprendre la structure et le développement des arbres, ainsi que leur impact sur la production de fruits. Par exemple, en analysant les données col-

lectées à partir du jumeau virtuel, les chercheurs ont pu identifier les zones de la canopée qui étaient sous-exposées à la lumière et donc moins propices à la production de fruits. Ces informations ont été utilisées pour optimiser la gestion des cultures en ajustant les pratiques agronomiques pour améliorer la production.

L'utilisation de la technologie Lidar 3D et des caméras dans le dispositif AgScan3D+ a permis de collecter des données tridimensionnelles hautement précises sur les arbres, ce qui a permis de créer un jumeau virtuel réaliste et détaillé. Cela démontre le potentiel des jumeaux numériques en agriculture pour fournir des informations précieuses sur la croissance des plantes, la santé des cultures et l'efficacité des pratiques agricoles. Bien que cette étude soit encore un exemple isolé de l'utilisation des jumeaux numériques en agriculture, elle met en évidence les possibilités passionnantes de cette technologie pour améliorer la gestion des cultures et optimiser les rendements agricoles.

Alves et al. (2019) a proposé une approche novatrice pour les systèmes de surveillance à distance des champs agricoles en utilisant la technologie développée dans les projets Sensing Change et Smart Water Management Platform pour créer un jumeau numérique spécifiquement adapté à l'agriculture. Ce système utilise des sondes pour recueillir des données sur la composition du sol, telles que l'humidité, la température. Cela permet aux agriculteurs de surveiller à distance les conditions du sol et de mieux comprendre les besoins de leurs cultures.

Cependant, il convient de noter que ce système ne comprend pas la phase de traitement ou la prise de décision dans sa plateforme. Cela signifie que bien que les données sur la composition du sol soient collectées et affichées, le système ne propose pas d'analyses ou de recommandations automatisées pour aider les agriculteurs à prendre des décisions éclairées sur la gestion de leurs cultures. Les agriculteurs devront donc interpréter les données eux-mêmes et prendre des décisions basées sur leur propre expertise.

D'un autre côté, l'étude menée par Skobelev et al. (2020) a développé un jumeau numérique dans le but de surveiller la qualité de la production végétale et de suivre leur croissance pendant la phase de développement. Ce jumeau numérique s'appuie sur l'utilisation de données

météorologiques, telles que la température, l'humidité, l'ensoleillement, ainsi que sur des ontologies et des technologies multi-agents pour modéliser la croissance des plantes. Cela permet de prédire et d'évaluer la qualité de la production végétale en temps réel, ce qui peut être très utile pour les agriculteurs dans la gestion de leurs cultures.

Dans un autre domaine d'application, notamment dans le contexte de l'agriculture urbaine, Jans-Singh et al. (2020) a proposé l'utilisation de jumeaux numériques pour gérer l'environnement intérieur d'une ferme hydroponique. Cette approche permet d'ajuster automatiquement la température et l'humidité en fonction des besoins des cultures, tout en optimisant l'utilisation de la lumière, un facteur clé de la croissance des plantes en milieu hydroponique. Pour cela, 89 variables ont été collectées à l'aide de capteurs, permettant ainsi de surveiller de près les conditions environnementales à l'intérieur de la ferme.

D'un autre côté, Ghandar et al. (2021) a proposé une approche novatrice de jumeau numérique pour apporter une assistance précieuse aux agriculteurs dans leur processus décisionnel en matière de gestion de leurs systèmes aquaponiques. Cette approche repose sur l'utilisation de modèles d'apprentissage automatique pour créer un modèle numérique qui simule les caractéristiques, les interactions et les dynamiques des systèmes aquaponiques réels des agriculteurs.

Le concept de jumeaux numériques suscite un intérêt croissant dans le domaine des greenhouses. En utilisant des modèles et des simulations basés sur des paramètres environnementaux tels que la température, l'humidité, la lumière et d'autres facteurs, les gestionnaires de greenhouses peuvent mieux comprendre les conditions de culture, optimiser les ressources et prendre des décisions éclairées pour améliorer la performance de la greenhouse. Les jumeaux numériques peuvent être utilisés pour optimiser la consommation d'énergie en simulant les besoins en chauffage, en ventilation et en éclairage de la greenhouse, en utilisant des données réelles pour ajuster les paramètres en conséquence. De même, en utilisant des modèles de croissance des plantes basés sur des données réelles, les gestionnaires de greenhouses peuvent simuler et anticiper les besoins en eau, en nutriments et en pesticides, ce qui peut contribuer à une utilisation plus efficace des ressources et à une augmentation de

la productivité des cultures. En somme, l'utilisation de jumeaux numériques offre des perspectives prometteuses pour optimiser la consommation d'énergie, accroître la productivité des cultures et améliorer la durabilité des opérations en greenhouse, tel que proposé par Howard et al. (2020).

Tsolakis et al. (2019) a évalué les performances et la compatibilité d'un véhicule terrestre autonome avec l'architecture d'une exploitation agricole à l'aide d'un simulateur. Cette évaluation a fourni aux exploitants agricoles des informations cruciales pour prendre une décision éclairée concernant l'achat de ce type d'équipement. Les résultats ont permis aux exploitants de comprendre les avantages potentiels de cette technologie, notamment la réduction des coûts de main-d'œuvre et de carburant, ainsi que l'amélioration de la qualité du travail effectué dans les champs. Cette étude a démontré comment l'évaluation des performances et de la compatibilité d'un véhicule terrestre autonome peut aider les exploitants agricoles à prendre des décisions idoines quant à son adoption.

Une étude récente Angin et al. (2020) a démontré l'utilité des jumeaux numériques dans plusieurs aspects de l'agriculture. Ils permettent notamment d'améliorer la détection des maladies, la gestion des mauvaises herbes et la nutrition des plantes. En surveillant en temps réel la santé des cultures, les jumeaux numériques permettent une détection précoce des maladies et des infections, ce qui facilite la mise en place de mesures de prévention et de traitement plus rapides, et entraîne une augmentation du rendement global des cultures. Par ailleurs, en simulant la croissance des mauvaises herbes dans un environnement virtuel, il est possible de prédire leur propagation et d'adopter des stratégies de gestion plus efficaces, tout en réduisant l'utilisation de produits chimiques nuisibles à l'environnement.

Dans une autre étude menée par Rawankar et al. (2018), les jumeaux numériques sont utilisés comme outil de régulation de la fertilisation. En exploitant les données en temps réel, ces jumeaux permettent d'ajuster les pratiques agricoles de manière plus précise, contribuant ainsi à une gestion plus durable des cultures.

En résumé, les jumeaux numériques présentent un potentiel prometteur pour améliorer la détection des maladies, la gestion de la croissance des plantes, le contrôle des mauvaises

herbes et l'optimisation de la nutrition des plantes en agriculture. Ils permettent de prendre des décisions éclairées et efficaces en exploitant les données en temps réel.

Par ailleurs, les jumeaux numériques peuvent également être utilisés dans le processus de stockage des récoltes. Ils facilitent la surveillance des conditions de stockage, permettant ainsi de préserver les produits sur une longue durée. En fournissant des informations en temps réel sur l'état des produits, les jumeaux numériques contribuent à minimiser les pertes. Des études telles que celle de Tagliavini et al. (2019) sur la qualité des mangues lors du stockage et celle de Kampker et al. (2019) sur la dégradation des pommes de terre lors de la récolte démontrent l'utilité des jumeaux numériques dans ce domaine.

Les travaux antérieurs ont souligné plusieurs limitations importantes des approches classiques de jumeaux numériques dans l'agriculture. Parmi ces limitations, on retrouve la complexité et la précision des modèles mathématiques utilisés. Les modèles traditionnels peuvent se révéler complexes à mettre en place et à calibrer, ce qui nécessite une expertise technique avancée. De plus, la précision de ces modèles peut être limitée, ce qui peut entraîner des erreurs dans les prédictions et les recommandations faites aux agriculteurs.

Une autre limitation des approches traditionnelles est liée à la disponibilité et à la qualité des données collectées. Les données agricoles sont souvent hétérogènes et peuvent provenir de différentes sources, ce qui rend difficile leur intégration et leur analyse. De plus, la qualité des données peut varier, ce qui peut entraîner des incohérences et des incertitudes dans les résultats des jumeaux numériques.

En outre, les approches traditionnelles peuvent être limitées dans leur adaptation à différents types de cultures ou de systèmes agricoles. Elles peuvent nécessiter des infrastructures coûteuses et des technologies complexes pour leur mise en œuvre, ce qui peut limiter leur applicabilité dans les zones rurales où les ressources sont souvent limitées.

Cependant, notre travail propose une architecture novatrice de système de monitoring, basée sur des sources d'énergie renouvelable, notamment les panneaux solaires. Cette approche assure une alimentation autonome du système de monitoring et permet également de prédire le rendement des cultures en prenant en compte la fiabilité des données collectées

à partir de diverses sources telles que capteurs et données satellitaires.

Pour renforcer davantage notre contribution, nous avons introduit une étape de clustering dans notre architecture. Cette étape permet de regrouper les données provenant de différentes sources en fonction de leur similarité, simplifiant ainsi leur comparaison et évaluation.

En utilisant un modèle de clustering amélioré, nous pouvons identifier plus efficacement les anomalies et les tendances dans les données agricoles collectées, ce qui améliore la compréhension des facteurs qui influencent le rendement des cultures. De plus, cela permet d'agréger naturellement des données hétérogènes, réduit les erreurs potentielles et garantit une meilleure qualité des données utilisées dans les jumeaux numériques.

Cette intégration du clustering améliore la fiabilité des informations fournies aux agriculteurs en éliminant les données aberrantes et les sources peu fiables. Ainsi, les agriculteurs bénéficient d'outils de prédiction plus précis et de confiance pour optimiser leurs pratiques agricoles, entraînant potentiellement un impact significatif sur leur productivité et leurs rendements.

En effet, dans cette thèse, nous visons à améliorer l'utilisation des jumeaux numériques dans l'agriculture en combinant un système de surveillance en temps réel des champs agricoles, des modèles de deep learning pour la prédiction du rendement du riz, ainsi qu'une combinaison de modèles de machine learning et de deep learning pour le clustering des données massives du champ agricole.

Tout d'abord, grâce à un système de surveillance en temps réel des champs agricoles, nous collectons une grande quantité de données, telles que les conditions météorologiques, le niveau d'eau, la nutrition des sols, la croissance des plantes, etc. Ces données sont essentielles pour comprendre l'environnement dans lequel le riz pousse et pour prendre des décisions éclairées en matière de gestion agricole.

Ensuite, nous utilisons des modèles de deep learning pour prédire le rendement du riz en se basant sur ces données collectées. Ces modèles sont entraînés à partir de données historiques de rendements de riz, de données climatiques et de données de gestion agricole.

Les prédictions de rendement du riz obtenues permettent aux agriculteurs d'estimer le rendement attendu du riz pour une saison donnée, ce qui leur permet de prendre des décisions éclairées en matière de gestion agricole.

L'avantage de la prédiction du rendement du riz est multiple. Tout d'abord, cela permet aux agriculteurs d'optimiser leur production agricole en ajustant leurs pratiques agricoles en conséquence. Par exemple, en connaissant à l'avance le rendement attendu du riz, les agriculteurs peuvent ajuster l'irrigation, la fertilisation et la protection contre les maladies pour maximiser la production agricole. Cela peut également aider à la planification des ressources et à la gestion des coûts, en évitant les pratiques excessives ou insuffisantes qui peuvent entraîner des pertes de rendement ou des coûts inutiles.

Le modèle de clustering des données est basé sur une combinaison de techniques de machine learning et de deep learning. Les résultats obtenus à partir de ce clustering peuvent être utilisés pour interpréter les rendements prédits, permettant ainsi une visualisation claire des données massives. De plus, ce modèle est capable de détecter les tendances, les ruptures et les anomalies dans les données, ce qui en fait un outil puissant pour l'analyse de données agricoles.

1.4 Formalisation du problème

Considérons un scénario où nous avons un champ agricole dont le rendement est mesuré à différents moments dans le temps, notés $t = 0, 1, 2, ..., T$. À chaque instant t, nous mesurons n variables différentes, telles que la température, l'humidité, la vitesse du vent, le pH du sol, la consommation d'eau, et d'autres variables qui peuvent influencer le rendement du champ. Pour représenter ces mesures, nous utilisons une série temporelle multivariée X_t, définie comme suit :

$$X_t = x_{1,t}, x_{2,t}, ..., x_{n,t} \tag{1.1}$$

où $x_{1,t}, x_{2,t}, ..., x_{n,t}$ sont les valeurs des variables à l'instant t caractérisant l'observation aléatoire X liée à un capteur donné. De plus, nous supposons que le rendement du champ à

l'instant t, noté Y_t, est également mesuré, reflétant l'état du champ agricole étudié.

Dans ce contexte, nous avons mis en place un système de surveillance qui collecte des données sur ces variables à chaque instant t et propose des actions en fonction de l'état du champ agricole. Nous utilisons ensuite ces données collectées pour prédire le rendement du champ dans le futur, noté $Y_{t+\delta t}$, où δt est un intervalle de temps d'intérêt spécifié souvent lié au champs étudié. La prédiction du rendement dépend des valeurs actuelles des variables X_t, des valeurs futures des variables $X_{t+\delta t}$, ainsi que du rendement actuel Y_t. Nous utilisons une fonction f pour modéliser cette relation de prédiction :

$$Y_{t+\delta t} = f(x_{i,(t_0,t_1,\ldots,t+\delta t)}, Y_t) \tag{1.2}$$

où $x_{i,(t_0,t_1,\ldots,t+\delta t)}$ représente les valeurs des variables x_i sur l'intervalle de temps t_0 à $t + \delta t$.

Nous avons observé la dépendance du rendement d'une part aux différentes saisons et d'autres part à la variation des variables elles-mêmes en fonction des saisons. En outre, pour mieux comprendre ces dépendances, nous utilisons une approche de regroupement, également appelée clustering temporel correspondant à une zone spatiale associée aux coordonnées géographiques de la région agricole fixée. Nous regroupons les différentes variables $x_{1,t}, x_{2,t}, ..., x_{n,t}$ ainsi que le rendement Y_t en K clusters, où K est un nombre prédéfini. L'objectif est de créer des groupes temporels de variables et de rendements similaires et homogènes à l'intérieur de chaque cluster, tout en étant différents des autres clusters. La mesure de similarité choisie devra prendre en considération la variation temporelle des variables.

L'objectif de notre thèse est de développer un système de monitoring du rendement d'un champ agricole capable d'optimiser la qualité du rendement dans des conditions extrêmes (manque d'eau, fertilisation du sol, environnement aride et forte dépendance alimentaire). Ce système se nourrit des données de capteurs en temps réel (de séries temporelles multivariées) tout en assurant la fiabilité des valeurs collectées.

Nous regroupons également les variables similaires en clusters homogènes en utilisant

des techniques compatibles avec les séries temporelles. Cette démarche nécessite la collecte précise de données provenant de différentes sources. De plus, notre approche intègre des mécanismes de contrôle qualité visant à garantir la fiabilité des données, ce qui améliore la précision des prédictions et permet aux agriculteurs d'optimiser leurs pratiques agricoles.

1.5 Nouvelle architecture proposée

En général, l'architecture des jumeaux numériques du système de surveillance des champs agricoles est standard et se compose de quatre couches principales. La première couche est la couche capteurs, intégrant divers capteurs ou objets intelligents utilisés pour la collecte et le suivi des données agricoles. Ces capteurs peuvent mesurer des paramètres tels que la température, l'humidité du sol, la pluviométrie, la luminosité et bien d'autres encore.

La deuxième couche est la couche réseaux, qui comprend les technologies responsables de la transmission des données collectées par les capteurs vers les systèmes de traitement et d'analyse. Les technologies de communication utilisées peuvent inclure le Wi-Fi, le Bluetooth, les réseaux cellulaires, ou d'autres protocoles de communication adaptés aux besoins spécifiques du système de surveillance agricole.

La troisième couche est la couche services, qui est responsable du traitement et de l'analyse des données collectées. Cette couche peut inclure des systèmes de stockage des données, des outils d'analyse des données, des algorithmes d'apprentissage automatique, et d'autres technologies nécessaires pour transformer les données brutes collectées en informations utiles pour les agriculteurs, comme des prévisions météorologiques, des modèles de croissance des cultures, des recommandations d'irrigation, etc.

Enfin, la quatrième couche est la couche application, qui comprend une interface utilisateur permettant d'afficher les données collectées et les résultats des traitements effectués. Les agriculteurs peuvent accéder à ces informations à partir d'un ordinateur, d'une tablette ou d'un smartphone, ce qui leur permet de prendre des décisions éclairées pour la gestion de leurs cultures. Les applications peuvent également permettre aux agriculteurs de confi-

gurer des alertes, de programmer des tâches de surveillance automatique, et d'interagir avec d'autres fonctionnalités du système de surveillance agricole.

En somme, l'architecture des digital twins du système de surveillance des champs agricoles se compose généralement de ces quatre couches : capteurs, réseaux, services et applications, qui travaillent ensemble pour collecter, transmettre, traiter et afficher les données agricoles de manière efficace et conviviale pour les agriculteurs.

Dans le cadre de cette étude, nous avons développé une architecture de digital twin pour un système de surveillance en temps réel des champs agricoles. Cette architecture est composée de plusieurs couches qui permettent la collecte, le traitement et l'analyse de données agricoles.

La première couche, la couche capteurs, est constituée de capteurs et d'objets qui permettent la collecte de données sur diverses variables telles que la température, l'humidité de l'air, l'humidité du sol, l'humidité des plantes, la salinité du sol, le niveau d'eau dans les plantes, la vitesse du vent (Bruit ambiant), Les composés organiques volatils totaux TVOC et la prise de vue de plantes à différents stades de croissance sous la forme d'images JPEG ou de petites séquences vidéos.

La couche transmission utilise la technologie de communication 4G pour transmettre les données collectées vers le cloud, sachant que les opérateurs sénégalaises offrent une bonne couverture dans la zone d'étude.

La couche service se charge du traitement des données collectées. Nous avons appliqué plusieurs traitements, tels que l'analyse de la stationnarité des données, la corrélation entre les différentes variables et la décomposition additive des données en trois composantes (tendance, saisonnalité et bruit). La couche service comprend également des modèles de prédiction du rendement et de clustering innovants. Nous avons utilisé des techniques de machine et de deep learning pour développer ces modèles, qui permettent de prédire avec précision le rendement de la période de culture et de regrouper les champs agricoles en fonction de leurs caractéristiques.

Enfin, la couche application permet la gestion et la visualisation des données collectées.

Nous avons proposé une plateforme qui affiche les informations de manière continue et qui permet de prédire le rendement de la période de culture en se basant sur la surface cultivée, les données météorologiques collectées et les indices de végétation tels que le NDVI, le NDWI, le NDSI, le NDMI et la salinité du sol collectés par les satellites MODIS et Landsat.

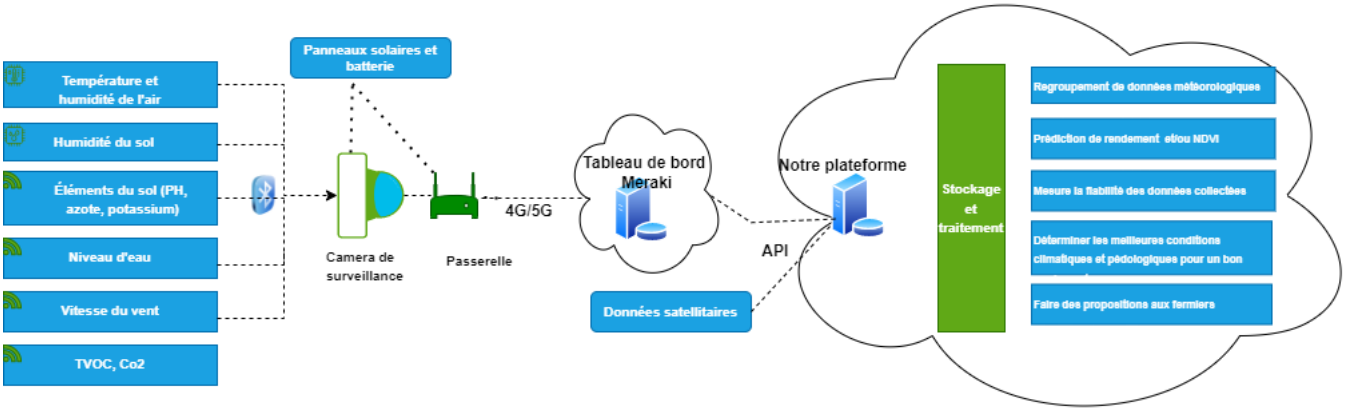


FIGURE 1.1 – Architecture du DT du systeme de surveillance du champs agricole

La Figure 1.1 offre une vue d'ensemble de notre système de surveillance. À ce stade, nous prévoyons d'utiliser des capteurs, des caméras, la plateforme Cisco Meraki, ainsi que des données satellitaires provenant du site web apps.sentinel-hub.com.

Chapitre 2

Données et métriques d'évaluation utilisées

Ce chapitre se concentre sur les données cruciales qui alimentent notre recherche ainsi que les métriques d'évaluation essentielles qui nous permettront de mesurer le succès de notre approche. Nous examinerons également les mesures de similarité qui sous-tendent notre méthodologie.

2.1 Données utilisées

Notre étude se concentre sur la Mauritanie, un pays d'Afrique subsaharienne caractérisé par un climat chaud persistant tout au long de l'année, une sécheresse persistante et une pénurie d'eau. Dans cette région, il est essentiel d'adopter une gestion efficace des ressources agricoles pour optimiser l'utilisation des ressources disponibles.

Cependant, étant donné que notre système n'a pas encore été déployé, nous devons nous appuyer sur des données collectées sur plusieurs années pour entraîner et tester nos modèles. Ces modèles seront utilisés pour prédire les rendements, regrouper les données et fournir des recommandations aux agriculteurs.

Obtenir des données précises pour effectuer des prédictions, regrouper et analyser les données agricoles, en particulier pour notre zone d'étude, représente un objectif qui ne peut être atteint qu'une fois notre système de surveillance déployé.

Dans un premier temps, nous avons consulté les informations disponibles sur le site du

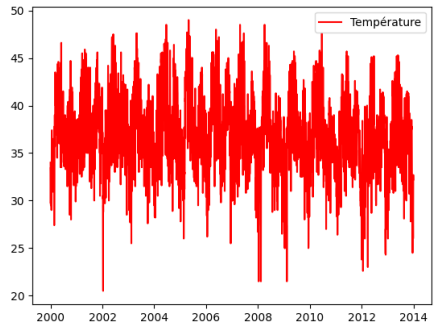
ministère de l'Agriculture mauritanien. Nous avons trouvé des données concernant les superficies cultivées en hectares, la production brute en tonnes et l'évolution des rendements (T/H) selon les différentes typologies et types de cultures. Cependant, ces données représentent en réalité la moyenne de la productivité de toutes les rizières de la Mauritanie, ce qui ne reflète pas précisément les rendements réels ni la quantité d'engrais gaspillée. Même si ces données ne donnent que la moyenne générale des valeurs, nous les avons utilisées pour entraîner et tester nos modèles de prédiction du rendement et de regroupement des données.

Nous avons également recherché des données météorologiques auprès de l'Office National de la Météorologie ONM (Figure 2.1). Cependant, les données disponibles se limitaient généralement au département de Rosso, sans refléter exactement le climat des champs agricoles, qu'ils soient proches ou éloignés du fleuve Sénégal. Dans ce contexte, il est crucial de mettre en place un système de surveillance pour chaque champ agricole afin d'obtenir des données précises.

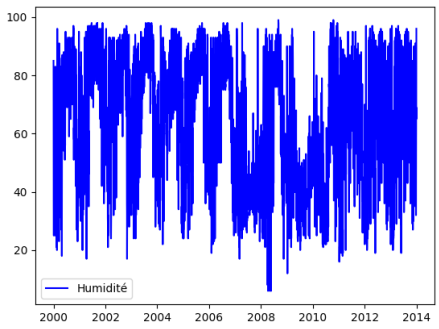
Enfin, nous avons également exploré l'utilisation de données satellitaires pour obtenir une vision plus précise de l'état des champs agricoles. Parmi les indicateurs utilisés, nous avons examiné l'indice de végétation par différence normalisée (NDVI), qui permet d'évaluer la santé et la couverture végétale, l'indice de l'eau par différence normalisée (NDWI), qui reflète la présence des plans d'eau, ainsi que l'indice d'humidité des plantes (NDMI) et l'indice de salinité du sol, qui fournissent des informations sur l'humidité et la salinité respectivement.

En général, les données collectées par les capteurs se présentent sous forme de séries temporelles multivariées. Or pour valider nos modèles, il est indispensable de les évaluer d'abord sur des données de la littérature utilisées pour évaluer des tâches similaires aux notre et dans un contexte de données massives. Pour cette raison, nous avons utilisé des données de séries temporelles multivariées telles que les données de l'ECG (électrocardiogramme) et les données de la Covid-19. Ces données ont été collectées à partir du site web Kaggle. Nous avons utilisé ces données pour entraîner nos modèles sur des données similaires à celles qui seront utilisées à l'avenir.

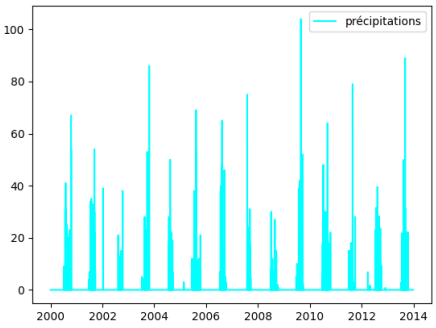
Les données de séries temporelles multivariées de l'ECG utilisées dans cette étude pro-



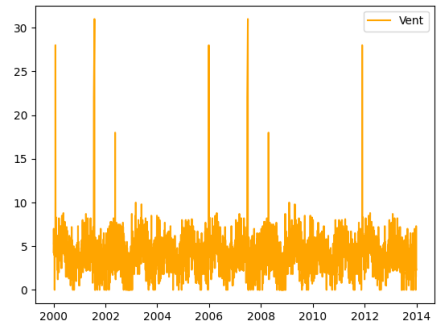
(a) Temperature



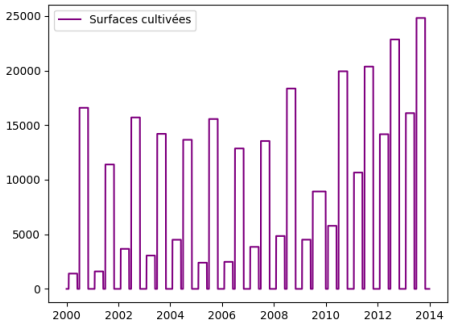
(b) Humidité



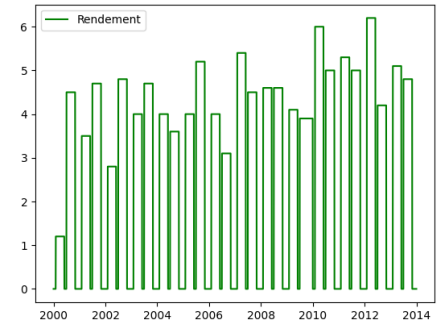
(c) Précipitation



(d) Vitesse du vent



(e) Surface cultivée



(f) Rendement du riz

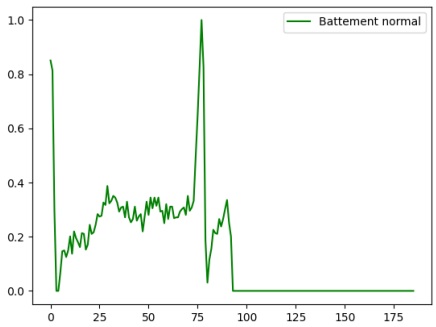
FIGURE 2.1 – Variation des différentes variables

viennent de deux ensembles de signaux cardiaques : la base de données MIT-BIH dédiée à l'analyse des arythmies et la base de données PTB utilisée pour le diagnostic de l'ECG. Ces signaux reflètent les schémas électrocardiographiques des battements cardiaques, qu'ils soient normaux ou présentant des anomalies. La base de données est composée de cinq catégories distinctes, incluant les battements normaux, les battements non identifiés, les battements ectopiques ventriculaires, les battements ectopiques supra-ventriculaires et les battements de fusion (Figure 2.2). Afin de garantir une représentation équilibrée, des quantités équivalentes de données ont été sélectionnées pour chacun de ces cinq clusters de l'ECG.

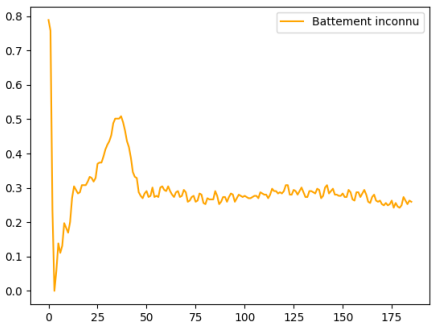
Les données de Covid-19 sont des séries temporelles provenant d'un vaste ensemble de patients anonymisés fourni par le gouvernement mexicain. Les données comprennent des informations sur le sexe, l'âge, les résultats des tests Covid, les antécédents médicaux (diabète, hypertension, etc.), le type de prise en charge (retour à domicile ou hospitalisation) et même la date de décès en cas de décès du patient. Ces données permettent de suivre l'évolution de la maladie dans le temps et d'analyser les facteurs de risque associés (Figure 2.3).

La zone d'étude dans laquelle nous avons décidé d'installer notre système est une rizière de 50 hectares située à Tekane, un village mauritanien dans le département de Rosso. Nous avons collecté plusieurs indices pour évaluer l'état de rendement du champ agricole au cours des 5 années précédentes : le NDVI (Indice de Végétation par Différence Normalisée) qui reflète l'état de la végétation, le NDWI (Indice de l'Eau par Différence Normalisée) pour mesurer les plans d'eau, et le NDMI (Indice de l'Humidité par Différence Normalisée) pour évaluer la sécheresse dans la rizière. De plus, nous avons utilisé l'indice de salinité du sol pour mesurer la quantité de sel présente dans les sols, car la salinisation des sols est l'un des processus de dégradation des terres les plus courants (Figure 2.5). Les coordonnées de la rizière sont 16°36'28.3"N 15°16'48.4"W, (16.607873, -15.280107), Mauritanie (Figure 2.4).

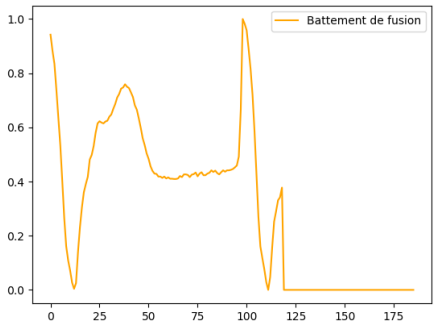
Pour évaluer les modèles de clustering proposés dans notre étude, nous avons utilisé quatre jeux de données synthétiques (Figure 2.6). Chacun de ces jeux de données présente



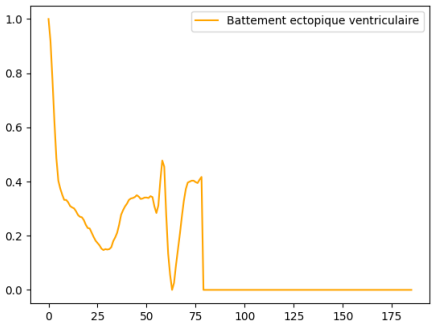
(a) Battement normal



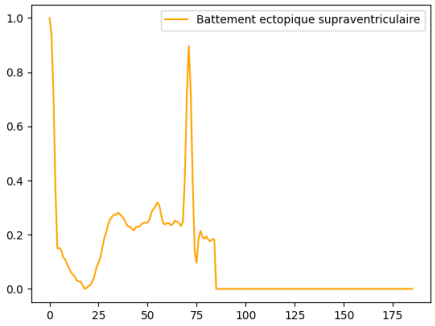
(b) Battement inconnu



(c) Battement de fusion



(d) Battement ectopique ventriculaire



(e) Battement ectopique supraventriculaire

FIGURE 2.2 – Exemple des donnees ECG

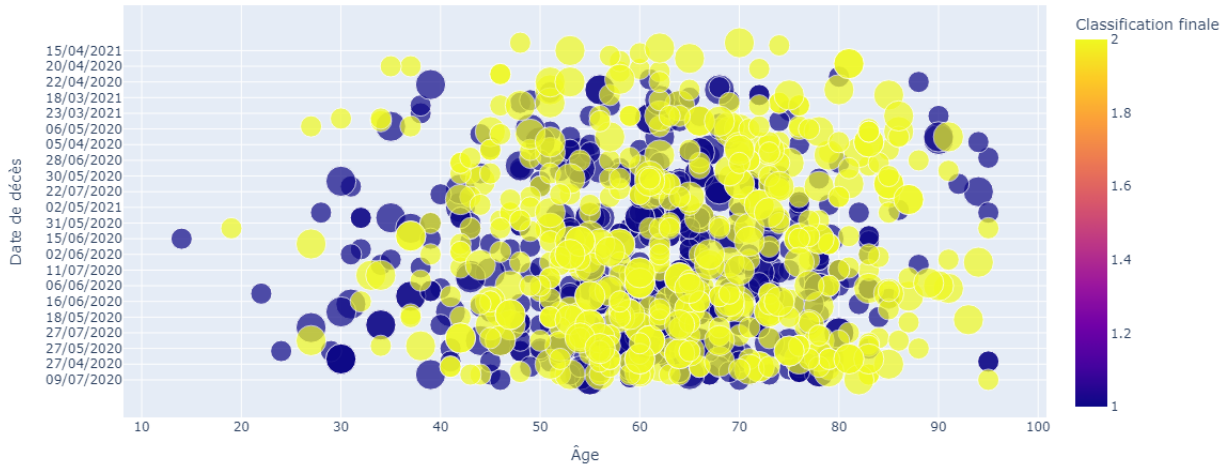


FIGURE 2.3 – Les données de Covid 19

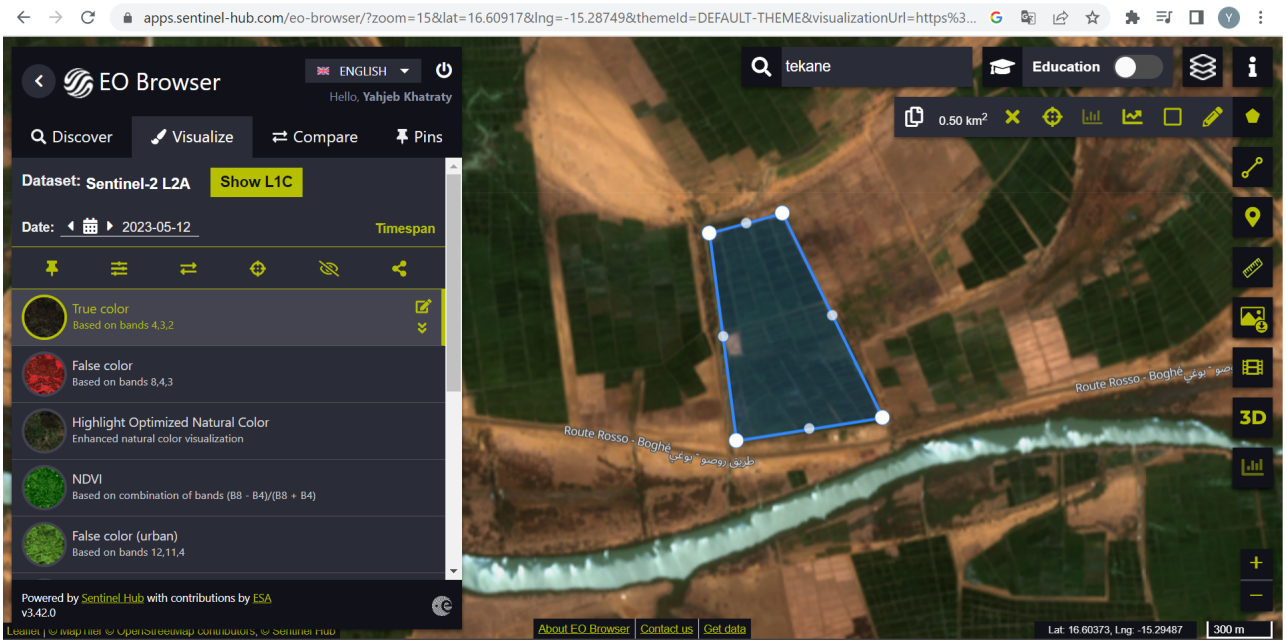
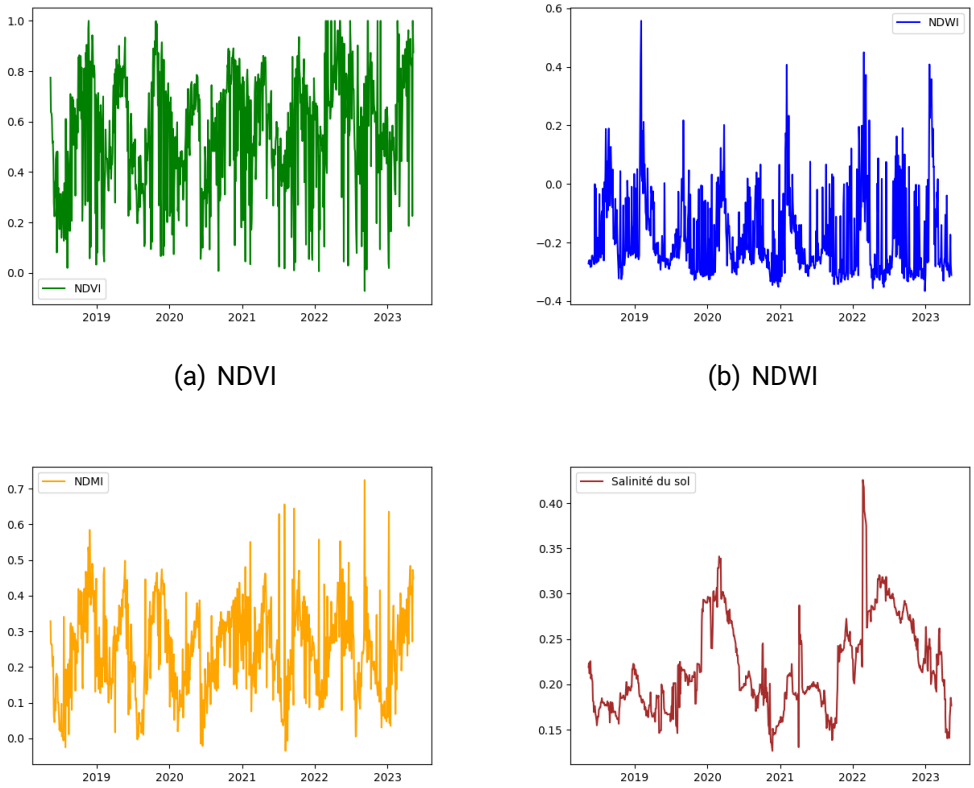


FIGURE 2.4 – le champ de l'étude à Tekane, en Mauritanie

des clusters avec différentes densités, ce qui nous a permis d'évaluer nos hypothèses de recherche l'efficacité de notre modèle face à des distributions de données variées. Les jeux de données synthétiques que nous avons utilisés sont les suivants : Spiral, Flame, Aggregation et Compound.

Nous avons choisi ces jeux de données synthétiques en raison de leurs caractéristiques distinctes. Par exemple, le jeu de données Spiral contient des clusters en forme de spirale,



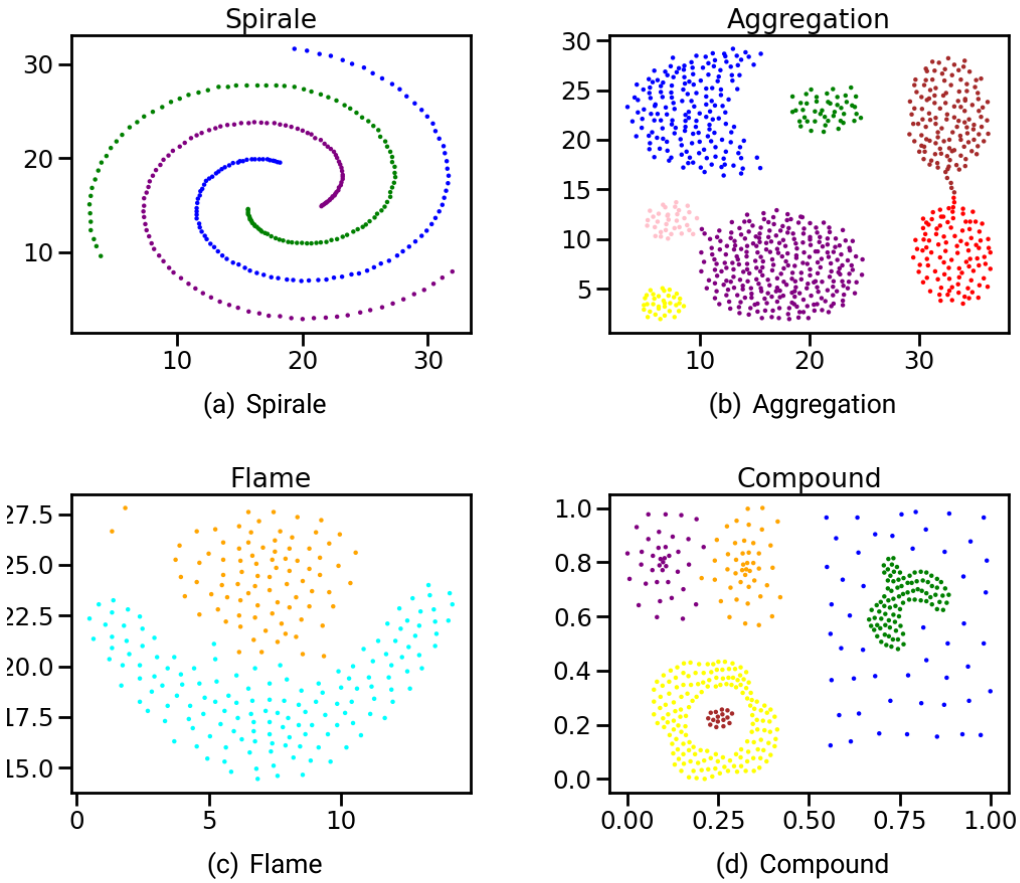
(a) NDVI (b) NDWI

(c) NDMI (d) Salinité du sol

FIGURE 2.5 – Données satelittaires de la rizére à Tekane

tandis que le jeu de données Flame présente des clusters en forme de flamme. Le jeu de données Aggregation présente des clusters bien séparés, tandis que le jeu de données Compound présente des clusters de formes complexes et interconnectés.

En utilisant ces différents jeux de données synthétiques, nous avons pu évaluer la capacité de notre modèle de clustering à détecter et regrouper correctement les différentes structures de clusters. Cette approche nous a permis d'obtenir une évaluation globale de la performance de notre modèle et de comparer son efficacité dans des scénarios de distributions variées de données.



(a) Spirale (b) Aggregation

(c) Flame (d) Compound

FIGURE 2.6 – Données synthétiques

2.2 Metriques d'evaluation

Dans notre étude, nous avons utilisé différentes métriques d'évaluation réparties en trois phases : l'évaluation de base des données collectées, l'évaluation des modèles de prédiction et l'évaluation des modèles de clustering.

Dans la première phase de notre étude, nous avons utilisé différentes métriques pour évaluer la compréhension des données collectées. Parmi ces mesures, un élément clé était l'évaluation de la stationnarité des données, pour laquelle le test d'**ADF** (Dickey-Fuller Augmenté) a été utilisé Akbulaev and Muradzada (2024).

L'expression du test ADF est la suivante :

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \sum_{i=1}^{p} \delta_i \Delta y_{t-i} + \epsilon_t \tag{2.1}$$

Dans cette équation 2.1 :

- Δy_t représente la différence entre la valeur de la série temporelle à l'instant t et sa valeur à l'instant $t-1$ (la première différence).
- α est une constante.
- β est le coefficient de la tendance linéaire.
- γ mesure l'impact de la valeur précédente de la série, y_{t-1}, sur la variation actuelle Δy_t.
- p désigne le nombre de retards (lags) utilisés dans le modèle.
- δ_i sont les coefficients des différences entre les valeurs de la série à différents retards.
- ϵ_t est le terme d'erreur à l'instant t, qui capture le bruit ou les perturbations non expliquées.

Le test ADF permet de déterminer si une série temporelle est stationnaire ou non. L'hypothèse nulle du test est que la série temporelle possède une racine unitaire, ce qui signifie qu'elle est non stationnaire. En rejetant cette hypothèse nulle, on conclut que la série est stationnaire. Le test compare le coefficient γ à zéro pour déterminer si la série temporelle possède une racine unitaire. Si γ diffère significativement de zéro, cela suggère que la série est non stationnaire. Une série stationnaire présente des propriétés statistiques constantes au fil du temps, facilitant ainsi l'analyse et la prédiction.

Ensuite, nous avons calculé plusieurs propriétés statistiques des données, telles que la **Moyenne**, la **Variance** et la **Covariance**. La moyenne est une mesure de tendance centrale qui donne une indication de la valeur moyenne des données. La variance quant à elle mesure la dispersion des données par rapport à leur moyenne. Enfin, la covariance évalue la relation linéaire entre deux variables, indiquant si elles évoluent ensemble ou de manière opposée.

Ces métriques statistiques nous permettent de mieux comprendre les caractéristiques des données collectées, notamment leur stabilité et leurs propriétés statistiques. Cela nous aide à préparer les données pour les étapes suivantes de l'analyse et à choisir les méthodes appropriées pour les modèles de prédiction et de clustering.

De plus, dans notre étude, nous avons effectué la décomposition des séries temporelles en différentes composantes afin de mieux comprendre leurs caractéristiques. La décomposition

des séries temporelles permet d'identifier les composantes qui contribuent à la variation des données.

La première composante est la composante de **Saisonnalité** $S(t)$, qui représente les variations périodiques régulières dans les données. Elle met en évidence les schémas et les cycles récurrents présents dans la série temporelle.

La deuxième composante est la composante de **Tendance** $T(t)$, qui indique la direction générale dans laquelle les données évoluent à long terme. Elle permet d'observer les mouvements globaux ou les tendances ascendantes ou descendantes des données.

Enfin, la troisième composante est la composante de **Résidu** $R(t)$, qui représente l'erreur ou le bruit restant après avoir retiré les composantes de saisonnalité et de tendance. Cette composante contient les variations aléatoires et les informations non expliquées par les deux premières composantes.

La décomposition des séries temporelles en ces différentes composantes nous permet de mieux analyser les schémas de variation, les tendances et les fluctuations aléatoires dans les données. Cela facilite également la modélisation et la prédiction des séries temporelles en prenant en compte les différentes composantes.

En outre, nous avons mesuré la **corrélation** entre les différentes variables de nos données. La corrélation est une métrique qui quantifie la relation linéaire entre deux variables. Elle permet de déterminer dans quelle mesure les variations d'une variable sont associées aux variations d'une autre variable.

Pour mesurer la corrélation, nous avons utilisé le coefficient de corrélation de Pearson. Ce coefficient varie entre -1 et 1. Une valeur de -1 indique une corrélation négative parfaite, c'est-à-dire que les variables évoluent de manière opposée de manière linéaire. Une valeur de 1 indique une corrélation positive parfaite, c'est-à-dire que les variables évoluent de manière linéaire dans la même direction. Une valeur proche de zéro indique une corrélation faible ou nulle, où il n'y a pas de relation linéaire évidente entre les variables.

La corrélation nous permet de comprendre les relations entre les différentes variables de nos données. Elle peut nous aider à identifier les associations ou les dépendances entre les

variables, ce qui peut être utile dans la sélection des caractéristiques pour les modèles de prédiction ou pour mieux comprendre les relations causales entre les variables.

Pour évaluer la performance de nos modèles de prédiction des rendements du riz, nous avons utilisé plusieurs métriques afin de quantifier l'exactitude de nos prédictions. Ces métriques comprennent :

- L'erreur moyenne de prédiction (MFE - Mean Forecast Error) : Cette métrique mesure la différence moyenne entre les valeurs prédites et les valeurs réelles. Elle permet de déterminer si nos prédictions sont systématiquement en-dessous ou au-dessus des valeurs réelles.

$$\text{MFE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i) \tag{2.2}$$

- L'erreur absolue moyenne (MAE - Mean Absolute Error) : Cette métrique calcule la moyenne des valeurs absolues des erreurs de prédiction. Elle nous indique l'erreur moyenne entre les valeurs prédites et les valeurs réelles, sans considérer la direction de l'erreur.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{2.3}$$

- L'erreur quadratique moyenne (MSE - Mean Squared Error) : Cette métrique calcule la moyenne des carrés des erreurs de prédiction. Elle accorde plus de poids aux grandes erreurs, ce qui permet de mieux prendre en compte les erreurs importantes dans l'évaluation de la performance du modèle.

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{2.4}$$

- L'erreur quadratique moyenne racine (RMSE - Root Mean Squared Error) : Cette métrique est simplement la racine carrée de l'erreur quadratique moyenne. Elle permet d'obtenir une mesure de l'erreur de prédiction dans la même unité que la variable cible, ce qui facilite son interprétation.

$$\text{RMSE} = \sqrt{\text{MSE}} \tag{2.5}$$

Où N est le nombre d'échantillons, y_i représente la valeur réelle, et $\hat{y}_i$ représente la valeur prédite.

Ces métriques nous permettent de mesurer l'écart entre nos prédictions et les valeurs réelles, ce qui nous aide à évaluer la précision de nos modèles de prédiction des rendements du riz. En utilisant ces métriques, nous pouvons comparer différentes approches de modélisation et sélectionner le modèle qui offre les meilleures performances en termes d'exactitude des prédictions.

Dans le cadre de notre étude, l'évaluation des regroupements obtenus a été réalisée à l'aide de plusieurs métriques spécifiques. Ces métriques fournissent des mesures quantitatives pour évaluer la qualité et la performance des clusters formés par les algorithmes de clustering. Les métriques utilisées peuvent être regroupées en plusieurs catégories, notamment les métriques de cohésion intra-cluster (pour mesurer la pertinence, l'homogénéité d'un cluster), de séparation inter-cluster (pour mesurer la séparabilité des clusters, ou la non redondance entre les clusters) et de similarité avec les étiquettes de classe réelles. Les métriques utilisées comprennent l'homogénéité, la complétude, la V-mesure, le coefficient de silhouette, l'indice de Rand ajusté (ARI) et l'indice d'information mutuelle ajusté (AMI).

— **Homogénéité** : mesure dans quelle mesure chaque cluster ne contient que des échantillons d'une seule classe c'est à dire à quel point chaque cluster est homogène en termes de caractéristiques dominantes. La valeur de l'homogénéité varie de 0 à 1, où 1 signifie une homogénéité parfaite. Elle est calculée en comparant la distribution conditionnelle des classes pour chaque cluster avec la distribution marginale des classes. Plus la valeur de l'homogénéité est élevée, plus les échantillons d'un même cluster appartiennent à une seule classe dont ci-dessous la définition de chacune de ces mesures.

$$\text{homogeneity} = 1 - \frac{H(C|K)}{H(C)} \tag{2.6}$$

L'entropie conditionnelle $H(C|K)$ mesure la quantité d'incertitude ou d'information moyenne nécessaire pour déterminer la classe d'un échantillon, sachant son cluster. Elle quantifie la pureté des clusters en évaluant à quel point les échantillons d'un même cluster appartiennent à une seule classe.

L'entropie de la classe $H(C)$ mesure la quantité d'incertitude ou d'information moyenne nécessaire pour déterminer la classe d'un échantillon, sans connaître son cluster. Elle quantifie la répartition des classes dans l'ensemble de données.

— Complétude (completeness) : La complétude reflète dans quelle mesure tous les échantillons d'une classe donnée sont regroupés dans le même cluster. La valeur de la complétude varie de 0 à 1, où 1 signifie une complétude parfaite. Elle est calculée en comparant la distribution conditionnelle des clusters pour chaque classe avec la distribution marginale des clusters. Plus la valeur de la complétude est élevée, plus les échantillons d'une même classe sont regroupés dans le même cluster. Autrement dit, la classe se confond avec un cluster.

$$\text{completeness} = 1 - \frac{H(K|C)}{H(K)} \tag{2.7}$$

L'entropie conditionnelle $H(K|C)$ mesure la quantité d'incertitude ou d'information moyenne nécessaire pour déterminer le cluster d'un échantillon, sachant sa classe. Elle quantifie à quel point les échantillons d'une même classe sont regroupés dans le même cluster.
L'entropie du cluster $H(K)$ mesure la quantité d'incertitude ou d'information moyenne nécessaire pour déterminer le cluster d'un échantillon, sans connaître sa classe. Elle quantifie la répartition des clusters dans l'ensemble de données.

— **V-mesure** : est la moyenne harmonique de l'homogénéité et de la complétude. La valeur de la V-mesure varie de 0 à 1, où 1 indique une meilleure qualité de regroupement. Une valeur élevée de la V-mesure indique à la fois une homogénéité et une complétude

élevées.

$$Vmeasure = 2 \cdot \frac{homogeneity \cdot completeness}{homogeneity + completeness} \tag{2.8}$$

— **Coefficient de silhouette SC** : Le coefficient de silhouette mesure à la fois la cohésion intra-cluster et la séparation inter-cluster. Il est calculé pour chaque échantillon et ensuite la moyenne est prise pour obtenir le coefficient de silhouette global. La valeur du coefficient de silhouette varie de -1 à 1, où une valeur proche de 1 indique une bonne séparation des clusters.

$$SC = \frac{1}{N} \sum_{i=1}^{N} \frac{b_i - a_i}{\max\{a_i, b_i\}} \tag{2.9}$$

où a_i représente la distance moyenne entre la donnée i et toutes les autres données du même cluster (cohésion), et b_i représente la distance moyenne entre la donnée i et tous les points d'un autre cluster voisin le plus proche (séparation). La distance utilisée peut être de différents types, telle que la distance euclidienne.

— **Indice de Rand ajusté (ARI)** : L'indice de Rand ajusté mesure la similarité entre les regroupements obtenus et les étiquettes de classe réelles, en prenant en compte le hasard. La valeur de l'ARI varie de -1 à 1, où une valeur proche de 1 indique une forte similarité entre les regroupements et les étiquettes de classe réelles.

$$ARI = \frac{\sum_{ij} \binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}} \tag{2.10}$$

Dans cette formule, n représente le nombre total d'échantillons, n_{ij} est le nombre de paires d'échantillons qui sont dans la même classe dans l'étiquetage réel et dans le même cluster dans le regroupement obtenu, a_i est le nombre d'échantillons dans la même classe que l'échantillon i dans l'étiquetage réel, et b_j est le nombre d'échantillons dans le même cluster que l'échantillon j dans le regroupement obtenu.

— **Indice d'information mutuelle ajusté (AMI)** : mesure la quantité d'information mutuelle

partagées entre les regroupements obtenus et les étiquettes de classe réelles, en prenant en compte le hasard. L'AMI est une extension de l'indice d'information mutuelle (MI) qui corrige les effets du hasard et fournit une mesure de similarité entre les regroupements et les étiquettes de classe réelles. L'AMI prend des valeurs entre 0 et 1, où une valeur de 1 indique une concordance parfaite entre les regroupements et les étiquettes de classe réelles, tandis qu'une valeur proche de 0 indique une absence de corrélation entre les deux.

$$\text{AMI} = \frac{MI - E(MI)}{\max(H(C) + H(K)) - E(MI)} \tag{2.11}$$

Dans cette formule, MI représente l'information mutuelle entre les étiquettes de classe réelles (C) et les regroupements obtenus (K). L'information mutuelle mesure la quantité d'information mutuelle partagées entre deux variables aléatoires.
$H(C)$ est l'entropie des étiquettes de classe réelles et $H(K)$ est l'entropie des regroupements obtenus. L'entropie mesure l'incertitude ou la désorganisation d'une variable aléatoire.
$E(MI)$ est l'espérance de l'information mutuelle entre deux variables indépendantes. Cela représente la valeur attendue de l'information mutuelle lorsque les regroupements sont effectués au hasard.

2.3 Mesures de simularité

L'utilisation de mesures de similarité est essentielle dans de nombreux domaines de l'analyse de données, notamment dans le domaine du clustering et de la classification. Ces mesures nous permettent de quantifier la proximité ou la similarité entre des objets ou des échantillons de données. Dans notre étude, nous avons opté pour deux mesures de similarité couramment utilisées : la distance euclidienne et la métrique Dynamic Time Warping (DTW).

La distance euclidienne est une mesure directe de similarité qui calcule la distance entre

deux points dans un espace euclidien en prenant en compte leurs coordonnées. Elle est largement utilisée dans les problèmes de clustering où la proximité spatiale des données est un facteur important.

$$\text{dist}(X, Y) = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2} \tag{2.12}$$

La métrique DTW, en revanche, est spécifiquement adaptée aux données de séries temporelles. Elle prend en compte les variations de vitesse et de longueur des séquences et permet un alignement flexible des séries temporelles. Cela nous permet de capturer les similarités malgré les différences de vitesse ou de retards dans les séries.

$$\text{DTW}(X, Y) = \min_{\text{chemin}} \left(\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}(X_i - Y_j)^2} \right) \tag{2.13}$$

Dans ces formules, X et Y représentent les séries temporelles à comparer, composées respectivement de n et m points. Le chemin fait référence à l'alignement optimal entre les points des deux séries, où chaque point de la série X est associé à un point correspondant de la série Y. La métrique DTW cherche le chemin qui minimise la somme des différences quadratiques entre les paires de points correspondants.

Prenons l'exemple de deux séries temporelles représentant les niveaux d'eau dans deux rivières, Rivière A et Rivière B, sur une période de 10 jours. L'objectif est de démontrer l'avantage de la distance dynamique de temps (DTW) par rapport à la distance Euclidienne pour comparer ces séries malgré les variations de vitesse et les décalages possibles.

Rivière A :

Jour 1 : 2.5 m

Jour 2 : 2.7 m

Jour 3 : 2.9 m

Jour 4 : 3.2 m

Jour 5 : 3.6 m

Jour 6 : 3.9 m

Jour 7 : 4.2 m

Jour 8 : 4.0 m

Jour 9 : 3.7 m

Jour 10 : 3.5 m

Rivière B :

Jour 1 : 2.0 m

Jour 2 : 2.2 m

Jour 3 : 2.5 m

Jour 4 : 2.8 m

Jour 5 : 3.1 m

Jour 6 : 3.4 m

Jour 7 : 3.6 m

Jour 8 : 3.8 m

Jour 9 : 4.0 m

Distance Euclidienne : Si nous utilisons la distance Euclidienne pour comparer ces séries,

nous obtiendrons la distance géométrique entre les points correspondants :

$$\sqrt{(2.5-2.0)^2+(2.7-2.2)^2+(2.9-2.5)^2+\ldots+(3.5-4.0)^2} \approx 2.78$$

La distance Euclidienne ne tient pas compte des décalages possibles ni des variations de vitesse, ce qui peut rendre la comparaison moins précise pour des séries non alignées.

DTW : Si nous utilisons la métrique DTW, nous devons passer par deux étapes : le calcul de la matrice de coût et le calcul du chemin optimal.

Étape 1 : Création de la Matrice de Coût : La matrice de coût est construite en calculant les différences au carré entre les niveaux d'eau de chaque jour entre les rivières A et B.

Chaque élément de la matrice représente la différence au carré entre les niveaux d'eau de la rivière A et de la rivière B pour les jours correspondants. Par exemple, l'élément dans la première ligne et la première colonne (0.25) représente la différence au carré entre les niveaux d'eau de la rivière A et de la rivière B pour le premier jour.

$$\begin{bmatrix}
0.25 & 0.49 & 0.64 & 1.44 & 4.84 & 9.61 & 12.96 & 16.00 & 14.44 & 11.56 \\
0.25 & 0.25 & 0.16 & 0.36 & 1.21 & 4.41 & 6.25 & 9.00 & 7.84 & 6.25 \\
0.25 & 0.09 & 0.04 & 0.25 & 0.64 & 1.96 & 2.89 & 4.00 & 4.41 & 4.00 \\
1.44 & 1.00 & 0.25 & 0.16 & 0.25 & 1.44 & 2.89 & 2.56 & 0.01 & 0.25 \\
4.84 & 4.41 & 2.56 & 0.64 & 0.25 & 0.04 & 0.36 & 0.16 & 0.36 & 1.44 \\
9.61 & 8.41 & 5.29 & 2.25 & 0.25 & 0.25 & 0.64 & 0.16 & 1.21 & 4.41 \\
12.96 & 11.56 & 7.84 & 4.00 & 0.36 & 0.81 & 0.00 & 0.64 & 2.25 & 5.29 \\
16.00 & 14.44 & 9.00 & 5.76 & 1.44 & 0.16 & 0.64 & 0.04 & 4.84 & 7.84 \\
14.44 & 13.00 & 8.41 & 5.29 & 1.96 & 0.36 & 0.36 & 0.64 & 0.09 & 2.56 \\
11.56 & 10.24 & 6.25 & 3.61 & 0.64 & 0.25 & 0.81 & 0.36 & 0.64 & 0.25
\end{bmatrix}$$

Étape 2 : Calcul du Chemin Optimal : En utilisant la matrice de coût, nous calculons le chemin optimal qui aligne les séries temporelles de manière à minimiser la distance.Nous commençons au coin supérieur gauche et choisissons les cases avec les valeurs minimales pour nous déplacer vers le coin inférieur droit. Imaginons que le chemin optimal soit :

$$(1,1) \to (2,2) \to (3,3) \to (4,4) \to (5,5) \to (6,6) \to (7,7) \to (8,8) \to (9,9) \to (10,9)$$

En utilisant ce chemin, la DTW serait :

$$\sqrt{(2.5-2.0)^2 + (2.7-2.2)^2 + (2.9-2.5)^2 + \ldots + (3.5-4.0)^2} \approx 2.42$$

Dans cet exemple, la valeur donnée par DTW est inférieure à la distance Euclidienne. Cela démontre l'avantage du DTW pour tenir compte des décalages possibles et des variations de vitesse, ce qui en fait une mesure plus appropriée pour la comparaison de séries temporelles non alignées.

En résumé, la métrique DTW permet de mesurer la similarité entre les séries temporelles en prenant en compte les différences de vitesse et les décalages temporels possibles. Cela en fait une mesure utile pour les tâches de clustering, d'appariement et de comparaison de séries temporelles.

Deuxième partie

Analyse et clustering des séries temporelles

Chapitre 3

Clustering des séries temporelles

Dans ce chapitre, nous plongerons dans le monde de l'analyse et du clustering des séries temporelles. Nous explorerons l'état de l'art en matière de clustering de ces données, en mettant en évidence des modèles de clustering spatio-temporel, des techniques de réduction de dimension, et nous évaluerons de manière comparative différentes approches.

3.1 État de l'art sur le clustering des séries temporelles

L'analyse et la compréhension des séries temporelles sont des tâches essentielles dans de nombreux domaines tels que la finance, l'économie, l'industrie et les sciences de la santé. Cependant, avec l'abondance croissante de données temporelles, il est devenu de plus en plus complexe d'extraire des informations significatives à partir de ces données. C'est là que le clustering des séries temporelles entre en jeu.

Le clustering des séries temporelles est une technique puissante qui permet de regrouper des séries temporelles similaires en fonction de leurs motifs, de leurs formes ou de leurs comportements. Cette approche offre une vue d'ensemble des données temporelles, facilitant ainsi la découverte de tendances, de structures et de comportements communs au sein des séries.

L'analyse et le regroupement des données temporelles présentent des défis uniques qui nécessitent des approches adaptées. Les techniques de clustering conventionnelles peuvent

être classées en deux catégories principales : le clustering hiérarchique et le clustering par partitionnement. Le clustering hiérarchique organise les données dans une structure arborescente, permettant d'explorer les relations de similarité à différentes échelles. D'autre part, le clustering par partitionnement divise les données en groupes distincts, où chaque élément appartient à un seul groupe. Les méthodes couramment utilisées incluent le clustering spectral, statistique et spatial.

Cependant, lorsqu'il s'agit de données temporelles, ces approches conventionnelles peuvent être limitées en raison de la nature séquentielle des séries temporelles et de leur dépendance temporelle. C'est là qu'intervient le clustering profond, une approche émergente qui exploite les capacités des réseaux neuronaux pour extraire des caractéristiques complexes et des motifs temporels à partir des données séquentielles. En utilisant des architectures de réseau spécifiquement conçues pour le clustering, le clustering profond offre de nouvelles perspectives pour le regroupement précis et interprétable des séries temporelles.

Parmi les approches spécifiques au clustering des séries temporelles, les variantes de DBSCAN (Density-Based Spatial Clustering of Applications with Noise) ont suscité un intérêt croissant. DBSCAN est un algorithme de clustering basé sur la densité qui peut identifier des clusters de formes arbitraires et gérer efficacement le bruit et les valeurs aberrantes. De plus DBSCAN n'a pas besoin de lui définir le nombre de clusters *a priori*

Dans notre étude, nous nous concentrons sur une variante novatrice de DBSCAN pour le clustering des séries temporelles, en combinant les avantages de DBSCAN avec les capacités de modélisation et d'apprentissage profond du clustering profond. Cette approche hybride vise à surmonter les limitations des approches conventionnelles et à offrir une meilleure précision et une meilleure interprétabilité dans le regroupement des séries temporelles.

Nous explorerons en détail les différentes variantes de DBSCAN adaptées aux séries temporelles, en mettant l'accent sur leurs caractéristiques, leurs avantages et leurs limites. Nous examinerons également les méthodes d'évaluation et les domaines d'application pertinents. En combinant les fondements solides de DBSCAN avec les avancées du clustering profond, notre approche contribue à l'état de l'art en matière de clustering des séries temporelles et

ouvre de nouvelles perspectives de recherche dans ce domaine stimulant.

Dans notre étude, nous avons examiné différentes variantes de l'algorithme de clustering DBSCAN. Parmi ces variantes, DBSCAN++ Jang and Jiang (2019) et SNG-DBSCAN Jiang et al. (2020) ont été développées pour améliorer les performances en termes de temps d'exécution. Ces versions cherchent à réduire la complexité temporelle de DBSCAN en introduisant des optimisations spécifiques. Cependant, elles peuvent encore souffrir de certaines limites, telles que des problèmes de regroupement incorrect des données de densité variable et une importance inégale accordée aux différents attributs lors du calcul des voisins d'un point.

D'autres travaux de recherche ont également exploré des améliorations alternatives pour DBSCAN. Par exemple, une approche basée sur la parallélisation a été proposée par Götz et al. (2015) pour accélérer le calcul de DBSCAN. D'autre part, certaines versions ont utilisé des implémentations GPU Andrade et al. (2013) pour traiter des ensembles de données massifs en termes de nombre d'enregistrements. Cependant, ces approches mettent moins l'accent sur l'augmentation du nombre des attributs.

En ce qui concerne l'adaptation de la densité, une variante appelée DC-SKCG Zhang et al. (2021) a été proposée pour remplacer le principe de densité de DBSCAN par une densité adaptative. Cette approche utilise un ϵ adaptatif pour la densité des points et tente de corriger les regroupements incorrects. Cependant, elle présente toujours des limites en termes de complexité temporelle.

D'autres tentatives d'amélioration ont été faites en utilisant des approches telles que l'espace d'influence k (IS_k) proposé par Cassisi et al. (2013) ou en utilisant le hachage sensible à la localité (LSH) Wu et al. (2007).

De plus, certaines variantes, comme MDST-DBSCAN Choi and Hong (2021), ont été développées pour prendre en compte les caractéristiques temporelles des données. Cependant, ces approches peuvent augmenter considérablement la complexité temporelle en fonction du nombre d'attributs.

Une autre approche, proposée par Yanenko and der Weberei (2019), a tenté de renforcer

la détection d'événements au sein des données des médias sociaux. Contrairement à l'algorithme ST-DBSCAN traditionnel, STS-DBSCAN intègre un nouveau paramètre, noté ϵ_3, en plus des seuils spatiaux (ϵ_1) et temporels (ϵ_2) lors du processus de regroupement. Ce paramètre supplémentaire représente la "distance sociale" entre les créateurs de données. En utilisant cette information sociale, l'approche vise à accroître la fiabilité des résultats en privilégiant des observations issues de sources diverses, plutôt que d'un seul contributeur.

En résumé, plusieurs variantes de DBSCAN ont été développées pour améliorer ses performances en termes de temps d'exécution, de précision de clustering et de prise en compte des caractéristiques des données. Cependant, chaque approche présente ses propres avantages et limites, et le choix de la méthode dépend des caractéristiques spécifiques des données et des objectifs du clustering."

L'étude menée par Kumar and Reddy (2016) présente une approche alternative pour améliorer le temps de calcul de DBSCAN. Ils ont introduit une technique basée sur les graphes de groupes. Cette technique vise à accélérer les opérations de recherche des voisins lors du processus de regroupement en exploitant la structure des données sous forme de graphes.

En ce qui concerne la variante DBSCRN Chowdhury and Amorim (2019), elle se distingue par l'utilisation du concept de K plus proches voisins (*K Nearest Neighbors*) au lieu des paramètres traditionnels tels que ϵ et le nombre minimum d'échantillons ($Minpts$) utilisés dans la version classique de DBSCAN. Malgré ces améliorations, tous les modèles basés sur DBSCAN souffrent encore d'une complexité temporelle élevée.

Le clustering profond peut être classé en deux types principaux : le clustering profond intégré et le clustering profond séparé. Dans le clustering profond intégré, l'algorithme de clustering est directement intégré aux réseaux neuronaux profonds. Cela permet une extraction de caractéristiques et un regroupement simultanés en utilisant la fonction de perte (LOSS function). Cette approche favorise une meilleure représentation des données et une meilleure performance de clustering. En revanche, dans le clustering profond séparé, un réseau neuronal profond est utilisé pour extraire les caractéristiques des données, qui sont ensuite soumises à des algorithmes de clustering classiques pour réaliser le regroupement

Alqahtani et al. (2021).

Le clustering profond est un domaine de recherche relativement récent et présente des défis spécifiques dans le contexte des séries temporelles. L'état de l'art actuel en matière de clustering profond pour les séries temporelles est encore limité, ce qui ouvre de nombreuses opportunités de recherche et de développement. Par conséquent, dans le cadre de notre étude, nous nous proposons de présenter une synthèse des derniers développements dans ce domaine.

Lors de notre revue de littérature, nous avons constaté que les modèles basés sur la densité, tels que DBSCAN, sont rarement utilisés dans le clustering profond des séries temporelles multivariées (DTSC). Au lieu de cela, une approche courante consiste à utiliser des autoencodeurs profonds pour extraire les caractéristiques des séries temporelles, puis à appliquer des algorithmes de clustering traditionnels sur les caractéristiques extraites. Cette approche permet de combiner les avantages des réseaux neuronaux profonds pour l'extraction des caractéristiques et des algorithmes de clustering classiques pour le regroupement Alqahtani et al. (2021).

Cependant, dans notre contribution, nous souhaitons explorer une approche alternative en utilisant une version améliorée de DBSCAN spécifiquement adaptée aux séries temporelles multivariées. Nous proposons d'appliquer un autoencodeur profond aux données de séries temporelles multivariées afin d'obtenir une représentation latente riche en information. Cette représentation latente sera ensuite utilisée comme entrée pour notre version améliorée de DBSCAN, qui intègre un ϵ adaptatif et diverses mesures de similarité pour le regroupement des données (Figure 3.1).

En plus d'explorer une version améliorée de DBSCAN adaptée aux séries temporelles multivariées, notre étude comprend une comparaison approfondie de différentes techniques de réduction de dimensionnalité, telles que le deep autoencodeur DAE, PCA, LDA UMAP et t-SNE. L'objectif de cette comparaison est d'évaluer la capacité de chaque méthode à préserver les informations pertinentes pour le clustering des séries temporelles multivariées, en mettant l'accent sur les caractéristiques temporelles.

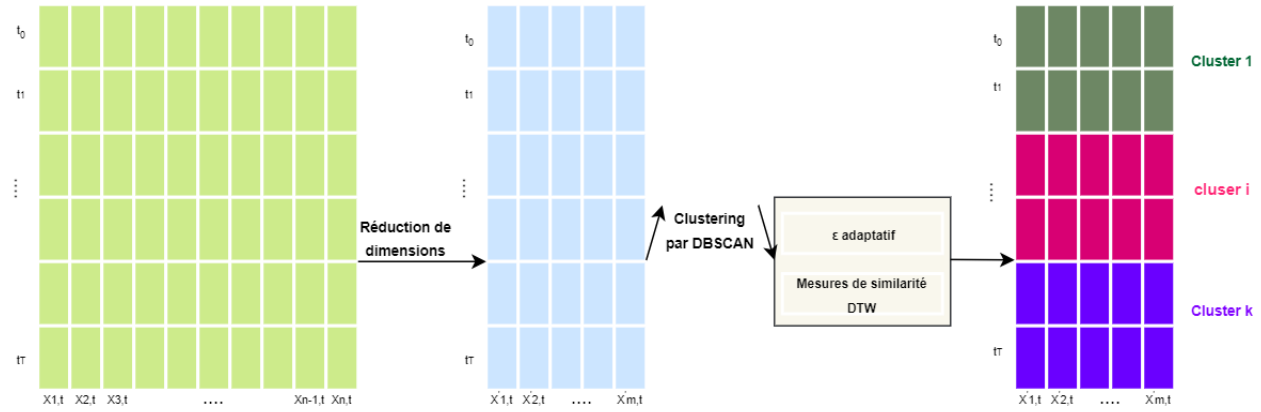


FIGURE 3.1 – Approche Innovante de Regroupement pour les Séries Temporelles Multivariées

L'autoencodeur profond est une technique de réduction de dimensionnalité basée sur l'apprentissage non supervisé. Il apprend à reconstruire les données d'entrée en passant par des couches cachées de dimensions réduites. L'autoencodeur est capable d'apprendre des représentations latentes qui capturent les caractéristiques importantes des séries temporelles, ce qui en fait une méthode pertinente pour le clustering.

PCA (Principal Component Analysis) est une méthode linéaire de réduction de dimensionnalité qui cherche à représenter les données dans un espace de dimensions inférieures en se basant sur les directions de variance maximale.

LDA (Linear Discriminant Analysis) est une technique de réduction de dimensionnalité supervisée qui vise à trouver une projection des données qui maximise la séparation entre les classes.

UMAP (Uniform Manifold Approximation and Projection) est une technique de réduction de dimensionnalité non linéaire qui tente de préserver les relations locales entre les points dans l'espace d'origine. Il est particulièrement adapté aux données présentant des structures complexes et non linéaires.

t-SNE (t-Distributed Stochastic Neighbor Embedding) est une autre technique non linéaire de réduction de dimensionnalité qui se concentre sur la préservation des relations de similarité entre les points. Il est souvent utilisé pour visualiser les données dans un espace de dimension réduite.

Lors de notre comparaison, nous évaluons la capacité de chaque technique à préserver

les informations temporelles et à capturer les structures et les relations entre les séries temporelles multivariées. Nous analysons également les performances de clustering obtenues à partir des représentations réduites par chaque méthode. Cette comparaison approfondie nous permettra de mieux comprendre les avantages et les limitations de chaque technique et de sélectionner la méthode la plus adaptée pour notre approche de clustering des séries temporelles multivariées.

L'un des défis majeurs dans le clustering des séries temporelles est la présence de clusters de densité variable, c'est-à-dire des clusters qui présentent des variations de densité à travers l'espace. Pour corriger ce problème, nous avons exploré différentes hypothèses pour trouver la meilleure valeur d'epsilon pour chaque point. Nous avons envisagé des approches telles que l'ajustement adaptatif de l'epsilon en fonction des caractéristiques locales des données et la définition d'une valeur d'epsilon différente pour chaque dimension temporelle. Cette recherche de la meilleure valeur d'epsilon vise à améliorer la précision et la qualité globale du clustering pour les séries temporelles multivariées.

En outre, nous avons également utilisé la métrique Dynamic Time Warping DTW (2.3) au lieu de la distance euclidienne pour mesurer la similarité entre les séries temporelles. La DTW est une mesure de similarité plus appropriée pour les séries temporelles, car elle est capable de gérer les variations de vitesse, de décalage et de distorsion temporelle entre les séquences. L'utilisation de la métrique DTW peut donc améliorer la précision du clustering en prenant en compte les caractéristiques temporelles spécifiques des données.

Enfin, afin d'améliorer davantage la qualité des regroupements obtenus, nous envisageons d'appliquer une technique de révision du clustering. Cette approche consiste à analyser la cohérence des clusters, à détecter les outliers et à fusionner les clusters similaires pour optimiser les résultats du clustering initial. En intégrant cette technique de révision du clustering, nous visons à obtenir des clusters plus cohérents, plus stables et de meilleure qualité.

En résumé, notre étude propose une approche novatrice pour le clustering des séries temporelles multivariées en utilisant une version améliorée de DBSCAN. Nous combinons

l'utilisation d'autoencodeurs profonds, d'une recherche de la meilleure valeur d'epsilon, de la métrique DTW et d'une technique de révision du clustering pour améliorer la performance et la précision du regroupement. En comparant notre approche avec d'autres variantes de DBSCAN et les méthodes de clustering profond existantes, nous espérons démontrer sa pertinence et son efficacité dans le contexte des séries temporelles multivariées.

3.2 Modèles de clustering spatio-temporel

La classification des algorithmes de regroupement s'articule autour de deux axes principaux : le regroupement conventionnel et le regroupement profond, chacun d'eux se subdivisant en catégories distinctes. Le regroupement conventionnel se divise en deux sous-catégories : le regroupement hiérarchique et le regroupement par partitionnement. Dans ce dernier, diverses approches telles que le regroupement spectral, statistique, spatial et spatio-temporel sont mises en œuvre. Par exemple, le DBSCAN est un algorithme spatial qui prend en compte la densité des données, tandis que le MDST-DBSCAN est spécialement conçu pour le regroupement spatio-temporel.

En parallèle, le regroupement profond se décline en deux orientations distinctes : le regroupement intégré et le regroupement séparé. Notre attention se porte sur la branche du regroupement profond séparé, où nous avons élaboré deux modèles novateurs : D-DBSCAN et D-MDST-DBSCAN, la lettre "D" faisant référence à leur base en apprentissage profond.

Cette taxonomie soigneusement tracée reflète l'évolution des méthodes de regroupement et leur adaptabilité à des données complexes. Les différentes catégories, qu'il s'agisse du regroupement conventionnel ou du regroupement profond, offrent des solutions distinctes pour la segmentation et la compréhension des motifs dans les données. Notre contribution se situe au cœur du regroupement profond séparé, où D-DBSCAN et D-MDST-DBSCAN se positionnent comme des piliers prometteurs, exploitant les avantages de l'apprentissage profond pour améliorer la capacité de détection de clusters dans des environnements variés (Figure 3.2).

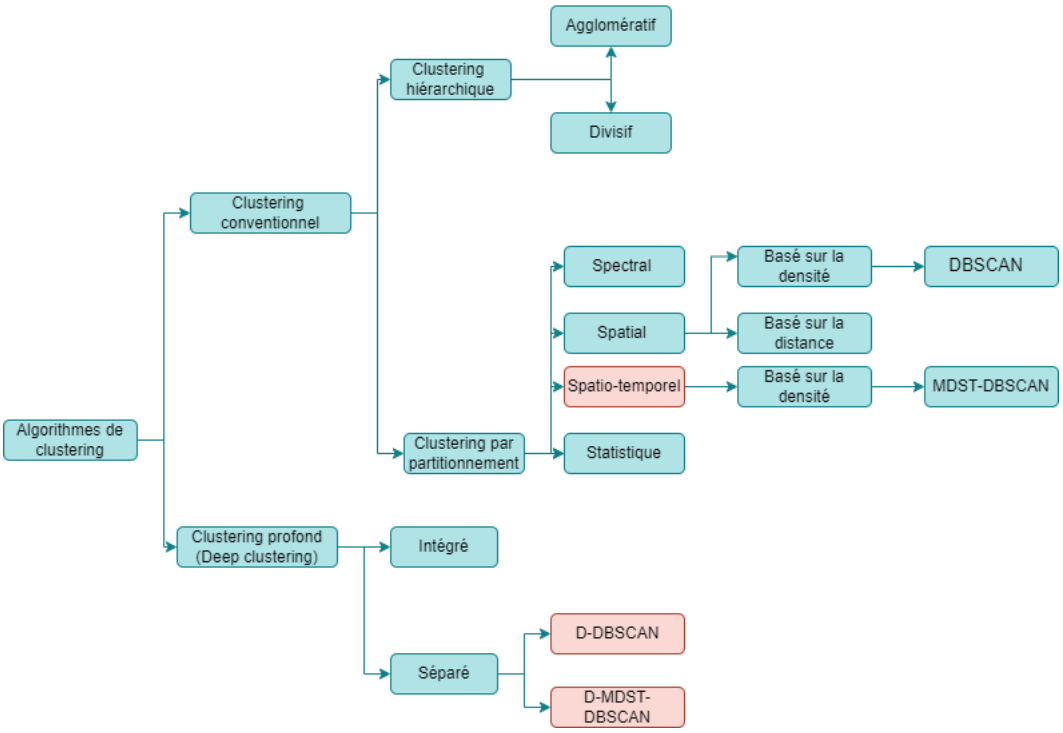


FIGURE 3.2 – Taxonomie des Méthodes de Regroupement Spatio-Temporel et Profond

3.2.1 DBSCAN

L'algorithme DBSCAN (Density-Based Spatial Clustering of Applications with Noise) est un modèle de clustering largement utilisé pour l'analyse de données où la répartition spatiale est prédominante. L'algorithme DBSCAN nécessite la spécification de trois paramètres : $\mathcal{X}$ pour représenter l'ensemble de données, ϵ pour définir le seuil de distance ou dit autrement l'étendu du voisinage, et $MinPts$ pour indiquer le nombre minimal de points requis dans le voisinage ϵ pour former un cluster (Figure 3.3). Ces paramètres doivent être adaptés spécifiquement aux données spatio-temporelles pour garantir des résultats précis. Dans cette partie, nous allons décrire en détail ces paramètres et les concepts clés associés à l'algorithme DBSCAN, afin de permettre une identification efficace des structures de clusters dans nos données.

Afin de mieux comprendre ces définitions et leurs implications, il est essentiel de clarifier certains concepts clés tels que le voisinage ϵ d'un point, les points centraux, l'atteignabilité en termes de densité et la connectivité en termes de densité. Ces définitions fournissent un

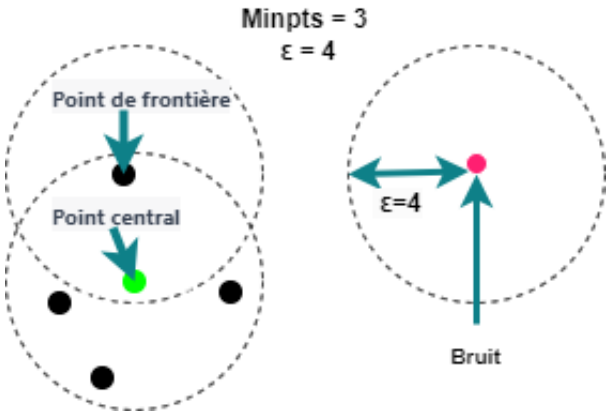


FIGURE 3.3 – Schéma Descriptif de DBSCAN

cadre théorique solide pour comprendre le fonctionnement de l'algorithme DBSCAN et permettent d'explorer ses performances et ses limites dans le contexte du clustering de données temporelles.

- Le paramètre ϵ définit le seuil de voisinage, représentant la distance maximale acceptée pour qu'un point soit considéré comme un point central et inclus dans un cluster.
- Le paramètre $MinPts$ indique le nombre minimal de points requis dans le voisinage ϵ d'un point central pour qu'il soit inclus dans un cluster. Des études antérieures suggèrent une valeur optimale d'environ deux fois la dimension du jeu de données.
- Le voisinage ϵ d'un point X_{t_1} est l'ensemble des points situés à une distance égale ou inférieure à ϵ du point X_{t_1}. Voisinage$(\epsilon, X_{t_1}) = \{X_{t_2} \mid \text{distance}(X_{t_1}, X_{t_2}) \leq \epsilon\}$
- Un point central est un point qui a un nombre de points dans son voisinage ϵ supérieur ou égal à $MinPts$. Point central$(X_{t_1}) \Leftrightarrow |\text{Voisinage}(\epsilon, X_{t_1})| \geq MinPts$
- Un point X_{t_1} est directement atteignable en termes de densité à partir d'un point X_{t_2} si X_{t_2} se trouve dans le voisinage ϵ de X_{t_1} et que X_{t_1} est un point central. Atteignabilité directe$(X_{t_1}, X_{t_2}) \Leftrightarrow X_{t_2} \in$ Voisinage$(\epsilon, X_{t_1}) \wedge$ Point central(X_{t_1}).
- Un point M est atteignable en termes de densité à partir d'un point N avec respect à ϵ et $MinPts$ s'il existe une séquence de points $M_1, M_2, ..., M_p$, où $M_1 = N$ et $M_p = M$, telle que chaque M_{j+1} est directement atteignable en termes de densité à partir de M_j. Atteignabilité$(M, N, \epsilon, MinPts) \Leftrightarrow \exists M_1, M_2, ..., M_p : M_1 = N \wedge M_p = M \wedge \forall j \in [1, p-1]$:

Atteignabilité directe(M_j, M_{j+1}).

— Un point X_{t_2} est connecté en termes de densité à partir d'un point X_{t_1} avec respect à ϵ et $MinPts$ s'il existe un point X_{t_3} dans l'ensemble de données tel que X_{t_1} et X_{t_2} soient atteignables en termes de densité à partir de X_{t_3} avec respect à ϵ.
Connectivité en densité$(X_{t_1}, X_{t_2}, \epsilon, MinPts) \Leftrightarrow \exists X_{t_3}$: Atteignabilité$(X_{t_1}, X_{t_3}, \epsilon, MinPts) \wedge$ Atteignabilité$(X_{t_2}, X_{t_3}, \epsilon, MinPts)$.

Entrées :

— $\mathcal{X} = \{x_1, x_2, \ldots, x_m\}$: Ensemble de données représentant les séries temporelles multivariées.

— ϵ : Rayon de voisinage pour déterminer la distance seuil.

— $MinPts$: Nombre minimal de points requis dans le voisinage ϵ pour former un cluster.

Sortie :

— Clusters : Ensemble de clusters formés à partir des séries temporelles.

Début de l'algorithme DBSCAN
Clusters $\leftarrow \{\}$
Visited $\leftarrow \{\}$
for $i = 1$ **à** m **do**
  if x_i n'est pas visité **then**
    Marquer x_i comme visité
    Voisins $\leftarrow$ TrouverVoisins(x_i, ϵ)
    if NombreDeVoisins(Voisins) $\geq MinPts$ **then**
      Cluster $\leftarrow$ NouveauCluster()
      Ajouter x_i à Cluster
      ÉtendreCluster(Cluster, x_i, Voisins, ϵ, $MinPts$)
      Ajouter Cluster à Clusters
    else
      Marquer x_i comme point de bruit
    end if

16: **end if**
17: **end for**
18: **return** Clusters
19: **Fin de l'algorithme DBSCAN**

Fonction ÉtendreCluster(Cluster, Point, Voisins, ϵ, $MinPts$) :

1: **for** chaque voisin v de Point dans Voisins **do**
2: **if** v n'est pas visité **then**
3: Marquer v comme visité
4: NouveauxVoisins $\leftarrow$ TrouverVoisins(v, ϵ)
5: **if** NombreDeVoisins(NouveauxVoisins) $\geq MinPts$ **then**
6: Ajouter NouveauxVoisins à Voisins
7: **end if**
8: **end if**
9: **if** v n'appartient à aucun cluster **then**
10: Ajouter v à Cluster
11: **end if**
12: **end for**

Fonction TrouverVoisins(Point, ϵ) :

1: Voisins $\leftarrow \{\}$
2: **for** $i = 1$ **à** m **do**
3: **if** distance(Point, x_i) $\leq \epsilon$ **then**
4: Ajouter x_i à Voisins
5: **end if**
6: **end for**
7: **return** Voisins

3.2.2 MDST-DBSCAN

L'algorithme MDST-DBSCAN est une extension du DBSCAN qui permet de traiter des données spatio-temporelles en incorporant une valeur ϵ spécifique à chaque dimension des données. Contrairement au DBSCAN traditionnel qui utilise un seul ϵ pour définir le seuil de voisinage dans toutes les dimensions, MDST-DBSCAN propose une approche plus flexible en utilisant plusieurs valeurs ϵ. Chaque valeur ϵ est associée à une dimension spécifique et contrôle la sensibilité du clustering dans cette dimension Choi and Hong (2021).

L'utilisation de plusieurs valeurs ϵ permet à MDST-DBSCAN de prendre en compte les variations de densité dans les différentes dimensions des données spatio-temporelles. Par exemple, dans un ensemble de données représentant à la fois la position géographique et le temps, les densités peuvent varier considérablement dans l'espace géographique par rapport au temps. En utilisant des valeurs ϵ adaptées à chaque dimension, l'algorithme peut capturer efficacement les clusters spatio-temporels et éviter les regroupements indésirables.

Algorithm 1 MDST-DBSCAN

Require: $\mathcal{X} = \{x_1, x_2, \ldots, x_m\}', \epsilon = \{\epsilon_{x_1}, \epsilon_{x_2}, \ldots, \epsilon_{x_m}\}, MinPts > 0$
Ensure: Clusters : Ensemble de clusters formés à partir des séries temporelles.

```
Clusters ← []
Visited ← []
for i = 1 à m do
    if x_i n'est pas visité then
        Marquer x_i comme visité
        Voisins ← TrouverVoisins(x_i, ε_{x_i})
        if NombreDeVoisins(Voisins) ≥ MinPts then
            Cluster ← NouveauCluster()
            Ajouter x_i à Cluster
            ÉtendreCluster(Cluster, x_i, Voisins, ε_{x_i}, MinPts)
            Ajouter Cluster à Clusters
        else
            Marquer x_i comme point de bruit
        end if
    end if
end for
return Clusters
```

L'utilisation de l'algorithme MDST-DBSCAN présente certains inconvénients, notamment en termes de complexité temporelle et de scalabilité.

L'un des principaux inconvénients de MDST-DBSCAN réside dans sa complexité temporelle, qui est quadratique par rapport à la taille de l'ensemble de données. La complexité de l'algorithme est de l'ordre de O(n^2), où n représente le nombre d'instances dans l'ensemble de données. Cette complexité est due au fait que MDST-DBSCAN calcule le voisinage d'un point en fonction de chaque dimension et de chaque valeur ϵ. Par conséquent, plus le nombre de dimensions et de valeurs ϵ est élevé, plus le temps d'exécution de l'algorithme sera long. Cette complexité temporelle peut rendre l'algorithme inefficace pour les ensembles de données massifs, nécessitant ainsi des optimisations pour améliorer sa performance.
Un autre inconvénient de MDST-DBSCAN réside dans sa scalabilité. L'algorithme peut rencontrer des difficultés lorsqu'il est appliqué à des ensembles de données de grande taille. En raison de sa complexité temporelle élevée, l'algorithme peut nécessiter des ressources computationnelles importantes, telles que la mémoire et la puissance de calcul, pour traiter efficacement de grandes quantités de données. De plus, l'augmentation du nombre de dimensions des données et du nombre de valeurs ϵ peut aggraver ces problèmes de scalabilité, limitant ainsi la capacité de l'algorithme à gérer de manière efficace et efficiente des ensembles de données massifs.
Il est important de prendre en compte ces inconvénients lors de l'utilisation de MDST-DBSCAN, en particulier pour les applications nécessitant le traitement de données spatio-temporelles volumineuses. Des techniques de réduction de dimension, des optimisations d'algorithme et des approches de parallélisation peuvent être explorées pour atténuer ces limitations et rendre MDST-DBSCAN plus adapté à la réalisation de clustering sur des ensembles de données massives.

D-DBSCAN (Density-Based Spatial Clustering with Data Reduction) est une version améliorée de l'algorithme DBSCAN qui intègre une étape de réduction de données pour améliorer l'efficacité et la précision du clustering. Cette méthode vise à résoudre les défis liés aux ensembles de données volumineux et complexes en réduisant la dimensionnalité des séries temporelles multivariées à l'aide d'un Deep Autoencodeur avant d'appliquer DBSCAN.

L'algorithme D-DBSCAN commence par réduire les données à l'aide d'un DAE, ce qui per-

Algorithm 2 D-DBSCAN : Density-Based Spatial Clustering with Data Reduction

Require: $\mathcal{X} = x_1, x_2, \ldots, x_m$: Ensemble de données représentant les séries temporelles multivariées.
Require: ϵ_{xi} : Rayon de voisinage adaptatif pour déterminer la distance seuil de chaque point.
Require: $Minpts$: Nombre minimal de points requis dans le voisinage ϵ_{xi} pour former un cluster.
Ensure: Clusters : Ensemble de clusters formés à partir des séries temporelles.

```
Effectuer la réduction des données via un DAE
Obtenir un nouveau jeu de données réduit X' = x_1, x_2, ..., x_n
Calculer l'ε_xi adaptatif pour chaque point de X' en utilisant la méthode de sélection de voisinage
Clusters ← []
Visited ← []
for i = 1 à n do
    if x_i n'est pas visité then
        Marquer x_i comme visité
        Voisins ← TrouverVoisins(x_i, ε_xi)
        if NombreDeVoisins(Voisins) ≥ Minpts then
            Cluster ← NouveauCluster()
            Ajouter x_i à Cluster
            ÉtendreCluster(Cluster, x_i, Voisins, ε_xi, Minpts)
            Ajouter Cluster à Clusters
        else
            Marquer x_i comme point de bruit
        end if
    end if
end for
return Clusters
```

met d'obtenir une représentation plus compacte des séries temporelles tout en préservant leurs caractéristiques discriminantes. Cette étape de réduction des données (Étape 1) permet de travailler avec un ensemble de données réduit $\mathcal{X}'$ (Étape 2), ce qui réduit la complexité computationnelle et améliore l'efficacité du clustering.

Ensuite, pour chaque point x_i dans $\mathcal{X}'$, l'algorithme calcule un rayon de voisinage adaptatif ϵ_{xi} (Étape 3). Cette adaptation de ϵ tient compte des variations locales de densité dans l'ensemble de données réduit, ce qui permet d'ajuster la sensibilité du clustering à la densité des points.

En appliquant les principes de DBSCAN à $\mathcal{X}'$, l'algorithme D-DBSCAN identifie les points centraux et forme des clusters en considérant les voisins de chaque point dans son voisinage

adaptatif ϵ_{xi}. Les points qui ne satisfont pas la condition du nombre minimal de voisins ($Minpts$) sont marqués comme du bruit.

Enfin, D-DBSCAN renvoie l'ensemble des clusters formés à partir des séries temporelles, intégrant ainsi la réduction de données et l'adaptation de ϵ pour une meilleure précision et une efficacité accrue du clustering.

En utilisant D-DBSCAN, les praticiens peuvent obtenir des clusters plus précis et significatifs à partir de séries temporelles multivariées, tout en réduisant le temps de calcul et en minimisant l'impact de la dimensionnalité élevée des données. Cette méthode offre ainsi une solution prometteuse pour l'analyse et l'exploration de données temporelles complexes dans divers domaines d'application.

Algorithm 3 D-MDST-DBSCAN

Require: $\mathcal{X} = \{x_1, x_2, \ldots, x_m\}$: Ensemble de données
Require: $\epsilon = \{\epsilon_{x_1}, \epsilon_{x_2}, \ldots, \epsilon_{x_m}\}$: Rayon de voisinage initial pour chaque point
Require: $Minpts$: Nombre minimal de points requis pour former un cluster.
Ensure: Clusters : Ensemble de clusters formés à partir des données.

```
Effectuer une réduction des données via un autoencodeur profond (DAE)
Obtenir un nouvel ensemble de données réduit 𝒳′ = {x_1, x_2, ..., x_n}
Calculer l'epsilon adaptatif (ϵ_xi) pour chaque point de 𝒳′
Clusters ← {}
for i = 1 à n do
    Points ← TrouverPointsDansVoisinage(x_i, ϵ_xi)
    if NombreDePoints(Points) ≥ Minpts then
        Cluster ← NouveauCluster()
        Ajouter x_i à Cluster
        for all x_j dans Points do
            if x_j est déjà dans un cluster then
                Fusionner(Cluster, ClusterDe(x_j))
            else
                Ajouter x_j à Cluster
            end if
        end for
        Ajouter Cluster à Clusters
    else
        Marquer x_i comme point de bruit
    end if
end for
return Clusters
```

Notre proposition, appelée D-MDST-DBSCAN (Density-Based Spatial Temporal Clustering

with Data Reduction), intègre une étape de réduction des données à l'aide d'un autoencodeur profond (DAE) pour améliorer l'efficacité et la précision du clustering. Voici comment cette proposition se distingue de l'algorithme MDST-DBSCAN de base :

L'algorithme D-MDST-DBSCAN commence par appliquer un autoencodeur profond (DAE) aux données temporelles. Cela permet de réduire la dimensionnalité des données tout en préservant leurs caractéristiques discriminantes. La réduction des données permet de travailler avec un ensemble de données réduit, ce qui réduit la complexité computationnelle et améliore l'efficacité du clustering. En réduisant la dimensionnalité des données, D-MDST-DBSCAN peut gérer des ensembles de données volumineux et complexes de manière plus efficace.

Dans D-MDST-DBSCAN, nous proposons d'utiliser des epsilon adaptatifs (ϵ_{x_i}) plutôt que des ϵ initiales fixes. Les epsilon adaptatifs tiennent compte des variations locales de densité dans les données réduites, ce qui permet d'ajuster la sensibilité du clustering à la densité des points. Contrairement aux epsilon initiales, qui sont les mêmes pour tous les points, les epsilon adaptatifs permettent d'adapter le rayon de voisinage pour chaque point en fonction de son environnement local. Cela permet de mieux capturer les variations locales de densité et d'obtenir des clusters plus précis.

En combinant la réduction des données avec l'utilisation d'epsilon adaptatifs, notre proposition D-MDST-DBSCAN améliore l'efficacité et la précision du clustering de données temporelles. La réduction des données facilite le traitement des ensembles de données volumineux, tandis que l'utilisation d'epsilon adaptatifs permet d'ajuster la sensibilité du clustering à la densité des points. Ces améliorations permettent d'obtenir des clusters plus précis et de mieux capturer les structures et les patterns spécifiques des données temporelles.

En résumé, notre proposition D-MDST-DBSCAN combine la réduction des données à l'aide d'un autoencodeur profond, l'utilisation d'epsilon adaptatifs pour ajuster la sensibilité du clustering à la densité des points, et l'utilisation d'une technique de revision de clustering (section 3.2.5) pour corriger les points mal regroupés. Ces améliorations visent à améliorer l'efficacité, la précision et la cohérence du clustering des données temporelles.

3.2.3 Choix de ϵ adaptatif pour densité hétérogène

Le choix d'une valeur appropriée pour le paramètre ϵ peut être une tâche complexe lors du regroupement de données. Bien que certaines approches Rahmah and Sitanggang (2016), Sander et al. (1998) déterminent ϵ de manière à ce que 90% des points aient une distance proche de cette valeur en utilisant l'algorithme des k plus proches voisins (k-NN), cette méthode peut se révéler inefficace lorsque la densité autour des points de données est hautement hétérogène. Ce problème est particulièrement fréquent dans le cas de séries temporelles multivariées irrégulières, où les intervalles de temps entre les observations sont inégaux.

Afin de résoudre ce défi, nous proposons un ensemble d'hypothèses visant à sélectionner un ϵ adaptatif pour chaque point lors du regroupement de données présentant une densité hétérogène. Dans ce contexte, nous notons la distance entre les points x_i et x_j par d_{ij}. Ces hypothèses offrent une approche alternative pour choisir un ϵ approprié, en tenant compte de la densité locale des points et de leurs voisins.

- **Hypothèse 1 (H1)** : Nous utilisons la distance de X_i à son voisin le plus éloigné selon ϵ pour définir son voisinage. Soit $X = x_1, x_2, ...x_m$ un ensemble de données.

$$\epsilon_{x_i} = Argmax_j\{d_{ij} = d(x_i, x_j) | d(x_i, x_j) \leq \epsilon\} \tag{3.1}$$

- **Hypothèse 2 (H2)** : Nous utilisons l'inverse de la densité locale d'un point x_i pour définir son voisinage. Soit ρ_{x_i} la densité locale de x_i où

$$\rho_{x_i} = \sum_{j=1}^{m} \sigma(d_{ij} - \epsilon) \tag{3.2}$$

et

$$\sigma(d_{ij} - \epsilon) = \begin{cases} 1 & \text{if } d_{ij} \leq \epsilon \\ 0 & \text{if } d_{ij} > \epsilon \end{cases} \tag{3.3}$$

Donc,

$$\epsilon_{x_i} = \frac{1}{\rho_{x_i}} \tag{3.4}$$

— **Hypothèse 3 (H3)** : Dans cette hypothèse, nous utiliserons la formule ϵ utilisée dans l'article de Zhang et al. (2021) : elle utilise la moyenne de la densité locale d'un point.

$$\epsilon_{x_i} = \epsilon\eta|\frac{(\rho_{x_i} - (\sum_{i=1}^{m} \frac{\rho_{x_i}}{m})}{(\sum_{i=1}^{m} \frac{\rho_{x_i}}{m})}| \tag{3.5}$$

où η est un paramètre qui contrôle l'intensité de l'ajustement et prend généralement une valeur de 0.8 Zhang et al. (2021).

— **Hypothèse 4 (H4)** : Nous utilisons la densité de domaine adaptative de x_i proposée par Chen and Philip (2019) au lieu de la densité locale ρ_{x_i} utilisée dans les hypothèses 1 et 2.

Nous avons choisi d'utiliser l'algorithme DBSCAN comme référence et d'appliquer diverses hypothèses à celui-ci. Nous avons appliqué les paramètres de la meilleure configuration obtenue avec DBSCAN à l'algorithme MDST-DBSCAN.

3.2.4 Techniques de réduction de dimension

La réduction de dimension des séries temporelles multivariées est une étape essentielle dans l'analyse de ces données. Les séries temporelles multivariées se composent de plusieurs variables qui évoluent dans le temps, ce qui peut entraîner une dimensionnalité élevée et complexe.

La réduction de dimension est nécessaire pour plusieurs raisons. Tout d'abord, une grande dimensionnalité peut rendre l'analyse des séries temporelles plus difficile et complexe. En réduisant la dimension, on peut simplifier la représentation des données et faciliter leur compréhension. Cela peut être particulièrement utile lorsqu'on cherche à extraire des informations significatives à partir des séries temporelles, telles que des tendances, des motifs ou des relations entre les variables.

Ensuite, la réduction de dimension permet de surmonter le problème de la malédiction de la dimensionnalité. Lorsque la dimension des données est élevée, les modèles d'apprentissage automatique peuvent avoir du mal à généraliser correctement et à faire des prédictions

précises. En réduisant la dimension, on réduit également la complexité des modèles et on améliore leur capacité à traiter les données.

De plus, la réduction de dimension peut contribuer à l'efficacité computationnelle. En réduisant la dimension des données, on réduit la quantité d'informations à traiter, ce qui peut accélérer les calculs et rendre l'analyse des séries temporelles plus rapide et plus efficace.

en outre, la réduction de dimension peut aider à visualiser et à interpréter les données. En réduisant la dimension, on peut projeter les données dans un espace de dimension inférieure, ce qui facilite la visualisation et la détection de schémas ou de clusters. Cela permet également de réduire la complexité des modèles graphiques et de rendre les résultats plus interprétables.

En résumé, la réduction de dimension des séries temporelles multivariées est essentielle pour simplifier l'analyse, surmonter la malédiction de la dimensionnalité, améliorer l'efficacité computationnelle et faciliter la visualisation et l'interprétation des données. Cela permet d'extraire des informations utiles et de prendre des décisions éclairées à partir de ces séries temporelles.

Dans cette étude, nous avons choisi d'utiliser différentes techniques de réduction de dimension pour réduire la dimension des données de séries temporelles multivariées. Ces techniques ont été sélectionnées en raison de leur popularité et de leur efficacité dans la réduction de dimension.

Nous avons opté pour l'utilisation de l'analyse en composantes principales (PCA) en raison de sa capacité à capturer la variance maximale des données et à identifier les structures linéaires. le PCA est une méthode couramment utilisée dans l'analyse des données multivariées et offre une approche linéaire de la réduction de dimension.

L'analyse discriminante linéaire (ADL) a également été choisie en raison de sa capacité à maximiser la séparation entre les classes dans les données. Cette méthode est pertinente dans le cas où l'on cherche à extraire les caractéristiques les plus discriminantes des séries temporelles multivariées en vue d'une classification ultérieure.

Nous avons également inclus l'utilisation de l'UMAP et du t-SNE, deux méthodes de réduction

de dimension non linéaires, en raison de leur capacité à capturer des structures complexes et à fournir des visualisations de haute qualité. Ces techniques sont particulièrement utiles pour explorer les relations non linéaires et les clusters dans les données de séries temporelles multivariées.

Aussi,Nous avons également utilisé des autoencodeurs en raison de leur capacité à apprendre des représentations compressées des données quelque soit les relations entre les données (linéaire ou non) en utilisant des architectures neuronales. Les autoencodeurs sont capables de capturer des relations non linéaires complexes et de reconstruire les données d'origine à partir de la représentation réduite.

En comparant ces techniques, nous pourrons évaluer leur efficacité et leurs performances dans la réduction de dimension des données. Nous nous intéresserons à plusieurs critères tels que la préservation de l'information, la capacité à capturer les structures sous-jacentes des données, la conservation des relations entre les données, et la capacité à détecter des clusters ou des motifs intéressants.

Au regard de ces paramètres, nous pourrons déterminer quelle technique de réduction de dimension est la plus adaptée à notre étude spécifique.

L'Analyse en Composantes Principales (PCA)

L'Analyse en Composantes Principales (PCA) est une technique de réduction de dimension largement utilisée. Son objectif est d'identifier les directions principales de variation dans un espace de caractéristiques de grande dimension et de projeter les données sur un sous-espace de dimension inférieure qui capture ces variations principales.

Le PCA cherche à trouver les directions dans lesquelles les données présentent la plus grande variance, ce qui signifie qu'elles capturent la plus grande quantité d'information contenue dans les données. Ces directions sont appelées les composantes principales. Le PCA ordonne les composantes principales en fonction de leur contribution à la variance totale des données, de la plus grande à la plus petite.

Une fois que les composantes principales sont identifiées, le PCA effectue une projec-

tion orthogonale des données sur l'hyperplan généré par ces composantes principales. Cela permet de réduire la dimension des données, en conservant les informations les plus significatives.

Le PCA est particulièrement utile lorsque les données ont une structure linéaire, c'est-à-dire lorsque les variations peuvent être modélisées par des combinaisons linéaires des caractéristiques. Cependant, pour les données non linéaires, le PCA standard peut ne pas être efficace. C'est là que le PCA à noyau entre en jeu. Le PCA à noyau utilise une fonction de transformation non linéaire pour projeter les données dans un espace de dimension supérieure où elles peuvent être linéairement séparables. Cela permet de capturer les structures non linéaires dans les données et d'effectuer une réduction de dimension plus efficace.Kurita (2019).

L'Analyse Discriminante Linéaire (ADL)

L'analyse discriminante linéaire (ADL) est une technique d'apprentissage automatique linéaire utilisée principalement pour la classification multi-classes. Son objectif est de projeter les points de données dans de nouvelles dimensions de manière à obtenir une séparation maximale entre les classes et une séparation minimale au sein de chaque classe.

L'ADL cherche à trouver des combinaisons linéaires des caractéristiques originales qui maximisent la distance entre les clusters de différentes classes et minimisent la distance entre les points de données à l'intérieur d'un cluster et leur centroïde respectif. Cette approche vise à obtenir une meilleure séparabilité des classes dans les nouvelles dimensions, facilitant ainsi la classification.

Pour réaliser cela, l'ADL calcule les vecteurs propres (ou axes discriminants) correspondants aux plus grandes valeurs propres d'une matrice de dispersion interclasse et d'une matrice de dispersion intraclasse. La matrice de dispersion interclasse mesure la dispersion entre les clusters de différentes classes, tandis que la matrice de dispersion intraclasse mesure la dispersion à l'intérieur de chaque classe.

Les vecteurs propres associés aux plus grandes valeurs propres de ces matrices sont

utilisés pour construire la transformation linéaire qui projette les données dans les nouvelles dimensions discriminantes. Les nouvelles dimensions sont ordonnées en fonction de leur capacité à maximiser la séparation entre les classes.

En résumé, l'ADL est une technique de réduction de dimension qui vise à améliorer la séparabilité des classes dans un problème de classification multi-classes. Elle utilise une transformation linéaire pour projeter les données dans de nouvelles dimensions discriminantes, où la distance entre les classes est maximisée et la distance à l'intérieur des classes est minimisée. Cela permet d'améliorer la performance de la classification en réduisant la complexité des données tout en préservant les informations discriminantes. Ye et al. (2004).

UMAP (Uniform Manifold Approximation and Projection)

UMAP est une technique de réduction de dimension non linéaire qui vise à représenter des données complexes dans un espace de dimension réduite tout en préservant leurs relations structurelles. Elle est basée sur l'idée de modéliser la topologie des données à partir d'une approximation d'une variété uniforme.

L'algorithme UMAP fonctionne en construisant un graphe de voisinage à partir des données d'origine, où les points similaires sont connectés par des arêtes. La similarité entre les points est généralement mesurée en utilisant la distance euclidienne ou une autre mesure de distance appropriée. Ensuite, UMAP utilise une approche d'optimisation pour trouver une représentation de basse dimension où les distances entre les points du graphe sont préservées autant que possible.

La méthode UMAP se distingue par son utilisation de techniques d'optimisation stochastique et de la théorie des graphes. Elle utilise une fonction d'objectif basée sur la minimisation de la divergence de Kullback-Leibler entre la distribution des probabilités du graphe d'origine et celle de l'espace de dimension réduite. Cette fonction d'objectif est optimisée en utilisant des méthodes d'approximation stochastique pour obtenir une projection efficace.

Un aspect important d'UMAP est sa capacité à capturer les structures complexes des données, y compris les clusters, les groupes et les connexions non linéaires entre les points.

Elle peut également être adaptée pour prendre en compte des paramètres tels que la distance entre les voisins, le nombre de voisins considérés et la dimension de l'espace réduit.

t-SNE (t-Distributed Stochastic Neighbor Embedding)

La méthode t-SNE (t-Distributed Stochastic Neighbor Embedding) est une technique de réduction de dimension non linéaire utilisée pour visualiser des données complexes à haute dimension. Elle est largement utilisée dans le domaine de la visualisation des données pour explorer et comprendre la structure des données.

L'algorithme t-SNE fonctionne en plusieurs étapes. Tout d'abord, il calcule les similarités entre les points de données dans l'espace d'origine à l'aide d'une mesure de similarité telle que la distance euclidienne. Ensuite, il construit une matrice de similarité conditionnelle qui représente la similarité relative entre les points. À partir de cette matrice, il calcule les probabilités de similarité conjointe entre les points dans l'espace de réduction de dimension en utilisant une distribution t-distribuée. Enfin, t-SNE minimise la divergence Kullback-Leibler entre les probabilités de similarité conditionnelle et conjointe en ajustant la position des points dans l'espace de réduction de dimension. L'algorithme est itératif et utilise des techniques d'optimisation pour trouver la meilleure configuration des points. Il produit une représentation visuelle des données dans un espace bidimensionnel ou tridimensionnel, où les points similaires sont regroupés et les points dissimilaires sont séparés. Cette visualisation permet d'explorer et de comprendre la structure des données de manière plus claire.

Les Autoencodeurs Profonds (DAE)

L'autoencodeur est une technique de réduction de dimension basée sur des réseaux de neurones artificiels non supervisés. Son principe fondamental est d'apprendre à coder les données dans un espace de dimensions réduites, appelé espace latent, et de les décoder pour retrouver leurs dimensions d'origine.

L'autoencodeur se compose de deux parties principales : l'encodeur et le décodeur. L'encodeur prend en entrée les données d'origine et les transforme dans l'espace latent de di-

mensions réduites. Le décodeur, quant à lui, reçoit les données codées dans l'espace latent et les décode pour les reconstruire dans les dimensions d'origine.

L'objectif de l'autoencodeur est de minimiser la différence entre les données d'entrée et les données de sortie reconstruites. Pour cela, il est entraîné en utilisant des données non étiquetées, c'est-à-dire sans informations de classe ou de catégorie. L'apprentissage de l'autoencodeur se fait en ajustant les poids des neurones de l'encodeur et du décodeur afin de minimiser une fonction de perte, généralement mesurée par l'entropie croisée binaire ou l'erreur quadratique moyenne.

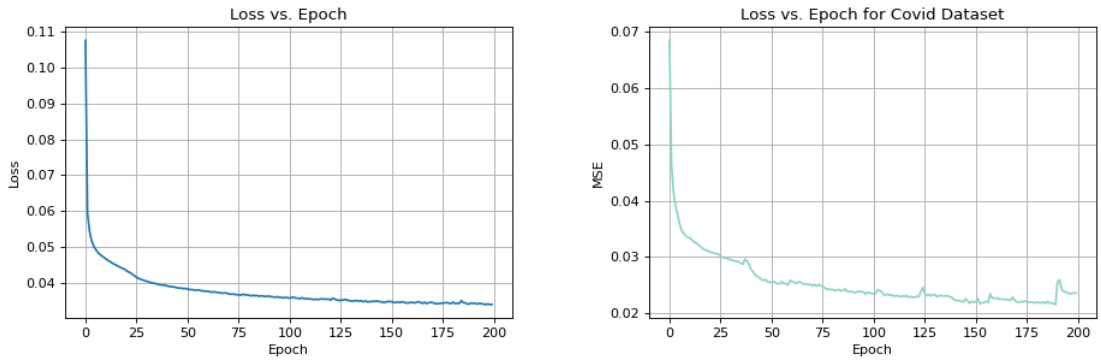


FIGURE 3.4 – Loss vs Epoque

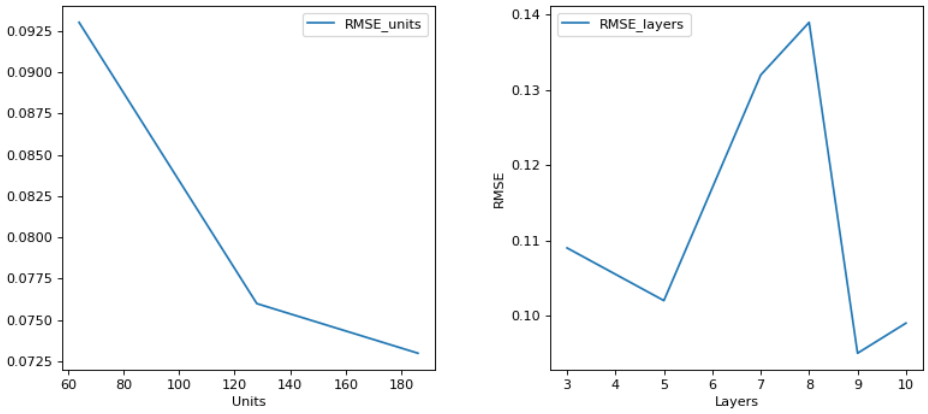


FIGURE 3.5 – Variation des hyperparamètres sur le jeu de données ECG

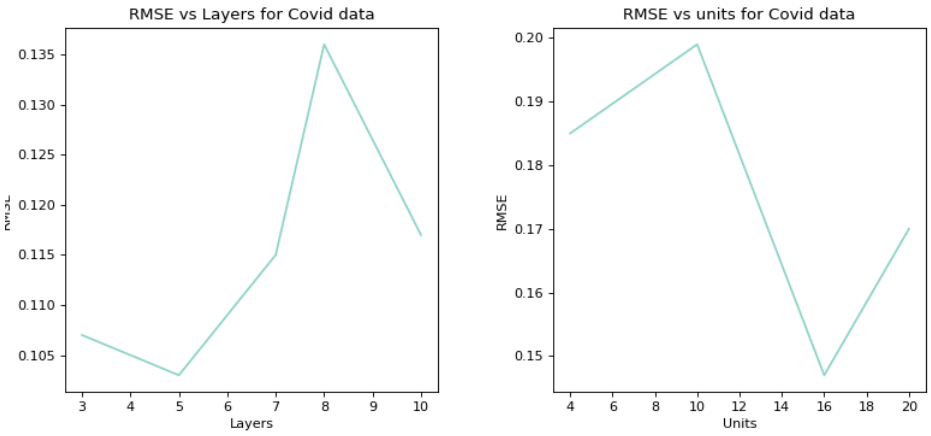


FIGURE 3.6 – Variation des hyperparamètres sur le jeu de données Covid

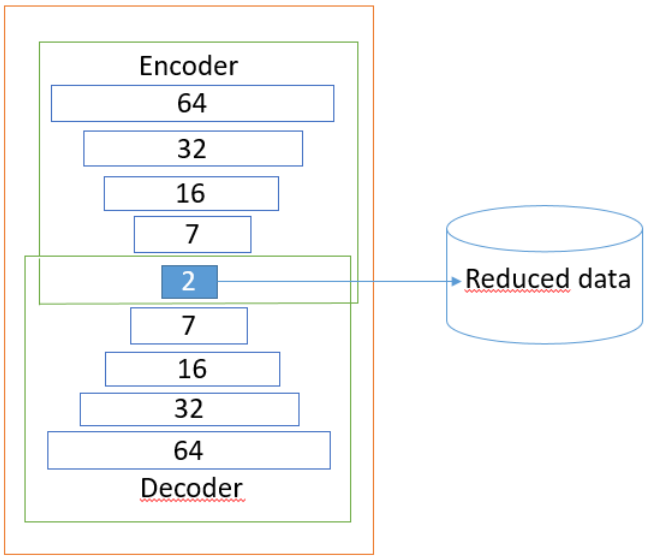


FIGURE 3.7 – Architecture de DAE pour les données ECG

3.2.5 Révision des clusters

Lors de la phase de regroupement en clusters, il est possible que de nouveaux points centraux modifient les résultats du regroupement des points déjà traités, ce qui peut entraîner des changements dans l'assignation des points aux clusters. Ce phénomène est connu sous le nom de conflit inter-cluster. Pour évaluer si de tels conflits se produisent, nous avons proposé l'utilisation d'une technique de révision des clusters Zhang et al. (2021).

L'objectif de cette technique est de corriger les erreurs de regroupement en se basant sur l'homogénéité des clusters obtenus. Le principe de la méthode repose sur les observations

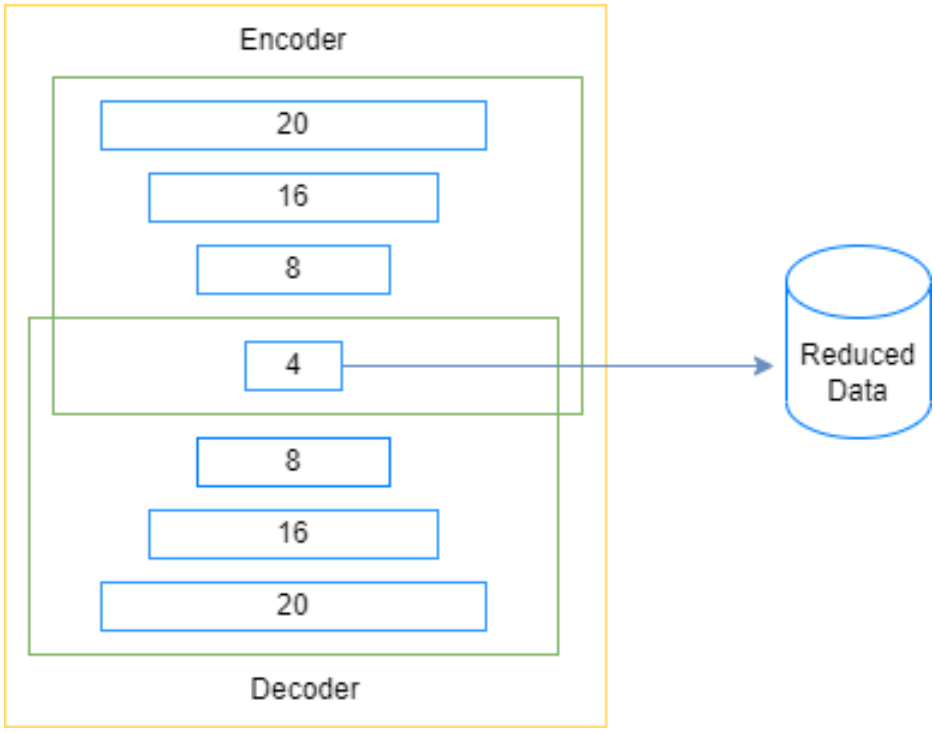


FIGURE 3.8 – Architecture de DAE pour les données Covid

suivantes : si la distance entre deux points x_i et x_j est inférieure à un seuil prédéfini ϵ, et que x_j est attribué à un cluster différent de celui de x_i, cela indique un conflit potentiel.

Pour résoudre ce conflit, une mesure de similarité appelée $Sim(x_i, x_j)$ est utilisée. Elle est calculée en comparant les ensembles des échantillons qui sont à la fois parmi les k plus proches voisins de x_i (représenté par $kNN(x_i)$) et parmi les k plus proches voisins de x_j (représenté par $kNN(x_j)$). Le paramètre K correspond au nombre de voisins les plus proches pris en compte dans le calcul.

Ensuite, en utilisant la valeur de similarité $Sim(x_i, x_j)$, une décision est prise pour réviser l'assignation des points aux clusters. Si $Sim(x_i, x_j)$ dépasse un seuil prédéfini α, cela signifie que les points x_i et x_j sont suffisamment similaires pour être attribués au même cluster. Dans ce cas, x_j est réassigné au cluster de x_i. Si $Sim(x_i, x_j)$ est inférieur à α, aucun changement n'est effectué et x_j reste dans son cluster d'origine.

Dans l'étude Zhang et al. (2021), une valeur seuil de $\alpha = 0.5$ est recommandée pour la correction des erreurs de regroupement. Cependant, il convient de noter que ce seuil peut être ajusté en fonction des caractéristiques spécifiques du jeu de données et des objectifs

de l'analyse.

En résumé, la technique de révision des clusters permet de détecter les conflits inter-cluster et de les résoudre en réassignant les points à des clusters appropriés. Elle utilise une mesure de similarité basée sur les plus proches voisins pour prendre des décisions de révision. Cette approche contribue à améliorer la qualité du regroupement et à obtenir des résultats plus fiables dans l'analyse de données.

3.3 Évaluation comparative entre différents modèles

Nous avons procédé à l'évaluation et à l'optimisation des modèles de clustering en utilisant des données étiquetées. Une fois les meilleurs modèles sélectionnés, nous les avons appliqués aux données provenant du champ agricole. Cependant, afin de simplifier notre analyse initiale, nous avons choisi de travailler avec des données qui présentaient des caractéristiques similaires. Cela nous a permis de mieux comprendre le comportement des modèles de clustering et d'identifier les approches les plus efficaces avant de les appliquer aux données agricoles, qui présentent des variations de densité spécifiques au domaine.

Nos modèles sont exécutés dans un environnement Colab équipé de 12 Go de RAM et 107 Go de disque dur, en utilisant les 100 unités de calcul offertes par Colab. Nous avons pris diverses mesures pour optimiser notre code. Au départ, nous avons mis en œuvre les modèles en tant que fonctions Python standards, puis nous avons adopté des techniques de programmation orientée objet pour réduire encore plus le temps de calcul. En outre, nous avons utilisé la bibliothèque joblibspark pour paralléliser les calculs et améliorer les performances.

Nous avons réalisé une analyse comparative entre les résultats obtenus en utilisant la méthode DBSCAN de la bibliothèque sklearn et notre propre implémentation de DBSCAN. L'implémentation de DBSCAN par nous-même est indispensable afin de l'adapter aux données spatio-temporelles. La comparaison a été effectuée en utilisant divers ensembles de données décrits dans le chapitre 2. Il est important de noter que les deux implémentations ont produit

des résultats identiques pour chaque ensemble de données utilisé.

Dans notre étude, nous avons appliqué le modèle DBSCAN classique à quatre ensembles de données synthétiques présentant des densités de clusters variables. Les résultats obtenus ont été analysés de manière approfondie pour évaluer la performance de l'algorithme.

Pour les ensembles de données Spiral et Aggregation (la section 2.1), l'algorithme DBSCAN a fourni des résultats remarquables. Les taux de bruit détecté étaient très faibles, avec seulement 0,32% de bruit détecté pour l'ensemble de données Spiral et 1,27% pour l'ensemble de données Aggregation. De plus, les métriques d'évaluation telles que l'homogénéité, la complétude, la V-mesure, l'ARI et l'AMI ont montré des valeurs élevées (2.2), indiquant une forte similitude entre les clusters prédits et les clusters réels pour ces ensembles de données.

Cependant, la performance de DBSCAN n'était pas reproductible pour les ensembles de données Flame et Compound (la section 2.1). Pour l'ensemble de données Compound, seuls 3 clusters ont été détectés au lieu des 6 attendus (Tableau 3.4), tandis que pour l'ensemble de données Flame, 3 clusters ont été identifiés au lieu de 2. De plus, pour ces ensembles de données, DBSCAN a identifié plus de 10% des données comme étant du bruit (Tableau 3.3, Figure 3.10).

Ces résultats soulignent l'impact de la densité de cluster sur la performance de l'algorithme DBSCAN. Les ensembles de données Spiral et Aggregation, avec des clusters bien définis et une densité uniforme, ont produit des résultats plus fiables. En revanche, les ensembles de données Flame et Compound, avec des structures plus complexes et des densités de clusters variables, ont présenté des difficultés pour l'algorithme DBSCAN.

Nous avons appliqué l'algorithme DBSCAN à l'ensemble de données ECG (voir la section 2.1), qui est une collection de séries temporelles multivariées massives comprenant 187 variables et plus de 87 000 enregistrements répartis en cinq clusters distincts. Pour réduire le temps d'exécution des modèles, nous avons limité le nombre d'enregistrements à 20 000 échantillons.

Malheureusement, DBSCAN n'a pas réussi à regrouper de manière satisfaisante cet ensemble de données (Figure 3.9), considérant près de la moitié des données ECG comme du

bruit (Tableau 3.4, première colonne). Des résultats similaires ont été observés lors de l'application de DBSCAN à l'ensemble de données COVID (voir la section 2.1), où 48,73% des données ont été identifiées comme du bruit. De plus, les métriques d'évaluation telles que l'homogénéité, l'ARI et l'AMI ont montré des scores très faibles, ce qui indique que les clusters détectés n'étaient pas homogènes et que certains points d'un même cluster pouvaient appartenir à différents clusters réels (Tableau 3.5).

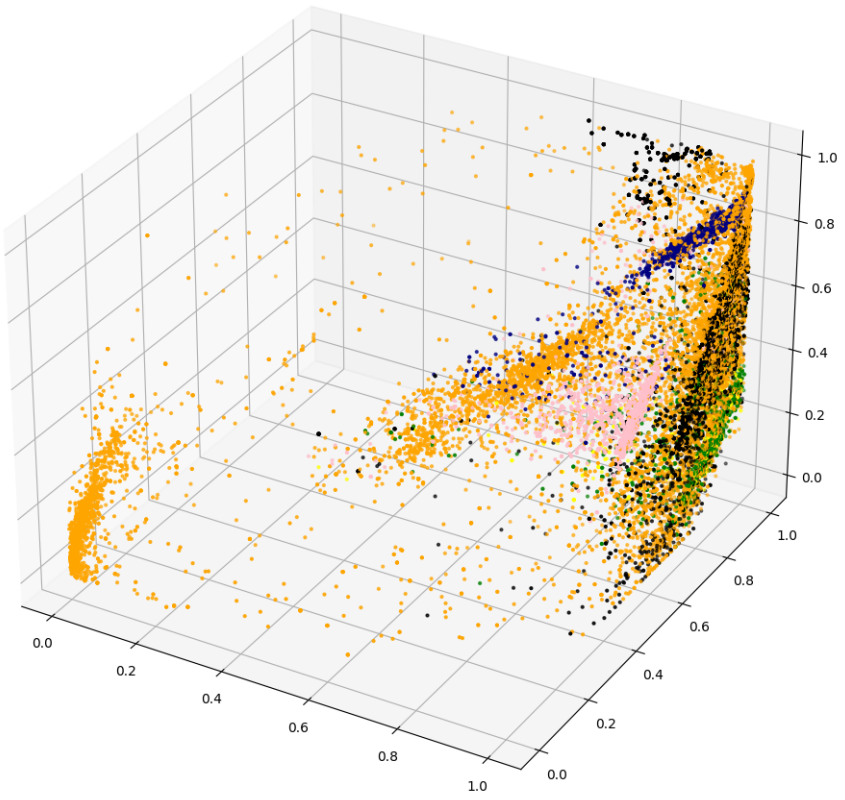


FIGURE 3.9 – dbscan sur ECG

Plusieurs facteurs peuvent expliquer cette faible performance de regroupement. Le grand nombre de variables utilisées, le choix de la mesure de similarité et les valeurs des hyperparamètres ϵ et $MinPts$ sélectionnées peuvent tous impacter les résultats. Pour tenter de remédier à ces problèmes, nous avons proposé de réduire la dimension en utilisant des techniques de réduction de dimension telles que PCA, LDA, UMAP et t-SNE et autoencodeur. De plus, nous avons exploré différentes valeurs pour ϵ et $MinPts$ sur une large plage et avons

utilisé deux mesures de similarité différentes, notamment la distance euclidienne et d'autres mesures adaptées aux séries temporelles comme le DTW.

Clustering après réduction de la dimension

Nous avons appliqué différentes méthodes de réduction de dimension à des ensembles de données variés, tels que ECG et COVID, qui présentaient un grand nombre de variables. En revanche, les ensembles de données plus petits tels que Aggregation, Flame, Spiral et Compound ne nécessitaient pas de réduction de dimension. Pour notre étude, nous avons commencé par appliquer les techniques de réduction sur les données ECG.

L'une des techniques utilisées était l'Analyse en Composantes Principales (PCA). Nous avons analysé la variation du pourcentage de variance expliquée en fonction du nombre de composantes, basée sur les 187 caractéristiques initiales. Nous avons choisi de conserver les composantes principales qui expliquaient au moins 80% de la variance, ce qui correspondait à neuf composantes principales. En visualisant les scores des composantes principales dans un espace bidimensionnel formé par les deux premières composantes (PC1 et PC2), nous avons pu observer l'influence des variables sur ces composantes principales. Certaines variables présentaient une corrélation positive, tandis que d'autres n'avaient aucune corrélation ou une corrélation négative.

Par la suite, nous avons appliqué la réduction de dimension par PCA aux données ECG en faisant varier le paramètre ϵ de 0,5 à 0,9. Cette méthode a permis de réduire considérablement le nombre de points identifiés comme du bruit, passant de 52% à seulement 1% lorsque ϵ était fixé à 0,9. Cependant, le nombre de clusters détectés a presque doublé par rapport au nombre réel de clusters. De plus, ces clusters présentaient une faible homogénéité et un indice d'AMI proche de zéro. Ces résultats indiquent que l'utilisation de PCA seule peut ne pas être adaptée à la réduction de dimension des données de séries temporelles multivariées (Tableau 3.6).

Nous avons utilisé l'Analyse Discriminante Linéaire (LDA) comme deuxième technique de réduction de dimension, en supposant une relation linéaire entre les caractéristiques. Nous

avons étudié les résultats de LDA en faisant varier les valeurs de ϵ de 0.03 à 0.05. Nous avons observé une réduction significative du bruit, avec un bruit réduit à seulement 2.49% lorsque nous avons utilisé une valeur d'epsilon de 0.03. Cependant, l'indice d'homogénéité des clusters était proche de zéro et le coefficient de silhouette était négatif. Ces résultats suggèrent que les points de frontière des différents clusters étaient mal séparés, ce qui conduisait à un regroupement médiocre. En conséquence, LDA a moins bien performé que PCA, car l'hypothèse de linéarité de LDA contredit la nature des données ECG (Tableau 3.6).

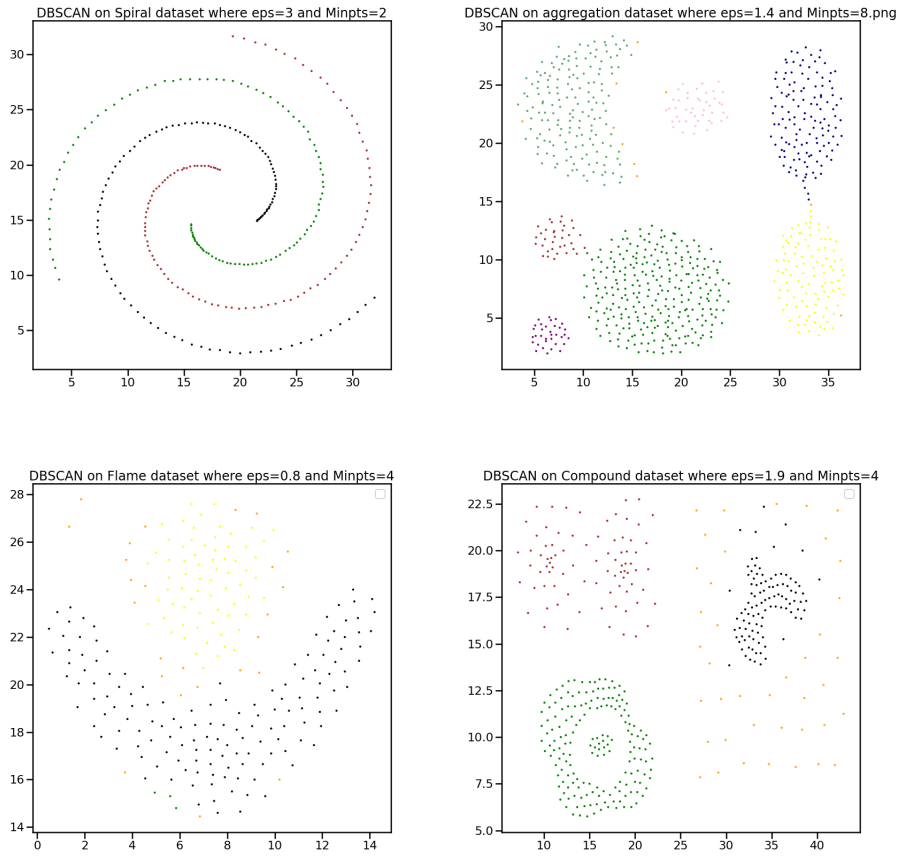


FIGURE 3.10 – DBSCAN sur des données synthétiques

Nous avons ensuite appliqué une troisième technique appelée "Deep Autoencoder" (autoencodeur profond). nous avons réalisé une normalisation des données en utilisant la mise à l'échelle MinMaxScaler, ce qui a permis de mettre les données dans une plage de valeurs entre 0 et 1.

L'autoencodeur profond est un réseau de neurones qui est entraîné pour reconstruire les

données d'entrée en passant par une couche de compression (l'encodeur) et une couche de décompression (le décodeur). La performance de l'autoencodeur est évaluée en mesurant la différence entre les données d'entrée et les données reconstruites, en utilisant des métriques telles que l'entropie croisée binaire et l'erreur quadratique moyenne.

Pour trouver la meilleure architecture d'autoencodeur, nous avons effectué des expérimentations en utilisant une combinaison de recherche en grille et de recherche aléatoire, mais les résultats étaient instables et le temps de calcul était élevé. Par conséquent, nous avons choisi d'effectuer une exploration manuelle en variant le nombre de couches, de neurones et d'époques (Figures3.5, 3.6 et 3.4).

Pour l'ensemble de données ECG, nous avons finalement identifié la meilleure architecture d'autoencodeur, composée de neuf couches avec deux neurones dans la couche de compression (couche cachée) et une formation sur quarante époques (Figure 3.7). Pour l'ensemble de données COVID, la meilleure architecture comprenait sept couches avec quatre neurones dans la couche de compression et une formation sur vingt-cinq époques (Figure 3.8).

Une fois les hyperparamètres optimisés, nous avons appliqué cette architecture d'autoencodeur profond aux données ECG. Les deux caractéristiques extraites par la couche de compression ont ensuite été utilisées comme entrée dans l'algorithme DBSCAN, en faisant varier les valeurs de ϵ de 0.01 à 0.03, avec un $MinPts$ fixé à 10. Cela a permis de réduire considérablement le bruit dans les clusters, avec seulement 0.58% de bruit détecté lorsque nous avons utilisé une valeur d'ϵ de 0.03. De plus, notre approche a réussi à estimer le même nombre de clusters réels avec un coefficient de silhouette élevé de 0.623, indiquant une bonne séparation entre les clusters (Tableau 3.6).

Dans notre exploration pour améliorer la représentation des données ECG à travers la réduction de dimension, nous avons appliqué l'algorithme UMAP (Uniform Manifold Approximation and Projection) (Tableau 3.2). Notre démarche s'est avérée être une expérience itérative visant à découvrir les paramètres optimaux pour une performance de regroupement optimale. Initialement, nous avons employé l'algorithme GridSearch dans le but de déterminer

TABLE 3.1 – Récapitulatif des combinaisons de paramètres de UMAP et des scores de silhouette

Combinaison	n_neighbors	min_dist	n_components	Score de silhouette
1	1000	0.1	10	0.077
2	800	0.3	8	0.0628
3	800	0.2	6	0.0644
4	600	0.5	10	0.05516
5	600	0.5	30	0.052
6	500	0.2	6	0.053
7	400	0.5	8	0.045
8	300	0.7	7	0.0425
9	300	0.3	5	0.047
10	400	0.1	8	0.0473
11	300	0.5	10	0.044
12	100	0.5	10	0.040
13	200	0.9	15	0.036
14	200	0.1	10	0.05344
15	10	0.5	10	-0.015
16	10	0.5	2	-0.074

les combinaisons de paramètres idéales en recourant à la validation croisée. Cependant, nous avons rapidement constaté que GridSearch n'était pas compatible avec nos données, ce qui nous a poussés à entamer une exploration manuelle.

Cette exploration a consisté en l'ajustement de trois paramètres clés : le nombre de voisins (n_neighbors), la distance minimale (min_dist), et le nombre de composantes (n_components). Pour chaque ensemble de paramètres, nous avons évalué les résultats en utilisant la métrique de silhouette, une mesure apte à évaluer la cohérence des groupements obtenus (Tableau 3.1). Bien que notre démarche ne soit pas structurée comme celle de GridSearch, nous avons méthodiquement enregistré les scores de silhouette associés à chaque combinaison de paramètres. Cette approche pragmatique nous a orientés dans nos choix, nous permettant de cibler progressivement les configurations générant des scores de silhouette plus élevés, attestant d'une meilleure qualité de regroupement.

Les résultats obtenus à partir des différentes configurations de réduction de dimension et de clustering (Tableau 3.1) mettent en évidence les difficultés liées à la réduction de dimension des données de séries temporelles multivariées à l'aide de UMAP. Parmi les configurations testées, aucune n'a réussi à produire des résultats satisfaisants en termes de performances de clustering et d'évaluation.

L'analyse des scores d'évaluation révèle des résultats peu concluants pour la plupart des configurations UMAP. Les valeurs de la métrique de silhouette (SC) indiquent que les clusters générés sont souvent mal définis, avec des valeurs négatives suggérant que les points sont dispersés et ne forment pas de groupes compacts. De plus, les scores d'homogénéité, de complétude et de V-mesure montrent que les clusters produits par UMAP ont du mal à capturer la structure réelle des données sous-jacentes. Cela se traduit par des taux élevés de points de bruit et des scores faibles en termes de qualité des regroupements.

Particulièrement, la configuration "UMAP (ϵ=0.2, Minpts=100)" présente des résultats particulièrement médiocres, avec une homogénéité, une complétude et une V-mesure très faibles. Ces résultats indiquent que UMAP a du mal à extraire les structures discriminantes et significatives des données de séries temporelles multivariées, ce qui conduit à des clusters peu cohérents et mal définis.

Cela suggère que, dans le contexte de données de séries temporelles multivariées comme les ECG, UMAP peut ne pas être la meilleure méthode de réduction de dimension pour capturer les motifs et les relations complexes présents dans les données. Des approches alternatives ou une combinaison de différentes techniques de réduction de dimension pourraient être nécessaires pour obtenir des résultats plus significatifs et informatifs en ce qui concerne la structure des données ECG multivariées.

La dernière technique de réduction de dimension que nous avons employée dans notre exploration des données ECG est l'algorithme t-SNE (Tableau 3.2) . Contrairement aux méthodes précédentes telles que l'UMAP, t-SNE se concentre spécifiquement sur la préservation des relations locales entre les points dans l'espace d'origine. Lors de notre tentative initiale d'application de t-SNE, nous avons été confrontés à une contrainte technique liée aux méthodes

d'implémentation proposées par la bibliothèque scikit-learn. Malheureusement, l'algorithme impose une limitation sur le nombre de dimensions (n_components), le restreignant à une valeur inférieure à 4 en raison de son utilisation d'approximations arborescentes.

Cette approche nous a permis de réduire notre ensemble de données ECG à seulement trois variables, en préservant les relations locales entre les points.

En explorant davantage, l'algorithme t-SNE a également été appliqué aux données ECG en variant les configurations de paramètres. Parmi ces diverses configurations, l'utilisation de t-SNE avec un ϵ de 1 et un Minpts de 20 a permis d'identifier trois clusters distincts. Néanmoins, il est crucial de noter qu'une majorité écrasante des points (99,66 %) a été catégorisée comme du bruit, ce qui soulève des inquiétudes quant à la validité de cette segmentation. Les indicateurs de performance dévoilent des scores particulièrement faibles en ce qui concerne l'homogénéité, la complétude, la V-measure et l'Adjusted Rand Index (ARI), mettant en évidence une correspondance limitée entre les clusters détectés et les groupes réels présents dans les données. Par ailleurs, l'Adjusted Mutual Information (AMI) ainsi que le Silhouette Coefficient (SC) mettent en avant une discordance et une séparation peu évidentes entre les clusters. Ces résultats témoignent du fait que t-SNE ne parvient pas efficacement à saisir les structures de regroupement, particulièrement dans le contexte des séries temporelles.

Nous avons tenté egalement d'augmenter la valeur de ϵ à 4 tout en réduisant la valeur de Minpts. Cependant, en comparant les résultats obtenus avec l'algorithme t-SNE aux données réelles, une divergence significative devient apparente. Contrairement aux cinq clusters distincts présents dans les données réelles, t-SNE parvient à identifier seulement trois clusters. Cette incohérence entre les nombres de clusters réels et prédits soulève des questions quant à la capacité de t-SNE à pleinement appréhender la complexité des données de séries temporelles. Les mesures d'évaluation renforcent cette constatation. L'homogénéité de 0,062 indique une adéquation limitée entre les clusters prédits et les clusters réels, tandis que la complétude de 0,393 révèle que certains membres des clusters réels sont disséminés parmi plusieurs clusters prédits.

TABLE 3.2 – DBSCAN avec UMAP et t-SNE sur les données ECG

	UMAP (ϵ=0.2, Minpts=100)	UMAP (ϵ=0.2, Minpts=20)	UMAP (ϵ=0.6, Minpts=20)	t-SNE (ϵ=1, Minpts=20)	t-SNE (ϵ=3, Minpts=10)	(ϵ=4, Minpts=10)
Nb-principal components	10	10	10	3	3	3
Nb-clusters	6	144	86	3	29	3
Noise points %	94.62	26.47	6.31	99.66	0.92	0.07
Homogeneity	0.055	0.744	0.941	0.003	0.220	0.062
Completeness	0.292	0.339	0.477	0.208	0.341	0.393
V-measure	0.092	0.466	0.633	0.007	0.268	0.107
ARI	0.003	0.133	0.513	0.000	0.063	0.012
AMI	0.092	0.463	0.631	0.006	0.266	0.106
SC	-0.142	0.115	-0.013	-0.380	-0.425	-0.121

En résumé, l'utilisation de l'auto-encodeur profond comme technique de réduction de dimension a permis d'améliorer significativement les résultats de regroupement.

Sur la base des résultats de notre expérience, nous avons déterminé que la méthode la plus efficace pour réduire les caractéristiques des données de séries temporelles multivariées est l'utilisation d'un autoencodeur profond. Par conséquent, nous avons décidé d'appliquer cette technique aux données spatio-temporelles de Covid.

Nous avons utilisé un autoencodeur à trois couches (Encodeur/Décodeur) pour réduire les données de Covid à quatre caractéristiques. Suite à une investigation empirique des hyperparamètres tels que $Minpts$ et ϵ, nous avons sélectionné une paire de valeurs ($Minpts = 70$, $\epsilon = 0.12$). En appliquant l'autoencodeur aux données de Covid, nous avons pu améliorer le regroupement par DBSCAN. Plus précisément, cette approche a réduit le pourcentage de points regroupés en tant que bruit de 48.73% à 2.63%. Les points restants ont été répartis dans trois clusters, ce qui est un nombre raisonnable compte tenu du nombre réel de clusters qui est de deux.

Bien que l'homogénéité (qui est proche de 10^{-3}) et l'indice AMI des clusters soient très bas, les résultats sont acceptables en termes de distance moyenne entre les points d'un cluster et ses clusters voisins, tant que le coefficient de silhouette (0.071) est positif.

Dans les expériences suivantes, notre objectif était de sélectionner la valeur adaptative de ϵ en se basant sur les quatre hypothèses (H1, H2, H3, H4) exposées dans la section 3.2.3. Pour évaluer l'efficacité de ces hypothèses, nous les avons appliquées à l'ensemble de données Compound, qui contient des clusters de densités variables. Les résultats, comme le montre la Figure 3.11, ont révélé qu'en utilisant un ϵ adaptatif conjointement avec un DBSCAN classique

basé sur H1 (appelé D-DBSCAN(H1)), nous avons obtenu un regroupement de grande qualité, avec des mesures d'ARI et d'AMI proches de 1.

Cependant, lorsque nous avons appliqué les hypothèses H2, H3 ou H4, nous avons obtenu soit des clusters de bruit, soit un seul cluster, indiquant que le ϵ adaptatif n'était pas pertinent dans ces cas. Par conséquent, nous avons décidé de continuer à utiliser l'hypothèse H1 pour les expériences ultérieures. La Figure 3.12 résume les résultats obtenus en utilisant D-DBSCAN avec H1 sur les autres ensembles de données.

Par la suite, nous avons appliqué D-DBSCAN avec l'hypothèse H1 D-DBSCAN(H1) aux ensembles de données. Sur les données synthétiques, nous avons réussi à regrouper avec succès les données en forme de spirale en évitant tout bruit détecté, et l'homogénéité des clusters a atteint sa valeur maximale de 1. De plus, D-DBSCAN(H1) a produit un regroupement parfait des données en forme de spirale, correspondant aux clusters originaux.

De manière individuelle pour chaque jeu de données, nous pouvons observer ce qui suit lors de la comparaison des modèles proposés. Pour les données de Compound, le modèle D-DBSCAN(H1), qui intègre l'hypothèse de l'epsilon adaptative , a identifié 5 clusters pour les données Compound. Le pourcentage de points de bruit a augmenté à 24,56 %, cependant, les mesures d'homogénéité (0,838) et de complétude (0,903) indiquent une séparation parfaite des clusters.

Le modèle MDST-DBSCAN classique a également identifié 5 clusters, cependant, il a généré un pourcentage élevé de points de bruit (22,56 %). Les mesures d'homogénéité (0,857) et de complétude (0,904) indiquent une séparation raisonnable des clusters, bien que certains points puissent être mal attribués (Figure 3.13).

Le modèle D-MDST-DBSCAN(H1), qui combine l'hypothèse de l'epsilon adaptative et le modèle MDST-DBSCAN, a identifié les 6 clusters réels pour les données Compound. Il a réduit le pourcentage de points de bruit à 19,30 %, et les mesures d'homogénéité (0,891) et de complétude (0,901) indiquent une légère amélioration de la séparation des clusters par rapport au modèle MDST-DBSCAN classique (Figure 3.14).

Le modèle DBSCAN(DTW), qui utilise la mesure de similarité DTW à la place de la distance

euclidienne, a identifié 5 clusters pour les données Compound. Il a réduit le pourcentage de points de bruit à 14,29 %. Les mesures d'homogénéité (0,842) et de complétude (0,927) indiquent une séparation des clusters raisonnable, suggérant que l'utilisation de la mesure de similarité DTW a été bénéfique pour améliorer la détection des clusters dans les données Compound.

Le modèle D-DBSCAN(H1, DTW), qui combine l'hypothèse de l'epsilon adaptative et la mesure de similarité DTW, a également identifié 5 clusters pour les données Compound. Il a réduit le pourcentage de points de bruit à 15,79 %. Les mesures d'homogénéité (0,828) et de complétude (0,918) indiquent une séparation des clusters raisonnable, mais il reste des zones d'amélioration.

Le modèle DBSCRN, qui utilise le KNN pour la détection des voisinages des points, a identifié un grand nombre de clusters (183 clusters) pour les données Compound.

Nous pouvons deduire que l'utilisation de différents modèles de clustering sur les données qui présentent une densité variable, a permis d'obtenir des résultats intéressants. Le modèle D-DBSCAN(H1) avec l'hypothèse de l'epsilon adaptative a donné une séparation parfaite des clusters, tandis que le modèle MDST-DBSCAN classique a obtenu une séparation raisonnable des clusters, mais avec un pourcentage élevé de points de bruit. Le modèle DBSCAN(DTW) utilisant la similarité DTW a également donné des résultats raisonnables.

Pour les données de séries temporelles multivariées ECG, le modèle DBSCAN classique a identifié 5 clusters pour les données ECG, avec un pourcentage élevé de points de bruit (52,22 %). Les mesures d'homogénéité (0,319) et de complétude (0,398) indiquent une séparation des clusters limitée.

Le modèle D-DBSCAN(H1) a également identifié 7 clusters pour les données ECG, avec un pourcentage de points de bruit réduit à 4,08 %. Les mesures d'homogénéité (0,018) et de complétude (0,094) indiquent une séparation des clusters insatisfaisante.

Le modèle MDST-DBSCAN classique a identifié 0 clusters, il a généré un pourcentage élevé de points de bruit (100 %) pour les données ECG. Cela suggère que le modèle a eu du mal à détecter des motifs significatifs dans les données.

Le modèle D-MDST-DBSCAN(H1) a identifié 8 clusters pour les données ECG, avec un pourcentage de points de bruit réduit à 5,32 %. Les mesures d'homogénéité (0,026) et de complétude (0,106) indiquent une séparation des clusters limitée.

Le modèle DBSCAN(DTW) a identifié 5 clusters pour les données ECG, avec un pourcentage de points de bruit de 2,28 %. Les mesures d'homogénéité (0,039) et de complétude (0,155) indiquent une séparation des clusters améliorée par rapport au modèle DBSCAN classique.

Le modèle D-DBSCAN(H1, DTW) qui combine H1, DAE et DTW, a identifié 5 clusters pour les données ECG, avec un pourcentage de points de bruit de 15,79 %. Les mesures d'homogénéité (0,016) et de complétude (0,090) indiquent une séparation des clusters limitée.

Enfin, le modèle DBSCRN a identifié un nombre élevé de clusters (1039 clusters) pour les données ECG. Cependant, les mesures d'homogénéité (0,802) et de complétude (0,190) indiquent une séparation des clusters imparfaite et une segmentation excessive des données.

En conclusion, l'utilisation de différentes variations de modèles de clustering sur les données Compound et ECG a montré des performances variables en termes de séparation des clusters. L'intégration de l'hypothèse de l'epsilon adaptative, de la réduction de dimension avec DAE et de l'utilisation de la mesure de similarité DTW a généralement amélioré les résultats, mais certains modèles ont encore des difficultés à filtrer le bruit et à obtenir une séparation précise des clusters. Des efforts supplémentaires sont nécessaires pour améliorer la performance et l'interprétabilité des méthodes de clustering appliquées à ces deux types de données.

De plus, nous avons également testé ces modèles de clustering sur un autre ensemble de données de séries temporelles multivariées liées au Covid, qui présentaient une nouvelle distribution, afin de généraliser la performance des modèles sélectionnés.

Pour le modèle DBSCAN classique, nous avons observé l'identification de 3 clusters, mais avec un pourcentage élevé de points de bruit (48.73%). Les mesures d'homogénéité (0.015) et de complétude (0.009) indiquent une faible séparation des clusters, avec une association limitée entre les clusters prédits et les étiquetages de référence. Les autres métriques telles

que la mesure V (0.011), l'ARI (0.021), l'AMI (0.011) et le coefficient de silhouette (SC) (0.043) confirment également une performance limitée du modèle.

Le modèle D-DBSCAN(H1) a réussi à identifier 3 clusters, mais avec un pourcentage de points de bruit considérablement réduit à 2.63%. Cependant, les mesures d'homogénéité (0.009) et de complétude (0.008) indiquent une faible séparation des clusters, avec une amélioration marginale par rapport au modèle DBSCAN classique. Les autres métriques telles que la mesure V (0.009), l'ARI (0.011), l'AMI (0.009) et le coefficient de silhouette (SC) (0.071) indiquent également une amélioration légère mais limitée des performances du modèle.

Le modèle MDST-DBSCAN a identifié un nombre élevé de clusters (53) pour les données Covid. Cependant, cela a été accompagné d'un pourcentage élevé de points de bruit (40.76%). Les mesures d'homogénéité (0.075) et de complétude (0.015) suggèrent une séparation modérée des clusters, mais avec une association limitée aux étiquetages de référence. Les autres métriques telles que la mesure V (0.024), l'ARI (0.014), l'AMI (0.018) et le coefficient de silhouette (SC) (0.246) indiquent une amélioration relativement significative des performances du modèle par rapport aux modèles DBSCAN classique et D-DBSCAN(H1).

En ce qui concerne le modèle D-MDST-DBSCAN(H1), nous avons constaté l'identification de 5 clusters, avec un pourcentage de points de bruit réduit à 3.08%. Les mesures d'homogénéité (0.008) et de complétude (0.010) indiquent une séparation limitée des clusters, avec une amélioration légère par rapport au modèle MDST-DBSCAN. Cependant, les métriques telles que la mesure V (0.009), l'ARI (-0.018), l'AMI (0.007) et le coefficient de silhouette (SC) (0.032) suggèrent des performances globalement inférieures au modèle MDST-DBSCAN.

Le modèle D-DBSCAN(H1,DTW) a identifié 2 clusters, avec un pourcentage de points de bruit de 1.46%. Les mesures d'homogénéité (0.020), de complétude (0.020) et la mesure V (0.020) indiquent une séparation raisonnable des clusters, avec une amélioration significative par rapport aux autres modèles. L'ARI (-0.004) indique une concordance légèrement inférieure aux étiquetages de référence, tandis que l'AMI (0.019) suggère une association relativement précise entre les clusters prédits et les étiquetages de référence. Le coefficient de silhouette (SC) présente une valeur élevée de 0.387, indiquant une séparation nette des

TABLE 3.3 – Spiral - Aggregation - Flame

	DBSCAN classique			D-DBSCAN (H1)			MDST DBSCAN et D-MDST DBSCAN (H1)		
	Spiral	Aggre	Flame	Spiral	Aggre	Flame	Spiral	Aggre	Flame
Nb-clusters	3	7	3	3	8	7	3	8	2
Points de bruit (%)	0,32	1,27	10,42	0	3,17	17,50	0	2,16	7,5
Homogénéité	1,000	0,990	0,906	1	0,970	0,815	1	0,983	0,894
Complétude	0,984	0,962	0,620	1	0,904	0,309	1	0,924	0,687
V-mesure	0,992	0,976	0,736	1	0,936	0,448	1	0,952	0,777
ARI	0,995	0,983	0,805	1	0,954	0,276	1	0,962	0,862
AMI	0,992	0,975	0,733	1	0,935	0,442	1	0,951	0,776
SC	-0,079	0,485	0,018	0,001	0,440	0,259	0,001	0,432	0,284

TABLE 3.4 – Compound-ECG

	DBSCAN		D-DBSCAN(H1)		MDST DBSCAN		D-MDST-DBSCAN (H1)		DBSCAN (DTW)		D-DBSCAN (H1, DTW)		DBSCRN	
	Comp	ECG	Comp	ECG	Comp	ECG	Comp	ECG	Comp	ECG	Comp	ECG	Comp	ECG
Nb-clusters	3	5	5	7	5	0	6	8	5	5	5	5	183	1039
Noise points %	10.03	52.22	24.56	4.08	22.56	100	19.30	5.32	14.29	2.28	15.79	3.4	0	0
Homogeneity	0.770	0.319	0.838	0.018	0.857	0	0.891	0.026	0.842	0.039	0.828	0.016	0.959	0.802
Completeness	0.951	0.398	0.903	0.094	0.904	1	0.901	0.106	0.927	0.155	0.918	0.090	0.296	0.190
V-measure	0.851	0.354	0.869	0.030	0.880	0	0.896	0.042	0.882	0.063	0.871	0.027	0.452	0.308
ARI	0.826	0.219	0.875	0.001	0.893	0	0.922	0.002	0.881	0.009	0.869	0.001	0.026	0.004
AMI	0.849	0.354	0.866	0.024	0.877	0	0.893	0.035	0.880	0.058	0.869	0.023	0.240	0.131
SC	0.460	0.006	0.060	0.552	0.082	0	0.087	0.060	0.440	0.431	0.425	0.577	-0.02	0.856

clusters.

En conclusion, parmi les modèles évalués, le modèle D-DBSCAN(H1,DTW) se distingue par ses performances supérieures. Il a réussi à identifier 2 clusters avec une séparation raisonnable et un faible pourcentage de points de bruit. Cependant, il est important de noter que les performances des modèles peuvent varier en fonction des spécificités du jeu de données et des métriques d'évaluation utilisées.

TABLE 3.5 – Données du Covid

	DBSCAN classic	D-DBSCAN (H1)	MDST-DBSCAN	D-MDST-DBSCAN (H1)	D-DBSCAN (H1,DTW)
Nb-clusters	3	3	53	5	2
Noise points %	48.73	2.63	40.76	3.08	1.46
Homogeneity	0.015	0.009	0.075	0.008	0.020
Completeness	0.009	0.008	0.015	0.010	0.020
V-measure	0.011	0.009	0.024	0.009	0.020
ARI	0.021	0.011	0.014	-0.018	-0.004
AMI	0.011	0.009	0.018	0.007	0.019
SC	0.043	0.071	0.246	0.032	0.387

TABLE 3.6 – DBSCAN(Minpts=10) avec PCA, LDA et DAE sur les données ECG

	PCA (ϵ=0.5)	PCA (ϵ=0.7)	PCA (ϵ=0.9)	LDA (ϵ=0.02)	LDA (ϵ=0.03)	LDA (ϵ=0.05)	DAE (ϵ=0.01)	DAE (ϵ=0.02)	DAE (ϵ=0.03)
Nb-principal components	9	9	9	4	4	4	2	2	2
Nb-clusters	36	16	4	9	4	17	16	7	5
Noise points %	10.63	3.52	1.09	4.29	2.49	7.77	3.13	1.07	0.58
Homogeneity	0.086	0.030	0.003	0.009	0.004	0.017	0.017	0.006	0.004
Completeness	0.179	0.137	0.056	0.062	0.049	0.074	0.094	0.081	0.085
V-measure	0.116	0.050	0.005	0.015	0.007	0.028	0.029	0.011	0.008
ARI	0.019	0.002	0.000	0.001	0.000	0.002	0.001	0.000	0.000
AMI	0.113	0.048	0.005	0.014	0.007	0.026	0.027	0.010	0.007
SC	-0.359	-0.148	0.168	-0.315	-0.011	-0.300	0.211	0.605	0.623

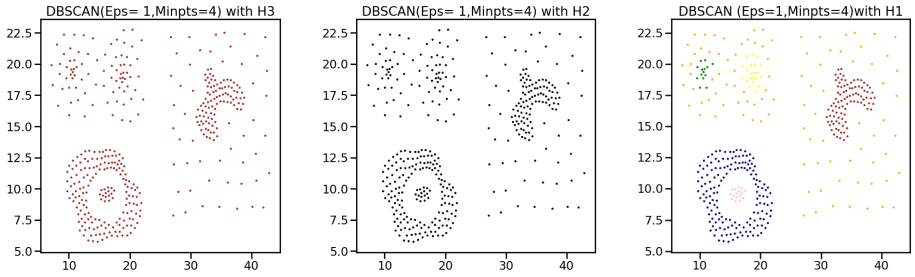


FIGURE 3.11 – DBSCAN avec ϵ adaptatif des hypothèses H_1, H_2 et H_3 sur les données de Compound

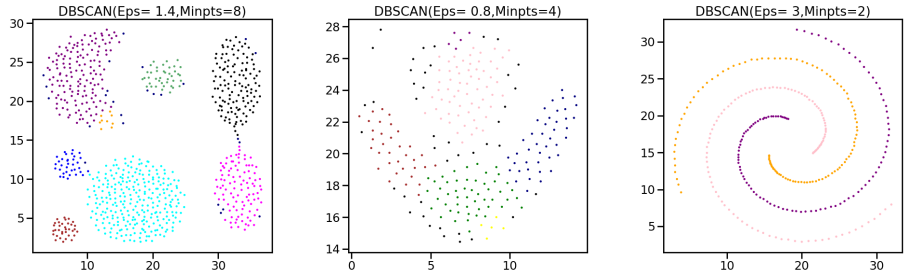


FIGURE 3.12 – DBSCAN avec ϵ adaptatif de l'hypothèse H_1 sur d'autres ensembles de données.

Pour améliorer l'efficacité de nos calculs, nous avons utilisé Joblib-Pyspark et comparé les temps d'exécution avec et sans cette méthode. Comme prévu, l'utilisation de Joblib-Pyspark a considérablement réduit les temps d'exécution pour tous les ensembles de données, en particulier pour les données de séries temporelles (Tableau 3.8).

Il est important de noter que l'application de MDST-DBSCAN à l'ensemble de données ECG nécessite un temps de calcul et des ressources considérables. Par exemple, pour un ensemble de données ECG comprenant 2500 échantillons, l'exécution de MDST-DBSCAN a pris environ 3 heures, tandis que D-MDST-DBSCAN a pu terminer en seulement 500 secondes.

Les meilleurs résultats en termes de temps d'exécution et de qualité de regroupement ont été obtenus en combinant la mesure de similarité DTW avec D-DBSCAN(H1). Cette combinaison a permis de réduire significativement les temps de calcul tout en maintenant la qualité du regroupement.

Enfin, pour améliorer davantage les performances des modèles utilisés dans cette étude, nous avons appliqué une technique de révision de regroupement en utilisant la technique de revision sur les ensembles de données Compound et Covid (référez-vous au Tableau 3.7). Après cette étape de révision, nous avons constaté que les modèles MDST-DBSCAN et D-MDST-DBSCAN présentaient respectivement 0,5% de données bruitées (mal regroupées), tandis que DBSCAN et D-DBSCAN(H1) ne présentaient que 0,2% et 0% de données bruitées. Sur l'ensemble de données Covid, MDST-DBSCAN avait 17% de données bruitées, D-MDST-DBSCAN avait 7,12%, DBSCAN avait 1,96%, tandis que D-DBSCAN(H1) n'avait que 0,16% de données bruitées. Ces résultats confirment que D-DBSCAN(H1) est le meilleur algorithme parmi les quatre modèles en termes de qualité du regroupement.

En conclusion, pour notre application de clustering et d'interprétation des données collectées dans notre champ agricole à l'aide de capteurs IoT, nous recommandons l'utilisation du modèle D-DBSCAN(H1) en raison de ses performances élevées en termes de temps d'exécution et de qualité du regroupement.

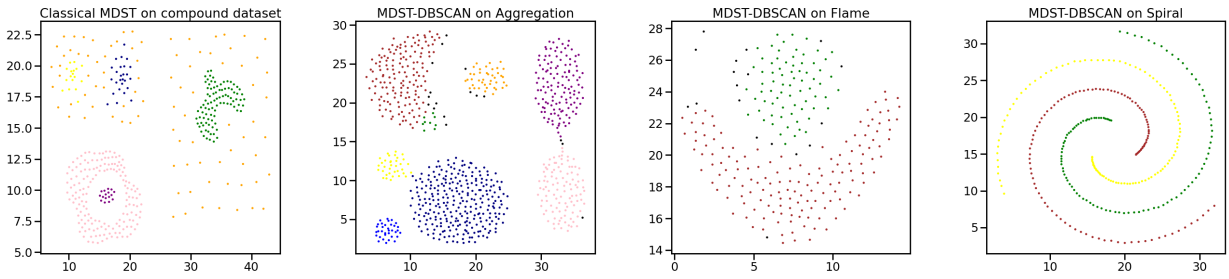


FIGURE 3.13 – MDST-DBSCAN classique

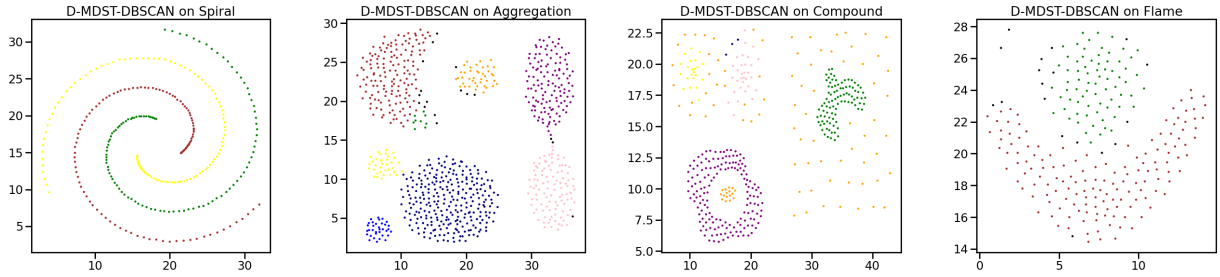


FIGURE 3.14 – D-MDST DBSCAN

TABLE 3.7 – Révision du regroupement (données mal regroupées)

Modéle	Compound	Covid
DBSCAN	0.2%	1.9%
D-DBSCAN	0%	0.16%
MDST-DBSCAN	0.5%	17%
D-MDST-DBSCAN	0.5%	7.12%

Nos modèles, D-DBSCAN et D-MDST-DBSCAN, surpassent les méthodes classiques DBSCAN et MDST-DBSCAN en termes de qualité de regroupement, mais leur temps d'exécution peut encore être optimisé en utilisant les techniques utilisées par Weng et al. (2021) pour réduire le temps de calcul.

En raison de leur capacité à optimiser le regroupement des données en termes de séparabilité entre les clusters, d'homogénéité des clusters et de pourcentage de bruit détecté, D-DBSCAN et D-MDST-DBSCAN peuvent être efficacement utilisés pour étiqueter des données agricoles spatiotemporelles.

TABLE 3.8 – Temps de calcul en secondes sans/avec Joblibspark

Taille	Sans Joblibspark			Avec Joblibspark		
	D-DBSCAN	MDST- DBSCAN	D-MDST-DBSCAN	D-DBSCAN	MDST -DBSCAN	D-MDST -DBSCAN
100	0.219	4.24	2.04	0.18	4.28	1.87
300	1.99	34.2	15.5	1.69	30.8	9.26
400	3.57	56.8	27	3.03	54.8	16.2
700	10.7	177	80	9.34	178	48.9
1000	22.8	368	164	22.9	357	100
2000	104	1464	689	66	1402	396
5000	877	9180	4740	420	8460	4560
10000	5760	>13140	>7200	1800	>13140	>7200
>20000	>7224	>14400	>14400	>7200	>14400	>14400

Chapitre 4

Prédiction de la qualité et de la quantité du rendement du riz

Ce chapitre se penche sur la prédiction cruciale de la qualité et de la quantité du rendement du riz. Nous examinerons l'utilisation de modèles statistiques ainsi que d'apprentissage automatique et profond pour cette prédiction. De plus, nous détaillerons l'architecture du modèle que nous utilisons et son application sur des données réelles.

4.1 Utilisation des modèles statistiques pour la prédiction

Au fil des années, l'utilisation de modèles statistiques basés sur l'auto-régression (AR) et la moyenne mobile (MA) a été une tendance parmi de nombreux chercheurs, comme illustré dans les etudes ci-dessous, par exemple, Reddy and Sureshbabu (2020) présente une étude qui se concentre sur la prédiction du rendement du riz en utilisant des modèles basés sur l'auto-régression (AR) et la moyenne mobile (MA). Dans cette étude, les chercheurs ont développé un modèle basé sur la méthode ARIMA (Auto-Regressive Integrated Moving Average) pour prédire la production saisonnière de riz pour quatre ans. Les données de production de riz ont été collectées dans le district d'Ananthapur entre 2008 et 2014. Le modèle ARIMA a montré des résultats précis. Cependant, il convient de noter que l'utilisation de l'ARIMA peut présenter des limites, notamment la sensibilité aux fluctuations saisonnières

et la nécessité de supposer la stationnarité des données. Ces limites doivent être prises en compte lors de l'application de l'ARIMA pour la prédiction du rendement du riz.

meme pour l'etude de Biswas (2021), se focalise sur la prédiction de la superficie et de la production de riz kharif au Bengale occidental en utilisant un modèle de séries chronologiques appelé ARIMA (p, d, q). Les données agricoles de l'Inde depuis 1962 sont utilisées pour cette étude. mais, il est important de noter que cette étude présente certaines limites et nécessite une expertise approfondie pour sélectionner un modèle ARIMA (p, d, q) spécifique afin de garantir que les valeurs résiduelles estimées à partir de ce modèle soient du bruit blanc

Cependant, avec les améliorations continues des modèles d'apprentissage automatique et d'apprentissage profond, de nouvelles contributions ont émergé dans le domaine de la prédiction de la production. Ces avancées ont permis d'explorer des approches plus sophistiquées et plus puissantes pour la prédiction des rendements, en exploitant des modèles tels que les réseaux de neurones et l'apprentissage profond. Ces modèles ont démontré leur capacité à capturer des relations complexes et non linéaires dans les données, ce qui peut conduire à des prédictions plus précises. Ainsi, ces nouvelles approches ont apporté des contributions significatives dans le domaine de la prédiction de la production agricole, ouvrant de nouvelles perspectives pour améliorer les résultats et la précision des prévisions.

4.2 Utilisation des modèles d'apprentissage automatique et profond pour la prediction

Plusieurs études ont utilisé des réseaux de neurones profonds pour prédire le rendement du riz. Par exemple, Satpathi et al. (2023) aborde le problème de la prévision du rendement des cultures de riz dans le Chhattisgarh, en Inde. Des prévisions de rendement précises sont importantes pour les politiques de sécurité alimentaire, le stockage des produits agricoles et la commercialisation. L'étude compare cinq modèles différents, dont la Stepwise Multiple Linear Regression (SMLR), le réseau neuronal artificiel (ANN), Least Absolute Shrinkage and

Selection Operator (LASSO), Elastic Net (ELNET) et la Ridge Regression , afin de déterminer le meilleur modèle de prévision du rendement du riz. Les modèles ont été calibrés et validés à l'aide de statistiques historiques sur le rendement en riz et de données météorologiques pour trois districts du Chhattisgarh. L'étude a révélé que l'ANN avait obtenu de bons résultats dans deux districts, tandis que l'ELNET et le LASSO avaient obtenu de meilleurs résultats dans le troisième district. Les paramètres utilisés pour évaluer les modèles étaient le coefficient de détermination (R^2) et l'erreur quadratique moyenne (RMSE). Pour le district de Raipur, le modèle le plus performant était l'ANN, avec des valeurs R^2 de 1 pour les ensembles de données d'étalonnage et de validation, et des valeurs RMSE de 0,002 à 0,003. Les autres modèles présentaient des valeurs R^2 inférieures et des valeurs RMSE plus élevées.

Pour le district de Surguja, le modèle le plus performant était également l'ANN, avec des valeurs R^2 de 1 pour l'étalonnage et proches de 1 pour la validation, et de faibles valeurs RMSE (0,004 à 0,214). Les autres modèles présentaient des valeurs R^2 inférieures et des valeurs RMSE plus élevées.

Pour le district de Bastar, les modèles les plus performants étaient ELNET et LASSO, avec des valeurs R^2 de 0,9 à 0,93 pour l'étalonnage et de 0,48 à 0,57 pour la validation, et des valeurs RMSE de 0,05 à 0,1. Les autres modèles présentaient des valeurs R^2 inférieures et des valeurs RMSE plus élevées. L'étude présente des limites, notamment l'hypothèse selon laquelle les facteurs biophysiques et les pratiques agricoles sont constants d'un district à l'autre et l'utilisation de données de rendement au niveau des districts, qui peuvent ne pas refléter avec précision la variabilité des rendements au niveau des champs.

Hu et al. (2023) examine le défi de la prévision des rendements des cultures dans des conditions climatiques changeantes à l'aide de modèles statistiques. Les choix de modèles sont importants en termes d'interprétabilité et de pouvoir prédictif. Le modèle d'ensemble bayésien (BM) proposé dans l'article est interprétable et présente un grand pouvoir explicatif et prédictif. Il décompose les données de rendement historiques pour estimer conjointement les tendances technologiques et les effets climatiques sur le rendement des cultures. L'étude utilise des données sur le rendement du maïs en Ohio, aux États-Unis, sur une période de 80

ans, de 1930 à 2010. Elle compare les performances de BM à celles d'autres modèles d'apprentissage automatique tels que ElasticNet, Neural Network, MARS, SVM, Random Forests et XGBoost. Les résultats montrent que BM excelle à la fois dans la prédiction et l'explication des rendements des cultures. Il détecte également deux changements technologiques dans les données sur les rendements du maïs en Ohio et met en évidence des réponses asymétriques non linéaires des cultures au climat, ainsi que des effets d'interaction non négligeables entre la température et les précipitations. Les métriques d'évaluation, telles que l'erreur absolue moyenne (MAE) et le coefficient de détermination (R-carré), confirment que BM surpasse les autres modèles en termes de valeurs MAE et R-carré. Cependant, L'etude a plusieurs limites lors de l'utilisation du BM notamment leur manque d'interprétabilité et la possibilité de produire des prévisions peu fiables lors de l'extrapolation à de nouvelles conditions climatiques. Les auteurs soutiennent que leur modèle d'ensemble bayésien (BM) constitue une alternative plus interprétable et plus fiable, mais ils reconnaissent également que le BM a ses propres limites, telles que la complexité informatique liée à l'ajustement et à l'évaluation de plusieurs modèles et la nécessité de connaissances spécifiques au domaine pour éclairer la structure du modèle.

Nous pouvons également trouver des prédictions pour d'autres céréales, dont les modèles peuvent être utilisés pour la prédiction de la production de riz. comme Islam et al. (2023), il a utilisé plusieurs modèles d'apprentissage automatique pour améliorer la précision de l'estimation du rendement des cultures au niveau du district népalais. Dans un premier temps, l'étude a utilisé la régression linéaire comme modèle de référence et a entraîné plusieurs modèles d'apprentissage automatique pour voir comment ils amélioraient la précision par rapport aux modèles de référence. L'étude a également appliqué l'un des modèles de régression arborescents populaires, XGBoost, et a comparé le résultat avec le modèle de référence. Enfin, l'étude a proposé un modèle d'ensemble de piles personnalisé en combinant quatre régresseurs arborescents différents : XGBoost, LightGBM, Gradient Boost et un modèle de forêt aléatoire. L'étude a révélé que le produit NDVI dérivé de la télédétection ne suffisait pas à lui seul pour une estimation précise du rendement des cultures. Après avoir intégré d'autres

variables météorologiques dans les modèles d'apprentissage automatique, la précision des estimations s'est considérablement améliorée. Tout comme le NDVI, les variables météorologiques relatives aux précipitations, à l'humidité du sol et à l'évapotranspiration présentaient également une forte association avec le rendement en riz. Le modèle d'ensemble empilé proposé a enregistré globalement moins d'erreurs que la régression linéaire de référence et les autres modèles d'apprentissage automatique. Le flux de travail méthodologique a atteint une précision moyenne de 92 % de l'estimation du rendement avec un RMSE 328,06 kg/ha et un MAE de 317,21 kg/ha. En raison du contexte physiogéographique unique et distinct de chaque district, une variation de la précision des estimations a pu être observée d'un district à l'autre.

L'étude présente quelques limites, notamment la disponibilité des données et l'absence d'informations factuelles suffisantes sur le terrain. La présence d'une couverture nuageuse importante au-dessus des districts de la ceinture du Teraï au Népal pendant la principale saison de croissance et de récolte a rendu difficile l'obtention de produits NDVI à haute résolution à partir des images Landsat/Sentinel 2. L'ensemble de données ne contient que 19 cas (de 2001 à 2019) pour chaque district, ce qui peut entraîner un surajustement du modèle. De plus, l'étude n'a pas pris en compte d'autres paramètres environnementaux susceptibles d'affecter le rendement des cultures, tels que les ravageurs et les maladies. Enfin, l'étude a été menée uniquement au Népal, et les résultats peuvent ne pas être généralisables à d'autres régions présentant des contextes physiogéographiques différents.

Jiya et al. (2023) vise à développer des modèles de prévision du rendement du riz pour l'État de Katsina au Nigéria, en utilisant des données climatiques et des données sur le rendement du riz, afin de faciliter la formulation de politiques, la planification et l'intervention efficaces pour la sécurité alimentaire dans l'État. L'étude a recueilli des données sur le rendement du riz dans l'État de Katsina, couvrant les années 1970 à 2017, provenant du Bureau des statistiques du Nigéria. En outre, les données climatiques pour la même période ont été obtenues à partir du portail de connaissances climatiques de la Banque mondiale. La régression logistique, le réseau neuronal artificiel, la forêt aléatoire, les arbres aléatoires et les algorithmes naïfs de Bayes ont été utilisés pour développer des modèles de prédiction du

rendement du riz à l'aide de cet ensemble de données. Les résultats révèlent que les forêts aléatoires et les arbres aléatoires présentaient de meilleures performances de classification pour la prédiction du rendement. L'évaluation de chaque modèle comprenait des mesures telles que le vrai positif, le faux positif, la précision, le rappel, la mesure F et la zone ROC. Parmi les modèles, Random Forest et Random Trees ont démontré des performances supérieures avec un taux TP de 1, suivis par Naïve Bayes avec un taux de 0,19, tandis que le réseau neuronal et la régression logistique ont tous deux atteint un taux de 0,75. Les modèles développés constituent un outil prometteur pour prévoir les rendements futurs en riz, facilitant ainsi la prise de mesures proactives visant à garantir la sécurité alimentaire de la population de l'État. Cependant,les forêts aléatoires et les arbres aléatoires présentent certaines limites à prendre en compte. Ils peuvent être sensibles au surajustement, ce qui peut entraîner une mauvaise généralisation des prédictions. De plus, ils peuvent être sensibles aux valeurs aberrantes, ce qui peut affecter la fiabilité des résultats. Leur complexité rend également l'interprétation des résultats plus difficile. De plus, ils dépendent de la configuration des paramètres, ce qui nécessite une sélection judicieuse.

Un autre aspect de la mesure de la qualité du riz consiste à détecter les maladies à travers l'analyse des feuilles de riz. Dans l'etude Islam et al. (2018), les auteurs présentent une méthode utilisant l'algorithme de Naive Bayes pour la classification des maladies foliaires du riz. Ils ont mis en place un processus en cinq étapes, comprenant l'acquisition d'images de feuilles, le prétraitement, l'extraction des proportions des pixels RVB, la détection des maladies et la classification. Les résultats obtenus ont démontré une précision satisfaisante de 90% pour la détection des maladies spécifiques. Cependant, il est important de noter certaines limites potentielles de cette méthode. Les performances peuvent être influencées par des facteurs tels que la lumière, le contraste et l'angle de capture des images. Il est donc essentiel de prendre en compte ces limitations lors de l'utilisation de modèles basés sur Naive Bayes pour la classification des maladies foliaires du riz. Des études supplémentaires pourraient être nécessaires pour évaluer la robustesse et l'adaptabilité de cette méthode dans différentes conditions environnementales.

L'etude Prajapati et al. (2017) présente un prototype de système de détection et de classification des maladies du riz. Les chercheurs ont utilisé un modèle de machine à vecteur de support (SVM) en se basant sur des images de feuilles de riz infectées. Les étapes du processus comprenaient la préparation d'une collection d'images de feuilles de riz, la suppression de l'arrière-plan de l'image, l'élimination du bruit, le redimensionnement et la segmentation de l'image. Ils ont ensuite extrait les parties de l'image présentant des signes de maladies en utilisant la méthode de regroupement K-means. Pour différencier chaque maladie, 88 caractéristiques liées à la couleur, à la texture et à la forme ont été extraites. En utilisant ces caractéristiques, un modèle SVM a été construit pour la classification des maladies. Les résultats ont montré une précision d'apprentissage de 93,33% et une précision de test de 73,33%. Bien que ces résultats indiquent une bonne performance pour certaines maladies, nous ne pouvons pas assurer la robustesse du système dans des conditions variables et explorer son application à d'autres maladies foliaires du riz.

L'étude menée par Shrivastava and Pradhan (2021) présente une approche similaire utilisant à la fois un modèle SVM et un modèle CNN pour la classification des maladies foliaires du riz. Les chercheurs ont collecté un ensemble de données de 619 images de maladies foliaires provenant d'une rizière en Inde. L'ensemble de données a été divisé de manière aléatoire en ensembles de formation et de test, avec différentes répartitions. Les résultats ont montré que la meilleure précision de classification atteignait 91,37% avec une répartition de 80% pour l'ensemble de formation et 20% pour l'ensemble de test. Ces résultats sont encourageants et démontrent l'efficacité des modèles SVM et CNN dans la classification des maladies foliaires du riz. Cependant, il convient de noter que cette étude se limite à un ensemble spécifique de maladies et qu'il est nécessaire de mener davantage de recherches pour évaluer la performance de ces modèles sur un plus large éventail de maladies foliaires du riz.

L'étude présentée par Joshi and Jadhav (2016) propose l'utilisation d'un modèle KNN pour la classification des maladies des feuilles de riz. Les chercheurs ont recueilli 115 images de feuilles de riz provenant de quatre maladies différentes, collectées à la station de recherche

agricole de Lonavala dans le Maharashtra, ainsi que sur Internet. Les images ont été divisées en ensembles de formation et de test, avec 70% des images utilisées pour la formation du modèle et 30% pour les tests. Les résultats ont montré une précision globale de 87,02% obtenue grâce à l'utilisation du modèle KNN. Ces résultats suggèrent l'efficacité du modèle KNN dans la classification des maladies des feuilles de riz. Toutefois, il est important de noter que cette étude se limite à un ensemble spécifique de maladies et qu'il serait intéressant de poursuivre les recherches pour évaluer la performance de ce modèle sur un plus large éventail de maladies foliaires du riz.

L'étude Ahmed et al. (2019) a réalisé une comparaison entre plusieurs classificateurs, dont les arbres de décision, la régression logistique, KNN et Naive Bayes, pour la détection et la classification de trois maladies des feuilles de riz. Pour cela, ils ont utilisé un ensemble de 480 images de feuilles atteintes de maladies provenant du référentiel d'apprentissage automatique de l'UCI. Des filtres ont été appliqués pour minimiser le temps de traitement et extraire les objets pertinents des images. L'ensemble de données a été divisé en 90 % pour l'entraînement et 10 % pour les tests, en utilisant une validation croisée à 10 reprises pour évaluer l'efficacité des algorithmes.

Les chercheurs ont évalué les performances des classificateurs en utilisant plusieurs indicateurs tels que le taux de vrais positifs (TPR), le taux de faux positifs (FPR), la précision (P), le rappel (R), la mesure F et l'aire sous la courbe ROC (AUC). Les résultats ont montré que, dans chaque cas, l'arbre de décision a obtenu les meilleures performances, avec une précision de 94,9 % pour la formation et de 97,9% pour les tests.

Ces résultats démontrent que l'arbre de décision est un classificateur prometteur pour la détection et la classification des maladies des feuilles de riz, surpassant les autres algorithmes évalués dans cette étude. Pourtant, les performances des modèles peuvent varier en fonction de la qualité et de la diversité des données utilisées, et des recherches supplémentaires pourraient être nécessaires pour évaluer la généralisation de l'arbre de décision sur d'autres ensembles de données et maladies foliaires du riz.

Guo et al. (2021) propose une étude comparative visant à prédire le rendement du riz

en utilisant différentes méthodes, notamment Random Forest (RF), régression linéaire multiple (MLR), réseau neuronal de rétropropagation (BP) et SVM. Les chercheurs ont utilisé un ensemble de données comprenant des informations climatiques, phénologiques et géographiques provenant de 75 stations agrométéorologiques dans le sud de la Chine. Les résultats ont montré que les méthodes d'apprentissage automatique, en particulier SVM et RF, surpassaient la méthode MLR en termes de performances de prédiction. Des corrélations significatives ont été observées entre les variables phénologiques, géographiques et climatiques et les rendements en riz, ce qui souligne leur importance pour la prédiction précise des rendements. La combinaison de la phénologie intégrée, des informations climatiques et géographiques a été identifiée comme la plus efficace pour prédire le rendement du riz. Les mesures de performance telles que le coefficient de détermination R^2 et l'erreur quadratique moyenne RMSE ont été utilisées pour évaluer les méthodes, et le SVM a montré la plus grande précision de prédiction avec un R^2 de 0,33 et un RMSE de 737 kg/ha. Il convient cependant de noter que cette étude présente certaines limites, telles que la dépendance des résultats aux variables sélectionnées et aux paramètres du modèle. De plus, la généralisation des résultats à d'autres régions géographiques et conditions climatiques peut être sujette à caution. Malgré ces limites, cette recherche met en évidence le potentiel des modèles d'apprentissage automatique pour la prédiction du rendement du riz en utilisant des données climatiques, phénologiques et géographiques.

Une étude similaire Mansaray et al. (2020) a exploré l'estimation de la biomasse sèche du riz en utilisant différentes méthodes de régression, notamment le Gradient Boosting Decision Tree (GBDT), le Support Vector Machine (SVM) et le k-Nearest Neighbors (KNN) [16]. Les chercheurs ont utilisé des données provenant du mode Sentinel-1A Interferometric Wide-swath (IW), comprenant trois ensembles de données représentant différentes polarisations. L'expérimentation a été menée dans le comté de Tongxiang, en Chine, sur une zone d'environ 715 km^2. Les résultats ont montré que la polarisation verticale transmise et verticale reçue (VV) était plus précise que la polarisation verticale transmise et horizontale reçue (VH) pour estimer la biomasse sèche du riz en utilisant les algorithmes sélectionnés. De plus, l'utilisa-

tion de la combinaison linéaire de VH et VV (VHVV) a conduit à des estimations de biomasse plus précises que l'utilisation de chaque polarisation indépendamment. En termes de performance des algorithmes de régression, l'utilisation de VHVV avec Random Forest (RF) a montré la plus grande précision (R^2 de 0,73 et RMSE de 462,4 g/m^2), suivie de VHVV avec k-NN (R^2 de 0,70 et RMSE de 484,1 g/m^2) sur l'ensemble de la période de végétation. Pour la phase de reproduction du riz, le k-NN a atteint la plus grande précision en utilisant l'ensemble de données VH, avec une corrélation de 0,72 et une RMSE de 362,4 g/m^2. Ces résultats soulignent l'importance de la sélection des polarisations appropriées et de l'utilisation de méthodes de régression adaptées pour une estimation précise de la biomasse sèche du riz. Bien que cette étude ait fourni des résultats encourageants pour l'estimation de la biomasse sèche du riz, il est important de prendre en compte certaines limites des modèles utilisés. Tout d'abord, les performances des modèles de régression peuvent varier en fonction des conditions spécifiques du site d'étude et des caractéristiques des données. Par conséquent, les résultats obtenus dans cette étude peuvent ne pas être directement généralisables à d'autres régions ou périodes de croissance du riz. De plus, l'utilisation de données radar Sentinel-1A IW peut être sensible à certains facteurs environnementaux, tels que la couverture nuageuse ou la présence d'obstacles physiques, ce qui peut affecter la précision des estimations. En outre, bien que les modèles de régression tels que le GBDT, le SVM et le KNN soient couramment utilisés, ils ont leurs propres hypothèses et limitations inhérentes. Par exemple, le KNN peut être sensible aux valeurs aberrantes et à la distribution des données, tandis que le SVM peut être sensible au choix des paramètres et à la nature linéaire ou non linéaire de la relation entre les variables. Par conséquent, une approche intégrée combinant différentes méthodes de régression et l'utilisation de données complémentaires pourrait potentiellement améliorer davantage la précision des estimations de la biomasse sèche du riz.

Yu et al. (2021), les chercheurs ont exploré différentes méthodes pour estimer le nombre d'épillets à partir d'images RVB. Ils ont comparé cinq approches différentes : la méthode de la surface, P-TRAP, Faster R-CNN, Cascade R-CNN et SSD. Les résultats ont mis en évidence la supériorité des méthodes d'apprentissage profond, notamment Faster R-CNN et Cascade

R-CNN, par rapport aux autres méthodes telles que la méthode de la surface, P-TRAP et SSD. Les méthodes R-CNN en cascade ont atteint un R^2 de 0,99 avec une erreur moyenne absolue en pourcentage MAPE inférieure à 1,42 %. De son côté, la méthode R-CNN plus rapide a obtenu un R^2 de 0,99 avec une MAPE inférieure à 1,68 %. En termes de temps de traitement, la méthode R-CNN plus rapide s'est révélée plus rapide que la méthode R-CNN en cascade.

De plus, l'étude a également comparé Resnet-50, Faster R-CNN, Cascade R-CNN, SSD et SVM pour la classification des riz indica et japonica à partir d'images radiographiques d'épillets. Les résultats ont montré que Resnet-50 avait la précision la plus élevée, atteignant 0,96. Cependant, il est important de noter que le SSD était le moins précis parmi les méthodes comparées.

Cependant, il convient de souligner certaines limitations des modèles utilisés dans cette étude. Par exemple, l'estimation du nombre d'épillets à partir d'images RVB peut être sensible à des facteurs tels que la qualité des images, l'éclairage et les variations de couleur, ce qui peut entraîner des erreurs dans les prédictions. De plus, les performances des méthodes d'apprentissage profond peuvent dépendre de la disponibilité d'un ensemble de données d'entraînement représentatif et de grande taille. En outre, le choix du modèle approprié doit tenir compte des caractéristiques spécifiques des données et des objectifs de la tâche.

En resumé, bien que les méthodes d'apprentissage profond, en particulier Faster R-CNN, Cascade R-CNN et Resnet-50, aient démontré de bonnes performances dans cette étude pour l'estimation du nombre d'épillets et la classification des riz indica et japonica, il est important de prendre en compte les limitations liées à la qualité des images, à la disponibilité des données d'entraînement et au choix du modèle approprié pour obtenir des résultats fiables.

Alors, les modèles de deep learning, tels que les réseaux de neurones profonds, ont démontré leur efficacité dans la prédiction des rendements et de la qualité des cultures. Ces modèles peuvent être entraînés sur de grandes quantités de données historiques, comprenant des informations sur les conditions de croissance, les caractéristiques du sol, les pratiques agricoles et les données climatiques. En apprenant à partir de ces données, les réseaux de neu-

rones profonds peuvent capturer des modèles complexes et non linéaires qui influencent le rendement et la qualité du riz.

Notre contribution consiste à :

- Proposer une nouvelle approche pour la prédiction du rendement du riz en utilisant une combinaison novatrice de données climatiques, de précipitations et de données satellitaires.
- Développer une nouvelle architecture basée sur les mémoires à long terme LSTM et les unités récurrentes Gated GRU pour améliorer la précision des prédictions. Ces modèles sont capables de capturer les dépendances temporelles complexes et de mieux exploiter les informations des données climatiques et satellitaires.
- Comparer ces modèles avec les modèles de référence simples et efficaces tels que KNN, ARIMA, VAR et VARMAX pour évaluer l'efficacité de nos propositions et déterminer si notre nouvelle approche permet d'obtenir des performances supérieures en termes de précision et d'exactitude dans la prédiction du rendement du riz.

4.3 Architecture du modèle utilisé

Nous avons choisi d'utiliser les modèles LSTM (Long Short-Term Memory) et GRU (Gated Recurrent Unit) pour nos prédictions, en les comparant aux modèles basés sur ARIMA. Les architectures de ces modèles sont les suivantes :

Pour le modèle GRU, nous avons utilisé une architecture avec plusieurs couches GRU. Le premier GRU a une couche de retour séquentielle avec une fonction d'activation tangente hyperbolique (tanh). Ensuite, nous avons ajouté deux autres couches GRU avec des configurations similaires. Enfin, nous avons ajouté une dernière couche GRU suivie d'une couche dense de sortie. Cette architecture nous permet de capturer les motifs temporels complexes dans les données (Figure 4.2).

Pour le modèle LSTM, nous avons également utilisé une architecture avec plusieurs couches LSTM. Chaque couche LSTM a une couche de retour séquentielle et utilise également la fonc-

tion d'activation tanh. Nous avons ajouté trois couches LSTM supplémentaires pour capturer les informations à long terme. Enfin, nous avons ajouté une couche dense de sortie. Ce modèle est capable de mémoriser les informations à court et long terme, ce qui est bénéfique pour la prédiction des séries temporelles (Figure 4.1).

Nous avons choisi ces architectures en fonction de leur capacité à capturer les motifs et les dépendances temporelles dans les données. Les LSTM et GRU sont connus pour leur capacité à traiter les séries temporelles et à prendre en compte les dépendances à long terme. De plus, ces modèles sont adaptatifs et peuvent apprendre à partir des données sans avoir besoin d'hypothèses spécifiques sur la structure du modèle .

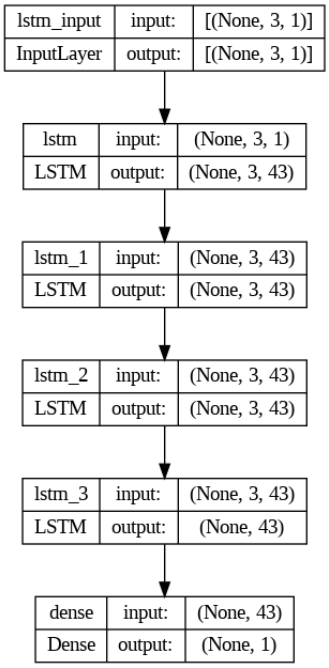


FIGURE 4.1 – Architecture du modèle LSTM

En plus des modèles LSTM et GRU, nous avons également utilisé des modèles ARIMA, VAR et VARMAX pour nos prédictions. Les modèles ARIMA (AutoRegressive Integrated Moving Average) sont des modèles populaires pour la prévision des séries temporelles. Ils combinent les composantes autorégressive (AR) et moyenne mobile (MA) avec une intégration (I) des données pour capturer les tendances et les motifs dans les séries temporelles.

$$(1 - \phi_1 B)(1 - B)^d X_t = c + (1 + \theta_1 B) Z_t \tag{4.1}$$

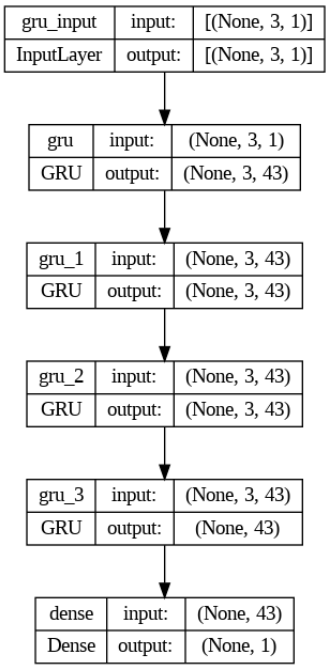


FIGURE 4.2 – Architecture du modèle GRU

Les modèles VAR (Vector Autoregression) sont une extension des modèles ARIMA pour les séries temporelles multivariées. Ils permettent de modéliser les relations entre plusieurs variables en utilisant des termes autorégressifs pour chaque variable. Les modèles VAR peuvent capturer les interactions entre les différentes variables et fournir des prédictions simultanées pour l'ensemble du système.

$$X_t = c + \phi_1 X_{t-1} + \ldots + \phi_p X_{t-p} + Z_t \tag{4.2}$$

Les modèles VARMAX (Vector Autoregressive Moving Average with Exogenous Variables) sont une extension des modèles VAR qui prennent également en compte les variables exogènes. Ces modèles sont utiles lorsque les variables exogènes ont une influence sur les variables endogènes et doivent être incluses dans le modèle de prédiction.

$$X_t = c + \Phi_1 X_{t-1} + \ldots + \Phi_p X_{t-p} + \Theta_1 Z_{t-1} + \ldots + \Theta_q Z_{t-q} + Z_t \tag{4.3}$$

Dans les équations des modèles VAR et VARMA, différentes variables sont utilisées pour

modéliser les séries temporelles multivariées. Voici une description de chaque variable :

X_t : Il s'agit d'un vecteur de variables temporelles. Ces variables peuvent représenter différentes mesures ou caractéristiques observées au fil du temps, telles que la température, l'humidité, la vitesse du vent, etc.

B : C'est l'opérateur de retard qui permet de décaler les valeurs des variables temporelles dans le temps. Il est utilisé pour prendre en compte les dépendances temporelles entre les observations passées et présentes.

$\phi_1, ..., \phi_p$: Ce sont les paramètres AR (AutoRégressifs). Ils représentent les coefficients qui pondèrent les valeurs précédentes des variables temporelles pour prédire leur valeur actuelle. Chaque ϕ_i correspond à un retard spécifique.

θ_1 : C'est le paramètre MA (Moyenne Mobile). Il représente le coefficient qui pondère le bruit blanc dans le modèle. Le bruit blanc est une séquence de variables aléatoires indépendantes et identiquement distribuées.

Z_t : Il s'agit du bruit blanc, qui représente le terme d'erreur dans le modèle. Il est généralement supposé avoir une moyenne nulle et une variance constante.

d : C'est l'ordre de différenciation, qui indique le nombre de fois que les valeurs des variables doivent être différenciées pour rendre la série temporelle stationnaire.

c : C'est une constante, également appelée terme constant ou intercept, qui représente le niveau moyen des variables temporelles.

$\Phi_1, ..., \Phi_p$: Ce sont les paramètres VAR (Vector Autoregression), qui représentent les coefficients qui pondèrent les valeurs précédentes des variables temporelles dans le modèle VAR.

$\Theta_1, ..., \Theta_q$: Ce sont les paramètres VARMA (Vector Autoregressive Moving Average), qui représentent les coefficients qui pondèrent les valeurs précédentes des variables temporelles et des termes d'erreur dans le modèle VARMA.

p : C'est l'ordre du modèle VAR, qui indique le nombre de retards utilisés pour prédire les valeurs des variables temporelles.

q : C'est l'ordre du modèle VARMA, qui indique le nombre de termes d'erreur utilisés pour

prédire les valeurs des variables temporelles.

Ces modèles basés sur ARIMA, VAR et VARMAX sont couramment utilisés pour la prévision des séries temporelles, en particulier lorsque les données présentent des tendances, des saisons ou des dépendances entre les variables. Ils sont plus simples que les modèles de deep learning tels que LSTM et GRU.

4.4 Application des modèles

4.4.1 Application des modéles sur des données de ROSSO

Dans notre étude, nous avons commencé par appliquer différents modèles de prédiction sur les données de l'ONM de Rosso (voir la section 2.1), qui comprennent des variables telles que la température, l'humidité, les précipitations, la vitesse du vent, le rendement et la surface cultivée. Cette approche nous a permis d'évaluer les performances de chaque modèle de manière générale et de déterminer leur adéquation pour la prédiction des données climatiques et agricoles.

Avant d'appliquer les modèles de prédiction, nous avons effectué une étape importante pour vérifier la stationnarité des données. La stationnarité est une propriété essentielle pour les modèles ARIMA, car ces modèles supposent que les données sont stationnaires, c'est-à-dire qu'elles ne présentent pas de tendance ou de motifs saisonniers. Pour vérifier la stationnarité, nous avons utilisé le test Augmented Dickey-Fuller (ADF) statistique (2.2).

Le résultat de l'ADF statistique est présenté dans le tableau précédent. Les valeurs de l'ADF statistique pour chaque variable, telles que l'humidité, la température, les précipitations, la vitesse du vent, le rendement et les zones cultivées, ont été calculées. Les valeurs de l'ADF statistique plus négatives indiquent une stationnarité plus forte des données. Dans notre cas, toutes les variables ont des valeurs d'ADF statistique inférieures aux valeurs critiques à 1%, 5% et 10% de signification, ce qui suggère que les données sont stationnaires.

Une fois que nous avons confirmé la stationnarité des données, nous avons procédé à la

sélection des ordres P, I, Q pour les modèles ARIMA. Les ordres P, I, Q déterminent le nombre de termes autorégressifs, de termes de différenciation et de termes de moyenne mobile dans le modèle ARIMA, respectivement. Pour sélectionner ces ordres, nous avons utilisé une approche basée sur l'analyse des autocorrélations partielles (PACF) et des autocorrélations (ACF) des données. Nous avons également utilisé des critères d'information tels que l'Akaike Information Criterion (AIC) et le Bayesian Information Criterion (BIC) pour choisir les ordres optimaux.

En ce qui concerne les modèles VAR et VARMAX, nous avons effectué une analyse similaire pour sélectionner les ordres optimaux. Nous avons examiné les autocorrélations et les autocorrélations partielles des résidus des modèles VAR et VARMAX pour choisir les ordres appropriés.

Dans l'ensemble, la vérification de la stationnarité des données et la sélection des ordres P, I, Q ont été des étapes cruciales dans notre étude. Cela nous a permis d'appliquer les modèles ARIMA, VAR et VARMAX de manière appropriée en garantissant la stationnarité des données. Cependant, pour les modèles KNN, LSTM et GRU, ces étapes ne sont pas nécessaires car ces modèles ne reposent pas sur l'hypothèse de stationnarité des données.

Les modèles ARIMA, VAR et VARMAX ont été évalués en fonction de leurs performances en termes d'erreurs de prédiction (MFE, MAE, MSE, RMSE 2.2). Les modèles ARIMA (1,0,q) avec q compris entre 1 et 15, ARIMA (1,0,q) avec q compris entre 16 et 25, ainsi que ARIMA (p,0,2) avec p compris entre 1 et 10, ont montré des valeurs similaires pour ces mesures. Cela suggère que ces modèles n'ont pas une grande variabilité de performance en fonction des paramètres choisis. Les modèles VAR avec différents ordres p (2, 10, 20, 120, 180) ont également affiché des performances similaires. On observe cependant une légère diminution des erreurs de prédiction à mesure que l'ordre p augmente. Le modèle VARMAX avec p=1 et q=1 a montré des valeurs similaires aux autres modèles VAR. Ces résultats indiquent que les modèles ARIMA, VAR et VARMAX ont des performances similaires mais relativement moins précises dans la prédiction des états similaires (Tableau 4.1).

En ce qui concerne les modèles KNN, le modèle KNN avec K=3 a présenté une erreur

Modèle	MFE	MAE	MSE	RMSE
ARIMA (1,0,q) (q=1,...,15)	-0.07	0.37	0.15	0.39
ARIMA (1,0,q) (q=16,...,25)	-0.08	0.38	0.16	0.40
ARIMA (p,0,2) (p=1,...,10)	-0.08	0.38	0.16	0.40
VAR (p=2)	-0.10	0.40	0.17	0.42
VAR (p=10)	-0.09	0.399	0.171	0.413
VAR (p=20)	-0.08	0.39	0.16	0.40
VAR (p=120)	-0.05	0.42	0.27	0.52
VAR (p=180)	-0.04	0.47	0.33	0.57
VARMAX (p=1,q=1)	-0.09	0.386	0.159	0.39
KNN (K=3)	0.031	0.070	0.014	0.121
KNN (K=10)	0.033	0.075	0.015	0.123
KNN (K=356)	0.016	0.115	0.020	0.144
LSTM	0.008	0.061	0.010	0.100
GRU	0.028	0.081	0.015	0.126

TABLE 4.1 – Performances des modèles de prédiction

de prédiction légèrement inférieure par rapport aux modèles KNN avec K=10 et K=356. Cela suggère que considérer moins de voisins donne de meilleurs résultats pour ce jeu de données. Cependant, les modèles KNN ont globalement affiché des erreurs de prédiction plus élevées par rapport aux modèles ARIMA, VAR et VARMAX. Ainsi, ils sont moins performants dans la prédiction des états similaires (Tableau 4.1).

En ce qui concerne les modèles LSTM et GRU, le modèle LSTM a affiché les meilleures performances parmi tous les modèles examinés. Il a montré les valeurs les plus faibles pour les mesures d'erreurs de prédiction (MFE, MAE, MSE, RMSE). Cela suggère que le modèle LSTM est capable de capturer efficacement les tendances et les motifs dans les données, ce qui conduit à des prédictions plus précises. Le modèle GRU, bien qu'il ait affiché des performances légèrement inférieures à celles du modèle LSTM, reste meilleur que les modèles ARIMA, VAR, VARMAX et KNN en termes d'erreurs de prédiction (Tableau 4.1).

En résumé, les modèles ARIMA, VAR et VARMAX ont des performances similaires mais relativement moins précises, tandis que les modèles KNN ont des performances inférieures.

En revanche, les modèles LSTM et GRU se distinguent avec des performances supérieures. Parmi eux, le modèle LSTM présente la meilleure précision dans la prédiction des états similaires. Par conséquent, il est recommandé d'utiliser le modèle LSTM pour les prédictions futures des états similaires, car il offre la meilleure précision dans ce contexte.

4.4.2 Mesure de la fiabilité des données IoT de ROSSO via données satellitaires

Lorsque nous travaillons avec des données provenant de l'Internet des objets (IoT), il est essentiel de prendre en compte la fiabilité de ces données. Les capteurs IoT peuvent parfois générer des mesures erronées ou bruitées en raison de divers facteurs tels que des erreurs de capteur, des interférences, des pannes matérielles, etc. Par conséquent, il est important de pouvoir évaluer la qualité et la fiabilité des données IoT pour garantir des prédictions précises et fiables. Dans ce contexte, l'utilisation de données satellitaires peut jouer un rôle crucial. Les données satellitaires fournissent une perspective globale et objective de l'environnement, ce qui peut aider à évaluer la validité des données de l'IoT. Par exemple, en utilisant des indices de végétation satellitaires tels que le NDWI et le NDMI, nous pouvons estimer l'utilisation de l'irrigation et comparer ces résultats avec les mesures fournies par les capteurs IoT. Si les mesures de l'IoT sont cohérentes avec les estimations satellitaires, cela peut renforcer la fiabilité de ces données.

Alors, notre but est de prouver si les données collectées par l'ONM reflètent fidèlement la réalité des champs agricoles à ROSSO de manière globale.

En calculant les coefficients de corrélation, nous sommes en mesure d'évaluer la relation linéaire entre les facteurs climatiques et les indices de végétation. Alors, Nous avons entamé l'analyse de la fiabilité des données IoT en calculant la corrélation entre différents facteurs climatiques et les indices de végétation. Nous avons remarqué que la plupart d'entre eux présentent des corrélations positives, variant de 0,1 à 0,58 (Figure 4.4).

Par exemple, nous avons observé une corrélation d'environ 0,1 entre la température et les

indices NDVI, NDWI et la salinité. De même, une corrélation d'ordre 0,1 a été constatée entre l'humidité de l'air et les indices NDVI, salinité et NDMI. En revanche, une corrélation d'ordre 0,5 a été observée entre l'indice NDVI et les indices NDMI et salinité.

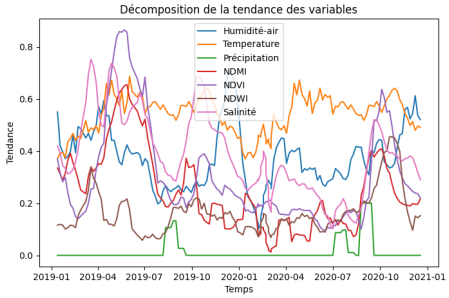
Ces relations suggèrent l'existence d'une relation entre l'humidité, la température et les indices satellitaires. Cependant, il est important de noter que ces corrélations sont relativement faibles. Cela est dû au fait que les indices de végétation utilisés sont calculés à partir d'un champ de ROSSO de manière aléatoire, tandis que les valeurs de l'ONM représentent la moyenne de toutes les régions de ROSSO. Par conséquent, ils ne reflètent pas précisément la réalité d'un champ spécifique, ce qui met en évidence l'utilité de notre idée de mettre en place un système de surveillance pour chaque champ agricole.

Certainement ! Après avoir calculé les coefficients de corrélation entre les facteurs climatiques et les indices de végétation, nous allons procéder à une analyse plus approfondie en mesurant la tendance, la saisonnalité et le bruit dans les variables collectées.

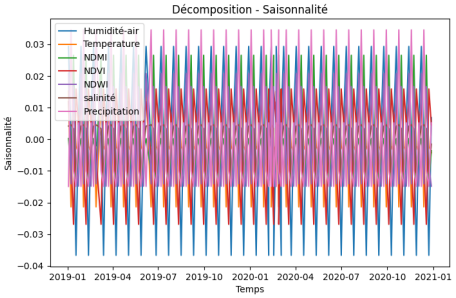
La mesure de la tendance nous a permis de déterminer s'il existe une évolution générale et progressive des variables au fil du temps. Cela nous aide à identifier les changements à long terme dans les conditions climatiques et à évaluer leur impact sur la croissance des cultures. Par exemple, en observant la tendance, nous avons noté une augmentation des valeurs de NDWI après les précipitations, ce qui reflète l'existence de plans d'eau après la pluie. De plus, la température reste élevée, ce qui représente la réalité de ces régions. Nous avons également observé que l'humidité de l'air et l'humidité des plantes varient simultanément entre octobre 2019 et octobre 2020, ce qui peut être expliqué par une augmentation des pratiques d'irrigation pendant cette période, contribuant ainsi à maintenir une humidité adéquate dans les champs (Figure 4.3(a)).

D'autre part, la décomposition de la saisonnalité nous a permis d'analyser les variations périodiques et saisonnières des variables collectées. Nous avons constaté que toutes les variables présentent la même période de saisonnalité, avec un décalage de quelques jours seulement (Figure 4.3(b)).

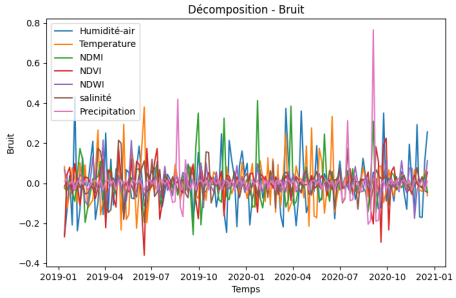
Enfin, l'évaluation du bruit dans les variables collectées nous a aidé à distinguer les va-



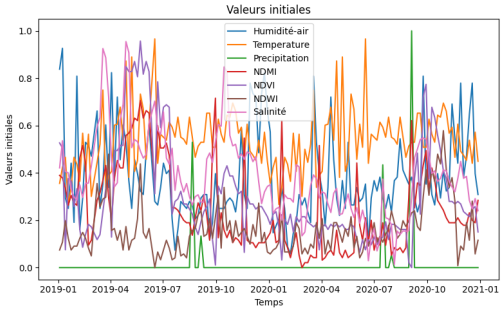
(a) Tendance



(b) Saisonnalité



(c) Bruit



(d) Valeurs initiales

FIGURE 4.3 – Décomposition des données de l'ONM et des données satellites

riations aléatoires des véritables tendances ou relations significatives. Par exemple, nous pouvons voir que les précipitations sont classées comme bruit, car cette zone n'a pas beaucoup de précipitations et est considérée comme une exception rare (Figure 4.3(c)).

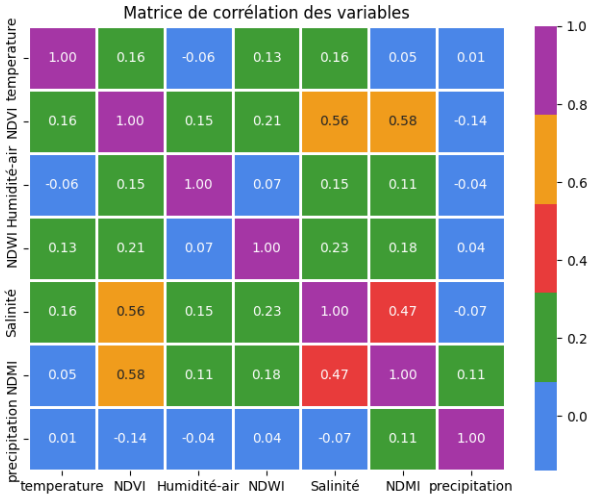


FIGURE 4.4 – Matrice de corrélation des variables

Troisième partie

Prédiction et optimisation du rendement des rizières en Mauritanie

Chapitre 5

Interface de monitoring administrateur

Dans ce chapitre, nous décrirons l'installation du système de monitoring, en précisant les matériels utilisés. Nous explorerons également l'interface Meraki, les données collectées, et les services qu'elle fournit, formant ainsi la base de notre système de surveillance administrateur.

5.1 Installation du système de monitoring

La transformation de notre architecture pour la gestion intelligente des rizières en un système réellement installé dans l'une des rizières de Mauritanie a été un défi de taille. Nous avons dû tenir compte des conditions climatiques et de la situation géographique des rizières, qui se trouvent généralement à une grande distance, le long des frontières avec le Sénégal.

Après une analyse approfondie, nous avons pris la décision de mettre en place notre système dans une rizière spécifique située dans la région de Tekane, à environ 63 km de Rosso en Mauritanie. Cette rizière présentait l'avantage d'avoir une durée d'exploitation de plus de 10 ans, ce qui témoignait de sa stabilité et de sa viabilité. Ce choix nous a permis de relever le défi en adaptant notre architecture à ce contexte spécifique et en fournissant une solution de gestion intelligente adaptée aux besoins de la rizière de Tekane.

5.1.1 Matériels utilisés

Nous avons décidé de travailler avec les appareils Meraki pour plusieurs raisons. Tout d'abord, leur précision est certifiée dans le domaine de l'intelligence artificielle, ce qui nous a donné confiance en leur capacité à collecter des données précises et fiables. Cette précision est essentielle pour garantir une gestion efficace des rizières et prendre des décisions éclairées.

Ensuite, la puissance des appareils Meraki a été un facteur déterminant dans notre choix. Nous avons besoin de dispositifs robustes et durables, capables de résister aux conditions climatiques difficiles et de fonctionner de manière optimale pendant de longues périodes. Les appareils Meraki répondent à ces exigences, ce qui nous permet de bénéficier d'une solution de surveillance à long terme, sans compromettre les performances .

En résumé, nous avons opté pour les appareils Meraki en raison de leur précision certifiée dans le domaine de l'IA et de leur puissance durable, qui nous permettent de relever les défis spécifiques liés à la gestion des rizières dans des conditions climatiques exigeantes.

Notre système de surveillance intelligent des rizières est composé d'un ensemble de matériaux et de capteurs spécifiques qui fonctionnent de manière synergique pour collecter et analyser des données clés. Les principales composantes de notre système comprennent (Figure 5.1) :

1. Caméra Meraki MV63 : Une caméra mini-dôme à objectif fixe offrant une résolution jusqu'à 4K, une vision nocturne avec éclairage IR jusqu'à 20 mètres, et des fonctionnalités avancées telles que l'indexation intelligente des mouvements et l'apprentissage automatique pour la détection d'objets. Elle est conçue pour une utilisation en extérieur et est certifiée IK10+ et IP67 pour résister aux environnements exigeants.

2. Capteur MT10 : Un capteur d'humidité et de température qui offre une précision de 2,5%. Il est alimenté par deux piles AA et dispose d'une connectivité Bluetooth pour une transmission sans fil des données. Il est conçu pour une utilisation en intérieur et sa durée de vie de batterie peut atteindre jusqu'à 5 ans.

3. Deux capteurs MT14 : Des capteurs de qualité de l'air qui mesurent des paramètres

tels que la qualité de l'air, la température, l'humidité, les TVOC (composés organiques volatils) et le bruit ambiant. Ils sont alimentés par quatre piles AA et prennent en charge la connectivité sans fil Bluetooth 4.2 LE.

4. Routeur MG21 Cellular Gateway : permet d'établir une connexion sans fil fiable et sécurisée avec les réseaux cellulaires. Ce routeur joue un rôle clé en convertissant le signal LTE provenant d'un fournisseur de services cellulaires en une connexion Ethernet, permettant ainsi la transmission des données collectées par nos capteurs vers la plateforme Meraki. Il est alimenté par un adaptateur secteur.
5. Deux adaptateurs : Des adaptateurs utilisés pour éviter que la caméra et le routeur ne soient exposés à des conditions de brillance excessive.
6. Batterie de 120 MA : Une batterie rechargeable utilisée comme source d'alimentation de secours pour le système en cas de panne d'électricité.
7. Régulateur : Un régulateur qui contrôle la charge de la batterie et assure une alimentation stable et fiable.
8. Panneau solaire de 250 watts : Un panneau solaire utilisé pour recharger la batterie en utilisant l'énergie solaire, ce qui permet une utilisation plus durable et respectueuse de l'environnement du système.

Ces composantes matérielles fonctionnent de manière coordonnée pour collecter des données précises sur la qualité de l'air, la température, l'humidité, les TVOC et le bruit ambiant dans les rizières. Les données sont ensuite transmises et analysées par notre système de surveillance intelligent, permettant ainsi une gestion plus efficace et une prise de décision éclairée dans le domaine de la culture du riz.

Nous avons utilisé une carte SIM sénégalaise de l'opérateur Orange pour connecter le routeur et assurer la transmission des données vers la plateforme Meraki. Cette carte SIM permet d'établir une connexion Internet via le réseau mobile, ce qui garantit une connectivité fiable et continue pour le système de surveillance des rizières. Grâce à cette connexion, les données collectées par les capteurs sont envoyées en temps réel vers la plateforme Meraki,



FIGURE 5.1 – Installation du système de monitoring

Équipement	Caractéristiques	Dimensions	Connectivité
Camera MV63X	Capteur d'image, résolution jusqu'à 4K, vision nocturne jusqu'à 20m	-	Sans fil - Bluetooth 4.2 LE
Capteur MT10	Capteur d'humidité et de température, plage de mesure : 0-95%, précision : 2,5%	26 mm x 65,9 mm x 117,1 mm	Bluetooth
Capteur MT14	Capteur de qualité de l'air	26 mm x 67 mm x 120 mm	Sans fil - Bluetooth 4.2 LE
Routeur MG21	Gateway cellulaire avec Ethernet	-	4G
Adaptateurs	Utilisés pour éviter les reflets du camera et du routeur	-	-
Batterie	Capacité de 120 mAh	-	-
Régulateur	Régule la tension électrique	-	-
Panneau solaire	Puissance de 250W	-	-

TABLE 5.1 – Caractéristiques du matériel utilisé

où elles sont traitées et analysées pour une gestion efficace des rizières.

5.2 Interface Meraki

L'interface Meraki joue un rôle essentiel dans notre système de gestion intelligent des rizières en Mauritanie. Cette interface conviviale et intuitive nous permet de visualiser et d'analyser les données collectées par nos capteurs, notamment la qualité de l'air, la température, l'humidité, les TVOC et le bruit ambiant. Grâce à l'interface Meraki, nous avons une vue d'ensemble de l'état de nos rizières en temps réel, ce qui nous permet de prendre des décisions éclairées pour optimiser nos opérations agricoles. Nous pouvons facilement configurer des alertes et des notifications pour être informés des variations importantes dans les conditions environnementales, ce qui nous aide à intervenir rapidement et à prévenir d'éventuels problèmes. De plus, l'interface Meraki offre des fonctionnalités avancées telles que l'analyse des données, la génération de rapports et la gestion centralisée, ce qui facilite la surveillance et la maintenance de notre système (Figure 5.2)

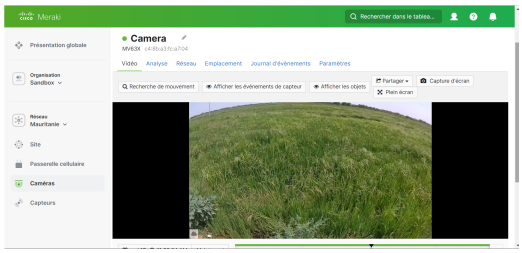
5.2.1 Données collectées

Chaque capteur utilisé collecte des données spécifiques qui sont essentielles pour la surveillance et l'analyse de l'environnement. Les données collectées par les capteurs comprennent des informations telles que la température, l'humidité, la qualité de l'air, le TVOC (composés organiques volatils totaux) et le bruit ambiant.

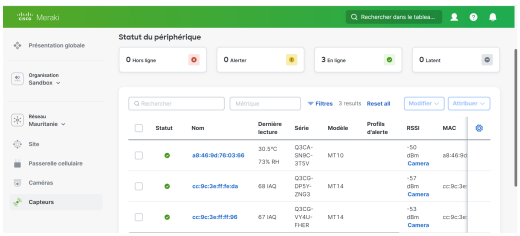
L'interface Meraki offre une visualisation claire et informative de ces données à travers



(a) Modem



(b) Camera



(c) Capteurs

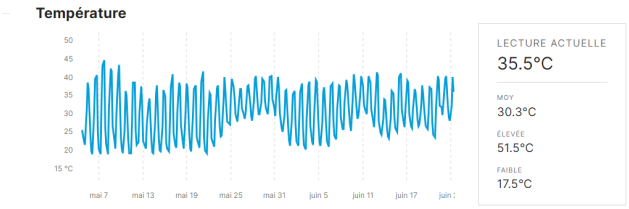
FIGURE 5.2 – Interface de Meraki

trois types de graphiques différents pour chaque variable mesurée. Ces graphiques présentent les valeurs maximales, minimales et moyennes des données collectées sur une période donnée (Figures 5.3 , 5.4 et 5.5). Cela permet aux utilisateurs de surveiller les fluctuations des paramètres environnementaux au fil du temps et d'identifier les tendances ou les anomalies potentielles.

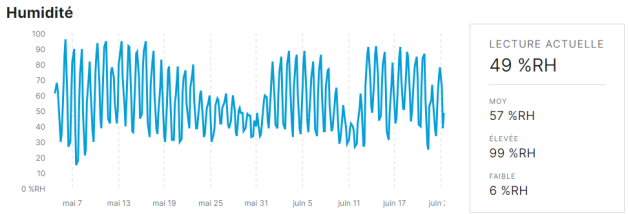
5.2.2 Services fournis par Meraki

L'interface Meraki propose également la possibilité de tracer des diagrammes psychrométriques. Ces diagrammes représentent les données de bulbe humide et de bulbe sec, qui décrivent les caractéristiques de l'air humide en termes de température, d'humidité relative, d'enthalpie spécifique et de point de rosée. Ces informations sont essentielles pour comprendre et analyser les propriétés de l'air dans un environnement donné, en particulier dans le contexte de la gestion de la qualité de l'air.

Deplus, notre interface Meraki offre une visualisation en temps réel de la connectivité des appareils utilisés dans notre système de monitoring. Nous avons accès à un tableau de bord intuitif qui affiche l'état de chaque appareil connecté, indiquant s'ils sont en ligne, hors ligne



(a) Temperature



(b) Humidité

FIGURE 5.3 – Données du capteur MT10

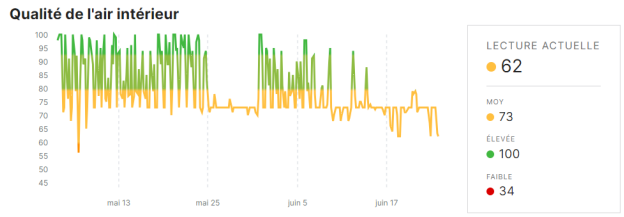
ou en mode veille.

Grâce à cette fonctionnalité, nous pouvons rapidement identifier les appareils qui rencontrent des problèmes de connectivité et prendre des mesures pour les résoudre. Notre interface Meraki affiche également des informations détaillées sur la force du signal et la qualité de la connexion de chaque appareil, nous permettant ainsi de diagnostiquer plus précisément les problèmes de connectivité.

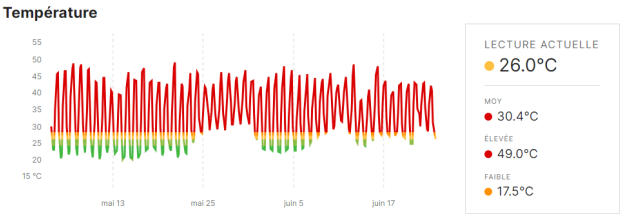
De plus, notre interface Meraki fournit des alertes en temps réel pour nous informer des interruptions de connectivité ou de tout autre problème rencontré par nos appareils. Ces alertes peuvent être personnalisées en fonction de nos préférences, nous permettant de recevoir des notifications par e-mail ou via d'autres canaux de communication.

En résumé, l'interface Meraki offre une vue en temps réel de la connectivité de nos appareils, nous permettant de surveiller et de gérer efficacement notre système de monitoring. Cela nous aide à maintenir la disponibilité et la fiabilité de nos appareils, garantissant ainsi une collecte continue et précise des données nécessaires à nos analyses et à nos prises de décision.

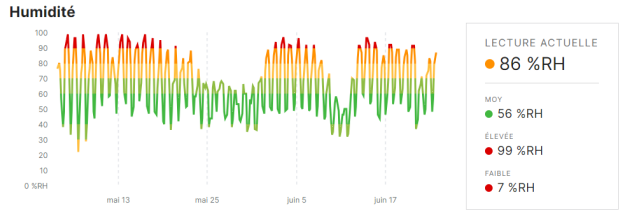
En outre, Notre interface Meraki va au-delà de la simple visualisation de la connectivité des appareils. Elle nous permet également de consulter les images en temps réel capturées



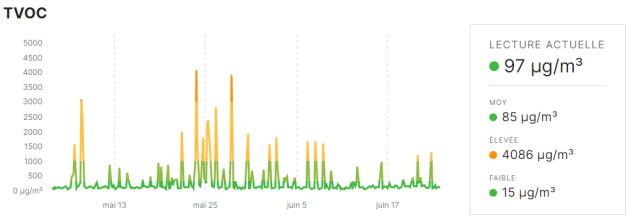
(a) Qualité de l'air



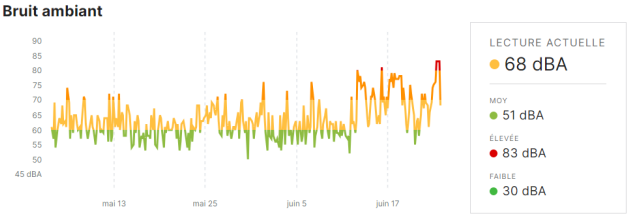
(b) Temperature



(c) Humidité

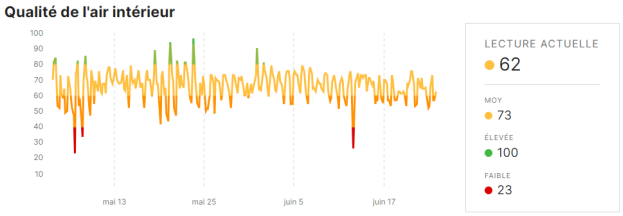


(d) TVOC

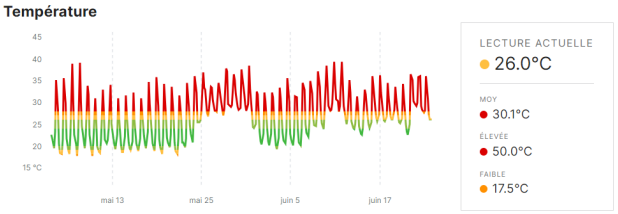


(e) Bruit ambiant

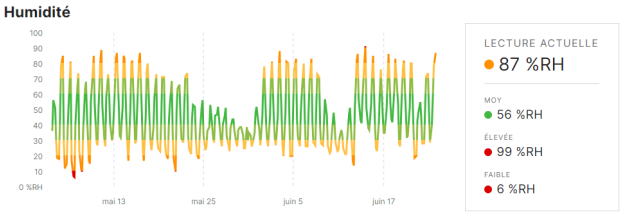
FIGURE 5.4 – Données du capteur MT14-1 (valeurs maximales)



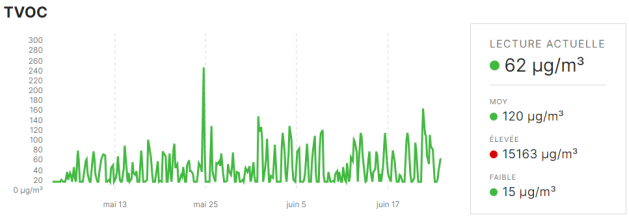
(a) Qualité de l'air



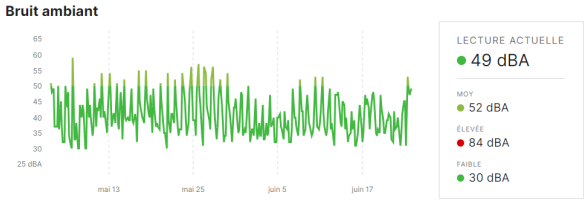
(b) Temperature



(c) Humidité



(d) TVOC



(e) Bruit ambiant

FIGURE 5.5 – Données du capteur MT14-2 (valeurs minimales)

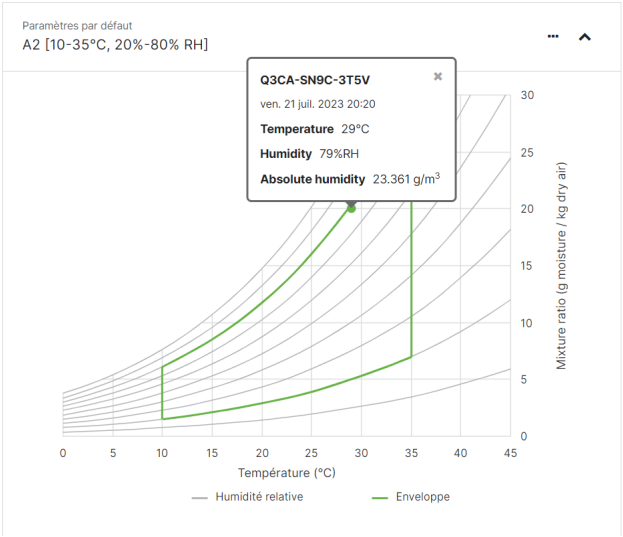


FIGURE 5.6 – graphe economie d'energie via MT14

par nos caméras sur le terrain. Grâce à cette fonctionnalité, nous pouvons observer en direct l'état des champs de riz et suivre leur évolution.

De plus, l'interface Meraki intègre des fonctionnalités avancées de détection de mouvement. Elle est capable de détecter les mouvements des plantes et des personnes présentes dans les champs. Cela nous permet d'obtenir des informations précieuses sur l'état de croissance des plantes, d'identifier d'éventuels problèmes ou anomalies, et de prendre des mesures appropriées en conséquence.

En détectant les mouvements des personnes, l'interface Meraki nous aide également à assurer la sécurité du site. Nous pouvons être alertés en cas de présence non autorisée ou de comportements suspects, ce qui nous permet de réagir rapidement et de prendre les mesures nécessaires pour protéger nos installations.

En résumé, notre interface Meraki nous offre la possibilité de consulter les images en temps réel, de détecter les mouvements des plantes et des personnes, et de prendre des mesures appropriées en fonction des informations recueillies. Cela nous permet de surveiller activement nos champs de riz, d'optimiser les opérations agricoles et de garantir la sécurité de notre site.

Chapitre 6

Interface de monitoring utilisateur

L'utilisation de l'interface administrateur est un outil précieux pour la surveillance en temps réel des rizières. Grâce à cette interface, les utilisateurs peuvent accéder aux données en temps réel provenant de divers dispositifs tels que des caméras et des capteurs de température, d'humidité de l'air, de TVOC, de bruit et de qualité de l'air. Cela leur permet de surveiller les conditions environnementales et de détecter rapidement les variations ou les problèmes potentiels.

Cependant, il est important de noter que l'interface administrateur a ses limites. Elle ne propose pas d'outils d'analyse avancés intégrés ni de fonctionnalités de recommandation spécifiques aux agriculteurs. Ainsi, bien qu'elle soit utile pour la surveillance en temps réel, elle ne permet pas d'exploiter pleinement le potentiel des données collectées.

Pour combler cette lacune, il serait nécessaire de développer un système complémentaire qui intègre des fonctionnalités d'analyse des données et de recommandation. Ce système pourrait utiliser des techniques avancées d'analyse de données et d'intelligence artificielle pour extraire des informations significatives des données collectées. En analysant ces données de manière approfondie, il serait possible de détecter des tendances, des modèles et des relations complexes.

Dans ce chapitre, nous allons présenter en détail notre interface utilisateur conçue spécifiquement pour l'analyse des données collectées. Cette interface joue un rôle central dans notre système, permettant aux utilisateurs d'exploiter pleinement les données recueillies à partir de différentes

sources.

L'une des fonctionnalités clés de notre interface est l'application de notre modèle de clustering aux données provenant des rizières de Tekane, collectées à la fois par notre système et par des satellites. Le clustering est une technique d'analyse de données qui regroupe les observations similaires en clusters afin de détecter des schémas et des structures dans les données. En appliquant notre modèle de clustering aux données de rizière, nous sommes en mesure de fournir des informations précieuses sur la distribution spatiale des caractéristiques des rizières, telles que la température, l'humidité du sol, la salinité, etc. Cela permet aux agriculteurs d'avoir une meilleure compréhension de la variabilité des conditions dans leurs rizières et d'ajuster leurs pratiques en conséquence.

En outre, notre interface offre la possibilité de prédire le rendement du riz en se basant sur divers facteurs, tels que la salinité du sol, l'humidité des plantes et d'autres facteurs climatiques. Ces prédictions sont générées à l'aide de modèles d'apprentissage automatique spécialement entraînés sur les données historiques de rendement de riz et sur les mesures des différents paramètres. Les agriculteurs peuvent ainsi obtenir des estimations de rendement précises, ce qui les aide à prendre des décisions éclairées concernant leurs cultures et leurs pratiques agricoles.

En plus de la prédiction du rendement du riz, notre système est capable de prédire d'autres facteurs importants tels que la salinité du sol, l'humidité des plantes et d'autres facteurs climatiques.

6.1 Exploitation des données IoT

6.1.1 Analyse des données provenant de dispositifs connectés à l'Internet des objets (IoT).

Pendant la période de contre-saison chaude CSC au Tekane, qui se déroule en mai et juin, nous avons installé notre système de surveillance dans les rizières. En examinant les

données collectées par nos capteurs, nous avons remarqué que la température atteint régulièrement 50°C au cours de la journée. La plus petite valeur de température enregistrée est de 17°C entre 04h00 et 08h00 du matin, après quoi la température commence à augmenter progressivement.

Ces conditions de température élevée peuvent avoir plusieurs effets sur la culture du riz pendant la période de CSC :

Stress thermique : Les températures élevées, atteignant jusqu'à 50°C, peuvent provoquer un stress thermique chez les plants de riz. Le stress thermique peut affecter la physiologie des plantes, perturber leur métabolisme et nuire à leur croissance et à leur développement. Les plantes peuvent montrer des signes de flétrissement, une réduction de la production de biomasse et une diminution du rendement.

Besoins en eau accrus : Les températures élevées entraînent une évaporation accrue de l'eau dans les rizières, ce qui peut augmenter les besoins en eau des plants. Il est important de veiller à ce que les rizières soient suffisamment irriguées pour maintenir une humidité adéquate dans le sol et éviter le stress hydrique des plantes.

Réduction de la durée du cycle de croissance : Les températures élevées peuvent accélérer le développement des plants de riz, réduisant ainsi la durée du cycle de croissance. Cela peut entraîner une maturation précoce des plants, une diminution de la période de remplissage des grains et une réduction du rendement global.

Risques accrus de maladies et de ravageurs : Les températures élevées peuvent favoriser la propagation de maladies fongiques et augmenter la reproduction des ravageurs spécifiques au riz. Il est important de surveiller régulièrement les cultures et de mettre en place des mesures de lutte intégrée pour minimiser les pertes potentielles dues aux maladies et aux ravageurs.

Il est crucial de mettre en œuvre des stratégies adaptées pour faire face aux effets de la température élevée pendant la période de CSC. Cela peut inclure l'utilisation de variétés de riz résistantes à la chaleur, l'ajustement des pratiques d'irrigation pour répondre aux besoins en eau accrus, la surveillance régulière des cultures pour détecter les signes de stress

thermique, et l'application de mesures de lutte intégrée pour prévenir les maladies et les ravageurs.

Nous avons également observé des niveaux d'humidité élevés. Les données collectées par nos capteurs ont révélé des variations importantes dans les niveaux d'humidité en fonction de l'heure de la journée.

Nous avons enregistré des valeurs d'humidité relativement basses, atteignant 7% à une occasion et 10% à plusieurs reprises au début du mois de mars. Cependant, nous avons également constaté que les valeurs d'humidité atteignent régulièrement 99% pendant les deux mois de la période de CSC, presque chaque jour entre 00 :00 et 08 :00.

Ces niveaux élevés d'humidité peuvent avoir un impact sur la culture du riz. Voici quelques effets potentiels :

Climat local : Le Tekane peut être situé dans une zone où le climat est caractérisé par une humidité élevée pendant la période de CSC. Des conditions météorologiques spécifiques, telles que la proximité de l'eau, la topographie ou les régimes de vents, peuvent contribuer à des niveaux d'humidité élevés dans la région.

Évapotranspiration : L'évapotranspiration, qui est le processus combiné d'évaporation de l'eau du sol et de transpiration des plantes, peut être importante pendant la période de CSC. Les rizières, avec leur couverture végétale dense, peuvent favoriser l'évapotranspiration et maintenir des niveaux d'humidité élevés dans l'air ambiant.

Irrigation : La pratique de l'irrigation dans les rizières peut également contribuer à maintenir une humidité élevée. L'irrigation régulière nécessaire pour la culture du riz peut maintenir le sol humide, ce qui augmente l'humidité dans l'air environnant.

Condensation nocturne : Pendant la nuit, lorsque les températures chutent, l'air peut se refroidir et atteindre son point de rosée, ce qui entraîne la formation de gouttelettes d'eau par condensation. Cela peut augmenter l'humidité relative de l'air pendant les premières heures du matin.

Il est important de noter que des niveaux d'humidité élevés peuvent influencer la croissance des cultures de riz et augmenter les risques de maladies et de ravageurs. Une gestion

adéquate de l'irrigation et de la surveillance des cultures peut être nécessaire pour maintenir un équilibre optimal entre l'humidité et les autres facteurs environnementaux afin de favoriser une croissance saine des plants de riz et d'assurer des rendements satisfaisants.

En outre, nous avons utilisé les mesures de bruit ambiant captées par nos capteurs pour évaluer les conditions environnementales dans les rizières, en particulier l'effet du vent. Étant donné que le champ agricole est éloigné de la route principale où il n'y a pas de voitures et qu'il n'y a pas d'habitants à proximité des rizières, nous pouvons raisonnablement considérer que le bruit ambiant enregistré représente principalement le bruit généré par le vent.

En examinant les données collectées, nous avons remarqué une variation irrégulière des valeurs de bruit ambiant, qui varient entre 50 et 76 dBA. Ces variations peuvent être interprétées comme étant directement liées à l'effet du vent sur les rizières.

Le bruit ambiant résulte des mouvements de l'air provoqués par le vent. Des valeurs plus élevées de bruit ambiant, telles que 76 dBA, peuvent indiquer une augmentation de la vitesse du vent, tandis que des valeurs plus basses, telles que 50 dBA, peuvent correspondre à une diminution de la vitesse du vent. Ces variations de bruit ambiant sont donc un indicateur indirect de l'activité et de la force du vent dans la zone d'étude.

L'effet du vent sur les rizières peut avoir plusieurs implications pour la culture du riz :

Dispersion des pollens : Le vent joue un rôle essentiel dans la dispersion des pollens des plantes de riz. Une vitesse de vent adéquate facilite la pollinisation croisée entre les plantes, ce qui favorise la fécondation et la production de grains.

Évapotranspiration : Le vent peut augmenter l'évaporation de l'eau dans les rizières. Cela peut avoir un impact sur les besoins en irrigation, l'humidité du sol et le stress hydrique des plantes.

Mouvement des ravageurs : Le vent peut influencer le mouvement des ravageurs qui affectent les cultures de riz, tels que les insectes nuisibles. Une forte brise peut perturber leur vol ou leur capacité à se déplacer d'une plante à une autre, ce qui peut réduire les risques d'infestation.

Il est important de surveiller et de comprendre les variations du bruit ambiant pour éva-

luer l'effet du vent sur les rizières. Cela peut aider les agriculteurs à prendre des décisions éclairées sur la gestion des cultures, notamment en adaptant les pratiques d'irrigation, en prévenant les problèmes de dispersion du pollen et en mettant en œuvre des stratégies de lutte intégrée contre les ravageurs.

Dans notre étude, nous avons également mesuré les TVOC (Composés Organiques Volatils Totaux) pendant les mois de mai et juin dans les rizières. Les TVOC sont des substances chimiques volatiles qui peuvent provenir de diverses sources, telles que les émissions industrielles, les produits chimiques agricoles, les combustibles fossiles et les activités humaines.

D'après nos mesures, nous avons constaté que la plupart du temps, les niveaux de TVOC étaient très faibles, se situant généralement entre 0 et 500 μg/m^3. Cette gamme de concentration suggère que l'air ambiant dans les rizières présente des niveaux de TVOC relativement bas, indiquant une bonne qualité de l'air. Cette observation peut indiquer une utilisation limitée de produits chimiques et une approche potentielle de culture plus respectueuse de l'environnement.

Des niveaux de TVOC bas dans les rizières peuvent être le résultat de pratiques agricoles durables, telles que l'utilisation minimale de pesticides et d'engrais chimiques, ou l'adoption de méthodes de lutte biologique contre les ravageurs et les maladies. Une gestion appropriée des ressources naturelles, une rotation des cultures, la préservation de la santé des sols et l'utilisation de techniques d'irrigation efficaces peuvent également contribuer à maintenir des niveaux de TVOC bas.

Cependant, nous avons identifié deux pics de TVOC pendant la période d'étude, atteignant des niveaux élevés allant jusqu'à 4000 μg/m^3. Ces pics plus élevés peuvent être attribués à des événements spécifiques, tels que des activités agricoles à proximité des rizières.

En resumé, nous avons utilisé un indicateur de la Qualité de l'Air qui prend en compte la température, l'humidité et les TVOC. Nous avons constaté que cet indicateur affiche toujours une valeur de 100, ce qui suggère que la qualité de l'air est constamment excellente. Cependant, il est important de noter que les jours où l'indicateur de la qualité de l'air a diminué jusqu'à 73, correspondent aux mêmes jours où des pics de TVOC ont été détectés. Cette

corrélation peut indiquer une utilisation accrue de produits chimiques pour lutter contre les ravageurs agricoles.

6.2 Intégration des données IoT et des données satellitaires

Dans notre étude de recherche, nous avons entrepris de combiner des données satellitaires, notamment le NDVI (Indice de Végétation par Différence Normalisée), le NDWI (Indice de l'Eau par Différence Normalisée), le NDMI (Indice de l'Humidité par Différence Normalisée) et la salinité du sol (voir la section 2.1), avec des données provenant de capteurs IoT, tels que l'humidité de l'air, la température de l'air, la qualité de l'air, les TVOC (Composés Organiques Volatils Totals) et le bruit ambiant. Ces données ont été collectées sur une période d'un mois, avec plusieurs relevés effectués tout au long de la journée, pour un total de 896 enregistrements.

L'objectif principal de notre étude était de comprendre les relations entre ces différentes variables et d'examiner comment elles se combinent pour fournir des informations plus complètes sur l'environnement étudié. Pour cela, nous avons d'abord analysé les corrélations entre toutes les variables collectées.

Les corrélations positives que nous avons trouvées nous ont permis de mettre en évidence certaines relations cohérentes entre les paramètres. Par exemple, nous avons observé une corrélation positive de 0,65 entre le NDVI et le NDMI, ce qui suggère qu'une augmentation du NDVI est associée à une augmentation du NDMI. De même, une corrélation positive de 0,19 entre le NDMI et le NDWI indique qu'une augmentation du NDMI est associée à une augmentation du NDWI. De plus, nous avons identifié une corrélation positive de 0,19 entre la salinité du sol et le CO2, ce qui suggère qu'une augmentation de la salinité du sol est associée à une augmentation des niveaux de CO2.

D'autre part, nous avons également identifié des corrélations négatives, qui mettent en évidence des relations inverses entre les variables. Par exemple, nous avons trouvé une forte corrélation négative de -0,92 entre l'humidité de l'air et la température de l'air, indi-

quant qu'une augmentation de l'humidité est généralement associée à une diminution de la température. De même, une corrélation négative de -0,65 entre la température de l'air et l'indice IAQ suggère qu'une augmentation de la température de l'air est associée à une diminution de la qualité de l'air. Enfin, nous avons relevé une corrélation négative de -0,28 entre le bruit ambiant et l'indice IAQ, indiquant qu'une augmentation du bruit ambiant est associée à une diminution de la qualité de l'air (Figure 6.1).

Ces résultats de corrélation nous ont fourni des indices précieux sur les interactions entre les différents paramètres environnementaux mesurés. Cette intégration de données à partir de sources variées, telles que les satellites et les capteurs IoT, nous a permis de mieux comprendre l'écosystème étudié et d'obtenir des informations plus complètes sur son fonctionnement. Ces découvertes pourraient avoir des implications importantes dans divers domaines tels que l'agriculture, la gestion de l'environnement, la surveillance des ressources naturelles et bien d'autres applications liées à la compréhension et à la préservation de notre environnement.

Ensuite, nous avons tenté de tracer la décomposition additive des données collectées dans le but d'analyser les tendances saisonnières et les fluctuations de bruit. Cependant, il a été observé qu'il n'y avait pas de saisonnalité marquée ni de bruit significatif dans les données.

Une hypothèse possible pour cette absence de saisonnalité et de bruit est liée à la durée de la période d'étude, qui ne couvrait qu'un mois. En général, les analyses de décomposition additive sont plus efficaces lorsqu'elles sont effectuées sur des séries chronologiques plus longues, permettant de mieux saisir les modèles saisonniers et les tendances sur une période de temps plus étendue.

Alors, il est important de noter que la durée de la période d'étude peut avoir un impact significatif sur les résultats des analyses temporelles. Si l'objectif est de détecter des modèles saisonniers ou des tendances plus subtiles, une période d'observation plus longue pourrait être nécessaire pour obtenir des conclusions plus fiables et représentatives.

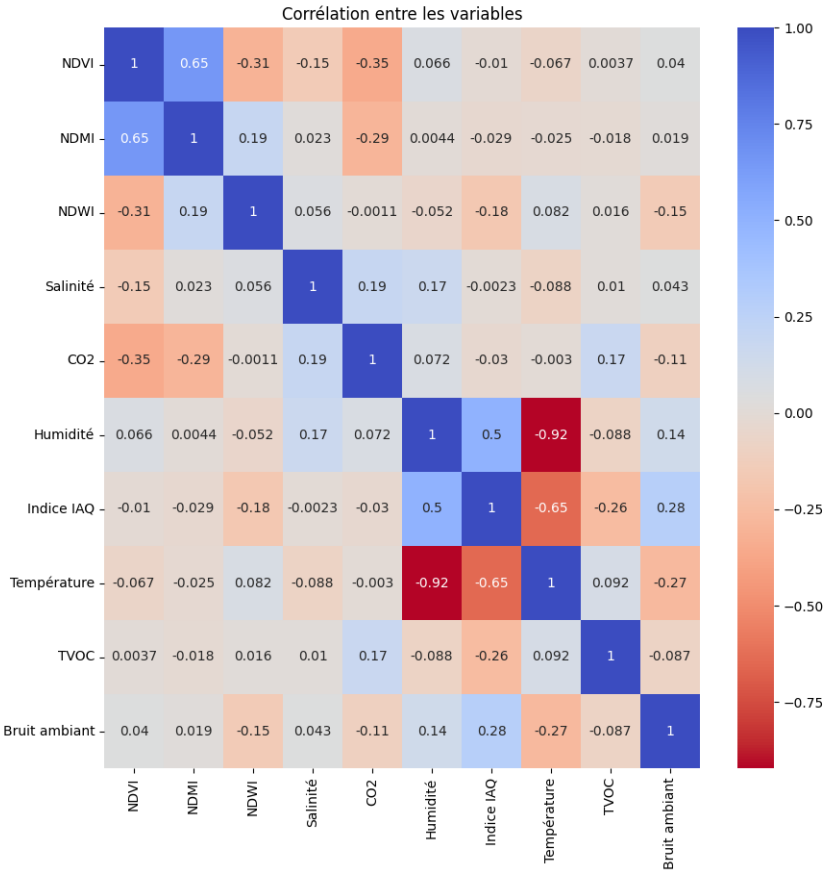


FIGURE 6.1 – Matrice de corrélation des indices satellitaires avec les variables collectées par notre système.

6.2.1 Clustering des données IoT et satellitaires

Nous avons appliqué notre modèle de clustering D-DBSCAN, en le combinant avec la métrique de similarité DTW et en utilisant l'hypothèse H1 pour choisir l'ϵ adaptatif pour chaque enregistrement de notre ensemble de données provenant d'un champ agricole de riz situé au Tekane, Rosso, Mauritanie.

L'ensemble de données que nous avons utilisé pour cette étude contient 10 variables importantes liées à la culture du riz dans cette région. Ces variables comprennent le NDVI, NDMI, NDWI, la salinité du sol, la température, l'humidité de l'air, le TVOC, le bruit ambiant et l'indice de la qualité de l'air.

L'objectif était de choisir des combinaisons de ces variables qui pourraient potentielle-

ment fournir des informations significatives sur l'état de votre champ de riz. Nous avons sélectionné les combinaisons en fonction de la pertinence des variables dans le contexte de la riziculture et de l'agriculture en général, ainsi que de leur capacité à donner une vue complète de l'environnement de champ de riz.

- **Température, Humidité , NDVI** : Cette combinaison permet d'évaluer l'impact des conditions climatiques (température et humidité) sur la santé de la végétation (NDVI) dans chaque cluster, ce qui permet de mieux comprendre comment les variations de température et d'humidité influencent la croissance des cultures (Figure 6.2(a)).
- **Salinité, Température, Humidité** : Ces variables sont cruciales pour évaluer les conditions environnementales du champ de riz, en particulier en relation avec la salinité du sol (Figure 6.2(b)).
- **NDVI, NDMI, NDWI** : Ces indices sont couramment utilisés pour évaluer la santé de la végétation, l'humidité du sol et la présence d'eau. Ils sont essentiels pour comprendre la croissance des cultures dans le champ de riz (Figure 6.2(c)).
- **NDVI, Salinité, Temperature** : Le NDVI est toujours inclus car il fournit une évaluation directe de la végétation. Nous avons ajouté la salinité pour évaluer son impact sur la végétation en fonction de la température (Figure 6.2(d)).
- **NDMI, NDWI, Humidité** : Ces combinaisons incluent des mesures d'humidité pour évaluer leur impact sur l'indice NDMI (humidité du sol et de la végétation) et l'indice NDWI (présence d'eau). Cela pourrait être important pour la santé globale de la végétation (Figure 6.2(e)).
- **NDWI, Salinité, CO2** : Cette combinaison explore comment la salinité du sol peut être liée à la présence d'eau (NDWI) et à la concentration de CO2, qui peut affecter la santé de la végétation (Figure 6.3(a)).
- **CO2, humidité, tvoc** : Cette combinaison examine les niveaux de CO2, d'humidité et de composés organiques volatils, qui peuvent être liés à la qualité de l'air et affecter la croissance des cultures (Figure 6.3(b)).
- **Temperature, tvoc, Bruit ambiant (vent)** : La température est une variable clé pour éva-

luer les conditions climatiques, et les niveaux de TVOC (composés organiques volatils) et de bruit ambiant peuvent influencer l'environnement global du champ de riz (Figure 6.3(c)).

- **NDVI, NDMI, CO2** : Cette combinaison associe la santé de la végétation (NDVI et NDMI) avec le niveau de CO2, ce qui pourrait aider à comprendre comment la santé de la végétation est influencée par la concentration de CO2 dans l'air (Figure 6.3(d)).
- **NDMI, Salinité, Bruit ambiant (vent)** : Cette combinaison explore comment la santé de la végétation (NDMI) est liée à la salinité du sol et au bruit ambiant, ce dernier pouvant avoir un impact sur la croissance des cultures (Figure 6.3(e)).

En analysant les figures représentant les différents clusters dans le champ de riz, nous pouvons observer qu'il y a trois clusters principaux identifiés, notés par les couleurs rouge, orange et jaune, ainsi que quelques points aberrants représentés par la couleur bleue (Figures 6.2 et 6.3).

Le cluster dominant (rouge) est caractérisé par des valeurs moyennes et denses de NDVI, indiquant une végétation saine et vigoureuse dans cette zone. De plus, ce cluster montre des variations significatives dans les mesures de NDMI, salinité, température et humidité de l'air, ce qui suggère des conditions environnementales diverses à l'intérieur de ce groupe. En outre, les valeurs de NDWI sont relativement faibles dans ce cluster, ce qui signifie qu'il y a une présence d'eau relativement limitée. Les points aberrants en bleu dans ce cluster sont caractérisés par des jours où le TVOC est élevé, ce qui peut indiquer une augmentation de la concentration de composés organiques volatils dans l'air ambiant, potentiellement liée à des facteurs environnementaux ou à des activités spécifiques.

Les clusters moins dominants sont représentés en orange et en jaune. Le cluster jaune est caractérisé par un NDVI proche de zéro, ce qui peut indiquer une absence ou une faible croissance de végétation dans cette zone. De plus, la salinité du sol est élevée dans ce cluster, ce qui peut être un facteur limitant pour la croissance des cultures de riz. En revanche, le cluster orange montre un NDVI élevé, suggérant une végétation plus dense et vigoureuse par rapport au cluster jaune. De plus, la salinité du sol est faible dans ce cluster, ce qui peut être

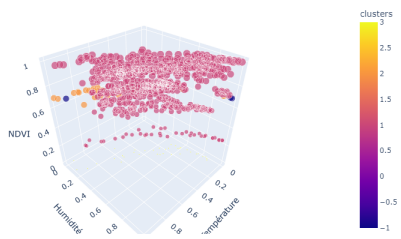
plus favorable à la croissance des cultures. Les valeurs de NDWI sont moyennes dans ce groupe, indiquant une présence d'eau modérée. Dans ce cas, la présence de valeurs élevées de NDWI ne reflète pas les pluies car le mois de mai et juin ne correspond pas à la période de pluie dans la région étudiée. Ainsi, la présence d'eau mise en évidence par le NDWI est plus probablement due à l'irrigation des cultures.

Il est important de noter que les points aberrants en bleu peuvent représenter des enregistrements atypiques ou des valeurs extrêmes qui se démarquent du comportement général des clusters principaux. Ces points pourraient être dus à des erreurs de mesure, à des conditions environnementales exceptionnelles ou à d'autres facteurs inattendus, tels que les variations des niveaux de TVOC.

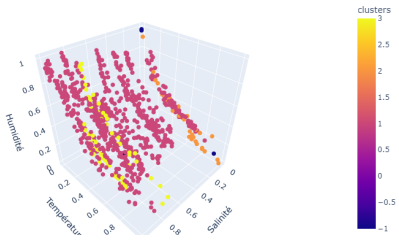
En résumé, les clusters obtenus à partir de l'analyse représentent différentes zones du champ de riz avec des caractéristiques distinctes. Le cluster dominant (rouge) montre une végétation saine avec des variations significatives dans les conditions environnementales. Les clusters moins dominants (orange et jaune) présentent des différences dans la santé de la végétation et la salinité du sol, indiquant des conditions environnementales contrastées. Les points aberrants en bleu méritent une attention particulière, car ils peuvent fournir des informations précieuses sur des événements ou des conditions exceptionnelles dans le champ.

Ces résultats offrent des perspectives importantes pour la gestion du champ de riz au Tekane, Rosso, Mauritanie. En utilisant les informations extraites des clusters, les agriculteurs pourraient adapter leurs pratiques agricoles en fonction des caractéristiques spécifiques de chaque zone, ce qui pourrait améliorer l'efficacité de la production agricole et la santé globale des cultures de riz. La présence d'eau mise en évidence par le NDWI étant plus probablement due à l'irrigation, cela pourrait également être pris en compte pour une meilleure gestion de la ressource en eau dans la région.

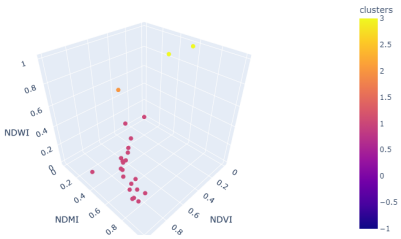
Comme chaque cluster présente des caractéristiques spécifiques en termes de santé de la végétation, de conditions environnementales et de salinité du sol, offrant ainsi des informations précieuses pour une meilleure gestion agricole. En se basant sur ces résultats, nous proposons des recommandations adaptées à chaque cluster pour aider les agriculteurs à op-



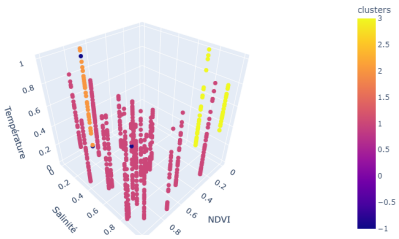
(a) Température, Humidité , NDVI



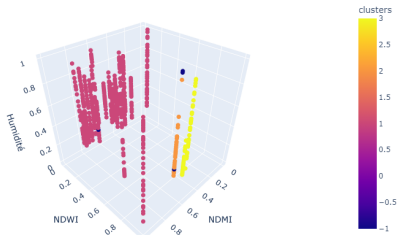
(b) — Salinité, Température, Humidité



(c) NDVI, NDMI, NDWI

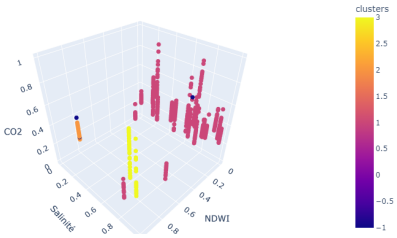


(d) NDVI, Salinité, Temperature

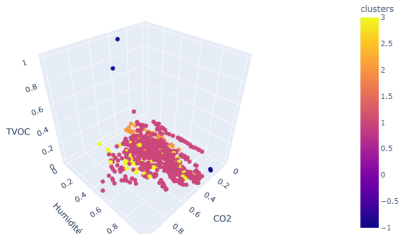


(e) NDMI, NDWI, Humidité

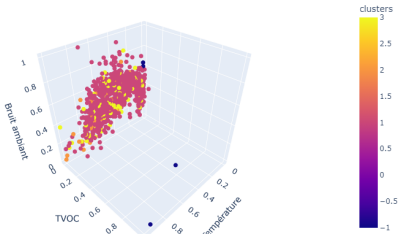
FIGURE 6.2 – Clustering des données satellitaires et IoT 1



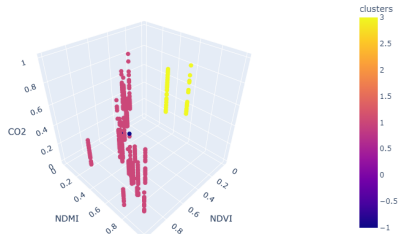
(a) NDWI, Salinité, CO2



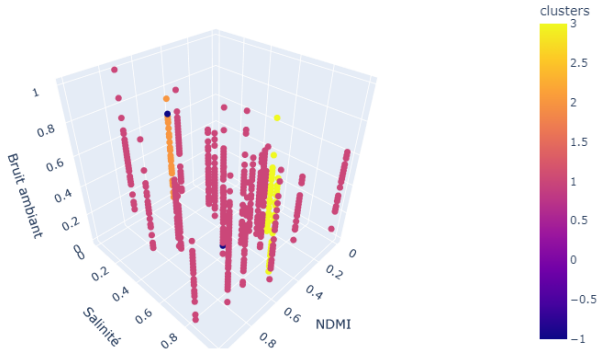
(b) CO2, humidité, tvoc



(c) Temperature, tvoc, Bruit ambiant (vent)



(d) NDVI, NDMI, CO2



(e) NDMI, Salinité, Bruit ambiant (vent)

FIGURE 6.3 – Clustering des données satellitaires et IoT 2

timiser leurs pratiques agricoles et à améliorer la productivité et la santé des cultures de riz dans cette région.

Recommandations pour chaque cluster

1. **Cluster Rouge (Végétation saine et conditions environnementales diverses) :**
   - Surveiller attentivement les variations de température et d'humidité de l'air, car elles semblent jouer un rôle essentiel dans la santé de la végétation. Des variations importantes peuvent nécessiter des ajustements dans les pratiques d'irrigation et de gestion de la culture.
   - Évaluer régulièrement la salinité du sol, car les variations dans cette mesure peuvent influencer la croissance des cultures. Si la salinité augmente, envisager des stratégies de gestion du sol pour réduire cet effet négatif.
   - Faire un suivi régulier du NDVI pour évaluer la santé globale de la végétation et détecter rapidement tout signe de stress ou de problèmes de croissance.
2. **Cluster Orange (Végétation dense avec faible salinité du sol) :**
   - Concentrer les efforts d'irrigation dans cette zone pour maintenir les niveaux d'humidité du sol et soutenir la croissance vigoureuse de la végétation.
   - Surveiller de près les variations de température et d'humidité de l'air pour ajuster les pratiques d'irrigation en fonction des conditions climatiques changeantes.
   - Étant donné que la salinité du sol est faible dans ce cluster, il peut être bénéfique de considérer des pratiques de gestion du sol pour maintenir cette condition favorable à la croissance des cultures.
3. **Cluster Jaune (Végétation faible et salinité élevée du sol) :**
   - Investiguer les raisons de la faible croissance de la végétation dans cette zone, qui pourrait être liée à la salinité élevée du sol. Des analyses approfondies du sol peuvent être nécessaires pour comprendre la nature du problème.
   - Considérer des pratiques de drainage améliorées pour réduire la salinité du sol et améliorer les conditions de croissance des cultures.

— Étant donné que le NDVI est proche de zéro dans ce cluster, il est important de surveiller de près l'évolution de la végétation dans cette zone et d'adopter des mesures correctives pour restaurer la santé des cultures.

4. **Remarques générales :**

   Pour les points aberrants (en bleu), qui sont caractérisés par des niveaux élevés de TVOC, il est essentiel de déterminer la source de ces composés organiques volatils. Des mesures de contrôle de la pollution et une évaluation de la qualité de l'air pourraient être nécessaires pour minimiser leur impact sur la santé des cultures.

   La présence d'eau mise en évidence par le NDWI étant plus probablement due à l'irrigation, les agriculteurs doivent optimiser l'utilisation de l'eau en fonction des besoins réels des cultures et des conditions climatiques pour une gestion durable de cette ressource précieuse.

En mettant en pratique ces recommandations, les agriculteurs pourront améliorer l'efficacité de leur gestion agricole, optimiser les rendements et assurer une meilleure santé globale des cultures de riz dans le champ au Tekane, Rosso, Mauritanie. Ces mesures adaptées à chaque cluster aideront à tirer le meilleur parti des ressources disponibles et à maximiser le potentiel de croissance des cultures, contribuant ainsi à une agriculture plus durable et rentable dans la région.

Parmi les données collectées par notre système de surveillance, ainsi que par notre Caméra Meraki MV63, nous avons obtenu des images en temps réel qui nous ont été extrêmement précieuses pour surveiller le champ de riz. Ces images nous permettent de détecter les personnes et les animaux qui entrent dans le champ, ce qui est essentiel pour assurer la sécurité et la protection des cultures. De plus, ces images ont été utilisées pour créer un dataset sur la croissance des plantes, offrant ainsi une vue détaillée et continue de l'évolution des cultures au fil du temps.

En analysant attentivement les images fournies, nous pouvons observer une maturation progressive des plantes jour après jour, ce qui est clairement identifiable par les changements de couleur des plantes et de leur densité. Ces changements indiquent la croissance

et le développement des cultures, nous permettant ainsi d'évaluer leur santé et leur vigueur au fil du temps. Ces informations précieuses nous aident à prendre des décisions éclairées concernant les pratiques agricoles, telles que le moment optimal pour la récolte.

Un autre avantage des images est la possibilité de détecter facilement les mauvaises herbes présentes dans le champ. Ces images fournissent une vue d'ensemble du champ, nous permettant d'identifier les zones où les mauvaises herbes peuvent être présentes et de prendre des mesures appropriées pour les éliminer afin de préserver la santé et la croissance des cultures de riz.

Les images collectées par notre système de surveillance, notamment grâce à la Caméra Meraki MV63, jouent un rôle clé dans notre étude de surveillance du champ de riz. Elles fournissent des données visuelles riches et précises, permettant une analyse approfondie de l'état des cultures et de l'environnement du champ. Grâce à ces images, nous pouvons surveiller en temps réel l'évolution des cultures, détecter les menaces potentielles, et prendre des décisions proactives pour optimiser les pratiques agricoles et maximiser les rendements dans le champ au Tekane, Rosso, Mauritanie. La Figure 6.4 illustre quelques exemples d'images capturées par la Caméra Meraki MV63, mettant en évidence l'importance de ces données visuelles pour notre étude.

6.2.2 Prediction du NDVI

Dans notre étude, nous avons utilisé un modèle de deep learning pour prédire le NDVI en se basant sur une combinaison de facteurs climatiques et d'indices de végétation. Cependant, notre ensemble de données était relativement restreint car nous n'avions pas encore collecté une grande quantité de données à partir de notre système. Ce modèle a été préalablement entraîné dans le quatrième chapitre de notre travail.

Pour contourner ce problème, nous avons divisé nos données en deux ensembles : un ensemble d'entraînement, représentant 80 % des données sur une période de deux mois, et un ensemble de test, représentant les 20 % restants.

Nous avons étudié l'évolution du NDVI sur une période de 5 jours, allant du 31 mai au

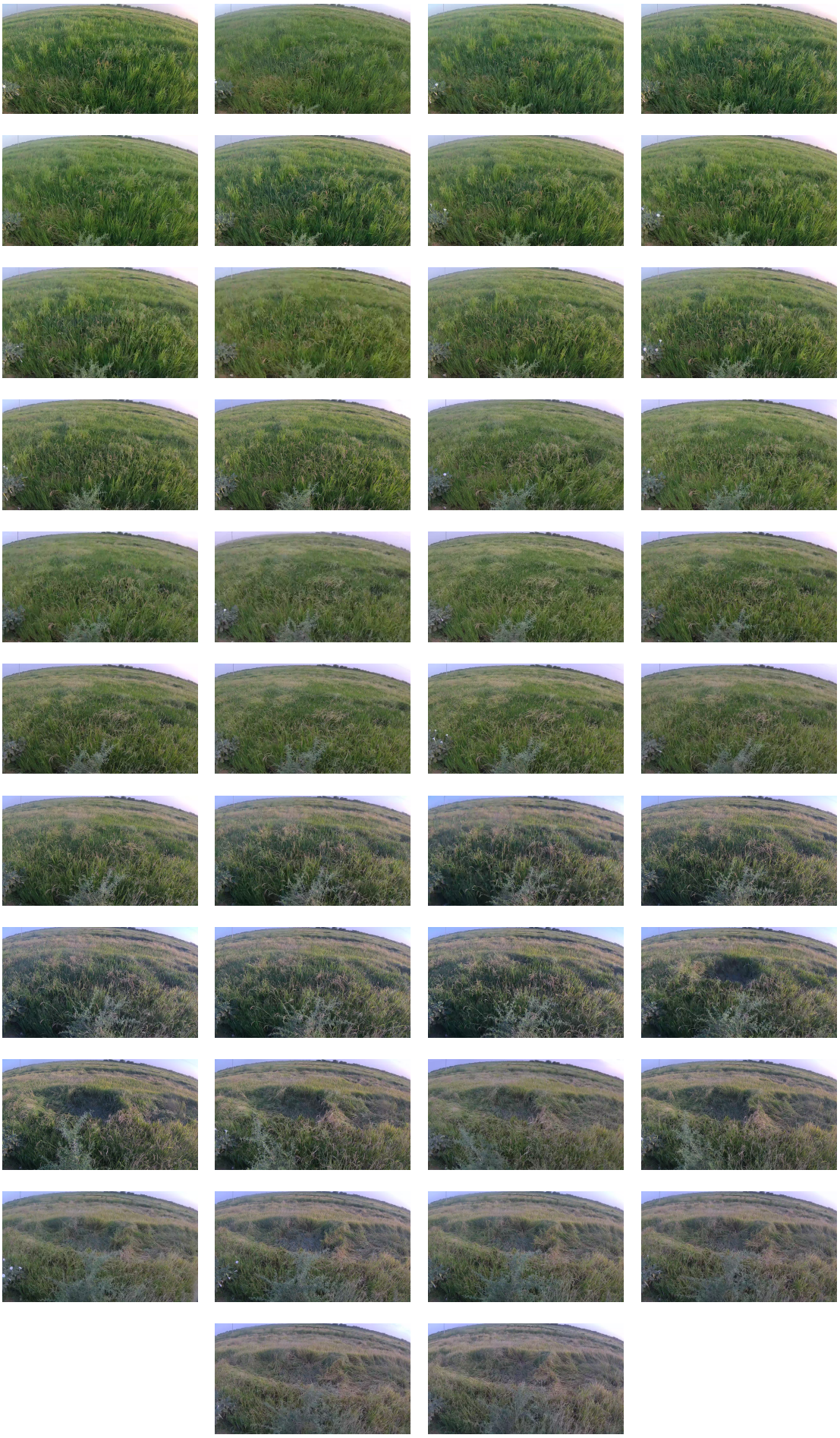


FIGURE 6.4 – Ensemble d'images du mois de mai et juin 2023

4 juin. Le premier jour dans l'ensemble de test correspondait au dernier jour où les plantes présentaient une couleur verte intense. Ensuite, le NDVI a progressivement diminué, traduisant le déclin de la couleur verte dans les plantes. Toutefois, le NDVI est resté élevé en raison de la date de récolte prévue pour ce champ, qui était fixée au 24 juin.

Nous avons réussi à prédire avec succès le NDVI pour les jours du 31 mai et du 1er juin jusqu'au 4 juin. Ces prédictions ont bien reflété le scénario de changement de phase dans la croissance des plantes dans le champ.

Malgré le manque de données initiales, nos résultats démontrent que notre modèle de deep learning a été capable de fournir des prédictions significatives concernant l'évolution du NDVI au fil du temps, contribuant ainsi à mieux comprendre le processus de croissance des plantes dans le champ étudié (Figure 6.5).

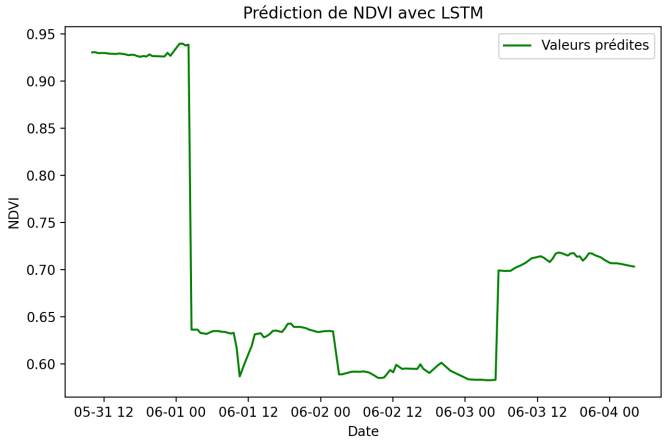


FIGURE 6.5 – Prédiction de l'indice de végétation NDVI

Notre étude précédente a mis en évidence une corrélation positive entre le NDVI et le rendement du riz. Cette découverte suggère que le NDVI peut être utilisé comme un indicateur du rendement. Selon les déclarations de la propriétaire du champ, le rendement réel était d'environ 4000 kg/hectare, ce qui confirme que notre estimation était proche de la réalité.

Afin de renforcer la fiabilité des résultats obtenus à partir des données satellitaires, nous prévoyons de créer un nouvel ensemble de données spécifiquement axé sur le rendement de cette zone. Ce nouveau jeu de données sera utilisé à la fois pour entraîner et tester notre modèle de prédiction. L'objectif est de pouvoir évaluer la précision de nos prédictions par

rapport aux valeurs réelles du rendement du riz dans la région étudiée.

En utilisant ce jeu de données spécifique au rendement, nous pourrons mieux évaluer la performance de notre modèle de prédiction et ainsi confirmer la robustesse de l'utilisation du NDVI comme substitut fiable pour estimer le rendement des cultures à l'aide de données satellitaires.

Conclusion Générale

La contribution de cette thèse couvre plusieurs axes majeurs de l'analyse des séries temporelles, chacun apportant une avancée significative dans ce domaine en pleine expansion.

Dans un premier axe, nous avons approfondi l'étude du clustering des données de séries temporelles, mettant l'accent sur les défis posés par les données massives et multivariées. Notre approche a conduit à l'amélioration du modèle DBSCAN grâce à l'intégration d'un Deep Autoencoder. Cette innovation a permis une réduction efficace de la dimension des données massives. De plus, nous avons introduit une mesure de similarité basée sur le Dynamic Time Warping (DTW) ainsi qu'une adaptabilité de l'epsilon dans le DBSCAN. Ces ajustements ont considérablement renforcé notre capacité à interpréter visuellement les résultats du clustering, ouvrant ainsi de nouvelles perspectives dans l'exploration de données massives et complexes.

Dans un second axe, notre contribution s'est articulée autour de l'optimisation d'un modèle de prédiction basé sur les réseaux LSTM (Long Short-Term Memory) pour les séries temporelles multivariées. Une comparaison minutieuse entre les approches statistiques et dynamiques a été entreprise pour démontrer la supériorité de notre modèle optimisé. Ce dernier s'est révélé capable de réaliser des prédictions d'une grande précision, se rapprochant sensiblement de la réalité. Cette avancée offre des perspectives considérables pour la modélisation et la prédiction des séries temporelles complexes, avec des applications potentielles dans divers domaines.

Un troisième axe novateur de notre recherche concerne la conception et la mise en œuvre

de "digital twins" pour les champs agricoles. Cette approche visionnaire s'est attaquée aux caractéristiques climatiques particulières, marquées par la sécheresse, le manque de précipitations, les températures élevées et les contraintes économiques des propriétaires, le tout dans un contexte géographique varié. Ce système de "jumeaux numériques" se distingue par sa capacité à évoluer et à être reproduit, ouvrant la voie à des applications dans d'autres régions et exploitations agricoles. De plus, son utilisation d'énergies renouvelables s'inscrit dans une démarche de durabilité environnementale.

Dans un quatrième axe, notre thèse a contribué à l'intégration harmonieuse de modèles de prédiction et de clustering dans les systèmes de monitoring. Cette synergie a permis d'optimiser la gestion des données collectées, ouvrant ainsi la voie à une utilisation plus efficiente des informations générées par ces systèmes.

Un cinquième axe majeur de notre travail a porté sur l'exploitation des dernières avancées en intelligence artificielle pour mesurer les données climatiques, en particulier celles générées par les dispositifs Meraki Cisco. Cette démarche novatrice ouvre de nouvelles perspectives pour la surveillance environnementale et la compréhension des phénomènes climatiques complexes.

Enfin, un sixième axe de notre recherche s'est concrétisé par l'élaboration d'une interface de monitoring utilisateur personnalisée. Cette interface offre la possibilité de visualiser en temps réel les résultats des analyses des données collectées par notre système de monitoring, améliorant ainsi la convivialité et l'utilité de notre approche.

En ce qui concerne les perspectives futures, elles varient en fonction de chaque axe de notre contribution. Nous envisageons d'améliorer encore davantage notre modèle de clustering D-DBSCAN en introduisant une fonction de perte lors de l'affectation des clusters. De plus, nous avons l'intention d'enrichir notre système en ajoutant de nouveaux capteurs pour mesurer un éventail élargi de facteurs, notamment l'état des sols avec un niveau de détail accru. Nous prévoyons également d'explorer d'autres technologies de transmission, notamment LORA, afin de réduire notre dépendance aux réseaux GSM.

Bibliographie

Kawcher Ahmed, Tasmia Rahman Shahidi, Syed Md Irfanul Alam, and Sifat Momen. Rice leaf disease detection using machine learning techniques. In *2019 International Conference on Sustainable Technologies for Industry 4.0 (STI)*, pages 1–5. IEEE, 2019.

Nurkhodzha Akbulaev and Imangulu Muradzada. Econometric analysis of foreign trade and economic growth relationship in azerbaijan and italy. *Multidisciplinary Science Journal*, 6 (1) :2024002–2024002, 2024.

Ali Alqahtani, Mohammed Ali, Xianghua Xie, and Mark W Jones. Deep time-series clustering : A review. *Electronics*, 10(23) :3001, 2021.

Rafael Gomes Alves, Gilberto Souza, Rodrigo Filev Maia, Anh Lan Ho Tran, Carlos Kamienski, Juha-Pekka Soininen, Plinio Thomaz Aquino, and Fabio Lima. A digital twin for smart farming. In *2019 IEEE Global Humanitarian Technology Conference (GHTC)*, pages 1–4. IEEE, 2019.

Guilherme Andrade, Gabriel Ramos, Daniel Madeira, Rafael Sachetto, Renato Ferreira, and Leonardo Rocha. G-dbscan : A gpu accelerated algorithm for density-based clustering. *Procedia Computer Science*, 18 :369–378, 2013.

Pelin Angin, Mohammad Hossein Anisi, Furkan Göksel, Ceren Gürsoy, and Asaf Büyükgülcü. Agrilora : a digital twin framework for smart agriculture. *J. Wirel. Mob. Networks Ubiquitous Comput. Dependable Appl.*, 11(4) :77–96, 2020.

Manolya Atalay. An overview of digital twin applications on smart agriculture. In *2023 International Balkan Conference on Communications and Networking (BalkanCom)*, pages 1–5. IEEE, 2023.

Ria Biswas. Study on arima modelling to forecast area and production of kharif rice in west bengal. *Cutting-edge Research in Agricultural Sciences*, 12 :142–151, 2021.

Carmelo Cassisi, Alfredo Ferro, Rosalba Giugno, Giuseppe Pigola, and Alfredo Pulvirenti. Enhancing density-based clustering : Parameter reduction and outlier detection. *Information Systems*, 38(3) :317–330, 2013.

Stefano Cesco, Paolo Sambo, Maurizio Borin, Bruno Basso, Guido Orzes, and Fabrizio Mazzetto. Smart agriculture and digital twins : Applications and challenges in a vision of sustainability. *European Journal of Agronomy*, 146 :126809, 2023.

Hao Chen, Chaoyang Fang, and Xin Xiao. Visualization of environmental sensing data in the lake-oriented digital twin world : Poyang lake as an example. *Remote Sensing*, 15(5) :1193, 2023.

Jianguo Chen and S Yu Philip. A domain adaptive density clustering algorithm for data with varying density distribution. *IEEE Transactions on Knowledge and Data Engineering*, 33(6) : 2310–2321 (Algorithm 3.1 , page 5), 2019.

Changlock Choi and Seong-Yun Hong. Mdst-dbscan : A density-based clustering method for multidimensional spatiotemporal data. *ISPRS International Journal of Geo-Information*, 10 (6) :391, 2021.

Stiphen Chowdhury and Renato Cordeiro de Amorim. An efficient density-based clustering algorithm using reverse nearest neighbour. In *Intelligent Computing-Proceedings of the Computing Conference*, pages 29–42. Springer, 2019.

Cherine Fathy and Hassan M Ali. A secure iot-based irrigation system for precision agriculture using the expeditious cipher. *Sensors*, 23(4) :2091, 2023.

Laura Foster, Katie Szilagyi, Angeline Wairegi, Chidi Oguamanam, and Jeremy de Beer. Smart farming and artificial intelligence in east africa : Addressing indigeneity, plants, and gender. *Smart Agricultural Technology*, 3 :100132, 2023.

Adam Ghandar, Ayyaz Ahmed, Shahid Zulfiqar, Zhengchang Hua, Masatoshi Hanai, and Georgios Theodoropoulos. A decision support system for urban agriculture using digital twin : A case study with aquaponics. *Ieee Access*, 9 :35691–35708, 2021.

Markus Götz, Christian Bodenstein, and Morris Riedel. Hpdbscan : highly parallel dbscan. In *Proceedings of the workshop on machine learning in high-performance computing environments*, pages 1–10, 2015.

Yahui Guo, Yongshuo Fu, Fanghua Hao, Xuan Zhang, Wenxiang Wu, Xiuliang Jin, Christopher Robin Bryant, and J Senthilnath. Integrated phenology and climate in rice yields prediction using machine learning methods. *Ecological Indicators*, 120 :106935, 2021.

Daniel Anthony Howard, Zheng Ma, Jesper Mazanti Aaslyng, and Bo Nørregaard Jørgensen. Data architecture for digital twin of commercial greenhouse production. In *2020 RIVF International Conference on Computing and Communication Technologies (RIVF)*, pages 1–7. IEEE, 2020.

Tongxi Hu, Xuesong Zhang, Gil Bohrer, Yanlan Liu, Yuyu Zhou, Jay Martin, Yang Li, and Kaiguang Zhao. Crop yield prediction via explainable ai and interpretable machine learning : Dangers of black box models for evaluating climate change impacts on crop yield. *Agricultural and Forest Meteorology*, 336 :109458, 2023.

Gaganpreet Singh Hundal, Chad Matthew Laux, Dennis Buckmaster, Mathias J Sutton, and Michael Langemeier. Exploring barriers to the adoption of internet of things-based precision agriculture practices. *Agriculture*, 13(1) :163, 2023.

Md Didarul Islam, Liping Di, Faisal Mueen Qamer, Sravan Shrestha, Liying Guo, Li Lin, Timothy J Mayer, and Aparna R Phalke. Rapid rice yield estimation using integrated remote sensing and meteorological data and machine learning. *Remote Sensing*, 15(9) :2374, 2023.

Taohidul Islam, Manish Sah, Sudipto Baral, and Rudra Roy Choudhury. A faster technique on rice disease detectionusing image processing of affected area in agro-field. In *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, pages 62–66. IEEE, 2018.

Jennifer Jang and Heinrich Jiang. Dbscan++ : Towards fast and scalable density clustering. In *International conference on machine learning*, pages 3019–3029. PMLR, 2019.

Melanie Jans-Singh, Kathryn Leeming, Ruchi Choudhary, and Mark Girolami. Digital twin of an urban-integrated hydroponic farm. *Data-Centric Engineering*, 1 :e20, 2020.

Mohd Javaid, Abid Haleem, Ibrahim Haleem Khan, and Rajiv Suman. Understanding the potential applications of artificial intelligence in agriculture sector. *Advanced Agrochem*, 2 (1) :15–30, 2023.

Heinrich Jiang, Jennifer Jang, and Jakub Lacki. Faster dbscan via subsampled similarity queries. *Advances in Neural Information Processing Systems*, 33 :22407–22419, 2020.

Eli Adama Jiya, Umar Illiyasu, and Mudashiru Akinyemi. Rice yield forecasting : A comparative analysis of multiple machine learning algorithms. *Journal of Information Systems and Informatics*, 5(2) :785–799, 2023.

Amrita A Joshi and BD Jadhav. Monitoring and controlling rice diseases using image processing techniques. In *2016 International Conference on Computing, Analytics and Security Trends (CAST)*, pages 471–476. IEEE, 2016.

A Kampker, V Stich, P Jussen, B Moser, and J Kuntz. Business models for industrial smart services–the example of a digital twin for a product-service-system for potato harvesting. *Procedia Cirp*, 83 :534–540, 2019.

K Mahesh Kumar and A Rama Mohan Reddy. A fast dbscan clustering algorithm by accelerating neighbor searching using groups method. *Pattern Recognition*, 58 :39–48, 2016.

Takio Kurita. Principal component analysis (pca). *Computer Vision : A Reference Guide*, pages 1–4, 2019.

Xin Liu, Du Jiang, Bo Tao, Feng Xiang, Guozhang Jiang, Ying Sun, Jianyi Kong, and Gongfa Li. A systematic review of digital twin about physical entities, virtual models, twin data, and applications. *Advanced Engineering Informatics*, 55 :101876, 2023.

Mirjana Maksimović. A faster path to sustainability : the use of digital twins. In *2023 22nd International Symposium INFOTEH-JAHORINA (INFOTEH)*, pages 1–6. IEEE, 2023.

Lamin R Mansaray, Kangyu Zhang, and Adam Sheka Kanu. Dry biomass estimation of paddy rice with sentinel-1a satellite data using machine learning regression algorithms. *Computers and Electronics in Agriculture*, 176 :105674, 2020.

Tsega Y Melesse, Chiara Franciosi, Valentina Di Pasquale, and Stefano Riemma. Analyzing the implementation of digital twins in the agri-food supply chain. *Logistics*, 7(2) :33, 2023.

Laura Melgar-García, David Gutierrez-Aviles, Maria Teresa Godinho, Rita Espada, Isabel Sofia Brito, Francisco Martínez-Álvarez, Alicia Troncoso, and Cristina Rubio-Escudero. A new big data triclustering approach for extracting three-dimensional patterns in precision agriculture. *Neurocomputing*, 500 :268–278, 2022.

Devika Menon, Bharath Anand, and Chiranji Lal Chowdhary. Digital twin : Exploring the intersection of virtual and physical worlds. *IEEE Access*, 2023.

Tamás Mizik. How can precision farming work on a small scale? a systematic literature review. *Precision agriculture*, 24(1) :384–406, 2023.

Peyman Moghadam, Thomas Lowe, and Everard J Edwards. Digital twin for the future of orchard production systems. *Multidisciplinary Digital Publishing Institute Proceedings*, 36 (1) :92, 2020.

Fazal E Nasir, Muhammad Tufail, Muhammad Haris, Jamshed Iqbal, Said Khan, and Muhammad Tahir Khan. Precision agricultural robotic sprayer with real-time tobacco recognition and spraying system based on deep learning. *Plos one*, 18(3) :e0283801, 2023.

Suresh Neethirajan and Bas Kemp. Digital twins in livestock farming. *Animals*, 11(4) :1008, 2021.

Sjoukje A Osinga, Dilli Paudel, Spiros A Mouzakitis, and Ioannis N Athanasiadis. Big data in agriculture : Between opportunity and solution. *Agricultural Systems*, 195 :103298, 2022.

HB Prajapati, JP Shah, and VK Dabhi. Detection and classification of rice plant diseases. intell. deci. technol. 11 (3), 357–373 (2017), 2017.

Warren Purcell, Thomas Neubauer, and Kevin Mallinger. Digital twins in agriculture : Challenges and opportunities for environmental sustainability. *Current Opinion in Environmental Sustainability*, 61 :101252, 2023.

Christos Pylianidis, Sjoukje Osinga, and Ioannis N Athanasiadis. Introducing digital twins to agriculture. *Computers and Electronics in Agriculture*, 184 :105942, 2021.

Nadia Rahmah and Imas Sukaesih Sitanggang. Determination of optimal epsilon (eps) value on dbscan algorithm to clustering data on peatland hotspots in sumatra. In *IOP conference series : earth and environmental science*, volume 31, page 012012. IOP Publishing, 2016.

Arpit Rawankar, Mayurkumar Nanda, Hemant Jadhav, Prem Lotekar, Rahul Pawar, Libin Sibichan, and Akshay Pangare. Detection of n, p, k fertilizers in agricultural soil with nir laser absorption technique. In *2018 3rd International Conference on Microwave and Photonics (ICMAP)*, pages 1–2. IEEE, 2018.

P Chandra Shaker Reddy and A Sureshbabu. An applied time series forecasting model for yield prediction of agricultural crop. In *Soft Computing and Signal Processing : Proceedings of 2nd ICSCSP 2019 2*, pages 177–187. Springer, 2020.

Mario San Emeterio de la Parte, José-Fernán Martínez-Ortega, Vicente Hernández Díaz, and Néstor Lucas Martínez. Big data and precision agriculture : a novel spatio-temporal semantic iot data management framework for improved interoperability. *Journal of Big Data*, 10(1) :1–32, 2023.

Jörg Sander, Martin Ester, Hans-Peter Kriegel, and Xiaowei Xu. Density-based clustering in spatial databases : The algorithm gdbscan and its applications. *Data mining and knowledge discovery*, 2(2) :169–194, 1998.

T Saranya, C Deisy, S Sridevi, and Kalaiarasi Sonai Muthu Anbananthen. A comparative study of deep learning and internet of things for precision agriculture. *Engineering Applications of Artificial Intelligence*, 122 :106034, 2023.

Anurag Satpathi, Parul Setiya, Bappa Das, Ajeet Singh Nain, Prakash Kumar Jha, Surendra Singh, and Shikha Singh. Comparative analysis of statistical and machine learning techniques for rice yield forecasting for chhattisgarh, india. *Sustainability*, 15(3) :2786, 2023.

Murali Krishna Senapaty, Abhishek Ray, and Neelamadhab Padhy. Iot-enabled soil nutrient analysis and crop recommendation model for precision agriculture. *Computers*, 12(3) :61, 2023.

Shivangi Sharma, Kirti Verma, and Palak Hardaha. Implementation of artificial intelligence in agriculture. *Journal of Computational and Cognitive Engineering*, 2(2) :155–162, 2023.

Vimal K Shrivastava and Monoj K Pradhan. Rice plant disease classification using color features : a machine learning paradigm. *Journal of Plant Pathology*, 103 :17–26, 2021.

Letícia Silva, Francisco Rodríguez-Sedano, Paula Baptista, and João Paulo Coelho. The digital twin paradigm applied to soil quality assessment : A systematic literature review. *Sensors*, 23(2) :1007, 2023.

Gaurvendra Singh, R Rajesh, Yash Daultani, and Subhas Chandra Misra. Resilience and sustai-

nability enhancements in food supply chains using digital twin technology : A grey causal modelling (gcm) approach. *Computers & Industrial Engineering*, 179 :109172, 2023.

Maulshree Singh, Rupal Srivastava, Evert Fuenmayor, Vladimir Kuts, Yuansong Qiao, Niall Murray, and Declan Devine. Applications of digital twin across industries : a review. *Applied Sciences*, 12(11) :5727, 2022.

Petr Skobelev, Vladimir Laryukhin, Elena Simonova, Oleg Goryanin, Vladimir Yalovenko, and Olga Yalovenko. Multi-agent approach for developing a digital twin of wheat. In *2020 IEEE International Conference on Smart Computing (SMARTCOMP)*, pages 268–273. IEEE, 2020.

Naftali Slob, William Hurst, Rick Van de Zedde, and Bedir Tekinerdogan. Virtual reality-based digital twins for greenhouses : A focus on human interaction. *Computers and Electronics in Agriculture*, 208 :107815, 2023.

G Tagliavini, Thijs Defraeye, and Jan Carmeliet. Multiphysics modeling of convective cooling of non-spherical, multi-material fruit to unveil its quality evolution throughout the cold chain. *Food and Bioproducts Processing*, 117 :310–320, 2019.

Naoum Tsolakis, Dimitrios Bechtsis, and Dionysis Bochtis. Agros : A robot operating system based emulation tool for agricultural robotics. *Agronomy*, 9(7) :403, 2019.

Shaoyuan Weng, Jin Gou, and Zongwen Fan. h-dbscan : A simple fast dbscan algorithm for big data. In *Asian Conference on Machine Learning*, pages 81–96. PMLR, 2021.

Yi-Pu Wu, Jin-Jiang Guo, and Xue-Jie Zhang. A linear dbscan algorithm based on lsh. In *2007 International Conference on Machine Learning and Cybernetics*, volume 5, pages 2608–2614. IEEE, 2007.

Olga Yanenko and An der Weberei. Introducing social distance to st-dbscan. In *Proceedings of the 22nd AGILE Conference*, 2019.

Jieping Ye, Ravi Janardan, and Qi Li. Two-dimensional linear discriminant analysis. *Advances in neural information processing systems*, 17, 2004.

Lejun Yu, Jiawei Shi, Chenglong Huang, Lingfeng Duan, Di Wu, Debao Fu, Changyin Wu, Lizhong Xiong, Wanneng Yang, and Qian Liu. An integrated rice panicle phenotyping method based on x-ray and rgb scanning and deep learning. *The Crop Journal*, 9(1) :42–56, 2021.

Rui Zhang, Tao Du, Shouning Qu, and Hongwei Sun. Adaptive density-based clustering algorithm with shared knn conflict game. *Information Sciences*, 565 :344–369, 2021.