République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université A. Mira de Bejaia
Faculté des Sciences Exactes
Département d'Informatique

***MEMOIRE DE MASTER PROFESSIONNEL***

En

Informatique

Option

*Administration et sécurité des réseaux*

**Thème**

# Nouvelle approche ANN pour le comblement des données manquantes observées par des cartes satellitaires TRMM

## Cas d'étude : carte pluviométrique du nord d'Algérie

Présenté par : MM. AMGHAR Amazigh & YAICHE Houssam

| | | | |
|---|---|---|---|
| Président | Dr H. EL BOUHISSI | Maître de conf.B | U.A/Mira Béjaia. |
| Promotrice | Dr Y. SAMIRA | Maître de conf.B | U. A/Mira Béjaia. |
| Co-encadreur | M A.AMIR | | |
| Examinateur | Dr A.MOULOUD | Maître de conf.B | U. A/Mira Béjaia. |

Bejaia, décembre 2020.

# Remerciements

Avant comme après tout, nous exprimons notre profonde gratitude à notre Seigneur, l'Éternel, pour nous avoir doté de la force nécessaire et de la détermination à soutenir notre persévérance au travail pour le mener à bonne fin. Nous adressons nos remerciements chaleureux aux personnes qui nous ont aidés à persévérer, en nous apportant leur aide et leurs encouragements de telle sorte que cette année universitaire soit enfin une réussite sans encombres. Nous témoignons de notre sincère gratitude à notre promotrice, Mme. YASSAD Samira, qui nous a orientés pour nous tirer d'affaire dans les moments de notre hésitation et accentués par les difficultés de cette année sans nulle pareille. Nous remercions aussi M. AIEB Amir qui, depuis l'Europe, avait pu nous inspirer notre thème et contribua beaucoup à nous déblayer nos pistes de recherche. Tout ce qui n'est pas encore disponible dans notre pays, notions et données, il a pu nous le compléter.

Nous remercions vivement Mesdames et Messieurs les membres du Jury pour avoir accepté de consulter notre étude et se prononcer sur sa recevabilité.

Nous remercions également les corps professoral de notre Faculté et notre Université ainsi que son personnel administratif qui ont rendu possibles des étapes clés de notre étude, au moment opportun, en répondant à nos requêtes avec plus de vision judicieuse que ce que nous pouvions atteindre. Outre l'aide fructueuse de nos professeurs, nous avions souvent besoin de notions qui étaient mieux traitées et connues chez d'autres facultés et ses professeurs. Ils n'ont jamais hésité à nous les expliquer en nous orientant au mieux à leur sujet pour en tirer le meilleur parti.

Certains moments ont été ardus, pour les dépasser, il n'y a rien de tel que l'expérience riche de patience et d'exemples que nous trouvions chez le personnel administratif. Par leur action soutenue se renouvellent ainsi les années de succès. Nous espérons aussi et surtout que nos chers parents se contenteront de notre travail et qu'ils acceptent notre humble gratitude pour tant de sacrifices hors de prix dont ils nous avaient entourés des décennies durant, armés de leur patience unique et leur stoïcisme à toute épreuve.

Enfin, nous adressons nos plus sincères remerciements à tous nos proches et amis, qui nous ont toujours encouragées au cours de la réalisation de ce mémoire.

# **Dédicaces**

*A mes chers parents, pour tous leurs sacrifices, leur amour, leur tendresse, leur soutien et leurs prières tout au long de mes études,*

*A mes chers sœurs Mounira, Maroua, et mes frères Bilal, Kheireddine pour leur appui et leur encouragement,*

*A toute ma famille et mes amis pour leur soutien tout au long de mon parcours universitaire,*

*Que ce travail soit l'accomplissement de vos vœux tant exprimés, et le fruit de votre soutien infaillible,*

*Merci d'être toujours là pour moi.*

***M. YAICHE Houssam.***

# **Dédicace**

*A mes chers parents, pour tous leurs sacrifices, leur amour, leur tendresse, leur soutien et leurs prières tout au long de mes études,*

*A mes chers frères, Abdelghani, Azzedine, Fares, Lamine, Djilali pour leur appui et leur encouragement,*

*A toute ma famille pour son soutien tout au long de mon parcours universitaire,*

*Que ce travail soit l'accomplissement de vos vœux tant formulés, le fruit de votre soutien infaillible,*

*merci d'être toujours là pour moi.*

***M. AMGHAR Amazigh***.

# Table des matières

## Chapitre 4 : validation et comparaisons

## Table des figures

## Liste des tableaux

## Liste des abréviations

A

ACP : Analyse Composantes Principale.

AI : Algorithme Itératif.

ANN : Artificiel Neural Network.

C

CERES : Clouds and the Earth's Radiant Energy System.

MCMC : Markovo Chaine Monté Carlo.

D

AD : Augmentation de Données.

DE : donnée estimée.

DM : Données Manquantes.

E

ERBE : Earth Radiation Budget Experiment.

F

FCS: Fully Conditional Spécification.

I

IDW : Inverce Distance Weighted

J

JAXA : Japan Aerospace Exploration Agency.

K

KDR : known Data Regression.

KDR-PCR : KDR avec régression en composante principale.

KDR-PLS : KDR avec de moindres carrés partiels.

L

LIS : Lightning Imaging Sensor.

M

MAE : Mean Absolute Error.

MAR : manquant au hasard.

MCAR : manquant complètement au hasard.

MDI : Missing Data Imputation.

MDI : Boite à outils Matlab.

N

NASA : National Aeronautics and Space Administration.

NIPALS : algorithme de régression des moindres carrés partiel itératif non linéaire.

NMAR : ne manquant pas au hasard.

P

PMP : Projection to the Model Plane.

PR : Precipitation Radar.

PCA : principal analysis composant.

Q

QQ plot : Quantile-quantile plot.

R

RMSE : Root Mean Square Error.

S

SPSS : Statistique Package for Social Sciences.

SIG : Système Information Géographique.

T

TMI : TRMM Microwave Imager.

TOA: top of the Earth's atmosphere.

TSR : Trimmed Scores Regression.

TRMM : Tropical Rainfall Measurement Mission.

V

VIRS : Visible and Infrared Scanner

W

WKNNI : Weighted K-Nearset Neigbhors Imputation.

# Introduction générale

## Introduction générale

La télédétection est, de nos jours, une des méthodes les plus importantes utilisées pour acquérir rapidement et directement des informations sur la surface de la terre. À la faveur du développement de la science de l'information environnementale, ces dernières années, les données de télédétection jouent un rôle déterminant dans de nombreux domaines de recherche. Or, à cause de la grande masse d'informations (big data), les images de télédétection ne s'en sortent pas toujours sans quelques souffrances, problèmes courants, où certaines parties se trouvent contaminées par le bruit de bande/pixels défectueux. Ils sont causés principalement par la non-réponse du capteur, tout comme ils peuvent être dus au décalage des capteurs. Les valeurs manquantes empêchées par la couverture nuageuse et les problèmes spécifiques aux capteurs en sont une autre raison. Les problèmes des données manquantes restent très préoccupants dans les cartes satellitaires. La mission de mesure des précipitations tropicales (TRMM), lancée en 1997, a entraîné des améliorations significatives en matière de gestion des pluies dans le monde. Ce satellite a été développé comme un projet conjoint entre le Japon et les États-Unis. Il est la première mission spatiale dédiée à la mesure de la chute des pluies. TRMM observera principalement la structure, le taux et la distribution des pluies dans les régions tropicales et subtropicales. Les données devraient jouer un rôle-clé aidant à comprendre les mécanismes du changement climatique mondial et surveiller les variations environnementales. Les observations par TRMM restent cependant parcellaires, limitées, souffrant souvent de carence de données, entre autres des pixels morts et des remparts des couches nuageuses ainsi que des chutes de neige etc. Dans la littérature, il existe plusieurs approches pour le comblement des données manquantes. Elles sont classées en trois types : méthode spatiale, temporelle et spectrale. À cet égard, il y a deux types d'imputations à distinguer: l'imputation simple et l'imputation multiple. L'imputation simple est celle dans laquelle la valeur manquante est remplacée par la moyenne de toutes les observations mesurées dans la même série. L'imputation multiple est un traitement utilisant plusieurs bases de données similaires. Il donnera de nombreux remplacements pour la même valeur non observée. Enfin, l'inférence ultérieure obtenue en combinant toutes les valeurs imputées. Ainsi, l'interpolation des pixels manquants de couleur est une opération importante dans le traitement d'image qui peut être utilisée pour ré-échantillonner l'image, que ce soit pour diminuer ou en augmenter la résolution. Plusieurs algorithmes d'interpolation couramment utilisés ont été suggérés, tels que

# Introduction générale

*kregeage*, voisins naturels: le *krigeage* peut produire des prédictions de valeurs non observées à partir d'observations de sa valeur à des emplacements mitoyens. Il confère des poids pour chaque point en fonction de sa distance par rapport à la valeur inconnue. En fait, ces prédictions sont traitées comme des combinaisons linéaires pondérées des valeurs connues. La méthode du voisin naturel propose une mesure pour le calcul des poids et la sélection des voisins interpolés. La solution que nous préconisions dans ce mémoire est basée sur une approche permettant de combler la perte ou le manque de données dans les cartes TRMM. Notre approche reprend les caractéristiques de l'approche existante, MDI et KNNI, pour étayer son efficacité expérimentale. La solution que nous proposons, ANN (Réseaux Neurones Artificiels), se compose de trois couches. 1. Couche d'entrée, l'information d'accès est une image TRMM avec données manquantes. À cet effet, nous produirons sous Matlab un algorithme qui permettra d'extraire les matrices RGB. 2. En utilisant des commandes Matlab qui agréent, nous transformerons ensuite les matrices RGB sous forme d'un seul tableau de colonnes. Dans la couche cachée, nous appliquerons des fonctions de sommation et de transfert avec KNNI et la régression comme méthode pour estimer les valeurs manquantes des matrices décimales RGB. 3. Dans la couche de sortie, nous utiliserons, toujours sous Matlab, un algorithme qui permettra de concaténer les matrices de pixels RGB afin de créer une nouvelle image TRMM traitée. Des tests statistiques descriptifs quantitatifs et qualitatifs seront appliqués, tels que RMSE, MAE, etc., qui serviront à étudier la performance de notre modèle proposé (ANN) selon les différents mécanismes de données manquantes (MAR, MCAR et NMAR), de même que pour en arrêter exactement le pourcentage de données (15% et 30%) afin d'étudier les tendances des erreurs.

Ce mémoire est organisé en quatre chapitres articulés dans l'ordre suivant :

- Le premier chapitre porte sur un aperçu général des cartes satellitaires, retraçant notamment l'utilité des différentes cartes TRMM. Dans sa deuxième partie, le chapitre aborde quelques facteurs fauteurs de manques de données sur les cartes satellitaires TRMM ;
- Le deuxième chapitre esquisse une vue générale, descriptive du traitement des absences de données chronologiques. Les ressources matérielles et logicielles concernant le traitement des données manquantes contribueront à définir les méthodes dédiées au traitement des données manquantes, méthode spatiale, temporelle et spectrale. La deuxième partie chapitre passe en revue quelques-unes des méthodes de comblement,

nous nous y étalerons un moment sur la description des méthodes d'imputation et d'interpolation ;

- Le troisième chapitre consiste à donner une vue de notre modèle proposé (ANN) utilisant les deux méthodes de comblement, KNNI et la régression, ainsi qu'une étude descriptive détaillée de cette dernière pour le comblement des manques de données dans l'imagerie TRMM, avec un exemple d'application, indiquant en passant les environnements de simulation ;
- Le quatrième chapitre traite d'une étude comparative entre le modèle proposé et les toolbox MDI et KNNI. Il comporte aussi une partie expérimentale, des tests de validation de notre modèle et, enfin, la conclusion.

Le mémoire se termine par une conclusion générale, ouvrant sur des perspectives de recherche d'approfondissement futures qui prendront le relai de notre étude.

# Chapitre 1

# Chapitre 1 : généralités et informations sur les cartes TRMM

## 1.1. Introduction

Les cartes sont un atout crucial pour communiquer la science du climat à un public diversifié, plusieurs satellite sont fréquemment utilisés pour surveilliez, stocker, traiter et visualiser les données dans l'imagerie climatiques. Dans ce chapitre, nous introduirons notre sujet sur les cartes. Nous présenterons information digitale utilité et des cartes et ça nécessité dans les études en big data, aussi nous présenterons les différentes cartes satellitaire TRMM (TMI, PR, VIRS…..etc.), en donnant ainsi une vue sur les facteurs les plus connus qui causent les données manquantes dans l'imagerie.

## 1.2. L'information digitale par cartes et sa nécessité dans l'étude en big data

De nos jours, la plupart des données de télédétection étant enregistrées en format numérique, presque toutes les interprétations et analyses d'images requièrent d'une façon ou d'une autre le traitement numérique. Celui-ci, peut recourir lors du traitement des images à divers procédés dont le formatage et la correction des données, le rehaussement numérique pour faciliter l'interprétation visuelle ou même la classification automatique des cibles et des structures entièrement sur ordinateur.

Le traitement numérique de l'imagerie de télédétection exige que les données soient enregistrées et disponibles dans un format numérique convenable pour l'entreposage sur disques ou cassettes informatiques.

Le traitement d'images numériques nécessite évidemment un système informatique (système d'analyse d'images) ainsi que l'équipement et les logiciels propres au traitement de telles données. Plusieurs systèmes de logiciels commerciaux ont été développés spécifiquement pour le traitement et l'analyse des images de télédétection.

Pour les besoins, nous regrouperons les fonctions de traitement des images communément disponibles en analyse d'images en quatre catégories [W1] :

### 1.2.1. Prétraitement

Lorsqu'une image est acquise par un capteur, elle contient fréquemment des erreurs géométriques et radiométriques, erreurs à corriger avec une précision dépendant du type d'application. Les opérations de prétraitement peuvent être divisées en corrections radiométriques et en corrections géométriques.

#### 1.2.1.1. Corrections radiométriques

Elles comprennent, entre autres, la correction des données déformées par les irrégularités du capteur, de ses bruits ou ceux de l'atmosphère, de la conversion des données, afin qu'elles puissent représenter précisément le rayonnement réfléchi ou émis par le capteur.

#### 1.2.1.2. Corrections géométriques

Elles comprennent la correction des distorsions géométriques dues aux variations de la géométrie Terre-capteur, de telle sorte que la transformation des données aboutisse sur les vraies coordonnées (ex. en latitude et longitude) sur la surface terrestre.

### 1.2.2. Fonctions de rehaussement

Leur but est d'améliorer l'apparence de l'imagerie pour aider l'interprétation et l'analyse visuelle. Les fonctions de rehaussement permettent l'étirement des contrastes pour augmenter la distinction des tons entre différents éléments d'une scène, ainsi que le filtrage spatial pour rehausser (ou éliminer) les patrons spatiaux spécifiques sur une image.

### 1.2.3. Transformations d'images

Opération similaire au rehaussement d'image, à ceci près qu'elle s'en distingue par la multiplicité d'application sur plusieurs bandes de données simultanément, combinant le traitement de plusieurs bandes spectrales. Des opérations arithmétiques (addition, soustraction, multiplication, division) sont faites pour combiner et transformer les bandes originales en de « nouvelles » images qui montrent plus clairement certains éléments de la scène. Nous aurons à examiner certaines de ces opérations incluant diverses méthodes de rapport de bande aussi appelé rapport spectral et un procédé appelé analyse des composantes principales utilisée pour mieux représenter l'information en imagerie multi-spectrale.

### 1.2.4. Classification et analyse d'images

Elles sont utilisées pour identifier et classifier numériquement des pixels sur une image. La classification est habituellement faite sur des banques de données multi-spectrales, et ce procédé donne à chaque pixel d'une image une certaine classe basée sur les caractéristiques statistiques de la valeur de l'intensité du pixel. Il existe une variété d'approches adoptées pour faire une classification numérique. Nous allons brièvement décrire les deux approches générales qui sont souvent utilisées [1]: la classification supervisée et la classification non supervisée.

#### 1.2.4.1. Classification supervisée

Lors de l'utilisation d'une méthode de classification supervisée, l'analyste identifie des échantillons d'images assez homogènes, représentatifs de différents types de surfaces (classes

d'information). Ces échantillons forment un ensemble de données-tests (zones d'entraînement). La sélection de ces données-tests est basée sur les connaissances de l'analyste, sa familiarité avec les régions géographiques et les types de surfaces présentes dans l'image. L'opérateur supervise donc la classification d'un ensemble spécifique de classes. Les informations numériques pour chacune des bandes et pour chaque pixel de ces ensembles sont utilisées pour que l'ordinateur puisse définir automatiquement les classes et ensuite reconnaître automatiquement des régions aux propriétés similaires à chaque classe.

### 1.2.4.2. Classification non supervisée

La classification non supervisée procède de la façon contraire. Les classes spectrales sont formées en premier, basées seulement sur l'information numérique des données. Ces classes sont ensuite associées, par un analyste, à des classes d'information si possible utiles. Des programmes appelés algorithmes de classification sont utilisés pour déterminer les groupes statistiques naturels ou les structures des données. Habituellement, l'analyste spécifie le nombre de groupes ou classes qui seront formés avec les données. De plus, l'analyste peut spécifier certains paramètres relatifs à la distance entre les classes et la variance à l'intérieur même d'une classe. Le résultat final de ce processus de classification itératif peut créer des classes que l'analyste voudra combiner, ou des classes qui devraient être séparées de nouveau. La classification (figure 1.1.) se fait habituellement sur des banques de données multi-spectrales (A), procédé qui donne à chaque pixel d'une image une certaine classe ou thème (B) basé sur les caractéristiques statistiques de la valeur de l'intensité du pixel.

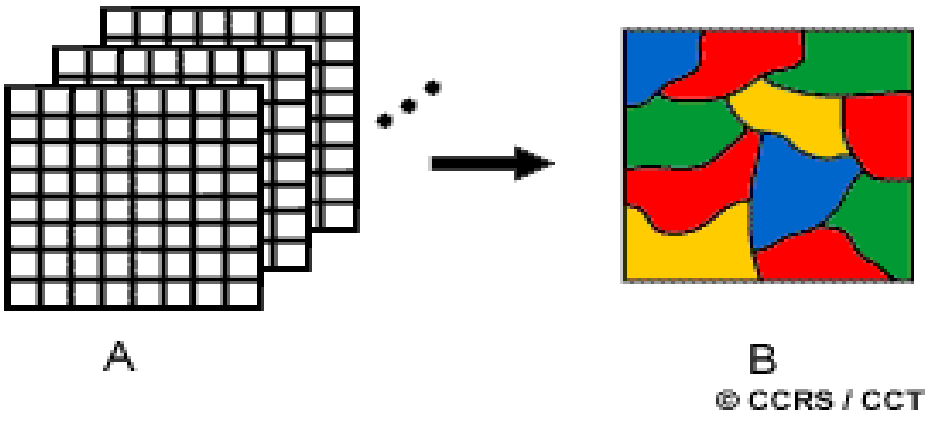


**Figure 1.1 :** La classification sur des banques de données multi-spectrales (A), procédé qui donne à chaque pixel d'une image une certaine classe ou thème (B) basé sur les caractéristiques statistiques de la valeur de l'intensité du pixel.

# Chapitre 1 : généralités et informations sur les cartes TRMM

## 1.3. Cartes météorologiques géo-localisées

Une carte météorologique est une carte sur laquelle les données de certains paramètres météorologiques sont pointées et analysées pour fournir un très grand nombre d'informations concernant l'état de l'atmosphère dans une région donnée. Sa localisation relative par rapport aux espaces voisins, ainsi que la localisation des éléments qu'il contient. Les cartes servent également à représenter des phénomènes géographiques, phénomènes dont la configuration spatiale produit du sens [W2].

Pour lire une telle carte, il faut comprendre les codes graphiques internationaux familiers aux météorologues. Par convention, chaque station d'observation météorologique est représentée par un cercle. Autour de ces cercles, des symboles indiquent température, type de précipitations, force du vent, type de nuages ou encore pression atmosphérique qui ont été mesurés à la station. Pour faciliter la lecture, les cartes comportent d'autres signes, comme les isobares, courbes reliant les points d'égale pression, ou des lettres qui indiquent les zones de haute (H) et de basse pression (L).

Les cartes météorologiques peuvent aussi indiquer les lignes le long desquelles s'affrontent différentes masses d'air ou les régions soumises à des précipitations. La lecture des cartes météorologiques est particulièrement utile aux aviateurs et aux marins, qui peuvent ainsi choisir des itinéraires loin des tempêtes.

## 1.4. Différentes cartes satellitaires TRMM

Le programme TRMM est issu de l'objectif commun à deux agences spatiales majeures, l'agence américaine NASA (National Aeronautics and Space Administration) et sa pendante japonaise, la JAXA (Japan Aerospace Exploration Agency) visant à estimer les pluies à partir de données satellitaires de manière précise entre les latitudes de 50° nord et sud, plus particulièrement dans les régions tropicales [2]. TRMM a été lancé de Tanegashima, au Japon, le 27 Novembre 1997, à 403 km d'altitude pour une fréquence de passage au-dessus d'un point de 16 fois par jour (NASA 2008).

Les principales missions scientifiques du TRMM [3] sont résumées comme suit :

- Surveiller quantitativement le taux de pluies tropicales et comprendre l'énergie et le cycle hydrologique terrestre ;

- Clarifier la condition réelle des changements temporels et spatiaux des précipitations tropicales ainsi que le mécanisme ayant effet sur la circulation atmosphérique, évaluer et développer le modèle numérique pour les reproduire et pouvoir les prédire ;
- Établir la méthode pour observer les précipitations depuis l'espace.

Ies cartes satellitaire TRMM est composé [3] de :

### 1.4.1. Radar de précipitations (PR)

Le radar de précipitation (PR) est le principal instrument à bord du TRMM [4]. Le plus innovant des cinq instruments TRMM, le PR est le premier radar de pluie (quantitatif) à voler dans l'espace. Les données PR seront essentielles pour obtenir le profil de hauteur du contenu des précipitations, à partir duquel le profil de dégagement de chaleur latente de la Terre peut être estimé. Le taux de pluie sera estimé à partir du facteur de réflectivité radar lorsque le taux de pluie est faible en appliquant les algorithmes conventionnels utilisés pour les radars au sol. Pour les fortes pluies, une correction d'atténuation de la pluie sera effectuée en utilisant l'atténuation totale du trajet des échos de surface terrestre ou marine.

Les principaux objectifs de l'instrument PR sont les suivants :

- fournir une structure pluviométrique tridimensionnelle ;
- réaliser des mesures quantitatives des taux de pluie sur terre et sur l'océan.

### 1.4.2. L’imageur micro-ondes (TMI)

Le (T-M-I) est un radiomètre micro-ondes passif multicanaux à double polarisation. Il a neuf canaux basés sur le capteur spécial micro-ondes/imager (SSM / I) [4]. L'instrument TMI fournira des données relatives aux taux de précipitations sur les océans, mais des données moins fiables sur surface terrestre, où les émissions de surface hétérogènes compliquent l'interprétation. Les données TMI combinées aux données du PR et celles du VIRS seront également utilisées pour dériver les profils de précipitations.

TMI fonctionne suivant un mode unique sans aucune redondance command-able. Il a essentiellement deux modes, OFF et ON. Deux calibreurs externes sur l'arbre stationnaire sont utilisés pour effectuer les calibrages pendant chaque rotation de l'instrument (balayage).

### 1.4.3. Scanner visible et infrarouge (VIRS)

L'instrument VIRS est un radiomètre à balayage transversal mesurant le rayonnement de la scène dans cinq bandes spectrales, fonctionnant dans le visible à travers les régions spectrales infrarouges [2]. VIRS est similaire aux instruments pilotés par d'autres satellites météorologiques de la NASA et de la NOAA. La comparaison des données d'hyperfréquence avec les données visibles et infrarouges VIRS devrait fournir les moyens par lesquels les précipitations seront estimées plus précisément que par les seules données visibles et infrarouges. L'instrument VIRS servira d'imageur de fond et fournira disposition et évolution des nuages, il autorise les observations passives par micro-ondes et radar [3]. Les données du VIRS seront reprises par les algorithmes d'estimation pluviométrique basée principalement sur les capteurs micro-ondes passifs et actifs.

### 1.4.4. Capteur de radiation terrestre (CERES)

L'expérience CERES contribuera à réduire l'une des incertitudes majeures dans la prévision des changements à long terme du climat de la Terre [5]. Les flux radiants au sommet de l'atmosphère terrestre (TOA) ont été mesurés par l'ERBE (Earth Radiation Budget Experiment), non seulement comme un champ indifférencié, mais avec une séparation raisonnable entre les flux provenant d'atmosphères claires et nuageuses. Il a également été découvert à partir d'ERBE que les nuages ont un effet plus important sur les flux TOA qu'on ne le pensait auparavant, mais les détails du processus ne sont pas encore entièrement compris. L'expérience CERES [6] tentera de fournir une meilleure compréhension de la façon dont les différents processus nuageux, tels que l'activité convective et la météorologie de la couche limite. Le CERES fournira également des informations pour déterminer le bilan de rayonnement de surface, important dans l'énergétique atmosphérique.

### 1.4.5. Capteur d'imagerie de la foudre (LIS)

Le LIS est un télescope à visée optique, un système d'imagerie à filtre permettant d'acquérir et d'étudier la distribution et la variabilité de la foudre nuage-sol sur la Terre. Les données LIS seront également utilisées avec les données PR, TMI et VIRS pour étudier la corrélation de l'incidence globale de la foudre avec les précipitations et d'autres propriétés des tempêtes [3]. Les données de l'instrument LIS peuvent être corrélées aux taux globaux, aux quantités et à la distribution des précipitations. LIS est un instrument à corde unique avec plusieurs points de défaillance uniques. LIS sera mis sous tension lors de l'activation initiale de l'instrument et restera alimenté dans cette configuration pendant toute la durée de la mission (sauf conditions

anormales échappant aux prévisions). Une représentation (figure 2.1) instruit sur le satellite TRMM et sa précision de capteurs le composant [3].

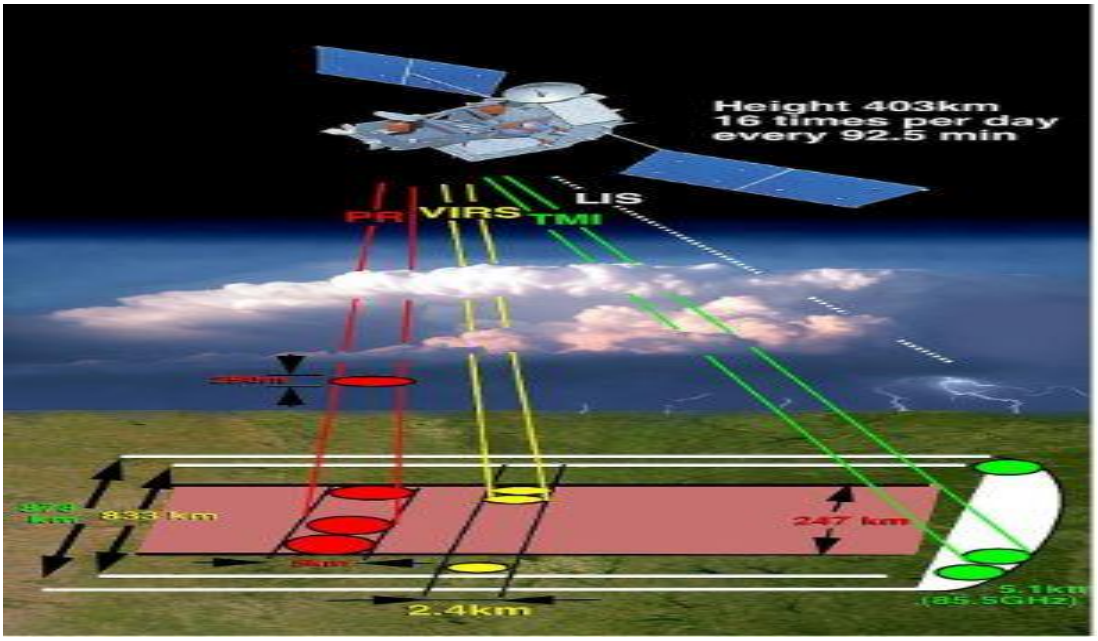


**Figure 1.1 :** représentation instruit sur le satellite TRMM et sa précision de capteurs le composant

## 1.5. Problèmes récurrents des études en big data utilisant l'information pixel pour les études climatiques et environnementales

Les problèmes de données manquantes dans les images climatiques sont nombreux et variés :

- Les satellites géostationnaires tournent à la même vitesse que celle de la planète bleue tournant sur elle-même. Les satellites défilants font le tour de notre planète généralement en un peu plus d'une heure et trente minute. Cela leur permet de survoler plusieurs points terrestres au cours d'une journée. Ils construisent ainsi des images par accumulation. Mais certains points océaniques restent occultés. La cause la plus fréquente du manque de données [W3], si ce n'est la principale, revient à l'hyper- ou infra-sensibilité des capteurs satellitaires aux conditions atmosphériques ;
- Les valeurs manquantes induites par la couverture nuageuse et les problèmes spécifiques aux capteurs en sont une autre de ces raisons limitatives de l'application d'images Landsat multi-temporelles dans la recherche environnementale. Les problèmes spécifiques aux capteurs sont

le premier facteur à l'origine des données manquantes dans Landsat. À titre d'exemple, en raison de l'échec du correcteur de ligne de balayage (SLC) en 2003, le capteur Landsat 7 Enhanced Thematic Mapper Plus (ETM +) a perdu environ 22% de pixels dans chaque sens [8]. La couverture nuageuse est le deuxième facteur entraînant absence de données, sinon rognure et troncature de l'intégrité des images Landsat. En moyenne, à tout moment, environ 35% de la surface terrestre mondiale est obscurcie par les nuages [7]. L'ampleur des données manquantes est encore plus élevée dans les analyses multi-temporelles en raison de la nature dynamique de la couverture nuageuse et des lacunes de SLC-off ;

- Les images de télédétection souffrent souvent des problèmes courants de parties contaminées par le bruit de bande/pixels défectueux [35]. Elles sont causées principalement par la non-réponse du capteur, les variations de gain relatif et/ou de décalage des capteurs, les erreurs d'étalonnage, etc.

## 1.6. Facteur cause des données manquantes dans l'imagerie

Maintes raisons peuvent se trouver à l'origine du manque de données. Exemple :

- Défaillance de l'équipement, des erreurs dans les mesures ou des défauts dans [9] l'acquisition de données ;
- Risques naturels, tels que glissements de terrain [9] ou encore éventuelle absence d'observateurs ;
- Pollution atmosphérique [10], [11] et [12], causée par les fumées d'usines, dont le monoxyde de carbone, le dioxyde de soufre et la concentration d'ozone ont été relevées à partir de stations de surveillance automatisées. Ces données contiennent généralement des valeurs manquantes qui peuvent entraîner un biais en raison de différences systématiques entre les données observées et non observées ;
- Couverture nuageuse [13], gros problème compliquant le traitement de l'image par la contamination des données par la présence nuages ;
- Outre les nuages, le brouillard interfère négativement dans le traitement de l'image, réduisant la visibilité à moins de 1 km :
- Couverture neigeuse saisonnière, est un autre facteur de perte de données dans l'imagerie [14].

## 1.7. Conclusion

Dans ce chapitre, nous avons présenté essentiellement une vue sur les cartes météorologique et les différente cartes satellitaire TRMM en donnant en particulier leurs caractéristiques et quelques problèmes rencontrés dans ce cadre. Nous avons également présenté les facteurs causant des données manquantes dans l'imagerie. Le contenu de ce chapitre en utilise pour aborder les chapitres restants.

# Chapitre 2

# Chapitre 2 : Traitement des données manquantes dans les images TRMM

## 2.1. Introduction

Le traitement des données manquantes dans l'imagerie climatique pose problème en informatique, imposant à ce titre une recherche méthodologique. Aujourd'hui, plusieurs méthodes sont déjà à l'œuvre, menant de telles analyses statistiques en présence d'observations incomplètes dans l'image climatique, permettant un appréciable saut qualitatif en termes de performance de résultats par rapport aux méthodes proposées antérieurement. Ce chapitre portera sur la présentation de quelques techniques appliquées au traitement des données manquantes. Nous définirons ensuite quelques méthodes d'imputation et d'interpolation dédiées au traitement des données manquantes dans l'imagerie climatique.

## 2.2. Données manquantes

Une donnée incomplète est celle pour laquelle la valeur de certain attribut reste inconnue, carence généralement due à l'absence d'observation, causée par perte ou défaut d'enregistrement. Les valeurs manquantes peuvent être de deux natures :

- Valeur manquante totale : c'est-à-dire qu'elle n'a pratiquement pas été observée ;
- Valeur manquante partielle : c'est-à-dire que l'observation a eu lieu, mais de façon incomplète.

Dans la littérature, de nombreuses études proposent des algorithmes conçus pour l'estimation de données manquantes produites pour le traitement des images climatiques. Ces études se penchent sur la question de résolution de ces problèmes d'ordre informatique en développant de nouvelles techniques. Aussi avons-nous opté pour une mise en évidence des principales caractéristiques des différentes méthodes. Nous présenterons les techniques les plus éprouvées pour donner sur le domaine une vue d'ensemble

## 2.3. Traitement des données manquantes dans les images climatiques

Malgré la quantité toujours croissante de données disponibles, avec l'émergence du big data, les problèmes de données manquantes dans les images climatiques restent très préoccupants et attendent une approche particulière. Cette partie avance une liste de solutions logicielles appropriées en ce qui concerne les observations pour le traitement des valeurs manquantes dans des ensembles de données réelles. La liste ne prétend cependant pas être complète. Les auteurs ne sont pas consacrés unanimement, pas plus qu'ils ne présentent de

comparaison exhaustive pour les résultats des différents outils logiciels en raison des efforts manifestement plus importants qu'auraient demandé de tels travaux.

Nous aurons à présenter dans cette partie trois stratégies de comblement de données manquantes en vue d'en élire celle la plus appropriée pour le traitement concerné.

### 2.3.1. Stratégie dépendant du mécanisme de données manquantes

Lorsqu'on souhaite analyser un jeu de données dont certaines sont manquantes, il est nécessaire d'appréhender les mécanismes et se fonder sur la probabilité qu'une donnée soit manquante. En général, le choix de chaque méthode de remplissage dépend de ces mécanismes de manque [15]. Pour gérer correctement les matrices de données incomplètes, il faut identifier le mécanisme responsable de la perte de données.

Trois mécanismes de traitement des données manquantes ont été décrits par LITTLE et RUBIN [16]: celui des données manquantes complètement au hasard (MCAR), manquantes au hasard (MAR) et manquantes sans interférence du hasard (MNAR).

- **Mécanisme MCAR (manquant complètement au hasard) :**

Ce type de manque survient entièrement au hasard, il ne se rapporte ni sur les valeurs observées ni sur celles manquantes. En termes simples, la probabilité qu'une observation soit manquante est indépendante à la fois des valeurs des autres variables et de la valeur de l'observation elle-même. Il s'écrit :

$$P(r|Xobs,Xmis)= P(r). \quad (2.1)$$

Tel que Xobs sont les données observées, Xmis est DM et (r) est la condition de distribution de DM.

- **Mécanisme MAR (manquant au hasard) :**

La probabilité qu'une observation soit manquante est également indépendante de la valeur de l'observation elle-même mais pas sur d'autres variables. Il s'écrit :

$$P (r|Xobs, Xmis)= p (r|Xobs). \quad (2.2)$$

- **Mécanisme NMAR (ne manquant pas au hasard) :**

Ce dernier cas survient lorsque la perte de données n'a rien d'aléatoire. Dans ce cas, la probabilité qu'une valeur soit manquante dépend de la valeur manquante elle-même. C'est:

$$P (r|Xobs, Xmis)=P (r|Xmis). \quad (2.3)$$

## Chapitre 2 : Traitement des données manquantes dans les images TRMM

L'organigramme représenté par la figure 2.1 [17] résume l'utilisation des méthodes d'imputation correspondant aux mécanismes de données manquantes.

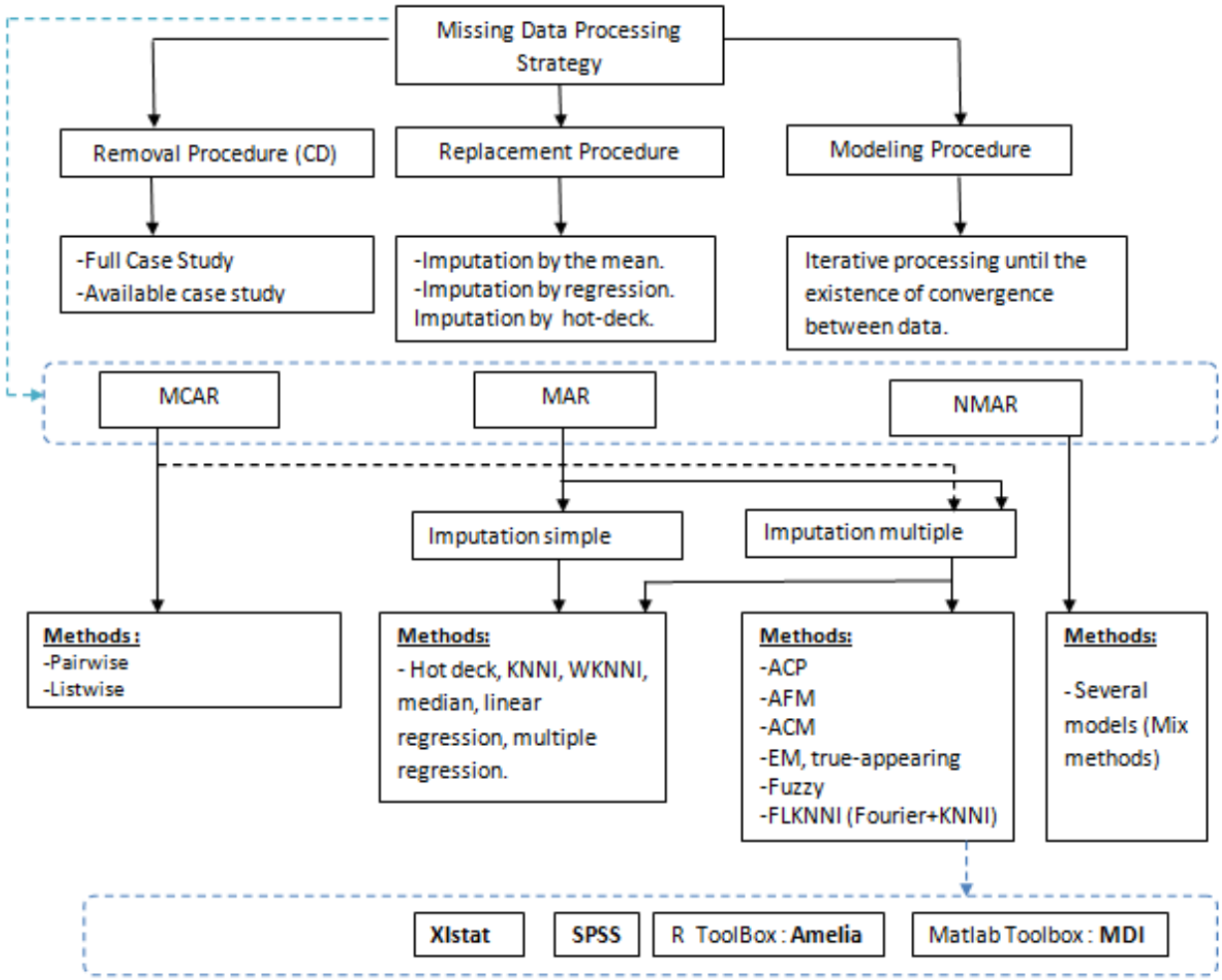


**Figure 2.1 :** Résumé des différentes stratégies de traitement des données manquantes.

Selon KLINE (1998), SONG et SHEPPERD, (2007) [18], le tableau 2.1 résume les trois stratégies possibles pour traiter des données manquantes, le recours aux procédures de :

- suppression ;
- remplacement ;
- modélisation de la distribution des données manquantes.

Différentes procédures de remplacement des données manquantes ont été élaborées au cours des années. Généralement, la différence entre ces méthodes s'estompe avec:

- une plus grande dimension de la base de données,
- un plus petit pourcentage des valeurs manquantes,

On distingue deux catégories de remplacement de la donnée manquante :

- imputation simple (imputation par la moyenne, imputation par la régression, etc.).
- imputation multiple.

**Tableau 2. 1:** Comparaison entre méthodes de comblement des données manquantes.

| Technique | Description | Avantage | Inconvénient |
|---|---|---|---|
| **Procédures de suppression** | | | |
| Étude des cas complets (Listwise deletion) [19, 20] | Supprime toutes les observations dont certaines valeurs sont manquantes. | Facile à utiliser. | Sacrifie une grande quantité de données et a un impact négatif sur les paramètres d'estimation (corrélation –régression). |
| Étude des cas disponibles (Pairwise deletion) [21, 20] | Crée une matrice de corrélation avec les valeurs disponibles. | Plus précis que la suppression listwise et préserve l'avantage des données. | Corrélations ou covariances biaisées. |
| **Procédures de remplacement** | | | |
| Imputation par la moyenne totale (Total mean substitution) [19,22] | Remplace par la moyenne des valeurs disponibles de la variable, toutes les valeurs manquantes pour la même variable. | Préserve la taille de la base de données et la rend facile à utiliser. | Sous-estimation de la variance et biaisement de la corrélation entre les variables. |
| Imputation par la moyenne de chaque classe (Subgroup mean substitution) [23, 24] | Remplace par la moyenne des valeurs disponibles de la variable de la même classe, toutes les valeurs manquantes pour la même variable et dans la même classe. | Ses résultats sont meilleurs par rapport à Imputation par la moyenne totale. | Sous-estimation de la variance et biaisement de la corrélation entre les variables. |

| | | | |
|---|---|---|---|
| Imputation par k-plus proche voisins (k-nearest-neighbors imputation) [25] | Remplace les valeurs manquantes par la valeur du k plus proche voisin dans l'ensemble de données. | Il ne fait aucune supposition quant à la distribution des données, et prend en considération la corrélation entre variables. | La difficulté réside dans le choix du paramètre k. |
| L'imputation hot deck (Hot-deck imputation) [26]. | Remplace une valeur manquante par la valeur de la même variable à partir d'un cas similaire dans l'ensemble de données. | Préserve les distributions des variables. | Risque d'altérer les relations entre les variables. |
| **Procédures de modélisation** | | | |
| Vraisemblance maximale (Maximum likelihood) [27]. | Les paramètres sont estimés par les données disponibles et les valeurs manquantes sont estimées en fonction des paramètres. | Augmentation de la précision si le modèle est correct. | Les hypothèses de la distribution exigée par la technique sont strictes. |
| Maximisation d'espérance (Expected maximisation) [28,29] | Processus itératif qui se poursuit jusqu'à ce qu'il y ait convergence dans les estimations des paramètres | Augmentation de la précision si le modèle est correct. | L'algorithme, trop complexe, met du temps à converger. |

### 2.3.2. Stratégie dépendant du pattern de données manquantes

ENDERS (2010) [30] souligne qu'il existe trois types de modèles de données manquantes, respectivement appelés uni-variés, monotones et arbitraires.

**2.3.2.1. Uni-variée :** les données manquantes sont dites uni-variées si une seule variable connait des données absentes. Ce cas est rarement observé en pratique.

**2.3.2.2. Monotone :** les données manquantes sont dites monotones si les variables peuvent être ordonnées en fonction de leurs données absentes.

**2.3.2.3. Arbitraire :** les données manquantes sont dites arbitraires s'il n'y a pas de structures dans les données manquantes. C'est-à-dire que les données manquantes sont réparties uniformément dans le jeu de données.

Nous voyons ci-après la configuration des données manquantes, rendues dans les cases gris clair qui les représentent. Le tableau 2.2. Décriras les choix des méthodes d'imputation suivant les patterns de manque de donnée

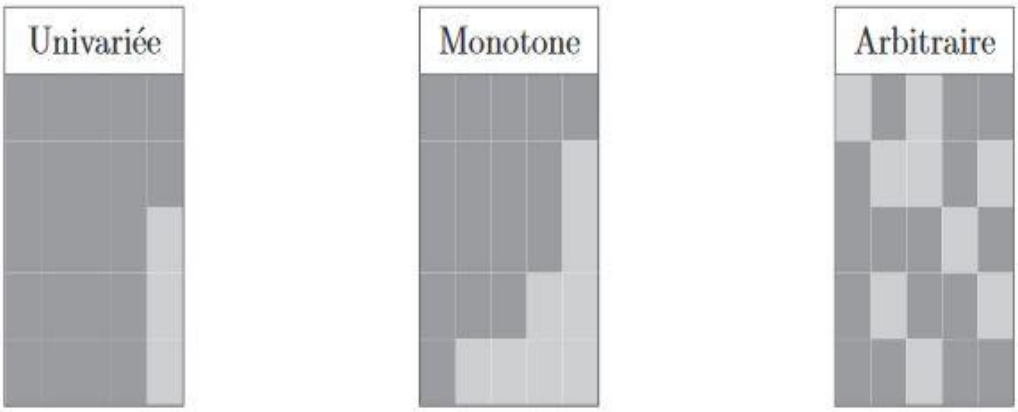


**Figure 2.2 :** Illustration des différents dispositifs de données manquantes.

**Tableau 2. 2 :** Choix des méthodes d'imputation en fonction des patterns, [31,57].

| Pattern de données manquantes | Type de variables à imputer | Méthodes d'imputation possibles |
|---|---|---|
| Monotone | Continue | ▪ Méthodes paramétriques qui reposent sur l'hypothèse de normalité multi-variée : régression monotone ; *«Prédictive mean matching»*.<br>▪ Méthode non paramétrique: score de propension. |
| | Ordinal | -Régression logistique. |
| Arbitraire | Continue | -« Markov Chain Monte Carlo » (MCMC) : imputation totale (*« full-data imputation »*) ou partielle<br>-FCS (*« Fully Conditional Spécification »*) régression.<br>-FCS *« Predicted mean matching »* |
| | Ordinal | -FCS *« logistic régression »* |
| Uni-varié | Continue | Régression linéaire bayésien |
| | Qualitative | Prédictive mean matching<br>Régression logistique |

### 2.3.3. Stratégie dépendants du pourcentage de donnée manquante

Selon Presti et al. [32], le choix des méthodes de comblement dépend de la proportion des données manquantes dans un échantillon, il est rapporté que le remplacement des variables manquantes en utilisant les données d'autre série complète et de trouver des bon résultats,

seulement s'il existe une corrélation. Le tableau 2.3 Nous montre le choix de la méthode adéquate pour le comblement des données manquantes.

**Tableau 2. 3:** Méthodes adéquates selon le pourcentage des données manquantes.

| Taux de données manquantes | Méthodes proposées pour traiter les données manquantes |
|---|---|
| <5% | Toutes les méthodes fonctionnent correctement. |
| [5%−10%] | En utilisant une valeur constante (ex. modèle simple),<br>les méthodes d'imputation simple (ex. régression).<br>plusieurs méthodes d'imputation sont de bons choix. |
| [10%−15%] | Les méthodes d'imputation simples pourraient être biaisées.<br>L'imputation multiple est fortement recommandée. |
| >15 % | Plusieurs méthodes d'imputation sont fiables. |

## 2.4. Programmes et bibliothèques pour le comblement de DM

Il existe des logiciels spécialisés pour le traitement et l'imputation des valeurs manquantes. Ces progiciels effectuent essentiellement l'imputation des valeurs manquantes et rendent les jeux de données complets pour des évaluations supplémentaires. Nous décrivons ici certains scripts et logiciels.

### 2.4.1. Présentation du package Amelia II pour langage R

Le script Amelia II permet aux utilisateurs d'imputer des ensembles de données incomplets. Il facilite aux analyses nécessitant des observations complètes l'utilisation de manière appropriée toutes les informations disponibles dans un ensemble de données permettant d'éviter biais, inefficacité ainsi que les estimations d'incertitude incorrectes pouvant découler de la suppression de toutes les observations partiellement faites. L'analyse avec Amelia II effectue des imputations multiples [33], c'est une approche polyvalente des données contenant des valeurs manquantes. Il a été démontré que les imputations multiples réduisent les biais tout en augmentant l'efficacité.

Amelia II combine l'algorithme EM classique avec une approche bootstrap pour obtenir les postérieurs (Figure 2.3) pour chaque tirage par EM. Nous amorçons les données pour simuler l'incertitude de l'estimation, puis exécutons l'algorithme EM pour trouver le mode du postérieur des données initialisées.

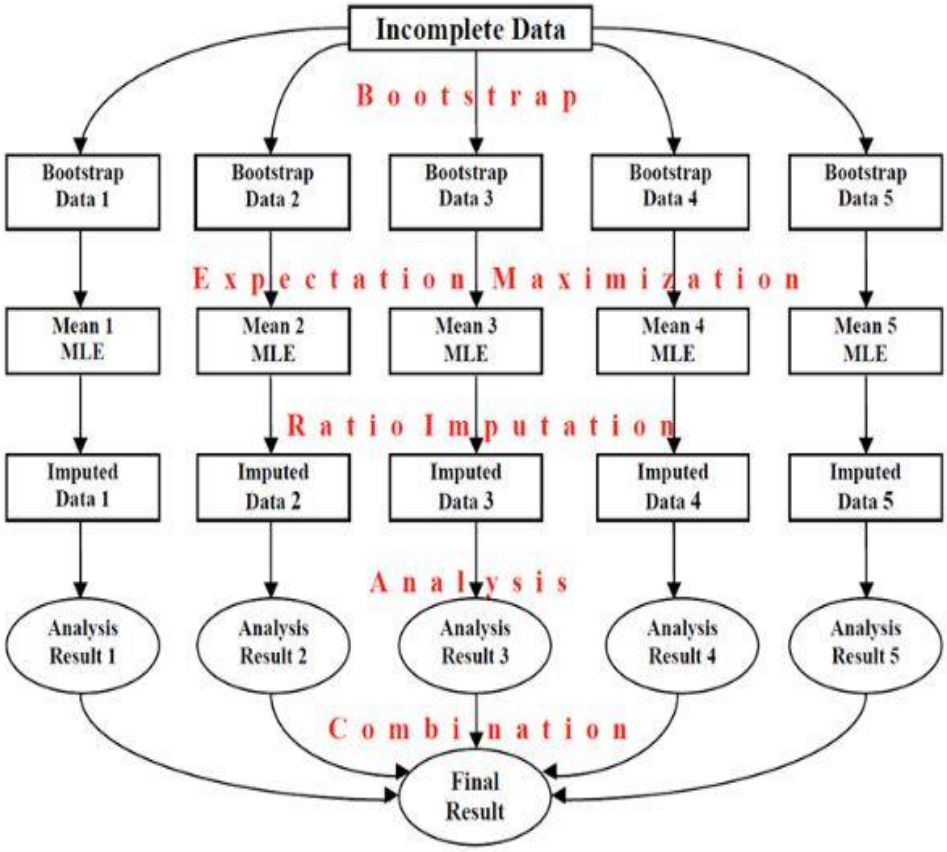


**Figure 2.3:** Représentation graphique des étapes de remplissage des manques de données à l'aide de la méthode EM [34].

### 2.4.2. Logiciel ArcGIS

ArcGIS est un ensemble de logiciels SIG réalisé par la société ESRI, regroupant des logiciels clients (ArcView, ArcEditor, ArcInfo et ArcExplorer) et des logiciels serveurs (ArcSDE et ArcIMS). Ils permettent de visualiser, interroger, analyser et mettre en page des données. Arc GIS fournit également des outils interactifs pour explorer, sélectionner, afficher, éditer, analyser, symboliser et classifier des données ou créer automatiquement la «boite à outils» d'ArcGIS. Arctoolbox regroupe l'ensemble des outils de géo-traitement utiles pour interpoler des cartes. De nombreuses méthodes dans ArcGIS fournissent des techniques pour interpolation de valeurs manquantes, dont l'outil Pondération par l'inverse de la distance (IDW), Krigeage, La méthode d'interpolation par voisins naturels (NN), Spline, etc.

## 2.5. Traitement de pixel manquant dans l'imagerie climatique

Une variété de méthodes de reconstitution d'informations manquantes pour l'imagerie de télédétection a été proposée pour résoudre le problème, les pixels manquants dans l'imagerie climatique doivent être estimés aussi précisément que possible. Il existe plusieurs méthodes selon la source d'information, la plupart des méthodes de reconstruction peuvent être classées en trois catégories [35] principales.

### 2.5.1. Méthodes spatiales

Les méthodes spatiales sont qualifiées pour reconstruire de petites zones manquantes ou des régions à texture régulière. Cette approche comprend plusieurs méthodes, par exemple les méthodes d'interpolation. L'avantage des méthodes d'interpolation est qu'elles sont simples et efficaces, mais ne peuvent pas reconstruire de grandes zones manquantes ou de zones de texture complexe.

### 2.5.2. Méthode spectrale

Les méthodes basées sur le spectre peuvent récupérer les données spectrales manquantes avec un haut niveau de précision grâce à l'utilisation de la forte corrélation entre les différentes données spectrales. Cependant, ces méthodes ne peuvent pas traiter une épaisse couverture nuageuse, car cela conduit à l'absence de toutes les bandes spectrales à des degrés divers.

### 2.5.3. Méthode temporelle

La méthode temporelle repose sur le fait que les données strictement chronologiques et présentent des fluctuations régulières. Par exemple, GAO et *al*. Proposa une méthode de cartographie d'angle tempo-spectral (TSAM) pour SLC-off capable de mesurer la similarité tempo-spectrale entre les pixels décrits dans la dimension spectrale et la dimension temporelle.

## 2.6. Traitement des données manquantes dans les images climatiques

Les méthodes de traitement appropriées aux images climatiques sont plusieurs, d'efficience et d'emploi variés. Nous passerons ici en revue deux d'entre-elles.

### 2.6.1. Méthode d'interpolation

Les techniques d'interpolation sont considérées comme des outils déterministes et géostatistiques. Les techniques déterministes créent des surfaces basées sur des points mesurés ou des formules mathématiques, tandis que les techniques d'interpolation géostatistique sont basées sur des statistiques et utilisées pour la modélisation de surface de prédiction plus avancée.

#### 2.6.1.1. Distance inverse pondérée(IDW)

IDW est une méthode déterministe d'interpolation spatiale basée sur la similitude ou la régularité dans une zone de recherche. IDW est basé sur l'hypothèse que les points échantillonnés plus proches des points prédits ont plus d'influence sur la valeur prévue que les points d'échantillonnage plus éloignés [36]. Ainsi, Les valeurs attribuées aux points inconnus sont calculées comme une moyenne pondérée des valeurs disponibles aux points connus. Elle est mise en œuvre dans de nombreux progiciels de systèmes d'information géographique (SIG) [37]. L'IDW est formulé comme suit :

$$Z(x) = \frac{\sum_{i=1}^{n} WiZi}{\sum_{i=1}^{n} Wi} \quad (2.4)$$

$$W(i) = \frac{1}{di^{-k}}$$

Où Z(x) est la valeur estimée à un point prédit. Zi est la valeur observée au point i. Wi est la valeur de poids attribuée au point i, et di est la distance entre le point i et le point prédit; de plus, k est la variable de puissance.

#### 2.6.1.2. Méthode de Krigeage

Le krigeage est une technique stochastique similaire à l'IDW [38], développée pour modéliser ces concepts. Méthode d'interpolation, le krigeage prédit de manière optimale les valeurs de données en utilisant des données prises à des emplacements connus à proximité. Il utilise des combinaisons linéaires de poids à des points connus pour estimer la valeur à un point inconnu (LUO et *al*. 2008). Cependant, dans cette méthode, la corrélation spatiale est prise en compte lors de l'estimation de la surface [39]. Cette corrélation est déterminée en utilisant la fonction de semi-variance comme indiqué dans l'Eq 2.5 où N (h) désigne le nombre de paires de points échantillonnés avec une distance h. La formulation complète de la méthodologie du Krigeage est fournie par plusieurs littératures.

$$y(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(Si) - h(Si+h)]^2 \quad (2.5)$$

Il existe plusieurs types de krigeage. Dans sa version ordinaire, elle est la méthode la plus courante supposant l'absence de moyenne constante pour les données sur une zone. Quant à sa version krigeage universelle, elle suppose qu'une tendance dominante existe dans les données et qu'elle peut être modélisée.

#### 2.6.1.3. Méthode Natural Neighbors

L'interpolation NN trouve le sous-ensemble d'échantillons d'entrée le plus proche d'un point de requête et leur applique des pondérations en fonction de zones proportionnées afin d'interpoler une valeur [39]. La méthode du voisin naturel propose une mesure pour le calcul des poids et la sélection des voisins interpolant. Les propriétés de base de NN sont locales, n'utilisant qu'un sous-ensemble d'échantillons qui entourent un point de requête, et les hauteurs interpolées sont garanties dans la plage des échantillons utilisés. NN s'adapte localement à la structure des données d'entrée, ne nécessitant aucune entrée de l'utilisateur concernant le rayon de recherche, le nombre d'échantillons ou la forme. Cela fonctionne aussi bien avec des données distribuées régulièrement qu'irrégulièrement. L'équation de base de la méthode NN s'écrit ainsi :

$$G\ (x,y) = \sum_{i=1}^{N} Wif(\ x\mathrm{i}\ ,Y\mathrm{i}) \qquad (2.6)$$

Tel que G (x, y) est l'estimation en (x, y), Wi sont les poids, f (xi, yi) sont les données connues en (xi, yi), et N le nombre total de points échantillonnés.

### 2.6.2. Méthodes d'imputation

Des méthodes de ce type, nous retenons ici trois.

#### 2.6.2.1. Méthode de la moyenne arithmétique

Cette méthode sert à remplacer chaque valeur manquante par la moyenne des valeurs observées pour cette variable. La moyenne de la variable est calculée à partir des valeurs présentes, sur lesquelles on se base pour restituer la valeur manquante de cette variable. Généralement utilisée pour imputer des données météorologiques et hydrologiques manquantes [16], cette méthode convient au traitement des données manquantes de débit obtenues par la moyenne des stations proches sélectionnées autour de la station cible. La valeur manquante estimée est donnée par :

$$Y(t) = \frac{1}{n}\sum_{i=1}^{N} Xi \quad (2.7)$$

Où $Y_t$ est la valeur projetée des données manquantes à la station cible ***t***, Xi sont les données observées aux stations proches ***i***, et ***n*** est le nombre de stations proches.

#### 2.6.2.2. Imputation pondérée des voisins les plus proches (WKNNI)

Il s'agit d'une autre méthode d'estimation, basée sur le coefficient de pondération de séries chronologiques similaires [40]. Dans les séries de données climatiques, nous avons utilisé la distance euclidienne entre les stations de similarité et la station de référence pour calculer ce coefficient, façon d'obtenir la meilleure estimation de DM, ce qui rend le résultat final déterminé par une moyenne pondérée de toutes les données voisines. Il est donné par l'équation suivante:

$$Z(x) = \frac{\sum_{i=1}^{n}(Pi*Wi)}{\sum_{i=1}^{n} Wi} \quad (2.8)$$

Tel que

$$W(i) = \frac{1}{dxi^{-k}}$$

Où *Px* est le DM observé dans la station de référence *x,* et *k,* le nombre de stations similaires ; *Pi,* données imputées, c'est-à-dire relatives aux précipitations observées sur la station voisine, *Wi*, coefficient de pondération égal à $dxi^{+k}$, *dxi* la distance euclidienne entre l'emplacement de la station voisine ***i*** et la station de référence ***x*** ; et ***K*** appelé distance de frottement.

#### 2.6.2.3. Imputation multiple

L'imputation multiple s'est distinguée en tant que méthode de traitement de données manquantes, elle fournit des inférences statistiques valides sous la condition MAR [41]. Elle a pour particularité d'imputer les données manquantes tout en reconnaissant l'incertitude associée aux valeurs imputées. Elle traite les ensembles de données manquantes complets que les procédures statistiques standard prennent ensuite en charge pour analyse. L'imputation multiple

a été mise en œuvre dans des logiciels tels que SAS [33]. Et le progiciel Amelia II [42]. Pour appliquer cette méthode, il faut suivre les étapes suivantes :(1) imputer les données manquantes ***m*** fois pour produire des ensembles de données complets; (2) analyser chaque ensemble de données en utilisant une procédure statistique standard ; (3) combiner les résultats en utilisant les formules de on écrit.

$$Px = \frac{1}{m}\sum_{i=1}^{m} \mathrm{I}i\ (Px, Pi) \tag{2.9}$$

## 2.7. Conclusion

Dans l'imagerie appliquée au domaine climatique, le traitement des données manquantes constitue un problème assez délicat à traiter. Les choix incontournables qui devraient être faits sont tributaires de facteurs ayant directement rapport à la maitrise infaillible qui n'est pas toujours pas acquise à ce jour. Il a été présenté au long de ce chapitre méthodes et logiciels intervenant dans le traitement des données manquantes, sous l'éclairage des présentations que l'on rencontre dans la littérature. Nous avons assez brièvement survolé quelques méthodes d'imputation et d'interpolation.

# Chapitre 3

## 3.1. Introduction

L'étude que nous venons d'exposer au chapitre précédent s'était assigné, avions-nous vu, l'objectif de construire un modèle assez robuste qui puisse aider à réduire, en attendant de surmonter l'aléa que représente les manques de données dans l'imagerie TRMM.

Le sujet, dans cette partie, portera sur la définition des *« boîtes à outils »*. Nous leur garderons leur terminologie originale anglaise, aisément reconnaissable puisqu'assez évocatrice : *« Toolbox »* MDI et KNNI, avec une description statistique quantitative et qualitative de chacune d'entre elles. Dans ce cadre, la solution que nous proposons pour estimer les données manquantes dans l'imagerie TRMM et d'appliquer un réseau artificiel de neurones ANN qui compose de trois couches : la couche d'entrée qui va collecter les données, une couche de sortie qui génère des informations calculées et une ou plusieurs couches cachées appropriées à connecter la couche d'entrée/sortie en utilisant les fonctions de deux méthodes KNNI et la régression appliquée dans les séries chronologiques à résoudre le problème de lacune de données.

Tout au long de ce chapitre, nous décrirons principalement le fonctionnement des toolboxes (MDI et KNNI) ainsi que les méthodes implémentées. Il s'ensuivra la formulation du comblement des données manquantes dans l'imagerie TRMM par les réseaux de neurones artificiels ANN, prenant les deux méthodes k-plus proche voisin (KNNI) et de régression, avec un exemple d'application.

## 3.2. Méthode de comblement

Pour l'heure, il existe déjà plusieurs méthodes de comblement des données manquantes. Cette partie entame la présentation de deux toolboxes (MDI et KNNI) avec étude comparative entre elles.

### 3.2.1. Définition de toolbox MDI

La toolbox MDI est une boite MATLAB graphique, appelée bibliothèque d'imputation de données manquantes (MDI), dédiée à la réalisation de jeux de données incomplets. Elle a été construite et intégrée dans MATLAB R2013 (Maths, Works, Sheborn, MA), testée dans ses nombreuses versions antérieures et postérieures (MATLAB 2010-2015). La toolbox MDI est disponible gratuitement à des fins académiques à l'adresse (http : //mseg.webs.upv.es), sous licence GNU. Les valeurs manquantes sont imputées en appliquant des méthodes de construction du modèle PCA avec des données manquantes. Les différentes méthodes implémentées dans MDI sont les suivantes [43] : la régression des scores ajustés TSR, régression des données connues KDR, KDR avec régression en composante principale KDR-

PCR, KDR avec de moindres carrés partiels KDR-PLS, projection sur le plan du modèle PMP, algorithme itératif IA, algorithme de régression des moindres carrés partiels itératifs, modifié, non linéaire NIPALS, avec augmentation de données AD.

### 3.2.2. Analyse en composantes principales (ACP)

PCA [43] est une méthode multi-variée visant à trouver le sous-espace où les données varient. Pour cela, les variables d'origine, généralement corrélées, sont compressées en un ensemble réduit de composants principaux non corrélés (PC). Le modèle PCA est le suivant:

$$X = TP^T + E \tag{3.1}$$

Où *P* est la matrice de chargement, contenant par colonnes les combinaisons linéaires des variables originales définissant l'espace latent; *T* est la matrice des scores, ayant par colonnes les *PC*; et *E* est la matrice d'erreur. La partition de données manquantes X = $[X^{\#}X^{*}]$, induite par ligne $X_i^T = [X_i^{\#T}\ X_i^{*T}]$, peut également être étendue au modèle PCA (équation 3.1), c'est-à-dire: $[X^{\#}X^{*}]=T[P^{\#T}P^{*T}]+E$ (3.2)

### 3.2.3. Méthodes d'imputation des données manquantes (MDI)

Les méthodes implémentées dans la boîte à outils MDI sont brièvement décrites dans cette section. Où les méthodes de cadre basées sur la régression pour la construction de modèles PCA ont été initialement proposées et comparées à d'autres approches classiques. La plupart d'entre elles sont également incluses dans la boîte à outils MDI.

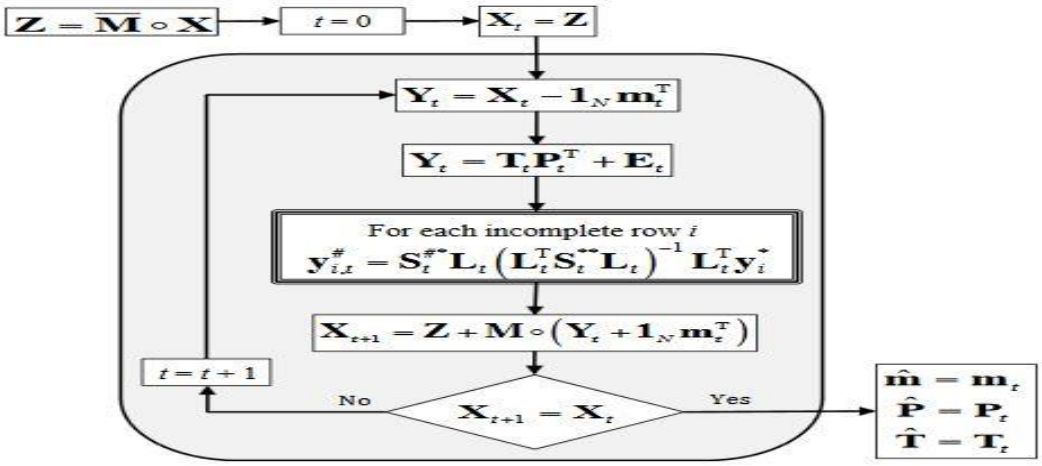


**Figure 3.1 :** Cadre basé sur la régression adaptée pour PCA-MB avec données manquantes.

| **Méthode** | **Clé matrice L** |
|---|---|
| TSR | $P^*$ |
| KDR | I |
| KDR-PCR | $V_{1:\rho}$ Matrice de vecteur propre de $\boldsymbol{S}^{**}$ et ρ ≤rang ($S^{**}$) |
| KDR-PLS | $W^*$, Matrice de chargement du modèle $PLS$ , $T_{PLS} = X^*W^*$ |

**Tableau 3.1 :** Matrice-clé L pour les méthodes basées sur la régression.

### 3.2.3.1. Méthodes de cadre basées sur la régression

Les méthodes basées sur la régression ont été récemment adaptées [45] suivant le contexte d'exploitation du modèle PCA (ME) [44]. Dans PCA-ME, on suppose qu'un modèle PCA soit déjà ajusté sur des données complètes et l'on souhaite analyser une nouvelle observation avec des valeurs manquantes. Les méthodes basées sur la régression sont: TSR, KDR, KDR-PCR et KDR-PLS.

Un organigramme (figure 3.1) représente la procédure de ces méthodes. Il commence à remplir les positions manquantes par des zéros, suite à quoi il estime, après centrage avec le vecteur moyen ***m***, un modèle PCA. Ensuite, pour chaque ligne incomplète ***i***, une estimation de valeur manquante est obtenue à partir de $S^{\#*}$ , $S^{**}$ et de la matrice-clé ***L*** qui est différente pour chaque méthode (Cf. tableau 1). Par la suite, les valeurs initiales manquantes sont remplacées par les estimations du modèle de régression. Si la solution entre la nouvelle matrice de données imputée et la précédente est inférieure à la tolérance établie, l'algorithme aura convergé. Sinon, une autre itération de l'algorithme sera effectuée (figure 3.1).

### 3.2.3.2. Projection sur le plan-modèle (PMP)

La méthode PMP a été adaptée conjointement avec les méthodes basées sur la régression [45]. À l'origine, elle fut d'abord proposée pour le problème PCA-ME [43], L'algorithme présenté à la figure 3.1, au lieu d'utiliser les matrices de covariance, à l'étape ***t***, impute les valeurs manquantes en utilisant uniquement la matrice de chargement:

$$y_{i,t}^{\#} = P_i^{\#}(P_i^{*T}P_i^{*})^{-1}\,P_i^{*T}y_{i,t}^{*} \tag{3.3}$$

### 3.2.3.3. Algorithme itératif (AI)

AI [43] impute les valeurs manquantes en utilisant la prédiction du modèle PCA. Par conséquent, IA suit également le schéma présenté dans la figure 3.1, quoiqu'en évitant de tenir compte de la case encadrée, c'est-à-dire que les valeurs manquantes sont remplacées par

l'estimation utilisant les chargements et les scores du modèle PCA ajusté avec l'imputation précédente:

$$X_{t+1} = Z + M \text{ o } (T_t P_i^T + 1_N m_t^T) \quad (3.4)$$

### 3.2.3.4. Régression des moindres carrés partiels itérative non linéaire modifiée (NIPALS)

NIPALS modifié [44] consiste à adapter l'algorithme pour traiter les lignes incomplètes de l'ensemble de données ***X*** en effectuant les régressions itératives par les données disponibles et en ignorant les valeurs manquantes.

### 3.2.3.5. Augmentation des données (AD)

AD [43], méthode d'imputation multiple, se distingue des autres méthodes essentiellement par la pluralité des valeurs qu'elle applique par itération à chaque donnée manquante (***M***). la figure 3.2 montre l'algorithme AD implémenter dans le toolbox MDI

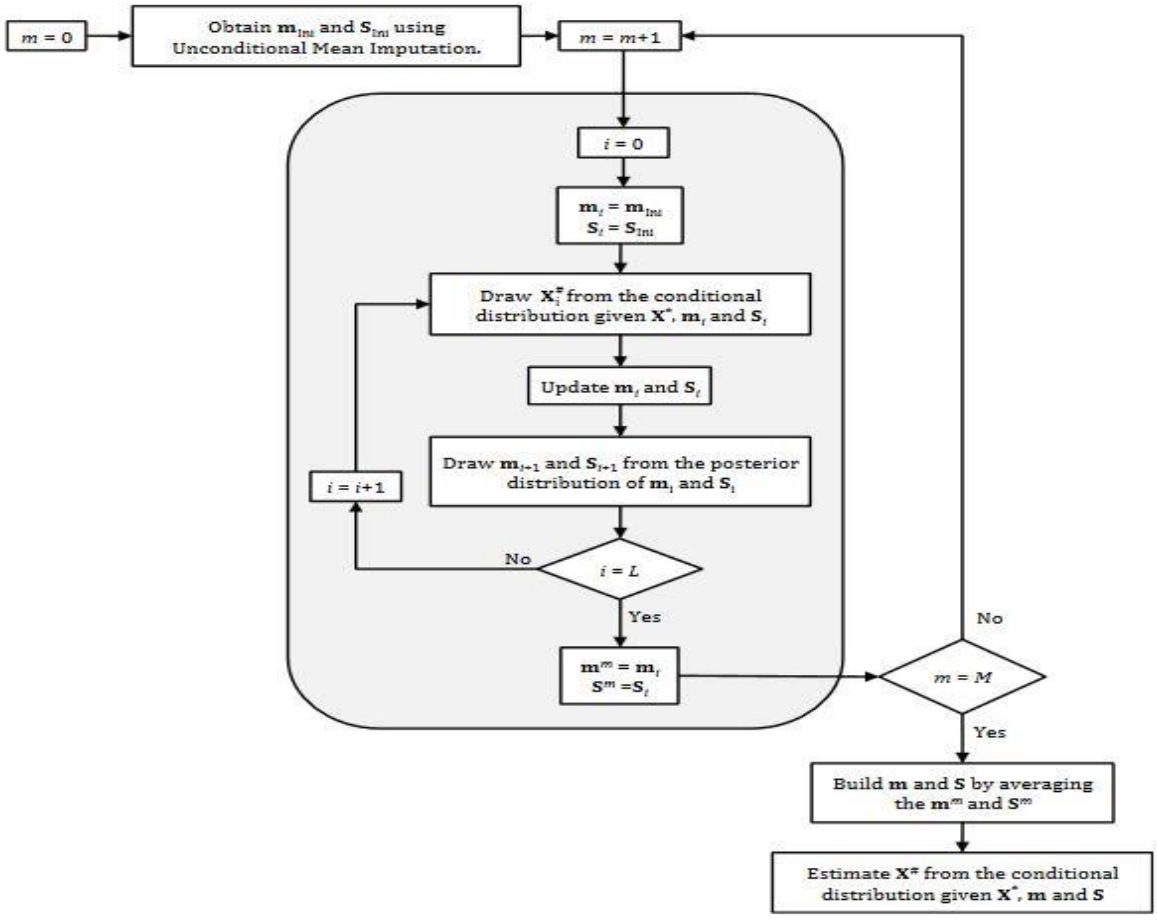


**Figure 3.2 :** Algorithme d'augmentation des données implémenté dans MDI Toolbox.

### 3.2.4. Mode opératoire

Le toolbox MDI est lancée en introduisant **MDIgui** dans la fenêtre de commande MATLAB. Comme indiqué (figure 3.3), la première étape consiste à sélectionner l'ensemble des données. Le bouton *Données* de l'espace de travail charge un jeu de données avec des DM depuis l'espace de travail MATLAB. Un ensemble de données précédemment stocké dans Excel peut être chargé en cliquant sur *Lire le fichier*. Le bouton *Utiliser l'exemple* ouvre une nouvelle fenêtre avec des exemples de données (figure 3.4). De cette façon, trois ensembles de données différents peuvent être sélectionnés avec trois pourcentages de valeurs manquantes : 10%, 30% et 60%. Le jeu de données simulé avec 30% de valeurs manquantes est sélectionné. Les méthodes d'imputation des données manquantes sélectionnées sont TSR, KDR, KDR-PCR, KDR-PLS, PMP, IA, NIPALS modifiés et DA. La méthode recommandée est TSR, car elle délivre un résultat suffisamment probant [45].

L'interface MDI permet de modifier les paramètres des différentes méthodes. De cette manière, le nombre d'itérations maximales effectuées par la méthode et la tolérance pour la convergence peuvent être modifiées à partir de leurs valeurs par défaut: 5000 itérations et une tolérance de$10^{-10}$. Ces paramètres sont actifs lorsque les méthodes sont basées sur la régression, notamment IA et NIPALS. Si AD est sélectionnée comme méthode d'imputation, ces paramètres sont désactivés et le nombre de chaînes de Markov et la longueur de chaîne sont activés. L'utilisateur pourrait ainsi modifier par défaut le nombre des itérations. Une fois la méthode et les paramètres introduits, la fenêtre *Aperçu des données* apparaît. Le modèle des valeurs manquantes et son pourcentage peuvent y être visualisés (figure 3.5). Les carrés rouges représentent les entrées manquantes dans l'ensemble de données alors que les valeurs disponibles sont rendues en blanc. Après avoir cliqué sur *Continuer*, deux barres de progression s'affichent consécutivement. La première montre la progression du calcul des variances tandis que la seconde affiche la progression du calcul des covariances. La fenêtre ***NumberComponents*** permet de sélectionner le nombre approprié de PC pour le modèle PCA. Trois graphiques sont présentés ici pour évaluer ce nombre (figure 3.6). Sur le côté gauche, le classique graphique en éboulis, avec des valeurs propres de la matrice de covariance estimée de $\boldsymbol{X}$. Au centre, le pourcentage cumulé de la variance expliquée. Il convient de noter que les deux graphiques sont obtenus sur la base d'une estimation par paire de la matrice de covariance de l'ensemble de données avec des valeurs manquantes :

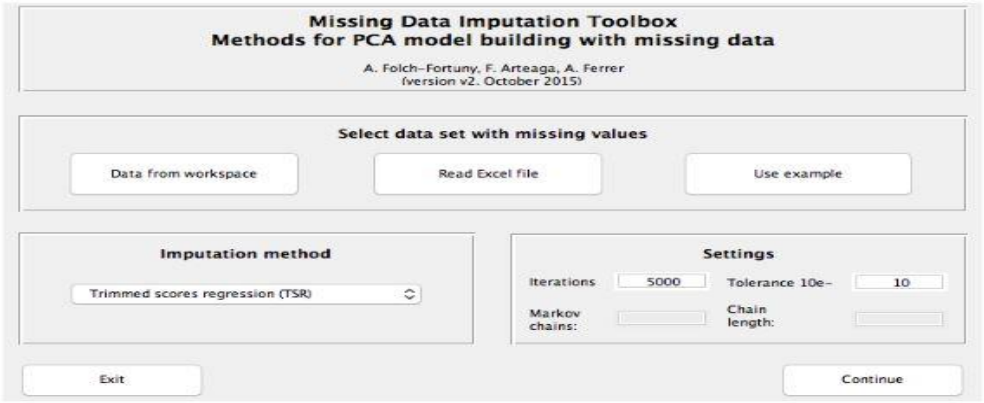


**Figure 3.3 :** Interface graphique de MDI pour la sélection des données, méthode et paramètres.

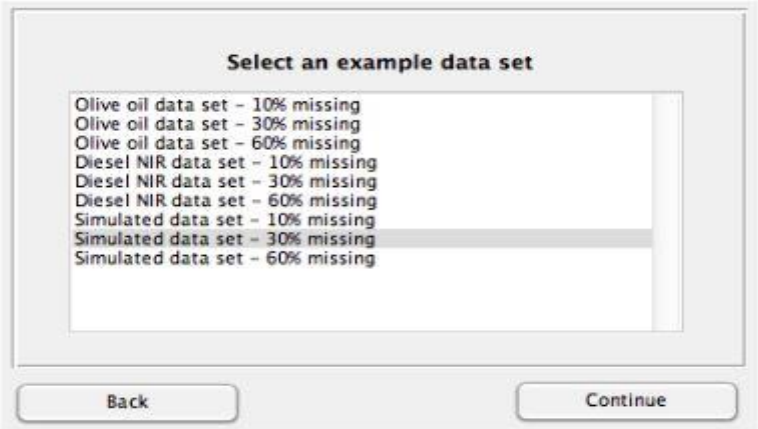


**Figure 3.4 :** Exemple de fenêtre de sélection de données.

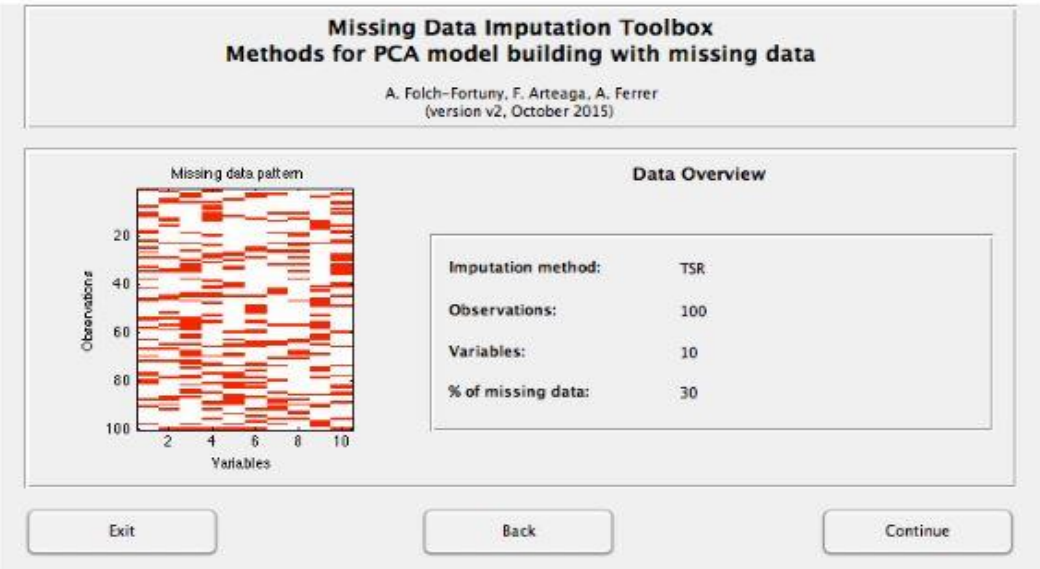


**Figure 3.5 :** Interface graphique pour l'aperçu des données.

Une troisième parcelle est incluse sur le côté droit. Ce tracé correspond aux résultats de l'algorithme de multiplication par colonne (**ckf**) pour estimer le nombre de PC en PCA, récemment proposé [46].

Dans cette étude, les informations fournies par les trois graphiques de la (figure 3.6) sont cohérentes, donc trois PC sont sélectionnés, car : i) il y a une énorme différence entre les valeurs propres ; ii) la variance expliquée cumulative avec trois composantes est d'environ 90%, et la variance expliquée avec 4 composantes est similaire; iii) la PRESSE est minimale en utilisant trois composants. Une fois le nombre de PC sélectionné, MDI Toolbox exécute la méthode d'imputation des valeurs manquantes. Le temps de calcul dépend de la méthode choisie. Habituellement, TSR et IA sont les méthodes les plus rapides, par contre DA et KDR sont les plus lentes. Deux barres de progression apparaissent simultanément (figure 3.7) [47] pendant que la boîte à outils effectue les imputations itératives. La barre supérieure montre le numéro d'itération en cours et s'exécute jusqu'à atteindre le nombre maximal spécifié d'itérations dans la fenêtre initiale de **MDIgui** (figure 3.3). La barre inférieure suggère une idée sur l'écart entre les itérations consécutives et la tolérance définie pour la convergence. Ceci est calculé comme $1 - \frac{d-l}{d}$ où $\boldsymbol{d}$ représente la différence quadratique moyenne entre les valeurs imputées dans les itérations consécutives et $\boldsymbol{l}$ est la tolérance spécifiée.

***ShowResults*** est la dernière fenêtre de Toolbox MDI (voir figure 3.8). Ici, les détails de l'imputation des données sont résumés: méthode d'imputation, itérations, tolérance et temps de calcul (en secondes). De plus, deux figures avec graphiques de chargement et des scores sont présentées pour faciliter l'interprétation graphique du modèle.

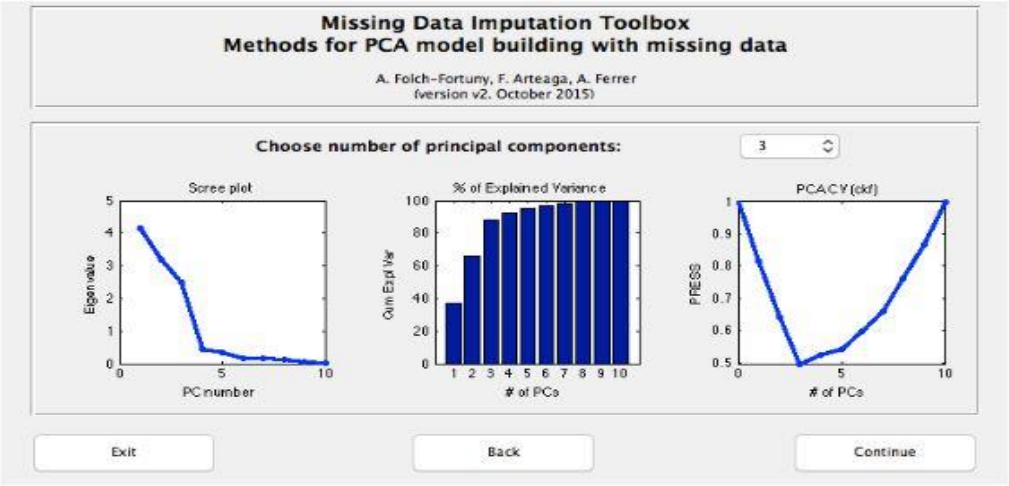


**Figure 3.6 :** Sélection du nombre de composantes principales, basée sur le graphe en éboulis (à gauche) et le diagramme à barres de la variance expliquée cumulative (au centre), et la validation croisée de l'ACP à l'aide de l'algorithme ckf.

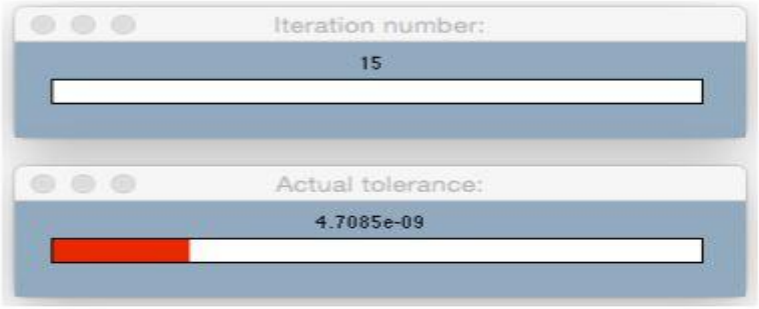


**Figure 3.7 :** Barres de progression reflétant la procédure d'imputation des données manquantes. Dans cet exemple, 15 des 5000 itérations ont été calculées (en haut), et la différence quadratique moyenne entre les valeurs imputées dans les itérations 14 et 15.

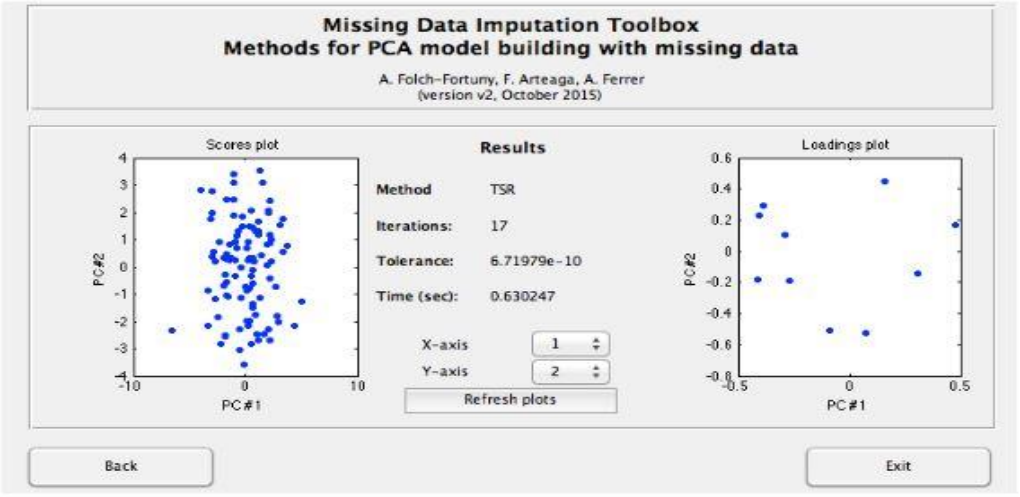


**Figure 3.8 :** Tracés des scores et des charges du modèle PCA ajustés sur les données imputées.

Enfin, la Toolbox MDI renvoie automatiquement une structure de données à l'espace de travail Matlab avec toutes les informations de l'imputation des données. Le modèle PCA résultant stocke ces données dans les champs *Chargements* et *Scores*, ainsi que la moyenne et les covariances des variables. De cette façon, les données sont reproduites comme suit: **Xreconstruit** = Moyenne + Scores × **ChargementsT.**

### 3.2.5. Validation de toolbox MDI

La boîte à outils MDI offre un ensemble d'outils pour effectuer l'imputation des données manquantes à l'aide de différentes méthodes d'imputation [43]. La boîte est implémentée dans Matlab, livrée avec une interface graphique élégante et intuitive avec une procédure guidée étape par étape qui rend la boîte à outils facile à utiliser et facile à intégrer aux routines Matlab existantes. Il n'y a pas beaucoup d'outils disponibles pour l'imputation des données manquantes et la boîte MDI est un ajout très apprécié à la boîte à outils de chimio-métriez.

### 3.2.6. Définition de toolbox KNNI

KNN est un algorithme d'apprentissage paresseux basé sur des instances. Il figure parmi les 10 meilleurs algorithmes d'exploration de données. L'apprentissage basé sur les instances est basé sur le principe que ces instances d'un ensemble de données existeront généralement à proximité d'autres observations ayant des propriétés similaires [58]. L'approche KNN a été étendue à l'imputation des données manquantes dans divers ensembles de données. Généralement, l'imputation KNN est un choix approprié lorsque nous n'avons aucune connaissance préalable de la distribution des données. Étant donné l'incomplétude de l'instance, cette méthode sélectionne ses ***k*** voisins les plus proches en fonction d'une métrique de distance et estime les données manquantes avec la moyenne ou le mode correspondant. La règle moyenne est impliquée afin de prédire les entités numériques manquantes et la règle de mode est utilisée pour prédire les caractéristiques catégoriques manquantes. La méthode d'imputation KNN ne crée pas de modèles prédictifs explicites, car l'ensemble des données d'apprentissage est utilisé comme modèle paresseux. En outre, cette méthode peut facilement traiter les cas avec plusieurs valeurs manquantes.

### 3.2.7. Application

La complétion par ***k*** plus proches voisins (***k-nearest neighbors***) consiste à exécuter l'algorithme suivant sous python qui modélise et prévoit les données manquantes.

```
#importer les bibliothèques nécessaires
import numpy as np
import pandas as pd

# importer la classe KNNimputer
from sklearn.impute import KNNImputer

# créer un jeu de données pour une matrice RVB
dict = {'R':[80, 90, np.nan],
        'G': [60, 65, 56],
        'B':[np.nan, 57, 80],

# creating a data frame from the list

  Before_imputation = pd.DataFrame(dict)

#print dataset before imputaion

print("Data Before performing imputation\n",Before_imputation)

# create an object for KNNImputer
imputer = KNNImputer(n_neighbors=2)
After_imputation = imputer.fit_transform(Before_imputation)
# print dataset after performing the operation
print("\n\nAfter performing imputation\n",After_imputation)

#Données avant de procéder à l'imputation

     R          G       B
0   80.0       60.0     NaN
1   90.0       65.0    57.0
2    NaN       56.0    80.0
3   95.0        NaN    78.0

#Après avoir effectué l'imputation

[[80.   60.   68.5 .]
 [90.   65.   57.  .]
 [87.5 56.   80.  .]
```

**Figure 3.9 :** Imputation des valeurs manquantes avec la méthode KNNI.

### 3.2.8. Validation de toolbox KNNI

***K-plus proche voisin*** est une méthode d'imputation efficace dans la manipulation des données manquantes. Les résultats qu'elle délivre sont sensiblement positif au regard d'autres méthodes d'approche utilisées dans l'imputation [59]. Les avantages de l'imputation **k-NNI** résident dans sa capacité à se passer de la création d'un modèle prédictif pour chaque attribut avec des valeurs manquantes dans l'ensemble de données, de traiter les instances avec plusieurs valeurs manquantes, de prendre en compte la structure de corrélation des données et de prédire les attributs qualitatifs et quantitatifs. L'algorithme K-NNI figure parmi les plus simples algorithmes d'apprentissage artificiels.

## 3.9. Exemple d'application de toolbox MDI et KNNI

### 3.9.1. Application de Toolbox MDI

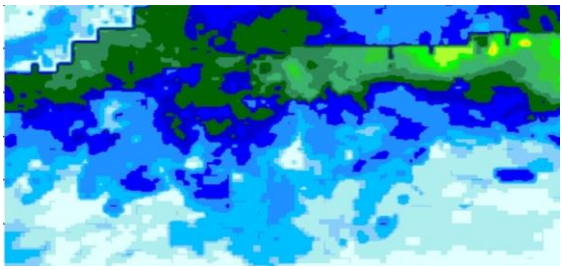

**Figure 3.10** : image réel.

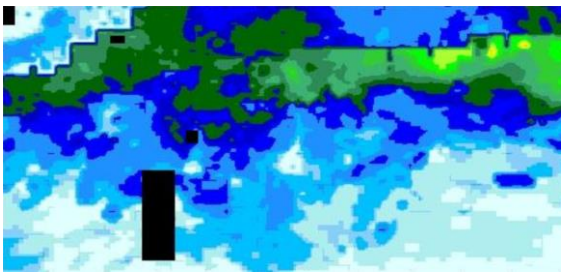

**Figure 3.11** : image avec 5% DM

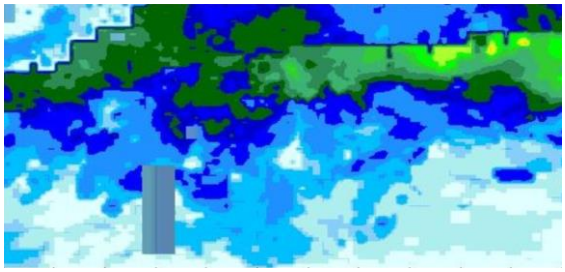

**Figure 3.12** : image traité par KDR.

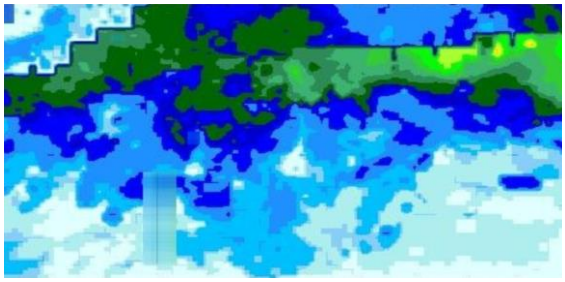

**Figure 3.13** : image traitée par TSR.

Dans cette partie, nous avons représenté une étude descriptive qualitative et quantitative pour un cas d'application de toolbox MDI, appliqué sur 5% de données manquantes, donné par la carte TRMM du nord algérien de la pluviométrie mensuelle de janvier 2019.

La toolbox est appliquée en fonction de deux méthodes, KDR ou TSR, précédemment définies. Dans la carte montrée par la **figure 3.11**, la zone noire représente les pixels manquants. Ils ont été restitués de manière visible afin de mieux comparer les différents résultats. Les quatre images de la figure prouvent que les deux méthodes peuvent traiter le manque de pixels, c'est le cas dans la zone de couleur vert et bleu foncé. Les deux méthodes de la toolbox sont performantes. Par contre, pour la zone bleu-clair, la méthode TSR est la meilleure qui puisse être appliquée. La **figure 3.14** montre les graphes de box plots des résultats obtenus par la toolbox MDI et les données réelles pour chaque matrice de couleurs RGB. Le graphe donne les trois quantiles Q1, Médian, Q3 et la valeur min, max et moyenne. Ces valeurs sont aussi données dans le **tableau 3.2,** avec d'autres critères statistiques, utilisés pour cette comparaison, comme la variance et la moyenne absolue de déviation. La **figure 3.14** et le **tableau 3.2** montrent que dans la matrice bleue, le modèle TSR est généralement plus performant que le modèle KDR, lorsqu'on est dans un cas de comparaison de données réelles, tel que la médiane

et la moyenne de la série de données réelles et celle obtenue par TSR, respectivement égales (94.67, 91.45) et (80.38, 87.00), contrairement au cas KDR où elles sont respectivement égales (24.42, 28.91). Les deux autres matrices (vert et rouge) montrent aussi que le modèle TSR est le plus performant. Par contre avec la matrice verte, le KDR montre une valeur moyenne plus proche des données réelles données par 50.30.

| | **Bleu** | | | **Vert** | | | **Rouge** | | |
|---|---|---|---|---|---|---|---|---|---|
| **Statistique** | **réelle** | **KDR** | **TSR** | **réelle** | **KDR** | **TSR** | **réelle** | **KDR** | **TSR** |
| Min | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| Max | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| Q1 | 89.9408 | 24.4278 | 80.3810 | 70.1961 | 17.9304 | 62.2463 | 12.0000 | 43.5685 | 36.2119 |
| Médian | 94.6746 | 28.9102 | 87.0090 | 92.9412 | 26.1992 | 84.5449 | 77.3333 | 46.1296 | 73.5178 |
| Q3 | 100.000 | 33.2309 | 91.9443 | 93.3333 | 53.5779 | 91.8149 | 80.0000 | 60.8389 | 82.5936 |
| Moyenne | 91.4596 | 33.8572 | 83.9846 | 77.7383 | 35.4526 | 74.3442 | 54.8937 | 50.3040 | 61.6904 |
| Variance | 301.9823 | 446.5368 | 219.0394 | 791.6059 | 507.1214 | 510.8793 | 1410.0697 | 246.0029 | 634.2306 |
| Écart-type | 17.377 | 21.131 | 14.799 | 28.135 | 22.519 | 22.602 | 37.550 | 15.684 | 25.183 |
| Moyenne des écarts absolus | 8.0853 | 13.7700 | 8.9104 | 21.3474 | 18.7605 | 18.0614 | 34.9766 | 11.0694 | 22.6139 |

**Tableau 3.2 :** Statistique descriptive des données RGB réelles et estimées par les méthodes de comblement de toolbox MDI, appliquées sur 5% de données manquantes.

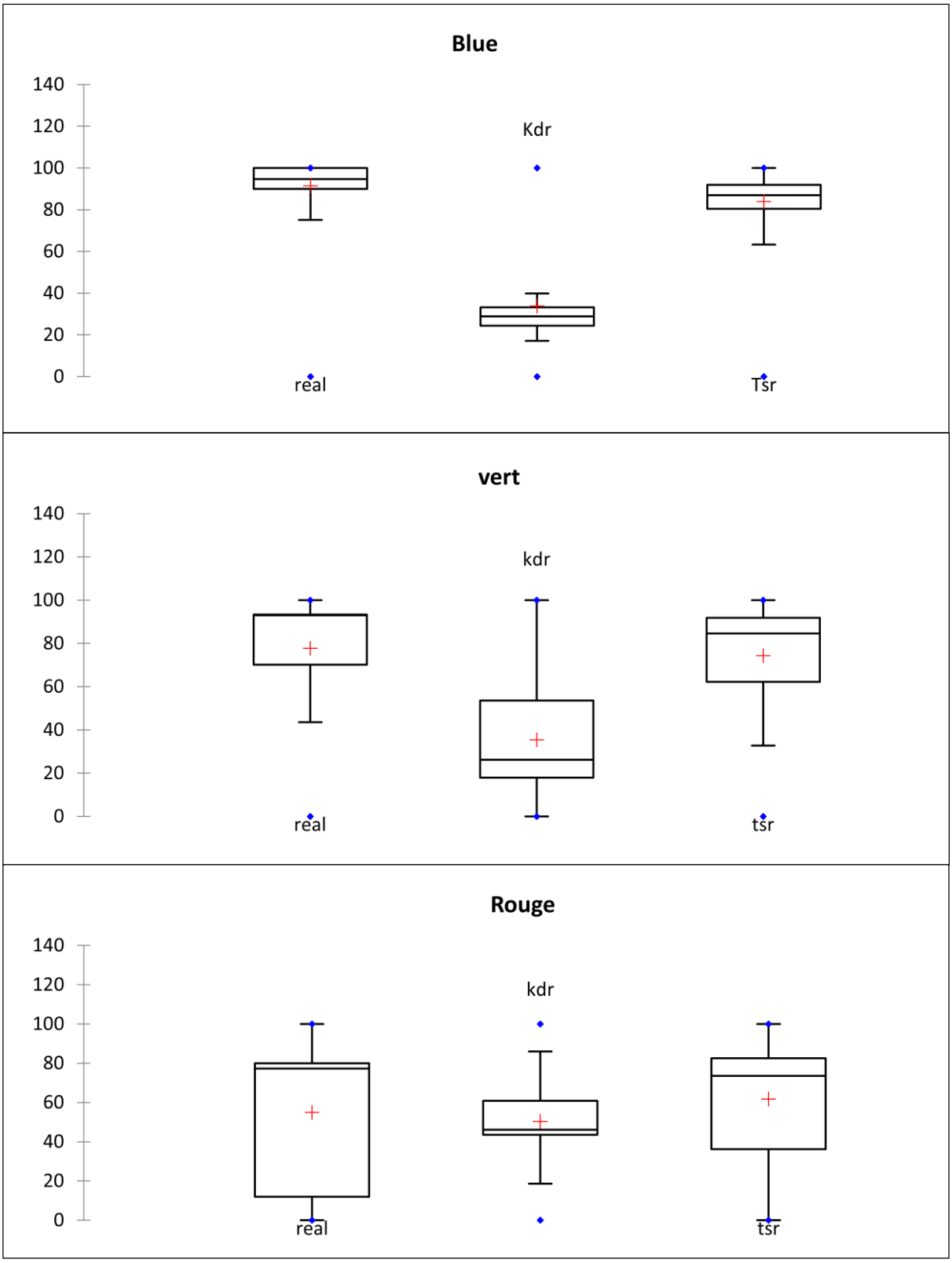


**Figure 3.14 :** Box plot des séries de données RGB réelles et estimées par la méthode KDR et TSR de toolbox MDI.

### 3.9.2. Toolbox k-plus proche voisin KNNI

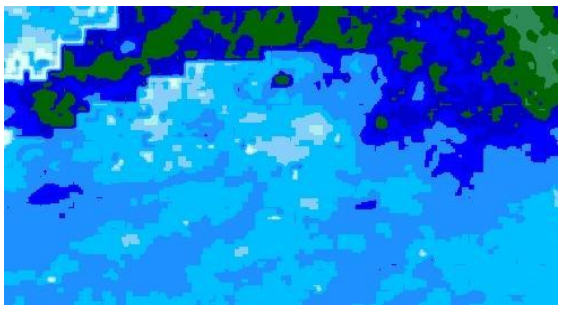

**Figure 3.15** : image réelle.

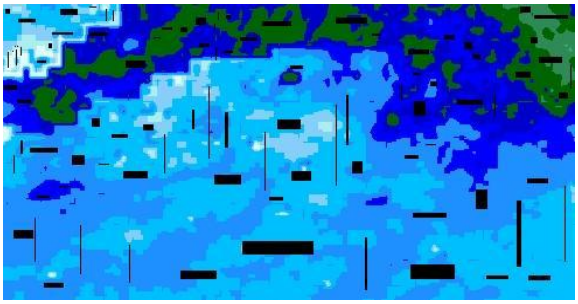

**Figure3.16** : image avec 10% de DM.

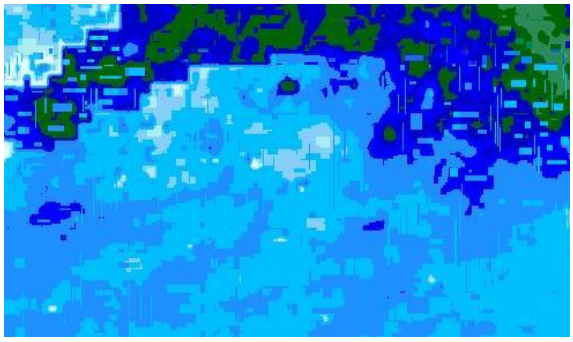

**Figure 3.17** : image traite avec KNNI.

Nous avons utilisé une carte pluviométrique TRMM, titrant 10% de données manquantes (DM), afin d'expliquer statistiquement le fonctionnement de KNNI toolbox. Les résultats de cette étude sont montrés par les figures suivantes : **3.15 ; 3.17 ; 3.18** ainsi que le **tableau 3.3**.

L'image montre que la méthode KNNI est plus convenable pour traiter les zones bleu-clair, ce qu'elle n'arrive pas à faire aux zones bleu-foncé et vert. Les graphes de box plots représentent la distribution des valeurs estimées dans chaque matrice RGB, par rapport aux données réelles des codes décimaux. Les valeurs certifient de la performance de la méthode pour traiter les valeurs manquantes en vert et rouge, lorsqu'on compare avec matrice bleue. La toolbox prouve avoir fourni le meilleur d'elle-même avec le vert plutôt que le rouge, à tel point que les différents critères, comme Q1, médiane et moyenne ont une grande similarité entre les valeurs estimées et réelles de couleur verte, données respectivement par la valeur (62.68, 72.91, 56.56) et (67.50, 79.58, 65.69). D'autre part, la série de donnée estimée de la matrice rouge montre une moyenne similaire avec les données réelles. L'écart-type standard et la moyenne des écarts absolus montrent aussi cette performance. Les valeurs montrent une différence remarquable entre les deux paramètres précédents obtenus par la série de données estimées et réelles de la matrice bleue, donnée respectivement par les valeurs (27.61, 20.42) et (22.54, 12.47).

| | Bleu | | Vert | | Rouge | |
|---|---|---|---|---|---|---|
| **Statistique** | **Réelle** | **KNNI** | **Réelle** | **KNNI** | **Réelle** | **KNNI** |
| Min | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| Max | 100.0000 | 100.0000 | 100.0000 | 100.0000 | 100.0000 | 100.0000 |
| Q1 | 99.6078 | 76.3346 | 60.0000 | 44.6793 | 0.0000 | 4.3191 |
| Médian | 99.6078 | 86.2923 | 67.5000 | 62.6896 | 1.1236 | 6.2240 |
| Q3 | 100.0000 | 87.8446 | 79.5833 | 72.9179 | 16.8539 | 8.7454 |
| Moyenne | 92.4986 | 74.9102 | 65.6947 | 56.5688 | 7.1997 | 8.2027 |
| Variance | 508.2894 | 557.7816 | 369.4504 | 469.7247 | 109.3952 | 74.8603 |
| Écart-type | 22.5453 | 23.6174 | 19.2211 | 21.6731 | 10.4592 | 8.6522 |
| Moyenne des écarts absolus | 12.4780 | 17.4273 | 13.7916 | 17.4579 | 7.8132 | 4.7506 |

**Tableau 3.3 :** Statistique descriptive des données RGB réelles et estimées par la toolbox KNNI, appliquée sur 10% de données manquantes.

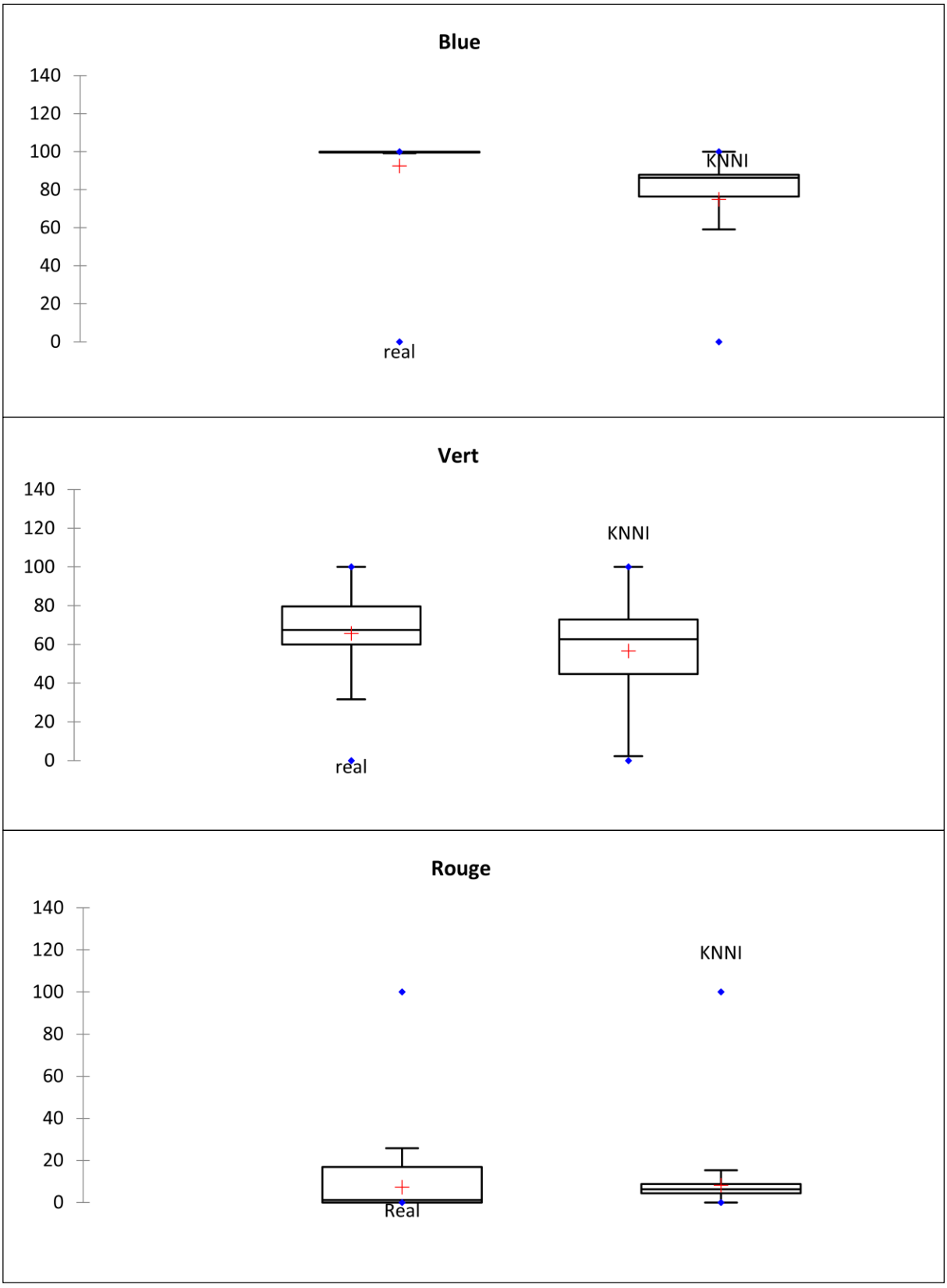


**Figure 3.18 :** Box plot des séries de données RGB réelles et estimés par la toolbox KNNI.

## 3.10. Approche proposée

Notre approche sert à appliquer un réseau artificiel de neurones ANN, où l'information d'entrée est une image TRMM avec données manquantes DM, en utilisant un programme sous Matlab pour extraire les matrices décimales RGB (rouge, vert et bleu). Chaque matrice colorée est ainsi transformée sous forme de vecteur ($C^{B}_{dm}$, $C^{G}_{dm}$ et $C^{R}_{dm}$). Suite à cela, dans la couche cachée, on traite les pixels de chaque vecteur en utilisant la fonction de sommation. On calculera de la sorte la moyenne pondérée de chaque vecteur RGB dans la fonction de transfert en utilisant la méthode de régression et des ***k-vois*** plus proches, estimant ainsi les valeurs manquantes de chaque matrice de couleur qui comblera le jeu de données. Ceci fait, on établira un programme Matlab permettant de concaténer les matrices RGB traitées, image TRMM traitée dans la couche de sortie.

La **figure 3.19** résume les déférentes étapes de notre proposition. Dans la phase d'entrée, c'est une image TRMM avec des données manquantes DM. L'extraction des matrices RGB-**DM$_{TRMM}$** (rouge, vert et bleu) dans cette partie se fait par le biais d'un algorithme, sous Matlab. Chaque matrice de couleur sera ensuite transformée en forme vecteur [un seul tableau de colonne ($C^{B}_{1-n-dm}$, $C^{G}_{1-n-dm}$ et $C^{R}_{1-n-dm}$)]. À ce niveau arrive le tour d'appliquer les fonctions de la méthode de régression et de ***k-plus proche voisin*** aux vecteurs RGB pour évaluer les valeurs manquantes des matrice décimales ( $M^{B}_{-DE}$ , $M^{G}_{-DE}$ et $M^{R}_{-DE}$), ce qui redonnera au jeu de données son architecture complète. En guise de finalisation ou de vérification, un autre algorithme sera conçu, toujours sous Matlab, qui sera employé à concaténer les images RGB-DE$_{TRMM}$ avec des données, de telle sorte à pouvoir estimer le résultat : une image TRMM ne manquant d'aucune donnée.

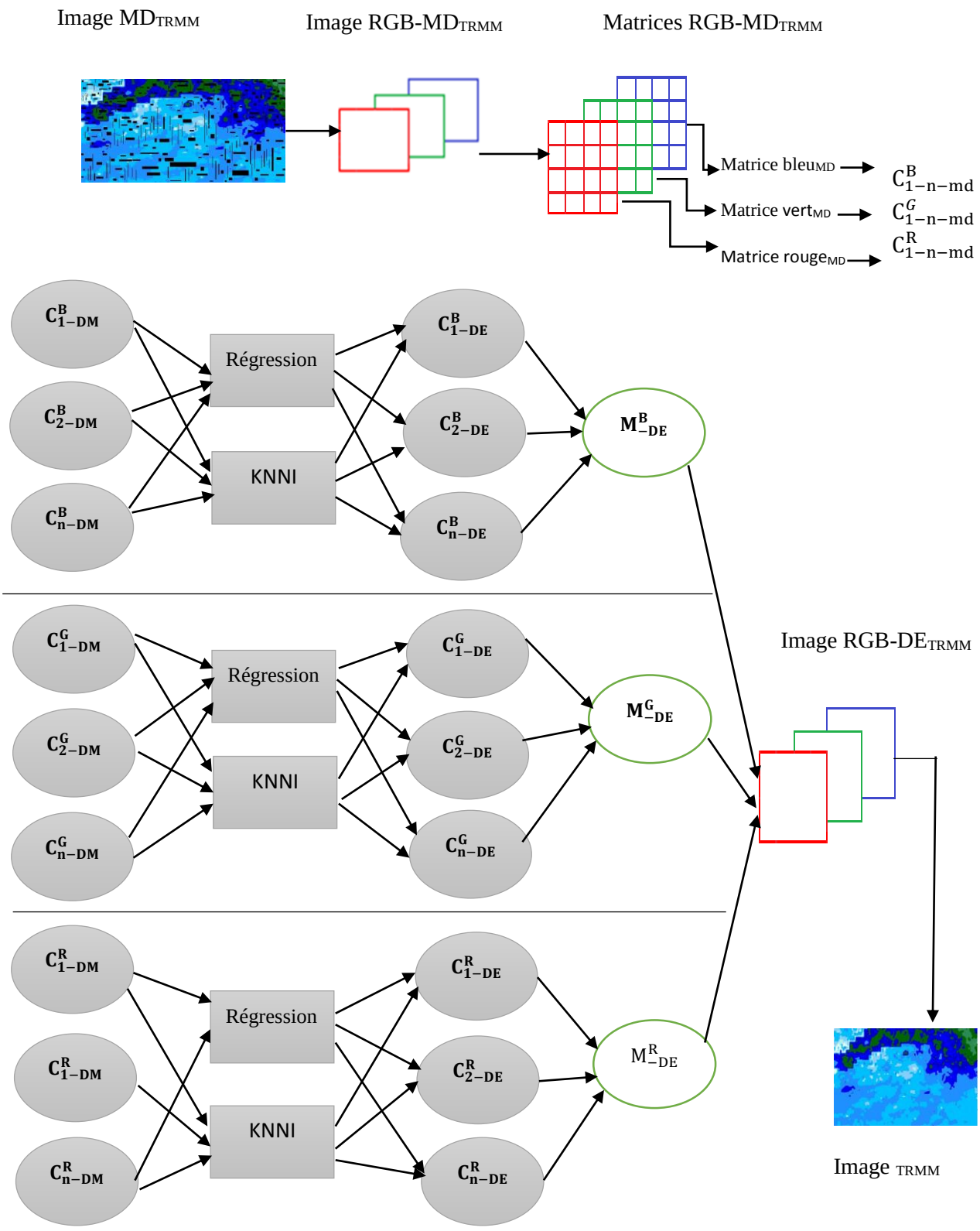


**Figure 3.19 :** schéma représentant le fonctionnement de la méthode ANN proposée.

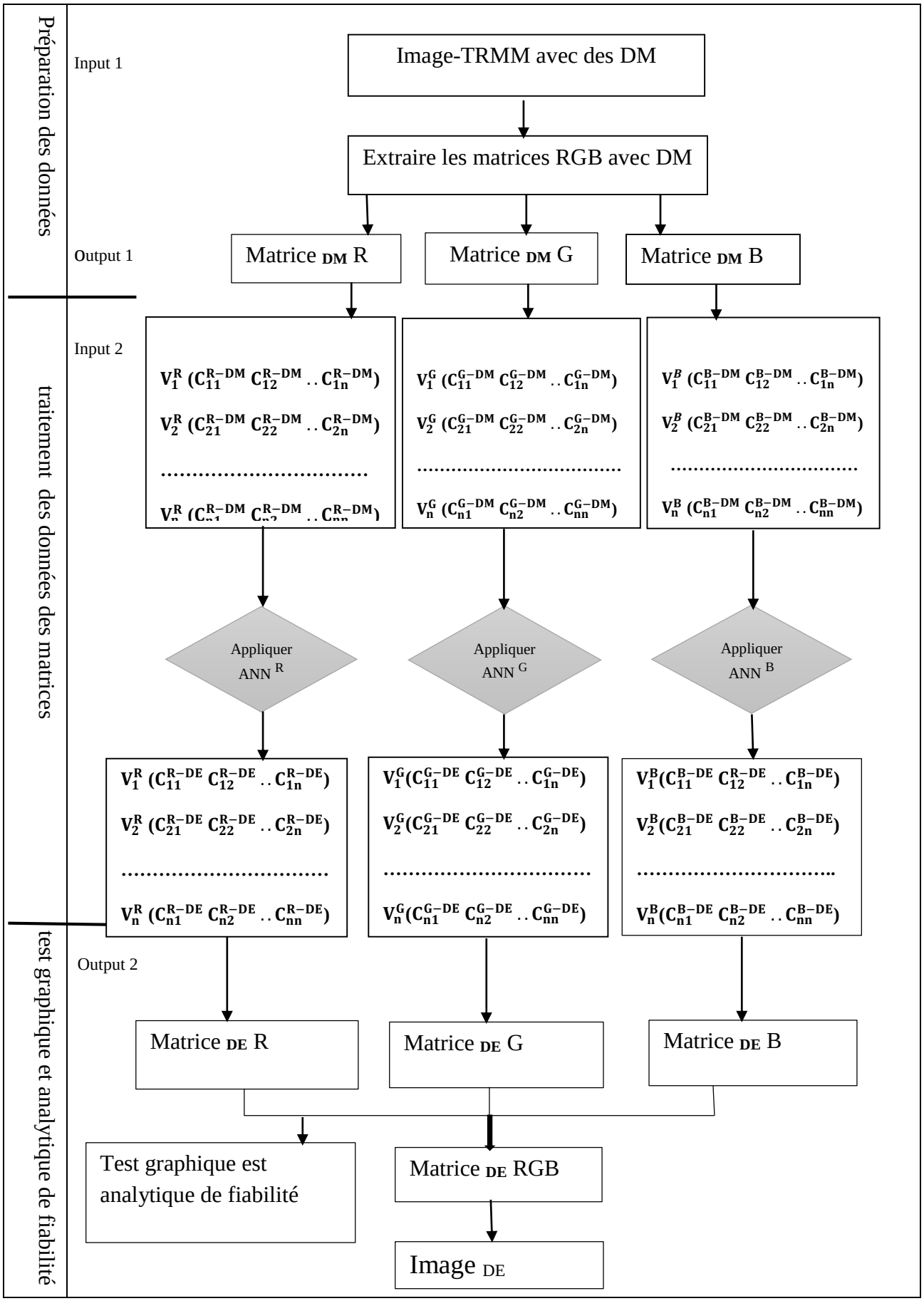


**Figure 3.20 :** organigramme de l'approche proposée.

L'organigramme de la **figure 3.20** représente les trois étapes de notre approche proposée pour estimer l'ampleur du manque de données dans les images TRMM.

Préparation des données : l'information d'entrée, ***« input »***, est une image TRMM manquant de données. Dans cette étape aura lieu l'extraction des matrices de couleur RVB accusant le manque de données, ce qui se réalisera par le moyen d'un algorithme mis au point sous Matlab. Le résultat ***« output »*** est : matrice $_{DM}$ R, matrice $_{DM}$ V et matrice $_{DM}$ B).

Traitement des données des matrices : dans cette partie (input 2), chaque matrice de couleur sera transformée sous forme vecteur, ce qui nous donnera, en principe, trois réseaux de neurone neurones artificiel ($ANN^R$, $ANN^V$ et $ANN^B$ ). Pour chaque ANN seront appliquées des fonctions estimatrices des données manquantes à chaque vecteur RVB, ce qui conduirait à rendre toute son envergure au jeu de données (output 2).

Test de fiabilité graphique et analytique : les données des vecteurs de chaque matrice de couleur seront estimées puis traitées avant d'être soumises à un autre algorithme (sous Matlab) qui permettra de concaténer les matrices RGB des trois couleurs pour déboucher sur le résultat final Image $_{DE}$. La fiabilité de notre proposition d'approche, celle d'un réseau, sera soumise, une fois de plus, à l'épreuve de tests graphiques et analytiques.

## 3.11. Traitement des données manquantes avec les ANN

### 3.11.1. Définition des ANN

Les réseaux de neurones artificiels sont un écheveau de connexions aux processeurs élémentaires fonctionnant en parallèle. Chaque processeur élémentaire calcule une sortie unique sur la base des informations qu'il reçoit [48]. Il est utilisé pour résoudre des problèmes complexes dans de nombreuses applications. La structure ANN se compose de trois couches [49]: une couche d'entrée, collectrice des données ; une autre de sortie, génératrice d'informations calculées, puis une ou plusieurs couches cachées appropriées pour connecter la couche d'entrée et de sortie. Un neurone est une unité de traitement de base d'un NN, remplissant deux fonctions: la collecte des entrées et la production pour la sortie. Chaque entrée est multipliée par les poids de connexion, ses produits et leurs biais sont ajoutés puis passés par une fonction d'activation pour produire une sortie comme indiqué sur la figure (3.21).

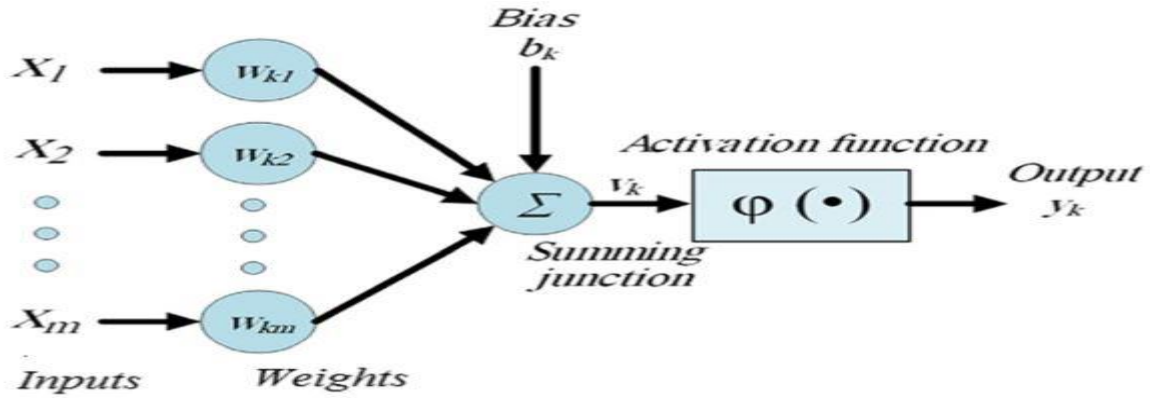


**Figure 3.21 :** Simple réseau de neurones artificiels RNA.

### 3.11.2. Application des ANN sur la matrice RGB

#### 3.11.2.1. Input layer

Les entrées sont des matrices RVB (rouge, vert, bleu) extraites grâce à un programme mis au point sous Matlab depuis une image TRMM (voir l'algorithme 1). La reconstitution de ces matrices s'obtient par l'intervention d'un algorithme créant une nouvelle carte déficitaire en données.

**Algorithme 1 :**

**im= imread('image.jpg');** lire une image RGB

**size (im) ; ans = 392 916 3** dimension de l'image.

**imRed =im(:,:,1);** extraction de la matrice rouge.

**imGreen=im(:,:,2);** extraction de la matrice green.

**imBlue =im(:,:,3);** extraction de la matrice bleue.

**subplot(2,2,1); imshow(im); axis image;** title('image originale 3d');

**subplot(2,2,2); imshow(imRed);** title('red');

**subplot(2,2,3); imshow(imGreen);** title('green');

**subplot(2,2,4); imshow(imBlue);** title('blue');

**Xlrange1='A1' Xlrange1 =A1**

**xlswrite('matRED.xlsx',imRed, 1,'A1');** exporter la matrice rouge dans un fichier Excel.

**xlswrite('matGREEN.xlsx',imGreen, 1,'A1');** exporter la matrice verte dans un fichier Excel.

**xlswrite('matBLUE.xlsx',imBlue, 1,'A1');** exporter la matrice bleue dans un fichier Excel.

### 3.11.2.2. Hidden layer

**A / Fonction de transfert**

Dans la fonction de transfert, la somme totale peut être comparée à un certain seuil pour déterminer la sortie neuronale. Si la somme est supérieure à la valeur-seuil, l'élément de traitement génère un signal. Si la somme des produits d'entrée et de poids est inférieure au seuil, aucun signal ne sera généré [50]. Différentes fonctions de transfert peuvent être utilisées comme fonction d'activation du neurone [51].

- **La fonction-seuil :** Entrecoupée de fortes discontinuités, sans linéarité, reçoit un seuil à son entrée. Plus précisément, une entrée négative ne passera pas le seuil, la fonction retournera alors la valeur 0 (on peut interpréter ce 0 comme signifiant *faux*). Une entrée positive ou nulle ne dépassera pas le seuil, et la fonction retournera à un (*vrai*).

- **La fonction linéaire :** tout à fait linéaire, exempte de tout changement de pente, assez simple, elle affecte directement son entrée à sa sortie. Elle se transcrit comme suit :

$$Y = s \tag{3.5}$$

Appliquée dans le contexte d'un neurone, dans le cas, la sortie du neurone correspond à son niveau d'activation dont le passage à zéro se produit lorsque $w^T x = b$.

- **Fonction de sigmoïde :** c'est un compromis intéressant entre les fonctions seuil et linaire. Son équation est donnée par :

$$y = \frac{1}{1+ exp^{-s}} \tag{3.6}$$

**B / Méthode implémentée dans ANN**

- **L'imputation par k-plus proche voisin (KNNI)**

KNNI est une méthode de traitement et l'imputation des données manquantes repose sur la recherche de ***k observations les plus proches*** (respectives de l'observation qui doit être imputée). Elle remplace la valeur manquante par la moyenne des ***k observations trouvées*** (ou par la valeur la plus fréquente parmi les ***k voisins les plus proches*** en cas de variable discrète) [54]. Les ***k voisins les plus proches*** sont déterminés lorsqu'on aura trouvé la distance entre les

cas complets et ceux incomplets qui comparent le degré de similitude entre eux. La distance de Manhattan et la distance euclidienne [55] ont été utilisées pour trouver les cas les plus proches.

- **Imputation par la méthode de régression**

C'est une méthode en deux étapes : d'abord, on estime les rapports entre les attributs, puis on emploie les coefficients de régression pour estimer la valeur manquante [18]. La condition fondamentale de l'utilisation de l'imputation par régression est l'existence d'une corrélation linéaire entre les attributs [56]. La technique suppose également que les valeurs soient MAR. Dans le contexte de valeurs manquantes, deux modèles de régression sont généralement appliqués : la régression linéaire et la régression logistique. Cette dernière est utilisée plutôt pour traiter les variables discrètes, laissant à la régression linéaire le traitement des variables continues.

**C / Environnement de simulation**

La mise en chantier de notre approche nous a conduits à élire les deux environnements de travail suivants :

- **Logiciel Matlab :** ce logiciel (Matrix Laboratoire) est spécialisé dans le domaine du calcul matriciel numérique produit par Math-Works (voir le site web http://www.mathworks.com/). Il est disponible sur plusieurs plateformes. Matlab est un langage alliant simplicité et efficacité, optimisé pour le traitement des matrices ou le calcul numérique. Un de ses avantages réside dans sa concision, qui fait défaut à d'autres langages (C, Pascal, etc.). Il optimise le code des programmes en utilisant des fonctions prédéfinies. Ses interfaces graphiques sont puissantes. On peut enrichir Matlab en ajoutant des "boîtes à outils" (toolbox), ensembles de fonctions supplémentaires profilées pour des applications particulières (traitement de signaux, analyses statistiques, optimisation, etc.). L'ordre d'exécution des instructions est déterminé par des structures de contrôle. Sur notre chemin pour réaliser notre expérience, Matlab n'avait pas tardé à s'illustrer comme le plus approprié.
- **Logiciel SPSS :** *in extenso* **S**tatistique **P**ackage for **S**ocial **S**ciences, est un logiciel utilisé pour l'analyse statistique des séries de données. Il est tout à fait fonctionnel sur un système Windows. Il peut être utilisé pour effectuer la saisie et l'analyse des données, de même que pour créer tableaux et graphiques. Il est capable de gérer de grandes quantités de données, aussi peut-il effectuer toutes les analyses couvertes dans les textes. De nombreuses méthodes dans SPSS fournissent au chercheur certaines techniques

statistiques pour estimer les valeurs manquantes. Il s'agit notamment de l'algorithme EM, régression, moyenne série, etc.).

- **Script Xlstat :** développé depuis une décennie environ, ce script a pour but de faciliter l'analyse statistique des séries de données. Il est out à la fois complet et convivial. Son interface s'appuie entièrement sur Microsoft Excel tant pour la récupération des données que pour la restitution des résultats. La plupart des outils de XLSTAT comporte un angle pour le traitement des données manquantes. Néanmoins, les méthodes disponibles sont peu nombreuses. Cet outil permet un prétraitement des données, complétant les manques avec des méthodes avancées. Il s'agit notamment de l'imputation par la moyenne, une approche de voisin le plus proche (KNNI), l'algorithme NIPALS et la méthode d'imputation multiple, etc.
- **Fonction de sommation**

La première étape de l'opération d'un élément de traitement consiste à calculer la somme pondérée de toutes les entrées. Mathématiquement, les entrées et les poids correspondants sont des vecteurs représentables (par i1, i2... In) et (w1, w2... Wn). L'entrée totale signale le point, ou produit interne, de ces deux vecteurs [50]. Cette fonction de sommation simpliste est trouvée en multipliant chaque composante du vecteur ***i*** par la composante correspondante du vecteur ***w*** puis en additionnant tous les produits. Input1 = i1 * w1, input2 = i2 * w2, etc., sont ajoutés comme input1 + input2 + . . . + inputn. Le résultat est un seul nombre, pas un vecteur à plusieurs éléments.

Il arrive que la fonction de sommation soit plus complexe que la simple somme d'entrées et de poids des produits. Les coefficients d'entrée et de pondération sont combinables de nombreuses manières avant de passer à la fonction de transfert.

- **Définition du coefficient de détermination $R^2$**

Le coefficient de détermination, également appelé coefficient de corrélation multiple, s'est bien établi dans l'analyse de régression classique [52]. Il est défini en tant que somme des carrés due à la régression divisée par la somme des carrés totaux. Habituellement, R2 est interprété comme représentant le pourcentage d'écart de la variable dépendante expliquée par la fluctuation des variables indépendantes. Il s'écrit [53] :

$$R^2=1-\frac{SCR}{SCT} \qquad (3.7)$$

SCR est la somme des carrés des résidus. SCT est la somme des carrés totaux.

### 3.11.2.3. Output layer

Après l'estimation des données manquantes dans la matrice décimale RGB (rouge, vert, bleu), ce sera le tour de l'étape d'application sous Matlab d'un algorithme ($2^e$) qui nous permettra de créer une image TRMM avec des données estimées.

**Algorithme 2**

**R= xlsread('matriceRED.xlsx') :** lire la matrice rouge.

**G= xlsread('matriceGREEN.xlsx') :** lire la matrice verte.

**B= xlsread('matriceBLUE.xlsx') :** lire la matrice bleue.

**my_image(:,:,1)=R :** image rouge.

**my_image(:,:,2)=G :** image verte.

**my_image(:,:,3)=B :** image bleue.

**imwrite(my_image,'Image1.jpg') :** image RGB trmm.

## 3.11. Exemple d'application de la méthode proposée ANN

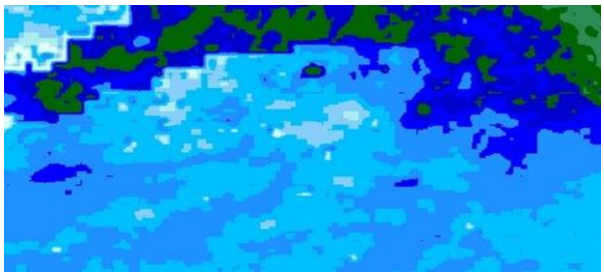

**Figure 3.22**: image réelle.

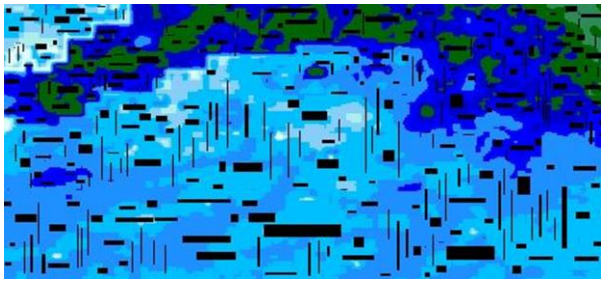

**Figure 3.23** : image avec 10% de DM.

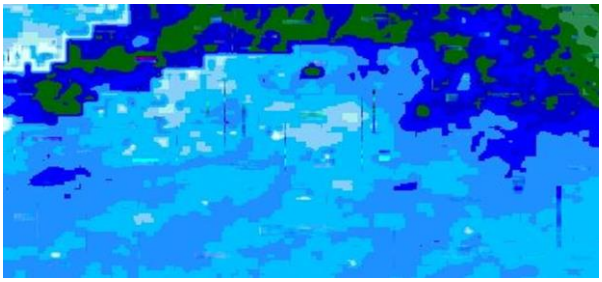

**Figure 3.24** : image traité avec la méthode de régression.

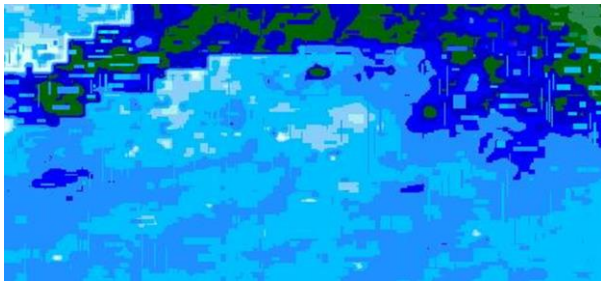

**Figure 3.25** : image traitée avec KNNI.

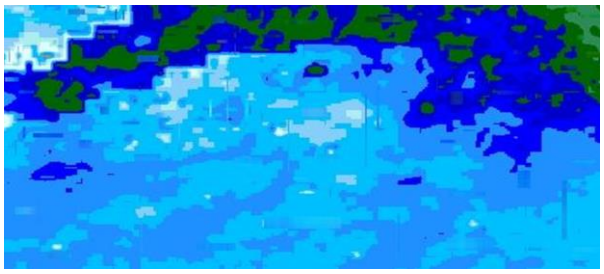

**Figure 3.26** : image traitée avec ANN

L'exemple d'application de la méthode proposée est donné par la **figure 3.26.** Des détails y sont donnés sur les résultats de chaque méthode appliquée sur 10% de données manquantes de l'image TRMM de la pluviométrie mensuelle du nord du pays, durant le mois de mai 2019. La figure montre l'image des données réelles, celle des données manquantes, l'image estimée par chaque méthode de transfert (KNNI et régression) ainsi que l'image finale du traitement obtenue par la proposition ANN. Dans cette figure, on peut observer que la méthode de régression est idéale pour traiter les pixels manquants dans la zone bleue foncée et verte. Par contre, la méthode KNNI convient mieux à traiter les pixels de la zone bleu clair. Dans cet exemple d'application, l'approche proposée donne la meilleure estimation, elle est capable de traiter les pixels de toutes les zones coloriées. La **figure 3.27** et les **tableaux 3.4, 3.5 et 3.6** montrent une étude comparative entre les valeurs décimales estimées et réelles pour chaque

matrice RGB, obtenues en utilisant les mêmes méthodes citées dans la figure précédente. Les graphes de Box plot montrent que la meilleure distribution des données estimées par rapport aux données réelles, est celle délivrée par la méthode ANN pour chaque matrice RGB. Les tableaux montrent que tous les critères utilisés pour cette comparaison sont d'une très forte similitude entre la méthode ANN et les données réelles. Lorsqu'on compare avec d'autres modèles, comme KNNI et régressif, les deux derniers s'avèrent plus performants pour traiter les pixels manquants. La méthode KNNI ne montre aucune tendance d'erreur d'estimation selon les résultats des trois matrices RGB, où la moyenne des erreurs absolues montre une différence entre les valeurs réelles et estimés égale à 1.

| | **Blue** | | | |
|---|---|---|---|---|
| **Statistique** | **régression** | **KNNI** | **ANN** | **Réelle** |
| Min | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| Max | 100.0000 | 100.0000 | 100.0000 | 100.0000 |
| Q1 | 91.3725 | 99.6078 | 78.0203 | 73.3607 |
| Médiane | 99.2157 | 99.6078 | 87.2687 | 86.6175 |
| Q3 | 100.0000 | 100.0000 | 88.7104 | 88.2068 |
| Moyenne | 84.9863 | 92.4986 | 76.6596 | 76.6379 |
| Variance | 886.0396 | 508.2894 | 486.4256 | 409.7966 |
| Écart-type | 29.7664 | 22.5453 | 22.0551 | 20.2434 |
| Moyenne des écarts absolus | 21.1438 | 12.4780 | 16.2435 | 15.0803 |

**Tableau 3.4 :** Statistique descriptive de la matrice bleue des données réelles et estimées par l'approche ANN.

| | **Vert** | | | |
|---|---|---|---|---|
| **Statistique** | **régression** | **KNNI** | **ANN** | **Réelle** |
| Min | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| Max | 100.0000 | 100.0000 | 100.0000 | 100.0000 |
| Q1 | 54.1176 | 60.0000 | 51.1367 | 55.7781 |
| Médiane | 60.0000 | 67.5000 | 67.0447 | 60.2430 |
| Q3 | 74.5098 | 79.5833 | 76.0791 | 63.0878 |
| Moyenne | 56.0290 | 65.6947 | 61.5532 | 58.7153 |
| Variance | 607.0472 | 369.4504 | 374.9876 | 53.4323 |
| Écart type | 24.6383 | 19.2211 | 19.3646 | 7.3097 |
| Moyenne des écarts absolus | 18.4576 | 13.7916 | 15.5275 | 5.0476 |

**Tableau 3.5 :** Statistique descriptive de la matrice verte des données réelles et estimées par l'approche ANN.

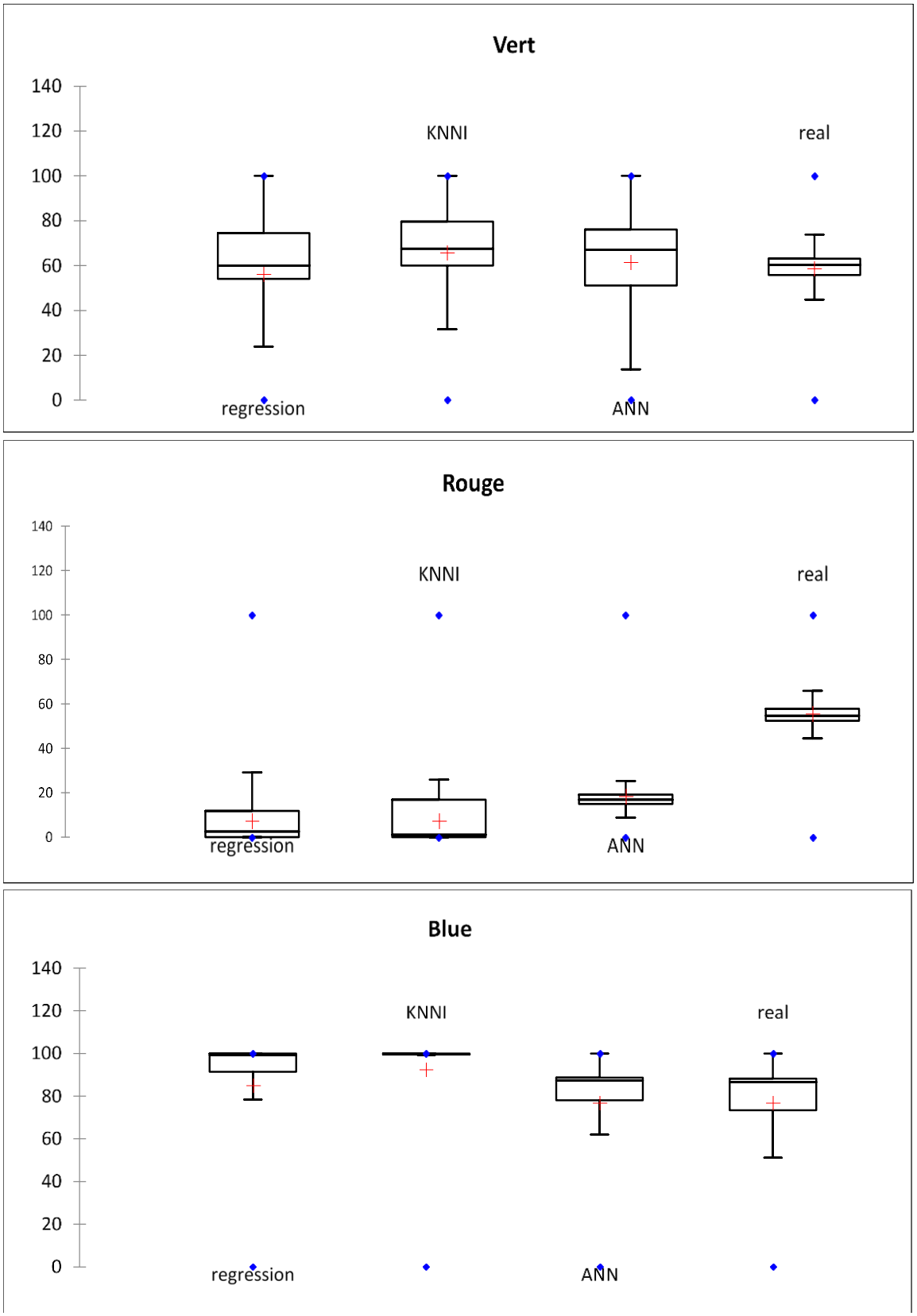


**Figure 3.27 :** Box plot des séries de données RGB réelles et estimés par l'approche ANN proposé.

| | Rouge | | | |
|---|---|---|---|---|
| **Statistique** | **Régression** | **KNNI** | **ANN** | **réelle** |
| Min | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| Max | 100.0000 | 100.0000 | 100.0000 | 100.0000 |
| Q1 | 0.0000 | 0.0000 | 14.9973 | 52.3906 |
| Médiane | 2.5316 | 1.1236 | 16.8886 | 54.7080 |
| Q3 | 11.8143 | 16.8539 | 19.1416 | 57.7862 |
| Moyenne | 7.2235 | 7.1997 | 18.3028 | 55.5728 |
| Variance | 142.7672 | 109.3952 | 64.0264 | 38.5775 |
| Écart-type | 11.9485 | 10.4592 | 8.0016 | 6.2111 |
| Moyenne des écarts absolus | 7.5881 | 7.8132 | 4.4387 | 4.1194 |

**Tableau 3.6:** Statistique descriptive de la matrice rouge des données réelles et estimées par l'approche ANN.

## 3.12. Conclusion

Ce chapitre a donné lieu à la présentation des toolbox MDI et KNNI, la description des environnements de leur simulation, présentation étayée d'exemples d'application de chacune des deux boîtes. Aussi nous sommes étalés sur le principe de fonctionnement de la méthode que nous proposons (ANN). La circonstance ne manque pas d'inviter également, pour plus de détails explicatifs, à présenter exhaustivement quelque étapes à suivre et des algorithmes à employer pour restituer les pixels manquants dans l'imagerie TRMM de façon à reconstituer intégralement le puzzle qu'est le jeu de données. Un aperçu exhaustif y est aussi donné à travers un exemple applicatif pour le réseau de fiabilité de l'approche que nous préconisons. Il sera à présent question, dans le chapitre prochain, de la validation de notre approche, appuyée sur une étude comparative entre les toolbox MDI et KNNI, attestant de sa fiabilité.

# Chapitre 4

# Chapitre 4 : Validation et comparaison

## 4.1. Introduction

Avant de nous lancer dans la plaidoirie pour la solution que nous avions mise au point, nous l'avions amplement éprouvée par des tests techniques jalonnant son étude descriptive quantitative et qualitative au sujet de sa performance selon les différents mécanismes du manque de données MAR, MCAR et NMAR. L'ensemble de ces dispositions, circonscrites à cette limite de pourcentage de données : 15% et 30%, représente à nos yeux la seule façon d'en savoir assez amplement sur sa faisabilité d'abord puis sa fiabilité. Tableaux et graphes contenus dans ce chapitre détaillent justement, étape par étape, les résultats des essais de notre approche en question. La comparaison a aussi porté sur d'autres solutions rencontrées dans la littérature relative au sujet, comme les toolbox MDI et KNNI.

Notre approche (ANN) ouvre sur la perspective de résolution de la problématique que posent, chacun à sa manière, les deux phénomènes que sont la perte et la carence de données dans les images TRMM. Notre méthode s'est avérée capable de remonter contextuellement des données, façon la plus convenable d'identifier l'instant précis où elles subirent la modification ayant creusé le manque. Elle est également capable d'esquisser assez fidèlement la structure originelle tronquée au niveau des images TRMM. Elle procède de la vérification de leur cohérence d'ensemble par la mise en interaction de toutes les données disponibles.

Le chapitre est organisé comme suit : une partie expérimentale, dans la section 4.2, décrivant les différents tests statistiques. La section suivante (4.3) donne lieu à deux comparaisons, l'approche proposée confrontée à la toolbox MDI et KNNI. Ensuite, nous présentons les résultats obtenus à fin de discussion pour clore le chapitre par une conclusion.

## 4.2. Partie expérimentale

Dans cette partie en va donner une présentation des différente teste statistique qui sont appliquer dans notre étude.

- **Définition de $R^2$ ajusté :** R-carré ajusté est une version modifiée du R-carré ajusté pour le nombre de prédicteurs dans le modèle. Le R-carré ajusté n'augmenterait que si le nouveau terme améliorait le modèle plus que ne le permettent les prévisions raisonnablement attendues [60]. Il diminuerait lorsqu'un prédicteur améliorait le modèle moins que prévu par hasard. Il est toujours inférieur au R-carré. Il s'écrit [53]:

$$R^2_{Adj} = 1 - \frac{N-1}{N-p-1}(1 - R^2) \quad (4.1)$$

$\boldsymbol{R^2}$ : Coefficient de détermination, $\boldsymbol{P}$ : nombre de prédiction, $\boldsymbol{N}$ : taille d'échantillon.

- **Définition de l'erreur absolue moyenne (MAE) :** elle mesure l'ampleur moyenne des erreurs dans un ensemble de prédictions, sans tenir compte de leur direction. Il s'agit de la moyenne sur l'échantillon de tests des différences absolues entre la prédiction et l'observation réelle où toutes les différences individuelles ont le même poids [61].il s'écrit [62] :

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}\left(|y_j - \hat{y}_j| \right) \quad (4.2)$$

Ou $y_j$ prédiction, $\hat{y}_j$valeur réelle et ***N*** nombre total de point de données.

- **Définition de l'Erreur quadratique moyenne (RMSE):** elle est une règle de notation quadratique qui mesure également l'ampleur moyenne de l'erreur [61]. C'est la racine carrée de la moyenne des différences carrées entre la prédiction et l'observation réelle il s'écrit [62]:

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)} \quad (4.3)$$

Ou $y_j$ prédiction, $\hat{y}_j$valeur réelle et ***N*** nombre total de point de données.

- **Analyse de régression :** C'est un test statistique graphique obtenu par la méthode du moindre carrée pour analyser les relations entre différentes variables (dépendantes et indépendantes). Elle permet de savoir le degré de corrélation via le coefficient de détermination ($R^2$). Les analyses de régression et de corrélation sont considérées comme un volet de méthodes analytiques multi-variables et elles sont utilisées dans des domaines très différents.
- **Quantile-quantile plot (QQ plot) :** Le tracé quantile-quantile (tracé Q-Q) est un outil graphique efficace pour analyser les fonctions de distribution [63]. Un diagramme QQ peut être utilisé pour comparer des échantillons entre eux ou un échantillon avec une distribution théorique.

## 4.3. Résultat et comparaisons

Le travail dans cette partie s'est appliqué sur un ensemble de paramètres statistiques de résultats de traitement de pixels manquants. Pour un taux de 15%, choisi selon trois mécanismes de données (MAR, MCAR et NMAR). L'objectif de cette étude est de classifier en deux phases pour :

- étudier les tendances des erreurs de notre modèle ANN ;
- prouver le degré de performance de modèle par rapport aux modèles KNNI et MDI.

Cette comparaison est basée techniquement sur le graphe des quantiles QQ qui montre le degré de régression entre le modèle expérimental et le modèle de distribution théorique de la loi normale comme référence. Différents critères de performance ont été appliqués pour quantifier cette comparaison, comme $R^2$, $R^2_{Adj}$, RMSE et MAE. La **figure 4. 1** montre que le modèle ANN a un ajustement proche de la distribution de données réelles, dans chaque cas des trois matrices RGB avec le manque de données MAR. L'intervalle de quantile de notre modèle montre la même distribution de valeurs sur le même ordre de grandeur avec les données réelles, ou le modèle est plus performant par rapport aux boites à outils KNNI et MDI.

Le **tableau 4.1** montre le degré de performance des trois modèles pour traiter les valeurs décimales manquantes de la matrice bleue des pixels, tel que $R^2$ et $R^2_{Adj}$ sont égaux, respectivement à 0.91et 0.90, lorsqu'on applique le modèle ANN. D'autre part, le **tableau 4.1** montre aussi que la toolbox KNNI est plus performante que la MDI. L'application des trois modèles sur les deux autres matrices RGB donne à constater les mêmes contraintes (15% DM et mécanisme MAR). Les modèles montrent les mêmes performances. Le RMSE et le MAE utilisés pour étudier la fluctuation des erreurs, qui peut affecter chaque modèle durant le traitement des trois matrices (**Figure 4.1**), montrent que l'intervalle des erreurs minimales est celui montré à chaque cas par l'approche ANN, intervalle donné respectivement par [15.33, 30.74] et [6.24, 2.51].

La **figure 4.2** et le **tableau 4.2** donnent la même étude de comparaison, appliquée sur les trois modèles de traitement (approche ANN, toolbox KNNI et toolbox MDI), sur une image TRMM de 15% pour suivre la distribution de données MCAR. Les graphes des quantiles QQ montrent que dans ce cas, le modèle ANN est toujours plus efficace, plus remarquable graphiquement dans le traitement de la matrice verte lorsqu'on compare avec les résultats de cette approche avec les toolbox KNNI et MDI.

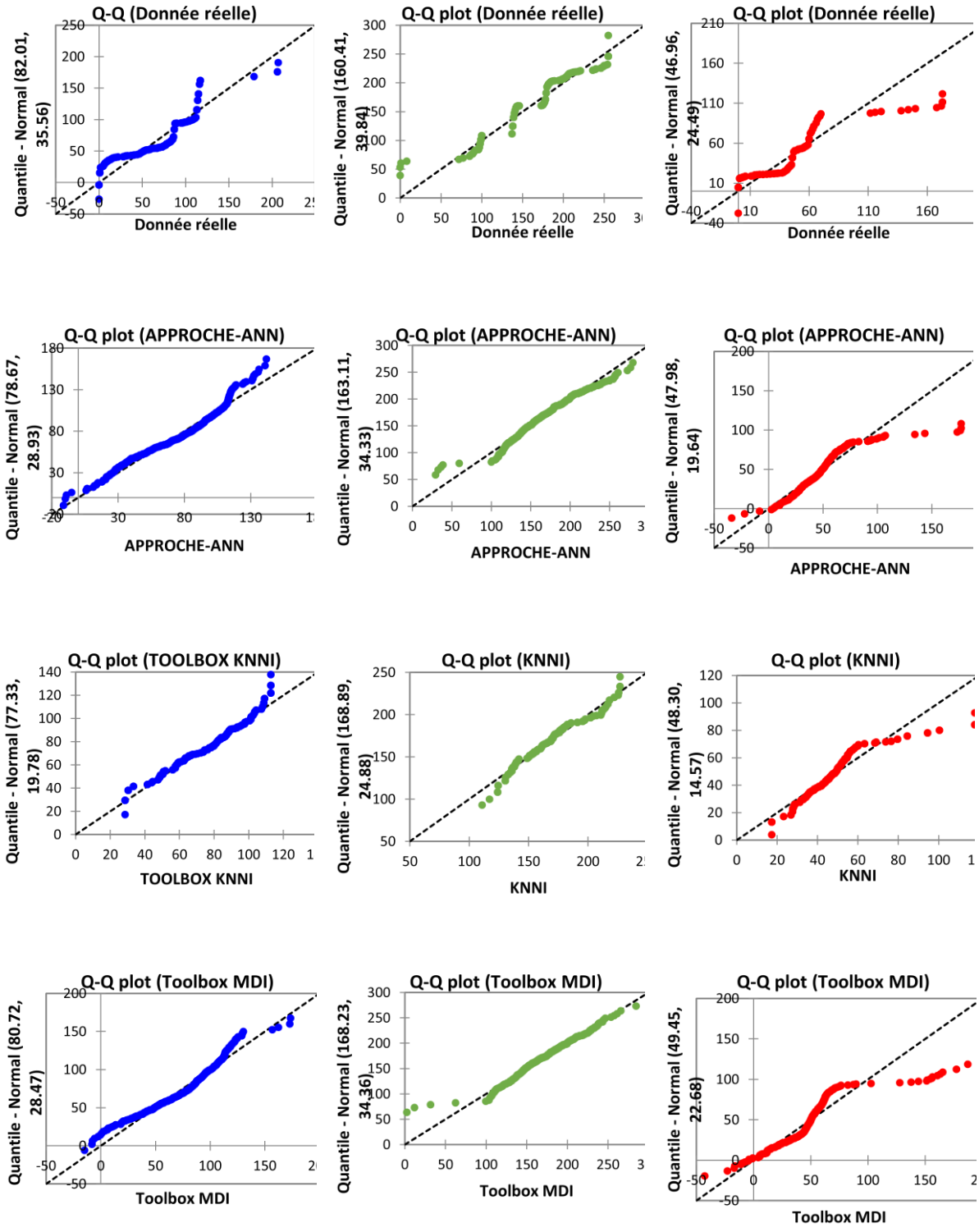


**Figure 4.1** : Graphe de QQ montre la distribution des quantiles expérimentaux des données estimés par le toolbox KNNI, MDI et l'approche ANN par rapport aux quantiles théorique de la loi normale. Cas : de 15% de donnée manquante, mécanisme MAR.

| MAR | Blue | | | | Vert | | | | Rouge | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Variable** | **Real** | **Approche-ANN** | **Toolbox-KNNI** | **Toolbox-MDI** | **Real** | **Approche-ANN** | **Toolbox-KNNI** | **Toolbox-MDI** | **Real** | **Approche-ANN** | **Toolbox-KNNI** | **Toolbox-MDI** |
| **Observations** | 604.00 | 604.00 | 604.00 | 604.00 | 604.00 | 604.00 | 604.00 | 604.00 | 604.00 | 604.00 | 604.00 | 604.00 |
| **Missing data (%)** | 0.00 | 15.00 | 15.00 | 15.00 | 15.00 | 15.00 | 15.00 | 15.00 | 15.00 | 15.00 | 15.00 | 15.00 |
| **Std. deviation** | 35.59 | 32.95 | 19.79 | 28.49 | 39.88 | 37.36 | 24.90 | 34.39 | 24.51 | 19.66 | 14.58 | 22.69 |
| **Mean** | 82.01 | 79.67 | 77.33 | 65.72 | 160.41 | 163.11 | 178.89 | 128.23 | 46.96 | 47.98 | 48.30 | 49.45 |
| **Maximum** | 207.00 | 173.87 | 112.90 | 141.58 | 255.00 | 269.81 | 227.15 | 284.65 | 173.00 | 177.13 | 117.88 | 187.36 |
| **Minimum** | 0.00 | 11.05 | 28.56 | 14.92 | 0.00 | 29.19 | 110.96 | 2.71 | 0.00 | -33.90 | 17.44 | -42.45 |
| **R2** | 1.00 | 0.91 | 0.67 | 0.61 | 1.00 | 0.93 | 0.60 | 0.63 | 1.00 | 0.94 | 0.69 | 0.67 |
| **R2Adj** | 1.00 | 0.90 | 0.66 | 0.59 | 1.00 | 0.91 | 0.59 | 0.62 | 1.00 | 0.92 | 0.67 | 0.66 |
| **RMSE** | 0.00 | 30.74 | 75.33 | 102.11 | 0.00 | 15.33 | 121.33 | 87.33 | 0.00 | 21.20 | 64.11 | 82.50 |
| **MAE** | 0.00 | 12.51 | 30.67 | 41.57 | 0.00 | 6.24 | 49.40 | 35.55 | 0.00 | 8.63 | 26.10 | 33.59 |

**Tableau 4.1 :** Statistiques descriptives des résultats de comblement de 15% de données manquantes, pour chaque matrice RGB par les toolboxes MDI, KNNI et la méthode ANN dans le cas du mécanisme MAR.

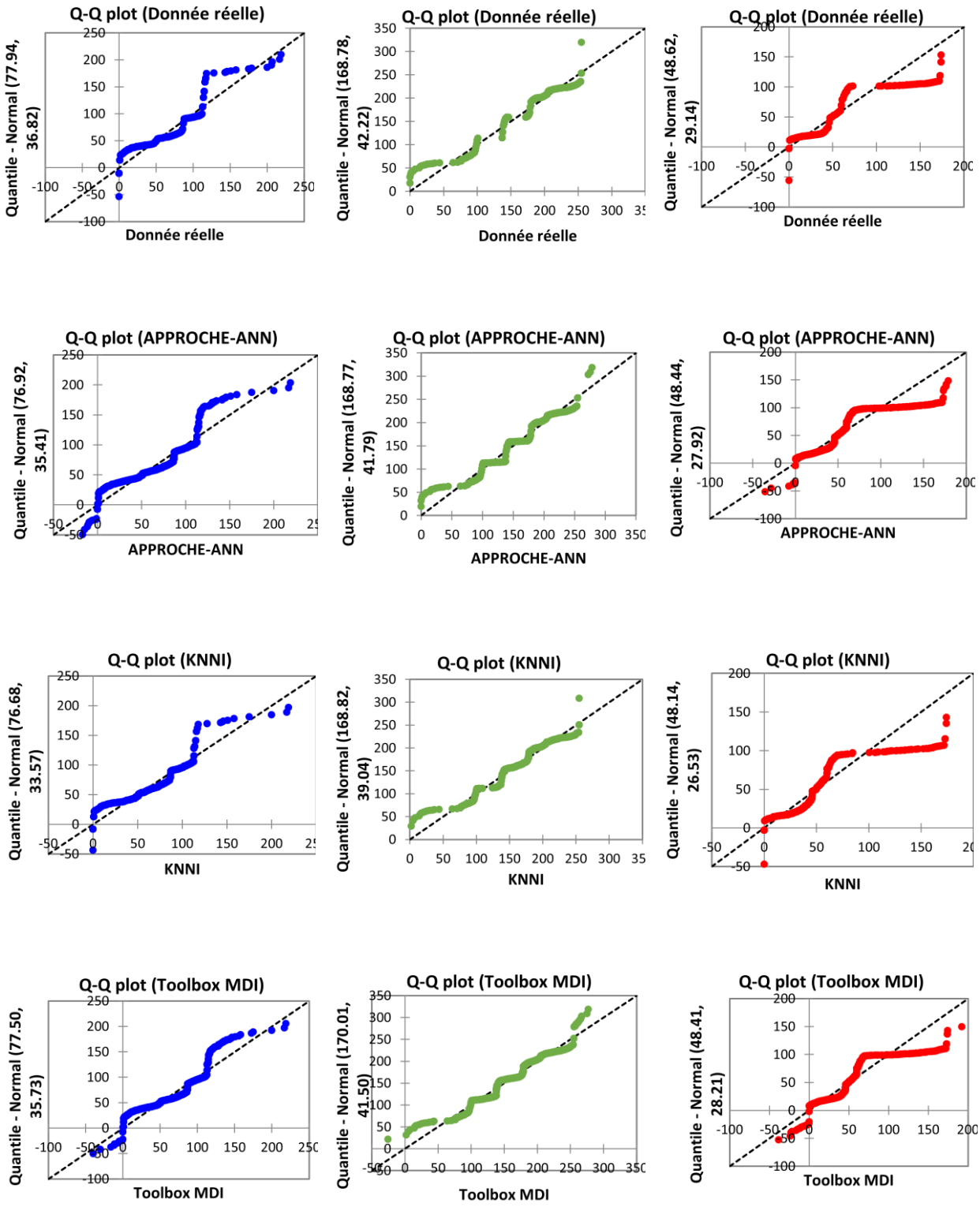


**Figure 4.2**: Graphe de QQ montre la distribution des quantiles expérimentaux des données estimé par le toolbox KNNI, MDI et l'approche ANN par rapport aux quantiles théorique de la loi normale. Cas : de 15% de donnée manquantes, mécanisme MCAR.

| MCAR | Blue | | | | Vert | | | | Rouge | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Variable | Real | Approche-ANN | Toolbox-KNNI | Toolbox-MDI | Real | Approche-ANN | Toolbox-KNNI | Toolbox-MDI | Real | Approche-ANN | Toolbox-KNNI | Toolbox-MDI |
| Observations | 604.00 | 604.00 | 604.00 | 604.00 | 604.00 | 604.00 | 604.00 | 604.00 | 604.00 | 604.00 | 604.00 | 604.00 |
| Missing data (%) | 0.00 | 15.00 | 15.00 | 15.00 | 15.00 | 15.00 | 15.00 | 15.00 | 15.00 | 15.00 | 15.00 | 15.00 |
| Std. deviation | 36.83 | 35.42 | 33.57 | 34.74 | 42.23 | 41.79 | 39.05 | 41.50 | 29.14 | 27.93 | 26.53 | 28.22 |
| Mean | 77.94 | 76.92 | 76.68 | 77.50 | 160.41 | 163.11 | 178.89 | 170.01 | 48.62 | 48.44 | 61.14 | 52.41 |
| Maximum | 219.00 | 219.00 | 221.00 | 221.00 | 255.00 | 269.81 | 227.15 | 277.06 | 174.00 | 178.99 | 185.00 | 180.61 |
| Minimum | 0.00 | 2.00 | 2.57 | 5.73 | 0.00 | 4.19 | 110.96 | 10.68 | 0.00 | 5.41 | 11.69 | 5.23 |
| R2 | 1.00 | 0.89 | 0.64 | 0.66 | 1.00 | 0.91 | 0.63 | 0.64 | 1.00 | 0.90 | 0.67 | 0.68 |
| R2Adj | 1.00 | 0.88 | 0.66 | 0.65 | 1.00 | 0.90 | 0.62 | 0.63 | 1.00 | 0.89 | 0.66 | 0.67 |
| RMSE | 0.00 | 41.50 | 75.02 | 62.77 | 0.00 | 18.75 | 66.32 | 57.12 | 0.00 | 21.66 | 44.17 | 42.01 |
| MAE | 0.00 | 20.36 | 36.80 | 30.80 | 0.00 | 9.20 | 32.54 | 28.02 | 0.00 | 10.63 | 21.67 | 20.61 |

**Tableau 4.2**: Statistique descriptives des résultats de comblement de 15% de données manquante, pour chaque matrice RGB par le toolbox MDI, KNNI et la méthode ANN dans le cas de mécanisme MCAR.

## Chapitre 4 : Validation et comparaison

Nous avons aussi appliqué les trois modèles sur 15% de données manquantes avec le type NMAR (**Figure 4.3**). Les graphes des quantiles QQ montrent que le modèle ANN et MDI sont plus satisfaisants que le modèle KNNI. L'approche ANN est la plus performante et elle plus remarquable graphiquement, où dans les trois cas, la méthode donne des valeurs décimales fortement similaires aux données réelles lorsqu'on prend comme critère statistique l'intervalle des quantiles. Les critères statistiques donnés par le **tableau 4.3** montrent le même degré de performance, où $R^2$ et $R^2_{Adj}$ montrent que dans les trois matrices, le modèle ANN détient les plus grandes valeurs, variant entre 0.93 et 0.97. Par contre, le modèle KNNI et MDI ont un intervalle variant respectivement comme suit : [0.61, 0.65] et [0.66, 0.71]. D'autre part, le RMSE et le MAE prouvent que le modèle ANN ne montre aucune tendance à l'erreur. Par contre, dans la matrice bleu et rouge, les trois modèles montrent une bonne performance de traitement. Le **tableau 4.2** favorise toujours l'approche ANN pour traiter les pixels manquants, distribués selon le mécanisme MCAR, où $R^2$ et $R^2_{Adj}$ donnés par les trois méthodes dans chaque matrice (bleue, verte et rouge) égalent respectivement 0.89, 0.64 et 0.66 **;** 0.91, 0.63 et 0.64 **;** 0.90, 0.66 et 0.67. Ce qui confirme que le modèle ANN est le plus performant. Dans le **tableau 4.3**, $R^2$ et $R^2_{Adj}$ montrent que le modèle ANN prouve le maximum de valeurs avec la matrice verte, il arrive respectivement jusqu'à 0.91 et 0.90. Les autres modèles y montrent des valeurs moindres. Par contre, le reste des matrices montrent l'inverse, sauf que dans tous les cas, $R^2$ et $R^2_{Adj}$ de l'approche ANN affiche les valeurs les plus élevées. RMSE et MAE ont également établi que les petits intervalles de valeurs sont observés avec le modèle ANN, et ils sont donné respectivement par [18.75, 41.50] et [9.20, 26.36].

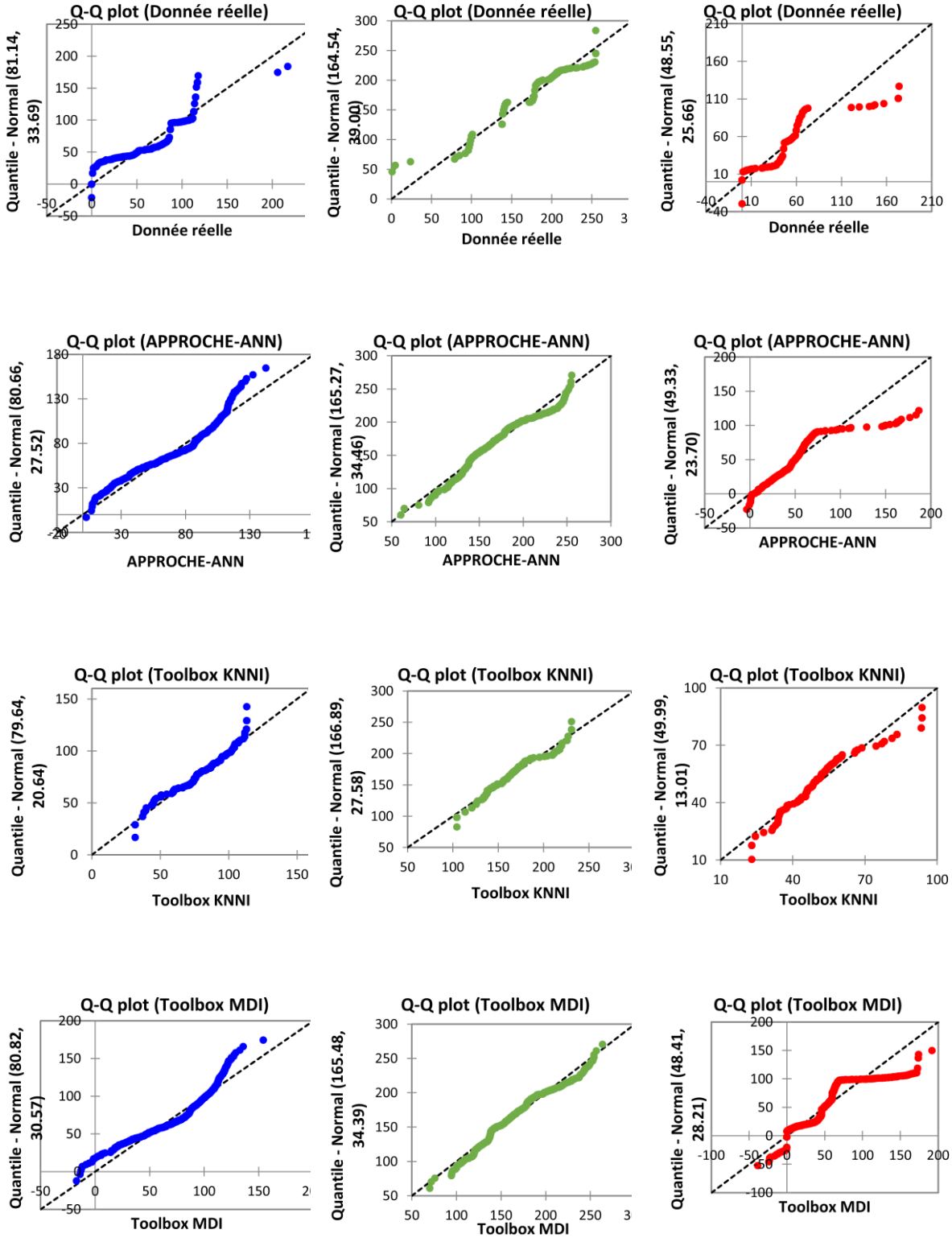


**Figure 4.3** : Graphe de QQ montre la distribution des quantiles expérimentaux des données estimé par le toolbox KNNI, MDI et l'approche ANN par rapport aux quantiles théorique de la loi normale. Cas : de 15% de donnée manquantes, mécanisme NMAR.

| NMAR | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Blue | | | | Vert | | | | Rouge | | | | | | | | |
| Variable | Real | Approche -ANN | Toolbox-KNNI | Toolbox-MDI | Real | Approche -ANN | Toolbox-KNNI | Toolbox-MDI | Real | Approche-ANN | Toolbox-KNNI | Toolbox-MDI | | | | | |
| Observations | 604.00 | 604.00 | 604.00 | 604.00 | 604.00 | 604.00 | 604.00 | 604.00 | 604.00 | 604.00 | 604.00 | 604.00 | | | | | |
| Missing data (%) | 0.00 | 15.00 | 15.00 | 15.00 | 15.00 | 15.00 | 15.00 | 15.00 | 15.00 | 15.00 | 15.00 | 15.00 | | | | | |
| Std. deviation | 33.72 | 31.66 | 41.59 | 27.54 | 39.03 | 37.36 | 34.42 | 27.60 | 25.68 | 27.66 | 35.01 | 21.02 | | | | | |
| Mean | 81.14 | 79.64 | 105.82 | 64.66 | 164.54 | 163.11 | 170.48 | 168.89 | 48.55 | 47.98 | 68.26 | 53.99 | | | | | |
| Maximum | 217.00 | 193.27 | 154.25 | 142.72 | 255.00 | 260.81 | 263.67 | 231.15 | 174.00 | 177.13 | 202.70 | 192.52 | | | | | |
| Minimum | 0.00 | 3.65 | 37.00 | 22.87 | 1.00 | 9.19 | 70.08 | 104.44 | 0.00 | 2.06 | 12.12 | 17.66 | | | | | |
| R2 | 1.00 | 0.95 | 0.64 | 0.70 | 1.00 | 0.94 | 0.63 | 0.68 | 1.00 | 0.97 | 0.65 | 0.71 | | | | | |
| R2Adj | 1.00 | 0.94 | 0.63 | 0.69 | 1.00 | 0.93 | 0.61 | 0.66 | 1.00 | 0.94 | 0.63 | 0.69 | | | | | |
| RMSE | 0.00 | 25.74 | 62.70 | 45.58 | 0.00 | 12.37 | 64.33 | 51.02 | 0.00 | 15.20 | 73.55 | 32.17 | | | | | |
| MAE | 0.00 | 10.02 | 24.41 | 20.13 | 0.00 | 4.82 | 25.04 | 19.86 | 0.00 | 5.92 | 28.63 | 12.52 | | | | | |

**Tableau 4.3** : Statistique descriptives des résultats de comblement de 15% de données manquante, pour chaque matrice RGB par le toolbox MDI, KNNI et la méthode ANN dans le cas de mécanisme MAR.

## 4.4. Analyse de tendance

En matière analytique, nous avions appliqué l'analyse régressive, relayée par l'analyse des résidus afin de pouvoir étudier les corrélations dans toute la série des données estimées et réelles de chaque vecteur de données ($X^{R}_{R}$, $X^{R}_{G}$, $X^{R}_{B}$), ($X^{E}_{R}$, $X^{E}_{G}$, $X^{E}_{B}$).

Nous avions choisi aléatoirement 30% de données manquantes dans une image TRMM prise sur le nord du territoire algérien. Les résultats de test sont montrés par la **figure 4.4**. Les graphes montrent que notre approche (ANN) est toujours la meilleure et la plus performante par rapport aux toolboxes MDI et KNNI.

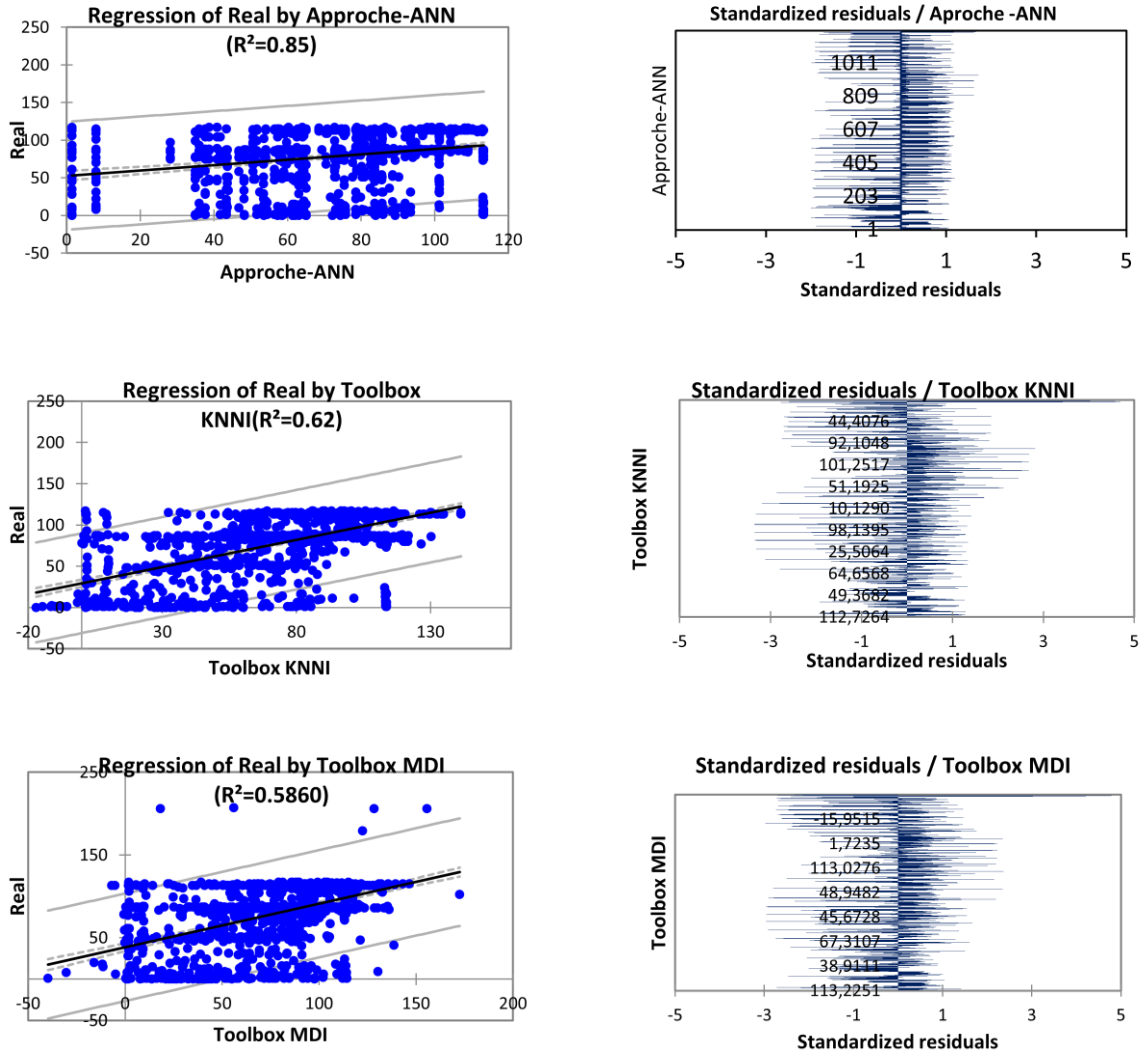


**Figure 4.4 :** Graphe de régression multiple des vecteurs de données estimés et réelles, suivre par l'analyse de résidus de chaque méthodes (KNNI, MDI et approche ANN), appliqué sur 30% de données manquantes choisis aléatoirement.

Le graphe de la régression montre que tous les pixels estimés rentrent dans l'intervalle de confiance du modèle de régression, ou $R^2$ égale 0.85. D'autre part, l'analyse des résidus standards, obtenus par la division des valeurs estimées sur les écarts obtenus entre les valeurs réelles et estimées pour chaque pixel, montre par le graphe une distribution homogène rentrant dans l'intervalle [-2, 2]. Par contre, le modèle KNNI et MDI montrent des tendances observées graphiquement par des points qui se trouvent en dehors de la courbe de tendance du graphe de régression. Au point comparatif, on peut déduire que la toolbox KNNI satisfait mieux que la toolbox MDI, sauf que l'analyse des résidus montre que tous les deux ont des tendances variant aléatoirement dans l'intervalle [-3, 3].

## 4.5. Conclusion

Les tests de faisabilité et de fiabilité sur la méthode que nous préconisons pour rétablir les données manquantes ou perdues sur les images satellitaires TRMM, ne nous a finalement pas pris en défaut. Au bout des explications relatant cette expérimentation, les résultats se sont avérés concluants par leurs performances selon différents mécanismes (MAR, MCAR et NMAR) et que le pourcentage de données est à 15% et 30%. L'avantage majeur qu'elle présente est l'exactitude maximale des données qu'elle parvient à repêcher de leur perte. Sa pertinence tient également dans ses possibilités d'application en diverses images TRMM. Faillibilité et partialité des méthodes existantes ont été à l'origine de notre finalité à élaborer notre modèle. Sa mise sur pied est prompte, son contrôle n'est pas trop exigeant en moyens matériels pas plus qu'en main-d'œuvre spécialisée. Enfin, de la façon dont elle s'est déroulée et jusqu'aux résultats qu'elle a livrés, la méthode ANN ne semble nullement souffrir de contradictions ou de risques. Ceci dit, il n'en demeure pas moins que sa mise en chantier sur le terrain d'applicabilité.

# Conclusion générale

# Conclusion générale

## Conclusion générale

L'objectif de notre étude dans ce mémoire, c'est le traitement de la perte de donnée dans les cartes TRMM avec données manquantes. Le problème de perte de données est récurrent. Comment et pourquoi se produisent ces pertes? La question conduirait dans un premier temps à élaborer une solution qui surmonterait le problème. En un deuxième temps, elle tâcherait de le porter sur le versant des avantages après l'avoir tourné à profit. Il s'agira de toutes les façons de réussir à percer le secret des mécanismes de ces pertes. Les mécanismes compris, il faudra ensuite se focaliser sur les pourcentages de DM en lui-même. Enfin, noter la façon précise par laquelle s'est signalée chaque absence de donnée.

Notre travail a porté sur une approche nouvelle pour traiter du problème de perte de données. L'approche n'arrive pas sur un terrain vierge, mais devra composer, cohabiter, voire surpasser les méthodes existantes: KNNI et MDI. Nous nous sommes assurés d'abord quant à la réalisabilité de notre approche (ANN), reprenant les deux méthodes, KNNI et la régression, appliquées au traitement des séries chronologiques. Ainsi assurés, nous nous sommes ensuite livrés à toute une batterie de tests statistiques descriptifs quantitatifs et qualitatifs (MAE, RMSE, $R^2$ etc.) pour attester de sa fiabilité. Appliquée sur divers pourcentages de manques de données (15% et 30%), induits par différents mécanismes de perte (MAR, MCAR et NMAR), ses preuves d'efficacité nous ont à juste raison paru concluantes. Nous ne prétendrons aucunement qu'elle aurait atteint le degré de satisfaction maximal, mais la possibilité est démontrée qu'elle peut bien l'atteindre dans de futures recherches et applications pratiques.

# Bibliographie

## Bibliographie

[1] Tankam, N. T., A. Dipanda, et al. (2010). TRAITEMENT NUMÉRIQUE D'IMAGES ET APPLICATIONS: Méthodes statistiques optimisées pour le traitement numérique des images de grandes tailles.

[2] Hong, Y., G. Tang, et al. (2019). "Remote Sensing Precipitation: Sensors, Retrievals, Validations, and Applications." Observation and Measurement; Li, X., Vereecken, H., Eds: 1-23.

[3] Center, E. O. (2001). "TRMM data users handbook." National Space Development Agency of Japan.

[4] Michot, V., D. Vila, et al. (2017). Performance of TRMM TMPA 3b42 v7 RT in replicating daily rainfall and rainfall regime in the Amazon Basin (2000-2013). XVIII SBSR.

[5] Kummerow, C., W. Barnes, et al. (1998). "The tropical rainfall measuring mission (TRMM) sensor package." Journal of atmospheric and oceanic technology **15**(3): 809-817.

[6] Simpson, J., R. F. Adler, et al. (1988). "A proposed tropical rainfall measuring mission (TRMM) satellite." Bulletin of the American meteorological Society **69**(3): 278-295.

[7] Fan, W., W. Ju, et al. (2013). "Method for reconstructing the pixel missing region on remote sensing images." Journal of Applied Remote Sensing **7**(1): 073536.

[8] Gao, G. and Y. Gu (2017). "Multitemporal Landsat missing data recovery based on tempo-spectral angle model." IEEE Transactions on Geoscience and Remote Sensing **55**(7): 3656-3668.

[9] Di Piazza, A. (2011). The problem of missing data in hydroclimatic time series. Applicationof spatial interpolation techniques to construct a comprehensive of hydroclimatic data in Sicily, Italy, Thèse de doctorat, IRIS, Université de Palerme.

[10] Norazian, M. N., Y. A. Shukri, et al. (2008). "Estimation of missing values in air pollution data using single imputation techniques."

[11] Choudhary, M. P. and V. Garg (2013). Causes, consequences and control of air pollution. All India seminar on methodologies for air pollution control, Jaipur, Rajasthan.

# Bibliographie

[12] Noor, N. M., A. S. Yahaya, et al. (2006). "The replacement of missing values of continuous air pollution monitoring data using mean top bottom imputation technique." Journal of Engineering Research & Education **3**: 96-105.

[13] La, H. P. and M. Q. Nguyen (2017). Reconstruction of Missing Imagery Data Caused by Cloudcover Based on Beyesian Neural Network and Multitemporal Images. International Conference on Geo-Spatial Technologies and Earth Resources, Springer.

[14] Aalstad, K., S. Westermann, et al. (2020). "Evaluating satellite retrieved fractional snow-covered area at a high-Arctic site using terrestrial photography." Remote Sensing of Environment **239**: 111618.

[15] Little R.J. and Rubin D.B. Statistical analysis with missing data. Hoboken, NJ : Wiley, 1987.

[16] Hamzah, F. B., F. MohdHamzah, et al. (2020). "Imputation methods for recovering streamflow observation: A methodological review." Cogent Environmental Science 6(1): 1745133.

[17] Schafer P., Joseph L., and J.W. Graham. Missing data : Our view of the state of theart. Psychological Methods, 7(2) :147, 2002.

[18] Bennane, A. (2010). Traitement des valeurs manquantes pour l'application de l'analyse logique des données à la maintenance conditionnelle, École Polytechnique de Montréal.

[19] Little R.J.A. and Rubin D.B. Statistical analysis with missing data. wiley. New York, 2002.

[20] Gleason, T. C. and R. Staelin (1975). "A proposal for handling missing data." Psychometrika **40**(2): 229-252.

[21] Young, R. and R. M. Rowell (1986). Cellulose: structure, modification and hydrolysis.

[22] Raaijmakers, Q. A. (1999). "Effectiveness of different missing data treatments in surveys with Likert-type data: Introducing the relative mean substitution approach." Educational and Psychological Measurement **59**(5): 725-748.

[23] Cramer, G., R. Ford, et al. (1976). "Estimation of toxic hazard—a decision tree approach." Food and cosmetics toxicology **16**(3): 255-276.

# Bibliographie

[24] Frane, J. W. (1976). "Some simple procedures for handling missing data in multivariate analysis." Psychometrika **41**(3): 409-415.

[25] Acurna, E. and C. Rodriguez (2004). The treatment of missing values and its effect in the classifier accuracy, classification, clustering, and data mining applications. Proceedings of the Meeting of the International Federation of Classification Societies (IFCS).

[26] Sinharay, S., H. S. Stern, et al. (2001). "The use of multiple imputation for the analysis of missing data." Psychological methods **6**(4): 317.

[27] Panduro, A., Y. C. Lin-Lee, et al. (1990). "Transcriptional and posttranscriptional regulation of apolipoprotein E, AI, and A-II gene expression in normal rat liver and during several pathophysiologic states." Biochemistry **29**(36): 8430-8435.

[28] Ruud, P. A. (1991). "Extensions of estimation methods using the EM algorithm." Journal of Econometrics **49**(3): 305-341.

[29] Johnson-Laird, P. N. (1988). The computer and the mind: An introduction to cognitive science, Harvard University Press.

[30] Kalkan, Ö. K., Y. Kara, et al. (2018). "Evaluating Performance of Missing Data Imputation Methods in IRT Analyses." International Journal of Assessment Tools in Education 5(3): 403-416.

[31] Horton, N. J. and S. R. Lipsitz (2001). "Multiple imputation in practice: comparison of software packages for regression models with missing variables." The American Statistician 55(3): 244-254.

[32] Presti, R. L., E. Barca, et al. (2010). "A methodology for treating missing data applied to daily rainfall data in the Candelaro River Basin (Italy)." Environmental monitoring and assessment **160**(1-4): 1.

[33] Honaker, J., G. King, et al. (2011). "Amelia II: A program for missing data." Journal of statistical software 45(7): 1-47.

[34] Takahashi, M. (2017). "Multiple ratio imputation by the EMB algorithm: Theory and simulation." Journal of Modern Applied Statistical Methods **16**(1): 34.

# Bibliographie

[35] Zhang, Q., Q. Yuan, et al. (2018). "Missing data reconstruction in remote sensing image with a unified spatial–temporal–spectral deep convolutional neural network." IEEE Transactions on Geoscience and Remote Sensing 56(8): 4274-4288.

[36] arbulescu, A., A. Bautu, et al. (2020). "Optimizing Inverse Distance Weighting with Particle Swarm Optimization." Applied Sciences 10(6): 2054.

[37] Wu, C.-Y., J. Mossa, et al. (2019). "Comparison of different spatial interpolation methods for historical hydrographic data of the lowermost Mississippi River." Annals of GIS 25(2): 133-151.

[38] Vybornova, Y. (2018). Application of spatial interpolation methods for restoration of partially defined images. CEUR Workshop Proceedings.

[39] Keskin, M., A. O. Dogru, et al. (2015). Comparing spatial interpolation methods for mapping meteorological data in Turkey. Energy systems and management, Springer: 33-42.

[40] Aieb, A., K. Madani, et al. (2019). "A new approach for processing climate missing databases applied to daily rainfall data in Soummam watershed, Algeria." Heliyon 5(2): e01247.

[41] Melgani, F., G. Mercier, et al. (2016). Recent methods for reconstructing missing data in multispectral satellite imagery. Applications+ Practical Conceptualization+ Mathematics= Fruitful Innovation, Springer: 221-234.

[42] Jerez, J. M., I. Molina, et al. (2010). "Missing data imputation using statistical and machine learning methods in a real breast cancer problem." Artificial intelligence in medicine 50(2): 105-115.

[43] Folch-Fortuny, A., F. Arteaga, et al. (2016). "Missing data imputation toolbox for MATLAB." Chemometrics and Intelligent Laboratory Systems **154**: 93-100.

[44] : Arteaga, F. and A. Ferrer (2002). "Dealing with missing data in MSPC: several methods, different interpretations, some examples." Journal of Chemometrics: A Journal of the Chemometrics Society **16**(8-10): 408-418.

[45] : Folch-Fortuny, A., F. Arteaga, et al. (2015). "PCA model building with missing data: New proposals and a comparative study." Chemometrics and Intelligent Laboratory Systems **146**: 77-88.

# Bibliographie

[46] Saccenti, E. and J. Camacho (2015). "On the use of the observation-wise k-fold operation in PCA cross-validation." Journal of Chemometrics **29**(8): 467-478.

[47] Camacho, J. and A. Ferrer (2014). "Cross-validation in PCA models with the element-wise k-fold (ekf) algorithm: practical aspects." Chemometrics and Intelligent Laboratory Systems **131**: 37-50.

[48] Touzet, C. (1992). les réseaux de neurones artificiels, introduction au connexionnisme.

[49] Mohamed, Z. E. (2019). "Using the artificial neural networks for prediction and validating solar radiation." Journal of the Egyptian Mathematical Society **27**(1): 47.

[50] Anderson, D. and G. McNeill (1992). "Artificial neural networks technology." Kaman Sciences Corporation **258**(6): 1-83.

[51] : Blayo, F. and M. Verleysen (1996). Les réseaux de neurones artificiels.

[52] Figueiredo Filho, D. B., J. A. S. Júnior, et al. (2011). "What is R2 all about?" Leviathan (São Paulo)(3): 60-68.

[53] : Bozdogan, H. (1987). ICOMP: A new model-selection criterion. 1. Conference of the international federation of classification societies.

[54] Hartini, E. (2017). "Implementation of missing values handling method for evaluating the system/component maintenance historical data." Jurnal Teknologi Reaktor Nuklir Tri Dasa Mega **19**(1): 11-18.

[55] Gimpy, M. (2014). "Missing value imputation in multi attribute data set." International Journal of Computer Science and Information Technologies **5**(4): 1-7.

[56] Hartini, E. (2018). "Classification of missing values handling method during data mining." SIGMA EPSILON-Buletin Ilmiah Teknologi Keselamatan Reaktor Nuklir **21**(2).

[57] Morisot, A. (2015). Méthodes d'analyse de survie, valeurs manquantes et fractions attribuables temps dépendantes: application aux décès par cancer de la prostate.

[58] Pan, R., T. Yang, et al. (2015). "Missing data imputation by K nearest neighbours based on grey relational structure and mutual information." Applied Intelligence **43**(3): 614-632.

# Bibliographie

[59 ] De Silva, H. and A. S. Perera (2016). Missing data imputation using Evolutionary k-Nearest neighbor algorithm for gene expression data. 2016 Sixteenth International Conference on Advances in ICT for Emerging Regions (ICTer), IEEE.

[61] : Vastrad, C. (2013). "Performance analysis of neural network models for oxazolines and oxazoles derivatives descriptor dataset." arXiv preprint arXiv:1312.2853.

[62] : Wang, W. and Y. Lu (2018). Analysis of the mean absolute error (MAE) and the root mean square error (RMSE) in assessing rounding model. IOP Conference Series: Materials Science and Engineering.

[60] : Karch, J. and D. van Ravenzwaaij (2020). "Improving on Adjusted R-squared." Collabra: Psychology 6(1).

[63] : Ghasemi, A. and S. Zahediasl (2012). "Normality tests for statistical analysis: a guide for non-statisticians." International journal of endocrinology and metabolism 10(2): 486.

[W1] https://www.rncan.gc.ca/cartes-outils-publications/imagerie-satellitaire-photos-aer/tutoriels-sur-la-teledetection/analyse-interpretation-dimages/traitement-numerique-des-images/9280#answer

[W2] http://www.youtube.com/user/cagira73#p/u/128/FUlyHCKfJIc.

[W3] https://blogrecherche.wp.imt.fr/2016/11/23/teledetection-ocean-donnees-manquantes/

**Résumé**

Ce travail a été réalisé en vue de concrétiser une nouvelle approche estimative des lacunes de pixels sur des images obtenues par télédétection. La cause des pertes étant les aléas climatiques ou la difficulté qu'éprouve le satellite à couvrir toute la surface du Globe, particulièrement les grandes étendues océaniques. Nous avions choisi les cartes pluviométriques mensuelles TRMM, obtenues pour le nord algérien. Nous nous étions basés, dans ce traitement, principalement sur l'aspect mathématique des réseaux de neurones (ANN). Une nouvelle architecture ANN (N-2-N) a pu être proposée sur l'exemple de la méthode KNNI et celle de régression, considérées comme fonction de transfert. Ces méthodes ont largement fait leurs preuves, scientifiquement, dans le traitement des données manquantes dans les sérés de données chronologiques à haute fréquence. Un ensemble de tests statistiques ont été appliqués sur les données estimées selon différents mécanismes de manques de données (MAR, MCAR et NMAR), choisis avec 5%, 15% et 30% de manque de données. Les tests que nous avons choisis sont $R^2$, $R^2_{Adj}$, RMSE, MAE, analyse de résidus, courbes de régression, nuage de point des quantiles (QQ plot) et le box plots. Les tests ont comparé les données estimées par rapport aux données réelles. Sur un autre volet, les résultats ont été confrontés avec ceux obtenus par les toolboxes KNNI et MDI. Les valeurs ont montré une très bonne performance de l'approche nouvelle. La tendance à l'erreur ne s'est manifestée nulle part durant la comparaison de tous les cas.

**Abstract**

This work is carried out as a purpose of a new approach in order to estimate missing pixels gaps, observed in the remote sensing image. The reasons of such loss are climatic problems or inability of the satellite to cover the whole earth land, especially oceans surfaces. We have chosen TRMM monthly rainfall maps of Algerian northern area. We are mainly based in this treatment on mathematical aspect of neural networks (ANN). Our new proposal (ANN) architecture of (N-2-N) has used KNNI and regression methods, considered as a transfer function. Scientifically, these methods have been already improved successfully while it is for handling missing data in chronological data sets of high frequencies. A set of statistical tests were applied on estimated data according to different cases of mechanisms leading on the misses (MAR, MCAR and NMAR), chosen by 5%, 15% and 30% of missing data. The tests that we have chosen are: $R^2$, $R^2_{Adj}$, RMSE, MAE, residuals analysis, regression curves, quantile point cloud (QQ plot) and box plots. The tests compared the estimated data with the real values. In a second step, they have also been confronted with the results obtained by the KNNI and MDI toolboxes. Accordingly to all these cases, the values have shown a very good performance of our approach, especially what was expected as the main point: total absence of error trends.