République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Farhat Abbas - Sétif 1 -

THESE

Présentée à la Faculté des Sciences

Département d'Informatique
Pour l'Obtention du Diplôme de

DOCTORAT EN SCIENCES

Option : Informatique

Thème

Extraction de connaissances à partir de données multi-spectrales : cas des images MSG

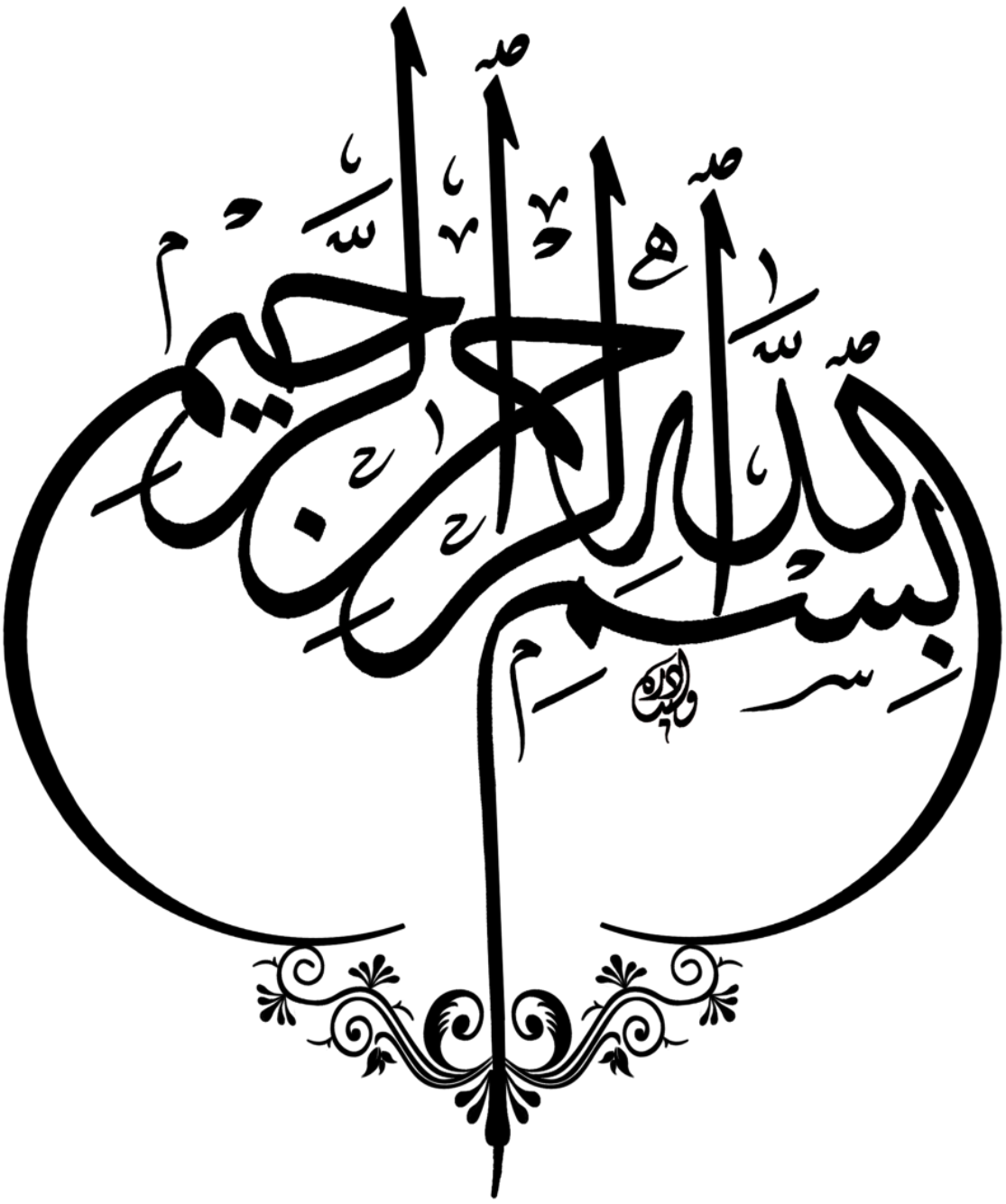
Présentée par

Bilal BOUAITA

Soutenu le : **24 / 09 / 2020**

Devant le jury composé de :

TOUAHRIA Mohamed	**Professeur à l'UFA-Setif-1**	**Président**
MOUSSAOUI Abdelouahab	**Professeur à l'UFA-Setif-1**	**Directeur**
BACHARI Nour El Islam	**Professeur à U.S.T.H.B, Alger**	**Co-directeur**
CHIKHI Salim	**Professeur à Constantine 2**	**Examinateur**
BOUZENADA Mourad	**M.C.A. à Constantine 2**	**Examinateur**


بسم الله الرحمن الرحيم

Dédicaces

Je dédie ce modeste travail à :

À la personne la plus chère à mes yeux, à ma chère mère qui a tout sacrifié pour ses enfants, qui a veillé à notre éducation, qui sans elle je ne serais pas ce que je suis,

À mon père, la personne que j'admire le plus au monde, pour tout ce qu'il m'a inculqué comme valeurs et principes,

À mon grand-père que Dieu me le garde,

À ma femme et mes enfants,

A toute ma famille,

A tous mes amis et mes collègues.

REMERCIEMENTS

Je remercie avant tout DIEU le tout puissant qui m'a donné le courage, la volonté et la patience pour réaliser ce modeste travail.

J'adresse mes vifs remerciements à mon directeur de thèse le Pr. Abdelouahab MOUSSAOUI de l'attention et du soutien qu'il a porté à mon travail, aussi pour toutes ses orientations et ses multiples conseils précieux pour mener à bien cette thèse.

Un grand merci à mon autre co-directeur de thèse le Pr. Nour El Islam BACHARI, je lui exprime ma profonde gratitude pour m'avoir fait profiter de ses connaissances.

J'exprime mes sincères remerciements à monsieur le Pr. Mohamed TOUAHRIA, qui a accepté de présider mon jury de soutenance.

J'exprime ma profonde gratitude à Pr. Salim CHIKHI et Dr. Mourad BOUZENADA de m'avoir fait l'honneur d'accepter de participer au jury et pour le temps consacré à la lecture de ce travail.

Je tiens à remercier chaleureusement tous les membres de l'équipe de la section de télédétection spatiale de l'institut royal météorologique de Belgique (IRM), et plus particulièrement le Dr. Nicolas CLERBAUX pour son accueil, sa disponibilité, son écoute et ses conseils qui m'ont permis d'avancer dans ma thèse.

Abstract

The Meteosat Second Generation (MSG) satellite transmits images of the observation of the earth every fifteen minutes in twelve different spectral bands. It allows us to follow phenomena that take place on the surface of the planet. Our work is about an important weather phenomenon is precipitation, to estimate precipitation from MSG images, most studies have not maximized the terabytes of data provided by the channels in this satellite, which are potentially rich in new resources that need to be exploited. Moreover, these studies classify pixels conventionally, where a pixel is considered either 100 % precipitant or 0 % (no-precipitant), whereas actually it cannot be classified in a clear and unambiguous way. To address this problem, we propose a method that exploits the images of the channels and constructs an estimation model in the form of fuzzy association rules to estimate the rainfall in Northeastern Algeria. Each rule is in if (condition)-then (conclusion) form, where the condition is a combination of the various fuzzy classes of MSG images, and the conclusion contains a single fuzzy class that represents the intensities of rain : no-rain, low, moderate, and high. The obtained results are compared with the data obtained by the Multi-Sensor Precipitation Estimate (MPE) product of the European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT).

Keywords : Data mining ; MSG images ; Apriori algorithm ; fuzzy association rules ; fuzzy c-means algorithm (FCM).

Résumé

Le satellite Météosat Seconde Génération (MSG) transmet des images de l'observation de la terre toutes les quinze minutes dans douze bandes spectrales différentes. Il nous permet de suivre des phénomènes qui se déroulent à la surface de la planète. Notre travail porte sur un phénomène météorologique important, les précipitations. Pour estimer les précipitations à partir des images MSG, la majorité des études effectuées ne profitent que des données de quelques canaux, ils n'exploitent pas suffisamment toutes les données fournies par ce satellite alors que ces téraoctets de données sont potentiellement riches en ressources inouïes qui demandent à être exploitées. De plus, ces études classifient les pixels à une classe d'une manière classique, par exemple un pixel est considéré 100 % précipitant où bien à 0 % non précipitant alors qu'on ne peut réellement le classifier d'une manière nette et précise. Pour cela, nous avons proposé une méthode qui exploite les images des canaux et construit un modèle sous la forme de règles d'association floues pour estimer les précipitations dans le nord-est de l'Algérie. Chaque règle est sous la forme de : si (condition) alors (conclusion), où la condition est une combinaison des différentes classes floues des images MSG, et la conclusion contient une seule classe floue qui représente l'intensité de précipitations : pas de précipitations, faible, modérée et forte. Les résultats obtenus sont comparés aux données obtenues par le produit MPE d'estimation des précipitations de l'organisation européenne pour l'exploitation de satellites météorologiques (EUMETSAT).

Mots clés : Data mining; les images MSG; algorithme Apriori; les règles d'association floues; algorithme c-moyennes floues (FCM).

ملخص

يرسل القمر الصناعي متيوسات الجيل الثاني (MSG) صورًا لمراقبة الأرض كل خمسة عشر دقيقة في اثني عشر نطاقًا طيفيًا مختلفًا. يسمح لنا بمتابعة الظواهر التي تحدث على سطح الكوكب. يركز عملنا على ظاهرة جوية مهمة وهي هطول الأمطار، معظم الدراسات المقدمة لتقدير هطول الأمطار من صورMSG لم تستغل إلا بعض البيانات لبعض قنوات هذا القمر الصناعي، لذلك فإن البيانات التي لم تستغل من المحتمل أن تكون غنية بالكثير من المعلومات المخبأة التي تحتاج إلى استخراجها والاستفادة منها. علاوة على ذلك، تصنف هذه الدراسات البيكسلات بشكل تقليدي، حيث يعتبر البيكسل إما ماطرا 100٪ أو 0٪ (لا مطر)، في حين أنه في الواقع لا يمكن تصنيفه بطريقة واضحة لا لبس فيها. لمعالجة هذه المشكلة، نقترح طريقة تستغل صور القنوات بإنشاء نموذج على شكل قواعد إرتباط غامضة لتقدير هطول الأمطار في شمال شرق الجزائر. كل قاعدة على شكل إذا (الشرط) إذن (النتيجة)، حيث الشرط هو مزيج من مختلف الفئات الغامضة من صور MSG، والنتيجة تحتوي على فئة واحدة غامضة تمثل شدة المطر: لا مطر، منخفض، معتدل، مرتفع. تم مقارنة النتائج التي تم الحصول عليها مع بيانات برنامج المنظمة الأوروبية لاستغلال سواتل الأرصاد الجوية لتقدير الهطول المتعدد.

الكلمات المفتاحية: التنقيب في البيانات، صورMSG ، خوارزمية Apriori، قواعد الارتباط الغامض، خوارزمية c-غامض (FCM).

Table des matières

Table des figures **xii**

Liste des tableaux **xiv**

Introduction générale **1**
1 Contexte et problématique 1
2 Objectifs et contributions 2
3 Contenu et organisation 3
4 Liste des publications 4

1 Les méthodes d'estimation des précipitations à travers des images MSG **5**
1 Introduction 5
2 Les satellites METEOSAT 5
2.1 Historique 5
2.2 METEOSAT de première génération 6
2.3 METEOSAT Seconde Génération (MSG) 7
2.3.1 Les canaux visibles 9
2.3.2 Les canaux vapeur d'eau 10
2.3.3 Les canaux infrarouges 11
2.4 Caractéristiques des images MSG 15
2.4.1 La résolution spatiale 16
2.4.2 La résolution spectrale 16
2.4.3 La résolution radiométrique 17
2.4.4 La résolution temporelle 18
3 Méthodes d'estimation des précipitations à partir des images satellitaires . 19
3.1 Les méthodes basées sur l'Infrarouges / Visibles (IR / VIS) 19
3.1.1 Méthodes d'indices du nuage 20
3.1.2 Méthodes bi-spectrales 21

3.1.3 Méthodes multi-spectrales 21
3.1.4 Méthodes basées sur le cycle de vie d'un nuage 23
3.1.5 Méthodes basées sur la modélisation des processus physiques des nuages 24
3.2 Les méthodes micro-ondes 24
3.3 Les méthodes hybrides (combinées) 26
4 Conclusion 26

2 Extraction de connaissances à partir des images satellitaires 27
1 Introduction 27
2 Extraction de Connaissances à partir de Données 28
2.1 Définition 28
2.2 Notions de bases 28
3 Extraction de connaissances à partir des images satellitaires 29
3.1 Définition 29
3.2 Les étapes d'extraction de connaissances à partir des images satellitaires 29
3.2.1 Étape d'acquisition 30
3.2.2 Étape du prétraitement 30
3.2.3 Étape de transformation 31
3.2.4 Étape du Datamining 31
3.2.5 Étape d'évaluation et présentation 31
4 Les principales taches du Datamining 32
4.1 Les méthodes selon les objectifs 32
4.1.1 La classification 32
4.1.2 La segmentation (le clustering) 32
4.1.3 L'association 33
4.1.4 La prédiction 33
4.2 Les méthodes selon le type d'apprentissage 33
4.2.1 Apprentissage supervisé 33
4.2.2 Apprentissage non supervisé 33
4.3 Les méthodes selon les types de modèles obtenus 34
4.3.1 Les méthodes d'explication et de prédiction 34
4.3.2 Les méthodes de description et de visualisation 34
5 Les principales méthodes du Datamining 34
5.1 Méthode des C-moyennes (C-Means) 34
5.2 Méthode des C-moyennes floues (Fuzzy C-Means (FCM)) 36

5.3 Méthode des arbres de décision 37
5.4 Méthode k plus proche voisins 38
5.5 Méthode des réseaux de neurones 39
5.6 Méthode des règles d'association 41
6 Conclusion 41

3 Extraction de connaissances par les règles d'association classiques et floues 42
1 Les règles d'association classiques 42
1.1 Introduction 42
1.2 Exemple de paniers de marché 43
1.3 Applications des règles d'association 44
1.4 Les avantages et les inconvénients des règles d'association 46
1.4.1 Les avantages 46
1.4.2 Les inconvénients 46
1.5 Concepts et définitions sur les règles d'association 46
1.5.1 Item 46
1.5.2 Itemset 47
1.5.3 K-Itemset 47
1.5.4 Support d'un itemset 47
1.5.5 Règle d'association 47
1.5.6 Confiance d'une règle d'association 47
1.5.7 Itemset fréquent 47
1.6 Les étapes de l'extraction de règles d'association 48
1.7 Extraction de règles d'association 49
1.7.1 Recherche de l'ensemble des itemsets fréquents par l'algorithme Apriori 50
1.7.2 Génération des règles d'association 53
2 Les règles d'association floues 54
2.1 Introduction 54
2.2 Concepts et définitions sur les règles d'association floues 54
2.2.1 Item flou 54
2.2.2 Itemset flou 54
2.2.3 Le degré d'un itemset flou 55
2.2.4 Le support d'un itemset flou 55
2.2.5 Règle d'association floue 56
2.2.6 La confiance d'une règle d'association floue 57
2.3 Extraction de règles d'association floues 57

2.4 Exemple d'xtraction de règles d'association floues 58
3 Conclusion 61

4 Estimation des précipitations à partir des images MSG en utilisant les règles d'association floues 62
1 Introduction 62
2 Description de la méthode 63
2.1 Création d'une base de données transactionnelle 64
2.2 Création d'une base de données transactionnelle floue 66
2.2.1 Détermination du nombre des items flous 66
2.2.2 Détermination des degrés d'appartenance correspondant à chaque partition (classe) floue de chaque item 68
2.3 Extraction de règles d'association floues 69
2.3.1 Identification de la liste des itemsets flous fréquents . . . 69
2.3.2 Génération de règles d'association floues 72
3 Résultats expérimentaux 73
3.1 Données utilisées 73
3.2 Construction du modèle 74
3.3 Validation du modèle 78
4 Conclusion 83

Conclusion générale 84

Bibliographie 86

Annexe : généralités sur la logique floue 96
1 Pourquoi la logique floue 96
2 Les ensembles flous 96
3 Fonction d'appartenance 97
4 Caractéristiques d'un ensemble flou 98
5 Opérations de base sur les ensembles flous 99

Table des figures

1.1 Le satellite METEOSAT de première génération (MFG) 7
1.2 Satellite METEOSAT Seconde Génération(MSG) 8
1.3 Image de MSG du canal VIS 0.6 du 01/04/2008 à 12 :00 [1] 10
1.4 Image de MSG du canal VIS 0.8 du 01/04/2008 à 12 :00 [1] 10
1.5 Image de MSG du canal WV 6.2 du 01/04/2008 à 12 :00 [1] 11
1.6 Image de MSG du canal WV 7.3 du 01/04/2008 à 12 :00 [1] 11
1.7 Image de MSG du canal NIR 1.6 du 01/04/2008 à 12 :00 [1] 12
1.8 Image de MSG du canal IR 3.9 du 01/04/2008 à 12 :00 [1] 12
1.9 Image de MSG du canal IR 8.7 du 01/04/2008 à 12 :00 [1] 13
1.10 Image de MSG du canal IR 9.7 du 01/04/2008 à 12 :00 [1] 13
1.11 Image de MSG du canal IR 10.8 du 01/04/2008 à 12 :00 [1] 14
1.12 Image de MSG du canal IR 12.0 du 01/04/2008 à 12 :00 [1] 14
1.13 Image de MSG du canal IR 13.4 du 01/04/2008 à 12 :00 [1] 15
1.14 Caractéristiques des images MSG 15
1.15 La résolution spatiale 16
1.16 La résolution spectrale 17
1.17 La résolution radiométrique 17
1.18 Les mesures radiométriques en fonction de l'objet ciblé dans une image visible 18
1.19 Les mesures radiométriques en fonction de l'objet ciblé dans une image infrarouge 18
1.20 Volume d'absorption et de diffusion en fonction du taux de pluie, pour les gouttes d'eau et pour les cristaux de glace, aux 3 différentes fréquences principales [117] 25

2.1 Les étapes du processus d'extraction de connaissances à partir des images satellitaires [131] 29
2.2 Les étapes de l'algorithme C-Means 35

2.3 Exemple d'un arbre de décision 38
2.4 Exemple d'un k plus proche voisins 39
2.5 Architecture générale d'un réseau de neurones 40
2.6 Les deux types principaux d'un réseau de neurones 41

3.1 Les étapes d'extraction des règles d'association [8] 49
3.2 Exmeple de recherche des itemsets fréquents 52
3.3 Partitionnement flou des attributs quantitatifs 59

4.1 La procédure générale de notre méthode proposée 63
4.2 Partie de la base de données transactionnelle des images MSG 65
4.3 Partitionnement floue des attributs visibles 66
4.4 Partition floue de l'attribut précipitations 67
4.5 Exemple d'extraction des itemsets flous fréquents en utilisant l'algorithme Apriori flou 71
4.6 La zone d'étude 74
4.7 Nombre de pixels appartenant à chaque classe flou de l'attribut précipitations pour chaque α-coupe 80
4.8 La racine carrée de l'erreur quadratique moyenne pour chaque α-coupe . . 82
4.9 Le taux moyen de validation pour chaque α-coupe 82
10 Exemple de la différence entre la logique booléenne et la logique floue . . 97
11 Principaux caractéristiques d'un sous ensemble flou 99

Liste des tableaux

1.1 Date de lancée des satellites METEOSAT [1] 6
1.2 Caractéristiques des trois canaux du capteur MVIRI du MFG [35] 7
1.3 Caractéristiques des canaux MSG [112] 9

3.1 Exemple d'une base de transactions binaire 43
3.2 Tableau de cooccurrence des produits 44
3.3 La complexité des calculs pour les règles d'association 48
3.4 Les règles extraites par l'algorithme Apriori 53
3.5 Degré d'un itemset flou 55
3.6 Différent type de calcul de la cardinalité flou 56
3.7 Exemple d'une base de transactions quantitatives 58
3.8 Base des degrés d'appartenances 59
3.9 Support flou des items 60

4.1 Quantité de précipitations en Algérie 67
4.2 Support de chaque item flou de l'attribut précipitations 75
4.3 Liste des règles d'association floues pour l'item flou [Précipitations, pas de précipitations] avec $Mfs = 0.14$ et $Mfc = 0.74$. 76
4.4 Liste des règles d'association floues pour l'item flou [Précipitations, modérée] avec $Mfs = 0.12$ et $Mfc = 0.80$. 77
4.5 Liste des règles d'association floues pour l'item flou [Précipitations, forte] avec $Mfs = 0.19$ et $Mfc = 0.65$. 77
4.6 Liste des règles d'association floues pour l'item flou [Précipitations, faible] avec $Mfs = 0.12$ et $Mfc = 0.95$. 78
4.7 Nombre de pixels de chaque item flou de l'attribut précipitations pour chaque α-coupe 79
4.8 Le taux moyen de validation et la racine carrée de l'erreur quadratique moyenne pour chaque α-coupe 81

9 Définitions des t-normes et t-conormes les plus utilisées 99

Introduction générale

1 Contexte et problématique

La télédétection est définie comme un ensemble de techniques permettant d'acquérir différentes informations sur la surface de la Terre, d'étudier et d'identifier des phénomènes sans frottement, en étudiant et en analysant le rayonnement électromagnétique réfléchie ou émis par la surface terrestre, ensuite, ces informations sont utilisées à des fins météorologique, géographique, géologique, océanographique, agriculture et génie civil [27, 52]. Ce rayonnement électromagnétique est enregistré par des caméras ou capteurs fonctionnant généralement à partir des avions ou des satellites [68]. Si la source de rayonnement est une source artificielle générée par un capteur tel que le radar, le processus s'appelle une "télédétection actif", et si le capteur recueille des rayons naturels qui sont les rayons du soleil tel que le capteur SEVIRI du satellite METEOSAT Seconde Générations (MSG) dans ce cas, le processus de télédétection est appelé "télédétection passive". Ces capteurs mesurent et enregistrent le rayonnement reçu, puis le convertissent en images numériques où chaque pixel représente la quantité de rayonnement reçue par une cellule de cible au sol. Nous nous sommes intéressés aux images qui proviennent du satellite météorologique MSG qui nous fournit quotidiennement toutes les quinze minutes des images dans douze canaux spectraux différents (visible, proche infrarouge, infrarouge thermique) [112].

Grâce à leur fréquence, leur qualité et leur précision, les images multi-spectrales du satellite MSG sont particulièrement utiles pour la prévision immédiate et à court terme. Elles permettent aux météorologues de reconnaître et de suivre le développement détaillé des phénomènes météorologiques chaque 15 minutes dans 12 bandes spectrales. Parmi les phénomènes importants qui confrontent les météorologues notamment de l'Office National de la Météorologie (ONM) en Algérie sont les précipitations [56]. Une bonne connaissance de la quantité et de la distribution des précipitations dans le temps et dans l'espace, est indispensable pour les autorités de notre pays à la prise de décision pour le développement durable et pour la prévention des catastrophes naturelles qui menacent notre vie et ruinent nos économies locales. Par exemple, le choix d'un site de construction de barrage, anticiper

la sécheresse et s'adapter à ses conséquences en particulier sur l'agriculture, le volume de précipitations fortement élevés pouvant entraîner des inondations ou des glissements de terrain… etc.

L'estimation des précipitations en utilisant les images satellitaires notamment les images MSG représente toujours un défi. La majorité des méthodes développées dans ce contexte [28, 9, 2, 55, 80, 20, 132, 85, 70, 69, 124] ont donné des résultats satisfaisants, néanmoins, ces méthodes profitent que les données de quelques canaux, ils n'exploitent pas bien toutes les données fournies par le satellite MSG qui sont devenues de plus en plus grandes avec le temps, alors que cette énorme quantité de données spectrales est potentiellement riche en ressources inouïes qui demandent à être exploitées. En plus, la classification des pixels précipitants et non précipitants par ces méthodes se fait d'une manière classique, un pixel est considéré 100 % précipitant où bien à 0 % non précipitant, ce qui n'est pas toujours vrai dans la réalité.

2 Objectifs et contributions

L'ONM dispose d'un volume très important d'images MSG, accumulées aux files du temps, ces téraoctets de données spectrales stockées sont susceptibles d'être riches en ressources incroyables qui doivent être exploitées. Nous avons besoin donc de méthodes et d'outils capables de les exploiter au maximum, de les analyser, de les représenter, de les classifier, d'en extraire les connaissances pertinentes et enfin de visualiser les résultats de cette extraction, cet ensemble d'outils est appelé le Datamining [40]. L'utilisation des techniques du Datamining, telles que les règles d'association, peut nous aider à découvrir et à extraire des relations intéressantes entre les données des images MSG afin d'estimer les précipitations [4]. Nous nous intéressons aux règles d'association car ce type de connaissance est descriptif (faciles à interpréter et son niveau explicatif ou sémantique est très élevé), prédictif et même décisif. Pour la classification des pixels précipitants et non précipitants à partir des images MSG, il est d'autant plus difficile de dire si un pixel est précipitant ou non et c'est justement là où on a introduit la théorie des ensembles flous proposée par Lotfi Zadeh [133] qui nous offre un moyen de gérer ces mesures de données qui sont inexactes. L'estimation des précipitations n'est pas aussi précise, c'est pourquoi nous avons jugé utile d'utiliser les règles d'association floues basées sur la théorie des ensembles flous plutôt que les règles d'association classiques où nous avons tout d'abord définit pour chaque variable linguistique un intervalle flou comme précipitation faible, modérée et forte. A cet effet, nous avons développé dans ce travail une méthode [23] qui exploite les images multi-spectrales issues des onze canaux du satellite MSG (sauf le canal HRV qui a des dimensions différentes

par rapport aux autres) afin d'extraire des corrélations entre les images de ces canaux pour l'estimation des précipitations en Algérie sous forme de règles d'association floues.

3 Contenu et organisation

Afin de présenter le travail que nous avons assigné, nous avons opté pour une organisation en quatre chapitres principaux :

— **Premier chapitre** : nous présentons dans ce chapitre, une description générale sur le satellite MSG puisque les images utilisées dans le cadre de notre travail de recherche issues de ce satellite, ainsi que les différentes caractéristiques de ces images. Ensuite, nous présentons les méthodes principales d'estimation de précipitations utilisant des données qui proviennent de satellites météorologiques notamment le satellite MSG.

— **Deuxième chapitre** : ce chapitre comprend le concept d'extraction de connaissances à partir des bases de données notamment les données images satellitaires, ainsi l'ensemble de méthodes et de techniques d'analyse de données pour l'extraction des connaissances comme : C-moyennes, C-moyennes floues, les règles d'association, les arbres de décisions, les réseaux de neurones,... etc.

— **Troisième chapitre** : ce chapitre sera consacré à l'extraction de connaissances par la méthode des règles d'association, ensuite nous abordons une extension par des règles d'association floues, puisque, nous avons réalisé et développée notre méthode sur la base de cette méthode.

— **Quatrième chapitre** : nous présentons notre méthode développée basée sur les règles d'association floues afin d'estimer les précipitations à partir des images MSG, et nous faisons une comparaison des résultats en l'appliquant à la région du Nord-Est de l'Algérie avec les données du produit d'estimation des précipitations (MPE) de l'organisation européenne pour l'exploitation de la météorologie Satellites (EUMETSAT).

Nous achevons notre travail par une conclusion générale et les perspectives possibles de nos futures travaux.

4 Liste des publications

Les travaux réalisés dans cette thèse sont soutenus par les publications suivantes :

— Bouaita, B., Moussaoui, A., & Bachari, N. E. I. "*Rainfall estimation from MSG images using fuzzy association rules*", Journal of Intelligent & Fuzzy Systems, 2019, vol. 37, no 1, p. 1357-1369.

— Bilal Bouaita, Abdelouahab Moussaoui, "*Extraction de connaissances par la méthode des règles d'association avec une nouvelle mesure*", deuxième conférence nationale de l'informatique destinée aux étudiants de graduation et de post-graduation (JEESI'12), Alger, Avril 2012.

Chapitre 1

Les méthodes d'estimation des précipitations à travers des images MSG

1 Introduction

Le satellite MSG a la capacité d'observer la Terre dans 12 canaux spectraux et de fournir des images toutes les 15 minutes. Ces images multi-spectrales sont considérées comme une source riches d'informations et fiables sur l'occupation du sol, et aussi sur l'évolution détaillée des phénomènes qui se déroulent à la surface de la planète. Parmi ces différents phénomènes, nous nous intéressons aux précipitations et ses estimations. C'est l'un des principales objectifs de la surveillance par le satellite météorologique MSG à partir de ses images, il représente une aspiration ambitieuse aux chercheurs scientifiques et un besoin indispensable aux différents domaines tel que la météorologie, la géoscience, l'hydrologie, l'agriculture… etc. Dans la première partie de ce chapitre, nous présentons ce satellite météorologique ainsi que les caractéristiques des images MSG fournies par ce dernier. Ensuite, nous présentons dans la deuxième partie les différentes approches d'estimation de précipitations.

2 Les satellites METEOSAT

2.1 Historique

La série des satellites géostationnaires de METEOSAT de première et seconde génération est une famille de satellites météorologiques Européens, ils permettent d'observer et d'amasser continuellement de l'information d'une zone précise du globe, ils sont situés à 36000 km d'altitude au-dessus de l'Équateur, ce qui assure une couverture complète de l'Afrique,

l'Europe et les extrémités de l'Asie et de l'Amérique méridionale. Ils ont été mis en place par l'organisation européenne pour l'exploitation des satellites météorologiques (EUMETSAT), le premier satellite de la première génération a été lancé en 1977, après son succès, il est suivi par 6 autres satellites. Tous les satellites de la première génération sont maintenant hors d'usage. Météosat-8 (MSG-1) est le premier satellite de seconde génération, il a été mis sur orbite en août 2002, il a été suivi par trois autres satellites afin de poursuivre la continuité opérationnelle des données METEOSAT. Les 4 stellites sont actuellement opérationnels [1]. L'historique de lancement des différents satellites METEOSAT est illustré dans le tableau 1.1.

TABLE 1.1 Date de lancée des satellites METEOSAT [1]

Nom du satellite	Date de lancée
METEOSAT-1	23 Novembre 1977
METEOSAT-2	19 Juin 1981
METEOSAT-3	15 Juin 1988
METEOSAT-4	06 Mars 1989
METEOSAT-5	02 Mars 1991
METEOSAT-6	19 Novembre 1993
METEOSAT-7	02 Septembre 1997
MSG-1 (METEOSAT-8)	28 Août 2002
MSG-2 (METEOSAT-9)	21 Décembre 2005
MSG-3 (METEOSAT-10)	5 Juillet 2012
MSG-4 (METEOSAT-11)	15 Juillet 2015

2.2 METEOSAT de première génération

La première génération de satellites METEOSAT (METEOSAT First Generation (MFG)) du METEOSAT-1 à METEOSAT-7, a fourni des observations météorologiques continues et fiables depuis l'espace, tous les pays du monde sont capables de recevoir les données de ce satellite et joue par conséquent un véritable rôle international. METEOSAT-1 à 7 sont maintenant tous hors service (1977 - 2017). Lorsqu'il était opérationnel, MFG (Figure 1.1) fournissait des images toutes les demi-heures dans trois canaux spectraux (Visible (VIS), Infrarouge (IR) et Vapeur d'eau (WV)) grâce au capteur METEOSAT Visible and InfraRed Imager (MVIRI). Le MVIRI tourne avec une vitesse de rotation de 100 tours par minute autour de son axe principal. Pour chaque tour, il balaye une ligne de l'image terrestre, toutes les 25 minutes, une image complète de 2500 lignes est acquise sur les bandes spectrales IR (longueurs d'onde infrarouges thermiques comprises entre 10.5 et 12.5 μm) et WV

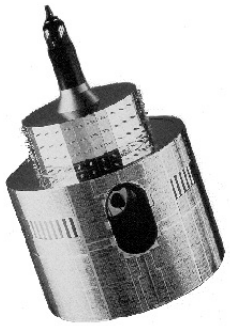


FIGURE 1.1 Le satellite METEOSAT de première génération (MFG)

(rayonnement de la bande d'absorption de la vapeur d'eau de 5.7 à 7.1 μm). En même temps, une troisième image VIS est acquise dans le domaine du visible au proche infrarouge (0.4 – 0.11 μm) par deux détecteurs avec une résolution deux fois plus fine que les détecteurs IR et WV. Une période de retrace et de veille de 5 min pour préparer le capteur à la prochaine acquisition des images, donc chaque acquisition complète le cycle de répétition dans 30 minutes, soit 48 acquisitions de données par jour. Les images complètes du disque complet sont 2500 × 2500 pixels pour les canaux IR et WV, résolution doublée (5000 lignes de 5000 pixels) pour le canal VIS, avec une résolution au sol au point sous-satellite de 5 km pour les deux images IR et WV, 2.5 km pour l'image VIS [35]. Les caractéristiques des trois canaux du capteur MVIRI sont illustrées dans le tableau 1.2.

TABLE 1.2 Caractéristiques des trois canaux du capteur MVIRI du MFG [35]

Nom du canal	Bande spectrale en micromètre (μm)	Résolution du pixel (Km)	Taille de l'image (pixels)
VIS	0.4 – 0.11	2.5 × 2.5	5000 × 5000
WV	5.7 – 7.1	5 × 5	2500 × 2500
IR	10.5 – 12.5	5 × 5	2500 × 2500

2.3 METEOSAT Seconde Génération (MSG)

Le satellite MSG [112] (figure 1.2) consiste en une série de quatre satellites météorologiques géostationnaires, cette famille de satellites dispose d'un capteur principal plus sophistiqué, c'est le Spinning Enhanced Visible and InfraRed Imager (SEVIRI), il fournit des images de la Terre toutes les 15 minutes dans 12 bandes spectrales différentes, contre 30 minutes et 3 bandes pour les METEOSAT de première génération. Ce rafraîchissement rapide des images permet aux prévisionnistes de suivre les phénomènes météorologiques dangereux à évolution rapide, telles que les orages et les tempêtes de neige. La résolution de

toutes les images fournies est de 3 km à l'exception des images du canal visible HRV qui a une résolution de 1 km. Le capteur SEVIRI tourne autour de son axe principal avec une vitesse de rotation 100 tours par minute comme le capteur MVIRI du MFG. Il permet un balayage complet de la Terre en environ 12,5 minutes, les 2 minutes 30 secondes suivantes sont consacrées au retour du miroir du SEVIRI à sa position initiale, donc, un cycle de répétition d'imagerie de 15 min, on obtient donc 96 acquisitions d'images par jour. Les images résultantes consistent en 3712 x 3712 pixels sauf pour l'image HRV (11136 x 5568 pixels). Un autre capteur embarqué sur les satellites MSG est le Geostationary Earth Radiation Budget experiment (GERB), conçus essentiellement pour effectuer des mesures précises du bilan radiatif de la Terre [60]. En plus de ces deux capteurs, le satellite MSG est doté d'équipements de télécommunication sophistiqués qui sont nécessaires pour l'exploitation et à la transmission des données images brutes à la station sol principale de Darmstadt en Allemagne, après un traitement, les images sont diffusées à une large communauté d'utilisateurs à travers le satellite MSG lui-même. Dans notre pays, elles sont captées au niveau de la station radiométrique de l'Office National de la Météorologie (ONM).

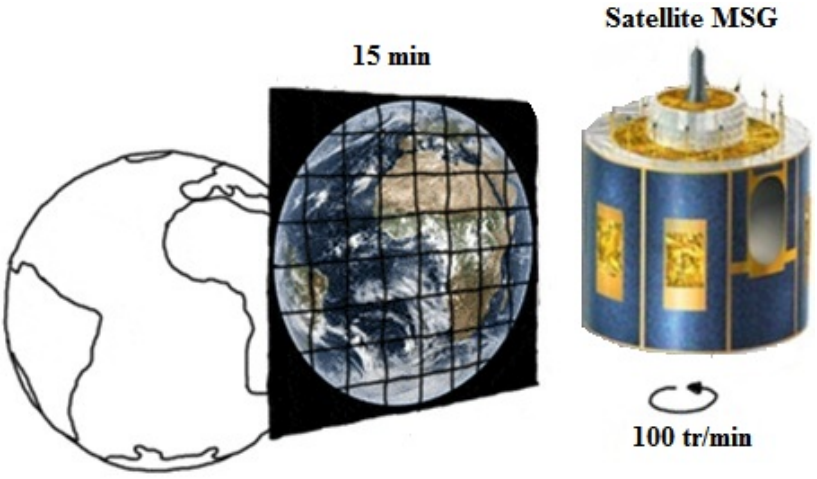


FIGURE 1.2 Satellite METEOSAT Seconde Génération(MSG)

Les douze canaux du capteur SEVIRI permettent de recueillir une grande diversité d'informations. Il y trois canaux qui fonctionnent sur le canal visible, deux sur les vapeurs d'eau et le reste sur les infrarouges (voir tableau 1.3). Les canaux fonctionnent jour et nuit, à l'exception des canaux du domaine visible qui ne sont bien évidemment exploitable que le jour. Nous présentons dans ce qui suit les fonctions des différents canaux du capteur SEVIRI du satellite MSG [112, 132, 6].

TABLE 1.3 Caractéristiques des canaux MSG [112]

Nom du canal	Bande spectrale en micromètre (μm)	Résolution spatiale (km)	Nombre de lignes par image	Nombre de pixels par ligne
HRV	0.5 - 0.9	1	11136	5568
VIS 0.6	0.56 - 0.71	3	3712	3712
VIS 0.8	0.74 - 0.88	3	3712	3712
NIR 1.6	1.50 - 1.78	3	3712	3712
IR 3.9	3.48 - 4.36	3	3712	3712
WV 6.2	5.35 - 7.15	3	3712	3712
WV 7.3	6.85 - 7.85	3	3712	3712
IR 8.7	8.30 - 9.10	3	3712	3712
IR 9.7	9.38 - 9.94	3	3712	3712
IR 10.8	9.80 - 11.80	3	3712	3712
IR 12.0	11.00 - 13.00	3	3712	3712
IR 13.4	12.40 - 14.40	3	3712	3712

2.3.1 Les canaux visibles

Le satellite MSG comporte 2 canaux visibles qui sont VIS 0.6 et VIS 0.8 (figure 1.3 et figure 1.4). Leur longueur d'onde se situe entre 0.56 - 0.71 μm et 0.74 - 0.88 μm respectivement. Ils englobent la totalité du domaine visible (i.e. le segment du spectre électromagnétique visible par l'œil humain). Ces images apparaissent totalement en noires pendant la nuit, et on ne peut pas les exploitées en raison de l'absence de rayonnements solaires. Chaque pixel de ces images acquises pendant la journée représente la quantité de lumière solaire réfléchie dans l'espace par les nuages ou la surface de la Terre. Dans ces images, la neige et les nuages semblent en blanc et les zones sans nuages en noir. Les nuages épais sont plus brillants que les nuages fins. Il est difficile de différencier les nuages de basse altitude et les nuages de haute altitude dans ces images visibles, pour les distinguer, il est nécessaire d'utiliser les images infrarouges. Ces deux canaux sont essentiels pour la détection des nuages, l'identification des scènes, la surveillance des surfaces terrestres et de la végétation. La résolution de ces images est de 3 km à la verticale du satellite. Le canal visible collecte 3712 lignes consistant chacune en 3712 pixels.

MSG contient aussi un canal visible à haute résolution (High-Resolution Visible (HRV)) qui est un canal visible à large bande spectrale (0.5 - 0.9 μm). Ce canal fournit des images de 11136 lignes de 5568 pixels, il regroupe les informations des deux autres canaux visibles pour donner plus de détails sur les surfaces terrestres et les petits nuages, sa résolution est de 1 km, il permet de mesurer le vent à l'altitude des nuages et de distinguer la texture des nuages.

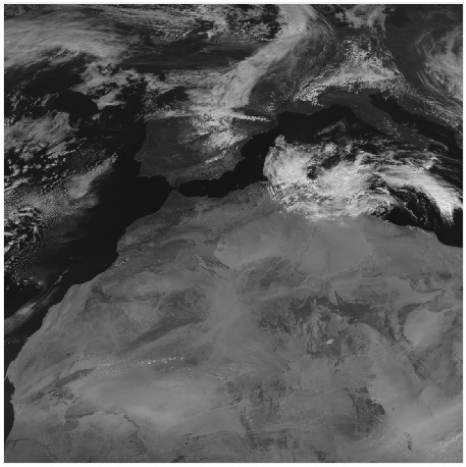


FIGURE 1.3 Image de MSG du canal VIS 0.6 du 01/04/2008 à 12 :00 [1]

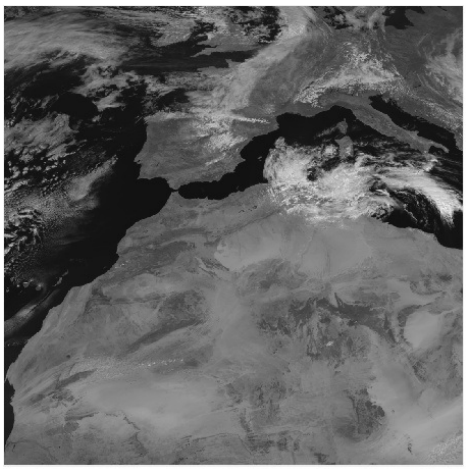


FIGURE 1.4 Image de MSG du canal VIS 0.8 du 01/04/2008 à 12 :00 [1]

2.3.2 Les canaux vapeur d'eau

Le MSG possède aussi deux canaux de vapeur d'eau : WV 6.2 et WV 7.3, ils fournissent des images (figure 1.5 et figure 1.6) jour et nuit de 3712 lignes de 3712 pixels. Ces images représentent une mesure du rayonnement infrarouge émis par la vapeur d'eau dans l'atmosphère. Cela aide à distinguer les zones sèches et les zones humides. Lorsque l'atmosphère est riche en vapeur d'eau, les rayons qui ont une longueur d'onde comprise dans l'intervalle [5.35 μm, 7.85 μm] sont absorbés, mais, quand l'atmosphère est pauvre en vapeur d'eau, ces rayons traversent l'atmosphère vers les deux canaux WV 6.2, WV 7.3. Plus l'atmosphère est chargée de vapeur d'eau moins ils la traversent. Les images fournies par ces deux canaux sont utilisées aussi pour déduire une éventuelle instabilité atmosphérique locale pouvant entraîner une convection et des orages violents.

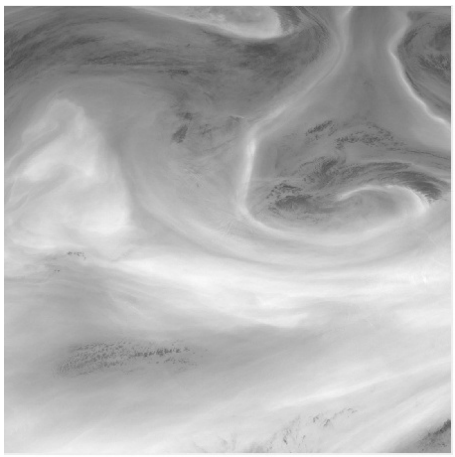


FIGURE 1.5 Image de MSG du canal WV 6.2 du 01/04/2008 à 12 :00 [1]

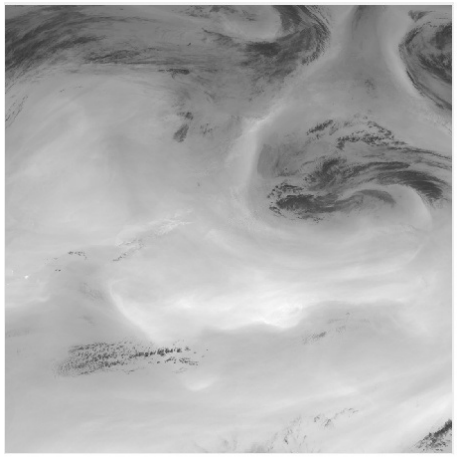


FIGURE 1.6 Image de MSG du canal WV 7.3 du 01/04/2008 à 12 :00 [1]

2.3.3 Les canaux infrarouges

MSG possède 7 canaux infrarouges, nous expliquons leurs caractéristiques ci-dessous :

— **Le canal NIR 1.6 :** ce canal proche infrarouge permet de distinguer la différence entre la neige et les nuages, entre les nuages de glace et d'eau. Les images de ce canal (figure 1.7) sont importantes pour l'aviation, du fait de ses capacités à détecter les nuages de glace. Ce canal apporte aussi des informations sur les aérosols et la végétation en combinaison avec les 2 canaux visibles VIS 0.6 et VIS 0.8.

FIGURE 1.7 Image de MSG du canal NIR 1.6 du 01/04/2008 à 12 :00 [1]

— **Le canal IR 3.9 :** ce canal est particulièrement utilisé pour détecter le brouillard et les nuages bas. Les images de ce canal (figure 1.8) sont également utiles pour mesurer la température nocturne des terres et des océans et pour détecter les incendies de forêt.

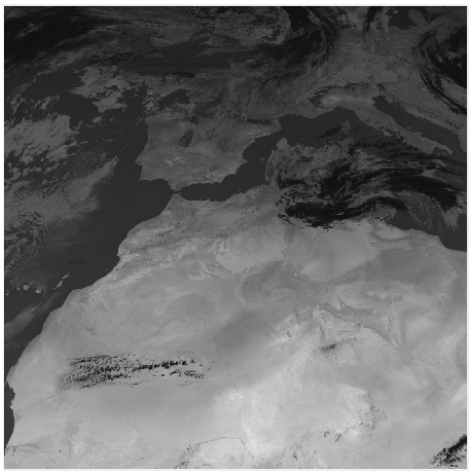


FIGURE 1.8 Image de MSG du canal IR 3.9 du 01/04/2008 à 12 :00 [1]

— **Le canal IR 8.7 :** les images du canal IR 8.7 (figure 1.9) sont principalement utilisées pour fournir des informations quantitatives sur les masses nuageuses minces des cirrus et la discrimination entre les nuages de glace et les nuages d'eau. Le canal IR 8.7 est également nécessaire pour l'identification de l'instabilité atmosphérique.

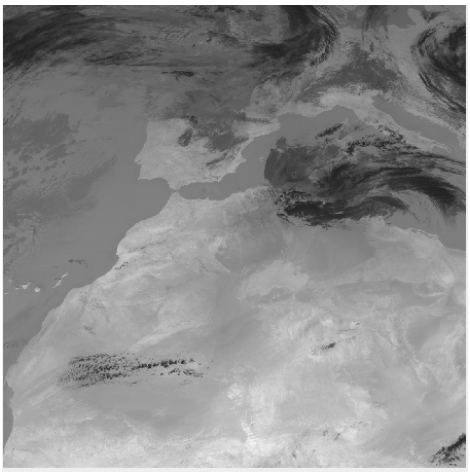


FIGURE 1.9 Image de MSG du canal IR 8.7 du 01/04/2008 à 12 :00 [1]

— **Le canal IR 9.7 :** ce canal est axé sur l'ozone. Il mesure la concentration d'ozone dans la basse stratosphère. Les images de ce canal (figure 1.10) sont utilisées pour surveiller l'ozone total et la hauteur de la tropopause. Il est susceptible de suivre les tendances de l'ozone en tant qu'indicateur des champs de vent à ce niveau.

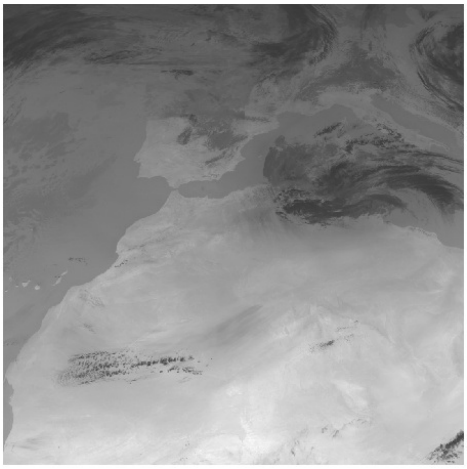


FIGURE 1.10 Image de MSG du canal IR 9.7 du 01/04/2008 à 12 :00 [1]

— **Les canaux IR 10.8 et IR 12.0 :** les images de ces deux canaux infrarouges thermiques (figure 1.11 et figure 1.12) permettent de fournir une observation continue sur les nuages et les sommets des nuages. Ainsi qu'une estimation de la température des nuages, des terres et des surfaces marines. Ces canaux sont également utilisés pour mesurer les effets atmosphériques dans les couches inférieures de l'atmosphère, détection des cirrus et déduction des quantités d'eau précipitable au-dessus de la mer. Ils servent aussi

au suivi des nuages pour déterminer les vents atmosphériques et estimer l'instabilité atmosphérique. Le canal IR 10.8 fait suite au canal IR de MFG.

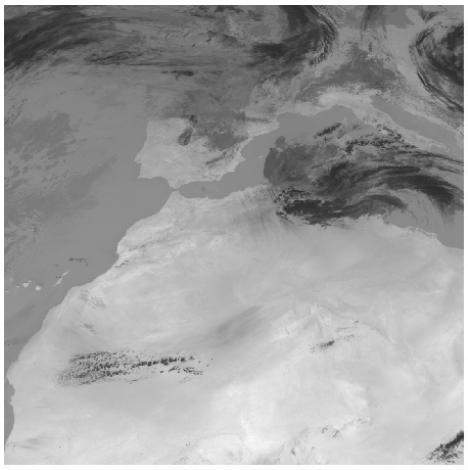


FIGURE 1.11 Image de MSG du canal IR 10.8 du 01/04/2008 à 12 :00 [1]

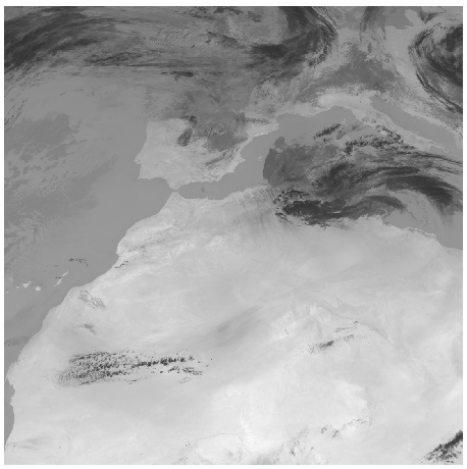


FIGURE 1.12 Image de MSG du canal IR 12.0 du 01/04/2008 à 12 :00 [1]

— **Le canal IR 13.4 :** ce canal se situe dans la bande d'absorption du CO2, les images de ce canal (figure 1.13) sont destinées à estimer l'instabilité atmosphérique et contribuent à fournir des informations sur la température de la basse troposphère, la détermination de la hauteur des nuages semi-transparents, à l'évaluation de la pression aux sommets des nuages.

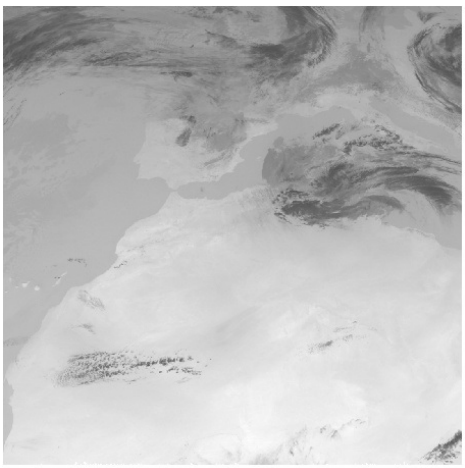


FIGURE 1.13 Image de MSG du canal IR 13.4 du 01/04/2008 à 12 :00 [1]

2.4 Caractéristiques des images MSG

Les données d'une image du satellite MSG sont plus qu'une image, ce sont des mesures d'énergie électromagnétique réfléchie ou émise par la surface de la Terre par le capteur SEVIRI. Les données d'image sont stockées dans un format d'une matrice de lignes et colonnes représente la surface totale balayée par le capteur. Un élément ou un carré de la matrice est appelé pixel, pour chaque pixel, la mesure est stockée sous forme de valeur numérique ou compte numérique (CN). Pour chaque bande de longueur d'onde particulière mesurée, une image (couche) distincte est stockée (figure 1.14).

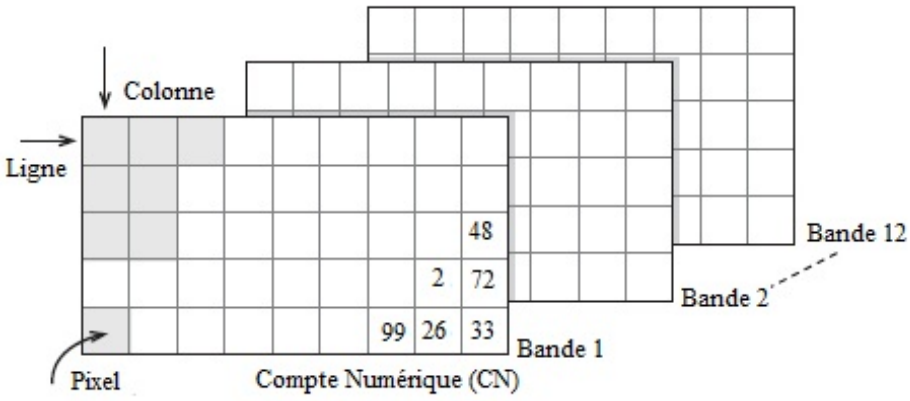


FIGURE 1.14 Caractéristiques des images MSG

Les images MSG sont caractérisées par leurs résolutions spatiale, spectrale, radiométrique et temporelle.

2.4.1 La résolution spatiale

Représente la mesure des objets les plus petits discernés par le capteur SEVIRI du satellite MSG ou alors la surface du sol représentée par chaque pixel [32]. Chaque pixel représente une superficie de 1 kilomètre sur 1 kilomètre pour l'image HRV et de 3 kilomètres sur 3 kilomètres pour les autres images (voir figure 1.15). La taille du pixel au sol est une caractéristique importante puisque c'est elle qui déterminera les éléments pouvant être distingués sur une image.

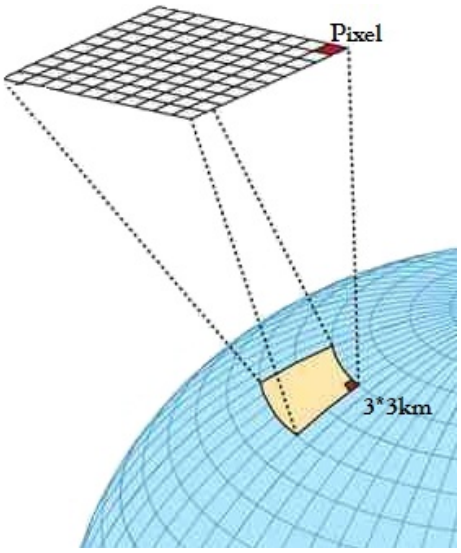


FIGURE 1.15 La résolution spatiale

2.4.2 La résolution spectrale

La résolution spectrale décrit le nombre et la largeur des bandes spectrales. L'intervalle de longueur d'onde spécifique dans le spectre électromagnétique varie d'une bande à une autre [108], par exemple l'intervalle du canal spectral VIS 0.8 = [0.74 – 0.88 μm]. Si l'intervalle est petit donc la résolution spectrale est fine sinon elle est grossière, par exemple l'eau ou la végétation ne nécessitent pas une résolution spectrale fine, à l'inverse des roches et minéraux qui réclament l'utilisation d'un intervalle de longueurs d'onde beaucoup plus fine (voir figure 1.16). Si la résolution spectrale est trop grossière, il ne sera alors plus possible de bien différentier les différents minéraux. Si le capteur contient plusieurs bandes spectrales (plus de 2 et moins de 20 bandes), elles sont appelées multi-spectrales et si le nombre de bandes spectrales est exprimé en centaines, elles sont appelées hyper-spectrales.

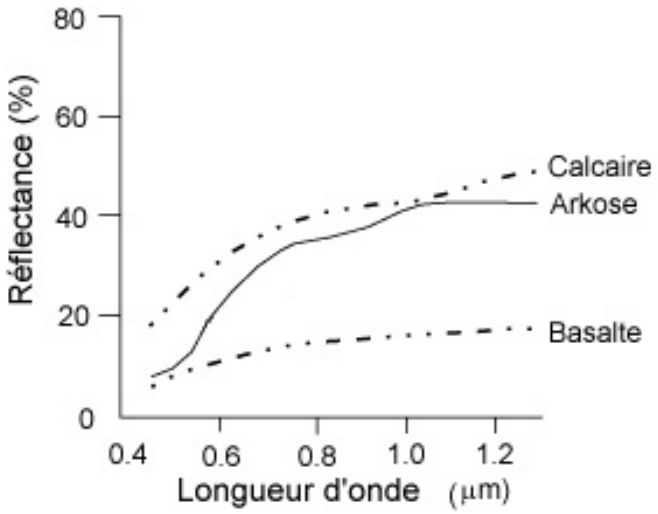


FIGURE 1.16 La résolution spectrale

2.4.3 La résolution radiométrique

La résolution radiométrique [68] est la précision avec laquelle le capteur SEVIRI divise la mesure énergétique qu'il reçoit dans chaque bande, exprimée en nombre de bits par lequel l'énergie enregistrée est répartie dans le fichier image. Pour 8 bits, il existe 2^8 (entre 0 et 255) valeurs radiométriques possibles, chaque valeur radiométrique est représentée par un niveau de gris, la valeur 0 indique les rayons non arrivés au capteur et donc avec une couleur noir dans l'image visible, puis ce numéro varie selon l'intensité des rayons jusqu'à ce qu'atteigne un nombre numérique maximal 255 qui représente la couleur blanche (voir figure 1.17).

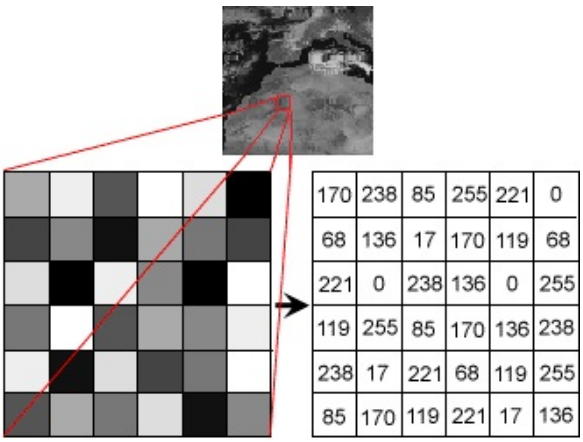


FIGURE 1.17 La résolution radiométrique

La mesure radiométrique (valeur d'un pixel) de l'image visible ou infrarouge est variée selon l'objet ciblé [10] comme montrent les deux figures (figure 1.18 et figure 1.19).

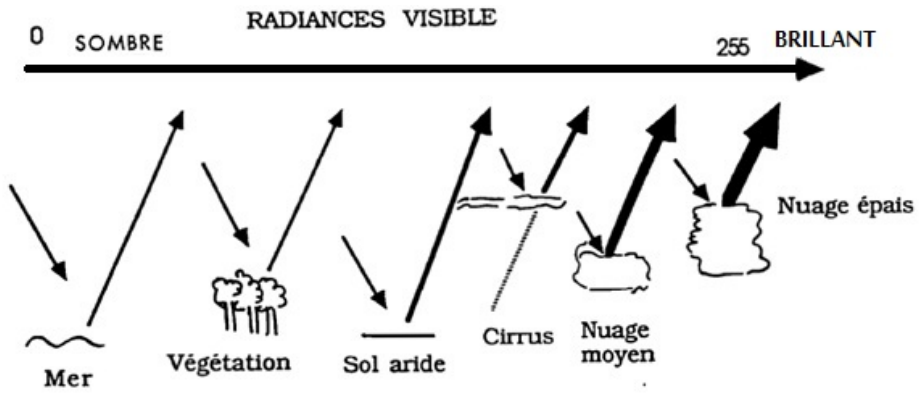


FIGURE 1.18 Les mesures radiométriques en fonction de l'objet ciblé dans une image visible

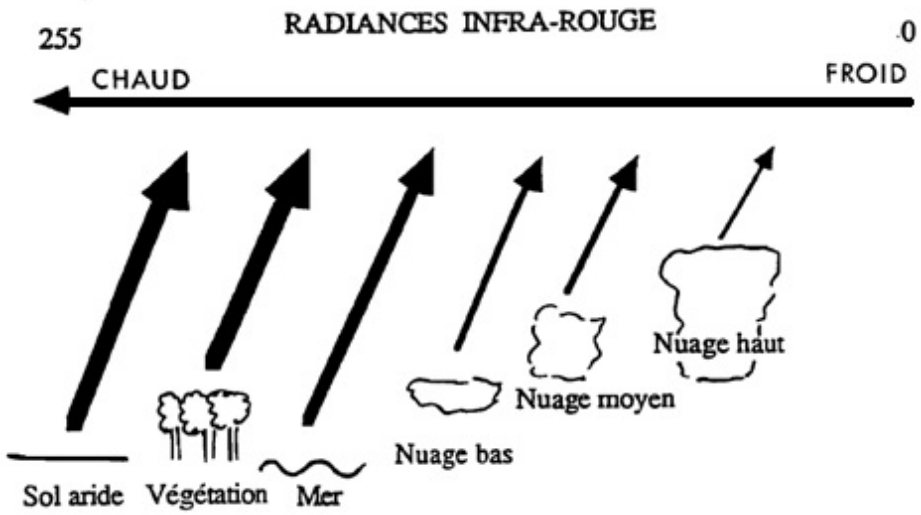


FIGURE 1.19 Les mesures radiométriques en fonction de l'objet ciblé dans une image infrarouge

L'intensité mesurée par le capteur SEVIRI est codée numériquement à l'origine sur 10 bits [112], pour des raisons de commodités, elle est codée sur ordinateur sur 8 bits (comme notre cas) ou sur 16 bits.

2.4.4 La résolution temporelle

La résolution temporelle est une mesure de cycle ou de la fréquence de répétition avec laquelle un capteur revisite la même partie de la surface de la Terre [108], c'est-à-dire le temps entre deux acquisitions d'images successives au même endroit sur Terre, le satellite MSG peut prendre des images de la même région toutes les 15 minutes.

3 Méthodes d'estimation des précipitations à partir des images satellitaires

Les précipitations sont un paramètre météorologique important difficile à évaluer et à mesurer, en particulier dans des zones telles que les déserts, les mers, les océans et les montagnes. La vaste couverture spatiale continue des satellites nous permet de fournir des données partout dans le monde en temps continu, plusieurs méthodes ont été développées pour estimer les précipitations en utilisant les données de ces satellites, elles sont divisées en trois catégories principales. Des méthodes [28, 9, 2, 55, 80, 20] qui utilisent les informations infrarouges / visibles (IR / VIS) pour trouver la relation entre les informations satellitaires et la quantité de précipitations observée mesurée sur le sol. Des méthodes micro-ondes [132, 85] qui utilisent les mesures satellitaires obtenues dans les Micro-ondes (MO). D'autres méthodes hybrides [70, 69, 124] qui exploitent les avantages des deux méthodes précédentes, en combinant les données IR/VIS et les données Micro-ondes. Nous nous focalisons sur les méthodes IR / VIS et en particulier les travaux qui ont été faits aux images multi-spectrales MSG utilisées dans le cadre de notre travail de recherche. Les radiomètres à Micro-ondes ne sont pas disponibles sur le satellite MSG mais on va l'expliquer en bref pour prendre une idée sur les méthodes combinées (MO-IR / VIS).

3.1 Les méthodes basées sur l'Infrarouges / Visibles (IR / VIS)

Les approches basées sur les images IR / VIS tentent de trouver une relation entre les observations de la quantité de précipitations sur le sol et les propriétés microphysiques des nuages notamment la température de leurs sommets et leurs épaisseurs. Les précipitations peuvent être déduites à partir des images VIS, où les sommets des masses nuageuses apparaissent plus brillants que les surfaces terrestres, cette brillance dépend de l'épaisseur des nuages, les nuages épais sont plus susceptibles d'être associés aux précipitations. Généralement, les données de bande visible seules sont rarement utilisées pour l'estimation des précipitations. La température du sommet des nuages peut être obtenue à partir de données IR, c'est un indicateur de leurs altitudes et potentiellement de leurs épaisseurs. Les précipitations plus abondantes ont tendance à être associées à des nuages plus grands et plus hauts avec des sommets plus froids. Les approches IR / VIS sont dites indirectes car la température aux sommets des nuages dans IR ou la brillance des nuages dans le VIS fournissent une évaluation indirecte des précipitations. Ces approches sont ensuite divisées en fonction des caractéristiques physiques et dynamiques des précipitations en cinq méthodes : 1) méthodes des indices nuageux, 2) méthodes bi-spectrales, 3) méthodes multi-spectrales, 4) méthodes

basées sur le cycle de vie d'un nuage et 5) méthodes basées sur la modélisation des processus physiques des nuages.

3.1.1 Méthodes d'indices du nuage

Ces méthodes utilisent des seuils déduits empiriquement pour la mesure de la température au sommet du nuage (indice nuageux) dans l'IR thermiques (10.5 à 12.5 μm) du satellite pour détecter la zone de pluie, à laquelle un taux de pluie fixe est attribué. Ces méthodes sont principalement utilisées dans les études climatologiques avec des données de précipitations agrégées dans le temps et dans l'espace. Le taux de précipitations assigné est basé sur une fonction de transfert statistique entre la température au sommet du nuage (température de brillance (TB)) dans le canal IR et le taux de précipitations mesuré avec des jauges conventionnelles ou un radar au sol. La méthode la plus utilisée est celle qui est développée par Arkin and Meisner [9], connue sous le nom Goes Precipitation Index (GPI), ils ont déterminé une corrélation entre les précipitations estimées par radar sur une zone de 2,5° x 2,5° (latitudes / longitude) et les pixels observés dans l'image IR thermique du satellite GOES sur une période de 12 heures chaque jour. Après plusieurs tests empiriques, ils ont découvert qu'un pixel peut être considéré pluvieux si sa température au sommet des nuages mesurée dans la bande infrarouge thermique est inférieure à un seuil de -38°C (235 K (Kelvin)), un taux de pluie constant de 3 mm/ h est attribué à tous les pixels d'image dont la température est inférieure à ce seuil. Les précipitations cumulées (PC) est données par l'équation 1.1 :

$$Si\ TB \leq 235\ k\ alors\ PC = F * \triangle T * 3\ mm/h,\ sinon\ PC = 0\ mm/h, \tag{1.1}$$

Où :

— F : la fraction de la zone couverte par les pixels inférieure à 235 K en IR.

— $\triangle T$:le temps en heures entre deux observations successives.

Cette technique est simple et utile pour l'estimation du taux de pluie à long terme et dans une zone étendue, elle est plus adaptée pour estimer les précipitations cumulées mensuelles que d'estimer des précipitations pour des périodes plus courtes. Kerrache et Schmetz [72] ont reconstruit le GPI à partir des images MFG. A partir de ces images, Menz and Zock [88], Todd et al. [122] ont utilisé le GPI avec succès en Afrique de l'Est. Une autre méthode a été développée par Ba et Nicholson [12], ils ont analysé l'activité convective et sa relation avec les précipitations, à l'aide d'un indice de convection basé sur les données IR du satellite MFG et les mesures de précipitations. Ils ont trouvé corrélations positives qui justifient l'utilisation de cet indice pour les estimations des précipitations annuelles et mensuelles.

3.1.2 Méthodes bi-spectrales

Les méthodes des indices nuageux utilisent uniquement l'information infrarouge pour l'estimation des précipitations, cette estimation peut être améliorée en incluant les informations provenant du canal visible. Ces méthodes tentent de combiner et coupler les données des canaux visible et infrarouge à la fois, elles sont basées sur la relation entre la température et la brillance des nuages mesurées dans la bande infrarouge thermique et visible respectivement, les nuages qui ont le plus de chance de précipitations doivent être à la fois froids et brillants qui sont des caractéristique des nuages cumulonimbus. Les précipitations sont moins probables avec des nuages brillants mais chauds (exemple : nuages stratiformes) ou vice versa des nuages moins brillants mais froids (exemple : nuages cirrus). Bellon et al. [15] développent la méthode (RAINSAT), ils ont éliminé les nuages froids mais pas très réfléchissants ou ceux qui sont très réfléchissants mais dont le sommet est relativement chaud, comme résultat le nombre de détection de précipitations fausses des techniques IR pures est réduit. Cheng et Brown [30] ont appliqué la méthode RAINSAT en utilisant les images MFG en Angleterre, et ont obtenu des résultats optimisés. King et al. [74], ont comparé les techniques bi-spectrales et infrarouges, ils ont montré qu'une amélioration considérable est obtenue par rapport aux techniques IR, en utilisant les données VIS.

3.1.3 Méthodes multi-spectrales

Contrairement aux méthodes bi-spectrales, les méthodes multi-spectrales prennent en compte les données provenant de plus de deux canaux spectraux. Ces données multi-spectrales visible, proche infrarouge ainsi que dans l'infrarouge vapeur d'eau et thermique obtenues à partir des capteurs embarqués sur des satellites géostationnaires ont été largement utilisées pour caractériser les nuages précipitants. Par exemple, Ba et Gruber [11] ont développé une approche qui combine les informations provenant de cinq canaux du satellite GOES (0.65 μm, 3.9 μm, 6.7 μm, 11.0 μm et 12.0 μm) pour la détection des zones de nuages précipitants. Seuls les nuages plus froids (TB_{IR11}) que les 230 K sont pris en compte pour le filtrage. Le canal 12.0 μm est utilisé conjointement avec le canal 11 μm pour estimer la température au sommet des nuages, cette température estimée est utilisée pour calculer l'émission thermique à 3.9 μm, qui est ensuite soustraite des mesures de ce canal pour obtenir le rayonnement solaire réfléchi dans la bande spectrale 3.9 μm, un seuil de 15.0 μm de rayon effectif des gouttelettes est utilisé comme limite inférieure des nuages précipitants. Tous les nuages ayant une réflectance visible (0.65 μm) supérieure à 0.40 sont pris en compte pour le filtrage. Le taux de pluie est attribué par le produit de la probabilité de pluie et du taux de

pluie moyen calculé en fonction de (TB_{IR11}), ensuite, il est ajusté par un facteur d'humidité (6.7 μm).

Nous citons ci-après quelques méthodes utilisant les données multi-spectrales du satellite MSG :

Wolters et al. [129] ont montré que l'utilisation simultanée de la température de brillance $TB_{IR10.8}$ et la différence de température de brillance $\triangle TB_{IR8.7-IR10.8}$ permet une identification plus précise des phases thermodynamiques des nuages. Ils ont constaté que la formation de cristaux de glace commence lorsque $TB_{IR10.8} < 238$ K et $\triangle TB_{IR8.7-IR10.8} > 0.25$ K. les nuages sont susceptibles de produire des précipitations lorsque des cristaux de glace sont présents au sommet des nuages.

Thies et al. [121] ont développé une technique pour déterminer les nuages précipitants en utilisant les images MSG pendant le jour. La méthode utilise les réflectances dans les canaux VIS0.6 et NIR1.6 pour obtenir des informations sur des propriétés des nuages telles que le rayon effectif et l'épaisseur optique du nuage, et les différences de température de brillance $\triangle TB_{IR8.7-IR10.8}$ et $\triangle TB_{IR10.8-IR12.0}$ pour avoir des informations sur la phase des nuages et classer les nuages en glace ou eau. Les mêmes auteurs ont développé une autre méthode [120] d'identification des nuages précipitants dans les systèmes stratiformes pendant la nuit à partir des images MSG, Ils ont exploité les différences de température de brillance suivant : $\triangle TB_{IR3.9-IR7.3}$, $\triangle TB_{IR3.9-IR10.8}$, $\triangle TB_{IR8.7-IR10.8}$ et $\triangle TB_{IR10.8-IR12.0}$.

Feidas et Giannakos [41] ont développé deux méthodes pour délimiter les zones de précipitations en utilisant les différences de température de brillance du canal IR10.8 avec les canaux WV6.2, IR8.7 et IR12.0. Ils ont également introduit deux techniques [42, 51] permettant de classifier les nuages précipitants convective et stratiforme en fonction des caractéristiques spectrales et texturales des données SEVIRI. Roebeling et Holleman [109] ont présenté une méthode pour détecter les précipitations et estimer les taux de pluie à l'aide des propriétés optiques et microphysiques des nuages, ils ont exploité les images extraites du satellite MSG dans les canaux visibles, proche infrarouge et l'infrarouge thermique.

Lazri et al. [77] ont introduit une nouvelle méthode basée sur un réseau neuronal artificiel (ANN) pour identifier les nuages précipitants pendant le jour et la nuit à partir des données du satellite MSG. Les entrées de l'ANN sont les données du satellite MSG, et les sorties sont les deux classes (précipitations, pas de précipitations) du radar de Sétif.

Ouallouche et al. [98] ont également présenté une méthode basée sur l'ANN pour délimiter les zones de précipitations, ils ont utilisé quatre paramètres des canaux infrarouges du satellite MSG et trois paramètres du satellite Tropical Rainfall Measuring Mission (TRMM) comme entrées de la méthode ANN, et les sorties sont les deux classes (précipitations, pas de précipitations) du radar de précipitation embarqué sur le satellite TRMM. Ils ont proposé

aussi une méthode [99] basée sur l'algorithme de forêts aléatoires (RF) en utilisant les données MSG. Le RF comprend deux parties principales : la classification et la régression. Le RF classe les pixels des images MSG en trois classes (pas de précipitations, convective et stratiforme), ensuite, la régression RF est utilisée pour attribuer un taux de précipitations aux pixels appartenant aux classes convectives et stratiformes.

Sehad et al. [114] ont développé aussi une méthode basée sur ANN pour identifier les nuages précipitants et de les classifier en deux classes convectifs et stratiformes pendant le jour et la nuit, les entrées de chaque ANN correspondent aux caractéristiques spectrales, texturales et temporelles déterminés pour chaque pixel des images MSG. Les sorties de chaque ANN sont les trois classes (pas de précipitations, convectif, stratiforme) du radar de Sétif.

Une nouvelle technique d'estimation des précipitations a été introduite par Bensafi et al. [18], qui est basée sur les k-plus proches voisins pondérés (WKNN) en utilisant les données multi-spectrales MSG pour déterminer le taux de précipitations d'un pixel parmi 16 taux prédéfinis observés sur le radar météorologique de Sétif.

3.1.4 Méthodes basées sur le cycle de vie d'un nuage

La base de ces méthodes est que l'intensité des précipitations varie en fonction de l'étape du cycle de vie des nuages, particulièrement les nuages convectifs qui sont responsables pour une part significative des précipitations et qui peuvent être différenciés, dans les images satellitaires, des autres types de nuages. L'observation de l'état d'évolution du nuage est faite par des données visibles et infrarouges successives provenant d'images satellites géostationnaires, parce qu'ils ont une résolution temporelle courte (15-30 min) en comparaison avec le cycle de vie des nuages. L'une des techniques les plus largement utilisées est connue sous le nom de Griffith-Woodley présenté par Griffith et al. [53], à partir d'une série d'images où les nuages sont très froids (IR$\leq$ 253 k), ils ont observé qu'après une séquence de mesures de l'aire des nuages convectifs (cumulonimbus) déterminée par les images VIS ou IR, qu'il y'a une relation entre cette aire nuageuse qui varient en fonction de l'évolution du nuage et la surface d'échos radar au sol. Negri et al. [95] ont présenté une technique (NAW, Negri – Adler – Wetzel) basée sur la méthode de Griffith-Woodley, après une analyse détaillée de cette méthode. Ils ont découvert que, au-delà de l'échelle temporelle de 1-3 heures, les précipitations sont beaucoup plus fortement liées avec la surface de nuages froids qu'avec les paramètres liés au cycle d'activité des nuages. Pour le même seuil de température (253 k), ils attribuent différents taux de pluie selon la température la plus froide, un taux de précipitations de 8 mm/h pour les 10% des pixels qui ont une température de brillance plus froide et un taux de 2 mm/h pour les 40% suivants pixels froids, les pixels restants (50%) sont considérés

comme non précipitants. En utilisant l'imagerie infrarouge de MFG, Levizzani et al. [79], et Amorati et al. [7] ont montré que la méthode NAW donne de bons résultats avec des systèmes convectifs étendus aux latitudes moyennes sur des intervalles spatio-temporels appropriés.

3.1.5 Méthodes basées sur la modélisation des processus physiques des nuages

Cette approche utilise des modèles de nuages afin de délimiter les nuages convectives et stratiformes du système. La majorité des systèmes précipitant sont d'origine convective, mais, des recherches [47, 67] ont montré que les précipitations stratiformes représentent environ 40% de la quantité totale des précipitations observées pour certains systèmes. Ils ont classifié les précipitations en deux grandes catégories : convectives et stratiformes. La technique CST (Convective-Stratiform Technique) développée par Adler et Negri [2] est la plus utilisée, elle permet d'estimer les précipitations à partir des nuages convectives et stratiformes en utilisant les données satellitaires infrarouges, les minima de température locaux dans le canal IR sont utilisés dans un premier temps pour identifier les noyaux de convection. Ils ont appliqué une fonction discriminante déterminée de manière empirique sur la base de données radar au sol pour exclure les minimums correspondant aux cirrus (sans pluie). Ensuite, un modèle de nuages est utilisé pour déterminer le taux de pluie convective, plus la température est froide plus le taux de pluie est élevé. Un seuil de 208 k est utilisé pour caractériser les zones stratiformes. Ces zones détectées obtiennent un taux de pluie fixe de 2 mm / h. En utilisant les images MFG, la méthode CST est validée et appliqué avec succès dans les régions tropicales [16, 17]. Cependant, des études faites par Negri et Adler [94], Pompei et al. [104] ont montré qu'elle a des limites dans les régions extratropicales. Pour cette raison, Reudenbach et al. [106] ont développé et amélioré la technique CST en ECST (Enhanced Convective-Stratiform Technique) dans ces régions, en incluant le canal de vapeur d'eau du MFG pour une distinction plus fiable entre les nuages convectifs et les nuages stratiformes et même les cirrus. Pour appliquer l'ECST au MSG, Thies et al. [121] ont examiné quels canaux de vapeur d'eau (WV) et IR du MSG sont les plus appropriés pour le remplacement. Il a été démontré que l'ECST est applicable au MSG, et que le meilleur accord est obtenu en utilisant les canaux WV7.3 et IR12.0.

3.2 Les méthodes micro-ondes

Contrairement aux observations visibles et infrarouges qui ne peuvent détecter que la couche supérieure des nuages, certains capteurs d'un satellite peuvent détecter le rayonnement micro-ondes (MO) émis et diffusés par les hydrométéores (gouttes d'eau et cristaux de glace) contenues dans les nuages. La possibilité de détecter et évaluer à distance les précipitations

à partir d'un rayonnement MO repose sur la compréhension de l'interaction directe entre ce rayonnement et les hydrométéores. Cette interaction dépend de la fréquence d'onde, et selon les propriétés de diffusion et de l'absorption/ émission du rayonnement par les hydrométéores. Spencer et al. [117] ont calculé ces propriétés dans trois fréquences d'onde principales (19, 36.5 et 86 GHz) qui ont été utilisées pour mesurer les précipitations (figure 1.20). Ils ont observé que les gouttes liquides absorbent et diffusent les MO, mais l'absorption est dominante aux basses fréquences (19 et 36,5 GHz), tandis que les particules de glace n'absorbent pas le rayonnement MO quel que soit sa fréquence, mais ils diffusent un MO important aux fréquences élevées (86 GHz). La diffusion et l'absorption augmentent avec la fréquence et le taux de pluie, et la diffusion par les particules de glace augmente beaucoup plus rapidement avec la fréquence 86 GHz que la diffusion par liquide.

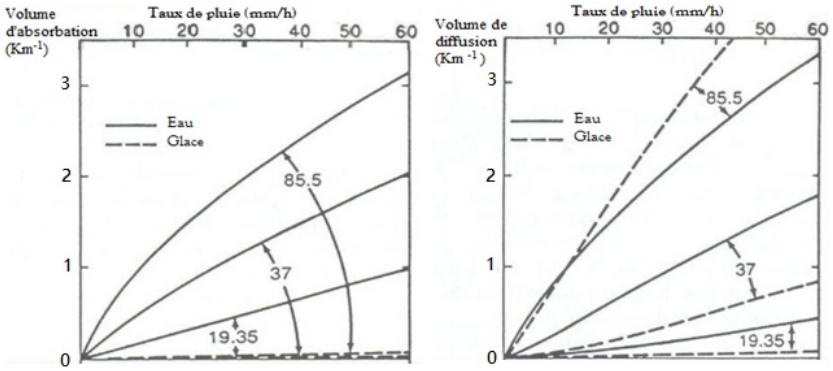


FIGURE 1.20 Volume d'absorption et de diffusion en fonction du taux de pluie, pour les gouttes d'eau et pour les cristaux de glace, aux 3 différentes fréquences principales [117]

Plusieurs méthodes d'estimation des précipitations utilisant les MO ont été développées, divisées en deux catégories principales, à savoir méthodes d'émission et celles de diffusion :

— Méthodes d'émission : sont des méthodes micro-ondes à basse fréquence, basées sur l'émission accrue par les gouttes de pluie à ces fréquences, des exemples de cette approche ont été développés par Ferraro et Marks [43] et Berg et Chase [19].

— Méthodes diffusion : sont des méthodes à haute fréquence, fondées sur l'atténuation par diffusion du rayonnement ascendant à ces fréquences par les particules de glace, des exemples de cette approche ont été développés par Ferraro et al. [44] et Grody [54].

L'avantage des méthodes MO est lié au fait que les interactions entre le rayonnement MO et les hydrométéores offrent une mesure plus directe des précipitations, mais, l'inconvénient le plus critique est qu'elles ont une faible résolution spatiale et temporelle en raison de leur grande longueur d'onde et de l'orbite non géostationnaire du satellite.

3.3 Les méthodes hybrides (combinées)

Ces techniques profitent des avantages des deux méthodes précédentes (IR/VIS et MO), en exploitant la bonne résolution spatiale et temporelle des données IR/VIS et la plus grande précision physique des méthodes MO. Plusieurs techniques qui combinent les données IR/VIS et MO ont été développées, la méthode la plus utilisée est celle de Jobard and Desbois [70], connue sous le nom de RACC (Rain And Cloud Classification), qui repose sur une procédure de classification automatique divisée en deux phases. Dans la première phase, la méthode utilise un ensemble d'images du canal IR du satellite MFG et les données MO du radiomètre SSM/I en coïncidence spatio-temporel pour obtenir des classes homogènes caractérisant les différents types de nuages. Dans la deuxième phase toutes les images IR sont utilisées. Chaque pixel de chaque image IR est attribué à une classe de nuages (dans la classification trouvé précédemment). Ensuite, des taux de précipitations obtenus en fonction de la valeur du paramètre MO de chaque classe sont attribués à ces images. Adler et al. [3] ont proposé la méthode de GPI ajusté (AGPI) qui corrige les estimations mensuelles de précipitations de GPI en utilisant un facteur d'ajustement basé sur les données MW et IR. Turk et al. [125] ont proposé une technique mixte MO-IR afin d'estimer les précipitations en temps quasi-réel, la technique consiste à combiner statistiquement des données MO satellitaires en orbite terrestre basse (SSM/I ou TRMM) avec des données coïncidant spatio-temporel de l'un des quatre satellites géostationnaires (GOES-Est / Ouest, GMS et METEOSAT) pour une analyse rapide et à jour des précipitations. Cette technique a été prise comme modèle par plusieurs chercheurs (Kidd et al. [73], Marzano et al. [85]), et par le produit MPE (Multisensor Precipitation Estimate) [63] développé au niveau de EUMETSAT, il représente un étalonnage (calibration) continu des températures de brillance IR du satellite METEOSAT par rapport aux taux de pluie du satellite SSM/I. Le produit MPE est constitué de cartes des taux de pluie en temps quasi réel pour chaque image MSG.

4 Conclusion

Dans ce chapitre, nous avons présenté le satellite MSG d'où proviennent les images sur lesquelles nous avons réalisé notre étude. Ainsi, nous avons décrit les caractéristiques principales de ces images MSG, radiométrique, spatiale, spectrale et temporelle. Ensuite, nous avons présenté les différentes méthodes d'estimation des précipitations à partir des images satellitaires qui ont été développées et en particulier celles qui ont été faites aux images multi-spectrales MSG. Dans ce qui suit, nous présentons les principales méthodes du datamining qui nous permettent l'extraction de connaissances à partir des images satellitaires.

Chapitre 2

Extraction de connaissances à partir des images satellitaires

1 Introduction

La majorité des méthodes d'estimation des précipitations à partir des images satellitaires présentées dans le chapitre précédent ont été basées sur l'exploitation des propriétés optiques et microphysiques des nuages telles que leurs épaisseurs optiques, leurs hauteurs, la température de leurs sommets, le rayon effectif des particules qui les composent (la taille des gouttelettes) et la phase thermodynamique des nuages (glace ou liquide)... etc. Ces méthodes ont donné des résultats satisfaisants, néanmoins, ils ne prennent pas en compte les corrélations spectrales et spatiales qui peuvent exister entre les pixels de l'image elle-même, ou bien entre les pixels des différentes images MSG. De plus, ces études classifient les pixels de façon conventionnelle, où un pixel est considéré comme 100 % précipitant ou 0 % (non précipitant), alors qu'en réalité il ne peut pas être classifié de façon claire et sans ambiguïté. De plus, ces méthodes n'utilisent que les données de quelque canaux, ils n'exploitent pas bien toutes les données fournies par ce satellite, alors que ces téraoctets de données sont potentiellement riches en ressources inouïes qui demandent à être exploitées. Les techniques Datamining permettent de les faire apparaître et d'extraire les informations utiles (les connaissances), et de trouver des corrélations cachées entre les données des différentes images MSG grâce à un certain nombre d'outils et techniques notamment des associations, des classifications, des regroupements (Clustering) et des prévisions. Dans ce chapitre, nous allons introduire les notions de bases essentielles de l'extraction de connaissances à partir de données images satellitaires.

2 Extraction de Connaissances à partir de Données

2.1 Définition

L'Extraction de Connaissances à partir de Données (ECD en français) ou Knowledge Discovery in Databases (KDD en Anglais) est défini comme le processus non trivial d'identification de modèles valides, potentiellement utiles et compréhensibles à partir de données volumineuses et complexes [40]. Ce processus doit être automatique ou (plus généralement) semi-automatique. L'objectif fondamental de l'ECD est de découvrir des connaissances utiles, valides, pertinentes et nouvelles à travers l'utilisation de différents algorithmes et méthodes, qui permettent l'amélioration des processus et faciliter la prise de décision. Une confusion subsiste encore entre le Datamining et l'ECD, le Datamining est l'une des étapes cruciales de l'ECD.

2.2 Notions de bases

Le processus d'ECD réside dans l'extraction de connaissances à partir des informations fournies par les données, ces trois notions (donnée, information et connaissance) sont définies comme suit :

— **Donnée** : les données sont des faits et des chiffres non traités, sans interprétation ou analyse ajoutée. Par exemple, 15, ce chiffre en soi n'a aucune signification particulière.

— **Information** : les informations sont des données qui ont été interprétées de manière à avoir un sens pour l'utilisateur. Par exemple, la température aujourd'hui à Constantine est égale à 15° donne un sens aux données et est donc considéré comme une information pour une personne qui suit la météorologie.

— **Connaissance** : la connaissance est une combinaison d'informations, d'expérience et de perspicacité pouvant bénéficier à l'individu ou à l'organisation. Par exemple, lorsque la température baissera à 10° et le taux d'humidité sera supérieur à 75%, il y a une probabilité de 80% qu'il pleut.

3 Extraction de connaissances à partir des images satellitaires

3.1 Définition

L'extraction de connaissances à partir des images est le processus de recherche et de découverte des modèles d'images significatifs, des connaissances implicites et des relations qui peuvent exister entre les données images dans une vaste collection d'images [131, 89].

3.2 Les étapes d'extraction de connaissances à partir des images satellitaires

La figure 2.1 montre les différentes étapes du processus d'extraction de connaissances à partir des images satellitaires. Tout d'abord, les images sont prétraitées et restaurées pour améliorer leur qualité. Ensuite, ces dernières subissent diverses transformations. L'étape suivante consiste à appliquer une ou plusieurs méthodes de Datamining selon les besoins de l'utilisateur et les données exploitées afin d'extraire des patterns et motifs, ces derniers seront évalués et interprétés pour obtenir des connaissances finales qui peuvent être exploitées et appliquées aux applications des différents domaines. Le processus d'extraction de connaissances à partir des images comprend les étapes suivantes :

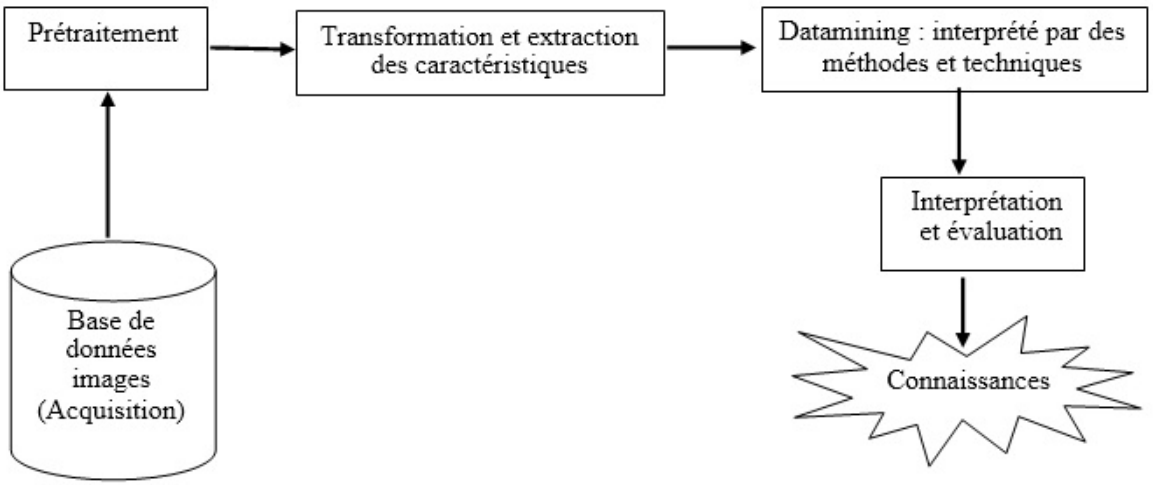


FIGURE 2.1 Les étapes du processus d'extraction de connaissances à partir des images satellitaires [131]

3.2.1 Étape d'acquisition

Le centre européen des opérations spatiales (ESOC), situé à Darmstadt en Allemagne reçoit des signaux sous forme d'images brutes à partir du satellite METEOSAT grâce à ses équipements de télécommunications, où chaque pixel représente la valeur numérique de l'énergie électromagnétique du rayonnement réfléchie ou émise par la surface de la Terre. Dans ce centre, ces images subissent tout d'abord un prétraitement (sera expliqué dans l'étape suivante), ensuite, elles sont stockées et transmises via METEOSAT vers les stations météorologiques munies d'un système de réception conçu pour ce type d'images [29], dans notre pays, elles sont captées au niveau de la station radiométrique de l'Office National de la Météorologie (ONM). Les pays utilisateurs d'images METEOSAT, notamment notre pays, sont chargés à leur tour de transmettre vers l'ESOC toutes les données météorologiques collectées au niveau de leurs centres météorologiques. Ces derniers sont équipés de stations automatiques jouant le rôle de plate-forme de collecte de données (DCP ou Data Collection Platform) et reliées par voie hertzienne au centre de Darmstadt par le biais de canaux de transmission du satellite METEOSAT.

3.2.2 Étape du prétraitement

Il existe deux caractéristiques principales des images satellitaires, géométriques et radiométriques. Les caractéristiques géométriques basées sur les coordonnées de l'image et les caractéristiques radiométriques décrivent le contenu réel de l'information dans l'image. Les données brutes des capteurs ne représentent pas avec précision les cibles au sol, car ces données sont sujettes à des distorsions lors du processus de balayage et de numérisation. De nombreux facteurs sont à l'origine de cette distorsion, tels que le mouvement du capteur pendant le balayage, la vitesse de la plate-forme, les reliefs du terrain, la courbure et la rotation de la Terre. Tous ces facteurs produisent des distorsions radiométriques et géométriques [13], par conséquent, les valeurs numériques des pixels d'une image numérique ne représentent pas complètement l'énergie reflétée par la cible au sol correspondante, et la localisation spatiale de la cible terrestre sur l'image numérique n'est pas strictement corrélée à sa position sur le sol. Donc, nous avons besoin de prétraiter ces données des images afin de garantir l'extraction d'une bonne connaissance. Les images reçues par le satellite MSG sont restaurées au niveau du centre de Darmstadt, ces images subissent un prétraitement qui consiste à corriger notamment les dégradations subies à cause des distorsions radiométriques et géométriques afin de procéder au traitement permettant d'améliorer les données et d'en extraire des connaissances. Après ce traitement et stockage au centre de Darmstadt, les images prétraitées sont transmises vers les différents centres météorologiques.

3.2.3 Étape de transformation

Les transformations des images impliquent généralement la manipulation de données de plusieurs bandes, qu'elles proviennent d'une seule image ou de plusieurs images de la même zone. Les transformations d'images génèrent de nouvelles images à partir de deux sources ou plus qui mettent en évidence des caractéristiques ou propriétés intéressantes, mieux que les images d'entrée originales. Les principales transformations peuvent être divisées en deux méthodes [86]. Des méthodes de transformation théoriques dans laquelle la conversion est effectuée par des calculs et des opérations arithmétiques de base (+, -,*, /), par exemple l'opération d'addition (+) sur une paire d'images est effectuée pour réduire le bruit et l'opération de soustraction (-) sur une paire d'images est effectuée pour rechercher les différences entre les images afin de détecter les changements. D'autres méthodes de transformation empiriques telles que l'Analyse en Composantes Principales (ACP), la transformation des images en couleur.

3.2.4 Étape du Datamining

Le Datamining est le cœur du processus de l'extraction de connaissances à partir des données. Il consiste à rechercher des corrélations et des modèles intéressants à partir des données volumineuses en utilisant un ensemble d'algorithmes et des méthodes issues de l'apprentissage automatique, de l'intelligence artificielle et de la statistique… etc. [71]. Chaque algorithme ou méthode a des paramètres et des tactiques d'apprentissage, donc, il faut comprendre les conditions dans lesquelles un algorithme du Datamining est le plus approprié à un problème donné, par exemple la classification ou la segmentation (la mise en cluster). Cela dépend principalement des objectifs du processus de l'extraction de connaissances, ainsi que des étapes précédentes. Il est possible de combiner plusieurs méthodes (Méthodes hybrides) pour essayer d'obtenir une solution optimale globale. Nous allons donner un aperçu général sur les principales méthodes et algorithme dans la suite du travail.

3.2.5 Étape d'évaluation et présentation

Cette étape consistera à évaluer la qualité des résultats et des modèles découverts par rapport aux objectifs définis du processus de l'extraction de connaissances à partir des images [71]. Par exemple, on fait des tests de la précision du modèle sur un jeu de données images indépendant qui n'a pas été utilisé pour créer le modèle. Ensuite, des techniques de visualisation et de représentation des connaissances sont utilisées pour présenter les connaissances extraites aux utilisateurs.

4 Les principales taches du Datamining

Le Datamining nous offre de nombreuses méthodes et techniques parmi lesquelles on peut choisir en fonction du type de données images et l'étude que nous souhaitons effectuer. Il nous permet d'effectuer les différentes taches selon les typologies des méthodes suivantes.

4.1 Les méthodes selon les objectifs

4.1.1 La classification

La classification des images peut être définie comme le processus de réduction d'une image en classes d'informations (exemple : nuages, eau, forêt, ... etc). La catégorisation des pixels d'image est basée sur leurs valeurs numériques / niveau de gris dans une ou plusieurs bandes spectrales, elle consiste à analyser les caractéristiques d'un pixel et lui assigner une classe parmi un ensemble prédéfini de classes. Plus précisément, on dispose en données du problème d'un ensemble de classes et de pixels, chacun d'eux étant déjà placé dans une classe qui lui convient au mieux. On dit alors que les pixels sont étiquetés. Le principe de la classification est de collecter le plus d'informations possibles à partir des classes connus. Ceci vise à assurer une bonne connaissance des classes et à comprendre, en quelque sorte, pourquoi deux pixels distincts sont séparés dans des groupes différents, ou au contraire rassemblés. Dans une même classe :

— 2 pixels d'une même classe se ressemblent le plus possible.

— 2 pixels de classes distincts diffèrent le plus possible.

— Le nombre des classes est prédéfini

4.1.2 La segmentation (le clustering)

La segmentation permet de partitionner automatiquement un ensemble de pixels en groupes homogènes (clusters) en fonction de leur similarité (c'est-à-dire qu'elles ont des niveaux de gris similaires), alors que les données de différents clusters doivent être relativement bien séparées (c'est-à-dire qu'elles ont des niveaux de gris très différents). Le principal avantage du clustering par rapport à la classification est qu'il n'y a pas de classe à expliquer ou de valeur à prédire définie a priori, on dispose au départ d'un ensemble de pixels non étiquetés. Typiquement, un clustering sera jugé satisfaisant si on obtient en sortie de la méthode des groupes de pixels homogènes (c'est-à-dire possédant des pixels les plus similaires possibles) qui sont les plus hétérogènes possibles entre eux.

4.1.3 L'association

Cette technique de la feuille de données d'image permet de trouver l'association entre les pixels de l'image spectrale elle-même, ou bien entre les pixels des différentes images spectrales, cette technique est plus utilisée dans l'analyse du panier de marché pour découvrir les associations et les corrélations entre les produits que les clients achètent fréquemment ensemble pour prédire leurs comportements au future.

4.1.4 La prédiction

La prédiction dans le datamining consiste à identifier la valeur d'un pixel en fonction d'autres valeurs des pixels associée. Ce n'est pas nécessairement lié à des événements futurs, mais les valeurs des pixels utilisées sont inconnues.

4.2 Les méthodes selon le type d'apprentissage

4.2.1 Apprentissage supervisé

L'apprentissage supervisé est effectué lorsque l'identité et la localisation de certaines des caractéristiques de l'image, telles que les zones nuageuse, les zones humides, les zones urbaines et les forêts, sont connues a priori grâce aux informations recueillies lors de visites sur le terrain. Il s'agit d'un processus de classification basé sur des informations sur les caractéristiques spectrales de la couverture terrestre dans la zone représentée précédemment, obtenu par le biais de visites sur le terrain, de cartes ou de photographies aériennes couvrant la zone. Ces modèles ou zones sont appelés sites d'entraînement et leurs caractéristiques spectrales servent à guider l'algorithme de classification. Ensuite, chaque pixel de l'image est comparé à chaque modèle et se voit attribuer la classe dont les propriétés sont les plus proches. En d'autres termes, il est affecté à une classe dont il a la plus grande probabilité d'être membre. En bref, l'apprentissage supervisé est le processus dans lequel l'apprenant reçoit des exemples d'apprentissage comprenant à la fois des données d'entrée et de sortie.

4.2.2 Apprentissage non supervisé

Si les identités des entités terrestres qui doivent être classées dans une scène ne sont pas généralement connues a priori en raison de l'absence des données de vérité au sol ou d'autres données justificatives, un apprentissage non supervisé est effectué. Une méthode ou algorithme automatique regroupe les pixels en différents clusters en fonction de certains critères statistiques. L'analyste, basé sur son expérience et sa connaissance de la scène pour nommer ces clusters. S'il n'y a que quelques pixels dans certains groupes, ces groupes

peuvent être supprimés ou fusionnés avec d'autres. De même, si certains groupes sont trop hétérogènes, ils peuvent être encore divisés. En bref, l'apprentissage non supervisé est le processus dans lequel l'apprenant reçoit des exemples d'apprentissage ne comprenant que des données d'entrée.

4.3 Les méthodes selon les types de modèles obtenus

4.3.1 Les méthodes d'explication et de prédiction

Ces méthodes fournissent un modèle explicatif et/ou prédictif à partir des données, ils utilisent ces données avec des résultats connus pour développer des modèles permettant de prédire les valeurs d'autres données. Ils cherchent à construire une relation entre les attributs à prédire et les attributs prédictifs.

4.3.2 Les méthodes de description et de visualisation

Les méthodes de description fournissent à l'analyste une vision synthétique et proposent des descriptions des données pour aider à la prise de décision. Les modèles descriptifs aident à la construction de modèles prédictifs.

5 Les principales méthodes du Datamining

5.1 Méthode des C-moyennes (C-Means)

C-Means [84] est l'une des méthodes d'apprentissage non supervisée la plus simple et courante du Datamining permettant de résoudre le problème bien connu du clustering. Il est également utile dans le domaine de l'analyse par télédétection [118], où des objets ayant des valeurs de spectre similaires sont regroupés ensemble sans aucune connaissance préalable. L'algorithme de C-moyennes comprend deux phases distinctes. Dans la première phase, il calcule les centres de clusters (centroïdes) où chaque centre est la valeur spectrale moyenne de tous les pixels dans le cluster, et dans la seconde phase, il attribue chaque pixel au groupe ayant le centre de gravité le plus proche. Il existe différentes méthodes pour définir la distance du centroïde le plus proche, l'une la plus utilisée est celle de la distance Euclidienne. Une fois le regroupement effectué, il recalcule le nouveau centroïde de chaque groupe. Sur la base de ce centre, une nouvelle distance Euclidienne est calculée entre chaque centre et chaque pixel, et assigne les pixels aux clusters qui ont une distance Euclidienne minimale. Chaque cluster de la partition est défini par ses pixels membres et par son centroïde. Le centre de gravité de chaque groupe est le point auquel la somme des distances de tous les

pixels de ce groupe est réduite. Donc, l'algorithme de clustering C-means est un algorithme itératif aide à partager un ensemble de pixels inconnu donné en un nombre de cluster (nc) fixe définis par l'utilisateur, chaque pixel sera attribué à un et un seul cluster. L'objectif de l'algorithme C-means est de minimiser la variabilité au sein du cluster en minimisant une fonction objective "J" définie par l'équation 2.1.

$$J(C,X) = \sum_{i=1}^{n} \sum_{j=1}^{nc} \|\mathbf{x_i} - \mathbf{c_j}\| \,, \tag{2.1}$$

Où $\|\mathbf{x_i} - \mathbf{c_j}\|$ est la distance entre l'un des n échantillons de pixel x_i et son centre de centroïde le plus proche c_j, afin de minimiser la fonction objective J, les étapes suivantes sont suivies :

1. Étape 1 : choisir le nombre de groupes nc, initialiser leurs centres de manière aléatoire.
2. Étape 2 : pour tous les pixels de l'image, procéder comme suit :
   — Calculer la distance Euclidienne entre le centre c_j et chaque pixel x_i d'une image.
   — Attribuer chaque pixel x_i au centre c_j le plus proche en fonction de la distance.
3. Étape 3 : mettre à jour les centres de cluster en calculant la moyenne des pixels appartenant à ce cluster.
4. Étape 4 : entre deux mises à jour consécutives : Arrêter, si aucun pixel ne change de groupe ou bien si les modifications dans les centres de cluster sont inférieures à une valeur spécifiée. Sinon, passer à l'étape 2.

La figure 2.2 illustre les différentes étapes de l'algorithme C-Means.

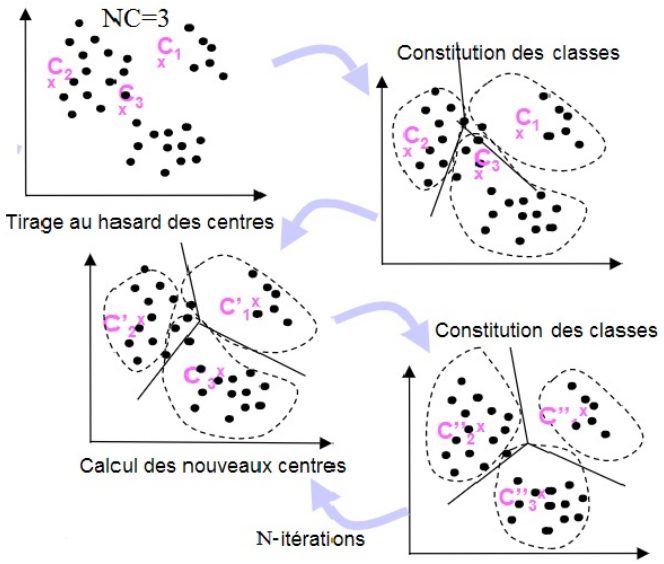


FIGURE 2.2 Les étapes de l'algorithme C-Means

5.2 Méthode des C-moyennes floues (Fuzzy C-Means (FCM))

Le FCM est une méthode développée par Dunn [39] et améliorée par Bezdek en 1981 [22]. Le FCM est une version floue de C-Means (Pour plus de détail sur la notion du flou, le lecteur peut voir l'annexe), il attribue des pixels aux clusters flous sans étiquettes, contrairement aux méthodes de clustering C-means, qui forcent les pixels à appartenir exclusivement à une seule classe, FCM permet aux pixels d'appartenir à deux ou plusieurs clusters avec différents degrés d'appartenance. Le FCM est basé sur la même idée de C-Means, c'est de trouver des centres de clusters en ajustant de façon itérative leurs positions et en évaluant une fonction objective comme C-Means, cependant, cela permet plus de flexibilité en introduisant la possibilité d'appartenances partielles aux clusters. La fonction objective de l'équation 2.1 est donc étendue à l'équation 2.2.

$$JF(C,X,U) = \sum_{i=1}^{n} \sum_{j=1}^{nc} u_{ij}^{m} \, \|\mathbf{x_i} - \mathbf{c_j}\| \, , \tag{2.2}$$

Où :

— $C = \{c_1, c_2, \ldots c_{nc}\}$: l'ensemble des centres des groupes.

— $X = \{x_1, x_2, \ldots x_n\}$: l'ensemble des pixels.

— $U = u_{ij}, (i = 1,2,\ldots,n), (j = 1,2,\ldots,nc)$ est la matrice $(n \times nc)$ des valeurs de degré d'appartenance floue comprise entre 0 et 1, et u_{ij} est le degré d'appartenance floue du $i^{\text{ème}}$ pixel x_i au centre du $j^{\text{ème}}$ cluster c_j, où la somme des degrés d'appartenance d'un $i^{\text{ème}}$ pixel à tous les clusters possibles étant égale à 1 ($\forall i, \sum_{j=1}^{nc} u_{ij} = 1$).

— m est un nombre réel supérieur à 1, il représente le degré de flou de la partition (généralement $m = 2$), quand ce degré égale 1, la partition tend vers une partition nette (i.e. non floue).

— $\|\mathbf{x_i} - \mathbf{c_j}\|$ est la distance entre le $i^{\text{ème}}$ pixel et le $j^{\text{ème}}$ cluster.

La fonction d'appartenance $JF(C,X,U)$ mise à jour en calculant les degrés d'appartenances u_{ij} et les centres de clusters c_j par les deux équation 2.3, 2.4 respectivement.

$$u_{ij} = \frac{1}{\sum_{k=1}^{nc} \left(\frac{\|\mathbf{x_i} - \mathbf{c_j}\|}{\|\mathbf{x_i} - \mathbf{c_k}\|} \right)^{\frac{2}{(m-1)}}} \tag{2.3}$$

$$c_j = \frac{\sum_{i=1}^{n} (u_{ij})^m x_i}{\sum_{i=1}^{n} (u_{ij})^m} \tag{2.4}$$

Les étapes de la segmentation d'image en c-moyennes floues sont les suivantes :

1. Étape 1 : initialiser les centres du cluster c_j (itération t = 1).
2. Étape 2 : calculer les degrés d'appartenance aux partitions floues selon l'équation 2.3.
3. Étape 3 : calculer la valeur de la fonction objective JF_t par l'équation 2.2.
   Si $t > 1$ et JF_{t+1} - $JF_t < \xi$ alors fin de l'algorithme, où ξ est un critère d'arrêt par exemple $\xi = 0.001$ dans [48, 25], sinon aller à l'étape 4.
4. Étape 4 : $(t = t + 1)$ calculer de nouveaux centres du cluster c_j en utilisant l'équation 2.4 et aller à l'étape 2.

En raison de la flexibilité de l'algorithme de classification FCM, plusieurs méthodes de segmentation d'images satellitaires basées sur le FCM ont été proposées dans la littérature [14, 31, 127].

5.3 Méthode des arbres de décision

La méthode des arbres de décision est une méthode d'apprentissage supervisé qui permet de répartir un ensemble de pixels en groupes homogènes selon des attributs spectraux discriminants en fonction d'un objectif fixé et connu. Un arbre de décision est composé de nœuds où les attributs sont testés avec des seuils déduits de façon empirique. Pour chaque nœud interne, le test utilise uniquement l'un des attributs spectraux. Les branches sortantes d'un nœud correspondent à tous les résultats possibles du test sur le nœud. Les feuilles d'un arbre de décision sont ses prédictions concernant les données spectrales à classifier. Donc, un arbre de décision est un arbre dont les nœuds internes sont des tests sur les attributs spectraux et dont les nœuds terminaux (feuilles) sont des catégories (classes) [82]. L'arbre de décision peut non seulement être exprimé par arbre, mais également par un ensemble de règles $SI - SINON$, où on fait une lecture hiérarchique de cet arbre de haut en bas, chaque route de la racine à la feuille, correspond à une règle où la condition est présentée par tous les tests des attributs des nœuds. Le résultat de la règle est présenté par la feuille dans la route. Les arbres de décision ont une forme de représentation simple, ce qui permet à l'utilisateur de comprendre facilement le modèle inféré. Différents algorithmes pour classifier les images de télédétection sont décrits dans la littérature [46, 50]. Un simple arbre de décision pour la classification des échantillons avec deux attributs d'entrée X_{IR} et X_{VIS} est donné dans la figure 2.3. Tous les pixels avec des valeurs radiométriques du pixel d'une image infrarouge X_{IR} <= 130 appartiennent à la classe « Nuage » quelle que soit la valeur du pixel auquel il correspond dans une image visible X_{VIS}, sinon, les pixels avec des valeurs X_{VIS} =<20 appartiennent à la classe « Mer », sinon, le reste des pixels sont classifie comme « Sol ».

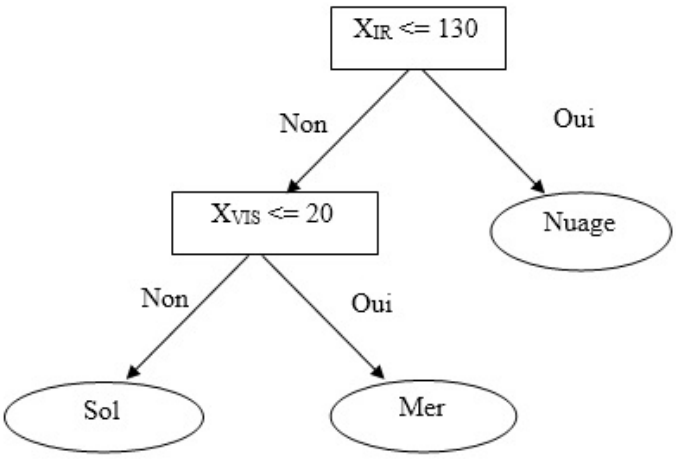


FIGURE 2.3 Exemple d'un arbre de décision

5.4 Méthode k plus proche voisins

K plus proche voisins (en anglais k Nearest Neighbors (kNN, k-NN)) est une technique de classification supervisée [83], le KNN se base uniquement sur les données d'apprentissage existantes déjà. Son principe est simple, pour prédire la classe d'un nouvel échantillon pixel (pixel à classer), il est comparé à tous les pixels d'entraînement. kNN affectera ce nouveau pixel à la classe majoritaire parmi ses k pixels plus proches voisins en utilisant une métrique de distance. Plusieurs fonctions permettent de calculer la distance entre l'échantillon observé et les échantillons d'apprentissage (Euclidien, Minkowski, Manhattan,... etc), en général, la distance Euclidienne est fréquemment utilisée. Les performances de la méthode kNN dépendent du choix du paramètre k. c'est un paramètre de la méthodologie, il est souvent choisi en fonction de l'expérience ou de la connaissance du problème de classification en question. Le processus de classification kNN est généralement basé sur les étapes suivantes :

— Déterminer le paramètre k nombre de pixels voisins les plus proches.

— Calculer la distance entre chaque pixel échantillon d'essai et tous les pixels échantillons d'entraînement.

— Trier la distance et déterminer les pixels voisins les plus proches en fonction du k-ème seuil.

— Déterminer la catégorie (classe) pour chacun des pixels voisins les plus proches.

— Utiliser la majorité simple de la catégorie des pixels voisins les plus proches comme valeur prédictive de la classification de l'échantillon pixel à tester.

Pour l'exemple de la figure 2.4, nous avons trois classes (C1, C2, C3) représentées par un ensemble d'échantillons d'apprentissage, l'objectif est de trouver une étiquette de classe pour l'échantillon testé x. Dans ce cas, nous utilisons la distance Euclidienne et une valeur de k = 5 voisins comme seuil. Sur les cinq plus proches voisins, 4 appartiennent à la classe C1 et 1 à la classe C3, x est donc attribué à C1 en tant que classe prédominante dans le voisinage.

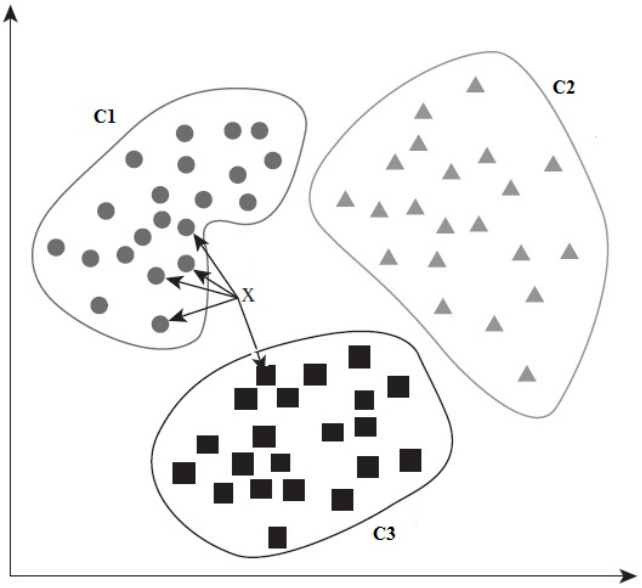


FIGURE 2.4 Exemple d'un k plus proche voisins

Différentes approches de classification basée sur la méthode k-NN utilisant des images satellitaires ont été développées dans la littérature [81, 96, 107].

5.5 Méthode des réseaux de neurones

Les réseaux de neurones artificiels [62] sont inspirés du fonctionnement des neurones biologiques, semblable au cerveau humain avec en entrée des images et en sortie la classification, un réseau de neurone est composé de neurones artificiels (ou unités de traitement) et d'interconnexions. Lorsque nous considérons un tel réseau sous forme de graphique (voir figure 2.5), les neurones peuvent être représentés sous forme de nœuds (ou de sommets) et les interconnexions en tant qu'arêtes (arcs). Un réseau de neurones est généralement composé d'une succession de couches dont chacune prend ses entrées sur les sorties de la précédente. Chaque couche est composée de neurones et les neurones de chaque couche sont reliés entre eux par des poids synaptiques, durant la phase d'apprentissage, le réseau de neurones permet de modifier les poids à attribuer à chaque liaison entre les neurones afin que les sorties du réseau de neurones soit aussi proche que possible de la cible souhaitée (classe souhaitée).

Les neurones qui ne sont ni en entrée ni en sortie sont appelés des neurones cachés. Chaque neurone dans la couche cachée calcule la somme pondérée de ses entrées et retourne une valeur en fonction de sa fonction d'activation (Par exemple la fonction d'activation sigmoïde). Cette valeur peut être utilisée soit comme une des entrées d'une nouvelle couche de neurones, soit comme un résultat qu'il appartient à l'utilisateur d'interpréter (classe, résultat d'un calcul, … etc). La figure 2.5 [126] montre un exemple de réseau de neurones avec p neurones au niveau de la couche d'entrée $(X_1, X_2, \ldots X_p)$ et k neurones au niveau de la couche cachée et un neurone au niveau de la couche de sortie (Y). Les W_{ij} sont respectivement les poids reliant la couche d'entrée à la couche cachée et cette dernière à la couche de sortie.

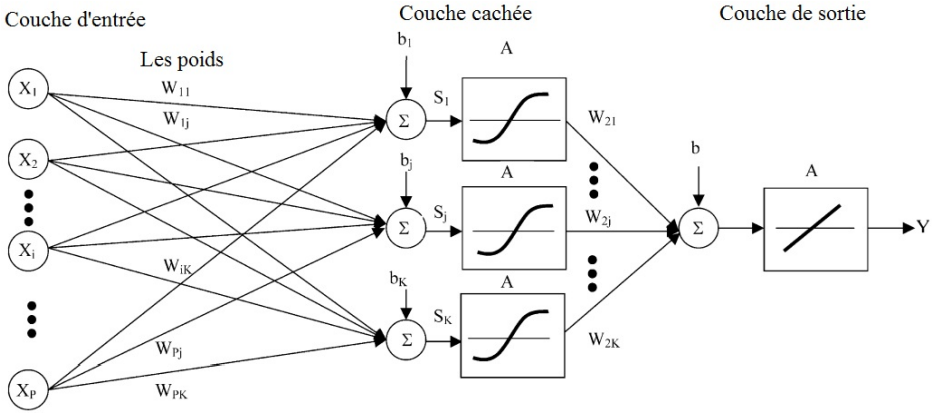


FIGURE 2.5 Architecture générale d'un réseau de neurones

Deux grands types de réseaux sont définis [116] (voir figure 2.6) :

— **Réseaux à propagation directe (les réseaux perceptron multicouches) :** sont l'une des classes des réseaux de neurones les plus importantes et les plus populaires dans les applications réelles. En règle générale, le réseau consiste en un ensemble d'entrées constituant la couche d'entrée , une ou plusieurs couches cachées de nœuds de calcul et, enfin, une couche de sortie de nœuds de calcul. Les couches sont connectées sous forme de graphe dirigé dans un seul sens (acyclique) entre les couches d'entrée et de sortie.

— **Réseaux récurrent :** dans ce type de réseaux les connexions entre les neurones (cachés ou visibles) forment un graphe dirigé d'une façon cyclique. Comme par exemple le réseau de Hopfield.

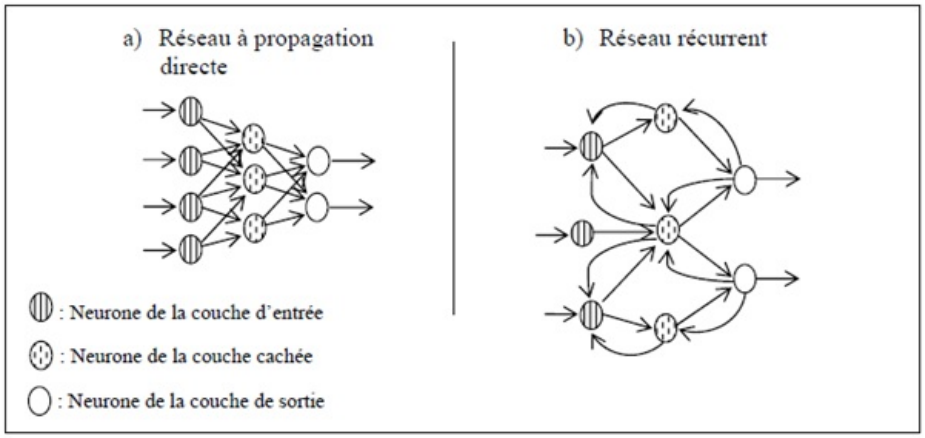


FIGURE 2.6 Les deux types principaux d'un réseau de neurones

5.6 Méthode des règles d'association

La méthode des règles d'association est l'une des méthodes importantes dans le domaine de la feuille de données [37]. Elle consiste à découvrir des associations et des corrélations intéressantes entre un vaste ensemble de données spectrales. Elle est largement utilisée dans l'analyse du panier du marché [4]. Nous ne détaillons pas davantage la méthode des règles d'association dans ce paragraphe car elle sera l'objet du chapitre suivant.

6 Conclusion

Dans cette partie, nous avons présenté le concept d'extraction de connaissances à partir des bases de données notamment celles des images satellitaires. Ensuite, nous avons décrit les principales méthodes et techniques du datamining appliquée sur ces images, où elles peuvent tout à fait suggérer des associations et des corrélations entre les données images pour extraire des informations cachées afin d'en tirer des connaissances. Parmi ces méthodes, nous avons opté pour la technique des règles d'association qui sera développée en détail dans le prochain chapitre, ensuite nous présentons une extension par des règles d'association floues, puisque, nous avons réalisé et développée notre méthode sur la base de cette technique.

Chapitre 3

Extraction de connaissances par les règles d'association classiques et floues

1 Les règles d'association classiques

1.1 Introduction

Parmi les différentes méthodes d'extraction automatique des connaissances à partir de données, nous nous concentrons sur la méthode des règles d'association, elle constitue un moyen pratique et efficace de découvrir et de représenter certaines dépendances et relations significatives entre attributs dans une base de données [34]. Elle a été introduite pour la première fois par Agrawal et al. [4] pour l'analyse des données des paniers de marché, ces «paniers» sont transformés en base de données composée d'un grand nombre de relevés de transaction, chaque enregistrement comprend tous les articles achetés par un client sur une seule transaction d'achat. La méthode consiste à rechercher quels groupes de produits qui sont fréquemment achetés ensemble, elle fournit des informations sous forme de règles, comme par exemple : "Si un client achète du lait, il achète du gâteau (70%) ". Cette règle signifie que 70% des clients qui achètent du lait ont également tendance à acheter du gâteau, cette découverte est très intéressante et peut être utilisée dans le but de prédire le comportement des clients (prédiction), comme elle peut aider un gestionnaire du supermarché à ranger ses étagères afin que le lait et le gâteau se trouvent à proximité (classification) parce que souvent sont achetés ensemble. Il peut aussi décider quelles sont les articles à mettre en promotion (les articles rarement achetés).

1.2 Exemple de paniers de marché

Afin de mieux comprendre la méthode, nous illustrons leur principe dans l'exemple suivant. Considérons les transactions (paniers de marché) obtenues d'un supermarché. Chaque transaction représente les articles achetés par un client.

— Ticket 1 : gâteau, lait, banane

— Ticket 2 : gâteau, café, lait, sucre

— Ticket 3 : café, lait, sucre

— Ticket 4 : café, lait, sucre, pomme

— Ticket 5 : gâteau, lait, banane, sucre

— Ticket 6 : gâteau, café, banane, sucre

Ces achats des clients peuvent être transformés en une base de transactions binaires, les lignes représentent les transactions et les colonnes représentent les différents produits existants dans le supermarché, dans l'intersection d'une ligne et d'une colonne on met la valeur "1" si un produit est acheté par un client, sinon on met la valeur "0" dans l'autre cas [33], chaque produit/item est représenté par un identifiant. Le tableau 3.1 représente la base de transactions binaire construite à partir des achats des clients mentionnés ci-dessus.

TABLE 3.1 Exemple d'une base de transactions binaire

Transactions Items	Item i_1 (Gâteau)	Item i_2 (Café)	Item i_3 (Lait)	Item i_4 (Banane)	Item i_5 (Sucre)	Item i_6 (Pomme)
t_1	1	0	1	1	0	0
t_2	1	1	1	0	1	0
t_3	0	1	1	0	1	0
t_4	0	1	1	0	1	1
t_5	1	0	1	0	1	1
t_6	1	1	0	0	1	1

À partir du tableau 3.1, on peut construire le tableau 3.2 qui indique la fréquence que deux articles ont été achetés simultanément. Le tableau 3.2 exprime une corrélation de cooccurrence entre les produits achetés, il nous permet de déterminer combien de fois deux produits ont étés achetés ensemble. Par exemple on peut observer que le produit i_1 (Gâteau) et i_3 (Lait) apparaissent simultanément dans 50% (3/6) des achats, alors que les deux produits i_4 et i_5 n'apparaissent jamais ensemble. A partir de la première observation on peut proposer les deux règles suivantes :

TABLE 3.2 Tableau de cooccurrence des produits

	Item i_1	Item i_2	Item i_3	Item i_4	Item i_5	Item i_6
Item i_1	4	2	3	1	3	2
Item i_2	2	4	3	0	4	2
Item i_3	3	3	5	1	4	2
Item i_4	1	0	1	1	0	0
Item i_5	3	4	4	0	5	3
Item i_6	2	2	2	0	3	3

1. Si un client achète le produit i_1 alors il achète aussi le produit i_3 ($i_1 \longmapsto i_3$).
2. Si un client achète le produit i_3 alors il achète aussi le produit i_1 ($i_3 \longmapsto i_1$).

La probabilité d'acheter les deux produit i_1 et i_3 ensemble est égale à 50%, cette mesure est appelée le support d'une règle, les deux règles précédentes ont un support de 50%. Le produit i_1 apparaît dans 4 achats, dans ces 4 achats le produit i_3 apparaît 3 fois. On peut remarquer aussi que le produit i_3 est acheté 5 fois, parmi ces achats le produit i_1 apparaît 3 fois, cette mesure est appelée la confiance d'une règle. La confiance de la première règle ($i_1 \longmapsto i_3$) est égale à 75% (3/4) alors que la deuxième règle ($i_3 \longmapsto i_1$) a une confiance de 60% (3/5), donc la règle 1 est mieux que la $2^{\text{ème}}$ règle.

1.3 Applications des règles d'association

Depuis la découverte de la méthode des règles d'association classiques ou binaires par Agrawal et al. [4], elle a montré son efficacité comme un outil d'aide à la décision à partir des bases de données. Elle a été largement utilisée dans nombreux domaines notamment le secteur commercial, selon [26, 103] la méthode permet aux établissements de vente de faire progresser ses ventes en approchant l'emplacement des produits vendus fréquemment ensemble, mettre en promotion les produits non fréquents, identifier les préférences de différents groupes de clients, produits et services adaptés à leur goût, anticiper leurs besoins futures, envisager les produits à leur offrir. . . etc.

Les règles d'association sont appliquées avec succès dans le domaine de la santé, à chaque consultation d'un patient, la majorité des établissements sauvegardent les informations relatives à leurs patients dans des bases de données. La découverte d'association entre ces données peut aider les médecins au diagnostic de quelques maladies et définition de traitement afin de guérir leurs patients. Par exemple, Tyler et al. [87] ont développé une méthode qui permet de prévoir les symptômes futurs possibles du patient en fonction de ses antécédents et de ses symptômes antérieurs et selon les symptômes de nombreux patients similaires (telles

que "symptôme 1 et symptôme 2 $\longmapsto$ symptôme 3 "). Les chercheures [115] ont proposé une méthode permettant de déterminer la probabilité d'une maladie particulière en identifiant les relations entre les symptômes. Les auteurs [92, 93] ont proposé des méthodes basées sur les règles d'association pour diagnostiquer la maladie cardiaque et détecter les facteurs qui contribuent à la maladie. L'application de la méthode a permis l'identification de populations à risque vis à vis de certaines maladies, par exemple les auteurs [105] ont extrait un schéma de risque du diabète de type 2 à partir d'une base de données d'études sur les lipides et le glucose de Téhéran.

Quelques phénomènes biologiques peuvent aussi se modéliser par une approche des règles d'association, par exemples les protéines sont des séquences composées de 20 types d'acides aminés. Chaque protéine a une structure tridimensionnelle unique, qui dépend de la séquence d'acides aminés, un petit changement de séquence peut modifier le fonctionnement de la protéine. Ils ont découvert que généralement les séquences d'acides aminés des protéines ne sont pas aléatoires. Les chercheurs [57] ont déchiffré la nature des associations entre différents acides aminés présents dans une protéine. Ces dernières années, quelques auteurs [78, 130] sont arrivés à utiliser les règles d'association avec succès pour découvrir les liaisons protéine-ADN. Les règles d'association ont été utilisées pour permettre la détection des facteurs génétiques et environnementaux d'une maladie [134].

La méthode est aussi appliquée dans le web (WebMining), les règles d'association extraites en fonction de l'historique des ventes effectuées en ligne par des internautes ont été utilisées pour l'aide à la conception et l'organisation des sites web, suggérer des combinaisons de produits à l'utilisateur Web [90]. Ces règles permettent de prédire le comportement des internautes pour accéder à une page web [49], déterminer les relations possibles entre les ensembles de pages consultées ensemble afin de permettre aux internautes de trouver rapidement les informations recherchées sur un site web.

Pour le réseau et la télécommunication, chaque jour un grand nombre d'alarmes est produit, [119, 61] ont réussi à appliquer les règles d'association afin de détecter et filtrer les alarmes non informatives en déterminant les différentes causes d'anomalies. Dans la télémaintenance la méthode est utilisée pour détecter et prédire les incidents potentiels des systèmes de production avant qu'ils ne surviennent [75].

Les bases de données spatiales sont utilisées dans les systèmes d'information géographiques, astronomique, cartographique et environnementaux. Par exemple, les règles d'association sont appliquées dans le Geominer pour prédire des phénomènes naturels et météorologiques [58]. D'autres domaines d'applications des règles d'association tels que l'analyse de données de recensements et statistiques [128], l'éducation [110] et le E-learning [91], ... etc.

1.4 Les avantages et les inconvénients des règles d'association

Parmi les avantages des règles d'association on peut citer :

1.4.1 Les avantages

— Les résultats des règles d'association sont faciles à interpréter et à comprendre.

— Leur niveau explicatif ou sémantique est très élevé par rapport aux autres techniques comme les réseaux de neurones, par exemple, qui sont souvent qualifiés de boites noires (niveau explicatif faible).

— La méthode permettant de découvrir des relations intéressantes entre les articles dans des bases de données volumineuses.

— Aucune hypothèse préalable est exigée, la méthode nécessite uniquement la liste des données des items.

— La méthode est facilement adaptable pour traiter des séries temporelles comme par exemple : si un client achète un computer alors il achètera une imprimante dans les 3 mois suivant.

— L'application des règles d'association dans nombreux domaines.

Malgré les avantages de la méthode, elle a aussi des inconvénients :

1.4.2 Les inconvénients

— Le nombre de règles d'association générées est très grand.

— Bien que le support minimum permet de diminuer les calculs, mais on peut éliminer des règles utiles.

— La recherche de règles d'association demande un temps considérable et un espace mémoire important.

— La quantité d'un produit acheté n'est pas prise en compte, la présence d'un produit dans une transaction exprime son achat, mais il n'indique pas sa quantité.

1.5 Concepts et définitions sur les règles d'association

1.5.1 Item

Un item désigne tout un article, objet ou un attribut appartenant à un ensemble 'I' fini d'éléments différents tel que $I = \{i_1, i_2, i_3, \ldots, i_d\}$, où d est le nombre d'items.

1.5.2 Itemset

Un itemset X noté $\{i_1 i_2 \dots i_m\}$ désigne un ensemble de m items où i_j est un item de I. Par exemple $\{i_2 i_6 i_9\}$ est un itemset composé de 3 items.

1.5.3 K-Itemset

Un k-itemset est une famille d'itemsets, où chaque itemset contient k items. Par exemple, l'itemset $\{i_2 i_6 i_9\}$ est un élément de la famille 3-Itemset.

1.5.4 Support d'un itemset

Soit $T = t_1, t_2, \dots, t_n$ une base de données de n transactions, où chaque transaction $t_i, (i = 1, 2, \dots, n)$ est représentée par un ensemble d'items de I (c'est-à-dire, $t_i \subseteq I$), le support d'un itemset X noté $(Supp(X))$ est le ratio de toutes les transactions contenant l'itemset X divisé par le total de toutes les transactions.

$$supp_{(X)} = \frac{|t_i \in T / X \subseteq t_i|}{|T|} \tag{3.1}$$

où $||$ désigne la cardinalité d'un ensemble.

1.5.5 Règle d'association

Une règle d'association est une relation d'implication sous la forme de : $X \longmapsto Y$, elle exprime le fait de si X alors Y, où $X, Y \subseteq I$, X et Y sont la condition et la conclusion de cette règle respectivement. X et Y n'ont pas un item commun (i.e. $X \cap Y = \emptyset$).

1.5.6 Confiance d'une règle d'association

La confiance de la règle d'association $X \Rightarrow Y$ est le ratio entre le nombre de transactions qui contiennent simultanément les deux itemsets X et Y divisé par le nombre de transactions contenant que l'itemset X.

$$conf_{(X \Rightarrow Y)} = \frac{|t_i \in T / (X \subseteq t) \cap (Y \subseteq t_i)|}{|t_i \in T / X \subseteq t_i|} \tag{3.2}$$

1.5.7 Itemset fréquent

Un itemset est fréquent si et seulement si son support est supérieur ou égal à un seuil minimal fixé préalablement par l'utilisateur [4]. Pour une base de données comportant d items,

le nombre d'itemsets possible en faisons les combinaisons entre les items sont présentés dans le tableau 3.3.

TABLE 3.3 La complexité des calculs pour les règles d'association

	d items	k=2	k=3	k=4
K-Itemsets possibles (C_d^k)	100	4 950	161 700	3 921 225
	10 000	5×10^7	1.7×10^{11}	4.2×10^{14}

On peut dire que plus le nombre d'items est grand, plus le nombre d'itemsets générés est très grand et donc un nombre de règles importants sera produits, l'extraction de ces règles est couteuse en temps de calcul et même pour l'espace de recherche et surtout dans des bases de données qui contiennent des milliers d'items. Le support minimum nous permet de minimiser les calculs et bien évidements le nombre de règles d'association, on garde uniquement les itemsets fréquents qui ont un support supérieur au support minimum. La majorité des algorithmes d'extraction de règles d'association utilisent ce seuil afin de réduire l'espace de recherche et éliminer autant que possible les itemsets non fréquents [64].

1.6 Les étapes de l'extraction de règles d'association

Le processus de découverte des règles d'association se fait en quatre phases principales [8] :

— **Prétraitement des données :** le prétraitement (ou la préparation) des données est nécessaire pour faciliter l'exploitation des règles d'association. Il consiste à choisir uniquement les données de la base qui sont intéressant pour l'utilisateur, cela permet de réduire le volume de données traitées. Après on converti les données en format base de données binaire. Cette étape garantit l'efficacité de l'extraction de règles d'association.

— **Extraction des itemsets fréquents :** dans cette étape on cherche tous les itemsets fréquents, c'est l'étape la plus compliquée nécessitant d'un temps d'exécution considérable et surtout un espace mémoire important, car le nombre d'itemsets fréquents est de complexité exponentielle avec le nombre d de l'ensemble d'items manipulés comme il est indiqué dans le tableau 3.3.

— **Génération des règles d'association :** trouver toutes les règles d'association à partir de l'ensemble des itemsets fréquents si son support et sa confiance satisfont un seuil minimal.

— **Visualisation et interprétation des règles d'association :** les connaissances découvertes (règles d'association) sont à la disposition de l'utilisateur, il peut les interpréter simplement pour une meilleure prise de décision.

La figure 3.1 résume les différentes étapes de l'extraction de règles d'association.

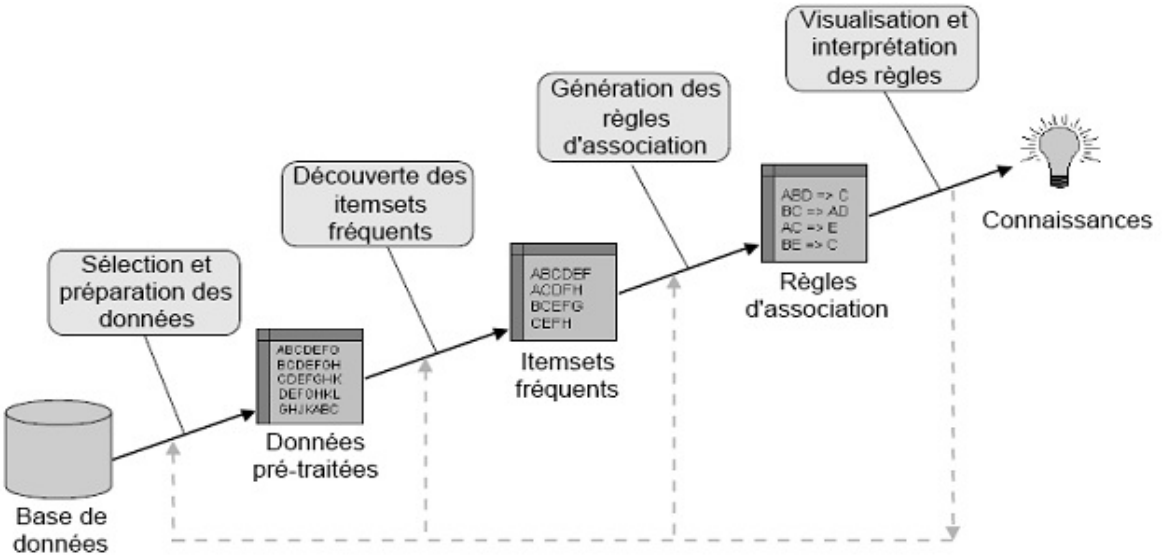


FIGURE 3.1 Les étapes d'extraction des règles d'association [8]

1.7 Extraction de règles d'association

Plusieurs algorithmes qui traite le problème de recherche de règles d'association ont été proposés dans la littérature tel que l'algorithme Apriori [5], l'algorithme Partition [111], l'algorithme Eclat [97], l'algorithme Sampling [123], l'algorithme Close [101], l'algorithme FP-growth [59], mais nous nous sommes focalisé, essentiellement, sur l'algorithme Apriori parce que c'est la base des autres algorithmes, et de même pour notre approche développée est fondée sur cet algorithme, il est considéré comme une référence dans ce domaine.

L'algorithme Apriori fournit un ensemble de règles basées sur deux mesures statistiques, le support et la confiance de la règle. Il a apporté une solution élégante au problème de l'extraction de règles, il produit un nombre important de règles, sélectionnant des règles fréquentes (intéressantes) qui ont un support et une confiance supérieurs respectivement aux deux seuils : support minimum (minsup) et confiance minimale (minconf) défini par l'utilisateur et ignorant des règles inutiles. C'est un algorithme par niveau, il calcule les itemsets fréquents en partant des itemsets les plus généraux vers les plus spécifiques. Ainsi, au niveau 1, il calcule les 1-itemsets fréquents (i.e de taille 1), passe au niveau 2 et continue ainsi jusqu'à trouver l'ensemble de tous les itemsets fréquents. L'idée générale de l'algorithme est basée sur la propriété d'anti-monotonie du support, si un itemset est fréquent alors tous ses sous-ensembles doivent également être fréquents, et inversement, tous les super-ensembles d'un itemset non fréquent sont obligatoirement non fréquents, cette propriété

permet de minimiser le temps d'exécution et l'espace de recherche. La découverte de règles d'association par l'algorithme Apriori se fait en deux étapes principales :

1.7.1 Recherche de l'ensemble des itemsets fréquents par l'algorithme Apriori

Au point de départ de l'algorithme Apriori, il faut fixer un seuil de support minimal (minsup) pour que seules les itemsets avec un support plus grand ou égal à ce seuil soient générés. La phase de génération se déroule en deux étapes, dans la première étape (étape de jointure), les itemsets de taille k sont générés en faisant une jointure des itemsets de taille k-1. Ensuite, dans la deuxième étape (étape d'élagage), tout itemset X généré par la phase de jointure est supprimé s'il existe un sous-ensemble de X qui est non fréquent (propriété d'anti-monotonicité). Par exemple si nous avons la liste L_3 des 3-itemsets : $L_3 = \{\{1\ 2\ 3\}, \{1\ 2\ 4\}, \{1\ 3\ 4\}, \{1\ 3\ 5\}, \{2\ 3\ 4\}\}$, l'étape de jointure donne comme résultat $C_4 = \{(1\ 2\ 3\ 4), (1\ 3\ 4\ 5)\}$, ensuite l'étape d'élagage va supprimer l'ensemble $\{1\ 3\ 4\ 5\}$ car le sous-ensemble $\{1\ 4\ 5\}$ n'est pas dans L_3.

Le pseudocode de l'algorithme Apriori est présenté dans l'algorithme 1.

Algorithm 1 algorithme Apriori

Données : T : base de transactions, minsup : support minimum
Sorties : LIF : liste des itemsets fréquents

1: $L_1 =$ liste des 1-itemsets fréquents
2: $k = 2$
3: **while** $L_{k-1} \neq \emptyset$ **do**
4: $C_k =$ Genere_Candidats (L_{k-1}) ▷ génération des itemsets candidats
5: **for** transaction $t \in T$ **do**
6: $C_t = Subset(C_k, t)$ ▷ les candidats contenus dans Candidats[k]
7: **for** candidat c dans C_t **do**
8: c.compteur++ ;
9: **end for**
10: **end for**
11: $L_k = \{$c $\in C_k$ / c.compteur $\geq$ minsup $\}$
12: k++
13: **end while**
14: $LIF = \bigcup_k L_k$ ▷ tous les itemsets fréquents trouvés

A chaque itemset, on associe un compteur noté c.compteur, pour stocker son support. Après, les 1-itemsets sont déterminés lors de la première passe sur la base de transaction T, en comptant les nombres d'apparitions de chaque item afin de déterminer ceux qui sont fréquents (L_1). Ensuite, la k^{ime} passe (k>=2) est composée de trois phases :

1. Les (k-1)-itemsets fréquents (L_{k-1}) trouvés lors de la (k-1)^{ime} passe sont utilisés pour générer les itemsets candidats de taille k, ce traitement est effectué dans la fonction (Genere_Candidats (L_{k-1})).
2. A chaque fois, la base de données T est parcourue afin de calculer la fréquence des candidats est comptée, en utilisant la fonction Subset qui fournit l'ensemble des itemsets C_t contenus dans une transaction.
3. On sélectionne les itemsets qui ont une fréquence supérieure ou égale au seuil minsup.

L'algorithme Apriori s'arrête quand il n'y a aucun nouveau candidat de taille supérieure à générer. Le pseudocode de la procédure (Genere_Candidats (L_{k-1})) de l'algorithme Apriori est le suivant.

Algorithm 2 algorithme de la procédure (Genere_Candidats (L_{k-1}))

```
1: for itemset X ∈ L_{k-1} do
2:     for itemset Y ∈ L_{k-1} do
3:         if (X[1] = Y[1]) ∧ (X[2] = Y[2]) ∧ ... ∧ (X[k-2] = Y[k-2]) ∧ (X[k-1]
           < Y[k-1]) then                ▷ X<Y, X et Y partagent leur k-2 premiers items
4:             c = X[1],..., X[k-1], Y[k-1]    ▷ étape de jointure : générer des candidats
5:             Insérer c dans C_k
6:         end if
7:         for all (k-1)-sous-ensembles s de c do                ▷ étape d'élagage
8:             if s ∉ L_{K-1} then
9:                 Supprimer c de C_k
10:            end if
11:        end for
12:    end for
13: end for
14: Retourner C_k
```

Exemple : l'application de l'algorithme Apriori sur la base de transactions de notre exemple (1.2) pour extraire la liste des itemsets fréquents se déroule comme indiqué sur la Figure 3.2. Les données de notre exemple est l'ensemble T de transactions ou d'achats : T = {{1, 3, 4}, {1, 2, 3, 5}, {2, 3, 5}, {2, 3, 5, 6}, {1, 3, 4, 5}, {1, 2, 4, 5}}, l'ensemble I d'items ou d'articles achetés : $I = \{1,2,3,4,5,6\}$, soit un seuil minimal de support minsup = 0.4.

On élimine les itemsets qui ont un support < minsup (Trame de fond en gris), comme par exemple l'itemset (2 4). On élague aussi les itemsets qui ont au moin un de ses sous-ensembles qui n'est pas fréquent (ecriture en Gras) comme par exemple l'itemset (1 3 4) parce que l'itemset (3 4) n'existe pas dans L_2. Donc la liste finale des itemsets fréquents *LIF* = $L_1 \cup L_2 \cup L_3$ = {(1), (2), (3), (4), (5), (1 3), (1 4), (1 5), (2 3), (2 5), (3 5), (2 3 5)}.

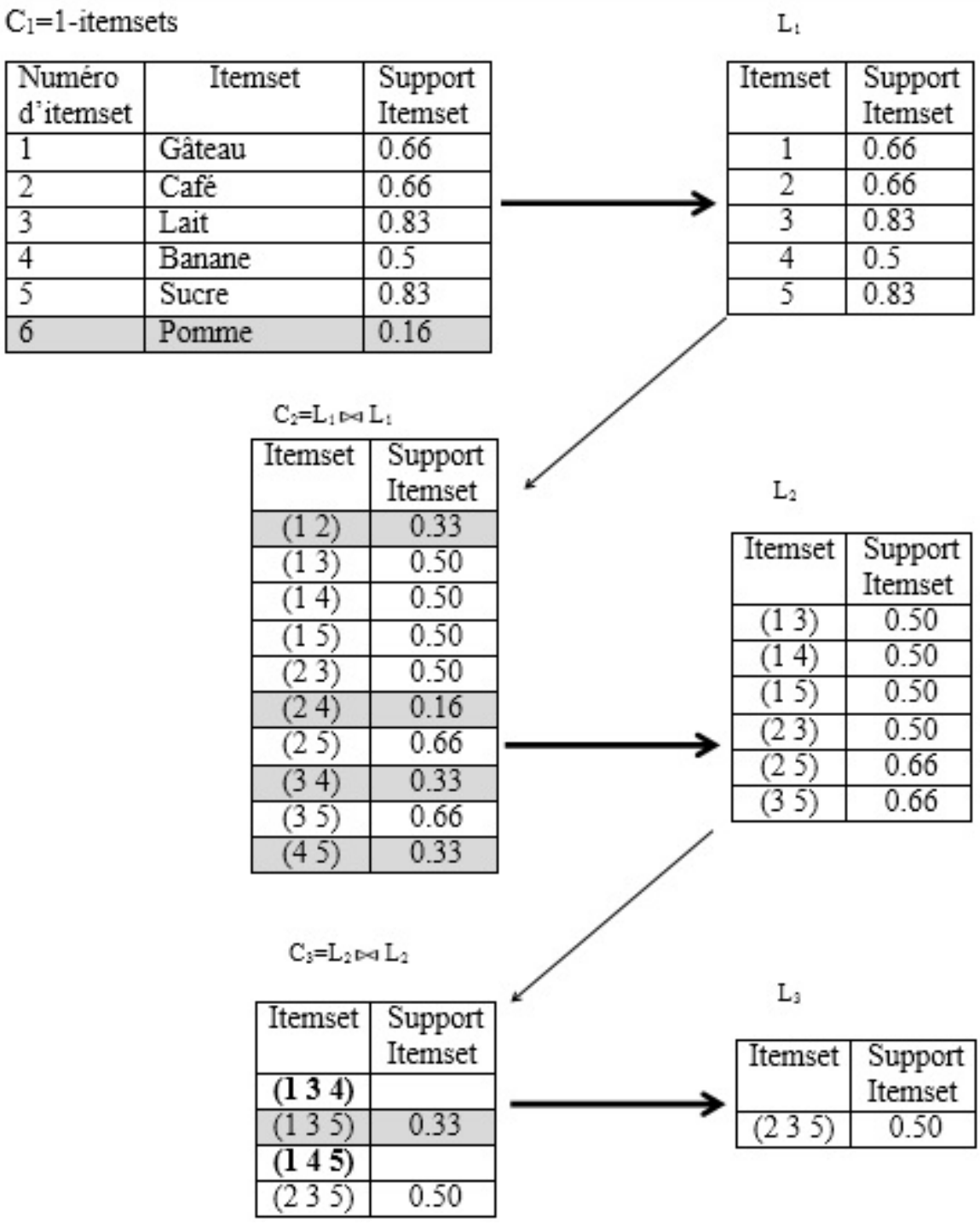


FIGURE 3.2 Exmeple de recherche des itemsets fréquents

1.7.2 Génération des règles d'association

A partir de la liste des itemsets fréquents LIF trouvés dans l'étape précédente, une liste de règles d'association LR sera générée, une règle est retenue si et seulement si sa confiance $\geq$ minconf définie par l'utilisateur, Agrawal et al. [5] ont proposé l'algorithme 3 qui permet de générer ces règles, son pseudocode est le suivant :

Algorithm 3 Génération des règles d'association

Données : LIF : liste d'itemsets fréquents, $minconf$: seuil minimal de confiance
Sorties : LR : liste de règles d'association générée

1: **for all** k-itemset fréquent $L_k \in LIF$ tel que $k \geq 2$ **do**
2: Générés tous les sous-itemsets S_m de L_k
3: **for all** S_m de L_k **do**
4: $Conf(R) = Supp(L_k)/Supp(L_k - S_m)$
5: **if** $Conf(R) \geq minconf$ **then**
6: $LR \leftarrow LR \cup R : (L_k - S_m) \longrightarrow S_m$
7: **end if**
8: **end for**
9: **end for**
10: Retourner LR

Pour chaque itemset fréquent L_k de taille supérieure à un, on génère un ensemble S_m ($m < k$) qui sont des sous-ensembles de L_k. Pour chaque S_m on calcule la confiance de la règle $R : (L_k - S_m) \longrightarrow S_m$, si sa confiance est supérieure ou égale à minconf (seuil minimal de confiance) on ajoute la règle R à LR (liste des règles d'association).
Pour l'exemple précédent l'ensemble des itemsets fréquents de taille $k \geq 2$ à partir de la liste LIF est : $\{(1\ 3), (1\ 4), (1\ 5), (2\ 3), (2\ 5), (3\ 5), (2\ 3\ 5)\}$. Pour un seuil minimal de confiance minconf= 0.70, les règles d'association extraites par l'algorithme 3 sont décrites par le tableau 3.4.

TABLE 3.4 Les règles extraites par l'algorithme Apriori

La règle	La confiance	La règle	La confiance
(1) $\longmapsto$ (3)	0.75	(3) $\longmapsto$ (5)	0.80
(1) $\longmapsto$ (4)	0.75	(5) $\longmapsto$ (3)	1
(4) $\longmapsto$ (1)	1	(2 3) $\longmapsto$ (5)	0.75
(1) $\longmapsto$ (5)	0.75	(2 5) $\longmapsto$ (3)	0.75
(5) $\longmapsto$ (1)	0.75	(3 5) $\longmapsto$ (2)	0.75
(2) $\longmapsto$ (3)	0.75	(2) $\longmapsto$ (3 5)	0.75

2 Les règles d'association floues

2.1 Introduction

L'algorithme Apriori d'Agrawal [5] et ses variantes ont été conçus pour des bases de données booléennes, (on met la valeur « 1 » dans la base des transactions si le produit est acheté sinon on met la valeur « 0 » dans l'autre cas), dans ce cas, la logique propositionnelle simple est suffisante pour exprimer les règles d'association. Mais nous n'avons pas pris en considération des critères comme par exemple la quantité des produits achetés, normalement un produit acheté pour une quantité donnée n'est pas considéré comme le même produit acheté avec une quantité différente, donc les algorithmes classiques d'extraction de règles d'association s'avèrent alors inadaptés pour traiter ce type d'informations. Pour cela, et pour pallier ce problème, les règles d'association floues sont apparues. C'est une extension des règles d'association basées sur la théorie des sous-ensembles flous proposée par Lotfi Zadeh [133], permettant de raisonner sur des attributs quantitatifs. Et ç'est ce que nous allons voir dans ce que suit

2.2 Concepts et définitions sur les règles d'association floues

En se basant sur la théorie des ensembles flous (dans l'annexe, nous avons donné des détails sur la logique floue et la théorie des sous-ensembles flous), les nouvelles définitions du support d'un itemset flou et la confiance d'une règle d'association floue sont les suivants.

2.2.1 Item flou

Un item flou est un couple [item, sous-ensemble flou], où on associer un item avec l'un des sous-ensembles flous de quantité associés. Cette item flou noté [x, a] correspondant à l'item x, associé au sous-ensemble flou a, défini sur l'univers du discours des quantités de x. Par exemple, [Température, froide] est un item flou où « froide » est un sous-ensemble flou défini par sa fonction d'appartenance.

2.2.2 Itemset flou

Un itemset flou noté par (X, A) est un ensemble d'item flou, où X =$\{x_1, x_2, \ldots, x_p\}$ est un ensemble de 'p' items et A = $\{a_1, a_2, \ldots, a_l\}$ est un ensemble de 'l' ensemble flous associés à ces items. Par exemple ([température, froide] [humidité, haute]) est un itemset flou avec deux items flous [température, froide] et [humidité, haute]. Nous disons qu'un itemset flous (X, A) n'est pas valide s'il contient au moins deux items flous du même attribut, par

exemple. ([Température, froide] [température, chaude]) n'est pas un itemset valide parce qu'il contient deux items flous du même attribut. Un K-itemset flou est l'ensemble d'itemsets flous contenant k items flous.

2.2.3 Le degré d'un itemset flou

Pour calculer le degré d'un itemset flou (X, A) dans une transaction trois approches ont été proposées :

— Delgado et al.[36], Dubois et al.[38], Hong et al.[65] ont choisi d'utiliser l'opérateur "min" comme une t-norm ($\top$), c'est le minimum entre les degrés d'appartenances des items flous qui compose l'itemset flou (X, A) dans une transaction.

— Fiot et al.[45] ont proposé d'utiliser l'opérateur "max" comme une t-conorme ($\perp$), c'est le maximum entre les degrés d'appartenances des items flous qui compose l'itemset flou (X, A) dans une transaction.

— Kuok et al.[76] ont préféré d'utiliser le produit ($\prod$) des degrés d'appartenances des items flous qui composent l'itemset flou (X, A) dans une transaction.

Soit le Tableau 3.5 qui contient les degrés d'appartenance de chaque item aux différents sous-ensembles pour chaque transaction t. cette table illustre un exemple de ces trois approches.

TABLE 3.5 Degré d'un itemset flou

Transaction	L'item x1		L'item x2			Degré de l'itemset flou ([x1, a1.3] [x2, a2.2])		
	a1.1	a1.2	a1.3	a2.1	a2.2	t-norm($\top$)	t-conorme($\perp$)	produit($\prod$)
t_1	0.1	0.1	0.8	0.1	0.9	0.8	0.9	0.72
t_2	0.1	0.4	0.5	0.5	0.5	0.5	0.5	0.25
t_3	0.1	0.7	0.2	0.7	0.3	0.2	0.3	0.06
t_4	0.6	0.4	0	0.75	0.25	0	0.25	0

2.2.4 Le support d'un itemset flou

pour calculer le support flou d'un itemset flou, quatre approches basées sur la cardinalité d'un itemset flou ont été proposées [45] :

— Comptage binaire : on compte toute transaction contenant l'itemset (X, A).

— Comptage seuillé : on comptabilise toutes les transactions pour lesquelles le degré d'un itemset (X, A) est supérieur à un seuil fixe (α-coupe).

— ∑comptage : sommer les degrés d'un itemset (X, A) de chaque transaction.

— ∑comptage seuillé : sommer les degrés d'un itemset (X, A) au-delà d'un seuil minimal fixe (α-coupe) de chaque transaction.

Pour l'exemple précédent, les différents calculs de cardinalité de l'itemset (X, A) = ([x1, a1.3] [x2, a2.2]) en utilisant un t-norm ($\top$) sont montrés dans le tableau 3.6.

TABLE 3.6 Différent type de calcul de la cardinalité flou

Transactions	Degré de l'itemset flou ([x1, a1.3] [x2, a2.2])			
t_1	0.8	0.8	0.8	0.8
t_2	0.5	0.5	0.5	0.5
t_3	0.2	0.2	0.2	0.2
t_4	0	0	0	0
Type de comptage	Binaire	Comptage seuillé (α-coupe = 0.4)	∑comptage	∑comptage seuillé (α-coupe = 0.4)
	3	2	1.5	1.3

A partir de ces comptages et selon les différentes méthodes de calcul du degré d'un itemset flou, on peut trouver plusieurs supports, par exemple, si on choisit l'opérateur t-norm ($\top$) avec un ∑comptage seuillé, le support d'un itemset flou (X, A) est calculé comme suit :

$$Supp_{(X,A)} \frac{\sum_{t_i \in T} \top(\alpha_A(t_i[X]))}{|T|} \tag{3.3}$$

Où $|T|$ est le nombre de transactions dans la base de données T, et α représente la fonction d'appartenance seuillée de l'équation suivante :

$$\alpha_A(t_i[X]) = \begin{cases} \mu_A(t_i[X]) \text{ si } \mu_A(t_i[X]) > \alpha - coupe \\ 0, \text{ sinon} \end{cases} \tag{3.4}$$

Où $\mu_A(t_i[X])$ est un vecteur de 'p' valeurs floues, p indique le nombre des items de l'itemset X, et chaque valeur floue représente le degré d'appartenance de chaque item $x_i(i = 1 \ldots p) \in X$ à ces ensembles flous associés $a_i(i = 1 \ldots l) \in$ A, et $\alpha - coupe$ est un seuil minimum d'appartenance définie par l'utilisateur.

2.2.5 Règle d'association floue

Une règle d'association floue est sous la forme de :Si "X est A", alors "Y est B". Où la partie "X est A" est la condition de la règle et la partie "Y est B" sa conclusion. On note la règle comme (X, A) $\longmapsto$ (Y, B), où X =$\{x_1, x_2, \ldots, x_p\}$ et Y =$\{y_1, y_2, \ldots, y_k\}$ sont

deux itemsets flous disjoints, tel que A = $\{a_1, a_2, \ldots, a_l\}$ et B = $\{b_1, b_2, \ldots, a_m\}$ sont des sous-ensembles flous associés aux itemsets flous X et Y.

2.2.6 La confiance d'une règle d'association floue

Si on choisit l'opérateur t-norm ($\top$) avec un $\sum$comptage seuillé, la confiance d'une règle d'association floue (X, A) $\longmapsto$ (Y, B) est calculée comme suit :

$$Conf_{(X,A)\longmapsto(Y,B)} \frac{\sum_{t_i \in T} \left[\alpha_A(t_i[X])\right] \top \left[\alpha_B(t_i[Y])\right]}{\sum_{t_i \in T} \alpha_A(t_i[X])} \tag{3.5}$$

Où $\alpha_A(t_i[X])$ et $\alpha_B(t_i[Y])$ ont la même définition décrit au-dessus.

2.3 Extraction de règles d'association floues

Tout d'abord, la base de données transactionnelle quantitative est transformée en une base de données transactionnelle de degrés d'appartenance, où chaque item /attribut est partitionné en plusieurs sous-ensembles flous afin de construire les items flous. Généralement, ces sous-ensembles flous et la forme des fonctions d'appartenance sont définis par des experts du domaine, mais, des fois cette expertise n'est pas toujours présente, soit à cause du manque des experts dans certain domaine ou bien à cause de la complexité du problème. Pour cela des algorithmes sont utilisés afin d'obtenir un partitionnement de chaque attribut en sous-ensembles flous tel que l'algorithme de C-moyenne floue (FCM) [22]. Une fois les sous-ensembles flous sont déterminés et la base de degrés d'appartenance est réalisée. C'est cette base qui sera utilisée afin d'extraire les règles d'association floues. Pour découvrir ces règles une variété d'approches a été développée.

Quelques algorithmes classiques d'extraction de règles d'association ont été fuzzifiés en intégrant les concepts de la théorie des ensembles flous. Par exemple l'algorithme de Pierrard et al. [102] repose sur l'algorithme close [101], l'algorithme de Hong et al. [65] est basé sur l'algorithme Apriori [5], les auteurs [66] ont proposés l'algorithme Apriori-TID flou fondé sur l'algorithme Apriori-TID [5], ces différents algorithmes développés génèrent toutes les règles d'association flous du même principe que ses anciens (c.-à-d. d'une manière classique), généralement, les différentes propositions étaient assez similaires, la différence est dans la manière de calculer le support et la confiance, par exemple Hong et al. [65] ont proposés d'utiliser l'opérateur "min" comme une t-norme avec la cardinalité ($\sum$comptage). Une autre méthode a été proposée par Kuok et al. [76], ils ont utilisés deux facteurs pour valider les règles d'association floues, un facteur de signification et un facteur de certitude. Une règle est valide si est seulement si ses deux facteurs sont supérieurs à des seuils définis

par l'utilisateur, le premier facteur est basé sur la notion du support, il a utilisé l'opérateur produit ($\prod$) avec un $\sum$comptage seuillé pour le calculer, le facteur de certitude est le ratio du facteur de signification de l'itemset constituant la règle sur le facteur de signification de la condition.

2.4 Exemple d'xtraction de règles d'association floues

— **Base de données**

Soit la base de transactions (voir tableau 3.7) qui comporte la quantité achetée pour chaque article (item).

TABLE 3.7 Exemple d'une base de transactions quantitatives

Transactions	Item1	Item2	Item3	Item4
t_1	1	1	5	4
t_2	1	2	4	3
t_3	3	2	4	1
t_4	2	2	5	2
t_5	1	3	3	2
t_6	2	3	4	1

— **Représentation des fonctions d'appartenances**

Tous d'abord il faut convertir la base de données quantitative en base de données de degrés d'appartenance. Les partitions des sous-ensembles flous de chaque item quantitatif et les fonctions d'appartenance pour chacun de ces sous-ensembles sont illustrées par la figure 3.3). Par exemple, l'item2 est découpé en 3 parties : peu (p), moyen (m), beaucoup (b), l'achat d'un seul item2 est considéré comme appartenant au sous ensemble flou "peu", 2 comme quantité "moyenne", 3 quantités de l'item2 correspondent à la fois une à une quantité "moyenne" et "beaucoup", 4 quantités de l'item2 et plus est considéré comme "beaucoup". A partir des fonctions d'appartenance ci-dessus, on définit la base transactionnelle floue (tableau 3.8), qui donne les degrés d'appartenance de chaque transaction pour chacun des sous-ensembles flous.

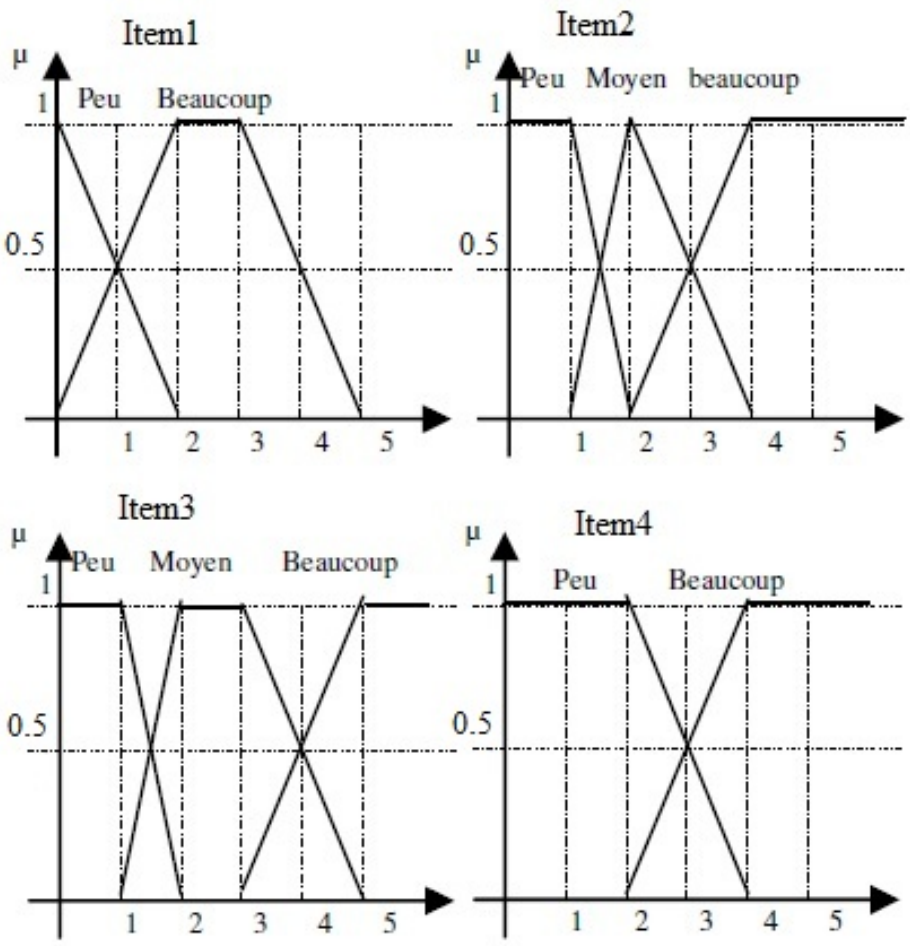


FIGURE 3.3 Partitionnement flou des attributs quantitatifs

TABLE 3.8 Base des degrés d'appartenances

Transactions	Item1		Item2			Item3			Item4	
	P	B	P	M	B	P	M	B	P	B
t_1	0.5	0.5	1	0	0	0	0	1	0	1
t_2	0.5	0.5	0	1	0	0	0.5	0.5	0.5	0.5
t_3	0	1	0	1	0	0	0.5	0.5	1	0
t_4	0	1	0	1	0	0	0	1	1	0
t_5	0.5	0.5	0	0.5	0.5	0	1	0	1	0
t_6	0	1	0	0.5	0.5	0	0.5	0.5	1	0

— **Recherche des items fréquents**

Les items flous correspondant à la base exemple sont les suivants : [Item1, peu], [Item1, beaucoup], [Item2, peu], [Item2, moyen], [Item2, beaucoup], [Item3, peu], [Item3, moyen], [Item3, beaucoup], [Item4, peu], [Item4, beaucoup].

Si on fixe les seuils α-coupe = 0,5 et un support minimal minsup = 0.45, les items flous fréquents sont (voir tableau 3.9) : [Item1, beaucoup], [Item2, moyen], [Item3, beaucoup] et [Item4, peu].

TABLE 3.9 Support flou des items

Item	Sous-ensemble flou	Support flou
Item1	Peu	0.25
	Beaucoup	0.75
Item2	Peu	0.166
	Moyen	0.666
	Beaucoup	0.166
Item3	Peu	0
	Moyen	0.416
	Beaucoup	0.583
Item4	**Peu**	0.75
	Beaucoup	0.25

— **Calcule du support et confiance d'une règle d'association floue**

Par exemple, le support et la confiance en utilisant l'équation 3.3 et 3.5 pour la règle [Item2, moyen] $\longmapsto$ [Item1, beaucoup]

$Support = \frac{min(0,0.5)+min(1,0.5)+min(1,1)+min(1,1)+min(0.5,0.5)+min(0.5,1)}{6} = \frac{0.5+1+1+0.5+0.5}{6}$
$= 0.583$

$Confiance = \frac{min(0,0.5)+min(1,0.5)+min(1,1)+min(1,1)+min(0.5,0.5)+min(0.5,1)}{0+1+1+1+0.5+0.5} = \frac{3.5}{4} = 0.875$

Ce qui signifie que : si achète moyen de l'item2 alors achète beaucoup de l'item1 dans 87.5% cas. Moyen de l'item2 et beaucoup de l'item1 sont achetés ensemble dans 58.3% dans la liste d'achats.

3 Conclusion

Dans ce chapitre, les règles d'association classiques et floues ont été présentées, ces règles montrent leurs utilités dans plusieurs domaines d'applications pour découvrir des relations cachées et des associations inattendues entre les attributs dans une base de données, pour l'extraction de ces règles, on s'est focalisé sur l'algorithme Apriori, parce que c'est la base de plusieurs algorithmes qui ont été développés. Dans le chapitre suivant nous présentons notre méthode développée basée sur les règles d'association floues afin d'estimer les précipitations à partir des images multi-spectrales du satellite MSG.

Chapitre 4

Estimation des précipitations à partir des images MSG en utilisant les règles d'association floues

1 Introduction

Dans ce chapitre, nous présentons une méthode développée pour l'estimation des précipitations appliquée au nord-est de l'Algérie en utilisant des images multi-spectrales du satellite MSG. Pour estimer les précipitations à partir de ces images, la majorité des études [28, 9, 2, 55, 80, 20, 132, 85, 70, 69, 124] qui ont été effectuées ne profitent que des données de quelques canaux, ils n'exploitent pas suffisamment toutes les données fournies par ce satellite, alors que ces téraoctets de données sont potentiellement riches en ressources inouïes qui demandent à être exploitées. De plus, ces études classifient et affectent les pixels à une classe d'une manière classique, par exemple un pixel est considéré 100 % précipitant où bien à 0 % non précipitant, alors qu'on ne peut réellement le classifier d'une manière nette et précise, à cet effet l'introduction du caractère flou est une nécessité, il nous a permis de traiter ces données incertaines d'une manière plus flexible où chaque pixel a un degré propre d'appartenance à une certaine classe. Pour cela nous avons proposé une méthode [23] qui exploite les images des 11 canaux (Sauf le canal HRV) et qui construit un modèle d'estimation sous la forme de règles d'association floues pour estimer les précipitations dans le nord-est de l'Algérie. Chaque règle est sous la forme si (condition) - alors (conclusion), où la condition est une combinaison des différentes classes floues d'images MSG, et la conclusion contient une seule classe floue qui représente l'intensité de la pluie : pas de précipitations, faible, modérée et élevée.

2 Description de la méthode

Notre méthode proposée se compose de trois étapes essentielles :

1. Création d'une base de données transactionnelle : les images MSG et MPE sont utilisées pour créer notre base de données initiale.
2. Création d'une base de données transactionnelle floue : qui est responsable de la transformation de la première base de données transactionnelle créée en base de données transactionnelle floue en utilisant les fonctions d'appartenance trapézoïdales et l'algorithme fuzzy c-means (FCM) [22].
3. Extraction de règles d'association floues : où les règles d'association sont extraites à l'aide d'une version étendue de l'algorithme Apriori original [5] en utilisant les définitions précédentes du support flou d'un itemset flou et la confiance floue d'une règle d'association floue.

La figure 4.1 illustre les différentes étapes de notre méthode.

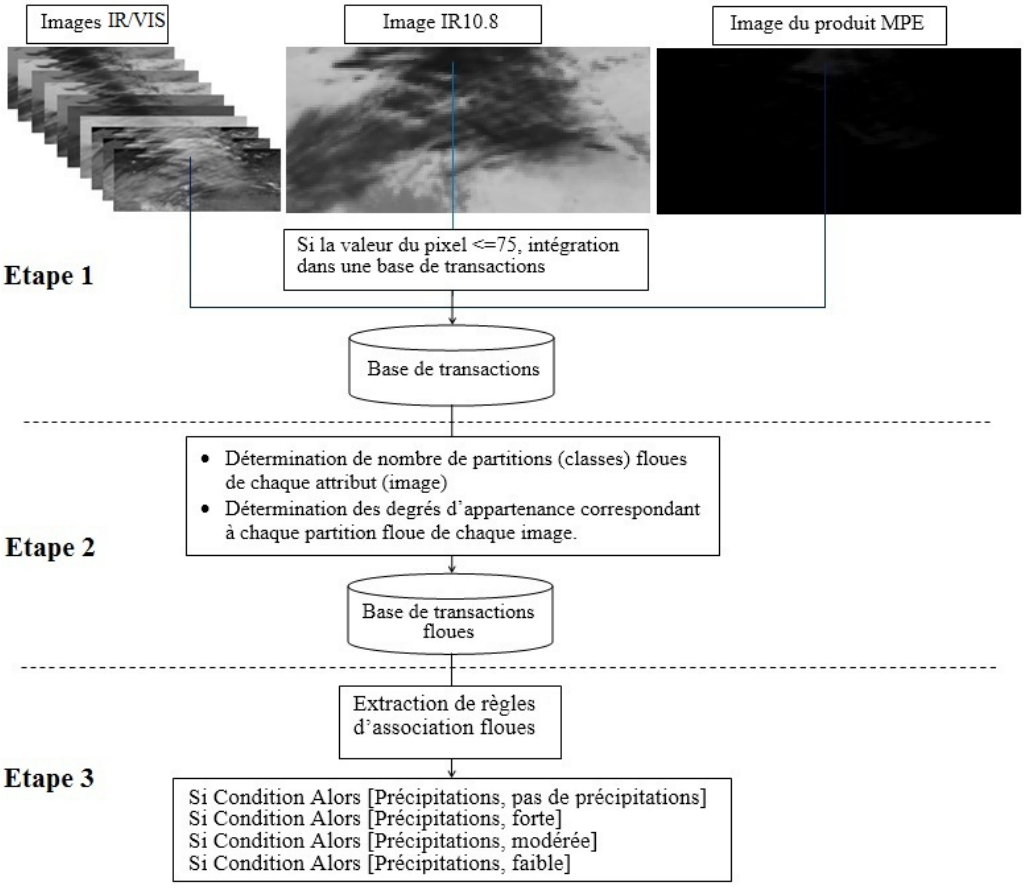


FIGURE 4.1 La procédure générale de notre méthode proposée

Dans ce qui suit, nous décrivons chaque étape en détail.

2.1 Création d'une base de données transactionnelle

Dans notre étude, pour créer la base de données transactionnelle, douze (12) attributs (items) sont considérés, où le premier attribut représente les images MPE (nommé précipitations) et les autres représentent les images MSG, sauf le canal HRV qui a des dimensions différentes par rapport aux autres. Les noms des canaux sont considérés comme des noms pour les attributs utilisés. Chaque image MSG est composée d'un ensemble de pixels, et chaque valeur de pixel représente l'intensité des rayons réfléchis par les nuages ou la surface de la Terre mesurée par le capteur SEVERI dans différents bandes spectrales [112]. L'intensité des pixels est codée numériquement sur 8 bits, par conséquent, sa valeur (c'est-à-dire son niveau de gris) est une valeur numérique compris dans l'intervalle [0, 255], où 0 représente la couleur noire et 255 indique la couleur blanche. Pour construire la base de données transactionnelle, nous assignons à l'attribut IR10.8, les valeurs de pixel de l'image du canal infrarouge thermique IR10.8 dont les températures sont inférieures à -35°C [11] un seuil de température qui représente une possibilité élevée d'avoir des précipitations. La relation entre la température et le compte numérique d'un pixel de l'image IR10.8 ($CN_{IR10.8}$) est calculée selon l'équation 4.1 [113].

$$T(^{\circ}C) = 55 - CN_{IR10.8}/2 \tag{4.1}$$

Ensuite, pour chaque $CN_{IR10.8}$ chargé à une date-heure (dt) avec les coordonnées pixels (x, y) correspondant à la transaction (t) de la base de données créée, nous chargeons les valeurs pixels des autres images MSG et les quantités de précipitations (QP) des images MPE dans leurs attributs associés pour la transaction de la même base de données (t) on prenant en compte leurs images, qui sont prises à la même date et heure dt. Leurs pixels associés ont les mêmes coordonnées du ($CN_{IR10.8}$). Selon EUMETSAT, la QP de chaque image MPE est défini par l'équation (4.2).

$$QP(mm/hour) = CN_{MPE} * 0.144, \tag{4.2}$$

Où CN_{MPE} indique la valeur du pixel de l'image MPE. La figure 4.2 représente une partie du notre base de données transactionnelle créée.

N_Pixel	DateTime	Coordonnées	VIS_0.6	VIS_0.8	IR_1.6	IR_3.9	WV_6.2	WV_7.3	IR_8.7	IR_9.7	IR_10.8	IR_12.0	IR_13.4	Précipitations(mm/H)
617	2016/11/07 08:00	(18,50)	185	210	95	94	34	34	36	40	35	33	33	5,040
618	2016/11/07 08:00	(20,50)	212	243	97	96	37	39	44	44	43	41	40	4,464
619	2016/11/07 08:00	(22,50)	205	236	91	92	41	43	48	46	47	45	42	4,896
620	2016/11/07 08:00	(24,50)	189	216	84	90	42	46	49	47	48	47	42	4,032
621	2016/11/07 08:00	(26,50)	178	199	83	90	42	45	46	46	45	44	42	4,320
622	2016/11/07 08:00	(28,50)	183	206	90	92	41	42	47	46	46	44	41	4,320
623	2016/11/07 08:00	(30,50)	161	184	70	92	43	48	56	50	54	52	45	1,872
624	2016/11/07 08:00	(32,50)	163	180	73	88	46	51	53	49	51	50	46	2,304
625	2016/11/07 08:00	(34,50)	146	168	60	92	47	53	62	53	61	59	52	0,432
626	2016/11/07 08:00	(36,50)	128	139	54	88	50	57	58	52	57	56	51	1,440
627	2016/11/07 08:00	(38,50)	144	156	70	86	44	49	46	46	45	45	42	4,176
628	2016/11/07 08:00	(40,50)	161	182	79	90	38	41	42	44	42	41	39	4,176
629	2016/11/07 08:00	(42,50)	179	202	90	92	38	41	45	44	43	41	40	4,176
630	2016/11/07 08:00	(44,50)	198	224	96	94	40	43	48	47	48	46	43	4,176
631	2016/11/07 08:00	(46,50)	196	222	91	96	43	48	54	49	54	52	47	1,728
632	2016/11/07 08:00	(48,50)	174	193	80	88	48	54	56	50	54	53	49	1,584
633	2016/11/07 08:00	(50,50)	185	209	88	88	44	51	56	51	56	54	48	1,440
634	2016/11/07 08:00	(52,50)	174	196	76	96	48	55	62	54	62	60	53	0,432
635	2016/11/07 08:00	(54,50)	168	190	70	92	51	60	67	55	67	64	55	0,288
636	2016/11/07 08:00	(56,50)	139	153	45	92	53	65	73	60	72	70	60	0,288
637	2016/11/07 08:00	(58,50)	135	149	50	90	52	63	64	56	62	61	56	0,432
638	2016/11/07 08:00	(60,50)	171	192	81	84	45	51	51	48	50	49	44	2,304
639	2016/11/07 08:00	(62,50)	126	139	49	84	48	56	59	52	58	56	52	1,152
640	2016/11/07 08:00	(64,50)	150	169	68	80	45	52	49	46	49	48	46	2,448

FIGURE 4.2 Partie de la base de données transactionnelle des images MSG

2.2 Création d'une base de données transactionnelle floue

Pour créer une base de données transactionnelle floue (BTF) à partir de la base de données transactionnelle initiale (BTI), nous divisons l'univers de chaque attribut quantitatif en plusieurs sous-ensembles flous. Chaque attribut de BTI est alors associé à un sous-ensemble flou pour former un item flou. Pour calculer les nouvelles valeurs floues du BTF, nous attribuons à chaque valeur originale de chaque attribut son degré d'appartenance aux différents ensembles flous trouvés correspondant à la partition de l'attribut, pour cela nous utilisant les fonctions d'appartenance trapézoïdales pour les attributs précipitations et visibles (VIS0.6, VIS0.8, NIR1.6), ainsi que l'algorithme FCM pour les attributs infrarouge (IR3.9, WV6.2, WV7.3, IR8.7, IR9.7, IR10.8, IR12.0 et IR13.4). Dans ce qui suit, nous déterminons le nombre des items flous de BTF (c'est-à-dire le nombre de sous-ensembles flous pour chaque attribut du BTI), aussi le degré d'appartenance des valeurs de l'attribut à chaque ensemble flou trouvé.

2.2.1 Détermination du nombre des items flous

Pour chaque attribut du BTI, le nombre de ses sous-ensembles sont définis comme suit :

(a) **Les attributs visibles**

La valeur du compte numérique pour ces canaux qui représente la brillance d'une image visible varie entre [0, 255], nous avons proposé de partager cet intervalle de brillance en 5 partitions floues comme suit : [très sombre, sombre, brillance moyenne, brillant, très brillant], nous avons utilisé 5 fonctions d'appartenance trapézoïdales pour les représenter comme il est indiqué dans la figure 4.3. Par conséquent, pour chaque attribut visible dans BTI, cinq items flous associés sont créés dans le BTF.

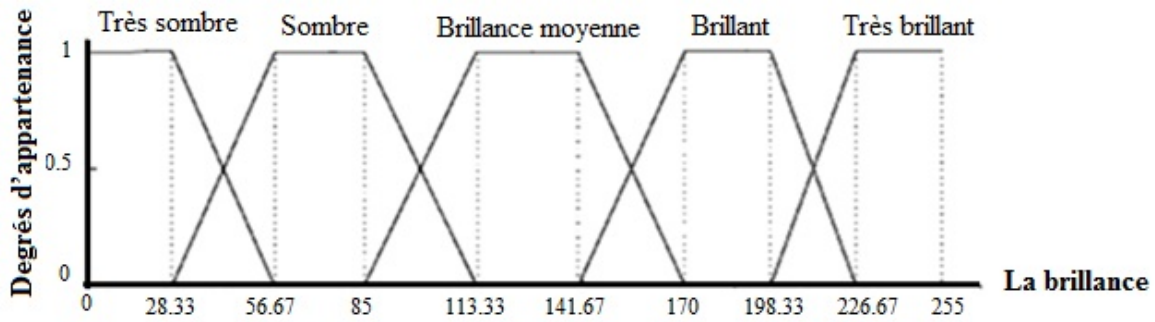


FIGURE 4.3 Partitionnement floue des attributs visibles

(b) **L'attribut précipitations**

D'après l'ONM les précipitations en Algérie sont considérées (voir tableau 4.1).

TABLE 4.1 Quantité de précipitations en Algérie

	Quantité de précipitations en millimètre par heure (mm/h)
Pas de précipitations	0
Faible	supérieur à 0 jusqu'à 3
Modérée	4 à 7
Forte	8 et plus

Nous avons utilisé 4 fonctions d'appartenance trapézoïdales pour représenter l'attribut précipitations comme indiqué dans la figure 4.4. Par conséquent, nous avons créé 4 items flous pour l'attribut précipitations.

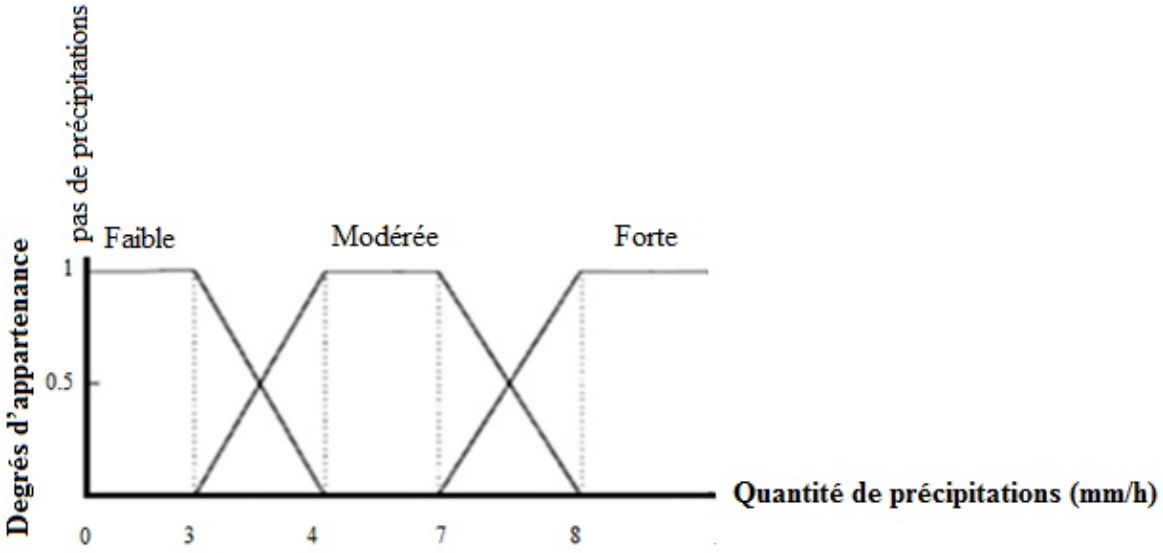


FIGURE 4.4 Partition floue de l'attribut précipitations

(c) **Les attributs infrarouges**

Pour l'attribut précipitations, le nombre des sous-ensembles flous est défini par les experts de l'ONM, alors que les attributs des canaux visibles sont déterminés en fonction de connaissance de la luminosité des images MSG (dans notre étude, nous proposons cinq sous-ensembles flous). Cependant, dans le cas des attributs du canal infrarouge, le nombre de sous-ensembles flous est ambigu, car nous ne disposons d'aucune information d'expert sur ces ensembles en terme linguistique. Par conséquent, nous devons utiliser des algorithmes flous pour définir le nombre de sous-ensembles pour chaque attribut de canal infrarouge. Dans notre étude, l'algorithme le plus connu FCM [22] et l'indice de validité [21] sont utilisés pour définir le nombre de sous-ensembles (clusters) pour chaque attribut infrarouge du *BT1*. La recherche du nombre

optimale du nombre de groupe s'effectue entre un nombre minimum ($C_{min} \geq 2$) et un nombre maximum (C_{max}) du groupe défini par l'utilisateur, par exemple selon Bezdek [100] le nombre de partitions est entre $C_{min} = 2$ et $C_{max} = \sqrt{n}$, où n est le nombre des éléments, alors que dans [25], $C_{min} = 2$ et $C_{max} = 10$. Ce processus de recherche se fait automatiquement, celui qui maximise l'indice de validité est retenu. Pour chaque attribut infrarouge IR_i de BTI, avec l'ensemble des éléments $E_i = \{x_1, x_2, \ldots x_n\}$, où i indique le $i^{\text{ème}}$ attribut infrarouge, n est le nombre des éléments de E_i, et $x_{t,(t=1,2\ldots n)}$ est la valeur du pixel de la $t^{\text{ème}}$ transaction pour l'attribut IR_i. Nous attribuons à IR_i le nombre de clusters C_i, où C_i est un nombre entier dans $[C_{min}^{IR_i}, C_{max}^{IR_i}]$, où $C_{min}^{IR_i}$ et $C_{max}^{IR_i}$ représentent le nombre minimal et maximal pour IR_i, respectivement. Pour définir le nombre optimal de clusters C_i pour chaque IR_i, les étapes suivantes sont suivies.

- Étape 1. pour $C_i = C_{min}^{IR_i}$ jusqu'à $C_{max}^{IR_i}$ faire
  - Appliquer l'algorithme FCM [22] en utilisant les C_i et E_i.
  - Calculer l'indice de validité (V_{C_i}) [21].
- Étape 2. le nombre optimal de clusters C_i est celui qui maximise le V_{C_i}.

2.2.2 Détermination des degrés d'appartenance correspondant à chaque partition (classe) floue de chaque item

Nous attribuons à chaque valeur de chaque attribut son degré d'appartenance aux différents sous-ensembles flous trouvés correspondant à la partition de l'attribut, pour notre application les lignes représentent les transactions et les colonnes représentent les différentes partitions floues trouvées. L'intersection d'une ligne et d'une colonne représente le degré d'appartenance de la valeur quantitative dans la partition floue. Pour déterminer ce degré d'appartenance nous avons utilisé 5 fonctions d'appartenance trapézoïdales (voir figure 4.3) pour les attributs des canaux VIS0.6, VIS0.8, NIR1.6 et 4 fonctions d'appartenance trapézoïdales (voir figure 4.4) pour l'attribut précipitations. Pour les autres attributs nous avons appliqué l'algorithme FCM [22] basé sur la minimisation d'une fonction objective. Les deux notations suivantes sont considérées pour notre application :

1. $[VR, FI]$ est un item flou pour l'attribut visible/précipitations VR avec son sous-ensemble flou FI. Par exemple, [VIS0.6, brillant] est un item flou de l'attribut visible VIS0.6 avec son sous-ensemble flou associée 'Brillant'.
2. $[IR, C(cr)]$ est un item flou pour l'attribut infrarouge IR avec son sous-ensemble flou C, où cr est une valeur réelle qui représente le centre du cluster C. Par exemple, $[IR12.0, C2(47.28)]$ est un item flou de l'attribut $IR12.0$ qui correspond au cluster $C2$, ayant la valeur centrale 47.28.

2.3 Extraction de règles d'association floues

Notre but est d'extraire des règles d'association floues sous forme de 'si *condition* alors *conclusion*', où la partie droite (*conclusion*) de chaque règle générée ne contient qu'un seul item flou, il s'agit de l'item flou [Précipitations, FI], où $FI \in \{$ faible, modérée, forte, pas de précipitations$\}$ sont les sous-ensembles flous associés à l'item précipitations. Pour extraire ces règles d'association, deux étapes sont suivies. Tout d'abord, une version étendue de l'algorithme Apriori [5] est utilisée pour trouver la liste des ensembles d'itemsets flous fréquents à partir de la base de données transactionnelles BTF. Deuxièmement, les règles d'association sont générées à partir des ensembles itemsets fréquents trouvés. Les descriptions de ces deux étapes sont présentées dans ce qui suit.

2.3.1 Identification de la liste des itemsets flous fréquents

Pour trouver la liste des itemsets flous fréquents ($LIFF$), nous adoptons l'algorithme Apriori dans notre estimation des précipitations. L'algorithme Apriori traite des ensembles des items classiques, alors que dans notre méthode, les items utilisés sont flous. Pour cela, une extension de l'algorithme Apriori est développée, en utilisant la formule de calcul du support flou donné à l'équation 3.3 au lieu d'utiliser un support normal. Notre algorithme développé appelé algorithme Apriori flou, son pseudocode est donné dans l'algorithme 4.

Algorithm 4 Algorithme Apriori flou

Input :
- BTF : base de données transactionnelle floue, Mfs : support minimum flou.
- α-coupe : seuil minimal d'appartenance à un sous-ensemble flou.

Output :
- $LIFF$: liste des itemsets flous fréquents.

1: L_1 = liste des 1-itemsets fréquents.
2: **for** $(k = 2\,; L_{k-1} \neq \emptyset\,; k++)$ **do**
3: $C_k \leftarrow Genere\text{-}Ccandidats(L_{k-1})$ ▷ les candidats k-itemsets flous.
4: **for all** transaction $t \in BTF$ **do**
5: $C_t = subset(C_k, t, \alpha\text{-coupe})$
6: **for all** candidates $c \in C_t$ **do**
7: $c.supp = +\top(\alpha_A(t[X])) \triangleright \top(\alpha_A(t_i[X]))$: est la valeur minimale des degrés d'appartenance des items flous (X, A) relatifs au itemset flous c dans la transaction t.
8: **end for**
9: **end for**
10: $L_k = \{c \in C_k | \frac{c.supp}{|T|} >= Mfs\}$ ▷ $|T|$: est le nombre total de transactions de BTF.
11: **end for**
12: $LIFF = \bigcup_k L_k$

Comme le montre l'algorithme 4, afin de déterminer la liste L_1 des 1-itemsets fréquents, un balayage de la base BTF est effectuée pour calculer le support de chaque item flou, nous gardons uniquement les items flou qui ont un support supérieur au support minimum flou. Ensuite, les étapes de 2 à 10 sont répétées jusqu'à ce que le critère d'arrêt soit stratifié (i.e. $L_k = \emptyset$). Pour chaque itération k, la procédure Genere-Ccandidats (L_{k-1}) construit l'ensemble C_k des k-itemsets flous candidats en joignant les itemsets flous fréquents de taille k-1 (L_{k-1}) avec lui-même. Dans l'ensemble C_k, on supprime chaque itemset flous candidat si au moins un de ses sous-ensembles d'itemsets flous de taille $k-1$ n'est pas fréquent, et nous supprimons aussi les itemsets flous non valides. Afin de trouver l'ensemble L_k à partir de l'ensemble C_k, tout d'abord, pour chaque transaction t dans la base de données BTF, un ensemble d'itemsets flous C_t est créé à partir de C_k en prenant en compte uniquement les itemsets flous de C_k qui satisfont un seuil minimal d'appartenance α-coupe, ensuite, pour chaque $c \in C_t$, nous calculons son support flou c.supp, et enfin le L_k est créé en gardant uniquement les itemsets flous c qui ont un support supérieur ou égal à Mfs. Afin d'illustrer l'algorithme Apriori flou (Algorithme 4), nous donnons un exemple à la figure 4.5.

Étant donné le BTF comme le montre la figure 4.5, où la première colonne indique le numéro de la transaction du BTF et les autres colonnes représentent trois éléments (Item1, Item2 et Item3) associés à leurs sous-ensembles flous FI1.1 jusqu'à FI3.2. Les deux paramètres Mfs et α-coupe sont définis avec les valeurs 0.25 et 0.5 respectivement. La première étape consiste à trouver la liste des 1-itemsets flous (L_1), qui contient tous les itemsets flous de la liste des candidats C_1 dont le support flou est supérieur à Mfs (voir les itemsets flous non colorés). Ensuite, chaque $C_k, k = 2,3,..$ est généré en utilisant sa liste d'itemsets flous fréquents l_{k-1}. Par exemple, la liste de candidats C_2 est générée en effectuant une jointure entre l'ensemble flou L_1 et lui-même (i.e. $C_2 = L_1 \bowtie L_1 = \{(1),(4),(6),(7)\} \bowtie \{(1),(4),(6),(7))\} = \{(1\ 4),(1\ 6),(1\ 7),(4\ 6),(4\ 7),(6\ 7)\}$), le dernier itemsets flous $(6\ 7)$ n'est pas un itemsets valide (voir la cellule dessinée d'une bordure pointillé) car les deux items 6 et 7 ont le même Item3. Par conséquent, nous le supprimons de C_2. Pour trouver L_2, nous calculons les supports flous de C_2, L_2 est créé en y ajoutant les itemsets flous de C_2, où leurs supports flous sont $k \geq Mfs$. Le même processus (jointure et suppression) est appliqué à la génération de C_3 en utilisant L_2. Où l'itemset flou $(1\ 6\ 7)$ est supprimé de C_3 car ce n'est pas un itemset flou valide, on élimine aussi l'itemset flou $(1\ 4\ 7)$ (voir la cellule dessinée avec une bordure en pointillés gras) car il contient l'itemset flou $(4\ 7)$ et qui n'existe pas à L_2. En résultat $LIFF = L_1 \cup L_2 \cup L_3 = \{(1),(4),(6),(7)\} \cup \{(1\ 4),(1\ 6),(1\ 7),(4\ 6)\} \cup \{(1\ 4\ 6)\}$.

Base de transactions floues BTF

Transactions	Item1		Item2			Item3	
	FI1.1	FI1.2	FI2.1	FI2.2	FI2.3	FI3.1	FI3.2
t_1	0.9	0.1	0.15	0.85	0	0	1
t_2	0.9	0.1	0	0.7	0.3	0.9	0.1
t_3	0.3	0.7	0.6	0.3	0.1	0.8	0.2
t_4	0.8	0.2	0	0.2	0.8	0.1	0.9
t_5	1	0	0.1	0.9	0	1	0

Mfs=0.25 et α-coupe= 0.5

C_1=1-itemsets flous

Numéro d'itemset flou	Itemset flou	Support Itemset flou
1	[Item1, FI1.1]	0.72
2	[Item1, FI1.2]	0.14
3	[Item2, FI2.1]	0.12
4	[Item2, FI2.2]	0.49
5	[Item2, FI2.3]	0.16
6	[Item3, FI3.1]	0.54
7	[Item3, FI3.2]	0.38

L_1

Itemset flou	Support Itemset flou
(1)	0.72
(4)	0.49
(6)	0.54
(7)	0.38

$C_2=L_1 \bowtie L_1$

Itemset flou	Support Itemset flou
(1 4)	0.49
(1 6)	0.38
(1 7)	0.34
(4 6)	0.32
(4 7)	0.17
(6 7)	

L_2

Itemset flou	Support Itemset flou
(1 4)	0.49
(1 6)	0.38
(1 7)	0.34
(4 6)	0.32

$C_3=L_2 \bowtie L_2$

Itemset flou	Support Itemset flou
(1 4 6)	0.32
(1 4 7)	
(1 6 7)	

L_3

Itemset flou	Support Itemset flou
(1 4 6)	0.32

FIGURE 4.5 Exemple d'extraction des itemsets flous fréquents en utilisant l'algorithme Apriori flou

2.3.2 Génération de règles d'association floues

Pour extraire la liste des règles d'association floues ($LRAF$), l'algorithme 5 est utilisé pour générer cette liste à partir de $LIFF$, qui est trouvée par l'algorithme Apriori flou.

Algorithm 5 Génération de règles d'association floues

Input :
- $LIFF$: liste des itemsets flous fréquents.
- Mfc : confiance minimale floue.
- α-coupe : seuil minimal d'appartenance à un sous-ensemble flou.

Output :
- $LRAF$: liste des règles d'association floues.

```
1:  for each k-itemsets flous fréquent L_k ∈ LIFF où k ≥ 2 do
2:      for each itemsets flous (X,A) ∈ L_k do
3:          if ∃[Prcipitations,FI] ∈ (X,A) then        ▷ FI ∈ {faible, modérée, forte,
                                                          pas de précipitations}.
4:              Générer les règles d'association floues RAF à partir de (X,A) comme suit :
                    (X,A) − [Prcipitations,FI] ⇒ [Prcipitations,FI]
5:              Calculer la confiance Conf_RAF de RAF en utilisant l'équation (3.5)
6:              if Conf_RAF ≥ Mfc then
7:                  LRAF = LRAF ⋃ RAF
8:              end if
9:          end if
10:     end for
11: end for
```

Comme il est indiqué dans le pseudocode de l'algorithme 5, on prend uniquement les itemsets flous (X,A) de la liste L_k (k $\geq$2) contenant l'item flou $[prcipitations,FI]$, où X représente la liste des attributs utilisés, et A leurs sous-ensembles flous associés et $FI \in \{$ faible, modérée, forte, pas de précipitations$\}$ sont les sous-ensembles flous associés à l'item précipitations. Pour chaque itemset flou (X,A) pris, la règle d'association floue RAF est générée comme suit : $(X,A) - [Prcipitations,FI] \Rightarrow [Prcipitations,FI]$ et calcule sa confiance floue $Conf_{RAF}$. Si $Conf_{RAF} \geq Mfc$, nous ajoutons RAF à la liste finale $LRAF$, où Mfc est la confiance minimale floue définie par l'utilisateur. Soit la liste $LIFF = L_1 \cup L_2 \cup L_3 = \{(1),(4),(6),(7)\} \cup \{(1\,4),(1\,6),(1\,7),(4\,6)\} \cup \{(1\,4\,6)\}$, généré par l'algorithme Apriori flou dans l'exemple présenté à la figure 4.5. En considérant que l'item2 est l'attribut précipitations, où leurs sous-ensembles flous associés sont FI2.1, FI2.2 et FI2.3, donc, les items flous de l'attribut précipitations sont : [Item2, FI2.1], [Item2, FI2.2] et [Item2, FI2.3] qui sont représentés par les nombres 3, 4 et 5 respectivement. À partir de $LIFF$, seuls les itemsets flous fréquents (X,A)= $\{(1\,4),(4\,6),(1\,4\,6)\}$ sont pris en considération pour générer les règles d'association floues, car L_1 est une liste de 1-itemsets flous, et $(1\,6),(1\,7)$ sont

des 2-itemsets flous qui ne contiennent pas au moins un item flou de l'attribut précipitations (i.e. un des nombres :3, 4 et 5). Donc les règles d'association floues générées sont $RAF = \{1 \Rightarrow 4, 6 \Rightarrow 4, (1 \wedge 6) \Rightarrow 4\}$, où $\wedge$ indique l'opérateur de conjonction flou. Chaque RAF est généré à partir de (X, A) en attribuant à la partie droite (conclusion) du RAF un seul item flou ($[prcipitations, FI]$) et la partie gauche (condition) de RAF le reste des items flous de (X, A). Par exemple, $(1 \wedge 6) \Rightarrow 4$ est une RAF générée à partir d'itemsets flou (1 4 6), où sa conclusion est l'item flou [Item2, FI2.2], qui est représenté par le nombre 4. En utilisant l'équation 3.5, où Mfc = 0.65 et α-coupe = 0.5, la confiance de chaque RAF générée est calculée et seules les RAF ayant une confiance $\geq Mfc$ sont considérées pour construire la liste finale $LRAF$. La confiance des RAF générées pour l'exemple sont : $Conf_{1 \Rightarrow 4} = 0.68$, $Conf_{6 \Rightarrow 4} = 0.59$ et $Conf_{(1 \wedge 6) \Rightarrow 4} = 0.84$, donc $LRAF = \{1 \Rightarrow 4, (1 \wedge 6) \Rightarrow 4\}$.

3 Résultats expérimentaux

Pour vérifier et valider la performance de notre proposition, deux bases de données réelles sont utilisées. La première pour l'apprentissage et construction de notre modèle, ensuite nous avons appliqué ce modèle sur la deuxième base de données pour valider notre méthode. Les données utilisées sont bien décrit dans la section 'données utilisées'. Ensuite, le modèle construit est présenté dans la section 'construction de modèle'. Dans la dernière section 'Validation du modèle', nous le validons en effectuant une comparaison avec le MPE, qui est un produit très utile pour estimer les précipitations.

3.1 Données utilisées

Dans notre recherche, nous avons étudié la région du Nord-Est de l'Algérie car c'est la région la plus précipitante d'après l'ONM. Elle se situe entre 34.5° Nord et 37 ° Nord, et entre 3.5° Ouest et 9° Est (voir figure 4.6). Deux bases de données sont créées en collectant les images SEVIRI de 11 canaux multi-spectraux du satellite MSG (à l'exception du canal HRV) et les images du produit MPE durant la période allant de du 1^{er} octobre 2016 au 31 mars 2017, où les images sont acquises toutes les 15 minutes de 8 :00 à 16 :00. Pour chaque heure, la première base de données (notée *TrainDB*) a été utilisée pour l'apprentissage du modèle développé. Elle est créée en attribuant à *TrainDB* les images acquises dans les minutes (00, 15, 30). La deuxième base de données (notée *ValidDB*) est exploitée pour valider notre proposition, elle est construite en enregistrant les images acquises à la 45 min sur *ValidDB*.

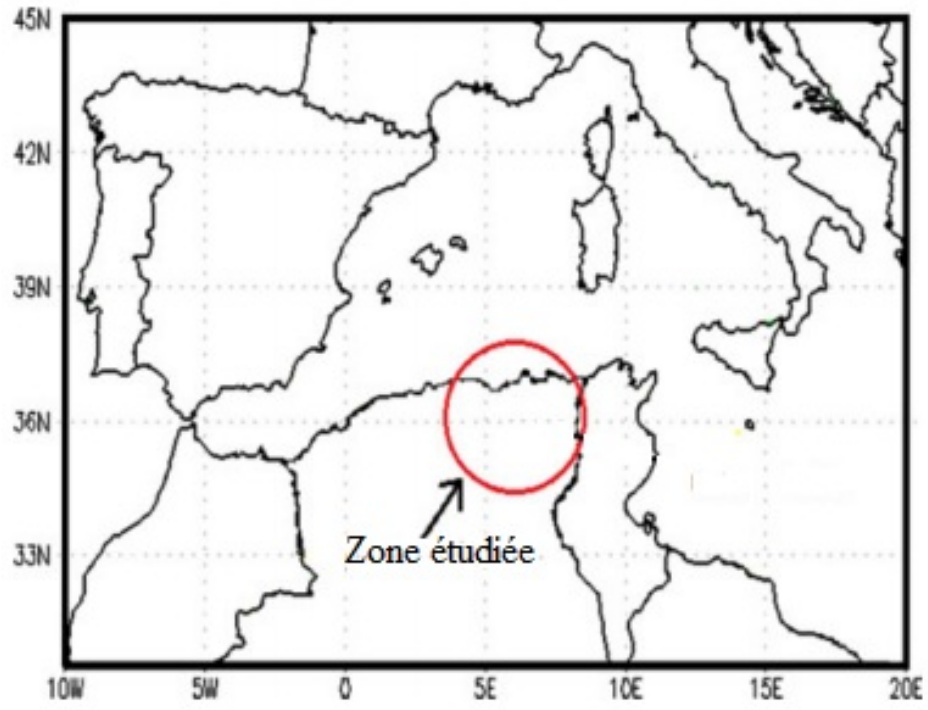


FIGURE 4.6 La zone d'étude

3.2 Construction du modèle

Pour extraire la liste des règles d'association floues (modèle), la base de données $TrainDB$ est utilisée pour l'apprentissage du modèle. L'algorithme FCM [22] et l'indice de validité [21] sont utilisés ensemble pour définir le nombre de sous-ensembles flous pour les attributs infrarouges et leurs valeurs floues (dégrées d'appartenance). Nous avons appliqué l'algorithme FCM avec les paramètres suivants :

— Un degré de flou $m = 2$.

— Une métrique pour déterminer les distances, la métrique utilisée est la distance Euclidienne.

— Epsilon ($\xi = 0.001$) comme un critère d'arrêt.

— C_{min} = 4 et C_{max} = 10.

Le modèle généré est construit à l'aide de l'algorithme Apriori flou donné dans l'algorithme 4 et l'extraction de règles d'association floues donnée dans l'algorithme 5, leurs paramètres sont les suivants : un support flou minimal Mfs = 0.1, une confiance floue minimale Mfc = 0.7, et un seuil minimal d'appartenance à un sous-ensemble flou α-coupe = 0.66. Dans l'algorithme Apriori flou, nous commençons par la génération de la liste 1-itemsets (L_1),

qui contient la première liste des items flous fréquents. Cette liste est déterminée en prenant uniquement les items flous qui ont un support flou supérieur à Mfs. Notre but est d'extraire des règles d'association floues sous forme de : Condition $\Rightarrow$ [Précipitations, FI], avec $FI \in$ {faible, modérée, forte, pas de précipitations} sont les sous-ensembles flous associés à l'item précipitations. La liste (L_1 devrait contenir tous les items flous de l'attribut précipitations). Cependant, et particulièrement pour notre zone étudiée, la fréquence des précipitations fortes ou modérées est très faible par apport à la fréquence de précipitations faibles ou pas des précipitations. Par conséquent, le support flou des deux items flous [précipitations, modérée] et [précipitations, forte] est faible par rapport aux supports flous des items flous [précipitations, faible] et [précipitations, forte]. Le tableau 4.2 présente le support trouvé de chaque item flou de l'attribut précipitations en appliquant notre méthode proposée sur la base de données $TrainDB$.

TABLE 4.2 Support de chaque item flou de l'attribut précipitations

Item flou	Support
[Précipitations, forte]	0.02
[Précipitations, modérée]	0.04
[Précipitations, faible]	0.36
[Précipitations, pas de précipitations]	0.56

D'après le tableau 4.2, le support des deux items flous [Précipitations, modérée] =0.04 et [Précipitations, forte] = 0.02 est inférieur au $Mfs = 0.1$ (ne sont pas des items flous fréquents), donc on n'aura jamais de règles d'association floues sous forme de condition $\Rightarrow$ [Précipitations, modérée] ou condition $\Rightarrow$ [Précipitations, forte], pour remédier ce problème, et construire un modèle avec une liste de règles d'association floues qui nous aide à décider de classifier chaque pixel (transaction d'une base de données) à sa classe de précipitations (c'est-à-dire , faible, modérée forte, et pas de précipitations), nous avons généré quatre sous-listes de règles d'association flous de l'attribut précipitations. La liste finale des règles d'association floues ($LRAF$) du modèle construit est le regroupement de ces quatre sous-listes floues générées. La base de données floue initiale est notée $FtrainDB$, qui est fournie à partir de $TrainDB$ en exécutant notre méthode proposée. Les étapes suivantes sont effectuées pour générer ($LRAF$).

— Étape 1. Pour i = 1 jusqu'à 4 faire // où i représente l'indice de chaque sous-liste floue générée.

— Étape 2. Calculez les supports flous pour les items flous de l'attribut précipitations.

— Étape 3. Triez par ordre croissant les items flous de l'attribut précipitations en fonction de leurs valeurs de support flou.

— Étape 4. Générez la sous-liste floue des règles d'association floues SL_i de l'item flou de l'attribut précipitation qui a le plus grand support flou en effectuant l'algorithme 4 et 5 sur $FtrainDB$, où les paramètres Mfs et Mfc sont définis avec des nouvelles valeurs.

— Étape 5. Mettre à jour $FtrainDB$ en éliminant les transactions de cet item flou dont le degré d'appartenance est supérieur à α-coupe.

— Étape 6. $LRAF = LRAF \bigcup SL_i$

Les tableaux (4.3, 4.4, 4.5, 4.6) représentent les quatre sous-listes floues du modèle construit (c'est-à-dire $LRAF$).

TABLE 4.3 Liste des règles d'association floues pour l'item flou [Précipitations, pas de précipitations] avec $Mfs = 0.14$ et $Mfc = 0.74$.

Règles d'association floues	Support flou	Confiance floue
21∧36 ⇒ 60	0.156	0.747
21∧40 ⇒ 60	0.155	0.769
21∧48 ⇒ 60	0.168	0.766
21∧52 ⇒ 60	0.173	0.759
21∧56 ⇒ 60	0.153	0.745
36∧40 ⇒ 60	0.143	0.757
36∧48 ⇒ 60	0.164	0.756
36∧52 ⇒ 60	0.175	0.754
36∧56 ⇒ 60	0.167	0.748
40∧48 ⇒ 60	0.198	0.775
40∧52 ⇒ 60	0.176	0.769
48∧52 ⇒ 60	0.203	0.766
48∧56 ⇒ 60	0.161	0.752
52∧56 ⇒ 60	0.174	0.751
21∧40∧48 ⇒ 60	0.145	0.768
21∧48∧52 ⇒ 60	0.155	0.760
36∧40∧48 ⇒ 60	0.141	0.757
36∧48∧52 ⇒ 60	0.160	0.755
36∧48∧56 ⇒ 60	0.147	0.751
36∧52∧56 ⇒ 60	0.158	0.750
40∧48∧52 ⇒ 60	0.175	0.769
48∧52∧56 ⇒ 60	0.159	0.752
36∧48∧52∧56 ⇒ 60	0.146	0.750

Nom des items flous :
21 :[NIR1.6,sombre], 36 :[WV7.3,C4(63.830)], 40 :[IR8.7,C4(73.203)]
48 :[IR10.8,C4(71.075)],52 :[IR12.0,C4(68.569)], 56 :[IR13.4,C4(60.506)]
60 :[Précipitations, pas de précipitations]

TABLE 4.4 Liste des règles d'association floues pour l'item flou [Précipitations, modérée] avec $Mfs = 0.12$ et $Mfc = 0.80$.

Règles d'association floues	Support flou	Confiance floue
30∧50 ⇒ 58	0.152	0.811
30∧54 ⇒ 58	0.168	0.819
30∧34∧46 ⇒ 58	0.124	0.805
30∧34∧50 ⇒ 58	0.137	0.815
30∧34∧54 ⇒ 58	0.142	0.820
30∧46∧54 ⇒ 58	0.125	0.803
30∧50∧54 ⇒ 58	0.144	0.814
30∧34∧46∧50 ⇒ 58	0.120	0.806
30∧34∧50∧54 ⇒ 58	0.132	0.817
30∧46∧50∧54 ⇒ 58	0.123	0.804

Nom des items flous :
30 :[WV6.2, C2(41.189)], 34 :[WV7.3,C2(45.983)], 46 :[IR10.8, C2(49.931)]
50 :[IR12.0, C2(47.281)], 54 :[IR13.4, C2(43.738)], 58 : [Précipitations, modérée]

TABLE 4.5 Liste des règles d'association floues pour l'item flou [Précipitations, forte] avec $Mfs = 0.19$ et $Mfc = 0.65$.

Règles d'association floues	Support flou	Confiance floue
22∧33∧41 ⇒ 59	0.193	0.654
22∧37∧41 ⇒ 59	0.203	0.653
22∧41∧45 ⇒ 59	0.196	0.655
22∧41∧49 ⇒ 59	0.194	0.656
22∧33∧37∧41 ⇒ 59	0.192	0.653
22∧33∧41∧45 ⇒ 59	0.191	0.654
22∧33∧41∧49 ⇒ 59	0.190	0.654
22∧37∧41∧45 ⇒ 59	0.195	0.654
22∧37∧41∧49 ⇒ 59	0.193	0.655
22∧41∧45∧49 ⇒ 59	0.193	0.654
22∧33∧37∧41∧45 ⇒ 59	0.190	0.653
22∧37∧41∧45∧49 ⇒ 59	0.192	0.654

Nom des items flous :
22 :[NIR1.6, brillance moyenne], 33 :[WV7.3, C1(31.673)]
37 :[IR8.7,C1(36.413)], 41 :[IR9.7,C1(44.919)], 45 :[IR10.8,C1(33.617)]
49 :[IR12.0,C1(30.129)], 59 :[Précipitations, forte]

TABLE 4.6 Liste des règles d'association floues pour l'item flou [Précipitations, faible] avec $Mfs = 0.12$ et $Mfc = 0.95$.

Règles d'association floues	Support flou	Confiance floue
12∧21 ⇒ 57	0.195	0.958
18∧31 ⇒ 57	0.135	0.984
21∧31 ⇒ 57	0.198	0.971
21∧35 ⇒ 57	0.150	0.957
21∧47 ⇒ 57	0.137	0.968
21∧51 ⇒ 57	0.131	0.964
21∧55 ⇒ 57	0.141	0.958
31∧35 ⇒ 57	0.148	0.970
31∧55 ⇒ 57	0.141	0.968
35∧47 ⇒ 57	0.131	0.967
35∧51 ⇒ 57	0.138	0.965
35∧55 ⇒ 57	0.150	0.962
39∧47 ⇒ 57	0.125	0.987
47∧51 ⇒ 57	0.155	0.971
47∧55 ⇒ 57	0.128	0.965
51∧55 ⇒ 57	0.140	0.964
35∧47∧51 ⇒ 57	0.120	0.966
47∧51∧55 ⇒ 57	0.122	0.965

Nom des items flous :
12 :[VIS0.6, brillance moyenne], 18 :[VIS0.8, brillant], 21 :[IR1.6, sombre]
31 :[WV6.2, C3(48.667)], 35 :[WV7.3, C3(55.399)], 39 :[IR8.7, C3(63.396)]
47 :[IR10.8, C3(61.174)], 51 :[IR12.0, C3(58.792)], 55 :[IR13.4, C3(52.706)]
57 :[Précipitations, faible]

3.3 Validation du modèle

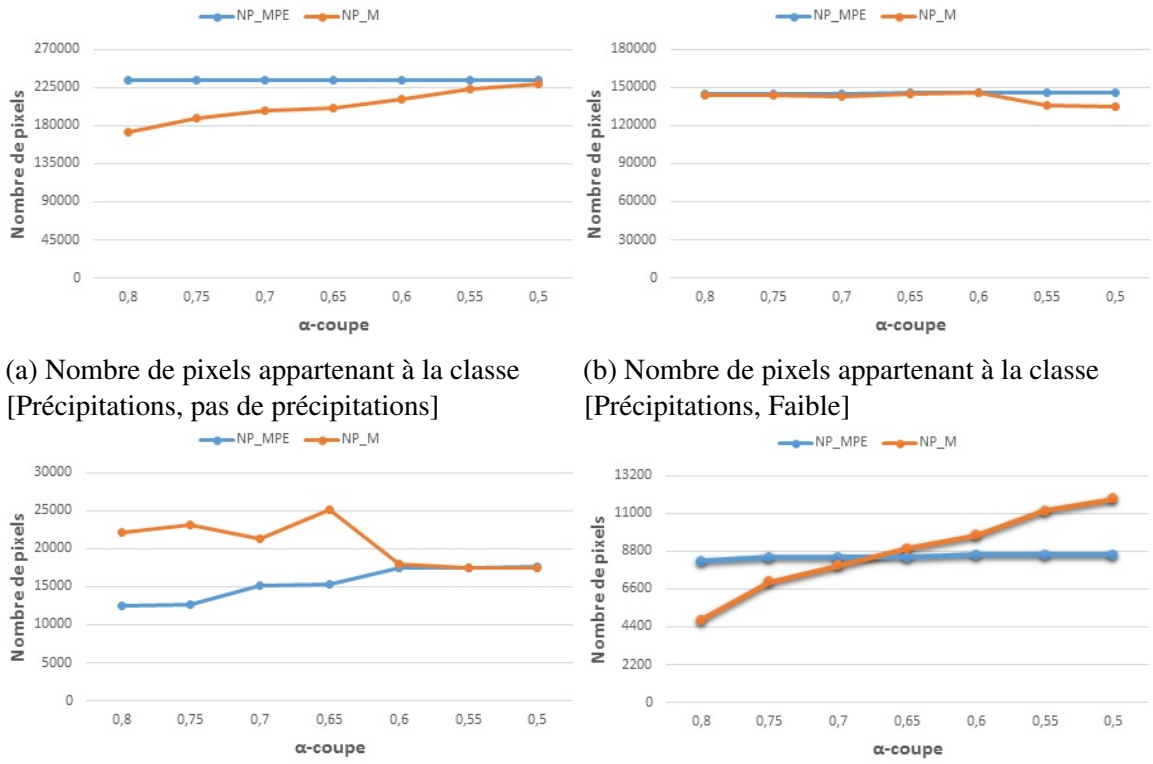
Pour évaluer les performances de notre méthode d'estimation des précipitations, nous comparons les résultats obtenus en appliquant le modèle construit sur la base de données *ValidDB* avec ceux obtenus par le produit MPE sur la même base de données *ValidDB*. Le produit MPE est choisi pour comparaison, car il est considéré comme une référence dans l'estimation de précipitations. Pour chaque item flou (classe floue) de l'attribut précipitations et on fait varier la valeur α-coupe de 0.80 à 0.50 d'une valeur décroissante de 0.05, nous calculons le nombre de pixels estimés par notre méthode (NP_M), nombre de pixels estimés par le produit MPE (NP_MPE) et le nombre de pixels valides obtenus par notre méthode (NP_V), seuls les pixels ayant un degré d'appartenance supérieur à α-coupe étant pris en compte. Pour chaque transaction de *ValidDB*, un pixel valide est un pixel qui a la même

classe floue que le pixel de l'image MPE (c'est-à-dire, une bonne classification). Le nombre de pixels NP_M, NP_MPE et NP_V calculés pour différentes valeurs de α-coupe sont présentés par le tableau 4.7.

TABLE 4.7 Nombre de pixels de chaque item flou de l'attribut précipitations pour chaque α-coupe

α-coupe	Nombre de pixels	[Précipitations, pas de précipitations]	[Précipitations, faible]	[Précipitations, modérée]	[Précipitations, forte]
0.80	NP_MPE	233705	145017	12455	8231
	NP_M	172635	144189	22187	4785
	NP_V	130125	101096	3006	3019
0.75	NP_MPE	233705	145017	12713	8429
	NP_M	189056	143734	23154	7006
	NP_V	141997	107691	4040	4448
0.70	NP_MPE	233705	145017	15145	8429
	NP_M	196901	142596	21265	7973
	NP_V	147502	111044	4783	5136
0.65	NP_MPE	233705	145645	15355	8429
	NP_M	201175	144545	25086	8932
	NP_V	150413	114186	6498	5517
0.60	NP_MPE	233705	145645	17399	8577
	NP_M	211581	145344	17998	9714
	NP_V	157322	120121	7405	5915
0.55	NP_MPE	233705	145645	17399	8577
	NP_M	223410	135800	17396	11145
	NP_V	165494	121336	7634	6215
0.50	NP_MPE	233705	146134	17648	8577
	NP_M	228945	134655	17426	11850
	NP_V	168754	124013	8531	6490

Les courbes de la figure 4.7 reflètent une comparaison entre le nombre de pixels pour chaque classe floue de l'attribut précipitations (pas de précipitations, faible, modérée, forte) pour des valeurs différentes de α-coupe par notre méthode et celui obtenu par le produit MPE.



(a) Nombre de pixels appartenant à la classe [Précipitations, pas de précipitations]

(b) Nombre de pixels appartenant à la classe [Précipitations, Faible]

(c) Nombre de pixels appartenant à la classe [Précipitations, modérée]

(d) Nombre de pixels appartenant à la classe [Précipitations, forte]

FIGURE 4.7 Nombre de pixels appartenant à chaque classe flou de l'attribut précipitations pour chaque α-coupe

Dans notre modèle, le nombre de pixels dans la figure 4.7a varie selon la valeur de α-coupe, nous remarquons que plus que le α-coupe diminue plus on aura une augmentation de nombre de pixels. Par contre le nombre de pixel non précipitant est fixe dans le produit MPE (233705, les pixels ayant une valeur = 0). Notre modèle converge vers le produit MPE lorsque α-coupe tend vers 0.55 et 0.50.

Pour le nombre de pixels dans la figure 4.7b (faible précipitation), il y a une juxtaposition totale entre le produit MPE et notre technique lorsque α-coupe $\in [0.80.6]$, une légère divergence lorsque α-coupe tend vers 0.55 et 0.50, cela s'explique par le fait que les pixels classifiés comme précipitation faible pour α-coupe ≥ 0.60 sont re-classifiés comme pas de précipitation lorsque α-coupe ≤ 0.55. On peut l'expliquer aussi par le fait que notre méthode sous-estime les pixels qui ont une faible précipitation et surestime ceux qui ont une forte précipitation lorsque α-coupe ≤ 0.55.

Dans la figure 4.7c, notre méthode dépasse le produit MPE sensiblement pour le nombre de pixels qui ont une précipitation modérée pour les valeurs de α-coupe ≥ 0.65. Cela

s'explique par le fait que notre méthode sous-estime les pixels non précipitants et ceux à forte précipitation. Pour α-coupe ≤ 0.60 il y a une juxtaposition totale d'estimation.

Dans la figure 4.7d, notre méthode et le produit MPE se convergent quand α-coupe ≥ 0.70, et cela expliqué par le fait que notre méthode surestime les pixels qui ont une précipitation modérée. Dans le cas contraire, notre méthode et le produit MPE se divergent sous contexte que notre méthode sous-estime les pixels qui ont des faibles précipitations.

Afin d'évaluer la performance de la méthode proposée par rapport au produit MPE les deux métriques suivantes sont utilisées :

1. La racine carrée de l'erreur quadratique moyenne (RMSE), elle est définit par :

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(NP\_M_i - NP\_MPE_i)^2}{n}} \tag{4.3}$$

2. Le taux moyen de validation (AVR), qui est calculé comme suit :

$$AVR = \frac{\sum_{i=1}^{n} NP\_V_i}{\sum_{i=1}^{n} NP\_M_i} \tag{4.4}$$

Pour chaque α-coupe, nous recherchons celui qui maximise le taux de validation et en même temps, celui qui minimise l'erreur d'estimation. Pour cela nous avons calculé pour chaque α-coupe le taux moyen de validation (AVR) et la racine carrée de l'erreur quadratique moyenne (RMSE). Les résultats obtenus sont présentés dans le tableau 4.8.

TABLE 4.8 Le taux moyen de validation et la racine carrée de l'erreur quadratique moyenne pour chaque α-coupe

α-coupe	RMSE	AVR
0.80	30971.02	0.69
0.75	22946.77	0.71
0.70	18695.30	0.72
0.65	16265.00	0.73
0.60	11062.00	0.75
0.55	5147.50	0.77
0.50	6426.25	0.78

Les deux histogrammes (voir figure 4.8, figure 4.9) suivants montrent bien la racine carrée de l'erreur quadratique moyenne et le taux moyen de validation pour chaque α-coupe.

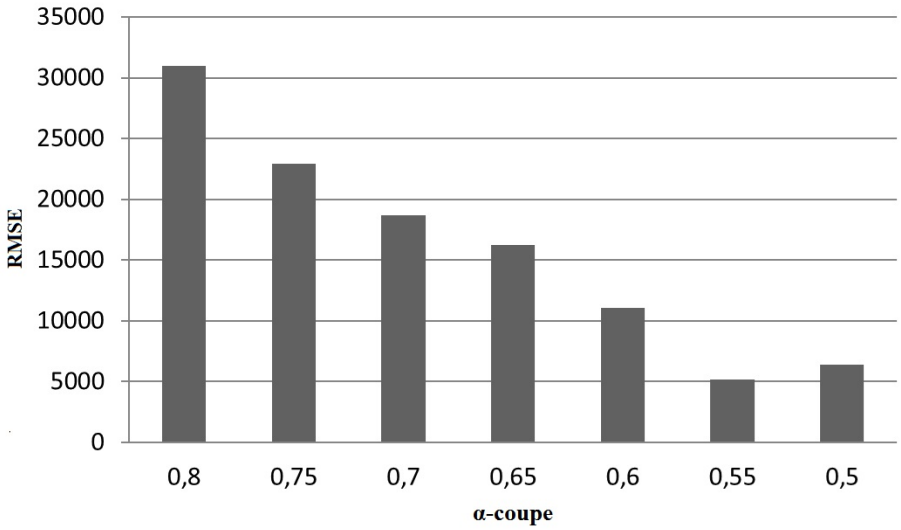


FIGURE 4.8 La racine carrée de l'erreur quadratique moyenne pour chaque α-coupe

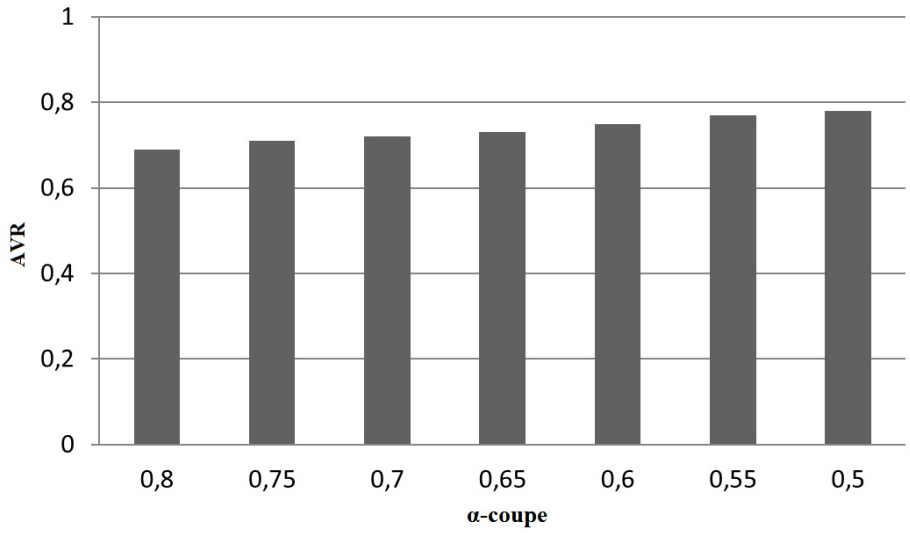


FIGURE 4.9 Le taux moyen de validation pour chaque α-coupe

Les résultats obtenus par notre méthode sont intéressants, comme montre la figure 4.9 et en variant la valeur de α-coupe de 0.80 à 0.50 par un pas décréments de 0.05, le taux moyen de validation est supérieur à 0.69. De plus, comme le montre la figure 4.8, pour les deux (2) valeurs de α-coupe 0.50 et 0.55, la racine carrée de l'erreur quadratique moyenne correspond à une valeur faible par rapport aux autres valeurs de α-coupe. La meilleure valeur

qui permet à notre méthode d'obtenir un bon modèle est α-coupe = 0.55, cette valeur nous permet de faire une bonne estimation tout en minimisant l'erreur d'estimation (5147.5) et en maximisant le taux de validation (0.77).

4 Conclusion

Dans ce chapitre, nous avons élaboré une méthode basée sur les règles d'association floues afin d'estimer les précipitations à partir des images multi-spectrales du satellite MSG, nous avons utilisé l'information infrarouge et visible de ce dernier afin d'extraire des corrélations entre les images des canaux MSG sous forme de (si condition alors conclusion), nous nous sommes intéressés aux règles d'association floues qui ont un seul attribut comme conclusion, il s'agit de l'attribut "précipitation". Une évaluation des règles d'association floues générées et expérimentées sur des images MSG du Nord-Est du notre pays, les résultats obtenus sont très satisfaisants, cela apparaît dans les deux métriques de comparaisons RMSE et AVR par rapport au produit MPE, nous avons obtenu un taux moyen de validation variant entre 0.69 et 0.79 avec une erreur moyenne d'estimation faible.

Conclusion générale

Dans cette thèse nous avons traité le problème d'extraction de connaissances à partir de grandes quantités de données multi-spectrales du satellite MSG. Premièrement, nous avons présenté ce satellite météorologique ainsi que les caractéristiques de ses images fournies toutes les 15 minutes dans 12 canaux spectraux différents. Ces images volumineuses accumulées aux files du temps permettent de recueillir une grande diversité d'information sur la planète et de suivre le développement détaillé des phénomènes météorologiques. Parmi ces phénomènes nous sommes intéressés aux précipitations et ses estimations parce que c'est l'un des objectifs des recherches importantes qui confrontent les météorologues, les hydrologues, les géologues et les agriculteurs... etc. Ensuite nous avons présenté un état de l'art des différentes méthodes d'estimation de précipitations à partir des images satellitaires qui ont été développées en particulier les images MSG. Ces méthodes ont donné des résultats satisfaisants, néanmoins, elles se limitent qu'à l'exploitation des images de quelques canaux et elles classifient les pixels de ces images d'une manière classique, où un pixel est considéré comme 100% précipitant ou 0% (non précipitant). En plus, elles ne prennent pas en compte les corrélations spectrales qui peuvent exister entre les pixels de l'image elle-même, ou bien entre les pixels des différentes images MSG. Pour cela nous avons pensé à l'utilisation des techniques de Datamining telles que les règles d'association floues par ce qu'elle nous permettent d'extraire les corrélations cachées et les informations utiles (les connaissances) dans une vaste collection d'images.

Ensuite, nous avons présenté une nouvelle méthode pour estimer les précipitations à partir des images de 11 canaux du satellite MSG et construit un modèle sous la forme de règles d'association floues. Chaque règle est sous la forme de : si (condition) alors (conclusion), où la condition est une combinaison des différentes classes floues des images MSG, et la conclusion contient une seule classe floue qui représente l'intensité de précipitations : pas de précipitations, faible, modérée et forte. L'importance de la méthode développée vient du fait qu'elle est capable d'exploiter une grande quantité de données diverses provenant de 11 canaux du satellite MSG. En plus l'introduction de la notion du flou nous a permet de traiter ces données incertaines de manière flexible. De plus, nous avons appliqué la méthode

proposée que dans le jour mais en peut l'appliquer aussi dans la nuit en utilisons uniquement les données des canaux infrarouges.

Pour valider notre méthode, des expérimentations pour estimer les précipitations ont été effectuées sur le Nord-Est de l'Algérie, une étude comparative entre la méthode élaborée et les données obtenues par le produit MPE de l'Organisation européenne pour l'exploitation de la météorologie Satellites, les résultats de l'étude montrent la performance de notre proposition qui nous a donné des résultats intéressants et satisfaisants.

Dans notre étude, les données du canal HRV ne sont pas prises en compte pour estimer les précipitations. Par conséquent, à l'avenir, nous prendrons en considération ces données en utilisant par exemple des méthodes de redimensionnement d'image pour qu'elle soit adéquate avec les autres images des canaux MSG. De plus l'ajout d'autres données météorologiques pertinentes telles que la pression et la vitesse du vent et utiliser les techniques d'apprentissage profond pour une meilleure estimation des précipitations. En plus, appliquer la méthode proposée pour détecter d'autres phénomènes météorologiques tels que les tempêtes de sable et suivre la progression de la désertification, ainsi que pour l'exploration des ressources non renouvelables (minéraux, pétrole, gaz), la surveillance de la végétation, la détection et le suivi des feux de forêt… etc.

Bibliographie

[1] https ://www.eumetsat.int. URL https://www.eumetsat.int/website/home/index.html.

[2] R. F. Adler and A. J. Negri. A satellite infrared technique to estimate tropical convective and stratiform rainfall. *Journal of Applied Meteorology*, 27(1) :30–51, 1988.

[3] R. F. Adler, A. J. Negri, P. R. Keehn, and I. M. Hakkarinen. Estimation of monthly rainfall over japan and surrounding waters from a combination of low-orbit microwave and geosynchronous ir data. *Journal of Applied Meteorology*, 32(2) :335–356, 1993.

[4] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. In *Acm sigmod record*, volume 22, pages 207–216. ACM, 1993.

[5] R. Agrawal, R. Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499, 1994.

[6] D. M. A. Aminou, A. Ottenbacher, B. Jacquet, and A. Kassighian. Meteosat second generation : on-ground calibration, characterization, and sensitivity analysis of the seviri imaging radiometer. In *Earth Observing Systems Iv*, volume 3750, pages 419–430. International Society for Optics and Photonics, 1999.

[7] R. Amorati, P. Alberoni, V. Levizzani, and S. Nanni. Ir-based satellite and radar rainfall estimates of convective storms over northern italy. *Meteorological Applications*, 7(1) : 1–18, 2000.

[8] M. L. Anne, M. T. Maguelonne, M. A. M. Rachid, M. B. Tassadit, M. E. M. Sami, and M. K. Mamadou. Fouille de données : Règles séquentielles. 2008.

[9] P. A. Arkin and B. N. Meisner. The relationship between large-scale convective rainfall and cold cloud over the western hemisphere during 1982-84. *Monthly Weather Review*, 115(1) :51–74, 1987.

[10] Y. Arnaud. Caractérisation des nuages précipitants en fonction de leur structure spatiale et de leur évolution temporelle en milieu sahélien à partir d'images meteosat. 1992.

[11] M. B. Ba and A. Gruber. Goes multispectral rainfall algorithm (gmsra). *Journal of Applied Meteorology*, 40(8) :1500–1514, 2001.

[12] M. B. Ba and S. E. Nicholson. Analysis of convective activity and its relationship to the rainfall over the rift valley lakes of east africa during 1983–90 using the meteosat infrared channel. *Journal of Applied Meteorology*, 37(10) :1250–1264, 1998.

[13] S. S. Baboo and M. R. Devi. Geometric correction in recent high resolution satellite imagery : a case study in coimbatore, tamil nadu. *International Journal of Computer Applications*, 14(1) :32–37, 2011.

[14] S. Bandyopadhyay. Satellite image classification using genetically guided fuzzy clustering with spatial information. *International Journal of Remote Sensing*, 26(3) : 579–593, 2005.

[15] A. Bellon, S. Lovejoy, and G. Austin. Combining satellite and radar data for the short-range forecasting of precipitation. *Monthly Weather Review*, 108(10) :1554–1566, 1980.

[16] J. Bendix. Adjustment of the convective-stratiform technique (cst) to estimate 1991/93 el nino rainfall distribution in ecuador and peru by means of meteosat-3 ir data. *International Journal of Remote Sensing*, 18(6) :1387–1394, 1997.

[17] J. Bendix. Precipitation dynamics in ecuador and northern peru during the 1991/92 el nino : a remote sensing perspective. *International Journal of Remote Sensing*, 21(3) : 533–548, 2000.

[18] N. Bensafi, M. Lazri, and S. Ameur. Novel wknn-based technique to improve instantaneous rainfall estimation over the north of algeria using the multispectral msg seviri imagery. *Journal of Atmospheric and Solar-Terrestrial Physics*, 183 :110–119, 2019.

[19] W. Berg and R. Chase. Determination of mean rainfall from the special sensor microwave/imager (ssm/i) using a mixed lognormal distribution. *Journal of Atmospheric and Oceanic Technology*, 9(2) :129–141, 1992.

[20] J. C. Bergès, I. Jobart, F. Chopin, and R. Roca. Epsat-sg : a satellite method for precipitation estimation. 2009.

[21] J. C. Bezdek. Cluster validity with fuzzy sets. 1973.

[22] J. C. Bezdek. Objective function clustering. In *Pattern recognition with fuzzy objective function algorithms*, pages 43–93. Springer, 1981.

[23] B. Bouaita, A. Moussaoui, and N. E. I. Bachari. Rainfall estimation from msg images using fuzzy association rules. *Journal of Intelligent & Fuzzy Systems*, 37(1) :1357–1369, 2019.

[24] B. Bouchon-Meunier and C. Marsala. *Logique floue : principes, aide à la décision*. 2003.

[25] M. Bouguessa, S. Wang, and H. Sun. An objective approach to cluster validation. *Pattern Recognition Letters*, 27(13) :1419–1430, 2006.

[26] B. Brühl, M. Hülsmann, D. Borscheid, C. M. Friedrich, and D. Reith. A sales forecast model for the german automobile market based on time series analysis and data mining methods. In *Industrial Conference on Data Mining*, pages 146–160. Springer, 2009.

[27] J. B. Campbell and R. H. Wynne. *Introduction to remote sensing*. Guilford Press, 2011.

[28] M. Carn and J.-P. Lahuec. Estimation des précipitations au niger au cours de la saison des pluies 1986 à partir de l'imagerie infrarouge météosat : bilan et critique des méthodes utilisées. *Veille Climatique Satellitaire*, (17) :40–48, 1987.

[29] C. CHEIKH. Développement d'un logiciel d'analyse simultanée des images meteosat et images noaa à partir de la station laar, 2008.

[30] M. Cheng and R. Brown. Delineation of precipitation areas by correlation of meteosat visible and infrared data with radar data. *Monthly weather review*, 123(9) :2743–2757, 1995.

[31] K.-S. Chuang, H.-L. Tzeng, S. Chen, J. Wu, and T.-J. Chen. Fuzzy c-means clustering with spatial information for image segmentation. *computerized medical imaging and graphics*, 30(1) :9–15, 2006.

[32] R. G. Congalton. 21 how to assess the accuracy of maps generated from remotely sensed data. *Manual of Geospatial Science and Technology*, page 403, 2010.

[33] T. Daurel. *Représentations Condensées d'Ensembles de Règles d'Association*. PhD thesis, Ph. D. thesis, L'Institut National des Sciences Appliquées de Lyon, 2003.

[34] M. De Cock, C. Cornelis, and E. E. Kerre. Elicitation of fuzzy association rules from positive and negative examples. *Fuzzy Sets and Systems*, 149(1) :73–85, 2005.

[35] I. Decoster, N. Clerbaux, E. Baudrez, S. Dewitte, A. Ipe, S. Nevens, A. Blazquez, and J. Cornelis. Spectral aging model applied to meteosat first generation visible band. *Remote Sensing*, 6(3) :2534–2571, 2014.

[36] M. Delgado, N. Marín, D. Sánchez, and M.-A. Vila. Fuzzy association rules : general model and applications. *IEEE transactions on Fuzzy Systems*, 11(2) :214–225, 2003.

[37] Q. Ding, Q. Ding, and W. Perrizo. Association rule mining on remotely sensed images using p-trees. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 66–79. Springer, 2002.

[38] D. Dubois, E. Hüllermeier, and H. Prade. A systematic approach to the assessment of fuzzy association rules. *Data Mining and Knowledge Discovery*, 13(2) :167–192, 2006.

[39] J. C. Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. 1973.

[40] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, et al. Knowledge discovery and data mining : Towards a unifying framework. In *KDD*, volume 96, pages 82–88, 1996.

[41] H. Feidas and A. Giannakos. Identifying precipitating clouds in greece using multispectral infrared meteosat second generation satellite data. *Theoretical and applied climatology*, 104(1-2) :25–42, 2011.

[42] H. Feidas and A. Giannakos. Classifying convective and stratiform rain using multispectral infrared meteosat second generation satellite data. *Theoretical and applied climatology*, 108(3-4) :613–630, 2012.

[43] R. R. Ferraro and G. F. Marks. The development of ssm/i rain-rate retrieval algorithms using ground-based radar measurements. *Journal of Atmospheric and Oceanic Technology*, 12(4) :755–770, 1995.

[44] R. R. Ferraro, N. C. Grody, and G. F. Marks. Effects of surface conditions on rain identification using the dmsp-ssm/i. *Remote Sensing Reviews*, 11(1-4) :195–209, 1994.

[45] C. Fiot, G. Dray, A. Laurent, and M. Teisseire. A la recherche des motifs séquentiels flous. *Actes des Rencontres Francophones sur la Logique Floue et ses Applications (LFA 04)*, pages 131–138, 2004.

[46] M. A. Friedl and C. E. Brodley. Decision tree classification of land cover from remotely sensed data. *Remote sensing of environment*, 61(3) :399–409, 1997.

[47] J. F. Gamache and R. A. Houze Jr. Water budget of a mesoscale convective system in the tropics. *Journal of the atmospheric sciences*, 40(7) :1835–1850, 1983.

[48] I. Gath and A. B. Geva. Unsupervised optimal fuzzy clustering. *IEEE Transactions on pattern analysis and machine intelligence*, 11(7) :773–780, 1989.

[49] R. Geetharamani, P. Revathy, and S. G. Jacob. Prediction of users webpage access behaviour using association rule mining. *Sadhana*, 40(8) :2353–2365, 2015.

[50] M. Ghose, R. Pradhan, and S. S. Ghose. Decision tree classification of remotely sensed satellite data using spectral separability matrix. *International Journal of Advanced Computer Science and Applications*, 1(5), 2010.

[51] A. Giannakos and H. Feidas. Classification of convective and stratiform rain based on the spectral and textural features of meteosat second generation infrared data. *Theoretical and applied climatology*, 113(3-4) :495–510, 2013.

[52] P. J. Gibson and C. H. Power. Introductory remote sensing : Digital image processing and applications. 2000.

[53] C. G. Griffith, W. L. Woodley, P. G. Grube, D. W. Martin, J. Stout, and D. N. Sikdar. Rain estimation from geosynchronous satellite imagery—visible and infrared studies. *Monthly Weather Review*, 106(8) :1153–1171, 1978.

[54] N. C. Grody. Classification of snow cover and precipitation using the special sensor microwave imager. *Journal of Geophysical Research : Atmospheres*, 96(D4) :7423–7435, 1991.

[55] B. Guillot. L'utilisation des satellites météorologiques pour l'estimation de la pluie en zone sahélo-soudanienne au centre de météorologie spatiale de lannion. 1990.

[56] B. Guillot. Projet epsat : Estimation des pluies par satellite. 1991.

[57] N. Gupta, N. Mangal, K. Tiwari, and P. Mitra. Mining quantitative association rules in protein sequences. In *Data Mining*, pages 273–281. Springer, 2006.

[58] J. Han, K. Koperski, and N. Stefanovic. Geominer : a system prototype for spatial data mining. *AcM sIGMoD Record*, 26(2) :553–556, 1997.

[59] J. Han, J. Pei, Y. Yin, and R. Mao. Mining frequent patterns without candidate generation : A frequent-pattern tree approach. *Data mining and knowledge discovery*, 8(1) :53–87, 2004.

[60] J. E. Harries, J. Russell, J. Hanafin, H. Brindley, J. Futyan, J. Rufus, S. Kellock, G. Matthews, R. Wrigley, A. Last, et al. The geostationary earth radiation budget project. *Bulletin of the American Meteorological Society*, 86(7) :945–960, 2005.

[61] K. Hatonen, M. Klemettinen, H. Mannila, P. Ronkainen, and H. Toivonen. Knowledge discovery from telecommunication network alarm databases. In *Proceedings of the Twelfth International Conference on Data Engineering*, pages 115–122. IEEE, 1996.

[62] P. D. Heermann and N. Khazenie. Classification of multispectral remote sensing data using a back-propagation neural network. *IEEE Transactions on geoscience and remote sensing*, 30(1) :81–88, 1992.

[63] T. Heinemann, A. Latanzio, and F. Roveda. The eumetsat multi-sensor precipitation estimate (mpe). In *Second International Precipitation Working group (IPWG) Meeting*, pages 23–27, 2002.

[64] J. Hipp, U. Güntzer, and G. Nakhaeizadeh. Algorithms for association rule mining—a general survey and comparison. *ACM sigkdd explorations newsletter*, 2(1) :58–64, 2000.

[65] T.-P. Hong, K.-Y. Lin, and S.-L. Wang. Fuzzy data mining for interesting generalized association rules. *Fuzzy sets and systems*, 138(2) :255–269, 2003.

[66] T.-P. Hong, C.-S. Kuo, and S.-L. Wang. A fuzzy aprioritid mining algorithm with reduced computational time. *Applied Soft Computing*, 5(1) :1–10, 2004.

[67] R. A. Houze Jr and E. N. Rappaport. Air motions and precipitation structure of an early summer squall line over the eastern tropical atlantic. *Journal of the Atmospheric Sciences*, 41(4) :553–574, 1984.

[68] J. R. Jensen. *Introductory digital image processing : a remote sensing perspective*. Prentice Hall Press, 2015.

[69] I. Jobard. Status of satellite retrieval of rainfall at different scales using multi-source data. In *MEGHA-TROPIQUES 2nd Scientific Workshop 2-6 July*, volume 28, 2001.

[70] I. Jobard and M. Desbois. Satellite estimation of the tropical precipitation using the meteostat and ssm/i data. *Atmospheric Research*, 34(1-4) :285–298, 1994.

[71] M. Kantardzic. *Data mining : concepts, models, methods, and algorithms*. John Wiley & Sons, 2011.

[72] M. Kerrache and J. Schmetz. A precipitation index from the esoc climatological data set. *ESA Journal*, 12(3) :379–383, 1988.

[73] C. Kidd, D. R. Kniveton, M. C. Todd, and T. J. Bellerby. Satellite rainfall estimation using combined passive microwave and infrared algorithms. *Journal of Hydrometeorology*, 4(6) :1088–1104, 2003.

[74] P. W. King, W. D. Hogg, and P. A. Arkin. The role of visible data in improving satellite rain-rate estimates. *Journal of Applied Meteorology*, 34(7) :1608–1621, 1995.

[75] M. Klemettinen, H. Mannila, and H. Toivonen. A data mining methodology and its application to semi-automatic knowledge acquisition. In *Database and Expert Systems Applications. 8th International Conference, DEXA'97. Proceedings*, pages 670–677. IEEE, 1997.

[76] C. M. Kuok, A. Fu, and M. H. Wong. Mining fuzzy association rules in databases. *ACM Sigmod Record*, 27(1) :41–46, 1998.

[77] M. Lazri, S. Ameur, J. M. Brucker, J. Testud, B. Hamadache, S. Hameg, F. Ouallouche, and Y. Mohia. Identification of raining clouds using a method based on optical and microphysical cloud properties from meteosat second generation daytime and nighttime data. *Applied Water Science*, 3(1) :1–11, 2013.

[78] K.-S. Leung, K.-C. Wong, T.-M. Chan, M.-H. Wong, K.-H. Lee, C.-K. Lau, and S. K. Tsui. Discovering protein–dna binding sequence patterns using association rule mining. *Nucleic acids research*, 38(19) :6324–6337, 2010.

[79] V. Levizzani, F. Porcù, and F. Prodiº. Operational rainfall estimation using meteosat infrared imagery : An application in italy's arno. *ESA Journal*, 14, 1990.

[80] V. Levizzani, J. Schmetz, H. Lutz, J. Kerkmann, P. Alberoni, and M. Cervino. Precipitation estimations from geostationary orbit and prospects for meteosat second generation. *Meteorological Applications*, 8(1) :23–41, 2001.

[81] Y. Li and B. Cheng. An improved k-nearest neighbor algorithm and its application to high resolution remote sensing image classification. In *2009 17th International Conference on Geoinformatics*, pages 1–4. Ieee, 2009.

[82] K.-C. Lu, D.-L. Yang, and M.-C. Hung. Decision trees based image data mining and its application on image segmentation. In *Proceedings of International Conference on Chinese Language Computing*, pages 81–86, 2002.

[83] L. Ma, M. M. Crawford, and J. Tian. Local manifold learning-based k-nearest-neighbor for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 48(11) :4099–4109, 2010.

[84] J. MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[85] F. S. Marzano, M. Palmacci, D. Cimini, G. Giuliani, and F. J. Turk. Multivariate statistical integration of satellite infrared and microwave radiometric measurements for rainfall retrieval at the geostationary scale. *IEEE transactions on Geoscience and remote sensing*, 42(5) :1018–1032, 2004.

[86] P. Mather. Computer processing of remotely sensed images,‖ john wiley, 2004.

[87] T. McCormick, C. Rudin, and D. Madigan. A hierarchical model for association rule mining of sequential events : An approach to automated medical symptom prediction. 2011.

[88] G. Menz. Regionalization of precipitation models in eastafrica using meteosat data. *International Journal of Climatology : A Journal of the Royal Meteorological Society*, 17(10) :1011–1027, 1997.

[89] N. Mishra and S. Silakari. Image mining in the context of content based image retrieval : a perspective. *International Journal of Computer Science Issues (IJCSI)*, 9 (4) :69, 2012.

[90] B. Mobasher, H. Dai, T. Luo, and M. Nakagawa. Effective personalization based on association rule discovery from web usage data. In *Proceedings of the 3rd international workshop on Web information and data management*, pages 9–15, 2001.

[91] E. Mwamikazi, P. Fournier-Viger, C. Moghrabi, and R. Baudouin. A dynamic questionnaire to further reduce questions in learning style assessment. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 224–235. Springer, 2014.

[92] J. Nahar, T. Imam, K. S. Tickle, and Y.-P. P. Chen. Association rule mining to detect factors which contribute to heart disease in males and females. *Expert Systems with Applications*, 40(4) :1086–1093, 2013.

[93] A. Natarajan, S. Subramanian, and K. Premalatha. A comparative study of cuckoo search and bat algorithm for bloom filter optimisation in spam filtering. *International Journal of Bio-Inspired Computation*, 4(2) :89–99, 2012.

[94] A. J. Negri and R. F. Adler. An intercomparlson of three satellite infrared rainfall techniques over japan and surrounding waters. *Journal of Applied Meteorology*, 32 (2) :357–373, 1993.

[95] A. J. Negri, R. F. Adler, and P. J. Wetzel. Rain estimation from satellites : An examination of the griffith-woodley technique. *Journal of climate and applied meteorology*, 23(1) :102–116, 1984.

[96] D. Nuzillard and C. Lazar. Partitional clustering techniques for multi-spectral image segmentation. *JCP*, 2(10) :1–8, 2007.

[97] Z. P. Ogihara, M. Zaki, S. Parthasarathy, M. Ogihara, and W. Li. New algorithms for fast discovery of association rules. In *In 3rd Intl. Conf. on Knowledge Discovery and Data Mining*. Citeseer, 1997.

[98] F. Ouallouche and S. Ameur. Rainfall detection over northern algeria by combining msg and trmm data. *Applied Water Science*, 6(1) :1–10, 2016.

[99] F. Ouallouche, M. Lazri, and S. Ameur. Improvement of rainfall estimation from msg data using random forests classification and regression. *Atmospheric Research*, 211 : 62–72, 2018.

[100] N. R. Pal and J. C. Bezdek. On cluster validity for the fuzzy c-means model. *IEEE Transactions on Fuzzy systems*, 3(3) :370–379, 1995.

[101] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Pruning closed itemset lattices for association rules. 1998.

[102] R. Pierrard, J.-P. Poli, and C. Hudelot. A fuzzy close algorithm for mining fuzzy association rules. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 88–99. Springer, 2018.

[103] T. Piton, J. Blanchard, F. Guillet, and H. Briand. Une méthodologie de recommandations produits fondée sur l'actionnabilité et l'intérêt économique des clients. 2011.

[104] A. Pompei, M. Marrocu, P. Boi, and G. Dalu. Validation of retrieval algorithms for the infrared remote sensing of precipitation with the sardinian rain gauge network data. *Il Nuovo Cimento C*, 18(5) :483–496, 1995.

[105] A. Ramezankhani, O. Pournik, J. Shahrabi, F. Azizi, and F. Hadaegh. An application of association rule mining to extract risk pattern for type 2 diabetes using tehran lipid and glucose study database. *International journal of endocrinology and metabolism*, 13(2), 2015.

[106] C. Reudenbach, G. Heinemann, E. Heuel, J. Bendix, and M. Winiger. Investigation of summertime convective rainfall in western europe based on a synergy of remote sensing data and numerical models. *Meteorology and Atmospheric Physics*, 76(1-2) : 23–41, 2001.

[107] E. Ricciardelli, F. Romano, and V. Cuomo. Physical and statistical approaches for cloud identification using meteosat second generation-spinning enhanced visible and infrared imager data. *Remote sensing of environment*, 112(6) :2741–2760, 2008.

[108] J. A. Richards and J. Richards. *Remote sensing digital image analysis*, volume 3. Springer, 1999.

[109] R. Roebeling and I. Holleman. Validation of rain rate retrievals from seviri using weather radar observations. In *EUMETSAT Meteorological Satellite Conference*, pages 8–11. Citeseer, 2008.

[110] C. Romero and S. Ventura. Educational data mining : a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6) :601–618, 2010.

[111] A. Savasere, E. R. Omiecinski, and S. B. Navathe. An efficient algorithm for mining association rules in large databases. Technical report, Georgia Institute of Technology, 1995.

[112] J. Schmetz, P. Pili, S. Tjemkes, D. Just, J. Kerkmann, S. Rota, and A. Ratier. An introduction to meteosat second generation (msg). *Bulletin of the American Meteorological Society*, 83(7) :977–992, 2002.

[113] F. Seddi and Z. Ameur. Estimation des precipitations en utilisant l'information multispectrale du satellite meteosat. *LARHYSS Journal P-ISSN 1112-3680/E-ISSN 2602-7828*, (9), 2011.

[114] M. Sehad, S. Ameur, J. M. Brucker, and M. Lazri. Msg seviri image segmentation using a method based on spectral, temporal and textural features. *Computers and Software*, page 618, 2014.

[115] G. Serban, I.-G. Czibula, and A. Campan. A programming interface for medical diagnosis prediction. *Studia Universitatis" Babes-Bolyai", Informatica, LI (1), pag*, pages 21–30, 2006.

[116] Y. Slimani. *Extraction et analyse de connaissances à partir du Web*. PhD thesis, 2018.

[117] R. W. Spencer, H. M. Goodman, and R. E. Hood. Precipitation retrieval over land and ocean with the ssm/i : Identification and characteristics of the scattering signal. *Journal of Atmospheric and Oceanic Technology*, 6(2) :254–273, 1989.

[118] B. Subbiah and S. C. Christopher. Image classification through integrated k-means algorithm. *International Journal of Computer Science Issues (IJCSI)*, 9(2) :518, 2012.

[119] M. Szczerba and A. Ciemski. Credit risk handling in telecommunication sector. In *Industrial Conference on Data Mining*, pages 117–130. Springer, 2009.

[120] B. Thies, T. Nauss, and J. Bendix. Delineation of raining from non-raining clouds during nighttime using meteosat-8 data. *Meteorol Appl*, 15 :219–230, 2008.

[121] B. Thies, T. Nauß, and J. Bendix. Discriminating raining from non-raining clouds at mid-latitudes using meteosat second generation daytime data. 2008.

[122] M. Todd, E. Barrett, M. Beaumont, and T. Bellerby. Estimation of daily rainfall over the upper nile river basin using a continuously calibrated satellite infrared technique. *Meteorological Applications*, 6(3) :201–210, 1999.

[123] H. Toivonen et al. Sampling large databases for association rules. In *Vldb*, volume 96, pages 134–145, 1996.

[124] F. Torricella, V. Levizzani, and M. Celano. Applications of a rainfall estimation technique based on mw and ir satellite data : Assessment of reliability of instantaneous rain rate maps in the mediterranean. *Institute of Atmospheric Sciences and Climate, National Research Council*, pages 1–6, 2003.

[125] F. J. TURK, G. D. ROHALY, H. JEFF, E. A. SMITH, F. S. MARZANO, A. MUGNAI, and V. LEVIZZANI. Meteorological applications of precipitation estimation from combined ssm/i, trmm and infrared geostationary. *Microwave Radiometry and Remote Sensing of the Earth's Surface and Atmosphere*, page 353, 2000.

[126] A. Vuckovic, D. Popovic, and V. Radivojevic. Artificial neural network for detecting drowsiness from eeg recordings. In *6th Seminar on Neural Network Applications in Electrical Engineering*, pages 155–158. IEEE, 2002.

[127] H. Wang and B. Fei. A modified fuzzy c-means classification method using a multiscale diffusion filtering scheme. *Medical image analysis*, 13(2) :193–202, 2009.

[128] W. Wang, Y. J. Wang, R. Bañares-Alcántara, Z. Cui, and F. Coenen. Application of classification association rule mining for mammalian mesenchymal stem cell differentiation. In *Industrial Conference on Data Mining*, pages 51–61. Springer, 2009.

[129] E. Wolters, B. Van Den Hurk, and R. Roebeling. Evaluation of rainfall retrievals from seviri reflectances over west africa using trmm-pr and cmorph. *Hydrology and Earth System Sciences*, 15(2) :437–451, 2011.

[130] M.-H. Wong, H.-Y. A. Sze-To, L.-Y. P. Lo, T.-M. C. Chan, and K.-S. Leung. Discovering binding cores in protein-dna binding using association rule mining with statistical measures. *IEEE/ACM transactions on computational biology and bioinformatics*, 12 (1) :142–154, 2014.

[131] H. Wynne, L. Mong, and J. Zhang. Image mining : trends and developments. journal of intelligent information systems. 2002.

[132] Z.-Y. Yin, X. Liu, X. Zhang, and C.-F. Chung. Using a geographic information system to improve special sensor microwave imager precipitation estimates over the tibetan plateau. *Journal of Geophysical Research : Atmospheres*, 109(D3), 2004.

[133] L. A. Zadeh. Fuzzy sets. *Information and control*, 8(3) :338–353, 1965.

[134] M. J. Zaki, J. T. Wang, and H. T. Toivonen. Biokdd 2002 : recent advances in data mining for bioinformatics. *ACM SIGKDD Explorations Newsletter*, 4(2) :112–114, 2002.

Annexe
Généralités sur la logique floue

1 Pourquoi la logique floue

La logique floue présente sur la logique classique (booléenne), la particularité de prendre en compte notre désir d'appréciation des nuances [24]. S'agissant par l'exemple de la taille d'un adulte, supposée comprise entre 1.50 m et 2.20 m (on appelle cet intervalle l'univers du discours), la logique booléenne à deux valeurs, Petit et Grand, va ranger les personnes de taille inférieur à un certain seuil, par exemple 1.72 m, dans l'ensemble petit et les autres dans l'ensemble grand, deux individus de tailles comparables, 1.715 m et 1.725 m seront, de ce fait, dans deux ensembles séparés. En logique floue, les ensembles grand et petit se recouvrent, en sorte qu'un individu de taille 1.72 m, sera grand à 60% et petit à 40%. Donc la logique floue offre des modes de raisonnement approximatifs. Ce dernier est le mode de raisonnement utilisé par les humains.

2 Les ensembles flous

Les ensembles flous sur laquelle s'appuie la logique floue ont été introduits par Lotfi Zadeh en 1965 [133]. Les ensembles flous peuvent être considérés comme une extension des ensembles classiques. Ce sont des ensembles permettant une appartenance partielle ou bien graduelle de leurs éléments. Il s'agit d'un degré d'appartenance qui indique l'emplacement de l'élément dans l'ensemble flou. Ainsi, un élément peut appartenir à plusieurs ensembles et cela est accompli par des degrés d'appartenance. Ces degrés sont des nombres réels dans l'intervalle [0..1]. On peut dire qu'un ensemble classique contient des éléments qui satisfont aux propriétés précises de l'appartenance tandis que l'ensemble flou contient des éléments qui satisfont aux propriétés imprécises de l'appartenance. Chaque ensemble flou est défini par sa fonction d'appartenance.

3 Fonction d'appartenance

En logique booléenne on appartient (100%) ou on n'appartient pas (0%) à un ensemble donné. En logique floue l'appartenance est décrite par leur degré d'appartenance (valeur de vérité) déterminé par la forme de la fonction d'appartenance qui délimite l'ensemble flou considéré. Dans l'exemple de la figure 10, l'univers du discours noté U = [1.50, 2.20]

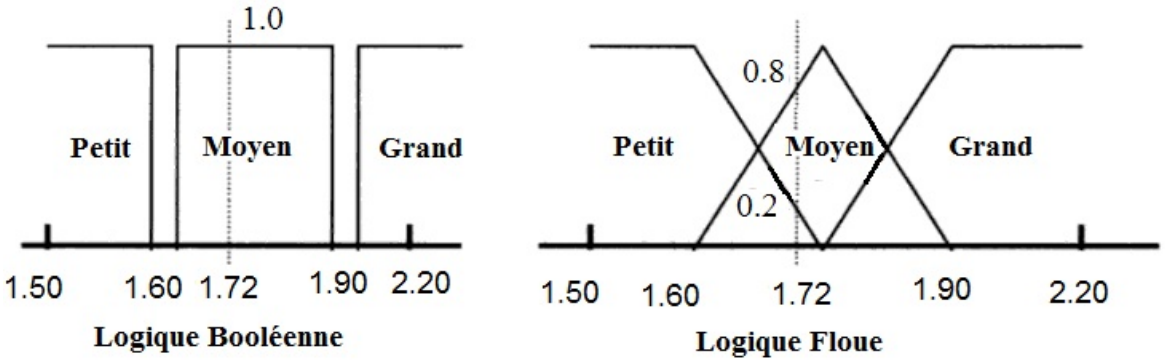


FIGURE 10 Exemple de la différence entre la logique booléenne et la logique floue

est décomposé en trois ensembles flous : grand, moyen, petit, le graphe de la fonction d'appartenance peut prendre différentes formes comme la forme triangulaire (l'ensemble moyen) ou trapézoïdale (ensemble petit et grand). La forme n'est pas exigée, on peut choisir d'autres, dans notre exemple, un individu de taille 1.72 m, en logique booléenne cet individu est considéré comme une personne de taille moyenne à 100%, parce que la fonction caractéristique de l'ensemble donne 1 pour les tailles appartiennent à l'intervalle [1.60, 1.90], mais, en logique flou cette personne de taille 1.72 m appartient à deux ensembles flous, moyen avec un degré d'appartenance 0.80, et petit avec un degré d'appartenance 0.20. Donc une partie floue A de U est caractérisée par une application de U dans [0,1]. Cette application, appelée une fonction d'appartenance et notée μ_A représente le degré de validité de la proposition « x appartient à A » pour chacun des éléments x de U. Si $\mu_A(x) = 1$, l'objet x appartient totalement à A, et si $\mu_A(x) = 0$, il ne lui appartient pas du tout. Pour un élément x donné, la valeur de la fonction d'appartenance $\mu_A(x)$ est appelée degré d'appartenance de l'élément x au sous-ensemble A [134].

$$\mu : x \in U \rightarrow \mu_A(x) \in [0,1]$$

Qui quantifie le degré d'appartenance de chaque élément x de U à A.

4 Caractéristiques d'un ensemble flou

Un ensemble flou est défini par sa fonction d'appartenance, et à partir de cette fonction on peut décrire plusieurs caractéristiques d'un ensemble flou, chaque ensemble flou possède les caractéristiques suivantes [24] :

— **Le support :**
pour un sous-ensemble flou A, nous définissons le support de A noté $Supp(A)$ comme l'ensemble de tous les éléments de U dont le degré d'appartenance à A est supérieures à zéro (non nul), En d'autres termes,

$$Supp(A) = \{x \in U / \mu_A(x) > 0\}$$

— **Le noyau :**
Le noyau d'un ensemble flou A noté $Noy(A)$ comme étant l'ensemble de tous les éléments de U dont le degré d'appartenance à A est égal à 1.

$$Noy(A) = \{x \in U / \mu_A(x) = 1\}$$

— **La hauteur :**
La hauteur d'un sous-ensemble flous A noté $h(A)$ est la valeur maximale atteinte par la fonction d'appartenance.

$$h(A) = \{x \in U / MAX(\mu_A(x))\}$$

— **La cardinalité :**
La cardinalité d'un sous-ensemble flou A noté $(||)$ est la somme des degrés d'appartenance de tous les éléments de U dans A. Elle est définie par :

$$|A| = \{x \in U / \sum \mu_A(x)\}$$

— **La coupe de niveau ou α-coupe :**
Un sous-ensemble floue A de U peut aussi être caractérisée par l'ensemble de ses α-coupes. Une α-coupe (en anglaisα-cut) d'un sous-ensemble floue A est le sous-ensemble net (classique) des éléments ayant un degré d'appartenance supérieur ou égal à α.

$$\alpha - coupe(A) = \{x \in U / \mu_A(x) \geq \alpha\}$$

La figure 11 illustre les principales notions définies ci-dessus.

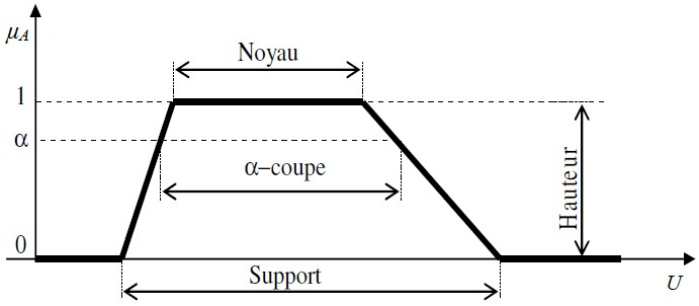


FIGURE 11 Principaux caractéristiques d'un sous ensemble flou

5 Opérations de base sur les ensembles flous

En logique floue, il existe deux opérations de base sur les ensembles flous : l'union, l'intersection [24]. Soit μ_A et μ_B des fonctions d'appartenance qui définissent les ensembles flous A et B, respectivement, sur l'univers U.

— **L'intersection :** l'intersection des deux ensembles flous A et B est un ensemble flou défini par la fonction d'appartenance.
$\forall x \in U \mu_{A\cap B}(x) = \top(\mu_A(x), \mu_B(x))$, Où $(\top)$ est un opérateur t-norme.

— **L'union :** l'union des deux ensembles flous A et B est un ensemble flou défini par la fonction d'appartenance.
$\forall x \in U \mu_{A\cup B}(x) = \perp (\mu_A(x), \mu_B(x))$, Où $(\perp)$ est un opérateur t-conorme.

En fait, il existe de nombreuses possibilités pour représenter les opérateurs t-norme $(\top)$ et t-conorme $(\perp)$. Les fonctions (MIN et MAX) sont dues à Zadeh et sont encore aujourd'hui les plus utilisées. Le tableau suivant récapitule quelques fonctions représentant les opérateurs $(\top)$ et $(\perp)$.

TABLE 9 Définitions des t-normes et t-conormes les plus utilisées

Nom des fonctions	t-norme $(\top)$	t-conorme $(\perp)$
Zadeh	$Min(x,y)$	$Max(x,y)$
Probabiliste	$X * y$	$x+y-x*y$
Lukasiewicz	$Max(x+y-1,0)$	$Min(x+y,1)$