Des données Satellitaires aux Connaissances : Modélisation des Incertitudes, Analyse et Interprétation

Wadii Boulila

Université de la Manouba

Ecole Nationale des Sciences de l'Informatique

Ecole doctorale STICODE

Rapport de synthèse

Présenté en vue de l'obtention d'une

Habilitation Universitaire

de l'Ecole Nationale des Sciences de l'Informatique

(Spécialité Informatique)

par

Wadii Boulila

Maître assistant à ISAMM - Chercheur au laboratoire RIADI-GDL

Docteur en informatique de l'ENSI

Ingénieur en informatique de l'EABA

Des données satellitaires aux connaissances : modélisation des incertitudes, analyse et interprétation

Soutenue le 15/08/2019 devant le jury composé de :

Président : Mme. Henda Ben Ghézala, Professeur, Université de la Manouba, Tunisie

Rapporteurs : Mr. Mohamed Hammami, Maître de Conférences, Université de Sfax, Tunisie

Mr. Christophe Claramunt, Professeur, Ecole Navale, France

Examinateurs : Mr. Imed Riadh Farah, Professeur, Université de la Manouba, Tunisie

Mme. Hajer Baazaoui, Maître de Conférences, Université de la Manouba, Tunisie

Remerciements

Mes sincères remerciements s'adressent tout d'abord à Monsieur Imed Riadh Farah, Professeur à l'Université de la Manouba en Tunisie, qui m'a accueilli dans son équipe de recherche et qui m'a offert diverses opportunités scientifiques qui ont contribuées significativement à mon enrichissement scientifique.

Mes profonds remerciements vont également à Monsieur Amir Hussain, Professeur à l'Université de Stirling en Royaume-Uni, qui m'a impliqué dans plusieurs projets de recherche et qui n'a cessé de m'encouragé.

Je remercie également Mme Henda Ben Ghézala, Professeur à l'Université de la Manouba en Tunisie, pour l'honneur qu'elle m'a accordé en présidant le jury de ma soutenance d'habilitation. Je tiens, aussi, à exprimer ma profonde reconnaissance à Monsieur Mohamed Hammami, Maître de Conférences à l'Université de Sfax en Tunisie, et Monsieur Christophe Claramunt, Professeur à l'Institut de Recherche de l'Ecole Navale en France, qui ont accepté la lourde tâche de rapporter mon habilitation. Mes remerciements vont également à Mme Hajer Baazaoui Zghal, Maître de Conférences à l'Université de la Manouba en Tunisie, pour avoir acceptée de faire partie du jury de la soutenance de mon habilitation.

Ce travail d'habilitation est le fruit de longues années de travail d'équipe et de collaborations scientifiques. A ce titre, je témoigne du mérite les doctorants que j'ai co-encadrés, nommément Ahlem Ferchichi, Muhamed Wael Farouq, Imen Chebbi, Zouhaira Ayadi et Yosra Hajjaji. Cela m'a fait très grand plaisir de participer à l'encadrement de leur thèse et je les remercie pour leur engagement et leur sérieux tout au long de ces années.

Mes sincères salutations vont également à tous mes collègues à l'ISAMM et au laboratoire RIADI en Tunisie, au CCSE en Arabie Saoudite, et au CogBID en Royaume-Uni.

Enfin, je remercie vivement ma femme, mes adorables enfants pour m'encourager et tolérer mes absences continues et répétitives. J'espère que ce mémoire soit une récompense de leurs sacrifices. Je n'oublie jamais à remercier et saluer fortement les membres de ma famille : ma mère, mon père, mes frères et ma sœur, ma belle-mère et mes belles-sœurs pour leur soutien moral constant et leurs encouragements.

Table des matières

Liste des figures

Liste des tableaux

Introduction générale

Contexte du travail

Ce document constitue une synthèse de nos travaux de recherche entamés depuis 2006 au sein du laboratoire RIADI [1] à l'ENSI [2] de l'Université de Manouba en Tunisie lors de mon mastère de recherche. Nous avons mené nos travaux de recherche en thèse de 2009 à 2012 entre le laboratoire RIADI et le département ITI [3] à Télécom Bretagne de l'Université de Rennes 1 en France [14].
Notre thèse porte sur l'extraction de connaissances spatiotemporelles incertaines afin de prédire les changements de l'occupation des sols (COS). Elle vise à tirer profit des connaissances issues des images satellitaires pour la prédiction de COS. Pour ce faire, nous avons proposé une modélisation des images intégrant des connaissances spatiotemporelles des objets extraits à partir de scènes d'images satellitaires. Ensuite, nous avons combiné la logique floue avec les systèmes experts pour la prédiction de COS. La logique floue permet de modéliser les incertitudes liées à la prédiction de COS, alors que les systèmes experts permettent de garantir des meilleures décisions concernant les COS. Ces derniers sont présentés sous forme de règles de décisions. La pertinence de ces règles est évaluée à l'aide d'un module de raisonnement à base des cas (RBC). Aussi, nous avons exploité la fusion de données pour combiner les décisions multiples concernant les COS et obtenir une meilleure prédiction de l'occupation des sols.
A la suite de nos travaux de thèse, nous avons élargi nos recherches dans le domaine de la modélisation des incertitudes et plus précisément la propagation et la réduction des incertitudes. En effet, plusieurs sources d'incertitude accompagnent les modèles de COS par exemples : les incertitudes liées aux paramètres du modèle et les incertitudes liées à la structure du modèle. Ces incertitudes sont de deux types : aléatoires et épistémiques. L'ignorance de ces incertitudes lors de la modélisation des COS peut affecter les décisions à prendre ; ce qui peut être coûteux, d'un point de vue environnemental et/ou économique. L'idée de la propagation des incertitudes est d'estimer le taux de variation de la sortie du modèle de COS suite aux modifications de ses pa-

1. Laboratoire de Recherche en Informatique Arabisé et Documentique Intégrée
2. Ecole Nationale des Sciences de l'Informatique
3. Image et traitement de l'information

ramètres d'entrée. Tandis que l'analyse de sensibilité permet de déterminer l'effet des variations des paramètres d'entrée sur la variation des résultats du modèle et à classer ces paramètres selon l'impact de leurs variations.
Parallèlement à ces travaux, nous avons eu l'occasion de monter une collaboration avec le laboratoire CogBID [4] à l'Université de Stirling en Royaume-Uni. Cette collaboration nous a permis à nous orienter aussi au domaine de la réduction des incertitudes dans le domaine de la bio-informatique. En effet, nous nous sommes concentré sur la compréhension du cancer du poumon en analysant les biomarqueurs biologiques potentiels de ce cancer suite à différentes solutions de traitement. Pour le faire, nous avons proposé une approche basée sur la fusion de données pour combiner des informations provenant de plusieurs échantillons ayant eu des traitements différents avec du plasma atmosphérique non-thermique. La fusion de données permettra d'obtenir une meilleure description des gènes liés au cancer des poumons.
Dans le but d'étendre nos travaux de recherche, nous nous sommes intéressé au domaine d'analyse et d'interprétation des images satellitaires. Ce domaine de recherche constitue toujours un centre d'intérêt pour la communauté de la télédétection. Le processus d'analyse et d'interprétation des images satellitaires est composé de plusieurs étapes. Dans ce travail, nous nous sommes concentré sur la segmentation sémantique des images. Cette étape consiste à associer, à chaque pixel, un label parmi un ensemble de classes prédéfinis : eau, végétation, urbain, sol nu, etc. C'est une étape clé et elle est, généralement, considérée comme une étape nécessaire pour toute tâche d'analyse et d'interprétation des images. Suite à l'étape de segmentation sémantique, nous avons proposé un processus d'intégration des données visant à construire un entrepôt de données satellitaires. De plus, nous avons proposé une étape de modélisation des données et une étape d'interprétation des données. A l'issue de ces étapes, nous pouvons faire appel à plusieurs modèles d'analyse et d'interprétation des données satellitaires pouvant être utilisés dans divers domaines de la télédétection.
Egalement, et avec l'apparition du concept des données massives "big data" en imagerie satellitaire, nous nous sommes orienté à ce domaine de recherche innovant. Au-delà du volume, les données satellitaires massives se caractérisent également par leur variété, leur vélocité, leur véracité et leur valeur. Les outils et les plateformes traditionnels dans le domaine de l'imagerie satellitaire sont devenus incapables de traiter ces données massives hétérogènes dans un délai réalisable. Plusieurs nouveaux défis apparaissent avec ces données tels que l'acquisition des données, l'intégration et le nettoyage, le stockage, le traitement, la réduction de la dimensionnalité et l'analyse des données. Les données massives proviennent de sources hétérogènes et autonomes à contrôle distribué et décentralisé. Par conséquent, il est souvent difficile de découvrir des connaissances utiles et des informations pertinentes à partir de ces données. Dans ce travail, nous nous sommes intéressé à la classification d'un gros volume d'images satellitaires. Pour ce faire, nous avons proposé une architecture basée sur Apache Spark. L'approche proposée utilise l'apprentissage profond basé sur le réseau neuronal convolutif pour faire la classification des images.

4. Cognitive Big Data and Cybersecurity

Dans ce qui suit, nous allons présenter les contributions de nos travaux qui seront détaillées tout au long de ce manuscrit.

Contributions de recherche

Nos contributions de recherche seront détaillées dans les prochains chapitres et peuvent se résumer sous forme de deux axes majeurs : le premier axe concerne la modélisation des incertitudes et le deuxième axe concerne l'analyse et l'interprétation des données satellitaires.

Axe 1 : Modélisation des incertitudes

Le premier axe, modélisation des incertitudes, résume deux contributions essentielles qui sont la propagation des incertitudes et la réduction des incertitudes.

Propagation des incertitudes

Nous avons proposé une approche de propagation qui prend en compte les incertitude de type aléatoire et épistémique, la corrélation entre les paramètres, les incertitudes liées aux paramètres et les incertitudes liées au modèle de COS. Pour ce faire, nous avons modélisé les deux types d'incertitudes liées aux paramètres d'entrée du modèle de COS [46]. Les incertitudes de type aléatoire sont modélisées tout d'abord par des distributions de probabilités [18]. Ensuite, ces distributions sont transformées en des structures évidentielles. Alors que les incertitudes de type épistémique sont modélisées par des structures évidentielles. D'autres part, nous avons étudié les corrélations entre les paramètres d'entrée et nous avons présenté notre méthode de calcul de la corrélation. Enfin, nous avons proposé une approche basée sur la théorie des fonctions de croyance pour propager les incertitudes liées aux paramètres. Pour les incertitudes liées au modèle, nous avons commencé par caractériser ces incertitudes et par la suite, faire la propagation de l'incertitude de la structure du modèle. Notre approche est validée en utilisant quatre modèles de COS qui sont DINAMICA, SLEUTH, CA-MARKOV et LCM. Les résultats obtenus confirment l'importance de la prise en compte des incertitudes aléatoires et épistémiques. Aussi, ces résultats confirment la nécessité de prendre en compte les corrélations entre les paramètres d'entrée lors de la propagation des incertitudes. Finalement, il est essentiel de considérer les deux sources d'incertitudes (liées aux paramètres d'entrée et liées aux modèles de COS) lors de la propagation et ceci afin de garantir des meilleures décisions pour les COS.

Réduction des incertitudes

La contribution de cette partie s'articule autour de la réduction des incertitudes. Nous avons proposé trois approches de réduction des incertitudes. La première approche

proposée utilise la méthode de sensibilité globale des dérivées (DGSM) pour réduire les incertitudes liées aux modèles de COS [17]. Le processus de réduction des incertitudes commence par appliquer une analyse de sensibilité qualitative basée sur la méthode de criblage de Morris pour déterminer les paramètres d'entrée incertains du modèle de COS. Ensuite, une étude de corrélation est faite pour déterminer les paramètres corrélés et les classer en groupes. Nous appliquons, ensuite, une analyse de sensibilité qualitative basée sur DGSM pour identifier les paramètres qui ont une grande influence sur le modèle de COS. La deuxième approche de réduction des incertitudes utilise l'analyse de sensibilité tout en se basant sur la théorie des fonctions de croyance [48]. Le processus de réduction des incertitudes commence par appliquer l'analyse de sensibilité pour identifier les paramètres les plus influents [47]. Ces paramètres sont, ensuite, estimés en utilisant les limites de confiance du Kolmogorov-Smirnov. La troisième approche est basée sur la fusion des données et a été expérimentée pour l'analyse des biomarqueurs biologiques potentiels du cancer du poumon [43]. L'idée de l'approche proposée est de combiner des informations provenant de plusieurs échantillons pour des états qui ont été soumis à des traitements différents avec du plasma atmosphérique non-thermique. Le but de la fusion est d'obtenir une seule décision pour chaque état. Cette décision permettra une meilleure description des gènes liés au cancer du poumon. Les trois approches proposées pour la réduction des incertitudes sont validées à travers plusieurs jeux réels de données. De plus, l'évaluation de ces approches par rapport aux méthodes existantes dans la littérature montre les bonnes performances des approches proposées.

Axe 2 : Analyse et interprétation des données satellitaires

Le deuxième axe, analyse et interprétation des données satellitaires, regroupe essentiellement deux contributions qui sont l'analyse des données satellitaires non massives et des données satellitaires massives.

Analyse et interprétation des données satellitaires non massives

Nous avons proposé une approche d'analyse et d'interprétation des images satellitaires permettant d'offrir une aide à la décision pour les utilisateurs dans plusieurs domaines de la télédétection. L'approche proposée est basée sur quatre étapes. La première étape est la segmentation sémantique des images satellitaires [15]. Le processus de la segmentation commence par calculer les caractéristiques des objets extraits des images satellitaires. Ces caractéristiques constituent l'entrée d'un réseau de neurones pour générer une structure permettant de classifier les objets issus des images satellitaires. Ensuite, la structure générée est utilisée pour effectuer une segmentation sémantique au niveau des pixels. La deuxième étape de notre approche a pour but de charger les données dans l'entrepôt de données. A ce niveau, de nombreuses opérations, généralement complexes, sont réalisées pour préparer les données à être chargées dans l'entrepôt de données [16]. Le schéma choisi pour l'entrepôt de données est le schéma en étoile. Lors de la troisième étape, l'étape de modélisation, nous avons établi la so-

lution multidimensionnelle d'analyse en se basant sur les besoins des utilisateurs du domaine de l'imagerie satellitaire. La dernière étape de notre approche fournit un ensemble d'information descriptives et prédictives qui constituent une source pour l'aide à la décision [21]. L'approche proposée est validée et comparée aux approches existantes dans la littérature. Les résultats montrent les bonnes performances de l'approche proposée.

Analyse et interprétation des données satellitaires massives

Nous allons commencer par élaborer une étude bibliographique sur les données satellitaires massives ou "big data". Nous nous sommes intéressé aux caractéristiques des données, à la chaîne de traitement de ces données allant de la collecte des données, du stockage, du nettoyage et de l'intégration, de l'analyse jusqu'à l'interprétation et la visualisation des données [25]. Ensuite, nous avons comparé deux plateformes bien connues de traitement de données massives à savoir Hadoop et Spark. Nous avons commencé par décrire les deux plateformes. La première, Hadoop, est conçue pour traiter les données massives non structurées dans un environnement distribué. Quant à la deuxième plateforme, Spark, elle est composée d'un ensemble de bibliothèques et utilise l'ensemble de données distribué résilient pour surmonter la complexité de calcul [27]. De plus, nous avons proposé une approche de classification des images satellitaires massives. Cette approche est basée sur une architecture Apache Spark distribuée permettant de stocker et traiter le gros volume d'images satellitaires. Ces images sont divisées et distribuées aux nœuds esclaves pour faire la classification. Cette tâche est assurée par un algorithme d'apprentissage profond qui est le réseau neuronal convolutif. Nous avons utilisé TensorFlow pour construire et former le modèle d'apprentissage en profondeur. De plus l'architecture d'apprentissage en profondeur a été distribuée sur les nœuds esclaves pour assurer le parallélisme de la classification des images.

Organisation du document

Ce manuscrit présente une synthèse de nos travaux de recherches. Il comporte essentiellement deux parties. La première partie porte sur la modélisation des incertitudes pour les modèles de COS avec une ouverture sur le domaine de la bio-informatique. La deuxième partie décrit l'analyse et l'interprétation des données satellitaires.

Partie 1 : modélisation des incertitudes

- Le chapitre 1 introduit, en premier lieu, la notion de COS, les approches de modélisation de COS et les paramètres d'entrées de ces approches de modélisation. En second lieu, une étude des incertitudes des modèles de COS est présentée. A ce niveau, nous détaillons les types des incertitudes (aléatoires et épistémiques), les sources des incertitudes (liées aux paramètres du modèle de COS ou liées au modèle lui-même). Dans la dernière partie de ce chapitre, nous présentons les zones d'étude à savoir la région du Caire en Egypte et la région de l'île de la Réunion.
- Le chapitre 2 s'intéresse à la propagation des incertitudes associées aux modèles de COS. Ce chapitre passe en revue sur les méthodes de propagation d'incertitude ainsi que sur les problématiques liées à la propagation des incertitudes. Ensuite, nous détaillons notre approche de propagation des incertitudes pour les modèles de COS. Cette approche est divisée en deux parties : propagation des incertitudes liées aux paramètres et propagation des incertitudes liées aux modèles eux-mêmes. L'expérimentation de l'approche proposée est illustrée dans la quatrième partie de ce chapitre.
- Le chapitre 3 détaille l'approche proposée pour la réduction des incertitudes. Nous nous détaillons la réduction des incertitudes pour le domaine de la prédiction de COS et pour le domaine de la bio-informatique. Pour le premier domaine, la réduction des incertitudes est assurée en utilisant l'analyse de sensibilité. Deux méthodes d'analyse de sensibilité sont utilisées, à savoir la méthode de sensibilité globale basée sur les dérivées et celle basée sur la théorie des fonctions de croyance. Pour le deuxième domaine, l'idée est d'analyser les biomarqueurs biologiques potentiels du cancer des poumons tout en utilisant la fusion des données et leur regroupement et tout en considérant des limites pour les groupes.

Partie 2 : analyse et interprétation des données satellitaires

- Le premier chapitre présente un état de l'art sur un ensemble des travaux s'intéressant à l'analyse et l'interprétation des images satellitaires et les problèmes à résoudre. Par la suite, nous détaillons notre approche proposée. Cette approche est composée d'un module de segmentation sémantique des images satellitaires et d'un module d'intégration des données satellitaires suivi d'une étape de modélisation des données et d'une étape d'analyse et d'interprétation. L'approche proposée est validée sur un ensemble de données réelles représentant l'île de la Réunion.
- Le deuxième chapitre s'intéresse à l'analyse des données massives "big data" en imagerie satellitaire. Il introduit le concept et les définitions des données massives ainsi que les plateformes les plus utilisées dans ce domaine à savoir Hadoop, Apache Spark, Apache Storm et HPCC (High-Performance Computing Cluster). Dans ce chapitre, nous présentons également une synthèse sur les travaux liés à

l'analyse de gros volumes d'images satellitaires et les problèmes à résoudre. La troisième partie de ce chapitre est réservée à l'approche proposée. Dans cette partie, nous détaillons l'architecture de notre approche ainsi que les différents modules qu'elle contient et les étapes de traitement à suivre.

Partie 1 : Modélisation des incertitudes

Chapitre 1

Changement de l'occupation des sols, incertitudes et données

Sommaire

1.1 Introduction

L'objectif de la télédétection est d'inférer des informations sur des objets se trouvant sur la terre à partir de mesures effectuées à distance, souvent depuis l'espace. Le processus d'inférence est toujours loin d'être parfait et il y a donc plusieurs éléments d'incertitude concernant les résultats obtenus par la télédétection. Vu sous cet angle, le problème de l'incertitude est au cœur des domaines liés à la télédétection. L'étude des incertitudes aide à comprendre leurs sources, leurs types et comment raisonner avec ces

incertitudes pour améliorer les processus de traitement et d'interprétation.
Dans ce chapitre, nous commençons par définir le concept de changement de l'occupation des sols et les approches de modélisation de ce changement. Ensuite, nous nous intéressons à l'étude des incertitudes dans les modèles de changement de l'occupation des sols. Dans la dernière partie de ce chapitre, nous présentons les zones d'études et des données qui seront utilisées tout au long de ce document.

1.2 Changement de l'occupation des sols (COS)

Dans cette partie, nous nous intéressons à définir le terme "changement de l'occupation des sols" que nous dénotons par COS dans le reste du document ; ensuite, nous présentons les modèles de COS.

1.2.1 Définition de COS

L'occupation des sols (urbain, végétation, sol nu, etc.) décrit les caractéristiques biophysiques de la surface des terres. La couverture du sol peut être déterminée en analysant l'imagerie satellitaire et aérienne. Les cartes de la couverture terrestre fournissent des informations pour aider les gestionnaires à mieux comprendre le paysage actuel. Il est souvent essentiel de construire des cartes sur plusieurs années afin de comprendre le changement au cours du temps. Ainsi, il est possible, tout en se basant sur les changements passés, de mieux comprendre les effets possibles des décisions actuelles avant leur mise en œuvre.
La recherche sur COS est de nature multidisciplinaire faisant intervenir divers domaines tels que le SIG (Systèmes d'Information Géographique), l'économie, la sociologie, la géographie, l'informatique et la démographie. Les COS peuvent être naturels suite à des phénomènes naturels tels que la déforestation, la désertification et l'inondation ou d'origine anthropique tels que les incendies et l'urbanisation.
L'étude des changements de l'occupation des sols fait souvent références à l'utilisation des terres (land use en anglais). Cette dernière décrit l'utilisation humaine (les activités humaines) des terrescomme : l'agriculture, les pâturages ou les plantations.

1.2.2 Approches de modélisation de COS

Avec la multiplication des phénomènes naturaux tels que la déforestation, sécheresse et l'étalement urbain, etc., le besoin pour des approches modélisatrices des COS est devenu essentiel afin de comprendre ces problèmes et pouvoir les résoudre. L'idée de la modélisation des COS est d'étudier les causes et les conséquences de ces changements afin de les bien assimiler [32].
Dans la littérature, plusieurs travaux ont été proposés pour simuler ou prédire les COS d'une région donnée. Ces modèles partent d'observations faites sur une région et produisent des aides à la prise des décisions concernant cette région. Plusieurs travaux

ont proposé des classifications de ces modèles de COS [94]. Ces travaux ont distingué plusieurs catégories de modèles : i) les modèles basés sur les statistiques, ii) les modèles basés sur les automates cellulaires, iii) les modèles basés sur les agents, iv) les modèles basés sur les chaînes de Markov, v) les modèles basés sur l'apprentissage automatique, et finalement vi) les modèles hybrides.
Dans notre étude, nous intéressons à cinq modèles de COS qui sont : DINAMICA, SLEUTH, CA-MARKOV, LCM et FS-FDT.

- **DINAMICA** : c'est un modèle de COS basé sur les automates cellulaires. Il présente des fonctions de transition, et intègre une approche de rétroaction spatiale permettant une simulation stochastique à étapes multiples. Il est basé sur la régression logistique pour déterminer les probabilités de transition dynamique spatiale. DINAMICA utilise une séquence d'étapes présentées dans la figure 1.1. Ce modèle a été utilisé dans plusieurs études telles que [106] [4].

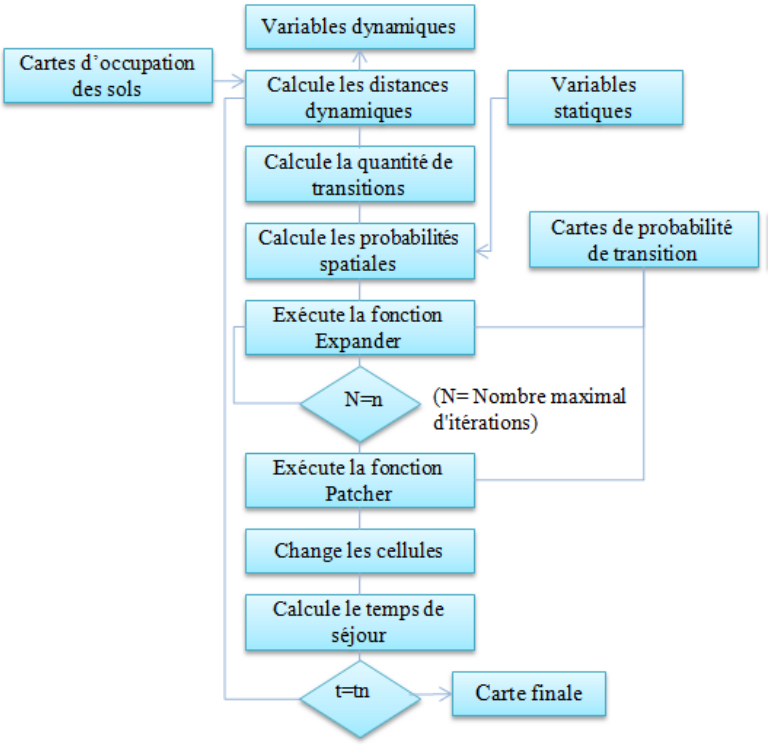


Figure 1.1 — Structure du modèle DINAMICA.

- **SLEUTH (Slope, Land cover, Exclusion, Urbanization, Transportation and Hillshade)** : c'est un modèle qui est basé sur les automates cellulaires pour simuler l'étalement urbain. Il utilise quatre types de règles de croissance : expansion par diffusion, continuité de l'urbain existant, le long des routes et création de nouveaux centres. Cinq coefficients de croissance (diffusion, propagation, race, résistance aux pentes et gravité de la route) contrôlent le comportement du modèle et déterminent les types de croissance déjà mentionnés (Figure 1.2).

SLEUTH a été utilisé dans plusieurs applications telles que [24] [10].

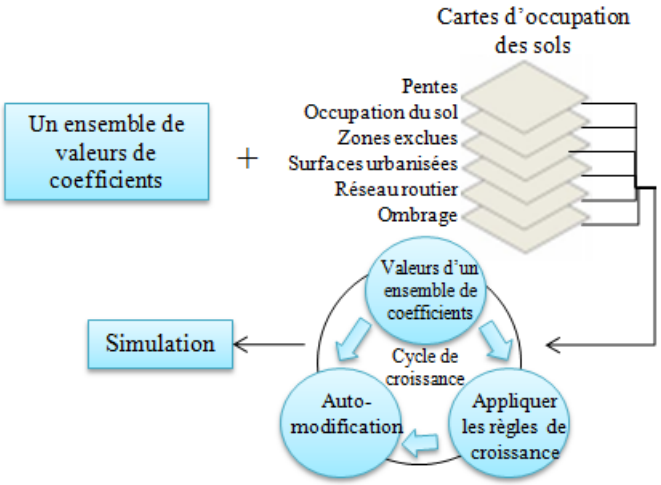


Figure 1.2 — Structure du modèle SLEUTH.

– **CA-MARKOV** : c'est un modèle adaptatif qui est basé sur les chaînes de Markov pour identifier les changements et les automates cellulaires (AC) pour allouer spatialement ces changements (Figure 1.3). Les chaînes de Markov servent pour créer une probabilité de transition et une matrice de zone de transition. Les ACs sont intégrés dans l'approche markovienne afin d'ajouter le caractère spatial au modèle. Des exemples d'applications peuvent être trouvés dans [117] [70].

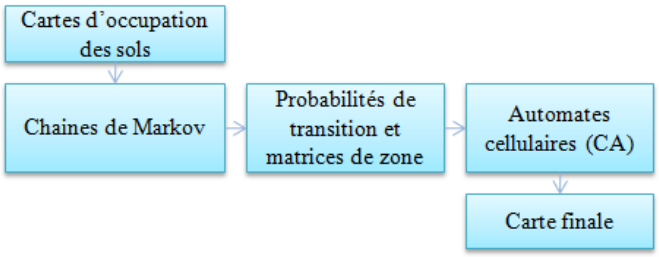


Figure 1.3 — Structure du modèle CA-MARKOV.

– **LCM (Land Change Modeler)** : ce modèle repose sur deux phases. La première permet d'analyser les changement pour construire de sous-modèles de transition entre les types de COS (Figure 1.4). La deuxième phase a pour but de modéliser les changements de COS en utilisant les chaînes de Markov et les cartes de transition obtenues par régression logistique ou par des algorithmes d'apprentissage automatique. Le modèle LCM est appliqué dans de nombreuses situations telles que [60] [1].
– **FS-FDT** : c'est un modèle qui permet de prédire les changements de COS en se basant sur les arbres de décisions flous [20]. FS-FDT est composé de quatre

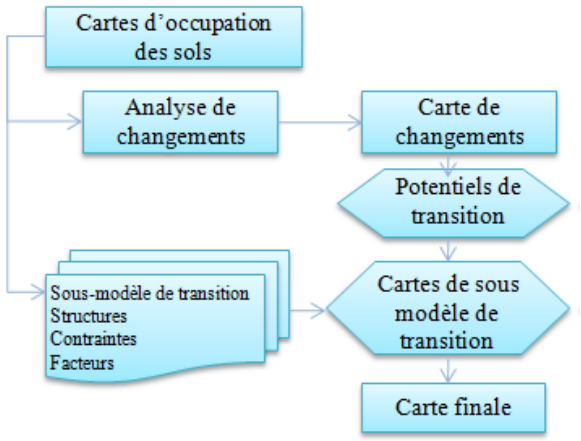


Figure 1.4 — Structure du modèle LCM.

étapes (Figure 1.5). La première étape a pour objectif de rechercher les objets similaires à un objet requête. Ensuite, les arbres de changements relatifs aux objets similaires trouvés sont construits. La troisième étape consiste à rechercher les arbres pertinents parmi l'ensemble des arbres trouvés. La dernière étape a pour objectif de combiner les arbres pertinents pour construire un arbre plus complet qui décrit les changements d'une région donnée. Le modèle FS-FDT est appliqué dans les situations suivantes [18] [17].

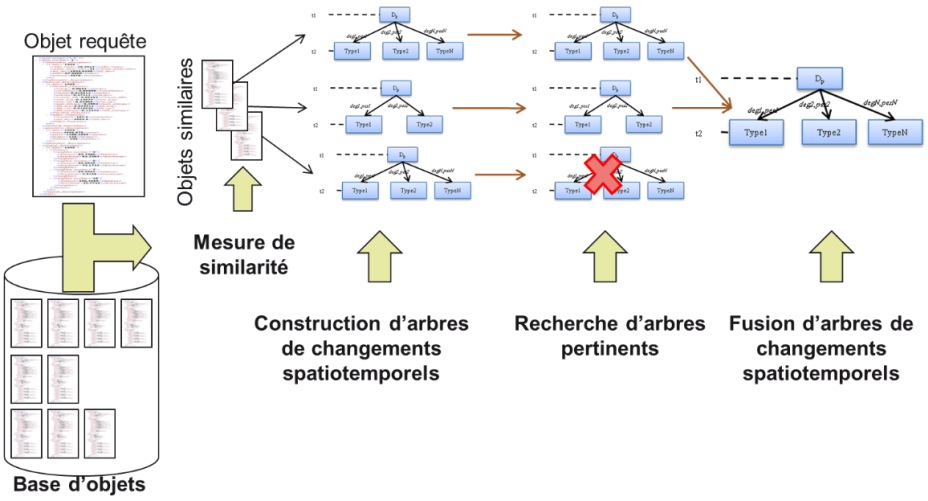


Figure 1.5 — Structure du modèle FS-FDT.

1.2.3 Description des paramètres d'entrées des modèles de COS

Les approches de modélisation de COS ont pour objectif de fournir une interprétation des changements de COS. Généralement, l'idée est d'étudier les changements passés pour fournir une simulation future des changements.
Dans un contexte de modélisation de COS, les objets contenus dans les images satellitaires peuvent être décrits par un ensemble évolutif de paramètres (par exemple forme, couleur, texture, taille et relations entre les objets) [20] [99]. Ces paramètres changent au cours du temps.
Dans nos travaux précédents, nous avons étudié les paramètres d'entrées des modèles de COS. Ces paramètres sont groupés en cinq familles : paramètres spectraux, paramètres de textures, paramètres de forme et de relations spatiales, paramètres de végétation et paramètres climatiques [17] [47] [46].

- **Paramètres spectraux** : ces paramètres fournissent des informations sur la composition spectrale d'un objet. Ces paramètres sont : les valeurs moyennes, les valeurs de déviation standard du vert (MVG, SDG), du rouge (MVR, SDR), de PIR (MVN, SDN), les bandes infrarouges à ondes courtes (MVS, SDS) et mono-spectrales (MVMB, SDMB) pour chaque objet de l'image.
- **Paramètres de textures** : ces paramètres fournissent des informations sur l'arrangement spatial des valeurs radiométriques d'un objet de l'image. Ces paramètres sont générés à partir de GLCM (matrice de co-occurrence de niveau de gris) et comprennent l'homogénéité (Hom), le contraste (Cont), la dissimilarité (Dis), l'entropie (Ent), l'écart-type (SD) et la corrélation (Cor).
- **Paramètres de forme et de relations spatiales** : ces paramètres fournissent des informations sur la forme, la position et les relations entre les objets d'une image. Ces paramètres sont : l'aire (A), la longueur / la largeur (LW), l'indice de forme (SI), l'arrondi (R), la densité (Den), les relations métriques (MR) et les relations de direction (DR) de chaque objet d'image.
- **Paramètre de végétation** : ce paramètre permet d'isoler les zones de végétation. Ce paramètre est le NDVI (Normalized Difference Vegetation Index).
- **Paramètres climatiques** : ces paramètres décrivent les conditions climatiques lors de la prise de l'image satellitaire. Ces paramètres sont : la température (Tem), l'humidité (Hum) et la pression (Pre).

Le tableau 1.1 décrit les paramètres d'entrée considérés dans ce travail pour les modèles de COS ainsi que leurs formules et gammes de valeurs.

1.3 Etudes des incertitudes dans les modèles de COS

Le processus de modélisation de COS est généralement entaché par des incertitudes [20] [54]. Ces incertitudes accompagnent le processus de modélisation dans ses différentes phases à savoir l'acquisition des données, le prétraitement, le traitement, l'analyse et l'interprétation.
Afin de modéliser les incertitudes qui accompagnent les modèles de COS, il faut tout

Paramètres	Description	Formule	Gamme
BI	c'est la moyenne des valeurs de gris de toutes les bandes pour chaque pixel.	$\frac{TM1+...+TM7}{7}$	$[C_k^{min}, C_k^{max}]$, C_k^{min} : plus faible valeur d'intensité et C_k^{max} : la valeur d'intensité la plus brillante
MB	La valeur moyenne de la bande spectrale bleue.	-	$[C_k^{min}, C_k^{max}]$
MR	La valeur moyenne de la bande spectrale rouge.	-	$[C_k^{min}, C_k^{max}]$
MG	La valeur moyenne de la bande spectrale verte.	-	$[C_k^{min}, C_k^{max}]$
MN	La valeur moyenne de la bande spectrale proche infrarouge.	-	$[C_k^{min}, C_k^{max}]$
Ctr	c'est la somme de la variance des carrés.	$\sum_{i,j=0}^{N-1} P_{i,j}(i-j)^2$, $P_{i,j}$: la matrice de co-occurrence, i et j : les niveaux de gris, N : le nombre de niveaux de gris	[0, 90]
Ent	Représente le caractère aléatoire de la répartition de l'intensité du niveau de gris .	$\sum_{i,j=0}^{N-1}(-lnP_{i,j})$	[0, 90]
ASM	Mesure l'uniformité de la texture.	$\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} P(i,j)^2$	[0, 90]
RF	Mesure l'écartement des objets	$a/l_{max} \times l_{min}$, a : zone du polygone, l_{max} : longueur du grand axe, and l_{min} : longueur du petit axe	[0,1]
EF	Explique comment un objet correspond à une ellipse.	-	[0,1]
SI	L'indice de forme décrit la lisibilité de la limite d'un objet.	$p/(4 \times \sqrt{a})$, p : périmètre du polygone	[1, ∞]
D	La densité est définie comme le rapport entre la surface de l'objet et son rayon [Navulur 2007].	$\sqrt{P_v}/(1+\sqrt{\sigma_X^2+\sigma_Y^2})$, $\sqrt{P_v}$: diamètre d'un objet carré,P_v pixels, et $\sqrt{\sigma_X^2+\sigma_Y^2}$: diamètre d'ellipse ajustée sur segment	[0, Selon la forme de l'objet de l'image]
A	La zone de l'objet identifié.	-	[0, taille de la scène]
NDVI	Principe de la forte absorbance de la végétation dans la réflectance rouge et PIR du spectre [Bonn et Rochon 1996].	$\frac{TM4-TM3}{TM4+TM3}$	[-1, 1]
Tem	Décrit le degré de chaleur ou de froid d'un objet.	-	$\mathbb{R}$
Hum	L'humidité est définie comme le déficit de pression de vapeur.	$\frac{e}{e_s} \times 100\%$, e : La pression partielle de la vapeur d'eau réelle, et e_s : La pression de la vapeur d'eau saturée	$\mathbb{R}$

Tableau 1.1 — Description des paramètres d'entrée.

d'abord comprendre la nature de ces incertitudes. Dans ce qui suit, nous allons présenter les types et les sources des incertitudes pour les modèles de COS.

1.3.1 Types des incertitudes

En littérature, plusieurs catégorisations ont été données aux incertitudes. Dans ce rapport, nous classons l'incertitude selon deux types : aléatoire et épistémique. L'incertitude est dite aléatoire lorsqu'elle est due à une variabilité naturelle, et épistémique si elle est due à un manque de connaissance [49] [46].

1.3.1.1 Incertitude aléatoire

Selon [97], l'incertitude aléatoire désigne la variabilité naturelle dans les populations connues (ou observables). Dans la littérature, plusieurs autres appellations sont accordées à ce type d'incertitudes telles qu'incertitude irréductible, variabilité et incertitude intrinsèque.
Dans notre contexte, les incertitudes sont considérées aléatoires si elles sont liées aux caractéristiques spectrales, de végétation et climatiques [68].

1.3.1.2 Incertitude épistémique

Selon [63], l'incertitude épistémique provient du manque de connaissances sur les phénomènes fondamentaux. Elle est relative à notre capacité à restreindre l'étendue de l'incertain grâce à l'approfondissement des connaissances (recueil de données, consultation d'experts, essais accélérés, etc.).
L'incertitude épistémique se caractérise par un manque ou une absence de données ou des informations sur un phénomène ou un paramètre physique.
Dans notre contexte, les incertitudes sont considérées épistémiques si elles sont liées aux caractéristiques de forme ou de texture [97].

Dans la littérature, plusieurs chercheurs ont conclu que les incertitudes épistémiques nécessitent un traitement différent de celui des incertitudes aléatoires. La distinction entre les deux types d'incertitudes demeure très importante. Ceci conduit généralement à l'application de méthodes différentes pour modéliser les deux types d'incertitudes. Ainsi, nous pouvons profiter des pouvoirs modélisateurs de chaque méthode pour modéliser chaque type d'incertitude (aléatoire et épistémique) et réduire les incertitudes qui accompagnent la prédiction de COS.

1.3.2 Sources des incertitudes

Les incertitudes liées aux modèles de COS parviennent de deux grandes sources : les incertitudes des paramètres du modèle et les incertitudes de la structure du modèle.

Paramètres	Sources d'incertitude
Paramètres spectraux	La réflectance spectrale de la surface, l'étalonnage du capteur, le bruit du capteur, l'effet de pixels mixtes, les nuages de sous-pixels, effet d'un changement dans l'emplacement de canal, l'enregistrement de pixels entre plusieurs canaux spectraux, le profil de la température et l'humidité atmosphériques, les conditions atmosphériques, les conditions de fonctionnement des instruments, les nuages, la topographie, la géométrie de visée, les particules de brume.
Paramètres de texture	L'interaction spatiale entre la taille de l'objet dans la scène et la résolution spatiale du capteur, l'effet de frontières, l'ambiguïté dans la distinction objet / arrière-plan.
Paramètres de forme	La comptabilisation de la position saisonnière du Soleil par rapport à la Terre, les conditions dans lesquelles l'image à été acquise, les changements dans l'illumination de la scène, les conditions atmosphériques, la géométrie d'observation.
Paramètres de végétation	La variation dans la luminosité de fond de sol, les bandes rouge et NIR, les perturbations atmosphériques, les perturbations atmosphériques en fonction de l'état de l'atmosphère et de la surface du sol au moment des deux acquisitions, la variabilité dans la structure de sous-pixel de la végétation.
Paramètres climatiques	La correction atmosphérique, le bruit du capteur, l'émissivité de la surface de la terre, les absorbeurs de gaz, les effets angulaires, l'imperfection de longueur d'onde.

Tableau 1.2 — Sources d'incertitude des paramètres d'entrée.

Il est essentiel de prendre en compte chacune de ces sources d'incertitude à part pour pouvoir améliorer les décisions sur les COS.

1.3.2.1 Incertitudes liées aux paramètres

Le premier type d'incertitude concerne les paramètres d'entrée du modèle de COS. Le type d'incertitude des paramètres d'entrée dépend des sources de ces incertitudes. Le tableau 1.2 décrit les paramètres d'entrée des modèles de COS et leurs sources d'incertitude [97] [68].

1.3.2.2 Incertitudes liées au modèle de COS

Le modèle de COS lui-même peut être sujet à des incertitudes. Ces incertitudes sont le plus souvent dues à des hypothèses ou des simplifications qui sont faites lors du développement de ces modèles [37]. Plusieurs structures d'un modèle peuvent être construites tout en changeant les hypothèses liées à ces structures. Cependant, aucune

d'elles ne pourra représenter le modèle physique réel avec certitude. L'objectif devient donc de choisir une meilleure structure du modèle parmi un ensemble de structures [58]. Ainsi, le choix d'une structure donnée d'un modèle implique des incertitudes liées à cette structure. Dans ce manuscrit, nous convenons d'appeler cette incertitude, incertitude de la structure du modèle. Cette dernière est généralement épistémique car elle provient de notre manque de connaissance.
Il est important de considérer ce type d'incertitudes afin de garantir une meilleure prise de décision concernant un modèle donné [100].

1.4 Présentation des zones d'études

Dans cette section, nous allons présenter les régions d'études qui sont la région du Caire en Egypte et la région de l'île de la Réunion. Aussi, nous allons décrire les données utilisées pour expérimenter notre approche.

1.4.1 Le Caire

La première région d'étude est le Caire,qui est la capitale de l'Egypte (Figure 1.6). C'est une méga-ville mondiale caractérisée par une importante croissance démographique. Cette croissance a provoqué un étalement urbain sur les zones agricoles autour du Caire. Ainsi, les plaines désertiques adjacentes ont subi un étalement urbain pour faire face à l'augmentation de la population. Cette dernière s'est augmentée de 6,4 millions en 1976 à 12,5 millions en 2006 pour la région du Caire [109]. D'autres part, la région du Caire témoigne la présence de principales installations et services du gouvernement. Ainsi, l'étude des COS demeure essentielle pour la région du Caire.

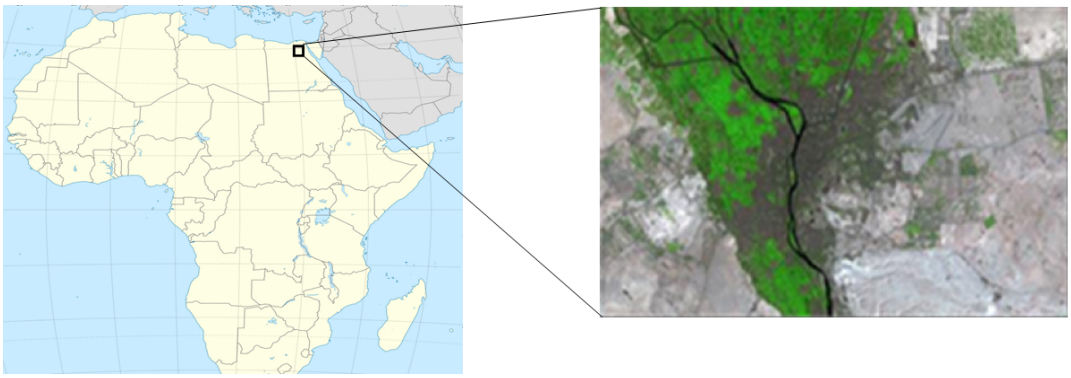


Figure 1.6 — Localisation du premier site d'étude.

Les images utilisées pour le premier site proviennent de la base USGS [1](United

1. https ://www.usgs.gov/

States Geological Survey) qui contient des images satellitaires prétraitées, multi-temporelles et multi-capteurs.

1.4.2 Ile de la Réunion

Le deuxième région d'étude est l'île de la Réunion (Figure 1.7). Cette région se trouve dans l'océan Indien à environ 700 kilomètres à l'est de Madagascar (21°7' à 19°40' sud, 55°13' à 61°13' est). La population de cette région est d'environ 865,826 habitants en 2008. Il s'agit d'une île d'origine volcanique de superficie 2 512 km^2 occupant une forme ovoïde, compacte et fortement accidentée. Dans notre étude, nous nous intéressons aux villes de Saint-Denis, le Port et Saint-Paul. Ces régions sont confrontées à une croissance démographique importante, à l'existence de zones naturelles protégées et à l'importance de l'agriculture. Ainsi, la modélisation de COS demeure très importante pour ce site d'étude.

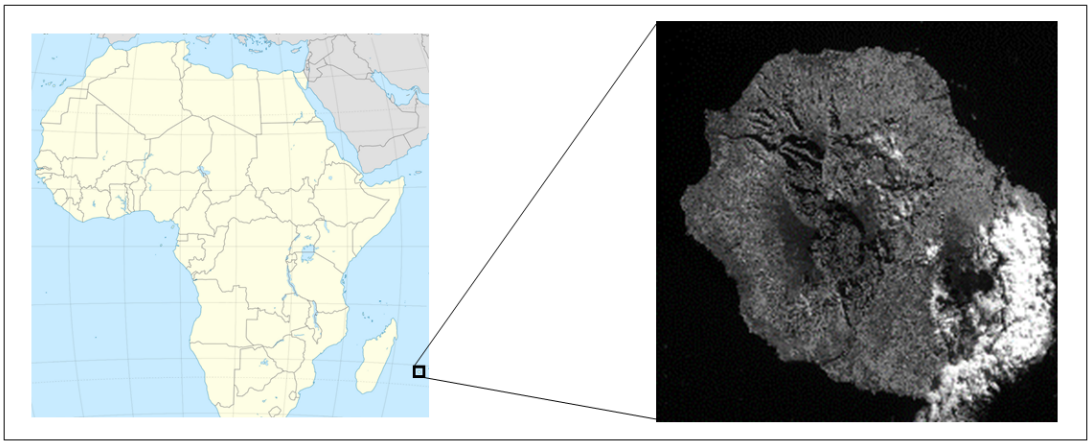


Figure 1.7 — Localisation du deuxième site d'étude.

Les images utilisées pour le deuxième site proviennent de la base Kalideos [2] Isle-Réunion du CNES [3]. Cette dernière renferme des images satellitaires corrigées radiométriquement, géométriquement et atmosphériquement. Aussi, ces images sont acquises à des différentes dates et provenant de plusieurs capteurs. Cette base a été choisie car elle contient des images satellitaires géoréférencées, parfaitement superposables et radiométriquement cohérentes. Ces caractéristiques de séries d'images permettent le suivi de COS.

2. http ://kalideos.cnes.fr
3. Centre National d'Etudes Spatiales (French Space Agency)

1.5 Conclusion

Dans cette première partie, nous nous intéressons à la modélisation des incertitudes pour les modèles de COS. Ainsi, nous avons défini le COS et les approches de modélisation qui seront utilisées dans cette première partie. Ensuite, nous avons décrit les types et les sources des incertitudes qui sont liées au COS. Enfin, les régions d'études et les données qui seront utilisées pour expérimenter nos approches sont présentés dans la dernière section de ce chapitre.
La modélisation des incertitudes liées aux modèles de COS est scind'ee en deux phases principales. La première phase est la propagation des incertitudes liées aux paramètres d'entrée et des incertitudes liées aux modèles eux-mêmes. Cette phase sera détaillée dans le chapitre suivant. La deuxième phase qui est la réduction des incertitudes liées aux modèles de COS sera présentée en détail dans le troisième chapitre.

Chapitre 2

Propagation des incertitudes

Sommaire

2.1 Introduction

Le présent chapitre s'intéresse à la propagation des incertitudes aléatoires et épistémiques associées aux données d'entrée des modèles de COS ainsi qu'aux structures de ces modèles. Notre approche de propagation des incertitudes est divisée en cinq étapes principales : 1) transformation des paramètres incertains dans le cadre des fonctions de croyance ; 2) analyse corrélationelle des paramètres ; 3) propagation de l'incertitude (aléatoire et épistémique) des paramètres d'entrée ; 4) caractérisation de l'incertitude des structures de modèles de COS et 5) propagation de l'incertitude associée aux structures de modèles de COS.

Une analyse correspondant à l'état de l'art relatifs aux méthodes de propagation d'incertitude ainsi que les problématiques à résoudre dans notre travail sera présentée dans la première partie de ce chapitre. Nous détaillons ensuite notre approche de propagation des incertitudes. Cette dernière est évaluée puis validée au niveau de la quatrième partie. Enfin, les conclusions et les perspectives relatives à notre travail font l'objet de la cinquième section.

2.2 Etat de l'art et problématiques étudiées

La propagation d'incertitude permet de mesurer l'effet des incertitudes des paramètres d'entrée et/ou les incertitudes du modèle lui-même sur les résultats finaux de ce modèle.
Une revue de la bibliographie met en évidence plusieurs méthodes de propagation d'incertitude. Par ailleurs, les travaux de plusieurs équipes de recherche montrent que ces dernières méthodes sont classées selon le type d'incertitude traité en deux familles : méthodes probabilistes et méthodes non probabilistes [23] [74] [76] [46].
Les méthodes probabilistes sont plus appropriées pour traiter les incertitudes aléatoires. Parmi ces méthodes, nous citons les méthodes d'échantillonnage [52], les méthodes basées sur l'expansion locale [62], la méthode de décomposition de Neumann [116], les méthodes basées sur l'intégration numérique telle que la méthode de réduction de dimension [119] et les méthodes basées sur les points les plus probables [30].
En ce qui concerne les méthodes non probabilistes, elles sont plus appropriées pour traiter les incertitudes épistémiques. Parmi ces méthodes, nous citons par exemple, la méthode des ensembles flous [120], la méthode des possibilités [39] et la méthode des fonctions de croyance [105].
Récemment, plusieurs travaux ont abordé la thématique de la propagation des incertitudes dans le domaine de l'imagerie satellitaire ([112] [18] [111] [113] [17] [47] [46]).
Dans la littérature, nous pouvons distinguer trois limites majeurs pour les travaux se rapportant à la propagation des incertitudes :

- La plupart des travaux traitent un seul type d'incertitude pour les paramètres d'entrée (aléatoire ou épistémique). Cependant, plusieurs types d'incertitude dans les modèles de COS peuvent accompagner les données.
- La corrélation entre les paramètres d'entrée reste un problème pour plusieurs études dans le domaine de COS [96]. D'autre part, les modèles de COS se caractérisent par un nombre important de paramètres corrélés. La négligence de la corrélation peut conduire à des décisions erronées.
- Un troisième point très important à considérer pour construire une approche de propagation d'incertitude est de considérer l'incertitude liée au modèle. En effet, aucun modèle de COS ne peut être fidèle à la réalité et il est souvent basé sur des simplifications et des approximations. Plusieurs modèles peuvent représenter un même système physique. De cette manière, l'incertitude provient de la sélection du modèle le plus fidèle au système physique parmi un ensemble de modèles. Dans ce travail, nous désignons cette incertitude par "incertitude de la structure

du modèle".

La démarche proposée concernant la propagation des incertitudes pour le modèle de COS sera détaillée ci-dessous. La démarche proposée permet de résoudre ces trois problèmes.

2.3 Approche proposée pour la propagation des incertitudes

Dans cette section, nous précisons notre approche de propagation des incertitudes liées : a) aux données d'entrée relatif aux modèles du COS et b) au modèle lui-même. Pour la propagation des incertitudes liées aux paramètres d'entrée, nous détaillons, dans un premier temps la modélisation des incertitudes. Ensuite, nous détaillons le concept de la corrélation existant entre les paramètres d'entrée. Ceci nous permettra, en troisième lieu, de propager les incertitudes correspondant aux données à travers le modèle de COS.

Le processus de propagation des incertitudes liées au modèle est divisé en deux étapes : la première concerne l'étude des incertitudes concernant la structure du modèle et la deuxième se focalise sur la propagation de ces incertitudes.

La figure 2.1 décrit l'architecture proposée pour la propagation des incertitudes.

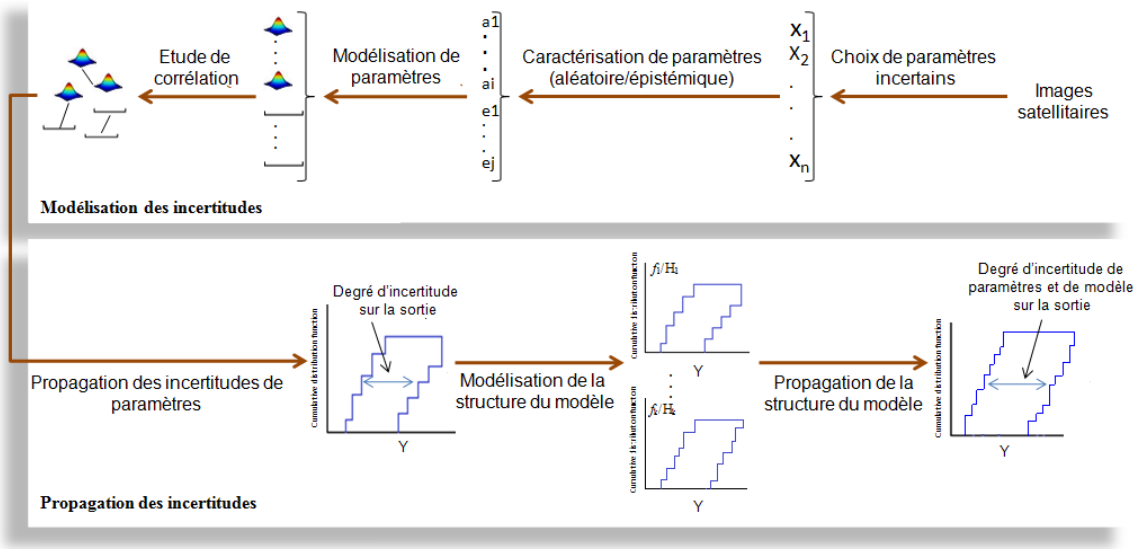


Figure 2.1 — Architecture proposée pour la propagation des incertitudes.

2.3.1 Propagation des incertitudes liées aux paramètres

Pour modéliser les incertitudes de type aléatoire liées aux paramètres d'entrée, nous avons utilisé la théorie des probabilités qui reste selon la littérature la théorie la plus appropriée pour ce type d'incertitude.
Cette section est scindée en trois parties. La première s'intéresse à la modélisation des incertitudes liées aux paramètres. La seconde partie concerne la corrélation entre les paramètres. La dernière partie se focalise sur la méthode de propagation des incertitudes relatives aux données d'entrée.

2.3.1.1 Modélisation des incertitudes liées aux paramètres

Les incertitudes liées aux paramètres peuvent être aléatoire ou épistémique.
Dans notre travail, la modélisation des paramètres de type aléatoire est assurée par des distributions normales de probabilité.
Les incertitudes liées aux paramètres de type épistémique sont modélisées par la théorie des fonctions de croyance. Chaque paramètre e est modélisé par un ensemble des intervalles de confiance qui concerne une structure des fonctions de masse (BPA) comme suit :

$$\{[e_1^L, e_1^U]/(m(1), [e_2^L, e_2^U]/(m(2), ..., [e_k^L, e_k^U]/(m(k), ...|k \in (1, 2, ..., P)\} \tag{2.1}$$

où P représente le nombre total de sous-intervalles de e et $m(k)$ désigne la valeur de BPA associée au k-ième sous-intervalle $[E_K^L, E_K^U]$. En cas de présence de structures de BPA différentes, la règle de combinaison est utilisée pour intégrer ces structures dans une structure BPA conjointe telle que $e_j/m(e_j)(j \in [1, 2, ..., P])$, où e_j est aussi un intervalle décrit comme $[e_K^L, e_K^U]$ et $m(e_j)$ est la valeur de BPA associée à e_j.

2.3.1.2 Analyse correlationelle des paramètres

La corrélation consiste à mesurer le degré d'association entre deux paramètres. Si ces deux paramètres sont corrélés, nous pouvons dans ce cas utiliser les informations de l'un pour prédire les valeurs de l'autre. Dans notre contexte, les paramètres d'entrée au modèle de COS sont généralement corrélés. Ainsi, il est important d'identifier l'effet de la corrélation des paramètres sur le changement de la sortie du modèle de COS.
Dans la littérature, le concept de corrélation a été largement discuté dans les statistiques classiques. En revanche, la problématique liée à la corrélation dans le contexte des données incertaines n'a pas été bien abordée.
Dans la partie suivante, nous évoquons de façon détaillée comment traiter la corrélation entre les paramètres aléatoires et épistémiques.

- **Corrélation entre les paramètres aléatoires**
  Le coefficient de corrélation, r, entre deux paramètres aléatoires X et Y

représentés par des distributions normales de probabilités est décrit par l'équation suivante :

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \tag{2.2}$$

où

$$\overline{x} = \mu = \sum_{i=1}^{n} x_i/n \quad et \quad \overline{y} = \mu = \sum_{i=1}^{n} y_i/n \tag{2.3}$$

- $-1 \leq r \leq 1$,
- si $r > 0$, X et Y sont positivement corrélés,
- si $r < 0$, sont négativement corrélés, et
- si $r = 0$, X et Y ne sont pas corrélés.

- **Corrélation entre paramètres épistémiques**
  Dans notre travail, la modélisation des paramètres épistémiques est faite en utilisant des intervalles de confiances.
  Soient $IX_i = [x_i^L, x_i^U]$ les intervalles observés pour le paramètre épistémique $e_1(X)$, où x_i^L et x_i^U sont respectivement la borne inférieure et la borne supérieure pour X_i. $IY_i = [y_i^L, y_i^U]$ les intervalles observés pour le paramètre épistémique $e_2(Y)$, où y_i^L et y_i^U sont respectivement la borne inférieure et la borne supérieure pour Y_i. Le facteur de corrélation r_{Intj} entre deux intervalles (IX, IY) est défini par l'équation suivante :

$$r_{Intj} = \frac{\sum_{i=1}^{n}(x_{ij} - \overline{x_j})(y_{ij} - \overline{y_j})}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \overline{x_j})^2}\sqrt{\sum_{i=1}^{n}(y_{ij} - \overline{y_j})^2}} \tag{2.4}$$

où

$$\overline{x_j} = \sum_{i=1}^{n} x_{ij}/n \quad et \quad \overline{y_j} = \sum_{i=1}^{n} y_{ij}/n \tag{2.5}$$

Le facteur de corrélation entre deux intervalles est un intervalle. Soient $r^L = inf_j\{r_{Intj} | j = 1, 2, ..., P\}$ et $r^U = sup_j\{r_{intj} | j = 1, 2, ..., P\}$, alors le coefficient de corrélation des intervalles de (IX, IY) est donné par $r = [r^L, r^U]$.
- Si $r^U > r^L \geq 0$, alors il y a une corrélation positive entre IX et IY
- Si $r^U < r^L \leq 0$, alors il y a une corrélation négative entre IX et IY
- Si $r^U = r^L = 0$, alors il n'a pas une corrélation entre IX et IY
- Si $r^L < 0 < r^U$, alors nous ne pouvons pas juger si IX et IY sont corrélés

2.3.1.3 Propagation de l'incertitude de paramètres

La propagation de l'incertitude de paramètres d'entrée du modèle de COS est réalisée par la théorie de la fonction de croyance. Cette théorie permet de considérer la corrélation entre les paramètres d'entrée [125].

Soit un modèle de changement (f) qui a des paramètres aléatoires notés a et des paramètres épistémiques notés e comme suit :

$$Y = f(a, e) \tag{2.6}$$

L'incertitude épistémique e est généralement représentée par des intervalles avec des fonctions de masse ou BPA, comme le présente l'équation suivante :

$$\{[e_1^L, e_1^U]/(m(1)), [e_2^L, e_2^U]/(m(2), ..., [e_k^L, e_k^U]/(m(k), ...|k \in (1, 2, ..., P)\} \tag{2.7}$$

Les paramètres aléatoires a sont décrits par une distribution normale de probabilités a (μ, σ), où μ est la moyenne et σ est l'écart type.
L'intervalle N pourra être discrétisé en des sous-intervalles $[a_i^L, a_i^U]$, $i \in [1, 2, ..., N]$. La fonction de masse sera, donc, donnée par l'équation suivante :

$$m(a_i) = \int_{a_i^L}^{a_i^U} f(x)dx, i \in [1, 2, ..., N]. \tag{2.8}$$

où a_i est défini comme $\{a_i|x \in [a_i^L, a_i^U]\}$ et $f(x)$ est la fonction de distribution de densité de probabilité de x.
Pour les paramètres aléatoires, les valeurs de BPA dans les intervalles considés sont égales à la surface au-dessous de la pdf. La propagation de l'incertitude pour ces paramètres est estimée par le produit cartésien [7] donné par l'équation suivante :

$$C = a \times e = \{c_{ij} = a_i \times e_j\} \tag{2.9}$$

où C représente le produit cartésien des paramètres incertains et c_{ij} est l'élément de C.

- **Cas de non corrélation entre les paramètres**
  Dans le cas où les paramètres sont corrélés, nous proposons d'utiliser le modèle ellipsoïdal [46]. La BPA conjointe est définie par l'équation suivante :

  $$m(c_{ij} \cap \Omega) = \frac{m(a_i) \times m(e_j)}{S}, c_{ij} \cap \Omega \neq 0 \tag{2.10}$$

  où S est un facteur permettant de normaliser les BPA à une valeur égale à 1, S est présenté par l'équation suivante :

  $$S = \sum_{c_{ij} \cap \Omega \neq 0} m(c_{ij}) \tag{2.11}$$

- **Cas de corrélation entre les paramètres**
  Après le calcul de la BPA conjointe de tous les paramètres incertains, la distribution de sortie $Y = f(a, e)$ est calculée par l'équation suivante :

  $$[Y_{min}, Y_{max}] = [\min_{\mathbf{x} \in c_{ij}} f(\mathbf{X}), \max_{\mathbf{x} \in c_{ij}} f(\mathbf{X})] \tag{2.12}$$

La valeur de *Bel* est égale à la somme de toutes les valeurs de BPA conjointes des éléments de X tout en considérant que la valeur de Y_{max} est inférieure à la valeur limite (Y_{limite}). La valeur de *Pl* est égale à la somme des BPA conjointes des éléments tout en considérant que la valeur Y_{limite} est supérieure à la valeur Y_{min}.

2.3.2 Propagation des incertitudes liées au modèle

Le modèle de cos peut subir des incertitudes au moment de la modélisation. Ces incertitudes liées à la structuration du modèle viennent s'ajouter aux incertitudes liées aux paramètres. Dans notre travail, nous proposons de modéliser aussi les incertitudes liées au modèle de COS.

2.3.2.1 Structure du modèle et incertitude

Les différentes structures du modèle de COS se distinguent chacune par un ensemble des hypothèses qui diffèrent d'une structure à une autre. Soit un ensemble de K hypothèses $\{H_i\}_{1\leq i\leq K}$ permettant d'obtenir K structures différentes du modèle de COS. Notons par $\{f_i\}_{1\leq i\leq K}$ ces différentes structures comme illustrée par la figure 2.2.

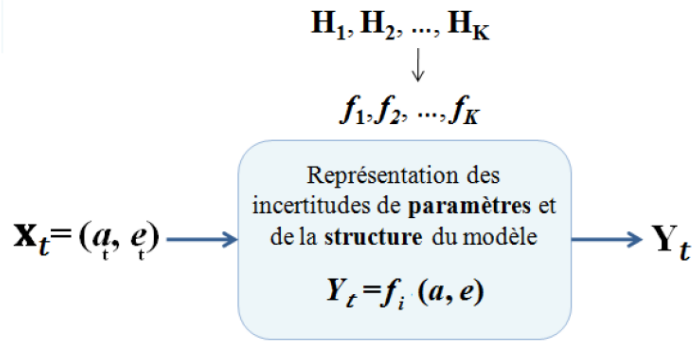


Figure 2.2 — Incertitude de la structure du modèle.

2.3.2.2 Propagation de l'incertitude de la structure du modèle

Les incertitudes liées à la structure du modèle sont propagées à l'aide de la théorie des fonctions de croyance. Supposons que nous avons un ensemble de structures du modèle $f_k(1 \leq k \leq K)$. Chacune de ces structures représente le modèle f_k tout en considérant un ensemble d'hypothèses. Les incertitudes liées aux paramètres sont propagées à à travers ces différentes structures du modèle ce qui permet d'obtenir un ensemble de représentations pour la variable de sortie Y. La différence entre ces représentations illustre l'effet des incertitudes de la structure du modèle. Pour modéliser

les incertitudes liées à ces structures, nous considérons que chaque structure est localisée dans un intervalle ayant une limite inférieure Bel et une limite supérieure Pl.
La combinaison de ces différentes représentations de fonctions de croyance et de plausibilité est donnée par les équations suivantes :

$$Bel^*(Y) = min(Bel_1(Y), Bel_2(Y), ..., Bel_K(Y)) \tag{2.13}$$

$$Pl^*(Y) = max(Pl_1(Y), Pl_2(Y), ..., Pl_K(Y)) \tag{2.14}$$

Les fonctions de croyance et de plausibilité obtenues permettent de prendre en considération les incertitudes de paramètres et de la structure du modèle dans la sortie.

La différence entre les valeurs de $Bel^*(Y)$ et $Pl^*(Y)$ représente le degré de l'incertitude des paramètres et de la structure du modèle de COS. Nous avons choisi d'utiliser la distance Kolmogorov-Smirnov (KS) [6] pour estimer les variations entre les valeurs de croyance et de plausibilité. La distance de KS est calculée comme suit :

$$d_{KS} = max_Y[Pl^*(Y)] - Bel^*(Y)] \tag{2.15}$$

D'après cette équation, la distance KS représente la différence entre les courbes de $Bel^*(Y)$ et $Pl^*(Y)$.

2.4 Expérimentations

Dans cette section, nous nous intéressons aux quatre étapes suivantes : la modélisation de paramètres incertains, l'étude de corrélation, la propagation des incertitudes de paramètres et la propagation des incertitudes de la structure du modèle. Nous proposons d'appliquer notre approche sur les quatre modèles de COS déjà décrits dans le premier chapitre qui sont DINAMICA, SLEUTH, CA-MARKOV et LCM.
Dans ce qui suit, nous allons prendre les paramètres suivants comme paramètres d'entrée pour les quatre modèles choisis :

- Paramètres spectraux : BI, MB, MR, MG et MN.
- Paramètres de texture : Ctr, Ent et ASM.
- Paramètres de forme : RF, EF, SI, D et A.
- Paramètre de végétation : NDVI.
- Paramètres climatiques : Tem et Hum.

Sur la base des 16 paramètres susmentionnés, quatre type de sols sont identifiés dans la zone du Caire à savoir urbain, agriculture, désert et eau.

2.4.1 Modélisation de paramètres incertains

Dans cette étude, nous ne considérons que les paramètres spectraux, de végétation et climatiques qui sont entachés par des incertitudes de type aléatoire. En effet, c'est la

variabilité naturelle qui influe le plus sur ces paramètres. Ces derniers sont représentés par des distributions normales de probabilité. Cependant, les paramètres de texture et de forme sont entachés par des incertitudes de type épistémiques. C'est le manque de connaissances qui est à l'origine de ces incertitudes. Les paramètres de texture et de forme sont décrits par des intervalles.
La figure 2.3 illustre les courbes représentant les paramètres d'entrée pour chaque type de l'occupation des sols.

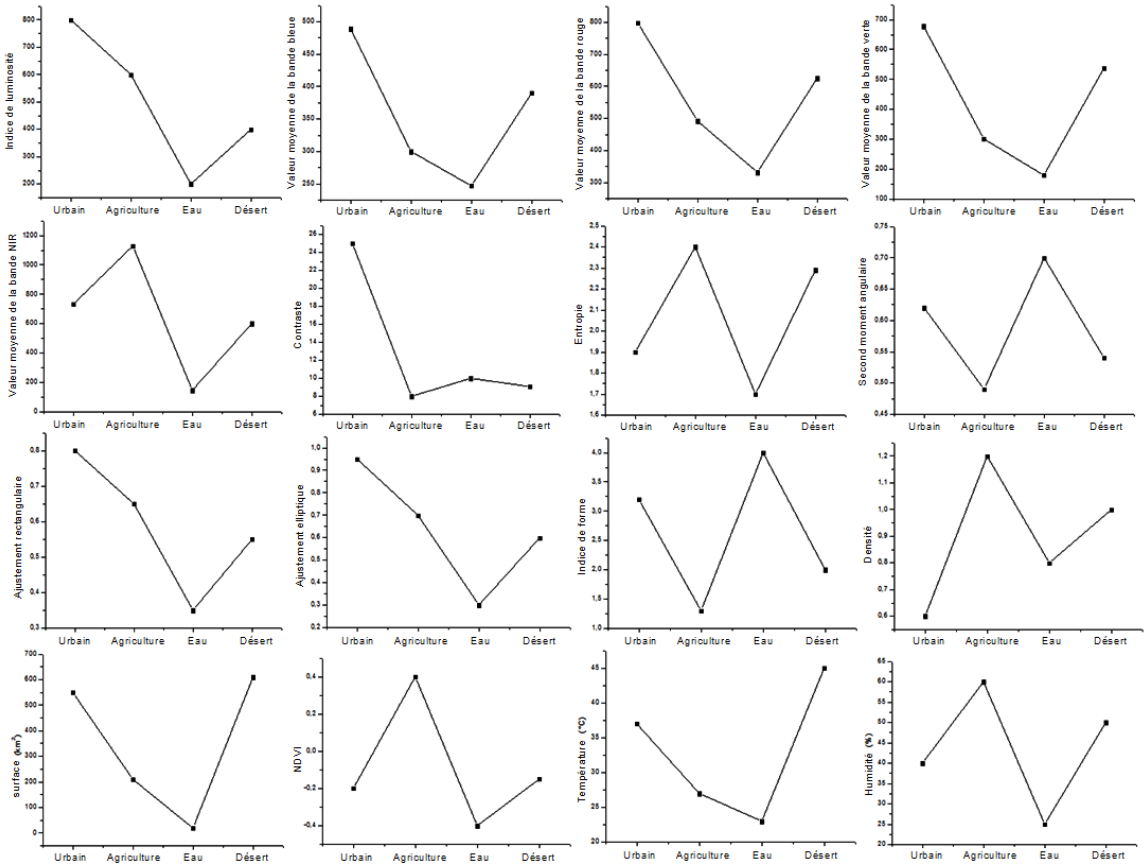


Figure 2.3 — Courbes représentant les paramètres d'entrée pour chaque type de l'occupation des sols.

Le tableau 2.1 illustre les valeurs incertaines de tous les paramètres aléatoires et épistémiques pour chaque type de l'occupation des sols.

2.4.2 Etude de la corrélation

Le but de l'analyse de corrélation est de mesurer la dépendance entre les paramètres des modèles de COS.
Le tableau 2.2 décrit les corrélations entre les paramètres d'entrée considérés dans cette étude. L'intensité de la couleur grise désigne l'importance de la corrélation entre

Paramètres	Moyenne				Ecart type				Intervalle			
	Urb	*Agr*	*Eau*	*Dst*	*Urb*	*Agr*	*Eau*	*Dst*	*Urb*	*Agr*	*Eau*	*Dst*
BI	792	602	200	398	170	214	34	72	-	-	-	-
MB	489	301	247	390	173	21	16	70	-	-	-	-
MR	798	492	330	626	317	49	37	158	-	-	-	
MG	678	302	180	538	274	48	33	181	-	-	-	-
MN	733	1132	147	602	282	117	22	214	-	-	-	-
Ctr	-	-	-	-	-	-	-	-	[24,26]	[7.5,9]	[9.5,10.5]	[8,8.5]
Ent	-	-	-	-	-	-	-	-	[1.7,2]	[2.2,2.5]	[1.5,2]	[2.25,2.3]
ASM	-	-	-	-	-	-	-	-	[0.6,0.65]	[0.47,0.5]	[0.7,0.75]	[0.5,0.55]
RF	-	-	-	-	-	-	-	-	[0.7,0.8]	[0.6,0.7]	[0.3,0.4]	[0.5,0.6]
EF	-	-	-	-	-	-	-	-	[0.9,1]	[0.6,0.7]	[0.2,0.3]	[0.4,0.5]
SI	-	-	-	-	-	-	-	-	[3,3.5]	[1,1.5]	[3.8,4.2]	[2,2.5]
D	-	-	-	-	-	-	-	-	[0.5,0.8]	[1,1.4]	[0.7,0.9]	[0.9,1.2]
A	-	-	-	-	-	-	-	-	[500,600]	[200,300]	[20,30]	[600,700]
NDVI	-0.19	0.43	-0.2	0.28	0.13	0.26	0.25	0.20	-	-	-	-
Tem	36.48	28.4	26.5	42.8	1.9	0.8	1.7	0.83	-	-	-	-
Hum	39.18	60	25.5	49	1.02	1.9	1.7	0.8	-	-	-	-

Tableau 2.1 — Valeurs des paramètres d'entrée pour chaque type de l'occupation des sols.

les paramètres. Une case de ce tableau avec une clouleur blanche signifie qu'il y a une faible corrélation entre les paramètres.
Nous pouvons constater que les paramètres MB/BI, MR/BI, MR/MB, etc, présentent une forte corrélation. Les paramètres Ctr/BI, ASM/BI, SI/BI, NDVI/MB, NDVI/MG, NDVI/Ctr, NDVI/Ent, NDVI/ASM, NDVI/A et Tem/D sont moins corrélés.

2.4.3 Propagation des incertitudes de paramètres

La propagation des incertitudes liées aux paramètres d'entrées des qutres modèles de COS déja cités est faite par la théorie des fonctions de croyance.
Afin de montrer le besoin de propager les incertitudes aléatoires et épistémiques à travers les modèles de COS, nous allons considérer le cas pour lequel tous les paramètres d'entrée sont entachés par des incertitudes de type aléatoire. Des distributions normales de probabilité sont calculées pour ces paramètres. La fonction de distribution cumulative (CDF) de la sortie de chaque modèle, tout en considérant seulement l'incertitude liées aux paramètres d'entrée, résulte de l'application de la théorie des fonctions de croyance.
La figure 2.4 montre les CDFs pour les quatre modèles de COS : DINAMICA, SLEUTH, CA-MARKOV et LCM en se basant sur 10.000 échantillons.

Pour montrer l'intérêt de la prise en compte des deux types d'incertitudes qui sont liées aux paramètres d'entrée du modèle de COS, les résultats précédents (cas de tous paramètres entachés par des incertitudes de type aléatoire) sont comparés avec ceux de la propagation des deux types d'incertitudes. La modélisation des incertitudes de type aléatoire est assurée par l'utilisation des distributions normales de probabilité, alors

	BI	MB	MR	MG	MN	Ctr	Ent	ASM	RF	EF	SI	D	A	NDVI	Tem	Hum
BI	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
MB	0.98	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-
MR	0.99	0.96	1	-	-	-	-	-	-	-	-	-	-	-	-	-
MG	0.98	0.94	0.95	1	-	-	-	-	-	-	-	-	-	-	-	-
MN	0.98	0.95	0.95	0.98	1	-	-	-	-	-	-	-	-	-	-	-
Ctr	0.25	0.85	0.83	0.81	0.79	1	-	-	-	-	-	-	-	-	-	-
Ent	0.11	0.76	0.72	0.81	0.72	0.89	1	-	-	-	-	-	-	-	-	-
ASM	0.29	0.85	0.83	0.87	0.79	0.97	-0.98	1	-	-	-	-	-	-	-	-
RF	0.12	0	0	0	0	0	0	0	1	-	-	-	-	-	-	-
EF	0.18	0	0	0	0	0	0	0	0.99	1	-	-	-	-	-	-
SI	0.26	0	0	0	0	0	0	0	0.89	0.90	1	-	-	-	-	-
D	0.06	0	0	0	0	0	0	0	0.98	0.99	0.98	1	-	-	-	-
A	0	0	0	0	0	0	0	0	0.76	0.76	0.78	0.76	1	-	-	-
NDVI	0.82	0.42	0.98	0.41	0.98	0.32	0.28	0.32	0.01	0.01	0.01	0.11	0.21	1	-	-
Tem	0.01	0	0	0	0	0	0	0	0	0	0	0.21	0	0.78	1	-
Hum	0.08	0	0	0	0	0	0	0	0	0	0	0.02	0	0.81	-0.80	1

Tableau 2.2 — Coefficients de corrélations entre les paramètres d'entrée du modèle de COS.

que la modélisation des incertitudes de type épistémiques est assurée par l'utilisation des intervalles. Ainsi, 10 000 échantillons sont pris pour les deux types d'incertitudes. Les CDF produites de quatre modèles de COS sont illustrées au niveau de la figure 2.5.

Les courbes de distributions de croyance et de plausibilité représentent l'intervalle d'incertitudes des sorties des modèles de COS. Les résultats obtenus dans la figure 2.5 indiquent que les quatre modèles sont plus sensibles à la considération de deux types d'incertitudes. D'autres part, les paramètres des modèles de COS sont généralement corrélés. La prise en considération de ces corrélations demeure essentielle dans l'étude de la propagation des incertitudes. Selon le tableau 2.2, SI est fortement corrélé avec A, D, RF et EF. D'autre part, nous avons des corrélations importantes entre le NDVI et Tem et Hum et entre les paramètres spectraux et les paramètres de texture.

Au niveau de la figure 2.6 est présentée la comparaison des distributions de sortie des quatre modèles de cos dans les deux cas à savoir la considération et la non considération des corrélations entre les paramètres. Les résultats obtenus confirment l'importance de la considération de la corrélation entre les paramètres. Néamoins, il est essentiel de déterminer la première source de variation de la sortie des modèles (ce qui nécessite une considération individuelle des paramètres ou une considération d'un groupe de paramètres ensemble).

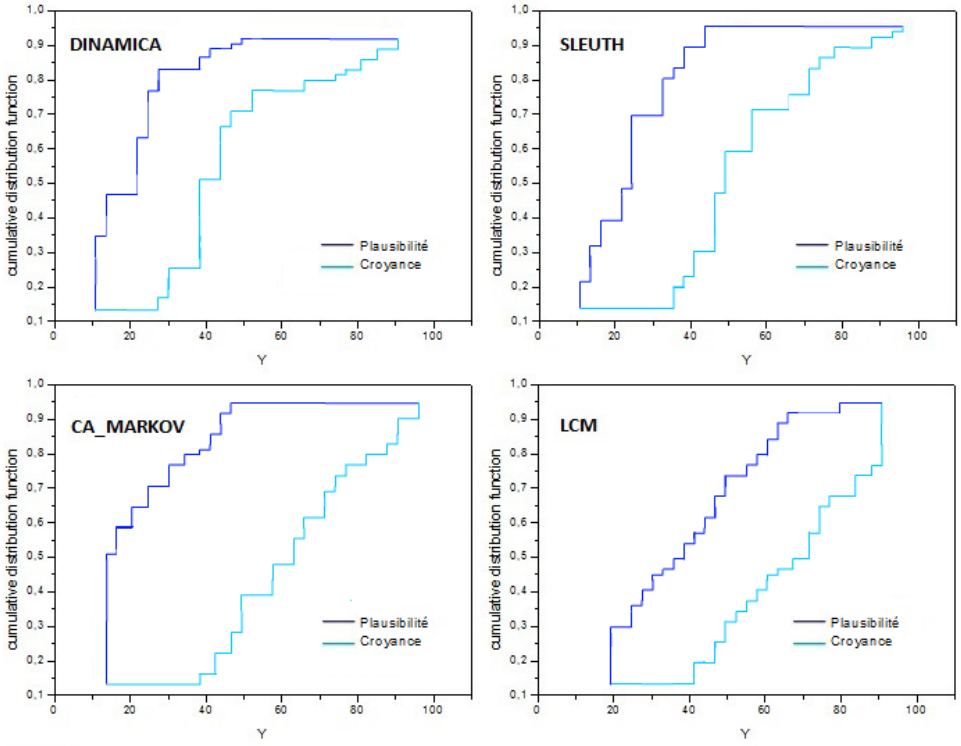


Figure 2.4 — Distributions de croyance des modèles de COS avec prise en compte de la propagation des incertitudes aléatoires de paramètres.

2.4.4 Propagation des incertitudes relatives à la structure de modèle

A fin de mettre en évidence la puissance de la modélisation des incertitudes relatives à la structure du modèle, trois structures différentes sont utilisés pour les modèles de COS considérés dans cette étude. Dans l'objectif de déterminer l'effet du changement de la structure de modèle sur le résultat de sortie, nous propageons les incertitudes des paramètres pour les trois structures et nous comparons les résultats obtenus. La différence dans le résultat illustre l'impact du changement de la structure sur le résultat final. Cette différence représente les incertitudes liées à la structure du modèle. La figure 2.7 présente le résultat de la propagation des incertitudes des paramètres pour les trois structures de chacun des modèles de COS : DINAMICA, SLEUTH, CA-MARKOV et LCM. Nous constatons que les résultats obtenus pour les quatre modèles sont très proches de fait qu'il n'y a pas un grand écart entre les hypothèses utilisées et les conditions d'origine. Propager les incertitudes liées aux structures des modèles permet de déterminer l'effet de chaque structure sur les résultats de sortie.

La figure 2.8 illustre une comparaison des distributions de sortie sous forme de croyance et de plausibilité pour les quatre modèles de COS. Nous considérons deux cas d'étude : propagation d'incertitudes liées aux paramètres et à la structure, et propagation d'incertitudes liées aux paramètres. Les résultats obtenus permet de constater

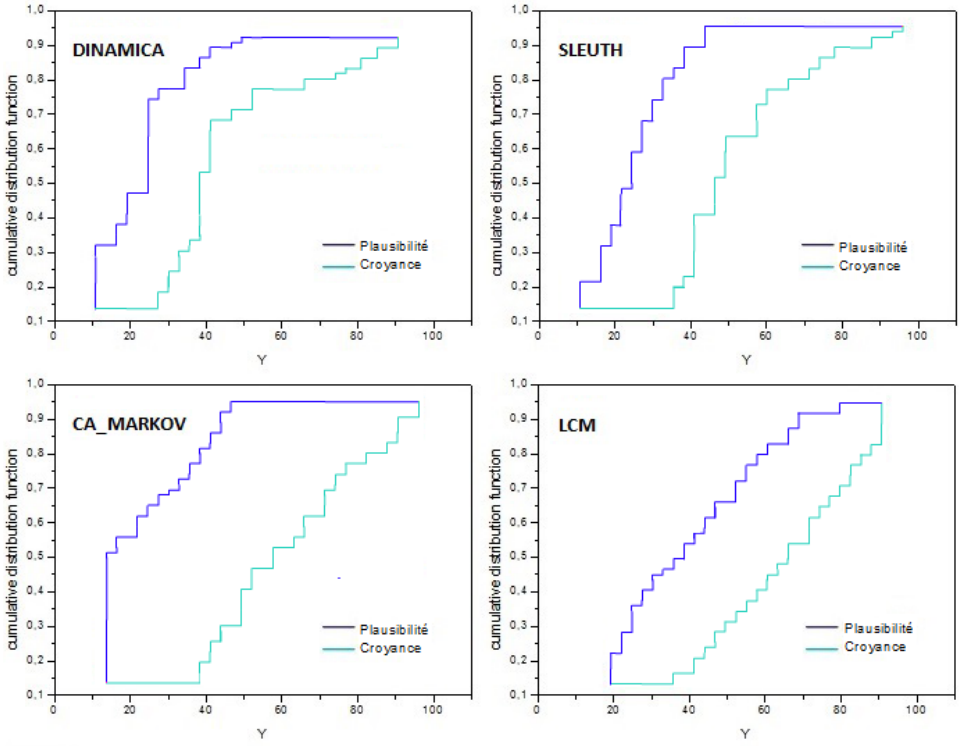


Figure 2.5 — Distributions de croyance et de plausibilité des modèles de COS avec prise en compte de la propagation des incertitudes de paramètres.

que la sortie des modèles de COS est plus influencée dans le cas de propagation des incertitudes des paramètres et de structure que celle des paramètres seuls.

2.5 Conclusions et perspectives

Dans le présent chapitre, un état de l'art sur les méthodes de propagation des incertitudes et les problématiques à résoudre sont détaillés. Ensuite, nous avons détaillé notre approche de propagation des incertitudes. L'approche proposée prend en considération les deux types d'incertitude, la corrélation entre les paramètres, les incertitudes liées aux paramètres et les incertitudes liées aux modèles de COS. Nous avons validé notre approche en utilisant quatre modèles de COS qui sont DINAMICA, SLEUTH, CA-MARKOV et LCM. Les résultats de l'approche proposée montrent l'importance de la prise en compte de deux types d'incertitude liées aux paramètres d'entrée et ce pour les quatre modèles de COS considérés. Par ailleurs, les résultats obtenus confirment l'importance de la prise en considération des corrélations entre les paramètres d'entrée lors de la propagation des incertitudes. Finalement, la modélisation des incertitudes de la structure du modèle est essentielle pour améliorer les décisions des modèles de COS. Dans ce travail, nous avons appliqué la propagation des incertitudes sur les quatre

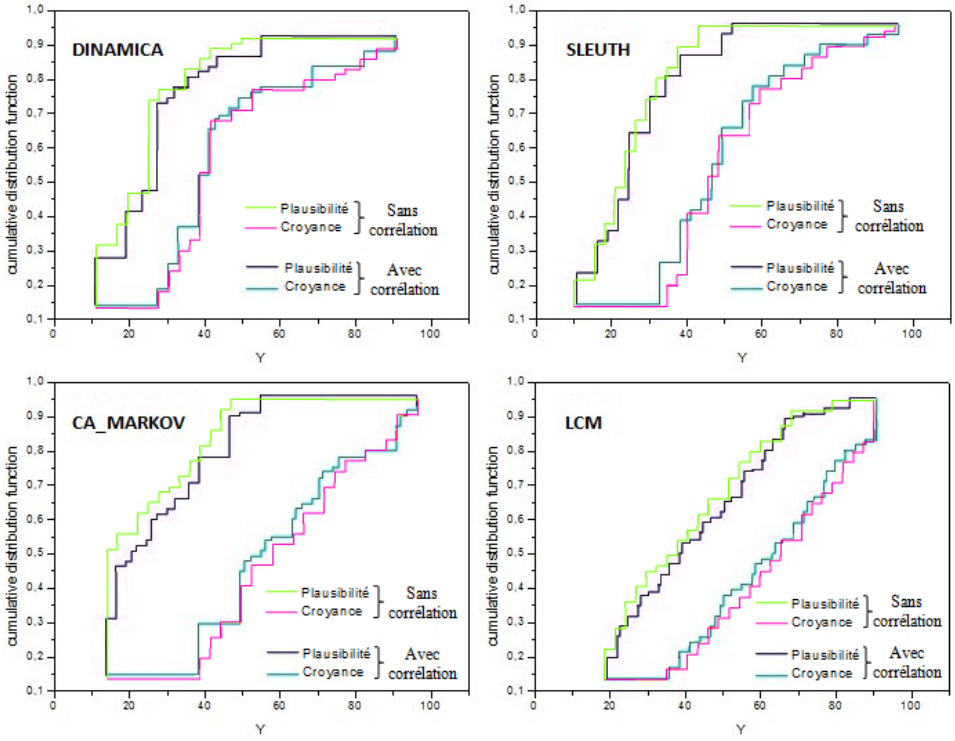


Figure 2.6 — Distributions de croyance et de plausibilité des modèles de COS.

modèles DINAMICA, SLEUTH, CAMARKOV et LCM. Ces modèles sont basés sur les automates cellulaires, la régression logistique et les chaines de Markov. Les travaux futurs devraient également appliquer l'approche proposée sur d'autres modèles très populaires de COS tels que les modèles basés sur l'apprentissage automatique.
De plus, l'apport principal de cette étude consiste à appliquer la propagation des incertitudes tout en considérant à la fois les diffèrent types d'incertitudes liées aux paramètres et au modèle lui-même et la corrélation entre les paramètres. Cependant, la complexité de l'approche proposée n'est pas considérée. Ainsi, il est essentiel d'aborder ce problème et d'étudier les coûts de calcul de l'approche proposée.

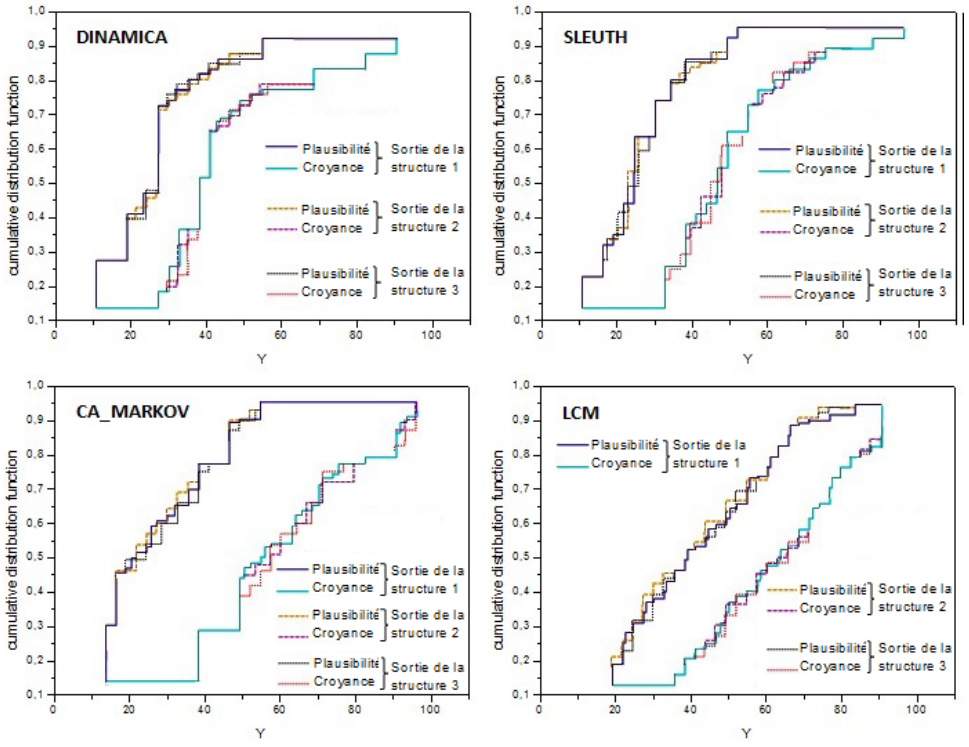


Figure 2.7 — Distributions de sortie des trois structures des modèles de COS.

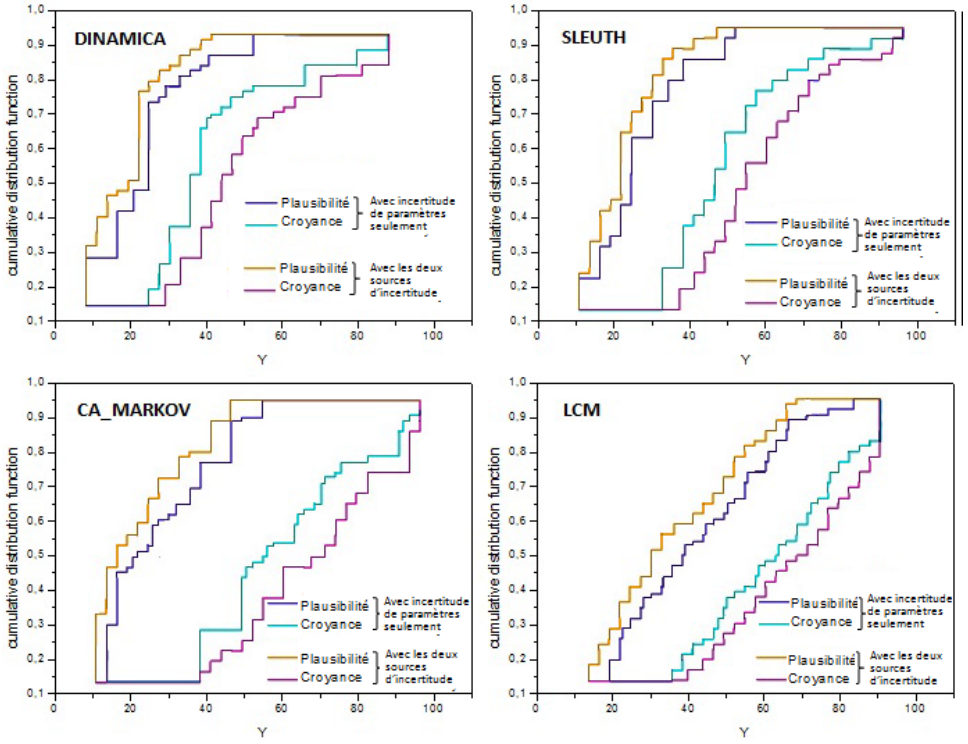


Figure 2.8 — Comparaison des distributions de sortie des modèles de COS avec propagation des incertitudes (paramètres avec/sans structure).

Chapitre 3

Réduction des incertitudes

Sommaire

3.1 Introduction

La modélisation des phénomènes physiques est généralement une tâche très complexe du fait du grand nombre de paramètres d'entrée, qui sont le plus souvent corrélés entre eux et entachés par plusieurs types d'incertitudes. Ceci affecte les modèles représentant les phénomènes physiques et par suite la qualité des décisions prises.
Dans ce chapitre, nous proposons de réduire les incertitudes dans deux domaines. Le premier correspond à la prédiction de COS, dont le but est de réduire les incertitudes liées d'une part aux paramètres d'entrée de modèles de COS et d'autre part aux modèles eux-mêmes. Le deuxième domaine est relatif à la bio-informatique et plus précisément à l'identification des biomarqueurs biologiques potentiels causant le cancer du poumon.
Nous commençons par décrire l'approche proposée pour la réduction des incertitudes dans le domaine de COS. Ensuite, nous détaillons l'approche proposée pour la réduction des incertitudes dans le domaine de la bio-informatique. La dernière partie sera consacrée à la présentation des conclusions et des perspectives.

3.2 Approche proposée pour la réduction des incertitudes dans le domaine de COS

Dans cette première partie, nous nous intéressons à la réduction des incertitudes dans le domaine de COS. Nous commençons par présenter un état de l'art sur les travaux se rapportant à la réduction des incertitudes dans le domaine de COS et les problématiques à résoudre. Ensuite, nous détaillons deux approches proposées pour la réduction des incertitudes dans le domaine de COS. La première approche correspond à la méthode de sensibilité globale basée sur les dérivées. La deuxième approche concerne la méthode des fonctions de croyance.

3.2.1 Etat de l'art et problématiques étudiées

L'analyse de sensibilité revient à étudier l'effet des perturbations et des incertitudes liées aux entrées sur la sortie d'un modèle. Elle étudie, qualitativement ou quantitativement, la manière dont les variations des entrées d'un modèle engendrent des variations de sa sortie. A ce niveau, l'analyse de sensibilité peut être utilisée pour varier les paramètres d'entrée du modèle de COS et identifier ceux qui ont une forte influence sur la sortie du modèle [18] [17] [48].
De nombreux travaux ont confirmé l'importance de l'utilisation de l'analyse de sensibilité pour comprendre l'effet des différentes sources d'incertitude sur les sorties du modèle [102] [111] [91] .
Dans la littérature, nous trouvons une multitude de méthodes d'analyse de sensibilité. Parmi ces méthodes, nous citons les méthodes de criblage [104], les méthodes d'analyse différentielle [57], les méthodes basées sur l'échantillonnage [61] et les méthodes basées sur l'entropie [83]. Ces méthodes déjà citées font partie des méthodes probabilistes.

D'autre part, des méthodes non probabilistes d'analyse de sensibilité sont proposées [73] [31] [122].
Une étude détaillée sur les différentes familles des méthodes d'analyse de sensibilité est présentée dans les travaux suivants [72] [98].
Afin de proposer une approche basée sur l'analyse de sensibilité pour les modèles de prédiction de COS, il faut résoudre les problèmes suivants :

- Les modèles de COS disposent le plus souvent d'un nombre important de paramètres d'entrée. Il est essentiel de réduire le nombre de ces paramètres sans affecter les résultats de la sortie du modèle.
- Les paramètres d'entrée des modèles de COS sont le plus souvent corrélés et entachés par plusieurs types d'incertitude. Il est important d'identifier les corrélations qui existent entre ces paramètres. De plus, toute modélisation et réduction des incertitudes des paramètres implique l'identification des types d'incertitudes qui sont liées à ces paramètres.
- Etudier les incertitudes qui sont liées au modèle lui-même. Ces incertitudes, si elles sont ignorées, vont influencer les décisions sur les changements de l'occupation des sols.

Dans nos travaux, nous avons choisi de travailler avec deux méthodes d'analyse de sensibilité à savoir :

1. la méthode de sensibilité globale basée sur les dérivées (DGSM)
2. la théorie des fonctions de croyance.

3.2.2 Méthode de sensibilité globale basée sur les dérivées (DGSM)

Dans notre première approche de réduction des incertitudes, nous nous intéressons uniquement aux incertitudes liées aux paramètres d'entrée du modèle de COS. Nous considérons les corrélations qui peuvent exister entre ces paramètres.
La figure 3.1 présente les étapes de l'approche proposée qui est basée sur la méthode de sensibilité globale des dérivées. Notre approche est divisée en quatre étapes principales :

1. analyse de sensibilité qualitative
2. mesure de corrélation
3. propagation d'incertitudes
4. analyse de sensibilité quantitative.

L'analyse de sensibilité qualitative vise l'identification de la liste des paramètres incertains alors que l'analyse de sensibilité quantitative détermine les paramètres les plus influents sur la sortie du modèle de COS.

3.2.2.1 Analyse de sensibilité qualitative

L'objectif de cette étape est d'identifier les paramètres incertains du modèle de COS. Pour cela, nous utilisons la méthode de criblage de Morris. Notre choix est jus-

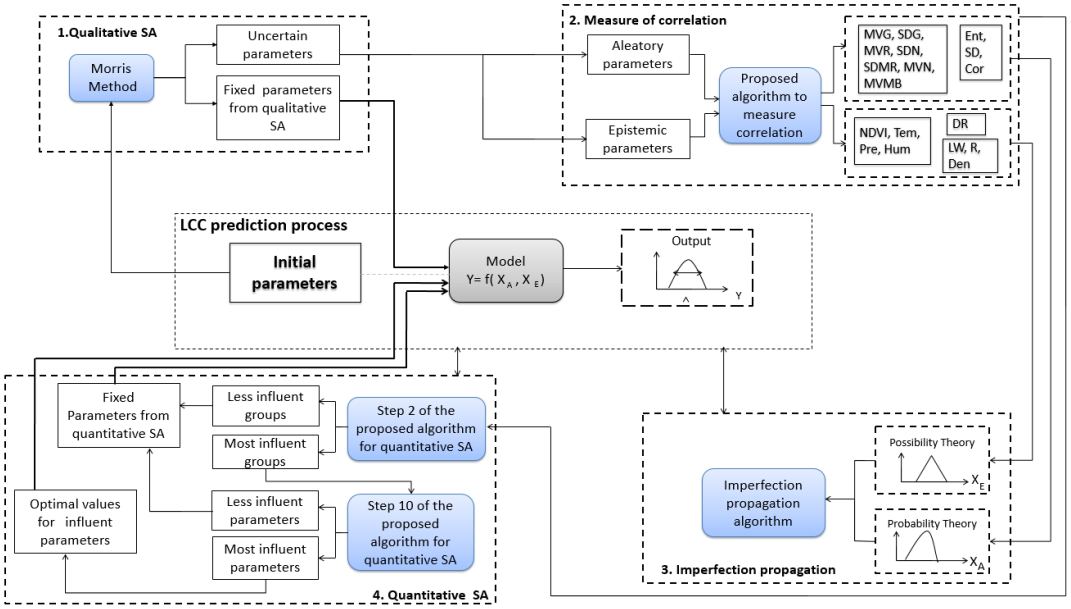


Figure 3.1 — Approche proposée de sensibilité globale basée sur les dérivées (DGSM).

tifié par sa large utilisation dans les problèmes d'analyse de sensibilité qualitative [72]. La méthode Morris est basée sur des perturbations des paramètres du modèle. Ainsi, pour chaque perturbation un effet élémentaire (EE) est introduit. Ce dernier représente la variation de la valeur de la sortie générée par la perturbation du $i^{\text{ème}}$ paramètre et est répétée r fois à différents endroits dans l'espace des paramètres.
Si nous dénotons par F notre modèle de COS et par $X = (X_1, X_2, ..., X_n, X_{n+1}, ..., X_K)$ les paramètres d'entrée au modèle F ; avec $X_(i \in [1..n])$ sont les paramètres ayant des incertitudes de type aléatoire et $X_(i \in [n+1..K])$ sont les paramètres ayant des incertitudes de type épistémiques. Nous supposons que l'espace des valeurs des paramètres d'entrée est discrétisé en p-niveau Ω.

$$Y = F(X_1, X_2, ..., X_n, X_{n+1}, ..., X_K) \tag{3.1}$$

L'effet élémentaire est défini par l'équation suivante :

$$EE_i(X) = \frac{Y(X_1, X_2, ..., X_{i-1}, X_i + \delta, X_{i+1}, ..., X_K)}{\delta} \tag{3.2}$$

Où
δ est la valeur comprise entre $\{\frac{1}{p-1}, ..., 1 - \frac{1}{p-1}\}$, p est le nombre de niveaux dans Ω , $X = (X_1, X_2, ..., X_K)$ toute valeur prise dans Ω de telle sorte que la valeur $(X + e_i)$ soit toujours présent dans Ω et e_i est un vecteur de zéros mais avec une unité au niveau de la $i^{\text{ème}}$ composante. Pour chacun des paramètres K, la mesure de sensibilité est résumée par la moyenne (μ^*) et l'écart-type (σ) de la distribution des EE respectivement définis

comme :

$$\sigma_i^* = \frac{1}{r}\sum_{j=1}^{r}|EE_i^{(j)}| \tag{3.3}$$

Et

$$\sigma_i = \sqrt{\frac{1}{r-1}\sum_{j=1}^{r}(EE_i^{(j)} - \mu_i^*)^2} \tag{3.4}$$

Où
*EE*s sont définis pour chaque paramètre d'entrée $EE_i(X^{(1)}), EE_i(X^{(2)}), ..., EE_i(X^{(r)})$ par un échantillonnage aléatoire de r points $X^{(1)}, X(2), ..., X^{(r)}$ (r est compris entre 4 et 10 dans [102]).
L'influence d'un paramètre d'entrée X_i est proportionnelle à la distance entre le point (μ_i^*, σ_i) et l'origine.
Il est à noter que les détails de l'application de la méthode de Morris à notre modèle de COS sont présentés au niveau de l'étude réalisée par Boulila et Collaborateurs [17].
La AS qualitative est utilisée pour identifier les paramètres influents. Des valeurs prédéfinies sont fixées pour les valeurs des paramètres non influents. Cela permet de limiter le nombre de paramètres d'entrée à prendre en compte.
Une fois les paramètres influents sont identifiés, l'étape suivante est de mesurer la corrélation entre ces paramètres afin de les regrouper en des groupes de paramètres corrélés.

3.2.2.2 Mesure de corrélation

Dans le contexte du COS, les modèles contiennent généralement un grand nombre de paramètres d'entrée corrélés. Par exemple, l'étalonnage radiométrique peut entraîner une modification des caractéristiques de la texture ou de la forme. En outre, les mesures radiométriques dépendent extrêmement du spectre et par conséquent de la bande et du capteur. De plus, les conditions météorologiques et atmosphériques peuvent être une source de variabilité des paramètres texturaux et spectraux. Ainsi, considérer que les paramètres d'entrée des modèles de COS sont indépendants peut entraîner des décisions erronées concernant les changements de l'occupation des sols. En présence d'incertitudes, l'étude de la corrélation entre les paramètres d'entrée devient encore plus difficile.
Nous commençons par identifier le type d'incertitude (aléatoire ou épistémique) pour chaque paramètre influent identifié. Le processus de mesure de la corrélation est basé sur le tau de Kendall comme illustré par l'algorithme 3.

En cas d'incertitude aléatoire, nous appliquons le tau de Kendall pour calculer les degrés de corrélation entre les paramètres caractérisés par des incertitudes aléatoires.
Soit $(X_i, X_j)_{1\leq i,j\leq n}$ un échantillon représentant des paires de valeurs prises parmi les paramètres aléatoires. Le nombre total de paires possibles est $n(n-1)/2$. Le coefficient

Algorithme 1 Mesure de corrélation

Entrée: Vecteur de paramètres épistémiques ou aléatoires
Sortie: Matrice de corrélation tau

Si les paramètres sont aléatoires **alors**
  Pour tout paramètres aléatoires **faire**
    Mettre les données dans des vecteurs à deux dimensions
    Calculer le tau de Kendall classique entre X_i et X_j ; $1 \leq i, j \leq n$
  Fin Pour
  Renvoyer la matrice du diagramme de dispersion du tau de Kendall entre deux paramètres aléatoires différents
Sinon
  Si les paramètres sont épistémiques **alors**
    Pour tout paramètres épistémiques **faire**
      Calculer le tau de Kendall dérivé entre X_i et X_j ; $n+1 \leq i, j \leq K$
    Fin Pour
    Renvoie la matrice du tau de Kendall dérivé entre deux paramètres épistémiques différents
  Fin Si
Fin Si

de corrélation entre deux paramètres est calculé comme suit :

$$\tau = \frac{4}{n-1} \sum_{j=1}^{n-1} V_{ij} - 1 \tag{3.5}$$

Avec $V_{ij} = \frac{card\{p,q:X_p<X_i,X_q<X_j\}}{n-2}, 1 \leq i, j \leq n$
Cependant, en présence d'incertitudes épistémiques, la valeur obtenue du tau de Kendall sera un intervalle $\tau = [\tau_L; \tau_U]$, où les valeurs de τ_L et τ_U sont définies en remplaçant V_{ij} dans l'équation 3.2.2.2 par les valeurs suivantes :

$$V_{ij,L} = \min_{\substack{X_p \in [X_{p,L}, X_{p,U}] \\ X_q \in [X_{q,L}, X_{q,U}]}} \frac{card\{p, q : X_p < X_i, X_q < X_j\}}{n-2} \tag{3.6}$$

et

$$V_{ij,U} = \max_{\substack{X_q \in [X_{p,L}, X_{p,U}] \\ X_q \in [X_{q,L}, X_{q,U}]}} \frac{card\{p, q : X_p < X_i, X_q < X_j\}}{n-2} \tag{3.7}$$

L'étape de corrélation permet de regrouper les paramètres qui sont en interaction des groupes et ceci pour chaque type d'incertitudes. Ensuite, nous utilisons une méthode hybride qui propage les incertitudes de type aléatoire et épistémique séparément [17]. Les paramètres aléatoires sont représentés par des distributions de probabilité et les paramètres épistémique sont représentés par des distributions de possibilité.

3.2.2.3 Analyse de sensibilité quantitative

Le but de l'étape d'analyse de sensibilité quantitative est de déterminer les paramètres les plus influents sur le modèle COS. Pour ce faire, nous avons choisi de travailler avec la méthode de sensibilité globale basée sur les dérivées (DGSM). DGSM combine entre les méthodes d'analyse de sensibilité locales et globales. Elle fournit les mêmes informations que les méthodes basées sur la variance, en particulier la méthode Sobol qui est la méthode de référence dans la littérature. De plus, elle a l'avantage d'être plus efficace en termes de temps de calcul et ne perd pas sa fiabilité pour les systèmes assez complexes.
Le but de l'analyse de sensibilité quantitative consiste à évaluer dans un premier lieu l'impact de chaque groupe de paramètres sur les incertitudes de sortie du modèle de COS. Ensuite, de déterminer les paramètres les plus sensibles dans chacun de ces groupes. Cette étape est subdivisée en trois sous-étapes : 1) calcul de l'indice de sensibilité total ; 2) calcul de l'indice de sensibilité individuel et 3) recherche des valeurs optimales des paramètres qui ont plus d'influence sur la sortie du modèle de COS.
Le pseudocode pour la méthode proposée est présenté dans l'algorithme 2.

Nous considérons notre modèle $Y = F(G)$, avec $G = (G_1, ..., G_{K^"})$, après le regroupement de paramètres est un vecteur de groupes indépendants. Chaque groupe est composé d'un ensemble de paramètres dépendants $(X_1, ..., X_{K^"})$ avec une distribution $(\mu_{X_1}, ..., \mu_{X_{K^"}})$.

– **Calcul de l'indice de sensibilité totale**

Dans cette étape, nous effectuons un AS quantitative pour identifier les groupes les plus influents. Tous les paramètres appartenant à un groupe doivent être déplacés simultanément avant l'évaluation du modèle de COS et le calcul des indices de sensibilité. Pour accomplir cette tâche, la méthode originale de DGSM est modifiée pour s'adapter à la nature des paramètres d'entrée du modèle de COS. Ainsi, nous étendons la méthode de DGSM afin de calculer l'indice de sensibilité total pour chaque groupe de paramètres. La variance de la sortie du modèle de COS qui est due à l'interaction entre ces paramètres est définie comme suit :

$$S_g = \int \left(\frac{\partial F(g)}{\partial x} \right)^2 d\mu(g) \tag{3.8}$$

Tous les groupes ayant une petite valeur de S_g auront une faible influence sur la sortie du modèle de COS et ces groupes sont considérés comme négligeables. Ainsi, les valeurs de tous les paramètres de ces groupes sont définies comme constantes. Cependant, pour les groupes ayant une valeur importante de l'indice de sensibilité totale, nous devons calculer l'influence de chaque entrée sur la sortie.

– **Calcul de l'indice de sensibilité individuel**

Le but de cette étape est de quantifier l'influence de chaque paramètre dans le groupe le plus important identifié lors de l'étape précédente. Nous utilisons la méthode

Algorithme 2 Algorithme proposé pour AS

Entrée: échantillon N-dimension
Sortie: Paramètres d'influence avec des valeurs optimales

Générer la matrice de permutation P
Pour each groupe g dans G **faire**
  Calculer l'indice de sensibilité totale S_g
  //Classer les groupes en fonction de l'indice de sensibilité total
  Si (S_g tend vers 0) **alors**
    Etiqueter le groupe G_g comme non influent
    Maintenir les paramètres contenus dans G_g comme constantes
  Sinon
    Si (S_g tend vers 1) **alors**
      Etiqueter le groupe G_g comme non influent
      Pour each paramètre x dans g **faire**
        Calculer l'indice de sensibilité DGSM individuel pour x : S_x
        // Classer les paramètres en fonction de l'indice de sensibilité individuel
        Si (S_x tend vers 0) **alors**
          Etiqueter le paramètre x comme non influent et est maintenu constant
        Sinon
          Si (S_x tend vers 1) **alors**
            Etiqueter le paramètre x comme influent
            // Rechercher la valeur optimale de x
            $x \leftarrow x + \triangle$
            Tant que ($x < x_{max}$ et $\triangle > \triangle_{min}$) **faire**
              $\triangle \leftarrow \triangle/2$
              Si (le résultat de prédiction obtenu est différent du résultat de prédiction initial) **alors**
                $x \leftarrow x - \triangle$
              Fin Si
              $x \leftarrow x + \triangle$
            Fin Tant que
          Fin Si
        Fin Si
      Fin Pour
    Fin Si
  Fin Si
Fin Pour
Renvoyer l'ensemble des paramètres d'influence et leurs valeurs optimales

DGSM classique [12] pour calculer l'indice de sensibilité individuel S_x qui détermine la contribution du paramètre x à la variation de sortie. Plus la valeur de Sx est grande, plus l'influence de x sur la sortie du modèle est grande.

– **Recherche des valeurs optimales des paramètres les plus influents**

Dans cette étape, nous recherchons les valeurs optimales des paramètres les plus influents. Nous changeons la valeur initiale du paramètre, puis nous comparons le résultat obtenu avec les résultats initiaux du modèle de COS. S'il y a une différence, nous chan-

geons à nouveau la valeur du paramètre. Sinon, si la différence est acceptable (un seuil est défini à cet effet), cette valeur du paramètre est considérée comme la valeur optimale. Les intégrales de l'équation précédente sont évaluées comme suit [81] :

$$S_g = \frac{1}{N}\sum_{j=1}^{N}(\frac{\partial F(g)^2}{\partial x}) \tag{3.9}$$

3.2.3 Théorie des fonctions de croyance

La deuxième méthode que nous avons utilisé pour analyser la sensibilité est la méthode des fonctions de croyance. Nous avons opté à utiliser la stratégie du pincement permettant d'analyser la sensibilité dans le cadre de la théorie des fonctions de croyance. Ce choix est justifié par le fait que la stratégie du pincement permet d'évaluer les incertitudes (aléatoires et épistémiques) avec considération des corrélations entre les paramètres du modèle [48]. Cette stratégie permet d'évluer la valeur de l'information empirique en mesurant l'incertitude du modèle avant et après l'opération de pincement [51] [61] [3]. La stratégie du pincement sera utilisée pour analyser les incertitudes des paramètres et de modèle. L'écart entre la croyance (Bel) et la plausibilité (Pl) reflète le degré d'incertitude de paramètres. L'incertitude dans la sortie du modèle Y est évaluée par la fonction *unc()* présentée par l'équation suivante :

$$unc(Y) := \|Bel - Pl\|_1 = \int_{-\infty}^{\infty} |Bel - Pl| dx \tag{3.10}$$

unc(Y) calcule l'aire du domaine compris entre *Bel* et *Pl*. Ainsi, elle caractérise le degré d'incertitude de la sortie Y du modèle. *unc(Y)* est égale à zéro lorsque les intervalles focaux sont réduits à des singletons. Par conséquent, nous pouvons déduire la sensibilité S_i d'un paramètre d'entrée i sur la sortie comme suit :

$$S_i = 1 - \frac{unc(Y_i)}{unc(Y)} \tag{3.11}$$

Plus la valeur S_i est grande, plus l'influence de l'incertitude du paramètre d'entrée i sur la variable de sortie Y est importante. Nous évaluons l'impact d'incertitude des paramètres d'entrée un par un. Ensuite, nous utilisons la valeur nominale de chaque paramètre par recours à des avis d'expert. L'évaluation des sensibilités est assurée par l'estimation des fonctions de croyance et de plausibilité. Un des avantages majeurs de cette technique de pincement est sa capacité de travailler avec plusieurs paramètres en même temps ; qui sera utile pour le cas de la corrélation entre les paramètres d'entrée du modèle. De la même manière, la mesure de sensibilité S_i est utilisée pour déterminer la structure du modèle de COS la plus appropriée (qui minimise les incertitudes liées au modèle). Ceci permettra d'évaluer l'effet des incertitudes liées à la structure du modèle sur les résultats à la sortie. La dernière étape du processus de réduction des incertitudes consiste à estimer les valeurs optimales des paramètres d'entrée les plus influents sur

la sortie et de fixer les autres à des valeurs nominales.
Les paramètres les plus influents sont estimés en utilisant les limites de confiance du Kolmogorov-Smirnov ; ce qui permettra d'affecter un niveau de confiance pour chaque paramètre. Les intervalles de confiance du Kolmogorov-Smirnov d'une structure évidentielle est donnée par l'équation suivante [50] :

$$Min(1, max(0, D(p) \pm stat(\alpha, n))) \tag{3.12}$$

où $D(p)$ est une fonction de distribution, α est le niveau de confiance, et $stat(\alpha, n)$ est la statistique de Kolmogorov pour un niveau de confiance 100*(1- α)% et n est le nombre des intervalles. Les valeurs de $stat(\alpha, n)$ ont été proposées par Lilliefors [85].

3.2.4 Expérimentation des approches proposées

L'expérimentation des approches proposées est divisée en deux parties : l'expérimentation de l'approche d'analyse de sensibilité basée sur la sensibilité globale des dérivées (DGSM) et l'expérimentation de l'approche basée sur la théorie des fonctions de croyance.

3.2.4.1 Analyse de sensibilité basée sur la sensibilité globale des dérivées (DGSM)

Nous nous intéressons dans cette section à l'application de l'approche DGSM sur un jeu de données réel et à évaluer sa performance. Nous appliquons l'approche proposée sur le modèle de COS "FS-FDT" présenté dans le premier chapitre. Les paramètres d'entrée à ce modèle sont aussi illustrés dans le premier chapitre. Le but de l'application étant le suivi des changements de l'objet "urbain" dans la région du Port de l'ile de la Réunion.
Nous commençons, cette section, par une brève description de la région d'étude, le modèle de prédiction de COS, les paramètres d'entrée à ce modèle et leurs sources d'incertitudes. Ensuite, nous présentons les résultats expérimentaux de l'application de l'approche sur un jeu de données réel. La dernière partie est dédiée à l'évaluation de l'approche proposée en la comparant avec des méthodes existantes.

1. **Description de la région d'étude**

La zone d'étude est la région du "Port". Cette région est située sur l'extrême ouest de l'île de la Réunion (Figure 3.2). La région du "Port" est caractérisée par un climat tropical et un développement démographique atypique.
Les images satellitaires utilisées pour les expérimentations, proviennent du satellite Spot 5 (Système Pour l'Observation de la Terre). La figure 3.3 montre des images satellites acquises le 24 juillet 2008 et le 12 août 2014. Les deux images ont une résolution spatiale de 10 m et une taille de 600X800 pixels.

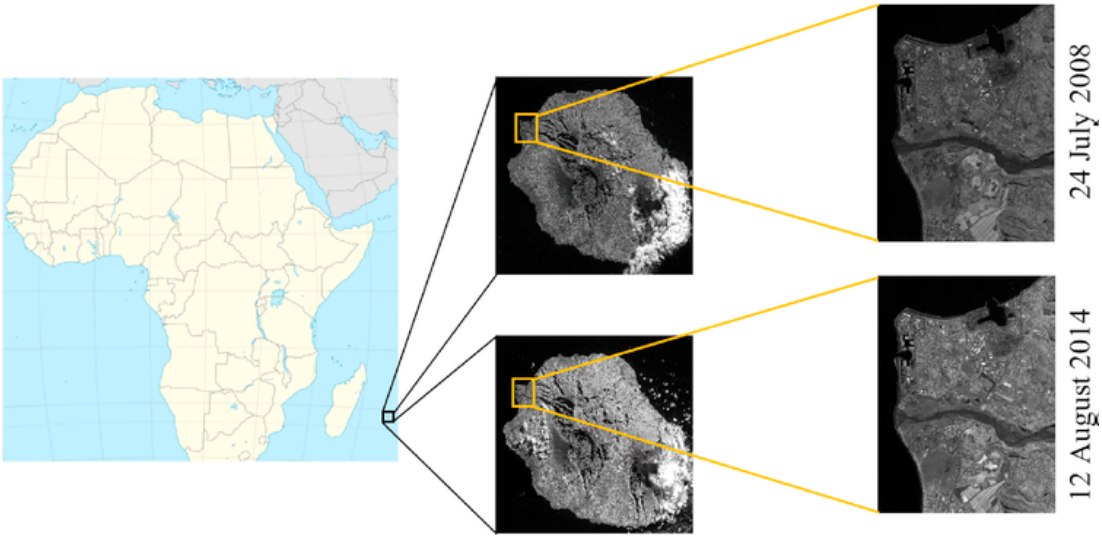


Figure 3.2 — Présentation de la région d'étude.

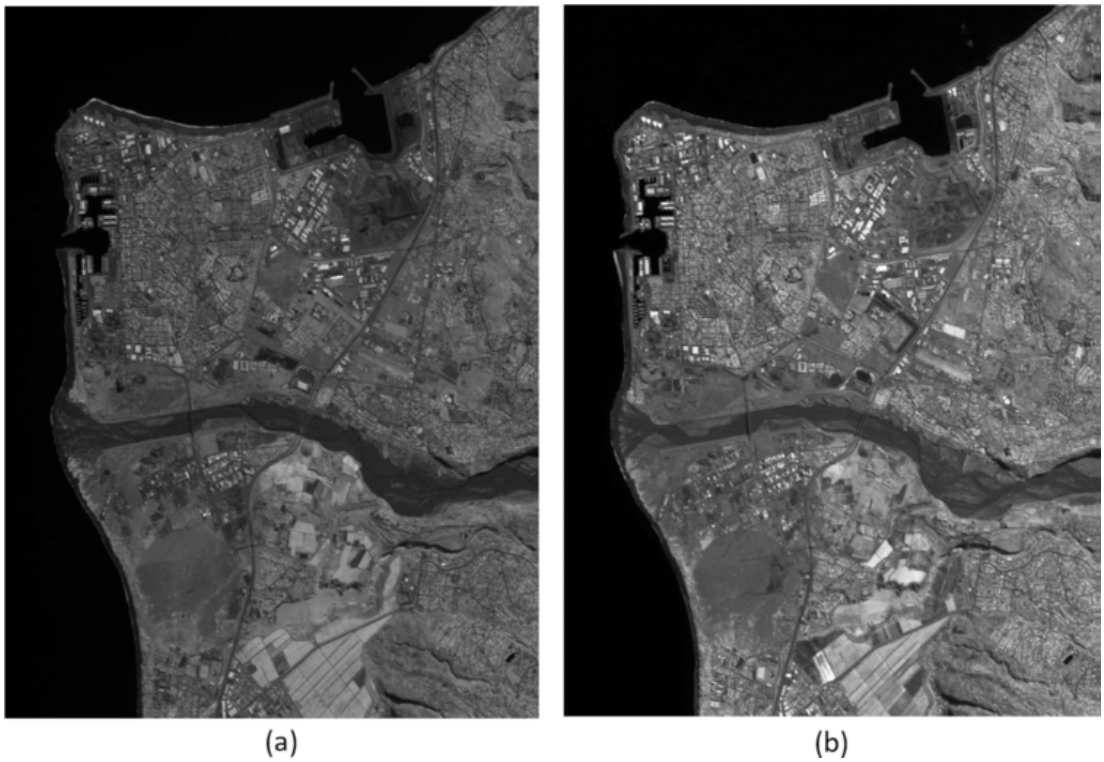


Figure 3.3 — Images satellites (a) et (b) acquises respectivement le 24 juillet 2008 et le 12 août 2014.

2. Application de l'approche proposée

Dans cette étude, nous nous intéressons à la prédiction de changements de la zone urbaine.

– **AS qualitative**

La méthode de criblage de Morris est appliquée pour identifier qualitativement la sensibilité des paramètres du modèle de prédiction de COS. Le nombre de répétitions r pour la méthode Morris est fixé à 40, de sorte que le modèle est exécuté 1120 fois ($rX[K+1]$ où K est le nombre de paramètres). Le résultat peut être représenté graphiquement par un tracé de dépistage où l'axe des abscisses représente les valeurs numériques des moyennes (μ^*) et l'axe des ordonnés représente les valeurs numériques des écarts-types (σ), comme indiqué dans la figure 3.4. L'influence de chaque paramètre est proportionnelle à la distance entre le point (σ, μ^*) et l'origine (0, 0).
A partir de la figure 3.4, nous constatons que les 9 paramètres (Hom, Cont, Dis, A, SI, MR, MVS, SDS et SDR) ont un effet négligeable et peuvent être considérés comme des paramètres fixes. Alors que les autres paramètres (NDVI, Ent, SD, R, Den, DR, MVMB, SDMB, Tem, Hum, Pré, Cor, LW, MVR, MVG, SDG, MVN, SDN) ont une influence importante sur les résultats de sortie. Seuls les paramètres ayant un effet important seront pris en compte pour les étapes suivantes. L'étape suivante consiste en une étude de la dépendance entre les paramètres d'entrée.

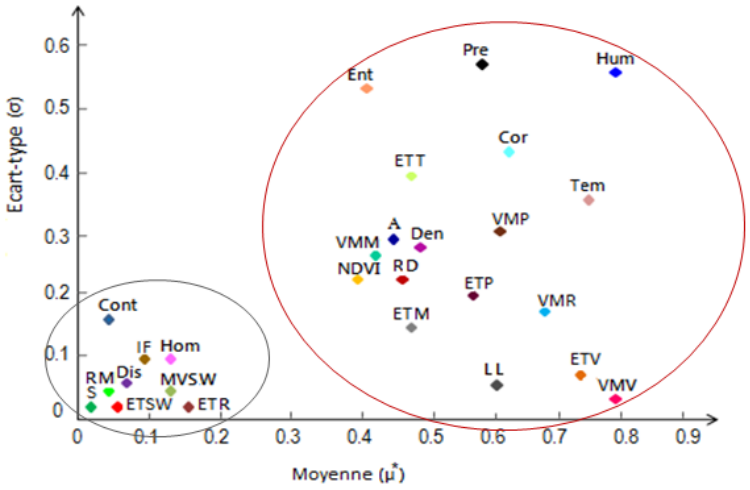


Figure 3.4 — Classification des paramètres en se basant sur la méthode Morris.

– **Mesure de corrélation**

Nous commençons par classer les paramètres sélectionnés dans la première étape en deux groupes :

Un groupe aléatoire contenant les paramètres de texture (Ent, ETT, Cor) et les paramètres spectraux (VMV, ETV, VMR, VMP, ETP, VMM, ETM).

Un groupe épistémique contenant les paramètres géométriques (LL, A, Den, RD), les paramètres climatiques (Tem, Hum, Pre) et le paramètre de végétation (NDVI).

Ensuite, nous appliquons l'algorithme de mesure de corrélation au vecteur contenant les 10 paramètres aléatoires et les 8 paramètres épistémiques.
La valeur absolue du coefficient de corrélation varie entre 0 et 1. Ici, nous choisissons un seuil égal à 0.5 pour évaluer la corrélation entre les paramètres.
Une paire de paramètres avec une valeur de coefficient de corrélation incluse dans l'intervalle suivant [0, ..., 0.5] sont considérés comme des paramètres corrélés. Tandis qu'une paire de paramètres avec un coefficient de corrélation compris dans l'intervalle]0.5, ..., 1] sont considérés comme des paramètres indépendants.
Pour les paramètres aléatoires, nous choisissons de représenter la sortie de l'algorithme en tant que diagramme de dispersion (Figure 3.5) et pour les paramètres épistémiques, le résultat est représenté comme une matrice de corrélation (Tableau 3.1).

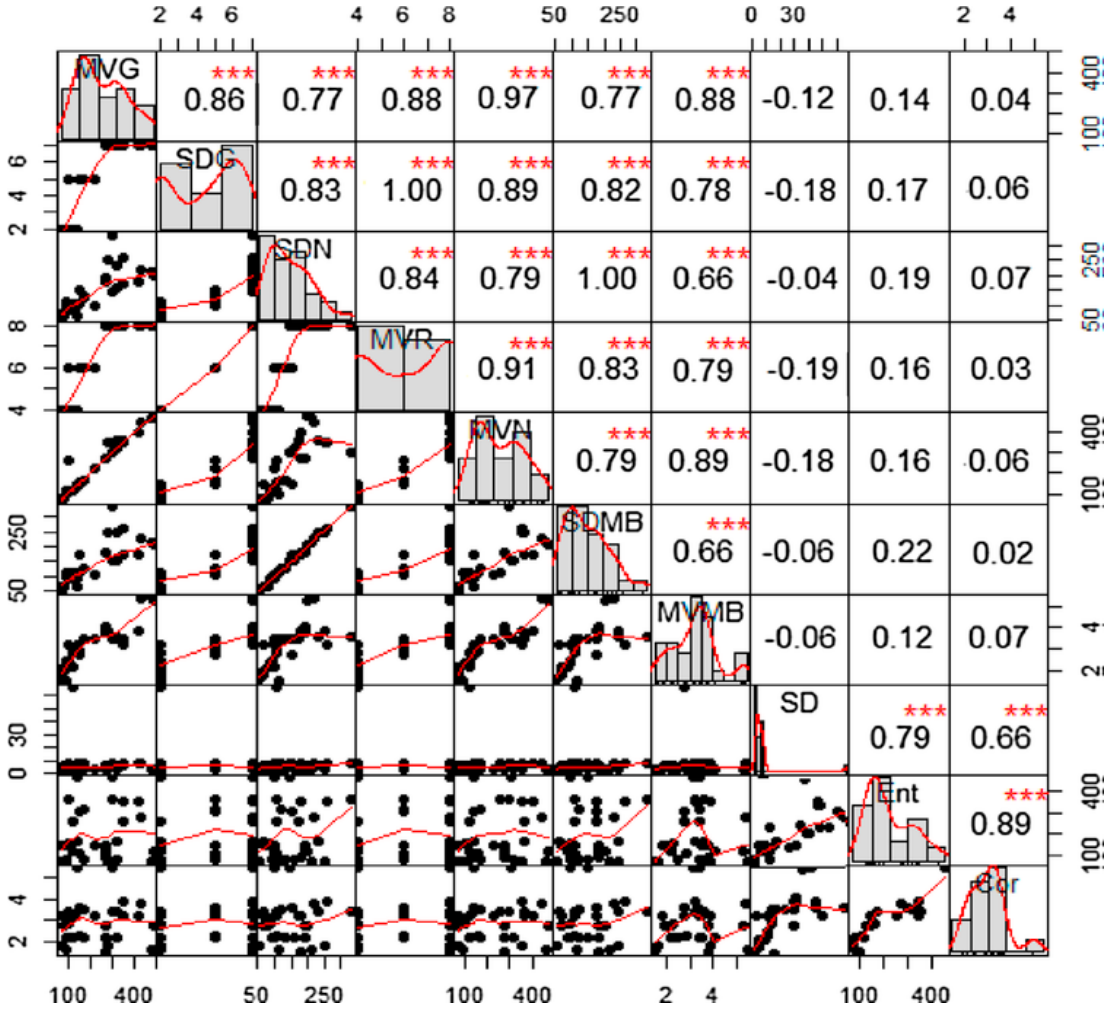


Figure 3.5 — Matrice de corrélation des dix paramètres aléatoires.

Dans la figure 3.5 :
- La diagonale représente les abréviations de chaque paramètre aléatoire et leurs distributions (histogrammes).
- Le panneau de gauche (en bas de la diagonale) montre le bivarié des diagrammes de dispersion entre chaque couple de paramètres. Ces graphiques permettent décrire la relation entre chaque paire de paramètres.
- Le panneau de droite (en haut de la diagonale) affiche les valeurs numériques du coefficient de corrélation et les astérisques représentant le degré de corrélation entre chaque paire de paramètres.
À partir de la figure 3.5, nous pouvons voir qu'il y a deux groupes de paramètres aléatoires, le premier (G1) inclut MVG, SDG, MVR, SDMB, SDN, MVMB et MVN tandis que le second (G2) inclut SD, Cor et Ent.
Le tableau 3.1 présente les coefficients de corrélation calculés pour les paramètres épistémiques. Nous notons que ces paramètres sont classés en trois groupes : le premier (G3) contient NDVI, Tem, Pre et Hum, le second (G4) contient LW, R et Den et le troisième (G5) contient un paramètre (DR).

	NDVI	Tem	Hum	Pre	LW	R	Den	DR
NDVI	1.00	0.90	0.83	0.78	0.18	0.16	0.03	0.47
Tem	0.90	1.00	0.79	0.89	0.13	0.16	0.07	0.31
Hum	0.83	0.79	1.00	0.65	0.05	0.22	0.06	0.54
Pre	0.78	0.89	0.65	1.00	0.06	0.12	0.08	0.16
LW	0.18	0.13	0.05	0.06	1.00	0.74	0.68	0.03
R	0.16	0.16	0.22	0.12	0.74	1.00	0.89	0.07
Den	0.03	0.07	0.06	0.08	0.68	0.89	1.00	0.13
DR	0.47	0.31	0.54	0.16	0.03	0.07	0.13	1.00

Tableau 3.1 — Matrice de corrélation des paramètres épistémiques.

– **AS quantitative**

Après propagation des incertitudes à travers le modèle de prédiction de COS, nous appliquons l'algorithme d'AS proposé pour calculer les indices de sensibilité totaux et individuels pour les 5 groupes de paramètres. La valeur des indices de sensibilité varie de 0 à 1. Plus sa valeur se rapproche de 0 plus le paramètre est dit sensible et inversement. Dans notre étude, nous choisissons un seuil égal à 0,5 pour la détermination des groupes sensibles et les paramètres les plus influents.
Les résultats obtenus de l'AS qualitative sont montrés sur la figure 3.6(a). Les résultats indiquent que G1, G2 et G5 sont les groupes des paramètres les plus sensibles, G3 et G4 sont les groupes les moins sensibles. Un groupe ayant un effet négligeable sur la sortie du modèle signifie que tous les paramètres de ce groupe ne sont pas influents sur la sortie du modèle de COS.

Pour les groupes qui ont une grande influence, nous devons connaître l'influence de chaque paramètre. Ainsi, nous calculons les indices de sensibilité individuels pour identifier les paramètres qui influencent le plus la sortie du modèle.
La figure 3.6(b), (c) et (d) montre que :
- Pour le groupe G1, (MVG, MVR et MVMB) sont les paramètres les plus influents,

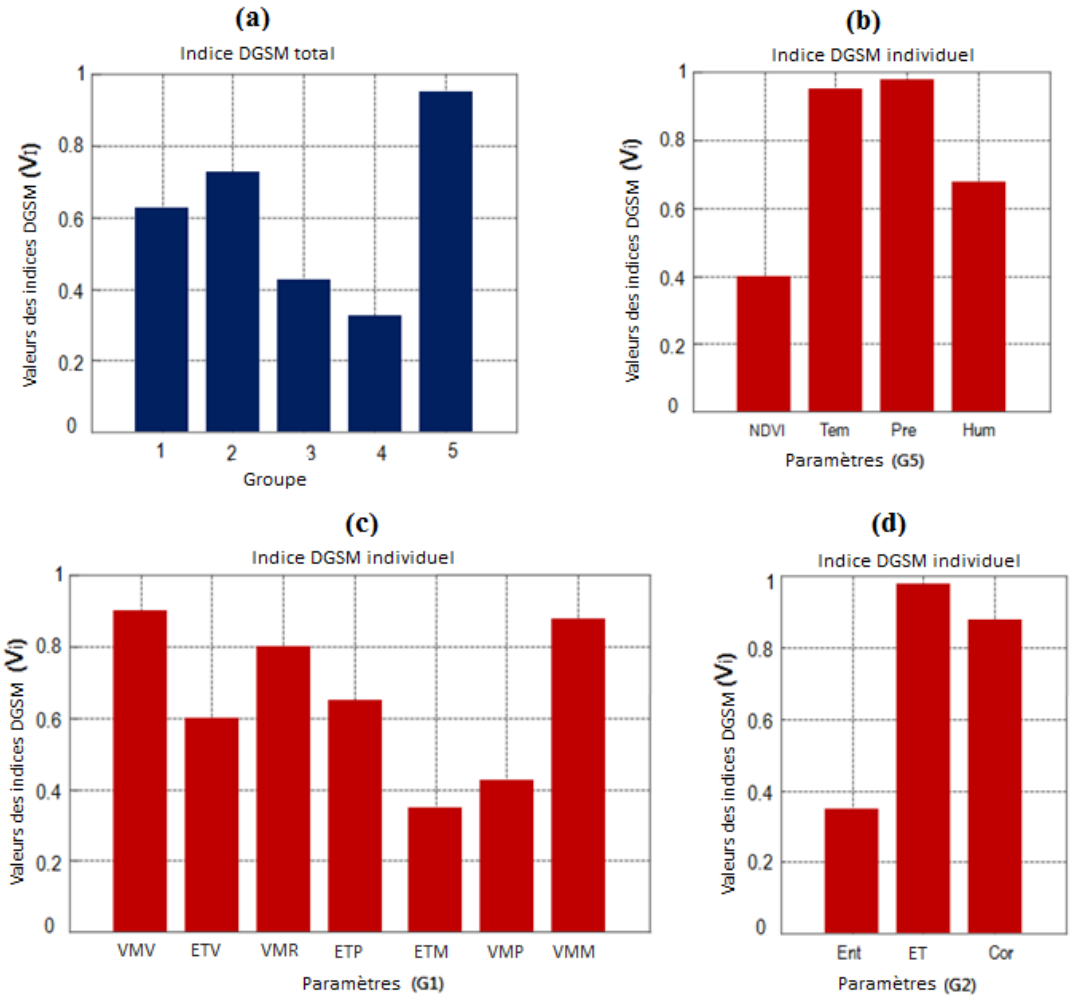


Figure 3.6 — Indices de sensibilité totaux et individuels de l'approche proposée.

tandis que les autres paramètres (MVN, SDG, SDN et SDMB) sont moins influents.
- Pour le groupe G2, (SD et Cor) ont une influence importante sur le modèle de prédiction de COS, tandis qu'Ent a une influence négligeable.
- Le G5 ne contient qu'un seul paramètre (DR) qui est un paramètre influent.

3. Evaluation de l'approche proposée

Le but de cette section est d'évaluer les performances de l'approche proposée.

– **Importance de la mesure de corrélation**

Pour évaluer le module de la mesure de corrélation, nous calculons les indices de sensibilité pour seulement 10 paramètres pour deux cas différents (avec et sans corrélation). Ensuite, nous comparons les résultats obtenus dans les deux cas.

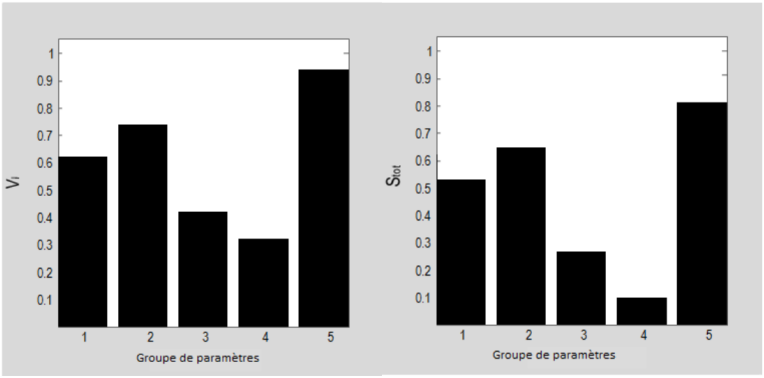
Paramètres	Tem	Pre	Hum	NDVI	Ent	DR	SDG	MVN	Cor	Den
SI (avec cor.)	0.97	0.99	0.7	0.4	0.37	0.3	0.6	0.42	0.9	0.25
SI (sans cor.)	0.77	0.85	0.68	0.34	0.36	0.18	0.55	0.19	0.76	0.12
Différence	0.2	0.14	0.02	0.06	0.01	0.12	0.05	0.23	0.14	0.13

Tableau 3.2 — Indices de sensibilité (SI) avec et sans corrélation.

Le tableau 3.2 illustre les indices de sensibilité de chacun des 10 paramètres entenant/sans-tenir compte de la dépendance entre les paramètres. La différence entre les indices de sensibilité dans les deux cas représente l'impact de la corrélation entre les paramètres sur le modèle de prédiction de COS.
Ceci prouve qu'ignorer la corrélation entre les paramètres entraîne des résultats qui sont parfois erronés. Ainsi, prendre en compte la corrélation entre les paramètres d'entrée du modèle de COS permettra une meilleure prise de décision sur la prédiction de COS.

– **Importance de l'AS quantitative**

Pour évaluer l'AS quantitative, nous comparons les résultats obtenus par notre approche avec ceux de la méthode de Sobol.
Les indices d'AS pour les cinq groupes basés sur les techniques de Sobol et de la DGSM sont présentés, respectivement, dans la figure 3.7 (A) et (B). Comme nous pouvons le constater, le classement des paramètres est toujours maintenu dans les deux méthodes, malgré qu'il y ait une petite différence en termes de valeurs numériques des indices de sensibilité calculés. Cependant, à partir de la figure 3.8, nous remarquons que le temps de calcul de DGSM est inférieur au temps de calcul de Sobol. Les résultats expérimentaux sont en faveur de l'utilisation du DGSM dans l'étude de l'AS.



(A) **(B)**

Figure 3.7 — Classement des paramètres de groupe en utilisant (A) DGSM et (B) Sobol.

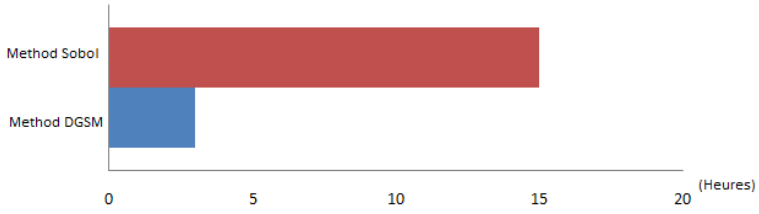


Figure 3.8 — Temps de calcul de la méthode DGSM et de la méthode Sobol.

– **Résultats de la prédiction de COS**

Nous comparons les performances de notre approche de prédiction de COS par rapport aux approches présentées dans les travaux relatifs à Boulila et collaborateurs et Ferchichi et collaborateurs [18] [47].
Le tableau 3.3 décrit les approches d'AS utilisées pour la comparaison.

Approche	Méthode d'AS	Dépendance	Type d'incertitude
Approche citée dans [13]	Sobol	-	Aléatoire
Approche citée dans [45]	Sobol	-	Aléatoire et épistémique

Tableau 3.3 — Approches d'AS utilisées pour la comparaison

Ces approches sont basées sur la méthode Sobol qui est considérée comme une méthode de référence pour l'AS dans la littérature.
Nous appliquons notre approche et les deux autres approches présentées dans le tableau 3.3 pour calculer le COS de la région du 'Le Port' en 2014.
La figure 3.9 présente les images de prédiction obtenues par les trois approches. Ces images sont comparées aux changements réels extraits de l'image de la vérité de terrain qui représente la même région et qui est acquise à la même date (Figure 3.9 (d)).

Le tableau 3.4 présente les erreurs pour la prédiction de LCC de la région 'Le Port' entre 2008 et 2014 pour l'approche proposée et les approches présentées dans [18] [47].

Approche	Erreur de prédiction de COS
Approche citée dans [13]	29.15%
Approche citée dans [45]	12.7%
Approche proposée	7.23%

Tableau 3.4 — Comparaison de l'erreur de prédiction de changements entre 2008 et 2014

A partir du tableau 3.4, nous notons que le taux d'erreur de l'approche proposée est inférieur au taux d'erreur de l'approche de Boulila et collaborateurs. [18] et l'approche de Ferchichi et collaborateurs [47] avec une différence de 21,92% et 5,47% respectivement. L'amélioration des résultats reflète l'importance de la considération des paramètres d'entrée corrélés du modèle de COS et aussi la modélisation des différents types d'incertitudes.
Notre approche produit de meilleurs résultats de prédiction de COS que les approches

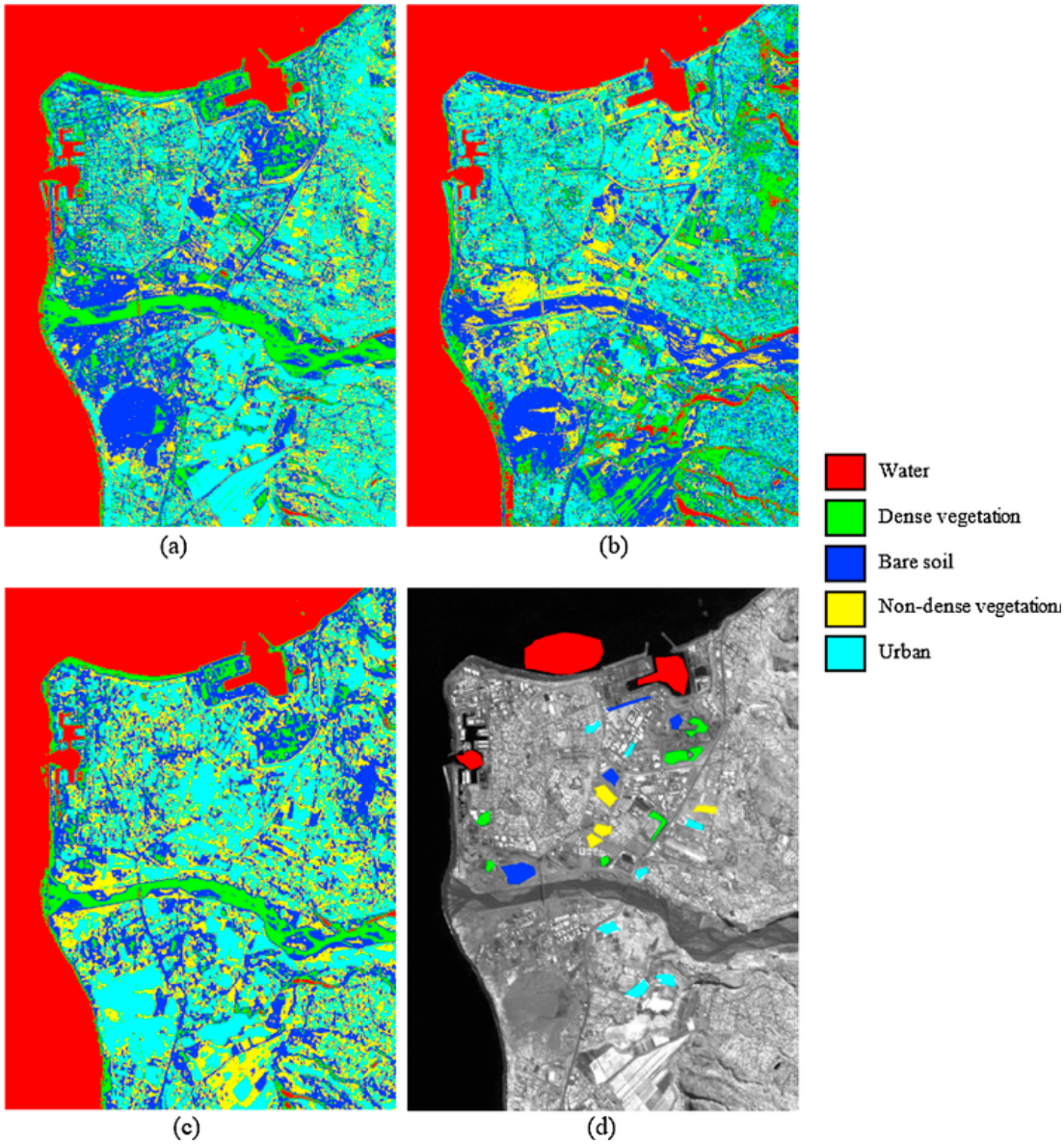


Figure 3.9 — Images de prédiction obtenues par : (a) l'approche proposée, (b) l'approche proposée dans Boulila et collaborateurs [18], (c) l'approche proposée dans Ferchichi et collaborateurs [47] et (d) l'image de la vérité du terrain.

déjà présentées. Ceci montre l'efficacité de notre approche pour réduire les incertitudes liées au processus de prédiction.
Aussi, nous proposons d'étudier les performances de notre approche en prédiction de COS. Nous avons comparé les résultats de la prédiction de changements de la région du Port entre 2008 et 2014 obtenus par notre approche et les deux méthodes déjà citée précédemment avec les changements réels des cinq types de couverture terrestre

(végétation dense, eau, sol nu, végétation non dense et urbaine).
Le tableau 3.5 illustre les pourcentages de changement obtenus pour chaque approche pour les cinq types d'occupation des sols.

Approche	Végétation dense	Eau	Sol nu	Végétation non dense	Urbain
Approche citée dans [13]	1.57	3.45	20.18	6.99	50.94
Approche citée dans [45]	2.33	2.69	40.97	6.92	38.31
Approche proposée	2.89	1.57	40.23	6.63	41.75
Changement réels	3.17	1.91	37.06	5.12	43.68

Tableau 3.5 — Pourcentages de changements prévus entre 2008 et 2014

Les résultats décrits dans le tableau 3.5 montrent que notre approche d'AS améliore la prédiction de COS par rapport aux approches existantes.
Nous notons que les résultats expérimentaux obtenus par l'approche proposée dans [47] qui traite à la fois les deux types d'incertitudes épistémique et aléatoire sont plus précis que les résultats obtenus par l'approche proposée dans [18] qui ne considère que l'incertitude aléatoire. Nous pouvons conclure qu'en tenant compte de la corrélation entre les paramètres et les deux types d'incertitudes (aléatoire et épistémique), nous pouvons obtenir des résultats plus efficaces et précis et donc améliorer les décisions du COS. La figure 3.10 représente les taux de convergence de l'erreur liée à la prédiction de COS des trois approches (approche proposée dans [18] et approche proposée dans [47]) selon la taille de l'échantillon. 50 000 échantillons de paramètres pour chaque approche sont générés. Nous notons que l'approche proposée basée sur une combinaison entre DGSM et Morris converge à 1000 simulations tandis que les deux autres approches basées sur la méthode de Sobol convergent à 20 000 simulations.
Cela prouve que la méthode DGSM nécessite moins de simulations et elle est plus rapide que la méthode de Sobol. Elle fournit un gain significatif en temps de calcul.

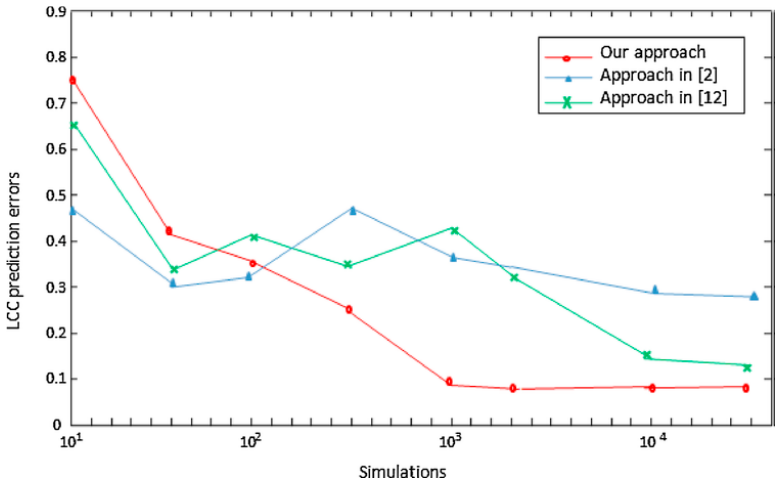


Figure 3.10 — Taux de convergence de l'erreur liée à la prédiction LCC pour notre approche, approche proposée dans [18] et approche proposée dans [47].

Afin de mieux évaluer les performances de l'approche proposée d'autres expériences sont effectuées. Nous avons estimé les coefficients de Kappa de l'approche proposée et de l'approche proposée dans [47] pour 14 périodes différentes avec un intervalle de six mois à partir de la date du 24 juillet 2008. Les vrais changements sont évalués à l'aide des images satellitaires pour chaque période de temps.
La figure 3.11 illustre la différence entre les coefficients Kappa (OK) obtenus par les deux méthodes pour chaque période de temps. Nous notons que l'approche proposée produit de meilleurs résultats de prédiction dans 11 cas par rapport à l'approche présentée par Ferchichi et collaborateurs [47].

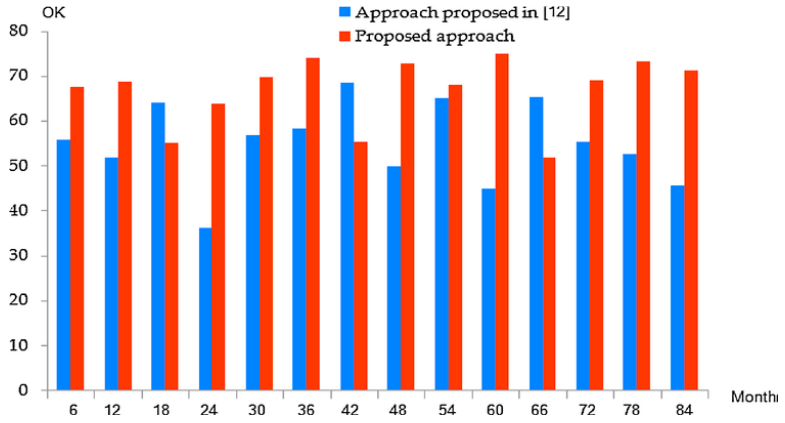


Figure 3.11 — Comparaison des coefficients kappa globaux pour l'approche proposée et l'approche proposée dans [47].

3.2.4.2 Analyse de sensibilité par la théorie des fonctions de croyance

L'expérimentation de notre approche d'analyse de sensibilité par la théorie des fonctions de croyance est divisée en deux parties : l'étude de la sensibilité pour déterminer les paramètres les plus influents des modèles de COS et l'estimation des valeurs optimales des paramètres influents.
Pour plus du détail sur l'expérimentation du processus d'analyse de sensibilité basée sur la théorie des fonctions de croyance, et la prédiction de COS pour la région du Caire et l'évaluation de l'approche proposée, le lecteur peut se référer à notre travail [48].

1. **Application de l'analyse de sensibilité basée sur la théorie des fonctions de croyance**

L'étude de l'analyse de sensibilité permet de déterminer les paramètres qui pourrons influencer la sortie des modèles de COS. Dans ce travail, l'analyse de sensibilité par la méthode de pincement est utilisée afin de mesurer l'influence des paramètres choisis sur les quatre modèles de COS : DINAMICA, SLEUTH, CA-MARKOV et LCM. Aussi,

la méthode de pincement servira pour tester l'effet de choix d'une structure donnée de modèle sur les resultats de COS finaux.

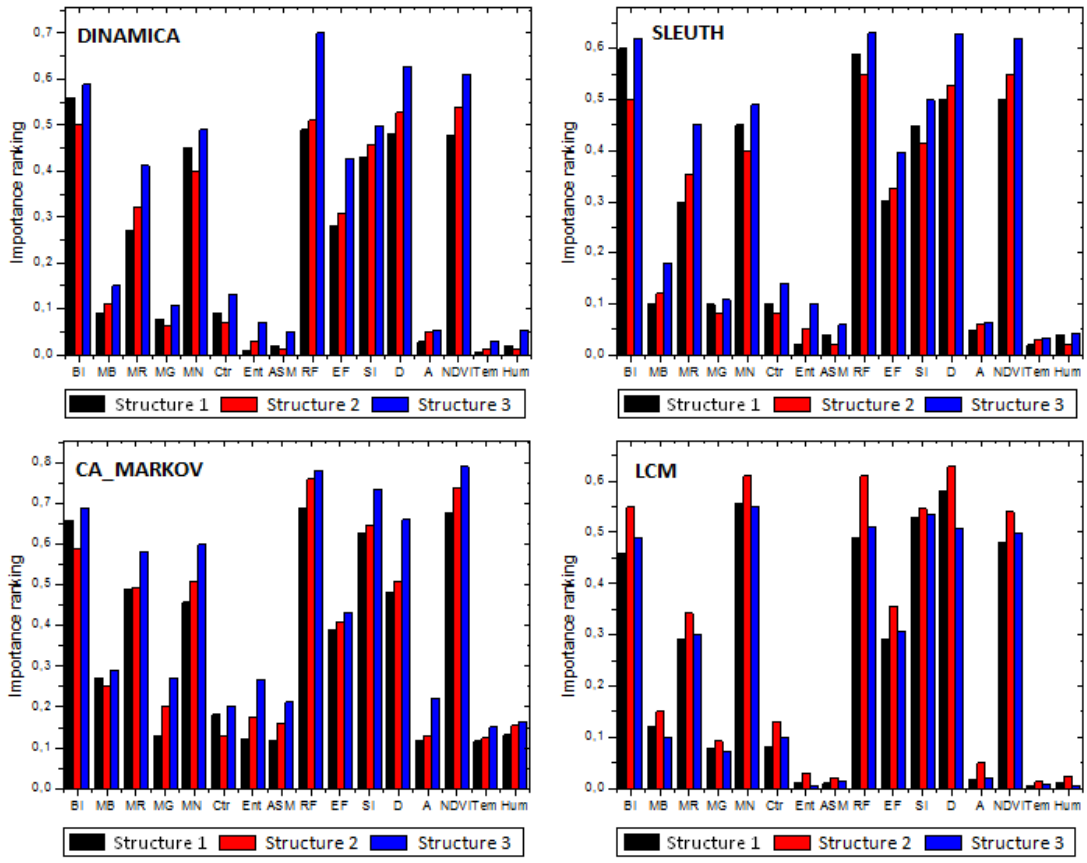


Figure 3.12 — Influence des paramètres d'entrée sur la sortie des modèles de COS.

La figure 3.12 présente l'influence des paramètres d'entrée sur la sortie des quatre modèles de COS :DINAMICA, SLEUTH, CA-MARKOV et LCM tout en considérant trois structures différentes pour chacun de ces modèles. Plus la barre est haute, plus l'influence de paramètre qui repésente cette barre sur la sortie de modèle est grande. Dans ce cas, nous pourrons dire que le modèle de COS dépend fortement de ce paramètre. Ainsi, l'étude des incertitudes liées à ce paramètre est cruciale pour garantir une meilleure décision.
La figure 3.12 permet aussi de constater que le changement de la structure des modèles de COS a une influence sur le choix des paramètres les plus influents sur ces modèles. Par exemple, la paramètre D est le plus influent sur le modèle LCM pour la structure 2, alors que MN sera considéré comme le paramètre le plus influent sur le même modèle si nous prenons le cas de la structure 3. Ces résultats confirment l'importance du choix de la structure la plus appropriée de modèle de COS afin de garantir une meilleure décision sur les COS.
D'autre part, nous remarquons que la corrélation entre les paramètres a une influence

importante sur la tendance de variation de ces paramètres. Par exemple, les paramètres RF, EF, SI et D ont des tendances de variation similaires et ceci quelque soit la structure de modèle de COS.

Le tableau 3.6 présente les huits paramètres les plus influents sur les modèles de COS : DINAMICA, SLEUTH, CA-MARKOV et LCM. Nous constatons qu'il n'y a pas un consensus sur les paramètres les plus influents entre les quatre modèles de COS. Egalement, les résultats obtenus montrent que l'analyse de sensibilité basée sur la méthode du picement permet déterminer les paramètres les plus influents pour les modèles de COS.

Rank	DINAMICA	SLEUTH	CA-MARKOV	LCM
1	BI	BI	RF	D
2	RF	RF	NDVI	MN
3	D	D	BI	SI
4	NDVI	NDVI	SI	RF
5	MN	MN	MR	NDVI
6	SI	SI	D	BI
7	EF	EF	MN	EF
8	MR	MR	EF	MR

Tableau 3.6 — Paramètres les plus influents pour les modèles de COS : DINAMICA, SLEUTH, CA-MARKOV et LCM.

2. Estimation des valeurs des paramètres les plus influents

Nous proposons d'utiliser les limites de confiance de Kolmogorov-Smirnov pour estimer les valeurs optimales des paramètres les plus influents.

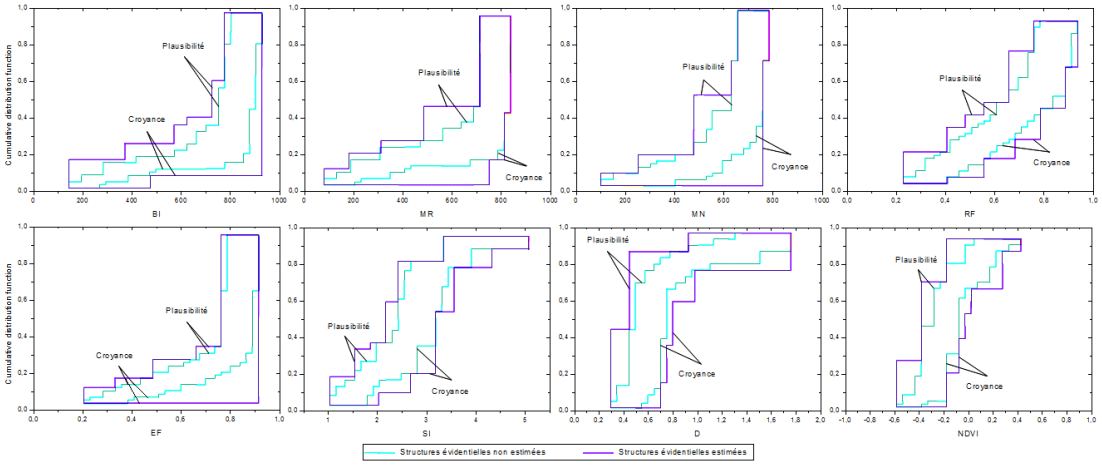


Figure 3.13 — Valeurs des paramètres les plus influents.

La figure 3.13 illustre les valeurs des paramètres les plus influents. La même méthode est utilisée pour estimer les structures les plus fidèles des modèles de COS.
Une fois les valeurs des paramètres influents sont estimées, nous appliquons la théorie des fonctions de croyance pour propager de nouveau les incertitudes à travers les modèles de COS et obtenir les nouvelles plages d'incertitude de prédiction.
La figure 3.13 présente une comparison des plages d'incertitude de sortie pour les 4 modèles de COS avant et après l'application de notre approche. Comme nous le remarquons, l'incertitude a un effet important sur les résultats de sortie des quatre modèles de COS. Ainsi, il est essentiel de modéliser les incertitudes qui accompagnent le processus de prédiction.

3.3 Approche proposée pour la réduction des incertitudes dans le domaine de la bio-informatique

Dans cette partie, nous nous intéressons à la réduction des incertitudes dans le domaine de la bio-informatique et plus particulièrement au cas du cancer du poumon. Il s'agit d'analyser les biomarqueurs biologiques potentiels par rapport aux différentes solutions de traitement. Il est à noter qu'une meilleure compréhension du cancer nécessite un examen approfondi des régulateurs. Les gènes jouent un rôle éminent dans la synthèse des protéines qui régulent la croissance des cellules tumorigènes. L'activité de certains gènes notamment la traduction ou la suppression, peut altérer et perturber l'âge de la cellule ou ses fonctions dont les cellules normales peuvent subir un processus de tumorigenèse.
Le contexte de ce travail se base sur le regroupement des données d'expression génétique qui aide à élucider les fonctions des gènes et à révéler la typologie des tumeurs [114]. En effet, les techniques de regroupement des données d'expression génétique ont pour objectif la projection des gènes individuels de grande dimension vers une dimension réduite optimale de groupes. Ceci est dans le but de déterminer les niveaux distincts d'expressions géniques, ce qui facilite la compréhension des fonctions des gènes.
Nous illustrons dans la partie suivante un état de l'art des travaux liés au groupement dans le contexte de l'expression génétique et les problèmes à résoudre, les données d'étude, l'approche proposée et finalement une expérimentation de cette approche.

3.3.1 Etat de l'art et problématiques étudiées

Dans la littérature, plusieurs travaux tentent d'appliquer des techniques de regroupement dans le contexte de l'expression génétique. Richards et collaborateurs [101], ont comparé les résultats de quatre algorithmes de regroupement, à savoir k-means, Chinese Restaurant Clustering (CRC), the Iterative Signature Algorithm (ISA) et une nouvelle variante d'ISA (memISA) sur des données d'expression du cerveau par puce à ADN. La comparaison repose sur trois mesures de performance, à noter la vitesse, la couverture des gènes et les GO-enrichissements. Bien que les enrichissements ISA et memISA

dépassent légèrement en GO-enrichissements k-means, cela se fait au détriment de la couverture et de la rapidité des gènes. Les auteurs rapportent que k-means dépasse les trois autres algorithmes avec une couverture génétique de 100%. Cependant, combiner k-means et ISA ou memISA améliorerait les performances du regroupement.
L'étude réalisée par Chen et collaborateurs présente une méthode de regroupement de données d'expression génétique basée sur un sous-espace régularisé par un graphe [28]. Le but de cette approche est de combiner à la fois les graphes de régularisation et le regroupement en sous-espaces pour modéliser la structure géométrique intrinsèque de l'espace de données. Afin de trouver une solution globale optimale pour le regroupement du sous-espace régularisé, les auteurs ont utilisé l'équation de Sylvester. L'approche est expérimentée moyennant huit ensembles de données d'expression génétique et comparée aux méthodes de regroupement de sous-espace, aux méthodes traditionnelles de regroupement et aux méthodes de regroupement basées sur la factorisation matricielle non négative.
Le travail effectué par Dutta et Collaborateurs explore l'utilisation d'une optimisation multi-objectifs basée sur des techniques de regroupement génétique [40]. L'objectif était de classer les gènes dans des groupes en fonction de leurs similitudes fonctionnelles et de leur pertinence biologique. Dans cette étude, les auteurs ont mis au point une mesure de la qualité permettant de calculer la qualité des groupes de gènes, qui est le score de confiance des interactions protéine-protéine. Dans cette approche, des expériences sont effectuées sur trois ensembles de données d'expression génétique dans la vie réelle et les résultats sont comparés aux techniques existantes.
Paul et Collaborateurs [95] ont proposé des annotations pour l'ontologie de gènes basées sur un algorithme de classification semi-supervisé. L'algorithme développé s'appelle le groupement relationnel flou GO. Dans cet algorithme, un gène peut être attribué à plusieurs groupes. L'algorithme utilise la connaissance biologique disponible sous la forme d'une ontologie de gènes, comme connaissance préalable avec les données d'expression génétique. Les connaissances antérieures contribuent à améliorer la cohérence des groupes concernant le domaine de connaissances. Les auteurs ont testé l'algorithme développé en se basant sur deux ensembles de données relatives à la levure (Saccharomyces cerevisiae). Les résultats obtenus ont été comparés par rapport à d'autres algorithmes de regroupement de pointe.
Comme il a été mentionné précédemment, le but de ce travail était d'analyser les biomarqueurs biologiques potentiels du cancer du poumon. Plusieurs états de traitement sont présents et nous disposons de trois échantillons pour chacun. Il est important à noter que suite à des incertitudes (par exemple des erreurs de mesures), les valeurs de mesure des biomarqueurs biologiques changent d'un échantillon à un autre. Il en découle des problèmes notamment celui relatif aux informations issues des différents échantillons permettant d'améliorer notre décision sur les biomarqueurs biologiques potentiels du cancer du poumon. Un autre problème à résoudre dans ce travail correspond au suivi du changement d'un état à un autre sur les gènes (en d'autres termes, le problème est de découvrir les gènes qui sont influencés par le traitement). Un traitement par plasma non-thermique est effectué sur les gènes ; les données utilisées dans ce travail sont décrites dans le paragraphe qui suit.

3.3.2 Présentation des données d'expression génétique

Le plasma est l'un des états fondamentaux de la matière, il a une propriété gazeuse (c'est-à-dire qu'il n'a pas de forme ni de volume définis) et, contrairement aux solides et aux liquides, il n'est pas dense [35]. Le plasma est créé par un processus appelé ionisation au cours du quel les atomes ou les molécules d'un gaz acquièrent une charge négative ou positive. Ceci se fait soit par chauffage ou lorsqu'ils sont soumis à un champ électromagnétique puissant à une température relativement très élevée. Ce processus provoque un gain ou une perte d'électrons, conduisant à la formation de particules chargées positivement ou négativement, appelées Ions. Le plasma peut être thermique ou non-thermique. Le plasma thermique a la même température pour les électrons, les ions et les neutres alors que pour le plasma non thermique, la température des électrons est supérieure à celle des ions et des neutres.
Les dernières avancées technologiques ont permis de concrétiser l'utilisation du plasma non-thermique dans le domaine médical. Le cancer apparaît comme un trouble de la fonction cellulaire de l'un des organes. Cela provoque une croissance anormale des cellules et pourrait même se propager d'un organe à un autre par le biais d'un processus appelé métastase. Hou et Collaborateurs [64] ont fourni le profil d'expression du gène cellulaire de la tumeur de l'adénocarcinome du poumon lors d'un traitement avec du plasma atmosphérique non-thermique. Les données consistent en des échantillons décrivant le transcriptome de la cellule tumorale dans les conditions suivantes :

- Trois échantillons dont les mesures sont prises avant de commencer le traitement.
- Trois échantillons dont les mesures sont prises après un traitement par du plasma non-thermique à courte exposition (la mesure est faite après 4 heures).
- Trois échantillons dont les mesures sont prises après un traitement par du plasma non-thermique à longue exposition (la mesure est faite après 1, 2 et 4 heures).

Ces données sont accessibles dans la banque de données GEO pour "*Gene Expression Omnibus*" du NCBI via le site (`https://www.ncbi.nlm.nih.gov/geo/`) et sous le numéro d'accès GEO GSE59997.
Le tableau 1.1 décrit les notations du jeu de données considéré dans notre étude.

NT	Groupe de échantillons mesuré avant de commencer le traitement
SE	Traitement par plasma non-thermique à courte exposition
LE après 1hr	Traitement par plasma non-thermique à longue exposition, mesuré après 1 heure du traitement
LE après 2hr	Traitement par plasma non-thermique à longue exposition, mesuré après 2 heures du traitement
LE après 4hr	Traitement par plasma non-thermique à longue exposition, mesuré après 4 heures du traitement

Tableau 3.7 — Notations de jeu de données.

3.3.3 Approche proposée

Dans cette section, nous détaillons notre approche de regroupement des données d'expression génétique. Le processus proposé est illustré dans la figure 3.14.

La première étape de ce processus est de vérifier si les données considérées présentent

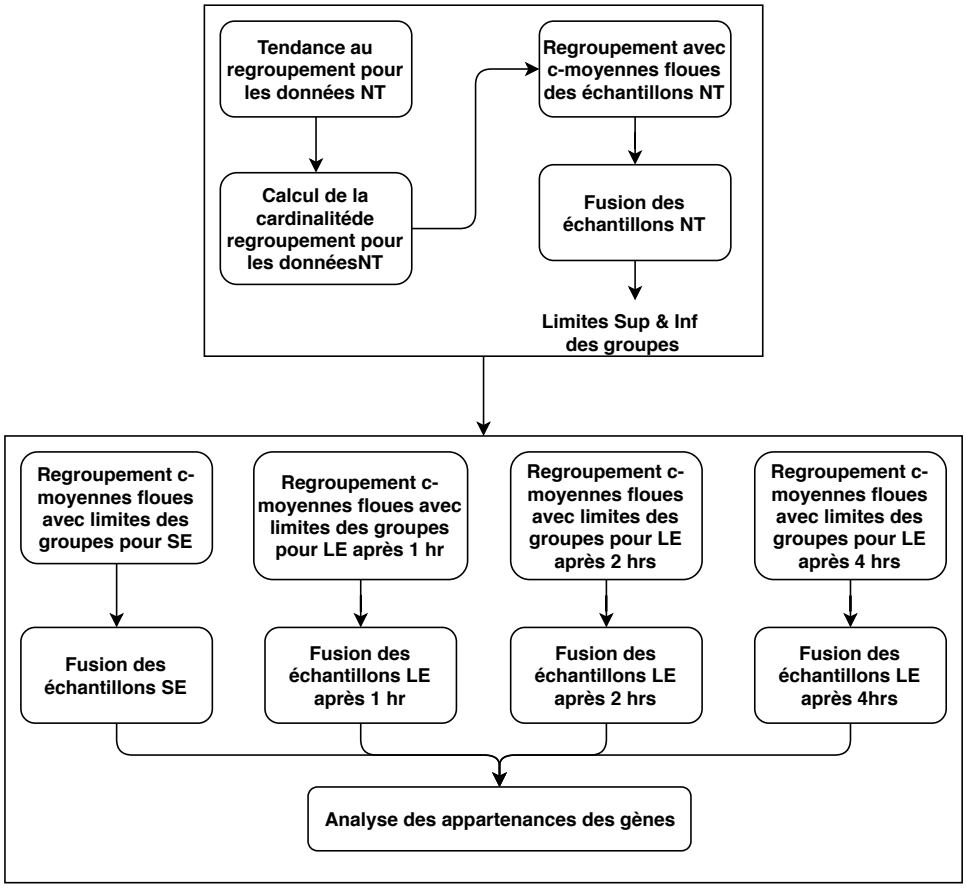


Figure 3.14 — Approche proposée.

une prédisposition intrinsèque à se regrouper en groupes distincts ou non. Cette étape est appelée tendance au regroupement (cluster tendency). Il s'agit d'utiliser trois techniques pour mesurer la tendance au regroupement dont deux statistiques nommées Hopkins et Cox-Lewis [8] [93] et une visuelle appelée l'évaluation visuelle de la tendance des groupes [9].
Il est à noter, que dans le cas de présence d'une structure de regroupement dans notre jeu de données, l'étape suivante consiste à déterminer la cardinalité de regroupement.
La troisième étape de l'approche proposée consiste à effectuer un regroupement avec c-moyennes floues pour les trois échantillons NT [118].
Le but de cette étape est d'estimer les masses nécessaires lors de la prochaine étape de fusion qui est basée sur la théorie des croyances. La fusion permet de combiner les trois échantillons NT en un seul résultat de regroupement contenant une meilleure description du jeu de données NT.
La dernière étape effectuée pour les échantillons NT se focalise sur la détermination des limites supérieure et inférieure pour chaque groupe obtenu après le processus de fusion. Ces limites seront utilisées pour le regroupement des échantillons SE et LE. En effet, le

traitement SE a pour objectif de déterminer les gènes qui ont changés d'appartenance entre les échantillons NT et SE. Par conséquent, les limites supérieure et inférieure pour chaque groupe d'échantillons SE sont prises identiques à celles des échantillons NT. Ensuite, nous effectuons un regroupement de c-moyennes floues tout en tenant compte de cette contrainte. La fusion par la théorie des croyances est appliquée au regroupement résultant des trois échantillons de SE. La dernière étape du traitement SE consiste à déterminer les gènes qui ont préservés leurs groupes durant les traitements NT et SE. Le même processus effectué sur le jeu de données SE est réalisé pour les jeux de données LE après 1h, après 2h et après 4h. L'idée de l'approche proposée est de combiner les informations provenant des trois échantillons pour chaque état (NT, SE, LE 1hr, LE 2hrs et LE 4hrs) en une seule décision pour chaque état. Cette décision permettra une meilleure description des gènes liés au cancer du poumon. Le processus de fusion des données assurera l'intégration des données provenant des trois cas afin de produire des informations plus cohérentes et précises sur les gènes liés au cancer du poumon.
La section suivante décrit les principales étapes du processus de fusion. La fusion d'informations peut être définie comme une combinaison des informations issues de plusieurs sources afin d'améliorer la prise de décision. Etant donné l sources $S_1, S_2, ..., S_l$, il s'agit de prendre une décision parmi n décisions possibles $d_1, d_2, ..., d_n$ [11].

3.3.3.1 Théorie des fonctions de croyance

Cette théorie permet la représentation de l'incertitude de type aléatoire et épistémique en utilisant des fonctions de masses (m), de croyances (Bel) et de plausibilités (Pl) [41] [42] [75].

– **Modélisation**

Les fonctions de masse sont définies sur tous les sous-ensembles de l'espace de discernement D (par exemple, $D = E_1, E_2, ..., E_n$) [11] [75].
La fonction de masse est définie telle que $m : 2^D \rightarrow [0, 1], m(\phi) = 0$ (en monde fermé) et $\sum_{A \subset D} m(A) = 1$
Les éléments focaux sont définis tels que $A \subset D$ tel que $m(A) > 0$.
La fonction de croyance Bel est une fonction totalement croissante de 2^D dans [0,1], telle que $Bel(\phi) = 0$, $Bel(D) = 1$ et

$$\forall A \in 2^D, Bel(A) = \sum_{B \subseteq A, B \neq \phi} \tag{3.13}$$

– **Combinaison**

Soit $m_j (j = 1...l)$ la fonction de masse définie pour la source j. La combinaison conjonctive des fonctions de masse est calculée en utilisant la règle orthogonale de Dempster

définie $\forall A \subseteq D$ par :

$$m(A) = (m_1 \oplus m_2 \oplus ... \oplus m_l)(A) = \frac{\sum_{B_1 \cap ... \cap B_l} m_1(B_1)m_2(B_2)...m_l(B_l)}{1 - K} \tag{3.14}$$

où

$$K = \sum_{B_1 \cap ... \cap B_l = \phi} m_1(B_1)m_2(B_2)...m_l(B_l) \tag{3.15}$$

K désigne le degré de conflit entre les différentes sources.

– **Décision**

La règle la plus utilisée pour la décision dans la théorie de l'évidence est le maximum de croyance :

$$x \in C_i \quad Si \quad Bel(C_i)(x) = max\{Bel(C_k)(x), 1 \leq k \leq n\} \tag{3.16}$$

Après avoir fusionné les résultats de groupement de données des 3 échantillons NT, l'idée est de voir quels sont les gènes qui ont gardé leurs groupements par rapport aux autres gènes qui ont changé de groupes. Il faut savoir que les gènes qui ont changé de groupes d'un état à un autre (NT, SE, LE 1hr, LE 2hrs et LE 4hrs) ont réagi positivement au traitement qu'ils ont reçu.
Ainsi, il est important de préserver les limites des groupes obtenus à l'état NT et de faire un regroupement en tenant compte de ces limites pour les autres états.

3.3.3.2 Regroupement avec c-moyennes floues en tenant compte des limites des groupes

L'objectif principal de l'algorithme de c-moyennes floues en tenant compte des limites des groupes est d'effectuer un regroupement pour les données SE et LE tout en considérant la plage de chaque groupe obtenu lors de l'étape de fusion des données NT.
Nous commençons par calculer les limites inférieures et supérieures des gènes appartenant aux données NT et faisant partie du même groupe. Ceci constitue les plages de chaque groupe des données NT. L'étape suivante consiste à appliquer l'algorithme c-moyennes floues tout en tenant compte des plages de chaque groupe pour les données SE et LE. Ensuite, nous appliquons la fusion basée sur la théorie des croyances pour obtenir un résultat final combinant les trois échantillons de chaque état.
L'étape finale de notre processus consiste à étudier la présence de changements des gènes entre les états NT, SE et LE. L'analyse de ce changement permet de révéler l'effet du traitement de plasma non-thermique sur le transcriptome de la tumeur.
Le tableau 3.8 et la figure 3.15 présentent la structure des groupes après la fusion pour chaque état et la migration des gènes entre les groupes pour les états NT, SE et LE après le traitement par plasma non-thermique.

Groupes	Fusion NT	Fusion SE	Fusion LE
1	$a_1 - b_1$	$a_1 - b_1$	$a_1 - b_1$
2	$a_2 - b_2$	$a_2 - b_2$	$a_2 - b_2$
.	.	.	.
.	.	.	.
n-1	$a_{n-1} - b_{n-1}$	$a_{n-1} - b_{n-1}$	$a_{n-1} - b_{n-1}$
n	$a_n - b_n$	$a_n - b_n$	$a_n - b_n$

Tableau 3.8 — Limites de regroupement après l'application de la fusion.

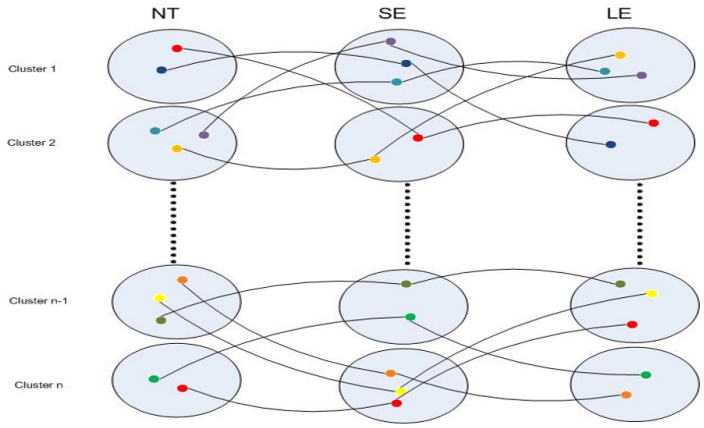


Figure 3.15 — Migration des gènes entre les groupes pour les états NT, SE et LE.

3.3.4 Expérimentation de l'approche proposée

Dans cette section, nous présentons l'expérimentation de notre approche de réduction des incertitudes appliquée dans le domaine de la bio-informatique. L'expérimentation est divisée en quatre parties :

1. tendance au regroupement et calcul de la cardinalité
2. fusion des échantillons NT
3. c-moyennes floues avec limites des groupes
4. fusion des données SE et LE

3.3.4.1 Tendance au regroupement et calcul de la cardinalité

Nous commençons par examiner la tendance au regroupement pour les données de l'état NT. Nous avons utilisé les deux indices statistiques (Hopkins et Cox Lewis) pour les trois échantillons NT. Le tableau 3.9 présente les résultats obtenus qui affirment la présence d'une structure de regroupement dans les données NT.

L'étape suivante consiste à déterminer le nombre de groupes présent dans les données NT. Pour ce faire, nous avons utilisé les indices de validation internes pour

	NT		
	Echantillons 1	Echantillons 2	Echantillons 3
Hopkins	0.91	0.98	0.62
Cox Lewis	1.34	1.48	1.2

Tableau 3.9 — Valeurs des indices de tendance au regroupement pour les données NT.

déterminer la cardinalité des groupes. Nous avons utilisé trois indices à savoir Calinski, Davies Bouldin et Silhouette pour rechercher le nombre de groupes [43]. Le résultat de calcul du nombre optimal est illustré au niveau de la figure 3.16. Le nombre optimal de groupes pour Calinski est de 10 groupes avec une valeur de Calinski de 3,75. Davis Bouldin atteint une valeur minimale légèrement supérieure à 0,51 pour 8 groupes. Le nombre optimal de groupes selon la silhouette égale 2 groupes.
Ensuite, nous combinons les valeurs des trois indices normalisés selon l'équation suivante :

$$normalizedIndex = \frac{index - \min(index)}{\max(index) - \min(index)} \tag{3.17}$$

La figure 3.16d présente le nombre optimal de groupes, soit 4, correspondant à la valeur d'indice la plus élevée légèrement inférieure à 0,6.

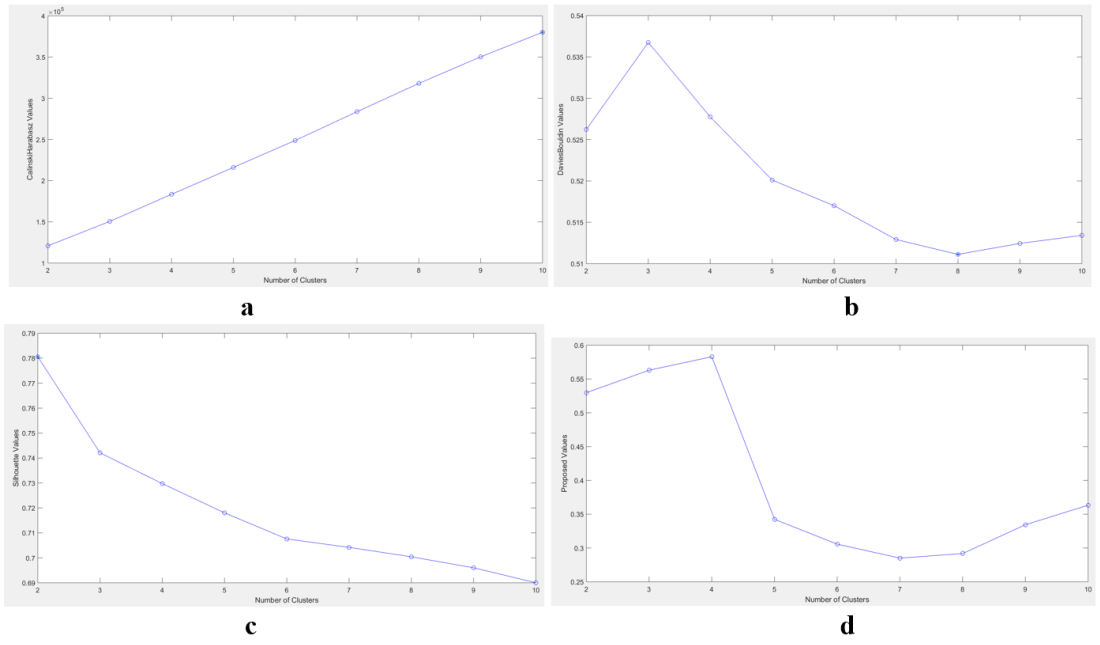


Figure 3.16 — Nombre optimal de groupes obtenu par la méthode de (a) Calinski, (b) Davies Bouldin, (c) Silhouette et (d) combinaison de ces trois méthodes.

Le même processus est répété pour les échantillons NT 2 et NT 3. Le nombre optimal de groupes est de 4.

3.3.4.2 Fusion des échantillons NT

La première étape consiste à effectuer un regroupement par c-moyennes floues pour les trois échantillons NT et ensuite appliquer la théorie de fusion. Le tableau 3.10 décrit les plages des 4 groupes et le nombre de gènes appartenant à chacun des groupes.

	Échantillon 1		**Échantillon 2**		**Échantillon 3**		**Fusion NT**	
	Inf/sup	**Nbr gènes**	**Inf/sup**	**Nbr gènes**	**Inf/sup**	**Nbr gènes**	**Inf/sup**	**Nbr gènes**
C1	2.7392 5.0842	19011	2.7392 5.1032	15888	2.7984 5.0892	19080	2.7635 5.0882	17000
C2	5.0846 7.2665	13999	5.1033 7.2575	14138	5.0896 7.2636	13821	5.0904 7.2614	14039
C3	7.2670 9.8090	11076	7.2576 9.7851	11189	7.2639 9.7898	11129	7.2618 9.9480	13867
C4	9.8102 15.1854	5309	9.7859 15.2003	5180	9.7901 15.1081	5365	9.9490 15.0773	4489

Tableau 3.10 — Limites inferieures et supérieures, et nombre de gènes pour le groupe de données NT.

3.3.4.3 C-moyennes floues avec limites des groupes et fusion des données SE et LE

Les limites des groupes après la fusion des échantillons NT seront prises comme contrainte pour l'application de l'algorithme c-moyennes floues. Cet algorithme est appliqué pour les trois échantillons des données SE et LE (1 hr, 2 hr et 4 hr). Ensuite, la fusion par la théorie des croyances est appliquée à ces données pour obtenir un regroupement final qui représente les données SE et LE (1 hr, 2 hr et 4 hr). Pour plus du détail sur la fusion de données et le nombre de gènes dans chaque groupe après cette étape, le lecteur peut se référer à notre travail [43].

	NT - SE 1hr	**NT - LE 1hr**	**LE 1hr - LE 2hr**	**LE 2hr - LE 4hr**
Gènes non stables	37162	36999	4039	4938

Tableau 3.11 — Nombre de gènes changeant de groupes.

Enfin, le nombre de gènes qui ont migré d'un groupe à un autre est présenté dans le tableau 3.11. Il est essentiel de préciser que cette comparaison de l'appartenance des gènes est basée sur les résultats obtenus après la fusion pour tous les états considérés (NT, SE et LE). Cette comparaison révèle des résultats intéressants à savoir 36999 gènes ont modifié leur appartenance de groupe de l'état NT à l'état LE après 1 h. Il y a un déclin éminent dans le nombre de gènes (4039 gènes) changeant d'appartenance au groupe de l'état LE après 1hr à l'état LE après 2hr. Par ailleurs, 899 autres gènes rejoignent les gènes 4039 et changent leurs appartenances de groupe de l'état LE après 2 heures à l'état LE après 4 heures.
La comparaison de notre travail avec les travails ultérieurs est présentée par Farouq et

collaborateurs [43].

3.3.4.4 Interprétation des résultats

Le cancer du poumon est défini au niveau moléculaire par des mutations et des altérations d'oncogènes, notamment AKT1, ALK, BRAF, EGFR, HER2, KRAS, MEK1, MET, NRAS, PIK3CA, RET et ROS1. Dans le tableau 3.12, un profil d'expression moléculaire des biomarqueurs biologiques du cancer du poumon après la fusion NT, LE après 1h, LE après 2h et LE après 4h est présenté.
Les gènes BRAF, RET et ROS1 montrent une transition de surexpression en passant de l'état NT à LE, puis une stabilisation dans les 3 états de LE. BRAF passe de C1 à C2 alors que RET et ROS1 passent de C2 à C3.
Les gènes MET, ALK et PIK3CA montrent une transition de surexpression en passant des états NT à LE, puis une stabilisation dans les 3 états du même groupe. Le MET passe de C1 à C3 et C4, ALK passe de C2 à C4 tandis que PIK3CA passe de C1 à C3. SME1 montre une transition de sous-expression en passant d'états de NT à LE, puis une stabilisation dans les 3 états. SME1 passe de C4 à C2.

Dans notre travail, nous avons pu améliorer le processus d'identification des biomarqueurs biologiques potentiels du cancer du poumon à travers la fusion de données (théorie de croyance) issues de plusieurs échantillons. Pour le suivi de changement des gènes d'un état à un autre, nous avons proposé un algorithme de regroupement avec c-moyennes floues en tenant compte des limites des groupes de l'état initial NT.

3.4 Conclusions et perspectives de recherche

Dans ce chapitre, nous avons proposé trois approches de réduction des incertitudes. La première approche proposée utilise la méthode de sensibilité globale des dérivées (DGSM) pour réduire les incertitudes liées aux modèles de COS. Cette approche est divisée en quatre étapes : 1) analyse de sensibilité qualitative, 2) mesure de corrélation, 3) propagation d'incertitudes et 4) analyse de sensibilité qualitative. Le processus commence par identifier les paramètres incertains. Puis il détermine les corrélations entre les paramètres incertains identifiés et regroupe ces paramètres en groupes. Ensuite, il fait la propagation d'incertitudes à travers le modèle de COS pour pouvoir finalement identifier les paramètres qui ont influence importante sur la sortie du modèle.
La deuxème approche de réduction des incertitudes est basée sur la théorie des fonctions de croyance. L'approche proposée est scindée en deux étapes : l'étude de la sensibilité pour identifier les paramètres les plus influents des modèles de COS et l'estimation des valeurs optimales des paramètres influents. La stratégie du pincement est utilisée pour générer une analyse de sensibilité et les limites de confiance du Kolmogorov-Smirnov sont utilisées pour déterminer les paramètres les plus influents.
La troisième approche est basée sur la fusion des données et a pour but d'analyser les biomarqueurs biologiques potentiels du cancer du poumon sur différentes solutions de

	Fusion NT	Fusion LE après 1h	Fusion LE après 2h	Fusion LE après 4h
AKT1	4	1	1	1
AKT1	1	1	1	1
AKT1	2	3	3	3
MET	1	3	3	3
MET	1	4	4	4
MET	4	4	4	4
MET	1	3	4	4
KRAS	1	3	3	3
KRAS	4	1	1	2
KRAS	1	2	2	2
KRAS	4	4	4	4
NRAS	4	1	1	1
NRAS	1	3	3	3
EGFR	1	3	3	3
EGFR	4	4	4	4
EGFR	4	1	1	1
EGFR	1	2	2	2
EGFR	2	3	3	3
EGFR	3	2	2	2
EGFR	1	1	1	1
EGFR	2	1	1	1
BRAF	1	2	2	2
ALK	2	4	4	4
ALK	2	4	4	2
PIK3CA	1	3	3	3
PIK3CA	1	1	1	1
SMEK1	4	2	2	2
RET	2	2	2	2
RET	2	3	3	3
RET	2	3	3	3
ROS1	2	3	3	1

Tableau 3.12 — Profil moléculaire des oncogènes NSLC.

traitement. L'idée de cette approche est de combiner des informations provenant de trois échantillons pour les états (NT, SE, LE 1hr, LE 2hrs et LE 4hrs) en une seule décision pour chaque état. Cette décision permettra une meilleure description des gènes liés au cancer du poumon.
Les trois méthodes proposées sont validées à travers plusieurs jeux réels de données. De plus, l'évaluation de ces méthodes par rapport aux méthodes existantes dans la littérature montre les bonnes performances des méthodes proposées.
Plusieurs perspectives peuvent être envisagées pour notre travail. Elles sont d'ordre méthodologique et expérimental :

– Sur le plan méthodologique : nous comptons d'élargir nos travaux dans le domaine de corrélation entre les paramètres. Aussi, nous proposons d'automatiser le processus d'identification de type d'incertitudes liées aux paramètres (aléatoire ou épistémique) et de prendre en compte la corrélation entre les paramètres mixtes

(corrélation entre paramètre épistémique et paramètre aléatoire). Par ailleurs, nous envisagerons d'appliquer d'autres méthodes de propagation d'incertitudes (telle que la théorie de possibilités) et d'analyse de sensibilité. En effet, comparer leurs résultats aidera à comprendre la méthode la plus appropriée pour un domaine donné. Enfin, nous proposons de mettre en ligne un outil de réduction d'incertitude. Cet outil permettra aux utilisateurs d'analyser les incertitudes liées à leurs modèles de simulation.

– Sur le plan expérimental, nous proposons appliquer et tester notre approche sur d'autres zones d'étude et dans d'autres domaines d'application. Le présent travail a été consacré à l'application de notre approche sur la prédiction de COS, que nous envisagerons d'étendre à d'autres applications telles que la classification et la segmentation, et à d'autres domaines comme la médecine et la biologie.

Partie 2 : Analyse et interprétation des données satellitaires

Chapitre 4

Analyse et interprétation des données satellitaires non massives

Sommaire

4.1 Introduction

Les données satellitaires sont utilisées dans plusieurs domaines tels que : la cartographie, la surveillance, l'aménagement du territoire, l'archéologie, les études environnementales, la gestion des ressources, etc. Cependant, la quantité de données satellitaires a considérablement augmenté en raison de l'évolution des capteurs satellitaires. Ainsi, il est nécessaire de disposer d'outils automatisés pour l'interprétation et l'analyse des données satellitaires.
Dans ce chapitre, nous proposons une approche d'analyse et d'interprétation des images satellitaires. L'approche proposée a pour but de fournir une aide à la décision dans plusieurs domaines de la télédétection en offrant des informations descriptives et prédictives. Le processus d'analyse et d'interprétation des données satellitaires est divisé en quatre étapes : la segmentation sémantique des images satellitaires, l'intégration des données, la modélisation des données et l'interprétation des données.
Dans ce chapitre, nous commençons par présenter un état de l'art sur les travaux se rapportant sur l'analyse et l'interprétation des images satellitaires et les problèmes à résoudre. L'approche proposée est détaillée dans la deuxième section. L'expérimentation et l'évaluation de notre approche sur un ensemble réel de données sont présentées dans la section 4. La dernière partie de ce chapitre présentera les conclusions et les perspectives de recherche.

4.2 Etat de l'art et problématiques étudiées

L'analyse et l'interprétation des données satellitaires aident les utilisateurs à comprendre de nombreux phénomènes liés à la terre et à fournir des décisions permettant de réduire les risques environnementaux afin d'améliorer la prise de décision concernant ces phénomènes. Les informations extraites des données satellitaires peuvent être utilisées dans de nombreux domaines telles que la prévision météorologique, la gestion des ressources, la planification régionale, la surveillance du trafic et l'évaluation des risques environnementaux.
De nombreuses tâches d'analyse et d'interprétation reposent sur la compréhension du contenu d'une image ou d'une scène. Dans la littérature, plusieurs techniques ont été proposées pour aider à comprendre le contenu des images satellitaires. Parmi ces techniques, nous énumérons la détection d'objets, la reconnaissance d'objets, la segmentation des images et la segmentation sémantique des mages. Souvent, ces techniques peuvent entraîner une certaine confusion. La détection d'objets dans les images satellitaires vise à déterminer si une image contient un ou plusieurs objet(s) appartenant à la classe d'intérêt et à localiser leurs positions [29]. La reconnaissance d'objets vise à détecter tous les objets dans les images satellitaires et à localiser leurs positions [36]. Pour la segmentation, l'image satellitaire est divisée en régions ; cependant, ces régions ne seront pas étiquetées [121]. La segmentation sémantique d'images a pour but d'étiqueter chaque pixel de l'image satellitaire en fonction d'une classe d'objets tels que les zones urbaines, les forêts, l'eau, etc. [124].

Toute tentative d'analyse et d'interprétation des données satellitaires fait face souvent à plusieurs défis :

- **La complexité des données satellitaires** : Les données satellitaires se distinguent par plusieurs caractéristiques par rapport aux autres types de données. Ces caractéristiques rendent les données satellitaires plus complexes à manipuler que plusieurs autres types de données. Parmi ces caractéristiques, nous pouvons citer la haute dimensionnalité des pixels [84]. En fait, chaque pixel de l'image peut être projeté dans un espace dans lequel plusieurs bandes existent et, par conséquent, il peut y avoir plusieurs interprétations (par exemple il peut appartenir à plusieurs types d'occupation des sols). De plus, le concept de voisinage dans les données satellitaires affecte le calcul de chaque pixel. Dans [86], les auteurs confirment que, dans un contexte de données satellitaires, les dépendances entre les pixels influencent le traitement et l'interprétation des images.
- **La modélisation des données satellitaires** : La tâche de modélisation des données satellitaires détermine la manière dont ces données seront représentées et interprétées. Les données satellitaires se présentent souvent sous différents formats : image, alpha-numérique, grille, carte, etc., ce qui complique la tâche de traitement de ces données. De plus, les outils d'interprétation des données satellitaires souffrent du problème de l'écart sémantique (semantic gap en anglais). L'écart sémantique est défini comme le problème de la description correcte des images lors du passage du bas niveau (pixel) au haut niveau (sémantique) [123]. Cet écart sémantique devient d'autant plus critique lorsque la quantité et la diversité des données satellitaires augmentent.
- **L'analyse et l'interprétation des données satellitaires** : L'analyse et l'interprétation des données satellitaires est une tâche très difficile. Il existe deux techniques principales couramment utilisées : l'interprétation manuelle et l'interprétation automatique. Le processus d'interprétation dépend de plusieurs opérations préliminaires telles que la segmentation, l'extraction des objets, le choix des attributs qui caractérisent les objets, etc. En raison du gros volume des données à manipuler, de la complexité de ces données, ainsi que de la nécessité de prendre des décisions en temps réel ou proche du réel, l'interprétation automatique des données satellitaires est sollicitée [77]. Cependant, le processus d'interprétation automatique est généralement une tâche difficile qui est soumise à de nombreux problèmes tels que la segmentation d'images, l'extraction des objets, le choix des attributs qui caractérisent les objets et l'extraction d'informations spatiales et thématiques utiles des objets d'intérêt [90].

Dans la littérature, plusieurs travaux ont été proposés pour analyser, interpréter et fournir une aide à la décision pour les utilisateurs dans le domaine de l'imagerie satellitaire.
Dempere-Marco et al. [34] ont proposé d'utiliser les automates cellulaires pour découvrir automatiquement les règles de transition dynamiques. L'objectif du modèle proposé est de récupérer les règles d'évolution de la dynamique urbaine dans le temps. Le système proposé tire profit des avantages d'un modèle d'automates cellulaires auto-adaptatif

intégré à un système immunitaire artificiel pour découvrir des règles de transition dynamiques. L'approche proposée par Dempere-Marco et al. est appliquée pour simuler la conversion des zones urbaines de la ville de Guangzhou, située au cœur du Delta de la rivière des Perles en Chine. Dans [69], Hwangbo et al. ont suggéré un système d'aide à la décision pour faciliter le choix de la méthode ou du schéma de classification optimale. Le système proposé est basé sur un raisonnement à base des cas pour assister les utilisateurs dans la tâche de classification de l'occupation des sols. Quatre caractéristiques sont déterminées pour assurer la récupération des cas : l'ensemble des données, l'emplacement, le climat et la classe. Fegraus et al. [44] ont présenté un tableau de bord environnemental basé sur des données de subsistance géoréférencées provenant d'enquêtes auprès des ménages et des données biophysiques recueillies à partir d'images satellitaires. Le système proposé a pour objectif de calculer diverses mesures du stress des écosystèmes. L'application de ce système consiste à proposer un système de gestion de la sécurité pour surveiller l'agriculture et les services écosystémiques en Tanzanie. Ai et al. [2] ont proposé un système dynamique d'aide à la décision. Le système proposé associe le SIG et les réseaux sociaux pour permettre la prise de décision en cas d'atténuation des risques liés aux tsunamis à Padang, en Indonésie. De nombreux acteurs peuvent utiliser le travail proposé par Ai et al. tels que les décideurs gouvernementaux, les responsables des politiques, les exécutants politiques et les habitants touchés par les catastrophes. L'objectif principal est de concevoir des cartes de risque de tsunami et des itinéraires d'évacuation rapides dans les régions exposées au risque de tsunami.
D'autres part, plusieurs travaux récents se sont orientés vers la segmentation d'images et plus précisément vers la segmentation sémantique d'images satellitaires. Parmi ces travaux, nous pouvons citer le travail du Zhang et al. [123] qui ont proposé de combler le fossé sémantique entre les caractéristiques visuelles de bas niveau et la sémantique de haut niveau des images. Dans cette étude, les auteurs ont développé une méthode de représentation de niveau intermédiaire basée sur les objets pour la classification sémantique. L'algorithme proposé est basé sur le sac de mots-visuels qui génère des fonctionnalités de niveau intermédiaire pour relier les deux niveaux. Dans [5], les auteurs ont présenté une approche basée sur une ontologie pour classifier les images satellitaires. Andrés et al. ont développé des règles spectrales pour une classification pixelaire des images Landsat. Le prototype proposé est couplé à un logiciel de traitement d'images open source lors de l'étape de pré-traitement et utilise un algorithme de raisonnement pour effectuer la classification des images. La principale limitation du travail présenté par Andrés et al. est liée au temps de traitement. Marmanis et al. [89] ont présenté un réseau neuronal convolutionnel profond pour la segmentation sémantique avec détection de contour. Les auteurs ont proposé de combiner la segmentation sémantique avec la détection sémantique informée de contour en ajoutant la détection de frontière à l'architecture codeur-décodeur.
Dans ce qui suit, nous allons présenter notre approche d'analyse et interprétation des données satellitaires non massives.

4.3 Approche proposée pour l'analyse et l'interprétation des données satellitaires non massives

L'approche proposée constitue un tableau de bord environnemental qui aide à la prise de décision. L'outil proposé peut avoir diverses applications dans différents domaines tels que la cartographie, la gestion des ressources et la planification régionale. Il offre une aide pour les utilisateurs afin de créer des analyses descriptives, prédictives et prescriptives.
Le processus proposé est divisé en quatre étapes principales, comme le montre la figure 4.1 :

1. Étape 1 : segmentation des images satellitaires
2. Étape 2 : intégration des données satellitaires
3. Étape 3 : modélisation des données satellitaires
4. Étape 4 : analyse et interprétation

L'objectif principal de la première étape est de faire une segmentation sémantique afin d'analyser et de comprendre le contenu des images manipulées. La deuxième étape a pour but de charger les données dans l'entrepôt de données. A ce niveau, de nombreuses opérations, généralement complexes, sont réalisées pour préparer les données pour être chargées dans l'entrepôt de données. Le défi consiste à intégrer et à consolider un volume important de données satellitaires dans un entrepôt de données unifié. Le schéma de l'entrepôt de données est choisi lors de l'étape de modélisation en fonction de l'exigence des utilisateurs du domaine de la télédétection. L'étape de modélisation nécessite l'identification de toutes les caractéristiques principales pour répondre aux exigences des différents utilisateurs du domaine de l'imagerie satellitaire. Une bonne modélisation permettra d'obtenir des décisions précises et pertinentes lors de l'étape d'analyse et d'interprétation.

Dans notre travail, nous proposons un système d'aide à la décision (SAD) qui peut être utilisé dans plusieurs domaines tels que : la classification de la couverture terrestre, la prédiction du changement de l'occupation des sols, la prévention des catastrophes, la planification régionale et la gestion des ressources. La fréquence du changement de données dans le contexte actuel est importante. Par conséquent, nous proposons un processus itératif et incrémentiel comme le montre la figure 2.

4.3.1 Segmentation sémantique des images satellitaires

La segmentation sémantique des images satellitaires est au cœur des préoccupations de recherche ces dernières années. Les méthodes de segmentation sémantique d'images soient agissent au niveau des pixels en classifiant chaque pixel indépendamment, soient tentent de regrouper les pixels en groupes et d'attribuer une étiquette à ces groupes.
La figure 4.3 décrit l'approche proposée. Le processus de la segmentation sémantique proposée est divisé en deux niveaux : 1) haut niveau et 2) bas niveau. Le premier niveau vise à assurer la construction du réseau de neurones à couches multiples (MLFFNN).

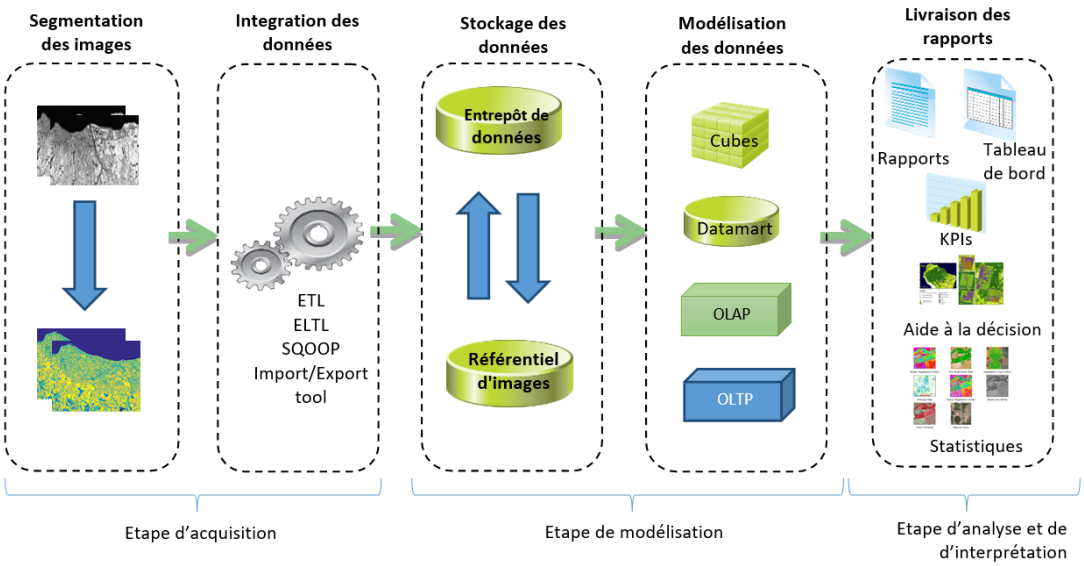


Figure 4.1 — Approche proposée pour l'analyse et l'interprétation des données satellitaires non massives.

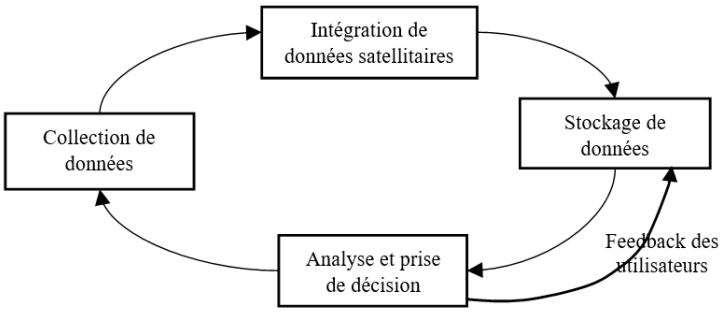


Figure 4.2 — Processus proposé de SAD.

Nous commençons par calculer les caractéristiques des objets extraits des images satellitaires. Ces caractéristiques constituent l'entrée du module MLFFNN pour générer une structure permettant de classifier les objets issus des images satellitaires. Dans la deuxième étape, la structure générée est utilisée pour effectuer une segmentation sémantique au niveau des pixels. Pour une image satellitaire d'entrée, une matrice centrée dans chaque pixel est prise en compte lors du calcul des caractéristiques associées à ce pixel. Les mêmes caractéristiques calculées au niveau de l'objet sont calculées au niveau pixel (nous calculons ces caractéristiques à partir de la fenêtre 3x3 entourant le pixel). Les valeurs calculées seront entrées dans le module MLFFNN pour déterminer la classe la plus similaire pour chaque pixel.

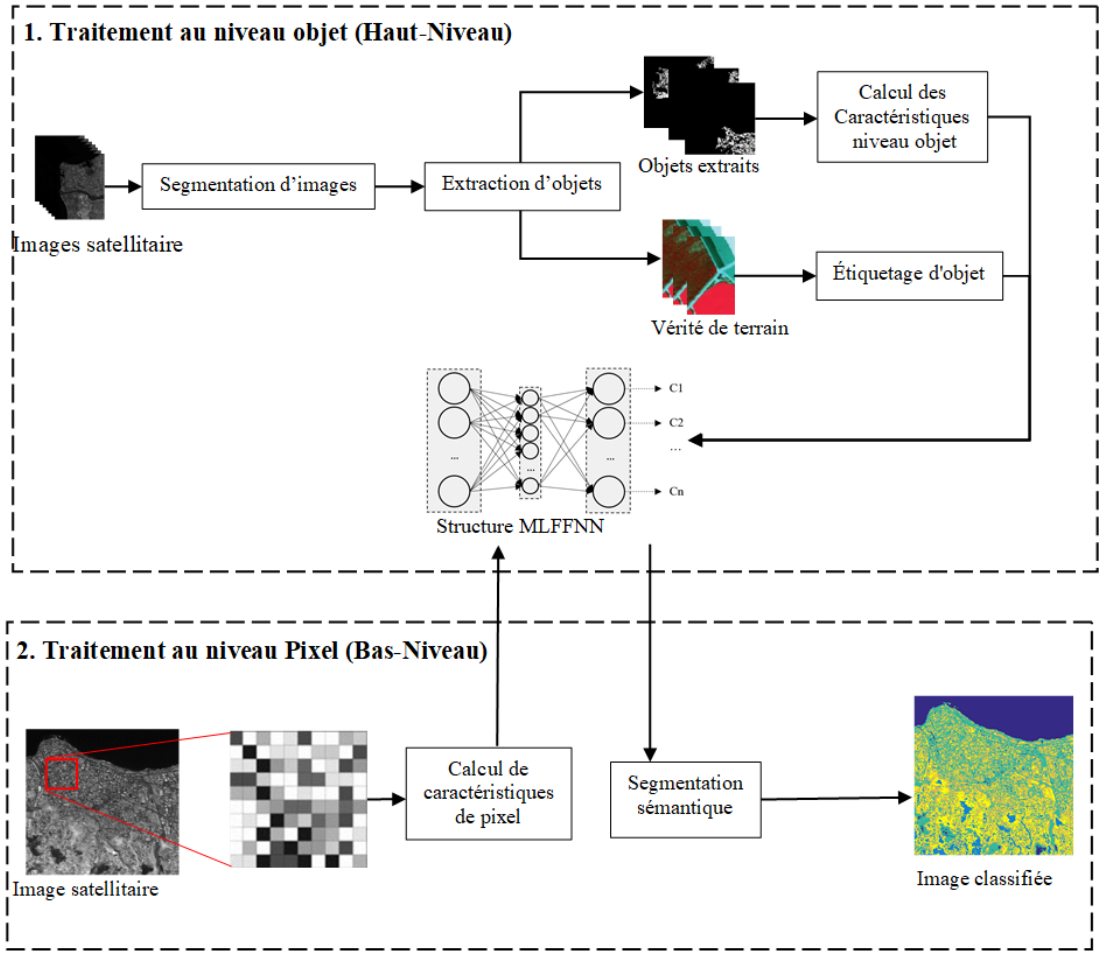


Figure 4.3 — Approche proposée pour la segmentation sémantique.

4.3.1.1 Traitement au niveau objet : haut niveau

Ce niveau vise à générer une structure neuronale qui sera utilisée pour la segmentation sémantique. Le processus de traitement au niveau objet est divisé en cinq étapes : a) segmentation de l'image, b) extraction des objets, c) calcul des caractéristiques des objets, d) étiquetage des objets, et e) génération de la structure du réseau de neurones.

1. **Segmentation des images satellitaires**

Le succès de l'interprétation des images est étroitement lié à la fiabilité de l'étape de la segmentation. Aujourd'hui, le problème de segmentation des images satellitaires est un sujet au cœur des recherches dans le domaine de l'image. De nombreux travaux ont été réalisés sur la segmentation d'images satellitaires. Parmi ces travaux, nous pouvons citer [19] [29]. Dans notre travail, la méthode k-means est utilisée pour seg-

menter les images [87]. Cette méthode peut être remplacée par toute autre méthode de segmentation.

2. **Extraction des objets**

Après la segmentation de l'image, nous obtenons un ensemble d'objets qui recouvrent l'ensemble de l'image manipulée. Le but l'étape de l'extraction des objets est de déterminer les objets significatifs contenus dans les images segmentées. Pour ce faire, deux paramètres sont choisis à savoir la connectivité et le nombre minimal de pixels (respectivement connectivité et minNumberPixels dans l'algorithme 1). Tous les objets connectés dans l'image segmentée qui ont un nombre de pixels inférieur à la valeur minNumberPixels ne sont pas pris en compte. Cette opération est appelée couverture de régions. Pour la connectivité, nous considérons le contexte de 8 pixels connectés (fenêtre 3x3 contenant le pixel et les voisins qui lui sont connectés horizontalement, verticalement et en diagonale). Une autre opération effectuée à cette étape consiste à supprimer les pixels isolés de l'image segmentée.

3. **Calcul des caractéristiques des objets**

Considérons un objet extrait d'une image satellitaire *img*. Les caractéristiques utilisées dans cette étude sont :

- La radiométrie du centre de l'objet

$$f_1 = img(centroide(obj)) \tag{4.1}$$

  Où *centroide* est la fonction qui calcule le centre d'un objet *obj*.
- Les cinq caractéristiques issues de la matrice de la co-occurrence des niveaux de gris GLCM d'un objet. Ces caractéristiques sont le contraste, la corrélation, l'énergie, l'homogénéité et l'entropie [59]. La GLCM calcule le nombre des différentes combinaisons de niveau de gris se produisant pour l'objet *obj*. Les caractéristiques issues de la GLCM donnent une mesure de la variation d'intensité au pixel d'intérêt.
  Considérons $p(i, j)$ l'élément qui a les coordonnées (i, j) dans la matrice GLCM.
  - Contraste : calcule l'intensité du contraste entre un pixel et son voisin. La formule du contraste est la suivante :

$$f_2 = \sum_{i,j} (i-j)^2 p(i,j) \tag{4.2}$$

  - Corrélation : mesure la corrélation entre un pixel et son voisin. La formule de la corrélation est la suivante :

$$f_3 = \sum_{i,j} p(i-j) \frac{(i-\mu)(j-\mu)}{\sigma^2} \tag{4.3}$$

Où μ est la moyenne de la GLCM, calculée comme suit $\mu = \sum_{i,j} p(i-j)i$, et $\sigma^2 = \sum_{i,j} p(i-j)(i-\mu)^2$

- Energie (connue aussi comme uniformité) : calcule la somme des éléments carrés dans le moment. La formule de l'énergie est la suivante :

$$f_4 = \sum_{i,j} (p(i,j))^2 \tag{4.4}$$

- Homogénéité : mesure la fréquence à laquelle la distribution des éléments de la matrice GLCM est proche de la diagonale. La formule de l'homogénéité est la suivante :

$$f_5 = \sum_{i,j} \frac{p(i,j)}{1+(i-j)^2} \tag{4.5}$$

- Entropie : quantifie le caractère aléatoire de la distribution d'intensité des niveaux de gris. La formule de l'entropie est la suivante :

$$f_6 = \sum_{i,j} -p(i,j)\ln(p(i,j)) \tag{4.6}$$

4. **Génération de la structure neuronale**

L'objectif de la structure neuronale est de créer un processus capable de déterminer la classe d'un objet extrait d'une l'image satellitaire en fonction de ses caractéristiques. Dans cette étude, nous avons choisi de travailler avec un réseau de neurones à couches multiples (MLFFNN) [115]. Notre choix est justifié par la capacité de MLFFNN à opérer sans l'assistance continue de l'utilisateur. De plus, MLFFNN réduit considérablement les efforts de calcul et les ressources en mémoire nécessaires pour stocker ses poids. Également, ce type de réseau de neurones est très robuste en présence d'incertitudes et de bruits, comme dans le cas des images satellitaires.
La figure 4.4 décrit l'architecture MLFFNN proposée. Les entrées sont les caractéristiques des objets. Nous avons une couche cachée, une couche de sortie et n sorties (différents types de couverture terrestre). w et b désignent respectivement le poids et le biais associé à la connexion entre les différentes unités des couches de réseau de neurones.

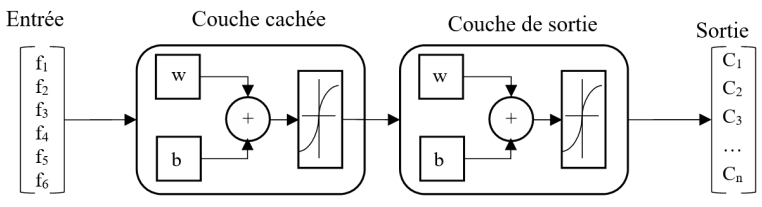


Figure 4.4 — Proposed MLFFNN architecture.

4.3.1.2 Traitement au niveau pixel : bas niveau

L'objectif du traitement de bas niveau est de déterminer la classe de chaque pixel d'une image d'entrée en fonction des caractéristiques calculées pour ce pixel. Ceci est assuré en utilisant la structure du réseau de neurones déjà construite.
Les caractéristiques décrites dans la section 3 sont des caractéristiques calculées au niveau objet et elles ne peuvent pas être calculées au niveau du pixel. Pour cela, nous considérons une matrice à 8 connexions centrée sur chaque pixel, comme illustrée dans la figure 4.5. Ensuite, les caractéristiques calculées au niveau objet peuvent être calculées sur cette matrice et ce pour chaque pixel dans l'image.

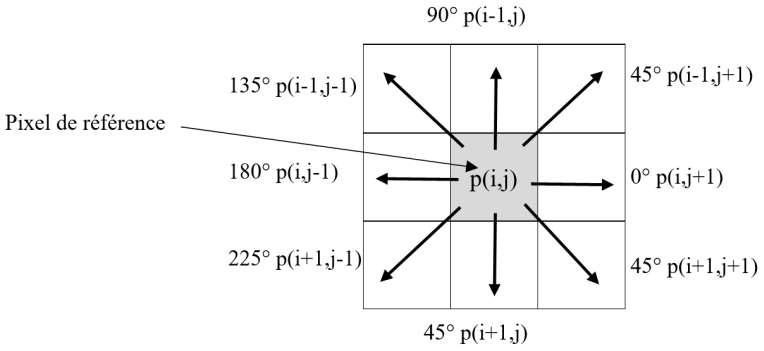


Figure 4.5 — La matrice à 8 connexions centrée sur le pixel de référence.

Une fois que le MLFFNN est formé, validé et testé, nous pouvons l'utiliser pour déterminer le type de la classe de chaque pixel. Les caractéristiques (radiométrie, contraste, corrélation, énergie, homogénéité et entropie) de chaque pixel pour une image d'entrée sont calculées. Ensuite, ces caractéristiques sont fournies à la structure MLFFNN. Ainsi, nous obtenons une segmentation sémantique de l'image.
La deuxième étape de l'approche proposée est de transférer les données satellitaires dans un entrepôt de données.

4.3.2 Processus d'intégration des données satellitaires

Le processus d'intégration vise à charger des données provenant de diverses sources généralement hétérogènes dans l'entrepôt de données. Dans notre travail, nous avons proposé un processus automatique et incrémental pour l'intégration des données satellitaires. Cela permet de mettre à jour l'entrepôt de données. Les données sont extraites des sources et mises initialement dans une zone de chargement (loading/staging area en anglais) puis dans l'entrepôt des données satellitaires. L'approche proposée évite d'inclure des données redondantes en identifiant les données qui ont été insérées ou mises à jour lors du dernier cycle de chargement des données. Parmi les opérations assurées par le processus d'intégration des données les trois opérations ETL (Extraction, Transformation, Chargement) acronyme de "Extract, Transform, Load" en anglais.
La figure 4.6 illustre le processus d'intégration des données satellitaires.

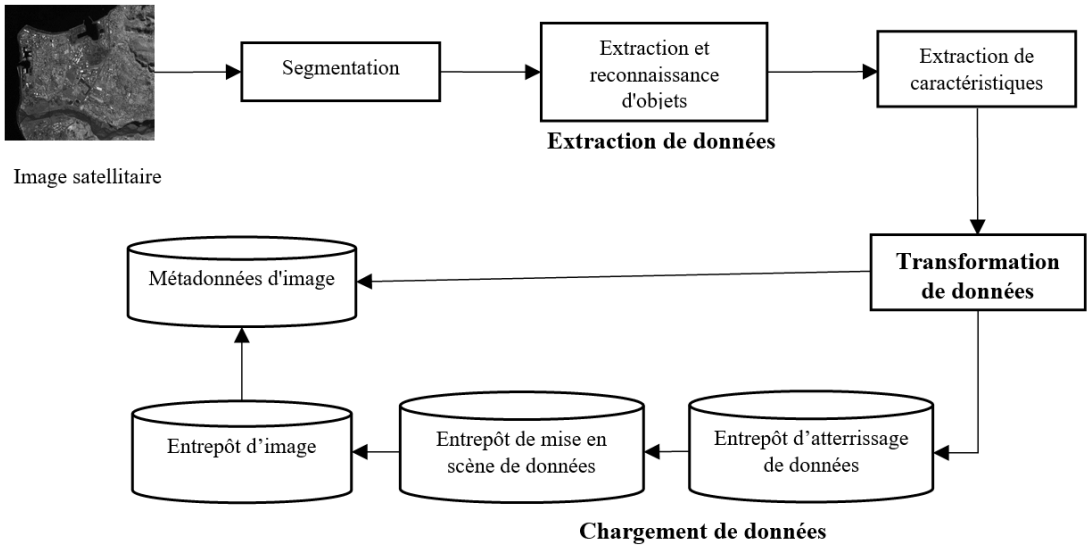


Figure 4.6 — Le processus d'intégration des données satellitaires.

4.3.2.1 Extraction des données

La première tâche du processus ETL est l'extraction des données. Elle vise à extraire les données utiles à partir d'images satellitaires. Pour ce faire, trois sous-tâches sont séquentiellement effectuées : 1) la segmentation, 2) la reconnaissance et l'extraction des objets, et 3) le traitement des caractéristiques. Les détails de ces étapes sont déjà décrits dans la section 4.3.1.
Les caractéristiques fixées pour décrire les objets extraits sont présentées dans la section 1.2.3 du chapitre 1 de la première partie.
Les caractéristiques considérées dans cette étude peuvent être modifiés en fonction du domaine d'application. Par exemple, pour le processus décisionnel dans le domaine médical, les caractéristiques peuvent être des caractéristiques du patient (nom, sexe, groupe sanguin, date de naissance, etc.), des caractéristiques de l'examen (équipement, type de la technologie radio, raison, etc.) ou des caractéristiques de l'hôpital (nom, adresse, ville, etc.).

4.3.2.2 Transformation des données

Cette tâche a pour but d'assurer le nettoyage, la validation et la conformité des données extraites des images satellitaires pour répondre aux exigences de l'entrepôt de données. Plusieurs opérations de transformation peuvent être effectuées telles que : scission, union, fusion, recherche et modifications (valeurs, types et structures).

4.3.2.3 Chargement des données satellitaires

Vu la taille importante des données à charger, nous avons opté pour un processus ETL en trois étapes. Les données satellitaires sont d'abord extraites, intégrées et chargées dans une zone d'atterrissage qui a une architecture similaire à celle des sources de données. Ceci permettra aux utilisateurs d'explorer, visualiser et analyser les données satellitaires avant de les valider. Ensuite, les données sont déplacées vers la zone de transfert qui a une architecture proche de l'entrepôt de données. La zone de transfert est nécessaire car le processus de lecture des données satellitaires est incrémentiel, et de nombreuses transformations et validations sont nécessaires avant de charger les données dans l'entrepôt.

4.3.3 Modélisation des données satellitaires

Nous avons choisi un schéma de modélisation en étoile pour notre entrepôt de données. Il contient une table de faits (appelée FactState) et sept tables de dimensions (DimObject, DimNDVI, DimTexturalFeatures, DimSpectralFeatures, DimDate, DimClimateFeatures, DimShapeFeatures). L'entrepôt de données conçu dans notre travail offre différents points de vue en fonction des caractéristiques spécifiques sélectionnées. Ainsi, nous pouvons chercher des objets similaires en fonction de caractéristiques texturales, spectrales, NDVI, climatiques ou de formes. Ceci peut être utile dans de nombreux cas, tels que le suivi des modifications de la couverture terrestre d'un objet en fonction de caractéristiques spécifiques ou la détermination des caractéristiques qui ont une influence sur le changement d'un objet.

La figure 4.7 illustre l'entrepôt de données proposé. La table des faits est placée au centre du schéma en étoile [78]. Il contient une clé étrangère issue des sept tables-dimensions et des mesures nécessaires pour plusieurs applications en imagerie satellitaire. Les tables de dimension entourent la table des faits et contiennent les clés primaires et les attributs décrivant les objets. L'architecture proposée pour l'entrepôt de données proposé permet d'obtenir plusieurs vues de cet entrepôt tout en choisissant un sous-ensemble de caractéristiques représentant les états d'un objet donné. Par conséquent, les utilisateurs peuvent évaluer les changements des objets en fonction des caractéristiques de l'objet (spectrales, texturales, climatiques ou de formes).

4.3.4 Interprétation des données satellitaires

L'objectif principal de l'étape d'analyse de données satellitaires est d'étendre les capacités de l'entrepôt de données en ajoutant des mesures métier utiles et des indicateurs de performance (KPI). Le modèle de données construit fournit une abstraction des domaines d'imagerie satellitaire en fonction des besoins des utilisateurs.
L'étape de création de rapports vise à fournir les informations tirées de la solution analytique. Cette étape résume l'ensemble des efforts déployés au cours du processus de modélisation. La tâche de création de rapports vise à fournir des représentations

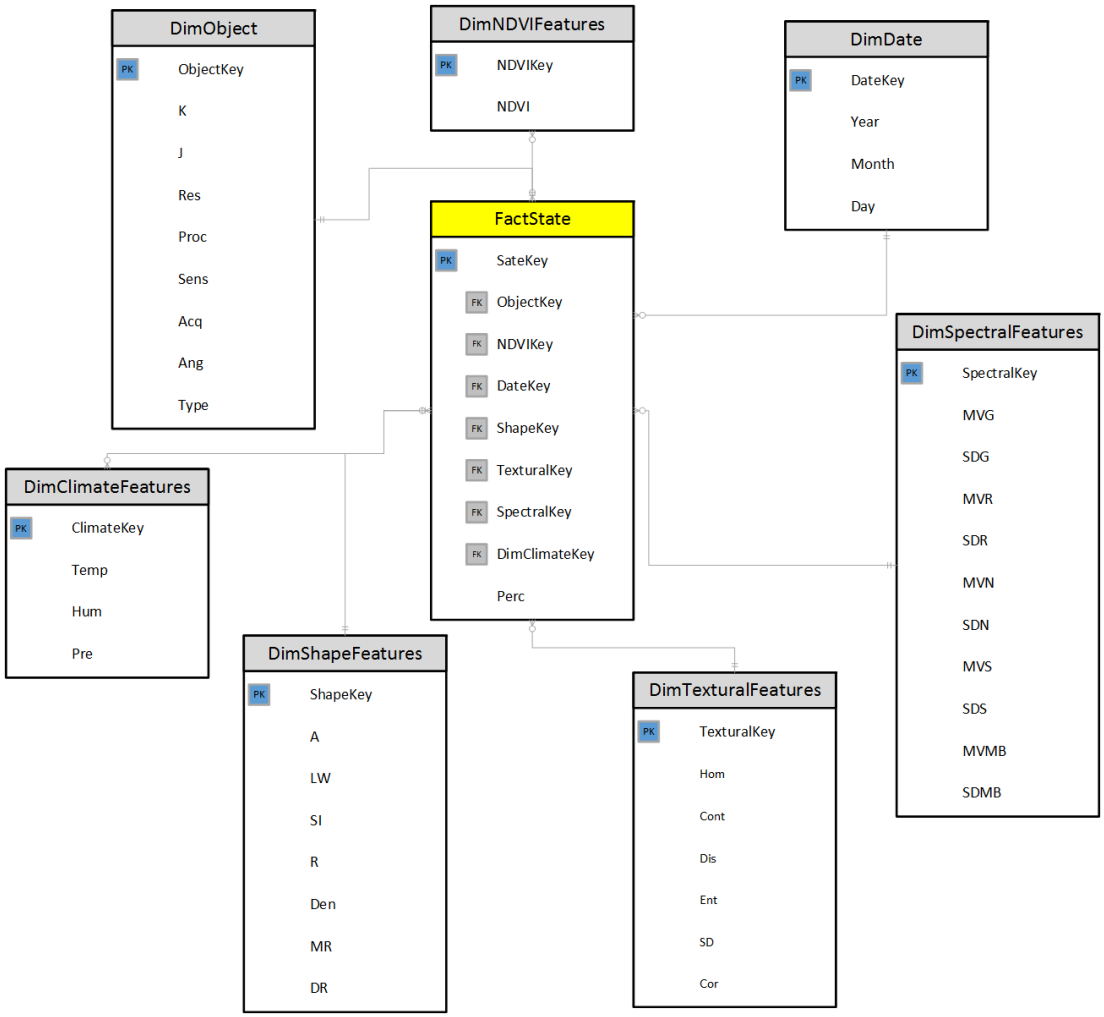


Figure 4.7 — Architecture de l'entrepôt de données.

visuelles et à donner des décisions informatives et prédictives. L'architecture proposée est divisée en trois étapes principales (Figure 4.8) : 1) moteur de requête sémantique, 2) traitement des rapports, et 3) rendu des rapports. La première étape vise à obtenir des données provenant essentiellement de deux sources (relationnelle et multidimensionnelle). Dans l'étape de création du rapport, le type de rapport en fonction des besoins de l'utilisateur est choisi. La dernière étape est le rendu du rapport. Son objectif est de choisir le mode de livraison des rapports (navigateur Web, messagerie électronique, application bureautique ou personnalisée).

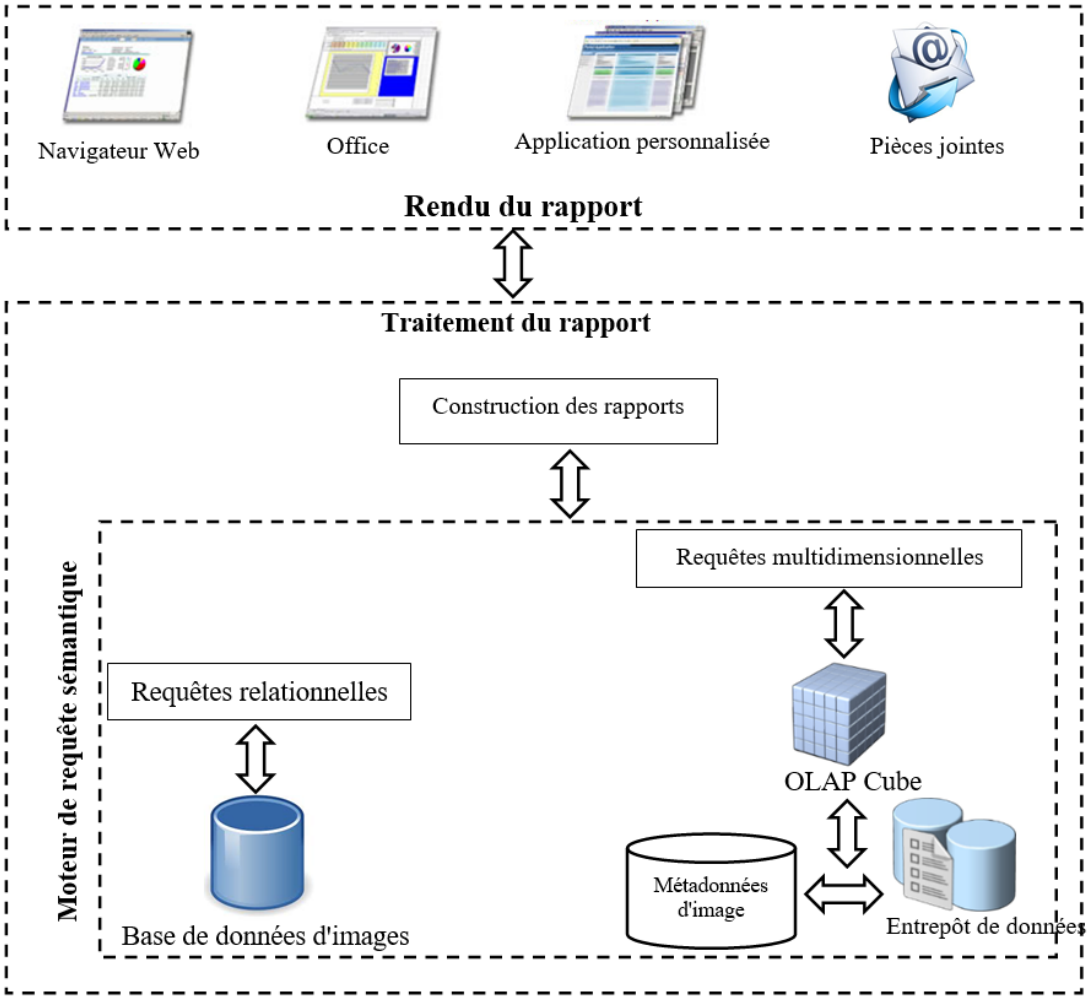


Figure 4.8 — Approche proposée pour la création de rapports.

4.4 Expérimentations

Cette section est divisée en deux parties : expérimentation de l'approche de segmentation sémantique des images satellitaires et expérimentation de l'approche d'analyse et d'interprétation.

4.4.1 Segmentation sémantique des images satellitaires

Afin de valider notre approche de segmentation sémantique, nous prenons un ensemble de données contenant 293 images représentant le deuxième site d'études (île de la Réunion) décrit dans le chapitre 1 de la première partie.
La résolution spatiale des images est de 10m par pixel. La taille de chaque image varie

de (3000x3000 pixels) à (6000x6000 pixels). Plusieurs imagettes sont extraites de ces images. Ensuite, une segmentation basée sur l'algorithme k-means est effectuée pour ces imagettes. Le tableau 4.2 décrit des échantillons des objets extraits des imagettes segmentées en tenant compte d'une connectivité de 8 pixels et d'un nombre minimal de 100 pixels.

Tableau 4.1 — Nombre d'échantillons pour chaque type d'occupation des sols

Classe	Type d'occupation des sols	Nombre d'échantillons
C1	Eau	45160
C2	Forêt	247120
C3	Urbain	151640
C4	Sol nu	131920
C5	Végétation non dense	270560

Tableau 4.2 — Nombre d'échantillons pour chaque type d'occupation des sols

La figure 4.9 présente un extrait d'échantillons pour chaque type d'occupation des sols.

La figure 4.10 décrit des images de vérité de terrain pour trois imagettes extraites de l'ensemble de données considéré. Pour obtenir les images de vérité de terrain, des informations ont été extraites par des experts sur les zones étudiées. Les polygones des régions étudiées de l'île de la Réunion sont numérisés pour obtenir les informations thématiques à l'aide d'une carte topographique à l'échelle 1/50000. Les informations topographiques permettent de déterminer les classes thématiques des zones étudiées. Cinq classes thématiques sont identifiées, à savoir : urbain, eau, forêt, sol nu et zones de végétation non dense.

4.4.1.1 Description du modèle neuronal

La validation du modèle neuronal est faite en utilisant nprtool fourni par Matlab R2008a [92]. Cet outil utilise une fonction nommée patternnet pour classifier les images satellitaires. Cette fonction est basée sur un réseau de neurones à propagation avant qui permet de classifier les pixels en fonction des classes cibles. Patternnet a trois paramètres d'entrée, hiddenLayerSizes, trainFcn et performFcn, et renvoie un réseau neuronal de reconnaissance de forme. HiddenLayerSizes est le nombre de couches cachées, fixé à 10 dans le présent document.
TrainFcn est la fonction d'entrainement fixée à trainbr dans notre travail. Trainbr est la régularisation bayésienne pour la rétropropagation du gradient. Il s'agit d'une fonction d'entraînement qui met à jour les valeurs de poids et de biais en fonction de l'optimisation de Levenberg-Marquardt [82]. performFcn est la fonction de performance fixée à crossentropy dans notre travail. Elle calcule la performance du MLFFNN en fonction des objectifs et des résultats.

Les données utilisées dans cet article sont divisées en 70% pour l'entrainement, 15% pour la validation et 15% pour les tests. L'objectif de l'ensemble de validation est de surveiller l'erreur de classification et d'arrêter l'entrainement avant que ne survienne un sur-apprentissage. L'ensemble de tests est ensuite utilisé pour évaluer la qualité de la classification. Le processus d'entraînement de MLFFNN est effectué de manière itérative 100 fois en utilisant un processeur graphique GeForce GTX 1080 de NVIDIA avec une mémoire de 8 Go.
La figure 4.11 décrit les performances du MLFFNN en montrant les erreurs d'apprentissage, de validation et de test. Nous pouvons conclure à partir de cette figure que la meilleure performance de validation a été atteinte à la période 108 avec un taux d'erreur de 0,047969. De plus, les courbes de validation et de test sont très similaires, ce qui implique qu'il n'y a pas eu de sur-apprentissage significatif [103].

4.4.1.2 Segmentation sémantique des images satellitaires

L'objectif de cette section est d'évaluer les performances de l'approche proposée pour la segmentation sémantiques d'images.
La figure 4.12 (à gauche) représente une imagette extraite de la base de données Kalideos décrite précédemment. Cette imagette ne fait pas partie du jeu de données d'entrainement. L'imagette est acquise le 31 janvier 2015 et provient du satellite SPOT 5 (Satellite Pour L'Observation de la Terre). L'imagette considérée a une résolution spatiale de 10 m et une taille de 800x500 pixels. La figure 4.12 (à droite) décrit la segmentation sémantique de l'imagette par l'approche proposée.
Les résultats de la segmentation sont comparés à l'image de vérité du terrain représentant la même région à la même date. La comparaison est effectuée en utilisant deux critères : la précision globale (OA) et le coefficient kappa (K). OA est la somme des pixels correctement classés divisée par le nombre total de pixels de l'image. K est une mesure de précision qui compare les résultats de classification proposés aux résultats réels. Il prend des valeurs de zéro à un (des valeurs plus élevées du coefficient kappa signifient une bonne classification). K est défini comme suit :

$$K = \frac{n \sum_{i=1}^{k} n_{ii} - \sum_{i=1}^{k} n_{i+} n_{+i}}{n^2 - \sum_{i=1}^{k} n_{i+} n_{+i}} \tag{4.7}$$

Où
k désigne le nombre de classes.
n est le nombre total de pixels dans les images.
n_{ii} est la somme des pixels correctement classés pour la classe i (le nombre de pixels appartenant à la classe i dans la vérité au sol qui ont également été classés dans la classe i dans l'image classée).
n_{i+} est la somme des pixels classés dans la classe i dans la classification d'image proposée.
n_{+i} est le nombre de pixels classés dans la classe i dans l'image de vérité du terrain.

		Image vérité de terrain					
		Eau	Forêt	Urbain	Sol nu	Végétation non-dense	Précision d'utilisateur
Image segmentée	Eau	94.31	0.45	2.05	2.29	0.9	94.31
	Forêt	0.09	89.05	2.75	3.01	5.1	89.05
	Urbain	1.3	0.26	90.85	4.56	3.03	90.85
	Sol nu	2.09	0.2	4.5	91.18	2.03	91.18
	Végétation non-dense	0.04	2.71	1.62	1.75	93.88	93.88
	Précision du Producteur	96.40	96.09	89.26	88.70	89.46	
Précision globale de classification=91.85, Kappa=0.8982							

Tableau 4.3 — Matrice de confusion de la segmentation proposée.

Le tableau 4.3 décrit la matrice de confusion de la segmentation proposée pour l'image présentée à la figure 4.12. Les lignes indiquent les classes de l'image de la vérité du terrain, tandis que les colonnes représentent les classes de la segmentation de l'image. Comme indiqué dans le tableau 4.3, l'approche proposée effectue une bonne segmentation de l'image avec un $OA = 91{,}85\%$ et un $K = 0{,}8982$.

4.4.1.3 Evaluation de l'approche proposée

Pour évaluer les performances de l'approche de segmentation d'images, nous comparons les résultats de notre approche avec des méthodes de classification existant dans la littérature. La comparaison inclut SVM (Machines à vecteur de support) [67] et la classification MLC (Maximum de vraisemblances) [66].
Le tableau 4.4 présente une comparaison de la classification d'images entre SVM, MLC et l'approche proposée en fonction de la précision globale et le coefficient kappa. Comme nous le remarquons, l'approche proposée dépasse les deux autres méthodes pour l'image présentée à la figure 4.12.

Méthode	OA	K
SVM	88.34	0.8590
MLC	86.11	0.8277
Approche proposée	91.85	0.8982

Tableau 4.4 — Comparaison de la classification d'images entre SVM, MLC et l'approche proposée

La figure 4.13 illustre la précision globale de la classification des images en fonction de la taille du jeu d'apprentissage pour les trois méthodes : SVM, MLC et la méthode proposée. La taille de l'ensemble d'apprentissage varie entre 100 et 800 000 échantillons. Nous pouvons remarquer que les trois méthodes ont été influencées positivement par la taille de l'ensemble d'apprentissage. Le SVM et l'approche proposée fournissent les meilleurs résultats dans tous les cas. L'OA passe de 79,8% (cas de 100 échantillons) à 87,4% (cas de 800 000 échantillons) pour la méthode SVM. Alors que l'OA passe de 71,5% (cas de 100 échantillons) à 84,1% (cas de 800 000 échantillons). pour la méthode MLC et de 74,2% (cas de 100 échantillons) à 91,6% (cas de 800 000 échantillons) pour la méthode proposée. En outre, l'approche proposée fournit les meilleurs résultats,

en particulier lorsque la taille de l'ensemble d'apprentissage devient plus importante (supérieure à 200 000).

La figure 4.14 décrit l'erreur de classification des images entre le SVM, le MLC et l'approche proposée. Nous pouvons noter que le SVM est moins sensible à la taille de l'ensemble d'apprentissage avec une différence de 7,6% entre la taille de 100 et 800 000 échantillons. La MLC arrive en deuxième position avec une différence de 12,6% et l'approche proposée en troisième place avec une différence de 17,4%.

La méthode SVM dépasse la méthode MLC dans toutes les situations, quelle que soit la taille de l'ensemble d'apprentissage. Bien que SVM donne de bons résultats pour la classification des images, un grand jeu d'apprentissage peut influer la précision de cette méthode. Cette observation est compatible avec les résultats rapportés dans la littérature [53].
La segmentation des images est une étape cruciale et nécessaire dans plusieurs tâches d'analyse et d'interprétation des images. Dans ce qui suit, nous présentons l'expérimentation de l'approche proposée pour l'analyse et l'interprétation des images.

4.4.2 Analyse et interprétation

Afin de valider notre approche, nous avons appliqué trois modèles différents pour analyser et interpréter les données satellitaires : le modèle de classification, le modèle d'arbre de décision et le modèle de règles d'association.

4.4.2.1 Classification

La première application de l'approche proposée est de regrouper les objets stockés dans l'entrepôt de données en fonction de leurs attributs.

La figure 4.15(a) présente un regroupement d'objets satellitaires dans l'entrepôt d'image en fonction de l'attribut NDVI. Nous obtenons pour chaque groupe un ensemble d'objets regroupés par leurs caractéristiques. La figure 4.15(b) décrit les différents états de l'attribut NDVI. Les groupes seront ombrés en conséquence de la modification de l'état NDVI. Les lignes entre deux groupes représentent la relation entre ces deux groupes (cela signifie que les deux groupes sont corrélés). L'intensité de la ligne montre le degré de corrélation ; si l'épaisseur est élevée, cela signifie qu'il existe une relation forte entre les deux groupes ou que ces deux grappes sont similaires.

La figure 4.16 présente une comparaison de groupes en fonction des différentes valeurs d'attributs. La figure 4.16(a) montre une comparaison de deux groupes (groupes 7 et 2) en fonction des attributs NDVI et de l'homogénéité. La figure 4.16(b) montre une comparaison d'un groupe et de son complément. Ceci aidera à comprendre les attributs qui identifient un groupe par rapport aux groupes d'objets.

4.4.2.2 Arbre de décision

Un deuxième type d'analyse et d'interprétation est fourni par l'approche proposée à savoir la détermination des types d'occupation des sols pour les objets extraits à partir des images satellites. L'un des avantages importants de l'approche proposée est d'étudier l'effet des attributs d'objets sur l'identification de la classe d'un objet donné. En outre, l'approche proposée fournit une appartenance floue à chaque type d'occupation des sols dans chaque nœud d'arbre. Plusieurs chemins peuvent être suivis pour classifier les objets. Chaque chemin fournit une appartenance floue de l'objet aux différents types d'occupation des sols. Ces chemins peuvent être traduits en règles de classification.

Dans l'exemple que nous avons pris, nous utilisons le NDVI et l'Homogénéité (Hom) pour classifier les objets. La figure 4.17(a) illustre l'arbre de classification floue basé sur NDVI et Hom. Comme nous le notons, à chaque nœud de l'arbre, nous avons une indication des types d'occupation des sols (couleurs dans les nœuds) auxquels un attribut donné peut conduire. La figure 4.17(b) illustre la translation de l'arbre de classification en une règle décrivant la classification de différents types d'occupation des sols. La figure 4.17(c) montre les différents types d'occupation des sols et leurs appartenances floues pour les attributs NDVI et Hom.

La figure 4.18 montre l'influence des attributs sur le COS. Ainsi, nous pouvons identifier quel attribut a la plus grande influence sur les COS (ici, nous notons que le pourcentage de changement dépend profondément de l'attribut NDVI).

4.4.2.3 Règles d'association

Une autre méthode d'analyse des données satellitaires proposée par notre approche est l'utilisation des règles d'association. Cette méthode peut être utilisée pour montrer les associations entre les attributs d'un objet satellitaire. Les règles d'association découvertes aident les interprètes et les décideurs à découvrir les attributs corrélés des objets et leurs impacts.
La figure 4.19 décrit la méthode des règles d'association utilisée pour identifier la corrélation qui peut exister entre les attributs ; associations qui conduisent à des COS bien déterminés. Ces règles illustrent les COS (pourcentage de changements), les différentes valeurs d'attributs impliquées dans les changements et la confiance pour ces changements (valeurs à côté des barres bleues).
La figure 4.20 présente l'impact des plages des valeurs de l'NDVI et l'Hom sur le changement d'un objet donné. Ceci permet de comprendre le lien entre les valeurs des attributs et les changements d'un objet donné.

4.5 Conclusions et perspectives de recherche

Les principaux problèmes soulevés dans ce chapitre sont : 1) la complexité des données satellitaires, 2) la modélisation des données satellitaires et 3) l'analyse et l'interprétation des données satellitaires.
Pour répondre à ces problèmes, nous avons proposé une approche en quatre étapes : segmentation sémantique des images satellitaires, intégration des données, modélisation des données, analyse et interprétation.
Afin d'expérimenter l'approche proposée, nous avons commencé par valider et évaluer l'approche de segmentation sémantique des images satellitaires. Ensuite, nous avons proposé d'utiliser trois modèles pour expérimenter l'approche proposée d'analyse et d'interprétation. Ces modèles (à savoir la classification, l'arbre de décision et les règles d'association) permettent d'exploiter les données stockées et préparées lors de la phase d'intégration et de la phase de modélisation des données satellitaires.
Plusieurs perspectives peuvent être envisagées à partir de notre travail :

- L'approche proposée peut être améliorée en considérant le problème de la détermination du nombre de nœuds de couches cachées et de leurs poids respectifs. Aussi, pour pouvoir calculer les attributs au niveau pixel, nous avons proposé d'utiliser une matrice à 8 connexions centrée sur ce pixel. L'effet de choix peut être discuté en changeant la dimension de la matrice et en évaluant les résultats de la classification obtenus.
- Une fois que nous avons réussi à extraire des informations utiles à partir des données satellitaires, le défi important est de savoir comment fournir ces informations à de nombreux types d'utilisateurs. Ainsi, il serait intéressant de proposer un cadre intelligent basé sur Internet des objets (en anglais Internet of Things, ou IoT) pour fournir de nombreux services environnementaux pour plusieurs types d'utilisateurs tels que les planificateurs régionaux, les utilisateurs travaillant dans les secteurs de la prévision climatique, la gestion des foules, la gestion des ressources, la gestion de l'utilisation / de la couverture des sols, la planification de l'étalement urbain, etc.
- Développement de techniques pour l'extraction de connaissances à partir des images satellitaires en utilisant de nouveaux algorithmes d'apprentissage automatique tels que l'apprentissage profond (en anglais deep learning), l'apprentissage par renforcement (en anglais reinforcement learning) et le calcul évolutif (en anglais evolutionary computation).

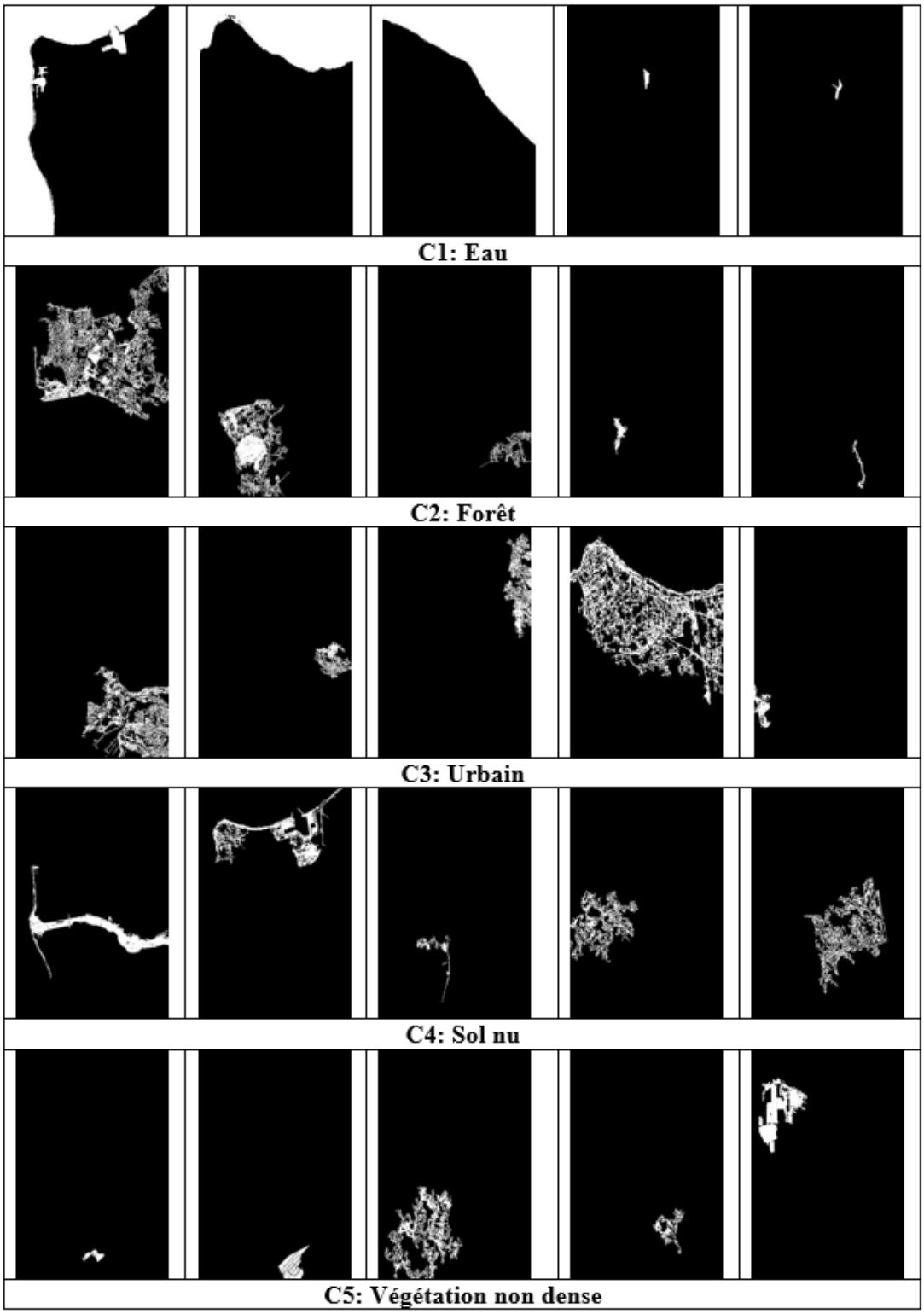


Figure 4.9 — Extrait d'échantillons pour chaque type d'occupation des sols.

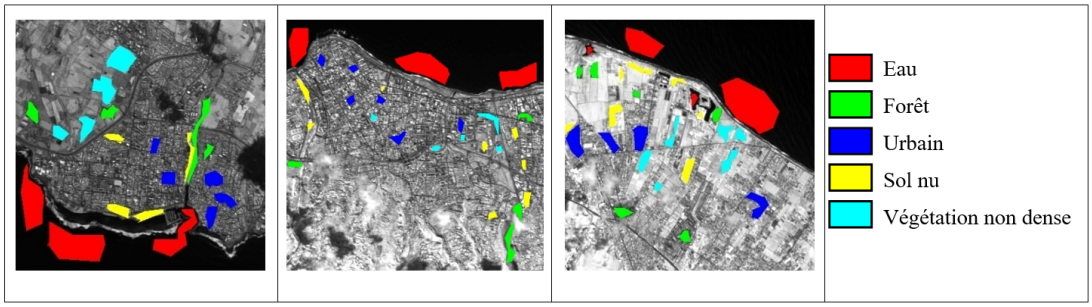


Figure 4.10 — Vérité de terrain pour trois imagettes extraites de l'ensemble des données considéré.

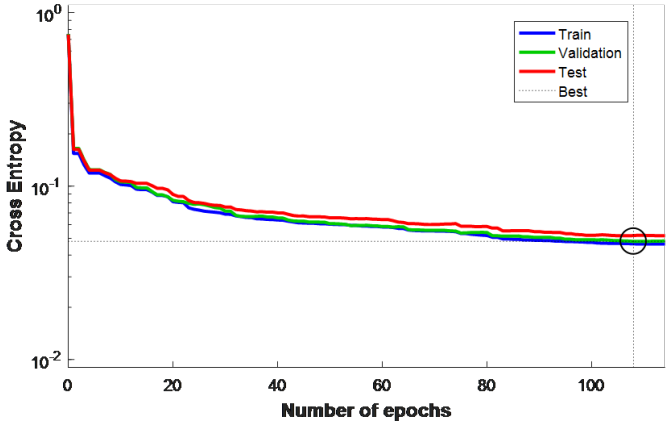


Figure 4.11 — Performance du modèle neuronal.

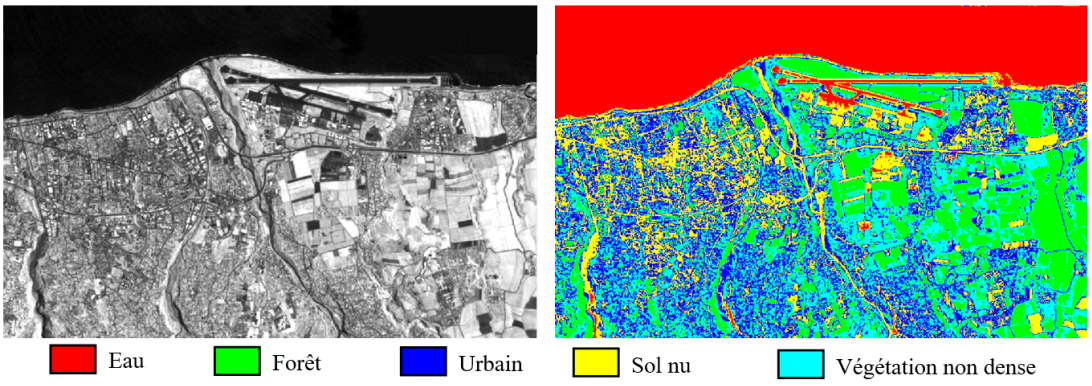
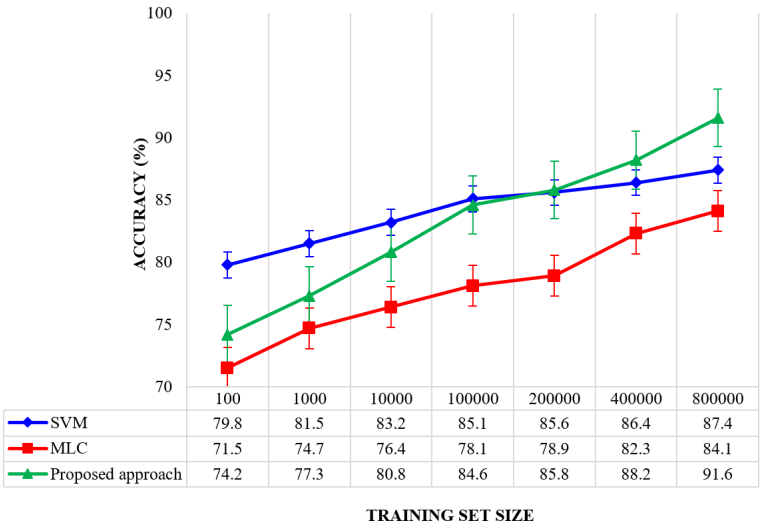


Figure 4.12 — Image satellite acquise le 31 janvier 2015 (à gauche) et la segmentation réalisée par l'approche proposée (à droite).



	100	1000	10000	100000	200000	400000	800000
SVM	79.8	81.5	83.2	85.1	85.6	86.4	87.4
MLC	71.5	74.7	76.4	78.1	78.9	82.3	84.1
Proposed approach	74.2	77.3	80.8	84.6	85.8	88.2	91.6

Figure 4.13 — Précision de la classification des images en fonction de la taille du jeu d'apprentissage.

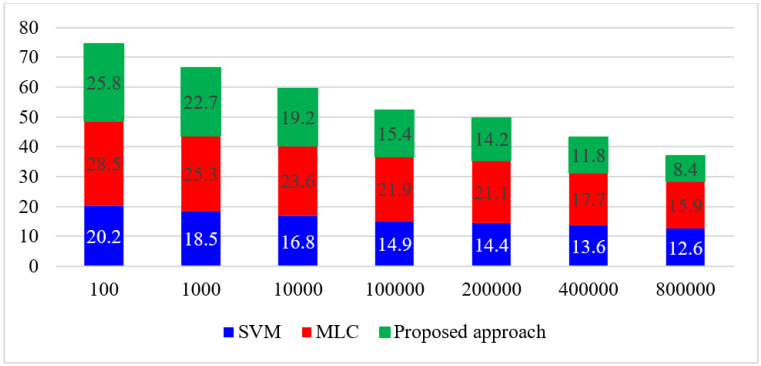


Figure 4.14 — Erreur de classification des images en fonction de la taille du jeu d'apprentissage.

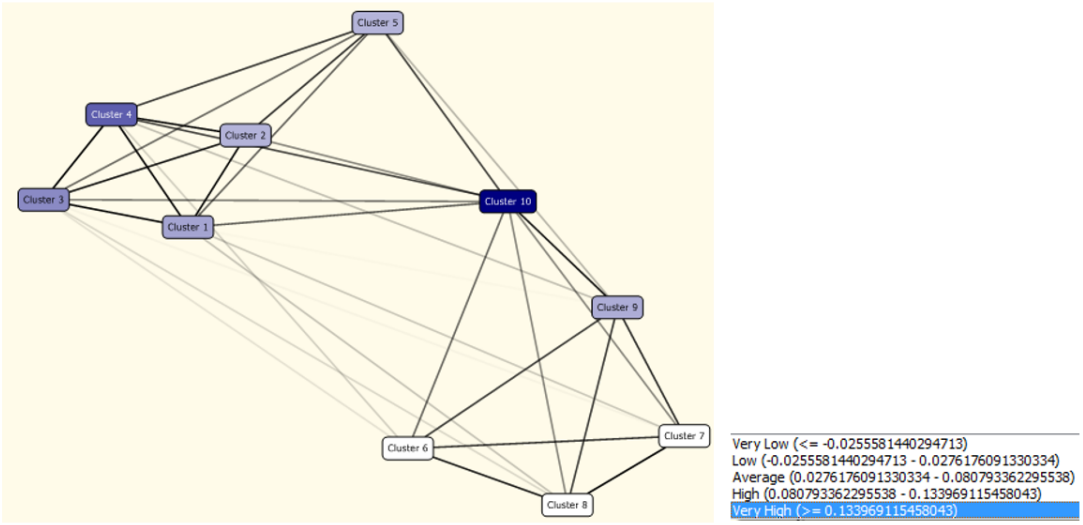


Figure 4.15 — (a) Regroupement d'objets satellitaires selon l'attribut NDVI et (b) différents états de l'attribut NDVI.

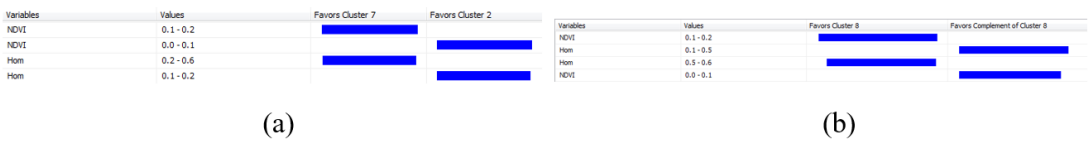
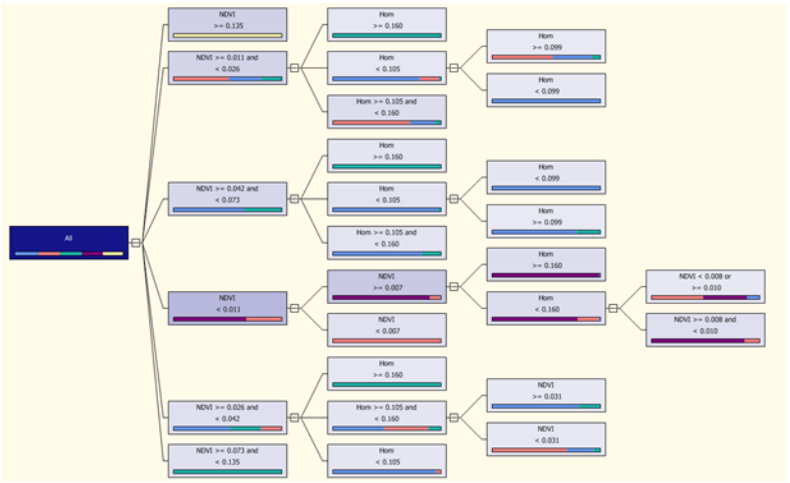
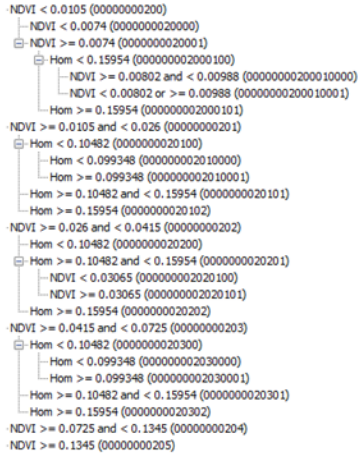
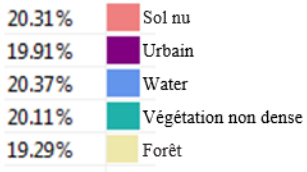


Figure 4.16 — Comparaison des groupes. (a) groupe par rapport à un autre. (b) groupe par rapport à son complémentaire.



(a) Identification de l'occupation des sols pour les attributs NDVI et Hom





(c) Types d'occupation des sols et leurs appartenances floues pour les attributs NDVI et Hom

Figure 4.17 — Identification du type d'occupation des sols des objets selon les attributs NDVI et Hom.

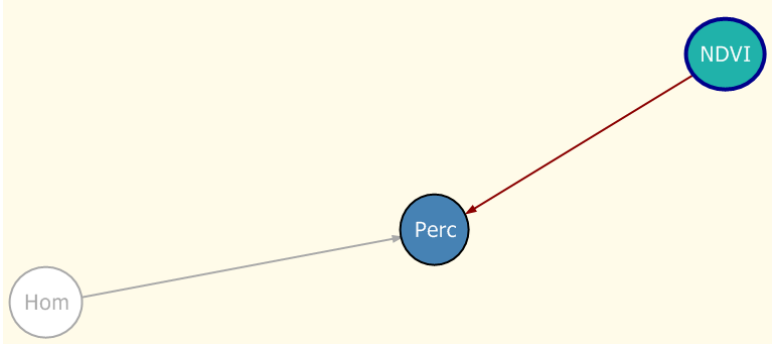


Figure 4.18 — Influence des attributs sur le COS.

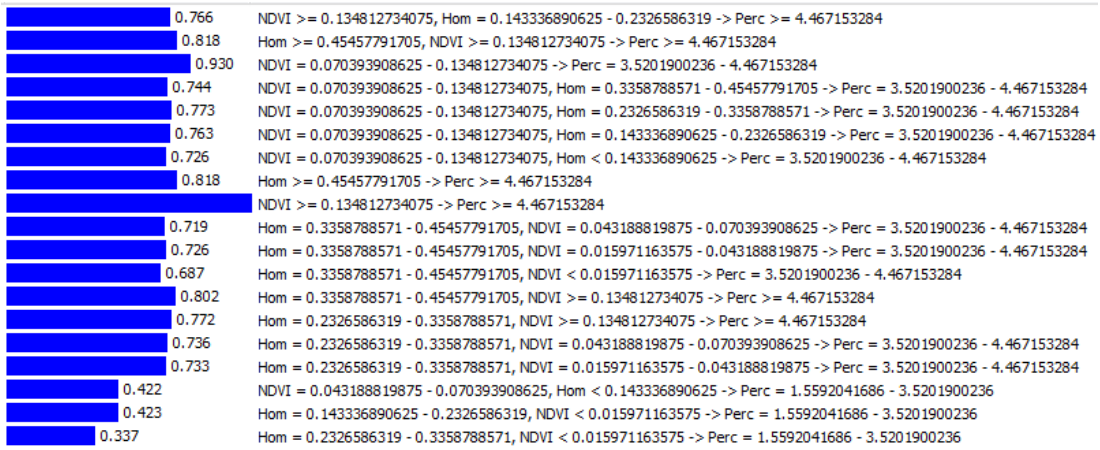


Figure 4.19 — Identification de COS en utilisant les règles d'association.

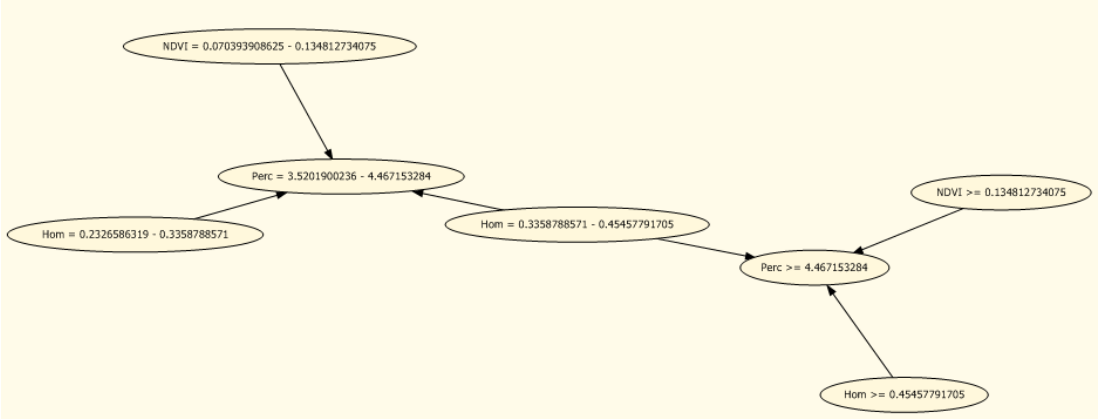


Figure 4.20 — Identification des types d'occupation des sols des objets.

Chapitre 5

Analyse et interprétation des données satellitaires massives

Sommaire

5.1 Introduction

La quantité croissante des données de télédétection a ouvert la porte à de nouveaux sujets de recherche difficiles. De nos jours, des efforts importants sont consacrés à l'analyse de gros volume de données satellitaires. Ce chapitre s'intéresse à la classification de gros volume d'images satellitaires. Pour ce faire, nous avons proposé une approche basée sur une architecture distribuée Apache Spark. L'approche proposée utilise l'apprentissage profond (le réseau neuronal convolutif) pour faire la classification des images.
Ce présent chapitre est structuré comme suit. La deuxième section décrit l'état de l'art et les problématiques étudiées. Dans cette section, nous commençons par décrire des

concepts et des définitions qui sont liés au domaine des données massives. Ensuite, nous présentons les plateformes les plus utilisés pour le domaine des données massives suivi d'une synthèse des travaux similaires et les problèmes à résoudre. La troisième section définit l'approche proposée. Dans cette section, nous détaillons les différents composants de cette approche à savoir le nœud maître, les nœud esclaves, le module d'apprentissage profond et le gestionnaire de cluster. La dernière partie de ce chapitre est consacrée pour les conclusions et les perspectives de notre travail.

5.2 Etat de l'art et problématiques étudiées

L'expansion continue du volume des données satellitaires rend les tâches de traitement, d'analyse et d'interprétation de ces données de plus en plus difficile. Cette expansion a fait apparaître un nouveau concept pour le domaine de l'imagerie satellitaire qui est le "big data" ou les données massives.
Il n'y a pas de définition universelle des données massives ou ce que nous appelons "big data". Dans la littérature, plusieurs définitions ont été annoncées :

- IDC [55] : les technologies big data décrivent une nouvelle génération de technologies et architectures conçues pour extraire économiquement des informations à partir d'un volume de données de très grandes tailles et d'une grande variété tout en permettant la capture à grande vitesse, ainsi que la découverte et / ou l' analyse.
- McKinsey Global Institute [88] "big data" fait référence à des ensembles de données dont la taille est au-delà de la capacité des outils logiciels de base de données permettant de capturer, stocker, gérer et analyser les données. Nous remarquons que cette définition du big data peut varier selon le secteur, selon les plateformes de traitement de données et la taille des ensembles de données. Avec ces mises en garde, le big data dans de nombreux secteurs aujourd'hui va de quelques dizaines de téraoctets à plusieurs pétaoctets (milliers de téraoctets).
- Oracle [33] : le big data sont des données caractérisées par 4 les attributs clés : volume, variété, vélocité et valeur.

Dans la littérature, le big data est généralement caractérisé par les 5Vs [25] :

- Volume : le big data renvoie aux données volumineuses. En fait, il ne doit pas être un certain nombre de pétaoctets à qualifier. Si les données sont devenues si grandes au point que vous avez du mal à les gérer, on parle alors du big data.
- Variété : elle renvoie à la diversification des données. Ceci est dû à la diversification des usages d'internet et du numérique. La variété est liée à la provenance des données, leur format et le domaine d'application.
- Vélocité ou vitesse : elle renvoie à la fréquence de génération, de partage et de capture de données.
- Véracité : elle réfère à l'exactitude des données. Elle désigne la crédibilité des données à traiter. Ces données sont issues de plusieurs sources ; il est souvent difficile de justifier son authenticité.

– Valeur : elle renvoie à la capacité de transformer les données en valeurs. Avoir un gros volume de données ne signifie pas toujours un gros volume de connaissances. Il est important que les entreprises établissent une analyse de rentabilisation pour toute tentative de collecte et de valorisation du big Data. Avant de se lancer aux domaines liés au big data, il faut faire une étude sur les coûts et les avantages de la collecte et de l'analyse des données pour faire en sorte que ces données récoltées soient finalement monétisées.

Actuellement, il existe plusieurs plateformes qui ont été proposées pour manipuler les données massives. Parmi lesquels nous citons Hadoop [56], Apache Spark [107], Apache Storm [108] et HPCC [65].

Dans notre travail, nous nous intéressons à la classification de gros volume d'images satellitaires. Ce domaine est en plein essor et plusieurs travaux ont été effectués dans le domaine des données massives et plus particulièrement en imagerie.
Cavallaro et al. [22] ont évalué les techniques de parallélisation lors du travail avec les données massives. Les auteurs ont conclu que le temps total de l'ensemble du processus d'analyse des images satellitaires peut être considérablement réduit en utilisant des méthodes de parallélisation, ce qui permet de l'utiliser même lorsque des techniques d'extraction et de sélection de caractéristiques et des méthodes d'analyse spatiale sont appliquées. Cavallaro et al. ont constaté que la valeur ajoutée de l'utilisation de techniques de parallélisation pour les données massives est plus élevée pour ceux ayant moins d'échantillons d'entraînement. Il est encore possible d'appliquer des techniques d'extraction de caractéristiques non seulement pour augmenter la précision d'un classificateur et ainsi réduire le volume de données mais également pour réduire le nombre de cycles de calcul nécessaires.
Dans [38], les auteurs ont présenté un cadre de traitement nommé ICP (Image Cloud Processing) pour faire face à l'explosion des données dans le domaine du traitement d'images. Dong et al. ont proposé de développer un système permettant d'exécuter plusieurs algorithmes de traitement d'images en parallèle. Le cadre proposé pour l'ICP comprend deux mécanismes, à savoir l'ICP statique (SICP) et l'ICP dynamique (DICP). Le SICP vise à traiter les données de grande tailles pré-stockées dans le système distribué, tandis que le DICP est proposé pour une entrée dynamique. Pour réaliser le SICP, deux nouvelles représentations des données nommées p-image et big-image sont conçues pour coopérer avec MapReduce. Le DICP est mis en œuvre via une procédure de traitement parallèle avec le mécanisme de traitement traditionnel du système distribué. L'approche proposée est validée à partir du jeu de données ImageNet.
Chebbi et al. [26] ont présenté un état de l'art sur les algorithmes les plus utilisés qui s'appliquent aux données spatiales, en particulier lorsque le volume d'images satellitaires devient très grand. Les auteurs ont proposé d'utiliser le cadre de travail Hadoop/MapReduce et HDFS en intégrant les outils de traitement d'images satellitaires OTB (en anglais Orfeo ToolBox) dans MapReduce. L'application de l'approche proposée consiste à classer les images satellitaires.
Dans [79], les auteurs ont proposé d'analyser de gros volumes d'images biomédicales.

Deux architectures sont proposées pour la classification des images biomédicales. Kouanou et al., 18 ont utilisé le cadre de travail Hadoop pour concevoir la première architecture et le cadre de travail Spark pour la deuxième. Les auteurs ont constaté que l'architecture Spark est plus complète, car elle facilite la mise en œuvre d'algorithmes avec ses bibliothèques intégrées. Dans leurs travaux futurs, les auteurs ont suggéré de développer / mettre en œuvre une application réelle de l'approche théorique qu'ils ont proposée en utilisant le cadre de travail Spark.
Imamoglu et al. [71] ont étudié l'efficacité d'utiliser le réseau neuronal convolutif (CNN) doté de caractéristiques récurrentes et de rétroaction pour la classification des centrales solaires sur des images satellitaires multispectrales à moyenne résolution. Les auteurs ont proposé un réseau CNN récurrent et un système CNN-Feedback basés sur un modèle à propagation avant pour la classification d'images multispectrales. L'expérimentation faite a montré que l'utilisation de signaux haut et bas (en particulier les caractéristiques récurrentes et de rétroaction) sur les CNN peut fournir une bonne représentation des images multispectrales, ce qui peut considérablement améliorer la précision de la classification. En tant que perspective de leur travail, Imamoglu et al. ont projeté d'étudier des réseaux récurrents tels que les réseaux convolutionnels LSTM (en anglais Long-Short Term Memory networks) ou GRU (en anglais Gated Recurrent Units) pour améliorer les résultats de leur approche. Selon l'étude de l'état de l'art, nous constatons que malgré les contributions élaborées pour l'analyse de gros volume de données dans le domaine de l'imagerie, peu de travaux se sont focalisés sur l'imagerie satellitaire et plus précisément en appliquant de nouvelles techniques d'apprentissage profond pour la classification des images satellitaires.
Les données massives désignent des contenus numériques de grande taille, hétérogènes et non structurés. Ces données sont difficiles à traiter à l'aide d'outils et de techniques de gestion de données traditionnels [110]. Ainsi, la proposition et le développement de nouveaux cadres et outils pour l'analyse et l'interprétation de données volumineuses en imagerie satellitaire sont devenus une nécessité pour les utilisateurs de la télédétection. Ainsi dans le cadre de la classification d'un gros volume d'images satellitaires, plusieurs problèmes peuvent être posés :

- Un cadre de travail permettant de stocker le gros volume de données satellitaires : les plateformes traditionnelles ne permettent pas de stocker, traiter et analyser un gros volume de données. Ces plateformes sont conçues pour travailler avec des données structurées. Cependant, les données massives peuvent se trouver sous plusieurs formes : structurées, semi structurées et non structurées. D'autres part, la nécessité d'avoir un temps de calcul proche du réel devient de plus en plus une exigence pour les utilisateurs des systèmes informatiques. Le besoin est donc double : besoin d'informations pertinentes et en temps rapide.
- La classification d'un gros volume d'images satellitaires : en général, il est difficile d'extraire des informations pertinentes à partir d'un gros volume de données. Le défi majeur est de découvrir des informations dans le bruit. Les techniques traditionnelles d'apprentissage automatique et d'exploration des données souffrent de nombreuses difficultés pour analyser et traiter efficacement les données massives

surtout en imagerie satellitaire.

Dans cette section, nous avons décrit quelques concepts qui sont importants pour s'initier au domaine des données massives ou "big data". Dans ce qui suit, nous allons présenter notre approche de classification de gros volumes d'images satellitaires.

5.3 Approche proposée pour l'analyse des données massives en imagerie satellitaire

Cette section décrit l'architecture et les étapes de l'approche proposée pour la classification de gros volumes d'images satellitaires. Elle est divisée en deux parties : architecture de l'approche proposée et étapes d'exécution de la classification des images satellitaires.

5.3.1 Architecture de l'approche proposée

L'approche proposée est basée sur une architecture de traitement de données massives en utilisant Apache Spark cluster. L'entrée du système proposé sont un gros volume d'images satellitaires et la sortie sont des images classifiées (Figure 5.1). Apache Spark suit une architecture maître/esclave avec deux démons principaux et un gestionnaire de cluster. Les deux démons sont nœud maître (processus maître ou processus pilote) et nœud esclave (nœud travailleur ou processus esclave). Nous parlons, généralement, de Spark cluster. Un cluster est un ensemble de machines connectées les unes aux autres. Un cluster Spark a un seul maître et plusieurs esclaves.

5.3.1.1 Nœud maître

C'est le point central et le point d'entrée de l'architecture Spark. Le SparkDriver est un coordinateur dès qu'il recevra les informations du maître Spark. Le Spark driver distribue les tâches de manière homogène aux exécuteurs et il reçoit également les informations de ces derniers. Spark driver contient divers composants tels que DAGScheduler (couche de planification d'Apache Spark qui implémente la planification par étapes), TaskScheduler (responsable d'envoyer des tâches au cluster, de les exécuter et de les réessayer s'il y a des échecs), BackendScheduler (soutien des divers gestionnaires de cluster) et BlockManager (cache local qui s'exécute sur chaque nœud d'un cluster Spark). Ces composants sont responsables de la traduction du code utilisateur de Spark en des tâches Spark exécutées sur le cluster.
Le nœud maître contient le SparkContext qui peut être qualifié de maître de l'application Spark. SparkContext permet au SparkDriver d'accéder au cluster via le gestionnaire de cluster. SparkContext permet de nombreuses fonctions telles que : obtenir la configuration actuelle du cluster pour exécuter ou déployer l'application, définir la nouvelle configuration, créer des objets, planifier des travaux, annuler des travaux, etc.

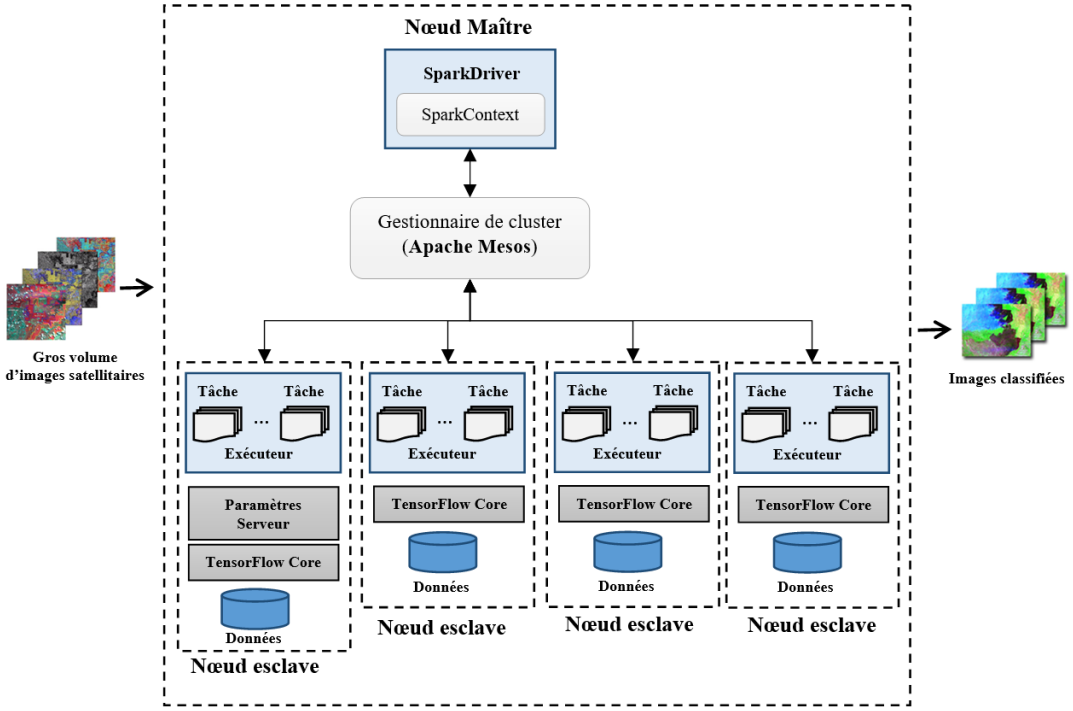


Figure 5.1 — Architecture de l'approche proposée.

5.3.1.2 Nœud esclave

C'est un nœud distribué responsable de l'exécution des tâches. Ce nœud contient des exécuteurs (executor en anglais). Les exécuteurs s'exécutent généralement pendant toute la durée de vie d'une application Spark. Ce phénomène est appelé "affectation statique des exécuteurs". Toutefois, les utilisateurs peuvent également opter pour des allocations dynamiques d'exécuteurs, dans lesquelles ils peuvent ajouter ou supprimer dynamiquement des exécuteurs Spark afin de correspondre à la charge globale de travail. Les tâches principales de l'exécuteur sont : 1) effectuer tout le traitement des données, 2) lire et écrire des données sur des sources externes, 3) stocker les résultats des calculs en mémoire, en cache ou sur des disques durs, et 4) interagir avec les systèmes de stockage.
Dans notre architecture, nous avons choisi de travailler avec cinq nœuds esclave.

5.3.1.3 Gestionnaire de cluster

C'est un service externe chargé d'acquérir des ressources sur le cluster Spark et de les affecter à des tâches Spark. Dans notre travail, nous avons choisi d'utiliser Apache Mesos comme gestionnaire de cluster pour la gestion de l'allocation et la désallocation de diverses ressources physiques. Notre choix est argumenté par le fait que le gestionnaire de cluster Mesos est hautement évolutif. Il détermine les ressources disponibles et

comment affecter les tâches d'un travail donné entre ces ressources lorsqu'une demande de travail arrive au maître Mesos. Ensuite, il envoie sa proposition au nœud maître pour que ce dernier l'accepte ou la refuse.

5.3.1.4 TensorFlow Core

TensorFlow est un système d'apprentissage automatique fonctionnant à grande échelle et dans des environnements hétérogènes. Il a été principalement créé pour construire et former des modèles d'apprentissage en profondeur. TensorFlow utilise des graphes de flux de données pour faire le calcul et pour représenter les états partagés, et les opérations qui mutent de ces états. Il mappe les nœuds d'un graphe de flux de données sur de nombreuses machines dans un cluster.
Dans notre travail, nous allons utiliser TensorFlow pour assurer la tâche de la classification des images satellitaires.

5.3.1.5 Noyau d'Apache Spark

Le noyau d'Apache Spark est constitué par les RDD (Jeux de données distribués résilients ou en anglais Resilient Distributed Datasets) qui constituent les principales caractéristiques de Spark.
RDD sont une collection d'objets distribués simples et immobiles. C'est un ensemble d'éléments tolérants aux pannes pouvant être exploités en parallèle. Chaque RDD est divisé en plusieurs partitions pouvant être calculées sur différents nœuds du cluster. Dans Spark, toutes les fonctions sont exécutées sur des RDD uniquement.
Spark RDD prennent en charge deux types d'opérations différents : transformations et actions. Une opération de transformation consiste à créer un nouvel ensemble de données à partir du RDD précédent. Parmi les opérations de transformation, nous citons : map, filter, mapPartitions, union, intersection, distinct et groupByKey. Par contre, une opération d'action renvoie une valeur au SparkDriver après avoir effectué le calcul sur l'ensemble de données. Parmi les operations d'action, nous citons reduce, collect, count, foreach, saveAsObjectFile et takeSample.
Parallèlement aux RDD, nous trouvons aussi les DataFrames (DF) qui sont des composants essentiels pour le noyau d'Apache Spark. Un DF est une collection distribuée de données organisées en colonnes nommées. Il est conceptuellement équivalent à une table dans une base de données relationnelle, mais avec des optimisations plus riches. Les DF peuvent être construits à partir d'un large éventail de sources telles que : des fichiers de données structurés, des tables dans Hive, des bases de données externes ou des RDD existants.

5.3.2 Etapes d'exécution de la classification des images satellitaires

L'algorithme 3 décrit les étapes de la classification des images satellitaires au sein d'un cluster Spark.

Algorithme 3 Etapes de la classification de gros volumes d'images satellitaires

Entrée: images satellitaires
Sortie: images satellitaires classifiées

// Etape 1 : Analyse de données volumineuses utilisant ML dans Spark
1: Convertir les images en niveaux de gris
2: Préparer un ensemble de données pour l'entraînement avec des étiquettes pour les classes de chaque pixel de l'image.
3: Entrer un jeu de données sous la forme d'un RDD
4: Convertir RDD en DF
5: Lire les caractéristiques et les étiquettes de DF
6: Exécuter le hot-encoding des caractéristiques non numériques
7: Indexer sous forme de chaîne les caractéristiques encodées
8: Assembler en vecteur les caractéristiques one-hot-encoded et les valeurs numériques
9: Convertir le vecteur assemblé en un pipeline
10: Adapter et transformer le pipeline en une forme appropriée pour que Spark puisse le lire
11: Entraîner le modèle en utilisant des fonctionnalités basées sur MLLib et en utilisant les données d'entraînement
12: Tester toutes les données pour obtenir une valeur de prédiction binaire de l'étiquette
// Etape 2 : Apprentissage en profondeur (CNN)
13: Ajouter les données de prédiction au fichier de jeu de données d'origine.
14: Former un perceptron multicouche (MLP) en utilisant les connaissances obtenues à l'étape précédente
15: Construire le MLP en répétant les étapes 2-10 de la phase 1.
16: Entraîner le CNN et réduire l'erreur de prédiction en utilisant le réseau de rétro-propagation

Dans le présent travail, nous avons utilisé le réseau neuronal convolutif (CNN) pour la classification des images [80].
Pour chaque image satellitaire, le processus SparkDriver qui s'exécute sur le nœud maître du cluster Spark planifie l'exécution du travail et négocie avec le gestionnaire de cluster. Il traduit les RDD dans le graphe d'exécution et les divise en plusieurs étapes. Il convertit implicitement le code contenant les transformations et les actions en un graphe acyclique dirigé (DAG). SparkDriver effectue également certaines optimisations, telles que les transformations de traitement en pipeline, puis convertit le DAG logique en un plan d'exécution physique avec un ensemble d'étapes. Après avoir créé le plan d'exécution physique, il crée de petites unités d'exécution physiques appelées tâches et ceci à chaque étape. Les tâches sont ensuite regroupées pour être envoyées au cluster Spark. SparkDriver communique ensuite avec le gestionnaire de cluster et négocie les ressources. Le gestionnaire de cluster lance ensuite les exécuteurs sur les nœuds de travail pour le compte du SparkDriver. Ce dernier envoie des tâches au gestionnaire de cluster en fonction du placement des données. Avant de commencer l'exécution, les exécuteurs s'inscrivent eux-mêmes auprès du SparkContext afin que celui-ci ait une vue globale de tous les exécuteurs. Les exécuteurs commencent à exécuter les différentes tâches assignées par le SparkDriver. À tout moment de l'exécution de l'application de classification des images, SparkDriver surveille l'ensemble l'exécution des différents exécuteurs. Il planifie également les tâches futures en fonction de l'emplacement des

données en cache.

5.4 Conclusions et perspectives de recherche

Dans ce chapitre, nous avons proposé une approche de classification des images satellitaires. L'approche proposée est basée sur une architecture maître/esclave Apache Spark. Elle est composée d'un nœud maître, de plusieurs nœuds esclaves et d'un gestionnaire de cluster. Afin de classifier les images, nous avons eu recourt à l'apprentissage profond en intégrant le cadre de travail TensorFlow avec Apache Spark. De plus, les étapes du mécanisme de classification de gros volumes d'images satellitaires ont été détaillées dans ce chapitre.
Ce travail pourra être étendu en considérant plusieurs pistes :

- Avec la présence de grandes quantités de données, un nouveau problème posé appelé "malédiction de la dimensionnalité" se pose. Il fait référence au traitement d'espaces de grandes dimensions et à la gestion d'un nombre important de configurations possibles. Ainsi, il est important d'étudier les méthodes de réductions des dimensions sans perdre le contenu informationnel que présente les données.
- Les techniques traditionnelles d'apprentissage automatique et d'exploration des données trouvent leurs limites pour analyser et traiter efficacement les données satellitaires volumineuses. Ainsi, il est important de creuser dans des pistes reliées à de nouvelles techniques d'apprentissage automatique telles que l'apprentissage profond. A ce niveau, deux types d'approches pouvant être explorées : ou bien une distribution des données entre les différents nœuds de l'architecture du big data ou bien une distribution des modèles de l'apprentissage profond entre ces différents nœuds. Pour la première approche, une copie complète du modèle serait appliquée aux données existant dans chaque nœud esclave. Ensuite, les résultats seront envoyés au nœud maître pour les combiner. Cependant, pour la deuxième approche, nous avons différentes copies du modèle dans chaque nœud esclave (par exemple, chaque couche du réseau de neurones profond peut être affectée à un nœud différent).
- Une fois que nous avons réussi à extraire des informations utiles à partir de données satellitaires volumineuses, le défi important est de savoir comment fournir ces données à de nombreux types d'utilisateurs. Ainsi, une bonne perspective de notre travail sera de proposer un cadre intelligent basé sur l'IoT pour fournir de nombreux services environnementaux pour plusieurs types d'utilisateurs tels que les planificateurs régionaux, les utilisateurs travaillant dans les secteurs de la prévision climatique, la gestion des foules, la gestion des ressources, la planification de l'étalement urbain, etc.

Bilan et perspectives

De nos jours, avec l'avènement de nouveaux capteurs en télédétection, les capacités d'acquisition des images satellitaires ne cessent d'évoluer. Nous nous retrouvons face à un besoin croissant aux nouvelles méthodes de traitement et d'analyse des images satellitaires à très haute résolution spatiale. La richesse du contenu de ces images est utile pour une vaste palette d'applications. Parmi ces applications, nous pouvons citer le changement de l'occupation des sols (COS). Durant nos travaux de recherche en thèse, nous avons proposé une approche de prédiction de COS. Comme les données satellitaires sont le plus souvent entachées par plusieurs types d'incertitudes et ces dernières se trouvent dans plusieurs niveaux du processus de prédiction de COS, nous avons proposé une gestion multi-niveaux des données incertaines. Ces niveaux sont relatifs aux données, à la prédiction et aux résultats. Cependant, la propagation des incertitudes d'un niveau à un autre n'a pas été étudiée dans le contexte de nos travaux de thèse. Suite à nos travaux de thèse, nous nous sommes concentré à l'étude de la propagation des incertitudes dans les modèles de COS.

Dans un premier temps, nous nous sommes intéressé à l'étude des différents types et sources d'incertitudes des modèles de COS. Nous avons pu distinguer deux types d'incertitudes à savoir les incertitudes aléatoires et épistémiques. Le premier type désigne la variabilité naturelle alors que le deuxième provient du manque de connaissances sur les modèles de COS. Ces incertitudes peuvent être liées aux paramètres d'entrée des modèles ou aux modèles eux-mêmes (structure des modèles). Pour faire face à la propagation de ces incertitudes dans les modèles de COS, nous avons proposé une approche basée sur la théorie des fonctions de croyance. Dans notre approche, les incertitudes aléatoires sont modélisées par des distributions de probabilités puis transformées en des structures évidentielles alors que les incertitudes épistémiques sont modélisées par des structures évidentielles. En second lieu, nous avons propagé les incertitudes (aléatoires/épistémiques) liées aux paramètres tout en considérant la corrélation entre ces paramètres à travers les modèles de COS. De même, les incertitudes liées à la structure des modèles sont également propagées par la théorie des fonctions de croyance. L'approche de propagation proposée a été validée à travers quatre modèles de COS à savoir DINAMICA, SLEUTH, CA-MARKOV et

LCM. Ces modèles diffèrent dans la façon dont ils simulent les COS. Les résultats de l'approche proposée montrent l'importance de la prise en compte des différents types d'incertitude liées aux paramètres d'entrée. Une autre constatation importante relevée par les expériences montre l'importance de la prise en considération des corrélations entre les paramètres d'entrée lors de la propagation des incertitudes. Finalement, nous pouvons affirmer que la modélisation des incertitudes de la structure du modèle est essentielle pour améliorer les décisions des modèles de COS.

Parallèlement à la propagation des incertitudes, nous avons dirigé nos efforts aussi sur la réduction des incertitudes. Nous avons proposé trois approches de réduction d'incertitudes. La première approche consiste à utiliser la méthode de sensibilité globale des dérivées (DGSM) pour réduire les incertitudes liées aux modèles de COS. Le processus de réduction des incertitudes commence par appliquer une analyse de sensibilité qualitative basée sur la méthode de criblage de Morris pour identifier les paramètres d'entrée incertains du modèle de COS. Les valeurs des paramètres non incertains sont fixées. Nous ne nous intéressons dans les étapes suivantes qu'aux paramètres incertains. Ensuite, une étude de corrélation est faite pour déterminer les paramètres corrélés et les classer en groupes. La dernière étape de notre deuxième approche est l'analyse de sensibilité qualitative basée sur DGSM pour déterminer les paramètres les plus influents sur la sortie du modèle de COS.

La deuxième approche de réduction des incertitudes est basée sur la théorie des fonctions de croyance. La stratégie du pincement est appliquée pour effectuer l'analyse de sensibilité dans le cadre de la théorie de croyance. Ce choix est justifié par le fait que la stratégie du pincement permet de quantifier des incertitudes représentées par des fonctions de croyance et de plausibilité en présence des incertitudes aléatoires/épistémiques et des corrélations entre les paramètres. L'approche proposée permet de hiérarchiser les sources d'incertitude de l'étude, c'est à dire de cibler les paramètres et les structures du modèle les plus influents sur la sortie nécessitant le plus de précision. Pour estimer les valeurs optimales des paramètres les plus influents sur la sortie des modèles de COS, nous avons eu recours aux limites de confiance du Kolmogorov-Smirnov.
L'expérimentation de trois approches de réduction a été faite à travers plusieurs jeux de test réels et plusieurs comparaisons ont été faites avec des approches existantes. Les résultats ont montré les bonnes performances des approches proposées en terme de réduction des incertitudes.

La troisième approche de réduction des incertitudesest appliquée au domaine de la bio-informatique et plus précisément à la compréhension du cancer des poumons. Pour ce faire, nous avons étudié les biomarqueurs biologiques potentiels du cancer du poumon sur différentes solutions de traitement. Plusieurs états de traitement par plasma non-thermique sont présents (NT, SE, LE 1hr, LE 2hrs et LE 4hrs). Pour chaque état, nous disposons de trois échantillons. Les valeurs de mesure des

biomarqueurs biologiques changent d'un échantillon à un autre, suite à des incertitudes (par exemple des erreurs de mesures). Le but est de combiner les informations issues de ces différents échantillons pour parvenir à une meilleure décision sur les biomarqueurs biologiques potentiels du cancer des poumons. Ceci est assuré en utilisant la fusion par la théorie des croyances. Un deuxième problème s'ajoute pour ce premier type d'application est de suivre le changement des gènes d'un état de traitement à un autre (en d'autres termes, le problème est de découvrir les gènes qui se sont influencés par le traitement). Pour répondre à ce problème, nous avons commencé par faire un regroupement des gènes dans le premier état de traitement NT. Ensuite, nous avons déterminé les limites supérieure et inférieure pour chaque groupe obtenu après le processus de fusion. Ces limites seront utilisées pour le regroupement des échantillons des autres états de traitement. Les gènes qui ont changé de groupe d'un état à un autre sont les gènes qui ont été influencés par le traitement par plasma non-thermique.

Notre deuxième axe de recherche est l'analyse et l'interprétation des images satellitaires. Cet axe de recherche se divise en deux volets. Dans le premier volet, nous nous sommes intéressé aux images satellitaires non massives. Alors que dans le deuxième volet, nous avons examiné le cadre de gros volumes d'images satellitaires.
Pour le premier volet, nous avons proposé une approche qui suit quatre étapes principales : 1) la segmentation sémantique des images satellitaires, 2) l'intégration des données satellitaires, 3) la modélisation des données satellitaires, et 4) l'analyse et l'interprétation.
La segmentation sémantique proposée est divisée en deux niveaux : 1) haut niveau et 2) bas niveau. Dans le premier niveau, nous commençons par calculer les attributs des objets extraits des images satellitaires. Ces attributs constituent l'entrée d'un module neuronal pour générer une structure permettant de classifier les objets issus des images satellitaires. Dans le deuxième niveau, la structure générée est utilisée pour effectuer la segmentation sémantique au niveau des pixels. Pour une image satellitaire d'entrée, une matrice centrée dans chaque pixel (une fenêtre 3x3 entourant le pixel) est prise en compte lors du calcul des caractéristiques associées à ce pixel. Les mêmes attributs calculés au niveau de l'objet sont calculés aussi au niveau pixel.
La deuxième étape qui est l'intégration des données a pour but de charger les données dans l'entrepôt de données. Plusieurs opérations sont appliquées pour préparer les données pour être chargées dans l'entrepôt de données. Le défi consiste à intégrer et à consolider un volume important de données satellitaires dans un entrepôt de données unifié. La troisième étape, la modélisation des données, a pour but de construire une solution d'analyse multidimensionnelle répondant aux exigences des différents utilisateurs du domaine de l'imagerie satellitaire. Le schéma choisi pour l'entrepôt de données est le schéma en étoile. La dernière étape de notre approche fournit un ensemble d'information descriptives et prédictives qui constituent une aide à la décision des utilisateurs dans différents domaines de la télédétection.

Dans le deuxième volet, nous avons choisi de faire une ouverture sur le domaine

des données satellitaires massives. D'abord, nous avons commencé par élaborer une étude bibliographique sur les données satellitaires massives, les caractéristiques de ces données et les plateformes les plus utilisées pour le stockage et le traitement de ces données. Ensuite, nous avons proposé une approche de classification des images satellitaires. L'approche proposée est basée sur une architecture Apache Spark se composant d'un nœud maître, de plusieurs nœuds esclaves et d'un gestionnaire de clusters. Les images d'entrée sont divisées et distribuées aux nœuds esclaves. Un cadre de travail basé sur TensorFlow est mis en place dans chacun de ces nœuds afin d'exécuter le réseau neuronal convolutif qui s'occupe de la classification de l'imagette reçue. Ceci permet d'assurer le parallélisme de la classification des images et diminue le temps de calcul tout en gardant une bonne précision de classification.

Perspectives de recherche

Les perspectives de nos travaux de recherche concernent deux nouveaux axes de recherche.

Le premier axe concerne le suivi spatiotemporel d'objets complexes en imagerie satellitaire en se basant sur les contraintes et les graphes. En effet, l'apparition des capteurs à très haute résolution spatiale a provoqué l'augmentation significative du détail dans les images satellitaires. Ainsi, nous pouvons distinguer deux types d'objets dans les images satellitaires : les objets simples et les objets composés. Les objets simples constituent la granularité la plus fine de l'image, alors que les objets complexes sont eux-mêmes composés d'un ensemble d'objets simples. En conséquence, la reconnaissance des objets en se basant uniquement sur des informations de bas niveau devient insuffisante. Dans ce cas, il faut reconnaître les objets (simples et complexes), leurs propriétés et l'organisation de ces objets (position, agencement spatial, etc.).
L'objectif du travail pour ce premier axe est de répondre aux trois problèmes suivants :1) construction des graphes spatiaux en se basant sur le réseau de contraintes pour extraire les objets complexes dans les images satellitaires, 2) construction des graphes d'évolution des objets complexes et les interpréter pour en déduire les changements de ces objets au cours du temps, 3) modélisation des incertitudes lors du suivi spatiotemporel des objets complexes. En outre, les relations spatiales peuvent être des relations simples (objet-objet) ou des relations complexes (objet-groupe ou groupe-groupe).

Le deuxième axe étudie l'analyse des données satellitaires massives en se basant sur l'IoT (Internet des Objets) et l'apprentissage profond. L'objectif de cet axe est de fournir des services environnementaux intelligents pour les utilisateurs dans les divers domaines de la télédétection. La combinaison entre les données satellitaires massives et l'IoT offrira la possibilité d'obtenir des informations précieuses à partir d'un grand

nombre de données collectées auprès de différentes sources et de les utiliser dans des services d'environnement réels au moyen d'une infrastructure IoT. Nous proposons alors d'explorer ce gros volume de données satellitaires en utilisant les nouvelles techniques d'apprentissage automatique telles que : l'apprentissage en profondeur, l'apprentissage par renforcement, le calcul évolutif, etc. Ces nouvelles techniques donnent des résultats intéressants lorsqu'il s'agit de traiter de gros volumes de données.
L'approche proposée est basée sur quatre étapes. La première étape vise à collecter des données environnementales multisources, dynamiques et hétérogènes. La deuxième étape consiste à nettoyer et à stocker ces données sur une grande plateforme de données. La troisième étape consiste à explorer les données satellitaires massives afin d'extraire des informations précieuses utiles pour de nombreux types d'applications. Cette étape est basée sur les nouvelles techniques d'apprentissage automatique. La quatrième étape a pour objectif de développer une infrastructure IoT sécurisée afin de fournir de nombreux services environnementaux aux citoyens et aux décideurs.
Afin de pouvoir développer une telle approche, nous devons prendre en compte les problèmes suivants :

- Analyse de données volumineuses : les données volumineuses proviennent de sources hétérogènes et dynamiques, tels que les images satellitaires, le système d'information géographique, les images aériennes et les médias sociaux. Ces données nécessitent de nombreuses tâches de prétraitement pour assurer une bonne qualité des données collectées. Parmi ces tâches, nous pouvons citer le nettoyage, la résolution des contradictions, la transformation, le filtrage et la validation. En outre, l'un des principaux problèmes liés aux données satellitaires massives est le stockage des données et leur traitement.
- Extraction efficace d'informations utiles pour les données volumineuses. En général, il est difficile d'obtenir des informations pertinentes sur une taille de données considérable. Le défi important est de découvrir d'information utiles parmi un gros volume de données. Les techniques traditionnelles d'apprentissage automatique et d'exploration de données trouvent leurs limites pour analyser et traiter efficacement les données satellitaires volumineuses.
- Prestation des services environnementaux aux citoyens et aux décideurs : le défi important est de savoir comment fournir l'information extraite à de nombreux types d'utilisateurs.

Bibliographie

[1] B. Abuelaish and M. T. C. Olmedo. Scenario of land use and land cover change in the gaza strip using remote sensing and gis models. *Arabian Journal of Geosciences*, 9(4) :274, 2016.

[2] F. Ai, L. K. Comfort, Y. Dong, and T. Znati. A dynamic decision support system based on geographical information and mobile social networks : A model for tsunami risk mitigation in padang, indonesia. *Safety science*, 90 :62–74, 2016.

[3] T. Ali, H. Boruah, and P. Dutta. Sensitivity analysis in radiological risk assessment using probability bounds analysis. *International Journal of Computer Applications*, 44(17) :1–5, 2012.

[4] C. M. D. Almeida, A. M. V. Monteiro, G. Camara, B. S. Soares-Filho, G. C. Cerqueira, C. L. Pennachin, and M. Batty. Gis and remote sensing as tools for the simulation of urban land-use change. *International Journal of Remote Sensing*, 26(4) :759–774, 2005.

[5] S. Andres, D. Arvor, I. Mougenot, T. Libourel, and L. Durieux. Ontology-based classification of remote sensing images using spectral rules. *Computers and Geosciences*, 102 :158–166, 2017.

[6] L. J. Bain and M. Engelhardt. *Introduction to probability and mathematical statistics*. Brooks/Cole, 1987.

[7] M. S. Balch. Mathematical foundations for a theory of confidence structures. *International journal of approximate reasoning : official publication of the North American Fuzzy Information Processing Society*, 53(7) :1003–1019, 2012.

[8] A. Banerjee and R. N. Dave. Validating clusters using the hopkins statistic. In *Fuzzy systems, 2004. Proceedings. 2004 IEEE international conference on*, volume 1, pages 149–153. IEEE, 2004.

[9] J. C. Bezdek, R. J. Hathaway, and J. M. Huband. Visual assessment of clustering tendency for rectangular dissimilarity matrices. *IEEE Transactions on Fuzzy Systems*, 15(5) :890–903, 2007.

[10] N. Bihamta, A. Soffianian, S. Fakheran, and M. Gholamalifard. Using the sleuth urban growth model to simulate future urban expansion of the isfahan metropo-

litan area, iran. *Journal of the Indian Society of Remote Sensing*, 43(2) :407–414, 2015.

[11] I. Bloch, A. Hunter, A. Appriou, A. Ayoun, S. Benferhat, P. Besnard, L. Cholvy, R. Cooke, F. Cuppens, and D. Dubois. Fusion : General concepts and characteristics. *International journal of intelligent systems*, 16(10) :1107–1134, 2001.

[12] E. Borgonovo, W. Castaings, and S. Tarantola. Moment independent importance measures : New results and analytical test cases. *Risk Analysis : An International Journal*, 31(3) :404–428, 2011.

[13] A. Bouatay, W. Boulila, and I. R. Farah. An approach for imperfection propagation : Application to land cover change prediction. In *International Conference on Artificial Intelligence and Soft Computing*, pages 637–648. Springer, 2014.

[14] W. Boulila. *Extraction de connaissances spatiotemporelles incertaines pour la prediction de changements en imagerie satellitale*. PhD thesis, Telecom Bretagne, 2012.

[15] W. Boulila. A top-down approach for semantic segmentation of big remote sensing images. *Journal of Earth Science Informatics*, pages 1–12, 2019.

[16] W. Boulila, M. Al-Kmali, M. Farid, and H. Mugahed. A business intelligence based solution to support academic affairs : case of taibah university. *Wireless Networks*, pages 1–8, 2018.

[17] W. Boulila, Z. Ayadi, and I. R. Farah. Sensitivity analysis approach to model epistemic and aleatory imperfection : Application to land cover change prediction model. *Journal of Computational Science*, 23 :58–70, 2017.

[18] W. Boulila, A. Bouatay, and I. R. Farah. A probabilistic collocation method for the imperfection propagation : Application to land cover change prediction. *JMPT*, 5(1) :12–32, 2014.

[19] W. Boulila, I. R. Farah, K. S. Ettabaa, B. Solaiman, and H. B. Ghezala. Spatio-temporal modeling for knowledge discovery in satellite image databases. In *CORIA*, pages 35–49, 2010.

[20] W. Boulila, I. R. Farah, K. S. Ettabaa, B. Solaiman, and H. B. Ghezala. A data mining based approach to predict spatiotemporal changes in satellite images. *International Journal of Applied Earth Observation and Geoinformation*, 13(3) :386–395, 2011.

[21] W. Boulila, I. R. Farah, and A. Hussain. A novel decision support system for the interpretation of remote sensing big data. *Earth Science Informatics*, 11(1) :31–45, 2018.

[22] G. Cavallaro, M. Riedel, M. Richerzhagen, J. A. Benediktsson, and A. Plaza. On understanding big data impacts in remotely sensed image classification using support vector machine methods. *IEEE journal of selected topics in applied earth observations and remote sensing*, 8(10) :4634–4646, 2015.

[23] M. Chakraborty, A. Skowron, M. Maiti, and S. Kar. Facets of uncertainties and applications. *ICFUA, Kolkata, India*, 125 :2194–1009, 2013.

[24] G. Chaudhuri and K. Clarke. The sleuth land use change model : A review. *Environmental Resources Research*, 1(1) :88–105, 2013.

[25] I. Chebbi, W. Boulila, and I. R. Farah. Big data : Concepts, challenges and applications. In *ICCCI 2015 7th International Conference on Computational Collective Intelligence*, pages 638–647, 2015.

[26] I. Chebbi, W. Boulila, and I. R. Farah. Improvement of satellite image classification : Approach based on hadoop/mapreduce. In *2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, pages 31–34, 2016.

[27] I. Chebbi, W. Boulila, N. Mellouli, M. Lamolle, and I. R. Farah. A comparison of big remote sensing data processing with hadoop mapreduce and spark. In *2018 4th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, pages 1–4, 2018.

[28] X. Chen and C. Jian. Gene expression data clustering based on graph regularized subspace segmentation. *Neurocomputing*, 143 :44–50, 2014.

[29] G. Cheng and J. Han. A survey on object detection in optical remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 117 :11–28, 2016.

[30] S.-K. Choi, R. Grandhi, and R. A. Canfield. *Reliability-based structural design*. Springer Science and Business Media, 2006.

[31] R. Chutia, S. Mahanta, and D. Datta. Sensitivity analysis of atmospheric dispersion model-rimpuff using the hartley-like measure. *Journal of applied mathematics and informatics*, 31(1-2) :99–110, 2013.

[32] S. Corgne. *Modelisation predictive de l'occupation des sols en contexte agricole intensif Application a la couverture hivernale des sols en Bretagne*. PhD thesis, Universite Rennes 2, novembre 2004.

[33] O. B. Data. 2018 (dernière visite le 23/06/2018). `https://www.oracle.com/bigdata/index.html`.

[34] L. Dempere-Marco, X.-P. Hu, S. L. MacDonald, S. M. Ellis, D. M. Hansell, and G.-Z. Yang. The use of visual search for knowledge gathering in image decision support. *IEEE Transactions on Medical Imaging*, 21(7) :741–754, 2002.

[35] D. Dhar. States of matter. *http ://arxiv.org/abs/0904.2664*, 5(4) :745–754, 2009.

[36] W. Diao, X. Sun, F. Dou, M. Yan, H. Wang, and K. Fu. Object recognition in remote sensing images using sparse deep belief networks. *Remote Sensing Letters*, 6(10) :745–754, 2015.

[37] O. Ditlevsen. Model uncertainty in structural reliability. *Structural safety*, 1(1) :73–86, 1982.

[38] L. Dong, Z. Lin, Y. Liang, L. He, N. Zhang, Q. Chen, X. Cao, and E. Izquierdo. A hierarchical distributed processing framework for big image data. *IEEE Transactions on Big Data*, 2(4) :297–309, 2016.

[39] D. Dubois and H. Prade. Possibility theory : qualitative and quantitative aspects. In *Quantified representation of uncertainty and imprecision*, pages 169–226. Springer, 1998.

[40] P. Dutta and S. Saha. Fusion of expression values and protein interaction information using multi-objective optimization for improving gene clustering. *Computers in biology and medicine*, 89 :31–43, 2017.

[41] I. R. Farah, W. Boulila, K. S. Ettabaa, and M. B. Ahmed. Multiapproach system based on fusion of multispectral images for land-cover classification. *IEEE Transactions on Geoscience and Remote Sensing*, 46(12) :4153–4161, 2008.

[42] I. R. Farah, W. Boulila, K. S. Ettabaa, B. Solaiman, and M. B. Ahmed. Interpretation of multisensor remote sensing images : Multiapproach fusion of uncertain information. *IEEE Transactions on Geoscience and Remote Sensing*, 46(12) :4142–4152, 2008.

[43] M. W. Farouq, W. Boulila, M. Abdel-aal, A. Hussain, and A. Badeh Salem. A novel multi-stage fusion based approach for gene expression profiling in non-small cell lung cancer. *IEEE Access*, 7 :37141–37150, 2019.

[44] E. H. Fegraus, I. Zaslavsky, T. Whitenack, J. Dempewolf, J. A. Ahumada, K. Lin, and S. J. Andelman. Interdisciplinary decision support dashboard : A new framework for a tanzanian agricultural and ecosystem service monitoring system pilot. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(6) :1700–1708, 2012.

[45] A. Ferchichi, W. Boulila, and I. R. Farah. An intelligent possibilistic approach to reduce the effect of the imperfection propagation on land cover change prediction. In *Computational Collective Intelligence*, pages 520–529. Springer, 2015.

[46] A. Ferchichi, W. Boulila, and I. R. Farah. Propagating aleatory and epistemic uncertainty in land cover change prediction process. *Ecological informatics*, 37 :24–37, 2017.

[47] A. Ferchichi, W. Boulila, and I. R. Farah. Towards an uncertainty reduction framework for land-cover change prediction using possibility theory. *Vietnam Journal of Computer Science*, 4(3) :195–209, 2017.

[48] A. Ferchichi, W. Boulila, and I. R. Farah. Reducing uncertainties in land cover change models using sensitivity analysis. *Knowledge and Information Systems*, 55(3) :719–740, 2018.

[49] S. Ferson and L. R. Ginzburg. Different methods are needed to propagate ignorance and variability. *Reliability Engineering and System Safety*, 54(2-3) :133–144, 1996.

[50] S. Ferson, V. Kreinovick, L. Ginzburg, and F. Sentz. Constructing probability boxes and dempster-shafer structures. Technical report, Sandia National Labs., Albuquerque, NM (US) ; Sandia National Labs, 2003.

[51] S. Ferson and W. T. Tucker. *Sensitivity in risk analyses with uncertain numbers.* Sandia National Laboratories Albuquerque, New Mexico, USA, 2006.

[52] G. Fishman. *Monte Carlo*. Springer Science and Business Media, 1996.

[53] G. M. Foody and A. Mathur. A relative evaluation of multiclass image classification by support vector machines. *IEEE Transactions on Geoscience and Remote Sensing*, 42(6) :1335–1343, 2004.

[54] D. A. Fordham, S. Haythorne, and B. W. Brook. Sensitivity analysis of range dynamics models (sardm) : quantifying the influence of parameter uncertainty on forecasts of extinction risk from global change. *Environmental Modelling and Software*, 83 :193–197, 2016.

[55] J. Gantz and D. Reinsel. The digital universe in 2020 : Big data, bigger digital shadows, and biggest growth in the far east. *IDC iView : IDC Analyze the future*, 2007(2012) :1–16, 2012.

[56] Hadoop. 2018 (dernière visite le 30/12/2018). `https://hadoop.apache.org/`.

[57] X. Haihua, Y. Xianchuan, H. Dan, and D. Sha. Sensitivity analysis of hierarchical hybrid fuzzy-neural network. *International Journal on Smart Sensing AND Intelligent Systems*, 8(3) :1837–1854, 2015.

[58] Y. Y. Haimes. *Risk modeling, assessment, and management*. John Wiley and Sons, 2015.

[59] R. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *IEEE Trans Syst Man Cybern :SMC 3*, pages 610–621, 1973.

[60] J. C. Helton, J. D. Johnson, and W. L. Oberkampf. An exploration of alternative approaches to the representation of uncertainty in model predictions. *Reliability Engineering and System Safety*, 85(1-3) :39–71, 2004.

[61] J. C. Helton, J. D. Johnson, C. J. Sallaberry, and C. B. Storlie. Survey of sampling-based methods for uncertainty and sensitivity analysis. *Reliability Engineering and System Safety*, 91(10-11) :1175–1209, 2006.

[62] G. B. Heuvelink, P. A. Burrough, and A. Stein. Propagation of errors in spatial modelling with gis. *International Journal of Geographical Information System*, 3(4) :303–322, 1989.

[63] F. O. Hoffman and J. S. Hammonds. Propagation of uncertainty in risk assessments : the need to distinguish between uncertainty due to lack of knowledge and uncertainty due to variability. *Risk analysis*, 14(5) :707–712, 1994.

[64] J. Hou, J. Ma, K. N. Yu, W. Li, C. Cheng, L. Bao, and W. Han. Non-thermal plasma treatment altered gene expression profiling in non-small-cell lung cancer a549 cells. *Bmc Genomics*, 16(1) :435, 2015.

[65] HPCC. 2018 (dernière visite le 30/12/2018). `https://hpccsystems.com/`.

[66] C. Huang, L. S. Davis, and J. R. G. Townshend. Unsupervised retraining of a maximum likelihood classifier for the analysis of multitemporal remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 39(2) :456–460, 2001.

[67] C. Huang, L. S. Davis, and J. R. G. Townshend. An assessment of support vector machines for land cover classification. *International Journal of Remote Sensing*, 23(4) :725–749, 2002.

[68] G. C. Hulley, C. G. Hughes, and S. J. Hook. Quantifying uncertainties in land surface temperature and emissivity retrievals from aster and modis thermal infrared data. *Journal of Geophysical Research : Atmospheres*, 117(D23) :1–18, 2012.

[69] J.-W. Hwangbo and K. Yu. Decision support system for the selection of classification methods for remote sensing imagery. *KSCE Journal of Civil Engineering*, 14(4) :589–600, 2010.

[70] C. Hyandye and L. W. Martz. A markovian and cellular automata land-use change predictive model of the usangu catchment. *International journal of remote sensing*, 38(1) :64–81, 2017.

[71] N. Imamoglu, P. Martinez-Gomez, R. Hamaguchi, K. Sakurada, and R. Nakamura. Exploring recurrent and feedback cnns for multi-spectral satellite image classification. *Procedia Computer Science*, 140 :162–169, 2018.

[72] B. Iooss and P. Lemaitre. A review on global sensitivity analysis methods. In *Uncertainty management in simulation-optimization of complex systems*, pages 101–122. Springer, 2015.

[73] L. M. Ivanov and R. Tokmakian. Sensitivity analysis of nonlinear models to parameter perturbations for small size ensembles of model outputs. *International Journal of Bifurcation and Chaos*, 21(12) :3589–3609, 2011.

[74] A. Jadidi, M. A. Mostafavi, Y. Bedard, and K. Shahriari. Spatial representation of coastal risk : A fuzzy approach to deal with uncertainty. *ISPRS International Journal of Geo-Information*, 3(3) :1077–1100, 2014.

[75] A. Kallel, S. Le Hegarat-Mascle, L. Hubert-Moy, and C. Ottle. Fusion of vegetation indices using continuous belief functions and cautious-adaptive combination rule. *IEEE Transactions on Geoscience and Remote Sensing*, 46(5) :1499–1513, 2008.

[76] R. Kang, Q. Zhang, Z. Zeng, E. Zio, and X. Li. Measuring reliability under epistemic uncertainty : Review on non-probabilistic reliability metrics. *Chinese Journal of Aeronautics*, 29(3) :571–579, 2016.

[77] H. A. Karimi. *Big Data : techniques and technologies in geoinformatics*. CRC Press, 2014.

[78] R. Kimball and M. Ross. *The data warehouse toolkit : The definitive guide to dimensional modeling*. John Wiley and Sons, 2013.

[79] A. T. Kouanou, D. Tchiotsop, R. Kengne, Z. D. Tansaa, N. M. Adele, and R. Tchinda. An optimal big data workflow for biomedical image analysis. *Informatics in Medicine Unlocked*, 11 :68–74, 2018.

[80] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105. the 25th International Conference on Neural Information Processing Systems, 2012.

[81] M. Lamboni, B. Iooss, A. L. Popelin, and F. Gamboa. Derivative based global sensitivity measures : general links with sobol indices and numerical tests. *Mathematics and Computers in Simulation*, 87 :45–54, 2013.

[82] K. Levenberg. A method for the solution of certain non-linear problems in least squares. *Quarterly of applied mathematics*, 2(2) :164–168, 1944.

[83] C. Li, W. Wang, J. Xiong, and P. Chen. Sensitivity analysis for urban drainage modeling using mutual information. *Entropy*, 16(11) :5738–5752, 2014.

[84] G. A. Licciardi and F. Del Frate. Pixel unmixing in hyperspectral data by means of neural networks. *IEEE transactions on Geoscience and remote sensing*, 49(11) :4163–4172, 2011.

[85] H. W. Lilliefors. On the kolmogorov-smirnov test for normality with mean and variance unknown. *Journal of the American statistical Association*, 62(318) :399–402, 1967.

[86] Y. Ma, L. Wang, P. Liu, and R. Ranjan. Towards building a data-intensive index for big data computing–a case study of remote sensing data processing. *Information Sciences*, 319 :171–188, 2015.

[87] 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*, pages 281–297. Oakland, CA, USA, 1967.

[88] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers. Big data : The next frontier for innovation, competition, and productivity. *Technical report, McKinsey Global Institute*, pages 1–20, 2011.

[89] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla. Classification with an edge : Improving semantic image segmentation with boundary detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 135 :158–172, 2018.

[90] L. Moller-Jensen. Classification of urban land cover based on expert systems, object models and texture. *Computers, environment and urban systems*, 21(3-4) :291–302, 1997.

[91] M. Mondal, P. Garg, N. Sharma, and M. Kappas. Cellular automata (ca) markov modeling of lulc change and sensitivity analysis to identify sensitive parameters. In *Proceedings of the 27th international cartographic conference*. 27th International Cartographic Conference, 2015.

[92] nprtool. 2018 (dernière visite le 08/01/2018). `https://www.mathworks.com/help/nnet/ref/nprtool.html`.

[93] E. Panayirci and R. C. Dubes. A test for multidimensional clustering tendency. *Pattern Recognition*, 16(4) :433–444, 1983.

[94] D. C. Parker, S. M. Manson, M. A. Janssen, M. J. Hoffmann, and P. Deadman. Multi-agent systems for the simulation of land-use and land-cover change : a review. *Annals of the association of American Geographers*, 93(2) :314–337, 2003.

[95] A. K. Paul and P. C. Shill. Incorporating gene ontology into fuzzy relational clustering of microarray gene expression data. *Biosystems*, 163 :1–10, 2018.

[96] M. A. Pena. Relationships between remotely sensed surface parameters associated with the urban heat sink formation in santiago, chile. *International Journal of Remote Sensing*, 29(15) :4385–4404, 2008.

[97] R. A. Peters. A new algorithm for image noise reduction using mathematical morphology. *IEEE transactions on Image Processing*, 4(5) :554–568, 1995.

[98] F. Pianosi, K. Beven, J. Freer, J. W. Hall, J. Rougier, D. B. Stephenson, and T. Wagener. Sensitivity analysis of environmental models : A systematic review with practical workflow. *Environmental Modelling and Software*, 79 :214–232, 2016.

[99] O. L. Puertas, C. Henriquez, and F. J. Meza. Assessing spatial dynamics of urban growth using an integrated land use model. application in santiago metropolitan area, 2010-2045. *Land use policy*, 38 :415–425, 2014.

[100] A. E. Raftery. Approximate bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika*, 83(2) :251–266, 1996.

[101] A. L. Richards, P. Holmans, M. C. O'Donovan, M. J. Owen, and L. Jones. A comparison of four clustering methods for brain expression microarray data. *BMC bioinformatics*, 9(1) :490, 2008.

[102] A. Saltelli, S. Tarantola, F. Campolongo, and M. Ratto. *Sensitivity analysis in practice : a guide to assessing scientific models*. John Wiley and Sons, 2004.

[103] O. Samuel, G. Asogbon, A. Sangaiah, P. Fang, and G. Li. An integrated decision support system based on ann and fuzzy ahp for heart failure risk prediction. *Expert Syst Appl*, 68 :163–172, 2017.

[104] M. Sanchez Canales, A. L. Benito, A. Passuello, M. Terrado, G. Ziv, M. Acuna, Vicenccnand Schuhmacher, and F. J. Elorza. Sensitivity analysis of ecosystem service valuation in a mediterranean watershed. *Science of the Total Environment*, 440 :140–153, 2012.

[105] G. Shafer. *A mathematical theory of evidence*, volume 42. Princeton university press, 1976.

[106] B. S. Soares Filho, G. C. Cerqueira, and C. L. Pennachin. Dinamica : a stochastic cellular automata model designed to simulate the landscape dynamics in an amazonian colonization frontier. *Ecological modelling*, 154(3) :217–235, 2002.

[107] Spark. 2018 (dernière visite le 30/12/2018). `https://spark.apache.org/`.

[108] Storm. 2018 (dernière visite le 30/12/2018). `http://storm.apache.org/`.

[109] K. Sutton and W. Fahmi. Cairo's urban growth and strategic master plans in the light of egypt's 1996 population census results. *Cities*, 18(3) :135–149, 2001.

[110] D. Talia. Clouds for scalable big data analytics. *Computer*, 46(5) :98–101, 2013.

[111] A. H. Tayyebi, A. Tayyebi, and N. Khanna. Assessing uncertainty dimensions in land-use change models : using swap and multiplicative error models for injecting attribute and positional errors in spatial data. *International journal of remote sensing*, 35(1) :149–170, 2014.

[112] P. H. Verburg, A. Tabeau, and E. Hatna. Assessing spatial uncertainties of land allocation using a scenario approach and sensitivity analysis : a study for land use in europe. *Journal of Environmental Management*, 127 :S132–S144, 2013.

[113] J. A. Verstegen, D. Karssenberg, F. van der Hilst, and A. P. Faaij. Detecting systemic change in a land use system by bayesian data assimilation. *Environmental Modelling and Software*, 75 :424–438, 2016.

[114] A. Wang, N. An, G. Chen, L. Liu, and G. Alterovitz. Subtype dependent biomarker identification and tumor classification from gene expression profiles. *Knowledge-Based Systems*, 146 :104–117, 2018.

[115] S. Wang, Y. Zhang, Z. Dong, S. Du, G. Ji, J. Yan, J. Yang, Q. Wang, C. Feng, and P. Phillips. Feed forward neural network optimized by hybridization of pso and abc for abnormal brain detection. *Int J Imaging Syst Technol*, 25(2) :153–164, 2015.

[116] X. Wang, S. Cen, and C. Li. Generalized neumann expansion and its application in stochastic finite element methods. *Mathematical Problems in Engineering*, 2013 :1–13, 2013.

[117] X. Xu, Z. Du, and H. Zhang. Integrating the system dynamic and cellular automata models to predict land use and land cover change. *International Journal of Applied Earth Observation and Geoinformation*, 52 :568–579, 2016.

[118] M.-S. Yang and Y. Nataliani. Robust-learning fuzzy c-means clustering algorithm with unknown number of clusters. *Pattern Recognition*, 71 :45–59, 2017.

[119] B. Youn, Z. Xi, L. Wells, and P. Wang. Enhanced dimension-reduction (edr) method for sensitivity-free uncertainty quantification. In *11th AIAA/ISSMO multidisciplinary analysis and optimization conference*, pages 69–77. Proceedings of 11th AIAA/ISSMO multidisciplinary analysis and optimization conference, 2006.

[120] L. A. Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy sets and systems*, 100(1) :9–34, 1999.

[121] A. Z. Zhang, G. Y. Sun, S. H. Liu, Z. J. Wang, P. Wang, and J. S. Ma. Multi-scale segmentation of very high resolution remote sensing image based on gravitational field and optimized region merging. *Multimedia Tools and Applications*, 76(13) :15105–15122, 2017.

[122] D. Zhang, W. Li, X. Wu, and T. Liu. An efficient regional sensitivity analysis method based on failure probability with hybrid uncertainty. *Energies*, 11(7) :1–19, 2018.

[123] J. Zhang, T. Li, X. Lu, and Z. Cheng. Semantic classification of high-resolution remote-sensing images based on mid-level features. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9(6) :2343–2353, 2016.

[124] C. Zheng, Y. Zhang, and L. Wang. Semantic segmentation of remote sensing imagery using an object-based markov random field model with auxiliary label fields. *IEEE Transactions on geoscience and remote sensing*, 55(5) :3015–3028, 2017.

[125] G. Zhi-qiang and O. Dennis. The temporal and spatial relationship between ndvi and climatological parameters in colorado. *Journal of Geographical Sciences*, 11(4) :411–419, 2001.