



AfIA

Association française
pour l'Intelligence Artificielle

APIA

*Conférence Nationale
sur les
Applications Pratiques de l'Intelligence Artificielle*

PFIA 2024



Table des matières

Ghislain Atemezing, Catherine Roussey Éditorial	5
Comité de programme	6
Session 1 : Articles longs originaux	7
M.-P. Doan, I. Joly SEM pour l'estimation des coûts de fabrication : une étude de cas sur les moules à injection ...	8
Session 1.1 : Apprentissage sur les images	18
A. Chemchem, L. Mohimont, F. Alin, L. A. Steffemel Estimation du Rendement du Mil Perlé (<i>Pennisetum glaucum</i>) par Machine Learning à l'aide d'Images Satellites	19
T. Goudemant, C. Szywala, B. Francesconi, M. Aubrun, E. Kervennic, I. Grenet, Y. Bobichon, M. Bellizzi, A. Girard Détection d'anomalies à bord de satellites pour la protection de l'environnement marin	27
Session 1.2 : Traitement automatique des langues	37
C. Bouvard, M. Ciancone, A. Gourru, M. Schaeffer Derby LLM : Évaluation comparative des approches RAG et fine-tuning	38
M. Chastan, R. Zaatour, C. Ouedraogo, K. Baati Mise en place d'un système de notation par aspects pour les articles de E-commerce en se basant sur les avis en français des consommateurs	48
A. Jourdan, Y. Le Nir, N. Girardin Détection automatique de similarités dans le cadre de la protection par Marque	57
N. Vautier, M. Héry, M. Miled, I. Truche, F. Bullier, A.-L. Guénet, G. Dubuisson Duplessis, S. Campano, P. Suignard Utilisation de LLMs pour la classification d'avis client et comparaison avec une approche classique basée sur CamemBERT	65
Session 1.3 : Graphes et trajectoires	74
A. Coppé, N. Prcovic Un algorithme de routage de navires générant des trajets précis et diversifiés	75
Y. Hamadi, G. Picard Résolution multicritère socialement acceptable du problème de réparation des contrats 4D dans le cadre de la gestion du trafic aérien sans pilote	83
E. Peyre, F. Amarger, N. Chauvat CapData Opéra : faciliter l'interopérabilité des données des maisons d'opéra	93
Session 2 : Articles déjà publiés à l'international	103
O. Labbani Narsis, E. Dujardin, C. Nicolle Intelligence Artificielle Modulaire et Hybride Orientée par les Objectifs - Application à la Simulation d'un Calculateur Booléen Plasmonique	104
N. Griselin, D. Dubois, P. Barbier Détection de défauts sur radiographies de pièces composites rotor grâce à l'IA	106
Session 3 : Articles prospectifs	108

J.-L. Farges, F. Perotto, G. Picard, C. Pralet, C. de Lussy, J. Guerra, P. Pavero, F. Planchou. Dépasser l'observation mono-mission de la Terre : utiliser le paradigme multi-agents pour fédérer de multiples missions	109
P. Feillet Pour une hybridation des Grands Modèles de Langue et de l'inférence de règles logiques	115
S. Sonfack Souchio, B. Kamsu-Foguem, L. Geneste Une représentation graphique multicouche des processus d'expertise	121
Session 4 : Posters et démonstrations	128
A. Abbadie, A. Lammoglia, M. Servalean Usages de l'IA dans l'enseignement supérieur : Enjeux et perspectives	129
F. Dama, R. Sleiman, S. Bellart Introduire l'IA dans la lutte contre la fraude : Comment choisir et convaincre ?	133

Éditorial

Conférence Nationale sur les Applications Pratiques de l'Intelligence Artificielle

La conférence nationale sur les Applications Pratiques de l'Intelligence Artificielle (APIA), soutenue par le Collège Industriel de l'AFIA ¹, est un lieu d'échanges entre les chercheurs académiques et les entreprises (industriels, entreprises de services, startups, etc). Ainsi, les chercheurs académiques et industriels partagent leurs expériences dans le domaine de l'Intelligence Artificielle (IA) autour de cas concrets, afin de débattre des différents verrous applicatifs qu'ils rencontrent. La communauté des chercheurs en IA se rencontre à APIA pour démocratiser l'utilisation de l'IA, pour présenter les méthodes mises en œuvre afin d'enrichir le potentiel applicatif des modèles et outils de l'IA, et pour partager les besoins naissants des entreprises.

Pour cette 10ème édition, le comité de programme a reçu des contributions pour illustrer des applications concrètes de l'IA sur les thèmes suivants :

- Ingénierie, partage et gestion des connaissances,
- Agents autonomes et systèmes multi-agents : simulation, planification, décision individuelle ou collective,
- Web sémantique, graphes de connaissances, graphes de propriétés, données ouvertes et liées,
- Traitement automatique du langage, terminologie, langage naturel contrôlé, recherche d'information, classification de texte, grand modèle de langage,
- Traitement du signal et de l'image, télédétection,
- Apprentissage, apprentissage par réseaux de neurones, apprentissage supervisé,
- hybridation des méthodes d'IA, explicabilité des méthodes d'apprentissage, apprentissage sur des modèles sémantiques, ...

L'ensemble des techniques d'IA s'applique à des domaines d'applications très variées comme l'analyse de trajectoires, la prédiction de coûts industriels, la détection de défauts ou d'anomalies, la prédiction de rendement, le suivi de produits, etc...

Un total de 25 soumissions d'articles ont été collectées sur la plate-forme de gestion de relecture EasyChair. Le comité de programme composé pour moitié d'industriels et d'académiques a retenu 17 articles (soit un taux d'acceptation de 68%) répartis comme suit :

- 10 articles en papiers longs,
- 2 articles en papiers déjà publiés à l'international,
- 3 articles prospectifs et 2 posters démonstrations.

APIA 2024 inclut deux conférences sélectionnées par le Collège Industriel de l'AFIA :

- *"Commonsense knowledge for trusted flexible Manufacturing"* par le Professeur Hedi KARRAY de l'Université de Technologie Tarbes Occitanie Pyrénées UTTOP,
- *"L'évolution de l'IA dans l'industrie depuis 10 ans"* par le professeur Jean-Gabriel GANASCIA de Sorbonne Université.

Plusieurs partenaires de PFIA 2024 présentent leurs avancées en IA lors de APIA 2024 :

- *"Le rôle et l'apport d'un tech lab innovation au sein d'une DSI"* par Adrian Ciocan et Dominique Meunier de la MAIF,
- *"Développement d'un Outil de Détection de Défauts des Éoliennes - Une Approche Basée sur les Réseaux Bayésiens"* par Ahmed Mabrouk d'Engie,
- *"TA KM ou l'ingénierie de la connaissance opérationnelle"* par Alain Berger d'Ardan et Sébastien Boblet de Technicatome.

Nous tenons à remercier ici tous ceux qui ont participé de près ou de loin au succès d'APIA 2024, le comité d'organisation de PFIA 2024 et l'ensemble de ses partenaires, les membres du comité de programme, le Collège Industriel de l'AFIA, les auteurs des articles, des posters et des démonstrations, Hedi Karray, Jean-Gabriel Ganascia et enfin tous les participants à la plate-forme.

Ghislain Atemezing, Catherine Roussey

1. <https://afia.asso.fr/les-colleges/le-college-industriel/>

Comité de programme

Présidence

- Ghislain Ateazing (ERA, Valenciennes);
- Catherine Roussey (MISTEA INRAE, Montpellier).

Membres

- Florence Amardeilh (Elzeard, Bordeaux);
- Fabien Amarger (Logilab, Toulouse);
- Nicolas Audebert (CNAM CEDRIC, Paris);
- Nathalie Aussenac-Gilles (IRIT, Toulouse);
- Alain Berger (Ardans, Montigny-le-Bretonneux);
- Sandra Bringay (LIRMM Université Paul Valéry Montpellier, Montpellier);
- Xavier Briottet (ONERA, Toulouse);
- Stéphan Brunessaux (Sensei Consult, Louviers);
- Patrice Buche (IATE INRAE, Montpellier);
- Davide Buscaldi (LIPN Université Sorbonne Paris Nord, Villetaneuse);
- Bruno Carron (Airbus Defence and Space, Grand Paris);
- Laurent Cervoni (Talan, Grand Paris);
- Caroline Chopinaud (Hub France IA, Paris);
- Gaël de Chalendar (CEA LIST, Saclay);
- Yves Demazeau (LIG CNRS, Grenoble);
- Sylvie Despres (LIMICS Université Sorbonne Paris Nord, Bobigny);
- Gayo Diallo (AHeaD ISPED Université de Bordeaux, Bordeaux);
- Valentina Dragos (Onera, Palaiseau);
- Guillaume Dubuisson Duplessis (EDF, Paris);
- Catherine Faron (I3S Université Côte d'Azur, Sophia Antipolis);
- Bernard Georges (Société Générale, Paris);
- Céline Hudelot (CentraleSupélec MICS, Gif-sur-Yvette);
- Dino Ienco (TETIS INRAE, Montpellier);
- Arnaud Lallouet (Huawei Technologies Ltd, Boulogne-Billancourt);
- Christine Largouët (IRISA, Rennes);
- Christelle Launois (Société Générale, Paris);
- Mustapha Lebbah (DAVID Université Paris-Saclay, Versailles);
- Dominique Lenne (HEUDIASYC Université de Technologie de Compiègne, Compiègne);
- Sylvain Mahé (EDF Recherche et Développement, Chatou);
- Céline Rouveirol (LIPN, Université Sorbonne Paris Nord);
- Françoise Soulié-Fogelman (Hub France IA, Paris);
- Élodie Thiéblin (Logilab, Toulouse);
- Brigitte Trousse (INRIA, Sophia Antipolis).

Session 1 : Articles longs originaux

SEM pour l'estimation des coûts de fabrication : une étude de cas sur les moules à injection

Minh-Phuoc. DOAN¹, Iragaël. JOLY²

¹ Université de Bordeaux, CNRS, IMS, UMR 5218, 33405 Talence, France

² Université Grenoble-Alpes, CNRS, INRAE, Grenoble INP, GAEL, Grenoble, France

minh-phuoc.doan@u-bordeaux.fr ; iragael.joly@grenoble-inp.fr

Résumé

L'estimation du coût est un défi pour le chiffrage des produits. Le coût final d'un produit est constitué de différents éléments associés à la conception, à la fabrication, à la distribution et aux processus de gestion de la fin de vie. Les méthodes d'estimation des coûts basées sur des données, qui prédisent le coût d'un produit sur la base des coûts d'autres produits similaires, connaissent un potentiel croissant en raison de leur niveau de précision de prédiction. Toutefois, pour certaines de ces méthodes, telles que le réseau de neurone, la précision dépend largement de la quantité et de la qualité des données disponibles. De plus, dans l'application de ces méthodes, la relation entre le coût et la caractéristique reste inexplicable. Cet article propose un modèle explicatif d'estimation des coûts, basé sur l'approche des modèles à équations structurelles (SEM). Des applications numériques, réalisées sur une étude de cas d'une entreprise de moule à injection, montrent le potentiel de cette méthode en termes de précision de prédiction et d'explicabilité.

Mots-clés

Modèles à équations structurelles ; Devis de produit ; Estimation des coûts ; Méthode fondée sur les données

1 Introduction

L'estimation des coûts de production, en particulier à un stade précoce du processus de développement du produit, offre à l'entreprise les avantages suivants [1] : i) être compétitive en proposant un prix de produit attractif à ses clients, ii) identifier les principaux facteurs de coût pour améliorer les processus et réduire les coûts à long terme, et iii) optimiser la conception du produit en intégrant la méthode d'estimation des coûts dans le système de conception. Dans un second temps, lors de la phase de production, une meilleure connaissance des coûts de production est une donnée essentielle du suivi de l'activité, de la stratégie de tarification et plus largement de l'activité de commercialisation du produit.

Les méthodes d'estimation des coûts, qui sont basées sur l'intelligence artificielle (IA), connaissent un potentiel croissant [1].

La Figure 1 présente un cadre conceptuel d'intelligence

Abstract

Estimating the cost is a challenge for product quotation. The product final cost is made up of different cost components associated with the design, the fabrication, the distribution and end-of-life management processes. The data-driven cost estimation methods, which predicts a product cost based on the costs of other similar products see an increasing potential in recent time because of its prediction accuracy level. However, for some of these methods, such as neural network, the accuracy depends wisely on the quantity and quality of available data and the cost-feature relationship is inexplicable. This article proposes an explicable cost estimation framework, which is based on the structural equation modelling (SEM) approach. Numerical experiments, performed on a case study of an injection mould company, show the potential of this method in in terms of the prediction accuracy and the explainability.

Keywords

Structural equation modelling; Product quotation; Cost estimation; Data-driven method

artificielle pour l'estimation du coût des produits en fabrication [2]. Ce cadre se compose de deux modules : le développement de produits, qui se concentre sur les phases de conception du produit et des processus le soutenant ; et l'analyse des coûts et des caractéristiques du produit, qui cherche à quantifier par l'analyse de données historiques de l'entreprise, les relations entre le coût du produit et ses caractéristiques et celles des processus. Ce dernier module peut être divisé en trois blocs : la transformation et la sélection des caractéristiques, le regroupement ou la classification des familles de produits, et la cartographie des coûts et des caractéristiques.

Au tout début du développement d'un produit, seules les spécifications du produit sont généralement disponibles. Elles fournissent certaines caractéristiques géométriques et fonctionnelles pour l'estimation des coûts [3]. Dans un second temps, ces spécifications permettent au concepteur de construire les modèles CAO du produit. Ces modèles précisent les caractéristiques géométriques et améliorent l'estimation des coûts [4], [5].

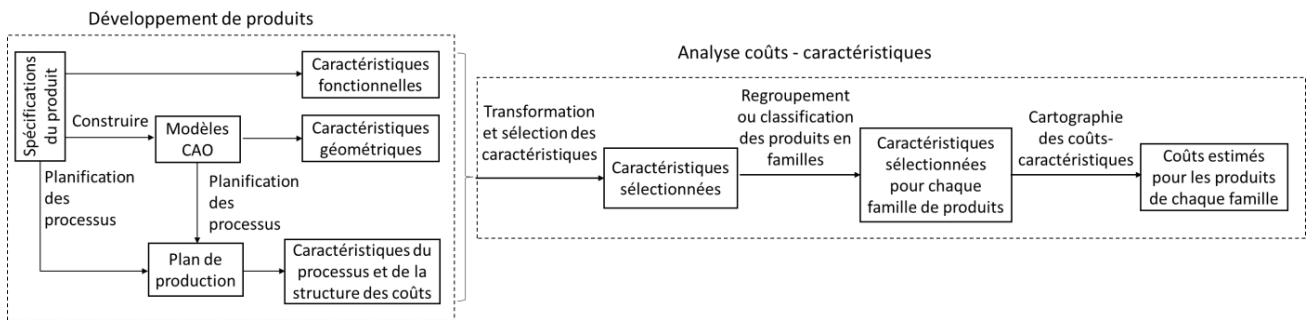


Figure 1. Représentation d'un système d'intelligence artificielle pour l'estimation du coût des produits dans l'industrie manufacturière

Les spécifications du produit et les modèles CAO sont ensuite utilisés pour créer le plan de production, en fournissant les caractéristiques du processus (par exemple, les technologies d'usinage) et les caractéristiques de la structure des coûts (par exemple, le coût des matériaux, le coût de la main-d'œuvre), ce qui est souvent appelé l'estimation des coûts à un stade ultérieur [5], [6].

Trois niveaux de décomposition du plan de production sont souvent considérés dans la littérature : 1) dans la décomposition simplifiée, le coût total est considéré comme la somme de groupes de coûts. Pour un produit assemblé, le plus souvent, le coût total du produit est considéré comme la somme des coûts de tous les composants et du coût d'assemblage (matériaux, main-d'œuvre, etc.) [7] ; 2) dans la décomposition traditionnelle, un assemblage est composé d'un ou de plusieurs éléments et chaque élément peut avoir des caractéristiques différentes, fabriquées par une ou plusieurs opérations d'usinage. Le coût estimé est donc la somme des coûts de toutes ces opérations de fabrication [7] ; 3) La décomposition détaillée (ou décomposition par activité) représente l'utilisation des ressources de l'organisation par les activités réalisées et associe ensuite le coût de ces activités aux résultats, tels que les produits et les services réalisés [8]. Les inducteurs de coûts sont identifiés pour chaque activité afin de calculer les coûts.

Dans la phase d'analyse coûts-caractéristiques, parmi les caractéristiques identifiées précédemment, certaines peuvent être transformées pour une meilleure comparaison des caractéristiques [9]. Ensuite, des techniques de sélection des caractéristiques, telles que la sélection par les experts [10] ou la sélection guidée par les données [11], sont appliquées pour identifier et sélectionner les caractéristiques les plus importantes ou ayant le plus d'impact pour l'étape suivante de l'estimation des coûts. Une méthode de regroupement ou de classification est parfois utilisée pour diviser les produits en familles présentant des caractéristiques similaires. Enfin, une méthode de mise en correspondance des coûts et des caractéristiques est ensuite utilisée pour estimer le coût du produit pour chaque famille. Le regroupement/la classification n'est pas toujours obligatoire pour l'estimation des coûts, mais il permet d'élaborer une méthode plus précise de mise en correspondance des coûts et des caractéristiques pour chaque famille de produits lorsqu'il existe différentes familles de

produits.

Cette analyse des relation coûts - caractéristiques peut être menée par l'appui des connaissances des experts du domaine ou par l'analyse des données historiques de produits similaires.

Les méthodes paramétriques et les méthodes analogiques sont deux familles courantes de méthodes basées sur les données. La méthode paramétrique est basée sur des formules mathématiques référencées, établies sur un grand nombre d'expériences statistiquement assez importantes pour pouvoir représenter la population de produits. Ces formules peuvent être construites à partir de l'exploration des données et/ou trouvées dans un manuel d'estimation des coûts, un manuel de référence ou une référence bibliographique [8], [10], [12]. La méthode analogique est basée sur l'analogie des données historiques de l'entreprise concernant des produits similaires. D'un point de vue statistique, il s'agit d'un échantillon d'une grande population de produits. Les méthodes courantes sont la régression linéaire [3], les forêts aléatoires [13], les k-plus proches voisins [12], [14] et les réseaux de neurones [9].

L'analyse des caractéristiques déterminantes du coût d'un produit et la prédiction de ce coût fondée sur les données repose sur une vision de la composition du coût de production. Plus particulièrement, les méthodes paramétriques formalisent les relations entre les caractéristiques des produits et des processus et les coûts de production. Ce formalisme est souvent hérité de la méthode d'estimation elle-même, comme par exemple la forme linéaire supposée entre caractéristiques et coût dans une régression linéaire, si ce n'est masqué ou trop complexe comme par exemple dans les réseaux de neurones multicouches.

Inspiré des trois niveaux de décomposition du plan de production présentés plus haut, cet article propose une discussion et une application d'une décomposition du coût d'un produit dans une approche fondée sur les données. Une attention particulière est donnée aux conséquences de la décomposition choisie et aux interrelations possibles entre les composants.

Le cas d'application est le chiffrage de moules pour injection plastique. Les données mises à disposition par une entreprise concernent les moules produits. Ils ont donc été chiffrés et les coûts ont ensuite été évalués lors de la production.

La méthode appliquée ici est la méthode SEM (ou modèle à équations structurelles), une méthode analogique axée sur les données, qui utilise un système de régressions linéaires multivariées pour modéliser les relations entre les composants du coût et les caractéristiques du produit [15]. Elle permet de tenir compte d'une décomposition a priori du coût du produit indiquée par les experts. De plus, la méthode SEM apporte une explication de la relation entre les composants du coût et les caractéristiques et peut être utilisée pour prédire simultanément les éléments de coût constituant le coût total, et pas seulement ce dernier.

La méthode SEM est couramment utilisée en biologie [16], en économétrie [17] et pour l'estimation des coûts des projets [18], mais à notre connaissance, cette méthode n'est pas encore utilisée comme méthode d'estimation des coûts des produits dans l'industrie.

La principale contribution de cet article est de proposer et d'évaluer la performance de plusieurs modèles d'estimation des coûts basés sur des données. Les modèles proposés correspondent à la régression linéaire et à plusieurs combinaisons de la méthode SEM avec plusieurs méthodes courantes de sélection des caractéristiques, y compris la sélection par un expert et par apprentissage automatique. L'article est organisé comme suit. La section 2 présente la revue de la littérature sur les méthodes d'estimation des coûts basées sur les données. La section 3 présente la méthode SEM et quelques méthodes courantes de sélection des caractéristiques. La section 4 présente les données et l'application des méthodes proposées dans une étude de cas d'une entreprise de moules à injection. La section 5 présente quelques conclusions et perspectives pour la poursuite de la recherche.

2 Revue de littérature

La sélection des méthodes d'analyse des coûts et des caractéristiques dépend de différents facteurs, tels que le stade de développement du produit, la disponibilité des données et de la référence paramétrique, la précision de la prédiction et le besoin d'explicitation de la relation entre les coûts et les caractéristiques. Pour l'estimation des coûts à un stade précoce, il est rare qu'il existe une référence en matière d'estimation des coûts performante en termes de précision de prédiction. En l'absence de données historiques, la méthode intuitive est généralement le seul choix possible. Si des données historiques sont disponibles, la méthode analogique est la plus utilisée en raison de sa précision.

Dans son manuel, [19] propose une méthode paramétrique pour prédire le coût du moule d'injection plastique et de la pièce fabriquée par le moule. Le coût total du moule est la somme de trois éléments de coût : le coût de la base du moule, le coût de toutes les cavités et le coût de leur personnalisation. Chacun de ces trois éléments de coût est évalué à l'aide d'une formule mathématique fondée sur l'expertise. [8] utilise la décomposition des coûts par activité pour estimer le coût du produit. Le coût total est divisé en centres de coûts et chaque centre de coûts est représenté par des inducteurs de coûts. Une

formulation mathématique est établie entre chaque centre de coûts et les inducteurs de coûts associés.

Les méthodes analogiques ont aussi été utilisées dans la littérature, mais à notre connaissance rarement pour le chiffrage de moules à injection. [20] estime le coût des moules industriels en combinant la méthode de sélection de l'analyse factorielle avec trois méthodes d'analyse de la relation coût - caractéristiques : les forêts aléatoires, les k plus proches voisins, et les réseaux de neurones. Les modèles sont assez performants avec une incertitude satisfaisante dans la prédiction comprise entre 10% et 20% selon le type de moule. [21] estime les coûts des produits dans l'industrie automobile en combinant la méthode de sélection par étapes avec six méthodes analogiques différentes de mise en correspondance des coûts et des caractéristiques : arbre de décision, ensembliste, réseau de neurones artificiels, régression SVM, régression linéaire et k plus proches voisins. Ils ont constaté que tous les modèles donnent une précision élevée de l'estimation des coûts avec le meilleur R^2 de 0,96.

Dans de nombreuses situations, l'explicabilité est, avec la précision de la prédiction, un autre facteur important pour le choix d'un modèle d'estimation des coûts. Pour améliorer l'explicabilité du modèle d'estimation des coûts, en particulier des modèles analogiques, les méthodes suivantes sont couramment utilisées dans la littérature : visualiser la relation entre le coût total et les caractéristiques associées [1], décomposer le coût total en groupes de coûts et établir la relation entre chacun de ces groupes et les caractéristiques du produit/processus [22], [23], ou décomposer le produit assemblé en composants, puis ces composants en caractéristiques/fonctions et estimer le coût de ces dernières [7], [24].

La sélection des caractéristiques relève de méthodes classées en deux catégories : la sélection par des experts et les méthodes d'apprentissage automatique (ou guidées par les données). Les méthodes de sélection par des experts utilisent les connaissances des experts pour sélectionner les caractéristiques [25]. [25] estiment le coût des pièces de moulage par injection de plastique. En interrogeant les experts du domaine, ils sont en mesure de redessiner le chemin que les experts ont suivi pour estimer le coût. En suivant ce chemin, ils peuvent identifier des facteurs de coût importants tels que le poids net du produit fini, la taille de la pièce moulée, etc.

La sélection par apprentissage automatique utilise l'analyse statistique des données pour sélectionner les variables. Une méthode courante est l'analyse de corrélation de Pearson, [3], l'analyse factorielle [26], l'analyse stepwise [11].

L'analyse de corrélation de Pearson est une méthode qui permet d'identifier les caractéristiques les plus étroitement corrélées avec la variable de sortie. [3] combine la sélection de la corrélation de Pearson et le réseau de neurones pour estimer le coût de fabrication des éléments de tuyauterie. Les corrélations de Pearson entre le coût et six caractéristiques, dont le poids, le type de soudage, le diamètre, la difficulté, les cavités et la classe, sont calculées. Seules quatre de ces caractéristiques dont le coefficient est supérieur à un seuil sont

sélectionnées.

L'analyse factorielle est une méthode qui permet de réduire un grand nombre de caractéristiques en un nombre plus restreint de facteurs importants [26]. [26] combine l'analyse factorielle avec l'optimisation par essaims de particules (PSO) et le réseau de neurones artificiels pour donner une estimation du coût du moulage par injection plastique.

La méthode de régression stepwise est utilisée pour sélectionner et ajouter progressivement des caractéristiques au modèle. Chaque fois qu'une caractéristique est ajoutée au modèle, un test statistique (généralement le F-test associé à une mesure de qualité d'ajustement) est utilisé pour valider le modèle. Si une caractéristique initiale n'est plus significative après l'introduction de nouvelles caractéristiques, elle est supprimée du modèle. Ce processus est répété jusqu'à ce qu'aucune caractéristique significative ne puisse être introduite dans le modèle de régression [11]. Pour la méthode stepwise, différentes mesures de qualité d'ajustement peuvent être utilisées, telles que la p-valeur, l'AIC et le BIC [15].

Une des limitations de l'application des modèles de régression linéaire à l'estimation des coûts réside dans la prise en compte de la décomposition du coût total. Les méthodes SEM apparaissent avantageuses dans cette perspective. [27] utilisent les SEM pour estimer les coûts d'une mission d'exploitation minière d'un astéroïde. Les données ont été collectées par le biais d'entretiens avec des ingénieurs spatiaux et des scientifiques internationaux de différents niveaux et domaines professionnels. [28] utilise les SEM, la régression multiple et les réseaux de neurones pour estimer les coûts d'un projet de construction de tunnels routiers. Les résultats montrent que le modèle SEM donne une meilleure précision de prédiction que celle de la régression multiple et aussi précise que celle du réseau de neurones. En outre, ce modèle permet à l'utilisateur de représenter visuellement les voies par lesquelles plusieurs variables affectent le coût d'un projet.

L'originalité de notre approche est de proposer une estimation fondée sur des données du coût de moules à injection plastique. Pour cela plusieurs modèles d'estimation des coûts par apprentissage automatique sont évalués : la régression linéaire multiple et méthode SEM. Les caractéristiques introduites dans les modèles sont sélectionnées par application de plusieurs méthodes de sélection concurrentes : la sélection par un expert et deux modèles de sélection par apprentissage automatique (sélections stepwise sur l'AIC et le BIC).

3 Méthodes

L'estimation du coût pour le chiffrage des produits soulève plusieurs défis. L'estimation du coût final d'un produit se compose d'un certain nombre de sous-coûts qui sont interdépendants. Les coûts de conception et de R&D varient en fonction de la complexité du produit, ce qui a une incidence sur les techniques et les coûts de fabrication. Par conséquent,

la prédiction du coût total doit tenir compte des interdépendances entre ses composants.

La régression linéaire est un outil classique et simple d'apprentissage automatique utilisé dans les méthodes d'estimation des coûts basées sur les données. Cette régression suppose que les termes aléatoires du modèle sont aléatoires et indépendants. Néanmoins, cette hypothèse d'indépendance peut conduire à des estimations et des prédictions peu fiables [29] et ne tient pas compte des interdépendances entre composant du coût total. La spécification habituelle de la régression linéaire multiple est une simple équation, $y = \alpha + \beta_1x_1 + \beta_2x_2 + \dots + \epsilon$, où y est le coût du produit à prédire en fonction des caractéristiques x_1, x_2, \dots et les paramètres à estimer, $\alpha, \beta_1, \beta_2, \dots$. Les erreurs de prédiction sont modélisées par $\epsilon \sim N(0, \sigma^2)$ qui sont supposées être indépendantes entre les produits.

Les techniques de régression multivariée, comme les modèles d'équations structurelles (SEM), permettent d'estimer le coût de fabrication et chaque composant ($j=1\dots J$) à l'aide d'une équation spécifique incluant une éventuelle covariance entre les termes aléatoires entre les équations ($cov(\epsilon_j, \epsilon_{j'}) \neq 0$) (voir Figure 2).

$$\begin{aligned} y_1 &= \alpha_1 + \beta_{11}x_1 + \beta_{12}x_2 + \dots + \epsilon_1 \\ y_2 &= \alpha_2 + \beta_{21}x_1 + \beta_{22}x_2 + \dots + \epsilon_2 \\ &\dots \\ y_J &= \alpha_J + \beta_{J1}x_1 + \beta_{J2}x_2 + \dots + \epsilon_J \end{aligned}$$

Un deuxième défi consiste à identifier et à sélectionner les caractéristiques pertinentes à inclure dans le modèle de régression. Les connaissances des experts donnent un aperçu de l'ensemble des variables à prendre en compte. Les méthodes itératives automatiques stepwise sont aussi appliquées ici pour sélectionner les caractéristiques en fonction de leur impact évalué sur un indicateur de performance (AIC ou BIC).

Dans cet article, l'analyse des données a été réalisée à l'aide du package Lavaan [30] de la version 4.3.1 de R (R Core Team, 2023). Nous comparons la régression linéaire multiple à la régression SEM avec différentes spécifications : en incluant ou non des interdépendances entre les équations de coût et avec différentes méthodes de sélection des caractéristiques. Cette comparaison est basée sur des mesures de performance classiques pour la prédiction quantitative. Les dépendances ou interdépendances des erreurs entre les équations de coût sont évaluées par un test de significativité de la matrice de variance et de covariance estimée des résidus.

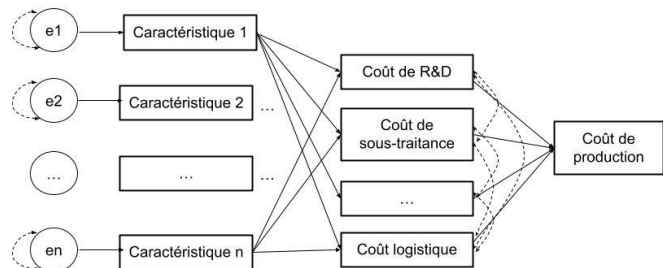


Figure 2. Structure générale du SEM du coût estimé

Dans le cadre des SEM, les indicateurs de performance habituels sont basés sur :

- Les R^2 exprimés en termes de résidus quadratiques du modèle.
- L'indice comparatif d'adéquation (CFI) et l'indice d'adéquation de Tucker-Lewis (TLI), qui sont des indicateurs évaluant l'adéquation entre les matrices de covariance attendues et observées des données ajustées en fonction de la taille de l'échantillon. Le CFI est compris entre 0 et 1. Une valeur de CFI plus élevée indique une meilleure adéquation du modèle. Dans la pratique, un CFI proche de 0,95 ou supérieur indique une adéquation acceptable. Le TLI peut être supérieur à 1. Un TLI supérieur à 0,90 est acceptable [15].
- L'AIC et le BIC sont basés sur le logarithme de la vraisemblance du modèle et ne sont ni bornés, ni normés. Entre plusieurs modèles estimés sur un même jeu de données, les valeurs les plus faibles caractérisent les meilleurs modèles.

4 Expérimentation et résultats

4.1 Données

Nous présentons une étude de cas d'une entreprise de moules à injection, qui conçoit et fabrique des moules à injection pour des clients du monde entier. Actuellement, le devis des moules est réalisé manuellement par des experts en moules qui suivent une procédure intégrée dans un fichier Excel. L'ensemble des données de l'entreprise utilisées pour les expériences consiste en 131 devis de moules acceptés par les clients entre 2016 et 2022. Un traitement préliminaire des données avec l'aide des experts en devis, comprenant le nettoyage, la transformation et la sélection des données, permet d'obtenir un ensemble de données approprié de 97 devis de moules pour les expériences. Dans ce nouvel ensemble de données, pour chaque moule, 32 caractéristiques d'entrée (variables explicatives) et 24 caractéristiques de sortie (variables expliquées) sont disponibles.

Les caractéristiques d'entrée peuvent être regroupées en plusieurs catégories : caractéristiques de la catégorie de moule (par exemple, type, catégorie et date de sortie), caractéristiques du moule (types d'injection, d'éjection et de régulation, nombre de cavités), caractéristiques du mécanisme du moule (nombre et types de cylindres mécaniques, nombre et types de cylindres hydrauliques), caractéristiques du produit (par exemple, type de produit, qualité de la surface, matériau) et autres caractéristiques (par exemple, méthode d'essai, garantie et livraison).

Les caractéristiques de sortie comprennent les coûts estimés par l'expert et les coûts réels mis en œuvre pour chacune des onze activités induites par les coûts (par exemple, l'achat, la conception, l'usinage brut et la finition). Dans cette étude, nous nous concentrons uniquement sur les coûts réels des moules et, avec l'aide des experts de l'entreprise, nous combinons les onze activités de coûts en six groupes (voir Tableau 1).

Tableau 1. Explication des activités induites par les coûts

Activité induite par les coûts	Explication
Achats	Coûts liés à l'achat de matériel et d'équipements
Etude	Coûts liés au processus de conception du produit
Sous-Traitance	Coûts liés aux opérations externalisées
Méthode	Coûts liés à l'industrialisation du produit (processus de fabrication et outils nécessaires)
Ebauche	Coûts associés aux opérations d'usinage d'ébauche et de finition
Ajustage	Coûts liés aux tests et ajustements

Définitions des variables explicatives (caractéristiques) :

Nbre_mvt_meca_simple (Variable numérique) : nombre de mouvements "simples" effectués par le moule.

Nbre_mvt_meca_double (Variable numérique) : nombre de mouvements "doubles" (plus complexes).

Nbre_mvt_rotatif (Variable numérique) : nombre de mouvements "rotatifs".

Nbre_total_point (Variable numérique) : nombre total de points d'injection pour le moule.

Type.de.produit_HT (valeur de référence) : l'entreprise réalise le produit pour la mise en production.

Type.de.produit_MIX : L'entreprise réalise en partie la mise en place de la production.

Type.de.produit_PROTO : L'entreprise réalise le prototype uniquement.

Type.injection_CAROTTE (Variable binaire) : le système utilise une carotte qui permet de traverser le bloc empreinte fixe jusqu'au plan de joint. Sa conicité favorise le démoulage. Sa valeur de référence est *Type.injection_Bloc_chaud* qui donne l'information que le système utilise un bloc chaud qui permet de maintenir la matière à température.

Ejection_Autre : la forme du vérin d'éjection est tubulaire ou en lame. Sa valeur de référence est *Ejection_Cylindrique* décrivant que la forme du mécanisme d'éjection est cylindrique.

Commande.ejection_Verin : Le mécanisme d'éjection se fait par vérin. Sa valeur de référence est *Commande.ejection_Asservissement_presse* qui décrit que le mécanisme est déclenché par Asservissement presse.

Regulation_Normal : La régulation de la température pour refroidir le système se fait de manière "normale" donc à plus basse température que la valeur de référence *Regulation_Chaud*.

Acier.moulant_H11 (Valeur de référence) : acier avec une ténacité légèrement supérieure aux aciers classiques.

Acier.moulant_110KG : acier plus souple.

Acier.moulant_Autre : acier non-conventionnel (pas 100KG, H11 ou H13).

Acier.moulant_H13 : l'acier le plus commun pour les outillages.

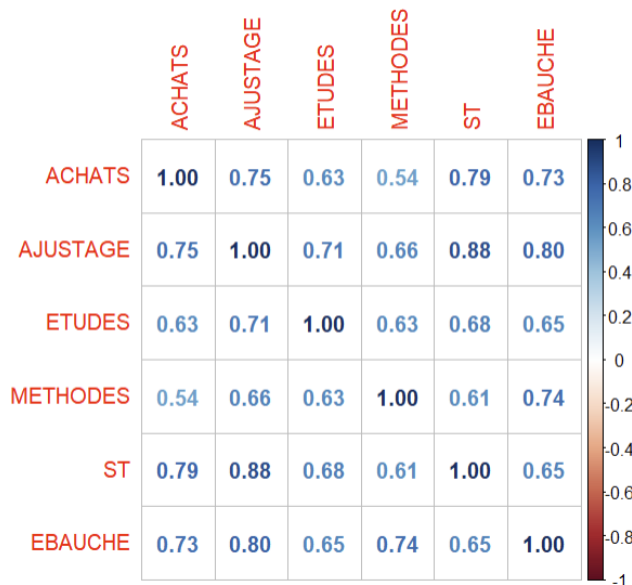


Figure 3. Matrice de corrélation entre les groupes de coûts

La Figure 3 montre que les 6 éléments de coût covarient principalement dans un schéma de corrélation positive, ce qui suggère que leurs prédictions peuvent ne pas être totalement indépendantes les unes des autres.

4.2 Résultats

La première étape consiste à sélectionner les caractéristiques pertinentes pour l'estimation et la prédiction des coûts. La sélection est basée sur l'expertise des ingénieurs de l'entreprise (modèle "M.Expert") et sur deux méthodes guidées par les données, une sélection stepwise basée sur les critères AIC et BIC (modèles "M.AIC" et "M.BIC").

Le modèle "Expert" comprend trois à cinq caractéristiques en fonction de l'élément de coût. Le nombre de mouvements mécaniques simples, de mouvements rotatifs et la taille sont des caractéristiques communes pour les 6 composantes de coûts.

Les modèles M.AIC et M.BIC comprennent jusqu'à sept caractéristiques, parmi lesquelles les caractéristiques sélectionnées par l'expert sont communes aux 6 composantes de coût.

La deuxième étape consiste à évaluer les deux spécifications avec et sans résidus interdépendants. Chaque modèle est estimé comme si les équations du système étaient indépendantes. Cela conduit à des modèles M.Expert.0, M.AIC.0, M.BIC.0 qui sont estimés avec la contrainte de résidus indépendants entre les équations.

Le Tableau 2 présente les indices d'adéquation globale des 6 modèles. En ce qui concerne la sélection des caractéristiques par l'expert ou par les données, nous observons que la méthode basée sur les données semble plus précise, avec un CFI et un TLI plus élevés, et un SRMS, un AIC et un BIC plus faibles. Les méthodes basées sur les données complètent les connaissances de l'expert en détectant des informations supplémentaires à partir des données.

L'indépendance des termes d'erreur entre les équations n'est pas confirmée par les résultats. Chaque modèle contraint à des covariances résiduelles nulles est associé à un ajustement plus faible par rapport à sa version concurrente comprenant une matrice de covariance non restreinte. Les tableaux 4, 5 et 6 (en annexe) présentent les matrices de covariances estimées dans les 3 modèles. Les covariances sont significativement différentes de zéro entre de nombreuses équations de coûts, ce qui confirme les dépendances entre ces estimations. Par exemple, les erreurs de prédictions des coûts d'achat, d'ajustage et d'études sont positivement reliées.

Tableau 2. Indicateurs de performance

	cfi	tli	srmr	aic	bic
M.Expert	0,93	0,68	0,09	1027	1158
M.Expert0	0,66	0,35	0,10	1208	1301
M.AIC	1,00	1,02	0,02	933	1103
M.AIC0	0,78	0,60	0,04	1097	1228
M.BIC	0,99	0,97	0,06	951	1072
M.BIC0	0,75	0,59	0,08	1120	1202

Le Tableau 3 compare les R^2 de chaque équation de coût et suggère que les méthodes basées sur les données améliorent de manière significative l'ajustement et la capacité de prédiction par rapport à la spécification du modèle expert.

La comparaison du R^2 entre les spécifications de la matrice de covariance avec et sans contrainte montre une légère diminution des indices d'ajustement lorsque les covariances estimées sont non nulles. Cependant, la comparaison des R^2 entre régression univariée et régression multivariée reste difficile. En effet, la tâche de prédiction d'un seul composant indépendant des autres est plus facile que la prédiction simultanée du système de composants. La performance devrait être évaluée sur l'ensemble du système.

Tableau 3. Comparaison des R^2

	Achat	Etude	Sous-Traitance	Méthode	Ebauche	Ajustage
M.Expert	0,61	0,37	0,61	0,47	0,47	0,47
M.Exper.0	0,69	0,43	0,72	0,49	0,58	0,50
M.AIC	0,78	0,69	0,44	0,59	0,74	0,66
M.AIC.0	0,79	0,70	0,45	0,59	0,75	0,67
M.BIC	0,74	0,63	0,36	0,51	0,68	0,63
M.BIC.0	0,77	0,67	0,38	0,53	0,71	0,65

La précision générale de la prédiction peut être évaluée avec les autres indicateurs présentés dans le tableau 2. Les modèles où les résidus interdépendants sont plus performants.

La comparaison des R^2 entre équations dans un système indique les qualités relatives de prédiction entre les composantes du coût. Pour le modèle final, M.AIC, la prédiction du coût d'achats est la plus précise ($R^2=0,78$), avec le coût d'ébauche ($R^2=0,74$), suivi du coût d'étude ($R^2=0,69$), d'ajustage ($R^2=0,66$), de sous-traitance ($R^2=0,44$).

Nous pouvons observer que l'estimation du coût d'étude est améliorée par les méthodes de sélection des caractéristiques basées sur les données, alors que les connaissances des experts semblent être plus précises pour le coût de sous-traitance.

Dans le meilleur modèle : M.AIC, les caractéristiques ont un impact significatif sur les différents composants du coût (Tableau 7 - Annexe).

La longueur du moule, le nombre de mouvements mécaniques "simples" sont des caractéristiques significatives dans tous les composants, avec des impacts positifs. Les variables continues sont standardisées, ce qui permet de comparer la taille des impacts. La longueur du moule a un impact plus important sur le coût de sous-traitance (ST) et des achats et plus faible sur le coût d'étude. Le nombre de mouvements mécaniques "simples" a un impact plus important sur le coût d'ajustage et d'étude et plus faible sur celui de la sous-traitance.

Le fait que le mécanisme d'éjection se fait par vérin ou non (*commande.ejection.verin*) a un impact significatif sur le coût d'ébauche, ce qui est cohérent pour l'expert, car il s'agit d'une option plus coûteuse pour l'usinage, qui est réalisée, soit pour des contraintes techniques, soit à la demande du client.

Pour le coût de la méthode, l'acier de moulage est significatif. On peut supposer que le H13 est un matériau aux propriétés similaires à celles du matériau le plus couramment utilisé, ce qui peut entraîner moins de difficultés pour la phase FAO et donc une réduction du coût de cette activité.

Il est surprenant de constater que les variables du type de régulation du système de refroidissement n'ont pas d'impact significatif sur les coûts. On s'attendait à ce que les coûts de cette activité soient plus élevés pour le type normal que pour le type chaud.

Le modèle estime que le type d'acier de moulage classique (110 KG et H13) a également un impact négatif significatif sur le coût d'étude par rapport au type de référence H11. Cependant, il est difficile d'établir un lien réel entre cette information et l'estimation selon la vision de l'expert.

La majorité des autres relations significatives sont cohérentes avec la vision d'expert.

5 Conclusion et perspectives

Dans cet article, nous proposons et évaluons plusieurs modèles d'estimation des coûts par apprentissage automatique correspondant aux combinaisons de la méthode SEM et de plusieurs méthodes courantes de sélection des caractéristiques, y compris la sélection par l'expertise et deux modèles de sélection par apprentissage automatique (sélection stepwise AIC et stepwise BIC). Des expériences numériques, réalisées sur une étude de cas d'une entreprise de moules à injection, ont montré tout d'abord que les modèles basés sur la combinaison des méthodes SEM avec la sélection des caractéristiques stepwise AIC donnent la meilleure estimation selon les mesures de performance classiques, telles que CFI, TLI, SRMR, AIC et BIC. Deuxièmement, la précision de la prédiction est différente pour chacun des six composants considérés du coût total.

Ensuite, certaines caractéristiques sont statistiquement importantes pour tous les éléments de coût (longueur du moule, le nombre de mouvements mécaniques "simples"), tandis que d'autres ne le sont que pour un ou plusieurs éléments. Ensuite, les inducteurs de coûts pour chaque élément de coût sont souvent cohérents avec la vision des experts en devis. D'autres recherches pourraient être menées sur les aspects suivants : 1) tester le modèle dans le contexte des big data en tenant compte de l'ajustement des hyperparamètres, 2) intégrer une méthode de regroupement comme K-mean, LCM ou Kamila dans le cadre de l'apprentissage automatique, 3) comparer la méthode SEM avec d'autres méthodes classiques d'apprentissage automatique et d'apprentissage profond comme les forêts aléatoires et les réseaux de neurones, et 4) prendre en compte d'autres facteurs de coûts liés à des phénomènes économiques comme l'inflation, la rareté des matériaux et la fluctuation des prix sur le marché.

L'accompagnement des experts de l'entreprise dès le début du développement de ces méthodes d'estimation des coûts nous permet de bien sélectionner les modèles pertinents, de bien préparer les données pour entraîner, tester, et valider les modèles, ainsi que d'interpréter les résultats obtenus. De plus, cet accompagnement permet à l'entreprise une compréhension complète de la méthode proposée. Les travaux futurs envisagés pour améliorer la performance de prédiction des modèles actuels sont de développer un outil plus convivial avec une interface visuelle facile à utiliser pour les opérateurs/techniciens. Enfin, la documentation de la méthode proposée devra être faite pour la rendre accessible aux personnels qui n'ont pas de compétence sur l'IA.

6 Remerciements

Cette recherche est financée par le projet Accel 4.0 Grenoble INP - Génie Industriel, le Programme Investissements Avenir, et le Plan de Relance, Nous tenons également à remercier Romain Hoarau, Maxime Cacace-Soret et Eloi Peretti pour leur aide dans la transformation et le nettoyage des données.

7 Références

- [1] S. Yoo and N. Kang, "Explainable artificial intelligence for manufacturing cost estimation and machining feature visualization," *Expert Syst. Appl.*, vol. 183, no. April, p. 115430, 2021, doi: 10.1016/j.eswa.2021.115430.
- [2] M. Doan, N. Pelouzet, and I. Joly, "Artificial intelligence for product cost estimation in manufacturing: review and perspectives," in *1er congrès SAGIP*, 2023, pp. 3–7.
- [3] O. Duran, J. MacIel, and N. Rodriguez, "Comparisons between two types of neural networks for manufacturing cost estimation of piping elements," *Expert Syst. Appl.*, vol. 39, no. 9, pp. 7788–7795, 2012, doi: 10.1016/j.eswa.2012.01.095.
- [4] D. T. Masel, W. A. Young, and R. P. Judd, "A rule-based approach to predict forging volume for cost estimation during product design," *Int. J. Adv. Manuf.*

- Technol.*, vol. 46, no. 1–4, pp. 31–41, 2010, doi: 10.1007/s00170-009-2108-6.
- [5] H. Zhang *et al.*, “A novel method based on a convolutional graph neural network for manufacturing cost estimation,” *J. Manuf. Syst.*, vol. 65, no. June, pp. 837–852, 2022, doi: 10.1016/j.jmsy.2022.10.007.
- [6] Z. Ding, Z. Jiang, Y. Liu, Y. Wang, and C. Li, “A Big Data based Cost Prediction Method for Remanufacturing End-of-Life Products,” *Procedia CIRP*, vol. 72, pp. 1362–1367, 2018, doi: 10.1016/j.procir.2018.03.129.
- [7] F. H’mida, P. Martin, and F. Vernadat, “Cost estimation in mechanical production: The Cost Entity approach applied to integrated product engineering,” *Int. J. Prod. Econ.*, vol. 103, no. 1, pp. 17–35, 2006, doi: 10.1016/j.ijpe.2005.02.016.
- [8] D. Ben-Arieh and L. Qian, “Activity-based cost management for design and development stage,” *Int. J. Prod. Econ.*, vol. 83, no. 2, pp. 169–183, 2003, doi: 10.1016/S0925-5273(02)00323-7.
- [9] F. Bodendorf, Q. Xie, P. Merkl, and J. Franke, “A multi-perspective approach to support collaborative cost management in supplier-buyer dyads,” *Int. J. Prod. Econ.*, vol. 245, no. November 2021, 2022, doi: 10.1016/j.ijpe.2021.108380.
- [10] S. Cavalieri, P. Maccarrone, and R. Pinto, “Parametric vs. neural network models for the estimation of production costs: A case study in the automotive industry,” *Int. J. Prod. Econ.*, vol. 91, no. 2, pp. 165–177, 2004, doi: 10.1016/j.ijpe.2003.08.005.
- [11] X. CHEN, J. HUANG, and M. YI, “Development cost prediction of general aviation aircraft using combined estimation technique,” *Chinese J. Aeronaut.*, vol. 34, no. 4, pp. 32–41, 2021, doi: 10.1016/j.cja.2020.07.024.
- [12] N. Sajadfar and Y. Ma, “A hybrid cost estimation framework based on feature-oriented data mining approach,” *Adv. Eng. Informatics*, vol. 29, no. 3, pp. 633–647, 2015, doi: 10.1016/j.aei.2015.06.001.
- [13] T. Buechler, M. Kolter, L. Hallweger, and M. F. Zaeh, “Predictive cost comparison of manufacturing technologies through analyzing generic features in part screening,” *CIRP J. Manuf. Sci. Technol.*, vol. 38, pp. 299–319, 2022, doi: 10.1016/j.cirpj.2022.04.012.
- [14] P. C. Chang, J. J. Lin, and W. Y. Dzan, “Forecasting of manufacturing cost in mobile phone products by case-based reasoning and artificial neural network models,” *J. Intell. Manuf.*, vol. 23, no. 3, pp. 517–531, 2012, doi: 10.1007/s10845-010-0390-7.
- [15] R. H. Hoyle, *Handbook fo Structural Equation Modelling*. The Guilford Press, 2023.
- [16] J. B. GRACE *et al.*, “Guidelines for a graph-theoretic implementation of structural equation modeling,” vol. 3, no. August 2012, pp. 1–44.
- [17] H. Li, Y. Liu, and K. Peng, “Characterizing the relationship between road infrastructure and local economy using structural equation modeling,” *Transp. Policy*, vol. 61, no. 1, pp. 17–25, 2018, doi: 10.1016/j.tranpol.2017.10.002.
- [18] F. M. Ottaviani and A. De Marco, “Multiple Linear Regression Model for Improved Project Cost Forecasting,” *Procedia Comput. Sci.*, vol. 196, no. 2021, pp. 808–815, 2021, doi: 10.1016/j.procs.2021.12.079.
- [19] D. O. Kazmer, *Injection Mold Design Engineering*. 2007. doi: 10.3139/9783446434196.fm.
- [20] S. Borzel and J. Frochte, “Case study on model-based application of machine learning using small CAD databases for cost estimation,” *IC3K 2019 - Proc. 11th Int. Jt. Conf. Knowl. Discov. Knowl. Eng. Knowl. Manag.*, vol. 1, no. Ic3k, pp. 258–265, 2019, doi: 10.5220/0007979802580265.
- [21] F. Bodendorf and J. Franke, “A machine learning approach to estimate product costs in the early product design phase: A use case from the automotive industry,” *Procedia CIRP*, vol. 100, pp. 643–648, 2021, doi: 10.1016/j.procir.2021.05.137.
- [22] M. Relich and R. Wójcik, “Decision Support Within Customized Products Using Computational Intelligence,” *IFAC-PapersOnLine*, vol. 55, no. 2, pp. 211–216, 2022, doi: 10.1016/j.ifacol.2022.04.195.
- [23] J. Y. Jung, “Manufacturing cost estimation for machined parts based on manufacturing features,” *J. Intell. Manuf.*, vol. 13, no. 4, pp. 227–238, 2002, doi: 10.1023/A:1016092808320.
- [24] E. M. Shehab and H. S. Abdalla, “Manufacturing cost modelling for concurrent product development,” *Robot. Comput. Integr. Manuf.*, vol. 17, no. 4, pp. 341–353, 2001, doi: 10.1016/S0736-5845(01)00009-6.
- [25] H. S. Wang, Y. N. Wang, and Y. C. Wang, “Cost estimation of plastic injection molding parts through integration of PSO and BP neural network,” *Expert Syst. Appl.*, vol. 40, no. 2, pp. 418–428, 2013, doi: 10.1016/j.eswa.2012.01.166.
- [26] Z. H. Che, “PSO-based back-propagation artificial neural network for product and mold cost estimation of plastic injection molding,” *Comput. Ind. Eng.*, vol. 58, no. 4, pp. 625–637, 2010, doi: 10.1016/j.cie.2010.01.004.
- [27] A. Probst, C. Nitzl, F. Kraus, and R. Förstner, “Cost estimation of an asteroid mining mission using partial least squares structural equation modelling (PLS-SEM),” *Acta Astronaut.*, vol. 167, no. November 2019, pp. 440–454, 2020, doi: 10.1016/j.actaastro.2019.07.032.
- [28] K. Petroutsatou and S. Lambropoulos, “Road tunnels construction cost estimation: A structural equation model development and comparison,” *Oper. Res.*, vol. 10, no. 2, pp. 163–173, 2010, doi: 10.1007/s12351-009-0061-7.
- [29] W. H. Greene, *Econometric Analysis*, vol. 8th editio. Pearson Education, 2018. doi: 10.1007/978-3-030-56239-7_5.
- [30] Y. Rosseel, “lavaan: An R Package for Structural Equation Modeling,” *J. Stat. Softw.*, vol. 48, no. 2, pp. 1–93, 2012.

8 Annexe

Tableau 4. Matrice de Variance covariance estimées du modèle M.Expert

	ACHATS	AJUSTAGE	ETUDE	METHODE	ST	EBAUCHE
ACHATS	0,33***	0,21***	0,16***	0,000	0,16***	0,18***
AJUSTAGE		0,61***	0,23***	0,19***	0,23***	0,29***
ETUDE			0,33***	-0,006	0,07	0,30***
METHODE				0,51***	0,20***	0,07
ST					0,46***	0,28***
EBAUCHE						0,52***

*** : pvalue <0.001 ** pvalue <0.01 * pvalue < 0.10

Tableau 5. Matrice de Variance covariance estimées du modèle M.AIC

	ACHATS	AJUSTAGE	ETUDE	METHODE	ST	EBAUCHE
ACHATS	0,22***	0,08***	0,13***	-0,03	0,12***	0,06***
AJUSTAGE		0,31***	0,19***	0,06	0,18***	0,16***
ETUDE			0,55***	0,15***	0,18***	0,14***
METHODE				0,41***	0,006	0,14***
ST					0,26***	0,02
EBAUCHE						0,33***

*** : pvalue <0.001 ** pvalue <0.01 * pvalue < 0.10

Tableau 6. Matrice de Variance covariance estimées du modèle M.BIC

	ACHATS	AJUSTAGE	ETUDE	METHODE	ST	EBAUCHE
ACHATS	0,24***	0,08***	0,16***	-0,004	0,11***	0,07***
AJUSTAGE		0,34***	0,21***	0,08	0,21***	0,16***
ETUDE			0,62***	0,20***	0,20***	0,16***
METHODE				0,47***	0,01	0,18***
ST					0,29***	0,01***
EBAUCHE						0,35***

*** : pvalue <0.001 ** pvalue <0.01 * pvalue < 0.10

Tableau 7. Résultat d'estimation des 6 équations de coûts du modèle M.AIC

Variable /Composant de coût	ACHATS	AJUSTAGE	ETUDE	METHODE	ST	EBAUCHE
Dim.outi.long	0.51***	0.46 ***	0.23*	0.46***	0.66***	0.37***
Nbre_mvt_meca_simple	0.21***	0.32***	0.32***	0.23**	0.15*	0.30***
Nbre_mvt_meca_double	-	0.16*	0.24**	0.23**	0.13*	0.27***
Nbre_mvt_rotatif	-	-	0.11	-	0.06	-
Nbre_total_point	0.13**			0.09	-	0.17***
Type.de.produit_HT	ref	ref	ref	ref	ref	ref
Type.de.produit_MIX	-	0.72**	-	-	0.76***	-
Type.de.produit_PROTO	-	-	-	-	-	-
Type.injection_BLOC_CHAUD	ref	ref	ref	ref	ref	ref
Type.injection_CAROTTE	-0.47***	-	-	-	-	-
Ejection_Cylindrique	ref	ref	ref	ref	ref	ref
Ejection_Autre	-	-0.07	-	-0.37**	-	-
Commande.ejection_Presse	ref	ref	ref	ref	ref	ref
Commande.ejection_Verin	0.66***	0.50***	-	-	0.17	0.55***
Regulation_Chaud	ref	ref	ref	ref	ref	ref
Regulation_Normal	-	-	-	-	-	-
Acier.moulant H11	ref	ref	ref	ref	ref	ref
Acier.moulant_110KG	-	-	-0.28	-	-	-
Acier.moulant_Autre	-	-0.40*	-	-0.51*	-	-0.48*
Acier.moulant_H13	-0.33**	-	-0.33*	-0.41*	-	-0.26*
R2	0.78	0.69	0.44	0.59	0.74	0.67

*** : pvalue <0.001 ** pvalue <0.01 * pvalue < 0.10

Session 1.1 : Apprentissage sur les images

Estimation du Rendement du Mil Perlé (*Pennisetum glaucum*) par Machine Learning à l'aide d'Images Satellites

A. Chemchem¹, L. Mohimont², F. Alin², L.A. Steffemel²

¹ ATOS - Pôle Data Driven Intelligence
Rue du Mas de Verchant, 34000 Montpellier, France

² Université de Reims Champagne-Ardenne,
Laboratoire LICIS - LRC CEA DIGIT

lamine.chemchem@atos.net, lucas.mohimont@univ-reims.fr,
francois.alin@univ-reims.fr, luiz-angelo.steffemel@univ-reims.fr

Résumé

L'estimation du rendement agricole joue un rôle crucial dans la poursuite des objectifs de développement durable des Nations Unies, représentant ainsi un outil essentiel dans la prise de décisions concernant les systèmes d'approvisionnement. Dans ce travail, nous nous intéressons à la prédiction du rendement du *Pennisetum glaucum*, aussi connu comme "mil à chandelle" ou "mil perlé". Connaître le potentiel de production le plus tôt possible permet de prendre des mesures préventives et éviter des défauts d'approvisionnement pour la population. Pour ce faire, nous croisons les données historiques de rendement des parcelles au Sénégal avec des données satellitaires couvrant trois phases différentes du cycle de vie du mil, grâce à des méthodes d'apprentissage automatique. En comparant différentes méthodes, nous avons obtenu des estimations de rendement assez précises 1 mois avant la récolte, avec un taux d'erreur qui ne dépasse pas 140 kg/ha.

Mots-clés

Rendement agricole, Télédétection optique, Apprentissage automatique

Abstract

Agricultural yield estimation contributes to many of the United Nations' sustainable development goals, and can be considered as a decision-making tool for a supply system. In this work, we are interested in predicting the yield of *Pennisetum glaucum*, also known as "pearl millet". Knowing the production potential of this cereal as early as possible enables authorities to take preventive measures and avoid supply shortages for the population. In this work, we cross-reference historical plot yield data in Senegal with satellite data covering three different phases of the millet life cycle, using machine learning methods. By comparing different methods, we obtained fairly accurate yield estimates 1 month before harvest, with an error rate of no more than 140 kg/ha.

Keywords

Crop yield estimation, Optical remote sensing, Machine learning

1 Introduction

Le "mil à chandelle", "mil perlé" ou simplement **mil** (*Pennisetum glaucum*) est une espèce de plantes annuelles de la famille des Poaceae (Graminées). Elle est cultivée comme céréale pour ses graines comestibles et joue un rôle important en tant que culture vivrière en Inde et au Pakistan, ainsi que dans le Sahel africain et dans des zones semi-arides. Connaître le potentiel de production fait partie des mesures permettant de garantir la sécurité alimentaire de la population en cas de baisse de production.

Cette étude vise à estimer les rendements de mil par télédétection optique dans la petite agriculture familiale au Sénégal. Ce projet s'inscrit dans le cadre de la sécurité alimentaire et des moyens de subsistance des populations, et ses résultats pourraient représenter un outil d'aide à la décision pour les services d'approvisionnement de plusieurs ODD (Objectifs de Développement Durable) adaptés par les Nations Unies. En particulier pour l'ODD 2, qui vise à "Éliminer la faim, assurer la sécurité alimentaire et une meilleure nutrition et promouvoir l'agriculture durable".

Le site choisi pour cette étude est situé dans la zone de l'ancien bassin arachidier du Sénégal, comme le montre la figure 1. Il représente la principale zone de production agricole du pays. Le bassin est constitué de sols ferrugineux tropicaux permettant une production agricole principalement composée de céréales sèches (mil, sorgho, maïs) et de légumineuses (arachide, niébé) cultivées seules ou en association.

Le jeu de données de cette étude a été partagé dans le cadre d'un défi de science des données organisé par Acta¹. Il se compose d'un total de 81 parcelles de mil réparties dans le site du bassin. Cette région a été choisie parce qu'il s'agit d'une zone d'intérêt de longue date pour plusieurs équipes de recherche, de sorte que nous disposons de connaissances

1. Instituts techniques agricoles <http://www.acta.asso.fr/>

sur le terrain, en plus de la base de données agronomique ou paysagère historique. Un autre avantage très important est la présence d'équipes de recherche sur le terrain pour assurer et coordonner la collecte de données.

Le climat de la région est unimodal avec une saison des pluies entre juillet et octobre. Cependant, la zone connaît une forte croissance démographique, à laquelle s'ajoutent une réduction des temps de jachère, un appauvrissement progressif des sols, ce qui conduit à une baisse des rendements observés en milieu rural, à une dégradation des ressources naturelles et à la perte de biodiversité [3].

Afin d'atteindre nos objectifs, nous mettons en œuvre des méthodes d'apprentissage automatique avec deux axes : régression et classification. Le premier axe d'étude consiste à prédire le rendement quantitatif avec des méthodes de régression pour chaque parcelle. Dans le deuxième axe, nous adaptons les méthodes de classification supervisée afin d'affiner encore la précision des prédictions. Le résultat final de cette étude pourrait être un OAD (outil d'aide à la décision) de surveillance des risques d'approvisionnement. Le reste du document est organisé comme suit : la section 2 passe en revue la littérature sur l'apprentissage automatique appliqués à l'agriculture intelligente. La section 3 décrit l'ensemble de données et le dispositif expérimental mis en place. La section 4 expose les résultats obtenus avec quelques analyses et discussions. Enfin, la section 5 conclut le document avec quelques perspectives.

2 État de l'Art

Les méthodes d'apprentissage automatique sont utilisées comme une nouvelle approche qui révolutionne les modèles classiques de prédiction du rendement. Toutefois, les approches basées sur l'apprentissage de données historiques nécessitent des données étiquetées pour fournir une modélisation précise. En effectuant une étape d'apprentissage sur les données étiquetées, ces méthodes sont capables de produire des modèles de prédiction quantitatifs appelés méthodes de régression. Par exemple, dans le cas de la prédiction du rendement, on veut prédire la quantité produite en tonnes par hectare, ou en kilogrammes par hectare. Il est aussi possible de construire un modèle de prédiction qualitatif via les méthodes de classification, permettant par exemple de prédire la classe de rendement attendue : faible, moyen ou élevé.

Dans la littérature, on peut citer le travail de [9], qui a comparé deux méthodes de régression : Perceptron multicouche (MLP) et Support Vector Regressor (SVR) pour la prédiction du rendement du maïs. En utilisant les données de l'indice de végétation amélioré (EVI) et des séries temporelles climatiques des dix dernières années, la méthode MLP mise en œuvre a atteint un score R^2 de 0,81. L'EVI et les données satellitaires en général sont utiles pour compenser le manque de données agricoles de détection sur le terrain. Malheureusement, nous n'avons pas pu accéder à la base de données de cette étude pour comparer nos méthodes, mais nous avons mis en œuvre les approches SVR et MLP sur notre ensemble de données.

Une autre recherche intéressante est celle de [14], dans laquelle les auteurs ont utilisé des séries temporelles de NDVI (Normalized Difference Vegetation Index) sur 10 jours avec des apports d'engrais chimiques comme caractéristiques d'apprentissage pour la prédiction du rendement du blé. Ils ont comparé différentes caractéristiques basées sur le NDVI, le NDVI cumulé et le NDVI cumulé à des dates significatives avec l'arbre de régression boosté (Boosted Regression Tree - BRT) et le SVR. Les deux techniques ont été utilisées pour la sélection des caractéristiques et la modélisation de la régression. Les meilleurs résultats ont été obtenus avec le BRT, avec une erreur calculée par RMSE (Root Means Square Error) inférieure à 0,2 tonne par hectare. Dans notre étude, nous avons utilisé l'indice NDVI avec cinq autres indices.

Dans [8], les auteurs ont utilisé des séries temporelles NDVI de 16 jours avec une résolution de 250 mètres et des observations climatiques SILO² pour la prévision du rendement du blé en Australie. Chaque échantillon de données est un pixel correspondant à un carré de 250m x 250m et différents modèles de base et d'ensemble ont été comparés sur la précision basée sur le pixel avec une validation croisée 5 fois. Les meilleurs résultats ont été obtenus par un SVR (Support vector Regressor) adapté utilisant le noyau RBF (radial basis function) qui atteint une erreur RMSE de 0,55t/ha et R^2 de 0,77.

Une autre étude de [12] se concentre sur l'axe de la classification supervisée afin de traiter les images satellites. Les auteurs ont utilisé les données NDVI, pédologiques et météorologiques pour prédire le rendement du maïs à l'intérieur d'un champ. Le champ a été divisé en 63 unités de traitement et la modélisation a été simplifiée à une classification binaire : classes de faible rendement et de rendement moyen à élevé. Cinq classificateurs supervisés différents ont été comparés avec une validation croisée 5 fois, et les meilleurs résultats ont été obtenus par le classificateur XG-Boost avec une précision de 95%. Cette étude nous a incités à réaliser une étude comparative des méthodes de classification les plus populaires.

Des méthodes basées uniquement sur l'historique de la production et les données météorologiques ont aussi montré des résultats prometteurs. Ainsi, [5] a pu obtenir des taux de précision supérieurs à 99% pour la culture du soja ou du riz, et de 98% pour la culture du maïs, en utilisant la méthode Random Forest sur des données issues de régions agricoles brésiliennes. Ce travail a toutefois bénéficié d'un historique sur plus de 20 ans de production agricole, ce qui n'est pas possible dans notre cas.

L'apprentissage profond peut également être utilisé pour prédire le rendement. Souvent, l'apprentissage profond nécessite de grands ensembles de données, mais dans certains cas, l'apprentissage par transfert peut être mis en œuvre pour compenser le manque de données en utilisant un modèle pré-entraîné. Les auteurs de [15] ont utilisé l'apprentissage par transfert pour former un modèle de mémoire à

2. SILO est une base de données du gouvernement du Queensland contenant des données climatiques quotidiennes continues pour l'Australie de 1889 à nos jours.

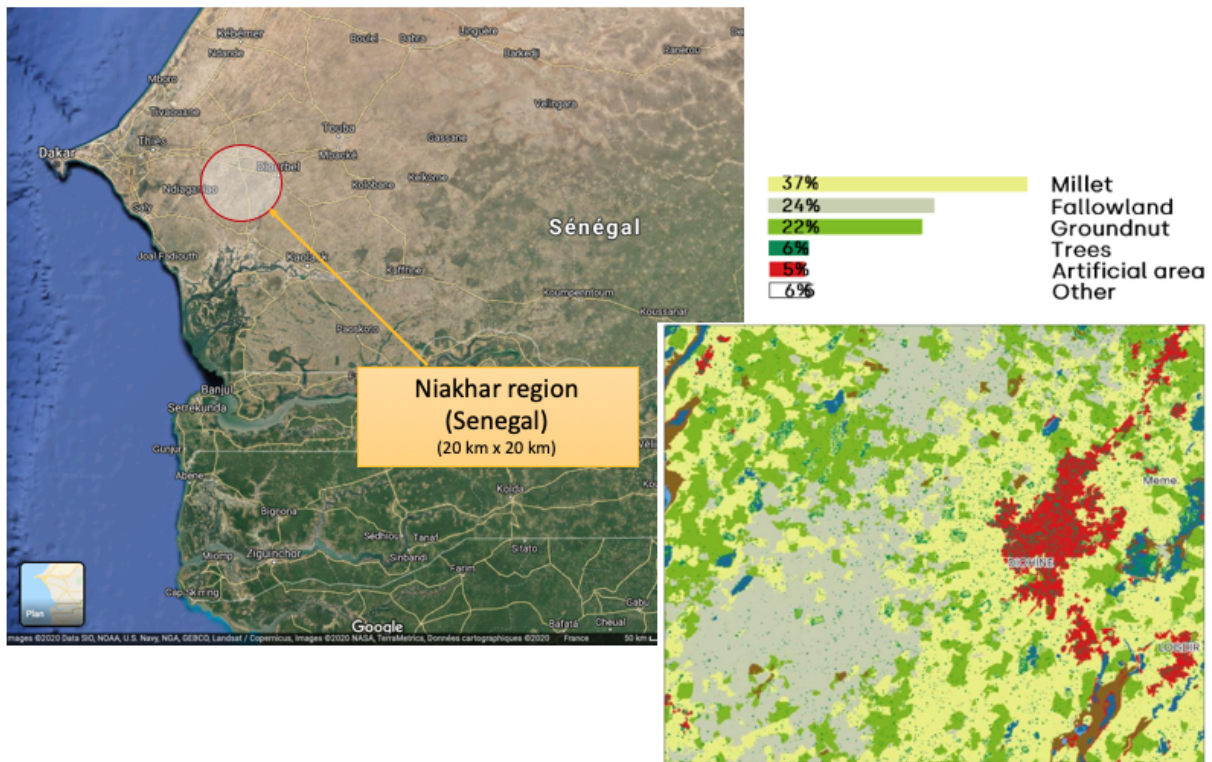


FIGURE 1 – Site étudié : région de Niakhar, ancien bassin producteur d'arachide, Sénégal [11].

long terme (LSTM) avec la réflectance MODIS³ et les séries temporelles de températures pour la prédiction du rendement du maïs au Brésil. Le modèle a d'abord été entraîné sur un plus grand ensemble de données, avec 1837 échantillons de récoltes en Argentine, puis le modèle a été réentraîné sur 336 données de récoltes brésiliennes. La LSTM avec apprentissage par transfert a obtenu de meilleurs résultats moyens que la LSTM sans apprentissage par transfert. C'est pourquoi nous prévoyons d'explorer cette méthode sur notre ensemble de données dans un travail futur.

3 Matériaux & Méthodes

3.1 L'ensemble de données

L'ensemble de données utilisé dans cette étude est récolté à partir de 81 parcelles de mil situées dans la région de Niakhar, au Sénégal. Ce jeu de données comprend des données historiques collectées sur les années 2017 et 2018, représentant le rendement du mil en Kg/ha, ainsi que les données satellitaires des parcelles correspondantes, comme le montre la figure 2. Il faut noter que les parcelles étudiées sont trop petites, pour cette raison nous ne pouvons pas réaliser une étude individualisée (*intra-field*).

Les données satellitaires extraites contiennent les indices de végétation suivants : **NDVI**, **MSAVI2**, **NDWI**, **CIGreen**, **GDVI** et **PSRINIR**. Ces indices permettent d'estimer les paramètres biophysiques et sont calculés à partir de la ré-

fectance de deux bandes spectrales, rouge (R) et proche infrarouge (NIR).

L'indice NDVI (Normalized Difference Vegetation Index) est le plus utilisé. Le calcul de cet indice est basé sur la réflectance de la chlorophylle dans le proche infrarouge et permet de suivre la biomasse intraparcellaire. L'indice SAVI (Soil-Adjusted Vegetation) est dérivé de cet indice et propose un ajustement avec une constante.

Plus tard, l'indice MSAVI2 (Modified Soil-Adjusted Vegetation Index) a été proposé par [13], il utilise une constante ajustée aux conditions locales.

Le NDWI (Normalized Difference Water Index) est basé sur le même principe que le NDVI et permet de surveiller l'état hydrique des cultures (Gao, 1996). Le NDWI est basé sur le pic d'absorption de l'eau dans une bande infrarouge de courte longueur d'onde.

Le CIGreen (Green Chlorophyll Index) est utilisé pour évaluer la teneur en chlorophylle des feuilles. Cet indice est sensible aux petites variations de chlorophylle.

Le GDVI (Generalized Difference Vegetation Index) est un indice dérivé du NDVI, particulièrement adapté aux zones arides où le couvert végétal est faible [16].

Finalement, le PSRINIR (Plant Senescence Reflectance Index Near Infra Red), proposé par [10], compare les caroténoïdes et les chlorophylles, identifiant ainsi la sénescence de la canopée (augmentation des caroténoïdes).

Les approches d'apprentissage automatique présentées dans cet article ont été mises en œuvre en utilisant le langage python avec la bibliothèque Scikit-Learn. L'optimi-

3. <https://visibleearth.nasa.gov/images/54078/modis-surface-reflectance>

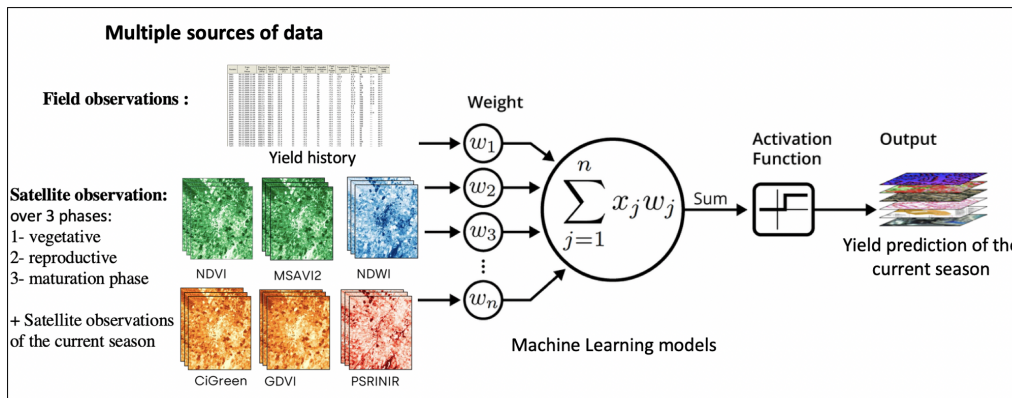


FIGURE 2 – Workflow général : De l'intégration des données à la prédiction du rendement.

sation des hyperparamètres a été accélérée grâce aux ressources du Centre de Calcul Régional ROMEO⁴ de l'Université de Reims Champagne-Ardenne.

3.2 Méthodologie

Les observations par satellite sont collectées au cours de trois périodes de croissance différentes, conformément au calendrier de culture du mil à chandelle (*Pennisetum glaucum*) expliqué dans la figure 3.



FIGURE 3 – Calendrier culturel du mil à chandelle dans l'ancien bassin arachidier, Sénégal [11].

Dans un premier moment, nous avons choisi d'estimer les rendements le plus tôt possible avant la récolte. Cela se fait en estimant le rendement uniquement avec des données de la phase végétative (5 mois avant la récolte), sans prendre en compte les données des phases reproductive et de maturation. Une deuxième expérience consiste à prédire les rendements 3 mois avant la récolte, utilisant cette fois-ci les données des phases végétative et reproductive. Finalement, la troisième expérience consiste à utiliser toutes les données disponibles, c'est-à-dire les données des trois phases végétation, reproduction et maturation, afin de prédire le rendement du mil environ 1 mois avant la récolte.

En outre, et pour tirer parti des méthodes d'apprentissage automatique, nous avons mis en œuvre des algorithmes pour estimer le rendement quantitatif, par le biais de méthodes de régression. Puis, dans un second temps, nous avons exploré les méthodes de classification supervisée afin de valider les résultats et d'être le plus précis possible. Pour chaque type d'approche d'apprentissage automatique, nous avons mis en place une étude comparative des algorithmes les plus prometteurs tels que décrits dans la section littérature. Il est à noter que, pour chaque approche d'apprentissage, les meilleurs hyperparamètres sont sélectionnés par

validation croisée et recherche en grille (*GridSearch*).

3.3 Méthodes pour la régression et leur évaluation

Dans cette partie, nous expliquons les étapes que nous avons suivies pour réaliser l'étude comparative des algorithmes de régression.

Prétraitement des données par mise à l'échelle des caractéristiques : cette étape est appliquée pour normaliser la plage des variables indépendantes de l'ensemble de données. Étant donné que la plage de valeurs des données brutes varie considérablement, les fonctions objectives ne fonctionneront pas correctement sans normalisation. Une autre raison pour laquelle la mise à l'échelle des caractéristiques est appliquée est que la descente de gradient converge beaucoup plus rapidement avec la mise à l'échelle des caractéristiques [7].

Dans notre implémentation, nous avons appliqué la normalisation min-max, qui est la méthode la plus simple et qui consiste à remettre à l'échelle la plage de caractéristiques dans un intervalle [0, 1] ou [-1, +1]. La sélection de la plage cible dépend de la nature des données, et puisque dans notre ensemble de données il n'y a pas de données négatives, nous les avons mises à l'échelle entre [0, 1].

Validation par répartition *train/test* : Fondamentalement, l'évaluation des approches d'apprentissage automatique se fait par la division de l'ensemble de données en deux ensembles, l'un appelé ensemble d'entraînement (*train*) et l'autre ensemble de test. Le premier contient les données avec les étiquettes utilisées pour construire le modèle, tandis que le second est utilisé pour tester les performances de ce modèle. Dans notre cas, comme nous ne disposons que de deux années d'historique de données, nous avons pris les données de 2017 pour former les modèles et celles de 2018 pour les tester.

Évaluation de la régression par R²-score et RMSE : nous avons évalué nos modèles de régression par les deux formules d'évaluation les plus connues : R²-score et RMSE. Le Score R² (appelé aussi R-carré) est une mesure statistique de la proximité des données par rapport à la

4. <http://romeo.univ-reims.fr>

droite de régression ajustée. Il est également connu sous le nom de coefficient de détermination ou de coefficient de détermination multiple pour la régression multiple [4]. En général, plus le R au carré est élevé, mieux le modèle s'adapte à vos données.

Le RMSE (Root Mean Squared Error) est la racine carrée de l'erreur quadratique moyenne (Mean Squared Error - MSE), qui est une fonction de risque correspondant à la valeur attendue de la perte d'erreur quadratique.

3.4 Méthodes de classification supervisée et leur évaluation

Après la mise en œuvre de l'étude comparative de régression, nous avons exploré les méthodes d'apprentissage automatique pour la classification supervisée, en transformant le problème de régression en un problème de classification multi-classes. Cette approche vise à affiner la précision de notre prédiction de rendement, de sorte qu'au lieu d'essayer de prédire une valeur quantitative, il nous suffit de prédire une classe parmi trois classes possibles : rendement faible, rendement moyen ou rendement élevé.

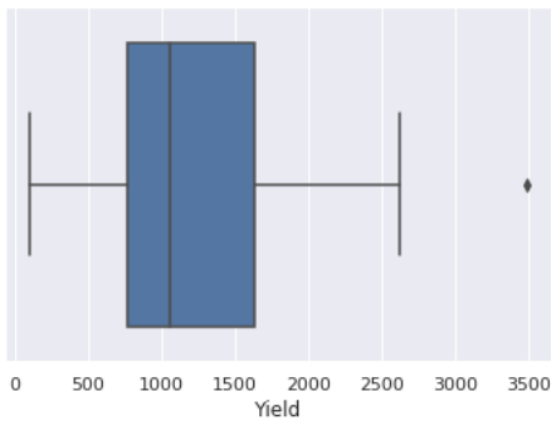


FIGURE 4 – Distribution des valeurs de rendement.

La distribution des valeurs de rendement se situe entre 107,9 kg/ha et 3488,9 kg/ha, comme le montre la figure 4. À partir de cette distribution, nous avons créé les trois classes de rendement suivantes :

- Classe "rendement bas" : si $700 \text{ kg/ha} > \text{rendement}$;
- Classe "rendement moyen" : si $700 \text{ kg/ha} \leq \text{rendement} < 1600 \text{ kg/ha}$;
- Classe "rendement haut" : si $\text{rendement} \geq 1600 \text{ kg/ha}$.

Le résultat de la distribution des classes obtenues est présenté dans la figure 5. Nous pouvons remarquer que les classes obtenues sont fortement déséquilibrées, ce qui nécessite une étape de prétraitement supplémentaire pour cet ensemble de données, avec la technique SMOTE (Synthetic Minority Over-sampling Technique). Dans SMOTE, les classes minoritaires sont suréchantillonnées en introduisant des instances synthétiques dans lesquelles chaque échantillon de classe minoritaire est prélevé. Les données générées sont insérées le long des segments de ligne reliant cer-

tains des k plus proches voisins de la classe minoritaire. Les voisins sont choisis au hasard parmi les k plus proches voisins en fonction de l'ampleur du suréchantillonnage nécessaire. Cinq voisins les plus proches sont actuellement utilisés dans la mise en œuvre de SMOTE [1] [2].



FIGURE 5 – Distribution des classes de rendement.

Afin de valider les résultats, nous avons utilisé les méthodes et métriques suivantes :

Validation des résultats par répartition train/test : De la même manière que pour l'étude comparative de régression, les données ont été réparties en deux groupes, l'un pour l'entraînement et l'autre pour le test. En raison de la faible quantité de données, l'ensemble d'entraînement couvre l'année 2017 tandis que l'ensemble de test correspond aux données de 2018.

Résultats Évaluation par F1-score & Accuracy : Nous avons évalué nos modèles de classification par les deux formules d'évaluation les plus connues : le score F et la précision (*accuracy*) de la classification. Le score F, également appelé mesure F, est basé sur les deux mesures principales : la précision (*precision*) et le rappel (*recall*). La précision est la proportion de cas que le sujet a classés comme positifs et qui étaient vraiment positifs (TP - *true positive*). Elle est équivalente à la valeur prédictive positive. Le rappel est la proportion de cas vraiment positifs qui ont été classés comme positifs par le modèle. Il est équivalent à la sensibilité.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Où TP est le nombre de vrais positifs, TN est le nombre de vrais négatifs, FP est le nombre de faux positifs et FN est le nombre de faux négatifs.

Les deux métriques sont souvent combinées sous la forme de leur moyenne harmonique [6] appelé F-Score. La métrique F-score peut être utilisée pour équilibrer la contribution des faux négatifs en pondérant le rappel par un paramètre $\beta \geq 0$. Dans notre cas, β est fixé à 1, le score F1 est alors égal à :

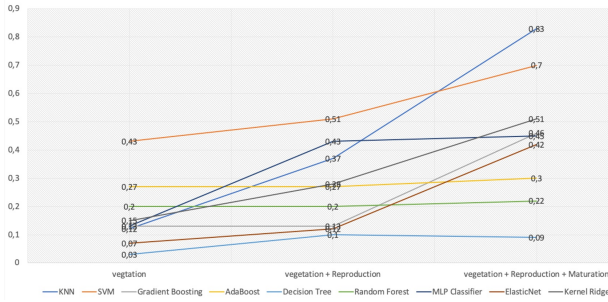


FIGURE 6 – Comparaison R^2 -score.

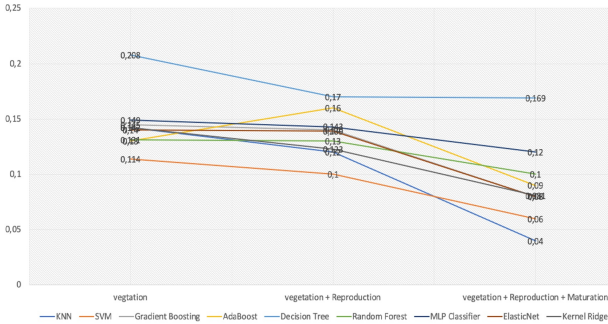


FIGURE 7 – Comparaison RMSE.

$$F1_score = \frac{2 \times recall \times precision}{precision + recall}$$

Finalement, nous utilisons la métrique *accuracy* (traduite par exactitude ou justesse), l'un des critères les plus connus pour évaluer les modèles de classification. D'une manière non formelle, elle se réfère à la proportion de prédictions correctes faites par le modèle. Sa formule est la suivante :

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

4 Résultats et Discussion

4.1 Le cas de la Régression

Après avoir effectué l'étape de prétraitement, les résultats obtenus à partir des méthodes de régression sont résumés dans le tableau 1.

En synthétisant les résultats obtenus et mentionnés dans le tableau, nous pouvons voir sur la figure 6 la comparaison du score R^2 des méthodes de régression au cours des trois phases de maturation du mil : végétative, reproductive et de maturation. De même manière, la figure 7 compare ces méthodes selon la métrique RMSE.

A partir de ces résultats, nous pouvons voir que les meilleurs scores pour les prédictions de rendement dans les phases végétative et reproductive sont donnés par la méthode SVM, alors que dans la phase de maturation les meilleurs scores sont donnés par la méthode du régresseur K plus proche voisin. Cela est en partie dû à la plus grande quantité de données disponibles lorsqu'on réduit le temps avant la récolte : plus nous ajoutons de nouvelles données

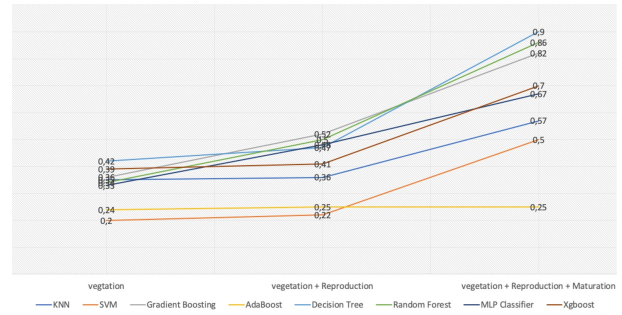


FIGURE 8 – Comparaison F-score.

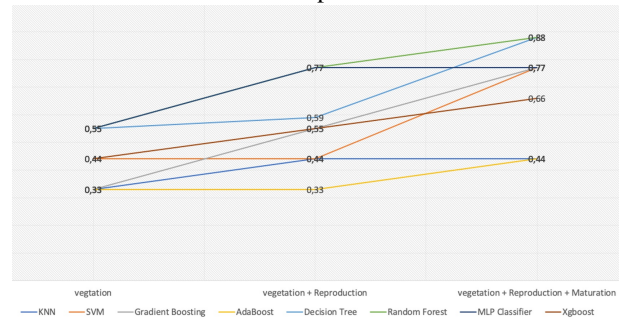


FIGURE 9 – Comparaison accuracy.

aux cycles de vie du mil, plus la précision des méthodes de régression augmente.

De ce fait, l'erreur de prédiction diminue de RMSE = 0,11, équivalent à une erreur de 400 kg/ha si la prédiction n'inclut que des données de la phase végétative, à un RMSE de 0,04, correspondant à une erreur de prédiction de 140 kg/ha lors de la phase de maturation.

4.2 Le cas de la Classification

Après l'étape de prétraitement, les résultats obtenus par les méthodes mises en œuvre en utilisant les méthodes de validation et d'évaluation expliquées précédemment sont résumés dans le tableau 2. En schématisant les résultats obtenus et mentionnés dans le tableau, nous pouvons voir sur la figure 8 la comparaison du score R^2 des méthodes de classification au cours des trois phases du cycle de vie du mil : végétative, reproductive et de maturation. De même, la figure 9 compare ces méthodes en évaluant leur précision.

D'après ces résultats, nous pouvons voir que les meilleurs scores pour la prédiction de la classe de rendement en utilisant uniquement les données végétatives (c'est-à-dire 5 mois avant la récolte) sont donnés par la méthode de l'arbre de décision, qui atteint une précision de 0,56. De même, dans la phase de maturation (lorsque l'on utilise tous les stades du cycle de vie du mil), les meilleures prédictions sont faites par le modèle d'arbre de décision avec 90% du score F et 88% de l'accuracy, suivi de près par la méthode Random Forest.

Comme nous nous y attendions, la précision de prédiction augmente de F-score = 0,42 pendant la phase végétative, à F-score = 0,90, ce qui signifie une erreur de prédiction de

Approche ML	Phase végétative		Végétative+reproduction		Toutes les trois phases	
	R2-score	RMSE	R2-score	RMSE	R2-score	RMSE
K-Nearest Neighbors	0.12	0.14	0.37	0.12	0.83	0.04
Support Vector Machine	0.43	0.11	0.51	0.10	0.70	0.06
Gradient Boosting	0.13	0.14	0.13	0.14	0.46	0.08
Ada Boosting	0.27	0.13	0.27	0.16	0.30	0.09
Decision Tree	0.03	0.20	0.10	0.17	0.09	0.16
Random Forest	0.20	0.13	0.20	0.13	0.22	0.10
ElasticNet	0.07	0.14	0.12	0.14	0.42	0.08
Kernel Ridge	0.15	0.14	0.28	0.12	0.51	0.08
Multi Layer Perceptron	0.13	0.19	0.43	0.11	0.45	0.09

TABLE 1 – Comparaison de performance pour les algorithmes de régression.

Approche ML	Phase végétative		Végétative+reproduction		Toutes les trois phases	
	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy
K-Nearest Neighbors	0.35	0.33	0.36	0.44	0.57	0.44
Support Vector Machine	0.20	0.44	0.22	0.44	0.50	0.77
Gradient Boosting	0.36	0.33	0.52	0.55	0.82	0.77
Ada Boosting	0.24	0.33	0.25	0.33	0.24	0.44
Decision Tree	0.42	0.56	0.47	0.59	0.90	0.88
Random Forest	0.34	0.55	0.50	0.77	0.86	0.88
Xgboost	0.39	0.44	0.41	0.55	0.70	0.66
Multi Layer Perceptron	0.20	0.55	0.50	0.77	0.72	0.77

TABLE 2 – Comparaison des performances des algorithmes de classification.

classe de 10% pendant la phase de maturation (c.-à-d. un mois avant la récolte).

4.3 Discussion

Les résultats obtenus dans cette étude se révèlent particulièrement satisfaisants, surtout lorsque l'on considère les limitations auxquelles nous avons dû faire face. Premièrement, notre ensemble de données d'entraînement était assez restreint, ne couvrant qu'une période de deux ans. Malgré cette contrainte temporelle, les performances de notre modèle ont été prometteuses, ce qui témoigne de son potentiel même avec des données limitées. De plus, un défi majeur auquel nous avons été confrontés était le manque de données météorologiques et de données de sol. Ces informations sont cruciales pour modéliser avec précision les rendements agricoles, mais malheureusement, leur disponibilité était limitée dans notre contexte. Malgré ces obstacles, les résultats que nous avons obtenus soulignent l'efficacité de notre approche méthodologique et suggèrent des possibilités futures pour améliorer encore davantage la prédiction des rendements agricoles dans des conditions de données similaires.

5 Conclusions et Perspectives

Cette étude montre l'application de méthodes d'apprentissage automatique en particulier afin d'améliorer l'estimation des rendements pour des paysages agricoles complexes, en utilisant des images satellites optiques à haute résolution spatiale et temporelle.

Dans un premier temps, nous avons exploré les méthodes

de régression pour obtenir des estimations assez précises, avec un R2 score qui atteint 0.83 un mois avant la récolte. Dans un deuxième temps, et pour affiner encore la précision des prédictions, nous avons mis en œuvre des méthodes de classification supervisée. Grâce à cela, nous avons obtenu de bonnes prédictions de rendement des cultures avec une précision de 90% pour la classe de rendement un mois avant la récolte. Enfin, nous pouvons dire que les résultats obtenus par cette étude sont vraiment satisfaisants, car nous ne disposons pas d'un grand ensemble d'entraînement (seulement deux années de données).

Comme perspective, nous avons l'intention d'enrichir cette base de données par l'historique des années subséquentes, et aussi d'essayer de généraliser cette application pour d'autres cultures telles que le blé ou le maïs. Nous prévoyons également de croiser les cartes satellites avec les cartes météorologiques afin d'implémenter des réseaux de neurones profonds et d'étudier leurs comportements pour la prédiction des rendements.

Références

- [1] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote : synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16 :321–357, 2002.
- [2] Amine Chemchem, François Alin, and Michaël Krajecki. Combining smote sampling and machine learning for forecasting wheat yields in france. In *2019 IEEE Second International Conference on Artificial Intelli-*

- gence and Knowledge Engineering (AIKE), pages 9–14. IEEE, 2019.
- [3] Agricultural Research for Development Cirad. Diversité paysagère et sécurité alimentaire en Afrique. <https://www.projects.igeo.fr/sites-d-etudes/>, 2018. [Online; accessed 26-February-2020].
- [4] Manal Fawzy, Mahmoud Nasr, Samar Adel, and Shacker Helmi. Regression model, artificial neural network, and cost estimation for biosorption of ni (ii)-ions from aqueous solutions by *potamogeton pectinatus*. *International journal of phytoremediation*, 20(4) :321–329, 2018.
- [5] Lilian Hollard, Angelica Durigon, and Luiz Angelo Steffanel. Machine learning forecast of soybean yields on south brazil. In *Workshop on Edge AI for Smart Agriculture (EAISA 2022)*, 2022.
- [6] George Hripcsak and Adam S Rothschild. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association*, 12(3) :296–298, 2005.
- [7] Sergey Ioffe and Christian Szegedy. Batch normalization : Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv :1502.03167*, 2015.
- [8] Elisa Kamir, François Waldner, and Zvi Hochman. Estimating wheat yields in australia using climate records, satellite image time series and machine learning methods. *ISPRS Journal of Photogrammetry and Remote Sensing*, 160 :124 – 135, 2020.
- [9] K. Kuwata and R. Shibasaki. Estimating crop yields with deep learning and remotely sensed data. In *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 858–861, July 2015.
- [10] Mark N Merzlyak, Anatoly A Gitelson, Olga B Chivkunova, and Victor YU Rakin. Non-destructive optical detection of pigment changes during leaf senescence and fruit ripening. *Physiologia plantarum*, 106(1) :135–141, 1999.
- [11] Babacar Ndao, Louise Leroux, Abdoul Aziz Diouf, Valerie Soti, and Bienvenu Sambou. A remote sensing based approach for optimizing the sampling strategies in crop monitoring and crop yield estimation studies. In Souleye Wade, editor, *Earth Observations and Geospatial Science in Service of Sustainable Development Goals*, pages 25–36, Cham, 2019. Springer International Publishing.
- [12] A. Nyéki, C. Kerepesi, B. Daróczy, A. Benczúr, G. Milics, A.J. Kovács, and M. Neményi. *Maize yield prediction based on artificial intelligence using spatio-temporal data*, chapter 124, pages 1011–1017. Wageningen Academic Publishers, 2019.
- [13] J. Qi, A. Chehbouni, A.R. Huete, Y.H. Kerr, and S. Sooroshian. A modified soil adjusted vegetation index. *Remote Sensing of Environment*, 48(2) :119–126, 1994.
- [14] M. Stas, J. Van Orshoven, Q. Dong, S. Heremans, and B. Zhang. A comparison of machine learning algorithms for regional wheat yield prediction using ndvi time series of spot-vgt. In *2016 Fifth International Conference on Agro-Geoinformatics (Agro-Geoinformatics)*, pages 1–5, July 2016.
- [15] Anna X. Wang, Caelin Tran, Nikhil Desai, David Lobell, and Stefano Ermon. Deep transfer learning for crop yield prediction with remote sensing data. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies, COMPASS '18*, New York, NY, USA, 2018. Association for Computing Machinery.
- [16] Weicheng Wu. The generalized difference vegetation index (gdvi) for dryland characterization. *Remote Sensing*, 6(2) :1211–1233, 2014.

Détection d'anomalies à bord de satellites pour la protection de l'environnement marin

T. Goudemant¹, C. Szywala¹, B. Francesconi¹, M. Aubrun¹, E. Kervennic¹, I. Grenet^{1,2},
Y. Bobichon¹, M. Bellizzi¹, A. Girard¹

¹ Institut de Recherche Technologique Saint Exupéry

² MyDataModels

30 avril 2024

Résumé

Cet article présente un pipeline d'intelligence artificielle dédié à la détection d'anomalies dans des images maritimes, conçu pour être déployé à bord de satellites d'observation de la Terre. L'approche proposée est innovante en ce qu'elle requiert peu de données annotées, est capable de fonctionner sur des cibles embarquées à faible puissance de calcul, et est générique pour tous types d'anomalies se différenciant de l'état normal de l'eau. Ces qualités font de cette méthode un outil adaptable à diverses missions, ce qui lui a valu d'être respectivement sélectionnée et présélectionnée dans deux concours organisés par l'Agence Spatiale Européenne pour intégrer de l'intelligence artificielle à bord de Φ sat-2 et IMAGIN-e.

Mots-clés

Surveillance de l'environnement marin, intelligence artificielle embarquée, télédétection, détection d'anomalies, apprentissage frugal.

Abstract

This article describes a comprehensive artificial intelligence pipeline for detecting anomaly in maritime images, designed to be run on-board Earth observation satellites. The proposed method is innovative as it requires only weakly labelled data, can run on low-power embedded targets, and is generic for all types of anomalies deviating from the normal water state. These assets make this pipeline easily adaptable to a wide range of missions, resulting in its respective selection and preselection in two European Space Agency competitive challenges aimed at integrating artificial intelligence into Φ sat-2 and IMAGIN-e.

Keywords

Marine environmental monitoring, embedded artificial intelligence, remote sensing, anomaly detection, frugal learning.

1 Introduction

Les océans abritent une large diversité d'espèces marines et constituent une composante vitale de la biodiversité mondiale. Les technologies avancées, telles que l'imagerie satellite et l'intelligence artificielle (IA), peuvent aider les

scientifiques à identifier et à faire face aux menaces environnementales qui pèsent sur cet écosystème.

À titre d'exemple, la détection précoce de proliférations d'algues par des satellites joue un rôle essentiel dans l'atténuation des impacts écologiques, mais aussi dans la préservation des pêcheries qui constituent une source essentielle de nourriture pour des milliards de personnes dans le monde. Une étude [10] a estimé à 82 millions de dollars l'impact annuel moyen des proliférations d'algues nuisibles aux États-Unis. L'identification précoce de la pollution d'origine humaine, comme les marées noires ou les déchets plastiques, revêt également une importance cruciale : elle permet la mise en place d'opérations de nettoyage efficaces, réduit les pertes industrielles et contribue à la compréhension de ces menaces [13].

La plupart des recherches dans le domaine de la protection marine se concentrent sur des menaces spécifiques : évaluation de la qualité de l'eau [19, 23], concentrations de solides en suspension liées aux activités humaines [6] ou non [14], marées noires [2], déchets plastiques flottants [15]. Ces méthodes reposent généralement sur des calculs d'indices à l'aide de méthodes simples, comme combiner plusieurs bandes spectrales pour identifier des matériaux spécifiques. Mais elles rencontrent des difficultés lorsqu'elles sont appliquées à d'autres satellites ou d'autres zones géographiques. L'utilisation de méthodes statistiques ou d'apprentissage automatique classiques [26, 21, 25] font face à cette même difficulté d'adaptation à de nouvelles missions. À l'inverse, les techniques d'apprentissage profond ont produit des résultats de pointe ces dernières années. Les méthodes supervisées [16, 21, 25] offrent une bonne capacité d'adaptation à de nouvelles missions. Elles ne permettent cependant pas de généraliser à d'autres cas d'utilisation aussi bien que les méthodes auto-supervisées [26, 20, 25] et non supervisées basées sur la mesure de l'erreur de reconstruction d'image [26, 20, 21, 25]. L'utilisation de ces techniques d'apprentissage profond implique cependant dans la plupart des cas soit des coûts de calcul élevés, soit une annotation laborieuse des bases de données.

Contrairement à ces approches, nous avons choisi d'exploiter la technologie satellitaire pour détecter tous les événements inattendus, dont diverses menaces environnemen-

tales (sédiments, proliférations d'algues, débris plastiques, marées noires, etc.), y compris ceux qui ne pourraient pas être caractérisés. Nous désignerons dès lors l'ensemble de ces événements par le terme *anomalies*, qui englobe toutes sortes d'éléments se différenciant d'une normalité définie [9], représentée dans notre cas par un ensemble d'images maritimes sans perturbation des écosystèmes.

Notre approche doit également répondre aux contraintes établies par les défis OrbitalAI proposés par l'Agence Spatiale Européenne (*European Space Agency*, ESA). L'objectif de ces défis est de faire progresser l'*edge computing* en orbite en sélectionnant des propositions innovantes d'IA pour les déployer à bord de deux missions de télédétection :

- Le satellite Φ sat-2, un CubeSat 6U en orbite héliosynchrone équipé d'une caméra d'une résolution spatiale de 5 mètres et de 8 bandes spectrales dans le visible et le proche infrarouge. Φ sat-2 embarque un processeur CogniSAT™ d'Ubotica qui intègre un accélérateur Intel® Movidius™ Myriad™ 2.
- IMAGIN-e (*ISS Mounted Accessible Global Imaging Nod-e*), une démonstration de *space edge computing* par Microsoft et Thales Alenia Space. Il s'agit d'installer une caméra hyperspectrale et des capacités de calcul à bord de la Station Spatiale Internationale (*International Space Station*, ISS). Le système comporte une caméra ayant une résolution spatiale de 50 mètres et 50 bandes spectrales, et un processeur Cortex-A72 à 16 cœurs.

L'intérêt de l'*edge computing* est de réduire la réactivité des systèmes d'observation de la Terre en traitant les images directement sur les satellites, plutôt que d'attendre qu'elles soient transmises puis traitées dans les stations au sol. Cela nécessite de modifier des algorithmes conçus pour une infrastructure terrestre afin de les rendre légers et compatibles avec les cibles matérielles embarquées, qui sont limitées en puissance de calcul et en capacité mémoire.

Après un processus compétitif de plus d'un an, notre solution a été sélectionnée par l'ESA pour faire partie des deux nouvelles expériences qui seront embarquées à bord de Φ sat-2 (la compétition continue sur IMAGIN-e).

Notre solution est décrite dans la suite de l'article :

La section 2 détaille le fonctionnement des modèles composant le pipeline et justifie leur sélection en fonction de facteurs identifiés.

La section 3 décrit les données utilisées pour l'entraînement des modèles, ainsi que les étapes de prétraitement entreprises pour garantir la représentation fidèle des caractéristiques des capteurs.

La section 4 couvre l'entraînement, la validation et les résultats algorithmiques. Elle décrit également le déploiement sur la cible Myriad™ 2 et sur des CPU représentatifs des deux missions, ainsi que les performances matérielles obtenues.

2 Méthode

Dans cette section, nous présentons l'architecture de notre solution, puis justifions la sélection des modèles d'IA qui la

composent.

2.1 Description de l'architecture

L'architecture de notre solution, présentée sur la Figure 1, est constituée de trois parties : un *Tiler* pour découper l'image en patches de 32×32 pixels, un *Pipeline IA* pour détecter et caractériser les anomalies, et un *Mosaic'er* pour visualiser les sorties des modèles.

Le *Pipeline IA*, au cœur de notre solution, est composé de quatre modèles d'IA successifs :

1. Un modèle qui filtre uniquement les patches marins pour traitement ultérieur. Pour cette fonction, nous avons choisi d'utiliser un modèle de segmentation de la mer obtenu par l'algorithme génétique et frugal ZGP (*Zoetrope Genetic Programming*) [4].
2. Un modèle qui détecte les menaces potentielles sur les zones maritimes en mesurant la distance entre chaque zone et une normalité apprise. Ce modèle est constitué d'un réseau de neurones entraîné à l'aide de la méthode d'apprentissage auto-supervisé SimCLR [5], qui projette les patches dans un espace latent plus robuste et...
3. ... d'un *Gaussian Mixture Model* (GMM) [27] qui compare les caractéristiques des patches à une distribution définissant l'état dit *normal* de l'écosystème marin. Il génère un score d'anomalie allant de zéro à un pour chaque patch et conserve ceux avec un score significatif pour l'étape suivante.
4. Un autre modèle ZGP [4] qui caractérise précisément certains types d'anomalies en cas de dépassement d'un seuil. Il segmente les pixels des patches en trois classes : marée noire, prolifération d'algues et sédiments. Une classe *Autres* est également incluse pour les pixels normaux d'un patch, ou ceux présentant un type d'anomalie inconnu.

La description détaillée et la justification du choix de chaque modèle sont présentées dans la section 2.2.

Le *Mosaic'er*, quant à lui, recombine les patches de sortie des trois modèles et génère :

- une carte de segmentation terre/mer/nuages après le premier ZGP ;
- une carte de chaleur illustrant l'amplitude de l'anomalie pour chaque patch après le GMM ;
- une carte de segmentation du type d'anomalie au niveau pixel après le deuxième ZGP.

Cette solution modulaire offre une approche globale et peut être utilisée de manière polyvalente pour diverses applications opérationnelles, telles que :

- La priorisation d'images : les images dont les scores d'anomalie sont les plus élevés sont téléversées en priorité par le satellite afin de permettre aux opérateurs de se concentrer uniquement sur les images renfermant des informations cruciales lors de la surveillance de vastes étendues marines.
- L'envoi d'alertes : en cas de détection d'un incident majeur en mer, tel qu'une marée noire, une alerte est immédiatement déclenchée afin de favoriser une

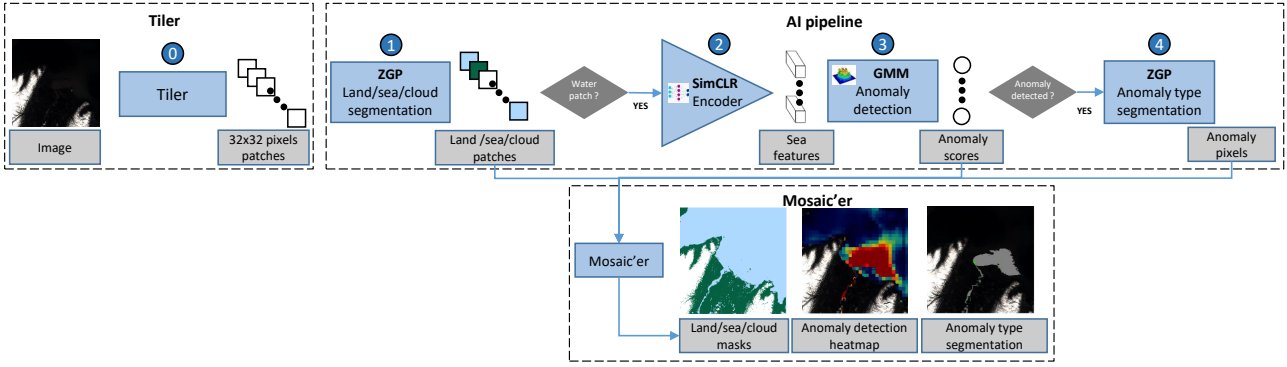


FIGURE 1 – Architecture du pipeline de détection d’anomalies. Elle est composée d’un *Tiler* (0) qui transforme l’image en patches. Les patches marins sont ensuite filtrés par le *Pipeline IA* (1), encodés dans un espace latent robuste (2), puis soumis à la détection d’anomalies (3). Les patches classés comme anomalies sont segmentés (4) afin de caractériser le type d’anomalie. Les sorties de chaque modèle peuvent être visualisées avec le *Mosaic’er* qui recombine les patches de sortie.

réaction rapide des autorités compétentes et d’atténuer les impacts environnementaux, sociaux et économiques.

2.2 Justification de l’architecture

Cette partie fournit des détails sur les différents modèles et justifie leur implantation à la lumière des limitations et objectifs décrits dans la section 1.

2.2.1 Segmentation marine et caractérisation des anomalies

Pour les étapes de segmentation marine et de caractérisation des anomalies, nous avons opté pour des modèles ZGP [4]. L’algorithme ZGP (*Zoetrope Genetic Programming*), développé par *MyDataModels*, intègre à la fois la programmation génétique et la régression symbolique. Son principe d’entraînement consiste à faire évoluer des individus représentés par des expressions mathématiques combinant des variables d’entrée et des constantes. Après des processus de mutation et de recombinaison sur plusieurs générations, le meilleur individu est sélectionné comme modèle final [4]. Les modèles obtenus sont des formules mathématiques simples combinant les variables d’entrée du modèle et des constantes numériques, avec des opérateurs mathématiques pouvant être algébriques, discrets, quadratiques ou trigonométriques.

Le choix de ZGP est motivé par ses nombreux avantages : frugalité en termes de données d’entraînement (environ mille pixels annotés par classe), rapidité d’inférence et facilité d’implantation dans les systèmes embarqués. En effet, sa mise en œuvre nécessite uniquement l’évaluation de la formule, et les deux modèles ZGP seront exécutés sur le processeur matériel en tant qu’étapes de prétraitement (pour la segmentation terre/mer/nuages) et de post-traitement (pour la caractérisation des anomalies).

2.2.2 Détection d’anomalies

Pour l’étape de détection d’anomalies, nous avons opté pour une méthode d’apprentissage profond auto-supervisé. Elle consiste à utiliser un encodeur pour extraire les caractéristiques de haut niveau d’une scène, combiné à un algorithme

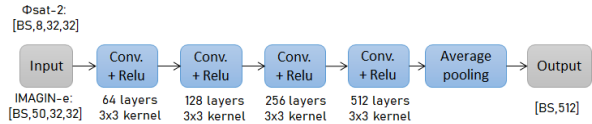


FIGURE 2 – Représentation de l’encodeur de patch composé de seulement quatre couches de convolution, et pouvant bénéficier d’une accélération matérielle. Les dimensions d’entrée et de sortie ainsi que la taille des couches apparaissent. BS (*Batch Size*) correspond à la taille des batches.

classique de détection d’anomalies. Ainsi, nous avons utilisé deux modèles pour cette étape : un encodeur SimCLR pour encoder chaque scène et un GMM pour leur attribuer un score d’anomalie.

L’approche SimCLR [5] est une méthode d’apprentissage auto-supervisé qui vise à représenter les patches de manière robuste en rapprochant les caractéristiques de patches similaires tout en éloignant celles de patches différents. C’est une méthode de référence en termes d’apprentissage de représentations, reconnue pour son efficacité avec des architectures de réseau de neurones petites et simples. Ainsi, notre choix s’est porté sur un petit encodeur, composé de seulement quatre couches de convolution (Figure 2). Ce modèle, avec une taille d’environ un million de paramètres, est environ dix fois plus petit que d’autres encodeurs légers courants tels que ResNet18 [8]. Une telle simplicité favorise l’adaptabilité à différentes missions en atténuant le risque de surapprentissage, tout en assurant la compatibilité avec divers accélérateurs matériels tels que l’Intel® Movidius™ Myriad™ 2 embarqué dans Φsat-2. De plus, son entraînement auto-supervisé permet une adaptation aisée à différents ensembles de données et ne requiert pas d’annotation.

Le GMM [27] est un algorithme probabiliste d’apprentissage automatique qui utilise une combinaison de gaussiennes pour modéliser la distribution des patches enco-

TABLE 1 – Caractéristiques des modèles du pipeline en termes de frugalité des données et d'embarquabilité.

Modèle	App.	Frugalité données	Type modèle	Exécution
Modèles ZGP	supervisé	1000 pixels annotés par classe	Formule math.	CPU pour pré-/post-traitement
Encodeur SimCLR	auto-supervisé	10 000 patches non annotés	Petit CNN (4 couches)	Accélération matérielle ou CPU
GMM	faiblement supervisé	1000 patches sans anomalie	Formule math.	CPU

dés. Pendant la phase d'apprentissage, les paramètres statistiques sont optimisés pour modéliser la distribution de patches représentant la normalité. Ensuite, lors de la phase de détection d'anomalies, les scores de vraisemblance produits par le GMM mesurent la déviation de chaque patch encodé par rapport à cette normalité. En fixant un seuil approprié, les anomalies peuvent être identifiées parmi les patches encodés. Le choix de cet algorithme est motivé par sa frugalité en termes de ressources informatiques et de données (environ mille patches d'entraînement avec pour seule annotation la classe *mer sans anomalie*). Notre approche, axée sur l'apprentissage de la distribution de la mer plutôt que sur celui des menaces, permet de détecter toutes les anomalies. En termes de déploiement, l'évaluation du modèle nécessite peu de ressources et peut être effectuée sur le CPU.

Le Tableau 1 récapitule les caractéristiques des modèles composant le pipeline. Nous avons choisi un modèle d'apprentissage profond auto-supervisé pour une bonne généralisation à différentes missions. L'encodeur et le GMM ont été sélectionnés pour la détection d'anomalies en raison de leurs faibles besoins en données étiquetées. Le modèle supervisé ZGP a été choisi pour sa capacité à caractériser précisément des anomalies en s'entraînant sur peu de données. Ajoutons que l'encodeur, qui est le modèle le plus gourmand en termes de capacités informatiques, a été réduit à quatre couches et que les autres modèles nécessitent très peu de ressources matérielles.

3 Base de données

Nous avons utilisé plusieurs ensembles de données d'images et d'annotations représentatives de Φ sat-2 et IMAGIN-e pour entraîner et tester notre pipeline. Nous les décrivons dans cette section.

3.1 Choix des données en fonction des exigences des modèles

Les modèles constituant notre pipeline nécessitent des ensembles de données adaptés à leurs spécificités d'entraînement respectives.

Pour l'encodeur qui requiert une importante quantité de

données non annotées représentatives des capteurs, nous avons choisi d'utiliser les ensembles de données Φ sat-2 et IMAGIN-e fournis par l'ESA dans le cadre du défi OrbitalAI, et désignés dans la suite de l'article sous le nom [DS_challenge] (section 3.2).

Pour le GMM, nous avons constitué un sous-ensemble de [DS_challenge] composé uniquement d'images de mer sans anomalies afin de l'entraîner à assimiler le concept de *normalité marine*. Puis, nous avons construit un ensemble de données contenant des anomalies, désigné sous le nom [DS_custom], qui a été créé en simulant des images Φ sat-2 et IMAGIN-e à partir de dates d'événements connus (section 3.3). Un sous-ensemble de ces images a été manuellement annoté pour une évaluation quantitative (section 4.2.2), tandis que les autres ont été utilisées pour une évaluation visuelle (section 4.2.1).

Le premier ZGP est entraîné sur des images annotées terre/mer/nuages afin de filtrer les patches marins, tandis que le deuxième ZGP est entraîné sur des images annotées d'anomalies afin de les caractériser. Pour entraîner et tester les ZGP, nous avons respectivement annoté manuellement un sous-ensemble de [DS_challenge] et utilisé un sous-ensemble de [DS_custom].

Des détails supplémentaires sur ces ensembles de données sont fournis dans ce qui suit.

3.2 DS_challenge : jeux de données fournis dans le cadre du défi OrbitalAI

Deux ensembles de données représentatifs de Φ sat-2 et d'IMAGIN-e ont été fournis dans le cadre des défis OrbitalAI de l'ESA. Ils comprennent 489 scènes et masques nuageux simulés à partir de produits Sentinel-2 L1C multispectraux convertis en unités de radiance en haut de l'atmosphère (*Top Of Atmosphere*, TOA). Dans le cas d'IMAGIN-e, 497 scènes supplémentaires ont également été simulées et fournies par l'ESA à partir de produits Level-1 PRISMA hyperspectraux. Nous n'avons pas utilisé ces dernières dans notre étude car le délai pour obtenir les produits PRISMA supplémentaires dont nous aurions eu besoin pour annoter des anomalies était incompatible avec le calendrier serré du défi. Chaque scène est divisée en patches de 256×256 pixels (sans chevauchement), ce qui permet d'obtenir respectivement 256 et 16 patches par scène avec 8 et 50 bandes spectrales VIS/NIR pour Φ sat-2 et IMAGIN-e.

Un sous-ensemble de [DS_challenge], comprenant environ 60 scènes, a été annoté manuellement avec des masques de segmentation terre/mer afin d'entraîner et de tester le premier ZGP dédié au filtrage des scènes marines. Un autre sous-ensemble, comprenant environ 70 patches, a été classifié manuellement en catégories telles que végétation, sol nu, neige, nuage et eau afin d'évaluer les capacités de représentation de l'encodeur. Enfin, un sous-ensemble de patches marins a été sélectionné pour définir l'état normal de la mer en prévision de l'entraînement du GMM.

Remarquons que pour entraîner l'encodeur et le GMM, les patches sont subdivisés en patches plus petits de 32×32 pixels. Nous obtenons ainsi environ 8 millions de patches pour Φ sat-2 et 500 000 pour IMAGIN-e.

3.3 DS_custom : jeu de données d’anomalies personnalisé

Des données supplémentaires ont été recueillies pour évaluer la performance du détecteur d’anomalies et entraîner le ZGP dédié à leur caractérisation. Ces données comprennent des événements de pollution extraits de la littérature scientifique, de contributions d’experts et de communiqués de presse : marées noires [12], proliférations d’algues [7, 11], sédiments et déchets plastiques [3, 24, 17]. Ces événements sont répertoriés dans le Tableau 2, accompagnés des coordonnées géographiques et des dates utilisés pour acquérir les images Sentinel-2 correspondantes.

Les images représentatives des capteurs de Φ sat-2 et IMAGIN-e ont été générées par simulation à partir de ces scènes Sentinel-2 (section 3.4). Les anomalies ont été annotées manuellement par des experts en traitement d’images, avec des masques de segmentation au niveau des pixels pour chaque type d’anomalie ainsi que pour les zones terrestres et marines. Les masques nuageux proviennent de Sentinel-2 et ont été récupérés par le simulateur.

Comme il était impossible de segmenter manuellement les images contenant des anomalies larges et diffuses telles que des proliférations d’algues, celles-ci ont été conservées dans un ensemble distinct dédié à l’évaluation visuelle des performances des modèles.

3.4 Présentation des simulateurs Φ sat-2 et IMAGIN-e

Des *notebooks* Python utilisant la bibliothèque EO-learn ainsi que des exécutables spécifiques ont été fournis dans le cadre d’OrbitalAI pour simuler des images représentatives de Φ sat-2 et IMAGIN-e :

- simulation d’images Φ sat-2 à partir de données Sentinel-2 L1C;
- simulation d’images IMAGIN-e à partir de données Sentinel-2 L1C ou de données PRISMA.

Ces *notebooks* permettent de définir une liste de demandes et de récupérer les données Sentinel-2 correspondantes rééchantillonnées depuis SentinelHub (avec les identifiants appropriés).

Les simulateurs permettent ensuite de produire des données de niveau 1A (radiances TOA sans correction des décalages spatiaux entre les bandes spectrales) ou de niveau 1C (réflectance TOA avec les bandes spectrales coréregistrées). Nous avons choisi d’entraîner les modèles de notre pipeline exclusivement sur des images L1C en raison du risque d’erreurs que pourrait générer la dérégistration entre bandes spectrales. À l’avenir, nous pourrions envisager d’adapter le pipeline à des données L1A.

L’architecture des simulateurs Φ sat-2 et IMAGIN-e est similaire, notamment avec les produits Sentinel-2. Le processus de traitement est illustré sur la Figure 3, qui met en évidence les étapes spécifiques à Φ sat-2 et à IMAGIN-e. À la fin du processus, l’image simulée peut être découpée selon les besoins des algorithmes d’apprentissage, et est sauvegardée au format Geotiff, incluant toutes les bandes et certains masques (nuages provenant des données Sentinel-2).

TABLE 2 – Dates, positions et types d’anomalies marines utilisées pour simuler les images de Φ sat-2 et IMAGIN-e à partir de celles de Sentinel-2.

ID	Date	Latitude	Longitude	Anomalie
AN1	07/08/17	20,8399	37,6159	Marée noire
AN2	26/02/23	11,2167	57,7782	Marée noire
AN3	07/08/17	20,8452	37,6104	Marée noire
AN4	27/08/17	20,8511	37,5657	Marée noire
AN5	08/09/17	-0,7802	37,7166	Algues
AN6	29/07/17	114,0258	22,256	Algues
AN7	29/07/17	114,0961	22,1866	Algues
AN8	14/05/18	-14,2314	65,6917	Sédiments
AN9	08/08/22	21,632	59,458	Algues
AN10	08/08/22	22,476	59,601	Algues
AN11	11/06/21	26,5247	39,0395	Plastique
AN12	21/06/21	26,5247	39,0395	Plastique
AN13	31/10/18	0,1420	5,6920	Plastique
AN14	31/10/18	0,0435	5,6154	Plastique
AN15	01/03/23	17,3799	55,7791	Marée noire
AN16	10/08/21	26,5247	39,0395	Plastique
AN17	27/09/21	28,152	121,216	Algues
AN18	08/03/21	1,75	105,17	Marée noire
AN19	02/04/21	1,75	105,17	Marée noire
AN20	18/03/21	2,5229	106,0052	Marée noire
AN21	02/04/21	2,0914	105,1947	Marée noire
AN22	02/05/21	2,4545	105,1541	Marée noire
AN23	10/11/20	17,4875	-88,175	Sédiments
AN24	15/11/20	17,4875	-88,175	Sédiments
AN25	30/11/20	17,4875	-88,175	Sédiments
AN26	28/08/21	35,1854	35,8728	Marée noire
AN27	02/09/21	35,1854	35,8728	Marée noire
AN28	07/09/21	35,3942	35,8363	Marée noire
AN29	07/02/19	5,6147	95,2998	Sédiments
AN30	09/02/19	5,6004	95,3226	Sédiments
AN31	18/01/22	-11,9581	-77,2195	Marée noire

4 Expériences menées

Cette section détaille l’entraînement des modèles, puis présente les résultats de détection pour diverses anomalies (plastiques, marées noires, proliférations d’algues, sédiments) sur deux missions (Φ sat-2 et IMAGIN-e) afin de mettre en avant la polyvalence de notre approche. Les performances de déploiement des modèles sur deux processeurs embarqués, ainsi que celles de l’encodeur sur l’accélérateur matériel de Φ sat-2, seront également présentées.

4.1 Paramètres d’entraînement des modèles

Pour entraîner le premier et le second ZGP, nous avons sélectionné de manière aléatoire 1000 pixels par classe respectivement dans [DS_challenge] et dans [DS_custom]

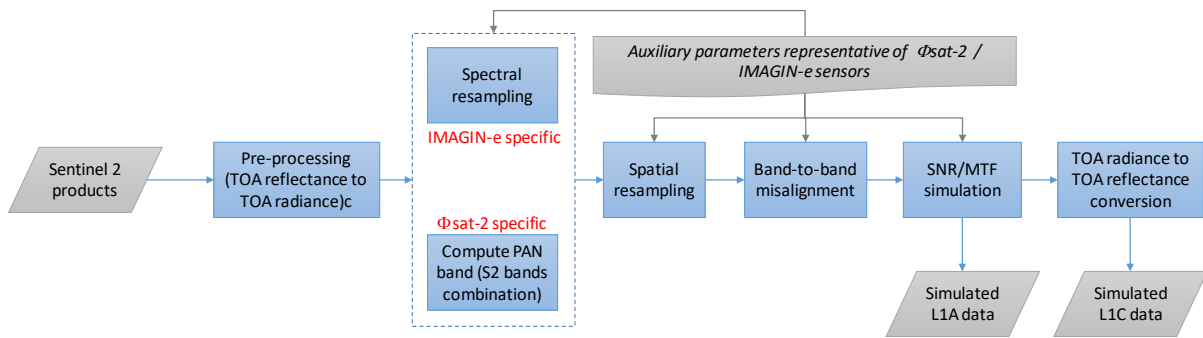


FIGURE 3 – Architecture des simulateurs IMAGIN-e et Φ sat-2, avec les spécificités des capteurs mises en évidence (rééchantillonnage spectral afin de correspondre au capteur hyperspectral d’IMAGIN-e, calcul de la bande panchromatique à partir des bandes multispectrales pour Φ sat-2). Plusieurs paramètres sont simulés lors des différentes étapes : résolution spatiale (50 m pour IMAGIN-e, 5 m pour Φ sat-2), décalages entre les bandes spectrales liés à l’acquisition, rapport signal sur bruit (*Signal-to-Noise Ratio*, SNR) et fonction de transfert de modulation (*Modulation Transfer Function*, MTF) des capteurs. Les simulateurs permettent de choisir le niveau du produit, avec des différences notables entre L1A et L1C (décalage entre les bandes plus important pour des données L1A car non corrigé, nécessité de convertir les données de radiance TOA en données de réflectance TOA si l’on souhaite simuler des données L1C).

(section 3). Cette sélection s’est faite en séparant méticuleusement les données d’entraînement et de test pour assurer l’indépendance entre ces deux phases. Les modèles sont ensuite entraînés de manière supervisée en utilisant l’algorithme de *MyDataModels* [4].

L’entraînement de l’encodeur s’inspire quant à lui d’une implantation PyTorch [22] qui vise à minimiser la perte d’*entropie croisée à température normalisée* [5] de l’architecture représentée sur la Figure 2. Les données d’entraînement comprennent 10 000 patches non étiquetés de 32×32 pixels issus de diverses scènes de [DS_challenge], en utilisant toutes les bandes spectrales. Les modèles ont été entraînés sur 80 époques, avec une taille de batch de 16 patches, un coefficient de température de 50 % et un taux d’apprentissage de 0,0003, en démarrant avec des poids aléatoires. Ce modèle prend environ une heure pour converger sur un processeur Intel® Xeon® Silver 4114 à quatre cœurs. L’architecture du modèle génère en sortie des vecteurs de 512 dimensions correspondant aux représentations encodées des patches dans un espace latent.

Enfin, le GMM a été entraîné sur 7 000 patches sans anomalies issus d’un sous-ensemble de [DS_challenge] et [DS_custom]. L’entraînement consiste à optimiser les caractéristiques de dix composantes gaussiennes de façon à ajuster leur distribution à celle des patches via un algorithme EM (*Expectation-Maximization*) implémenté à partir de la bibliothèque scikit-learn [18]. Le score d’anomalie fourni pour chaque patch de test est calculé à partir de la densité de la composante gaussienne la plus probable.

Les modèles entraînés pour les deux défis seront disponibles sur [1].

4.2 Résultats algorithmiques

Nous présentons dans cette partie une validation visuelle des sorties de notre solution pour chacune des deux missions (section 4.2.1), suivie d’une évaluation quantitative

de la performance algorithmique sur quatre images d’anomalies (4.2.2).

4.2.1 Évaluation qualitative

Pour valider nos modèles, nous analysons leur sortie avec le *Mosaic’er* (section 2.1). La Figure 4 montre les résultats obtenus sur deux images pour les missions Φ sat-2 et IMAGIN-e. Dans les deux cas, le premier ZGP segmente correctement la mer, les patches marins sont analysés par l’encodeur puis le GMM détecte bien l’anomalie, qui est ensuite classifiée correctement (prolifération d’algues) par le deuxième ZGP. Pour confirmer l’absence de faux positifs, ce processus d’évaluation visuelle a été, pour chaque mission, reproduit sur des images supplémentaires sans anomalie.

4.2.2 Évaluation quantitative

Nous avons évalué quantitativement les performances des ZGP et du GMM en utilisant des métriques classiques telles que la précision, le rappel, le *F1-score* et l’aire sous la courbe précision-rappel (AUROC). Ces métriques ont été calculées en comparant les sorties des modèles à des vérités terrain annotées manuellement. Nous avons utilisé deux images par mission, totalisant 35,6 millions de pixels ou 34,8 milliers de patches (32×32 pixels par patch). Les performances algorithmiques sont regroupées dans le Tableau 3, et la Figure 5 montre la visualisation des résultats des modèles pour les quatre images.

Pour évaluer les performances du premier ZGP, nous nous sommes concentrés exclusivement sur la segmentation des pixels marins, puisque l’objectif de ce modèle est d’envoyer uniquement des patches de mer à l’encodeur. Cette tâche affiche des performances très élevées, avec des scores de précision et de rappel dépassant en général 97 %.

Nous avons ensuite évalué la performance en détection du GMM indépendamment des résultats du modèle précédent. Pour ce faire, nous l’avons fait tourner sur des patches dont

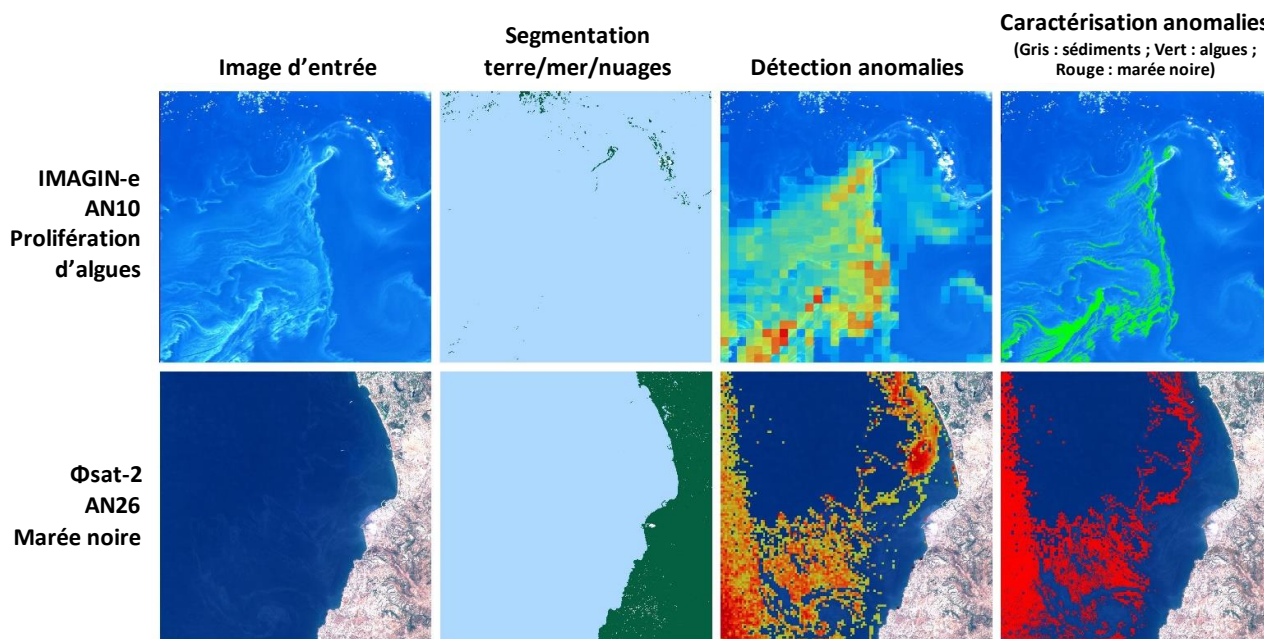


FIGURE 4 – Validation visuelle des sorties des ZGP et du GMM pour IMAGIN-e sur AN10 (en haut) et pour Φsat-2 sur AN26 (en bas) (cf. Tableau 2). Dans l'ordre : visualisation RGB de l'image d'entrée, segmentation terre/mer/nuages par ZGP, carte de chaleur des anomalies par GMM (rouge : score proche de 1 ; bleu : score proche de 0), et caractérisation des anomalies par ZGP. La prédiction du modèle est superposée à l'image originale pour la détection et la caractérisation.

TABLE 3 – Performances algorithmiques des modèles sur les quatre images de test (Figure 5). V.T. = Vérité Terrain.

Image	V.T.	Préc.	Rap.	F1-score	AUROC
Segmentation mer, ZGP (V.T. = pixels mer)					
#1 AN13	11,1M	96,6 %	99,9 %	98,2 %	–
#2 AN5	6,1M	99,1 %	83,1 %	90,4 %	–
#3 AN23	558K	98,8 %	96,4 %	97,5 %	–
#4 AN27	1042K	97,7 %	97,9 %	97,8 %	–
moy.	33,3M	98,1 %	94,3 %	96,0 %	–
Détection anomalies, GMM (V.T. = patches anom.)					
#1 AN13	17	47,8 %	64,7 %	55,0 %	44,2 %
#2 AN5	4501	74,8 %	99,8 %	85,5 %	87,4 %
#3 AN23	71	50,0 %	61,9 %	55,3 %	60,8 %
#4 AN27	33	37,7 %	69,7 %	48,9 %	38,6 %
moy.	4622	52,6 %	74,0 %	61,2 %	57,7 %
Caractérisation anomalies, ZGP (V.T. = pixels anom.)					
#1 AN13	2K	100 %	99,7 %	99,8 %	–
#2 AN5	4M	100 %	99,8 %	99,9 %	–
#3 AN23	37K	100 %	97,3 %	98,6 %	–
#4 AN27	6K	100 %	96,5 %	98,2 %	–
moy.	4M	100 %	98,3 %	99,1 %	–

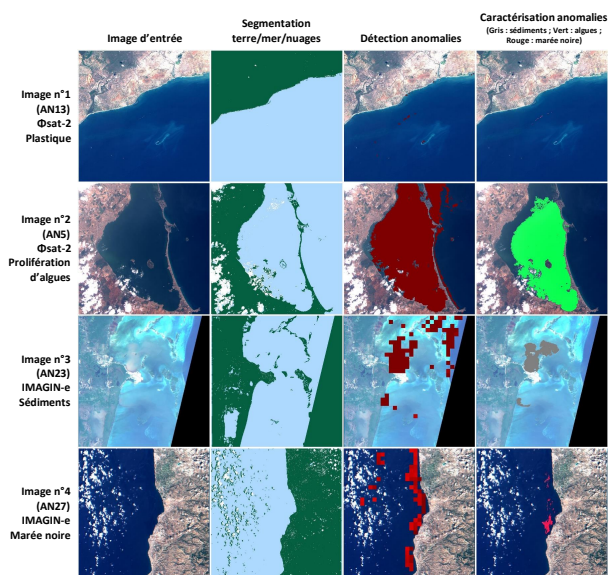


FIGURE 5 – Exemples de prédictions sur quatre images de test (AN13, AN5, AN23, AN27, cf. Tableau 2) pour Φsat-2 (lignes 1 et 2) et IMAGIN-e (lignes 3 et 4). Colonnes 1 : images ; Colonnes 2 à 4 : sorties reconstruites des modèles, avec une visualisation similaire à celle de la Figure 4.

la vérité terrain contenait plus de 80 % de mer. Pour classifier un patch comme anomalie, nous avons sélectionné des

seuils de détection optimisant la performance du modèle. La performance moyenne du modèle tourne autour de 60 % en termes de rappel et de précision. Cette performance est

dégradée par le fait que le critère de 80 % pour l'identification de la mer conduit à inclure des patches comprenant des lignes côtières qui génèrent des faux positifs et réduisent la précision. D'autre part, la prédiction des anomalies au niveau patch, comparée à l'annotation au niveau pixel, entraîne simultanément la diminution de la précision et du rappel — en particulier pour de petits objets tels que les marées noires. Notons que les cas d'utilisation présentés dans la section 1, et notamment la priorisation du téléversement des images, ne nécessitent ni une précision ni un rappel très élevé. Pour une mesure de performance indépendante du seuil de détection choisi, nous fournissons également dans le Tableau 3 les résultats de l'aire sous la courbe précision-rappel (AUROC).

Nous avons également évalué le second ZGP indépendamment des résultats du modèle précédent : nous avons effectué la caractérisation des anomalies sur des pixels annotés comme tels dans la vérité terrain, plutôt qu'en se basant sur les prédictions du GMM. Ceci a conduit à des performances élevées, atteignant près de 100 % en précision et en rappel. Ces résultats sont illustrés sur la Figure 5.

4.3 Performances matérielles

Dans cette dernière partie, nous présentons le déploiement de nos modèles sur des cibles matérielles représentatives des deux missions. Pour Φ sat-2, nous avons déployé l'encodeur de patch sur la cible Myriad™ 2, et les autres modèles ZGP / GMM sur un processeur ARM Cortex-A53. Pour IMAGIN-e, qui ne dispose pas d'accélérateur matériel embarqué, nous avons utilisé un CPU ARM Cortex-A72 pour le déploiement et l'évaluation des performances de l'ensemble des modèles.

4.3.1 Déploiement sur cibles matérielles

Déploiement sur Φ sat-2. La plateforme d'exécution du satellite Φ sat-2 est le Ubotica CogniSAT-XE1™. Il est construit à partir de l'accélérateur Intel® Myriad™ 2, conçu pour l'inférence de réseaux de neurones sur des cibles embarquées. À des fins de validation, nous avons utilisé le Movidius Neural Compute Stick 1 (NCS1), qui abrite l'unité de traitement de vision Myriad™ 2 (VPU). Myriad™ 2 comprend douze cœurs SHAVE (cœurs de traitement vectoriel sur 128 bits) pour les calculs des réseaux de neurones et prend en charge les arithmétiques 16 et 32 bits en virgule flottante. Avec 2 Mo de mémoire intégrée, elle peut être configurée pour prendre en charge différentes combinaisons d'instructions et de données en fonction de la charge de travail. Le déploiement sur Myriad™ 2 utilise l'outil logiciel OpenVINO 2020.3, qui permet de charger un modèle ONNX et de le convertir au format du moteur d'inférence avec le *Model Optimizer*. Nous avons validé le déploiement de l'encodeur sur la cible en comparant, pour plusieurs patches d'entrée, la cohérence des sorties (patches encodés) avant et après le déploiement. Les ZGP et le GMM ont quant à eux été déployés sur le processeur ARM Cortex-A53 de la carte d'évaluation Xilinx Zynq® UltraScale+™ MPSoC ZCU104. Un code Python instrumenté a été utilisé pour exécuter les trois modèles,

TABLE 4 – Performances matérielles des modèles.

Modèle	Cible	Données traitées	Latence (ms)	Débit par patch
ZGP (seg. terre/ mer/nuages)	ARM A53	1 patch (32 × 32 pixels)	11,3	90,3 Kpixels/s
	ARM A72		4,6	224,9 Kpixels/s
Encodeur (taille batch = 128)	Myriad™ 2	1 patch (32 × 32 pixels)	16,6	61,5 Kpixels/s
	ARM A72		57,4	17,8 Kpixels/s
GMM (détection anomalies)	ARM A53	1000 patches encodés	29,3	34,5 Kpatches/s
	ARM A72		15,7	67,1 Kpatches/s
ZGP (caract. anomalies)	ARM A53	1 patch (32 × 32 pixels)	13,1	77,8 Kpixels/s
	ARM A72		5,9	174,4 Kpixels/s

mesurant simultanément leurs temps d'exécution moyens. Nous avons également confirmé que les sorties de ces trois modèles étaient identiques avant et après le déploiement.

Déploiement sur IMAGIN-e. La cible matérielle à bord de l'ISS comporte un processeur ARM Cortex-A72 à 16 cœurs et dispose de 32 Go de mémoire SDRAM DDR4. Les deux ZGP, l'encodeur et le GMM ont été déployés sur le processeur ARM Cortex-A72 bicœur de la carte d'évaluation Xilinx Versal™ VCK190. Le même code Python instrumenté a été utilisé pour exécuter les quatre modèles, mesurant simultanément leurs temps d'exécution moyens. Nous avons confirmé que toutes les sorties étaient identiques avant et après le déploiement.

4.3.2 Métriques d'inférence

Pour évaluer les performances matérielles de nos modèles, nous avons mesuré le temps d'exécution moyen (latence) et le débit par patch pour chaque mission (Tableau 4).

Pour les ZGP exécutés sur le CPU, le débit est d'environ 90 Kpixels/s sur ARM A53 et 225 Kpixels/s sur ARM A72.

Concernant l'encodeur de patch exécuté sur Myriad™ 2, nous avons étudié l'influence de la taille du batch sur la latence et le débit. Les résultats, représentés sur la Figure 6, indiquent une utilisation optimale de la cible pour des batches de 128 patches. Au-delà, la latence et le débit restent constants (respectivement 16,6 ms/patch et 61,5 Kpixels/s) indépendamment de l'arithmétique choisie (virgule flottante de 16 ou 32 bits). Cependant, sans accélération matérielle, la latence chute à 18 Kpixels/s sur ARM A72.

Le GMM présente une remarquable capacité de débit, traitant jusqu'à 34 Kpatches encodés par seconde pour ARM A53 et 67 Kpatches pour ARM A72. Remarquons que ce modèle fonctionne sur des patches encodés de 32 × 32 pixels en entrée de l'encodeur. Par conséquent, ce débit

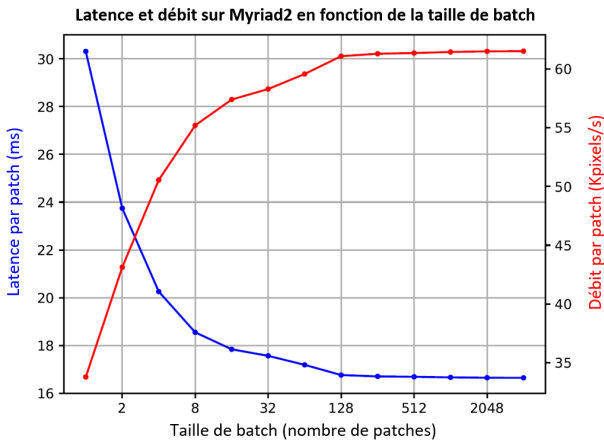


FIGURE 6 – Latence et débit de l’encodeur en fonction de la taille de batch. Optimum : batchs de 128 patches.

dépasse significativement celui des ZGP.

Remarquons que les exigences en débit ne sont pas uniformes tout au long de la chaîne de traitement : le premier ZGP doit classifier tous les pixels, l’encodeur et le GMM ne traitent que les patches marins, et le dernier ZGP ne traite que les patches marins identifiés comme des anomalies. De plus, le GMM opère sur des patches encodés de 32×32 pixels.

Ainsi, le goulot d’étranglement de la chaîne est soit le premier ZGP, soit l’encodeur, selon la proportion de patches marins dans l’image. Pour accélérer le premier ZGP, prédire seulement deux classes (mer/non-mer) au lieu de trois (terre/mer/nuages) pourrait réduire le temps d’inférence de 33 %. Il pourrait également être envisagé de le déployer sur un accélérateur matériel tel que Myriad™ 2. Enfin, remarquons que si la cible matérielle dispose d’un CPU multi-cœur, toute la chaîne peut être exécutée simultanément.

5 Conclusion & Perspectives

Cet article présente une solution originale combinant un réseau de neurones auto-supervisé avec des modèles frugaux pour la détection d’anomalies dans des images satellites. En se concentrant sur l’apprentissage d’une distribution de la normalité plutôt que sur des menaces spécifiques, cette approche améliore la généralité de la détection, permettant l’identification d’une grande variété d’événements. Nous avons illustré cette polyvalence en traitant trois types d’événements qui affectent les écosystèmes marins : les proliférations d’algues, les marées noires et les inondations de sédiments.

Nous avons démontré l’efficacité de notre approche et son adaptabilité en ayant été respectivement sélectionnés et pré-sélectionnés sur les deux défis de l’Agence Spatiale Européenne pour embarquer de l’intelligence artificielle à bord de Φ sat-2 et IMAGIN-e. La validation sur des données représentatives des missions a montré de bonnes performances en détection et caractérisation des anomalies. De plus, le déploiement sur l’accélérateur matériel Intel® My-

riad™ 2 et les processeurs ARM Cortex-A53/A72 a permis d’évaluer les performances de latence et de débit de chaque modèle et de déterminer les facteurs limitants.

À l’avenir, nous prévoyons d’améliorer notre solution en intégrant davantage de données provenant de différentes sources. L’objectif sera d’améliorer le champ de la normalité marine et d’affiner la caractérisation des performances algorithmiques. De plus, nous envisageons de tester la sensibilité de notre solution à différents phénomènes provenant des capteurs tels que le bruit ou le décalage entre bandes spectrales. Notre objectif ultime demeure de démontrer le potentiel et la réactivité de notre solution grâce à son déploiement opérationnel sur Φ sat-2. Enfin, nous espérons également faire partie des deux équipes qui pourront déployer leur solution sur la mission IMAGIN-e.

Remerciements

Le travail présenté dans cet article est mené dans le cadre du projet IRMA (IA pour la Réactivité Mission basée sur l’Analyse image) de l’IRT Saint Exupéry. Ce projet vise à développer des briques technologiques basées sur l’intelligence artificielle pour la planification de missions et le traitement des données, à bord ou au sol, avec l’objectif de répondre aux nouveaux besoins du marché de l’observation de la Terre par satellites, tels que la transformation de données brutes en informations interprétées, et la gestion réactive de constellations de satellites. IRMA s’est également donné pour objectif d’illustrer l’efficacité de ces technologies au travers de démonstrations de boucle système, que ce soit en laboratoire ou en orbite.

Les auteurs expriment leur gratitude envers les partenaires industriels et académiques du projet : Thales Alenia Space, Activeeon, MyDataModels, Geo4i, JoliBrain, et l’Université Côte d’Azur. Nous remercions également l’Agence Spatiale Européenne qui nous a permis de participer aux défis OrbitalAI.

Références

- [1] Marine anomaly detection presentation. <https://www.irt-saintexupery.com/fr/marine-anomaly-detection/>.
- [2] Rami Al-Ruzouq, Mohamed Barakat A Gibril, Abdallah Shanableh, Abubakir Kais, Osman Hamed, Saeed Al-Mansoori, and Mohamad Ali Khalil. Sensors, features, and machine learning for oil spill detection and monitoring : A review. *Remote Sensing*, 12(20) :3338, 2020.
- [3] Lauren Biermann, Daniel Clewley, Victor Martinez-Vicente, and Konstantinos Topouzelis. Finding plastic patches in coastal waters using optical satellite data. *Scientific Reports*, 10 :5364, 04 2020.
- [4] Aurélie Boisbunon, Carlo Fanara, Ingrid Grenet, Jonathan Daeden, Alexis Vighi, and Marc Schoenauer. Zoetrope genetic programming for regression. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 776–784, 2021.

- [5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.
- [6] Hongtao Duan, Zhigang Cao, Ming Shen, Dong Liu, and Qitao Xiao. Detection of illicit sand mining and the associated environmental effects in china's fourth largest freshwater lake using daytime and nighttime satellite images. *Science of the total environment*, 647 :606–618, 2019.
- [7] Diego Gómez, Pablo Salvador, Julia Sanz, and Jose Luis Casanova. A new approach to monitor water quality in the menor sea (spain) using satellite data and machine learning methods. *Environmental pollution*, 286 :117489, 2021.
- [8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.
- [9] Xing Hu, Chun Xie, Zhe Fan, Qianqian Duan, Dawei Zhang, Linhua Jiang, Xian Wei, Danfeng Hong, Guoqiang Li, Xinhua Zeng, Wenming Chen, Dongfang Wu, and Jocelyn Chanussot. Hyperspectral anomaly detection using deep learning : A review. *Remote Sensing*, 14(9), 2022.
- [10] EB Jewett, CB Lopez, Q Dortch, SM Etheridge, and LC Backer. Harmful algal bloom management and response : Assessment and plan. 2008.
- [11] Morteza Khalili and Mahdi Hasanlou. Harmful algal blooms monitoring using sentinel-2 satellite images. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2019.
- [12] Polychronis Kolokoussis and Vassilia Karathanassi. Oil spill detection and mapping using sentinel 2 imagery. *Journal of Marine Science and Engineering*, 6 :4, 2018.
- [13] Christos A Kontovas, Harilaos N Psaraftis, and Nikolaos P Ventikos. An empirical analysis of iopcf oil spill cost data. *Marine pollution bulletin*, 60(9) :1455–1466, 2010.
- [14] Huizeng Liu, Qingquan Li, Tiezhu Shi, Shuibo Hu, Guofeng Wu, and Qiming Zhou. Application of sentinel 2 msi images to retrieve suspended particulate matter concentrations in poyang lake. *Remote Sensing*, 9(7) :761, 2017.
- [15] Nikolai Maximenko, J Arvesen, G Asner, J Carlton, M Castrence, Luca Centurioni, Yi Chao, John Chapman, Ved Chirayath, P Corradi, et al. Remote sensing of marine debris to study dynamics, balances and trends. *White Paper, Decadal Survey for Earth Science and Applications from Space*, 22, 2016.
- [16] J. Mifdal, N. Longépé, and M. Rußwurm. Towards detecting floating objects on a global scale with learned spatial features using sentinel 2. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, V-3-2021 :285–293, 2021.
- [17] Dimitris Papageorgiou, Konstantinos N. Topouzelis, Giuseppe Suaria, Stefano Aliani, and Paolo Corradi. Sentinel-2 detection of floating marine litter targets with partial spectral unmixing and spectral comparison with other floating materials (plastic litter project 2021). *Remote. Sens.*, 14 :5997, 2022.
- [18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, 12 :2825–2830, 2011.
- [19] Miguel Potes, Gonçalo Rodrigues, Alexandra Marchã Penha, Maria Helena Novais, Maria João Costa, Rui Salgado, and Maria Manuela Morais. Use of sentinel 2–msi for water quality monitoring at alqueva reservoir, portugal. *Proceedings of the International Association of Hydrological Sciences*, 380 :73–79, 2018.
- [20] Oliver Rippel and Dorit Merhof. Anomaly detection for automated visual inspection : A review. *Bildverarbeitung in der Automation : Ausgewählte Beiträge des Jahreskolloquiums BVAu 2022*, pages 1–13, 2023.
- [21] Hongjun Su, Zhaoyue Wu, Huihui Zhang, and Qian Du. Hyperspectral anomaly detection : A survey. *IEEE Geoscience and Remote Sensing Magazine*, 10(1) :64–90, 2021.
- [22] thunderInfy. Simclr pytorch implementation. <https://github.com/thunderInfy/simclr>, 2013.
- [23] Kaire Toming, Tiit Kutser, Alo Laas, Margot Sepp, Birgot Paavel, and Tiina Nõges. First experiences in mapping lake water quality parameters with sentinel-2 msi imagery. *Remote Sensing*, 8(8) :640, 2016.
- [24] Konstantinos Topouzelis, Apostolos Papakonstantinou, and Shungudzemwoyo P. Garaba. Detection of floating plastics from satellite and unmanned aerial systems (plastic litter project 2018). *International Journal of Applied Earth Observation and Geoinformation*, 79 :175–183, 2019.
- [25] Yichu Xu, Lefei Zhang, Bo Du, and Liangpei Zhang. Hyperspectral anomaly detection based on machine learning : An overview. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15 :3351–3364, 2022.
- [26] Jie Yang, Ruijie Xu, Zhiquan Qi, and Yong Shi. Visual anomaly detection for images : A survey. *arXiv preprint arXiv :2109.13157*, 2021.
- [27] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection : A survey. *arXiv preprint arXiv :2110.11334*, 2021.

Session 1.2 : Traitement automatique des langues

Derby LLM : Évaluation comparative des approches RAG et fine-tuning

Christophe Bouvard¹, Mathieu Ciancone¹, Antoine Gourru², Marion Schaeffer¹

¹ Wikit, Lyon, France

² Laboratoire Hubert Curien, UMR CNRS 5516, Saint-Etienne, France

{christophe, mathieu, marion}@wikit.ai, antoine.gourru@univ-st-etienne.fr

Résumé

Les grands modèles de langage ont récemment été largement exploités dans les agents conversationnels, où l'injection de connaissances pour des domaines d'applications spécifiques est un enjeu crucial. Nous comparons deux approches : le fine-tuning et la génération augmentée de récupération. Nous évaluons ces techniques pour deux cas d'usage différents avec des métriques automatiques et la préférence humaine. Bien que la pertinence des réponses soit proche, la fidélité et la préférence humaine avantagent la génération augmentée de récupération.

Mots-clés

LLM, RAG, fine-tuning, benchmark, évaluation.

Abstract

Large language models have recently been widely exploited in conversational agents, where knowledge injection for specific application domains is a crucial issue. We compare two approaches : fine-tuning and retrieval augmented generation. We evaluate these techniques in two different use cases leveraging automatic metrics and human preference. Although the answer relevance is close, faithfulness and human preference favour retrieval augmented generation.

Keywords

LLM, RAG, fine-tuning, benchmark, evaluation.

1 Introduction

Les grands modèles de langage (*Large Language Models*, LLM) sont aujourd'hui utilisés dans diverses applications et projets de recherche. Leur pré-entraînement leur permet d'acquérir les règles morpho-syntaxiques d'une ou plusieurs langues et une grande quantité d'informations factuelles, leur donnant les connaissances nécessaires à la réalisation de tâches dans des domaines variés où le traitement automatique du langage naturel en est le cœur. Cependant, les connaissances du modèle sont statiques et évoluent seulement lorsqu'il est ré-entraîné. Elles sont également limitées à l'étendue du contenu des données d'entraînement, généralement non spécialisées, et peuvent ainsi manquer d'expertise sur des domaines spécifiques.

Ce manque de connaissances conduit à des hallucinations, c'est-à-dire un contenu généré qui n'a pas de sens ou qui n'est pas conforme au contenu source fourni [8]. Les hallucinations compliquent l'exploitation des LLM dans des cas où des compétences spécifiques sur des domaines fermés sont nécessaires, comme par exemple pour les agents conversationnels. Pourtant, les LLM restent largement utilisés pour leur capacité de génération de texte qui est un véritable atout. C'est pourquoi les travaux récents [3, 5, 12] cherchent à spécialiser les modèles grâce à des connaissances spécifiques au cas d'usage.

Différentes techniques ont été proposées pour injecter de la connaissance dans les LLM. La première est le fine-tuning [13, 20, 22], qui consiste à poursuivre l'entraînement du modèle sur une tâche et/ou un domaine spécifique pour le spécialiser. La seconde est la génération augmentée de récupération (*Retrieval Augmented Generation*, RAG) [10] où une étape de recherche d'information permet d'extraire du contenu pertinent pour répondre à la question. Ce contenu est ensuite associé à la question pour générer la réponse à l'aide du LLM. Ces techniques possédant chacune leurs avantages et inconvénients, le choix de l'une plutôt que l'autre s'avère complexe.

Nous proposons une comparaison des performances des techniques de RAG et de fine-tuning en français avec le modèle open-source *Mistral 7B* [9]. Nous utilisons des corpus spécifiques accessibles publiquement, parfaitement représentatifs d'un cas d'application d'agent conversationnel. Les métriques étant un réel enjeu pour l'évaluation, nous utilisons deux approches complémentaires : une arène de modèles permettant à des utilisateur-riche-s de voter en aveugle pour leur réponse préférée, ainsi que des métriques automatiques basées sur la pertinence des réponses et leur fidélité aux informations contenues dans les données. L'ensemble des données et du code produit est accessible sur demande auprès des auteur-riche-s.

2 Travaux antérieurs

De récents travaux ont comparé la RAG au fine-tuning d'un LLM [3, 5, 12]. D'après les analyses, la RAG obtient de meilleurs résultats que le fine-tuning. Par exemple, le mo-

dèle *gpt-3.5-turbo-0613* d'OpenAI¹ est comparé lorsqu'il est interrogé seul, fine-tuné ou avec de la RAG [5]. Les données utilisées portent sur un domaine spécifique, avec 100 questions dont les réponses ne sont pas disponibles dans la connaissance interne du modèle pour l'évaluation. Le fine-tuning est effectué avec le format de données rôle-contenu d'OpenAI et la quantité n'est pas précisée, tout comme pour la RAG. Les résultats sont obtenus en comptant les réponses correctes, les faux négatifs et les faux positifs.

Ce type de comparaison est étendu à d'autres modèles [12], par exemple *Mistral 7B* [9], *Llama2 7B* [17] et *Orca2 7B* [11]. Les modèles obtiennent également de meilleures performances en utilisant seulement la RAG plutôt qu'en combinant fine-tuning et RAG sur divers domaines plus génériques du dataset Massive Multilingual Language Understanding Evaluation (MMLU) [7]. Les auteur·rice·s de [12] justifient cela par la difficulté d'apprentissage de nouvelles informations factuelles des LLM lors d'un fine-tuning non-supervisé. Cependant, la répétition de nombreuses variations d'un même fait pendant cette étape peut atténuer cette difficulté. Les données utilisées pour la RAG et le fine-tuning sont extraites de Wikipedia, mais la quantité n'est pas précisée. Cette fois, la métrique utilisée est la précision de la log-vraisemblance des scores attribués par le modèle. Les modèles *Vicuna* [23] et *GPT-4*² ont également été testés sur des domaines spécifiques [3]. La quantité de données utilisées pour la RAG et le fine-tuning est à nouveau non précisée dans l'article. De nombreuses métriques y sont présentées comme la cohérence, la pertinence, la fidélité (*groundedness*) et la complétude, toutes évaluées par un LLM. La mise en pratique et le coût de chacune des techniques est aussi discutée.

Les travaux précédemment évoqués sont effectués sur des textes en anglais, nous proposons donc une comparaison des performances des techniques de RAG et de fine-tuning sur des textes en français. Nous discutons plus particulièrement des problèmes dus, pour le fine-tuning, aux spécificités des textes contenus dans les deux datasets utilisés, et pour la RAG des problèmes de récupération des fragments réellement utiles.

3 Méthodologie

Dans cette section, nous décrivons le processus expérimental mis en place pour comparer la RAG et le fine-tuning d'un LLM.

3.1 La compétition

Nous avons créé deux jeux de données différents qui illustrent le cas d'application d'un agent conversationnel de support pour des usager·ère·s.

Département de la Drôme Le premier jeu de données est issu du site internet du département de la Drôme³. 196 contenus de pages ont été collectés via l'API ouverte sur ledit site. Chaque contenu HTML a été converti en Mark-

down avec *markdownify*⁴ puis découpé en 357 fragments (*chunks*) de façon automatique en s'appuyant sur les titres. Les pages initiales ont une taille moyenne de 1973 caractères avec un écart type de 2898 caractères alors que les fragments obtenus ont une taille moyenne de 1109 caractères avec un écart type de 1483 caractères. Le processus de découpage du contenu en fragments réduit presque de moitié la taille des documents considérés bien que l'écart type nous indique une grande variabilité dans la taille des fragments. 80 questions ont été créées manuellement en sélectionnant de façon aléatoire des fragments de texte et en rédigeant une question dont la réponse est contenue dans le fragment considéré.

Télécom Saint-Etienne Le second jeu de données a été obtenu grâce au site internet de Télécom Saint-Etienne⁵. 502 pages Web ont été collectées par *web scraping*. Le contenu HTML a de nouveau été converti en Markdown avec *markdownify* pour obtenir 1114 fragments. Les pages initiales ont une taille moyenne de 5166 caractères avec un écart type de 38100 caractères alors que les fragments obtenus ont une taille moyenne de 2336 caractères avec un écart type de 4401 caractères. Cette fois, le processus de découpage réduit de plus de moitié la taille des documents considérés et diminue l'écart-type de façon drastique, bien qu'il reste assez important ce qui dénote ici aussi une grande variabilité dans la taille des fragments. 100 questions ont été annotées manuellement en sélectionnant de façon aléatoire des fragments et en rédigeant une question dont la réponse est contenue dans le fragment considéré.

A noter que le découpage des fragments est effectué en s'appuyant sur l'arborescence du plan. Les sous-parties sont découpées individuellement et les titres des parties et sous-parties sont ajoutés au début du fragment. L'ensemble des fragments a été utilisé pour le fine-tuning du modèle et pour la recherche d'information de la RAG.

3.2 Les joueurs

Pour comparer les différentes techniques d'injection de connaissances, nous utilisons le modèle *Mistral 7B* (plus précisément *Mistral-7B-Instruct-v0.2*) avec différentes configurations.

Mistral 7B *Mistral 7B* [9] est un LLM à 7 milliards de paramètres développé par l'entreprise Mistral AI. Comme l'ensemble des modèles de langage, il est entraîné sur un très grand corpus de documents (principalement constitué de données ouvertes du Web) en anglais, sur une tâche de prédiction du mot suivant. Son architecture repose sur un Transformer [18] de type *decoder*. Il dépasse les performances de modèles plus grands comme *Llama 13B* [17] sur de nombreux benchmarks d'évaluation [9] et intègre plusieurs innovations permettant notamment d'accélérer le calcul de l'attention. Le modèle que nous utilisons est instruit, c'est-à-dire qu'il est ensuite entraîné à suivre des instructions. Nous avons fait le choix de fine-tuner ce modèle plutôt que la version non instruite de manière à pouvoir com-

1. <https://platform.openai.com/docs/models/gpt-3-5>

2. <https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo>

3. <https://www.ladrome.fr/>

4. <https://github.com/matthewwithanm/python-markdownify>

5. <https://www.telecom-st-etienne.fr/>

parer des modèles équivalents, or le modèle instruit possède de meilleures compétences en RAG. La poursuite du pré-entraînement (tâche de prédiction du prochain token) pourrait dégrader les compétences acquises grâce à la phase d'alignement, mais il est difficile dans ce contexte de modéliser la tâche d'injection de connaissance comme un objectif de préférence compatible avec les optimisations RLHF ou DPO.

Procédure de fine-tuning Pour injecter de nouvelles informations dans le modèle de langage, nous adoptons l'approche classique de continuation du pré-entraînement sur la tâche de génération conditionnelle (*Mistral* étant un modèle auto régressif). Cette même approche a été utilisée par les auteur-riche-s de [3]. Elle consiste à entraîner le modèle avec un faible pas d'apprentissage sur la tâche de prédiction du mot suivant sur l'ensemble du corpus contenant les informations que l'on souhaite injecter dans le modèle.

Pour ce faire, nous utilisons le protocole de parallélisation DeepSpeed⁶ développé par Microsoft pour affiner le modèle sur 3 *Titan RTX* de 24 Gb de VRAM avec l'approche Zero Stage 3 [14]. Les jeux de données étant relativement petits, l'entraînement dure seulement deux heures pour les données du département de la Drôme et quatre heures pour les données de Télécom Saint-Etienne. Nous utilisons les hyperparamètres classiques [3] : un pas d'apprentissage de $2e-5$ sur deux itérations d'apprentissage (*epochs*) en lots de 16 fragments de texte et un planificateur cosinus (*cosine scheduler*), avec un ratio d'échauffement (*warmup*) de 10% des pas d'entraînement.

Nous utilisons les outils mis à disposition par Hugging Face, notamment le modèle *mistralai/Mistral-7B-Instruct-v0.2*, et la fonction *Trainer* de la librairie Transformers⁷.

Nous rappelons que le but est d'injecter la connaissance dans le modèle. Nous ne disposons donc pas de jeu de données de validation : le contenu informationnel des textes qui n'a pas été pas vu dans l'entraînement ne peut pas être intégré dans la connaissance du LLM. Pour cette raison, nous avons adopté les hyperparamètres de bases cités plus haut pour le fine-tuning du modèle. Néanmoins, une inspection qualitative semble montrer qu'un pas d'apprentissage ou un nombre de passes d'apprentissage trop élevés mènent à une dégradation de la qualité du contenu généré par le modèle (contenu non structuré, hallucination) alors qu'à l'inverse les réduire empêche le modèle d'apprendre le contenu spécifique. Un exemple typique de sur-apprentissage observé dans nos expériences pour le jeu de données de La Drôme (pas d'apprentissage trop élevé) : *Question* : "Présente-moi le projet pour le Pont du Tram" *Réponse* : "La Drôme est le Département de la Drôme. Le Département de la Drôme est la Drôme du Département de la Drôme.", ou encore *Question* : "Quelle est la plus haute chaîne de montagne de France?" *Réponse* : "Le massif du Vercors.". Ce point central est encore une question ouverte et a été peu exploré par les travaux précédents. Ce problème s'écarte du champ d'étude du présent article mais fera l'objet de futures contri-

butions. Lorsque le modèle est fine-tuné, il est ensuite interrogé pour répondre aux questions avec une température à 0 et le prompt suivant :

Département de la Drôme <s> [INST] Ceci est une conversation entre un humain et un assistant du département de la Drôme. L'assistant donne uniquement sa réponse. Sa réponse doit être courte et en français. La réponse est toujours en lien avec le département de la Drôme.
{question} [/INST]

Télécom Saint-Etienne <s> [INST] Ceci est une conversation entre un humain et un assistant de l'école d'ingénieur Télécom Saint-Etienne. L'assistant donne uniquement sa réponse. Sa réponse doit être courte et en français. La réponse est toujours en lien avec Télécom Saint-Etienne.
{question} [/INST]

Génération augmentée de récupération (RAG) La génération augmentée de récupération est fondée sur trois étapes principales :

1. Préparation de la base de données à exploiter pour la recherche d'information (sous-étapes 1a et 1b de la figure 2);
2. Récupération des éléments de contexte proches de la requête de l'utilisateur-riche (sous-étapes 2a, 2b et 2c);
3. Génération de la réponse (sous-étapes 2d et 2e).

Le pré-requis est la préparation des données spécifiques au cas d'application. Il s'agit de documents dont les données textuelles sont extraites, nettoyées, découpées en fragments (*chunking*), vectorisées (création d'*embeddings*, le cas échéant) puis indexées dans une base de données.

À l'arrivée d'une requête formulée par l'utilisateur-riche, la tâche de récupération (*retrieval*) consiste en la recherche d'information liée à la phrase exprimée. Elle est opérée par exemple via une méthode de pondération (e.g. BM25 [15] ou TF-IDF [16]), ou via une méthode de calcul vectoriel [1] (e.g. similarité cosinus entre *embeddings*), voire une hybridation de ces méthodes.

Ensuite, à partir du contexte constitué par la tâche de récupération, la tâche de génération implique un LLM pour synthétiser la réponse.

Dans ces travaux, nous implémentons le processus de RAG avec la solution commerciale Wikit Semantics⁸. Le modèle d'*embeddings text-embedding-ada-002* a été utilisé pour la partie récupération car il affiche de très bonnes performances pour cette tâche en français d'après le leaderboard MTEB⁹. Le modèle *Mistral-7B-Instruct-v0.2* (ou *mistral-tiny*) décrit précédemment a également été utilisé pour l'étape de génération. Pour l'étape de récupération, les 10 fragments ayant la similarité cosinus la plus élevée avec la question sont conservés pour former le contexte lors de la génération. La génération est faite avec une température à 0 et le prompt suivant :

6. <https://github.com/microsoft/DeepSpeed>

7. <https://huggingface.co/docs/transformers/index>

8. <https://www.wikit.ai/fonctionnalites/wikit-semantic>

9. <https://huggingface.co/spaces/mteb/leaderboard>

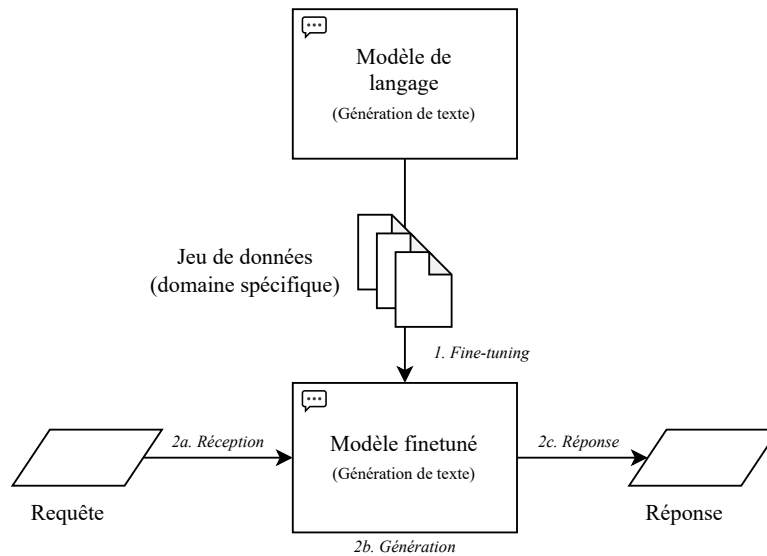


FIGURE 1 – Fine-tuning (1) et génération via le LLM (2)

Système Vous êtes un assistant utile qui répond aux questions à propos du département {de la Drôme | de l'école d'ingénieur Télécom Saint-Etienne}.

Humain Répondre à la question en utilisant uniquement le contexte ci-après. Si la question ne peut pas être répondue avec le contexte, répondre "Je ne sais pas".

Humain Contexte : {contexte}

Humain {question}

3.3 L'arbitrage

Deux stratégies différentes ont été utilisées de façon complémentaire pour évaluer la RAG et le fine-tuning : le vote en aveugle par des annotateur-ric-e-s pour la technique de génération favorite et des métriques calculées automatiquement.

3.3.1 L'arène de modèles

L'évaluation de la préférence des réponses est réalisée via une interface utilisateur-ric-e dédiée présentée figure 3, inspirée de la Chatbot Arena¹⁰. L'écran principal présente le contexte de la question, la question, et les deux réponses générées respectivement par le LLM fine-tuné et par la RAG. L'ordre d'affichage des réponses est aléatoire, sans indication de l'approche de génération. Les boutons mis à disposition des évaluateur-ric-e-s pour l'évaluation des réponses sont :

- "A est meilleure" pour indiquer que la réponse présentée à gauche est préférée ;
- "B est meilleure" pour celle de droite ;
- "Match nul" lorsque l'évaluateur-ric-e considère que les deux réponses se valent ;
- "Aucune" si aucune des réponses n'est considérée comme pertinente.

Les votes ont été réalisés par trois personnes expertes du domaine des agents conversationnels mais non expertes des domaines d'application utilisés pour les expérimentations. A titre d'illustration, nous présentons en Figures 4, 5, et 6 quelques cas annotés.

3.3.2 Métriques automatiques

De nombreux outils sont disponibles pour évaluer les réponses générées par un LLM, en particulier pour la RAG, comme RAGAS [6], Athina Evals¹¹ ou encore UpTrain¹². Nous nous intéressons aux métriques pouvant s'appliquer aussi bien au fine-tuning qu'à la RAG, qui ne nécessitent pas d'autres annotations que celles dont nous disposons déjà, et qui rendent compte de la qualité de réponse des LLM dans le cas d'application à des agents conversationnels.

Pertinence La première métrique que nous sélectionnons est la pertinence, qui traduit le terme *answer relevance* largement répandu dans la littérature [6]. Avec cette métrique, nous cherchons à mesurer à quel point la réponse générée par le LLM est pertinente au vu de la question posée. La véracité des informations apportées n'est pas vérifiée, cette métrique se concentre sur le périmètre de la question, en vérifiant que la réponse le couvre entièrement. Pour calculer la pertinence d'une réponse, nous nous inspirons de l'implémentation d'Athina Evals qui utilise un prompt contenant une question, une réponse et des instructions pour qu'un LLM évalue avec une valeur booléenne si la réponse donnée répond à la question posée, en justifiant l'évaluation avec des explications. Nous modifions le prompt pour obtenir un score plutôt qu'une valeur booléenne indiquant si oui ou non la réponse est pertinente par rapport à la question. Un LLM de plus grande taille génère donc un score et une explication qui représente la pertinence de la réponse

11. <https://github.com/athina-ai/athina-evals>

12. <https://docs.uptrain.ai/predefined-evaluations/overview>

10. <https://chat.lmsys.org/>

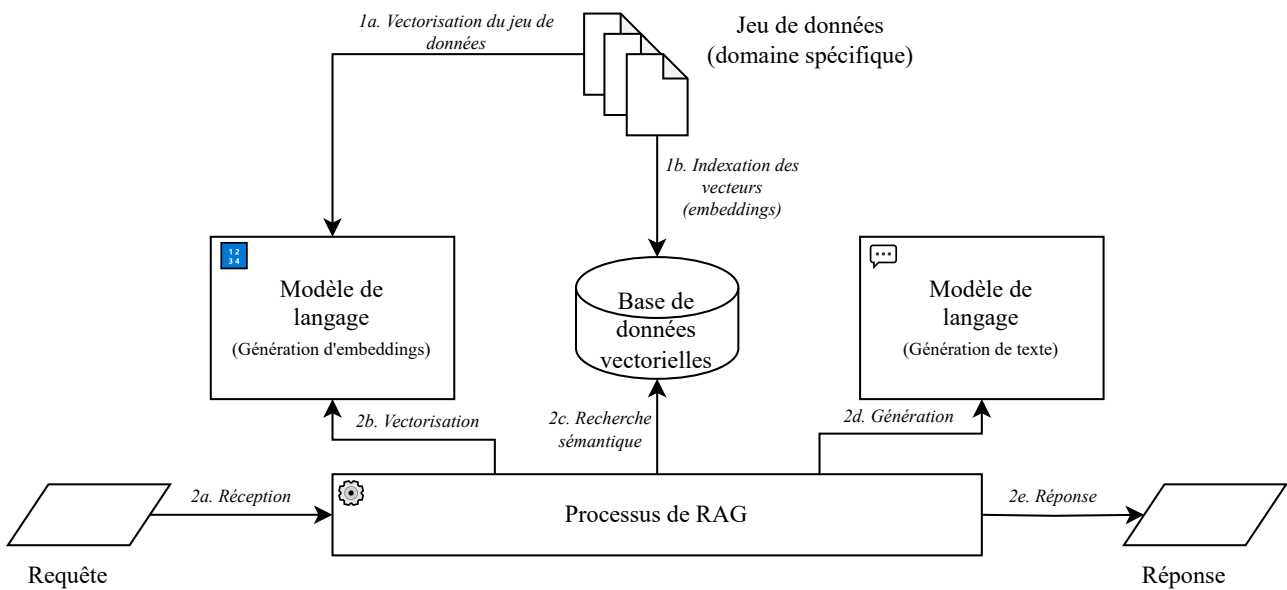


FIGURE 2 – Indexation des données (1) et génération via le processus de RAG (2)

Contexte

Séances publiques

Pour vous rendre et assister aux séances publiques

Salle Marius Moutet Conseil départemental de la Drôme Hôtel du Département 26 avenue du président Herriot 26000 VALENCE

Réponse A

Où se déroulent les séances publiques ?

Les séances publiques du Conseil départemental de la Drôme se déroulent à la Salle Marius Moutet, Conseil départemental de la Drôme, Hôtel du Département, 26 avenue du président Herriot, 26000 VALENCE.

Réponse B

Où se déroulent les séances publiques ?

Les séances publiques se déroulent au Conseil Départemental, 26 avenue du Président Herriot 26026 Valence Cedex 9.

Evaluation

A est meilleure
 B est meilleure
 Match nul
 Aucune

Évaluateur·rice : Christophe

Code : D26

FIGURE 3 – Interface de l'arène d'évaluation.

Question : *Qu'est ce que la cartoucherie ?*

Contexte (extrait) : *[...] l'ancienne cartoucherie nationale de Bourg-lès-Valence est devenue pôle d'excellence régional de l'image animée. En plus de nombreux studios, la Cartoucherie accueille l'école de La Poudrière (reconnue par le Ministère de la Culture) et l'ECAS (école du studio TeamTo, spécialisée dans la formation à l'animation 3D).*

Réponse Fine-tuning : *Une cartoucherie est un site d'exploitation de la tourbe, utilisé pour la fabrication des cartouches de fusil. (pour plus d'informations : <https://www.ladrome.fr/mon-quotidien/economie/les-entreprises/les-entreprises-de-la-filiere-de-la-tourbe/>)*

Réponse RAG : *La Cartoucherie est un pôle d'excellence régional de l'image animée situé à Bourg-lès-Valence en Drôme. Il abrite de nombreux studios de production de films d'animation, l'école de La Poudrière, et l'ECAS (école du studio TeamTo), spécialisée dans la formation à l'animation 3D.*

FIGURE 4 – Exemple de question-réponses dont la réponse préférée est celle de la RAG. Ce premier exemple montre une bonne formulation de la réponse par la RAG, alors que celle issue du LLM fine-tuné comporte une hallucination montrant que l'étape de fine-tuning n'a pas réussi à spécialiser suffisamment le LLM, ainsi qu'un lien hyper-texte invalide, ce qui est un problème assez récurrent avec le fine-tuning.

générée par un LLM de plus petite taille. Nous utilisons le modèle d'OpenAI *gpt-3.5-turbo* avec le prompt suivant :

Système *You are an expert at evaluating whether the response answers specifically what the user is asking about, and covers all aspects of the user's query. You are not checking for correctness, or factual accuracy. You are only checking if the response answers the user's query.*

Humain *Consider the following :*
user's query : {question}
response : {response}

Humain *Determine if the response answers specifically what the user is asking about, and covers all aspects of the user's query. Provide a brief explanation of why the response does or does not answer the user's query sufficiently, labeled as 'explanation', leading up to a score between 0 and 5 labeled as 'result'.*

Humain *Return a JSON object in the following format :*
'result' : result, 'explanation' : explanation.

Comme le score obtenu est compris entre 0 et 5, il est ensuite normalisé. L'explication demandée au modèle lui permet de générer des scores plus proches du résultat attendu, probablement grâce au principe des *chain-of-thought* [21]. De même, le modèle génère des scores plus variés lorsqu'on impose que le score soit entre 0 et 5 plutôt qu'entre 0 et 1 (où le modèle a tendance à ne générer que des 0 ou des 1).

Fidélité La seconde métrique que nous considérons est la fidélité, qui traduit le terme *faithfulness* utilisé dans la littérature [6]. Cette mesure représente la véracité factuelle de la réponse générée par le LLM. La fidélité peut facilement être liée à la détection d'hallucinations. La plupart des outils disponibles utilisent également un LLM pour l'évaluer. Dans le but de limiter notre utilisation de très gros modèles de langues, nous avons cherché à effectuer cette tâche sans LLM. Nous utilisons donc spaCy¹³ pour extraire les passages d'intérêt de la réponse et des fragments identifiés comme pertinents pour répondre à la question. Les passages d'intérêts sont les entités identifiées par spaCy grâce à son module de reconnaissance d'entités pré-entraîné, ainsi que les adresses de sites internet, les adresses e-mail et nombres (tels que les codes postaux ou numéros de téléphones). Soit $PI_{response}$ l'ensemble des passages d'intérêts de la réponse et $PI_{fragment}$ l'ensemble des passages d'intérêts du fragment pertinent correspondant. Le score de fidélité est alors calculé d'après l'équation 1 comme la proportion des passages d'intérêts de la réponse présents dans le fragment pertinent. Les passages d'intérêts étant des chaînes de caractères, la correspondance entre les passages d'intérêts de la réponse et ceux du fragment pertinent est calculée par égalité des chaînes de caractères.

$$fidélité = \frac{card(PI_{response} \cap PI_{fragment})}{card(PI_{response})} \quad (1)$$

13. <https://spacy.io/>

Question : <i>Quelle est l'activité de la société Sick ?</i>
Contexte (extrait) : <i>Télécom Saint-Etienne signe une Chaire Innovation avec SICK France [...] En cohérence directe avec les enseignements du cycle ingénieur, SICK vient apporter dans le cadre de cet accord son expertise dans le domaine de la captation, de la donnée et de l'intelligence artificielle auprès de nos étudiants. [...]</i>
Réponse Fine-tuning : <i>La société SICK produit des capteurs pour la captation de données.</i>
Réponse RAG : <i>Sick is a leading global supplier of sensor solutions for industrial applications.</i>

FIGURE 5 – Exemple de question-réponses dont la réponse préférée est celle générée par le LLM fine-tuné. Cet exemple montre une génération en anglais de la RAG, et une formulation du LLM qui est préférée par les évaluateur-riche-s.

4 Résultats

Dans cette section, nous analysons les résultats des comparaisons de la RAG et du fine-tuning sur le jeu de données du département de la Drôme et celui de Télécom Saint-Etienne.

4.1 Résultats des annotations de l'arène

Nous analysons d'abord la cohérence des évaluations effectuées par les annotateur-riche-s.

Comme indiqué sur la figure 7, les 3 annotateur-riche-s sont en accord total sur la réponse préférée dans 68% des cas. Dans 29% des cas, deux des trois annotateur-riche-s sont en accord. Nous pouvons donc conclure à un bon accord global entre les 3 annotateur-riche-s. Ces chiffres sont étayés par les scores du Kappa de Cohen [2] calculés pour les 3 paires d'annotateur-riche-s. Les scores obtenus pour chacune des paires sont peu variables, avec une valeur moyenne de 0.55 indiquant un accord modérément élevé entre chaque annotateur-riche-s. Notons que ce score, bien qu'il représente un indicateur utile, souffre d'une fiabilité diminuée lorsque les classes étudiées sont déséquilibrées[19], comme explicité dans le prochain paragraphe.

TABLE 1 – Accord inter-annotateur-riche-s

	P1/P2	P2/P3	P1/P3
Kappa de Cohen	0.545	0.539	0.577

Les réponses générées par la RAG sont largement préférées par les annotateur-riche-s d'après la figure 8. En effet, sur les 180 questions du jeu de données, la réponse de la RAG est en moyenne préférée dans $67\% \pm 3\%$ des cas. En revanche, la réponse générée par le modèle fine-tuné n'est préférée en moyenne que dans $3\% \pm 0\%$ des cas. Ce résultat s'explique par le fait que la réponse du modèle fine-tuné souffre d'un manque de justesse notable. Souvent, le modèle génère une

<p>Question : <i>Quand auront lieu les projets d'ingénierie de 2023 ?</i></p>
<p>Contexte (extrait) : <i>Lancement des projets d'ingénierie (PING) pour nos étudiants ingénieurs en 2e année [...] 16 mars 2023 [...] Les projets d'ingénierie sont lancés! [...] Du 14 mars au 23 juin, les étudiants répartis en équipe de 4 à 8 personnes vont travailler à la réalisation de leur projet à travers 3 étapes-clés : 1 phase de cadrage et 2 "sprints". [...] Au total, les étudiants travailleront l'équivalent de 700 heures réparties en séances de travail de mars à juin. [...]</i></p>
<p>Réponse Fine-tuning : <i>Les projets d'ingénierie de 2023 auront lieu du 20 avril au 14 juin. Pour plus d'informations : <https://www.telecom-st-etienne.fr/entreprise/etre-accompagne/projet-dingenierie/> Partagez Précédent (pour plus d'informations : https://www.telecom-st-etienne.fr/telecom-saint-etienne-fete-erasmus-days-2023/) Suivant (pour plus d'informations : https://www.telecom-st-etienne.fr/telecom-saint-etienne-acteur-du-projet-europeen-digital-creativity-for-digital-maturity-dc4dm/)</i></p>
<p>Réponse RAG : <i>Les projets d'ingénierie de 2023 ont commencé le 16 mars et s'achèveront en juin. Les étudiants travailleront l'équivalent de 700 heures sur leur projet, réparties en trois étapes-clés : une phase de cadrage, un premier sprint et un deuxième sprint. Les projets seront clôturés par une journée de soutenances de projets le 25 juin. Selon la situation sanitaire, les Projets Ingénierie se tiendront en présentiel et/ou en distanciel.</i></p>

FIGURE 6 – Exemple de question-réponses dont aucune des réponses n'est préférée par les évaluateur-ric-e-s. Ce dernier exemple montre une question-réponse liée à une période de temps. Les deux approches proposent des dates incorrectes.

réponse qui semble vraisemblable car elle reprend des références et éléments de langage du jeu de données. Cependant, les informations données sont fausses et ne correspondent pas à l'information qui peut être trouvée dans les documents. Au contraire, dans le cas de la RAG, le modèle a tendance à paraphraser précisément le document pertinent pour répondre à la question, lorsque ce dernier a bien été extrait par l'étape de récupération. Dans $23\% \pm 5\%$ des cas, aucune des deux réponses n'a été jugée satisfaisante pour être utile à l'utilisateur, à cause d'un problème de fond ou de forme (réponses fausses, rédigées en anglais ou incomplètes par exemple).

4.2 Résultat de l'évaluation automatique

Pour l'évaluation grâce aux métriques calculées automatiquement, nous utilisons les réponses générées par la RAG et par le fine-tuning sur chacun des jeux de données. Les résultats sont présentés dans la table 2. Comme l'évalua-

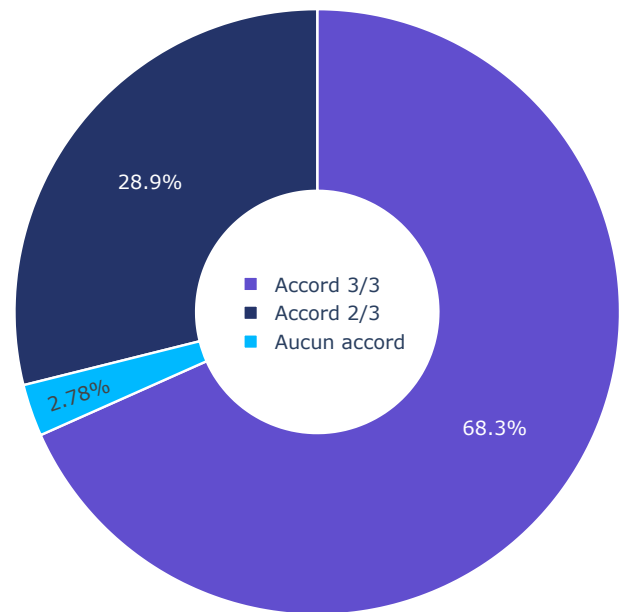


FIGURE 7 – Proportion d'accord entre les annotateur-ric-e-s

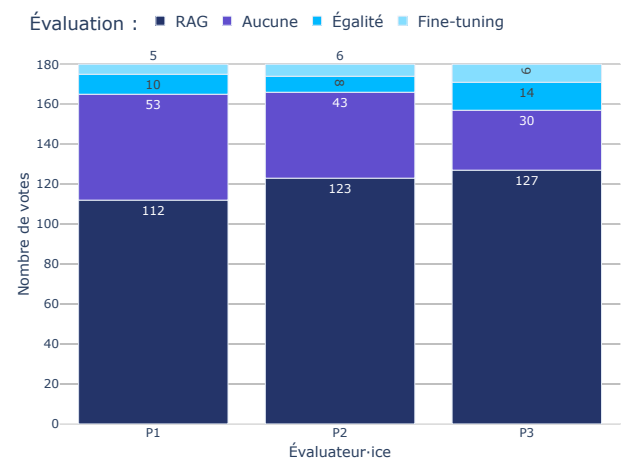


FIGURE 8 – Proportion de votes de préférence de réponse aux 180 questions du jeu de données selon la méthode utilisée pour la génération pour chacun-e des annotateur-ric-e-s.

tion de la pertinence (*answer relevance*) est faite avec un LLM, les résultats obtenus ne sont pas déterministes. Une même réponse peut donc se voir attribuer un score différent d'une génération à une autre. Pour obtenir un score robuste, la pertinence de chaque réponse a été évaluée 10 fois. Le score de pertinence attribué à chaque modèle dans la table 2 est donc la moyenne de la pertinence de chacune des réponses pour les 10 générations. Le calcul de la fidélité (*faithfulness*) est déterministe car la détection des passages d'intérêts avec spaCy est le même d'une exécution à l'autre pour une même phrase. La fidélité est donc calculée une seule fois pour chaque génération. Le score de fidélité présenté pour chaque technique dans la table 2 est la moyenne de la fidélité pour chaque réponse générée.

TABLE 2 – Pertinence et fidélité des réponses pour le fine-tuning et la RAG sur chaque dataset.

Dataset	Technique	Pertinence	Fidélité
Département de la Drôme	Fine-tuning	0.74	0.23
	RAG	0.81	0.56
Télécom St-Etienne	Fine-tuning	0.68	0.37
	RAG	0.79	0.66

Quel que soit le jeu de données, la RAG obtient de meilleurs résultats que le fine-tuning pour la pertinence et pour la fidélité. Les réponses des deux techniques semblent plus pertinentes sur les données du département de la Drôme mais plus fidèles sur celles de Télécom Saint-Etienne. Cela peut s'expliquer par le fait que le jeu de données de la Drôme contient plus d'informations complexes à restituer pour le LLM comme des numéros de téléphone, des adresses e-mail ou des adresses de sites internet. Les fragments de la Drôme contiennent donc plus de passages pertinents à retrouver dans la réponse générée, ce qui semble limiter les scores obtenus pour la fidélité.

Les scores de pertinence sont élevés pour les deux techniques, ce qui suggère que même si la réponse n'est pas exacte, le texte généré est cohérent avec la question posée et tend bien à y répondre.

Les scores de fidélité sont eux plus faibles, ce qui suggère que les informations factuelles transmises sont facilement erronées bien que la formulation de la réponse soit cohérente avec la question posée. Des résultats aussi bas que ceux du fine-tuning sont particulièrement problématiques pour le cas d'usage d'agents conversationnels car la réponse générée n'apporte pas d'informations fiables aux utilisateur.rice.s.

5 Discussion et conclusion

Les jeux de données utilisés dans cette étude ont plusieurs particularités notables. La première est qu'ils contiennent des informations spécifiques non vues par les modèles pré-entraînés couramment utilisés (comme *Mistral*). Ce type de donnée a fait l'objet d'études antérieures comparant RAG et fine-tuning, comme présenté précédemment. La deuxième spécificité n'a quant à elle pas encore été étudiée dans les approches existantes : les documents sont en français. Or,

même si *Mistral* peut générer du contenu dans cette langue, ce modèle est spécialisé pour la langue anglaise [9]. Enfin, les documents des corpus ont des structures particulièrement standardisées (par exemple titre + contenu + section "plus d'information <URL>" pour Télécom Saint-Etienne).

5.1 Injection de connaissances par fine-tuning

En raison des spécificités présentées plus haut, le fine-tuning n'a pas permis d'injecter convenablement les connaissances. En effet, de nombreux faits n'apparaissent qu'une seule fois dans le jeu de données (cf. figure 6), ce qui complique l'intégration de l'information. L'augmentation du nombre d'itérations d'apprentissage (*epochs*) permettrait de pallier à ce problème. Cependant, si le modèle est entraîné trop longtemps, il sur-apprend les données de spécialisation et perd ses capacités à générer du langage cohérent. Il est donc difficile de présenter les faits plusieurs fois au modèle sans lui faire perdre ses capacités de pré-entraînement. Pour résoudre ce problème, une piste intéressante consisterait à générer des paraphrases des documents du corpus spécifique, de manière à augmenter les données en réduisant le sur-apprentissage [12]. Néanmoins, une étude approfondie serait nécessaire : les paraphrases doivent être suffisamment diverses et ne pas altérer les informations essentielles contenues dans les documents. De plus, nous avons observé que la structure standardisée amplifie le phénomène de sur-apprentissage et d'hallucination. Sur le jeu de données Télécom Saint-Etienne, par exemple, le LLM termine quasiment toujours sa génération par "Pour plus d'information : <URL>". Ce phénomène indique clairement que le modèle a sur-appris. Il génère aussi systématiquement un lien, qui souvent semble lié à la réponse mais qui mène rarement à une page existante. Enfin, nous avons également évalué le fine-tuning au moyen de méthodes frugales, notamment la méthode QLora [4]. Cette approche consiste à quantifier le modèle, c'est-à-dire réduire la précision de stockage des paramètres du réseau sur 4 bits et ajouter des *adapters*. Ces couches supplémentaires sont entraînées sur la tâche pendant que les paramètres du réseau pré-entraîné sont gelés. Cette approche permet de réduire de façon drastique le nombre de paramètres entraînés. Néanmoins, nous avons observé une perte en génération : la log vraisemblance négative du contenu est 10 fois plus élevée en moyenne avec cette approche. L'utilisation de méthodes *parameter efficient* devra être plus longuement explorée dans le futur mais ne semble, pour l'instant, pas être une alternative satisfaisante sur ce type d'application.

5.2 Optimisation et enrichissement de la technique de RAG

La technique de RAG utilisée en comparaison au fine-tuning représente l'une des méthodes disponibles pour contextualiser la génération d'une réponse par un LLM. Cependant, elle repose sur de nombreux paramètres dont la modification pourrait modifier les résultats. La première étape de découpage du contenu en fragments est primordiale car elle permet d'obtenir du texte propre et cohérent,

porteur d'une information claire pour l'étape de recherche d'information. Optimiser cette étape permet donc d'améliorer la récupération qui améliore à son tour la génération. Lors de l'étape de recherche d'information, la technique de similarité et le modèle de représentation du langage choisis ont tous les deux une grande influence sur les résultats. Nous pouvons par exemple envisager d'utiliser un modèle de représentation du langage spécialisé sur la tâche de récupération d'informations ou sur les domaines spécifiques de nos données. Actuellement, la récupération extrait 10 fragments de texte qui sont ensuite utilisés comme éléments de contexte. Dans de futurs travaux, nous souhaitons réduire ce nombre de fragments pour conserver uniquement ceux contenant des informations relatives à la question posée et ainsi limiter la consommation de ressources en diminuant de façon drastique la taille du prompt. Des étapes supplémentaires pourront également être ajoutées comme le *re-ranking*, qui consiste à réordonner les fragments identifiés par ordre de pertinence pour être présentés au LLM. Enfin, une optimisation du prompt de génération de réponse est envisagée pour s'adapter au mieux au format attendu (langue, longueur de la réponse, précision, etc.).

5.3 Conclusion

Ces travaux nous ont permis de mettre en lumière les points suivants :

1. Bien que la pertinence des réponses soit proche pour les modèles RAG et fine-tunés, l'approche RAG est largement préférée dans une évaluation de type "Arena".
2. La RAG est peu dégradée d'un dataset à l'autre alors que l'impact sur le fine-tuning est plus important : la qualité des données semble donc avoir un impact plus important sur le fine-tuning.
3. Le fine-tuning est moins efficace que la RAG dans le cas de données spécialisées où les informations sont peu répétées.

Notre analyse possède également quelques limitations. Le modèle Mistral 7B n'est pas spécialisé sur la génération de contenu en français. Nous n'avons évalué qu'un seul modèle de langage, néanmoins, le coût d'entraînement peut s'avérer très élevé, notamment quand le nombre de paramètres du modèle croît fortement. Ces axes de recherche feront l'objet de travaux futurs.

Enfin, les observations résultant de ce travail permettent de donner de nouvelles directions de recherche dans le domaine de l'application des grands modèles de langage sur les données expertes. Nous envisageons d'étudier plus finement l'impact de la taille des jeux de données sur les résultats ou encore de l'approche de fine-tuning (par exemple par la mise en forme de la tâche de continuation comme une tâche d'instruction).

Remerciements

Ces travaux s'inscrivent dans le programme QABot, qui bénéficie du dispositif R&D Booster financé par la Région

Auvergne-Rhône-Alpes et BPIFrance que nous remercions ici.

Références

- [1] Tolgahan Cakaloglu, Christian Szegedy, and Xiaowei Xu. Text embeddings for retrieval from a large knowledge base. In *International Conference on Research Challenges in Information Science*, volume 385 of *Lecture Notes in Business Information Processing*, pages 338–351. Springer, 2020.
- [2] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1) :37–46, 1960.
- [3] Angels de Luis Balaguer, Vinamra Benara, Renato Luiz de Freitas Cunha, Roberto de M. Estevo Filho, Todd Hendry, Daniel Holstein, Jennifer Marsman, Nick Mecklenburg, Sara Malvar, Leonardo Nunes, Rafael Padilha, Morris Sharp, Bruno Leonardo Barros Silva, Swati Sharma, Vijay Aski, and Ranveer Chandra. Rag vs fine-tuning : Pipelines, tradeoffs, and a case study on agriculture. *ArXiv*, abs/2401.08406, 2024.
- [4] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora : Efficient finetuning of quantized LLMs. *Advances in Neural Information Processing Systems*, 36, 2024.
- [5] Jennifer Dodgson, Nanzheng Lin, Julian Peh, Akira Raphael Janson Pattirane, Alfath Daryl Alhajir, Eko Ridho Dinarto, Joseph Lim, and Syed Danyal Ahmad. Establishing performance baselines in fine-tuning, retrieval-augmented generation and soft-prompting for non-specialist llm users. *ArXiv*, abs/2311.05903, 2023.
- [6] Shahul ES, Jithin James, Luis Espinosa Anke, and Steven Schockaert. Ragas : Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - System Demonstrations*, pages 150–158, 2024.
- [7] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *Proceedings of the 9th International Conference on Learning Representations, ICLR, 2021*.
- [8] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models : Principles, taxonomy, challenges, and open questions. *ArXiv*, abs/2311.05232, 2023.
- [9] Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Re-

- nard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *ArXiv*, abs/2310.06825, 2023.
- [10] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NeurIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc.
- [11] Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Coda, Clarisse Simoes, Sahaj Agrawal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, Hamid Palangi, Guoqing Zheng, Corby Rosset, Hamed Khanpour, and Ahmed Awadallah. Orca 2 : Teaching small language models how to reason. *ArXiv*, abs/2311.11045, 2023.
- [12] Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. Fine-tuning or retrieval? comparing knowledge injection in LLMs. *ArXiv*, abs/2312.05934, 2023.
- [13] Evani Radiya-Dixit and Xin Wang. How fine can fine-tuning be? learning efficient language models. In *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020*, volume 108 of *Proceedings of Machine Learning Research*, pages 2435–2443, 2020.
- [14] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero : Memory optimizations toward training trillion parameter models. In *SC20 : International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE, 2020.
- [15] S. E. Robertson and Sparck K. Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3) :129–146, 1976.
- [16] Karen Sparck Jones. *A statistical interpretation of term specificity and its application in retrieval*, page 132–142. Taylor Graham Publishing, GBR, 1988.
- [17] Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2 : Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288, 2023.
- [18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30 : Annual Conference on Neural Information Processing Systems*, pages 5998–6008, 2017.
- [19] Anthony J Viera, Joanne M Garrett, et al. Understanding interobserver agreement : the kappa statistic. *Family Medicine*, 37(5) :360–363, 2005.
- [20] Zhongwei Wan, Xin Wang, Che Liu, Samiul Alam, Yu Zheng, Jiachen Liu, Zhongnan Qu, Shen Yan, Yi Zhu, Quanlu Zhang, Mosharaf Chowdhury, and Mi Zhang. Efficient large language models : A survey. *arXiv preprint arXiv :2312.03863*, 2024.
- [21] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35 : Annual Conference on Neural Information Processing Systems, NeurIPS 2022*, 2022.
- [22] Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. Instruction tuning for large language models : A survey. *arXiv preprint arXiv :2308.10792*, 2023.
- [23] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36 : Annual Conference on Neural Information Processing Systems, NeurIPS*, 2023.

Mise en place d'un système de notation par aspects pour les articles de E-commerce en se basant sur les avis en français des consommateurs

Mathias Chastan¹, Rania Zaatour¹, Cheik Ouedraogo¹, Karim Baati¹

Teolia Consulting

8 rue de l'Hôtel de Ville, 92200 Neuilly-sur-Seine, France

mathias.chastan@teolia.fr, rania.zaatour@teolia.fr, cheik.ouedraogo@teolia.fr, karim.baati@teolia.fr

2024

Résumé

Un très grand volume de données d'avis sur les articles consommés sont mis en ligne chaque jour. Ce sont des informations sur des produits ainsi que sur leurs différentes caractéristiques spécifiques. A l'encontre de la majorité des travaux de recherche qui se sont penchés sur l'évaluation globale des produits à partir des avis collectés, nous proposons, dans cet article, un système de notation par aspects. Le système suggéré s'appuie sur une analyse poussée d'avis rédigés en français, et permet d'extraire les différentes propriétés des articles évalués ainsi que l'opinion utilisateur afférente à chaque propriété. La validation technique de ce système est opérée sur des avis relatifs à des articles issus d'un site de E-commerce français.

Mots-clés

Traitement automatique du langage naturel, Analyse sentimentale par aspects, Avis produits, Notation, Extraction d'aspects

Abstract

A very large volume of review data on consumed articles is put online every day. These data provide information about products as well as their various specific characteristics. In contrast to the majority of research works that have focused on the global evaluation of products based on the collected reviews, we propose, in this paper, an aspect-based scoring system. This latter is based on a thorough analysis of reviews written in French. It allows to extract the different properties of assessed items as well as the user opinion related to each of these properties. The technical validation of this system is performed on reviews related to the products suggested by a French E-commerce website.

Keywords

Natural language processing, Aspect based sentiment analysis, Product reviews, Scoring, Aspect mining

1 Introduction

Avec l'expansion des plateformes de e-commerce, le shopping en ligne est devenu une méthode incontournable de consommation. De très grands volumes de données générées par les utilisateurs sont mis en ligne chaque jour. Il s'agit d'informations sur les préférences et les opinions des utilisateurs au regard de différents produits, souvent dirigées vers des caractéristiques spécifiques de ces articles. Bien que ces textes regorgent d'informations, ils sont souvent non structurés et demandent un effort d'analyse de la part des utilisateurs ou des enseignes, afin de parvenir à extraire l'information dont chacun a besoin.

Le domaine de l'intelligence artificielle [3, 4] a vu émerger ces derniers temps de nouvelles techniques de traitement du langage naturel (en anglais *Natural Language Processing* (NLP)) qui permettent d'assurer l'analyse sentimentale du langage humain. En d'autres termes, ces techniques permettent d'extraire le ressenti global d'un utilisateur sur un sujet à partir d'un message oral ou écrit qu'il a produit.

Dans le cas d'avis soumis en ligne, ces méthodes servaient, dans un premier temps, à assigner un score global à chaque produit évalué. Néanmoins, avec le temps, il s'est avéré que le score global est réducteur, étant donné qu'il ne prend pas en compte les opinions sur les différentes propriétés de l'article évalué, et que ces opinions peuvent être d'intensité et de polarité différentes. En effet, un client peut trouver que la qualité d'un produit est satisfaisante mais que son prix est trop élevé. Dans ce cadre, des techniques d'analyse sentimentale plus poussées permettent de répondre à cette problématique en effectuant une analyse par aspects. Le travail présenté dans cet article se joint à ces techniques et propose un système d'analyse sentimentale par aspects qui s'assoit sur les avis en français soumis en ligne. Le système suggéré fait usage de différents modèles de NLP pré-entraînés ainsi qu'un algorithme de règles pour extraire et noter automatiquement la satisfaction client au regard des différentes propriétés d'un article.

La suite de cet article est scindée en 5 sections. La section 2 porte sur une synthèse des travaux antérieurs ayant traité la

même problématique. Ensuite, dans la section 3, nous présentons notre système tout en détaillant les modèles ainsi que les données utilisés. Quant à la section 4, elle est dédiée à la discussion et aux applications pratiques de notre système. Finalement, la section 5 conclut ce travail et en dégage des perspectives pour des travaux futurs.

2 Etat de l'art

Une première remarque que l'on peut soulever en évoquant les travaux antérieurs en lien avec l'analyse sentimentale des avis en ligne, est que bon nombre de travaux ont visé l'attribution de scores globaux pour les produits évalués. [7, 18]. Par opposition à ces études, le travail suggéré dans cet article fait appel à l'analyse sentimentale en vue d'assigner des scores par aspects.

Un aspect est défini comme toute caractéristique ou propriété d'un produit pouvant influencer le choix d'un client potentiel (e.g., qualité, confort, couleur, prix) [22]. Un sentiment est défini comme une connaissance ou une impression qu'un client a sur un produit.

L'analyse sentimentale a pour objectif de déterminer l'opinion ou l'impression qui se trouve derrière le langage naturel. En se fondant sur cette définition, nous pouvons définir l'analyse sentimentale par aspects comme le processus qui permet de déterminer l'opinion ou l'impression associée à chaque caractéristique d'un produit. Elle est généralement fondée sur trois étapes : (1) l'extraction des aspects et des mots sentiments, (2) le regroupement des aspects, et (3) la notation des mots sentiments.

1. L'extraction des aspects et des dépendances consiste à extraire les caractéristiques mentionnées par les consommateurs dans leurs avis en ligne et les sentiments qui y sont associés. Le mot sentiment est le mot qui décrit l'opinion du consommateur sur une caractéristique. Ces aspects et mots sentiments sont cachés dans les phrases et pour les retrouver un *Part Of Speech tagger* (*PoS tagger*) est utilisé. Un *PoS tagger* permet d'extraire la nature d'un mot dans une phrase. Pour ce faire, quatre bibliothèques sont principalement utilisées, à savoir «NLTK»¹, «Stanford Core», «NLP suite», et «Spacy». Plus de notion sur ces bibliothèques sont disponibles dans [1].
2. Le regroupement des aspects permet de regrouper sous un seul groupe, les mots désignant le même aspect. Ce regroupement prend en compte la sémantique de l'aspect dans l'avis client. Cette étape requiert l'utilisation d'un modèle de plongement lexical tel que «Word2vec» [9].
3. La dernière étape vise à conférer un score de sentiment aux mots sentiments. Pour cela, il faut tout d'abord trouver la définition du mot par rapport au contexte. Elle est réalisée à travers une analyse sentimentale dont l'objectif est de juger le sentiment présent dans un écrit.

Il existe de nombreuses manières qui permettent d'effectuer une analyse sentimentale. Par exemple, dans [6], Banerjee et Pedersen font appel à l'algorithme «Lesk» et à ses variantes pour désambiguïser le sens des mots, pour ensuite déterminer le score de sentiment à partir de «SentiWordNet»², un jeu de données qui associe un score de sentiment à chaque définition d'un mot [5].

Il est également possible d'utiliser d'autres modèles d'analyse sentimentale tels que celui de Google pour calculer le score et la polarité d'un sentiment. Le modèle Google est avantageux puisqu'il détecte le contexte de la phrase de manière automatique et existe en langue française. C'est pourquoi nous l'avons considéré pour la mise en place de notre système.

Plusieurs articles décrivant des méthodes d'analyse sentimentale par aspects ont suivi les étapes précitées. A titre d'exemple, dans [22], les auteurs ont fait usage de l'analyse sentimentale afin d'analyser des avis sur des téléphones mobiles, dans [16], la méthode suggérée recommande des restaurants en fonction de l'analyse sentimentale des avis par aspects, et les travaux présentés dans [15] et [20] traitent des données de e-commerce. Les auteurs de [21] ont élaboré une méthode qui intègre la dépendance entre la cible (que nous nommons dans notre article "aspect") et la catégorie d'aspect (que nous appelons famille d'aspect ou classe d'aspect) dans la classification des sentiments. A l'encontre de la méthode suggérée dans [21], l'approche adoptée par notre article se limite à la classification des sentiments basée sur la classe d'aspect et le mot-sentiment.

D'un autre côté, la méthode exposée dans le cadre de nos travaux se positionne dans le contexte de recherches antérieures telles que résumées dans les articles [8] [13]. Ces articles examinent diverses techniques utilisées pour la préparation des données d'entrée [8] et l'extraction des aspects [13] ainsi que la classification des sentiments et l'évaluation des résultats. Pour la préparation des données d'entrée, plusieurs éléments peuvent être extraits tels que le contexte, l'aspect, la catégorie d'aspect et le mot-sentiment. Les techniques employées pour cela incluent le Part-of-speech tagging et le dependency parsing [8]. Concernant la classification des sentiments, nous avons opté pour le modèle de machine learning de Google dans le cadre de ce travail, mais d'autres modèles auraient également pu être envisagés. Des approches alternatives, telles que les modèles basés sur les connaissances ou hybrides [8], auraient pu donner des résultats comparables. L'article [8] suggère également différentes mesures de performance qui peuvent être utilisées pour évaluer les résultats. Dans le cadre de notre article, nous avons choisi d'utiliser l'erreur quadratique moyenne, bien que d'autres mesures auraient pu être prises en considération.

1. Natural Language Toolkit : <https://www.nltk.org/>

2. <https://github.com/aesuli/SentiWordNet/blob/master/papers/LREC06.pdf>

3 Présentation du système

En dépit du fait que plusieurs travaux antérieurs aient proposé des méthodes performantes pour l'analyse sentimentale par aspects, nous avons été contraints de ne pas considérer bon nombre d'entre elles vu qu'elles sont pré-entraînées sur des textes en anglais.

En effet, pour faire usage de ces méthodes, il faut passer par une étape onéreuse de traduction. Par ailleurs, un retour en arrière consistant à récupérer les résultats en français ne sera pas possible avec de telles méthodes.

Ainsi, le système que nous proposons est fondé sur l'utilisation de différents algorithmes gratuits ainsi qu'un algorithme payant de Google pour la notation des sentiments. La figure 1 présente le modèle conceptuel du système.

Dans les sous sections suivantes, nous détaillons les données utilisées ainsi que les différents modèles composant notre système, à savoir le modèle d'extraction des aspects et des mots sentiments, le modèle de regroupement des aspects, et le modèle de calcul du score associé aux aspects.

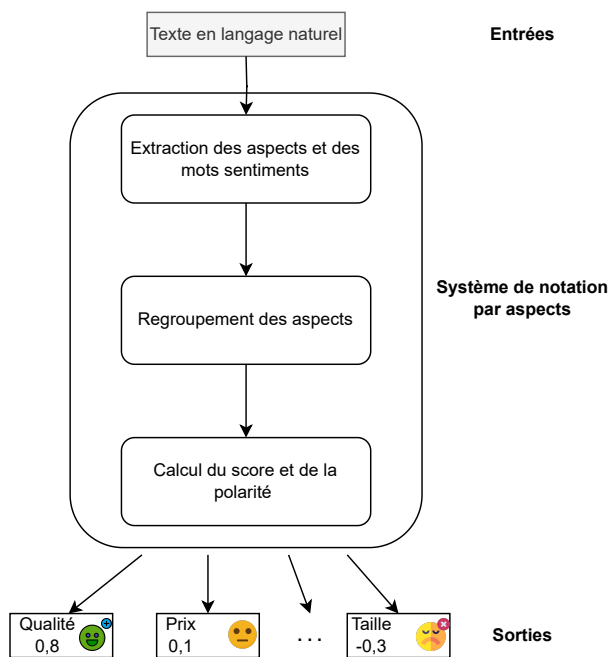


FIGURE 1 – MODÈLE CONCEPTUEL DU SYSTÈME.

3.1 Présentation du corpus de données utilisé

Nous présentons dans cette section les corpus de données existants et celui que nous avons utilisé pour la mise en place de notre système.

3.1.1 Corpus de données existant

Comme présenté dans la section 2, des solutions ont été proposées pour la notation des avis clients en s'appuyant sur des corpus données déjà annotés. Nous pouvons citer le corpus de données « *Amazon Reviews : Unlocked Mobile* » proposé par [17] et repris dans les travaux de [22] pour noter les avis clients sur les téléphones. Ce corpus contient plus

de 400 milles avis clients en anglais, suivants une annotation binaire où 1 correspond à un avis positif et 0 à un avis négatif. Nous pouvons également citer le corpus de données proposé par [5] où l'auteur a manuellement noté les avis clients sur les restaurants suivant un certain nombre d'aspects (e.g., la nourriture, l'environnement, le service). Malheureusement ces données ne sont pas en libre accès.

A notre connaissance, la plupart des corpus de données annotés, comme ceux que nous venons de présenter, permettant d'entraîner des modèles de notation automatique des avis clients qui sont soit en anglais, soit pas en libre accès. D'où l'intérêt de créer notre propre corpus de données en français qui nous servira également pour évaluer les performances de notre système.

3.1.2 Corpus de données utilisé

Les données utilisées pour développer et évaluer notre système sont les avis clients d'un échantillon de 81 produits d'un site de E-commerce français. Les produits sélectionnés pour notre cas d'usage sont des robes. Les 113 premiers avis ont été étiquetés à la main pour pouvoir tester et évaluer nos modèles. Pour ce faire, nous avons lu chaque avis, avons identifié les aspects importants, et avons ensuite attribué un score dans l'intervalle $[-1, 1]$ à chaque aspect identifié en fonction de sa perception de l'expérience client.

Par exemple, sur l'avis client « super jolie mais taille court. », nous avons identifié deux aspects : la « robe » avec le mot sentiment « jolie », et la « taille » avec le mot sentiment « court ». Il leur a ensuite attribué les scores respectifs de 0,9 et de $-0,3$.

Au terme de cette phase d'étiquetage de notre jeu de données, nous avons identifié les 6 principales caractéristiques suivantes, 160 fois :

- robe
- taille
- coupe
- qualité
- tissu
- matière

Quant au modèle de vectorisation, il a été entraîné avec des corpus de données composés de texte qui représentent des actualités en ligne et qui sont en libre accès sur³.

3.2 Modèle d'extraction des aspects et des mots sentiments

Pour effectuer l'analyse sentimentale par aspects, nous avons commencé par extraire les aspects et les mots sentiments des avis. Pour rappel, un aspect est une caractéristique du produit cible qui est dans notre cas d'étude une robe. Des caractéristiques possibles de ce produit sont la qualité, la coupe, le tissu, le prix et la couleur. Pour la mise en place du modèle, ces caractéristiques doivent être définies en amont.

D'un autre côté, les mots sentiments sont les termes qui portent sur le jugement subjectif du créateur de l'avis (e.g., belle, mauvaise, satisfait, etc.). La positivité ou la négativité

3. <https://wortschatz.uni-leipzig.de/en/download/French>

du sentiment associé à ce type de terme peut être généralement notée.

Le modèle d'extraction des aspects repose sur l'analyse des dépendances syntaxique des mots d'un avis et sur un ensemble de règles métiers sous la forme d'un algorithme.

3.2.1 Analyse de dépendances syntaxiques et étiquetage des avis

L'étiquetage morpho-syntaxique a pour objectif d'attribuer une étiquette grammaticale à chaque mot d'une phrase. L'étiquette grammaticale, aussi appelée *Part of Speech* (PoS), est le rôle du mot dans la phrase. Par ailleurs, l'analyse de dépendances permet de déterminer les relations syntaxiques entre les mots.

La revue [1] détaille différentes solutions existantes d'étiquetage morpho-syntaxique et d'analyse de dépendances. Ce travail a mis en avant la pertinence de la bibliothèque Python gratuite « Spacy »⁴ pour une meilleure précision globale. De ce fait, nous avons opté pour cette bibliothèque afin d'aborder la phase d'étiquetage et d'analyse des dépendances syntaxiques.

Spacy utilise des règles spécifiques à chaque langue pour déterminer le rôle de chaque mot. De plus, un système de transition non monotone [12] basé sur l'analyse de dépendances pseudo-projectives [14] est invoqué afin de déterminer les liens syntaxiques.

3.2.2 Algorithme d'extraction

L'algorithme d'extraction permet d'extraire les aspects et les mots sentiments dans un avis client. Pour ce faire, un groupe de règles fondé sur du *PoS tagging* et les dépendances entre les différents mots a été défini. Notre attention s'est portée sur les rôles syntaxique des aspects et les mots sentiments dans les avis clients et les dépendances entre ces derniers.

Le tableau 1 présente quelques exemples de rôles syntaxiques considérés lors de la mise en place de notre algorithme. Cette logique nous permet de considérer un avis comme une suite de *PoS* et de dépendances.

TABLE 1 – EXEMPLE DE DÉFINITION D'ACRONYMES D'ÉTIQUETAGES ET DE DÉPENDANCES SYNTAXIQUES

Acronyme	Signification
ADV	Adverbe
ADJ	Adjectif
NOUN	Nom
CCONJ	conjonction de coordination
amod	Modificateur d'adjectif/ d'un nom
root	Racine de la phrase

Pour illustrer le processus de mise en place de nos règles d'extraction, prenons l'étiquetage morpho-syntaxique associé à l'avis client suivant : « super jolie mais taille court ». Le *PoS tagging* de cet avis est le suivant :

```
super (ADV) : advmod  —>  jolie (ADJ)
jolie (ADJ) : root    —>  jolie (ADJ)
mais (CCONJ) : cc     —>  taille (NOUN)
taille (NOUN) : conj  —>  jolie (ADJ)
court (ADJ) : amod    —>  taille (NOUN)
```

Pour cet exemple nous pouvons remarquer que le *PoS tagging* associé au mot *jolie* est (ADJ) et sa dépendance est « root ». Le tag « root » indique que l'adjectif est directement en racine de la phrase et fait donc référence au produit (robe). Une de nos règles métiers mises en place va extraire *robe* comme aspect (grâce au « root ») et « jolie » comme son mot sentiment. Dans la suite de cet avis le mot *court* dont le tag est (ADJ) a une dépendance de type « amod » avec *taille* dont le tag est (NOUN). Cette dépendance de tag conduit également à une autre règle d'extraction.

En analysant les différents avis clients présents dans notre corpus de données, nous avons établi au total 41 règles métiers qui permettent d'extraire les aspects et les mots sentiments. Ces règles ont été formalisées sous la forme d'un algorithme dont un aperçu est présenté par l'algorithme 1, où w^i désigne un mot présent dans l'avis client. La première condition de l'algorithme 1 correspond à l'extraction de l'aspect *robe* et du mot sentiment *jolie* et la deuxième condition permet l'extraction de l'aspect *taille* et du mot sentiment *court*.

Algorithme 1 : Algorithme d'extraction des aspects et des sentiments

Entrées : les mots w^i des avis clients

Sorties : aspect, sentiment

initialisation

Pour chaque w^i faire ;

si $PoS(w^i) = ADJ$ **alors**

si $dep(w^i) = root$ **alors**

 aspect $\leftarrow robe$;

 sentiment $\leftarrow w^i$;

finsi

sinon

si $dep(w) = amod$ and $PoS(w+1) = NOUN$ **alors**

 aspect $\leftarrow w^{i+1}$;

 sentiment $\leftarrow w^i$

finsi

finsi

L'un des enjeux majeurs auxquels fait face notre algorithme est la difficulté à couvrir un maximum de type d'avis. En effet, les avis des utilisateurs en ligne sont souvent non structurés et syntaxiquement incorrects ce qui compromet souvent l'efficacité d'un tel algorithme. Dans le cas de notre système, une liste de types d'avis et de leurs constructions syntaxiques a été réalisée. Grâce à ce travail préalable, il a été possible de couvrir une vingtaine de types d'avis. Une liste des aspects (robe, taille, matière, coupe, qualité, couleur, tissu) à extraire a été également définie puis utilisée comme paramètre de l'algorithme.

4. Spacy : <https://spacy.io/>

3.2.3 Évaluation du modèle d'extraction d'aspect

Pour évaluer notre modèle d'extraction d'aspect nous utilisons le jeu de données de test présenté dans la sous-section 3.1.2. Pour rappel, ce jeu de données de test est composé de 113 avis clients qui ont été labélisés à la main suivant la méthode présentée dans la sous-section 3.1.2. Dans cette collection de donnée, nous avons pu extraire 6 aspects (robe, taille, matière, coupe, qualité, tissu) 160 fois.

Pour l'évaluation du modèle d'extraction des aspects, nous avons utilisé la justesse globale (OA⁵) et la justesse moyenne (AA⁶) du modèle comme métriques. Selon [11], la justesse globale mesure la proportion de prédictions correctes dans l'ensemble de données de test du modèle. Elle représente le rapport entre le nombre de prédictions correctes et le nombre total d'échantillons d'entrées. Quant à la justesse moyenne, elle permet de mesurer la justesse de prédiction par classe. Les équations 1 et 2 présentent les formules de calcul respectives de la justesse globale et de la justesse moyenne, où # indique le nombre.

$$OA = \frac{\# \text{ d'aspects extraits correctement}}{\# \text{ d'aspects dans le jeu de données}} \quad (1)$$

$$AA = \sum_{i=1}^n \frac{\# \text{ d'aspects extraits correctement de la classe } i}{\# \text{ d'aspects dans la classe } i} \quad (2)$$

En assimilant nos opérations d'extractions à des prédictions, notre modèle présente une justesse globale OA de 1,14 et une justesse moyenne AA de 0,44. La valeur de OA supérieure à 1 indique que le modèle a extrait plus d'aspects lors de la prédiction que celle annotés. En effet l'utilisation de l'algorithme d'extraction sur le jeu de données nous retourne 185 aspects. Cependant la valeurs de AA qui mesure la justesse du modèle par aspect, nous montre que le modèle peine à retrouver les aspects que nous avons défini. Une analyse qualitative de l'extraction automatique montre que des mots tel que « décollété » ou « textile » ont été extraits comme aspects. Même si ces mots ne sont pas les aspects que nous avons définis, elles s'y rapprochent. Par exemple « textile » est proche de « tissu ». D'où la nécessité de mettre en place un modèle de regroupement des aspects qui est présenté dans la sous-section suivante.

3.3 Regroupement des aspects

3.3.1 Le modèle de regroupement des aspects

Comme montré dans la section précédente, le nombre de caractéristiques mentionnées dans les avis au regard d'un produit donné est généralement trop élevé pour réaliser une analyse pertinente. Les synonymes et les aspects similaires doivent être ainsi regroupés. Par exemple, pour les avis sur les robes, les aspects « couleur » et « teinte » doivent être associés. Pour ce faire nous utilisons le plongement lexical. Le plongement lexical ou vectorisation des mots est une technique d'apprentissage qui permet de représenter chaque mot par un vecteur de nombres réels. Cette technique a été

décrite en détails dans [10]. L'avantage d'une telle représentation est qu'elle permet d'établir la proximité contextuelle des mots [15]. En effet, des mots sémantiquement proches auront des représentations vectorielles proches.

Des modèles tels que « Word2vec » [9] ou « BERT » [2] permettent de réaliser le plongement lexical des mots en anglais. Cependant, à notre connaissance, il n'existe pas de modèle gratuit pré-entraîné en français. Par conséquent, nous avons créé notre propre modèle de plongement lexical. Pour ce faire, nous avons utilisé le modèle « Word2vec » que nous avons ré-entraîné avec un corpus de données en français. Comme indiqué dans la section 3.1.2, le corpus de données utilisé pour l'entraînement est composé de sept millions de phrases issues de l'actualité française. Voici un exemple de phrase type de ce corpus : « 20h30 : devant les députés socialistes Après les journaux télévisés, le premier ministre, Myriam El Khomri et Emmanuel Macron se sont retrouvés à l'Assemblée nationale devant les députés du groupe socialiste pour un séminaire ».

L'environnement mis en place pour l'apprentissage est une machine virtuelle Google Cloud à 36 Go de RAM. Le modèle « Word2vec » a été paramétré sur la version « CBOW » avec une taille de vecteur de 300. Le modèle ainsi entraîné est utilisé pour créer le plongement lexical des aspects extraits automatiquement, et par la suite, calculer sa similarité avec un aspect défini au préalable.

La figure 2 présente le principe de fonctionnement du regroupement des aspects.

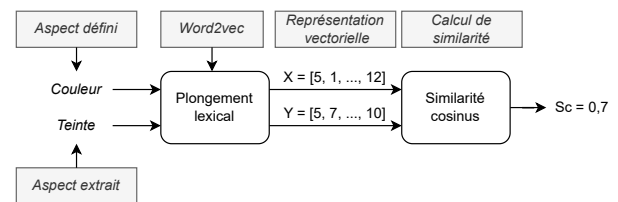


FIGURE 2 – PROCESSUS DE REGROUPEMENT DES ASPECTS.

La fonction de similarité offerte par le modèle « Word2vec » est basée sur le calcul de la similarité cosinus (S_c) entre deux vecteurs. L'équation 3 présente la formule utilisée pour calculer la S_c entre deux vecteurs X et Y [19].

$$S_c(X, Y) = \frac{X^T Y}{\|X\| \|Y\|} \quad (3)$$

où

$$\|X\| = \sqrt{\sum_{i=1}^n x_i^2} \quad (4)$$

et

$$\|Y\| = \sqrt{\sum_{i=1}^n y_i^2} \quad (5)$$

5. Overall accuracy

6. Average Accuracy

Lors de l'entraînement du modèle, nous avons procédé au nettoyage du corpus de texte afin d'améliorer sa qualité, et ce en supprimant la ponctuation. Pour rappel, notre modèle permet de mesurer la proximité entre deux aspects en produisant un score de similarité dans l'intervalle $[-1, 1]$. Pour utiliser le modèle dans notre système, il faut mettre en place un score de similarité minimum α à partir duquel nous pouvons considérer deux mots comme synonymes afin de pouvoir les regrouper.

3.3.2 Choix du seuil de similarité α

Dans l'optique de déterminer le score minimum de similarité α à partir duquel nous pouvons regrouper deux aspects, nous avons réalisé une étude empirique avec notre jeu de données de test présenté dans la sous-section 3.1.2. Nous avons appliqué à notre jeu de données le modèle de regroupement des aspects en faisant varier la valeur de α entre 0, 1 et 1, avec un pas de 0, 1.

Nous considérons le regroupement d'aspects comme un problème de classification multiple, les classes étant les aspects que nous souhaitons regrouper autour des aspects principaux définis dans la sous-section 3.1.2. Ceci nous permet d'évaluer les performances obtenues avec les différentes valeurs de α pour ensuite ne retenir que celle qui nous permet d'avoir les meilleurs résultats. Le tableau 2 présente la valeur de AA obtenue avec les différentes valeurs de α .

Cette étude empirique nous a permis de conclure que la valeur $\alpha = 0, 4$ est la valeur optimale avec laquelle le modèle se trompe le moins sur le regroupement des aspects. Par la suite, nous avons opté pour cette valeur dans la mise en place du système.

TABLE 2 – PERFORMANCE DE LA MÉTHODE EN FONCTION DU PARAMÈTRE DE SIMILARITÉ α

Similarité α	AA
1	0.620
0,9	0.625
0,8	0.626
0,7	0.626
0,6	0.626
0,5	0.626
0,4	0.74
0,3	0.589
0,2	0.589
0,1	0.589

3.4 Calcul du score et de la polarité d'un sentiment

3.4.1 Description du modèle d'analyse des sentiments

Le modèle d'analyse de sentiments a pour objectif d'affecter un score compris entre -1 et 1 à un aspect en fonction du mot sentiment qui lui est associé. Pour trouver le score de sentiment, nous avons fait le choix d'utiliser le modèle « *sentiment analysis* » de Google.

Notons, dans ce cadre, qu'il existe tout un système Google dédié à l'analyse sentimentale par aspect mais, pour des raisons de coût et d'efficacité, nous nous sommes limités à la composante de ce système qui est dédiée à l'analyse sentimentale.

Le modèle « *sentiment analysis* » prend comme entrée un texte et fournit en sortie un score $\beta \in [-1, 1]$. Notre calcul de score et de polarité est fondé sur ce modèle. Dans notre cas d'étude, nous utilisons comme entrée du modèle « *sentiment analysis* » l'aspect et le mot sentiment que nous extrayons de l'avis client grâce à nos modèles présentés dans la section 3.2. La sortie de notre modèle est un score exprimant la polarité associée à l'aspect qui a été précédemment extrait.

3.4.2 Illustration et évaluation du modèle d'analyse de sentiments

Pour illustrer le fonctionnement et évaluer les performances du modèle d'analyse de sentiments, nous utilisons le même jeu de données de tests que celui présenté dans la sous-section 3.3.2.

Le tableau 4 présente l'étiquetage et la notation effectués manuellement sur les avis présentés dans le tableau 3. La colonne *score* représente le score affecté à l'aspect en s'appuyant sur son mot sentiment.

TABLE 3 – EXEMPLES D'AVIS

Id	Avis
1	Très belle robe mais pas une grande qualité
2	La taille est bonne par contre attention mesdames si vous êtes grande ! Elle est hyper courte ... J'ai du la renvoyer car ça faisait vraiment trop court. Damage

TABLE 4 – ÉTIQUETAGE À LA MAIN

Id	Aspect	Mot-sentiment	Score
1	robe	belle	0,9
1	qualité	grande	-0,6
2	taille	bonne	0,8
2	robe	courte	-0,6

A partir de ces exemples, nous remarquons que le premier avis est un avis facile pour l'analyse sentimentale. En effet, la construction syntaxique de la phrase est juste et précise. En revanche, le deuxième avis est plus complexe au regard de notre modèle vu que la phrase n'est ni en français correct ni précise.

Dans le cas de l'étiquetage à la main, l'auteur est capable de comprendre que le premier avis exprime deux polarité d'opinions différentes. L'opinion positive est sur la robe en général et l'opinion négative est sur la qualité. L'auteur est aussi capable d'estimer que l'opinion positive sur la robe est plus forte que celle négative sur la qualité et ce, au travers de l'usage du superlatif « très ». Quant au deuxième avis, il est plus difficile à analyser. En effet, deux opinions opposées sont exprimées sur ce qui semble être la même

caractéristique. L'opinion négative « hyper courte » semble porter sur la robe dans le cas où la personne est de grande taille.

Le tableau 5 présente les résultats de l'analyse sentimentale associée aux deux exemples du tableau 3.

TABLE 5 – PRÉDICTIONS DE LA MÉTHODE PROPOSÉE

Id	Aspect	Mot-sentiment	Score
1	robe	belle	0,9
1	qualité	grande	-0,4
2	taille	bonne	0,9
2	robe (mesdames)	grande	-0,6

Le système est capable d'extraire les aspects et sentiments et de les noter correctement pour le premier avis. Cependant, pour le deuxième exemple, la solution proposée fonctionne moins bien. En effet, le modèle d'analyse sentimentale est programmée pour sélectionner seulement le sentiment, s , le plus fort sur un même aspect :

$$s = \max(|score(mot^i)|) \quad (6)$$

D'un autre côté, on s'aperçoit, par le biais des exemples, que le mot « mesdames » a été extrait comme un synonyme de robe par le modèle d'extraction des aspects. Cette erreur provient du modèle de regroupement « Word2Vec ». Ceci s'explique par le fait que, dans le jeu de données d'entraînement les mots « mesdames » et « robes » doivent être souvent associées. L'aspect robe aurait dû être extrait pour « elle est hyper courte » et non pas pour « mesdames si vous êtes grande ! ». Une des limites de l'analyse d'avis consommateurs en ligne est que ces avis peuvent être syntaxiquement faux, ce qui rend leur analyse très difficile, voire impossible.

Nous avons utilisé l'erreur quadratique moyenne (EQM), exprimée par l'équation 7, pour évaluer la performance de l'analyse de sentiment de notre jeu de test, où \hat{y}_i sont les scores prédits par le modèle, y_i sont les scores attribuées par l'auteur, et n représente le nombre d'aspects.

$$EQM = \frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n} \quad (7)$$

L'erreur quadratique moyenne permet de mesurer les écarts entre les valeurs réellement observées y_i et les valeurs prédites par le modèle \hat{y}_i . Plus elle est proche de 0 plus le modèle est précis.

Nous avons obtenu une $EQM = 0,19$ pour notre jeu de données de test. Ce qui nous permet d'affirmer que nos résultats sont prometteurs.

4 Discussion et applications pratiques du système

Dans cet article, nous proposons un système de notation des avis clients par aspects pour les produits des sites de e-commerce afin d'aider les enseignes à améliorer leur offres de produits et de services, et également guider les clients potentiels à affiner leurs choix lors de leurs achats.

Quoiqu'il existe des systèmes adressant la même problématique que la nôtre, ces derniers tendent à fournir une note globale sur un avis client, et les rares proposant des notes par aspects sont destinés aux avis clients rédigés en anglais uniquement [20].

Le système que nous proposons est destiné aux avis clients rédigés en français et est capable d'extraire automatiquement les aspects importants ayant été prédéfinis au préalable. Les enjeux majeurs actuels autour de la satisfaction client, dont les attentes sont de plus en plus hautes, et la concurrence forte entre les différentes enseignes de commerce montrent la nécessité d'un tel système.

Notre système est fondé sur des techniques de traitement automatique du langage naturel et sur de l'analyse sentimentale. Il est composé de trois modèles principaux : (a) un modèle d'extraction des aspects d'un produit ainsi que des mots sentiments reflétant l'opinion sur ces caractéristiques, (b) un modèle de regroupement d'aspects qui permet de regrouper les caractéristiques extraites d'un avis dans des classes, et (c) un modèle de calcul du score et de la polarité d'un sentiment qui permet d'attribuer un score entre -1 et 1 à un aspect extrait.

L'ensemble de ces modèles a été implémenté et évalué à partir d'un jeu de données de tests qui a été décrit dans la sous-section 3.1.2. Les résultats de l'évaluation des modèles sur notre cas d'usage (notation des robes) montrent des résultats prometteurs pour une future industrialisation. Un cas d'application pratique de notre système a été effectué sur un jeu de données composé de 1944 avis clients suivants 77 types de robes. La figure 3 illustre la sortie de notre système sous forme de diagramme radars pour 3 robes suivant 8 caractéristiques, à savoir la matière, la robe, la couleur, la taille, la qualité, le tissu, et la coupe. Les notes pour chaque aspect représentent la moyenne en fonction des différentes occurrences présentées par la figure 4.

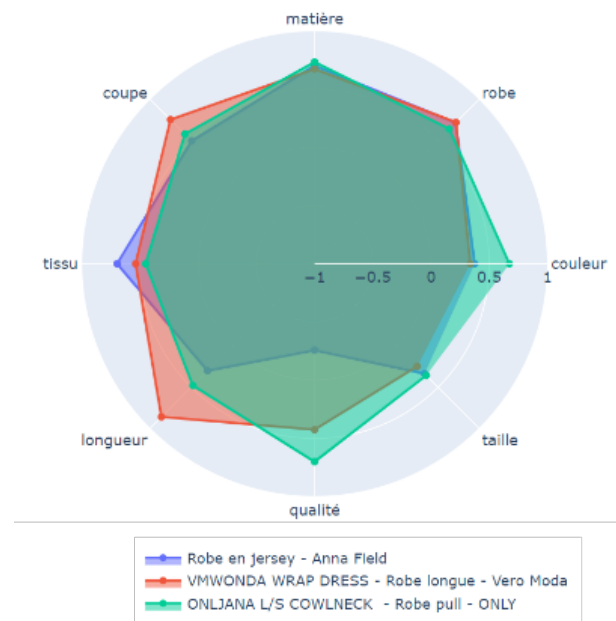


FIGURE 3 – DIAGRAMME RADAR DES ROBES.

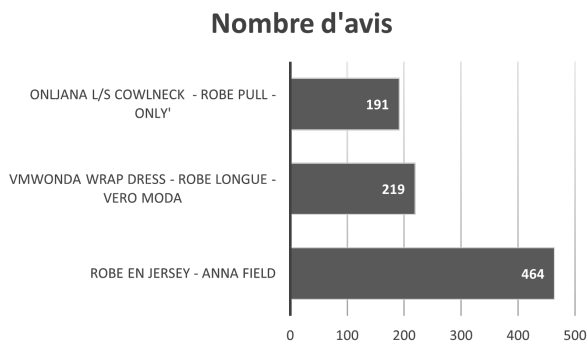


FIGURE 4 – OCCURRENCES DES ROBES DANS LES AVIS.

Ce diagramme radar offre une visibilité sur les différentes notes associées aux principales caractéristiques des robes dans les avis clients. Il représente un bon indicateur sur lequel les clients potentiels pourraient se baser pour effectuer leur choix ou encore les enseignes pour améliorer un des aspects sur lequel ils sont moins compétitifs.

5 Conclusion, limites et perspectives

Le système de notation par aspects présenté dans cet article est un système prometteur qui permet de traiter des textes en français sans avoir recours à un traducteur.

Deux populations peuvent être visées par ce système : les futurs consommateurs afin de pouvoir affiner leurs choix de produits selon leurs préférences, et les enseignes des produits notés afin d'améliorer leurs offres en considérant les différents aspects auxquels les consommateurs se sont intéressés.

Cependant, la performances encourageante du système proposé ne doit pas être un frein pour des améliorations futures. En effet, pour améliorer cette solution, un modèle de notation des sentiments en français pourra être entraîné. Cela permettra à notre système d'être entièrement indépendant au regard de tout modèle payant. L'utilisation du modèle Google a un coût d'environ 0,0010€ par 1000 caractères. C'est le seul modèle payant de notre algorithme. En outre, il est possible d'augmenter le nombre de types d'avis couverts par la méthode ce qui entraîne une véritable valeur commerciale pour le système suggéré.

Finalement, nous pourrions également envisager des perspectives dont l'objectif sera de comparer la méthode présentée dans cet article avec les méthodes récentes d'IA Générative (GPT, BARD, GROK, MISTRAL) en utilisant les jeux de données annotés SemEval.

Remerciements

Merci à notre entreprise, Teolia⁷, de nous avoir soutenu dans ce projet de recherche. Un grand merci, en particulier, à tous les consultants Data de Teolia qui ont participé au développement des différents algorithmes.

7. <https://www.teolia.fr/>

Références

- [1] Fouad Nasser A Al Omran and Christoph Treude. Choosing an nlp library for analyzing software documentation : a systematic literature review and a series of experiments. In *2017 IEEE/ACM 14th international conference on mining software repositories (MSR)*, pages 187–197. IEEE, 2017.
- [2] Shivaji Alaparathi and Manit Mishra. Bert : A sentiment analysis odyssey. *Journal of Marketing Analytics*, 9(2) :118–126, 2021.
- [3] Karim Baati. Hybridization of adaboost with random forest for real-time prediction of online shoppers' purchasing intention. In *Hybrid Intelligent Systems : 20th International Conference on Hybrid Intelligent Systems (HIS 2020), December 14-16, 2020*, pages 234–241. Springer, 2021.
- [4] Karim Baati and Mouad Mohsil. Real-time prediction of online shoppers' purchasing intention using random forest. In *Artificial Intelligence Applications and Innovations : 16th IFIP WG 12.5 International Conference, AIAI 2020, Neos Marmaras, Greece, June 5–7, 2020, Proceedings, Part I 16*, pages 43–51. Springer, 2020.
- [5] Stefano Baccianella, Andrea Esuli, Fabrizio Sebastiani, et al. Sentiwordnet 3.0 : an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec*, volume 10, pages 2200–2204, 2010.
- [6] Satanjeev Banerjee and Ted Pedersen. An adapted lesk algorithm for word sense disambiguation using wordnet. In *Computational Linguistics and Intelligent Text Processing : Third International Conference, CILCling 2002 Mexico City, Mexico, February 17–23, 2002 Proceedings 3*, pages 136–145. Springer, 2002.
- [7] Bhavana R Bhamare, P Jeyanthi, and R Subhashini. Aspect level sentiment analysis approaches. In *2019 5th International Conference On Computing, Communication, Control And Automation (ICCCUBEA)*, pages 1–5. IEEE, 2019.
- [8] Gianni Brauwere and Flavius Frasinca. A survey on aspect-based sentiment classification. *ACM Computing Surveys*, 55(4) :1–37, 2022.
- [9] Kenneth Ward Church. Word2vec. *Natural Language Engineering*, 23(1) :155–162, 2017.
- [10] Yoav Goldberg and Omer Levy. word2vec explained : deriving mikolov et al.'s negative-sampling word-embedding method, 2014.
- [11] Ahmed Taha Haouari, Labiba Souici-Meslati, Fadila Atil, and Djamel Meslati. Empirical comparison and evaluation of artificial immune systems in inter-release software fault prediction. *Applied Soft Computing*, 96 :106686, 2020.
- [12] Matthew Honnibal and Mark Johnson. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 conference on empi-*

tical methods in natural language processing, pages 1373–1378, 2015.

- [13] Ambreen Nazir, Yuan Rao, Lianwei Wu, and Ling Sun. Issues and challenges of aspect-based sentiment analysis : A comprehensive survey. *IEEE Transactions on Affective Computing*, 13(2) :845–863, 2020.
- [14] Joakim Nivre, Johan Hall, Jens Nilsson, Gülşen Eryiğit, and Svetoslav Marinov. Labeled pseudo-projective dependency parsing with support vector machines. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 221–225, 2006.
- [15] Debanjan Paul, Sudeshna Sarkar, Muthusamy Cheliah, Chetan Kalyan, and Prajit Prashant Sinai Nadkarni. Recommendation of high quality representative reviews in e-commerce. In *Proceedings of the eleventh ACM conference on recommender systems*, pages 311–315, 2017.
- [16] IKCU Perera and HA Caldera. Aspect based opinion mining on restaurant reviews. In *2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCI)*, pages 542–546. IEEE, 2017.
- [17] M Rathan, Vishwanath R Hulipalled, KR Venugopal, and LM Patnaik. Consumer insight mining : aspect based twitter opinion mining of mobile phone reviews. *Applied Soft Computing*, 68 :765–773, 2018.
- [18] Kim Schouten and Flavius Frasincar. Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 28(3) :813–830, 2015.
- [19] Sergios Theodoridis and Konstantinos Koutroumbas. *Pattern recognition*. Elsevier, 2006.
- [20] Satuluri Vanaja and Meena Belwal. Aspect-level sentiment analysis on e-commerce data. In *2018 International conference on inventive research in computing applications (ICIRCA)*, pages 1275–1279. IEEE, 2018.
- [21] Hai Wan, Yufei Yang, Jianfeng Du, Yanan Liu, Kunxun Qi, and Jeff Z Pan. Target-aspect-sentiment joint detection for aspect-based sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9122–9129, 2020.
- [22] Ye Yiran and Sangeet Srivastava. Aspect-based sentiment analysis on mobile phone reviews with lda. In *Proceedings of the 2019 4th International Conference on Machine Learning Technologies*, pages 101–105, 2019.

Détection automatique de similarités dans le cadre de la protection par Marque

A. Jourdan¹, Y. Le Nir¹, N. Girardin²

¹ CYU Cergy Paris University, ETIS

² LIPSTIP

astrid.jourdan@cyu.fr, yannick.le-nir@cyu.fr, ngirardin@lipstip.com

Résumé

Une Marque permet de protéger un signe pour des produits et services décrits sous la forme de mots ou de groupe de mots appelés items. Lorsqu'un titulaire de marque identifie une marque similaire à sa marque enregistrée, il peut former une opposition. Les produits et services de chaque marque étant décrits par des centaines d'items, l'opposition nécessite une analyse manuelle complexe de la similarité entre produits et services des deux marques. L'objectif est de proposer une méthode automatique pour trouver les similarités entre marques en utilisant les embeddings des LLMs sur les items. La mesure de similarité cosinus est ensuite intégrée avec d'autres features dans un processus d'apprentissage supervisé pour prédire la similitude. Un historique de plus de 120 000 décisions des Offices de Marques Européens permet d'évaluer la qualité des résultats

Mots-clés

Propriété Intellectuelle, Marques, Embeddings, Similarités, Apprentissage supervisé

Abstract

A trademark protects a sign for goods and services, which are described in words or groups of words called items. When a trademark owner identifies a mark similar to his registered trademark, he can file an opposition. Since the goods and services of each trademark are described by hundreds of items, opposition requires a complex manual analysis of the similarity between the goods and services of the two trademarks. The aim is to propose an automatic method for finding similarities between trademarks using LLM embeddings on items. The cosine similarity measure is then integrated with other features in a supervised learning process to predict similarity. A history of over 120,000 European Trademark Office decisions is used to assess the quality of the results

Keywords

Intellectual property, Trademarks, Embeddings, Similarities, Supervised learning

1 Introduction

La Marque est un titre de propriété industrielle permettant une fois enregistré de protéger un signe (par exemple un nom, un logo, ...) pour des produits et services spécifiques dans une liste d'items (mots ou groupe de mots). Par exemple, la marque enregistrée SO'NUM de Digital Aquitaine protège le signe SO'NUM pour une utilisation dans environ 50 produits et services comme : Matériel informatique; logiciels; Produits de l'imprimerie; Conseil et accompagnement en matière de développement de l'économie numérique et de la digitalisation des entreprises.

Une opposition de marque vise à démontrer qu'une marque nouvellement créée porte atteinte à une ancienne marque. Pour cela, une partie essentielle vise à démontrer dans quelle mesure deux marques sont similaires ou susceptibles de créer une confusion chez les consommateurs. Cela implique notamment une analyse complexe du chevauchement potentiel en termes de produits et services. Par exemple, il faut déterminer en quoi l'item "appareils et instruments photographiques" est similaire à l'item "appareils pour la transmission du son" alors que l'item "stylos à bille" est différent de l'item "billes de stylo". Une opposition de marque peut nécessiter la comparaison de centaines de produits et services, sur la base de critères d'évaluation et de la jurisprudence (les décisions passées des offices d'enregistrement ou tribunaux). Par exemple, la comparaison de deux marques protégeant chacune 50 produits et services nécessite l'évaluation de 2500 couples d'items. Les praticiens spécialisés dans l'opposition de marque ont besoin d'outils spécifiques pour comparer rapidement des produits et services.

Il existe de nombreuses études utilisant l'intelligence artificielle pour étudier la similitude entre marques. La plupart concerne la comparaison des logo ([9], [1], [8], [13], [2]) ou des noms d'un point de vue syntaxique, sémantique ou phonétique, ou une combinaison des deux ([11], [10], [3], [12]).

À notre connaissance seuls Shmatkov et al. [7] se sont intéressés à la comparaison des produits et des services. Leur approche consiste à vectoriser les items puis utiliser une mesure de similarité cosinus ou distance de Tanimoto. Les modèles de vectorisation (*embeddings*) utilisés

sont spécifiques à la vectorisation de mots et n'intègrent que très peu le contexte de la phrase. La comparaison des items repose entièrement sur la mesure de similarité calculée et n'intègre aucune autre information telle que l'office d'enregistrement ou l'origine de la décision. Leur étude porte exclusivement sur la classe 45 de la classification de Nice. Cette classification a été instituée par l'Arrangement de Nice en 1957, et une nouvelle édition est publiée tous les cinq ans. Il s'agit d'une classification internationale de produits et de services aux fins de l'enregistrement des marques. Elle comprend actuellement 45 catégories.

Dans ce travail, nous proposons d'utiliser des *embeddings* spécifiques aux phrases grâce aux récents modèles basés sur les *Transformers* [6], puis d'intégrer la mesure de similarité calculée dans un modèle d'apprentissage de classification avec d'autres caractéristiques sur les décisions dans le but de déterminer la similarité ou dissimilarité des items pour toutes les catégories de la classification de Nice. La section 2 présente le jeu de données utilisé. Dans la section 3, nous comparons les différents modèles de *sentence embeddings* pour en extraire des similarités sémantiques des items décrivant les services ou produits. La section 4 décrit le processus d'apprentissage automatique et compare les performances des modèles de classification utilisés. Une analyse des résultats est effectuée dans la section 5, puis nous terminons avec quelques perspectives pour des travaux futurs.

2 Présentation et pré-traitement des données

L'office de marques européen EUIPO met à disposition des données concernant des pratiques ou décisions sur des paires d'items prises par différents offices européens (<https://euipo.europa.eu>) sur la période 2012 à 2023. Le jeu de données constitué regroupe 107570 pratiques et décisions renseignées par :

- un couple d'items décrivant les produits ou les services,
- la catégorie de la classification de Nice de chacun des deux items (45 valeurs),
- une décision de degré de similitude donnée par un tribunal ou un office d'enregistrement de marques,
- le tribunal ou l'office d'enregistrement des marques ayant rendu le verdict (17 valeurs),
- l'origine de la décision, e.g. Tribunal de justice, Chambre de recours, Pratique de l'office,... (6 valeurs).

Une information supplémentaire est disponible dans le jeu de données concernant le raisonnement justifiant la décision. L'objectif étant d'automatiser le travail de l'expert, cette information n'a pas été retenue pour la suite de l'étude mais sera utilisée pour expliquer certains résultats.

Le degré de similitude donnée par l'office des marques comprend 6 niveaux : Identique, Degré élevé, Similaire, Degré faible, Dissemblable. L'objectif est d'aider l'expert en proposant un premier tri automatique permettant de détecter une possible similitude, et ce, quel que soit son degré. Pour cela nous créons une variable Similitude avec deux

Nom du modèle	Performance	Taille (Mo)
all-mpnet-base-v2	69.57	420
all-distilroberta-v1	68.73	290
all-MiniLM-L12-v2	68.70	120
all-MiniLM-L6-v2	68.06	80

TABLE 1 – Modèles d'*embeddings*

classes :

- Semblable si la décision est Identique, Degré élevé, Similaire, Degré faible
- Dissemblable sinon

Les deux classes sont déséquilibrées : Semblable = 15846, Dissemblable = 91724. Pour équilibrer les classes, 15846 décisions sont choisies aléatoirement dans la classe Dissemblable. Le jeu de données est ainsi réduit à 31692 décisions avec autant d'observations dans chacune des deux classes. L'objectif est de construire un modèle permettant de déterminer la variable Similitude à partir des autres caractéristiques du jeu de données.

3 Similarité sémantique

Les items décrivant les services ou les produits sont des données textuelles qu'il est nécessaire de traiter au préalable. Une approche classique consiste à utiliser une représentation vectorielle des items (*embeddings*) puis à en extraire une mesure de similarité. La plus utilisée est la mesure de similarité cosinus,

$$\text{cosinusSim}(E_{t_i}, E_{t_j}) = \frac{E_{t_i} \cdot E_{t_j}}{\|E_{t_i}\| \|E_{t_j}\|},$$

où E_{t_i} et E_{t_j} sont les *embeddings* des paires d'items t_i et t_j présentes dans le jeu de données.

Concernant l'*embedding* des items, il existe une multitude de modèles. Nous avons mené une étude comparative de différents modèles afin de déterminer leur impact sur la mesure de similarité des items. Nous avons sélectionné quatre modèles (distilroberta, llm_minilm6, minilm12, mpnet) depuis la librairie *SentenceTransformers* issue des travaux de [6]. Ils ont été pré-entraînés sur plus d'un milliard de paires d'apprentissage. Ils sont considérés comme les plus performants pour la tâche de vectorisation d'une phrase et ils sont de taille variable comme le montre le tableau 1.

Nous avons également testé deux autres modèles plus légers, word2vec (w2v) pré-entraîné sur GoogleNews [4] et un autre LLM, minillm3, entraîné à détecter des paraphrases sur un jeu de données de moins de 100 millions de paires, et ne faisant que 61 Mo. Le modèle word2vec donnant des *embeddings* de mots, nous avons pris la moyenne des vecteurs des mots constituant l'item.

La figure 1 montre que tous les modèles permettent d'obtenir une mesure de similarité cosinus plus élevée pour la classe Semblable, ce qui confirme qu'il est pertinent d'utiliser cette mesure pour discriminer les items semblables et dissemblables. L'étalement du cosinus pour la classe Semblable est certainement dû au regroupement des classes initiales (Identique, Similaire, Degré élevé, Degré faible).

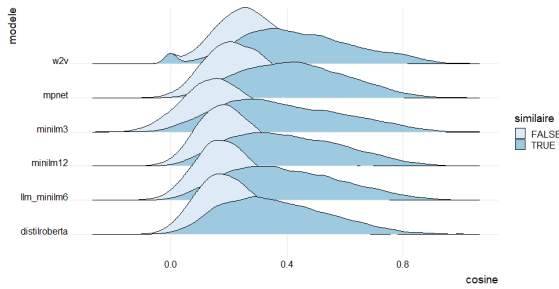


FIGURE 1 – Distribution du cosinus en fonction du modèle et des classes Semblable (TRUE) et Dissemblable (FALSE)

Pour choisir le meilleur *embedding*, on compare leur score de prédiction en comparant la classe observée et la classe obtenue en seuillant le cosinus. Les figures 2, 3, 4 et 5 montrent les scores obtenus en fonction du seuil.

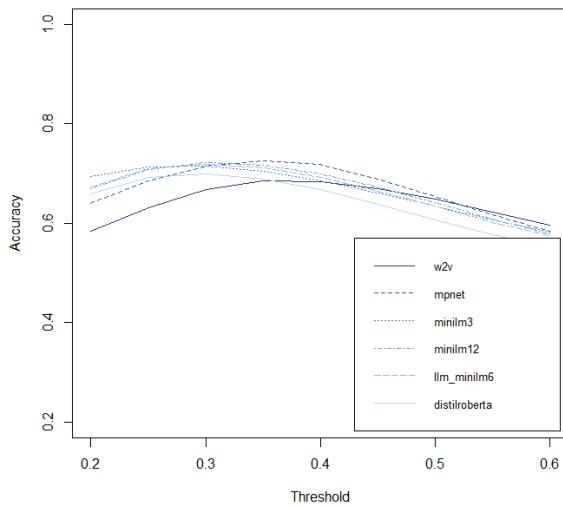


FIGURE 2 – Exactitude en fonction du seuil

On note que les modèles w2v et mpnet donnent un taux de bien classés légèrement meilleur pour la classe Semblable (*Recall*) mais sont nettement moins performants pour la classe Dissemblable (*Specificity*). Or seuls les items détectés semblables bénéficieront d'un traitement manuel. Il est donc primordial de bien prédire les items dissemblables car ils seront écartés de l'étude. Les autres modèles *Transformers* donnent des résultats identiques. On retiendra le plus léger et rapide minilm3, qui nous permettra de comparer efficacement un grand volume de paires d'items.

Le meilleur compromis entre taux de bien classés pour la classe Semblable et la classe Dissemblable est atteint pour un seuil de 0.275 comme le montre la figure 6.

Avec ce seuil, le modèle minilm3 permet d'atteindre un score de 70% de bien classés quelle que soit la classe (Ligne "Sans modèle" du tableau 2).

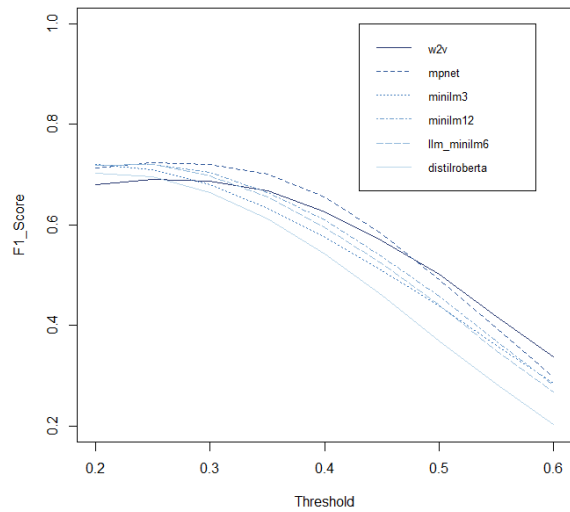


FIGURE 3 – F1_score en fonction du seuil

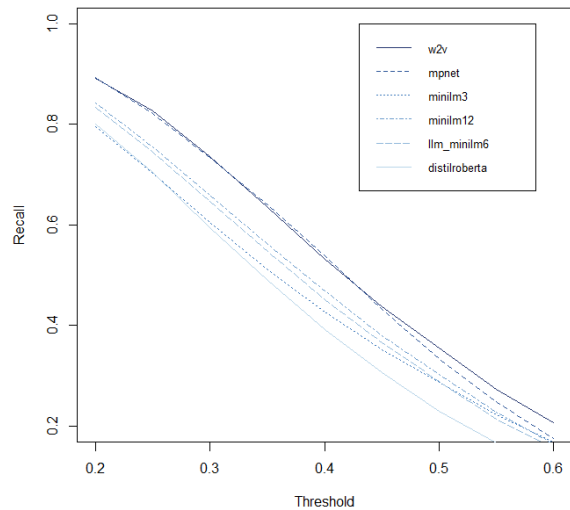


FIGURE 4 – Rappel en fonction du seuil

4 Modèles d'apprentissage

La détection d'items similaires repose essentiellement sur la similitude sémantique des items mais peut aussi être influencée par des paramètres complémentaires tels que l'office qui a rendu la décision ou les catégories de la classification de Nice des items. L'objectif de ce paragraphe est de déterminer un modèle d'apprentissage de classification de la variable Similitude avec en entrée la mesure de similarité cosinus, l'office, l'origine, les catégories de la classification de Nice des items 1 et 2 et une nouvelle variable binaire indiquant si la catégorie est identique pour les deux items. Le schéma du processus mis en place est donné dans la figure 7.

Les variables office, origine, catégorie de la classification de Nice de l'item 1 et de la classification de Nice de l'item 2 sont catégorielles avec 6, 17, 45 et 45 modalités respec-

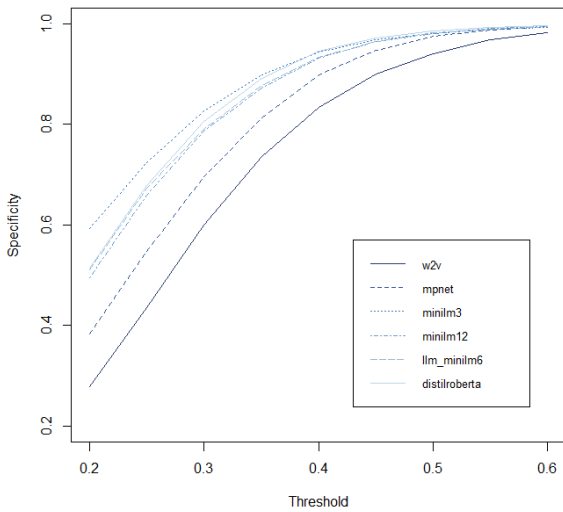


FIGURE 5 – Specificité en fonction du seuil

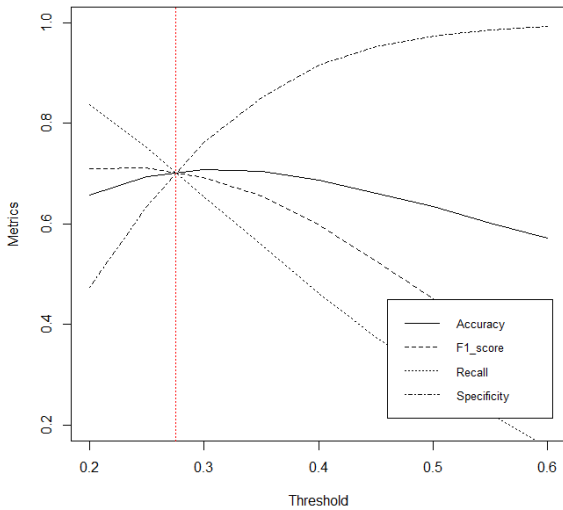


FIGURE 6 – Métriques en fonction du seuil pour le modèle minilm3

tivement. Chaque modalité est encodée en binaire pour obtenir un nombre total de 115 variables en entrée. Il existe un descriptif textuel des catégories de la classification de Nice. Afin de ne pas alourdir le modèle, nous avons préféré ne pas utiliser de modèle d’*embedding* et de traiter ces variables comme catégorielles.

Les modèles d’apprentissage utilisés sont un arbre de décision avec une profondeur maximale de 10, une régression logistique, une forêt aléatoire avec 100 arbres et 2 variables testées à chaque nœud, un réseau de neurones (MLP) avec une couche cachée constitué de 5 neurones (586 poids) et un autre avec 15 neurones (1756 poids) avec la fonction d’activation sigmoïde.

Les modèles sont construits sur une base d’apprentissage contenant 70% du jeu de données choisi aléatoirement, et sont comparés sur une base de test avec les 30% restants.

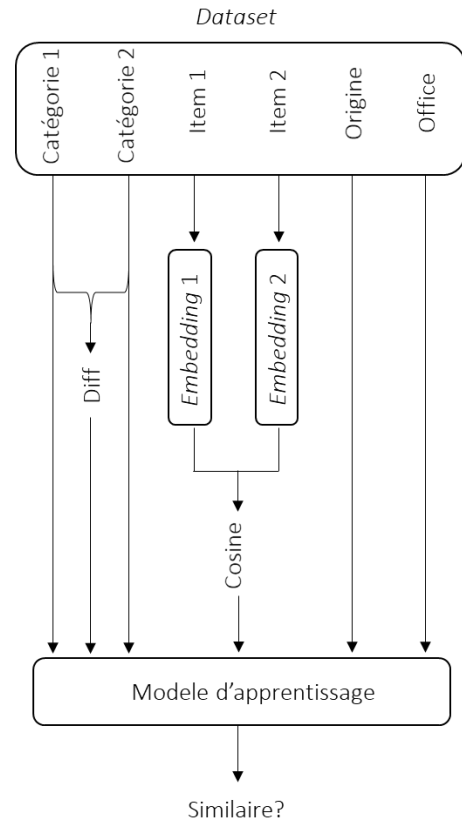


FIGURE 7 – Schéma général

Il y a trois niveaux d’aléa dans le processus : au niveau du re-équilibrage des classes, lors de la construction des bases d’apprentissage et de test, le caractère stochastique de certains modèles (Forêts aléatoires, MLP). Afin de quantifier l’incertitude due à ces aléas et vérifier la robustesse des résultats, le processus est répété 20 fois. Les figures 8, 9, 10 et 11 représentent la distribution des 20 métriques obtenues pour chaque modèle et le tableau 2 donne leurs valeurs moyennes. Les boxplots permettent d’appréhender la robustesse des résultats qui restent très stables à l’exception du rappel et de la spécificité par l’arbre de décision.

Modèle	Exact	F1	Rappel	Spec
Sans modèle	0.71	0.71	0.71	0.70
Arbre de décision	0.79	0.78	0.76	0.81
Régression logistique	0.83	0.83	0.81	0.85
Forêt aléatoire	0.85	0.85	0.84	0.86
MLP (5)	0.84	0.84	0.83	0.85
MLP (15)	0.84	0.84	0.83	0.85

TABLE 2 – Tableau des valeurs moyennes des métriques

A l’exception de l’arbre de décision, tous les modèles permettent d’améliorer le résultat obtenu sans modèle de 10 à 15% sur toutes les métriques. Le test statistique de l’analyse de la variance montre clairement qu’il n’y a pas de différence significative pour les différentes métriques entre

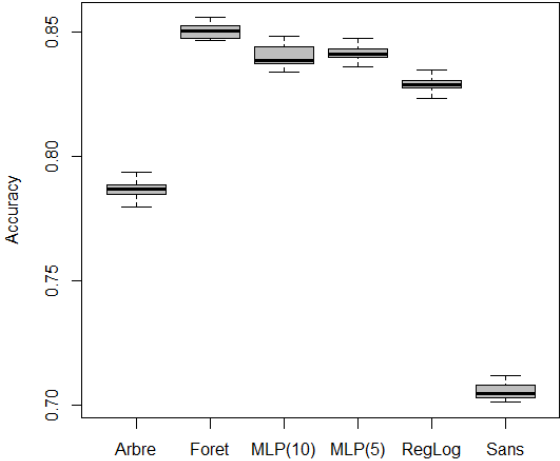


FIGURE 8 – Boxplots de l’exactitude

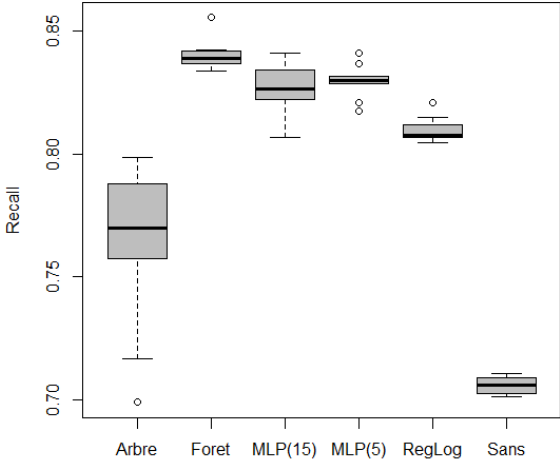


FIGURE 10 – Boxplots du rappel

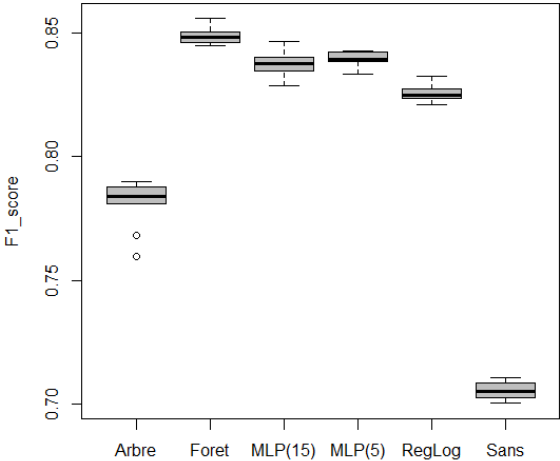


FIGURE 9 – Boxplots du F1_score

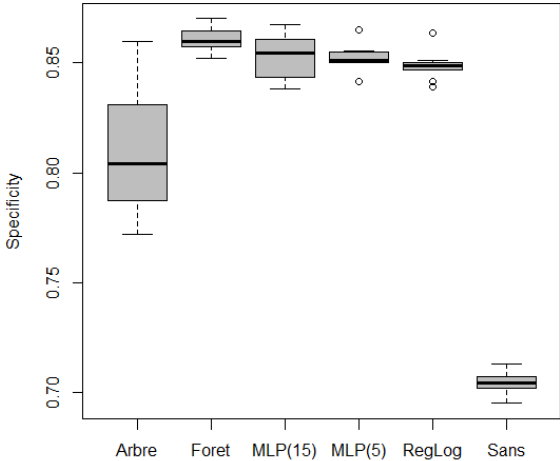


FIGURE 11 – Boxplots de la Spécificité

la régression logistique, la forêt aléatoire et les deux réseaux de neurones (p-valeur inférieure à 10^{-11}). Cependant graphiquement, on peut voir des métriques légèrement plus élevées pour la forêt aléatoire. De plus ce modèle a l'avantage d'être partiellement interprétable (figure 12). Bien que moins performants que la régression logistique, la forêt aléatoire ou les réseaux de neurones, l'arbre de décision améliore aussi les métriques. C'est un modèle qui reste intéressant car il permet d'expliquer entièrement la classification d'une nouvelle instance.

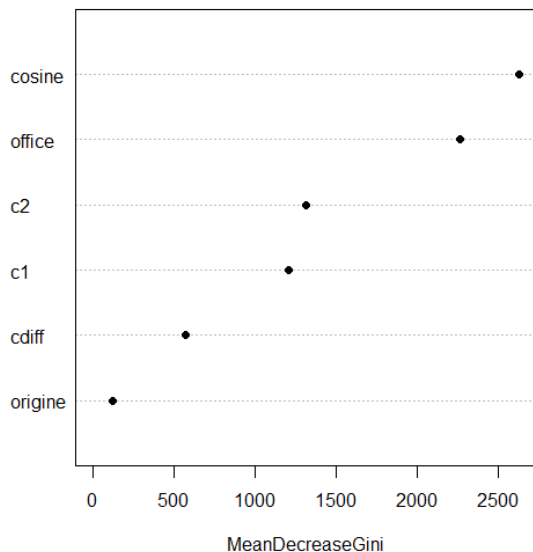


FIGURE 12 – Importance des variables dans la forêt aléatoire

5 Analyse des résultats

Impact de l'office. L'importance des variables de la forêt aléatoire montre que les variables déterminantes pour la classification sont le cosinus et l'office, suivies dans une moindre mesure par les catégories de la classification de Nice des deux items (figure 12). Le fait que la variable office ait une importance élevée pourrait indiquer un biais selon qui a rendu le jugement. Nous avons donc procédé à une validation croisée en entraînant le modèle sur 16 offices et en testant sur le 17ème office. Les résultats présentés dans la figure 13 montrent que cela n'est pas le cas. Ils sont à peu près stables sauf pour la Hongrie, l'Estonie et la Bulgarie. Ces derniers présentent une dissymétrie entre le rappel et la spécificité avec un rappel assez bas (50%-70%). Cela peut s'expliquer par une arrivée plus tardive de ces pays dans l'Union Européenne, donc moins de temps pour harmoniser les pratiques.

Erreur selon la classification de Nice. La répartition des erreurs au sein des catégories de la classification de Nice (Figure 14) montre un déséquilibre.

Cinq catégories présentent une erreur inférieure à 5% dont les trois meilleures : 15 (*Musical instruments ; music stands and stands for musical instruments ; measuring sticks*), 27

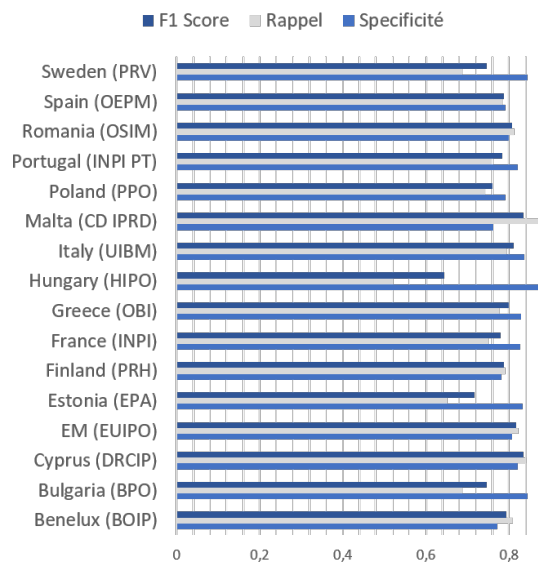


FIGURE 13 – Métriques par Office

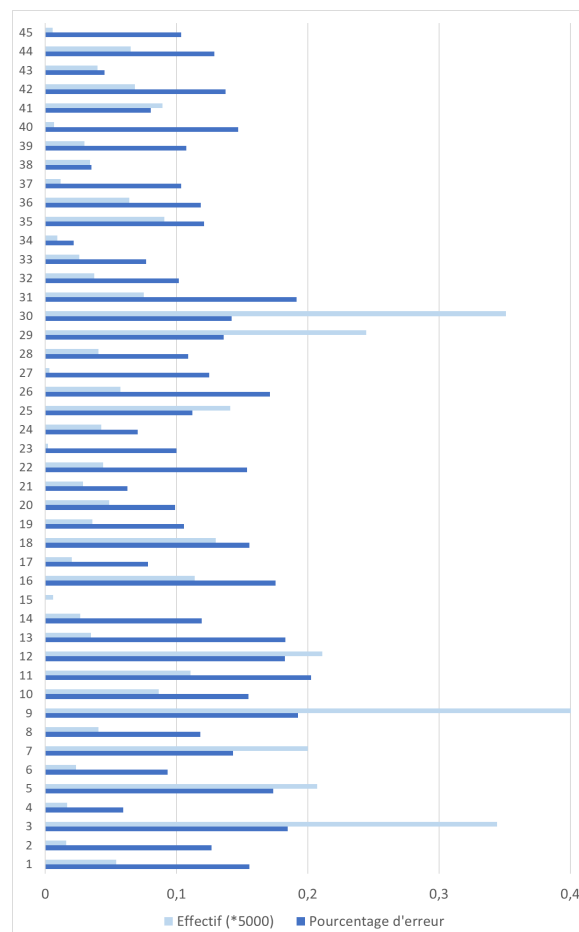


FIGURE 14 – Pourcentage d'erreur par catégorie

(*Carpets, mats, linoleum and other floor coverings ; wall hangings, not of textile materials*), 34 (*Tobacco and to-*

bacco substitutes; cigarettes and cigars; electronic cigarettes and smokers' sprays; smokers' articles; matches). En revanche, une dizaine de catégories ont un taux d'erreur supérieur à 15%, dont les trois premières : 9 (*Scientific, research, navigational, geodetic, photographic, cinematographic, audiovisual, optical, weighing, measuring, signaling, detection, testing, inspection, rescue and teaching apparatus and instruments; apparatus and instruments for conducting, distributing, transforming, accumulating, regulating or controlling the distribution or consumption of electricity; apparatus and instruments for recording, transmitting, reproducing or processing sound, images or data; recorded or downloadable media, software, blank digital or similar recording and storage media; mechanisms for prepayment machines; cash registers, calculating devices; computers and computer peripherals; diving suits, diving masks, diving ear plugs, nose clips for divers and swimmers, diving gloves, breathing apparatus for underwater swimming; fire extinguishers.*), 11 (*Appliances and installations for lighting, heating, cooling, steam production, cooking, drying, ventilation, water distribution and sanitary installations*), 31 (*Raw and unprocessed agricultural, aquacultural, horticultural and forestry products; raw and unprocessed seeds; fresh fruit and vegetables, fresh herbs; natural plants and flowers; bulbs, seedlings and seeds; live animals; animal feed and beverages; malt.*). Les catégories présentant un fort taux d'erreur semblent recouvrir un plus large champ de produits ou de services que celles avec une faible erreur. Cela pourrait expliquer les difficultés du modèle à être performant sur des catégories très larges. Une solution pourrait être d'entraîner un modèle spécifique à chacune de ces catégories.

Erreur selon le degré de similitude. Pour les items de la classe Semblable, le taux d'erreur du modèle de forêt aléatoire est 16% (table 2). Si on examine plus en détail ce taux selon la classe d'origine (avant regroupement), on observe qu'il est naturellement proportionnel au degré de similitude (Figure 15).

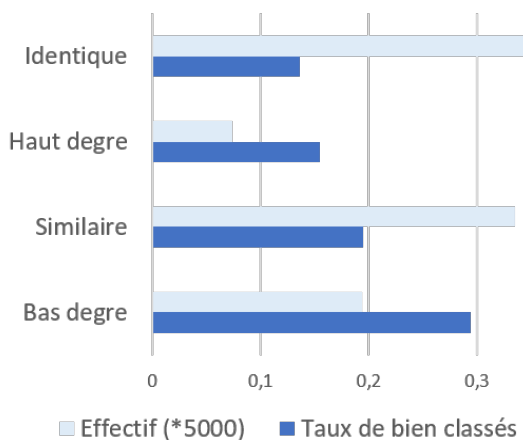


FIGURE 15 – Pourcentage d'erreur par degré de similitude

Dans la classe Identique le taux d'erreur n'est pas nul (6%)

car il ne s'agit pas d'items identiques syntaxiquement mais désignant un même produit ou service comme par exemple, *Marketing campaigns* et *Television advertising*, ou *Articles for smokers* et *Cigarette filters*. Le taux d'erreur atteint 29% pour la classe Bas degré. Il pourrait être intéressant d'utiliser un modèle d'apprentissage multiclassés et ne pas procéder au regroupement initial des classes. Il n'est pas certain qu'un tel modèle améliore le taux d'erreur de la classe Bas degré car celui-ci traduit une incertitude du décideur pouvant dépendre du contexte. Par exemple *Travelling bags* et *Bags for sports* ont été jugés de faible degré de similitude car « les sacs de voyage sont essentiellement des valises et ne couvrent donc pas les sacs de sport » mais qu'ils ont tout de même des points de similitude : « 1. Nature, 3. Méthodes d'utilisation, 6. Canaux de distribution, 8. Producteur / Fournisseur ».

6 Conclusion

Dans ce travail, nous avons proposé un outil d'aide à la décision pour l'opposition de marques avec une détection automatique de la similitude entre items. Cette approche présente deux originalités. L'utilisation de *sentence embeddings* basés sur les *Transformers* a amélioré la détection de similarité sémantique des items habituellement calculée à partir de *word embeddings*. L'intégration de la mesure de similarité cosinus avec d'autres caractéristiques sur les décisions dans un modèle d'apprentissage a augmenté de 10 à 15% les différentes métriques de performance de classification. Il demeure cependant un certain nombre d'erreurs qui ne peuvent s'expliquer ni par une proximité sémantique, ni par une autre caractéristique. Par exemple, les items *Preserved fruit* et *Rice-based snacks* sont détectés dissemblables avec notre modèle mais ils sont considérés comme similaires par l'office d'enregistrement des marques. La décision est argumentée par le fait que les deux peuvent être essentiellement des en-cas. Afin de palier ce problème, nous envisageons dans de futurs travaux de *fine-tuner* un LLM avec notre jeu de données, et de tester des *embeddings* plus frugaux, que nous pourrions créer à partir de notre jeu de données dans une approche similaire à SINr [5].

Références

- [1] Hayfa Alshwaish, Yousef Al-Ohali, and Abeer Al-Nafjan. Trademark image similarity detection using convolutional neural network. *Applied Sciences*, 12(3), 2022.
- [2] Jiangzhong Cao, Yunfei Huang, Qingyun Dai, and Wing-Kuen Ling. Unsupervised trademark retrieval method based on attention mechanism. *Sensors*, 21(5), 2021.
- [3] Yingchi Liu, Quanzhi Li, Changlong Sun, and Luo Si. Similar trademark detection via semantic, phonetic and visual similarity information. In *SIGIR '21: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2025–2030, July 2021.

- [4] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.
- [5] Thibault Prouteau, Victor Connes, Nicolas Dugué, Anthony Perez, Jean-Charles Lamirel, Nathalie Camelin, and Sylvain Meignier. SINr : Fast Computing of Sparse Interpretable Node Representations is not a Sin! In *Advances in Intelligent Data Analysis XIX, 19th International Symposium on Intelligent Data Analysis, IDA 2021*, number 12695 in Lecture Notes in Computer Science, pages 325–337, Porto, Portugal, April 2021. Springer, Cham. Frontier Prize (best paper).
- [6] Nils Reimers and Iryna Gurevych. Sentence-bert : Sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing*, 2019.
- [7] Daniyil Shmatkov, Oleksii Gorokhovatskyi, and Nataliya Vnukova. Elaborative trademark similarity evaluation using goods and services automated comparison. In *Computational Linguistics and Intelligent Systems 2023 (Volume III)*, Kharkiv, Ukraine, April 2023.
- [8] Amy J.C. Trappey, Charles Trappey, and Sam C.C. Lin. Detecting trademark image infringement using convolutional neural networks. In *Advances in Transdisciplinary Engineering*, October 2019.
- [9] Amy J.C. Trappey, Charles V. Trappey, and Samuel Shih. An intelligent content-based image retrieval methodology using transfer learning for digital ip protection. *Advanced Engineering Informatics*, 48 :101291, 2021.
- [10] Charles V. Trappey, Amy J.C. Trappey, and Sam C.-C. Lin. Intelligent trademark similarity analysis of image, spelling, and phonetic features using machine learning methodologies. *Advanced Engineering Informatics*, 45 :101120, 2020.
- [11] Osman Tursun, Simon Denman, Sabesan Sivapalan, Sridha Sridharan, Clinton Fookes, and Sandra Mau. Component-based attention for large-scale trademark retrieval. *IEEE Transactions on Information Forensics and Security*, 17 :2350–2363, 2022.
- [12] Julie Tytgat, Guillaume Wisniewski, and Adrien Bétrancourt. Apprentissage automatique pour la surveillance de marques. In *Conférence Nationale sur les Applications Pratiques de l'Intelligence Artificielle*, Saint-Etienne, France, June 2021.
- [13] Dmitry Vesnin, Dmitry Levshun, and Andrey Chechulin. Trademark similarity evaluation using a combination of vit and local features. *Information*, 14(7), 2023.

Utilisation de LLMs pour la classification d'avis client et comparaison avec une approche classique basée sur CamemBERT

N. Vautier¹, M. Héry¹, M. Miled³, I. Truche⁴, F. Bullier³, A.L. Guénet²
G. Dubuisson Duplessis², S. Campano¹, P. Suignard¹

¹ EDF Lab Paris Saclay, SEQUOIA

² EDF Commerce, Direction des Systèmes d'Information et du Numérique (DSIN)

³ AI&Data

⁴ EY

nicolas.vautier, marc.hery, guillaume.dubuisson-duplessis, anne-laure.guenet, sabrina.campano,
philippe.suignard@edf.fr

Résumé

Les cas d'usage courants de la relation client chez EDF comme le routage d'un document à la bonne personne, la catégorisation de documents, font actuellement appel à une tâche de classification automatique supervisée. Dans ce contexte, cet article compare deux approches pour la classification de commentaires de satisfaction et e-mails de clients : une approche classique mise en production à EDF basée sur CamemBERT fine-tuné sur des données EDF et une approche plus récente basée sur des LLMs (Large Language Models). Pour cette 2ème approche, 3 stratégies de prompting sont testées (zero-shot, few-shot, et keyword prompting) avec plusieurs LLMs "open-weights" : Mistral, Mixtral, NeuralHermès et Phi3 sur des tâches de classification de référence de 2 types différents : classification binaire et multilabel. En plus de ces stratégies, l'impact du pré-traitement sur les textes en entrée des LLMs a été évalué, ainsi que l'apport de leur fine-tuning spécifiquement pour les tâches. S'il ressort de ces tests que les performances des LLMs non fine-tunés sont en deçà de celles des approches CamemBERT, l'étude apporte des enseignements sur l'impact du prompting, du pré-traitement des textes et l'apport du fine-tuning dans l'utilisation de ces modèles.

Mots-clés

Grands modèles de langue, ingénierie du prompt, classification de texte.

Abstract

Common customer relationship use cases at EDF, such as routing a document to the right person and document categorization, currently require a supervised automatic classification task. In this context, this article compares two approaches for the classification of satisfaction comments and customer emails : a classic approach put into production at EDF based on CamemBERT fine-tuned on EDF data and a more recent approach based on LLMs (Large Language Models). For this 2nd approach, 3 prompting strate-

gies are tested (zero-shot, few-shot, et keyword prompting) with several open-weights LLMs : Mistral, Mixtral, NeuralHermès and Phi3 on reference classification tasks of 2 different types : binary classification and multilabel. In addition to these strategies, the impact of pre-processing on the LLM input texts was evaluated, as well as the contribution of their fine-tuning specifically for the tasks. If it appears from these tests that the performances of non-fine-tuned LLMs are below those of CamemBERT approaches, the study provides lessons on the impact of prompting, text preprocessing and the contribution of fine-tuning in the use of these models.

Keywords

Large language models, prompt engineering, text classification.

1 Introduction

Chaque mois, la gestion de la relation client chez EDF Commerce produit un volume important de données textuelles, issues tant des interactions des clients (par exemple, courriels et réponses ouvertes dans des enquêtes de satisfaction) que des observations formulées par les conseillers (telles que les commentaires de contacts). Ces informations, principalement rédigées en français, se caractérisent par leur richesse et diversité. Elles englobent une variété de formats, depuis des expressions spontanées jusqu'à des réponses plus structurées, comme celles enregistrées via des formulaires. De plus, elles révèlent une hétérogénéité notable en termes d'orthographe, de syntaxe et de niveau de langue qui soulève de véritables défis pour les explorer efficacement [7]. Ces données sont utilisées pour répondre au mieux aux attentes de nos clients en suivant le cadre réglementaire du « règlement général sur la protection des données » (RGPD) [6]. En outre, elles sont exploitées dans de nombreux cas d'usage visant à optimiser la relation client (e.g., sur le canal e-mail [8] ou sur le canal téléphonique [9]).

Les cas d'usage courants de la relation client comme le pilotage d'une activité basée sur des données texte (e.g., emails, réponses ouvertes à des enquêtes), le routage d'un document à la bonne personne, la catégorisation de documents font appel à une tâche de classification classique de l'apprentissage supervisé. La tâche de classification consiste à attribuer à chaque entrée de données une catégorie spécifique parmi un ensemble prédéfini. Une tâche de classification a le bénéfice d'être conceptuellement simple tout en offrant un cadre d'évaluation robuste via des métriques bien établies (comparativement moins subjectif que des tâches dites « génératives »).

L'avènement récent des "grands modèles de langue" (LLMs) comme GPT3, GPT4 [1], puis des modèles "open-weights", comme Mistral [14, 15], est en train de bouleverser le monde du traitement automatique du langage. Outre les nouvelles possibilités offertes par les LLMs (e.g., résumé de texte, chatbots, ...), ils sont également applicables sur des tâches de classification plus traditionnelles. A ce sujet, des résultats contradictoires ont été rapportés sur l'apport des LLMs pour les tâches de classification. D'un côté, les capacités d'apprentissage dans des scénarios en zéro/few-shot sont soulignées [16] tandis que de l'autre ces capacités sont remises en cause en raison de problèmes de contamination [18, 2, 5].

Les travaux présentés dans cet article visent à mieux cerner l'impact des LLMs sur des tâches de classification pour des cas d'usage industriels. Tout d'abord, nous nous intéressons à des tâches de classification sur des données privées qui ne sont pas sujettes à des problèmes de contamination. Cela nous permet d'évaluer de manière plus robuste le réel apport des LLMs sur des tâches de classification. Ensuite, nous comparons les LLMs à des modèles de référence actuellement employés industriellement, dont un modèle fondation similaire à CamemBERT [20] spécialisé sur les données de la relation client d'EDF Commerce. Nous présentons les principaux enseignements de cette comparaison, et nous discutons également de l'impact sur la méthodologie des projets impliquant des tâches de classification.

Le plan de l'article est structuré par les parties décrites ci-après. La Section 2 pointe les travaux connexes les plus saillants. Les données sont présentées en Section 3. La Section 4 forme le cœur de l'article, elle décrit les expérimentations réalisées et les résultats obtenus. Ces résultats sont discutés en Section 5, puis la Section 6 clôt cet article en soulignant les principales conclusions et en identifiant quelques perspectives prometteuses.

2 Travaux connexes

De manière très schématique, une approche classique pour classer des textes consiste à les représenter sous une forme vectorielle, que ce soit avec des sacs de mots (*bags of words*) ou avec des plongements de mots (*word embeddings*), puis à entraîner un classifieur de type régression logistique, SVM ou autre.

Les manières de réaliser ces plongements ont beaucoup évolué ces dernières années : dans un premier temps non

contextuels, comme Word2Vec [21] ou GloVe [23], ils sont devenus contextuels, en s'appuyant sur l'architecture *transformer*, comme BERT [4], CamemBERT [20] ou FlauBERT [17]. De manière encore plus récentes, les LLMs offrent également une possibilité de produire des plongements [12], possibilité qui n'a pas été investiguée dans cet article.

L'arrivée des LLMs permet d'envisager de nouvelles tâches, notamment en sciences sociales [26], mais également de réaliser des tâches historiques de classification, ce que nous avons testé dans notre étude. Une récente approche consiste à demander directement au LLM via une instruction en langue naturelle dans quelle catégorie il classerait les documents fournis en entrée, avec une approche en *zero shot* (sans exemple) ou en *few shot* (avec des exemples) :

- Chae et Davidson [3] montrent que les résultats obtenus avec des LLMs (ici GPT 3) surpassent les approches traditionnelles de machine learning utilisant des *bags of words* et des *word embeddings*, mais qu'un BERT fine-tuné sur le corpus complet reste compétitif ;
- Sun et al. [24] obtiennent des résultats largement supérieurs avec des LLMs en utilisant une stratégie de prompt plus sophistiquée consistant à demander au LLM d'extraire les informations importantes du texte à classer (mots clés, phrases, informations contextuelles et sémantiques, etc.) avant d'en déduire la catégorie du texte.

Fields et al. proposent un état de l'art sur les LLMs utilisés pour la classification de texte [11]. Ils comparent les performances de différents modèles sur 15 tâches, et notent que les LLMs récents sont plus efficaces sur la tâche de question / réponse, tandis que sur des tâches de classification de texte les modèles les plus performants incluent des modèles BERT mais également des modèles qui ne sont pas basés sur une architecture *transformer*. Dans tous les cas, une des difficultés consiste à bien formuler l'instruction ou *prompt* à adresser au LLM, comme le rapportent Liu et al. [19] dans leur revue des différentes méthodes de *prompting*.

Contrairement aux travaux précédents qui portent sur des corpus en anglais, notre étude se focalise sur des données privées composées d'avis clients au sens large, écrits en français, ce qui constitue un cas d'usage réel, avec toutes les particularités que cela peut représenter (fautes d'orthographe, écrits parfois de manière télégraphique, etc.).

3 Données

Nous présentons dans cette section les deux grands types de données utilisées pour cette expérimentation : des données issues de réponses à des enquêtes de satisfaction, et des données issues d'e-mails clients à destination d'un conseiller. Nous présentons également la procédure de désidentification appliquée à ces données.

3.1 Données de satisfaction

Le corpus est composé d'un échantillon d'un peu plus de 3000 réponses ouvertes à une enquête de satisfaction. En

effet, après clôture d'une demande ou réclamation, certains clients sont invités à répondre à plusieurs questions. Dans ce questionnaire, le client doit, entre autres, indiquer son niveau de satisfaction puis répondre à une question en fonction de celui-ci. Si le client est "très satisfait" ou "pas satisfait", la question posée sera : "Quelles sont les raisons de votre satisfaction/insatisfaction?". Si le client est "assez satisfait" ou "peu satisfait", la question posée sera : "Qu'aurait pu faire votre conseiller pour améliorer votre satisfaction?". C'est lors de cette réponse que le client aborde potentiellement les irritants (thématiques) que l'on veut détecter. Afin de préparer l'entraînement de modèles pour la détection de ces irritants clients, une double annotation avec ré-annotation en cas de désaccord a été réalisée. La procédure d'annotation était multilabel, impliquant ainsi l'attribution d'une ou plusieurs des 13 thématiques existantes à chaque réponse. Cette classification nous aide, in fine, à mieux comprendre les irritants exprimés et permet donc l'amélioration continue de l'expérience client.

Pour cet article, 4 thématiques ont été retenues, présentant des niveaux de difficulté variable sur la tâche de classification. Des classifications binaires (présence ou non du label correspondant à la thématique) et multilabels sont à la fois réalisées. Les jeux de données binaires sont indiqués par le nom suivant : "Satisfaction-b2b-[thématique]", le jeu de données multilabel par le nom "Satisfaction-b2c-MULTI". Les thématiques retenues pour l'étude et les caractéristiques des jeux de données associés sont décrits dans le tableau 1. Ces thématiques correspondent à des plaintes quant au traitement de la demande par le service client. COUPURE DE COMMUNICATION : Le client se plaint d'avoir subi une coupure téléphonique pendant l'appel. PROBLEME NON RESOLU : le client se plaint du fait que le problème n'est toujours pas résolu malgré la clôture de la demande. REPONSE NON DESIREE : la réponse est légitime du point de vue de EDF, mais celle-ci ne satisfait pas le client. SUIVI DE LA DEMANDE : manque de communication à propos de l'évolution de la demande client.

3.2 Données d'e-mails

Un corpus d'e-mails est considéré. Il s'agit d'e-mails entrants écrits par un client à destination d'un conseiller. Ce jeu de données est utilisé pour améliorer le routage et le pilotage de l'activité. Les sollicitations clients dans ces e-mails sont très variées de par leur nature, leur vocabulaire ou encore leur taille. Le corpus est composé d'un peu moins de 6000 e-mails annotés et va servir à modéliser le ressenti dans un e-mail. Le ressenti est représenté par trois catégories. La catégorie URGENCE dans laquelle le client demande une action rapide du conseiller, la catégorie RELANCE dans laquelle le client relance suite à une première demande sans réponse, la catégorie MÉCONTENTEMENT dans laquelle le client exprime son mécontentement quant à une situation donnée. Une double annotation a été réalisée pour parvenir à ce premier corpus avec ré-annotation des désaccords. La procédure d'annotation était multilabel permettant d'attribuer plusieurs ou aucun ressenti à un même e-mail.

Les jeux de données sont traités comme des problèmes de classification binaire (leur nom suit la forme "Ressenti-b2b-[thématique]") et multilabel ("Ressenti-b2b-MULTI"). Les caractéristiques du jeu de données "Ressenti-b2b" sont présentées dans le tableau 1.

3.3 Désidentification des données texte

En vertu du règlement général sur la protection des données (RGPD), l'ensemble des données texte a été désidentifié et ne comportent aucune donnée à caractère personnel telles que des noms, prénoms, adresses, numéros [6]. Les données à caractère personnel sont substituées par le type de l'entité (e.g., "Je suis Monsieur [*person*] (numéro de client : [*num*]), et je vous écris pour manifester mon mécontentement au sujet de ma dernière facture pour mon logement situé à [*localisation*]").

4 Méthodes et expérimentations

4.1 Méthodes

4.1.1 Modèles

Pour évaluer la qualité des classifications obtenues par les approches LLM, nous cherchons à les comparer avec l'approche de référence : un modèle CamemBERT spécifiquement entraîné pour la classification sur chaque jeu de données.

Modèle baseline CamemBERT Le modèle CamemBERT EDF Commerce est un modèle pré-entraîné sur la base du modèle CamemBERT [20] et spécialisé pour la gestion de la relation client de la direction Commerce d'EDF. Le modèle a été pré-entraîné sur un ensemble de données textes désidentifiées en français provenant de différentes sources internes (e.g., emails, réponses libres à des enquêtes de satisfaction, commentaires conseiller) reflétant les nuances et la complexité de la communication dans notre secteur d'activité. Il est important de noter que les données à caractère personnel directement identifiantes ont été exclues du pré-entraînement (e.g., nom, prénom, localisation, numéro client). Le pré-entraînement a été réalisé en utilisant la technique de masquage aléatoire (MLM) qui permet au modèle d'apprendre les relations sémantiques entre les mots et les phrases.

LLMs considérés Quatre LLMs ont été utilisés pour chaque expérimentation :

- **Mistral-7B-Instruct-v0.1** [14], le premier LLM produit par la société Mistral, spécialisé sur un jeu de données d'instructions conversationnelles, sorti en septembre 2023, avec une taille de 7 milliards de paramètres ;
- **NeuralHermes-2.5**¹, un modèle fine-tuné à partir de Mistral-7B, réalisé par le français Maxime Labonne ;
- **Mixtral-8x7B-Instruct-v0.1** [15], de la société Mistral, un modèle composé d'une mixture de 8 experts, sorti en décembre 2023 ;

1. <https://huggingface.co/mlabonne/NeuralHermes-2.5-Mistral-7B>

	Type	Nb. classes	Nb. documents	Nb. mots (moy.)	Ratio de déséquilibre
Jeu de données					
Ressenti-b2b-MECONTENTEMENT	binaire	2	4494	82 +/- 93	0.04
Ressenti-b2b-RELANCE	binaire	2	4494	82 +/- 93	0.07
Ressenti-b2b-URGENCE	binaire	2	4494	91 +/- 100	0.07
Satisfaction-b2c-COUPURE_DE_COMMUNICATION	binaire	2	2553	27 +/- 35	0.03
Satisfaction-b2c-PROBLEME_NON_RESOLU	binaire	2	2610	27 +/- 33	0.18
Satisfaction-b2c-REPONSE_NON_DESIREE	binaire	2	2678	29 +/- 37	0.14
Satisfaction-b2c-SUIVI_DE_LA_DEMANDE	binaire	2	2611	28 +/- 36	0.13
Ressenti-b2b-MULTI	multilabel	4	4494	91 +/- 100	0.04
Satisfaction-b2c-MULTI	multilabel	5	3810	28 +/- 37	0.02

TABLE 1 – Résumé des jeux de données utilisés pour les expérimentations. Le nombre moyen de mots correspond au nombre moyen de mots calculés sur l’ensemble des documents de chaque corpus (+/- l’écart type). Le ratio de déséquilibre est calculé comme le rapport entre le nombre de documents appartenant à la plus petite classe et le nombre total de documents du corpus.

— **Phi-3-mini-128k-instruct** [10], modèle léger à l’état de l’art publié par Microsoft ;

Nous avons sélectionné ces modèles en raison de leurs bonnes performances sur des cas d’usage métier EDF, notamment sur des tâches de question réponse et de résumé. Ces derniers sont sortis en *open-weights* à la fin de l’année 2023 et au début de l’année 2024, cela signifie que les poids de ces modèles sont accessibles, contrairement aux données qui ont servi à leur entraînement. Cela permet de les utiliser sur infrastructure de calcul interne, et de préserver la confidentialité des données. Par ailleurs, utiliser les versions les plus "légères" de ces modèles à 3.8 et 7 Md de paramètres, permet de se rapprocher d’une utilisation en contexte industriel, où les moyens de calcul et les temps requis pour exécuter les traitements sont plus contraints.

4.1.2 Découpage entraînement/validation/test

Pour pouvoir comparer les métriques entre la baseline CamemBERT et les LLMs, les découpages déjà utilisés pour l’entraînement et l’évaluation des modèles CamemBERT ont été repris dans la mesure du possible. Pour les expérimentations binaires, les plis ont pu être repris tels quels, avec une distribution stratifiée : avec un ratio apprentissage/test de 85%/15% pour les jeux de données d’emails et 80%/20% pour les jeux de données de satisfaction. Pour les expérimentations multilabel, les plis utilisés pour CamemBERT n’étant pas disponibles, une stratification itérative a été appliquée dans les mêmes proportions afin de conserver la même distribution de labels unitaires et des associations de labels entre jeux de données d’apprentissage et de test. On notera qu’il n’est pas pertinent d’avoir un ensemble de validation pour les évaluations des LLMs non fine-tunés, puisque seule l’inférence sur le jeu de test nous permet de se comparer à la baseline Camem. Pour les expérimentations de fine-tuning des LLMs en revanche, il a été nécessaire de redécouper l’ensemble d’apprentissage pour obtenir un ensemble de validation : le ratio du pli est alors fixé à 85%/15% et le découpage se fait par stratification simple (cas binaire) ou itérative (cas multilabel).

4.1.3 Métriques d’évaluation considérées

Chaque jeu de données étant déséquilibré, seules des métriques adaptées au déséquilibre des classes ont été utilisées. Les métriques choisies sont identiques à celles utilisées pour l’évaluation des modèles CamemBERT industrialisés. Pour les classifications binaires le coefficient de corrélation de Matthews (MCC) et le F1-score ont été utilisés. Pour les classifications multilabel, seul le F1-score a été utilisé, le MCC n’étant pas défini. Pour le cadre multilabel, le F1-Score a été calculé à l’aide de ScikitLearn [22] dans sa version "macro" calculant un F1-score global obtenu comme une moyenne des F1-score des différentes classes.

4.1.4 Cadre d’évaluation

De nombreuses évaluations ayant été nécessaires, nous avons choisi les librairies python Luigi² pour la chaîne de traitement, plis, entraînement et évaluation, ainsi que de la librairie MLFlow [25] pour l’enregistrement des paramètres et des métriques résultant de chaque expérimentation. Concernant l’environnement de calcul, nous avons utilisé 3 GPUs d’une plateforme Nvidia DGX H100 80Go.

4.2 Expérimentations

Nous présentons dans cette section les 3 stratégies explorées lors des expérimentations menées. Elles sont évaluées sur chaque modèle décrit Section 4.1.1, et chaque jeu de données décrit Section 3. Ces 3 stratégies sont les suivantes : quelle méthode de *prompting* utiliser, quel prétraitement appliquer sur les textes et enfin, quel gain peut-on espérer en spécialisant ces modèles avec du *fine-tuning*.

4.2.1 Stratégies de prompting

Le *prompt* désigne l’ensemble des instructions formulées en langage naturel en entrée d’un LLM afin d’obtenir le résultat désiré. Nous cherchons à concevoir un *prompt* qui contienne le maximum d’informations tout en étant le plus clair possible. Le *prompt* doit présenter le problème de manière concise, énoncer les instructions d’entrée et de sortie

2. <https://github.com/spotify/luigi>

au LLM et présenter ou non des exemples en fonction de la stratégie de *prompting* adoptée.

Dans le cadre de notre étude, 3 stratégies ont été retenues : *zero-shot*, *few-shot*, et *keyword prompting*. Nous souhaitons mesurer l'impact de celles-ci sur les performances de la classification.

Zero shot prompting Le *zero-shot prompting* consiste à demander au LLM de classer un texte sans lui donner d'exemple. Seulement une définition de la catégorie est utilisée. Après plusieurs essais, la stratégie *zero-shot* que nous avons retenue est la suivante :

"Tu dois répondre par oui ou par non afin de savoir si un client a fait l'objet de [thématique] ou non. Par [thématique], on veut dire : [définition]. Voici ce qu'a dit le client : [Verbatim du client]. Ce client a-t-il fait l'objet de [thématique] ?"

Keywords prompting La stratégie de *keyword prompting* diffère du *zero-shot prompting* par l'ajout de mots clés :

"Tu dois répondre par oui ou par non afin de savoir si un client a fait l'objet de [thématique] ou non. Par [thématique], on veut dire : [définition]. Si ce que le client dit est en rapport avec un des termes de la liste alors il y a [thématique] : [liste de mots clés]. Si ce que le client dit n'est pas en rapport avec un des termes de la liste précédente alors tu dois répondre non. Voici ce qu'a dit le client : [Verbatim du client]. Ce client a-t-il fait l'objet de [thématique] ?"

Voici un exemple de prompt pour la catégorie "Coupure de communication" :

"Tu dois répondre par oui ou par non afin de savoir si un client a fait l'objet d'une coupure de la communication ou non. Par coupure de la communication, on veut dire que le conseiller a raccroché au nez du client ou que la conversation a été interrompue à cause d'un problème technique. Si ce que le client dit est en rapport avec un des termes de la liste alors il y a coupure de communication : "communication coupée", "coupure téléphonique", "coupure communication", "raccroche tout seul", "raccroche au nez", "téléphone coupe", "conseiller raccroche". Si ce que le client dit n'est pas en rapport avec un des termes de la liste précédente alors tu dois répondre non. Voici ce qu'a dit le client : 'Verbatim du client'. Ce client a-t-il fait l'objet d'une coupure de la communication ?"

Few shot prompting La dernière stratégie de *prompting* vient en complément de la précédente. En plus d'ajouter des mots clés, on insère également dans le prompt des exemples concrets de commentaires client issus de notre plan d'annotation.

Voici un exemple pour la catégorie "Mécontentement" :

"Tu dois répondre par oui ou par non afin de savoir si un mail client évoque un mécontentement. Le mécontentement est une situation exprimée comme désagréable par le client et pour laquelle il n'attend pas forcément une explication, une solution ou toute autre forme de reconnaissance. Si ce que le client dit est en rapport avec un des termes de la liste alors il y a mécontentement : "mécontent", "résiliation", "concurrence", "ironie", "jugement négatif", "menaces", "Ce n'est pas normal de votre part!", "Je vous demande donc de me rembourser les sommes que

votre entreprise m'a indûment prélevées sur mon compte bancaire.", "Aujourd'hui nous recevons ENCORE une facture avec toujours l'ancienne adresse e faire le nécessaire AU PLUS VITE", "Pourriez vous faire ces modifications dans les plus brefs délais afin que je garde de bonne relation avec edf entreprises". Si ce que le client dit n'est pas en rapport avec un des termes de la liste précédente alors tu dois répondre non. Voici ce que dit le client : 'Verbatim du client'. Ce client évoque-t-il un mécontentement ?"

Prompting en multilabel Dans le cadre d'une classification multilabel, nous réutilisons les 3 stratégies précédentes pour la description des classes à détecter, en précisant que tout texte n'appartenant pas aux classes définies doit être catégorisé "AUTRE". Ainsi, pour le cas multilabel, le modèle n'est pas tenu de choisir un label quand aucun ne correspond aux données. Dans le cas multilabel, l'instruction passée en prompt est : "Chaque texte peut appartenir à plusieurs classes. Donne le résultat sous la forme d'une liste, par exemple : [CLASS_X, CLASS_Y]."

4.2.2 Pré-traitement des textes

Les textes de nos jeux de données étant déjà anonymisés, les entités nommées telles que les personnes physiques et morales, les montants, les numéros de téléphone etc. ont au préalable été remplacées par des balises d'anonymisation. Ainsi, la phrase : "une baisse d'un montant de 10 euros pour le contrat du site de Bordeaux" est présente dans nos corpus sous la forme : "une baisse d'un montant de __MONTANT__ pour le contrat du site de __LOCALISATION__". Ce pré-traitement pouvant impacter la compréhension du LLM, nous avons expérimenté un pré-traitement effectuant l'opération inverse, en remplaçant ces balises par des entités toujours factices mais vraisemblables pour le modèle. Ainsi, si le pré-traitement est appliqué, la phrase précédente sera sous la forme suivante : "une baisse d'un montant de 4 euros pour le contrat du site de Paris".

4.2.3 Fine-tuning

Pour évaluer le potentiel d'amélioration des performances des LLMs sur ces expérimentations, une spécialisation des modèles (*fine-tuning*) a été réalisée sur chaque jeu de données. Pour cela, nous avons constitué nos jeux d'apprentissage en encapsulant les exemples des ensembles d'apprentissages initiaux dans les prompts des 3 stratégies de *prompting* décrites précédemment. Le résultat attendu est ensuite concaténé au prompt afin de former la base d'apprentissage. Le jeu de test est formé en prenant le prompt indiquant la tâche de catégorisation et le texte à catégoriser, sans ajouter le résultat, que le modèle devra prédire correctement. Notre jeu de données est divisé en deux pour obtenir les jeux d'apprentissage (85%) et de validation (15%) qui serviront à *fine-tuner* nos modèles. Pour ces expérimentations de *fine-tuning*, la même méthode a été appliquée avec le même paramétrage. La méthode LoRa a été ici utilisée sur les modules d'attention des modèles pour accélérer l'apprentissage avec le moins de perte de performance possible, comme préconisé par les auteurs de [13]. Lors de l'apprentissage, un même taux d'apprentissage ($10e-4$) et un nombre variable d'époques (entre 3 et 8) ont été utili-

	Fine-tuning	Binaire				Multilabel			
		Zero-shot	Keyword	Few-shot	Baseline	Zero-shot	Keyword	Few-shot	Baseline
Mistral-7B	Non	0.38	0.35	0.35	/	0.22	0.21	0.24	/
	Oui	0.51	0.67	<u>0.72</u>	/	0.58	0.58	0.64	/
NeuralHermes-2.5	Non	0.38	0.48	0.51	/	0.29	0.29	0.39	/
	Oui	0.41	0.44	0.42	/	0.52	<u>0.60</u>	0.57	/
Mixtral-8X7B	Non	0.44	0.54	0.51	/	0.40	0.50	0.42	/
	Oui	-	-	-	/	-	-	-	/
Phi-3-mini-128k	Non	0.28	0.33	0.40	/	0.23	0.28	0.29	/
	Oui	0.53	0.60	0.65	/	0.53	0.58	0.55	/
CamemBERT	/	/	/	/	0.79	/	/	/	0.50

TABLE 2 – Résultats des différentes expérimentations : binaire = tâches de classification binaire, multilabel = tâches de classification multilabel. Chaque cellule correspond à la moyenne des scores obtenus sur tous les jeux de données pour une configuration (modèle fine-tuné ou non + tâche + stratégie de prompting). La métrique utilisée est le F1-score. En gras /souligné : meilleur score /second meilleur score par type de classification. CamemBERT représente le modèle baseline CamemBERT EDF Commerce.

	Fine-tuning	Ressenti-b2b-BINAIRE				Ressenti-b2b-MULTI				Satisfaction-b2c-BINAIRE				Satisfaction-b2c-MULTI			
		Zero-shot	Keyword	Few-shot	Baseline	Zero-shot	Keyword	Few-shot	Baseline	Zero-shot	Keyword	Few-shot	Baseline	Zero-shot	Keyword	Few-shot	Baseline
Mistral-7B	Non	0.40	0.34	0.42	/	0.28	0.27	0.28	/	0.37	0.36	0.30	/	0.17	0.15	0.20	/
	Oui	0.78	0.65	<u>0.79</u>	/	<u>0.80</u>	0.77	0.83	/	0.31	<u>0.69</u>	0.67	/	0.36	0.38	0.44	/
NeuralHermes-2.5	Non	0.39	0.41	0.52	/	0.40	0.37	0.51	/	0.37	0.52	0.49	/	0.18	0.22	0.24	/
	Oui	0.57	0.64	0.58	/	0.77	0.83	0.70	/	0.29	0.29	0.30	/	0.28	0.38	0.44	/
Mixtral-8X7B	Non	0.49	0.59	0.55	/	0.35	0.45	0.31	/	0.40	0.51	0.49	/	0.24	0.27	0.26	/
	Oui	-	-	-	/	-	-	-	/	-	-	-	/	-	-	-	/
Phi-3-mini-128k	Non	0.29	0.32	0.39	/	0.25	0.33	0.31	/	0.28	0.34	0.41	/	0.21	0.22	0.27	/
	Oui	0.63	0.70	0.67	/	0.69	0.76	0.69	/	0.46	0.52	0.64	/	0.37	0.40	<u>0.41</u>	/
CamemBERT	/	/	/	/	0.81	/	/	/	0.79	/	/	/	0.77	/	/	/	0.22

TABLE 3 – Résultats détaillés : binaire = tâches de classification binaire, multi = tâches de classification multilabel. Chaque cellule correspond à la moyenne des scores obtenus pour une configuration donnée. Par exemple, Ressenti-b2b-BINAIRE correspond à la moyenne des scores obtenus sur les jeux de données "mécontentement", "relance" et "urgence". La métrique utilisée est le F1-score. En gras /souligné : meilleur score /second meilleur score par type de classification. CamemBERT représente le modèle baseline CamemBERT EDF Commerce.

sés. Le modèle ayant obtenu le meilleur score sur le jeu de validation est conservé.

4.3 Résultats

Analyse globale Les résultats de la classification binaire et multilabel sont présentés dans le Tableau 2 qui contient la moyenne des résultats sur tous les jeux de données par stratégie de prompt.

Les meilleurs résultats sont mis en évidence pour chaque type de classification (binaire et multilabel) et sont à comparer avec le modèle baseline CamemBERT représentant le modèle CamemBERT EDF Commerce. Pour la tâche de classification binaire, la baseline est nettement supérieure avec un F1-score de 0.79 contre 0.72 pour la version fine-tunée de Mistral-7B sur la stratégie *few-shot*. En classification multilabel, Mistral-7B fine-tuné en *keyword* obtient un F1-score de 0.64, soit 14 points de plus que la baseline CamemBERT. A noter que Phi-3-mini-128k fine-tuné maintient un bon niveau sur les prompts keyword et few-shot

par rapport à sa taille.

Impacts de la stratégie de prompting Les stratégies de *keyword prompting* et *few-shot prompting* ont affiché des performances supérieures par rapport au *zero-shot prompting* avec en moyenne un gain de 3.2 points pour la partie non fine-tunée, et un gain de 5.3 points pour la partie fine-tunée. L’ajout de mots clés et/ou d’exemples a donc un impact sur la capacité des LLM à classer correctement les verbatims client.

Impacts du fine-tuning En moyenne, le fine-tuning a augmenté les performances de nos tâches de classification, toutes configurations confondues, de 18 points. Il est important de noter que le fine-tuning du modèle NeuralHermes-2.5 n’a pas fonctionné pour toutes les configurations de la classification binaire. Le fine-tuning du modèle est difficile et le processus de fine-tuning fait apparaître une divergence au niveau de la fonction objectif pour certains cas. Nous ne sommes pas parvenus à en identifier la cause.

Impacts du pre-processing L'utilisation de notre pré-traitement des textes n'a pas eu d'impact sur la moyenne des résultats obtenus. Il faudra réaliser une étude plus approfondie afin de préciser ces résultats.

Analyse détaillée Les résultats détaillés sont présentés dans le Tableau 3. Pour les 2 jeux de données (Ressenti et Satisfaction), le fine-tuning améliore les performances des modèles, sauf pour NeuralHermes-2.5 sur le corpus de Satisfaction. Le jeu de données Satisfaction-b2c-MULTI donne les moins bon résultats avec 0.22 de f1-score pour la baseline et 0.44 pour les meilleurs LLM. Les meilleures performances sont atteintes par Mistral-7B et NeuralHermes-2.5 (0.83 de F1-Score) sur Ressenti-b2b-MULTI. Plusieurs hypothèses peuvent expliquer ces différences : la taille des jeux de données, la longueur des documents qui les constitue, l'intention derrière le texte rédigé par le client (mail libre ou réponse à une question) mais aussi par la nature même des labels. En effet, les 3 labels du corpus de Ressenti sont très corrélés entre eux de manière positive. Un mail comportant un caractère de relance aura de forte chance d'apparaître avec un mail à caractère d'urgence etc. Ce qui n'est pas le cas pour les 4 labels étudiés sur le corpus de Satisfaction.

En résumé, nous avons montré la capacité des LLMs à se rapprocher, voire dépasser, les scores obtenus avec la baseline. Cependant la volatilité des résultats reflète la difficulté à désigner une stratégie de prompting comme étant la stratégie générique à appliquer pour chaque jeu de données. On retiendra la capacité des LLMs à être plus performants sur des tâches jugées plus complexes avec plusieurs labels/classes comparés à CamemBERT.

5 Discussion

Les résultats obtenus dans la section précédente montrent des performances pour les LLMs inférieures au modèle CamemBERT Commerce pour les tâches de classification binaire. Cela peut trouver plusieurs explications.

Tout d'abord, dans le cadre de l'utilisation d'un LLM non fine-tuné, concevoir un "bon" prompt pour une tâche donnée n'est pas intuitif. Il s'agit d'un processus itératif ayant demandé plusieurs essais. Nous avons constaté par expérimentation que tout doit être explicité dans le prompt pour obtenir de bons résultats, et que des détails importants pour la classification peuvent facilement être omis. Cette utilisation se différencie d'une tâche de classification avec un modèle fine-tuné sur des données de référence, où des connaissances sont implicitement contenues dans les classes attribuées aux données d'entraînement, et n'ont donc pas besoin d'être explicitées. Cette observation est étayée par le fait qu'ajouter des exemples (*few-shot*) ou des mots-clés (*keyword*) dans le prompt conduit à de meilleurs résultats qu'une approche en *zero-shot*.

Les résultats obtenus pourraient également avoir été impactés par la fragilité des LLMs face à la forme du prompt. De légères variations dans l'instruction (ex : minuscules vs

majuscules) induisent des résultats différents. De même, le format de la séquence de texte prédite en sortie des modèles présente des variabilités, même lorsque le format de sortie souhaité est explicité dans l'instruction fournie en entrée au modèle. Par exemple, il a été observé plusieurs noms de classes prédites contenant des caractères indésirables qui compliquent le parsing du résultat : "SOUSCRIPTION_CONTRAT" (Mixtral) ou "SOUSCRIPTION_CONTRACT" (NeuralHermes) au lieu du nom précisé dans l'instruction "SOUSCRIPTION_CONTRAT". De même, l'instruction "Donne le résultat sous la forme d'une liste, par exemple : [CLASS_X, CLASS_Y]." peut aboutir dans de très rares cas au résultat "CLASS_X" au lieu d'un nom de classe attendu.

Concernant la comparaison des performances des LLMs à celles du modèle métier CamemBERT EDF Commerce, des différences dans l'entraînement des modèles pourraient avoir avantaagé le modèle métier. Le modèle CamemBERT EDF Commerce a été pré-entraîné sur un grand volume de données métier EDF, ce qui pourrait l'avoir favorisé au détriment des LLMs malgré leur taille, y compris dans la configuration de fine-tuning. Les LLMs n'ont en effet pas été pré-entraînés sur données métier avant d'être fine-tunés sur la tâche. En bref : spécialiser des modèles plus petits semble offrir de meilleurs résultats que les LLMs non fine-tunés testés dans ces travaux.

Enfin, sur l'aspect multilabel, il est important de préciser que les modèles CamemBERT multilabels ne sont pas utilisés de façon opérationnelle à cause de leur plus faible performance (pouvant s'expliquer par de possibles corrélations entre les classes, ou encore par le problème de stratification multilabel sur des classes déjà fortement déséquilibrées). Au lieu de cela, un système composé de plusieurs classifieurs binaires leur est préféré, présentant de meilleures performances mais ayant également l'inconvénient d'être plus difficile à maintenir en production. Les résultats des LLM finetunés sur les tâches multilabel montrent qu'il est possible d'obtenir des performances presque équivalentes avec un seul modèle, un avantage de taille pour une utilisation en production.

6 Conclusion et perspectives

Cet article nous a permis de tester des LLMs et de comparer les résultats obtenus sur une tâche de classification avec une chaîne de traitement en production basée sur un modèle CamemBERT finetuné sur des données Commerce. De ces différentes expérimentations, nous pouvons conclure que :

- CamemBERT EDF obtient les meilleurs résultats en classification binaire, même s'il existe presque toujours une configuration avec LLM qui obtient un résultat équivalent (notamment en fine-tunant le LLM). En classification multilabel, ce sont les LLMs fine-tunés qui se montrent les plus performants, illustrant peut-être la plus grande capacité de généralisation de ces modèles.
- Les scores obtenus par les LLMs sont proches du modèle CamemBERT Commerce, surtout quand ils sont

fine-tunés. Par contre, nous observons un impact faible des stratégies utilisées (pré-traitement et prompting *zero shot*, *keyword* et *few shot*), l'avantage des LLMs étant leur performance correcte quelles que soient ces stratégies. Leur inconvénient est la difficulté à trouver une stratégie pour améliorer ces résultats.

- Trouver le meilleur *prompt* possible est une tâche complexe qui nécessite parfois plusieurs essais.

Pour améliorer les premiers résultats obtenus, plusieurs perspectives s'offrent à nous. Tout d'abord nous pourrions fine-tuner le modèle Mixtral-8x7B, ce que nous n'avons pas fait par manque de temps. Ensuite, il serait possible de poursuivre le pré-entraînement des LLMs sur les données Commerce, comme EDF Commerce l'a fait pour spécialiser CamemBERT et produire le CamemBERT EDF. Des améliorations pourraient être apportées sur la construction des prompts, notamment au niveau du formatage de la sortie du LLM afin de mieux exploiter ses réponses. Plusieurs options de fine-tuning peuvent également être envisagées pour encore augmenter les scores. Nous pourrions mesurer l'impact des différents hyper-paramètres comme la température. Au vu de l'instabilité des modèles, nous pourrions lancer plusieurs runs par configuration pour avoir une estimation plus fiable des performances. Enfin, pour généraliser les résultats obtenus dans l'article, il serait souhaitable de tester ces approches sur des benchmarks plus larges d'avis clients.

Remerciements

Nous remercions chaleureusement toutes les personnes qui sont intervenues de près ou de loin sur ce projet : Sofiane Kerroua, Mathilde Poulain, Mathilde Jeuland, Aurore Hamimi, Oualid Akhsass, Marwen Touzi, Laura Rouhier, Sonia Audheon, Dominique Manzoni-Quantin, François Raynaud.

Références

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv :2303.08774*, 2023.
- [2] Simone Balloccu, Patrícia Schmidová, Mateusz Lango, and Ondřej Dušek. Leak, cheat, repeat : Data contamination and evaluation malpractices in closed-source llms. *arXiv preprint arXiv :2402.03927*, 2024.
- [3] Youngjin Chae and Thomas Davidson. Large language models for text classification : From zero-shot learning to fine-tuning. *Open Science Foundation*, 2023.
- [4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 4171–4186, 2019.
- [5] Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, and Ge Li. Generalization or memorization : Data contamination and trustworthy evaluation for large language models. *arXiv preprint arXiv :2402.15938*, 2024.
- [6] Guillaume Dubuisson Duplessis, Elliot Bartholme, Sofiane Kerroua, Mathilde Poulain, Ahès Roulier, and Anne-Laure Guénet. Désidentification de données texte produites dans un cadre de relation client. In *Actes de la 27eme conférence Traitement Automatique des Langues Naturelles (TALN) – démonstrations*, pages 10–13, 2020.
- [7] Guillaume Dubuisson Duplessis, François Bullier, and Anne-Laure Guénet. Démonstration : exploration sémantique de données texte de la relation client. In *9ème Conférence Nationale sur les Applications Pratiques de l'Intelligence Artificielle APIA@ PFIA2023*, pages 103–106, 2023.
- [8] Guillaume Dubuisson Duplessis, Sofiane Kerroua, Ludivine Kuznik, and Anne-Laure Guénet. Cameli@ : analyses automatiques d'e-mails pour améliorer la relation client. In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019. Volume IV : Démonstrations*, pages 623–626, 2019.
- [9] Guillaume Dubuisson Duplessis, Manon Richard, and Anne-Laure Guénet. Segmentation de phases de dialogue dans des retranscriptions de conversations de centres d'appels. In *9ème Conférence Nationale sur les Applications Pratiques de l'Intelligence Artificielle APIA@ PFIA2023*, 2023.
- [10] Marah Abdin et al. Phi-3 technical report : A highly capable language model locally on your phone, 2024.
- [11] John Fields, Kevin Chovanec, and Praveen Madiraju. A survey of text classification with transformers : How wide ? how large ? how long ? how accurate ? how expensive ? how safe ? *IEEE Access*, 12 :6518–6531, 2024.
- [12] Matthew Freestone and Shubhra Kanti Karmaker Santu. Word embeddings revisited : Do llms offer something new ? *arXiv preprint arXiv :2402.11094*, 2024.
- [13] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, and Weizhu Chen. Lora : Low-rank adaptation of large language models. *CoRR*, abs/2106.09685, 2021.
- [14] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv :2310.06825*, 2023.
- [15] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas,

- Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv :2401.04088*, 2024.
- [16] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35 :22199–22213, 2022.
- [17] Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. Flaubert : Unsupervised language model pre-training for french. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*, 2020.
- [18] Changmao Li and Jeffrey Flanigan. Task contamination : Language models may not be few-shot anymore, 2023.
- [19] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict : A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9) :1–35, 2023.
- [20] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villamonte de la Clergerie, Djamé Seddah, and Benoît Sagot. Camembert : a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.
- [21] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.
- [22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, 12 :2825–2830, 2011.
- [23] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove : Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.
- [24] Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. Text classification via large language models. *arXiv preprint arXiv :2305.08377*, 2023.
- [25] Matei A. Zaharia, Andrew Chen, Aaron Davidson, Ali Ghodsi, Sue Ann Hong, Andy Konwinski, Siddharth Murching, Tomas Nykodym, Paul Ogilvie, Mani Parkhe, Fen Xie, and Corey Zumar. Accelerating the machine learning lifecycle with mlflow. *IEEE Data Eng. Bull.*, 41 :39–45, 2018.
- [26] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can large language models transform computational social science? *Computational Linguistics*, pages 1–53, 2023.

Session 1.3 : Graphes et trajectoires

Un algorithme de routage de navires générant des trajets précis et diversifiés

Alexandre Coppé¹, Nicolas Prcovic¹

¹ Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France

6 mai 2024

Résumé

Nous présentons un algorithme déterminant avec précision des trajectoires optimales d'un navire dans un contexte multi-objectif et dynamique, où il faut notamment prendre en compte le temps de trajet et la consommation de carburant, dans des conditions météorologiques qui varient pendant le trajet. Notre approche combine deux algorithmes récents, NAMOA*-TD et WRM, nous permettant d'obtenir un panel de trajectoires (sous-ensemble du front de Pareto) précises et diversifiés parmi lesquels un utilisateur peut choisir. Les premières expérimentations effectuées à partir de données météorologique réelles nous permettent de montrer l'efficacité de cette approche.

Mots-clés

Routage de navires, chemin optimal dans un graphe, recherche multi-objective

Abstract

We present an algorithm that determines optimal ship trajectories in a multi-objective and dynamic context, where factors such as travel time and fuel consumption, amidst varying meteorological conditions along the route, need to be considered. Our approach combines two recent algorithms, NAMOA*-TD and WRM, enabling us to generate a set of precise and diversified trajectories (a subset of the Pareto front) from which a user can choose. Initial experiments conducted using real meteorological data demonstrate the effectiveness of this approach.

Keywords

Ship routing, optimal path in a graph, multi-objective search.

1 Introduction

Plus de 80% du volume des échanges commerciaux mondiaux se traite par la mer. En 2018, au sein du transport maritime, 40% des frais opérationnels étaient absorbés par le coût du carburant lors d'un voyage. Une petite amélioration, aussi minime soit-elle, peut avoir de grandes répercussions au niveau des coûts. Le routage des navires en fonction de la météo est donc un domaine qui génère un grand

intérêt pour des raisons écologiques et économiques (cf figure 1).

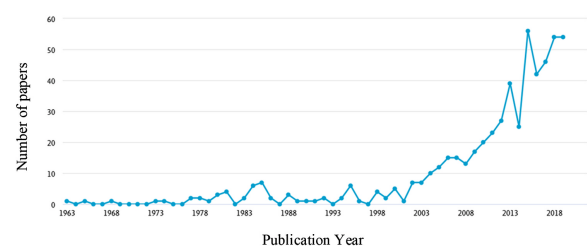


FIGURE 1 – Nombre de publication sur l'optimisation de trajet des navires. Source : Scopus February 2020.

L'objectif de nos travaux consiste à trouver un trajet de navire entre deux ports, en arrivant avant une date donnée, en optimisant la consommation de carburant et en prenant en compte les contraintes environnementales, tout ceci dans un contexte qui varie dynamiquement au cours du trajet (courants, météo).

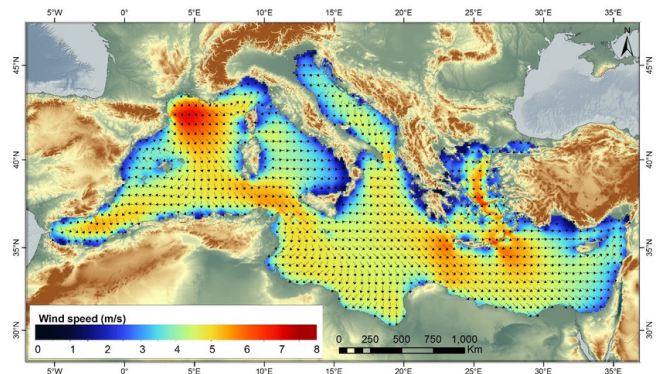


FIGURE 2 – Exemple de grille météo. Les flèches indiquent la direction du vent et les couleurs sa vitesse.

En plus de la géographie maritime, des lieux et dates de départ et d'arrivée, nous devons tenir compte des courants maritimes et de la météo. Ceux-ci nous sont données sous forme d'un pavage du globe découpé en cellules rectangulaires à l'intérieur desquelles le courant et le vent sont considérés comme étant uniformes (cf figure 2). Ces données nous permettent de calculer les différents coûts (temps,

carburant) liés au trajet qui sont fonctions du vent et des courants à un moment donné. Ces prévisions ne sont valables que pour une période donnée (typement, pendant 6h) et, pour un trajet, il nous faut récupérer les prévisions pour les périodes incluses entre la date de départ et d'arrivée du navire.

Bien que la modélisation la plus réaliste d'un trajet de navire soit une courbe continue, la numérisation des données implique leur discrétisation. L'espace et le temps numérisés seront donc un ensemble de lieux et de dates dont le nombre et la précision du repérage influenceront sur l'efficacité des algorithmes et la qualité des solutions produites. Bien souvent, les lieux sont répartis uniformément sur une grille rectangulaire et forment un maillage couvrant toute la zone de déplacements possibles en se calquant sur la grille météo.

L'approche la plus classique est de représenter les lieux par un graphe dont les arcs relient les sommets correspondant aux lieux les plus proches (cf fig 3). Les arcs sont étiquetés par un vecteur de coûts (contenant au moins ceux du temps et du carburant utilisés pour joindre les deux lieux).

Des outils calculant des trajets dans ce contexte existent déjà et sont utilisés par des entreprises de routage maritimes. Cependant les solutions en usage pèchent par au moins deux aspects d'après les utilisateurs avec lesquels nous avons discutés :

- Les trajectoires fournies sont trop grossières lorsqu'on est proche des côtes.
- Un seul trajet est souvent proposé alors que, dans un contexte multi-objectif, il y a en a souvent un grand nombre possibles (dites Pareto-optimales, cf figure 4). Or, un commandant de navire peut avoir ses idées préconçues sur le meilleur trajet à effectuer donc il aura du mal à accepter une unique solution qui diffère trop de ce qu'il estime a priori être le mieux. D'où l'utilité de lui offrir un panel diversifié de solutions possibles parmi lesquelles il en trouvera une qui a plus de chances de lui convenir.

Le plan de cet article est le suivant. Nous commençons par donner une définition formelle du problème de recherche d'un trajet optimal dans un contexte multi-objectif et dynamique. Puis, nous présentons les algorithmes existants permettant de traiter ce problème. Ensuite, nous présentons notre approche, basée sur deux algorithmes récents, qui nous permet d'obtenir des trajectoires diversifiées et aussi précises que l'on veut, afin de satisfaire aux demandes des utilisateurs du domaine. Enfin, nous présentons des expérimentations à partir de données réelles fournies par l'entreprise de transport maritime avec laquelle nous collaborons.

1.1 Définition formelle du problème

Considérons un graphe orienté et muni d'une fonction de coût c défini par : $G = (N, A, c)$ où $N = \{x_1, \dots, x_n\}$ est un ensemble fini de sommets et $A \subset N \times N$ un ensemble fini de $|A|$ arcs de la forme (x_i, x_j) . à chaque arc $a \in A$ est associé un vecteur de fonction de coûts à valeurs dans \mathbb{N} de la forme $c(a) = (c_1(a), \dots, c_q(a))$ où q est le nombre de valuations du graphe, et c_i la fonction qui attribue à un arc

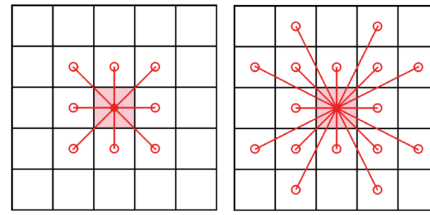


FIGURE 3 – Exemple de voisinage à 8 et 16 voisins où les sommets correspondent à des lieux répartis uniformément dans une grille rectangulaire, utilisé dans [1] .

du graphe sa i^{eme} valuation. $s \in N$ est le sommet source du graphe (départ), et $p \in N$ est le sommet puits du graphe (arrivée). Une solution est un chemin partant du sommet source et finissant par le sommet puits.

On s'intéresse aux chemins optimaux entre deux sommets du graphe. Soit CH_{ij} l'ensemble des chemins de x_i vers x_j dans le graphe G . On définit un vecteur de coût C qui à un chemin ch_{ij} associe l'évaluation des coûts pour parcourir ce chemin. On cherche à identifier le (ou les) chemin(s) $ch_{ij} \in CH_{ij}$ tels qu'il n'existe pas de chemin $ch'_{ij} \in CH_{ij}$ vérifiant $C(ch'_{ij}) < C(ch_{ij})$. L'opérateur $<$ compare des vecteurs, ce qui implique un ordre partiel sur le coût des chemins.

2 Recherche d'un chemin optimal dans un graphe

Il existe de nombreux algorithmes de recherche de chemins optimaux dans un graphe valué, qui se déclinent notamment selon s'ils sont mono ou multi-objectifs et si le(s) coût(s) des arêtes évoluent au cours du temps.

Tous ces algorithmes supposent que le trajet se fait à vitesse constante. La prise en compte d'un changement de vitesse ou de puissance du moteur au cours du trajet inclut une complexité encore plus grande qui n'est jamais pris en compte en pratique.

2.1 Algorithmes mono-objectifs

L'optimisation mono-objectif consiste à n'optimiser qu'un seul critère et permet d'obtenir une seule solution optimale. L'algorithme le plus connu est celui de Dijkstra [2] qui utilise une approche gloutonne pour obtenir une solution optimale en temps $O(n \log n)$, où n est le nombre de sommets du graphe.

L'algorithme A^* [4] est une variante de Dijkstra qui a été définie initialement pour traiter le cas où le graphe (trop grand voire infini) est défini en compréhension. Il évalue un chemin en construction non seulement en prenant en compte le coût du chemin déjà parcouru (comme Dijkstra) mais aussi une sous-évaluation heuristique du coût du chemin restant à parcourir.

En pratique A^* permet de sélectionner plus rapidement le meilleur chemin et de diminuer notablement le temps de calcul pour obtenir le chemin le moins coûteux.

2.2 Algorithmes multi-objectifs

Les algorithmes multi-objectifs doivent optimiser plusieurs objectifs dont aucun objectif n'est dominant par rapport aux autres.

Dans le cas qui nous occupe, il nous faut prendre en compte le temps de trajet, le coût du carburant, l'impact environnemental, l'usure du navire, le confort des voyageurs (minimiser les changements de cap et de vitesse), etc.

Dès lors que l'on a plus un seul objectif, on n'a plus unicité de la solution optimale (plus précisément : une seule des solutions parmi celles qui ont le coût optimal) mais un nombre potentiellement exponentiel de vecteurs de coûts dont aucun ne domine l'autre, qu'on appelle *Front de Pareto*.

Front de Pareto Lorsqu'on a plusieurs critères, on ne peut dire qu'une solution est meilleure qu'une autre que si elle l'est sur tous les critères. On dit alors qu'elle la *domine*. Certaines solutions sont incomparables : une solution peut être meilleure sur un critère et moins bonne sur un autre.

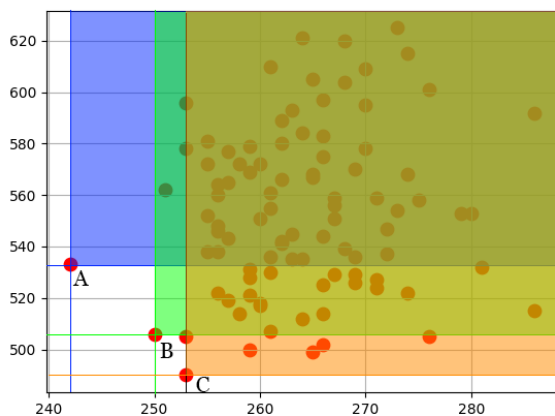


FIGURE 4 – Ensemble de solutions. On a en bleu, vert et orange les zones dominées par les sommets A, B et C, qui constituent un front de Pareto.

L'ensemble des solutions dominées par aucune autre solution s'appelle le *front de Pareto*. Une solution *pareto-optimale* a un vecteur de coût tel qu'il n'existe pas une solution alternative pour laquelle tous les éléments du vecteur de coût seraient meilleurs.

Un problème multi-objectif présente la difficulté qu'un algorithme le résolvant ne doit pas prendre en compte uniquement le coût de la meilleure solution pour éliminer d'autres candidats potentiels mais tous les vecteurs de coût du front de Pareto en cours de constitution. Le problème devient ainsi NP-difficile et donc potentiellement de complexité exponentielle.

MOA* [6] (Multi-Objective A*) et NAMOA* (New Approach to MOA*) [5] sont des variantes généralistes de A* prenant en compte le multi-objectif. MOA* prend en compte notamment le fait qu'un sommet est l'aboutissement de plusieurs chemins de vecteurs de coûts différents,

qu'il faut prendre en compte pour évaluer les coûts des chemins aboutissant aux successeurs de ce sommet (ses extensions). NAMOA* améliore MOA* en considérant des extensions de chemins partiels plutôt que des extensions de sommets.

Scalarisation La scalarisation consiste à combiner linéairement différents critères pour en former un seul. Si on a trois critères c_1 , c_2 et c_3 , la scalarisation consiste à définir la fonction $C(c_1, c_2, c_3) = \alpha_1 \cdot c_1 + \alpha_2 \cdot c_2 + \alpha_3 \cdot c_3$, où les α_i sont des coefficients à fixer qui déterminent l'importance relative de chaque coût c_i .

En choisissant adéquatement les coefficients (par recherche dichotomique), on peut obtenir toute l'enveloppe convexe d'un front de Pareto en un temps polynomial. Il nous manquera juste les solutions qui sont dans les "cavités" de l'enveloppe convexe.

La scalarisation a l'immense avantage de permettre une résolution rapide mais, comme elle ne trouve que les solutions de l'enveloppe convexe du front de Pareto, il faut que celle-ci contienne un "bon" sous-ensemble de solutions par rapport au contexte.

Dans la suite de l'article, nous appellerons *Dijkstra scalaire*, un algorithme de Dijkstra effectuant du multi-objectif en scalarisant les coûts.

Algorithmes dépendant du temps Lorsque le coût des arêtes d'un graphe change au cours du temps (par exemple à cause de la météo), la propriété qu'un sous-chemin d'un chemin optimal est optimal devient fautive (cf figure 5). On ne peut plus s'en servir pour éliminer des chemins partiels sous-optimaux. Il faut utiliser d'autres critères plus faibles.

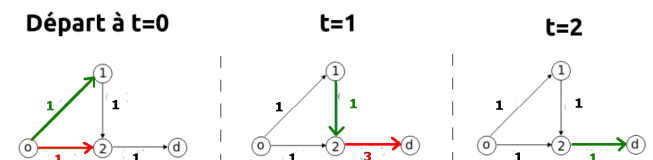


FIGURE 5 – Exemple de trajet pour lequel arriver plus tard en un sommet permet de minimiser un coût global de trajet : passer au temps 2 au sommet 1 permet de gagner en coût (1+1+1 au lieu de 1+3).

Dans le contexte du routage maritime, nous pouvons citer l'algorithme de Venetti [7] et NAMOA*-TD [1]. Pour permettre l'élimination de chemins partiels candidats, ces algorithmes ont deux critères :

- Le coût est supérieur à celui d'une solution déjà trouvée.
- Le coût est supérieur à celui d'un chemin partiel arrivant au même sommet à la même date.

NAMOA*-TD, une version "time-dependent" de NAMOA*, s'avère beaucoup plus efficace en pratique de l'algorithme de Venetti. Ceci est notamment dû au fait que, contrairement à Venetti, NAMOA*-TD utilise une heuristique d'estimation du coût du chemin restant à parcourir. Cette heuristique s'avère être une évaluation suffisamment proche du réel pour détecter tôt les chemins candidats qui

seront dominés.

Bien qu'apparaissant comme une méthode efficace dans le cadre du routage maritime international en permettant le calcul de trajets réalistes en moins de deux minutes, NAMOA*-TD a pourtant des limites qui l'empêchent d'être pleinement satisfaisant dans le contexte qui nous occupe. Le maillage rectangulaire et régulier qui détermine la forme des trajectoires n'est pas assez précis lors de la navigation proche des côtes ou dans des zones maritimes étroites (par exemple, la Manche) : la zone maritime peut être plus étroite que les mailles. La solution consistant à réduire la taille des mailles mène à une forte augmentation du temps de calcul. Par exemple, diviser la taille des mailles (rectangulaires) selon ses deux dimensions multiplie par quatre le nombre de sommets du graphe dans un contexte où la complexité temporelle des algorithmes est exponentielle en ce nombre de sommets.

En pratique, l'outil utilisé par l'armateur que nous avons consulté diminue les tailles des mailles dans certaines zones où il était constaté a posteriori que c'était nécessaire. L'inconvénient est que le graphe doit se construire en quelque sorte "à la main" et qu'il est difficile de juger précisément de la taille des mailles en fonction des lieux.

Or, il existe une approche récente, qui permet de fixer les lieux de passages d'un navire à un degré de précision arbitrairement grand, que nous allons pouvoir adapter à notre contexte.

2.3 L'approche de Weather Routing Metaheuristic

Weather Routing Metaheuristic (WRM) [3] constitue une approche originale du routage de navire qui ne fixe pas a priori les lieux de passages et permet de générer des trajets dont les lieux de passage peuvent se trouver n'importe où sur la surface de navigation à un degré de précision aussi élevé que l'on souhaite.

Le graphe se détermine en générant aléatoirement n lieux dans une zone donnée et en créant une arête entre deux lieux si la distance les séparant est inférieure à une constante donnée. Chaque segment liant deux lieux peut traverser plusieurs cellules rectangulaires de la grille météo indiquant les directions et forces des courants et des vents. On découpe donc le segment en sous-segments dont les extrémités sont sur les frontières entre les cellules, chaque cellule indiquant un courant et une vitesse de vent dont la direction et la force sont constantes. On calcule le vecteur de coûts pour chacun de ces sous-segments selon un modèle donné par l'armateur en fonction du navire, puis on fait la somme de ces vecteurs pour obtenir le vecteur de coûts de tout le segment.

Le fait qu'on puisse choisir n à l'unité près permet de choisir précisément la taille de l'instance et donc le temps d'exécution de la méthode. Les coordonnées des lieux sont déterminés avec la précision que l'on veut. Le fait que les lieux soient tirés au hasard (de manière uniforme) plutôt que fixés selon une grille régulière n'empêche pas la densité des points de rester à peu près uniforme sur toute la zone de navigation.

L'algorithme procède par itérations en cherchant un trajet dans le graphe puis en sélectionnant la zone géographique proche des lieux du trajet trouvé puis en générant aléatoirement n lieux dans cette zone restreinte pour former un nouveau graphe. Au fur et à mesure, la zone de proximité se réduit de plus en plus, ce qui affine la trajectoire progressivement (cf figure 6). Dans WRM, c'est un algorithme incomplet mono-objectif minimisant la consommation de carburant qui est utilisé mais rien n'empêche l'utilisation d'une autre procédure mono ou multi-objectif, complète ou incomplète. Si la procédure est multi-objectif, elle génère plusieurs chemins à chaque itération et il faut en choisir un seul pour l'itération suivante.

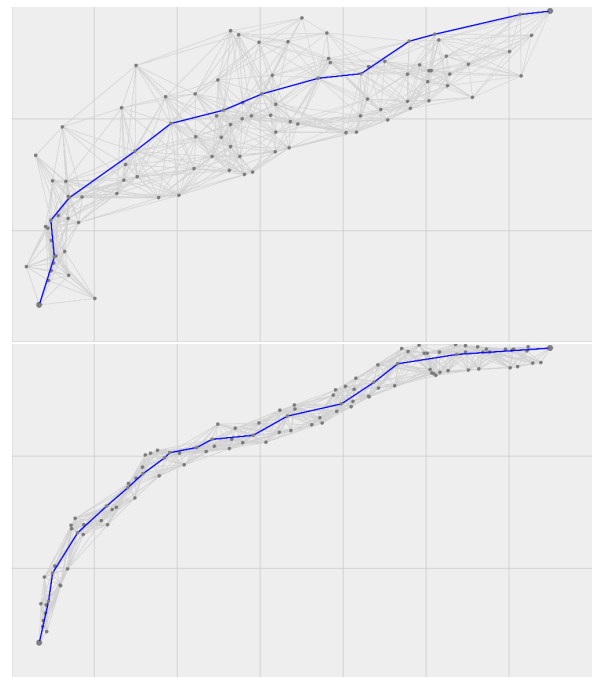


FIGURE 6 – Étape n (en haut) et étape $n + 1$ (en bas). La zone géographique des lieux possibles de l'étape $n+1$ est déterminée par la trajectoire trouvée à l'étape n . Images extraites de [3]

3 Recherche de trajets précis et diversifiés

Nous avons vu que NAMOA*-TD permettait d'obtenir un front de Pareto et donc tous les trajets possibles correspondant à une instance de problème donné mais avec une précision insuffisante dans un contexte de temps de calcul limitée. D'un autre côté, WRM permet de ne générer qu'un seul trajet mais avec un degré de précision important. Pour obtenir des trajets précis et diversifiés dans un temps court, nous proposons l'approche simple suivante.

- Utiliser un algorithme multi-objectif (NAMOA*-TD ou un Dijkstra scalaire) afin de générer un front de Pareto pour avoir des trajets diversifiés (mais peu précis).

- Sélectionner un petit nombre k de trajets du front de Pareto qui diffèrent le plus possible les uns des autres.
- Pour chacun des k trajets obtenus précédemment : lancer WRM en générant le premier graphe avec des lieux pris dans la zone de proximité du trajet. On espère ainsi raffiner progressivement une solution pour abaisser ses coûts.

Pour que le temps de calcul reste court, il faut que la grille de lieux utilisée par NAMOA*-TD sur une grille rectangulaire aient des mailles plus grosses que si NAMOA*-TD était utilisé seul afin de laisser du temps au reste de l'algorithme. Nous obtenons des trajets encore moins précis qu'avant mais ce n'est pas gênant car ils sont destinés à être raffinés juste après.

Pour la sélection des k trajets du front de Pareto, nous nous basons sur les vecteurs de coûts et faisons en sorte que leurs valeurs soient le plus uniformément réparties.

Concernant la recherche de chemins pareto-optimaux pendant une itération de WRM, nous testerons deux algorithmes :

- NAMOA*-TD pour être complet (au détriment du temps d'exécution)
- Dijkstra scalaire avec un coût égal à une scalarisation des différents coûts afin d'être rapide (au détriment de la complétude).

WRM est censé resserrer progressivement la zone où peuvent apparaître les sommets du prochain graphe autour du trajet actuel. Mais dans les cas où le trajet actuel n'est pas meilleur que le précédent, nous nous autorisons à élargir ponctuellement cette zone afin d'augmenter la chance de trouver un meilleur trajet lors de l'itération suivante.

4 Expérimentations

Les données des courants et météo sont issus de fichiers GRIB correspondant à des données réelles. Les algorithmes ont été testés sur une plage météo de 10 jours en prenant plusieurs trajets différents commençant à différents horaires. Nous avons utilisé un ordinateur 64 bits cadencé à 2.1 Ghz et ayant 192 Go de RAM.

Nous n'avons pour l'instant considéré que deux critères : le temps et le carburant consommé. Le modèle de consommation nous a été fourni par un armateur et ne constitue qu'une approximation pour un seul type de ses navires.

Les résultats des tests que nous présentons maintenant concernent un trajet de Boston vers Lisbonne à une date donnée. Les autres résultats que nous avons obtenus à d'autres dates et entre d'autres ports ne sont pas qualitativement différents.

Les tests sont tous fait avec la même graine aléatoire pour obtenir une reproductibilité des résultats et une comparaison équitable entre les différents paramétrages de l'algorithme.

Les distances sont en degrés. Pour déterminer la zone de recherche autour d'un trajet par WRM, nous fixons une distance initiale d autour du trajet initial, puis nous multiplions cette distance par 0.75 quand la solution a été améliorée et

par 1.25 quand il n'y a pas eu d'amélioration. Nous ne faisons que 10 itérations.

La figure 7 présente le résultat de la première partie de l'algorithme qui consiste à exécuter NAMOA*-TD à partir d'une grille rectangulaire.

Pour la deuxième partie, WRM utilise à nouveau NAMOA*-TD.

Les figures 8, 9 et 10 indiquent le résultat de l'application de WRM avec différentes valeurs initiales de d .

Nous constatons que plus d est grand, meilleurs sont les résultats.

Sur les différents tests, nous constatons que les solutions ont tendance à converger vers des minima locaux qui sont différents de ceux obtenus par NAMOA* lors de la première partie. Par ailleurs, plus on laisse d'espace pour diverger autour d'une solution (ie, plus d est grand), plus elles sont au final de bonne qualité (cf figure 11).

Par contre, les temps de calcul de NAMOA*-TD sont trop long (plusieurs centaines voire milliers de secondes) pour que l'approche soit acceptable en pratique. C'est pourquoi nous avons alors voulu vérifier dans un deuxième temps si la substitution de NAMOA*-TD par un Dijkstra scalaire sur chacune des parties de l'algorithme permettait d'obtenir de bons résultats (cf figure 12).

Il s'avère que les résultats obtenus sont similaires en termes de qualité des solutions mais beaucoup plus rapides à obtenir.

5 Conclusion et perspectives

Nous avons proposé une méthode de génération de trajets maritimes diversifiés et précis dans un contexte dynamique de météo et de courants changeant au cours du temps.

Notre approche est en deux phases : une phase de génération de trajets diversifiés mais peu précis et donc pas optimaux, suivie d'une phase de raffinement des lieux de passage des trajets permettant d'améliorer les objectifs (temps et consommation de carburant).

Sur nos premières expérimentations, nous avons constaté que pour chacune des deux phases, un algorithme rapide mais incomplet (Dijkstra scalaire) était plus pertinent qu'un algorithme complet. En effet, les temps de calculs étaient grandement réduits tandis que la qualité des solutions étaient préservées.

Alors que nous pensions au départ utiliser un algorithme incomplet et rapide (Dijkstra scalaire) pour améliorer les solutions d'un algorithme complet mais trop lent (NAMOA*-TD), nous nous sommes rendu compte qu'il était plus efficace de générer rapidement des solutions très approximatives car la qualité de celles-ci n'influaient pas sur la qualité des solutions finales.

Les résultats que nous avons obtenus demandent à être encore consolidés par un meilleur modèle de consommation de carburant et par la prise en compte d'autres critères (notamment environnementaux). L'applicabilité de notre approche dépend aussi de certaines contraintes réelles qui nous restent à connaître précisément : le temps maximum alloué pour l'ensemble des calculs et les ressources infor-

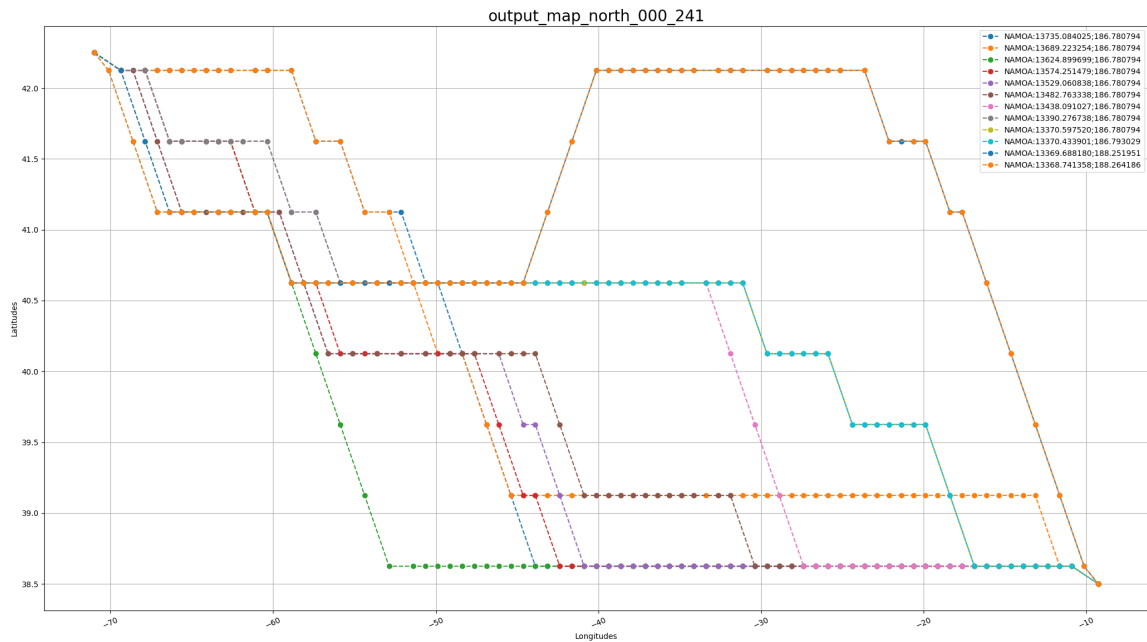


FIGURE 7 – Ensemble initial de solutions trouvées par NAMOA* dans une grille de 800 sommets. Les couples de valeurs apparaissant en haut à droite indiquent, pour chaque trajet, sa consommation et sa durée (en heures).

matiques utilisés pour ces calculs. De ceux-ci dépendent le degré de précision final des trajets que notre algorithme produit et donc de son acceptabilité.

Par ailleurs, comme pour tous les algorithmes de recherche de trajets maritimes, nous avons supposé que la vitesse était constante. Or, on sait que la consommation de carburant augmente quand la puissance du moteur change, ce qui advient quand on maintient une vitesse constante alors que les courants ou la météo évoluent. Si on veut minimiser la consommation de carburant, c’est la puissance du moteur qu’il faut rendre constante, ce qui peut impliquer un changement de vitesse. Prendre en compte cet aspect est une voie d’amélioration possible de notre approche.

6 Remerciements

Ce travail est soutenu par Bpifrance dans le cadre du projet PIA Transformation Numérique du Transport Maritime (TNTM).

Références

- [1] Estelle Chauveau. *Optimisation des routes maritimes : un système de résolution multicritère et dépendant du temps*. Theses, Aix-Marseille Université (AMU), April 2018.
- [2] E.W. Dijkstra. *A short introduction to the art of programming*. 1971.
- [3] Stéphane Grandcolas. A metaheuristic algorithm for ship weather routing. *SN Operations Research Forum*, 2022.
- [4] Peter E. Hart, Nils J. Nilsson, and Bertram Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics*, 4(2) :100–107, 1968.
- [5] Lawrence Mandow and José-Luis Pérez de-la Cruz. Multiobjective a* search with consistent heuristics. *J. ACM*, 57 :27 :1–27 :25, 2010.
- [6] Bradley S. Stewart and Chelsea C. White. Multiobjective a*. *J. ACM*, 38 :775–814, 1991.
- [7] Aphrodite Veneti, Charalampos Konstantopoulos, and G. Pantziou. Continuous and discrete time label setting algorithms for the time dependent bi-criteria shortest path problem. 2015.

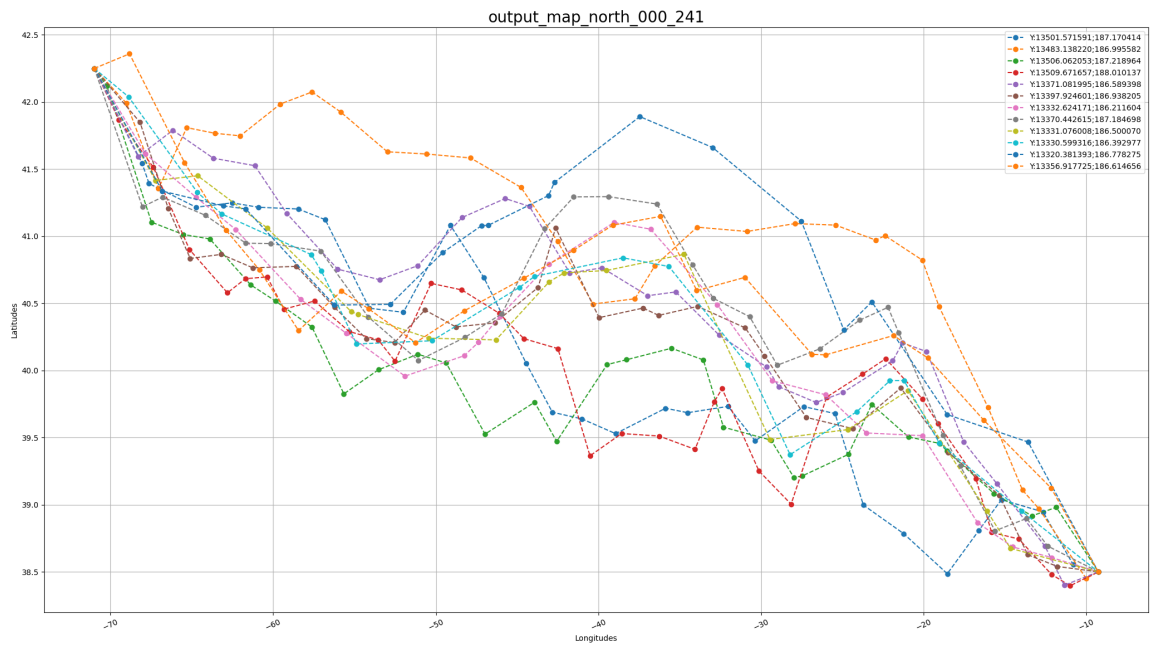


FIGURE 8 – Amelioration sur 10 iterations de la solution initiale avec $d=0.5$ et 200 sommets en utilisant NAMOA*-TD.

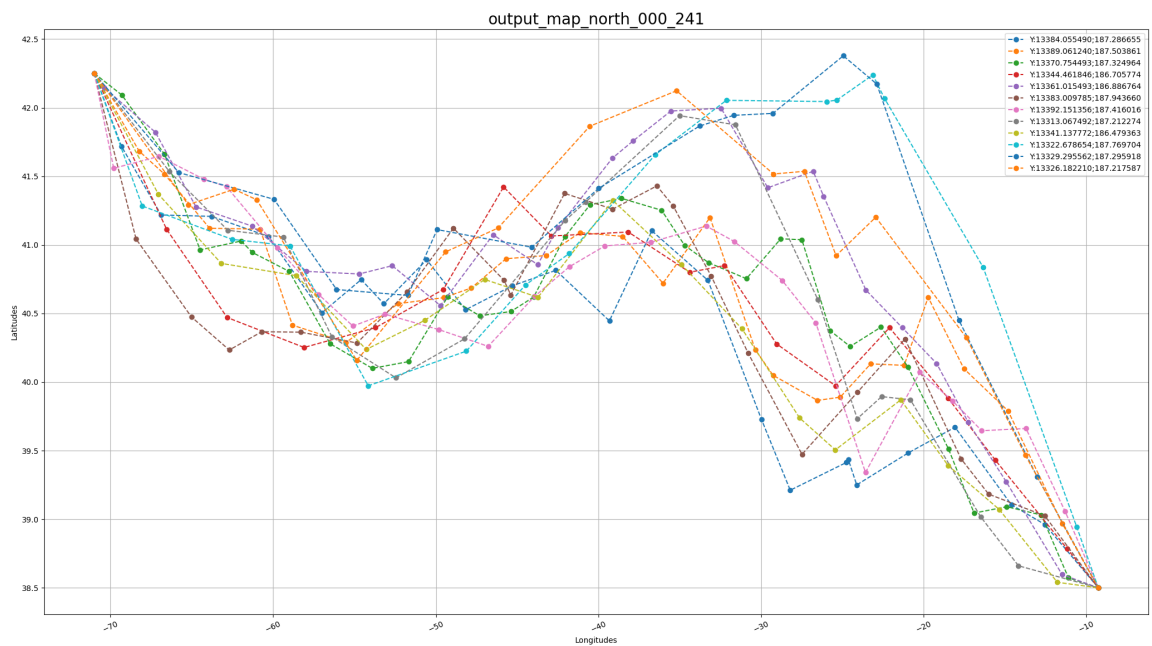


FIGURE 9 – Amelioration sur 10 iterations de la solution initiale avec $d=1.0$ et 200 sommets en utilisant NAMOA*-TD.

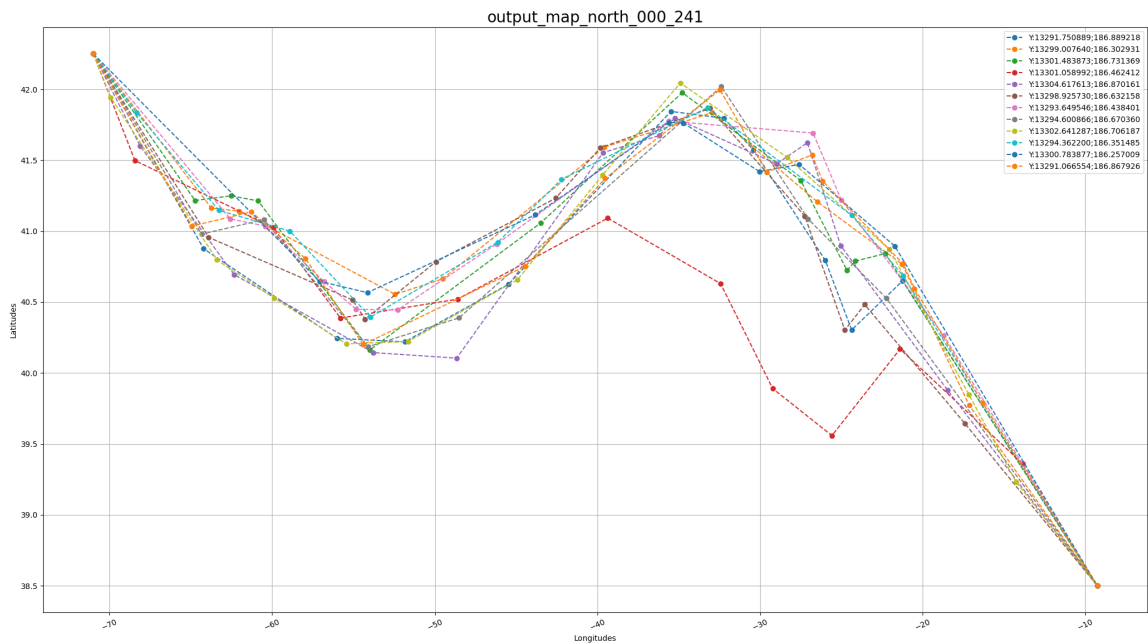


FIGURE 10 – Amélioration sur 10 iterations de la solution initiale avec $d=5.0$ et 200 sommets en utilisant NAMOA*-TD.

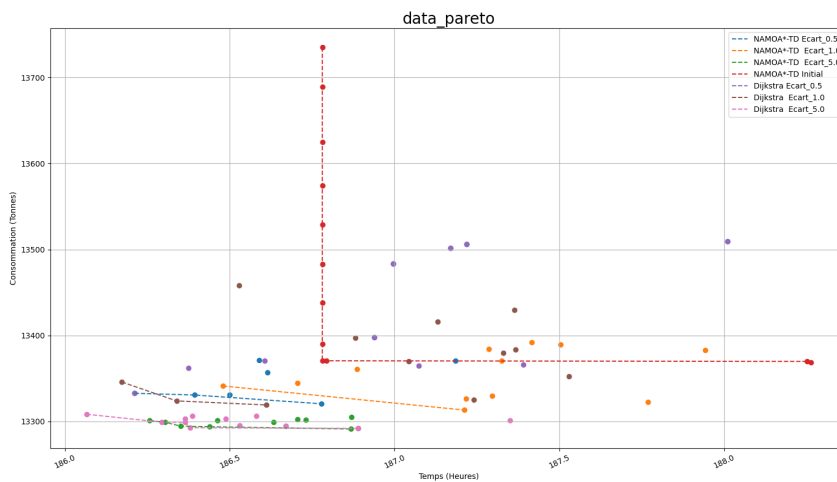


FIGURE 11 – Le front de Pareto pour l'étape 1 et ceux de l'étape 2 avec les trois valeurs de d et les deux types d'algorithmes (NAMOA*-TD et Dijkstra scalaire).

Première partie	NAMOA* (12 chemins)		Scalarisation (2 chemins)
Temps (s)	1610.12		7.743
Seconde partie	NAMOA*	Dijkstra	Dijkstra
D = 0.5	493.284	17.892	7.112
D = 1.0	1169.74	34.475	8.647
D = 5.0	> 20 min.	565.256	127.749

FIGURE 12 – Comparaison des temps d'exécution pour chacune des parties selon la méthode utilisée (NAMOA*-TD ou Dijkstra scalaire)

Résolution multicritère socialement acceptable du problème de réparation des contrats 4D dans le cadre de la gestion du trafic aérien sans pilote

Youssef Hamadi¹, Gauthier Picard²

¹ Tempero, France

² DTIS, ONERA, Université de Toulouse, France

youssefh@tempero.tech
gauthier.picard@onera.fr

Résumé

Nous utilisons des systèmes multi-agents pour résoudre les conflits entre les trajectoires de vol planifiées (contrats 4D) des drones dans le trafic urbain. Cela permet de garantir la sécurité et la qualité du service tout en tenant compte de la facilité du contrôle humain. Nous explorons différents mécanismes de prise de décision (recherche dans les graphes, enchères, etc.) et critères d'évaluation (délai, qualité, énergie) pour ces agents. Nous simulons un scénario de surveillance avec des perturbations pour tester l'efficacité de ces approches.

Mots-clés

Gestion du trafic aérien sans pilote; drone; Coordination; DCOP; Enchères SSI; Optimisation multicritère

Abstract

We use multi-agent systems to solve conflicts between planned flight paths (4D-contracts) of drones in urban traffic. This ensures safety and service quality while considering ease of human control. We explore different decision-making mechanisms (graph search, auctions, etc.) and evaluation criteria (delay, quality, energy) for these agents. We simulate a surveillance scenario with disruptions to test the effectiveness of these approaches.

Keywords

Unmanned Air Traffic Management; UAV; Coordination; DCOP; SSI Auctions; Multi-Criteria Optimization

1 Introduction

La mobilité aérienne urbaine (UAM) vise à créer des systèmes de transport sûrs utilisant des aéronefs électriques innovants pour le transport de passagers ou de marchandises dans les zones urbaines et suburbaines. Sa réussite nécessite une stratégie cohérente, capable d'harmoniser efficacement l'efficacité et la sécurité. Ce nouveau système fait face à plusieurs défis, tels que le développement d'infrastructures, l'intégration avec les systèmes de gestion du trafic aérien existants, les obstacles réglementaires et l'acceptation publique. L'acceptation publique de la gestion du trafic aérien sans pilote (UTM) est la mesure dans laquelle les individus et les communautés sont prêts à accepter et à soutenir l'intégration des aéro-

nefs sans pilote dans leur vie quotidienne. Elle est influencée par des facteurs tels que la sécurité, la protection de la vie privée et l'impact environnemental. L'acceptation publique est cruciale pour l'adoption à grande échelle des systèmes UTM, d'où l'importance de développer des technologies qui la favorisent. Dans le domaine de l'UTM, la gestion du trafic aérien (ATM) est un défi majeur. Elle se caractérise par ses complexités techniques et son rôle essentiel dans la réponse à la préoccupation principale de l'acceptation publique, à savoir garantir la sécurité [25].

Contrairement aux systèmes traditionnels conçus pour les aéronefs conventionnels à voilure fixe opérant sur de longues distances à des altitudes élevées sous la surveillance de contrôleurs aériens, l'ATM doit superviser efficacement une multitude de petits aéronefs souvent hautement automatisés ou autonomes opérant à des altitudes plus basses dans un espace aérien complexe et encombré. Ces caractéristiques nécessitent un changement d'approche substantiel, soulignant la transition vers des systèmes ATM de plus en plus autonomes et pilotés par logiciel. Le souci de la sécurité, principale exigence d'acceptation par le public, impose de nouvelles contraintes à ces systèmes. En particulier, la clarté et la compréhensibilité des solutions de déconfliction deviennent essentielles et jouent un rôle crucial dans la confiance que le public accorde à ces systèmes avancés.

La première étape vers la définition de systèmes ATM intégrés pilotés par logiciel a commencé en 2018 avec l'Administration fédérale de l'aviation des États-Unis (FAA). Cela a marqué l'introduction du premier concept mondial d'opérations (ConOps) conçu pour la gestion du trafic des systèmes d'aéronefs sans pilote (UTM) [6]. L'UTM représente le cadre de gestion permettant de faciliter les opérations multiples de drones menées au-delà de la ligne de visée visuelle (BVLOS) dans des zones où les services traditionnels de trafic aérien ne sont pas disponibles.

Dans le cadre du système UTM, illustré dans la Figure 1, une relation de collaboration existe entre les opérateurs de drones et la FAA pour déterminer et transmettre les conditions de l'espace aérien en temps réel. La FAA propose des restrictions et des paramètres instantanés aux opérateurs d'UAS (Unmanned Aircraft Systems), qui assument la responsabilité de gérer leurs opérations en toute sécurité dans le cadre de ces lignes directrices, sans services directs de contrôle du trafic aérien de la part de la FAA. La communication et la coordination entre la FAA, les opérateurs de drones et les autres parties

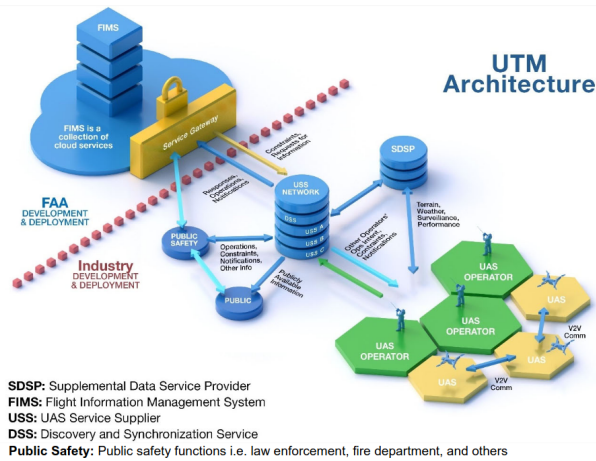


FIGURE 1 – Architecture haut niveau pour l’UTM, par la FAA [7]

prenantes s’effectuent principalement par l’intermédiaire d’un réseau décentralisé de systèmes hautement automatisés utilisant des interfaces de programmation d’applications (API), par opposition à la communication vocale traditionnelle entre les pilotes et les contrôleurs aériens [7].

Ce système est piloté par la communauté, où les opérateurs et les entités offrant des services de soutien opérationnel sont chargés d’organiser, d’exécuter et de superviser les opérations. La FAA établit un ensemble de règles de conduite et, dans ce cadre fédéré, la gestion coopérative des opérations des systèmes d’aéronefs sans pilote (UAS) est rendue possible. Cette collaboration est facilitée par des fournisseurs de services tiers (USSP) grâce à des échanges d’informations en réseau. L’UTM est spécialement conçu pour répondre à la demande croissante et aux attentes associées à un large éventail d’opérations, chacune caractérisée par une complexité et un risque accrus. Il fonctionne dans le cadre d’un marché ouvert et concurrentiel de fournisseurs de services, ce qui favorise l’innovation. Les services de l’USS sont interopérables afin de garantir que l’écosystème de l’UTM réponde aux besoins de la communauté des opérateurs UAS. Ce document de travail présente également une série de scénarios qui posent d’importants problèmes de coordination. La résolution de ces défis présente des obstacles scientifiques importants, comme le montre le document [9], et ils servent à définir des scénarios pour l’évaluation de solutions potentielles. Ici, nous nous concentrons sur la proposition UTM de la FAA, mais d’autres initiatives existent dans le monde entier, comme [21, 1, 3]. Elles diffèrent en ce qui concerne la catégorisation des drones, la division de l’espace aérien, les niveaux de sécurité et les responsabilités des prestataires de services [12, 13], mais convergent sur de nombreux aspects, en particulier en ce qui concerne la coordination entre les USSP et les UAS, qui est l’objet du présent document.

Dans cet article, nous nous concentrons particulièrement sur les interactions entre les drones et les opérateurs (partie gauche de la figure 1) afin de résoudre les conflits entre les trajectoires. En effet, la FAA identifie des scénarios dans lesquels les contrats 4D (plans de vol prédéterminés) sont remis en question

lorsque des trajectoires non prévues sont ajoutées à l’espace aérien, comme le suggère le scénario v2-2, illustré dans la Figure 2, dans lequel un hélicoptère d’évacuation médicale doit survoler la zone et, par conséquent, génère des conflits avec les trajectoires des drones engagés, ou contrats 4D [6, p. 40]. Ce scénario a été abordé dans des travaux antérieurs, à l’aide de l’optimisation distribuée des contraintes [17], mais sans tenir compte des exigences de sécurité des drones fondées sur l’énergie et de l’explicitation des critères de décision collectifs. Pour surmonter ces limitations, nous apportons les contributions suivantes. (i) Nous formalisons le problème, appelé problème de réparation des contrats 4D (4D-CRP), qui consiste à décider des actions correctives à effectuer collectivement pour résoudre les conflits entre les trajectoires 4D engagées par les drones, afin d’assurer à la fois la sécurité et la qualité du service. (ii) Nous concevons plusieurs méthodes de résolution pour résoudre la 4D-CRP : (a) un solveur centralisé basé sur la recherche de graphes pour trouver la meilleure séquence d’actions correctives à effectuer pour réparer un ensemble de trajectoires ; (b) un protocole d’enchères séquentielles distribuées à un seul élément dans lequel chaque agent détermine la meilleure séquence qu’il peut exécuter lui-même, ainsi que son coût, de sorte que l’agent dont le coût est le plus faible effectue la correction. (iii) Nous étudions plusieurs critères d’évaluation des actions, basés sur le retard résultant, les points de passage manqués résultant, la consommation d’énergie résultant, mais aussi les concessions passées, afin d’instaurer une certaine équité entre les opérateurs, si cela est souhaité. (contrairement à [17] qui n’a évalué qu’un seul critère d’évaluation). (iv) Nous avons évalué des combinaisons de mécanismes de coordination et de politiques d’évaluation afin d’évaluer les performances de ces approches dans le cadre d’un scénario tiré du ConOps FAA [6]. Afin de donner la priorité au principal critère d’acceptation par le public, à savoir la sécurité démontrée et compréhensible, nos méthodes et nos mécanismes de coordination sont conçus pour garantir en permanence la clarté et la compréhensibilité des solutions de déconfliction.

La section 2 présente brièvement et analyse les travaux connexes. La section 3 fournit des détails sur un scénario illustratif. La section 4 présente le problème, les modèles de drone et d’énergie, et détaille les critères d’évaluation utilisés pour évaluer les actions, qui sont utilisés par certaines méthodes de solution, présentées dans la section 5. Dans la section 6, nous évaluons expérimentalement certaines combinaisons de politiques et de méthodes de solution, sur des scénarios avec des trajectoires générées de manière aléatoire, des incidents et des trajectoires d’urgence. Enfin, nous concluons l’article et traçons quelques perspectives dans la section 7.

2 Travaux connexes

L’UTM a connu un effort de recherche intensif ces dernières années. Voici quelques travaux connexes qui ont abordé des problèmes proches de la 4D-CRP.

2.1 Approches centralisées

Dans [15], les auteurs introduisent une nouvelle approche centralisée pour aborder la déconfliction tactique dans l’UAM en utilisant l’optimisation mathématique. Leur méthode a un double

objectif : elle sert d'outil pour modéliser et résoudre les conflits tactiques tout en offrant des informations précieuses pour évaluer les futures infrastructures potentielles pendant la phase de conception. Ils exploitent le concept des corridors UAM pour établir une définition mathématique précise de la séparation des véhicules [23]. Leurs solutions visent à minimiser l'écart global par rapport aux plans de vol programmés, nécessaire pour éviter les violations de la séparation par paire. Cette stratégie de déconfliction intègre à la fois des ajustements en vol, réalisés par des modifications de vitesse, et des retards au sol impliquant des décollages programmés. Pour évaluer leur approche, les auteurs mènent des expériences impliquant divers scénarios de conflit dans des topologies synthétiques de réseaux UAM.

Dans leur étude, [2] s'intéresse à l'avenir des services de taxis aériens et introduit des méthodes mathématiques et des algorithmes novateurs visant à réaliser des opérations efficaces à grande échelle. Leur approche englobe deux aspects essentiels. Premièrement, ils adoptent une perspective centrée sur le passager qui introduit des classes de voyage et facilite l'échange d'informations entre les différents modes de transport, améliorant ainsi la qualité des services et garantissant des trajets multimodaux plus fluides. Deuxièmement, ils proposent une solution complète pour l'acheminement et la recharge des vols, qui non seulement minimise les coûts opérationnels immédiats, mais prolonge également la durée de vie de la batterie à long terme en réduisant les recharges gourmandes en énergie. En particulier, leur recherche se penche sur l'espace de conception des services de taxis aériens, en découvrant des solutions pour répondre aux problèmes d'équité dissimulés.

2.2 Approches décentralisées

L'étude [10] se concentre sur le modèle Sendai 2030, une projection réaliste de l'utilisation des drones pour les livraisons dans une zone spécifique du Japon. Ce cas implique jusqu'à 21 000 demandes d'opérations UAS sur une durée de service de 13 heures, ce qui représente un défi pour les méthodes de détection et de résolution des conflits avant le vol (CDR). Les auteurs proposent une méthode de réservation de l'espace aérien utilisant des trajectoires 4D et transposent le problème de la CDR avant le vol en un problème de recherche de chemin multi-agents (MAPF). Ils évaluent à la fois le traitement des demandes d'opérations UAS selon le principe du premier arrivé, premier servi (FCFS) et le traitement "par lots" et comparent leur débit. [14] s'attaquent au défi de la planification des trajectoires dans des scénarios où les demandes sont soumises par des UAS gérés par différents opérateurs. Ils introduisent un nouveau concept, à savoir le problème de la planification généralisée des trajets pour la gestion du trafic des UAS. Dans ce problème, les trajectoires des UAS sont définies par les opérateurs à l'aide de séquences de groupes de points de cheminement, et la trajectoire résultante doit traverser au moins un point de cheminement de chaque groupe. Ce problème est particulièrement pertinent pour les applications impliquant de multiples stations de recharge et des lieux de ramassage/dépose répartis dans une zone de vol désignée. Leur solution génère efficacement des trajectoires sans collision, ce qui permet un CDR avant le vol. Les résultats de la simulation démontrent l'efficacité et l'extensibilité de leur méthode, même lorsque le nombre de demandes et la taille du

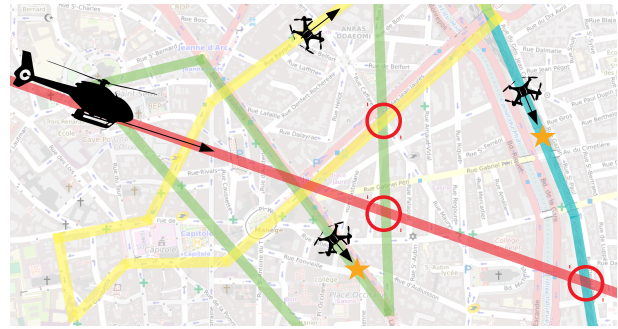


FIGURE 2 – 3 drones suivant leurs trajectoires (vert, bleu et jaune) gérant quelques incidents (étoiles orange), un hélicoptère médical sur sa trajectoire d'urgence (rouge), et des conflits identifiés (cercles rouges), d'après [17].

graphe augmentent.

[18] explorent les stratégies de gestion des conflits pour les drones autonomes. Ils examinent les procédures de déconfliction qui évitent les manœuvres horizontales et étudient deux types de structures d'espace aérien : monocouche et multicouche. Pour évaluer l'efficacité des différents paradigmes de déconfliction, ils simulent une journée d'opérations de drones dans une zone géographique. Leurs principales contributions comprennent l'introduction de nouveaux schémas de déconfliction pour les drones et une évaluation des conceptions de l'espace aérien à une ou plusieurs couches. Leurs conclusions indiquent qu'une approche centralisée ne permet pas toujours d'obtenir les meilleures performances.

[17] utilise des méthodes de résolution DCOP pour coordonner la décision des actions à sélectionner pour réparer les contrats 4D conflictuels. Pour chaque conflit, l'action qui minimise le nombre de conflits, le nombre de points de passage manqués et le retard qui en résulte est sélectionnée. Bien que cette approche ait montré de bonnes performances par rapport à un solveur centralisé, la solution ne prend pas en compte l'énergie des drones (alors qu'il s'agit d'une information cruciale lorsqu'il s'agit de véhicules ayant une autonomie énergétique très limitée) ni les décisions collectives multicritères et guidées par les concessions (certains drones peuvent devoir effectuer des actions et donc perdre en qualité de service, alors que d'autres n'ont pas fait de concession sur leurs contrats).

3 Scénario illustratif

Nous nous inspirons du document ConOps de la FAA [6] et du scénario utilisé dans [17], comme le montre la figure 2. Il est nécessaire de réviser les contrats des drones afin d'éviter les trajectoires contradictoires. Les trajectoires sont demandées à l'UTM, qui fournit alors les points de repère 4D que les drones doivent suivre pour accomplir leurs tâches, ainsi que des marges de tolérance présentées comme des tubes de sécurité. Les trajectoires initiales fournies par l'UTM sont garanties sans conflits spatiaux et temporels, mais doivent être demandées plusieurs heures avant le début des opérations.

Cependant, dans certaines situations, les drones peuvent rencontrer des difficultés à respecter leur trajectoire désignée.

Par exemple, des vents imprévisibles peuvent modifier considérablement ces trajectoires, en particulier pour les drones petits et légers [16]. En outre, des tâches spécifiques peuvent nécessiter qu'un drone maintienne sa position afin d'exécuter des routines de surveillance. Enfin, des trajectoires hautement prioritaires peuvent être introduites dans l'espace aérien en cas d'urgence. Les contrats 4D ne gèrent pas ces trajectoires conditionnelles et non déterministes, et doivent donc être mis à jour lorsque de tels événements se produisent.

Nous considérons ici les drones ayant des trajectoires de surveillance prédéfinies, qui peuvent avoir besoin de rester en vol stationnaire pour observer une scène imprévisible, interrompant ainsi temporairement leur trajectoire existante, comme dans [17]. Nous examinerons également certains événements d'urgence, dont la destination n'est pas non plus connue à l'avance, qui sont hautement prioritaires et ne peuvent être négociés, comme dans [6]. Ces deux types d'événements généreront très probablement des conflits entre les trajectoires. Ce scénario s'inscrit dans la couche UAS de la figure 1, où les drones peuvent s'engager dans un échange d'informations direct par le biais de la communication V2V, recevoir des commandes directives de la part des opérateurs et accéder à des informations en temps réel (par exemple sur les trajectoires, les conflits, les procédures d'urgence) de la part de certains USSP. Les mécanismes de coordination proposés seront mis en œuvre par plusieurs drones (semi-)autonomes ou opérateurs UAS, en fonction de l'autonomie des véhicules.

4 Concepts et modèle du problème

Cette section présente les modèles de base utilisés dans cette étude, à savoir les trajectoires 4D, les drones et les services UTM nécessaires pour mettre en œuvre les actions de déconfliction entre les drones. Nous réutilisons principalement les concepts et les notations de [17].

4.1 Trajectoires 4D

Définition 1. Une trajectoire ω est un ensemble $W \subset \mathbb{R}^4$ de 4D points $w = (x, y, z, t)$ où x et y sont des coordonnées sur le plan 2D (ou des coordonnées GPS), z est l'altitude, et t le temps.

Définition 2. Un segment $l = (w_s, w_e)$ est une ligne entre deux points consécutifs opportuns de la même trajectoire.

Seuls les segments horizontaux et verticaux sont autorisés. Les drones volent sur des plans en suivant des segments horizontaux ou d'un plan à l'autre en utilisant des segments verticaux. L'espace aérien est divisé en plusieurs plans séparés par une hauteur constante, notée z_{sep} . Sur chaque plan, les routes sont libres, c'est-à-dire qu'il n'y a pas de réseau structuré prédéfini.

Définition 3. Les tubes de sécurité sont des volumes définis horizontalement (pour les dimensions x et y), verticalement (pour z) et temporellement (pour t), par $\tau = (h, v, t)$, autour de chaque segment d'une trajectoire ω , notée $\sigma_\tau(\omega)$.

Définition 4. La trajectoire ω_1 est en conflit avec la trajectoire ω_2 lorsqu'il existe au moins un point des segments de ω_1 à l'intérieur du tube de sécurité de $\sigma_\tau(\omega_2)$.

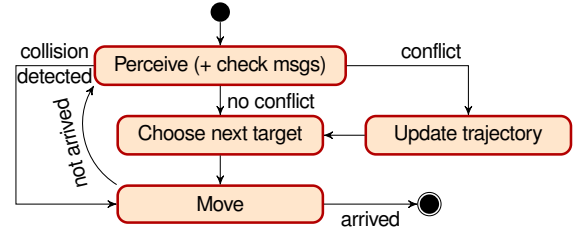


FIGURE 3 – Le modèle générique de drone.

Nous redirigeons vers [17, p.3] pour les conditions d'inclusion d'un point à l'intérieur d'un tube de sécurité. Notez que lorsque $\tau = (h, v, t)$ est le même pour tout l'espace aérien, si ω_1 est en conflit avec $\sigma_\tau(\omega_2)$, alors ω_2 est en conflit avec $\sigma_\tau(\omega_1)$.

4.2 Modèle de drones

Notre scénario considère un ensemble U de drones, en ajoutant l'état de charge au modèle de [17]. Il est primordial de surveiller attentivement l'état de charge de la batterie pendant les opérations des drones pour assurer la sécurité, la réussite de la mission et la préservation de l'équipement. Maximiser le niveau de la batterie garantit la sécurité des vols et le respect des réglementations aéronautiques [5]. Elle permet également d'optimiser l'efficacité des missions futures et l'intégrité des données en évitant les arrêts inopinés en cours de vol.

Définition 5. Un drone est défini par un tuple $u = (p, s, d, c, \omega)$, où $p = (x, y, z, t) \in \mathbb{R}^4$ est sa position, $s = (h, v, a) \leq (h_{max}, v_{max}, a_{max}) \in \mathbb{R}^3$ représente sa vitesse horizontale actuelle (en $m.s^{-1}$), verticale (en $m.s^{-1}$) et angulaire ($rad.s^{-1}$). $d \in [0, 2\pi]$ définit sa direction actuelle, c est son état de charge actuel et ω est sa trajectoire/contrat 4D.

Dans des conditions nominales, les drones changent de vitesse et de direction afin d'atteindre leur prochain point de cheminement à temps ou de retourner dans leur tube de sécurité s'ils ont été forcés d'en sortir (par exemple, en raison d'un vent fort ou pour éviter un obstacle). Lorsqu'il y a une scène à observer, un drone reste à la même position (il garde un point de cheminement à sa position actuelle) pour observer la situation pendant un temps donné, en effectuant un vol stationnaire. Il reprend ensuite sa trajectoire nominale. [17] modélise ce comportement de suivi de cible comme un modèle de mouvement simple, non détaillé ici, mais nécessaire pour simuler et évaluer les conséquences des actions correctives. En utilisant ce comportement de suivi de cible, les drones suivent une boucle classique percevoir-décider-agir, décrite dans la figure 3. Les drones interagissent avec les services UTM (par exemple, géolocalisation, détection des conflits, révision des contrats 4D) et avec les drones voisins (soit par V2V, soit par USSP).

Il est important d'ajouter la notion d'état de charge c . Lorsqu'un drone est en dessous d'un seuil d'énergie donné, il retire sa trajectoire pour atterrir en toute sécurité à la position sûre la plus proche [22]. Nous utilisons un modèle énergétique simple, qui consiste en une fonction de consommation définie comme suit :

Définition 6. La consommation d'énergie pour une vitesse 3D donnée et une durée dt , est donnée par

$$E = (c_f + c_v \max(0, v) + c_h (h^2)) dt \quad (1)$$

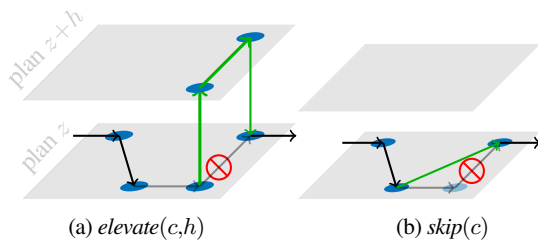


FIGURE 4 – Actions correctives ajoutant et supprimant des points de passage (cercle) et des segments (flèches) pour résoudre le conflit c (signe rouge) : les points de passage et les segments retirés sont transparents, les points de passage et les segments ajoutés sont en vert.

où le paramètre c_f est la consommation d'énergie fixe pour l'activité non inertielle du drone (calcul, communication, vol stationnaire, etc.), le paramètre c_v (resp. c_h) correspond à l'énergie pour les déplacements verticaux (resp. horizontaux).

Notez que la composante de vitesse angulaire a est considérée comme négligeable par rapport à c_f , et n'est pas utilisée pour évaluer la consommation d'énergie. Ces paramètres peuvent être déduits des caractéristiques des drones (puissance, poids, etc.). Ce modèle énergétique sera utilisé pour (i) les simulations utilisées dans l'évaluation expérimentale, où les vecteurs de vitesse sont les vitesses réelles des drones, et (ii) les évaluations des mesures correctives, où la vitesse est fixée à la vitesse de croisière des trajectoires.

4.3 Actions correctives

Pour répondre aux événements imprévus, les trajectoires doivent être corrigées afin d'éviter les conflits et donc les collisions. Avec les contrats 4D, l'espace d'action est immense et la résolution des conflits peut être complexe et difficile à comprendre. Par conséquent, comme dans la gestion traditionnelle du trafic aérien, nous donnons la priorité à des actions simples mais efficaces. De plus, la structure de l'espace aérien que nous considérons consiste en des plans séparés par une distance prédéterminée, les conflits consistent en des intersections sur le même plan. De plus, les drones sont capables d'effectuer des vols stationnaires. Compte tenu de ces contraintes opérationnelles, trois mesures correctives ont été identifiées : *postpone*, *elevate* et *skip*. [17].

Ces corrections simples favorisent l'évolutivité, la clarté et la prévisibilité : *postpone*(c,d) consiste à retarder tous les points de passage du contrat après une date donnée pour une durée donnée d , de manière à pouvoir éviter un conflit c , sans modifier la structure spatiale du graphe ; *elevate*(c,h) consiste à ajouter des points de passage pour changer de plan (soit vers le bas, soit vers le haut de la distance h), de manière à contourner l'emplacement du conflit c , puis à revenir à son plan initial ; *skip*(c) retire le segment sur lequel se produit le conflit c , en supprimant son point de passage de départ. Les changements structurels résultant de *elevate* et *skip* sont illustrés dans la Figure 4.

Pour réparer un ensemble de trajectoires conflictuelles, des actions correctives séquencées peuvent s'avérer nécessaires, car toute action corrective peut générer des conflits entre segments

après l'instant du conflit.

Proposition 1. *Il existe toujours une séquence de postpone actions correctives, avec une durée fixe d , qui résout tous les conflits pour un ensemble de trajectoires, dans l'hypothèse d'une énergie illimitée.*

Démonstration. Nous le prouvons en générant une séquence d'actions *postpone* d'une durée d , de sorte que tous les drones, à l'exception d'un drone u , reportent suffisamment longtemps pour que la trajectoire de u atteigne sa destination sans conflit (il faut autant de postpone que le temps restant pour que u atteigne sa destination divisé par d). Nous répétons le processus jusqu'à ce que tous les drones aient été traités. Cela revient à séquencer toutes les trajectoires, l'une après l'autre. \square

Toutefois, cela n'est vrai que dans le cas d'une énergie illimitée, car les trajectoires résultantes peuvent nécessiter plus d'énergie que celle qui est disponible. Une telle situation, dans laquelle tous les drones sauf un doivent attendre, n'est pas optimale, et même si elle garantit l'arrêt du processus, dans la pratique, de nombreuses autres séquences peuvent aboutir à des situations sans conflit. Dans le présent document, nous tiendrons compte des ressources énergétiques limitées, tout en supposant que les drones ont la capacité d'atterrir en toute sécurité à partir de leur position actuelle, si nécessaire [22].

Problème (4D-CRP). *Étant donné un ensemble de drones U , le Problème de réparation de contrats 4D (ou 4D-CRP) consiste à trouver un ensemble d'actions correctives pour résoudre tous les conflits entre les trajectoires des drones de U , tout en minimisant le coût global des actions correctives.*

La proposition 1 garantit qu'il existe au moins une solution à tout 4D-CRP. Toutefois, il est très probable que plusieurs solutions existent et que certaines soient meilleures que d'autres. Nous devons donc introduire des critères d'évaluation pour déterminer le coût des mesures correctives. En outre, ces critères devraient aider les opérateurs à comprendre les raisons qui sous-tendent la sélection des mesures correctives.

4.4 Évaluer le coût des actions correctives

Les actions correctives ont un impact différent sur les trajectoires : *postpone* ajoute du retard, *skip* retire un point de passage tout en conservant les mêmes points temporels, et *elevate* ajoute des points et du retard. Ces actions dégradent donc la qualité de service, c'est-à-dire le respect du contrat 4D initial. Mais plus important encore, du point de vue de la sécurité, ces actions peuvent générer d'autres conflits lors de la résolution de certains autres.

Nous considérons donc les fonctions suivantes pour évaluer le coût de l'action a quel que soit le drone qui l'exécute.

$\kappa_c(a)$ est la différence entre le nombre initial de conflits avant d'effectuer l'action a et les conflits dans l'ensemble de trajectoires résultant.

$\kappa_b(a)$ est la consommation d'énergie résultant de l'exécution de l'action a , ce qui nécessite d'évaluer l'énergie supplémentaire requise pour exécuter la nouvelle trajectoire à l'aide de l'équation 1.

$\kappa_d(a)$ est le délai résultant de l'exécution de l'action a . Pour $postpone(c, d)$ (resp. $skip(c)$) c 'est d (resp. 0), et pour $elevate(c, h)$, c 'est le temps de vol vers le haut et vers le bas de h .

$\kappa_w(a)$ est le nombre de points de passage manqués, c'est-à-dire 1 pour $skip$, 0 sinon.

Pour mettre en œuvre une évaluation multi-objectifs, nous considérons les critères de manière lexicographique, par exemple l'ordre $\kappa_c \succ \kappa_w$ signifie que nous préférons les actions avec un minimum de conflits, puis pour un nombre égal de conflits, nous préférons les actions retirant moins de points de passage. Pour entrer dans le cadre d'optimisation de la 4D-CRP, la minimisation peut être remplacée par *optimisation lexicographique*, puis encodée dans une somme pondérée équivalente [24]. Nous utiliserons cette technique dans nos algorithmes. Notons que le coût résultant d'un processus de résolution peut ensuite être décodé en un vecteur lexicographique, à des fins d'explication. Dans [17], seul l'ordre $\kappa_c \succ \kappa_w \succ \kappa_d$ est pris en compte, ce qui est limité pour deux raisons. Premièrement, l'énergie n'est pas prise en compte dans l'évaluation de l'action, alors qu'il pourrait s'agir d'un critère déterminant pour décider quel drone doit agir et donc consommer de l'énergie. Cette limitation peut être résolue en utilisant le critère κ_b . Deuxièmement, les actions passées des drones ne sont pas prises en compte lors de l'évaluation des actions. Par conséquent, il peut arriver que certains drones effectuent plus d'actions que d'autres, ce qui conduit à des situations injustes. Pour pallier cette limitation, nous proposons d'utiliser des critères liés aux concessions passées :

$\overline{\kappa_b}(u)$ est l'énergie totale concédée au cours des actions correctives passées menées par u .

$\overline{\kappa_d}(u)$ est le retard total concédé au cours des actions correctives menées par u .

$\overline{\kappa_w}(u)$ est le nombre total de points de passage retirés au cours des actions correctives passées menées par u .

Pour garantir la sécurité, nous considérerons dans nos expériences les critères lexicographiques avec κ_c comme priorité absolue ($\kappa_c \succ \kappa$ pour tout $\kappa \neq \kappa_c$).

5 Méthodes de résolution

Nous présentons trois algorithmes que nous avons mis en œuvre pour résoudre le 4D-CRP. Nous proposons des algorithmes de sélection d'actions séquentielles, afin de sélectionner des actions correctives de manière réactive. De plus, en raison des conséquences spatio-temporelles des actions, une approche pertinente consiste à considérer les conflits dans un ordre chronologique, ce qui correspond à la nécessité pour les actions correctives d'être compréhensibles pour les opérateurs de surveillance humains.

5.1 Recherche par parcours de graphe

La détermination de l'ensemble des actions à effectuer pour résoudre la 4D-CRP peut être modélisée comme un problème de recherche [19], où : (i) Chaque état $s = \langle \Omega, \hat{\kappa} \rangle$ est composé d'un ensemble de trajectoires $\Omega = \{ \omega, \forall u = (p, s, d, c, \omega) \in U \}$, et d'une cartographie $\hat{\kappa} : u \mapsto \overline{\kappa}(u)$ associant à chaque drone u dans U un certain coût d'évaluation évalué par les critères $\overline{\kappa}$, qui peuvent être n'importe quel critère spécifique à u (comme

$\overline{\kappa_b}$, $\overline{\kappa_d}$ ou $\overline{\kappa_w}$) ou n'importe quelle combinaison lexicographique de ces critères ; (ii) Un état initial s_0 est constitué des trajectoires actuelles et des concessions actuelles faites par chaque drone u jusqu'à présent ; (iii) Un état but est tout état s où Ω est sans conflit - dont l'existence est garantie par la Proposition 1 ; (iv) La fonction *successeur* consiste à effectuer une action a pour résoudre le premier conflit (dans l'ordre chronologique) dans l'état $s = \langle \Omega, \hat{\kappa} \rangle$ qui résulte dans l'état $s' = \langle \Omega', \hat{\kappa}' \rangle$, où Ω' est le nouvel ensemble de trajectoires (seule une trajectoire ω a été mise à jour), et $\hat{\kappa}'$ est mis à jour avec les coûts d'évaluation pour le drone u effectuant l'action a ; (v) La fonction de coût pour l'exécution de l'action a dans l'état $s = \langle \Omega, \hat{\kappa} \rangle$ peut être n'importe laquelle des fonctions d'évaluation d'action susmentionnées (voir la section précédente) calculée en utilisant soit a , soit a et $\hat{\kappa}$.

En raison de la nature cumulative du coût (même pour les fonctions lexicographiques codées en sommes pondérées équivalentes) et de la nature non informée du problème (pas de limite supérieure informative en raison d'événements imprévisibles), nous optons pour une *recherche à coût uniforme* (UCS) [19], qui explore d'abord les états résultant d'actions dont le coût cumulatif du chemin est minimal. Nous notons $ucs(P)$ le résultat de ce processus de recherche sur 4D-CRP P . Il renvoie une séquence d'actions A^P et un coût κ^P .

La solution est une séquence d'actions visant à atteindre des trajectoires sans conflit, qui peuvent mettre à jour des segments dans un avenir lointain, par rapport à la fréquence des événements générateurs de conflits. Garantir une telle absence de conflit à long terme peut s'avérer inutile ; ainsi, limiter à la fois l'horizon temporel (en ne considérant que les k premiers points de chaque trajectoire) et la profondeur de la recherche (c'est-à-dire une séquence de taille k) sont deux moyens faciles d'améliorer le temps de calcul, sans pour autant dégrader la qualité.

Il convient de noter que cette solution est entièrement centralisée, ce qui nécessite de fournir un service à l'ensemble de l'espace aérien pour installer la correction de trajectoire.

5.2 Enchères séquentielles

Considérons maintenant une approche plus distribuée, où chaque drone impliqué dans le premier conflit dans l'ordre chronologique calcule une séquence d'actions qu'il peut effectuer par lui-même sur les trajectoires conflictuelles, puis partage cette solution avec ses voisins en conflit, et enfin la séquence à coût minimal est sélectionnée pour résoudre le conflit. Cette approche locale peut s'appuyer sur la même procédure de recherche que précédemment, mais limitée aux actions d'un seul drone, et est notée $ucs(P, u)$ pour le drone u , et renvoie une séquence d'actions et son coût, et donc son vecteur lexicographique.

Cette approche, illustrée dans l'algorithme 1, revient à mettre en œuvre des enchères séquentielles à un seul élément (SSI), classiquement utilisées dans l'attribution des tâches [20]. Ici, les éléments sont des conflits et les offres sont évaluées en résolvant le 4D-CRP P en limitant les actions à un seul drone u , en appelant $ucs(P, u)$. Une fois qu'un gagnant est déterminé pour un conflit donné, la séquence d'actions gagnante est exécutée sur les trajectoires Ω , en appelant $apply(a, \Omega)$. Bien qu'elle ne soit pas entièrement coordonnée, cette approche locale présente l'avantage de permettre à chaque drone de calculer une séquence

Algorithme 1 : Enchères séquentielles**Data** : Un 4D-CRP P , avec drones dans U et trajectoires Ω **Result** : une séquence d'actions $A = \{a_1, \dots, a_n\}$

```

1  $A \leftarrow \emptyset$ 
2 while  $\Omega$  a des conflits do
3    $c \leftarrow$  premier conflit par ordre chronologique
4    $U^c \leftarrow \{u \text{ impliqué dans } c\}$ 
5   for each  $u \in U^c$  do concurrently
6      $A_u^c, \kappa_u^c \leftarrow \text{ucs}(P, u)$  // miser
7     // envoyer  $A_u^c, \kappa_u^c$  au commissaire
8    $w \leftarrow \text{argmin}_{u \in U^c} \{\kappa_u^c\}$ 
9     // détermination du gagnant
10   $A \leftarrow AU \cup A_w^c$ 
11  for each  $a \in A_w^c$  do  $\Omega \leftarrow \text{apply}(a, \Omega)$ 
12  // mise à jour des contrats
13 return  $A$ 

```

simple sans autorité centrale de contrôle. Cependant, un *service d'enchères* (AS) est nécessaire pour recueillir les offres et déterminer les gagnants. Ce service peut potentiellement être fourni par un drone, un USS ou être décentralisé en utilisant des approches basées sur le consensus, comme CBBA [4].

5.3 Optimisation sous contraintes distribuée

Nous considérons ici une autre approche distribuée, où le 4D-CRP est codé comme un problème d'optimisation de contraintes distribuée (DCOP), comme inspiré par [17].

Nous rappelons brièvement qu'un DCOP est un tuple $\langle \mathcal{A}, \mathcal{X}, \mathcal{D}, \mathcal{C}, \mu, f \rangle$. Ici, $\mathcal{X} = \{x_{u,a,c}\}$ sont des variables binaires concernant le drone u qui effectue ou non l'action a pour résoudre le conflit c ; $\mathcal{D} = \{\mathcal{D}_{x_{u,a,c}} = \{0,1\}\}$ est un ensemble de domaines binaires, un domaine par variable; L'élément \mathcal{C} est un ensemble de contraintes douces appartenant à deux catégories, à savoir les coûts unaires des actions et les contraintes "au plus un" empêchant le même conflit d'être résolu plus d'une fois. $\alpha: \mathcal{X} \rightarrow \mathcal{A}$ associe chaque variable $x_{u,a,c}$ à l'agent u ; L'élément f est une fonction objective, représentant le coût global (somme) d'une affectation de variable complète.

Nous renvoyons le lecteur à l'article [17], pour un modèle DCOP détaillé. Il est important de noter que notre modèle diffère à deux égards. Tout d'abord, nous ne considérons qu'un seul conflit à résoudre à la fois de manière itérée, alors que [17] considère tous les conflits en cours en même temps. Cela signifie que, comme pour la méthode basée sur les enchères, les agents se coordonnent séquentiellement pour résoudre un seul conflit avant de résoudre le suivant. Comme pour les deux méthodes précédentes, ceci est motivé par le fait que la résolution des conflits dans un ordre chronologique est plus efficace que la résolution des conflits à long terme avec un meilleur coût, mais qui peut être inutile lors de la résolution des conflits à court terme, tout en conduisant à des solutions plus pratiques et explicites. Deuxièmement, [17] n'a considéré qu'une seule fonction objective, équivalente à l'optimisation lexicographique de $\kappa_c \succ \kappa_w \succ \kappa_d$. Ici, nous évaluerons les performances d'une formulation DCOP sur plusieurs fonctions objectives (voir expériences). Nous tirerons également parti de l'étude réalisée

dans [17], en utilisant uniquement l'algorithme Asynchronous Forward Bounding (AFB) [8], qui a affiché les meilleures performances parmi une grande famille d'algorithmes DCOP.

6 Évaluation expérimentale

Nous comparons ici les algorithmes proposés, combinés à des critères d'évaluation, sur des ensembles de trajectoires perturbées par des incidents et des trajectoires d'urgence.

Les expériences sont mises en œuvre en Java et exécutées sur un processeur Intel(R) Xeon(R) E5-2660 v3 à 20 cœurs @ 2,60 GHz, 62 Go de RAM, Ubuntu 18.04.5 LTS, avec une JVM OpenJDK 11.0.9. Nous exécutons 30 instances d'ensembles de trajectoires et d'incidents générés aléatoirement pour chaque taille de flotte $n = \{5, 10, 15, 20, 25\}$, et nous traçons les valeurs moyennes, avec un intervalle de confiance de $[0.05, 0.95]$. Le solveur DCOP utilisé est l'implémentation AFB de la bibliothèque FRODO [11]. Les valeurs générées aléatoirement sont choisies uniformément dans les intervalles fournis. Le temps de calcul indiqué est le temps de simulation mono-CPU.

Nous rappelons les paramètres de génération du scénario de [17]. La zone est de 1000m par 1000m, avec des plans d'espace aérien verticaux à 20m, 40m et 60m. $h_{max} = 18m.s^{-1}$, $v_{max} = 6m.s^{-1}$, $a_{max} = \Pi/2rad.s^{-1}$, $\Delta h_{max} = \Delta v_{max} = 6m.s^{-2}$, $\Delta a_{max} = \Pi/2rad.s^{-2}$. La vitesse initiale est fixée à $(0,0,0)$. Les trajectoires initiales du drone sont générées de manière aléatoire et incrémentale en choisissant au hasard 30 points séparés par une distance de 200m à 1000m, éventuellement sur des plans séparés. Les trajectoires ont un segment supplémentaire pour revenir à leur point d'origine. Les tubes de sécurité sont définis par $(h, v, t) = (30m, 15m, 10s)$. Des trajectoires d'urgence de 10 sont générées, partant d'une position aléatoire à la frontière de la zone, se déplaçant vers un point aléatoire à une distance d'au moins 1km du point initial, puis un point d'attente pour 360s, et enfin un chemin de retour vers son origine. Des incidents supplémentaires se produisent avec 5% de chance toutes les secondes, à proximité d'un drone choisi aléatoirement avec une distribution normale $\mathcal{N}(0.5, 0.15)$ mise à l'échelle et coupée en $[0, n]$ pour déséquilibrer la charge entre les drones devant effectuer un vol stationnaire pendant 30s à 120s. Nous utilisons le même modèle énergétique que celui défini par l'équation 1, et équipons les drones de batteries de 30000mAh (provenant de missions de plus d'une heure).

Nous évaluons : **ucs**, qui résout les conflits avec le solveur centralisé basé sur la recherche graphique; **ssi**, qui résout les conflits avec les enchères séquentielles d'un seul article; **sdcop**, qui résout les conflits (un par un) avec AFB. Les drones peuvent effectuer les actions suivantes : *postpone*(c, d) avec $d \in \{20, 40, 60\}$, *elevate*($c, \pm 20$), et *skip*(c). Ces actions sont évaluées à l'aide de certains critères lexicographiques, qui ont tous κ_c en premier (pour garantir la sécurité), et utilisent toujours le hasard pour départager les ex-aequo. Dans la figure, nous les notons comme suit : **b** $\equiv \kappa_c \succ \kappa_b$, **d** $\equiv \kappa_c \succ \kappa_d$, **w** $\equiv \kappa_c \succ \kappa_w$, **wd** $\equiv \kappa_c \succ \kappa_w \succ \kappa_d$, **bwd** $\equiv \kappa_c \succ \kappa_b \succ \kappa_w \succ \kappa_d$, **b concession** $\equiv \kappa_c \succ \overline{\kappa_b} \succ \kappa_b$, **d concession** $\equiv \kappa_c \succ \overline{\kappa_d} \succ \kappa_d$, and **w concession** $\equiv \kappa_c \succ \overline{\kappa_w} \succ \kappa_w$.

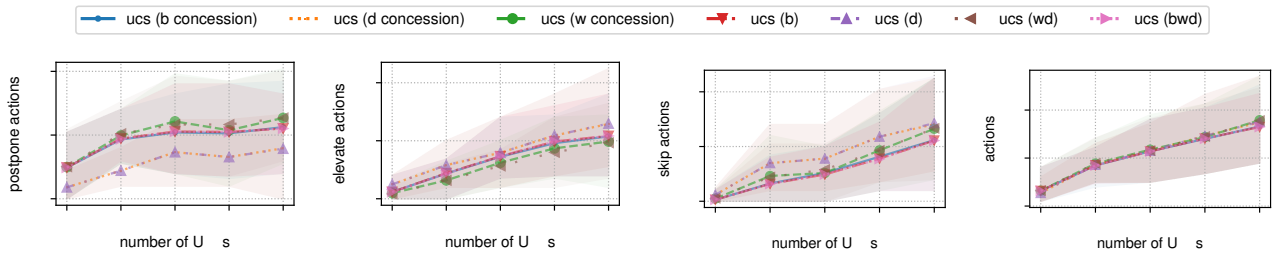


FIGURE 5 – Décisions prises avec les différents critères d'évaluation des coûts lorsqu'ils sont utilisés avec le solveur UCS.

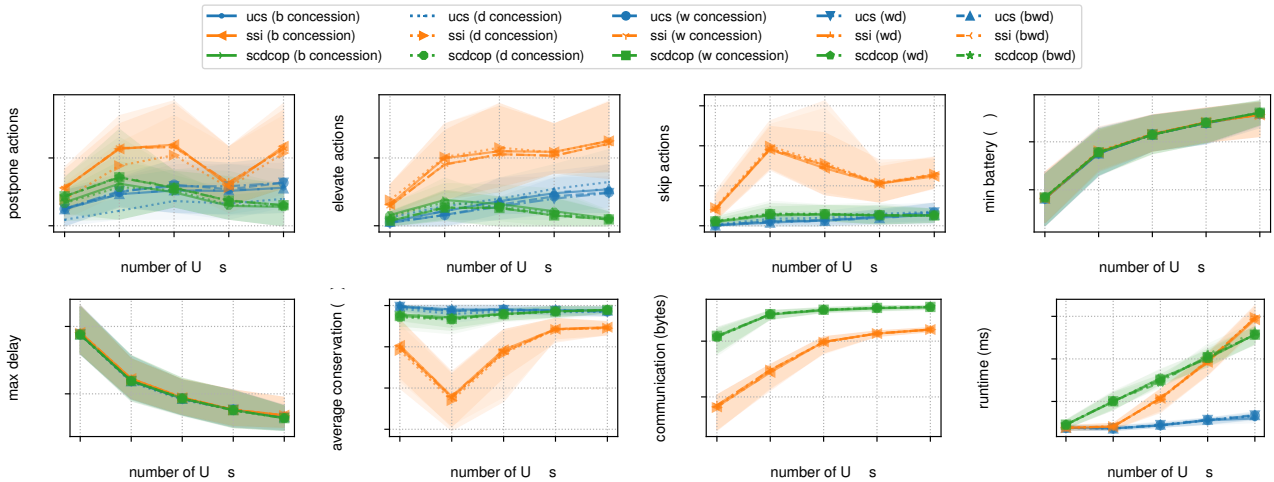


FIGURE 6 – Valeurs moyennes sur 30 instances pour plusieurs mesures de performance avec un nombre croissant de drones.

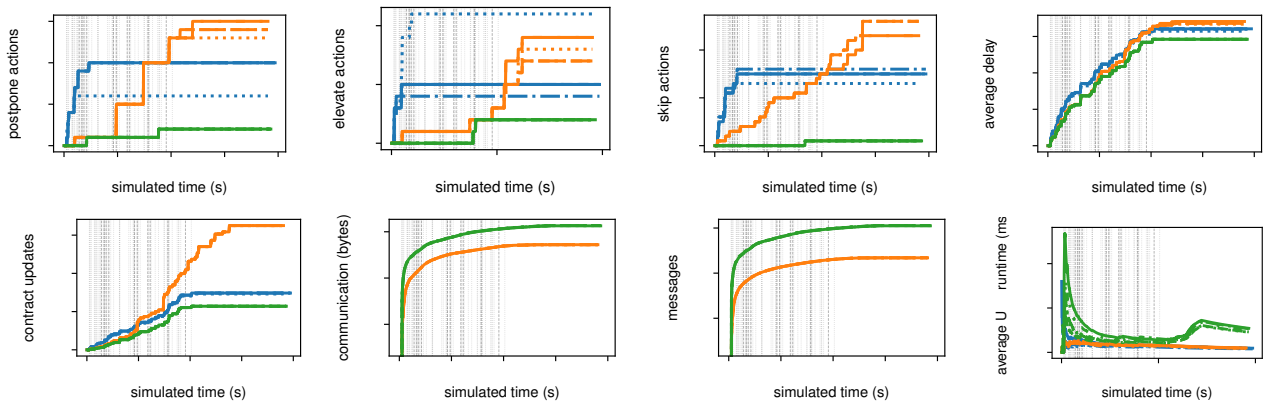


FIGURE 7 – Résultats d'une simulation avec 25 drones et 10 procédures d'urgence (en pointillés gris) et 46 incidents (en pointillés gris).

6.1 Effets des critères sur les choix d'action

Pour comprendre les conséquences de l'utilisation de certains critères d'évaluation par rapport à d'autres, nous analysons le nombre de fois où les trois actions (*postpone*, *elevate*, *skip*) sont déclenchées pour chacun de ces critères d'évaluation, comme l'illustre la Figure 5. Pour des raisons de lisibilité, nous présentons ici les résultats du solveur centralisé UCS seul. Nous pouvons observer de légères divergences dans les actions préférées résultant de l'utilisation de différents critères. Les critères *d* et *d concession* préfèrent utiliser les actions *postpone* et promouvoir les actions *skip* puis *elevate* afin de réduire les délais. *w*, *wd* et *w concession* favorisent *elevate* pour conserver autant de points de passage que possible. Enfin,

en ce qui concerne les critères liés à la batterie, *b*, *bwd* et *b concession* tendent à réaliser des compromis entre les deux familles susmentionnées en ce qui concerne *postpone* et *elevate*, et évitent étonnamment d'utiliser les actions *skip*.

Il est à noter que, malgré la densité de l'espace aérien que nous générons, relativement peu d'actions correctives sont déclenchées (environ 20, ce qui est encore extrême dans le domaine UTM, où la gestion de plus de 2 conflits à la fois est très improbable), quel que soit le critère d'évaluation. L'espace libre est très flexible et peu sujet aux conflits, contrairement à l'espace aérien structuré. Ainsi, malgré la forte dynamique et la fréquence élevée des événements imprévisibles, peu de conflits sont générés. Compte tenu de l'horizon court utilisé dans ces simulations, les mécanismes de concession ne ressortent pas vraiment, car les

agents n'ont pas assez de temps pour accumuler des concessions afin d'en tirer profit lors de la résolution des conflits. Cependant, nous pensons que ce mécanisme pourrait être utile dans des contextes plus longs, même avec des flottes multiples.

6.2 Comparaison des solveurs 4D-CRP

Nous allons maintenant comparer les performances des différents algorithmes dans le même contexte, illustré dans la Figure 6. Nous analysons en particulier le nombre d'actions déclenchées par type, comme dans la section précédente, dans les trois premières figures de la première ligne. Il est clair que ssi déclenche beaucoup plus d'actions correctives de tout type, et donc de nombreuses révisions du contrat 4D. Cela est dû à sa façon de résoudre chaque conflit en utilisant une séquence d'actions, à l'aide d'une version locale de ucs, qui peut s'avérer inutile à l'avenir. Bien qu'il soit très rapide dans des contextes plus restreints (figure en bas à droite), il nécessite presque 8 fois moins de partage d'informations que sdcop (deuxième figure en partant du bas à gauche). Curieusement, ssi éprouve des difficultés dans certains contextes (taille 10) : il génère de nombreuses actions *skip*. Cela signifie que ssi ne peut pas trouver de séquences d'actions mono-agent capables de sauvegarder des waypoints, en utilisant *elevate* ou *postpone*; cela est probablement dû au fait que les agents ne peuvent pas trouver de bonnes séquences de *postpone*, puisqu'ils sont déjà tous aux prises avec l'impact du délai de surveillance, et qu'ils ne peuvent pas s'équilibrer. sdcop tend également à déclencher plus d'actions que ucs, mais dans un ordre limité par rapport à ssi. En effet, sdcop ne déclenche qu'une action par conflit détecté de manière séquentielle. Il ne génère donc pas localement une séquence d'actions potentiellement inutiles, comme le fait ssi. sdcop enregistre autant de points de passage que ucs dans des contextes plus larges. Il est intéressant de noter qu'avec des flottes plus importantes, le temps d'exécution de sdcop augmente linéairement par rapport à ssi et ucs, étant donné qu'il ne prend en compte que les séquences d'actions uniques. Enfin, il est particulièrement intéressant de constater que tous les algorithmes améliorent l'équité des délais et des batteries : plus il y a d'agents, plus il y a d'actions permettant d'équilibrer l'allocation des délais (délai maximum) et les dépenses d'énergie (niveau minimum de batterie) face à un événement (alors que ces deux éléments sont proportionnels au nombre d'agents).

Enfin, la figure 7 montre l'évolution de certaines mesures au cours de l'exécution d'un scénario. Il est intéressant de noter que ucs répare principalement les conflits au début du scénario, ce qui se traduit par des trajectoires plus robustes aux conflits à venir; tandis que sdcop déclenche peu d'actions tout au long de la ligne temporelle. Les séquences de ssi sont souvent révisées jusqu'à la fin du scénario. Cependant, même les décisions mono action (pourtant très pratiques à appréhender du point de vue de l'opérateur) requièrent une grande quantité de données et de messages. Notez que la bosse à la fin de la figure en bas à droite (durée d'exécution moyenne des drones) est due au fait que certains drones sont arrivés et ne sont pas comptés dans la somme globale, augmentant ainsi la valeur moyenne.

7 Conclusion

L'approche de cet article consiste à résoudre le problème de réparation de contrat (4D-CRP) à chaque fois que les drones détectent des conflits. Nous avons étudié plusieurs solveurs : ucs, ssi, et sdcop. En outre, plusieurs fonctions d'évaluation du coût des actions ont été définies, sur la base des conséquences immédiates et des concessions passées. Nous avons également intégré le critère de l'énergie dans la décision, ce qui favorise les actions visant à économiser la batterie, conformément aux exigences réglementaires [5]. Nous avons évalué plusieurs combinaisons de solveurs et de critères, sur un espace aérien conflictuel. Dans l'ensemble, les opérateurs de drones et les gestionnaires d'UTM (tels que la FAA) peuvent choisir entre plusieurs mécanismes flexibles et compréhensibles pour coordonner les corrections de trajectoire, allant de la centralisation à la décentralisation. Chacun de ces mécanismes a des impacts opérationnels distincts (charge de communication, temps de calcul, changements d'altitude, etc). En outre, les parties prenantes de l'UTM ont le choix entre plusieurs critères, afin d'améliorer l'acceptabilité et de mettre en place des décisions explicables, consistant en des séquences d'actions avec des évaluations ordonnées correspondantes. Pendant le déploiement, ces priorités pourraient être fixées par consensus social, en tenant compte des préférences de la communauté touchée, telles que la minimisation du vol stationnaire, etc.

Nos contributions ont été évaluées sur des scénarios comportant un nombre limité de conflits. L'extension à des contextes plus vastes, sur plusieurs heures, avec de nombreuses itérations de drones, pourrait mettre davantage en évidence les avantages de la prise en compte des concessions, en particulier dans les scénarios impliquant des flottes hétérogènes exploitées par différents acteurs. Cela pourrait également être le cas entre plusieurs flottes aux caractéristiques et missions hétérogènes, exploitées par des acteurs distincts. D'autres approches fondées sur le marché (et non coopératives) doivent encore être étudiées. Nous constatons que l'élaboration de longues séquences d'actions, avec une connaissance locale limitée (ssi), peut s'avérer inefficace en cas de conflits fréquents. Cependant, dans des environnements plus stables, avec une prédiction possible de la densité de l'espace aérien, de telles séquences pourraient potentiellement être efficaces, comparées aux algorithmes myopes (sdcop). En outre, le passage à des espaces aériens structurés, connus pour être plus conflictuels en raison des effets de congestion, pourrait nécessiter l'utilisation d'approches basées sur des concessions et des décisions locales, comme nous l'avons proposé. Enfin, étant donné que, dans un avenir proche, l'UTM sera géré par des opérateurs, des expériences humaines en boucle pourraient être nécessaires pour adapter les explications des mesures correctives afin d'optimiser la compréhension et la réactivité de l'utilisateur. En résumé, l'acceptabilité sociale en constante évolution, le paysage changeant des algorithmes en réponse aux futures informations sur l'espace aérien urbain (par exemple, plus de prévisions sur les opérations futures) et la prise en compte des flottes avec des préférences potentielles de déconflition interne impliquent que les algorithmes de déconflition UTM doivent être capables d'évoluer en tandem avec l'avenir des opérations et de l'acceptation des UTM. Ces travaux constituent une première étape

dans la définition de cette catégorie algorithmique émergente.

Références

- [1] Airbus and Boeing. A new digital era of aviation : The path forward for airspace and traffic management. Technical report, Airbus, 2020.
- [2] Mehdi Bennaceur, Rémi Delmas, and Youssef Hamadi. Passenger-centric urban air mobility : Fairness trade-offs and operational efficiency. *Transportation Research Part C : Emerging Technologies*, 136 :103519, 2022.
- [3] Carlos Capitán, Héctor Pérez-León, Jesús Capitán, Ángel Castaño, and Anibal Ollero. Unmanned aerial traffic management system architecture for u-space in-flight services. *Applied Sciences*, 11(9), 2021.
- [4] Han-Lim Choi, Luc Brunet, and Jonathan P. How. Consensus-based decentralized auctions for robust task allocation. *IEEE Trans. Robotics*, 25(4) :912–926, 2009.
- [5] European Union Aviation Safety Agency (EASA). Easy access rules for unmanned aircraft systems (regulation (eu) 2019/947 and regulation (eu) 2019/945). Technical report, European Union Aviation Safety Agency (EASA), 2022.
- [6] Federal Aviation Agency. Utm concept of operations version 2.0. Technical report, Federal Aviation Agency, 2022.
- [7] Federal Aviation Agency. Unmanned aircraft systems traffic management (utm) implementation plan. Technical report, Federal Aviation Agency, 2023.
- [8] Amir Gershman, Amnon Meisels, and Roie Zivan. Asynchronous forward-bounding for distributed constraints optimization. In *Proceedings of the 2006 Conference on ECAI 2006 : 17th European Conference on Artificial Intelligence August 29 – September 1, 2006, Riva Del Garda, Italy*, page 103–107, NLD, 2006. IOS Press.
- [9] Youssef Hamadi. Optimization for urban air mobility. In Ilias S. Kotsireas and Panos M. Pardalos, editors, *Learning and Intelligent Optimization*, pages 1–8, Cham, 2020. Springer International Publishing.
- [10] Florence Ho, Rúben Geraldes, Artur Gonçalves, Bastien Rigault, Atsushi Oosedo, Marc Cavazza, and Helmut Prendinger. Pre-flight conflict detection and resolution for uav integration in shared airspace : Sendai 2030 model case. *IEEE Access*, 7 :170226–170237, 2019.
- [11] Thomas Léauté, Brammert Ottens, and Radoslaw Szymanek. FRODO 2.0 : An open-source framework for distributed constraint optimization. In *Proceedings of the IJCAI'09 Distributed Constraint Reasoning Workshop (DCR'09)*, pages 160–164, Pasadena, California, USA, July 13 2009. <https://frodo-ai.tech>.
- [12] J. Lieb and A. Volkert. Unmanned aircraft systems traffic management : A comparison on the faa utm and the european corus conops based on u-space. In *2020 AIAA/IEEE 39th Digital Avionics Systems Conference (DASC)*, pages 1–6, 2020.
- [13] Tim McCarthy, Lars Pforte, and Rebekah Burke. Fundamental elements of an urban utm. *Aerospace*, 7(7), 2020.
- [14] Rafael Papa, Ionut Cardei, and Mihaela Cardei. Generalized path planning for utm systems with a space-time graph. *IEEE Open Journal of Intelligent Transportation Systems*, 3 :351–368, 2022.
- [15] Mercedes Pelegrín, Claudia D'Ambrosio, Rémi Delmas, and Youssef Hamadi. Urban air mobility : from complex tactical conflict resolution to network design and fairness insights. *Optimization Methods and Software*, pages 1–33, 08 2023.
- [16] Gabriele Perozzi, Denis Efimov, Jean-Marc Biannic, and Laurent Planckaert. Trajectory tracking for a quadrotor under wind perturbations : sliding mode control with state-dependent gains. *Journal of the Franklin Institute*, 355(12) :4809–4838, 2018.
- [17] Gauthier Picard. Trajectory coordination based on distributed constraint optimization techniques in unmanned air traffic management. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS-22)*, pages 1065–1073. IFAAMAS, 2022.
- [18] Valentin Polishchuk. Centralized and distributed utm in layered airspace. 2018.
- [19] Stuart Russell and Peter Norvig. *Artificial Intelligence : A Modern Approach*. Prentice Hall, 3 edition, 2010.
- [20] Adrian Schoenig and Maurice Pagnucco. Evaluating sequential single-item auctions for dynamic task allocation. In Jiuyong Li, editor, *AI 2010 : Advances in Artificial Intelligence*, pages 506–515, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.
- [21] SESAR joint Undertaking. European atm master plan : Roadmap for the safe integration of drones into all classes of airspace. Technical report, SESAR, 2020.
- [22] Md Shah Alam and Jared Oluoch. A survey of safe landing zone detection techniques for autonomous unmanned aerial vehicles (uavs). *Expert Systems with Applications*, 179 :115091, 10 2021.
- [23] Savvy Verma, Victoria Dulchinos, Robert Dan Wood, Amir Farrahi, Richard Mogford, Megan Shyr, and Rania Ghatas. Design and analysis of corridors for uam operations. In *2022 IEEE/AIAA 41st Digital Avionics Systems Conference (DASC)*, pages 1–10, 2022.
- [24] Ronald R. Yager. On the analytic representation of the leximin ordering and its application to flexible constraint propagation. *European Journal of Operational Research*, 102(1) :176–192, 1997.
- [25] Pavan Yedavalli and Jessie Mooberry. An assessment of public perception of urban air mobility (uam). *Airbus UTM : Defining Future Skies*, pages 2046738072–1580045281, 2019.

CapData Opéra : faciliter l'interopérabilité des données des maisons d'opéra

E. Peyre¹, F. Amarger², N. Chauvat²

¹ Réunion des Opéras de France, ROF

² Logilab

eudes.peyre@rof.fr
prenom.nom@logilab.fr

Résumé

Le projet CapData Opéra, mené à l'initiative de la ROF¹ utilise les technologies du Web sémantique comme fondement d'une solution de structuration et de diffusion des données culturelles capable de répondre aux enjeux de développement des publics, de soutien à la création artistique et d'accès à la culture.

Cette solution de mutualisation permet d'interroger les données produites par plusieurs acteurs du domaine pour, par exemple, connaître la programmation et la circulation d'une œuvre ou d'une production entre plusieurs maisons d'opéra.

Pour faciliter la publication des données produites par chaque maison d'opéra, la ROF propose une ontologie du domaine, des référentiels, une infrastructure de publication, des outils et de l'accompagnement humain.

Dans cet article, nous présentons les objectifs et les moyens mis en œuvre pour fédérer des données hétérogènes, nous faisons un retour sur expérience en abordant les aspects techniques et la gestion, et nous décrivons les résultats actuels et les perspectives de ces travaux.

Mots-clés

ROF, Opéras, RDF, Interopérabilité

Abstract

The "CapData Opéra" project, initiated by ROF (Réunion des Opéras de France - French Opera Association) and supported by the French Ministry of Culture, uses Semantic Web technologies to share cultural data with the public and the artistic community.

The aim is to aggregate data produced by various domain actors to make it globally searchable. This highlights previously invisible data, such as the exchange of creative works and performances between opera houses. To achieve this, an ontology has been designed to define a common vocabulary and implement data interoperability objectives. This ontology is aligned with schema.org, and we are working to align additional models. A set of SHACL rules has been created to validate the data before publication. A dedicated tool, Rodolf, has been developed to monitor the

RDF publishing process. This tool is used to execute the process and track which sources have been uploaded to the SPARQL endpoint, including upload times and any errors encountered. Exporting RDF data can be challenging for institutions unfamiliar with Semantic Web technologies, so a dedicated Software Development Kit (SDK) has been developed to assist web developers in exporting CapData RDF data even if they lack experience in this area.

In this presentation, we aim to share with the SWIB community the objectives and solutions we have found to federate heterogeneous data. We will present feedback on this project, focusing on technical and management aspects, and then describe the results we have achieved and the future of this project.

Keywords

ROF, Opera, RDF, Interoperability

1 Introduction

Les structures et les établissements culturels du spectacle vivant ont connu à la fin des années 1990 une profonde révolution liée aux transformations induites par le développement d'Internet et du numérique. Deux décennies plus tard, les enjeux liés à la diffusion et aux échanges des données produites prennent une importance majeure.

Si les stratégies et solutions développées par les politiques culturelles et les acteurs de ce secteur, dont les maisons d'opéra, ont été menées de manière relativement homogène et précoce en réponse aux enjeux de démocratisation, de création, de valorisation et de médiation auprès des publics, les problématiques liées à la gestion et au partage des données produites n'ont émergé que plus récemment.

Pour mettre en valeur leur programmation et interagir efficacement avec leur public, les services au sein des maisons d'opéra sont amenés à manipuler quotidiennement de nombreux outils numériques, tel que des CMS (Content Management System) avec lesquels sont gérés les informations diffusées sur leur site web, des logiciels dédiés à la gestion de billetterie, à la production ou encore de multiples réseaux sociaux. La faible interopérabilité de ces systèmes, couplée à la mise en place de modélisations spécifiques au sein de chaque établissement, conduit à une complexité im-

1. Réunion des Opéras de France

portante dès lors que les données doivent être échangées ou croisées avec d'autres acteurs.

Par exemple, une requête permettant d'obtenir la programmation lyrique ou chorégraphique des maisons d'opéra pour la saison 2023/2024 reste aujourd'hui sans réponse satisfaisante. Ceci est dû au fait que les données, et surtout les modélisations des données, ne sont pas uniformisées.

Dans ce contexte, la Réunion des Opéras de France (ROF), réseau national des maisons d'opéra, scènes et compagnies lyriques, développe au sein de la mission ressources et développement numérique, le projet CapData Opéra. Porté en réponse aux enjeux et besoins identifiés auprès de ses membres et politiques culturelles, ce projet vise au développement d'une solution mutualisée et hautement répliquable afin de favoriser l'interopérabilité, l'échange de données et leur valorisation auprès des publics.

Dans cet article, nous aborderons le sujet du partage de données et de l'interopérabilité en mettant en lumière les différentes facettes du projet CapData Opéra. Nous débuterons par une présentation du projet et son contexte, en expliquant les objectifs visés, le choix de l'architecture et les solutions déployées pour répondre aux besoins détectés auprès des maisons d'opéra. Nous détaillerons ensuite les outils spécifiquement développés pour faciliter la réalisation de ce projet, en soulignant le rôle essentiel de l'ontologie dans le processus de création et de mise en œuvre. Enfin, nous conclurons en évoquant les perspectives futures du projet, dont l'importance d'une approche partagée, transversale et multi-échelles.

2 Contexte et besoins

Chaque maison d'opéra produit et diffuse sur son site web, ses réseaux sociaux et auprès de la presse et partenaires, sa programmation artistique et des ressources médias (vidéos, photographies, audios et textes) à destination des publics. En leur sein, les services, dont ceux de production ou de communication, produisent des données et des métadonnées de programmation et médias qui peuvent être similaires, mais qui sont souvent saisies à de multiples reprises et stockées dans des bases de données ne permettant que trop faiblement l'échange d'information (bases silotées et faiblement interopérables).

L'absence d'une stratégie généralisée et commune de standardisation, d'identification des données et de mise en place de référentiel au sein des maisons d'opéra et plus largement des lieux de programmation du spectacle vivant, représente un véritable frein au développement de la découvrabilité et à la diffusion des créations artistiques et des contenus auprès des publics.

Il n'existe en effet pas d'identifiant normalisé pour la gestion des productions de spectacle vivant, alors que le secteur du livre utilise l'ISBN² au niveau international et l'industrie musicale dispose de l'ISRC³ utilisé par exemple pour identifier les morceaux diffusés sur les plateformes en ligne[9]. La généralisation de l'usage d'identifiants publics,

2. Voir <https://www.isbn-international.org/>

3. Voir <https://isrc.ifpi.org/>

comme l'ISNI⁴ pour les artistes, permettrait par exemple de simplifier la diffusion des mentions et la gestion des droits lors d'une diffusion, ainsi que de développer la transparence lors de la diffusion des médias auxquels ils participent, comme l'ont illustré les expériences de diffusion des données de programmation des maisons d'opéra sur les espaces « #Culture chez nous », ou au sein de l'application du Pass Culture.

La diffusion des représentations programmées par les maisons d'opéra passe aujourd'hui par une succession de saisies manuelles : les services sont invités à saisir pour chaque réutilisation leurs offres d'événement, site web, billetteries, agendas culturels, applications. Malgré l'existence d'une API pour le Pass Culture, la faible structuration des données et interopérabilité au sein des systèmes conduit les services une nouvelle fois à une saisie majoritairement manuelle. L'enquête préliminaire a fait ressortir une moyenne de six doublons de saisie pour les données de programmations, traduisant un fort besoin, pour les équipes et pour la visibilité des contenus, de mise en place d'une solution efficace et puissante pour l'exposition des données. La faible exposition des données des maisons d'opéra sur les services innovants de diffusion musicale représente également un frein considérable à l'émergence de nouvelles expériences susceptibles de répondre aux besoins et usages du public.

Le développement de données structurées et leur exposition depuis les établissements culturels apparaissent donc comme un levier extrêmement puissant pour accroître la visibilité des contenus lyriques et chorégraphiques au sein des nouveaux modes de diffusion, dont les plateformes de *streaming*, enceintes connectées et services innovants. Ce même constat apparaît dans les « freins structurels technologiques à dépasser » de la « Mission exploratoire sur les métavers »[1] commandée par le ministère de l'Économie, des Finances et de la Relance, le ministère de la Culture et le secrétariat d'État chargé de la Transition numérique et des Communications électroniques.

Outre les problématiques d'échanges transversaux, de doublons de saisies des données, l'analyse comparative des services existants a permis d'observer trois freins supplémentaires à l'exposition et la diffusion des données culturelles :

1. Faute de donnée standardisée et exposée depuis les maisons d'opéra, une majorité de services numériques externes est contrainte de demander aux services des maisons une saisie manuelle supplémentaire ou un export spécifique non standardisé des données. Ce fonctionnement entraîne une surcharge importante de travail pour les équipes et limite le rayonnement des données culturelles.
2. Dans le cas minoritaire où le service externe propose une API pour collecter les données produites par les maisons, l'absence de standardisation des données impose des développements spécifiques, faiblement répliquables et un coût financier pour chaque structure. Ce constat fait ressortir un coût important pour les finances publiques sans bénéfices

4. Voir <https://isni.org/>

de répliquabilité et de ruissellement à l'ensemble des établissements culturels.

3. Pour combler ces problématiques d'absence de standardisation et d'exposition des données, plusieurs agrégateurs et services utilisent, la technique du *scraping* pour collecter les données de programmation. Cette technique, à l'impact environnemental négatif, n'est pas satisfaisante et implique des développements spécifiques non pérennes pour chaque établissement culturel.
4. Enfin, la protection de la souveraineté des établissements culturels sur les données qu'ils produisent apparaît comme prioritaire. Celle-ci passe par le développement de leur capacité à disposer et à exposer leurs données en toute autonomie.

La mise en place d'une solution mutualisée a pour objectif d'optimiser les coûts d'investissement et de fonctionnement, tout en favorisant son déploiement au sein des maisons d'opéra et potentiellement des établissements intéressés du secteur du spectacle vivant.

3 Projet CapData Opéra

Initié en 2022, le projet CapData Opéra est porté par la ROF en partenariat avec l'Opéra National de Bordeaux, le groupe de travail numérique de la ROF, et le réseau TMNlab. Une première expérimentation a été réalisée dans le cadre de l'appel à projets "Découvrabilité en ligne des contenus culturels francophones"[7] en 2023. Forte de cette première phase, la ROF s'est associée à 6 maisons d'opéras et au réseau TMNlab pour lancer le projet CapData Opéra - France 2030[12], qui s'est inscrit dans le Programme d'investissements d'avenir (PIA4) - "*Expériences augmentées du spectacle vivant*", une opération soutenue par l'Etat et opérée par la Caisse des Dépôts. Ce nouveau chantier vise à déployer et à industrialiser la solution à plus grande échelle. Ces deux projets font appel à plusieurs prestataires et expertises techniques afin d'assurer la mise en place de la solution de valorisation des données auprès des publics. Ils visent à proposer des solutions déployables au sein des maisons d'opéra participantes, qui peuvent déverrouiller les principaux freins à l'échange et la réutilisation des données. Chaque étape de la chaîne de circulation des données, de leur structuration à leur exposition et réutilisation, fait l'objet d'un travail spécifique et de la mise en place de solutions hautement répliquables, incluant des outils techniques, un accompagnement des partenaires et des prestataires, ainsi qu'une documentation appropriée.

L'exemple des maisons d'opéra illustre les défis rencontrés dans la gestion des données inter et intra sectorielle. Historiquement, chaque maison d'opéra a élaboré son propre système d'information, caractérisé par des schémas et des formats de données hétérogènes.

Depuis 2004, les technologies du Web Sémantique, promues par le World Wide Web Consortium (W3C)⁵, offrent une voie vers une interopérabilité accrue grâce

5. Voir <https://www.w3.org/>

aux standards de la famille RDF (Resource Description Framework)[11]. L'adoption d'un format tel que le RDF, et l'utilisation d'une ontologie, ou modèle de données commun, facilitent l'intégration des données en permettant à chaque contribution de s'aligner sur cette ontologie unifiée, assurant ainsi l'interopérabilité et la circulation des données.

Outre la facilitation de l'échange de données, cette approche soutient la souveraineté de chaque maison d'opéra sur ses propres données, en lui permettant de contrôler la manière dont elles sont partagées. Dans ce sens, le choix de l'architecture sélectionnée vise à répartir de manière équilibrée et dans le respect de la souveraineté de chaque partenaire l'enjeu de responsabilité de publication. Les maisons participantes publient elles-mêmes leurs données au format RDF, se rapportant spécifiquement à leurs besoins et stratégies de diffusion. Ces données sont ensuite collectées au sein d'un entrepôt SPARQL. La gestion mutualisée de cet entrepôt favorise l'interrogation des informations issues de l'ensemble des maisons d'opéra participantes au projet.

Cette démarche met en lumière l'importance du Web Sémantique dans le renforcement de l'interopérabilité entre systèmes d'information hétérogènes. Elle souligne également le rôle crucial de standards ouverts et partagés, comme le RDF, dans la construction d'un écosystème de données cohérent et efficace, bénéfique à l'ensemble du secteur culturel.

3.1 L'ontologie

Le développement d'une ontologie commune a pour objectif de favoriser le partage, les réutilisations et la découvrabilité des contenus culturels auprès des publics. Préalables à l'étape d'exposition des données et d'élaboration de connecteurs pour la diffusion des données depuis les systèmes d'informations respectifs des maisons participantes, les travaux ontologiques permettent également de définir le périmètre de connaissances partagées entre les maisons d'opéra et plus largement avec le secteur culturel. Sa mise en place permet de simplifier les échanges, tant au sein des services des maisons d'opéra qu'auprès des collectivités publiques et des industries culturelles et créatives (ICC).

La réalisation d'un état de l'art sur les ontologies existantes et en capacité de répondre aux besoins des maisons d'opéra et des politiques culturelles a permis de détecter deux ontologies candidates.

La première est l'ontologie schema.org⁶, qui présente de nombreux avantages. Tout d'abord, son approche englobante via l'adoption de définition large des concepts, apparaît comme très efficace pour répondre à de nombreuses situations, tout particulièrement pour le partage et la description des événements (schema.org/Event). Elle semble donc particulièrement adaptée pour la diffusion et la découvrabilité des dates de représentations. De plus sa documentation et son utilisation par les principaux moteurs de recherche pour l'indexation des contenus du web l'ont rendu très populaire auprès des équipes de communication et prestataires en charge des sites web.

6. https://schema.org

Trois écueils nous ont conduits à poursuivre la phase de recherche ontologique. Premièrement, bien que cette ontologie décrive, de manière détaillée, un événement, une organisation ou même une œuvre, il apparaît que les étapes préliminaires et nécessaires à l'élaboration d'un spectacle ne sont que partiellement décrites. La notion de production, essentielle au sein du spectacle vivant, car englobant l'ensemble des actions menant à la représentation, telles que la conception des décors et des costumes ou la gestion des distributions, sont absentes de schema.org. Deuxièmement, le concept de producteur apparaît trop large pour une description fine et essentielle du rôle et de l'implication juridique de chaque partie prenante dans l'élaboration d'une production. Troisièmement, l'absence de traduction se révèle être un frein dans la capacité de représenter finement la vision développée par les politiques culturelles et de la représentation de la diversité.

La seconde ontologie que nous avons considérée est IFLA-LRM[15], qui est une référence en termes de gestion de connaissances dans le monde de la Culture de manière générale. Les concepts représentés sont bien plus proches de ce que nous souhaitons représenter pour le projet CapData Opéra. Si la notion d'œuvre y apparaît comme centrale, l'ontologie est néanmoins élaborée pour répondre aux objectifs de la gestion des ressources bibliographiques, ce qui ne nous a pas semblé parfaitement adapté aux besoins et à la description du spectacle vivant.

Les spectacles vivants, incluant le théâtre, la danse, la musique live, et d'autres formes d'art performance, nécessitent en effet des informations et des métadonnées détaillées et intrinsèques à la description d'une production artistique ou d'un spectacle, par exemple de ses décors, costumes, montages, aux effectifs et compositions des formations, voir également des publics.

Les recherches sur l'ontologie ont également mis en lumière l'existence d'initiatives et travaux de recherches similaires à l'international, dont ceux du groupe de "Performing Arts Information Representation Community Group", néanmoins celui-ci semble inactif depuis quelques années. Cet intérêt à l'échelle internationale et visible lors de la journée "Rendez-vous France-Québec sur la découvrabilité des contenus culturels francophones" [4] de l'édition 2023 du MTL Connecte, a permis d'entamer une réflexion sur les enjeux et la pertinence d'une action coordonnée, voire mutualisée, du chantier ontologique.

Nous avons fait le choix de développer l'ontologie CapData Opéra afin de proposer une description des connaissances qui répond aux besoins détectés auprès des maisons d'opéra et plus largement du secteur du spectacle vivant. Nous l'avons voulue complémentaire des autres modèles et avons mis en place directement dans l'ontologie des alignements vers l'IFLA-LRM et le schema.org. Nous permettons ainsi une représentation fidèle aux besoins du domaine, tout en rendant aisée l'utilisation de l'ontologie de référence dans le domaine culturel et de l'ontologie de référence pour l'indexation par les moteurs de recherche sur le Web.

Pour faciliter la réutilisation de cette ontologie par les maisons d'opéra, nous la documentons et la publions

à l'URL <https://ontologie.capdataopera.fr>. Cette page permet de représenter les différentes versions de l'ontologie (actuellement la version 1.7) et la date de publication. Il y a, pour chaque version, une documentation générée par Widoco[5] et une représentation graphique générée avec WebVOWL[10]. De plus, les URI utilisées dans l'ontologie pointent sur ce site, ce qui favorise la négociation de contenu et la récupération des formats HTML ou RDF suivant la requête HTTP.

En parallèle de cette étape, les référentiels et vocabulaires contrôlés, utilisés au sein de l'ontologie, ont fait l'objet d'une exposition sur l'entrepôt SPARQL dédié au projet. Cette action participe à accroître l'interopérabilité et la découvrabilité via l'usage de définitions et de vocabulaires partagés au sein du réseau et plus largement du secteur culturel.

Enfin, des règles SHACL de vérification sont présentes sur le même site, avec une documentation adaptée⁷, pour permettre aux maisons d'opéra de valider leurs données.

L'un des avantages du processus de modélisation et des outils mis en place autour de la publication de l'ontologie est sa grande agilité. Nous avons pu, en effet, confronter la modélisation de manière concrète aux besoins des équipes des maisons participantes, prestataires, services externes et utilisateurs. Une modification de l'ontologie était rapidement intégrée dans un outil d'export de données en RDF grâce à la documentation disponible et aux outils de vérification des données. La réalisation manuelle des alignements a permis d'améliorer de manière itérative le modèle. Cette approche nous a permis de nous rendre compte très rapidement des écueils et de pouvoir corriger l'ontologie de manière agile afin d'obtenir une version stabilisée et intégrable au sein des connecteurs et des applications développées en parallèle.

3.2 Le suivi de production

Nous considérons, comme présenté en introduction de cette section, que la maison d'opéra gère la publication de ses propres données RDF. Cette publication prend la forme d'un fichier RDF disponible à une URL donnée contenant l'intégralité des données. Ce fichier est mis à jour régulièrement par un export régulier de la part des maisons d'opéra. Pour permettre l'interrogation de toutes ces données, nous souhaitons les récupérer pour les publier dans un entrepôt SPARQL dédié. Pour cela, un script s'exécute tous les jours pour récupérer ces fichiers mis à disposition derrière les URL de chaque maison. Un ensemble de traitements de nettoyage de données sont appliqués, comme l'effacement des espaces avant et après les valeurs littérales, la détection de l'utilisation d'une valeur à la place d'un identifiant de référentiels (par exemple le code pays), etc. Ces traitements sont appliqués systématiquement à chaque récolte des données. De plus, certaines données sont alignées sur les référentiels fournis par la ROF afin de faciliter l'interopérabilité. Enfin, les données sont validées en utilisant les règles SHACL liées à l'ontologie. De cette manière nous obtenons quatre graphes différents pour chaque maison d'opéra lors

7. générée grâce à <https://shacl-play.sparna.fr/play/doc>

d'une récolte de données :

- un graphe contenant les données initiales ;
- un graphe contenant les données nettoyées ;
- un graphe contenant les alignements avec les données de la ROF ;
- un graphe contenant les triplets du rapport de validation SHACL⁸.

Ces graphes sont ensuite envoyés dans un entrepôt SPARQL mutualisé pour toutes les maisons d'opéra et administré par la ROF (<https://sparql.capdataopera.fr/>).

De cette manière, toutes les données sont mises au même endroit et différents graphes nommés permettent de récupérer les données qui nous intéressent. De plus, il devient trivial de faire des requêtes entre plusieurs maisons d'opéra à partir du moment où toutes les données sont sur le même entrepôt. Le choix de cette architecture répond également à une analyse approfondie des coûts de maintenance à moyen et long terme. L'étude de faisabilité du projet a en effet mis en lumière le coût non soutenable que représentait la mise en place d'API au sein de chaque maison d'opéra pour la gestion de la diffusion et la récupération des données. Une approche non mutualisée engendrait une démultiplication des coûts financiers pour les établissements et les collectivités publiques.

Afin de mettre en place une boucle rétroactive bénéfique pour les maisons d'opéra, nous avons écrit des requêtes SPARQL pour qu'elles récupèrent leurs données, nettoyées, alignées et validées. Il leur est donc possible de réinsérer ces données dans les systèmes d'information d'origine pour augmentant la qualité de leurs données. Par exemple en ajoutant pour une personne les identifiants ROF, ISNI ou ARK issus de l'alignement.

4 Les outils

Au cours du projet, nous avons poursuivi une démarche visant à industrialiser toute la chaîne de production et nous avons constaté des manques parmi les outils disponibles sous une licence de logiciel libre. Nous avons essayé de cartographier cette situation à travers l'approche SemGraph⁹ qui, pour chaque étape de la chaîne, suggère des outils possibles. Nous avons développé ou fait évoluer certains outils quand nous l'avons jugé utile et que cela était dans nos moyens.

4.1 Publication de l'ontologie

Lorsque nous avons souhaité publier l'ontologie pour la rendre disponible sur le Web, nous avons identifié des portails comme OntoPortal[8] pour héberger nos modèles et avons envisagé de simplement les publier derrière un serveur Web standard. Néanmoins, nous avons souhaité pouvoir gérer plus finement les versions, avoir une documentation et de la négociation de contenu qui permettent d'accéder à la documentation directement, tout cela intégré à nos

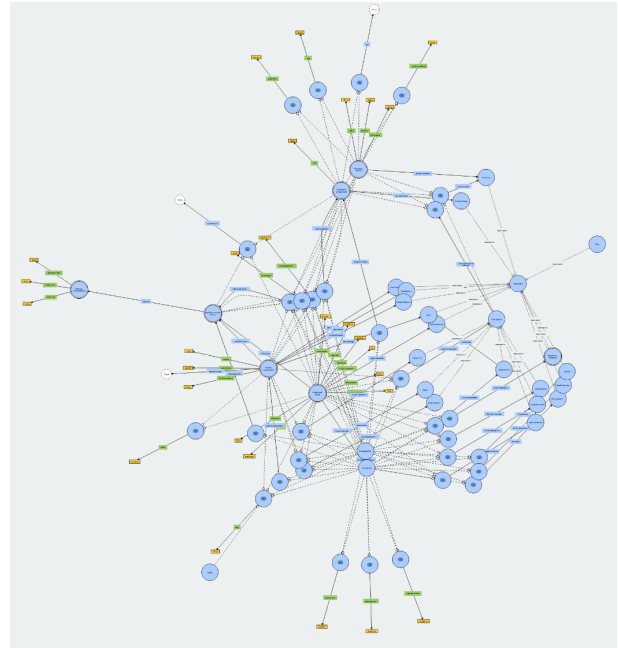


FIGURE 1 – L'ontologie CapData Opéra représentée avec WebVOWL

outils de déploiement continus habituels. Nous avons développé un script qui permet, lorsque l'on met à jour un entrepôt de gestion de version de code qui contient une ontologie, de générer la documentation, les règles SHACL, la documentation des règles SHACL et qui publie tout cela derrière un serveur Web automatiquement. Le déploiement continu permet de mettre à jour le site <https://ontologie.capdataopera.fr/>.

Le script utilise différents outils pour permettre la génération de tous ces éléments. Tout d'abord, nous utilisons Widoco[5] pour la génération de la documentation à partir des métadonnées de l'ontologie et de sa modélisation. Le rendu graphique est très lisible et proche de ce qui existe déjà dans d'autres projets, ce qui en fait une documentation simple à prendre en main. Cet outil utilise WebVOWL[10] pour avoir une représentation graphique de l'ontologie. Comme on peut le voir sur la figure 1, cette représentation permet d'avoir un aperçu global de ce qui est présent dans l'ontologie, et donc elle est particulièrement utile pour la découverte de l'ontologie, mais il est difficile d'en détecter les détails.

Nous avons utilisé la génération de documentation SHACL en utilisant l'outil proposé dans SHACL-play¹⁰. La documentation générée (comme nous pouvons l'observer sur la figure 2) permet de se rendre compte très facilement de ce qui est attendu et avoir un rapport valide lors de la validation SHACL des données.

Tous ces outils sont utilisés dans notre processus d'intégration continue proposé dans GitLab¹¹ et le résultat est dé-

8. <https://www.w3.org/TR/shacl/#validation-report>

9. Voir <https://semgraph.logilab.fr>

10. <https://shacl-play.sparna.fr/play/doc>

11. <https://docs.gitlab.com/ee/ci/>

ref:Collectivite
<https://ontologie.capdataculture.fr/v1/owli#Collectivite>

• Closed shape

Property name	URI	Expected value	Card.	Description
ref:description		xsd:string	0..*	
ref:afeurFonction		ref:Fonction	0..*	
ref:pageWeb		xsd:anyURI	0..*	
ref:openAgenda		owl:Thing	0..*	
ref:neaFereAjout		xsd:string	0..*	
ref:catalogeSourceAgence		ref:Collectivite	0..*	
ref:catalogeSourceDate		xsd:dateTime	0..*	
ref:isni		xsd:string	0..*	
ref:siret		xsd:string	0..*	
ref:catalogeSourcePays		ref:LieuGeographique	0..*	
ref:statutJuridique		ref:statutJuridique	0..*	
ref:nea		xsd:string	0..*	

FIGURE 2 – Documentation SHACL

ployé en utilisant les GitLab Pages¹². Chaque fois qu'une modification dans l'ontologie est effectuée, tout le processus est automatiquement relancé grâce à l'intégration continue et le résultat est accessible grâce au serveur Web proposé dans les GitLab Pages.

Ce processus d'intégration et de déploiement continu pour la publication d'ontologie est un réel atout qui peut être utilisé dans d'autres projets dès lors qu'une ontologie doit être maintenue.

Un besoin qui a été prégnant tout le long du projet a été de pouvoir vérifier ce qui a été exporté dans l'entrepôt SPARQL. Nous avons commencé à explorer les données exportées par l'intermédiaire d'un certain nombre de requêtes SPARQL pour voir le résultat. Cette solution a vite montré ses limites, car il n'a pas été simple d'écrire les requêtes SPARQL permettant de tout voir facilement et rapidement. Nous avons alors utilisé l'outil SparqlExplorer¹³ qui permet de parcourir l'ensemble des données d'un entrepôt SPARQL pour découvrir les données qui y sont présentes.

La figure 3 présente la page d'accueil du SparqlExplorer une fois que l'on a spécifié l'entrepôt SPARQL à explorer. Nous pouvons observer la liste des classes et le nombre d'instances associés à chacune des classes, et un champ de recherche, qui permet de chercher parmi les littéraux.

Lorsque nous cliquons sur une URI, nous affichons la vue présentée sur la figure 4. Cette vue permet de lister l'ensemble des triplets concernant cette URI et de pouvoir filtrer ces triplets (ici un filtre a été appliqué avec la valeur "nom"). De cette manière, il est possible de parcourir les triplets pour observer ce qui a vraiment été exporté et donc de s'assurer que le résultat correspond bien à ce qui est attendu.

Nous avons intégré l'outil YASGUI[13] pour interroger l'entrepôt SPARQL grâce à une interface plus pratique à utiliser que l'interface proposée par Virtuoso. Cette interface, visible sur la figure 5, comporte une option pour partager un lien vers une requête SPARQL. Ce lien a beau-

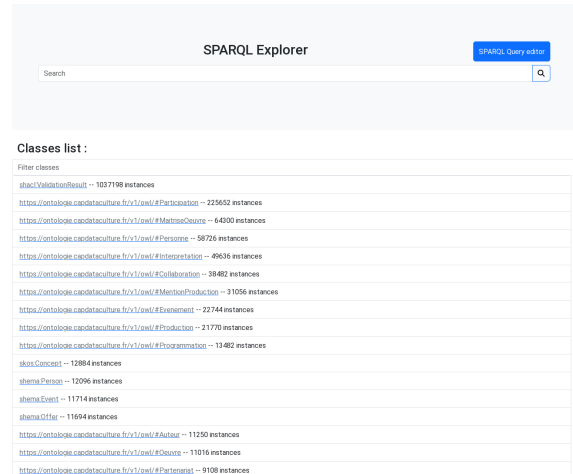


FIGURE 3 – Page d'accueil du SparqlExplorer

<http://capdataculture.fr/graph/identifieur/36135>
 Found 146 triples with http://capdataculture.fr/graph/identifieur/36135

nom	subject	predicate	object	graph
	http://capdataculture.fr/graph/identifieur/36135	https://ontologie.capdataculture.fr/v1/owli/#FeurFonction	https://opera-bordeaux.com/taxonomy/term/305	http://capdataculture.fr/graph/IMPORT
	http://capdataculture.fr/graph/identifieur/36135	https://ontologie.capdataculture.fr/v1/owli/#FeurFonction	https://opera-bordeaux.com/taxonomy/term/305	default
	http://capdataculture.fr/graph/identifieur/36135	https://ontologie.capdataculture.fr/v1/owli/#MembreCoeur	http://capdataculture.fr/graph/identifieur/3622	http://capdataculture.fr/graph/IMPORT
	http://capdataculture.fr/graph/identifieur/36135	https://ontologie.capdataculture.fr/v1/owli/#MembreCoeur	http://capdataculture.fr/graph/identifieur/3622	default
	http://capdataculture.fr/graph/identifieur/36135	https://ontologie.capdataculture.fr/v1/owli/#nom	Alvaro	http://capdataculture.fr/graph/IMPORT
	http://capdataculture.fr/graph/identifieur/36135	https://ontologie.capdataculture.fr/v1/owli/#nom	Alvaro	http://capdataculture.fr/graph/SYBRACUSSE/DFE
	http://capdataculture.fr/graph/identifieur/36135	https://ontologie.capdataculture.fr/v1/owli/#nom	Alvaro	http://capdataculture.fr/graph/SYBRACUSSE
	http://capdataculture.fr/graph/identifieur/36135	https://ontologie.capdataculture.fr/v1/owli/#nom	Alvaro	default
	http://capdataculture.fr/graph/identifieur/36135	https://ontologie.capdataculture.fr/v1/owli/#nom	Alvaro	default

FIGURE 4 – Liste de triplets dans le SparqlExplorer

12. <https://docs.gitlab.com/ee/user/project/pages/>

13. <https://sparqlexplorer.app/>

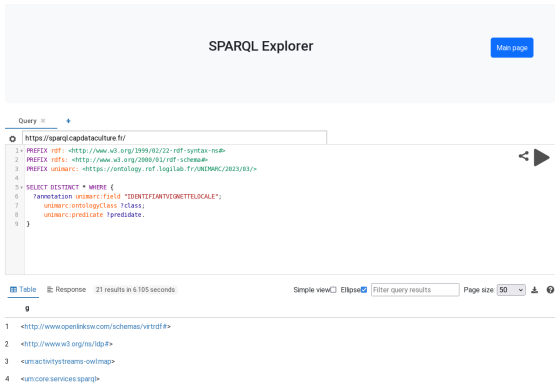


FIGURE 5 – YASGUI dans le SparqlExplorer



FIGURE 6 – Exécution des différentes étapes de récupération des données

coup été utilisé durant le projet et a grandement simplifié les échanges et les collaborations pour communiquer entre partenaires la présence au l'absence de certaines informations dans le graphe.

4.2 Assemblage et publication du graphe

Une fois que chaque maison d'opéra a publié ses données à une adresse URL de son choix, il faut les récupérer pour les agréger dans le même entrepôt SPARQL. Cette récupération s'effectue quotidiennement pour s'assurer d'avoir des données à jour dans l'entrepôt. Pour cela, nous avons mis en place une tâche récurrente avec le mécanisme d'intégration continue de notre forge logicielle¹⁴. Nous pouvons ensuite suivre l'exécution de chaque tâche et regarder le résultat obtenu à chaque étape.

La figure 6 permet de voir les différentes étapes d'une mise à jour de l'intégralité des données. Sur cet exemple, seules trois sources de données sont présentes : Les données de l'Opéra National de Bordeaux, les données de Music Story¹⁵ et les données de la Réunion des Opéras de France (ROF). Cette dernière source requiert plusieurs étapes car nous interrogeons directement une API pour récupérer les données que nous transformons ensuite en RDF. Nous sommes en train d'étudier le transfert de cette responsabilité vers l'équipe qui gère les données de la ROF.

La solution trouvée ici permet de mettre en lumière l'importance de l'adoption de la solution. Dès qu'un fournisseur de données n'adopte pas les technologies préconisées

14. GitLab <https://docs.gitlab.com/ee/ci/>

15. Pour la valorisation notamment sur les plateformes de streaming <https://music-story.com/fr/>

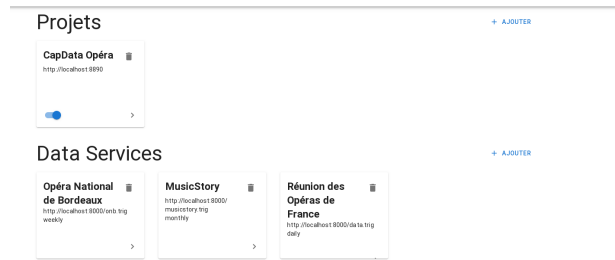


FIGURE 7 – Exemple de l'application de suivi de production

par le projet, cela demande un traitement dédié et spécifique pour cette source. Par exemple, ici, nous avons accès à une API, que nous avons utilisée pour générer le graphe RDF contenant l'intégralité des données. Ce traitement n'a pas été simple à mettre en place à cause de la complexité de l'API¹⁶. Ce travail a nécessité beaucoup d'échanges avec l'équipe en charge de la gestion des données. Il est apparu que les développements sont beaucoup plus fluides lorsque chaque gestionnaire de données gère la transformation en RDF de ses propres données.

Afin de faciliter les développements et de permettre aux fournisseurs de données d'être les plus autonomes possible, nous avons initié le développement d'une application de suivi de production qui permettra de suivre l'état de chaque récolte spécifiquement et d'avoir accès aux journaux d'import pour savoir comment il s'est déroulé. L'application enverra les données dans l'entrepôt SPARQL si la récolte s'est correctement déroulée. De cette manière, les maisons d'opéra seront autonomes dans la publication de leurs données. Elles pourront ajouter elle-même la source de données dans l'application de suivi de production et corriger les erreurs qui seront remontées dans les journaux suite à la vérification de la conformité des données en utilisant les règles SHACL. Cette application donnera une vision claire de ce qui a été importé dans l'entrepôt SPARQL. Elle constitue une étape importante dans la phase d'industrialisation du projet CapData Opéra.

Comme nous pouvons le voir dans la figure7, nous pouvons définir un projet, ici "CapData Opéra" et différentes sources ("Opéra National de Bordeaux", "Music Story" et "Réunion des Opéras de France").

La figure 8 montre un exemple de l'ajout d'une recette dans l'application de suivi de production. Cette recette permet d'identifier une source de données à importer pour le projet, l'URI du graphe dans lequel nous souhaitons envoyer les données et le processus à appliquer sur les données. Ce processus pourra être modifié dans le code pour permettre des traitements particuliers, comme transformer du CSV, ou utiliser une API, etc. Des erreurs d'import pour une source ne bloqueront pas l'import des autres sources, ce qui permet une plus grande flexibilité. Il sera alors possible d'intégrer

16. Basée sur une modélisation UNIMARC[2]

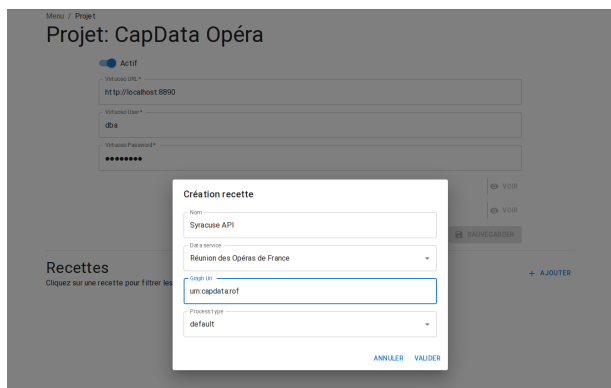


FIGURE 8 – Ajout d'une recette dans l'application de suivi de production

un plus grand nombre de maisons d'opéras plus facilement. Le graphe final est publié dans un entrepôt RDF interrogeable en SPARQL grâce au logiciel Virtuoso¹⁷ et parcouru avec un navigateur web en utilisant SparqlExplorer, comme présenté dans le chapitre précédent.

5 Responsabilités et périmètres

Nous avons présenté le processus de publication des données des différentes maisons d'opéra pour permettre l'interopérabilité entre ces données. L'architecture et les outils présentés facilitent l'autonomisation du processus, tout en assurant la pleine souveraineté des maisons d'opéra et structures participantes sur leurs données. La mise en place d'une solution mutualisée et la coordination des développements des outils développés dans le cadre du projet permet de répartir les responsabilités.

Tout d'abord dans le processus de publication des données, la mission numérique de la ROF en lien avec les besoins détectés auprès des groupes de travail dont celui des référents numériques et des échanges et des réflexions avec les partenaires, assure la maintenance de l'ontologie. Cette maintenance nécessite de considérer les besoins en modélisation des différentes maisons d'opéra, de les transcrire dans l'ontologie et de faire un suivi de versions de l'ontologie pour communiquer autour des changements. C'est pour cela que nous avons adopté pour la gestion rapide de l'ontologie un déploiement continu et documenté. De plus, une gestion de version de l'ontologie est intégrée directement dans l'URI de celle-ci, ce qui permet de pouvoir communiquer chaque fois qu'une modification importante dans l'ontologie a lieu en annonçant la publication d'une nouvelle version.

Un certain nombre de référentiels sont publiés par la ROF directement dans l'entrepôt SPARQL du projet en utilisant le vocabulaire SKOS (comme la liste des fonctions lors d'une participation d'une personne à une production). Il est important de pouvoir proposer des référentiels uniques pour tout le projet afin d'assurer une interopérabilité importante entre les données des différentes maisons d'opéra.

17. <https://virtuoso.openlinksw.com/>

Enfin, il est également nécessaire de proposer une infrastructure permettant d'héberger les outils proposés, dont principalement l'ontologie, l'application de suivi de production et l'entrepôt SPARQL. La coordination menée par le réseau ROF permet d'assurer une continuité et la stabilité des services tout en adoptant une approche mutualisée.

Au-delà de l'aspect technique, l'accompagnement et la constitution d'un socle commun de connaissances auprès des maisons participantes et des prestataires se sont révélés être une clé majeure dans la conduite et la réussite du projet. Les simples publications de l'ontologie et de l'entrepôt SPARQL seraient complètement insuffisantes pour assurer leur intégration et déploiement.

La constitution d'une documentation adaptée de l'ontologie et la communication s'avère stratégique pour le déploiement de la solution. Plusieurs présentations publiques ont eu lieu au cours de l'année 2023. Dans ce sens, nous avons souhaité développer une application de suivi de production la plus simple possible. Nous organisons aussi des réunions, ateliers et sessions de travail pour faciliter au maximum l'appropriation des concepts et des technologies que nous avons mis en œuvre ici. Les technologies du Web Sémantique restent méconnues par de nombreux prestataires du domaine culturel, la mise en place de salons de messagerie instantanée favorisant les échanges et questions, se révèlent être des leviers particulièrement puissants. Près de 5 salons dédiés à la gestion et coordination des chantiers ont été mis en place, générant plus de 10 000 interactions sur l'année 2023.

Tout en favorisant la mutualisation, le choix d'une architecture permet une grande agilité, assure le respect de la souveraineté et la pleine responsabilité des maisons, partenaires et services externes participant au projet. Ils conservent une complète autonomie dans le choix d'ouverture, de diffusion, de réutilisation et de l'enrichissement de leurs données. L'Opéra National de Bordeaux a ainsi la possibilité d'exposer ses données de programmation tout en enrichissant sa base avec des données identifiées et exposées au sein de l'entrepôt par un ou plusieurs membres des structures participantes.

6 Conclusion

Élaboré sur la base des standards du web sémantique, le projet CapData Opéra a déployé une solution mutualisée fondée sur une architecture et des outils hautement répliquables. L'expérimentation et les premiers résultats liés confirment la pertinence de l'approche choisie en vue de simplifier l'échange, la gestion et la découvrabilité des données des maisons d'opéra et des autres structures participantes.

En complète adéquation avec les politiques culturelles, l'interopérabilité des données répond à de multiples besoins détectés auprès des maisons d'opéra et enjeux du spectacle vivant, dont le développement de la découvrabilité des œuvres, des artistes et plus largement des arts lyriques et chorégraphiques auprès des publics. D'autres acteurs culturels nous semblent en effet être sur la même ligne que la

nôtre. Nous pouvons citer le monde des marionnettistes avec qui nous sommes en contact, le monde du théâtre qui porte des initiatives comme la publication des données des Registres de la Comédie Française[6], le ministère de la Culture qui conduit des ateliers dans le cadre de la deuxième génération de la feuille de route "Politique données et contenus culturels"[14] ou encore les réflexions du groupe de travail "Ouverture des données"[3] animé par le réseau du TMNlab.

La modélisation d'une ontologie et son adoption au sein de systèmes d'information hétérogènes sont des actions complexes. L'expérimentation souligne les rôles essentiels de la coordination et de l'accompagnement des établissements partenaires et prestataires. Si la question de l'héritage des logiciels et des processus humains existants est aujourd'hui bien connue, elle nécessite une attention toute particulière pour l'intégration de nouveaux modèles.

Le projet a été l'occasion d'éprouver un certain nombre d'outils du Web Sémantique et d'identifier les fonctionnalités manquantes ou les besoins pour lesquels les outils restent à concevoir. Nous avons mis en place la génération de la documentation de l'ontologie et son déploiement continu par l'intermédiaire d'un entrepôt de code, une application de suivi de production et un outil de navigation dans le graphe final. Nous prévoyons d'améliorer ces outils, mais surtout de les rendre plus génériques pour qu'ils puissent être utilisés dans d'autres projets.

L'expérimentation menée dans un premier temps avec l'Opéra National de Bordeaux est en cours d'industrialisation et de déploiement auprès de six maisons d'opéra dont l'Opéra National de Bordeaux, Théâtre du Châtelet, l'Opéra de Rennes, l'Opéra Comique, l'Opéra national Capitole Toulouse et l'Opéra de Limoges. Cette approche permet d'affiner progressivement les différents chantiers et outils : ontologie, connecteurs, applications, documentations et services dédiés à la valorisation.

Nous avons commencé à étudier la possibilité de valoriser ces données aux travers de services de valorisation dédiés, par exemple via un prestataire permettant de faire le lien avec les plateformes de streaming, ou encore un système de gestion d'agenda partagé pour la publication automatique des événements.

L'expérimentation, le constat d'un besoin présent partagé par un grand nombre d'établissements du spectacle vivant et plus largement du secteur culturel et l'émergence d'initiatives similaires à l'internationale, soulignent le besoin et la pertinence d'une démarche coordonnée des recherches et actions.

L'adoption d'une approche coordonnée met en lumière le chantier essentiel de la gouvernance. Les modèles de fonctionnement internationaux de l'IFLA ou du schema.org sont ainsi précieux d'enseignement. Les opportunités offertes par une telle approche sont nombreuses, tant pour la mutualisation des coûts financiers, le développement d'outils partagés en capacité de simplifier et d'assurer de manière pérenne leurs adoptions et déploiement, tout en répondant aux besoins et enjeux transversaux du secteur culturel.

Références

- [1] Adrien Basdevant, Camille François, and Rémi Ronfard. *Rapport de la mission sur le développement des métavers*. PhD thesis, Ministère de la Culture (France), 2022.
- [2] Permanent UNIMARC Committee et al. *Unimarc authorities format manual*. 2023.
- [3] Groupe de travail TMNlab. *Living lab - Ouverture des données*, 2023.
- [4] Mission franco-québécoise sur la découvrabilité en ligne des contenus culturels francophones. *Table ronde "Normaliser la diversité des données culturelles : est-ce possible? Rendez-vous France-Québec"*, Montréal Connecte. <https://www.youtube.com/watch?v=3HbgAUUNUiw>, 2023.
- [5] Daniel Garijo. *Widoco : a wizard for documenting ontologies*. In *The Semantic Web—ISWC 2017 : 16th International Semantic Web Conference, Vienna, Austria, October 21-25, 2017, Proceedings, Part II 16*, pages 94–102. Springer, 2017.
- [6] Charline Granger and Fabien Amarger. *Les registres de la comédie-française sur le web de données liées : de l'hétérogénéité de données vers des données quantitatives en rdf*. In *Conférence Nationale sur les Applications Pratiques de l'Intelligence Artificielle APIA@ PFIA2023*, number 2023, pages 63–71. AFIA-Association Française pour l'Intelligence Artificielle, 2023.
- [7] Direction générale des médias et des industries culturelles (DGMIC). *Découvrabilité en ligne des contenus culturels francophones*, 2022.
- [8] Clement Jonquet, John Graybeal, Syphax Bouaz-zouni, Michael Dorf, Nicola Fiore, Xeni Kechagioglou, Timothy Redmond, Ilaria Rosati, Alex Skrenchuk, Jennifer L Vendetti, et al. *Ontology repositories and semantic artefact catalogues with the ontoportal technology*. In *International Semantic Web Conference*, pages 38–58. Springer, 2023.
- [9] Julie Knibbe. *Les données dans la musique : Enjeux et stratégies d'investissement*. 2023.
- [10] Steffen Lohmann, Vincent Link, Eduard Marbach, and Stefan Negru. *Webvowl : Web-based visualization of ontologies*. In *Knowledge Engineering and Knowledge Management : EKAW 2014 Satellite Events, VISUAL, EKMI, and ARCOE-Logic, Linköping, Sweden, November 24-28, 2014. Revised Selected Papers*. 19, pages 154–158. Springer, 2015.
- [11] Frank Manola, Eric Miller, Brian McBride, et al. *Rdf primer. W3C recommendation*, 10(1-107) :6, 2004.
- [12] Eudes-Emmanuel Peyre and Groupe de travail numérique ROF. *Capdata Opéra - France 2030*. <https://www.rof.fr/rof/capdata-opera.aspx>, 2022.

- [13] Laurens Rietveld and Rinke Hoekstra. The yasgui family of sparql clients 1. *Semantic Web*, 8(3) :373–383, 2017.
- [14] Ministère de la Culture (SNUM) Service du numérique. Ateliers préliminaires à la deuxième génération de la feuille de route "Politique des données et contenus culturels", 2024.
- [15] Maja Žumer. Ifla library reference model (ifla lrm)—harmonisation of the frbr family. *KO Knowledge Organization*, 45(4) :310–318, 2018.

Session 2 : Articles déjà publiés à l'international

Intelligence Artificielle Modulaire et Hybride Orientée par les Objectifs : Application à la Simulation d'un Calculateur Booléen Plasmonique

Ouassila Labbani Narsis¹, Erik Dujardin², Christophe Nicolle¹

¹ CIAD UR 7533, Université de Bourgogne, UB, F-21000 Dijon, France

² ICB CNRS UMR 6303, Université de Bourgogne, UB, F-21000 Dijon, France

{ouassila.narsis, erik.dujardin, cnicolle}@u-bourgogne.fr

Cet article résume les travaux de recherche publiés lors d'une conférence internationale [5].

L'intelligence artificielle hybride est en plein essor, notamment les approches combinant des ontologies et des modèles d'apprentissage automatique [2]. Les modèles hybrides proposés dans la littérature promettent des systèmes performants en matière d'apprentissage à partir de données, tout en possédant également des connaissances structurées et des capacités d'inférence leur permettant de prendre des décisions éclairées. De manière globale et simplifiée, les modèles d'hybridation peuvent être regroupés en trois grandes catégories : *Learning and Reasoning System*, *Semantic Data Mining* et *Learning-Enhanced Ontology* [4].

La catégorie **Learning and Reasoning System** regroupe les travaux combinant l'apprentissage automatique et les ontologies pour résoudre des problèmes complexes et effectuer des tâches spécifiques dans le même domaine. Ces systèmes se divisent en deux principales sous-catégories : *Expert System Embedded Learning* et *Hybrid Application*. La sous-catégorie *Expert System Embedded Learning* utilise l'apprentissage automatique comme un sous-programme pour analyser et apprendre à partir des données en vue d'améliorer le fonctionnement global du système, tandis que la sous-catégorie *Hybrid Application* propose des approches qui relient l'apprentissage automatique aux connaissances ontologiques pour classifier les données, peupler l'ontologie ou raisonner sur les données pour analyser, valider et corriger les résultats d'apprentissage. D'autres approches transforment les entités ontologiques en données manipulables par les algorithmes d'apprentissage automatique et utilisent les relations sémantiques de l'ontologie pour former des réseaux neuronaux experts.

Les études de la catégorie **Semantic Data Mining** visent à intégrer des connaissances du domaine dans le processus d'apprentissage automatique afin d'améliorer son efficacité. Elles se divisent en deux principales sous-catégories : *Informed Machine Learning* et *Ontologies Explain Black-Box*. La sous-catégorie *Informed Machine Learning* intègre les connaissances préalables dans différentes phases de l'apprentissage automatique, telles que la sélection, l'extraction ou l'augmentation des caractéristiques, ainsi que dans

le choix de la structure du modèle. Les connaissances préalables sont représentées dans une ontologie et utilisées pour enrichir le processus d'apprentissage. La sous-catégorie *Ontologies Explain Black-Box* a pour objectif d'ajouter a posteriori de l'explicabilité aux modèles d'apprentissage en utilisant les connaissances d'une ontologie. L'explication peut être sur la prise de décision globale par le modèle ou localement pour chaque individu.

La catégorie **Learning-Enhanced Ontology** regroupe les travaux visant à améliorer l'utilisation des ontologies grâce à l'apprentissage automatique. Elle se divise en trois sous-catégories principales. La première est l'*Ontology Learning* où la création et la maintenance des ontologies sont (en partie) automatisées grâce à l'apprentissage machine. La seconde sous-catégorie, *Ontology Mapping*, cherche à améliorer l'alignement des ontologies grâce à l'apprentissage automatique, en garantissant une interopérabilité entre les deux modèles. La troisième sous-catégorie *Learning-based Reasoning* regroupe les approches visant à faciliter le raisonnement déductif d'une ontologie grâce à l'apprentissage automatique.

Cette étude des travaux existants met en évidence l'importance et la nature évolutive des approches d'hybridation en intelligence artificielle intégrant l'apprentissage automatique et le raisonnement ontologique. Malgré les avancées dans ce domaine, cette combinaison soulève le défi de leur intégration efficace due à la rigidité des modèles existants [2, 6]. Ces derniers, souvent basés sur un pipeline séquentiel et une ontologie globale, peinent à s'adapter à des contextes variés et à évoluer selon les besoins [1]. Il est souvent difficile de modifier ces approches existantes pour tenir compte des nouveaux besoins ou des changements dans les données et les connaissances.

Pour répondre à ces limites, nous proposons une approche modulaire et hybride visant à améliorer l'adaptabilité et la flexibilité par l'intégration de multiples modèles d'apprentissage et d'ontologies selon les spécificités des tâches. Cette stratégie a pour but de créer des modèles d'intelligence artificielle hybrides centrés sur les objectifs plutôt que sur l'ensemble du système. Elle vise également à faciliter la maintenance et l'évolution des différentes tâches de

l'application de manière indépendante, tout en simplifiant le choix des technologies d'intelligence artificielle adaptées à chaque module. Ce concept est illustré dans le cadre de la conception d'une unité arithmétique et logique plasmonique complexe et reconfigurable (projet [ANR DALHAI](#)). Il est appliqué dans le développement d'une architecture hybride qui visant à découvrir les meilleures configurations géométriques (formes), ainsi que les paramètres physiques d'excitation (propagation lumineuse), afin de détecter les portes logiques appropriées. Il s'agit dans ce cas d'un problème d'optimisation multi-objectifs qui implique une modélisation précise et une évaluation attentive des solutions proposées [3]. La résolution de ce type de problèmes implique l'utilisation de méthodes d'optimisation et de techniques d'apprentissage automatique pour trouver des solutions possibles et de les évaluer en fonction de critères de performance spécifiques et de contraintes du domaine. Notre architecture modulaire et hybride est composée de modèles d'apprentissage automatique, pour la génération de solutions optimales, et de modèles de raisonnement ontologique pour vérifier si la solution sélectionnée est réalisable selon les connaissances et les contraintes physiques définies par les experts du domaine. Dans ce cas, la partie d'apprentissage automatique peut être vue comme un *générateur* qui propose des solutions, tandis que la partie ontologique peut être assimilée à un *discriminateur* qui évalue la faisabilité de la solution et, si nécessaire, propose des ajustements au modèle d'apprentissage. Nous avons choisi d'utiliser un algorithme génétique évolutionnaire [7], adapté aux problèmes d'optimisation multi-objectifs, pour optimiser la forme et les paramètres d'excitation. Selon les portes logiques souhaitées, l'algorithme génère une forme sur laquelle sera appliquée une simulation numérique de la propagation du champ laser sur la structure en utilisant l'outil *PyGDM*. Cet algorithme inclue un deuxième algorithme évolutionnaire dont le rôle est d'optimiser les paramètres d'excitation (la position, la polarisation et la phase du laser) pour une forme particulière. Chaque solution fournie par l'algorithme génétique doit respecter les contraintes de la physique définies par les experts du domaine afin d'assurer la reproduction des résultats en expérimentation. Nous avons identifié trois principaux types de connaissances à mettre en place pour la définition de : (1) la forme, (2) les paramètres d'excitation, et (3) les paramètres de détection des portes logiques. Ceci nous a conduit à la définition de trois ontologies pouvant être développées et évoluées séparément. La première, *Ontologie de forme*, rassemble les connaissances et les contraintes liées à la description de la forme selon les contraintes imposées par la physique. La deuxième ontologie, *Ontologie de paramètres d'excitation*, concerne les paramètres d'excitation (points d'entrée, paramètres de laser, points de sorties, etc.). Elle correspond aux connaissances physiques liées à la faisabilité réelle de l'expérimentation, comme l'emplacement des deux lasers qui doivent respecter une certaine distance l'un par rapport à l'autre, ou la différence entre deux polarisations permettant d'encoder les états binaires en entrée, sans ambiguïté. La troisième ontologie, *Ontologie de porte logique*, est uti-

lisée pour capturer les connaissances sur les portes logiques (leur table de vérité) et leur assemblage (dans le cas où la fonction logique requiert plusieurs valeurs binaires simultanées en sortie). Son objectif est de vérifier l'assemblage correct entre les points d'entrée et de sortie afin de s'assurer que le résultat obtenu correspond bien aux portes logiques souhaitées. Le résultat de raisonnement des trois ontologies peut être également utilisé pour apporter des ajustements aux paramètres de modèle d'apprentissage, permettant ainsi d'avoir un modèle de raisonnement hybride performant, efficace et plus rapide, favorisant l'utilisation des techniques d'intelligence artificielle les mieux adaptées sur un sous-ensemble de données et de connaissances spécifiques à chacune de ses étapes de fonctionnement. Les résultats de l'application de notre approche modulaire et hybride dans le cadre de ce projet sont accessibles à [ce lien](#).

Remerciements

Les auteurs souhaitent remercier le projet DALHAI (Design of plasmonic ALU by Hybrid Artificial Intelligence), financé par l'Agence Nationale de la Recherche (ANR-20-CE240001 DALHAI).

Références

- [1] Fernando Almaguer-Angeles, John Murphy, Liam Murphy, and A Omar Portillo-Dominguez. Choosing machine learning algorithms for anomaly detection in smart building iot scenarios. In *2019 IEEE 5th World Forum on Internet of Things (WF-IoT)*, pages 491–495. IEEE, 2019.
- [2] Anna Breit, Laura Waltersdorfer, Fajar J Ekaputra, Marta Sabou, Andreas Ekelhart, Andreea Iana, Heiko Paulheim, Jan Portisch, Artem Revenko, Annette ten Teije, et al. Combining machine learning and semantic web : A systematic mapping study. *ACM Computing Surveys*, 2023.
- [3] Yann Collette and Patrick Siarry. *Optimisation multiobjectif : Algorithmes*. Editions Eyrolles, 2011.
- [4] Sarah Ghidalia, Ouassila Labbani Narsis, Aurélie Bertaux, and Christophe Nicolle. Combining machine learning and ontology : A systematic literature review. *arXiv preprint arXiv :2401.07744*, 2024.
- [5] Ouassila Labbani Narsis, Erik Dujardin, and Christophe Nicolle. Objective-driven modular and hybrid approach combining machine learning and ontology. In *2023 15th International Congress on Advanced Applied Informatics Winter (IIAI-AAI-Winter)*, pages 300–305. IEEE, 2023. <https://hal.science/hal-04573042>.
- [6] Yuan Ren, Jeff Z Pan, and Yuting Zhao. Towards scalable reasoning on ontology streams via syntactic approximation. *Proc. of IWOD*, 2010.
- [7] Ryoji Tanabe and Hisao Ishibuchi. A review of evolutionary multimodal multiobjective optimization. *IEEE Transactions on Evolutionary Computation*, 24(1) :193–200, 2019.

Détection de défauts sur radiographies de pièces composites rotor grâce à l'IA

N. Griselin¹, D. Dubois², P. Barbier¹

¹ AIRBUS Helicopters

² ALTEN

nicolas.griselin@airbus.com

Résumé

Lors du contrôle qualité des pièces composites rotor, chaque pièce est systématiquement inspectée au moyen d'une méthode de contrôle non-destructive, basée sur l'analyse visuelle de clichés radiographiques prises sur ces pièces. Dans cet article, une approche automatique de détection des défauts grâce à l'utilisation de l'IA est utilisée. L'application au cas industriel est présentée avec une précision pour prédire des défauts d'environ 85%, un Recall de 96% et un F1-score de 90%. Cet article sera présenté lors du prochain ERF 2024 (<https://hal.science/hal-04588729>).

Mots-clés

Détection, défaut, radiographie, composite, IA, segmentation, Mask-RCNN.

Abstract

During the quality control of rotor composite parts, each part is systematically inspected using a non-destructive control method, based on visual analysis of X-ray shots taken on the parts. In this article, an automatic approach for defects detection using AI is used. Application to the industrial case is presented with precision to predict defects of about 85%, a recall of 96% and a F1-score 90%. This article will be presented during the next ERF 2024 (<https://hal.science/hal-04588729>).

Keywords

Detection, defect, x-ray, composite, AI, segmentation, Mask-RCNN.

1 Introduction

L'état de l'art et les progrès les plus récents dans le domaine de l'IA appliqué à la détection des défauts ont été décrit [1], offrant une compréhension complète de l'utilisation des techniques d'apprentissage pour la classification, la détection et/ou la segmentation des défauts.

La classification est largement utilisée dans diverses applications telles que l'étiquetage d'images, la récupération d'images basée sur le contenu et la catégorisation d'images. Il sert de base à de nombreux systèmes de vision par ordinateur et a été utilisé dans notre travail précédent [2] où nous avons décrit une approche en trois étapes pour détecter les défauts dans les images de radiographie de pièces composites à l'aide de modèles simples d'intelligence artificielle (CNN) pour la classification des images.

Toutefois, l'utilisation d'une telle approche est limitée et ne permet pas de définir le nombre de défauts ainsi que la localisation et la caractérisation précise de chaque défaut.

Les techniques de segmentation dans le traitement de l'image et la vision par ordinateur, par contre, permettent de s'affranchir de ces limitations. Ces méthodes peuvent jouer un rôle important dans les tâches nécessitant une localisation et une segmentation détaillées des objets, comme l'analyse d'images médicales et le comptage des objets. La segmentation par instance et la segmentation sémantique sont deux techniques de segmentation importantes utilisées dans la vision par ordinateur pour l'analyse et la compréhension de l'image. Bien que la segmentation sémantique et la segmentation par instance fonctionnent au niveau du pixel pour diviser les images en régions significatives, elles diffèrent dans leur niveau de granularité. La segmentation sémantique se concentre sur l'attribution d'étiquettes de classe à des pixels basés sur des catégories sémantiques, tandis que la segmentation d'instance va plus loin en fournissant des masques de segmentation séparés pour chaque instance objet dans la même classe. Les deux techniques sont des outils précieux dans la vision par ordinateur et ont leurs applications respectives en fonction des exigences spécifiques de la tâche à accomplir. Les figures 1 et 2 illustrent la différence entre la segmentation sémantique et la segmentation par instance appliquée à notre étude.

Bien que la segmentation sémantique puisse améliorer la localisation des défauts sur clichés radiographiques sans faire de distinction entre les différents cas d'objets appartenant à la même classe (voir figure 1), elle ne permet pas de compter le nombre de défauts et de caractériser chaque défaut. D'autre part, la segmentation par instances (voir figure 2) permet de classer chaque pixel en catégories sémantiques et de distinguer les instances d'objets individuelles de la même classe. Cela nous donne la possibilité d'identifier et de différencier chaque instance objet présente dans l'image et ainsi caractériser chaque défaut en fonction de leur longueur et de leur largeur. Les progrès récents de l'apprentissage profond, notamment avec l'avènement de réseaux neuronaux convolutifs (CNN) et d'architectures comme R-CNN, ont grandement amélioré la précision et l'efficacité des modèles de segmentation [3].

Dans ce travail, nous avons mis en place une approche de segmentation par instance en utilisant un modèle Mask R-CNN pour détecter les défauts mais aussi pour pouvoir

localiser et caractériser chaque défaut dans la radiographie.

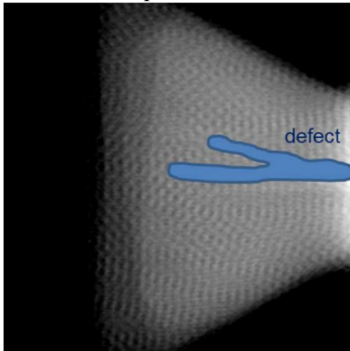


Fig. 1 : Segmentation sémantique

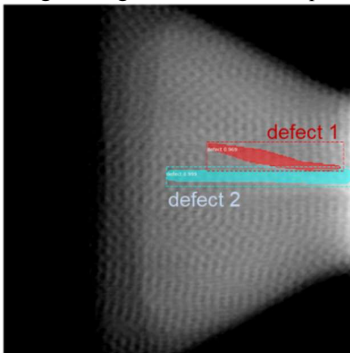


Fig. 2: Segmentation par instance

2 Méthodologie

Dans cette étude, nous avons utilisé la même base de données que dans notre étude précédente [2]. Elle se compose d'images 16 bits, en niveau de gris, avec une taille de 1024 x 1024 pixels, contenant un ou plusieurs défauts de tailles et localisations variables. Ces images ont été extraites du fichier DICOM d'origine collecté lors de la prise des clichés radiographiques de la pièce composite et converties en format PNG pour réduire la taille du fichier image sans perte de qualité. Il y a 345 images avec un ou plusieurs défauts et 200 images sans aucun défaut. Après extraction de l'image au format PNG, on applique une méthode de prétraitement de l'image pour enlever les bordures inutiles, normaliser le niveau de gris, inverser l'échelle d'intensité et filtrer la valeur des pixels de l'image (fenêtrage) afin de pouvoir visualiser clairement les défauts. Nous avons ensuite effectué la labélisation de tous les défauts de notre ensemble de données : chaque défaut d'une image est localisé et labélisé comme étant un « défaut » afin de distinguer les différents objets de la même classe. A partir de la localisation de chaque défaut, nous avons créé un masque de segmentation au niveau des pixels concernés. Ce masque indique quels pixels de l'image appartiennent à cette instance particulière. En attribuant une labélisation unique et en générant des masques de segmentation, les algorithmes de segmentation par instance peuvent délimiter et séparer avec précision les objets individuels dans la même classe. Un deuxième traitement a été effectué sur toutes les images (et tous les masques) consistant en une rotation d'image pour avoir tous les défauts dans la même direction, suivie d'un recadrage d'image (réduction de la

taille de l'image à 300 x 300 pixels) pour optimiser davantage la précision du modèle.

Pour la formation du modèle, nous n'avons considéré que les images avec des défauts. L'ensemble de données résultant (345 images présentant des défauts) a été divisé en trois groupes : un ensemble de données d'entraînement (60 %), un ensemble de données de validation (20 %) et un ensemble de données de test (20 %). Nous avons ensuite utilisé l'ensemble des données d'entraînement et de validation pour entraîner un modèle Mask R-CNN, avant de l'appliquer à notre ensemble de données de test.

3 Résultats

L'application de notre modèle Mask R-CNN sur notre ensemble de données test a fourni des bons résultats. Le modèle a été appliqué sur les 69 images de l'ensemble de données test. Une même image pouvant comporter plusieurs défauts, 132 défauts différents ont ainsi été considérés pour les données test. Les résultats sur ces données sont présentés dans la matrice de confusion ci-dessous (voir table 1).

Table 1. Matrice de confusion pour données test

		Prédiction	
		Sans défauts	Avec défauts
Référence	Sans défauts		FP = 23
	Avec défauts	FN = 6	VP = 126

FP = Faux Positif, FN = Faux Négatif, VP = Vrai Positif.

Lors du calcul des critères de performance, nous obtenons une précision pour la détection des défauts d'environ 85%, un *Recall* de 96% et un *F1-score* d'environ 90%.

4 Travaux en cours

Nous travaillons actuellement sur l'optimisation du modèle afin d'améliorer ses performances. Les développements ultérieurs se concentreront également sur l'explicabilité du modèle ainsi que sur son intégration au sein du processus qualifié en suivant les préconisations émises par l'EASA [4]. D'autres défauts seront également inclus dans l'algorithme de détection dans les travaux futurs. Cet article sera présenté lors du prochain ERF 2024.

5 Références

- [1] Z. Ren, F.Z. Fang, N. Yan et Y. Wu, State of the Art in Defect Detection Based on Machine Vision, *Vertical International Journal of Precision Engineering and Manufacturing-Green Technology*, 2022.
- [2] N. Griselin et P. Barbier, Defects Detection in Rotor Composite Parts using AI, *Vertical Flight Society 79th Annual Forum Proceedings*, 2023.
- [3] Z. Ren, F.Z. Fang, N. Yan et Y. Wu, Deep CNN-based visual defect detection: Survey of current literature, *Computers in Industry* 148, 2023.
- [4] EASA Concept Paper: First usable guidance for Level 1&2 machine learning applications, *easa.europa.eu/ai*, 2023.

Session 3 : Articles prospectifs

Dépasser l'observation mono-mission de la Terre : utiliser le paradigme multi-agents pour fédérer de multiples missions

Jean-Loup Farges¹, Filippo Perotto¹, Gauthier Picard¹, Cédric Pralet¹,
Cyrille de Lussy², Jonathan Guerra², Philippe Pavero², Fabrice Planchou²

¹ DTIS, ONERA, Université de Toulouse, France
first.last@onera.fr

² Airbus Defence and Space, Toulouse, France
first.last@airbus.com

Résumé

Nous envisageons ici une couche de fédération multi-agents pour coordonner des systèmes composés de multiples missions d'observation de la Terre indépendantes. L'objectif de cette fédération est de permettre aux clients qui demandent des acquisitions de grandes zones d'accéder facilement à plusieurs constellations de satellites et de sites de communication pour composer et télécharger leurs acquisitions, en un temps réduit par rapport aux demandes conventionnelles non coordonnées. Nous identifions plusieurs pistes et défis scientifiques liés aux approches multi-agents, telles que la coordination, la planification et l'apprentissage, pour mettre en œuvre deux fonctionnalités clés de la fédération : (i) la détermination de la faisabilité et la répartition de la charge sur des missions multiples pour la couverture de grandes zones, et (ii) la réservation de stations de communication.

Mots-clés

Satellites d'observation de la Terre, fédération de missions, couverture de grandes zones, réservation de stations de communications.

Abstract

We envision a multi-agent federation layer to coordinate systems composed of independent Earth observation missions. The goal of this federation is to allow clients requesting acquisitions of large areas to easily access several constellations of satellites and communication sites to compose and download their acquisitions, in a reduced time compared to conventional uncoordinated requests. We identify several scientific tracks and challenges related to agent-based approaches such as coordination, planning and learning, to implement two key federation functions : (i) multi-mission coverage feasibility and dispatching and (ii) communication site booking.

Keywords

Earth observation satellites, mission federation, large area coverage, communication booking.

1 Introduction

Le marché de l'*observation de la Terre*, en Anglais *Earth Observation* (EO), exige des systèmes satellitaires plus rapides et plus efficaces pour répondre aux besoins croissants de réactivité, de revisite et de capacités multi-missions¹ [10]. Les applications telles que la surveillance des frontières, la surveillance maritime et les interventions en cas de catastrophes nécessitent une programmation rapide des satellites afin de fournir des images en temps voulu. En outre, les utilisateurs recherchent des solutions évolutives et rentables pour soutenir leurs efforts en matière de durabilité environnementale et agricole. Pour répondre à ces exigences, les nouveaux systèmes EO sont équipés d'instruments et de satellites avancés, permettant des acquisitions haute résolution, de grandes capacités de stockage et d'acquisition rapide d'images sur plusieurs sites ; ce qui induit une amélioration significative de la performance.

Cependant, les utilisateurs et les opérateurs de satellites exigent des systèmes plus réactifs et plus souples. Les systèmes EO traditionnels, souvent constitués d'un ou deux satellites seulement, ne disposent pas de capacités de revisite et de réactivité suffisantes pour fournir des images dans les délais. Les utilisateurs ont besoin d'accéder à de multiples moyens d'acquisition, y compris à d'autres satellites, instruments, et même à des missions non dédiées à l'EO, pour répondre efficacement à leurs besoins. Dans cet article, nous nous concentrons particulièrement sur la chaîne amont du segment sol, qui comprend le contrôle du satellite, la programmation de la mission et la validation des images, qui sont spécifiques aux missions et étroitement intégrés aux systèmes satellitaires. Cette approche cloisonnée (voir Figure 1) est optimisée pour les missions individuelles et devient problématique lorsqu'il s'agit de gérer plusieurs systèmes d'observation de la Terre. Le manque d'interopérabilité entre ces systèmes empêche une utilisation efficace de leur capacité globale.

Pour répondre au besoin pressant d'une solution unifiée per-

1. Une *mission* consiste généralement en un seul satellite ou une constellation de satellites gérés par un seul opérateur et un ensemble de moyens de communication pour les liaisons de données montante et descendante.

Dépasser l'observation mono-mission de la Terre : utiliser le paradigme multi-agents pour fédérer de multiples missions

mettant un accès transparent aux différents systèmes d'information, le projet DOMINO-E [9] vise à développer une couche de fédération (voir Figure 2) qui permet aux opérateurs de missions EO de coordonner et de hiérarchiser leurs besoins en imagerie à travers différents systèmes, en utilisant le paradigme multi-agents. Alors que les systèmes multi-agents ont déjà été identifiés comme une approche pertinente pour les constellations EO [18], cet article propose une nouvelle vision pour orchestrer des constellations multiples. La conception modulaire, flexible et résiliente de l'architecture proposée devrait permettre l'intégration de services supplémentaires pertinents sans compromettre les performances du système.

Nous nous intéressons en particulier aux défis liés à deux services clés pour les opérations multi-missions :

- *Faisabilité de la couverture et répartition sur de multiples missions* (Section 2) : en cas de couverture d'une surface importante à effectuer dans une période donnée, une seule mission ne suffit souvent pas pour accomplir la tâche dans les délais et avec la qualité requise. Nous proposons donc un système multi-agents qui permet une sélection dynamique entre plusieurs systèmes de satellites pour la couverture.
- *Réservation de communication multi-mission* (Section 3) : Les systèmes terrestres s'appuient actuellement sur des antennes dédiées pour le contrôle des satellites, le transfert de données et le téléchargement d'images. Le choix des antennes est généralement fixe pendant toute la durée de vie du système. Nous proposons un système dynamique de réservation d'antennes, qui s'appuie sur le segment terrestre en tant que service (*Ground Segment as a Service*, ou GSaaS), afin de garantir une capacité adéquate de liaison montante de plans et descendante de données et d'optimiser les coûts.

Les solutions centralisées risquent de ne pas fonctionner pour mettre en œuvre ces services complexes, car les informations clés sont dispersées. Nous devons donc mettre en place une coopération entre les composants et les parties prenantes, en suggérant que les techniques de planification multi-agents, d'allocation des ressources, d'enchères et d'apprentissage seront cruciales pour son développement.

2 Fédération de l'observation

Cette section aborde le problème de décision suivant : "Comment diviser une grande zone et attribuer ces subdivisions à différentes missions afin de minimiser le temps d'acquisition et de maximiser la qualité des images, même si nous ne connaissons pas la charge de travail future des missions ?" Cette décision doit être prise à chaque fois qu'une nouvelle demande est reçue ou qu'une mission met à jour ses dates d'acquisition.

2.1 Composition du système

Le problème étudié concerne un ensemble de K utilisateurs finaux, appelés *clients*, qui demandent des images sur de

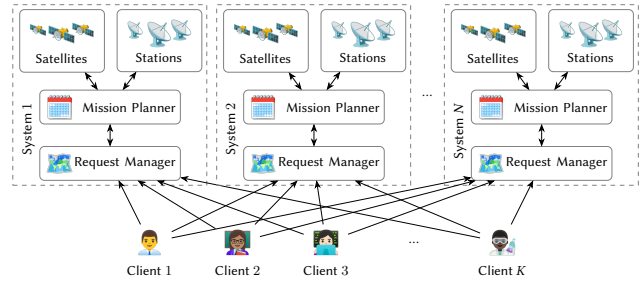


FIGURE 1 – L'architecture conventionnelle est verticale et les utilisateurs doivent interroger et envoyer des demandes à chaque système séparément.

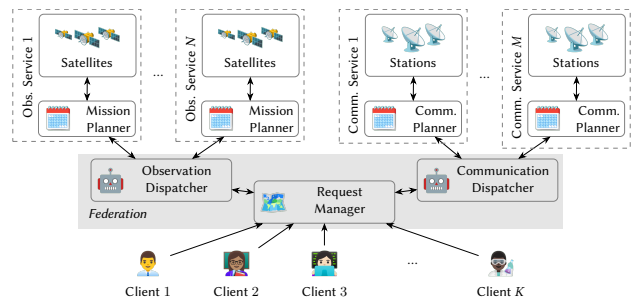


FIGURE 2 – L'architecture DOMINO-E appelle différents systèmes en tant que services, optimisant les demandes combinées de plusieurs clients vers plusieurs systèmes de manière transparente.

grandes zones de la surface de la Terre. Il implique également un ensemble de N agents d'observation indépendants, appelés *missions*, développés au fil des ans pour répondre à des besoins d'observation spécifiques. Fondamentalement, chaque mission dispose d'un ensemble de satellites en orbite basse capables de collecter des images sur des mailles élémentaires, où chaque maille couvre généralement quelques dizaines de kilomètres carrés, et chaque mission dispose de son propre *planificateur* pour déterminer les activités de ses satellites en fonction des demandes d'observation qu'elle reçoit.

L'objectif global est alors de réaliser les observations demandées par les K clients en utilisant les N missions d'observation disponibles. Pour ce faire, l'approche conventionnelle consiste à ce que chaque client envoie ses demandes d'observation directement à la mission de son choix (voir Figure 1). Pour aller au-delà de ce paradigme, la *couche de fédération* recevra les demandes d'observation des clients, enverra des tâches d'observation aux missions, récupérera les données d'observation et livrera les images aux clients (voir Figure 2). Ainsi, chaque client bénéficie d'un accès transparent à de nombreuses ressources satellitaires sans avoir à se soucier des détails. En outre, la couche de fédération peut exploiter au mieux les ressources disponibles pour accélérer considérablement l'observation d'une grande zone en la divisant en plusieurs sous-zones qui peuvent être observées en parallèle par des missions distinctes, et fusionner les demandes d'observation sur des

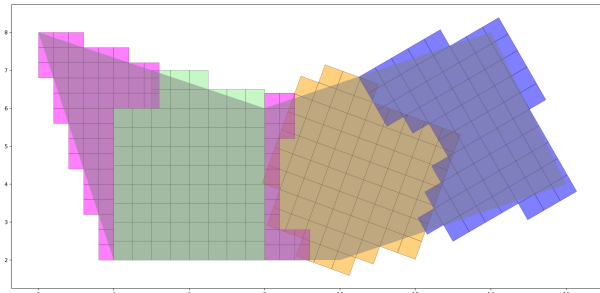


FIGURE 3 – Couverture multi-missions d’une grande zone (en gris) à l’aide de mailles de quatre missions différentes (une couleur par mission).

zones qui se chevauchent pour réduire la charge globale.

2.2 Décisions et problèmes multi-agents

Pour définir la couche de fédération, une première étape consiste à définir comment les différents agents interagissent. Une première approche consisterait à concevoir une architecture mettant en œuvre une *négociation multi-agents* [12], où la couche de fédération enverrait des requêtes aux missions tandis que ces dernières accepteraient ou non de faire des observations sur certaines sous-zones. Ensuite, chaque sous-zone serait effectivement attribuée à une mission spécifique. Cependant, un tel schéma de négociation n’est pas totalement compatible avec la présence de clients qui attendent des images, en particulier si le temps de réponse d’une mission est élevé pendant la phase de négociation. En effet, l’évaluation de l’impact de l’inclusion d’une nouvelle observation dans le programme d’une mission est un calcul coûteux, ou peut nécessiter un modèle de substitution efficace [22].

Pour limiter le nombre d’interactions entre les agents, une alternative pourrait consister à mettre en œuvre des *enchères combinatoires* [7], comprenant une première phase où la couche fédération demande aux missions de faire des offres sur les sous-zones d’observation possibles, et une seconde phase où les tâches d’observation sont réparties en fonction des offres reçues. Mais une telle approche nécessiterait le développement d’un système d’enchères pour chaque mission, ce qui peut s’avérer délicat et coûteux.

C’est pourquoi nous envisageons une architecture dans laquelle la couche de fédération saute l’étape de l’appel d’offres et attribue directement les tâches d’observation sur la base de sa connaissance actuelle de la capacité et de la charge de chaque mission. Dans ce cas, la communication avec les missions individuelles est minimale, mais la définition de la couche de fédération envisagée soulève plusieurs défis détaillés ci-après.

2.3 Défis pour la couche de fédération

2.3.1 Problèmes fortement combinatoires

Pour prendre des décisions de répartition, la couche de fédération doit s’attaquer à un problème hautement combinatoire (des centaines ou des milliers de demandes, des

dizaines ou des centaines de satellites, de nombreuses façons de diviser la zone de chaque demande en un ensemble de sous-zones, etc.). Sur ce point, il est nécessaire d’étudier à la fois les stratégies de répartition à gros grain où de grandes sous-zones sont allouées aux missions, et les stratégies de répartition à grain fin où des maillages détaillés sont pris en compte. Le raisonnement au niveau des mailles peut conduire à de meilleures décisions de répartition, mais augmente la difficulté du problème d’optimisation à résoudre. Un exemple de stratégie de couverture multi-mission possible définie au niveau de la maille est illustré dans la Figure 3. La couche de fédération doit également gérer des objectifs multiples, notamment la minimisation du temps de réalisation de la couverture et la minimisation de la charge totale du système.

Dans une autre direction, la couche de fédération doit prendre en compte la compatibilité entre les exigences associées à chaque demande et les capacités des agents disponibles, puisque les satellites de deux missions distinctes peuvent être équipés d’instruments ayant des résolutions ou des caractéristiques spectrales différentes.

2.3.2 Besoin d’un modèle des autres agents

Pour attribuer les tâches d’observation, la couche de fédération doit gérer un modèle des capacités actuelles de chaque mission. L’une des difficultés réside dans le fait qu’un tel modèle peut être imprécis, en particulier pour les *missions externes* dont les ressources ne sont pas sous le contrôle direct du gestionnaire du système. Et même pour les *missions propriétaires*, la couche de fédération n’a pas un contrôle total en raison de la présence d’un moteur de planification spécifique associé à chaque mission. C’est pourquoi la couche de fédération devrait idéalement *apprendre un modèle de haut niveau des capacités* de chaque mission. En outre, chaque mission peut recevoir directement des demandes urgentes de haute priorité qui peuvent retarder la réalisation d’observations de faible priorité sur de vastes zones, et pour cela la couche de fédération devrait envisager d’exploiter un modèle de la densité des demandes de haute priorité dans une zone donnée, basé par exemple sur des données de demandes historiques.

2.3.3 Gestion des incertitudes

Dans le problème considéré, il existe plusieurs sources d’incertitude, notamment en ce qui concerne la présence de nuages, qui entraîne des images ratées – par exemple jusqu’à 50% des images acquises par les satellites opérationnels peuvent être ratées en raison d’une couverture nuageuse excessive [11]. Sur ce point, un défi est de pouvoir exploiter à la fois les prévisions météorologiques à court terme pour préférer des solutions de répartition où une sous-zone donnée est attribuée à une mission disposant d’un satellite qui survole cette zone pendant une période ensoleillée, et les données météorologiques historiques pour identifier les régions dont les opportunités d’observation ne doivent pas être manquées. Pour ce faire, l’une des ambitions est d’apprendre un modèle de la récompense à long terme fournie par une décision de répartition. Un autre mécanisme clé consiste à construire une couche de fédération

qui effectue de manière itérative la redistribution des tâches, afin de mettre automatiquement à jour la stratégie de couverture en fonction de l'état réel de l'exécution.

2.4 Travaux connexes

Dans la littérature du domaine, plusieurs contributions traitent de l'observation de grandes zones [17, 15, 14, 2], mais il y a un fossé entre ces contributions et la couche de fédération recherchée – par exemple, le besoin de gérer des demandes multiples, le besoin de gérer la charge actuelle des missions, ou le besoin de gérer la couverture nuageuse. D'un point de vue plus large, la définition du service de couverture multi-mission est liée à plusieurs questions de recherche fondamentale telles que "comment construire un modèle de substitution des capacités d'un ensemble d'agents parmi lesquels un ensemble d'objectifs doit être réparti?", "comment estimer la récompense à long terme associée à une allocation de tâches donnée compte tenu de diverses sources d'incertitude?", ou "comment exploiter une approche d'*intelligence artificielle hybride* combinant l'optimisation et l'apprentissage automatique pour la fédération d'un ensemble de ressources?".

3 Fédération des communications

Pour la partie communication, nous adaptons l'utilisation de stations de communication non propriétaires en utilisant le paradigme GSaaS [3, 16], où les clients réservent des ressources de communication auprès d'un fournisseur de service de segments sols, sur une base de paiement à l'utilisation, au lieu de construire et exploiter leurs propres stations. Nous abordons ici le problème de décision suivant : "*Comment attribuer des paquets de fenêtres de communication aux satellites afin de répondre aux exigences en matière de flux de données, de minimiser le brouillage et de minimiser les coûts induits par les services de réservation?*" Une telle décision doit être prise régulièrement, plusieurs jours à l'avance, ou lorsqu'une demande urgente nécessite le téléchargement de données très rapidement.

3.1 Composition du système

Le problème considéré implique un ensemble de N satellites (détenus par des agents-missions), à la recherche de fenêtres de communication pour le transfert de données en liaison montante et descendante, et un ensemble de M agents de sites de communication indépendants, appelés sites, mettant en œuvre des interfaces GSaaS.

Au sein de la couche de fédération, un agent de réservation de sites de communication devrait être chargé de fournir des opportunités de communication aux satellites des missions fédérées (voir Figure 2). Il doit gérer les communications entre N satellites et M sites au sol, chaque site comprenant une ou plusieurs stations. Cette gestion doit être basée sur le concept de *contact*, c'est-à-dire la communication entre un satellite et un site pendant un intervalle de temps. Cet agent doit :

1. calculer tous les contacts potentiels (la Figure 4 illustre le concept de contact potentiel),

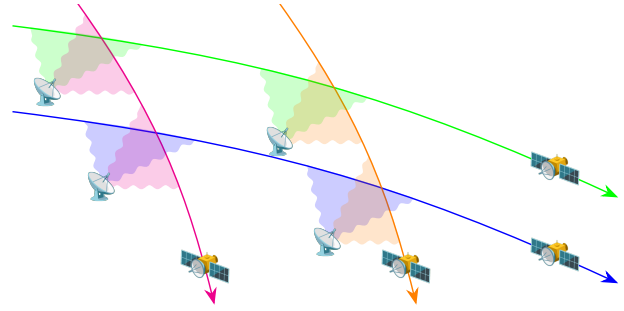


FIGURE 4 – Quatre satellites, chacun ayant des contacts potentiels avec deux stations au sol.

2. sélectionner les contacts à réserver, et
3. réserver les contacts sélectionnés.

Pour le point (1), les missions doivent fournir les orbites de leurs satellites et un propagateur d'orbite doit calculer la position de chaque satellite en fonction du temps et enregistrer les événements correspondant à l'entrée ou à la sortie du cône de visibilité de chaque station. Ces calculs fournissent, pour chaque satellite i , L_i contacts potentiels. Pour le point (3), l'activité de réservation de l'agent peut être purement réactive et basée sur des règles prédéfinies qui doivent être respectées lors de l'interaction avec chaque fournisseur GSaaS, sur la base d'un niveau de service convenu.

3.2 Décisions et problèmes multi-agents

La sélection des contacts à réserver dépend des besoins de communication de chaque mission d'observation de la Terre. La réservation des contacts sélectionnés s'effectue en adressant des demandes aux agents associés à chaque réseau de stations sols. La réservation d'un contact n'est pas toujours couronnée de succès car le site demandé peut ne pas avoir une capacité suffisante. Par conséquent, l'agent de réservation peut itérer les activités (2) et (3) au cours des cycles de négociation avec les agents des sites.

La sélection des contacts peut être basée sur la résolution d'un problème d'optimisation dont les variables, $x_{i,l} \in \{0, 1\}$, correspondent à l'utilisation du contact $l \in \{1, \dots, L_i\}$ par le satellite i , pour $i \in \{1, \dots, N\}$. Pour chaque satellite i , l'objectif est de satisfaire K_i besoins de contact, où chaque besoin k est défini par une durée de contact requise $D_{i,k}$, une liste de sites candidats, et une fenêtre temporelle. Chaque besoin de chaque satellite doit être satisfait, ce qui conduit à $\sum_{i=1}^N K_i$ contraintes inégalité.

Deux critères entrent en ligne de compte pour la sélection des contacts : le coût total, C , et le niveau total de conflit et de brouillage, J . Le coût est la somme des coûts des contacts sélectionnés, soit C , i.e. $C = \sum_{i=1}^N \sum_{l=1}^{L_i} c_{i,l} x_{i,l}$ où $c_{i,l}$ est le coût du l ème contact du satellite i . En ce qui concerne les conflits et le brouillage, ils peuvent se produire lorsque les fenêtres temporelles de deux contacts de deux satellites sur le même site se chevauchent. Cela peut s'écrire comme suit : $J = J_{\text{out}} + J_{\text{in}}$ avec $J_{\text{out}} = \sum_{i=1}^N \sum_{l=1}^{L_i} f_{i,l} x_{i,l}$

et $J_{in} = \sum_{i=1}^{N-1} \sum_{l=1}^{L_i} \sum_{j=i+1}^N \sum_{m=1}^{L_j} b_{i,l,j,m} x_{i,l} x_{j,m}$, où $f_{i,l}$ caractérise le conflit et le brouillage entre le contact l du satellite i et les satellites n'appartenant pas à la fédération. $b_{i,l,j,m}$ caractérise le conflit et le brouillage entre le contact l du satellite i et le contact m du satellite j . Si ces contacts ne se chevauchent pas ou s'ils ne sont pas liés au même site, $b_{i,l,j,m} = 0$. Dans le cas contraire, la valeur de $b_{i,l,j,m}$ dépend des caractéristiques du site. Si le site possède une station et donc une seule antenne, il s'agit d'un conflit et $b_{i,l,j,m}$ est le rapport de l'intersection des fenêtres temporelles sur leur union. Si le site possède plusieurs antennes, il peut y avoir brouillage lorsque l'angle entre les satellites i et j vus du site est inférieur à une valeur critique. Dans ce cas, $b_{i,l,j,m}$ peut être calculé comme le rapport entre la durée pendant laquelle cette condition est remplie et la durée de l'union des deux contacts. Les critères de coût et de brouillage doivent être combinés à l'aide d'une stratégie d'optimisation lexicographique, dans laquelle un critère est optimisé en premier et les égalités sont rompues en optimisant le second critère.

3.3 Défis pour la couche de fédération

3.3.1 Problèmes fortement combinatoires

Pour un agent de réservation, le problème de la sélection des contacts, c'est-à-dire des créneaux à demander au GSaaS, est hautement combinatoire, en raison du grand nombre de satellites et de contacts, et multicritères. Du côté centralisé, plusieurs options algorithmiques sont possibles, telles que la recherche locale, la recherche arborescente et la programmation en nombres entiers ou par contraintes. On peut également considérer qu'il s'agit d'un problème *d'allocation de ressources multi-agents* ou *de planification multi-agents* [6, 21]. Cependant, des méthodes de résolution spécifiques doivent être conçues pour s'adapter à la prise en compte de centaines de sites et de dizaines de satellites pour des horizons temporels supérieurs à une semaine. En outre, en raison des interférences possibles entre les contacts attribués, le problème de réservation devient un problème d'attribution multi-agents fortement couplé, où l'utilité d'un seul agent dépend fortement des lots de contacts des autres agents.

3.3.2 Besoins d'un modèle des autres agents

L'activité de réservation peut induire un *problème d'apprentissage* de la probabilité de rejet d'une demande de contact. Ce problème est également lié à l'apprentissage des termes $f_{i,l}$ sur la base de l'ensemble des contacts acceptés et rejetés. Pour la même raison que pour les activités d'observation, les activités de communication nécessitent *l'apprentissage du comportement et de la charge de chaque service GSaaS*, et pourraient être traitées comme un *problème d'apprentissage par renforcement multi-agents* (MARL) [1]. En effet, l'agent de réservation pourrait apprendre le comportement de chaque fournisseur de GSaaS en fonction des résultats des demandes antérieures, et pourrait également construire un modèle de substitution pour évaluer rapidement les probabilités d'acceptation de chaque demande de réservation en fonction de la charge du site,

compte tenu du fait que d'autres clients peuvent émettre des demandes pour des fenêtres temporelles concurrentes.

3.3.3 Comportements stratégiques

L'activité de réservation peut être influencée par des questions stratégiques. Par exemple, on peut envisager réserver le plus tôt possible pour augmenter la probabilité d'acceptation de la demande de réservation ou, au contraire, réserver le plus tard possible pour avoir une meilleure idée des besoins réels de communication en fonction de l'utilisation des satellites. Là encore, les techniques d'apprentissage par renforcement pourraient être utiles pour optimiser ces décisions.

3.4 Travaux connexes

Le problème de l'ordonnancement des téléchargements des satellites d'une constellation vers un réseau de stations au sol dédié est bien couvert par la littérature [5, 4, 24, 23, 19]. D'autres visions fédérées ont été développées dans plusieurs pays. Par exemple, ESTRACK est un réseau mondial de stations au sol qui soutient les missions de l'ESA [8]. Il se compose de 13 stations situées en Europe, en Afrique, en Australie et en Amérique du Sud. D'autres pays disposent également de réseaux similaires, comme les États-Unis, la Chine, la Russie, le Japon et l'Inde. Cependant, le problème de l'optimisation des besoins de communication de plusieurs constellations utilisant plusieurs réseaux de stations au sol avec des conditions d'accès différentes est un nouveau problème résultant du développement de plusieurs réseaux GSaaS concurrents tels que AWS [20] et KSAT [13].

4 Conclusion

Nous avons proposé une approche multi-agents pour orchestrer les activités d'observation de la Terre. Cette approche réduit le besoin d'intervention manuelle et favorise les décisions autonomes pour la programmation des observations et l'attribution des communications. Il reste cependant plusieurs défis à relever, en raison de la complexité combinatoire des problèmes et de la difficulté de modéliser le système complexe à l'aide d'algorithmes d'apprentissage. La communauté des chercheurs en IA et en systèmes multi-agents pourra ainsi s'approprier ces défis.

Remerciements

Ce travail fait partie du projet DOMINO-E qui a reçu un financement du programme Horizon Europe de l'Union européenne pour la recherche et l'innovation dans le cadre de la convention de subvention n°101082230.

This work is part of the DOMINO-E project which received funding from the European Union's Horizon Europe Programme for Research and Innovation under Grant Agreement n°101082230.

Références

- [1] Stefano V. Albrecht, Filippos Christianos, and Lukas Schäfer. *Multi-Agent Reinforcement Learning : Foundations and Modern Approaches*. MIT Press, Cambridge, MA, 2024.
- [2] M. Barkaoui and J. Berger. A new hybrid genetic algorithm for the collection scheduling problem for a satellite constellation. *Journal of the Operational Research Society*, 71(9) :1390–1410, 2020.
- [3] Elisa Carcaillon and Berylia Bancquart. Market perspectives of ground segment as a service. In *71st International Astronautical Congress (IAC)*, Online, 10 2020. International Astronautical Federation (IAF).
- [4] Jeremy Castaing. Scheduling downloads for multi-satellite, multi-ground station missions. In *28th Annual AIAA/USU Conference on Small Satellites*, Logan, Utah, 2014. AIAA/USU.
- [5] Hao Chen, Baorong Zhai, Jiangjiang Wu, Chun Du, and Jun Li. A satellite observation data transmission scheduling algorithm oriented to data topics. *International Journal of Aerospace Engineering*, 2020 :1–16, 2020.
- [6] Yann Chevalere, Paul Dunne, Endriss Ullé, Lang Jérôme, Lemaître Michel, Nicolas Maudet, Julian Padgett, Steve Phelps, Juan Rodríguez-Aguilar, and Paulo Sousa. Issues in multiagent resource allocation. *Informatica*, 30, 01 2006.
- [7] Peter Cramton, Yoav Shoham, and Richard Steinberg. *Combinatorial Auctions*. The MIT Press, Cambridge, MA, 2006.
- [8] Sylvain Damiani, Holger Dreihahn, Jörg Noll, Marc Niézette, and Gian Paolo Calzolari. Automated allocation of esa ground station network services. In *International Workshop on Planning and Scheduling for Space (IWPSS'06)*, Baltimore, MD, USA, 2006. self-published.
- [9] DOMINO-E Consortium. DOMINO-E – Your access to multi-mission Earth observation, 2024.
- [10] EUSPA. Eo and gnss market report. Technical report, European Union Agency for the Space Programme (EUSPA), 2022.
- [11] Adrien Hadj-Salah, Rémi Verdier, Clément Caron, Mathieu Picard, and Mikaël Capelle. Schedule earth observation satellites with deep reinforcement learning, 2019.
- [12] Sarit Kraus. *Strategic Negotiation in Multiagent Environments*. MIT Press, Cambridge, MA, USA, 2001.
- [13] KSAT. Ground network services, 2023.
- [14] C. Lenzen, M. Dauth, T. Fruth, A. Petrak, and E. Gross. Planning area coverage with low priority. In *International Workshop on Planning and Scheduling for Space (IWPSS)*, pages 80–88, Online, 2021. self-published.
- [15] A. Maillard, S. Chien, and C. Wells. Planning the coverage of solar system bodies under geometric constraints. *Journal of Aerospace Information Systems*, 18(5) :289–306, 2021.
- [16] Louis Nguyen. Ground stations as a service (gsaas) for near real-time direct broadcast earth science satellite data. Technical report, NASA, 2012.
- [17] X. Niu, H. Tang, and L. Wu. Satellite scheduling of large areal tasks for rapid response to natural disaster using a multi-objective genetic algorithm. *Journal of Disaster Risk Reduction*, 28 :813–825, 2018.
- [18] Gauthier Picard, Clément Caron, Jean-Loup Farges, Jonathan Guerra, Cédric Pralet, and Stéphanie Rousel. Autonomous agents and multiagent systems challenges in earth observation satellite constellations. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '21*, page 39–44, Richland, SC, 2021. International Foundation for Autonomous Agents and Multiagent Systems.
- [19] Marco Schmidt. *Ground station networks for efficient operation of distributed small satellite systems*. PhD thesis, Universität Würzburg, 2011.
- [20] Amazon Web Services. Aws ground station, 2023.
- [21] Yoav Shoham and Kevin Leyton-Brown. *Multiagent Systems : Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press, USA, 2008.
- [22] S. Tuli, G. Casale, and N. R. Jennings. Gosh : Task scheduling using deep surrogate models in fog computing environments. *IEEE Transactions on Parallel and Distributed Systems*, 33(11) :2821–2833, nov 2022.
- [23] Rafael Vazquez, Federico Perea, and Jorge Galán Vioque. Resolution of an antenna–satellite assignment problem by means of integer linear programming. *Aerospace Science and Technology*, 39 :567–574, 2014.
- [24] Andrea J Writt. *Optimization of CubeSat Ground Stations for Increased Satellite Numbers*. PhD thesis, Monterey, CA ; Naval Postgraduate School, 2018.

Grands Modèles de Langage (Large Language Models) et Règles Logiques pour une Automatisation Décisionnelle Avancée

P. Feillet¹

¹ IBM France Lab

feillet@fr.ibm.com

Résumé

Les Grands Modèles de Langage (Large Language Models) impactent tous les secteurs d'activité en promettant une nouvelle expansion de l'automatisation dans tous les métiers. L'enjeu pour les organisations est de bénéficier des gains de productivité annoncés par l'IA générative tout en sécurisant le raisonnement algorithmique dans la prise de leurs décisions critiques.

Il est opportun d'évaluer les qualités des LLMs et des moteurs d'inférence de règles logiques ainsi que leur hybridation pour préparer une meilleure automatisation des décisions.

Mots-clés

IA hybride, Règles Logiques, Large Language Models, IA Générative, Extraction de connaissances

Abstract

Large Language Models (LLMs) are impacting all sectors of activity by promising a new expansion for automation. The challenge for organizations is to benefit from the productivity gains announced by generative AI while securing algorithmic reasoning in critical decision-making.

It is timely to evaluate the qualities of LLMs and rule inference engines and prepare their hybridization for improved decision-making automation.

Keywords

Hybrid AI, Rules, Logic, Large Language Models, Generative AI, Knowledge extraction

1 Introduction

L'intelligence artificielle générative perturbe le monde de l'IA en ouvrant de nouvelles perspectives d'interactions Humain Machine et repousse les limites des tests de Turing.

Étant donné les progrès récents une question se pose : les LLMs seuls peuvent-ils suffire à automatiser les décisions critiques dans nos organisations ? Sont-ils compétents pour raisonner avec confiance afin de prendre une décision à fort impact pour les entreprises et citoyens ?

2 Les Grands Modèles de Langage

2.1 Définition générale

En résumé, un grand modèle de langage compresse les informations lues à partir d'un corpus de textes d'entraînement pour générer de nouveaux textes à partir d'une requête donnée. Basé sur une architecture de réseau de neurones, son comportement est statistique. Il prend une séquence de « tokens » (groupes de caractères) exprimés dans la requête et produit une autre séquence de « tokens » qui sont les plus probables en regard de sa base documentaire d'entraînement. Ce n'est pas dans sa construction un algorithme de raisonnement ; il ne repose pas sur des mécanismes logiques et peut être vu comme un perroquet stochastique [11].

Les modèles de langage à grande échelle (LLMs) se sont multipliés depuis 2022, disponible en source fermé ou ouvert et disponible en « Model as a Service » ou localement [8].

2.2 Apprentissage par renforcement à partir de rétroaction humaine

Au-delà de leur apprentissage auto-supervisé utilisant des corpus textuels, certains LLMs bénéficient d'un apprentissage par renforcement par rétroaction humaine (Reinforcement Learning from Human Feedback). Cette approche d'apprentissage par renforcement repose sur les commentaires et évaluations humaines pour guider l'apprentissage du modèle sur la base d'une évaluation humaine de ses résultats.

2.3 Chaîne de pensée

Le "Chain-of-thought prompting" (en français, requête par chaîne de pensée) est une méthode qui vise à améliorer les capacités de raisonnement des grands modèles de langage (LLM). Cette technique fonctionne en décomposant la résolution d'un problème en modélisant les étapes successives d'un processus de pensée.

Cette approche popularisée par LangChain [7] permet au

modèle non seulement d'arriver à des réponses plus précises, mais également de développer une méthode de résolution de problèmes plus structurée et explicative, rendant les réponses générées plus compréhensibles pour les utilisateurs. Cette décomposition garde néanmoins des limites en termes de raisonnement.

2.4 « Retrieval-Augmented Generation »

Le RAG pour "Retrieval-Augmented Generation", est une technique conçue pour améliorer les réponses du modèle en intégrant des informations extraites de documents externes. Ce processus se décompose en deux étapes principales :

- Récupération (Retrieval) : Dans cette première étape, le modèle utilise une requête (généralement basée sur la question ou la tâche posée par l'utilisateur) pour chercher et récupérer des informations pertinentes à partir d'une base de données de documents. Cette base de données peut être constituée de textes provenant de diverses sources telles que des livres, des articles ou des sites internet.
- Génération (Augmented Generation) : Ensuite, le modèle utilise les informations récupérées comme contexte supplémentaire pour générer une réponse. Ce faisant, le modèle peut fournir des réponses plus précises, informatives et basées sur des données concrètes, plutôt que de se baser uniquement sur ce qu'il a appris durant son entraînement.

L'approche RAG est particulièrement utile pour les tâches qui nécessitent des réponses détaillées ou des informations spécifiques qui ne sont pas nécessairement contenues dans le corpus d'entraînement du modèle. Elle permet aux modèles de langage d'enrichir les requêtes afin de rendre la génération plus performante. Toutefois elle n'influe pas sur les qualités intrinsèques de raisonnement du LLM.

3 Les moteurs d'inférence de règles

Depuis plusieurs décennies, les entreprises s'appuient sur des solutions logicielles pour automatiser leurs décisions critiques dans le but de déterminer des éligibilités réglementaires, approuver des prêts et souscrire à des offres de services. Au-delà des applications encodant en Java, Javascript ou autre langage, la logique métier, les systèmes de gestion de règles ont été adoptés par les industriels, notamment par les services financiers. Ces systèmes permettent le développement, le test, la simulation, le déploiement et la maintenance de politiques métier en s'appuyant sur des règles logiques. Cette approche repose sur la capture de connaissances et leur formalisation. Les experts métier définissent une ontologie avec des prédicats logiques et des structures de données représentant leur domaine. Ces règles, communément formulées sous forme d'instructions Si <conditions> Alors <Actions>, d'arbres ou tables de décision, sont évalués par un moteur d'inférence causale capable d'instancier et de chaîner l'exécution dans un contexte de données structurées, automatisant ainsi un chemin de raisonnement.

Cette technologie a mûri au fil des ans, permettant de capturer des modèles de décision complexes avec des milliers de règles et de tables de décision. Des produits comme IBM Operational Decision Manager (ODM) [3] ont prouvé leur capacité à capturer des politiques complexes et exécuter plus d'un milliard de décisions quotidiennement tout en prenant quelques millisecondes pour réaliser chaque décision sur une CPU.

Les moteurs de règles comme ceux d'IBM ODM et ADS permettent de prescrire un micro-flot de raisonnement, afin par exemple de procéder à la validation d'un dossier d'emprunt, puis au calcul d'éligibilité, et enfin de calculer les modalités de remboursements et d'assurance en fonction du profil, de la politique en cours et des taux bancaires. Ces micro-flots peuvent être des DAG (Direct Acyclic Graph) ou bien autoriser des cycles dans des flots de règles (ruleflows). Ils organisent les étapes de raisonnements, et évaluent dans chacune des tâches rencontrées un ensemble de règles.

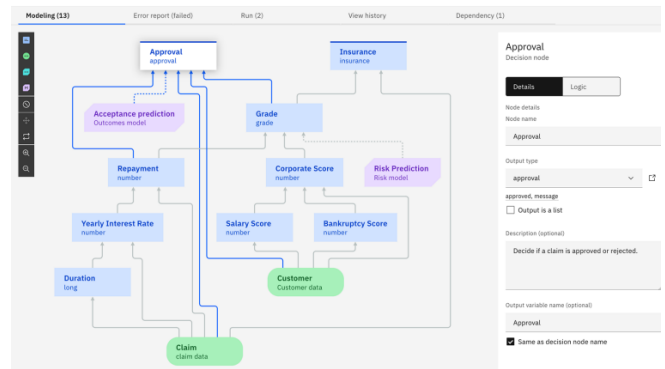


Figure 1: Un modèle de décision défini comme un graphe acyclique dirigé dans IBM ADS

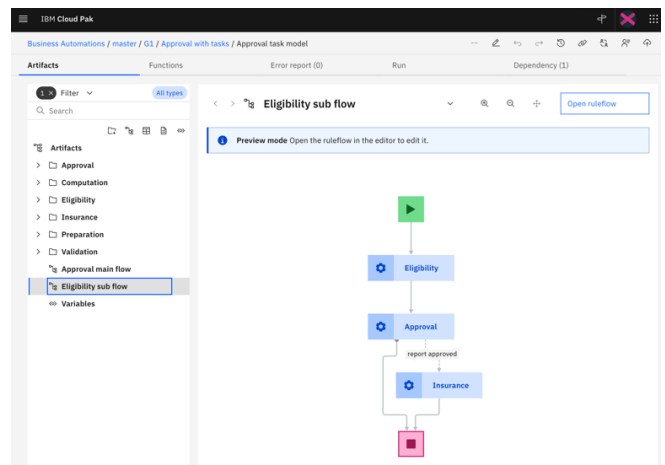


Figure 2: Un micro-flow appelé Ruleflow pouvant contenir un cycle dans IBM ADS

	Loan amount	Rate	Insurance required
1		0	false
2		0.001	true
3		0.003	true
4	A	0.005	true
5	B	0	false
6	B	0.003	true
7	B	0.002	true
8	B	0.007	true

Figure 3: Une table de décision regroupant des règles dans IBM ADS

4 Les exigences d'automatisation décisionnelle

4.1 Quelles sont les exigences clés pour la prise de décision en entreprise ?

L'automatisation des décisions en entreprise implique l'utilisation de la technologie et des logiciels pour systématiser les processus de prise de décision au sein d'une organisation. Les critères clés pour une automatisation réussie des décisions en entreprise sont les suivants :

- Exactitude et fiabilité : le système doit être précis dans le traitement des données et la prise de décisions pour assurer la confiance des utilisateurs et des parties prenantes.
- Évolutivité : la solution doit être capable de gérer de grands volumes de données et de demandes de prise de décision sans dégradation significative des performances.
- Flexibilité et adaptabilité : les environnements économiques et réglementaires sont dynamiques, et les exigences en matière de prise de décision peuvent changer. Le système doit être suffisamment flexible pour s'adapter à de nouvelles règles commerciales, politiques et réglementations sans nécessiter de coûts majeurs de développement.
- Prise de décision en temps réel : dans certains scénarios, des capacités de prise de décision en temps réel ou quasi-temps réel sont cruciales. Le système doit traiter les données et fournir des décisions dans des délais acceptables.
- Transparence et auditabilité : les décisions d'entreprise impactent souvent des processus critiques, et les parties prenantes ont besoin de comprendre comment les décisions sont prises. Le système doit fournir des explications claires pour les décisions, et il doit être auditable à des fins de conformité réglementaire.
- Sécurité et confidentialité des données : puisque l'automatisation des décisions traite de données sensibles, la sécurité et la confidentialité des données sont primordiales. Le système doit employer des mesures de sécurité robustes pour protéger les données contre des accès ou manipulations non autorisés.
- Observations des performances : le système doit

disposer de capacités de mesures et de rapport pour suivre les performances des processus de prise de décision et identifier les domaines à améliorer.

- Maitrise des coûts : la considération du coût du système par rapport à ses avantages est essentielle. La solution doit offrir un bon retour sur investissement et s'aligner sur les contraintes budgétaires de l'organisation.

Dans l'ensemble, une solution réussie d'automatisation des décisions en entreprise doit s'aligner sur les objectifs de l'organisation, rationaliser les processus de prise de décision et contribuer à une efficacité et une productivité accrue.

4.2 Les LLMs atteignent-ils seuls toutes ces exigences ?

Bien que certains LLMs montrent des résultats impressionnants et certaines capacités de raisonnement, ils échouent tout aussi facilement lors de la répétition de l'expérience, ou lors de légères modifications dans le requête. Vous pouvez expérimenter ce comportement à double face avec le « chatbot » de commande de pizza présenté dans le tutoriel OpenAI de DeepLearning.ai [1]. Selon les essais, le chatbot fournit le résultat attendu ou un résultat surprenant, même avec seulement un léger changement de requêtes.

Un autre défi avec les LLMs est leur limite maximale de « tokens », qui restreint la quantité de contexte qu'ils peuvent gérer. Une technique pour résoudre ce problème est d'appliquer une approche « Retrieve Augment Generate » [10] ou d'étendre l'apprentissage du LLM par fine-tuning sur un corpus complémentaire.

Par ailleurs, l'apprentissage par renforcement avec retour humain appliqué aux LLMs pour en atténuer leurs réponses purement statistiques peut s'avérer incomplet voir contre-productif dans certaines tâches [9].

4.3 Pouvons-nous les utiliser en combinaison avec des moteurs de décision basés sur des règles ?

Les moteurs d'inférence de règles permettent l'automatisation de prises de décisions. Ils reposent sur la capture d'un savoir-faire métier dans une ontologie et la spécification non ambiguë de la logique formalisant l'expertise sur des données structurées. Cela garantit une prise de décision déterministe, robuste et transparente en échange d'un effort de capture et de formalisation de la connaissance.

Le défi réside dans la fusion de ces technologies pour capitaliser sur leurs points forts, à l'instar de la création de matériaux composites qui surpassent les propriétés individuelles de chaque élément.

Nous proposons différentes méthodes combinant les LLMs avec des moteurs de règles.

5 Hybridations LLM et règles logiques

5.1 Compréhension du langage naturel suivie par un raisonnement basé sur des règles

Dans cette approche, nous utilisons d'abord un modèle de langage basé sur l'apprentissage automatique (LLM) pour comprendre le texte brut et en extraire des données structurées. Ensuite, nous passons ces données structurées extraites à un moteur de règles logiques pour raisonner de manière déterministe sur ces données, potentiellement combinées avec des informations supplémentaires provenant de référentiels d'entreprise.

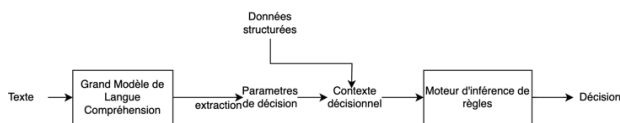


Figure 4: Compréhension du langage naturel suivie par un raisonnement basé sur des règles logiques

Avantages : L'intégration séquentielle d'un LLM dédiée à la compréhension du langage naturel afin d'en extraire des entités structurées suivie par l'exécution d'un moteur de règles est simple à mettre en place. Le pipeline LLM-Moteur de règles implique de transmettre les résultats du LLM comme paramètres d'entrée pour alimenter le contexte du moteur de règles.

Inconvénients : La performance de l'ensemble dépend de l'aptitude du LLM à extraire correctement les données structurées. Pour se protéger des cas où celles-ci s'avèrent incomplètes ou mal formatées, il est crucial de mettre en place des garde-fous avec des vérifications, des valeurs par défaut ou des heuristiques de repli pour alimenter la phase de raisonnement. L'extraction du contexte décisionnel doit se révéler robuste via une ou plusieurs appels au LLM, et éventuellement par une chaîne de pensée. Elle peut s'appuyer sur des interactions utilisateur dans le cadre d'une expérience conversationnelle.

5.2 Raisonnement basé sur des règles suivi par la génération de langage naturel avec un LLM

Dans cette intégration, un moteur de règles prend d'abord une décision basée sur des données structurées. La décision structurée est ensuite transmise à un modèle de langage qui en fait déduire une transformation dans un texte de langage comme un résumé ou un courriel.

Cette hybridation est particulièrement utile pour fournir des explications, justifications, résumer les décisions d'entreprise et les communiquer. Le contenu généré peut être adapté en fonction de du profil du destinataire et la communication visée.

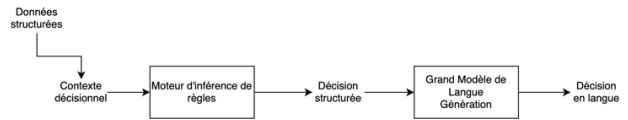


Figure 5: Raisonnement basé sur des règles suivi par une génération de texte en langage naturel avec un LLM

Avantages : Cette approche offre l'avantage de tirer parti des puissantes capacités de génération de langage naturel (NLG) fournies par le LLM pour créer des textes bien formulés pour communiquer des décisions automatisées.

Elle peut être facilement mise en œuvre facilement via une ingénierie de requête « prompt engineering » en passant la décision structurée au LLM.

Inconvénients : Alors que les règles se concentrent sur le raisonnement avec des données structurées, le LLM se concentre sur les tâches de traitement du langage naturel (NLP). Cette séparation nécessite une mise au point minutieuse pour gérer les variations de génération tant dans leurs formats que dans leurs contenus. Des tests automatisés du résultat de génération de langage naturel (NLG) sont nécessaires pour mesurer la performance de cette hybridation ainsi que la couverture des erreurs et variations.

5.3 Raisonnement basé sur des règles pilotant le traitement du langage naturel avec le LLM

Le moteur d'inférence de règles agit ici comme le moteur principal, invoquant le modèle de langage à la demande. Le moteur de règles pilote l'évaluation logique et appelle dynamiquement le LLM pour déléguer deux tâches :

- Traitement du texte reçu dans son contexte de décision pour la compréhension (NLU) et l'extraction de données structurées.
- Générer du texte (NLG) pour produire, par exemple, un résumé de la décision automatisée en texte brut.

Cette intégration se poursuit dans la continuité de l'appel de tout modèle d'apprentissage automatique depuis IBM ODM [3] ou IBM ADS [2] [12], pour considérer les probabilités de risque ou d'opportunité lors d'une prise de décision. De même, le LLM peut être appelé à partir d'une règle, soit à distance, soit localement, en fonction de son facteur de forme.

Il est possible d'envisager cette hybridation comme une implémentation de chaînes de pensée par un moteur logique, en considérant que le LLM est appelé à chaque étape ou le traitement du langage intervient.

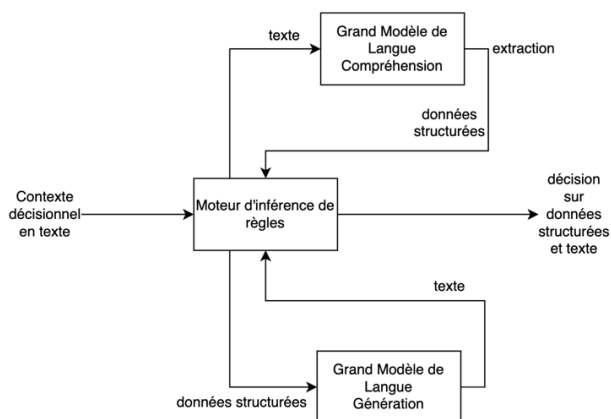


Figure 6: Raisonnement basé sur des règles pilotant le traitement du langage naturel avec le LLM

Avantages : Le contrôle est assuré par le moteur de règles permet d’orchestrer les étapes de raisonnement, et procéder par déduction. Les LLMs sont appelés à la demande en fonction du chemin de raisonnement pour réaliser les tâches de NLP.

Inconvénients : Les moteurs de raisonnement et LLMs sont étroitement couplés, nécessitant une intégration fine. Les données structurées utilisées dans la prise de décision doivent s’aligner avec les tâches de NLP. Des garde-fous appropriés sont nécessaires pour gérer la frontière entre données structurées et non structurées pour garantir la qualité et fiabilité de l’hybridation.

5.4 Extraire des règles métier depuis du texte avec un LLM pour les exécuter dans un moteur logique

Cette approche utilise un modèle de langage pour extraire des éléments d'automatisation, y compris des règles logiques, des modèles de données et des signatures fonctionnelles, à partir de politiques d’entrepris exprimées en texte brut. Ces éléments extraits sont ensuite utilisés pour générer un projet d'automatisation dans un moteur de règles comme ceux d’IBM ADS ou ODM. Cette approche a déjà été prototypée avec succès avec IBM ADS en passant par la génération d’un descripteur pivot en JSON.

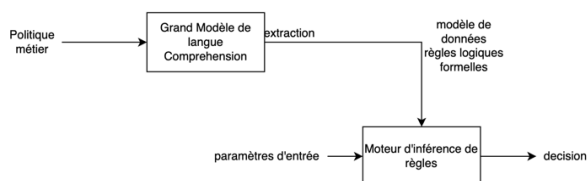


Figure 7: Extraire des règles logiques du texte avec un LLM et exécuter ces règles dans un moteur d’inférence

Avantages : Tirer parti du LLM comme outil d'extraction de connaissances permet d'extraire des règles et une ontologie de données sous-jacente. Les règles extraites, une fois révisées par un humain, permettent d’automatiser des décisions avec traçabilité et déterminisme. C’est un chemin prometteur pour assister l'extraction de connaissances expertes et capitaliser sur des moteurs d’inférence pour l’exécution des décisions.

Inconvénients : Cette approche nécessite une chaîne de requêtes efficaces, éventuellement un modèle réglé finement pour réaliser des extractions pertinentes de la logique, pour tout domaine d’expertise et formulation de la politique métier. Elle nécessite également des compétences et outils pour extraire efficacement la connaissance à automatiser, et en assurer la maintenance lorsque les documents sources évoluent ou sur la base de retours opérationnels.

5.5 Règles pour apporter un raisonnement fiable dans un agent conversationnel

Un grand modèle de langage (LLM) est ici utilisé pour piloter l'expérience conversationnelle, gérant les tâches de traitement du langage naturel (NLP). Le LLM délègue à un moteur de décision basé sur des règles pour appliquer des décisions logiques.

Cette intégration exige que le chatbot reconnaisse, pendant la conversation, quand déclencher un service de décision. Le chatbot guide le dialogue pour fournir le contexte et invoque le moteur de décision basé sur des règles lorsque tous les paramètres d'entrée sont définis. Le moteur de décision renvoie des paramètres de sortie, qui sont ensuite restitués dans la conversation par la génération de langage naturel (NLG).

IBM travaille activement à incorporer ce schéma pour apporter des capacités de prise de décision à Watson Orchestrate [6]. Par ailleurs les clients peuvent déjà développer des outils dans des solutions open-source comme LangChain pour invoquer des décisions basées sur des règles à partir d'un chatbot [4].

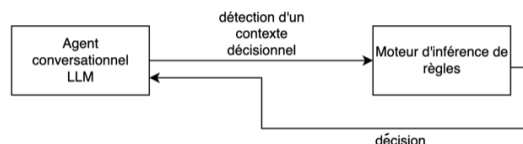


Figure 8: Décision à base de règles logiques injectée pour raisonner dans un agent conversationnel

Avantages : Cette approche permet de bénéficier d’une expérience utilisateur conversationnelle tout en déléguant le raisonnement à des moteurs déterministes spécialisés.

Inconvénients : La mise en œuvre de cette approche nécessite une détection performante des décisions à déléguer, de synchroniser un contexte entre les 2 moteurs, et de gérer les cas d'erreur à la frontière entre les domaines de données structurées et non structurées. Les défis peuvent inclure le traitement de différents formats de données et des informations de contexte incomplètes.

6 Conclusion

Nous avons exploré le pouvoir de transformation des modèles de langage (LLMs) dans l'automatisation des décisions d'entreprise. Les LLMs offrent des capacités impressionnantes dans le traitement du langage. En étant combinés à un apprentissage par renforcement à partir de rétroaction humaine, à des chaînes de pensées ou du RAG ils améliorent leurs performances.

Ils manquent néanmoins de compétences pour un raisonnement fiable et répétable afin de répondre aux exigences strictes associées à la prise de décision critique. Pour combler cette lacune, nous avons introduit cinq approches d'hybridation des LLMs avec des moteurs de raisonnement basés sur des règles logiques.

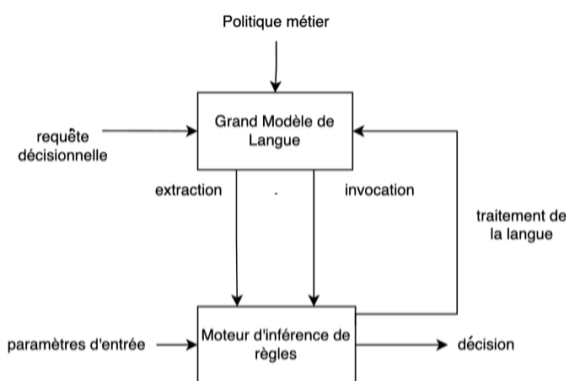


Figure 9: Récapitulatif des combinaisons LLM et règles logiques

Ces motifs d'intégration incluent l'invocation de LLM avant ou pendant l'exécution des règles, pour le traitement du texte brut en complément des données structurées, des règles pilotant le traitement du texte, l'extraction par LLM de règles logiques à partir de documents, et l'utilisation de services de décision basés sur des règles dans des agents conversationnels.

En combinant l'IA générative et symbolique, ces hybridations d'IA visent à promouvoir de nouveaux usages et démocratiser les solutions d'automatisation de prise de décision tout en satisfaisant les enjeux réglementaires et éthiques des entreprises.

7 Références

- [1] Deep Learning OpenAI courses : <https://www.deeplearning.ai/short-courses/chatgpt-prompt-engineering-for-developers/>
- [2] IBM ADS : <https://www.ibm.com/products/automation-decision-services>
- [3] IBM Operational Decision manager : <https://www.ibm.com/products/operational-decision-manager>
- [4] IBM ODM with LangChain : <https://community.ibm.com/community/user/automation/blogs/laurent-grateau1/2023/06/09/integrating-odm-with-large-language-model>
- [5] IBM watsonx : <https://www.ibm.com/watsonx>
- [6] IBM watsonx Orchestrate : <https://www.ibm.com/products/watson-orchestrate>
- [7] LangChain : <https://arxiv.org/html/2402.06196v1>
- [8] Large Language Models: A survey : <https://arxiv.org/html/2402.06196v1>
- [9] Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback : <https://arxiv.org/abs/2307.15217>
- [10] Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks : <https://arxiv.org/abs/2005.11401>
- [11] Stochastic parrot : https://en.wikipedia.org/wiki/Stochastic_parrot
- [12] Two ways of integrating Machine Learning in IBM ADS : <https://www.youtube.com/watch?v=7ZJqt5bexS8>

Un graphe multicouche pour la représentation des processus d'expertise

Serge SONFACK SOUNCHIO¹, Bernard KAMSU-FOGUEM², Laurent GENESTE²

¹ Robert Bosch Sp. z o. o. Jutrzenki 105, POLAND

² Laboratoire Génie de Production/UFTMiP 47, Avenue d'Azereix, BP 1629, 65016 Tarbes Cedex, France

Résumé

Les processus d'expertise sont une approche utilisée par les experts pour comprendre des problèmes complexes sur la base d'hypothèses. Il permet d'explorer les explications possibles d'un problème de façon itérative. Cette approche est généralement utilisée pour élucider des problèmes lorsque l'on n'a pas assez de connaissances sur le problème et permet ainsi de les résoudre, d'en tirer des leçons et de prendre des décisions. Cependant, bien que ce processus ait été normalisé au niveau européen pour encadrer les experts dans leurs activités, les processus d'expertise sont encore peu formalisés et les artefacts produits sont difficiles à comprendre et à réutiliser en raison de leur structure en graphes complexes nommés graphes exploratoires d'hypothèses. Cette étude propose une nouvelle formalisation des graphes exploratoires d'hypothèses en graphes multicouche, capitalisant ainsi différents types d'éléments structurants (questions, hypothèses, et connaissances) sur chaque couche et facilitant leur utilisation et leur compréhension par les humains et les machines. Un cas d'utilisation est proposé pour illustrer la représentation en graphe multicouche pour les graphes exploratoires d'hypothèses et la manière dont ils sont interrogés.

Mots-clés

Processus d'expertise, graphe multicouche, représentation des connaissances.

Abstract

An expertise process is an approach used by experts to understand complex problems based on hypotheses. It aids in exploring all possible explanations of a problem and unlocks its understanding. This approach is used for elucidating issues when confronted with limited knowledge regarding a problem and enables experts to solve problems, learn lessons, or make decisions. However, although this process has been standardized at the European level for human experts' guides, the Expertise Process is still poorly computerized. The artifacts produced during the processes are challenging to understand and reuse because of its complex graph structure called hypotheses exploratory graphs. This study formalizes hypotheses exploratory graphs in a multilayered graphs representation, thus capturing different types of knowledge at each layer and making it easier for humans and machines to use and understand.

A use case is carried out to illustrate the proposed layered representation of hypotheses exploratory graphs and how they are queried.

Keywords

Expertise process, Multilayer graph, Knowledge representation.

1 Introduction

Le processus d'expertise est une approche utilisée dans le cadre de l'expertise afin d'explorer les explications possibles des problèmes dans des conditions de connaissances limitées. Il implique des experts de différents domaines qui collaborent pour augmenter les possibilités d'étudier tous les indices pertinents pouvant mener à l'explication et la résolution du problème en question. Ce processus commence par des questions et se développe au fur et à mesure que d'autres questions et d'hypothèses associées sont exprimées, puis validées ou non, sur la base d'observations de l'environnement du problème. Principalement, les hypothèses formulées par les experts sont incertaines, et les déclarations prédictives [4] émanent de leur expérience. En revanche, les observations peuvent être des éléments de connaissance dérivés d'informations objectives provenant d'équipements de mesure, de capteurs ou d'observations générales de phénomènes perçus dans l'environnement lié au problème. En outre, les observations peuvent être effectuées à l'aide de connaissances spécifiques au domaine provenant d'êtres humains et d'autres systèmes [7].

Même si le processus d'expertise est guidé par des cadres normatifs tels que la norme française *NF X50 - 110* « Qualité des activités d'expertise » et par la norme européenne *CSN EN 16775* « Activités d'expertise exigences générales pour les services d'expertise », il n'est pas doté d'une représentation formelle que les humains et les machines peuvent facilement exploiter [8].

Les travaux de [2, 3] utilisent des techniques de représentation des connaissances telles que les graphes conceptuels et le web sémantique pour démontrer comment le savoir-faire humain peut être partagé ou réutilisé. L'étude [11] a proposé pour la représentation des processus d'expertise une structure de graphe qui s'est avérée complexe à comprendre et à réutiliser en raison de son intégration de com-

posants hétérogènes comme des connaissances supplémentaires, des hypothèses et des questions.

Cette étude élabore un formalisme de graphe multicouche pour représenter et interroger les processus d'expertise basés sur des hypothèses. L'approche proposée personnalise le cycle de raisonnement du processus d'expertise fondé sur des hypothèses, en intégrant la construction d'une couche spécifique du graphe multicouche à chaque étape du cycle. Le graphe multicouche obtenu offre une grande expressivité à partir des couches séparées et une compréhension de la connaissance du processus d'expertise à différents niveaux. Le reste de ce document est structuré comme suit. La section 2 présente les processus d'expertise et les graphes exploratoires des hypothèses. La section 3 définit la représentation en graphe multicouche et la manière dont elle est utilisée pour représenter les processus d'expertise. La section 4 conclut cette étude.

2 Cadre général

2.1 Processus d'expertise

L'expertise est une activité cognitive menée par des agents possédant des connaissances de haut niveau dans le même domaine ou dans des domaines différents, acquises par l'expérience ou l'apprentissage. Elle vise à fournir une compréhension d'une situation obscure ou des réponses à des problèmes complexes. Les processus d'expertise sont des lignes directrices normalisées françaises et européennes pour une méthodologie exploratoire basée sur des questions et des hypothèses que les experts suivent pour explorer les explications possibles d'un problème complexe. Ces processus prennent également en compte l'acquisition de connaissances que les experts utilisent pour valider ou réfuter les hypothèses. La figure 1 montre une trace générale d'un processus d'expertise avec des branches correspondant aux tentatives d'exploration et les nœuds du graphe aux points du processus où des questions, des hypothèses et des connaissances sont fournies [9, 12].

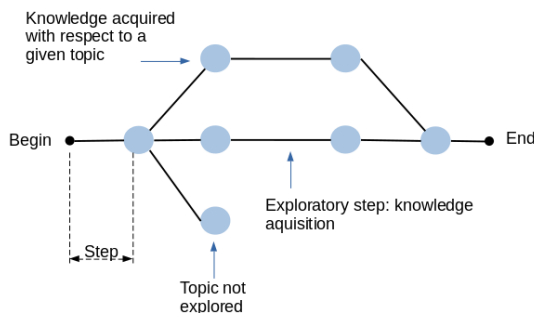


FIGURE 1 – Représentation générale d'un processus d'expertise de [13] montrant les étapes d'exploration et les nœuds de sujet

2.2 Hypotheses exploratory graphs (HEG)

Le processus de construction d'un graphe exploratoire d'hypothèses (GHE) à partir d'un processus d'expertise

quel que soit le domaine (assurance de responsabilité civile, immobilier, automobile, aéronautique) est un processus collaboratif et itératif homme-machine. Chaque itération comporte quatre étapes dans lesquelles les compétences et les capacités cognitives humaines contribuent aux deux premières : poser des questions et exprimer des hypothèses, qui sont des pensées incertaines et prédictives pouvant servir de base au raisonnement, à la résolution de problèmes ou à la recherche de relations plausibles entre plusieurs variables d'une circonstance [6]. Les deux autres étapes impliquent l'ingestion de connaissances supplémentaires et un raisonnement défaitiste de la part de la machine. Pour que le calcul soit possible, la logique d'hypothèse et son mécanisme de raisonnement défaisable, connu sous le nom d'extension, ont été adoptés pour la représentation des connaissances/observations et des hypothèses supplémentaires [10].

La figure 2 illustre les étapes du cycle de raisonnement d'hypothèse homme-machine, qui se déroule comme suit : (1) définir l'itération initiale avec son observation initiale; (2) poser des questions dans le cadre de cette itération : Ces questions aident à comprendre le problème; (3) exprimer des hypothèses pour chaque question de l'itération; (4) collecter des observations : les hypothèses ont par défaut un statut inconnu; (5) raisonner et changer le statut des hypothèses; (6) établir une nouvelle itération et recommencer pour chaque hypothèse posée jusqu'à satisfaction ou contrainte de temps.

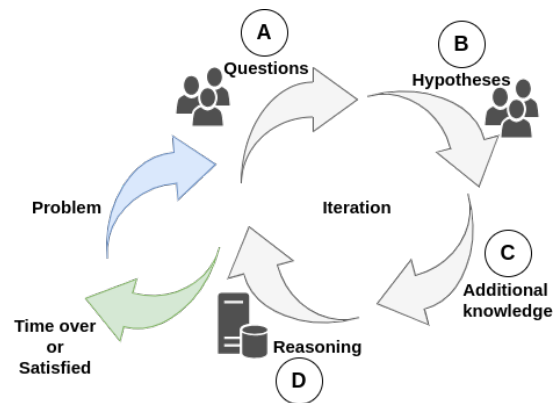


FIGURE 2 – Cycle de raisonnement du processus d'expertise des hypothèses montrant ses étapes itératives et ses conditions d'arrêt, extraites de [11]

2.3 Raisonnement et apprentissage HEG

La construction du graphe exploratoire d'hypothèses (HEG) est basée sur un mécanisme de *raisonnement faisable* et de *hypothèse du monde ouvert* qui influencent l'état des hypothèses (valides ou inconnues) d'une itération à l'autre au cours du processus d'expertise. Par conséquent, on s'attend à ce que l'état d'une hypothèse change au cours des différentes itérations en raison du raisonnement défaitiste et des connaissances supplémentaires. Le raisonnement défaitiste et les connaissances supplémentaires sont des éléments essentiels du processus d'expertise. L'hypo-

thèse d'un monde ouvert dans la construction du HEG soutient le fait que la non-validité d'une hypothèse est peut-être due à un manque de connaissances complètes qui pourraient la rendre valide.

3 Graphe d'hypothèses multicouche

Les connaissances formalisées à partir de cette méthodologie ont la vue synoptique présentée dans la figure 3. Cette représentation intègre les connaissances utilisées pour résoudre le problème, y compris les questions, les hypothèses et les observations, ainsi que les étapes suivies pour résoudre le problème.

- **Couche 0** : représente les itérations du processus, y compris les connaissances utilisées à chaque itération.
- **Couche 1** : représentation graphique des questions posées au cours du processus exploratoire et de leur cheminement.
- **Couche 2** : représentation graphique des hypothèses émises par les experts en fonction des questions posées à chaque itération.

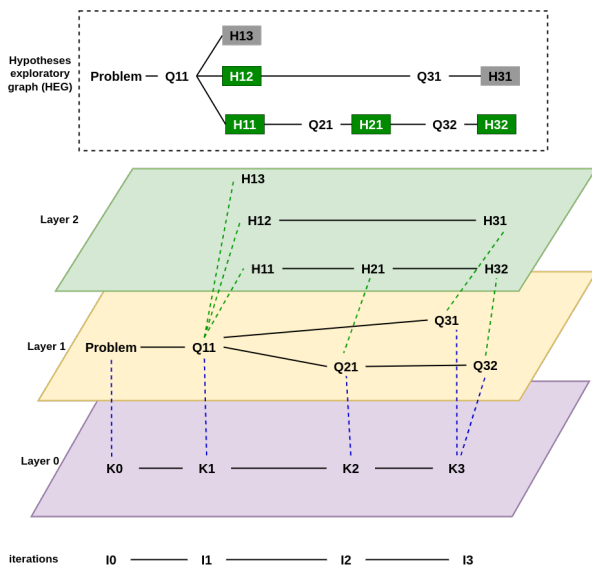


FIGURE 3 – Vue multicouche du graphe exploratoire des hypothèses présenté dans l'illustration : la couche 0 correspondant à la couche de connaissances, la couche 1 à la couche de questions et la couche 2 à la couche d'hypothèses.

3.1 Formalisation

La définition formelle du graphe exploratoire d'hypothèses multicouche est définie dans cette section sur la base du graphe multicouche proposé par [1]. $G = (V, E, L)$ où $V = \{V_K, V_Q, V_H\}$: correspondant aux sommets de connaissances (V_K), aux sommets de questions (V_Q) et aux sommets d'hypothèses (V_H). $E = \{E_1, E_2\}$: correspondant à l'interaction entre les nœuds des différentes couches avec $\{E_1, E_2\}$: correspondant à l'interaction entre les nœuds des différentes couches. $E_1 = \{(v_k, v_q) \in V_K * V_Q\}$

pour les arêtes entre la couche des connaissances et la couche des questions,

$E_2 = \{(v_q, v_h) \in V_Q * V_H\}$ pour les arêtes entre la couche des questions et celle des hypothèses. $L = \{L_0, L_1, L_2\}$: correspondant à la couche de connaissances (L_0), la couche de questions (L_1) et la couche d'hypothèses (L_2). La couche de connaissances (L_0), la couche de questions (L_1) et la couche d'hypothèses (L_2). $L_0 = (V_K, E_K)$ fait référence aux nœuds et aux arêtes de la couche L_0 . $L_1 = (V_Q, E_Q)$ désigne les nœuds et les arêtes de la couche L_1 . $L_2 = (V_H, E_H)$ fait référence aux nœuds et aux arêtes de la couche L_2 , avec les arêtes de connaissance (E_K), les arêtes de questions (E_Q) et les arêtes d'hypothèses (E_H).

3.2 Construction

La construction du processus d'expertise des hypothèses à plusieurs niveaux est basée sur le même cycle de raisonnement du processus d'expertise des hypothèses que celui décrit dans la figure 2. Toutefois, à chaque étape d'une itération, seul le graphe de l'instance d'étape correspondante est mis à jour. Par conséquent, dans l'étape (A), le graphe des questions de la couche 1 est mis à jour; dans l'étape (B), le graphe des hypothèses de la couche 2 est mis à jour; et dans l'étape (C), le graphe des connaissances de la couche 0 est mis à jour. Dans l'étape (D), le raisonnement est effectué sur la base de l'état actuel de chaque couche, de l'état des hypothèses (valides ou inconnues), et les doutes sont modifiés en conséquence.

3.3 Illustration

Pour montrer comment l'approche ci-dessus fonctionne, nous l'avons illustrée par un cas réel dans une entreprise de fabrication. Pour cette illustration, il a été demandé à des experts d'utiliser l'approche proposée pour rechercher des explications sur les raisons pour lesquelles un article avait été retourné par des clients.

- **Itération 0** : Correspond au problème initial.
- **Itération 1** :
 - Question : Pourquoi les produits KW831 ont-ils été rejetés par les clients ?
 - Hypothèses :
 - $h_{1,1}$
 - **Hypothèse** : Il est *presque certainement vrai* que cela est dû à des outils de mesure défectueux.
 - $h_{1,2}$
 - **Hypothèse** : Il est *très probable* que cela soit dû au non-respect du plan de fabrication.
 - $h_{1,3}$
 - **Hypothèse** : Il est *très probable* que cela soit dû à un serrage excessif de ses pièces.
 - Observation :
 - Certains opérateurs n'ont pas été formés à l'utilisation des outils de mesure, et certains

- n'ont donc pas pu mesurer correctement les composants du KW831.
 - Les outils de mesure sont neufs et ont été testés avant d'être utilisés ; ils ne sont donc pas défectueux.
 - Raisonnement :
 - $h_{1,1}$
 - **Hypothèse** : Il est *presque certainement vrai* que cela est dû à des outils de mesure défectueux.
 - **Statut** : Inconnu.
 - $h_{1,2}$
 - **Hypothèse** : Il est *presque certainement vrai* que cela est dû à une non conformité avec le plan de production.
 - **Statut** : Valide.
 - $h_{1,3}$
 - **Hypothèse** : Il est *presque certainement vrai* que cela est dû au non-respect du plan de fabrication.
 - **Statut** : Inconnu.
- Remarques** : Pour le processus de raisonnement, les hypothèses avec un statut *Inconnu* sont celles qui n'ont pas été soutenues par les observations, tandis que les hypothèses avec *Valide* sont celles qui sont cohérentes avec les observations. Ce mécanisme est utilisé à chaque itération.
- **Itération 2** :
 - Question : Pourquoi ces KW831 récemment fabriqués ont-ils été mal serrés ?
 - Hypothèses :
 - $h_{2,1}$
 - **Hypothèse** : Il est *probablement vrai* que c'est parce que les opérateurs ont mal fait le travail.
 - Observation :
 - Seuls les KW831 récemment fabriqués sont rejetés par les clients.
 - Raisonnement :
 - $h_{1,1}$
 - **Hypothèse** Il est *presque certainement vrai* que cela est dû à des outils de mesure défectueux.
 - **Statut** : Inconnu.
 - $h_{1,2}$
 - **Hypothèse** : Il est *très probable* que cela soit dû au non-respect du plan de fabrication.
 - **Statut** : Inconnu.
 - $h_{1,3}$
 - **Hypothèse** : Il est *très probable* que cela soit dû à un serrage excessif de ses pièces.
 - **Statut** : Inconnu.
 - $h_{2,1}$
 - **Hypothèse** : Il est *probablement vrai* que cela soit dû à un serrage excessif de ses pièces.
 - **Statut** : Inconnu.
- **Itération 3** :
 - Question : Pourquoi les dimensions des pièces du KW831 n'ont-elles pas été respectées ?
 - Hypothèses :
 - $h_{3,1}$
 - **Hypothèse** : Il est *probablement vrai* que cela est dû à des erreurs de mesure.
 - Question : Pourquoi ces opérateurs nouvellement recrutés ne sont-ils pas bons ?
 - $h_{3,2}$
 - **Hypothèse** : Il est *certainement vrai* que les opérateurs n'aient pas été bien formés sur la chaîne de production.
 - Observations :
 - Il y a des opérateurs nouvellement recrutés, ils pourraient donc mal monter ou mesurer les composants du KW831.
 - Les opérateurs ont travaillé sous pression afin de livrer les produits KW831 à temps, il est donc possible qu'il y ait des erreurs de fabrication.
 - Les opérateurs nouvellement recrutés sont des travailleurs inexpérimentés.
 - Raisonnement :
 - $h_{1,1}$
 - **Hypothèse** : Il est *presque certainement vrai* que cela est dû à des outils de mesure défectueux.
 - **Statut** : Inconnu.
 - $h_{1,2}$
 - **Hypothèse** : Il est *presque certainement vrai* que cela est dû au non-respect du plan de fabrication.
 - **Statut** : Valide.
 - $h_{1,3}$
 - **Hypothèse** : Il est *presque certainement vrai* que cela est dû au non-respect du plan de fabrication.
 - **Statut** : Valide.
 - $h_{2,1}$
 - **Hypothèse** : Il est *très probable* que c'est parce que les opérateurs ont mal fait le travail.
 - **Statut** : Valide.
 - $h_{3,1}$
 - **Hypothèse** : Il est *très probable* que c'est parce que les opérateurs ont mal fait le travail.
 - $h_{3,2}$
 - **Hypothèse** : Il est *certainement vrai* que les opérateurs n'ont peut-être pas été bien formés sur la chaîne de production.
 - **Statut** : Valide.

Cette illustration est présentée sur la figure 4.

La figure 3 montre respectivement les première, deuxième et troisième couches du graphe multicouche du processus

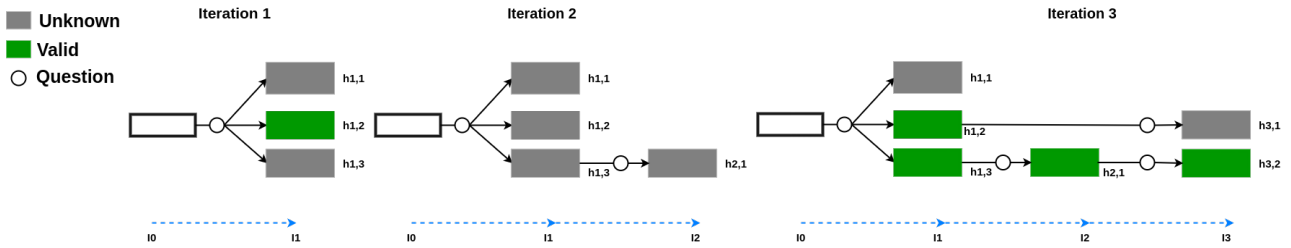


FIGURE 4 – Graphe exploratoire des hypothèses (HEG) de toutes les itérations de l'exemple illustratif de [12], montrant comment les états des hypothèses changent d'une itération à l'autre. L'état *Inconnu* est utilisé à la place de *Invalide* en raison de l'hypothèse *Monde ouvert*.

Block 1 – Query iterations with at least one valid hypothesis (Valid iterations)

```
kgtk query -i HEG_Hedge.tsv --match '()-[:Follows]->(i), (i)-[:HasQuestion]->()-[:HasHypothesis]->()-[:Status]->(v)' --return 'distinct i as `Iteration`,v as `Status`' --where 'v=kgtk_unstringify("Valid")'
```

Block 2 – Sortie des itérations valides pour l'exemple de graphe exploratoire d'hypothèses multicouches (Requête dans le bloc 1)

Iteration	status
I1	Valid
I2	Valid
I3	Valid

d'expertise conçu à l'aide de la boîte à outils du graphe de connaissances [5] (KGTK).

Le KGTK offre des moyens simples et efficaces de créer des graphes, tels que la tabulation et les valeurs séparées par des virgules. En outre, il fournit un langage d'interrogation statistique et basé sur le cryptage.

3.3.1 Requêtes sur le graphe exploratoire d'hypothèses multicouche

- **Itérations valides** : itérations comportant au moins une hypothèse valide. La requête dans le bloc 1 extrait du graphe exploratoire multicouche des hypothèses (MHEG) les étapes valides du problème comprises par les experts. Le bloc 1 est le résultat de la requête.
- **Hypothèses valides** : La requête du bloc 3 récupère toutes les hypothèses valides exprimées au cours du processus exploratoire. Le résultat de la requête est indiqué dans le bloc 3. Le résultat de la requête est présenté dans le bloc 4
- **Chemin valide** : La requête dans le bloc 5 récupère les chemins valides du graphe des couches d'hypothèses. Le résultat de cette requête est affiché dans le bloc 6.

Block 3 – Recherche d'hypothèses valides

```
kgtk query -i HEG_Hedge.tsv --match '()-[:HasHypothesis]->(h)-[:Status]->(v)' --return 'h as `Hypotheses`,v as `Status`' --where 'v=kgtk_unstringify("Valid")'
```

Block 4 – Sortie des causes possibles de l'exemple de graphe exploratoire d'hypothèses multicouches (requête dans le bloc 3)

Hypotheses	Status
It is almost certainly true that it is due to non-compliance with the manufacturing plan	Valid
It is almost certainly true that it is due to the over-tightening of its parts	Valid
It is highly likely that it is because operators poorly did the work	Valid
It is certainly true that operators may not have been well trained on the production line	Valid

Block 5 – Interroger le chemin d'accès valide (Chemin d'accès valide)

```
kgtk query -i HEG_Hedge.tsv --match '(h)-[:Hypothesis]->(j), (h)-[:Status]->(v), (j)-[:Status]->(v)' --where 'v=kgtk_unstringify("Valid")'
```

Block 6 – Sortie des chemins valides de l'exemple de graphe exploratoire d'hypothèses multicouches (requête dans le bloc 5)

```
id node1 label node2 Graph id node1 label node2
  Graph id node1 label node2 Graph
e20 It is highly likely that it is because
  operators poorly did the work Hypothesis
  It is certainly true that operators may not
  have been well trained on the production
  line Hgraph e40 It is highly likely that it
  is because operators poorly did the work
  Status Valid Hgraph e39 It is certainly
  true that operators may not have been well
  trained on the production line Status Valid
  Hgraph
e19 It is almost certainly true that it is due
  to the over-tightening of its parts
  Hypothesis It is highly likely that it is
  because operators poorly did the work
  Hgraph e27 It is almost certainly true that
  it is due to the over-tightening of its
  parts Status Valid Hgraph e40 It is highly
  likely that it is because operators poorly
  did the work Status Valid Hgraph
```

Block 7 – Chemin existant

```
kgtk paths --path-file pairsH.tsv --path-mode
  NONE --path-source source --path-target
  target -i H_Hedge.tsv
```

- **Explication possible du problème** : à partir de la couche 2 (le graphe des hypothèses uniquement), les explications possibles peuvent être interrogées en fournissant les derniers nœuds et le nœud initial du problème. Ces explications possibles correspondent à des chemins d'hypothèses valides allant du nœud initial à un nœud final. La requête dans le bloc 7 récupère les chemins depuis *H32* qui correspond à la dernière hypothèse valide jusqu'à *P* le nœud initial (Problème). L'explication du problème examiné dans cet article est le chemin produit par le KGTK sous la forme $p0-0-0$, $p0-1-1$, $p0-2-2$,

Block 8 – Sortie du chemin existant dans le graphe exploratoire d'hypothèses multicouches (requête dans le bloc 7)

node1	label	node2	id
Problem		h11	problem--0
Problem		h12	problem--1
Problem		h13	problem--2
h12		h31	h12--3
h13		h21	h13--4
h21		h32	h21--5
p0	0	e3	p0-0-0
p0	1	e4	p0-1-1
p0	2	e5	p0-2-2

qui correspond au *Problème - H13 – H21 – H32*.

3.3.2 Calcul du doute

Basé sur les valeurs de possibilité linguistique, il est possible de calculer le doute global d'un graphe exploratoire d'hypothèses multicouches (MHEG) ou celui pour chaque itération [11]. Ce doute est calculé comme suit :

- Le doute à une itération est donné par :
Le doute maximal sur toutes les hypothèses à l'itération donnée.
Par exemple le doute à l'itération *II* est $max_doubt\{$ "presque certainement vrai", "très vraisemblable", "très vraisemblable" $\} =$ "très vraisemblable",
ce qui signifie que cette itération "très vraisemblable" d'être vraie.
- Le doute sur le chemin valide d'un graphe exploratoire d'hypothèses multicouches (MHEG) est :
Le doute d'un chemin valide est égal au doute minimum des hypothèses appartenant à ce chemin. D'après l'exemple illustré dans cette étude, le doute de l'une des explications possibles (*H31*, *H21*, *H32*) est égal à :
 $min_doubt\{$ "Très probable", "Probablement vrai", "Certainement vrai" $\} =$ "Très probable".
Cette valeur correspond au doute de cette explication possible du problème évalué.

4 Conclusion

Cette étude présente des graphes exploratoires d'hypothèses (HEG) pour les processus d'expertise et comment les experts les construisent en collaboration en suivant le cycle de raisonnement du processus d'expertise des hypothèses. Il répond au besoin d'une représentation formelle plus expressive de ces HEG et propose une représentation graphique multicouche sur la structure existante. La représentation multicouche proposée est construite à partir d'un cycle de raisonnement de processus d'expertise d'hypothèse adapté. Un cas illustratif montre comment fonctionne l'approche proposée et comment les connaissances sont récupérées à l'aide du langage de requête kyper à partir de la boîte à outils du graphe de connaissances.

À l'avenir, nous avons l'intention d'étudier la construction d'un logiciel de processus d'expertise avec une interface d'interaction humaine intuitive entre les experts et le graphe multicouche proposé.

Références

[1] Guillermo Blanco and Anália Lourenço. A multi-layered graph-based framework to explore behavioural phenomena in social media conversations. *International Journal of Medical Informatics*, 179 :105236, 2023.

[2] Patrice Buche, Julien Couteaux, Julien Cufi, Sébastien Destercke, and Alrick Oudot. Integrating collective know-how for multicriteria decision support in agri-

- food chains—application to cheesemaking. *Frontiers in Artificial Intelligence*, 6 :1145007, 2023.
- [3] Patrice Buche, Bernard Cuq, Jerome Fortin, and Clément Sipieter. Expertise-based decision support for managing food quality in agri-food companies. *Computers and Electronics in Agriculture*, 163 :104843, 2019.
- [4] Patrick J Hurley. *A concise introduction to logic*. Nelson Education, 2014.
- [5] Filip Iliovski, Daniel Garijo, Hans Chalupsky, Naren Teja Divvala, Yixiang Yao, Craig Rogers, Ronpeng Li, Jun Liu, Amandeep Singh, Daniel Schwabe, and Pedro Szekely. KGTK : A toolkit for large knowledge graph manipulation and analysis. In *International Semantic Web Conference*, pages 278–293. Springer, 2020.
- [6] Xia Jing, James J Cimino, Vimla L Patel, Yuchun Zhou, Jay H Shubrook, Sonsoles De Lacalle, Brooke N Draghi, Mytchell A Ernst, Aneesa Weaver, Shriram Sekar, et al. Data-driven hypothesis generation among junior clinical researchers : A comparison of a secondary data analysis with visualization (viads) and other tools. *medRxiv*, pages 2023–05, 2023.
- [7] Pierre Marquis, Odile Papini, and Henri Prade. *Représentation des connaissances et formalisation des raisonnements*. Cepadues Editions, 2014.
- [8] Paul Peyrouy. La norme nf x 50-110 «qualité en expertise» : situation actuelle et perspectives. *Revue Experts*, 90 :36–39, 2010.
- [9] Pierre Siegel, Andrei Doncescu, Vincent Risch, and Sylvain Sené. Vers une représentation des systèmes dynamiques booléens en logique des hypothèses. In *Journées d’Intelligence Artificielle Fondamentale 2017*, page 14p, 2017.
- [10] Pierre Siegel, Vincent Risch, Syvain Sené, and Andrei Doncescu. Logique modale des hypothèses, systèmes dynamiques booléens et réseaux de gènes. *JIAF*, page 43.
- [11] Serge Sonfack Souchio. *Modélisation et exploitation des connaissances pour les processus d’expertise collaborative*. PhD thesis, 2022.
- [12] Serge SONFACK SOUNCHIO and Laurent GENESTE. A hypotheses-driven framework for human-machine expertise process. Available at SSRN 3982968.
- [13] Serge Sonfack Souchio, Laurent Geneste, and Bernard Kamsu Foguem. Modeling and sharing knowledge in expertise processes. *Enterprise Interoperability IX : Interoperability in the Era of Artificial Intelligence*, 10 :241, 2023.

Session 4 : Posters et démonstrations

Usages de l'IA dans l'enseignement supérieur : Enjeux et perspectives

A. Abbadie¹, A. Lammoglia², M. Servajean^{3,4}

¹ Université Paul Valéry Montpellier 3, UR LHUMAIN

² Université Paul Valéry Montpellier 3, UR LAGAM

³ Université Paul Valéry Montpellier 3, Université de Montpellier, CNRS, AMIS

⁴ Université de Montpellier, CNRS, UMR LIRMM

axelle.abbadie@univ-montp3.fr ; adrien.lammoglia@univ-montp3.fr ; maximilien.servajean@lirmm.fr

1 Introduction

Cette communication vise à présenter un exemple de contenu de formation innovant entièrement dédié à la compréhension des outils d'IA et à la proposition de bonnes pratiques d'usage dans le milieu universitaire que ce soit pour les enseignants ou les étudiants. Le contenu initial, nommé Alvéole, a été produit lors de la conception d'un module de formation en ligne asynchrone totalement inédit dédié à l'IA. Nous parlons ici des Briques en Humanités numériques, un des chantiers du programme ANR Nexus, porté par l'Université Paul Valéry Montpellier 3.

Au-delà d'un simple résumé de l'Alvéole "L'IA dans l'enseignement et la pédagogie" notre objectif est ici de communiquer sur les enjeux et les craintes liés au développement d'outils puissants à base d'IA. Notre souhait est d'une part de démystifier l'IA, d'en expliquer les enjeux et les limites et de montrer comment nous appréhendons ces innovations technologiques en termes de pédagogie et de recherche dans le milieu universitaire.

1.1 IA partout, compétence d'avenir

Aujourd'hui les IA génératives ont véritablement trouvé leur place dans toutes les strates professionnelles. Bien qu'il soit nécessaire d'en critiquer certaines applications, il nous semblait opportun de prendre un peu de hauteur face aux débats, et proposer une véritable compétence d'usage de ces nouveaux outils.

En effet, nombre d'études vantent les potentialités offertes par l'IA générative pour les entreprises, qu'il s'agisse d'aide à la prise de décision [1], [2], [3], [4] ou de productivité [3], [5], [6]. Il est évident que de nouvelles opportunités d'emploi sont créées et nécessitent des compétences techniques à acquérir avec des formations spécialistes du traitement des données et de l'apprentissage [7]. Cependant, il semble indispensable de former les futurs travailleurs à des usages que l'on peut qualifier de profanes [8], [9].

1.2 Construire une stratégie d'usage est bénéfique

Nombre d'initiatives de formation non-spécialistes ont vu le jour dans le milieu universitaire. On observe parfois une

meilleure compréhension de l'IA générative, et les étudiants semblent développer une "approche plus détaillée" de leur travail [10], ce que nous comprenons comme la construction de stratégie d'usage. Une autre initiative, dans une formation à la programmation qui présente l'IA comme un assistant, expose pour les étudiants une meilleure compréhension de la pensée "informatique" et, de manière plus étonnante, une amélioration de leur sentiment d'efficacité personnelle et de leur motivation [11].

1.3 Un module dédié à la pédagogie au sein de la Brique IA

Les Briques en Humanités numériques¹ sont des modules de formation, dont la granularité est la suivante : une Brique est composée de 8 chapitres, eux-mêmes constitués de 4 Alvéoles (cf. Figure 1). Une Alvéole équivaut à une heure de formation. L'étudiant travaille en ligne de manière asynchrone et est évalué en fin de semestre par un QCM sur l'ensemble de la brique.

Dans le dernier chapitre, nommé "Les implications sociales et environnementales de l'IA", nous avons conçu l'Alvéole "L'IA dans l'enseignement et la pédagogie"² qui est accessible en libre accès. L'objectif était de proposer une revue des usages pédagogiques à la fois pour les pédagogues (enseignants et ingénieurs pédagogiques) et les étudiants en contexte académique.

A partir de ce module nous cherchons plus particulièrement à sensibiliser aux innovations pédagogiques initiées en milieu universitaire, à initier la construction d'une stratégie d'usage de l'IA générative responsable et proposer des cas pratiques reproductibles pour l'étudiant. Une problématique majeure aujourd'hui puisque nous n'avons pas les réponses : nous mettons l'accent sur la distinction entre une utilisation frauduleuse, et une utilisation responsable utile à l'étudiant.

¹ <https://moodle-humanitesnumeriques.univ-montp3.fr>

² Disponible sur inscription auprès du lien précédent.

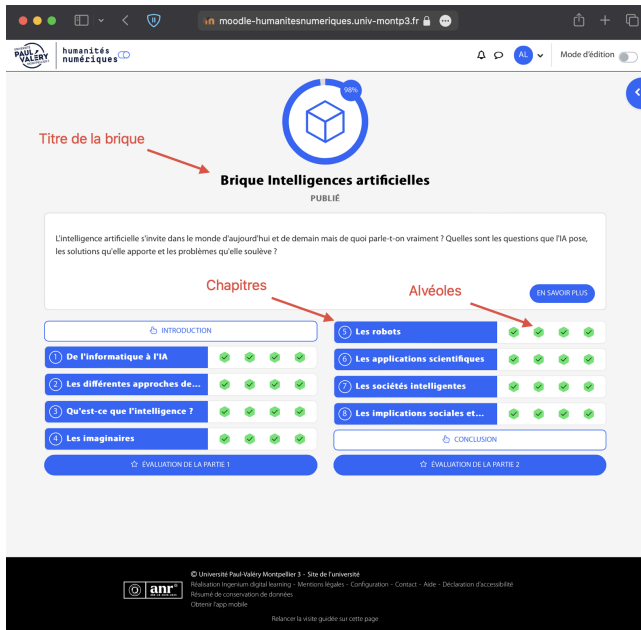


Figure 1 - Structure de la brique Intelligence Artificielle sur la plateforme des humanités numériques

2 De nouveaux outils pour les enseignants et ingénieurs pédagogiques

Nous vous proposons une rapide revue des contenus de cette alvéole, avec un focus sur l'intention pédagogique sous-jacent à chaque item. Il transparait chaque fois l'intention de présenter un usage institutionnel, pour promouvoir un usage personnel.

2.1 Cas d'écoles

Le cours propose trois exemples : La plateforme OptiPerf du Cours Gallien (1), le projet TIR-IA porté par l'Université de Laval et l'Université de Montréal (2) et l'initiative Auth+ du prestataire de service Sikanai partenaire de nombreuses universités (3).

Le premier exemple (Optiperf) présente l'IAG comme un compagnon de révision. Le deuxième (TIR-IA) est une ouverture sur les potentialités d'une IA plus complexe pour un tutorat automatisé. Le troisième (Auth+) montre d'autres façons de déceler la fraude dans une production pour un examen avec un nouvel examen sur-mesure généré par l'IA. Ces trois exemples sont une nouvelle opportunité de montrer aux étudiants que les universités et écoles se saisissent de ces technologies pour proposer une pédagogie différenciée, tant sur la préparation à un examen, que sur l'accompagnement des étudiants ou dans la prévention du plagiat et de la fraude.

2.2 Conception assistée

Un item complet est consacré aux différentes étapes de production d'un cours, de la définition des compétences à transmettre, passant par le scénario pédagogique et la production des supports. Chaque étape est détaillée comme pouvant être assistée par l'IA générative, et dans quelles

limites.

Nous portons l'attention sur la régulation de l'humain dans le processus de production du cours, et plus spécifiquement sur la prévention du *ghost writing*. Ce dernier est à la source d'une préoccupation importante des chercheurs quant à la perte de créativité dans les universités et l'intégrité académique [12], [13].

2.3 Pédagogie différenciée, tutorat et auto-formation

Deux items se succèdent pour présenter ces trois thématiques. Afin que l'étudiant comprenne comment l'IAG peut l'aider à monitorer son apprentissage, et proposer un support orienté sur les contenus. Des applications complexes, notamment de cognition augmentée, sont présentées afin d'être le plus exhaustif possible. La trame de fond expose l'IAG comme un assistant pour toutes les étapes d'apprentissage : s'initier à un jargon, organiser ses tâches d'apprentissage, traiter en profondeur un sujet, ou simplement améliorer l'accessibilité pédagogique. À cette fin, les travaux de Psyché & Ruer et ceux de Hou & Fidopiastis [14], [15] sur l'apprentissage adaptatif intelligent sont présentés, ainsi que ceux de Rospigliosi [16] sur les environnements interactifs d'apprentissage. Ceci, dans la perspective de leur offrir une méthodologie pour s'appropriier les applications de ces travaux.

3 Des outils attractifs pour les étudiants

3.1 Mauvaises et bonnes pratiques

Il aurait été irresponsable de présenter l'utilisation de l'IA générative sans en définir un cadre d'usage éthique et éclairé. Les mésusages, triches et fraudes sont explicités pour que l'étudiant puisse discerner les limites à ne pas franchir. Un item est également dédié au plagiat et à la propriété intellectuelle, un autre aux hallucinations de l'IA générative. La vigilance est toujours convoquée contre le ghost-writing.

Aujourd'hui, la question réglementaire des usages de l'IAG n'est pas encore résolue. D'abord parce que la propriété intellectuelle des produits générés ne saurait être définitivement attribuée. Ensuite parce que tous les usages ne sont pas détectables, la détection révèle simplement une "probabilité" d'utilisation de l'IAG et l'appréciation de la détection est à la discrétion de l'évaluateur. Pour cette raison, les conditions générales d'utilisation des détecteurs, comme Compilatio³, affirment que la seule détection ne peut suffire à sanctionner un étudiant suspecté.

³ <https://www.compilatio.net/magister-plus>

Les indicateurs visibles dans le rapport d'analyse Magister+

Aujourd'hui, l'accès au détecteur d'IA Compilatio se fait via l'[offre Magister+](#) qui fournit un **rapport combiné de détection de textes suspects**.

Plusieurs indicateurs sont mis en avant pour **faciliter la correction** d'un document et **s'assurer de son authenticité** :

- Un **pourcentage global de textes suspects** composé des similitudes identifiées, du score de rédaction par IA et des zones avec tentative d'altération de texte.

84%

- **Le détail du pourcentage global de textes suspects** :
 - score de similitudes (plagiat potentiel),
 - score de rédaction par IA (ghostwriting potentiel),
 - score de "langue non-reconnu" (tentative potentielle d'altération de texte).

Provenance des sources détectées : contenus générés par une IA, sources web, vos documents déjà analysés, documents d'un utilisateur de votre établissement ou d'un établissement partenaire.

Visualisation des passages directement dans le document.

Quelle preuve visuelle de l'utilisation de l'IA pour générer du texte ?

Si pour la détection de similitudes le logiciel Compilatio Magister+ met en avant la preuve de ressemblance grâce à la **vue face-à-face**, le logiciel **ne peut pas apporter de preuve visuelle à la détection de passages générés par IA**.

Figure 2 - Capture d'un article du centre d'aide de Compilatio sur l'objectivité de la détection IA⁴

Une fois les mauvaises pratiques évoquées, plusieurs items sont consacrés aux bonnes pratiques, avec une orientation en faveur de l'élaboration d'une stratégie d'usage. Les derniers items sont une ouverture à l'utilisation des outils qui intègrent IA générative et recherche documentaire, et une invitation à proposer sa stratégie d'utilisation à un enseignant si la finalité de l'usage est un devoir.

3.2 Initiation au prompt engineering pour un usage profane mais éclairé

Tous les étudiants inscrits sont évalués dans leur cursus principal par des productions académiques. Chacun craint un usage frauduleux de l'IA, alors nous proposons des utilisations pratiques concrètes de l'IA générative tout au long de la production d'un contenu (en l'occurrence, un article de blog). De la même manière qu'il a été présenté plus tôt un ensemble de stratégies d'usage à destination des concepteurs

4

<https://support.compilatio.net/hc/fr/articles/17432934803345-Le-d%C3%A9tecteur-d-IA-Compilatio-fournit-il-une-preuve-objective-de-l-utilisation-d-IA-g%C3%A9n%C3%A9rative-de-texte-dans-un-document>

pédagogiques, il est maintenant proposé des prompts pour chaque étape de production de l'article de blog (cf. F. Le fil d'ariane de cet item est de considérer l'IA comme un assistant à la prise de décisions, l'idéation, la structuration, la production et la correction d'une production écrite.

Articuler son projet
avec une intelligence artificielle générative

Analyse du bccdm | **Ideation** | Critique des idées | Développer ses idées | Évaluer la progression | Produire | Réviser

Prompt Engineering

Si vous avez suivi la Brique d'Introduction aux Humanités Numériques, vous avez été initié(e) aux ambiguïtés qui peuvent perturber la compréhension de l'intelligence artificielle.

Lorsque l'on s'attelle à l'ingénierie de commande, on tente de réduire les ambiguïtés et de préciser le comportement attendu par l'intelligence artificielle.

Ce qui revient à dire à l'IA pourquoi vous la sollicitez, ce qu'elle doit vous donner, comment et en jouant quel rôle.

Avant de commencer... suite ▶

Figure 2 - Extrait de la ressource "Articuler son projet"

3.3 Veille et outils recommandés

Nous recommandons aux étudiants des outils d'intelligence artificielle générative qui proposent un résultat sourcé. En effet, l'une des plus grandes critiques quant à l'usage de l'IA générative, et à raison, est la confusion sur les données d'entraînement utilisées pour générer un contenu. Nos recommandations sont donc sur Perplexity et Consensus, des outils orientés vers la curation de contenus. Ces deux outils citent leurs sources et invitent l'utilisateur à les explorer. L'étudiant découvre ensuite d'autres outils qui utilisent l'IA générative pour d'autres applications que générer un écrit, tels que Elicit ou Rabbit Research. Ainsi, l'IA est utilisée pour analyser les contenus et répondre à une requête ou révéler des similarités.

4 Perspectives

Cette Alvéole expérimentale a bien sûr l'ambition d'initier les étudiants à des pratiques plus éclairées de l'IA. Cependant, nous espérons qu'il s'agit d'une première pierre à un édifice plus grand. Cet espoir a bien sûr son lot d'enjeux et d'écueils.

4.1 Démocratisation des bonnes pratiques

Il s'agit là de l'ambition la plus grande de cette Alvéole : voir les bonnes pratiques en matière d'utilisation de l'IA remplacer les mauvaises. Si les initiatives expérimentales sur l'IA générative en formation ont montré de bons résultats, nous pouvons espérer que les étudiants prennent goût à une utilisation consciente et responsable de ce nouvel outil. Ces futurs travailleurs pourraient bénéficier de compétences pratiques à valoriser pour leur insertion professionnelle, et plus humblement, être en capacité de solliciter l'IA générative avec un but raisonnable et adapté.

4.2 Débattre de l'inclusion de l'IA à l'Université

À la suite de la publication de cette Brique, de nombreux débats se sont ouverts dans les emails entre les enseignants

concepteurs de la Brique IA et d'autres professionnels de l'Université. Il est probable, et souhaité, que cette Alvéole ait des conséquences sur le comportement des étudiants et soulève des questions chez les enseignants sur le campus et ailleurs. Ce débat est nécessaire pour délimiter les pratiques autorisées, recommandées, déconseillées ou interdites. Nous pensons que cette Alvéole, orientée vers les usages, permet de questionner les opinions clivées sur l'IA en contexte académique. Les échanges emails qui se sont ouverts depuis ont montré que pour appréhender une stratégie d'usage, il est nécessaire de comprendre le fonctionnement même de l'IA générative, et que les simples détecteurs d'IA ne suffisent pas à remettre en question la validité d'une production étudiante assistée par un modèle de langage. Par la même occasion un atelier ouvert à tous a été organisé pour débattre de l'utilisation des outils à base d'IA au sein des parcours universitaires.

4.3 Innovation et usages : lutte contre l'obsolescence

À peine quelques mois après la publication de cette Alvéole, le modèle de langage d'OpenAI a déjà publié une nouvelle version, et certaines recommandations en matière de prompt-engineering nécessitent des ajustements. Dans le même temps, de nouveaux outils, tels que SciSpace, ont vu le jour et offrent de nouvelles perspectives qu'il serait intéressant de présenter aux étudiants. Ces deux exemples illustrent le grand écueil d'un module de formation orienté vers les usages sur des innovations critiques : les usages et technologies évoluent vite, en comparaison aux contenus théoriques, les enseignements pratiques nécessitent une actualisation plus fréquente. Nous recommandons des sources de veille aux étudiants, afin que chacun puisse s'informer en l'attente de la mise à jour annuelle du module. Plus largement, la question de la mise à jour des contenus de formation dans un monde où l'utilisation des outils à base d'IA serait généralisée est un débat majeur.

5 Références

- [1] Q. Bi, « Analysis of the Application of Generative AI in Business Management », *Adv. Econ. Manag. Res.*, vol. 6, n° 1, Art. n° 1, juin 2023, doi: 10.56028/aemr.6.1.36.2023.
- [2] E. L. Chuma et G. G. De Oliveira, « Generative AI for Business Decision-Making: A Case of ChatGPT », *Manag. Sci. Bus. Decis.*, vol. 3, n° 1, p. 5-11, juill. 2023, doi: 10.52812/msbd.63.
- [3] Dr. J. J. Jeya, Dr. S. R. Ratna, et Dr. G. Gangadevi, « Generative AI Boosts Business Productivity », in *Decision Strategies and Artificial Intelligence Navigating the Business Landscape*, 1^{re} éd., San International Scientific Publications, 2023. doi: 10.59646/edbook3/009.
- [4] N. A. Parikh, « Empowering Business Transformation: The Positive Impact and Ethical Considerations of Generative AI in Software Product Management -- A Systematic Literature Review ». arXiv, 5 juin 2023. doi: 10.48550/arXiv.2306.04605.
- [5] E. Brynjolfsson, D. Li, et L. Raymond, « Generative AI at Work ». arXiv, 23 avril 2023. doi: 10.48550/arXiv.2304.11771.
- [6] S. Noy et W. Zhang, « Experimental evidence on the productivity effects of generative artificial intelligence », *Science*, vol. 381, n° 6654, p. 187-192, juill. 2023, doi: 10.1126/science.adh2586.
- [7] R. Tailor, S. Jain, et A. Kamble, « A Review paper on the Impact of Artificial Intelligence on the Job Market », *Int. J. Adv. Res. Sci. Commun. Technol.*, p. 68-73, mai 2023, doi: 10.48175/IJARST-10724.
- [8] S. Ali, D. DiPaola, R. Williams, P. Ravi, et C. Breazeal, « Constructing Dreams using Generative AI », 2023, doi: 10.48550/ARXIV.2305.12013.
- [9] A. Mills, M. Bali, et L. Eaton, « How do we respond to generative AI in education? Open educational practices give us a framework for an ongoing process », *J. Appl. Learn. Teach.*, vol. 6, n° 1, Art. n° 1, juin 2023, doi: 10.37074/jalt.2023.6.1.34.
- [10] P. Cotroneo et J. Hutson, « Generative AI tools in art education: Exploring prompt engineering and iterative processes for enhanced creativity », *Metaverse*, vol. 4, n° 1, p. 14, juin 2023, doi: 10.54517/m.v4i1.2164.
- [11] R. Yilmaz et F. G. Karaoglan Yilmaz, « The effect of generative artificial intelligence (AI)-based tool use on students' computational thinking skills, programming self-efficacy and motivation », *Comput. Educ. Artif. Intell.*, vol. 4, p. 100147, 2023, doi: 10.1016/j.caeai.2023.100147.
- [12] T. Rasul *et al.*, « The role of ChatGPT in higher education: Benefits, challenges, and future research directions », *J. Appl. Learn. Teach.*, vol. 6, n° 1, Art. n° 1, mai 2023, doi: 10.37074/jalt.2023.6.1.29.
- [13] A. Smolansky, A. Cram, C. Radulescu, S. Zeivots, E. Huber, et R. F. Kizilcec, « Educator and Student Perspectives on the Impact of Generative AI on Assessments in Higher Education », in *Proceedings of the Tenth ACM Conference on Learning @ Scale*, Copenhagen Denmark: ACM, juill. 2023, p. 378-382. doi: 10.1145/3573051.3596191.
- [14] M. Hou et C. Fidopiastis, « A generic framework of intelligent adaptive learning systems: from learning effectiveness to training transfer », *Theor. Issues Ergon. Sci.*, vol. 18, n° 2, p. 167-183, mars 2017, doi: 10.1080/1463922X.2016.1166405.
- [15] V. Psyché et P. Ruer, « L'apprentissage adaptatif intelligent », *Rev. Tableau*, vol. 8, n° 4, 2019.
- [16] P. A. Rospigliosi, *Interactive learning environments: all is change-as usual*, vol. 31, n° 9. Taylor & Francis, 2023, p. 5421-5423.

Ce travail a bénéficié d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre du Programme d'Investissements d'Avenir portant la référence (ANR-18-NCUN-0025).

Introduire l'IA dans la lutte contre la fraude : Comment choisir et convaincre ?

F. Dama¹, R. Sleiman¹, S. Bellart¹

¹ Centre de Recherche et d'Innovation, Talan France

fatoumata.dama@talan.com, rita.sleiman@talan.com, steve.bellart@talan.com

Résumé

La lutte contre le blanchiment constitue une priorité pour les institutions financières et les conduit à exercer une surveillance permanente. Face aux limites des outils traditionnels, les algorithmes d'Intelligence Artificielle (IA) représentent une opportunité en permettant la conception de dispositifs performants et flexibles. Ces travaux réalisent un benchmark des modèles d'IA (ML/DL) de détection de fraudes financières et discutent de leur intégration en pratique.

Mots-clés

Blanchiment d'argent, Machine Learning, Deep Learning, XAI (Intelligence Artificielle eXplicable), Données financières, Augmentation de données

Abstract

Combating money laundering is a priority for financial institutions, leading them to implement continuous monitoring. Given the limitations of traditional tools, Artificial Intelligence (AI) algorithms represent an opportunity by enabling the design of effective and flexible systems. This work benchmarks AI (ML/DL) models for financial fraud detection, and discusses their practical integration.

Keywords

Money Laundering, Machine Learning, Deep Learning, XAI (Explainable Artificial Intelligence), Financial Data, Data Augmentation

1 Introduction

La lutte contre le blanchiment de capitaux et le financement du terrorisme (LCB-FT) représente un enjeu majeur pour la stabilité financière et la sécurité internationale. Il est estimé que 2 à 5% du PIB mondial est compromis par des activités de blanchiment d'argent chaque année, mettant en lumière l'urgence et la gravité de ce fléau, comme le montre certains modèles d'estimations [7]. Dans ce contexte, un cadre juridique international, incluant les recommandations du Groupe d'action financière (GAFI), a été établi pour contraindre les institutions financières à adopter des mesures de vigilance renforcée.

L'une de ces mesures essentielles est la norme *Know Your Customer* (KYC) [14], qui oblige les institutions financières

à vérifier l'identité de leurs clients et à évaluer et surveiller en continu les risques associés à ces derniers. La conformité aux exigences KYC est cruciale pour prévenir le blanchiment d'argent, le financement du terrorisme et d'autres formes de fraude financière.

Toutefois, les systèmes LCB-FT actuels, principalement fondés sur des règles prédéfinies, se confrontent à des limites notables. Leur rigidité et leur incapacité à s'ajuster aux tactiques de blanchiment en constante évolution se traduisent par un taux élevé de faux positifs, nuisant à l'efficacité des processus de détection. Dans ce cadre, l'intelligence artificielle (IA) se présente comme une solution prometteuse, offrant des capacités de détection améliorées et adaptatives grâce à des algorithmes avancés d'apprentissage automatique et de *deep learning*. Néanmoins, leur adoption pour la détection de fraude se confronte à plusieurs défis, notamment la nécessité de données financières souvent restreintes et déséquilibrées, et la complexité de comprendre le fonctionnement souvent opaque des modèles d'IA les plus performants.

Notre étude vise à évaluer et comparer diverses méthodes d'IA dans le domaine du LCB-FT, en analysant leur performance sur des jeux de données réels et synthétiques. Nous évaluons la capacité de détection des modèles et leur aptitude à réduire les faux positifs qui représentent un coût opérationnel important pour les établissements financiers. En outre, l'explicabilité intrinsèque des modèles est examinée et discutée pour assurer la transparence et renforcer la confiance des utilisateurs, un aspect crucial dans l'adoption de systèmes d'IA dans des domaines aussi sensibles que la LCB-FT.

La structure de cet article comprend une revue des travaux antérieurs pertinents, la présentation de notre méthodologie et de nos résultats, suivie d'une discussion sur les implications et les perspectives futures de notre recherche.

2 Travaux antérieurs

Au cours de la dernière décennie, plusieurs travaux de recherche se sont intéressés à l'utilisation de l'IA pour concevoir de nouveaux outils de détection de fraudes financières dont le blanchiment de capitaux [11][4]. Une analyse de la littérature permet d'identifier deux grands groupes de modèles : les modèles supervisés entraînés sur des données labélisées (normale/suspecte) et les modèles non-supervisés

qui cherchent à séparer les données en différents groupes homogènes [4].

Dans cette étude nous nous intéressons aux modèles supervisés qui sont entraînés à détecter les *patterns* indicatifs de blanchiment spécifiés dans la base d'apprentissage. L'application de ces modèles dans la LCB-FT se divise principalement en deux grandes classes, à savoir les modèles de *Machine Learning* traditionnels (ML) et les modèles de *Deep Learning* (DL). Les modèles classiques de ML comprennent des algorithmes comme les arbres de décision [15], les modèles ensemble par Bagging (Random Forest, Extra Trees) [17], les arbres boostés (XGB, LGBM, CatBoost) [1], et les modèles probabilistes (Logistic Regression, Naive Bayes) [12]. De l'autre côté, les modèles de DL incluent des modèles comme les Perceptrons multicouches (MLP), les Auto-Encodeurs, et les graphes de réseaux de neurones (GNN, GCN)[16, 3].

D'après les études recensées, les modèles reposant sur des arbres sont largement utilisés en comparaison avec les modèles de DL. Ceci s'explique par plusieurs facteurs : d'une part, ces modèles génèrent des règles explicites utilisées pour effectuer les prédictions (normale/suspecte); d'autre part, les modèles de DL sont perçus comme des boîtes noires, ce qui complique l'interprétation de leurs prédictions, un défi significatif dans un domaine aussi critique.

Malgré l'existence de plusieurs études abordant l'apport de l'IA à la LCB-FT, ces dernières présentent plusieurs limites qui nécessitent d'être prises en considération dans des prochaines études. En effet, la majorité des études se concentrent sur l'utilisation de types de modèles bien déterminés, ce qui limite la comparaison et l'évaluation des performances entre les modèles cités. En plus, les conditions expérimentales souvent hétérogènes d'une étude à l'autre compliquent la généralisation des résultats et l'évaluation fiable des différentes approches. Ceci souligne l'importance de recherches plus approfondies qui explorent une plus grande diversité d'algorithmes avec une standardisation des protocoles expérimentaux.

Ainsi, bien que l'étude de l'usage de modèles d'apprentissage automatique dans la lutte contre la fraude a déjà été traité dans la littérature [9] en y montrant notamment les défis à relever pour avoir des modèles efficaces, notre étude vise à compléter ces travaux, en proposant une comparaison entre 17 modèles d'IA, tout en ayant un regard sur les difficultés vis-à-vis de leur adoption. Nos travaux intègrent donc également l'aspect explicabilité peu discuté dans la littérature, mais fondamental pour des raisons de conformité, de transparence et pour l'approbation de ces approches dans un secteur aussi critique.

3 Expérimentations

3.1 Modèles et Datasets

Dans nos expérimentations, nous avons considéré **17 modèles** de détection de fraude dont 13 modèles de *Machine Learning* et 4 modèles *Deep Learning*.

Modèles ML. Arbre de décision, les modèles ensemble par Bagging (Random Forest, BaggingClassifier et Extra

Trees), les arbres boostés (AdaBoost, Logitboost, XGBoost, LightGBM et CatBoost), les modèles probabilistes (Naive Bayes et Logistic Regression), le modèle Support Vector Machine (SVM) et le modèle des K plus proches voisins (KNN).

Modèles DL. Les modèles Perceptron et Perceptron multicouches (MLP) avec 1, 2 ou 3 couches cachées.

Les modèles ont été évalués sur 5 jeux de données réelles et synthétiques, présentés dans le tableau 1. Les données sont des transactions bancaires (transferts, paiements, dépôts, ...) labélisées normales (la classe 0) ou suspectes (la classe 1). Remarquons la faible proportions de transactions suspectes dans les 5 jeux de données. Il est bien connu que le déséquilibre entre les classes peut être source de biais dans les modèles. Afin d'éviter un tel biais, la méthode d'échantillonnage SMOTE (*Synthetic Minority Over-sampling Technique*) [6] a été utilisée pour équilibrer les données en générant des transactions suspectes supplémentaires grâce à une technique d'interpolation linéaire.

Dataset	n_{feat}	n_{obs}	P_{fraude}
CreditCard	30	284 807	0.2%
Ethereum	48	9 840	22%
Bitcoin	167	203 796	10%
IBM ALMSim	8	250 000	0.7%
Mobile Money	11	250 000	3%

TABLE 1 – Description des datasets. De gauche à droite : le nombre de features, d'observations et le ratio de fraudes (<https://www.kaggle.com/datasets>). Les 3 premiers datasets sont composés de données réelles et les 2 derniers sont composés de données synthétiques.

3.2 Protocole expérimental

Les 17 modèles considérés dans nos expérimentations ont été évalués sur les 5 datasets précédemment présentés. Chaque dataset a été découpé en deux parties : un jeu d'entraînement (80%) et un jeu de test (20%). Les hyperparamètres des différents modèles (la profondeur des arbres, le taux d'apprentissage, les termes de régularisation, le nombre K de voisins proches, ...) ont été calibrés par la méthode de la validation croisée à 5 champs. Les modèles ensemble sont composés de 200 arbres de décision.

Les métriques d'évaluation considérées sont la **fiabilité** et l'**efficacité opérationnelle**, définies ci-dessous. La fiabilité décrit la capacité du modèle à identifier les transactions suspectes. Tandis que l'efficacité opérationnelle décrit sa capacité à réduire le volume des fausses alarmes (faux positifs) et réduire par la même occasion le coût opérationnel lié au traitement manuel des alarmes.

$$\text{Fiabilité (Recall)} = \frac{\text{Vrais Positifs}}{\text{Vrais Positifs} + \text{Faux Négatifs}} \quad (1)$$

$$\text{Efficacité (Precision)} = \frac{\text{Vrais Positifs}}{\text{Vrais Positifs} + \text{Faux Positifs}} \quad (2)$$

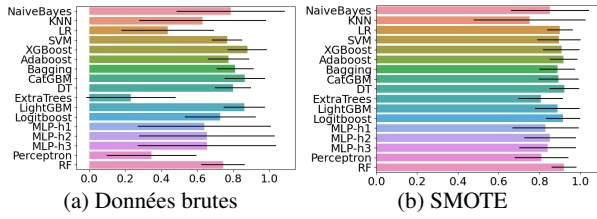


FIGURE 1 – Moyennes et écart-types des scores de fiabilité (calculés sur 5 datasets) obtenus par les modèles entraînés sur : (a) les données brutes; (b) les données équilibrées par la méthode SMOTE.

3.3 Résultats et Analyse

3.3.1 Données brutes versus SMOTE

La figure 1 présente les scores de fiabilité obtenus par les modèles testés avec les deux modalités : (a) modèles entraînés sur les données brutes (déséquilibrées); et (b) modèles entraînés sur les données équilibrées par la méthode SMOTE pour obtenir 50% de cas de fraudes.

Les résultats montrent une amélioration notable des capacités de détection des modèles (entre 3 – 57%) lorsque les données d’apprentissage sont équilibrées. Par ailleurs, les arbres de décision boostés (XGBoost, CatGBM et LightGBM) montrent une certaine robustesse au déséquilibre entre les classes avec une amélioration de 3%.

3.3.2 Cartographie des modèles

La figure 2 présente la moyenne des performances obtenues par les 17 modèles testés sur les 5 datasets considérés. Les jeux d’apprentissage ont été préalablement équilibrés par la méthode SMOTE.

Les résultats montrent une bonne capacité de détection des transactions suspectes pour l’ensemble des modèles testés avec un score de fiabilité supérieur à 75%. Par ailleurs, chaque modèle testé surpasse certains systèmes experts (basés sur des règles prédéfinies) avec un score d’efficacité opérationnelle au moins égal à 15% contre seulement 5% pour ces derniers. Les arbres de décision boostés obtiennent les meilleures performances (jusqu’à 90% de fiabilité et d’efficacité opérationnelle) et surpassent significativement les modèles de *Deep Learning*, corroborant l’avantage de ces modèles sur des données tabulaires [8]. Il reste à comparer ces résultats avec les systèmes experts actuels, mais cela est difficile, les informations à leur propos n’étant pas accessibles en raison de la sensibilité du secteur.

Du point de vue de l’explicabilité, les modèles à ensembles d’arbres offrent une meilleure interprétabilité que les modèles basés sur les réseaux de neurones, malgré la supériorité habituelle de ces derniers en termes de précision. Nos expériences démontrent une exception à cette règle générale, avec une performance supérieure des modèles ensemblistes, surtout ceux issus du *boosting*. Bien que les arbres de décision uniques soient intuitivement compréhensibles, la complexité des modèles ensemblistes en diminue la transparence. Néanmoins, l’avantage en précision des ensembles sur les arbres individuels est significatif. Nous sommes donc encouragés à continuer d’explorer l’explica-

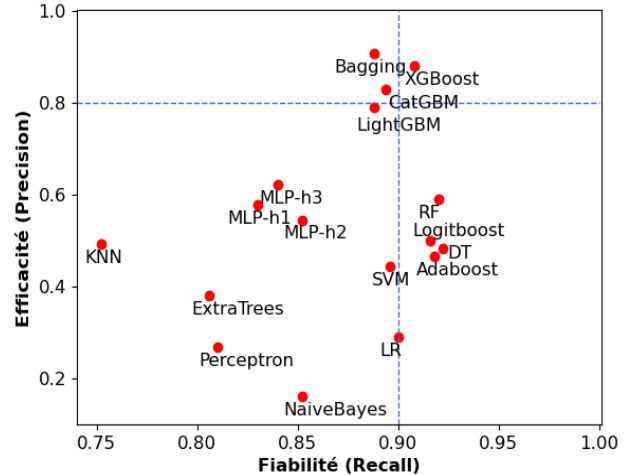


FIGURE 2 – Moyennes des performances (fiabilité et efficacité) obtenues par les 17 modèles testés sur les 5 datasets considérés.

bilité des modèles ensemblistes dans le cadre de la LCB-FT, en utilisant des outils d’explicabilité avancés pour faciliter leur acceptation par les experts.

4 Conclusion

Dans cette étude, nous avons évalué la performance de 17 modèles d’apprentissage automatique, incluant 13 modèles traditionnels de *Machine Learning* et 4 modèles de *Deep Learning*, testés sur 5 ensembles de données financières.

Les résultats obtenus montrent que les modèles testés possèdent une bonne capacité de détection des transactions suspectes (correspondant à du blanchiment d’argent). De plus, ces modèles se sont révélés plus efficaces que certains systèmes experts. Notamment, les arbres de décision boostés se sont distingués en atteignant jusqu’à 90% de fiabilité et d’efficacité opérationnelle. Ces modèles se révèlent être des outils précieux pour les institutions financières cherchant à améliorer leurs outils de détection des activités frauduleuses et optimiser le coût opérationnel de leurs systèmes de surveillance.

5 Perspectives et suite du projet

Le travail précédent s’est concentré sur l’évaluation de différents modèles d’apprentissage automatique dans la LCB-FT. D’après nos résultats, il apparaît que les modèles à ensembles d’arbres constituent une catégorie particulièrement prometteuse pour des analyses approfondies. Bien que nos recherches actuelles démontrent le potentiel de ces IA, l’intégration de ces technologies dans un domaine aussi sensible que la finance pose plusieurs défis. Il est crucial que ces modèles gagnent la confiance des experts du secteur et fassent l’objet d’une vérification rigoureuse avant leur déploiement. L’usage d’outils d’explicabilité pour l’IA peut jouer un rôle clé dans ce processus.

La recherche en explicabilité des modèles d’apprentissage automatique est récente et dynamique ayant déjà généré une multitude d’approches, chacune présentant ses avantages et ses inconvénients [10] [2]. Nous prévoyons pour la suite

d'exploiter certains de ces outils pour analyser les connaissances extraites par ces modèles. En particulier, nous envisageons d'utiliser des méthodes agnostiques au modèle (LIME, SHAP et Anchors), ainsi que des outils spécifiquement conçus pour les modèles à ensembles d'arbres (TreeSHAP et PyXAI).

L'objectif principal est de déterminer et d'examiner les règles ou *patterns* dans les données qui conduisent à la classification d'une action comme frauduleuse d'après les modèles d'IA. Nous proposons d'analyser ces règles à l'aide de méthodes statistiques, en examinant la longueur (le nombre d'assertions à respecter pour obéir à la règle) et la couverture (le nombre d'instances qui y obéissent) des règles, ainsi que leur contenu, en les comparant avec des connaissances établies et en sollicitant l'avis d'experts pour obtenir des évaluations constructives. Cette démarche vise à déterminer quel est la meilleure façon de les générer et si les règles calculées sont cohérentes ou si elles révèlent de nouveaux patterns de fraude qui sont plausibles mais auparavant non identifiés.

Finalement, nous soutenons que l'adoption généralisée des méthodes d'apprentissage automatique dépendra non seulement de leur performance statistique (bien que leur optimisation reste un défi important), mais aussi de la capacité à être comprises et contrôlées par les utilisateurs [13]. Les outils d'XAI améliorent la confiance en ces systèmes et nous pensons que développer des approches permettant aux utilisateurs d'interagir avec les systèmes d'IA et de les corriger en cas d'erreurs ou d'imprécisions est un enjeu important à leur mise en œuvre pratique. Nos travaux futurs se concentreront également sur l'intégration d'outils d'ajustement interactif [5], afin de renforcer l'efficacité et la fiabilité envers ces approches dans la LCB-FT.

Références

- [1] A Ahmed. Anti-money laundering recognition through the gradient boosting classifier. *Academy of Accounting and Financial Studies Journal*, 25(5), 2021.
- [2] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. Explainable artificial intelligence (xai) : What we know and what is left to attain trustworthy artificial intelligence. *Information fusion*, 99 :101805, 2023.
- [3] Zhiyuan Chen, Waleed Mahmoud Soliman, Amril Nazir, and Mohammad Shorfuzzaman. Variational autoencoders and wasserstein generative adversarial networks for improving the anti-money laundering process. *IEEE Access*, 9 :83762–83785, 2021.
- [4] Zhiyuan Chen, Le Dinh Van Khoa, Ee Na Teoh, Amril Nazir, Ettikan Kandasamy Karupiah, and Kim Sim Lam. Machine learning techniques for anti-money laundering (aml) solutions in suspicious transaction detection : a review. *Knowledge and Information Systems*, 57 :245–285, 2018.
- [5] Sylvie Coste-Marquis and Pierre Marquis. Rectifying binary classifiers. In *The 26th European Conference on Artificial Intelligence (ECAI'23)*. IOS Press, 2023.
- [6] Alberto Fernández, Salvador Garcia, Francisco Herrera, and Nitesh V Chawla. Smote for learning from imbalanced data : progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research*, 61 :863–905, 2018.
- [7] Joras Ferwerda, Alexander van Saase, Brigitte Unger, and Michael Getzner. Estimating money laundering flows with a gravity model-based simulation. *Scientific Reports*, 10(1) :18552, 2020.
- [8] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems*, 35 :507–520, 2022.
- [9] Prince Grover, Julia Xu, Justin Tittelfitz, Anqi Cheng, Zheng Li, Jakub Zablocki, Jianbo Liu, and Hao Zhou. Fraud dataset benchmark and applications, 2023.
- [10] Riccardo Guidotti, Anna Monreale, Salvatore Ruggeri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5) :1–42, 2018.
- [11] Dattatray Vishnu Kute, Biswajeet Pradhan, Nagesh Shukla, and Abdullah Alamri. Deep learning and explainable artificial intelligence techniques applied for detecting money laundering—a critical review. *IEEE access*, 9 :82300–82317, 2021.
- [12] Mark E Lokanan. Predicting money laundering using machine learning and artificial neural networks algorithms in banks. *Journal of Applied Security Research*, 19(1) :20–44, 2024.
- [13] Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. Human-in-the-loop machine learning : a state of the art. *Artificial Intelligence Review*, 56(4) :3005–3054, 2023.
- [14] Plaid. What is kyc? financial regulations to reduce fraud. <https://plaid.com>, 2022.
- [15] Omri Raiter. Applying supervised machine learning algorithms for fraud detection in anti-money laundering. *Journal of Modern Issues in Business Research*, 1(1) :14–26, 2021.
- [16] Mark Weber, Giacomo Domeniconi, Jie Chen, Daniel Karl I Weidele, Claudio Bellei, Tom Robinson, and Charles E Leiserson. Anti-money laundering in bitcoin : Experimenting with graph convolutional networks for financial forensics. *arXiv preprint arXiv :1908.02591*, 2019.
- [17] Wai Weng Lo, Gayan K Kulatilleke, Mohanad Sarhan, Siamak Layeghy, and Marius Portmann. Inspection-1 : Self-supervised gnn node embeddings for money laundering detection in bitcoin. *arXiv e-prints*, pages arXiv–2203, 2022.

