AfIA

Association française
pour l'Intelligence Artificielle

RJCIA

23es Rencontres des Jeunes Chercheurs en Intelligence Artificielle

PFIA 2025

Table des matières

Danai Symeonidou
Éditorial 5

Comité de programme 6

Session 1 7

Clément Contet, Umberto Grandi, Jérôme Mengin
Minimal supports for tournament winners 8

Sena Ozgunay, Louise Travé-Massuyès, Jean-Michel Loubes, Raul Sena Ferreira
Detecting anomalies using graph neural networks : a review 16

Houdhem Assoudi, Vincent Armant, Amira Mouakher, Emmanuel Roux, Jean-Christophe Desconnets, Victor N.Mose
Combining knowledge graphs and generative ai to explain the consequences of climatic events on the health of maasai herders 24

Célia Blondin, Joris Guérin, Kelly Iganaki, Guilherme Longo, Laure Berti-Equille
Hierarchical classification for automated image annotation of coral reef benthic structures 27

Session 2 33

Baptiste Darnala, Clément Jonquet, Florence Amardeilh, Konstantin Todorov
Prédiction des séquences culturales par apprentissage automatique : une comparaison entre grandes cultures et maraîchage diversifié 34

Jacques Aimé Razafindragy, Vincent Armant, Lalaina Fanomezantsoa, Jean-Christophe Desconnets, Eric Delaitre, Aimé Richard Hajalalaina, Hasina Rakotonirainy, Thérèse Libourel
Vers une approche hybride combinant intégration de données et IA pour la reconstitution d'un réseau d'adduction d'eau potable 42

Oussama Mechhour, Sandrine Auzoux, Clément Jonquet, Mathieu Roche
MAEVa : une approche hybride pour la mise en relation des variables expérimentales agroécologiques 45

Session 3 53

Hamid Ahaggach, Lylia Abrouk
Approches hybrides basées sur l'IA symbolique et l'apprentissage automatique : application au domaine automobile 54

Heni Lachouak, Youssef Mourchid, Yuehua Ding,Pierre-Grégoire Guinot
Apprentissage profond séquentiel pour la prédiction/anticipation de l'hypotension artérielle non-invasive par le signal de pléthysmographie 56

Ariel Guerra-Adames, Océane Dorémus, Marta Avalos-Fernandez, Vianney Jouhet, Emmanuel Lagarde, Cédric Gil-Jardiné
Génération de données de santé synthétiques : état de l'art en vue de la classification des mécanismes des traumatismes 59

Noama Adra, Lamine Bougueroua, Jean-Charles Huet, Faten Chakchouk
Graphe de connaissances pour la modélisation des projets de recherche équestre 62

Session 4 69

Farah Ben Slama, Frédéric Armetta
Grands modèles de langage et exécution d'algorithmes : application à une fonction arithmétique. 70

Hussam Ghanem, Samir Jabbar, Christophe Cruz
CoA-Text2OWL : *a chain-of-agents framework for ontology learning* 79

Lorenzo Ferlito, Ophélie Bugnon, Thomas Matagne, Baudouin Dafflon
Véhicules autonomes : simulation multi-agents pour explorer l'importance de la structuration de réseaux 86

Éditorial

23es Rencontres des Jeunes Chercheurs en Intelligence Artificielle

Les Rencontres des Jeunes Chercheurs en Intelligence Artificielle (RJCIA) se sont déroulées les deux premiers jours de l'édition 2025 de la Plate-Forme Intelligence Artificielle (PFIA), qui a eu lieu du 30 juin au 4 juillet 2025 à Dijon.

Les RJCIA sont destinées aux jeunes chercheurs en IA, doctorants ou titulaires d'un doctorat depuis moins d'un an. L'objectif de cette manifestation est double :

1. Permettre aux jeunes chercheurs préparant une thèse en Intelligence Artificielle, ou l'ayant soutenue depuis peu, de se rencontrer et de présenter leurs travaux, et d'ainsi former des contacts avec d'autres jeunes chercheurs et d'élargir leurs perspectives en échangeant avec des spécialistes d'autres domaines de l'intelligence artificielle, et
2. Former les jeunes chercheurs à la préparation d'un article, à sa révision pour tenir compte des observations du comité de programme, et à sa présentation devant un auditoire de spécialistes, leur permettant ainsi d'obtenir des retours de chercheurs de leur domaine ou de domaines connexes.

Pour cette édition de RJCIA nous avons eu l'honneur de recevoir Luis Galarraga, chercheur au centre de recherche IRISA/Inria de Rennes. Ses recherches se situent au carrefour de trois axes : la fouille de motifs, les graphes de connaissances et l'IA explicable. Luis Galarraga a donné une conférence invitée intitulée « Data- and Human-aware Explainable AI ».

Concernant les contributions scientifiques, 14 articles ont été acceptés et constituent le contenu de ces actes. Ces articles sont de deux types, 7 articles longs présentent des contributions originales dans les thèmes de la conférence et 7 articles courts qui présentent des premières pistes de travail dans un contexte clairement établi.

Danai Symeonidou

Comité de programme

Présidence

— Danai Symeonidou, INRAE Montpellier ;

Membres

— Lylia Abrouk - LIB, Dijon ;
— Vincent Armant - IRD, Montpellier ;
— Sandra Bringay - LIRMM, Montpellier ;
— Francesca Bugiotti - Université Paris-Saclay, LISN ;
— Davide Di Pierro - Université de Montpellier ;
— Luis Galárraga - Inria, Rennes ;
— Alexis Guyot - Université de Montpellier, LIRMM ;
— Nathalie Hernandez - IRIT ;
— Liliana Ibanescu - Université Paris-Saclay, AgroParisTech ;
— Pierre Larmande - IRD, Montpellier ;
— Melanie Munch - INRAE ;
— Pierre-Henri Paris - Université Paris-Saclay ;
— Nathalie Pernelle - LIPN, Université Sorbonne Paris Nord ;
— Joe Raad - Université Paris-Saclay, LISN ;
— Catherine Roussey - INRAE, Montpellier ;
— Cassia Trojahn - UT2J & IRIT ;
— Katerina Tzompanaki - CY Cergy Paris University ;
— Felipe Vargas-Rojas - IRD, Montpellier ;
— Nadia Yacoubi - Université Claude Bernard (Lyon 1) ;

Session 1

Minimal Supports for Tournament Winners

Clément Contet[1], Umberto Grandi[1], Jérôme Mengin[1]

[1] Université de Toulouse, IRIT

{clement.contet, umberto.grandi, jerome.mengin}@irit.fr

Résumé

Les tournois sont des modèles largement utilisés pour représenter des notions de dominance par paire entre des candidats, des alternatives ou des équipes. Les solutions de tournois associent un ensemble d'alternatives gagnantes à un tournoi. Les règles de Copeland et de Borda sur les tournois non pondérés et pondérés respectivement en sont deux examples classiques. Nous étudions le problème de l'identification des supports minimaux pour un gagnant de tournoi en trouvant des sous-tournois minimaux sur lesquels le gagnant du tournoi gagne indépendamment de l'achèvement du sous-tournoi, c'est-à-dire qu'il est un gagnant nécessaire du sous-tournoi. Dans cet article, nous présentons des algorithmes polynomiaux pour calculer les supports minimaux pour les gagnants de Copeland et de Borda des tournois, et nous fournissons des limites sur la taille des plus petits supports minimaux à la fois théoriquement et empiriquement.

Mots-clés

Choix Social Computationnel, Vote, Solutions de Tournoi, Gagnant Nécessaire.

Abstract

Tournaments are widely used models to represent pairwise notions of dominance between candidates, alternatives, or teams. Tournament solutions associate a set of winning alternatives with a tournament, notable examples including the Copeland and the Borda rules on unweighted and weighted tournaments respectively. We study the problem of identifying minimal supports for a tournament winner by finding minimal sub-tournaments on which the tournament winner wins independently of the sub-tournament completion, i.e., is a necessary winner of the sub-tournament. In this paper we present polynomial algorithms to compute minimal supports for Copeland and Borda tournament winners, and provide bounds on the size of the smallest minimal supports both theoretically and empirically.

Keywords

Computational Social Choice, Voting, Tournament Solutions, Necessary Winner.

1 Introduction

Tournaments are well-known mathematical structures to represent pairwise contests between alternatives, be they a set of candidates in an election, different teams in a sport league, or competing political proposals. Tournaments have an intuitive and appealing graphical form, representing the alternatives as nodes of a complete directed graph, with the edges' weights storing the number of pairwise contests won by an alternative against another one. A number of tournament solutions have been proposed and studied in the literature to aggregate the information represented in such graphs and select a set of winning alternatives (for an introduction see Brandt *et al.* [4] and Fischer *et al.* [13]).

But what justifies the choice made by a given tournament solution to its users? The classical answer from the literature in social choice and welfare economics (see, e.g., Arrow *et al.* [1]) is to present a set of desirable axiomatic properties or an intuitive mathematical structure that constitutes the normative justification of a tournament solution. For instance, the well-known Borda rule can be characterised as a function electing the winners of the closest unanimous preference profiles in terms of swap distance [12]. In this paper we take a different perspective and we introduce the concept of *minimal support (MS)* for tournaments, which encapsulates the core reasons why a tournament winner belongs to the winner set. We define minimal supports of a winning alternative w for a given tournament solution S on tournament G as the set of all minimal sub-tournaments of G for which w is a necessary winner, i.e., w is the winner of S in all completions of the minimal support.

Minimal supports can provide tools to better understand the reasons behind the choice of a winning alternative. This makes MSs interesting to study from the perspective of explainable AI which aims at helping human users to comprehend and trust the results and output created by AI systems. Indeed, MSs can be viewed as an adaptation of concepts from a recent stream of papers on formal explanations in machine learning (see, e.g., [21, 8, 15]), that provides definitions and algorithms to answer counterfactual questions such as "why a is the result of classifier F on input I?"—known as abductive explanations—or "why b is *not* the result of classifier F on input I?"—known as contrastive explanations—by identifying minimal sets of input features that have

the potential of reverting the classifier result, or that, if left unchanged, can guarantee that the same outcome is reproduced.

Minimal supports also have important connections with bribery, a well-studied problem in computational social choice. In a bribery problem an external attacker faces the problem of changing an election outcome by inducing voters to change their vote (usually by using monetary transfers) while keeping the structure of the election untouched (number of voters, candidates, decision rule...). Since minimal supports identify sub-tournaments where the original tournament winner is a necessary winner, the election can be protected by preventing the attacker from changing the elements of any minimal support of the winner, thus ensuring that the winner does not drop out of the winner set. With such a move no destructive attack on the considered winner can be successful and no constructive attack for a losing alternative can be successful (modulo some standard assumptions on tie-breaking).

Our Contribution After introducing some preliminary notions in Section 2 we define the notion of minimal supports in Section 3 and draw connections with various notions both in computational social choice and explainable AI. We refine our definition of MS by following two principles: to maximise the user readability minimal supports we only consider smallest (i.e., minimal wrt. cardinality) minimal supports; and to focus the attention of our algorithms on the relevant information, we maximise the number of victories from the tournament winner that are present in minimal supports. We show that computing minimal supports maximising these two principles can be done in polynomial time for both the Copeland rule on unweighted tournaments and the Borda rule on weighted ones (see Section 4). Our results contrast with general findings on minimal explanations, for which the problem of finding smallest explanations is often intractable (in its general form the decision problem is Σ_2^{P}-hard [18]). We also provide characterisations of smallest minimal supports and precise bounds on their size, showing that our algorithms provide human-readable structures that are as close as possible to the relevant information contained in the tournament. Finally, in Section
refsec:conclusions we conclude and we discuss important connections with recent work on tournament solutions and on abductive explanations in voting.

1.1 Context and Related Work

Recent position papers in computational social choice—and, more broadly, on the use of digital and AI tools in democratic applications—advocate for simplicity in the development of collective decision making tools, and for the improvement of broadly used mechanisms (see, e.g., Grossi *et al.* [14] and Lev [17]). Our paper aims at contributing to this line of work, by providing human-readable explanations supporting the winner of two simple tournament solutions such as the Copeland and the Borda rule.

Abductive explanations have recently been considered in voting with rankings for the Borda rule [7], resulting however in rather large objects that hinder their usability as explanatory tools. On the other hand, contrastive explanations in tournaments correspond to the destructive reversal sets defined by Brill *et al.* [5] in their effort to define collective rankings from tournament solutions. While contrastive explanations, i.e., answers to "*Why not?*" questions, are great tools to show why a losing alternative was not selected, they do not provide a global backing for the selection of one specific alternative among many others. In other words, explaining individually why each losing alternative did not win is not necessarily a good explanation of why the winner was elected [19].

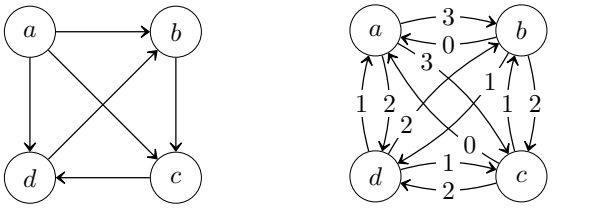
Bribery is a classical problem in computational social choice. It has been studied under various settings ranging from cases where votes can be freely changed to cases where a limited set of modifications of the original ballots are allowed (see, e.g., [10] and [9]). The setting considered in this paper, where each pairwise comparison can be individually changed even when breaking the transitivity of the original preferences, was introduced by Faliszewski *et al.* [11] as *microbribery*. As bribery is a well-studied problem, the dual problem of defending against it could have been investigated in similar depth. However, in line with the original argument of Bartholdi *et al.* [3], many papers rely on the intractability of computing an effective bribery to protect against it. To the best of our knowledge only one paper covers the problem of defending against bribery: Chen *et al.* [6] studies the computational complexity of protecting against both constructive and destructive bribery showing that the protecting against bribery is generally harder than the bribery problem.

Finally, a closely related paper is the work of Aziz *et al.* [2] who study the problem of determining if an alternative is a necessary (resp. possible) winner of a partial tournament, i.e., whether an alternative belongs to the winner set in all (resp. some) completions of a tournament with missing information.

2 Preliminaries

Tournaments are well-studied mathematical structures used to represent compactly pairwise comparisons among alternatives. In the following presentation we use the framework of Aziz *et al.* [2]. Formally, a *partial tournament* is a pair $G = (\mathcal{C}, E)$ where $\mathcal{C}$ is a nonempty finite set of candidates or alternatives and $E \subseteq \mathcal{C} \times \mathcal{C}$ is an asymmetric relation on $\mathcal{C}$, i.e., $(y, x) \notin E$ whenever $(x, y) \in E$. A *tournament* T is a partial tournament $(\mathcal{C}, E)$ for which E is also complete, i.e., either $(x, y) \in E$ or $(y, x) \in E$ for all distinct $x, y \in \mathcal{C}$. We also define the notions of *out-going edges* of a candidate x in E as $E^{x\rightarrow} = \{(x, y) : y \in \mathcal{C}, (x, y) \in E\}$ and its *in-going edges* as $E^{\rightarrow x} = \{(y, x) : y \in \mathcal{C}, (y, x) \in E\}$.

Example 1. *Consider tournament* $G = (\mathcal{C}, E)$ *in Figure 1a. In our notation, we have:* $E^{a\rightarrow}=\{(a, b), (a, c), (a, d)\}$, $E^{b\rightarrow}=\{(b, c)\}$, $E^{\rightarrow a} = \{\}$ *and* $E^{\rightarrow b}=\{(a, b), (d, b)\}$.



(a) Unweighted tournament G (b) 3-weighted tournament G_w

Figure 1: Example of a tournament and a weighted tournament.

Let n be a strictly positive integer. A *partial n-weighted tournament* is a pair $G = (\mathcal{C}, \mu)$ where $\mathcal{C}$ is a nonempty finite set of candidates and $\mu : \mathcal{C} \times \mathcal{C} \to \{0, \dots, n\}$ a weight function such that for each pair $(x, y) \in \mathcal{C} \times \mathcal{C}$ with $x \neq y$, $\mu(x, y) + \mu(y, x) \leq n$. Since loops are not allowed, we assume $\mu(x, x) = 0$ for all candidate x. A *(complete) n-weighted tournament* satisfies for each pair $(x, y) \in \mathcal{C} \times \mathcal{C}$ with $x \neq y$, $\mu(x, y) + \mu(y, x) = n$.
Naturally, (unweighted) tournaments can be seen as 1-weighted tournaments.
We will also use an equivalent definition that presents weighted tournaments as multigraphs. A *multiset* is a pair (A, μ) where A is the underlying set and $\mu : A \to \mathbb{N}$ a multiplicity function giving the number of occurrences of the elements in A. The cardinality $|(A, \mu)|$ is defined as $\sum_{e \in A} \mu(e)$. For multisets such as $(\{e_0, e_1, \dots, e_n\}, \mu)$, we will adopt the common notation $\{e_o^{\mu(e_0)}, e_1^{\mu(e_1)}, \dots, e_n^{\mu(e_n)}\}$ and we will omit the superscript in case of a multiplicity of 1. A *multigraph* is a pair $G = (V, E)$ where V is a nonempty finite set of vertices and E a multiset of elements of $V \times V$. A partial n-weighted tournament $G = (\mathcal{C}, \mu)$ can naturally be represented as a multigraph $(\mathcal{C}, (\mathcal{C} \times \mathcal{C}, \mu))$, simply denoted $(\mathcal{C}, E)$ where E is now a multiset. The (multi)sets of in-going and out-going edges are naturally defined for multigraph. In our figures we will omit edges with a weight/multiplicity of 0.

Example 2. *Consider the weighted tournament $G_w = (\mathcal{C}, E_w)$ in Figure 1b. In multiset notation we have: ${E_w}^{b\to}$=$\{(b, c)^2, (b, d)\}$, ${E_w}^{\to b} = \{(a, b)^3, (c, b), (d, b)^2\}$, ${E_w}^{a\to}$=$\{(a, b)^3, (a, c)^3, (a, d)^2\}$ and ${E_w}^{\to a} = \{(d, a)\}$.*

A tournament solution is a function S mapping a tournament $G = (\mathcal{C}, E)$ to a non-empty set of candidates $S(G) \subseteq \mathcal{C}$. In this paper we focus on two simple and widely studied rules. In the case of unweighted tournaments, the *Copeland rule* selects alternatives based on the number of other alternatives they dominate. Formally, the *Copeland score* of an alternative x in tournament $G = (\mathcal{C}, E)$ is $\sigma_{\text{Cop}}(x, G) = |E^{x\to}|$. Copeland rule selects as winners alternatives that have a maximal Copeland score. It is *Condorcet consistent*, i.e., if an alternative beats all others then it will be the unique element in the winner set. In weighted tournaments, the *Borda rule* selects alternatives that maximise the total weight of their outgoing edges. Define the *Borda score* of an alternative x in weighted tournament $G = (\mathcal{C}, \mu)$ as $\sigma_{\text{Borda}}(x, G) = \sum_{y \in C} \mu(x, y)$.

3 Minimal Supports for Tournaments

This section adapts the definitions from the literature on formal explanations to the case of unweighted and weighted tournaments, and propose two principles refining our search for minimal supports. Even if the focus of the paper is on Copeland and Borda, definitions in this section apply to any tournament solution.

3.1 MSs for Unweighted Tournaments

Abductive reasoning applied to tournament solutions explain the winner of a tournament by exhibiting a minimal subset of the tournament such that the winner remains unchanged. Let $G = (\mathcal{C}, E)$ and $G' = (\mathcal{C}', E')$ be two partial tournaments, we say that G' is an *extension* of G, denoted $G \subseteq G'$, if $\mathcal{C} = \mathcal{C}'$ and $E \subseteq E'$. If E' is complete, G' is called a *completion* of G, and we write $[G]$ for the set of completions of G. To obtain a formal definition we borrow from the literature the concept of necessary winner [16]:

Definition 1. *Given a partial tournament G, a tournament solution S and a candidate $c \in \mathcal{C}$, c is a* necessary winner *of S in G if for every completion $G' \in [G]$ we have that $c \in S(G')$. We write $c \in NW_S(G)$.*

We now define minimal supports for tournaments:

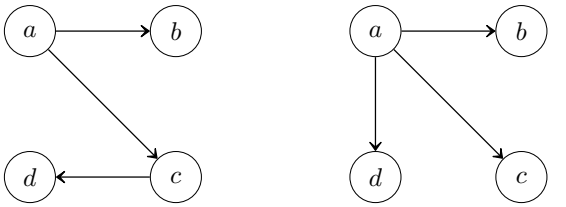
Definition 2. *Given a tournament $G = (\mathcal{C}, E)$, a tournament solution S and a winning candidate $w \in S(G)$, an **minimal support (MS)** for $w \in S(G)$ is a partial tournament $G' \subseteq G$ such that:*

(a) $w \in NW_S(G')$

(b) G' is $\subseteq$-minimal, i.e., all partial tournaments $G'' \subseteq G'$ are such that $w \notin NW_S(G'')$.

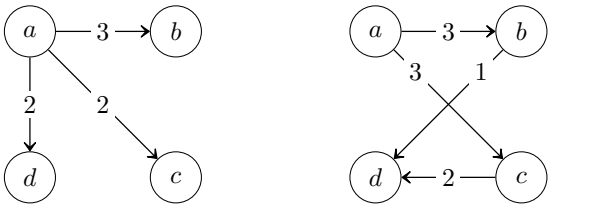
Partial tournaments only satisfying (a) are called weak minimal supports. Since the set of candidates in an MS is the same as the one of the original tournament, we will refer to an MS only by its set of edges.

Example 3. *Consider tournament G in Figure 1a. Both $\mathcal{X} = \{(a, b), (a, c), (c, d)\}$ and $\mathcal{Y} = \{(a, b), (a, c), (a, d)\}$ represented in Figure 2 are MSs for $a \in \text{Cop}(G)$. What $\mathcal{X}$ shows is that a beats at least two other candidates (b and c) while every other candidate has at least one defeat, hence at most two victories, preventing them from obtaining a Copeland score strictly better than a. Hence, a is a necessary winner of $\mathcal{X}$, and $\mathcal{X}$ is minimal because removing any edge results opening the possibility for an undefeated candidate different from a. The case of $\mathcal{Y}$ is more compelling, as it directly shows that a is a Condorcet winner.*



(a) $\mathcal{X}=\{(a,b),(a,c),(c,d)\}$ (b) $\mathcal{Y}=\{(a,b),(a,c),(a,d)\}$

Figure 2: MSs for $a \in \mathrm{Cop}(G)$ from Figure 1a.



(a) $\mathcal{I}=\{(a,b)^3,(a,c)^2,(a,d)^2\}$ (b) $\mathcal{J}=\{(a,b)^3,(a,c)^3,(b,d),(c,d)^2\}$

Figure 3: MSs for $a \in \mathrm{Borda}(G_w)$ from Figure 1b.

3.2 MSs for Weighted Tournaments

The case of weighted tournaments can be treated analogously by simply defining an inclusion relation for weighted tournaments. Let $G = (\mathcal{C}, \mu)$ and $G' = (\mathcal{C}', \mu')$ be two partial n-weighted tournaments, we say that G' is an *extension* of G, denoted $G \subseteq G'$, if $\mathcal{C} = \mathcal{C}'$ and $\mu(x,y) \leq \mu'(x,y)$ for all $(x,y) \in \mathcal{C}\times\mathcal{C}$. A *completion* of G is a complete n-weighted extension, and both Definitions 1 and 2 can be applied.

Example 4. *Consider G_w in Figure 1b. As represented in Figure 3, both $\mathcal{I} = \{(a,b)^3,(a,c)^2,(a,d)^2\}$ and $\mathcal{J} = \{(a,b)^3,(a,c)^3,(b,d),(c,d)^2\}$ are MSs for $a \in \mathrm{Borda}(G_w)$. While $\mathcal{I}$ is focused on showing that candidate a has sufficient support against any other candidate, $\mathcal{J}$ shows that candidate a is preferred by everyone (recall that $n = 3$) to both b and c, and that candidate d has too much incoming weight to challenge a as the winner. Both sub-tournaments are inclusion-minimal, as decreasing any of the weights by 1 will create an incumbent candidate that can potentially beat a.*

3.3 Smallest and Win-Maximising MSs

For a given tournament there might be an exponential number of different minimal supports, since MSs are an instance of a Sperner family [22]. Therefore in this section we introduce two criteria to select the most promising MSs. First, we focus our attention on the size of the support, as conciseness is an important factor in both settings of explainability and bribery. MSs are by definition irreducible, so they already optimize in a set-theoretic way for the criterion. However, not all MSs are equal in cardinality. For unweighted tournaments, we will consider *smallest minimal supports* (*SMS*) which are MSs such that no MS exists that is strictly smaller cardinal-wise (i.e., in number of edges). Defining the size of a weighted tournament is less straightforward, as they are based on a complete graph. We compare weighted subtournaments using the overall sum of the edges' weights, with the idea that this will result in supports containing the smallest number of pairwise comparisons to be locked, which are in a way the counterpart of the swap operations carried out in bribery. Formally, if G is a weighted tournament, let $|G| = \sum_{(x,y)\in\mathcal{C}\times\mathcal{C}} \mu(x,y)$ (corresponding to the ℓ_1 norm of the weight vector of G). Thus, SMSs for weighted tournaments are defined as SMSs such that no MS exists that is strictly smaller in ℓ_1-norm.

Looking back at Example 3, both MSs $\mathcal{X}$ and $\mathcal{Y}$ have the same cardinality and are SMSs for $a \in \mathrm{Borda}(G)$. However, $\mathcal{X}$ relies less on a's good performances as it does no use the fact that a is a Condorcet winner. This motivates us to introduce a second tournament-specific criterion to further refine MSs and favors MSs highlighting the winner's good performance instead of relying on others' bad performance.

Definition 3. *Given a tournament $G = (\mathcal{C}, E)$ and a winning candidate $w \in \mathrm{Borda}(G)$, the* win count *of a MS $\mathcal{X}$ for $w \in \mathrm{Borda}(G)$, denoted $\mathrm{WC}(\mathcal{X})$, is $|\mathcal{X}^{w\to}|$, i.e., the number of pairwise comparisons won by the tournament winner that are present in the MS.*

The multiset notation for weighted tournaments comes useful, as Definition 3 applies if E is a set of edges or a multiset, as in weighted tournaments. In the latter, what we maximise is $\sum_{c\in\mathcal{C}} \mu_{\mathcal{X}}(w,c)$, the total number of victories from the target winner w in the MS $\mathcal{X}$. A SMS maximising the win count, called *maxwin-SMS*, is a SMS such that no SMS with a larger win count exists.

3.4 Score Margins and Necessary Winners

In this section we present useful results characterising necessary winners in terms of the score margins of the winning candidate, a notion that will be used in subsequent sections to construct our algorithms for the computation of minimal supports. Our results are for weighted tournaments and the Borda rule, but they also apply to Copeland by setting $n = 1$.

Let us denote the Borda score of candidate c for the tournament G as $\sigma_{\mathrm{Borda}}(c,G)$. When the context will make the considered tournament clear, we will shorten this notation to $\sigma_{\mathrm{Borda}}(c)$. Idem for the Copeland rule.

Definition 4. *Given n voters, a partial n-weighted tournament $G = (\mathcal{C}, E)$, and a candidate $c \in \mathcal{C}$, we call $\sigma^{min}_{\mathrm{Borda}}(c,G) = min_{G'\in[G]}\sigma_{\mathrm{Borda}}(c,G')$ (resp. $\sigma^{max}_{\mathrm{Borda}}(c,G) = max_{G'\in[G]}\sigma_{\mathrm{Borda}}(c,G')$) the **minimal** (resp. **maximal**) **score** that candidate c can obtain in any completion of G.*

We then give a formula to compute $\sigma^{min}_{\mathrm{Borda}}$ and $\sigma^{max}_{\mathrm{Borda}}$ in partial tournaments.

Lemma 1. *Given n voters, a partial n-weighted tournament $G = (\mathcal{C}, E)$ and a candidate $c \in \mathcal{C}$, $\sigma^{min}_{\text{Borda}}(c, G) = |E^{c\rightarrow}|$ and $\sigma^{max}_{\text{Borda}}(c, G) = n(|\mathcal{C}| - 1) - |E^{\rightarrow c}|$.*

Proof. $\sigma^{min}_{\text{Borda}}(c, G) = |E^{c\rightarrow}|$ as the minimum number of pairwise comparisons to be won by c is the number of pairwise comparisons already won by c. $\sigma^{max}_{\text{Borda}}(c, G) = n(|\mathcal{C}| - 1) - |E^{\rightarrow c}|$ as the maximum number of pairwise comparisons to be won by c corresponds to the total number of pairwise comparisons to be played by c less those already lost. □

Then, we can introduce the notion of score margin, the main notion that we will used in the proofs of this paper.

Definition 5. *Given n voters, a partial n-weighted tournament $G = (\mathcal{C}, E)$, a pair of candidates $c, c' \in \mathcal{C}$, the **score margin** of candidate c w.r.t. c' for G is $\delta_{\text{Borda}}(c, c', G) = \sigma^{min}_{\text{Borda}}(c, G) - \sigma^{max}_{\text{Borda}}(c', G)$.*

From now on we will consider the score margins of the winning candidate w for the Borda rule, writing $\delta(c, G)$ as a shorthand for $\delta_{\text{Borda}}(w, c, G)$.
We can then provide a characterisation for necessary winner in terms of score margin, in line with an analogous result by Konczak and Lang [16] in voting.

Proposition 2. *Given n voters, a partial n-weighted tournament $G = (\mathcal{C}, E)$, and a candidate $w \in \mathcal{C}$, $w \in NW_{Borda}(G)$ if and only if for all $c \in \mathcal{C} \setminus \{w\}$, $\delta(c, G) \geq 0$.*

Proof. $\boxed{\Longleftarrow}$ is straightforward. $\boxed{\Longrightarrow}$ To prove this direction, for every losing candidate c, we build a completion G_c of G where $\sigma_{\text{Borda}}(w, G_c) = \sigma^{min}_{\text{Borda}}(w, G)$ and $\sigma_{\text{Borda}}(c, G_c) = \sigma^{max}_{\text{Borda}}(c, G)$. Since w is a necessary winner, w is a winner of G_c. Hence $\sigma_{\text{Borda}}(w, G_c) \geq \sigma_{\text{Borda}}(c, G_c)$ i.e. $\delta(c, G) \geq 0$. To build G_c we proceed as in the proof of Lemma 1 and complete G such that w loses all remaining contests while c wins them all. Let $E_c = E \cup \{(x, w)^{n-\mu(w,x)} : x \in \mathcal{C}\} \cup \{(c, y)^{n-\mu(y,c)} : y \in \mathcal{C}\}$, complete E_c (in any way) to get E'_c. G_c is the tournament $(\mathcal{C}, E'_c)$. □

4 Computing and Characterising Minimal Supports for the Copeland and Borda Rules

In this section we provide bounds and algorithms for the computation of maxwin-SMSs for the Copeland rule in unweighted tournament and the Borda rule in weighted tournaments.

4.1 Lower Bounds on the Size of MSs

We begin by providing a lower bound on the minimal number of pairwise comparisons in a MS for the Borda rule.

Proposition 3. *Given n voters, a complete n-weighted tournament $G = (\mathcal{C}, E)$ with $|\mathcal{C}| = m$, and a winning candidate $w \in \text{Borda}(G)$, for all MS $\mathcal{X}$ for $w \in \text{Borda}(G)$, we have:*

$$|\mathcal{X}| \geq (m-1)(n(m-1) - \sigma_{\text{Borda}}(w))$$

Proof. By definition, if $\mathcal{X}$ is a MS for $w \in \text{Borda}(G)$ then $w \in \text{NW}_{\text{Borda}}(\mathcal{X})$. Summing the inequalities from Proposition 2 we have:

$$\begin{aligned}
&\sum_{c \neq w} \left(\sigma^{min}_{\text{Borda}}(w, G) - \sigma^{max}_{\text{Borda}}(c, G)\right) \geq 0 \\
\Longrightarrow\ &(m-1)|\mathcal{X}^{w\rightarrow}| - \sum_{c \neq w} (n(m-1) - |\mathcal{X}^{\rightarrow c}|) \geq 0 \\
\Longrightarrow\ &(m-1)|\mathcal{X}^{w\rightarrow}| - (m-1)n(m-1) + \sum_{c \neq w} |\mathcal{X}^{\rightarrow c}| \geq 0 \\
\Longrightarrow\ &\sum_{c \neq w} |\mathcal{X}^{\rightarrow c}| \geq (m-1)(n(m-1) - |\mathcal{X}^{w\rightarrow}|) \\
\Longrightarrow\ &|\mathcal{X}| \geq (m-1)(n(m-1) - |\mathcal{X}^{w\rightarrow}|)
\end{aligned}$$

To conclude, note that $\mathcal{X}^{w\rightarrow} \subseteq E^{w\rightarrow}$ thus $|\mathcal{X}^{w\rightarrow}| \leq |E^{w\rightarrow}| = \sigma_{\text{Borda}}(w)$, obtaining the desired bound. □

With a similar proof we can obtain a lower bound for the Copeland rule on (unweighted) tournaments, since the Borda rule coincides with the Copeland rule in the case of 1-weighted tournaments.

Corollary 4. *Given a complete tournament $G = (\mathcal{C}, E)$ with $|\mathcal{C}|$=m and a winning candidate $w \in \text{Cop}(G)$, for all MS $\mathcal{X}$ for $w \in \text{Cop}(G)$, we have:*

$$|\mathcal{X}| \geq (m-1)(m - 1 - \sigma_{\text{Cop}}(w))$$

4.2 Algorithms to Compute Maxwin-SMS

We now provide polynomial algorithms to compute a smallest MS which also maximises win count. We first start with the simpler case of the Copeland rule on (unweighted) tournaments and then move on to the case of the Borda rule on weighted tournaments.

Proposition 5. *Given a complete tournament $G = (\mathcal{C}, E)$ with $|\mathcal{C}| = m$ and a winning candidate $w \in \text{Cop}(G)$, there exists an algorithm which computes a maxwin-SMS for $w \in \text{Cop}(G)$ in $\mathcal{O}(m^2)$.*

Algorithm 1: FINDMAXWINSMS-COP

Data: Complete tournament $G = (\mathcal{C}, E)$, winning candidate $w \in \text{Cop}(G)$
Result: a maxwin-SMS $\mathcal{X}$

```
X ← E^{w→}
for c ∈ C \ {w} do
    while |X^{→c}| < m − 1 − σ_Cop(w) do
        X ← X ∪ {(c', c)} for some (c', c) ∈ E \ X
return X
```

Proof. Consider Algorithm 1. Its underlying idea is to start by initialising a MS $\mathcal{X}$ in line 1 with all winner's victories. Then, in the loops at lines 2 to 4, add enough defeats to $\mathcal{X}$ for candidates different from w, so that these candidates cannot win in any completion of $\mathcal{X}$. To show the correctness of Algorithm 1 we begin by showing that it ouputs an SMS. If $\sigma_{\mathrm{Cop}}(w) = m-1$, Algorithm 1 returns $E^{w\to}$ which is minimal (else the winning candidate can possibly have one defeat while an undefeated losing candidate exists). If $\sigma_{\mathrm{Cop}}(w) \leq m-1$, then for every losing alternative $c \neq w$, before the while loop on line 3, $|\mathcal{X}^{\to c}|$ is at most one since it can only potentially contain (w,c) if it is present in E. We know that $\sigma_{\mathrm{Cop}}(c) \leq \sigma_{\mathrm{Cop}}(w)$, so c has more defeats than w and $|E^{\to c}| \geq m-1-\sigma_{\mathrm{Cop}}(w)$. Edges from $E^{\to c}$ are added one by one to $\mathcal{X}$ in the while loop at lines 3 and 4, until c has exactly $m-1-\sigma_{\mathrm{Cop}}(w)$ defeats in $\mathcal{X}$. Proposition 2 is thus satisfied and $w \in \mathrm{NW}_{Cop}(G)$, so $\mathcal{X}$ is a weak MS. Every edge in $E^{w\to}$ is in $\mathcal{X}^{\to c}$ for some $c \neq w$; thus the number of edges in $\mathcal{X}$ at the end of the algorithm is exactly $(m-1)(m-1-\sigma_{\mathrm{Cop}}(w))$. By Proposition 4 no smaller MS exists, therefore $\mathcal{X}$ is an SMS for $w \in Cop(G)$. Finally, since all winner's wins are included in the returned SMS, it is a maxmin-SMS. □

We now show that Proposition 2 generalises to the case of computing a maxwin-SMS for the Borda rule on weighted tournaments. First, we need to show the following lemma showing the existence of a SMS which is either included in $E^{w\to}$ or contains it.

Lemma 6. *Given n voters, an n-weighted tournament $G = (\mathcal{C}, E)$ with $|\mathcal{C}| = m$, and a winning candidate $w \in \mathrm{Borda}(G)$, there exists a SMS $\mathcal{X}$ for $w \in \mathrm{Borda}(G)$ such that $\mathcal{X} \subseteq E^{w\to}$ or $E^{w\to} \subseteq \mathcal{X}$.*

We are now ready to show the main result of this section.

Proposition 7. *Given n voters, a complete n-weighted tournament $G = (\mathcal{C}, E)$ with $|\mathcal{C}| = m$, and a winning candidate $w \in \mathrm{Borda}(G)$, there exists an algorithm which computes a maxwin-SMS for $w \in \mathrm{Borda}(G)$ in $\mathcal{O}(nm^2)$.*

Proof. Consider Algorithm 2. In line with Algorithm 1 we start by selecting all winner's wins (for loop lines 2 and 3). If it is not enough to make w a necessary winner, we proceed the same way as in Algorithm 1 (lines 4 to 7) else we have a weak MS that needs to be minimised to obtain a SMS (lines 7 to 10).

We now show that Algorithm 2 returns a maxwin-SMS. In combination with lines 1 to 3, the first branch of the if on line 4 precisely corresponds to the generalization of Algorithm 1. Indeed, if $\exists c \in \mathcal{C} \setminus \{w\}$, $\delta(c, \mathcal{X}) < 0$ then $E^{w\to}$ is not a weak MS and it contains no MS. Thus, Lemma 6 gives the existence of a SMS containing $E^{w\to}$ which is then completed like in Algorithm 1 to obtain a SMS of size $(m-1)(n(m-1)-\sigma_{\mathrm{Borda}}(w))$.

Regarding the second case corresponding to the second branch of the if on line 4, we need to show that it gives us an SMS. First note that since $\exists c \in \mathcal{C} \setminus \{w\}$, $\delta(c, \mathcal{X}) < 0$ is violated, $\mathcal{X}$ is a weak MS and with Lemma 6, we have the existence of a SMS included in $E^{w\to}$. We will show that when the guard of the while on line 10, i.e., $\delta(c_0, \mathcal{X}) \geq 2 \wedge \min_{c\in\mathcal{C}\setminus\{w\}} \delta(c, \mathcal{X}) \geq 1$ is not met, we land on a SMS. If $\delta(c_0, \mathcal{X}) < 2$, $\forall c \in \mathcal{C} \setminus \{w\}$ then $\delta(c, \mathcal{X}) \in \{0, 1\}$. Hence, $\sum_{c\neq w} \delta(c, \mathcal{X}) \leq m-1$. Additionally, observe that $\mathcal{X} \subseteq E^{w\to}$ and that each element of $E^{w\to}$ i.e. (w, c') for some candidate $c' \neq w$ contributes m to $\sum_{c\neq w} \delta(c, \mathcal{X})$ (a contribution of 2 in the term associated to c' and 1 for the $m-2$ candidates different from c' and w). Thus, for any subset $\mathcal{X}'$ of $E^{w\to}$ strictly smaller (cardinal-wise) than $\mathcal{X}$, we have $\sum_{c\neq w} \delta(c, \mathcal{X}') \leq -1$. $\exists c \in \mathcal{C} \setminus \{w\}$, $\delta(c, \mathcal{X}') \leq -1$ and w is not a necessary winner. Hence, no strictly smaller MS included in $E^{w\to}$ exists, and $\mathcal{X}$ is a SMS.

If $\min_{c\in\mathcal{C}\setminus\{w\}} \delta(c, \mathcal{X}) < 1$, we show that the sequence of choices proposed in Algorithm 2 is optimal at every step regarding the maximisation of $\min_{c\in\mathcal{C}\setminus\{w\}} \delta(c, \mathcal{X})$ which implies that the algorithm optimally reduces the explanation and returns a SMS. We proceed by induction. Let us note $\mathcal{X}_i$ the value of $\mathcal{X}$ after the i^{th} iteration. Suppose that after i iterations, the guard condition is still met and that $\min_{c\in\mathcal{C}\setminus\{w\}} \delta(c, \mathcal{X}_i)$ is maximal. If $c_0 \neq \operatorname{argmin}_{c\in\mathcal{C}\setminus\{w\}} \delta(c, \mathcal{X}_i)$, $\min_{c\in\mathcal{C}\setminus\{w\}} \delta(c, \mathcal{X}_{i+1}) = \min_{c\in\mathcal{C}\setminus\{w\}} \delta(c, \mathcal{X}_i) - 1$ which is inevitable and thus optimal. Otherwise, if $c_0 = \operatorname{argmin}_{c\in\mathcal{C}\setminus\{w\}} \delta(c, \mathcal{X}_i)$, all score margins are the same and $\min_{c\in\mathcal{C}\setminus\{w\}} \delta(c, \mathcal{X}_{i+1})$ is independent of the $i+1^{\text{th}}$ chosen candidate. Hence, when the second condition of $\delta(c_0, \mathcal{X}) \geq 2 \wedge \min_{c\in\mathcal{C}\setminus\{w\}} \delta(c, \mathcal{X}) \geq 1$ is violated, no other strictly longer sequence of choices do not violate it. The algorithm returns a smallest (cardinal-wise) MS included in $E^{w\to}$ which is a SMS since there exists a SMS included in $E^{w\to}$.

Regarding the maximisation of win counts, the first case is clear. For the second case, since the returned SMS is included in $E^{w\to}$, it is optimal among SMSs. □

Algorithm 2: FINDMAXWINSMS-BORDA

Data: Complete n-weighted tournament $G = (\mathcal{C}, \mu)$, winning candidate $w \in \mathrm{Borda}(G)$
Result: a maxwin-SMS $\mathcal{X} = (\mathcal{C}, \mu_{\mathcal{X}})$

1 $\mu_{\mathcal{X}}(\cdot,\cdot) \leftarrow 0$
2 **for** $c \in \mathcal{C} \setminus \{w\}$ **do**
3 $\mu_{\mathcal{X}}(w, c) \leftarrow \mu(w, c)$
4 **if** $\exists c \in \mathcal{C} \setminus \{w\}$, $\delta(c, \mathcal{X}) < 0$ **then**
5 **for** $c \in \mathcal{C} \setminus \{w\}$ **do**
6 **while** $\delta(c, \mathcal{X}) < 0$ **do**
7 $\mu_{\mathcal{X}}(c', c) \leftarrow \mu_{\mathcal{X}}(c', c) + 1$ for some c' s.t. $\mu_{\mathcal{X}}(c', c) < \mu(c', c)$
8 **else**
9 $c_0 \leftarrow \operatorname{argmax}_{c\in\mathcal{C}\setminus\{w\}} \delta(c, \mathcal{X})$
10 **while** $\delta(c_0, \mathcal{X}) \geq 2 \wedge \min_{c\in\mathcal{C}\setminus\{w\}} \delta(c, \mathcal{X}) \geq 1$ **do**
11 $\mu_{\mathcal{X}}(w, c_0) \leftarrow \mu_{\mathcal{X}}(w, c_0) - 1$
12 $c_0 \leftarrow \operatorname{argmax}_{c\in\mathcal{C}\setminus\{w\}} \delta(c, \mathcal{X})$
13 **return** $\mathcal{X}$

As observed previously, Algorithm 1 can be seen as a special case of Algorithm 2 on 1-weighed tournaments. In this particular case the else branch of the if on line 4 would be useless as it would only be triggered when the winning candidate is a Condorcet winner but it would not enter in the while loop on line 10 and return all pairwise comparisons won by the winner.

4.3 Bound tightness and Characterisation of SMSs

In this section we show that the lower bounds on the size of MSs are almost tight. We characterise the set of SMSs and show that SMSs share a core structure based on the wins of the considered winner.
First, for Copeland, we show with the proof of Proposition 5 that the bound of Proposition 4 is tight unless there is a Condorcet winner.

Proposition 8. *Given a complete tournament $G = (\mathcal{C}, E)$ with $|\mathcal{C}| = m$ and a winning candidate $w \in \mathrm{Cop}(G)$ with a Copeland score of $\sigma(w)$, for all SMS $\mathcal{X}$ for $w \in \mathrm{Cop}(G)$:*

- *if w is not a Condorcet winner then $|\mathcal{X}|=(m-1)(m-1-\sigma(w))$*
- *if w is a Condorcet winner and $|\mathcal{X}|=m-1$.*

Similarly, by inspecting the proof of Proposition 7 we are able to show that the bound of Proposition 3 is tight except for tournaments where the winner wins by a very large margin. In that case, additional constraints have to be taken into account and prevents us from reaching the lower bound.

Proposition 9. *Given n voters, a complete n-weighted tournament $G = (\mathcal{C}, \mu)$ with $|\mathcal{C}| = m$, and $w \in \mathrm{Borda}(G)$ a winning candidate with a Borda score of $\sigma(w)$, for all SMS $\mathcal{X}$ for $w \in \mathrm{Borda}(G)$:*

- *if exist a SMS $\mathcal{Y}$ cointaining $E^{w\to}$ then $|\mathcal{X}|=(m-1)(n(m-1)-\sigma(w))$*
- *else $|\mathcal{X}|=\max(\lceil n(m-1)^2/m\rceil, n(m-1) - \min_{c\neq w}\mu(w,c))$.*

Algorithms 1 and 2 return SMSs maximising the number of pairwise comparisons won by the winner. Naturally, one might ask what the other SMSs look like. That is why we provide a full characterisation of all the SMSs of a given tournament.

Proposition 10. *Given n voters, a complete n-weighted tournament $G = (\mathcal{C}, E)$, and a winning candidate $w \in \mathrm{Borda}(G)$, for all SMS $\mathcal{X}$ for $w \in \mathrm{Borda}(G)$,*

(i) if $\exists c_0, c_1 \in \mathcal{C}\setminus\{w\}$, $c_0 \neq c_1$ s.t. $\delta(c_0, E^{w\to}) < 0$ and $\delta(c_1, E^{w\to}) < 0$ then $E^{w\to} \subseteq \mathcal{X}$

(ii) else if $\exists! c_0 \in \mathcal{C}\setminus\{w\}$ s.t. $\delta(c_0, E^{w\to}) \leq 0$ then $|E^{w\to}\setminus\mathcal{X}| \leq \min_{c\in\mathcal{C}\setminus\{w,c_0\}}\delta(c, E^{w\to})$ and exists a SMS containing $E^{w\to}$

(iii) else for all SMS $\mathcal{Y}$ for $w \in \mathrm{Borda}(G)$, $|\mathrm{WC}(\mathcal{X}) - \mathrm{WC}(\mathcal{Y})| \leq 1$.

For Copeland, SMSs contains all pairwise comparisons won by w except when w is a Condorcet winner: in that case, there are SMSs that contain all but one win, but the maxwin-SMS contains them all.

Proposition 11. *Given a complete tournament $G = (\mathcal{C}, E)$ and a winning candidate $w \in Cop(G)$, for all SMS $\mathcal{X}$ for $w \in \mathrm{Cop}(G)$:*

(i) if w is not a Condorcet winner then $E^{w\to} \subseteq \mathcal{X}$

(ii) if w is a Condorcet winner, then $|E^{w\to}\setminus\mathcal{X}| \leq 1$, $|\mathcal{X}|=m-1$ and $E^{w\to}$ is the maxwin-SMS

Finally, thanks to the previous characterisation, we show that all SMSs share a similar structure centered around the winner's victories by bounding the difference in terms of win count between SMSs.

Proposition 12. *Given n voters, an n-weighted tournament $G = (\mathcal{C}, \mu)$, and a winning candidate $w \in \mathrm{Borda}(G)$, for all SMSs $\mathcal{X}, \mathcal{Y}$ for $w \in \mathrm{Borda}(G)$, we have $|\mathrm{WC}(\mathcal{X}) - \mathrm{WC}(\mathcal{Y})| \leq \max(1, \frac{n}{2})$.*

5 Conclusion and Future Work

In this paper we defined minimal supports for tournament winners, by identifying specific sub-tournaments that can then be employed either as minimal explanations supporting the choice of the winner or as protection sets against bribery. For the case of the Copeland and the Borda rule we presented polynomial algorithms for computing specific MSs that satisfy two desirable principles, characterised the set of shortest MSs.
The most pressing direction for future research is to apply the present study to tournament solutions that are more elaborate than the Borda rule—with the main open problem being the existence of polynomial algorithms for the computation of SMSs. Moreover, our work has interesting connections with recent work on tournament solutions that should be further explored. First, Maiti and Dey [20] showed that any query algorithm necessitates $\Omega(n^2)$ queries to compute the Copeland winner in the worst case [20, Theorem 3.1]. Since knowing a minimal support allows the computation of the winner without querying the entire graph, their result can be obtained as a corollary of our Proposition 4. Our Proposition 3 shows an analogous novel result for weighted tournaments. Second, the computation of margin of victory of candidates by Brill *et al.* [5] uses a form of formal explanation known as contrastive, which is dual to the minimal supports we considered (known as abductive explanations). While bounds on the size of contrastive and abductive explanations cannot be obtained one from the other, algorithms for their computation are often intertwined. Rephrasing Brill *et al.*'s research in terms of contrastive explanation can shed new light on their results and suggest further advances towards the development of tournament solutions.

References

[1] Kenneth J Arrow, Amartya Sen, and Kotaro Suzumura. *Handbook of social choice and welfare*. Elsevier, 2010.

[2] Haris Aziz, Markus Brill, Felix Fischer, Paul Harrenstein, Jérôme Lang, and Hans Georg Seedig. Possible and necessary winners of partial tournaments. *Journal of Artificial Intelligence Research*, 54:493–534, 2015.

[3] John J Bartholdi, Craig A Tovey, and Michael A Trick. The computational difficulty of manipulating an election. *Social choice and welfare*, 6:227–241, 1989.

[4] Felix Brandt, Markus Brill, and Paul Harrenstein. Tournament solutions. In Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D. Procaccia, editors, *Handbook of Computational Social Choice*, pages 57–84. Cambridge University Press, 2016.

[5] Markus Brill, Ulrike Schmidt-Kraepelin, and Warut Suksompong. Refining tournament solutions via margin of victory. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

[6] Lin Chen, Ahmed Imtiaz Sunny, Lei Xu, Shouhuai Xu, Zhimin Gao, Yang Lu, Weidong Shi, and Nolan Shah. Computational complexity characterization of protecting elections from bribery. *Theoretical Computer Science*, 891:189–209, 2021.

[7] Clément Contet, Umberto Grandi, and Jérôme Mengin. Abductive and contrastive explanations for scoring rules in voting. In *Proceedings of the 27th European Conference on Artificial Intelligence (ECAI)*, 2024.

[8] Adnan Darwiche and Auguste Hirth. On the reasons behind decisions. In *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI)*, 2020.

[9] Edith Elkind, Piotr Faliszewski, and Arkadii Slinko. Swap bribery. In *Proceedings of the 2nd International Symposium on Algorithmic Game Theory (SAGT)*, 2009.

[10] Piotr Faliszewski, Edith Hemaspaandra, and Lane A Hemaspaandra. How hard is bribery in elections? *Journal of Artificial Intelligence Research*, 35:485–532, 2009.

[11] Piotr Faliszewski, Edith Hemaspaandra, Lane A Hemaspaandra, and Jörg Rothe. Llull and copeland voting computationally resist bribery and constructive control. *Journal of Artificial Intelligence Research*, 35:275–341, 2009.

[12] Daniel Farkas and Shmuel Nitzan. The Borda rule and Pareto stability: A comment. *Econometrica*, 47(5), 1979.

[13] Felix Fischer, Olivier Hudry, and Rolf Niedermeier. Weighted tournament solutions. In Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D. Procaccia, editors, *Handbook of Computational Social Choice*, pages 85–102. Cambridge University Press, 2016.

[14] Davide Grossi, Ulrike Hahn, Michael Mäs, Andreas Nitsche, Jan Behrens, Niclas Boehmer, Markus Brill, Ulle Endriss, Umberto Grandi, Adrian Haret, Jobst Heitzig, Nicolien Janssens, Catholijn M. Jonker, Marijn A. Keijzer, Axel Kistner, Martin Lackner, Alexandra Lieben, Anna Mikhaylovskaya, Pradeep K. Murukannaiah, Carlo Proietti, Manon Revel, Élise Rouméas, Ehud Shapiro, Gogulapati Sreedurga, Björn Swierczek, Nimrod Talmon, Paolo Turrini, Zoi Terzopoulou, and Frederik Van De Putte. Enabling the digital democratic revival: A research program for digital democracy. *CoRR*, abs/2401.16863, 2024.

[15] Alexey Ignatiev, Nina Narodytska, Nicholas Asher, and João Marques-Silva. On relating 'why?' and 'why not?' explanations. *CoRR*, abs/2012.11067, 2020.

[16] Kathrin Konczak and J. Lang. Voting procedures with incomplete preferences. In *Proceedings of the Multidisciplinary IJCAI Workshop on Advances in Preference Handling*, 2005.

[17] Omer Lev. Towards a more burkean approach to computational social choice. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence (AAAI)*, 2024.

[18] Paolo Liberatore. Redundancy in logic i: Cnf propositional formulae. *Artificial Intelligence*, 163(2):203–232, 2005.

[19] Peter Lipton. Contrastive explanation. *Royal Institute of Philosophy Supplements*, 27:247–266, 1990.

[20] Arnab Maiti and Palash Dey. Query complexity of tournament solutions. *Theoretical Computer Science*, 991:114422, 2024.

[21] Andy Shih, Arthur Choi, and Adnan Darwiche. A symbolic approach to explaining bayesian network classifiers. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, 2018.

[22] Emanuel Sperner. Ein satz über untermengen einer endlichen menge. *Mathematische Zeitschrift*, 27(1):544–548, 1928.

Detecting Anomalies Using Graph Neural Networks: A Review

Sena Ozgunay[1,2], Louise Travé-Massuyès[1], Jean-Michel Loubes[2,3], Raul Sena Ferreira[4]

[1] LAAS-CNRS, University of Toulouse, CNRS, Toulouse, France
[2] Institut de Mathématiques de Toulouse, France
[3] INRIA, France
[4] Continental Automotive France, Toulouse, France

sozgunay@laas.fr, louise@laas.fr, jean-michel.a.loubes@inria.fr, raul.ferreira@continental.com

Résumé

La détection d'anomalies vise à identifier des comportements atypiques au sein des systèmes complexes. Parmi les différentes approches développées dans ce domaine, les réseaux de neurones graphiques (GNN) se distinguent par leur efficacité. Dans cet article, nous proposons une revue des méthodes fondées sur les GNN pour la détection d'anomalies, et introduisons une nouvelle taxonomie, construite autour des mécanismes de prédiction d'anomalies utilisés.

Mots-clés

Détection d'anomalies, réseaux de neurones graphiques, séries temporelles.

Abstract

Anomaly detection is the process of identifying unusual behaviors in systems. In this wide-ranging field, graph neural networks (GNNs) are highly effective compared to the other proposed approaches in the literature. This article summarizes the representative GNN-based methods for anomaly detection and proposes a novel taxonomy based on how these methods predict anomalies.

Keywords

Anomaly detection, graph neural network, time series.

1 Introduction

Real-world systems are generally complex and include vast numbers of interconnected sensors. Anomaly detection is an important task but is difficult in such complex systems since capturing relationships between sensors and detecting sensors deviating from these is challenging [10]. Graph Neural Networks (GNNs) [13] have emerged as a robust AI method for anomaly detection in a wide range of application domains, including spam detection [4], fraud detection [11], cybersecurity [40], industrial system supervision [10], etc. In a financial transaction network, where nodes represent user accounts and edges represent transactions, an anomaly may correspond to a fraudulent account that interacts abnormally with others, such as making high-frequency or high-value transactions to unrelated accounts [38].

GNNs operate on graph-structured data, which can be constructed from various types of raw data such as tabular data, images, or time series, when a graph input is not available, and they aim to learn data representations (or embeddings) by capturing complex relationships from the graph. In many real-world scenarios, time series data constitute an essential setting for anomaly detection since anomalies may occur over time. Time series often involve multiple signals from system sensors that may exhibit spatial dependencies among sensors and temporal dependencies across different time steps [23]. In an industrial sensor network, time series data from different sensors can be modeled as a graph, where nodes represent sensor variables, and edges reflect their spatial or temporal correlations. An anomaly might occur as a sudden deviation in one sensor's behavior that breaks the normal relationship with others over time [10]. In all settings, GNNs offer a unified framework for capturing the underlying spatial or temporal patterns and identifying abnormal behaviors.

The survey [24] reviews GNN-based anomaly detection techniques. [23] analyzes GNN methods for different time series tasks, including anomaly detection. Moreover, [26] proposes a survey of anomaly detection approaches in graph data based on different types of anomalies, graphs, and graph-based methods, including GNNs. Unlike the recent survey [43] that summarizes different types of GNN-based anomaly detection methods based primarily on data distribution assumptions (such as whether normal or anomalous data boundaries are clearer), we propose an embedding-centric taxonomy focused on how the learned embeddings by GNNs are utilized for anomaly detection independently of the original anomaly boundary assumptions and the input raw data nature. Specifically, we classify representative GNN-based methods for anomaly detection based on whether embeddings are directly used for classification, for target value prediction, or for embedding similarity to identify anomalies, offering a complementary perspective to the existing literature. We also discuss the particular case of anomaly detection with GNN when the inputs are time series.

The rest of this paper is organized as follows. Section 2

gives a background on graphs and GNNs. Section 3 proposes the taxonomy based on different GNN methods for anomaly detection. Section 4 discusses the particular case of anomaly detection with GNNs where the inputs are time series, and Section 5 presents the conclusions and future directions.

2 Background

A graph is denoted as $G = (V, E)$, where V is the set of nodes, and E is the set of edges. Graphs can be presented by an adjacency matrix $A \in \mathbb{R}^{|V|\times|V|}$ where $A_{i,j} = 1$ if $(v_i, v_j) \in E$ and $A_{i,j} = 0$ otherwise. For a node $v_i \in V$, the set of neighbors $N(v_i)$ contains the nodes adjacent to v_i. A graph can have attributes or feature information on nodes/edges. Generally, a node feature matrix is denoted as $X \in \mathbb{R}^{|V|\times d}$, where d is the number of attributes [17]. A graph can be directed or undirected; that is, edges of directed graphs are directed from one node to another [7]. The graph can be homogeneous or heterogeneous. If only one type of node exists and one type of edge, the graph is homogeneous; otherwise, it is heterogeneous [29]. In a temporal aspect, a graph can be static or dynamic. In dynamic graphs, input features or the graph structure vary over time, reflecting the changing relationships and attributes within the network, while in static graphs, the graph structure remains fixed [17].

GNNs process graph-structured data to capture complex relationships and extract significant system information. The key principle behind GNNs is aggregating information from a node's local neighborhood to learn node representations or embeddings. That is, a node's state is affected by the states of its neighboring nodes.

For a given graph $G = (V, E, X)$, where $X \in \mathbb{R}^{|V|\times d}$ is the input feature matrix, x_i is the d-dimensional input node feature vector of v_i, GNN learns the node representation h_i for each $v_i \in V$. During a L-layer GNN training, the node embedding at the l-th GNN layer $h_i^{(l)}$ is updated based on aggregated information from v_i's neighborhood $N(v_i)$. This update can be formalized as follows:

$$m_i^{(l)} = \text{AGG}^{(l)}\left(\left\{h_j^{(l-1)}, \forall v_j \in N(v_i)\right\}\right), \quad (1)$$

$$h_i^{(l)} = \text{UPDATE}^{(l)}\left(h_i^{(l-1)}, m_i^{(l)}\right) \quad (2)$$

where UPDATE and AGG are differentiable functions with AGG as permutation invariant that can operate on multiple inputs (e.g., element-wise sum, mean, or max). This function takes node embeddings of v_i's neighbors as input and computes a message $m_i^{(l)}$ by aggregating the information coming from neighbors. Then, the UPDATE function combines the message and previous v_i's node embedding to obtain the current node embedding.

We note that the initial embeddings at $l = 0$ are input node features, i.e., $h_i^{(0)} = x_i$. After L iterations, the output of GNN is a set of node embeddings $\{h_i^{(L)}, \forall v_i \in V\}$ where each obtained node embedding includes information about its L-hop neighborhood.

Regarding the Equation 1, a simple GNN message-passing mechanism [13] can be constructed as follows:

$$h_i^{(l)} = \sigma\left(W_{self}^{(l)} h_i^{(l-1)} + W_{neigh}^{(l)} \sum_{v_j \in N(v_i)} h_j^{(l-1)} + b^{(l)}\right) \quad (3)$$

where $W_{self}^{(l)}, W_{neigh}^{(l)} \in \mathbb{R}^{d_{(l)}\times d_{(l-1)}}$ are trainable weight matrices with $d_{(l)}, d_{(l-1)}$ embedding dimensions of the corresponding layer and σ is a non-linearity function. $b^{(l)} \in \mathbb{R}^{d_{(l)}}$ is the bias term, which is often neglected. This simple GNN formulation can be decomposed as:

$$m_i^{(l)} = \sum_{v_j \in N(v_i)} h_j^{(l-1)}, \quad (4)$$

$$\text{UPDATE}^{(l)}(h_i^{(l-1)}, m_i^{(l)}) = \sigma(W_{self}^{(l)} h_i^{(l-1)} + W_{neigh}^{(l)} m_i^{(l)}) \quad (5)$$

Among the different GNN formulations, classical GNN layers are Graph Convolutional Networks (GCNs) [25], GraphSAGE (Graph Sample and Aggregate) [16], and Graph Attention Networks (GATs) [36], each with the functions for aggregation and message computation.

GCN and GraphSAGE suppose that all neighbors $v_j \in N(v_i)$ are equally important to the node v_i, that is, the messages coming from neighbors have the same weighting factor (importance), which is based on the node degree. However, some neighbors might be more relevant in some cases than others. Thus, Graph Attention Networks (GATs) introduce attention mechanisms to learn different weights for different neighbors [36]. The node embedding of v_i at the l-th GAT layer is defined as:

$$h_i^{(l)} = \sigma\left(\sum_{v_j \in N(v_i)} \alpha_{ij} W^{(l)} h_j^{(l-1)}\right) \quad (6)$$

where α_{ij} are attention weights across a pair of nodes v_i, v_j. To determine α_{ij}, we first compute attention coefficients e_{ij}, which indicate the importance of v_j to node v_i:

$$e_{ij} = \text{a}(W^{(l)} h_i^{(l-1)}, W^{(l)} h_j^{(l-1)}) \quad (7)$$

where a is an attention mechanism (e.g., a single-layer neural network). Then we normalize e_{ij} using the softmax function to obtain the final attention weight α_{ij}:

$$\alpha_{ij} = \text{softmax}_j(e_{ij}) = \frac{\exp(e_{ij})}{\sum_{v_k \in N(v_i)} \exp(e_{ik})} \quad (8)$$

3 Anomaly Detection with GNNs

In graphs, anomalies may appear at the node-level, edge-level, or graph-level. At the node level, an anomaly can be seen on node attributes as a significantly different characteristic. Regarding neighborhood interaction, an anomaly can be present in structural patterns such as unusual degrees of connections or unexpected labels, etc. An edge can be anomalous at the edge-level if it has an abnormal edge weight compared to the distribution of the

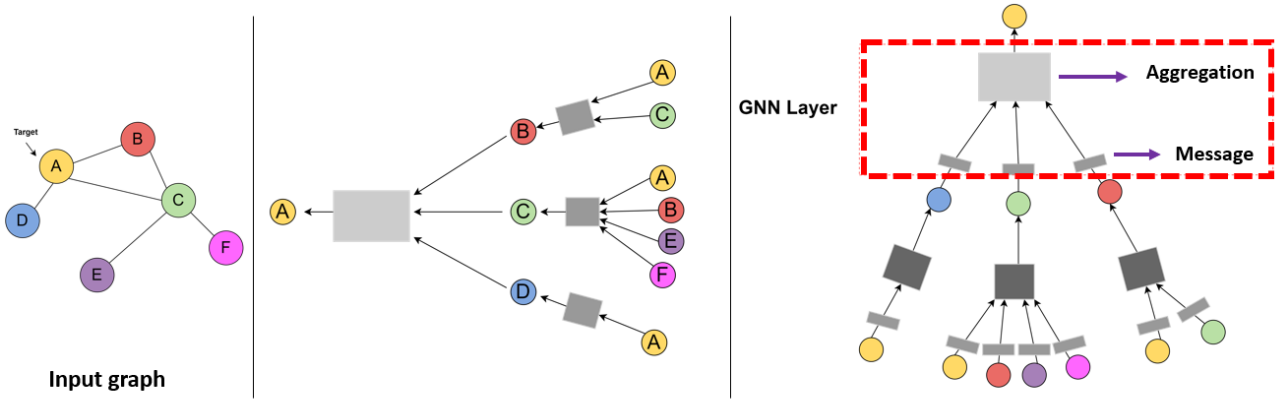


Figure 1: **Overview of GNN's two-layer version of a message-passing mechanism.** It shows the aggregation of messages from A's local neighbors on the graph, which are B, C, and D. Concept adapted from [17].

other edge weights. The discovery of an unexpected link or absence of any edge might also be pointed out as an anomaly. At the graph-level, a graph can be anomalous if it has unusual structural patterns or unexpected changes compared to other input graphs [26].

In anomaly detection, the training dataset may include normal and anomalous instances (fully or partially labeled) or only normal instances. In the former case, GNN can directly learn to identify instances as normal or anomalous based on known labels, while in the latter, GNN learns the expected behavior. The embeddings generated by GNN can be fed into the prediction phase using a classifier that performs a prediction function to decide whether an instance is anomalous. Further, the embeddings may be used to predict a value of a specific instance. Moreover, they can be compared with each other through similarities between them to infer anomaly information. In this survey we introduce a taxonomy that reflects these three ways of leveraging the embeddings learned by GNN for anomaly detection: via direct prediction, comparison with target values, or evaluation of embedding similarities.

3.1 Direct Prediction of Anomalies

The first category focuses on direct prediction of anomalies using a classifier function. Most methods in this category perform a training process under supervision, i.e., supervised (labeled normal and abnormal data) or semi-supervised (only labeled standard data). Depending on the task level, a given entity (node, edge, or graph)'s learned representation by GNN is used as input by a classifier that decides whether the entity is anomalous.

[4], **GEM** [29], and **HACUD** [20] aim to detect spam profiles/malicious accounts in social networks by considering a graph where accounts are nodes and communications between accounts are edges. Learned node embeddings by GNN-based models are passed to a classifier to predict whether an account is malicious.

SemiGNN [38] combines GNN with attention mechanisms to detect fraud in financial transaction networks that contain different types of relationships. Final user embeddings are obtained by combining the learned relation-specific embeddings. Then the users are classified as fraudulent or not. Likewise, **CARE-GNN** [11] aims to detect camouflaged fraudsters by applying a similar approach on a multi-relation graph. On the other hand, **DeepHGNN** [39] detects anomalies in program executions by modeling the different relationships (events) between different node types (programs, files, etc.). It models the program's behavior by an attention-based GNN model and checks whether the given event corresponds to that program. **GAS** [28] identifies spam reviews using GCNs to model the interactions between reviews, users, and products. GAS learns embeddings for each review, which are then fed to a classifier to decide whether it is a spam comment. **iDetective** [46] introduces an attributed heterogeneous information network to identify key players in hack forums. Learned user embeddings are fed to Support Vector Machine (SVM) to classify the user as a key player (anomalous) or a non-key player (standard). **MVAN** [35] employs GNNs to detect anomalous players with fraudulent money trading activities in online games by leveraging multiple player relationships.

AddGraph [49] models a dynamic graph and identifies abnormal changes in the graph's connections for telecom and social networks. Learned node embeddings are then used to classify edges as anomalous or normal based on the computed edge anomaly score at the current timestamp.

3.2 Target Value Prediction

The GNN-based methods in this category predict future time series values by learning the expected behavior of systems. Due to that, they can be considered one-class methods since the training dataset contains observations from only one class (normal). An anomaly can be predicted via a forecasting-based model that predicts the target value at $t+1$ based on historical normal data on a temporal window $[t-w,\ldots,t]$. When the predicted value differs too much from the observed one, an anomaly is detected [34].

This step usually involves a single fully connected layer or Multilayer perceptron (MLP).

Anomalies can also be identified by a reconstruction-based model, which aims to reconstruct given inputs. The model can approximate well the normal inputs, but it fails to reconstruct abnormal ones [23].

MTAD-GAT [48], **GTAD** [14], and **HAD-MDGAT** [52] apply jointly a reconstruction module and a forecasting module after obtaining node embeddings from a GAT-based model. For each node, an anomaly score is defined based on the predicted value by the forecasting module and the reconstruction probability obtained from the reconstruction module. Similarly, [12] aims to detect leaks in water distribution networks (WDNs), where nodes are sensors and edges are pipes. With partially observed nodal pressures, the model reconstructs and predicts complete nodal pressure values, respectively. Then, edge-wise prediction errors are computed to detect leakages. On the other hand, **GReLeN** [45] uses only a reconstruction module after combining Variational Auto Encoder (VAE) and an attention-based GNN for learning graph structure.

GTA [7], **GDN** [10], **CST-GL** [50], and **STGDNN** [15] use a forecasting approach and aim to learn a graph structure along with the GNN-based model. For learning node embeddings, while GTA applies a transformer-based GNN, GDN and STGDNN use a GAT-based model.

DyGraphAD [21] models a multivariate time series as a set of dynamic correlation graphs. At the forecasting step, it predicts the dynamic correlation graph and the sensor values used to compute an anomaly score for the timestamp. On the other hand, **DGNN** [5] leverages dynamic neighborhood subgraphs based on correlations, and a GAT-based model to forecast sensor values. [27] also employs a forecasting approach, but for finding anomalies in interconnected traffic flows. The model uses GNN with attention mechanism to extract a context of the flows based on their past traffic activity and then forecast the next value of the flow. **GSC-MAD** [47] forecasts sensor values by analyzing changes in learning graph structure. An anomaly score at the timestamp is calculated by combining prediction deviations and differences between the new and stable graph structures. **GST-Pro** [51] deals with missing values by modeling dynamic graph neural controlled differential equations on multi-sensor data to learn embeddings for each sensor, which are then used for forecasting sensor values.

3.3 Embedding Similarity

In the third category, the focus is put on learned embeddings. Anomalies are detected by analyzing how similar or dissimilar embeddings are. This can be done using a clustering approach where nodes with similar embeddings are grouped into clusters, or using similarity metrics where the distance or similarity between node embeddings can indicate abnormality.

MatchGNet [40] applies a graph-matching approach based on the similarities between known benign and unknown programs. For a given unknown program, MatchGNet computes its graph embedding, and then the cosine similarity between the ones of the existing programs is computed. Based on a threshold, the unknown anomalous programs are identified. From normal data, **OCGNN** [41] learns the node embeddings to find a minimal hypersphere that encloses them. The node is detected as anomalous if a given node's embedding is outside of this hypersphere. **GraphSAD** [6] aims to detect a sequence of abnormal observations in a time series split into subsequences. An abnormal subsequence is detected by comparing the learned embedding with reference subsequence embeddings based on its neighbors.

While our proposed taxonomy offers a general perspective on how GNN-learned embeddings can be used for anomaly detection, the specific nature of the input data can significantly influence how these methods are structured. In particular, time series data often exhibit introduce temporal dependencies that require appropriate graph construction and adaptation of GNN architectures. In the following section, we analyze a subset of the surveyed GNN-based anomaly detection methods that consider time series as input, and discuss how they represent spatial and temporal dependencies within graph-based frameworks.

4 Anomaly Detection with GNNs in Time Series

Time series is a key data type in most systems that record masses of data over time [18, 23]. Formally we denote a time series X as:

$$X = \{X_1, X_2, ..., X_T\} \in \mathbb{R}^{N \times T} \tag{9}$$

where $X_t = (x_{t,1}, x_{t,2}, ..., x_{t,N})$ is the vector of values across N variables at time t, and $x_{t,i}$ is the value of the i-th variable at time t [44]. Among the methods mentioned in Section 3, [29, 35, 49, 42] and all the methods presented in Section 3.2 focus on anomaly detection in multivariate time series, while [6] deals with univariate time series.

Anomalies in time series may occur as a single point (e.g., at a given time step), referred to as point anomalies, or as a subsequence that deviates significantly from the expected behavior [44], referred to as collective anomalies [3]. For a long time series, fixed-length samples generated by a sliding window of length T are used as inputs. For point anomalies, the objective of the anomaly detection in multivariate time series is to decide if a point at a given timestamp i is an anomaly or not, i.e., compute an output vector $y \in \mathbb{R}^T$, where $y_i \in \{0, 1\}$ indicates whether the timestamp i is an anomaly [48].

In complex systems, where time series reveal the evolution of measurements in multiple sensors, anomalies may arise not only from the characteristics of variables but also from the relationships between variables [48]. Due to the graph's capacity to exhibit complex data structures, GNNs have emerged as a powerful tool for revealing complex relationships on non-Euclidean data [18]. GNNs can capture the links between variables within a multivariate time series (spatial dependency) and dependencies between

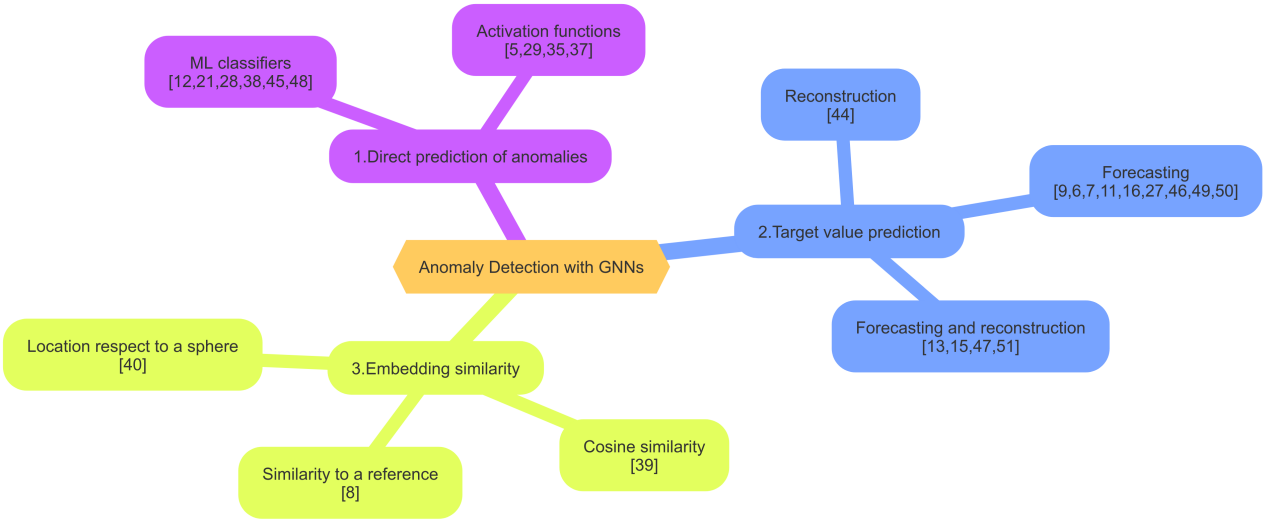


Figure 2: **Taxonomy of the representative GNN-based methods for anomaly detection.** Direct prediction of anomalies can be done by interpreting the output layer of the GNN where an activation function (softmax, ReLU, etc.) is applied, or the learned embeddings can be fed into a separate classifier (MLP, logistic regression, SVM, etc.). The methods in the second category can use a forecasting or reconstruction-based approach. Further, some methods combine these two techniques. The third category contains methods that compare the similarity of the learned embeddings of GNN to detect anomalies. They can compare by using the cosine similarity, the similarity to a reference, or the similarity based on a defined sphere location.

different time points (temporal dependency). A recent survey by [23] offers a comprehensive overview of GNN methods in time series for several major tasks, including anomaly detection. Other recent surveys [18, 44] review the time series anomaly detection techniques using graphs and deep learning methods (including GNNs).

A graph structure can be learned directly from input time series data or constructed by computing several metrics or functions, such as spatial proximity, attribute similarity, etc. Spatial proximity aims to compute proximity between nodes' locations based on, for example, the shortest path between nodes [23]. Meanwhile, attribute similarity determines the similarity between node attributes and then connects similar nodes. This can be done, among other ways, by computing Euclidean distance, cosine similarity, Pearson Correlation Coefficients (PCC), or dynamic time warping (DTW) [31] between time series [18, 23].

Multivariate time series data can be introduced as a *spatial-temporal graph*. It is a set of graphs $\mathcal{G} = \{\mathcal{G}_1, \mathcal{G}_2, ..., \mathcal{G}_T\}$ with $\mathcal{G}_t = (A_t, X_t)$ a graph at time t where $A_t \in \mathbb{R}^{N \times N}$ is the adjacency matrix and $X_t \in \mathbb{R}^{N \times d}$ is the node-feature matrix with d being the feature dimension [23]. The adjacency matrix A_t may remain fixed (static case) or evolve in time (dynamic case). Thus, spatial-temporal graphs allow GNN to capture in an explicit way both:

- Spatial (inter-variable) dependencies: Relationships between nodes (e.g., sensors, devices).
- Temporal (intra-variable) dependencies: Evolving patterns over time (e.g., periodic behaviors or sequential dependencies) [44].

Several techniques exist in the spatial-temporal graph structures to integrate temporal dependency into GNNs. One approach is to include time-specific features for each node, for example, the average or variance of past values [22]. Another approach uses a sliding window, where the data can be divided into overlapping fixed-length windows, creating a separate graph for each window [48]. This allows the GNN to learn spatial relationships in each window while implicitly considering the temporal context. Another approach is to insert some hybrid architectures that combine GNNs with temporal modeling methods such as RNN [1], LSTM [19], or GRU [8], allowing them to model time dependencies explicitly. Moreover, Temporal GAT (TGAT) [9], which employs a self-attention mechanism with a functional time encoding, or Temporal Convolutional Networks (TCNs) [2], which use a 1D fully-convolutional network (FCN) [30] architecture and causal convolutions, might be applied to extract temporal dependencies. In such hybrid architectures, the spatial and temporal layers can be combined directly (e.g., using GCNs for spatial learning and GRUs for temporal patterns) or by stacking the temporal layer on the GNN spatial layer (e.g., processing a GRU after a GNN) [23, 44].

[12, 27], GEM [29], GDN [10], and GSC-MAD [47] can extract spatial dependencies between nodes, but they ignore temporal dependencies between different timestamps. Meanwhile, AddGraph [49] extends GCN to a temporal GCN with an attentional GRU to capture the nodes' long- and short-term temporal patterns. Similarly, GReLeN [45] applies an architecture combined with GRU and GNN. However, MTAD-GAT [48] employs a GAT

model as a spatial-temporal encoder with spatial and temporal GAT layers. The spatial GAT learns spatial dependencies between sensors, while the temporal GAT extracts temporal relationships from a given input in a sliding window; it considers a fully connected graph where timestamps in the sliding window are nodes. Concatenated learned embeddings across layers are fed into a GRU as a forecasting module to capture long-term temporal dependencies as in HAD-MDGAT [52]. However, in HAD-MDGAT, GRU output is used in forecasting and reconstruction modules for model inference. Besides, Event2Graph [42] extends the TGAT architecture to model the temporal dependency on the node level. At the same time, GTA [7] employs a transformer-based architecture within a hierarchical causal convolution to extract temporal dependencies. GTAD [14] uses a TCN to extract temporal information and a variant of GAT for spatial dependencies. The obtained embeddings are fed into the forecasting and reconstruction models, which both contain a TCN module to avoid losing temporal information.

DyGraphAD [21] applies DTW for constructing dynamic correlation graphs for each input subsequence from a multivariate time series that computes similarities between pairwise time series. Then, embeddings of these graphs and representations of subsequences are learned via a temporal-based GNN architecture. CST-GL [50] combines GCN and TCN for spatial and temporal convolutions, respectively. Similarly, DGNN [5] uses GRU, a forecasting module to predict values at the next time step and extract temporal dependency. Further, STGDNN [15] extends the GDN [10] architecture as a spatial-temporal GNN framework by inserting spatial-temporal relationships in the graph structure. GST-Pro [51] exploits spatial and temporal dependencies by modeling Dynamic Graph Neural Controlled Differential Equations.

5 Conclusion

In this paper, we reviewed the representative GNN-based methods for anomaly detection and proposed a taxonomy based on how these use learned embeddings for predicting anomalies. We discussed the specific case of the GNNs for anomaly detection, where the inputs are time series.

This review highlights that GNNs are well-suited solutions for the anomaly detection task due to their capacity for modeling complex systems. The message-passing technique allows GNNs to generate embeddings that contain rich neighborhood information. These learned embeddings are then used in many ways to identify anomalies. Our taxonomy helps researchers better understand anomaly detection using GNNs by clarifying the differences in the ways anomalies are identified using the generated embeddings from GNNs.

In the anomaly detection, it's crucial to understand why a particular event is flagged as anomalous and how to validate the relevance of these detected anomalies [33]. Therefore, as future work, we intend to study the anomaly detection in industrial production lines using GNNs, and explain the detected anomalies using counterfactual explanations by focusing on alternative scenarios [37]. That is, we aim to generate counterfactual explanations that would indicate what changes to an anomalous instance would make it be considered normal [33]. Another research direction is to explore how to determine which system components are malfunctioning through model-based diagnosis [32].

Acknowledgments

This work is supported by the AI Interdisciplinary Institute ANITI funded by the France 2030 program under the Grant agreements n°ANR-19-PI3A-0004 and n°ANR-23-IACL-0002.

References

[1] Oludare Isaac Abiodun, Aman Jantan, Abiodun Esther Omolara, Kemi Victoria Dada, Nachaat AbdElatif Mohamed, and Humaira Arshad. State-of-the-art in artificial neural network applications: A survey. *Heliyon*, 4(11), 2018.

[2] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *CoRR*, abs/1803.01271, 2018.

[3] Mohammad Braei and Sebastian Wagner. Anomaly detection in univariate time-series: A survey on the state-of-the-art. *CoRR*, abs/2004.00433, 2020.

[4] Anshika Chaudhary, Himangi Mittal, and Anuja Arora. Anomaly detection using graph neural networks. In *2019 international conference on machine learning, big data, cloud and parallel computing (COMITCon)*, pages 346–350. IEEE, 2019.

[5] Bowen Chen, Hancheng Lu, Yuang Chen, Haoyue Yuan, and Minghui Wang. Dgnn: Dynamic graph neural networks for anomaly detection in multivariate time series. In *SEKE*, pages 415–420, 2023.

[6] Weiqi Chen, Zhiqiang Zhou, Qingsong Wen, and Liang Sun. Time series subsequence anomaly detection via graph neural networks, 2023.

[7] Zekai Chen, Dingshuo Chen, Xiao Zhang, Zixuan Yuan, and Xiuzhen Cheng. Learning graph structures with transformer for multivariate time-series anomaly detection in iot. *IEEE Internet of Things Journal*, 9(12):9179–9189, 2021.

[8] Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder–decoder approaches. In *SSST@EMNLP*, 2014.

[9] da Xu, chuanwei ruan, evren korpeoglu, sushant kumar, and kannan achan. Inductive representation learning on temporal graphs. In *International Conference on Learning Representations*, 2020.

[10] Ailin Deng and Bryan Hooi. Graph neural network-based anomaly detection in multivariate time series. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 4027–4035, 2021.

[11] Yingtong Dou, Zhiwei Liu, Li Sun, Yutong Deng, Hao Peng, and Philip S Yu. Enhancing graph neural network-based fraud detectors against camouflaged fraudsters. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 315–324, 2020.

[12] Garðar Örn Garðarsson, Francesca Boem, and Laura Toni. Graph-based learning for leak detection and localisation in water distribution networks. *IFAC-PapersOnLine*, 55(6):661–666, 2022.

[13] Marco Gori, Gabriele Monfardini, and Franco Scarselli. A new model for learning in graph domains. In *Proceedings. 2005 IEEE international joint conference on neural networks, 2005.*, volume 2, pages 729–734. IEEE, 2005.

[14] Siwei Guan, Binjie Zhao, Zhekang Dong, Mingyu Gao, and Zhiwei He. Gtad: Graph and temporal neural network for multivariate time series anomaly detection. *Entropy*, 24(6):759, 2022.

[15] Ziyu Guo, Weiyang Kong, and Yubao Liu. Spatial-temporal graph deviation neural networks for anomaly detection in time series. In *Journal of Physics: Conference Series*, volume 2735, page 012016. IOP Publishing, 2024.

[16] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

[17] William L Hamilton. *Graph representation learning*. Morgan & Claypool Publishers, 2020.

[18] Thi Kieu Khanh Ho, Ali Karami, and Narges Armanfard. Graph-based time-series anomaly detection: A survey. *arXiv preprint arXiv:2302.00058*, 2023.

[19] S Hochreiter. Long short-term memory. *Neural Computation MIT-Press*, 1997.

[20] Binbin Hu, Zhiqiang Zhang, Chuan Shi, Jun Zhou, Xiaolong Li, and Yuan Qi. Cash-out user detection based on attributed heterogeneous information network with a hierarchical attention mechanism. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 946–953, 2019.

[21] Xiangheng Huang, Ningjiang Chen, Ziyue Deng, and Suqun Huang. Multivariate time series anomaly detection via dynamic graph attention network and informer. *Applied Intelligence*, 54(17–18):7636–7658, June 2024.

[22] Guangyin Jin, Yuxuan Liang, Yuchen Fang, Zezhi Shao, Jincai Huang, Junbo Zhang, and Yu Zheng. Spatio-temporal graph neural networks for predictive learning in urban computing: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

[23] Ming Jin, Huan Yee Koh, Qingsong Wen, Daniele Zambon, Cesare Alippi, Geoffrey I Webb, Irwin King, and Shirui Pan. A survey on graph neural networks for time series: Forecasting, classification, imputation, and anomaly detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[24] Hwan Kim, Byung Suk Lee, Won-Yong Shin, and Sungsu Lim. Graph anomaly detection with graph neural networks: Current status and challenges. *IEEE Access*, 10:111820–111829, 2022.

[25] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

[26] Prabin B Lamichhane and William Eberle. Anomaly detection in graph structured data: A survey. *arXiv preprint arXiv:2405.06172*, 2024.

[27] Hamid Latif, José Suárez-Varela, Albert Cabellos-Aparicio, and Pere Barlet-Ros. Detecting contextual network anomalies with graph neural networks. In *Proceedings of the 2nd on Graph Neural Networking Workshop 2023*, pages 25–30, 2023.

[28] Ao Li, Zhou Qin, Runshi Liu, Yiqun Yang, and Dong Li. Spam review detection with graph convolutional networks. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 2703–2711, 2019.

[29] Ziqi Liu, Chaochao Chen, Xinxing Yang, Jun Zhou, Xiaolong Li, and Le Song. Heterogeneous graph neural networks for malicious account detection. In *Proceedings of the 27th ACM international conference on information and knowledge management*, pages 2077–2085, 2018.

[30] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[31] Chotirat Ann Ratanamahatana and Eamonn Keogh. Everything you know about dynamic time warping is wrong. In *Third workshop on mining temporal and sequential data*, volume 32. Citeseer, 2004.

[32] Raymond Reiter. A theory of diagnosis from first principles. *Artificial intelligence*, 32(1):57–95, 1987.

[33] Max Schemmer, Joshua Holstein, Niklas Bauer, Niklas Kühl, and Gerhard Satzger. Towards meaningful anomaly detection: The effect of counterfactual explanations on the investigation of anomalies in multivariate time series. *CoRR*, abs/2302.03302, 2023.

[34] Sebastian Schmidl, Phillip Wenig, and Thorsten Papenbrock. Anomaly detection in time series: a comprehensive evaluation. *Proceedings of the VLDB Endowment*, 15(9):1779–1797, 2022.

[35] Jianrong Tao, Jianshi Lin, Shize Zhang, Sha Zhao, Runze Wu, Changjie Fan, and Peng Cui. Mvan: Multi-view attention networks for real money trading detection in online games. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2536–2546, 2019.

[36] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

[37] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

[38] Daixin Wang, Jianbin Lin, Peng Cui, Quanhui Jia, Zhen Wang, Yanming Fang, Quan Yu, Jun Zhou, Shuang Yang, and Yuan Qi. A semi-supervised graph attentive network for financial fraud detection. In *2019 IEEE international conference on data mining (ICDM)*, pages 598–607. IEEE, 2019.

[39] Shen Wang, Zhengzhang Chen, Ding Li, Zhichun Li, Lu-An Tang, Jingchao Ni, Junghwan Rhee, Haifeng Chen, and Philip S Yu. Attentional heterogeneous graph neural network: Application to program reidentification. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pages 693–701. SIAM, 2019.

[40] Shen Wang and S Yu Philip. Heterogeneous graph matching networks: Application to unknown malware detection. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 5401–5408. IEEE, 2019.

[41] Xuhong Wang, Baihong Jin, Ying Du, Ping Cui, Yingshui Tan, and Yupu Yang. One-class graph neural networks for anomaly detection in attributed networks. *Neural computing and applications*, 33:12073–12085, 2021.

[42] Yuhang Wu, Mengting Gu, Lan Wang, Yusan Lin, Fei Wang, and Hao Yang. Event2graph: Event-driven bipartite graph for multivariate time series forecasting and anomaly detection. In *AMLTS*, 2022.

[43] Ziqi Yuan, Qingyun Sun, Haoyi Zhou, Minglai Shao, and Xingcheng Fu. A comprehensive survey on gnn-based anomaly detection: taxonomy, methods, and the role of large language models. *International Journal of Machine Learning and Cybernetics*, pages 1–26, 2025.

[44] Zahra Zamanzadeh Darban, Geoffrey I Webb, Shirui Pan, Charu Aggarwal, and Mahsa Salehi. Deep learning for time series anomaly detection: A survey. *ACM Computing Surveys*, 57(1):1–42, 2024.

[45] Weiqi Zhang, Chen Zhang, and Fugee Tsung. Grelen: Multivariate time series anomaly detection from the perspective of graph relational learning. In *IJCAI*, pages 2390–2397, 2022.

[46] Yiming Zhang, Yujie Fan, Yanfang Ye, Liang Zhao, and Chuan Shi. Key player identification in underground forums over attributed heterogeneous information network embedding framework. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 549–558, 2019.

[47] Zhen Zhang, Zhiqiang Geng, and Yongming Han. Graph structure change-based anomaly detection in multivariate time series of industrial processes. *IEEE Transactions on Industrial Informatics*, 2024.

[48] Hang Zhao, Yujing Wang, Juanyong Duan, Congrui Huang, Defu Cao, Yunhai Tong, Bixiong Xu, Jing Bai, Jie Tong, and Qi Zhang. Multivariate time-series anomaly detection via graph attention network. In *2020 IEEE international conference on data mining (ICDM)*, pages 841–850. IEEE, 2020.

[49] Li Zheng, Zhenpeng Li, Jian Li, Zhao Li, and Jun Gao. Addgraph: Anomaly detection in dynamic graph using attention-based temporal gcn. In *IJCAI*, volume 3, page 7, 2019.

[50] Yu Zheng, Huan Yee Koh, Ming Jin, Lianhua Chi, Khoa T Phan, Shirui Pan, Yi-Ping Phoebe Chen, and Wei Xiang. Correlation-aware spatial–temporal graph learning for multivariate time-series anomaly detection. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[51] Yu Zheng, Huan Yee Koh, Ming Jin, Lianhua Chi, Haishuai Wang, Khoa T Phan, Yi-Ping Phoebe Chen, Shirui Pan, and Wei Xiang. Graph spatiotemporal process for multivariate time series anomaly detection with missing values. *Information Fusion*, 106:102255, 2024.

[52] Liwen Zhou, Qingkui Zeng, and Bo Li. Hybrid anomaly detection via multihead dynamic graph attention networks for multivariate time series. *IEEE Access*, 10:40967–40978, 2022.

Combining Knowledge Graphs and Generative AI to explain the consequences of climatic events on the health of Maasai herders

Houdhem Assoudi[1], Vincent Armant[1], Amira Mouakher[1], Emmanuel Roux [1], Jean-Christophe Desconnets[1] and Victor N.Mose[2]

[1] ESPACE-DEV UMR 228, IRD, UM, UPVD, UA, UG, UR, Montpellier, France
[2] African Conservation Centre, Kenya

houdhem.assoudi@ird.fr

Résumé

Ce projet de doctorat débuté en janvier 2025, vise à concevoir un système fondé sur les connaissances pour expliquer l'impact des événements climatiques extrêmes et des changements environnementaux sur la santé des communautés Maasai au Kenya et en Tanzanie. Il repose sur l'utilisation de graphes de connaissances pour structurer et relier des données hétérogènes (santé, climat, pratiques locales) et sur un système de génération augmentée par récupération (RAG). Le graphe de connaissances sert de référentiel structuré, permettant d'extraire et d'organiser les informations pertinentes. Le RAG exploite ces connaissances pour produire des explications adaptées aux différents utilisateurs (communautés locales, chercheurs, décideurs), garantissant ainsi des réponses contextualisées, traçables et compréhensibles. Menée dans le cadre du projet MOSAIC*, cette recherche combine intelligence artificielle, web sémantique et santé environnementale pour renforcer la compréhension scientifique et soutenir une prise de décision éclairée.*

Mots-clés

Ontologies, graphes de connaissances, génération augmentée par récupération, IA explicable, événements climatiques extrêmes, indicateurs de santé, communautés Maasai

Abstract

This PhD project, started in January 2025, aims to develop a knowledge-driven system to explain the impact of extreme climatic events and environmental changes on the health of Maasai communities in Kenya and Tanzania. It leverages knowledge graphs to structure and interconnect heterogeneous data sources (health, climate, local practices) and integrates a Retrieval-Augmented Generation (RAG) system. The knowledge graph serves as a structured repository, facilitating the extraction and organization of relevant information. The RAG system utilizes these structured insights to generate context-aware, traceable, and comprehensible explanations tailored to different stakeholders, including local communities, researchers, and policymakers. Conducted within the MOSAIC *project, this research combines artificial intelligence, semantic web technologies, and environmental health studies to enhance scientific understanding and support informed decision-making.*

Keywords

Ontologies, knowledge graphs, retrieval-augmented generation, explainable AI, severe climate events, health indicators, Maasai communities

1 Introduction

Environmental changes and extreme climatic events, such as prolonged droughts, increasingly threaten human health, particularly in vulnerable regions. Climate variability has been linked to malnutrition, infectious disease outbreaks, and mental health disorders, particularly in pastoralist communities that depend on fragile ecosystems for their livelihoods [10]. In East Africa, Maasai communities face severe challenges, including malnutrition, infectious diseases, and socio-economic instability due to declining pasture availability and water scarcity. A study published in PLOS ONE [6] revealed that four out of five Maasai households in Tanzania experience severe food insecurity, with nearly 60% of Maasai children being chronically malnourished. To address these complex interactions requires an interdisciplinary approach that integrates environmental, epidemiological, and socio-economic data.

This research, conducted within the MOSAIC[1] Horizon Europe project, adopts a Planetary Health perspective to develop an explainable knowledge system that models climate-driven health impacts and ensures accessibility for scientists, policymakers, and local communities. While previous studies have examined climate-health links in East Africa [8], structured and explainable AI (XAI) applications in this context remain underexplored.

Most existing approaches [2] focus on data collection and retrieval, but few systems offer context-aware, explainable insights tailored to different user needs. There is a pressing need for methodologies that go beyond simple information retrieval to provide structured, interpretable, and user-adaptive knowledge representations [1]. This re-

[1] MOSAIC stands for Multi-site application of Open Science in the creAtion of healthy environments Involving local Communities, which runs from January 2024 to December 2027.

search addresses this gap by leveraging knowledge graphs and retrieval-augmented generation (RAG) [7] to enhance health-related decision support in the context of environmental change.

2 Research Objectives

This PhD research aims to develop a knowledge-based system to explain how severe climate events and environmental changes affect the health of Maasai communities in East Africa. It is conducted within the MOSAIC initiative, which promotes Open Science, transdisciplinary research, and actionable knowledge for vulnerable regions. The main objectives are:

— **Structuring and formalizing knowledge:** Develop a structured representation of data and domain knowledge from both scientists and local communities, leveraging ontologies and knowledge graphs, with a long-term goal of automating this process in future iterations.

— **Context-aware information retrieval and explanation:** Design and implement a RAG system that interacts with the knowledge graph to retrieve relevant information and generate explainable, user-adaptive responses for diverse stakeholders (e.g., local communities, healthcare professionals, researchers).

— **Development of an adaptive system architecture:** Build a scalable and flexible architecture that can evolve with new data inputs and user feedback.

3 Methodology and related work

The integration of ontologies and knowledge graphs has been widely explored in the environmental and health domains [2]. Similarly, recent surveys on RAG systems underscore their increasing role in enhancing knowledge accessibility and explainability [3]. However, few approaches have applied these technologies to examine the impact of drought and environmental changes on human health in East Africa, particularly within Maasai communities. Notably, there is a lack of systems that effectively combine scientific knowledge with insights from local communities while providing clear, user-adaptive explanations. This research addresses this gap by developing a structured, explainable AI system that integrates formalized environmental and health knowledge with context-aware reasoning and retrieval mechanisms.

3.1 Co-construction of knowledge graph

The knowledge graph is being developed through participatory workshops involving researchers, domain experts, and local communities. This collaborative approach ensures the incorporation of multidimensional knowledge relevant to the Maasai context, including ecological, socio-economic, and stakeholder relationships. By grounding the representation in real-world insights, the resulting knowledge graph supports various reasoning and decision-making tasks while remaining contextually relevant and interpretable. To facilitate its integration into a broader AI system, we are now focusing on the technical aspects, specifically investigating how modular RAG architectures [3] can interact with structured knowledge to enable adaptive, user-specific responses.

3.2 Benchmarking criteria for RAG systems

To ensure the effectiveness of the RAG system within this project, we identify key benchmarking criteria that align with structured knowledge representation and explainability. Specifically, our analysis focuses on RAG systems or frameworks that exhibit the following capabilities:

— Seamless integration with structured knowledge sources, such as ontologies and knowledge graphs [4].

— Transparency and interpretability of the retrieval process.

— The ability to model and reason over causal relationships.

— Adaptability of generated responses based on user profiles and contextual factors.

— Awareness of evolving knowledge, ensuring the system reflects data updates in real-time.

— A modular design that allows integration with domain-specific components.

3.3 Evaluation framework

The evaluation of the proposed system is structured into two levels: component-level and system-level assessment. We review state-of-the-art evaluation methodologies in knowledge graphs, RAG systems, and XAI [3] to inform the selection and adaptation of appropriate metrics and benchmarks. At the component level, we assess:

1. **Knowledge graph quality:** coverage, consistency, and relevance.

2. **Retriever performance:** precision, recall, and transparency.

3. **Generator effectiveness:** fluency, factual accuracy, and adaptability.

System-level evaluation examines the overall capability of the system to provide relevant, context-aware responses tailored to different user profiles. Additionally, we evaluate the system's sensitivity to knowledge updates and the clarity of generated explanations. To enhance robustness, the evaluation framework will evolve iteratively, incorporating expert validation and user feedback in later phases [9, 5].

4 Conclusion and future directions

This research develops a knowledge-based system integrating structured knowledge and adaptive generation to explain climate-related health impacts in Maasai communities. Grounded in collaborative modeling and modular design, the system aims to provide explainable and context-aware insights.

4.1 Current focus

We are testing modular RAG configurations, evaluating retrievers such as the G retriever [4] alongside LLaMA 3.2[2] and DeepSeek-R1[3]. To refine responses, we explore iterative generation and self-evaluation strategies, including BERTScore [11] and LLM-as-a-judge frameworks [12].

4.2 Investigative directions

Key questions shaping system development include:

— How to adapt responses to different user profiles?

— Can multi-LLM evaluation enhance relevance and clarity?

— How do modular configurations shape reasoning and explainability?

— How can the system extend MOSAIC project outputs through external modules?

4.3 Fieldwork and knowledge collection

An upcoming field mission to Kenya will refine the ontology and knowledge graph, validate system outputs, and identify key health indicators. This will ensure scientific rigor while aligning with local needs.

Acknowledgment

This study has been funded by the MOSAIC Horizon Europe project (grant number: 101137398).

References

[1] V. Armant, A. Mouakher, F. Vargas-Rojas, D. Symeonidou, J. Guérin, I. Mougenot, and J. C. Desconnets. Can knowledge graphs and retrieval-augmented generation be combined to explain query/answer relationships truthfully? In *Proceedings of the DAO-XAI 2024 Data meets Ontologies in Explainable AI co-located with the 27th European Conference on Artificial Intelligence (ECAI 2024)*, volume 3833, 2024.

[2] V. Armant, F. Vargas-Rojas, V. Agazzi, J. C. Desconnets, I. Mougenot, V. Beretta, S. Debard, D. Symeonidou, A. Mouakher, J. Guérin, T. Catry, and E. Roux. Leveraging Knowledge Graphs for Earth System Dataset Discovery. In *Proceedings of the International Semantic Web Conference*, pages 271 – 288, Baltimore (Maryland), United States, 2024.

[2] https://ollama.com/library/llama3.2

[3] https://ollama.com/library/deepseek-r1

[3] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, H. Wang, and H. Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2, 2023.

[4] X. He, Y. Tian, Y. Sun, N. Chawla, T. Laurent, Y. LeCun, X. Bresson, and B. Hooi. G-retriever: Retrieval-augmented generation for textual graph understanding and question answering. *Advances in Neural Information Processing Systems*, 37:132876–132907, 2024.

[5] V. K. Kommineni, B. König-Ries, and S. Samuel. From human experts to machines: An LLM supported approach to ontology and knowledge graph construction. *arXiv preprint arXiv:2403.08345*, 2024.

[6] David W Lawson, Monique Borgerhoff Mulder, Margherita E Ghiselli, Esther Ngadaya, Bernard Ngowi, Sayoki GM Mfinanga, Kari Hartwig, and Susan James. Ethnicity and child health in northern tanzania: Maasai pastoralists are disadvantaged compared to neighbouring ethnic groups. *PloS one*, 9(10):e110447, 2014.

[7] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yihand T. Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[8] V. Ongoma, H. Chen, and C. Gao. Projected changes in mean rainfall and temperature over east africa based on CMIP5 models. *International journal of Climatology*, 38(3):1375–1392, 2018.

[9] C. Shimizu and P. Hitzler. Accelerating knowledge graph and ontology engineering with large language models. *Journal of Web Semantics*, page 100862, 2025.

[10] N. Watts, M. Amann, N. Arnell, S. Ayeb-Karlsson, K. Belesova, W. Cai, et al. The 2018 report of the lancet countdown on health and climate change: shaping the health of nations for centuries to come. *The Lancet*, 392(10163):2479–2514, 2018.

[11] T. Zhang, V. Kishore, F. Wu, K. Weinberger, and Y. Artzi. BERTscore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*, 2019.

[12] L. Zheng, W. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al. Judging LLM-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

Hierarchical Classification for Automated Image Annotation of Coral Reef Benthic Structures

Célia Blondin[1], Joris Guérin[1], Kelly Iganaki[2], Guilherme Longo[2], Laure Berti-Equille[1]

[1] Institut de Recherche et Développement, IRD UMR Espace-Dev
[2] Universidade Federal do Rio Grande do Norte

celia.blondin@ird.fr

Résumé

L'annotation automatisée des images benthiques est essentielle pour surveiller et protéger efficacement les récifs coralliens face au changement climatique. Les approches actuelles d'apprentissage automatique ne parviennent pas à capturer la nature hiérarchique des organismes benthiques recouvrant le substrat récifal, c'est-à-dire les niveaux taxonomiques des coraux et leur état de santé. Pour pallier cette limitation, nous proposons d'annoter les images benthiques à l'aide d'une classification hiérarchique. Des expériences menées sur un jeu de données spécifiquement créé provenant d'un récif corallien du nord-est du Brésil montrent que notre approche surpasse les classifieurs "plats", en améliorant à la fois les mesures de F1-score et de F1-score hiérarchique d'environ 2%, quels que soient les volumes de données d'entraînement. De plus, cette méthode hiérarchique s'aligne plus étroitement sur les objectifs écologiques (calcul des taux de couverture pour les différentes espèces et groupes taxonomiques supérieurs (familles, genres...)). La version originale de l'article est disponible à l'adresse suivante: https://arxiv.org/abs/2412.08228.

Mots-clés

Classification Hiérarchique, Annotation d'image benthique

Abstract

Automated benthic image annotation is crucial to efficiently monitor and protect coral reefs against climate change. Current machine learning approaches fail to capture the hierarchical nature of benthic organisms covering reef substrata, i.e., coral taxonomic levels and health condition. To address this limitation, we propose to annotate benthic images using hierarchical classification. Experiments on a custom dataset from a Northeast Brazilian coral reef show that our approach outperforms flat classifiers, improving both F1 and hierarchical F1 scores by approximately 2% across varying amounts of training data. In addition, this hierarchical method aligns more closely with ecological objectives. The first version of the article is available at: https://arxiv.org/abs/2412.08228.

Keywords

Hierarchical Classification, Benthic Image Annotation

1 Introduction

Coral reefs support 25% of marine species [12] and provide essential ecosystem services such as coastal protection, fisheries, and tourism [17]. However, they are increasingly threatened by climate change, notably through coral bleaching events [13, 3]. To guide conservation efforts, coral reef monitoring has transitioned from in-situ measurements to photographic surveys [5], where images are manually annotated to estimate benthic cover proportions.

In response to these challenges, comprehensive coral reef surveying and monitoring have become essential for implementing appropriate and timely conservation actions [4].

Manual annotation is time-consuming and prone to inter-analyst variability [1]. Automated methods, like CoralNet [16], leverage deep learning to classify small image patches (Figure 1), reducing annotation effort. Our collaboration with marine ecologists revealed a gap between ML models and ecological needs. While models use detailed labels (e.g., Bleached Palythoa which combines coral type, genus, and health status, Figure 2), ecologists often aggregate data differently, such as grouping all corals regardless of species or health to compute cover proportions.

To address these challenges, we explore Hierarchical Classification (HC) [11] for benthic annotation. By leveraging taxonomy, HC allows partially correct predictions and improves performance for underrepresented categories. Additionally, hierarchical metrics [7] better reflect ecological practices. Our approach seeks to enhance both classification accuracy and the ecological relevance of automated coral reef monitoring.

2 Methodology

Hierarchical Classification (HC) structures classes into a tree, aligning with natural taxonomies [15, 11] (Figure 2). In HC, a global approach involves training a single model that handles the entire hierarchy simultaneously, but this strategy is beyond the scope of this paper. Instead, we adopt a top-down strategy, assigning a dedicated classifier to each node in the hierarchy, as this design aligns more naturally with the ecological structure of our problem.

Building an HC tree requires answering several design questions: Can a prediction be composed of multiple la-

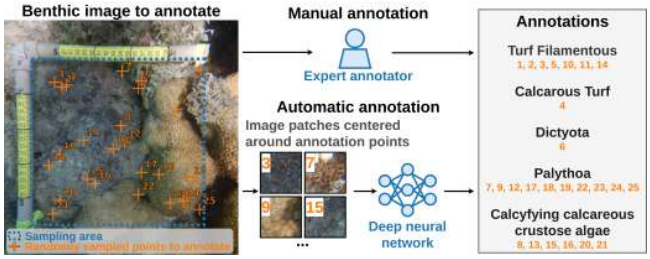


Figure 1: **Benthic image annotation process** – A predetermined number of random points are sampled within a fixed area and annotated with the corresponding organism or substrate. The relative frequencies are used as proxies for cover proportions of benthic categories within the sampled area.

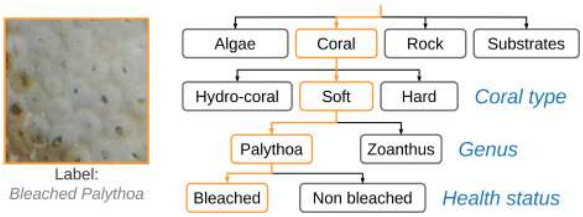


Figure 2: **Hierarchical structure of benthic labels** – This conceptual diagram illustrates the multi-level categories used for classifying benthic organisms and substrates in coral reef ecosystems.

bels? Can multiple paths lead to the same label? Can predictions stop at intermediate nodes? As the goal of this work is to validate the relevance of HC for benthic image annotation, we chose the most simple tree design possible where the final prediction is always a single leaf node, and there is only one path to each label. This is the simplest way to enable a fair comparison with flat classification, which can only predict leaf nodes and lacks the ability to select a path for each label. The design of the tree (Appendix A)is the result of our collaboration between ML experts and marine ecologists to ensure that both computational constraints and biological relevance are respected.

Although this initial step represents an additional effort, it helps aligning the annotation process with biological classification principles and facilitates subsequent ecological analyses involving grouping labels at various taxonomic levels. This step can be relatively straightforward for biological classification; indeed, the easiest tree to use is the taxonomy tree (e.g., order, phylum). However, in our case, although we rely on the taxonomy tree, we do not use all nodes for each species. We include taxonomic levels only when needed—for instance, to distinguish between several species of the same genus—and also incorporate health status, which may cause some categories to overlap. For example, the label "Bleached" could be confused with "Palythoa" if the image shows a bleached Palythoa. We adopted this approach to retain the original labels without re-annotating images, even when such label ambiguities could arise.

Hierarchical Classifier. For image annotation, 224x224 patches are extracted and passed through CoralNet's EfficientNet B0 backbone [2] as for the flat classifier. We then employ a *Local Classifier per Parent Node* (LCPN) strategy using MLPs with two hidden layers (200 and 100 neurons), implemented with the *hiclass* library [8]. LCPN is a methodology that assigns a dedicated classifier to each non-leaf node in the hierarchy. We chose this strategy based on insights from [10], where the authors compared several hierarchical approaches—including Local Classifier per Node (LCN), Local Classifier per Level (LCL), a hybrid LCN+LCPN method, and some global classifiers. While LCN and the hybrid approach yielded the best results with MLPs, LCPN was not evaluated. We opted to explore LCPN because it offers a more intuitive and modular structure, especially when dealing with complex taxonomies.

3 Experiments

We assess the potential of HC for benthic image annotation and provide a reproducible Github repository[1].

3.1 Experimental Setup

Dataset. We use a dataset from Rio do Fogo, Brazil, containing 1,549 images annotated with 38,725 points. The dataset is highly imbalanced: 11 labels cover 95% of the annotations, while 24 have fewer than 20 samples (Appendix B). A test set (10%) was randomly selected, preserving the class distribution. Test patches were chosen independently, without considering image-level coherence.

Baseline. We compare a standard flat classifier (MLP with two hidden layers: 200, 100) as in CoralNet [2] with a top-down hierarchical approach that integrates taxonomy during training.

Evaluation Metrics. Traditional classification metrics (e.g., F1-score) treat all errors equally, while hierarchical metrics [7] consider ecological relevance. Misclassifying a coral as "Algae" impacts all taxonomic levels, unlike a misclassification within the same genus. Therefore, hierarchical classification metrics account for the varying severity of classification errors based on their position in the hierarchy. We evaluate both F1-score and hierarchical F1-score to assess the benefits of HC in capturing hierarchical relationships. We therefore report both the standard F1-score and the hierarchical F1-score to evaluate the benefit of HC in capturing taxonomic relationships.

The standard F1-score is defined as follows, where TP denotes true positives, FP false positives, and FN false negatives:

$$\text{Precision} = \frac{TP}{TP+FP}, \quad \text{Recall} = \frac{TP}{TP+FN},$$

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

[1] https://github.com/celia-bl/hierarchical_classifying_corals_dataset.git

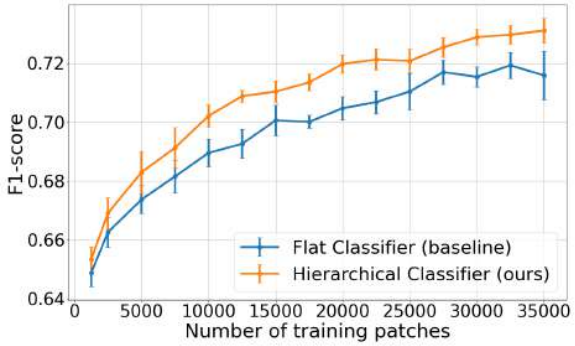


Figure 3: F1-score

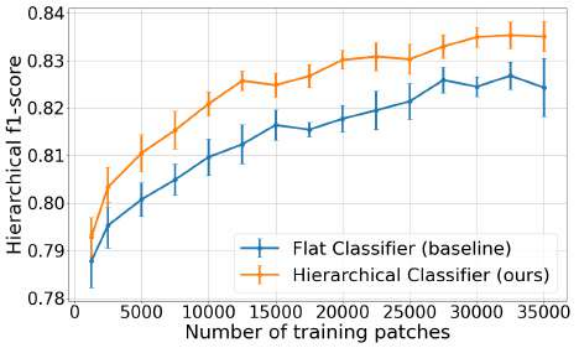


Figure 4: Hierarchical F1-score

Hierarchical extensions of these metrics, initially proposed by [6], are defined as:

$$hP = \frac{\sum_i |\alpha_i \cap \beta_i|}{\sum_i |\alpha_i|}, \; hR = \frac{\sum_i |\alpha_i \cap \beta_i|}{\sum_i |\beta_i|},$$

$$hF = 2 \cdot \frac{hP \cdot hR}{hP + hR}.$$

where:

- α_i is the set of predicted labels for example i, including the predicted leaf and all its ancestors;
- β_i is the corresponding set of true labels, including the true leaf and its ancestors;
- the sums are computed over all test examples.

3.2 Results

Our experimental results are displayed in Figure 3 and 4, where we compare the F1-score and hierarchical F1-score for the flat and hierarchical classifiers when the number of training patches increases. It shows that the hierarchical classifier always outperforms the flat classifier, with performance gains ranging from 1% (for small training sets) to 2% (for larger sets).
Interestingly, the performance gap between the two approaches remains similar for the hierarchical F1-score, contrary to our expectation that this metric would more severely penalize errors early in the tree structure for the flat classifier. This suggests that the flat classifier may implicitly capture some of the hierarchical relationships between classes. If the flat classifier was making largely implausible errors (errors with a high distance in the tree), we would expect a more substantial difference in the hierarchical F1-score.

4 Conclusion

This study demonstrates that hierarchical classification improves benthic image annotation, outperforming flat classifiers in both F1-score and hierarchical F1-score. However, the flat classifier's robustness suggests that it learns some class relationships. Further analysis of error types could refine our understanding of HC advantages.
Applying HC to datasets like Moorea Labeled Corals [9], which has a simpler hierarchy, could help contextualize these results. While tree construction requires effort, it prevents label inconsistencies and supports ecological analyses. Automating this process could enhance class discrimination [14].
Although hierarchical classification offers certain advantages, the modest accuracy improvement (2%) comes at the cost of increased computational complexity, including a 17% longer training time due to the need to train multiple classifiers. Future work will explore alternative HC methods, such as global approaches, to optimize this trade-off.

References

[1] Oscar Beijbom, Peter J Edmunds, Chris Roelfsema, Jennifer Smith, David I Kline, Benjamin P Neal, Matthew J Dunlap, Vincent Moriarty, Tung-Yung Fan, Chih-Jui Tan, et al. Towards automated annotation of benthic survey images: Variability of human experts and operational modes of automation. *PloS one*, 10(7):e0130312, 2015.

[2] Qimin Chen, Oscar Beijbom, Stephen Chan, Jessica Bouwmeester, and David Kriegman. A new deep learning engine for coralnet. In *Proc. of the IEEE/CVF international conference on computer vision*, pages 3693–3702, 2021.

[3] C. Mark Eakin, Hugh P. A. Sweatman, and Russel E. Brainard. The 2014–2017 global-scale coral bleaching event: insights and impacts. *Coral Reefs*, 38(4):539–545, August 2019.

[4] Graham J Edgar and Rick D Stuart-Smith. Systematic global assessment of reef fish communities by the reef life survey program. *Scientific Data*, 1(1):1–8, 2014.

[5] Paul L Jokiel, Kuulei S Rodgers, Eric K Brown, Jean C Kenyon, Greta Aeby, William R Smith, and Fred Farrell. Comparison of methods used to estimate coral cover in the hawaiian islands. *PeerJ*, 3:e954, 2015.

[6] Svetlana Kiritchenko, Stan Matwin, Richard Nock, and A. Fazel Famili. Learning and evaluation in the presence of class hierarchies: Application to text categorization. In Luc Lamontagne and Mario Marchand, editors, *Advances in Artificial Intelligence*, pages 395–406. Springer.

[7] Aris Kosmopoulos, Ioannis Partalas, Eric Gaussier, Georgios Paliouras, and Ion Androutsopoulos. Evaluation measures for hierarchical classification: a unified view and novel approaches. *Data Mining and Knowledge Discovery*, 29:820–865, 2015.

[8] Fábio M. Miranda, Niklas Köhnecke, and Bernhard Y. Renard. HiClass: a Python Library for Local Hierarchical Classification Compatible with Scikit-learn. *Journal of Machine Learning Research*, 24(29):1–17, 2023.

[9] Moorea Coral Reef LTER and Peter Edmunds. MCR LTER: Coral Reef: Computer Vision: Moorea Labeled Corals, 2019.

[10] Antonio R. S. Parmezan, Vinicius M. A. Souza, Arpita Seth, Indrė Žliobaitė, and Gustavo E. A. P. A. Batista. Hierarchical classification of pollinating flying insects under changing environments. 70:101751.

[11] Carlos N Silla and Alex A Freitas. A survey of hierarchical classification across different application domains. *Data mining and knowledge discovery*, 22:31–72, 2011.

[12] David Souter, Serge Planes, Jérémy Wicquart, Murray Logan, David Obura, and Francis Staub. Status of Coral Reefs of the World: 2020, 2020.

[13] Mark D Spalding and Barbara E Brown. Warm-water coral reefs and climate change. *Science*, 350(6262):769–771, 2015.

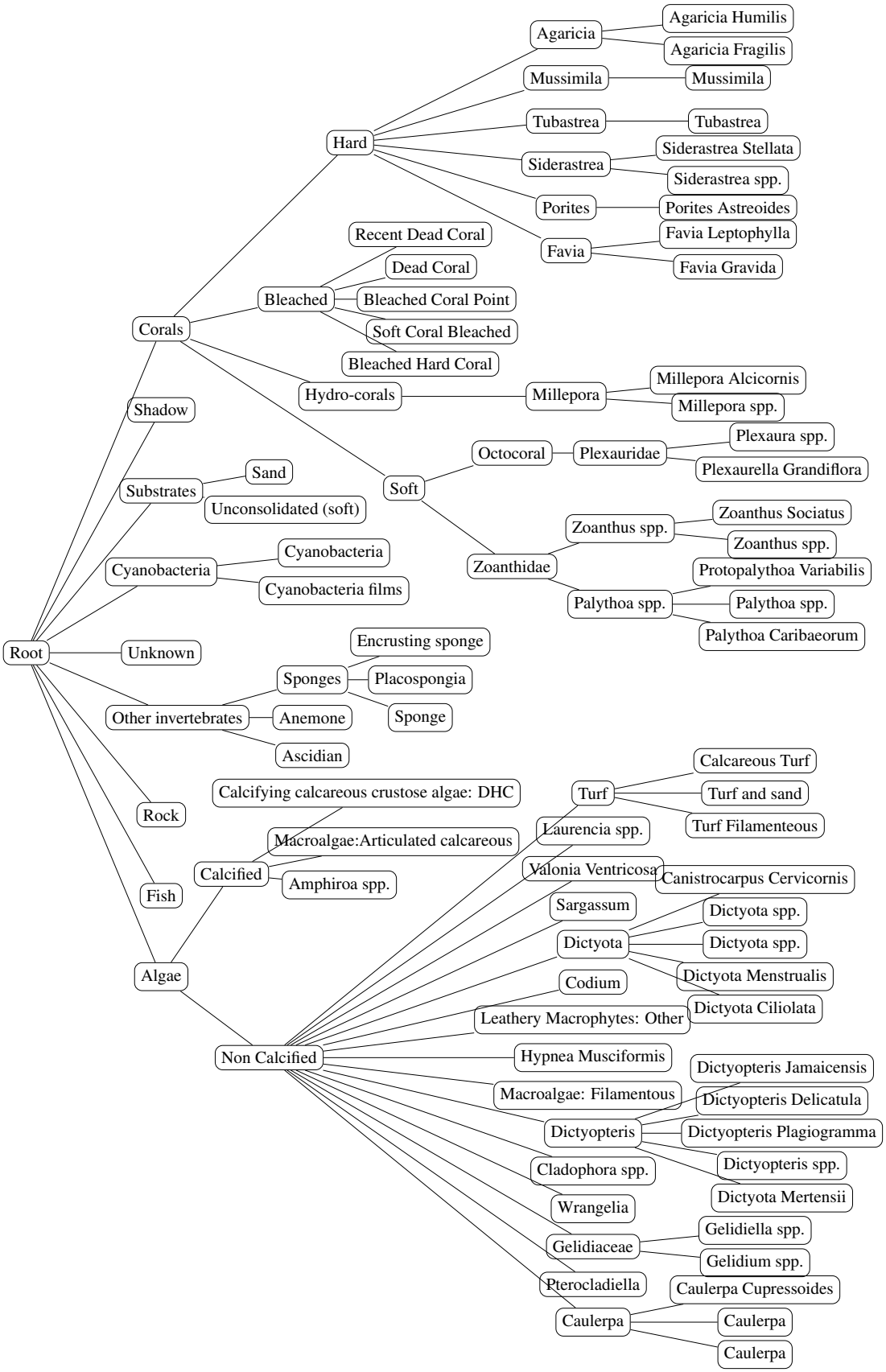
[14] Nitish Srivastava and Russ R Salakhutdinov. Discriminative Transfer Learning with Tree-based Priors. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

[15] Jonatas Wehrmann, Ricardo Cerri, and Rodrigo Barros. Hierarchical multi-label classification networks. In *International conference on machine learning*, pages 5075–5084. PMLR, 2018.

[16] Ivor D Williams, Courtney S Couch, Oscar Beijbom, Thomas A Oliver, Bernardo Vargas-Angel, Brett D Schumacher, and Russell E Brainard. Leveraging automated image analysis tools to transform our capacity to assess status and trends of coral reefs. *Frontiers in Marine Science*, 6:222, 2019.

[17] Anna J Woodhead, Christina C Hicks, Albert V Norström, Gareth J Williams, and Nicholas AJ Graham. Coral reef ecosystem services in the anthropocene. *Functional Ecology*, 33(6):1023–1034, 2019.

A Hierarchical Tree



B Class Distribution

Labels	Number of Patches	Labels	Number of Patches
Palythoa spp.	7671	Gelidiella spp.	29
Sand	6622	Millepora Alcicornis	28
Turf and sand	4981	Dictyopteris Plagiogramma	21
Turf Filamenteous	4938	Recent Dead Coral	17
Calcifying calcareous crustose algae: DHC	3629	Sargassum	12
Dictyota spp.	3443	Protopalythoa Variabilis	11
Zoanthus Sociatus	2211	Siderastrea spp.	10
Calcareous Turf	1366	Hypnea Musciformis	9
Siderastrea Stellata	774	Leathery Macrophytes: Other	7
Dictyota Mertensii	694	Amphiroa spp.	6
Shadow	681	Plexaurella Grandiflora	5
Dictyota Ciliolata	187	Palythoa Caribaeorum	4
Cyanobacteria films	172	Sponge	4
Unknown	158	Valonia Ventricosa	4
Soft Coral Bleached	158	Anemone	3
Porites Astreoides	132	Dictyota Menstrualis	3
Bleached Coral Point	107	Tubastrea	2
Rock	101	Coral	2
Dictyopteris Delicatula	95	Gelidium spp.	2
Dictyopteris spp.	81	Fish	2
Laurencia spp.	76	Dead Coral	2
Wrangelia	56	Encrusting sponge	1
Substrate: Unconsolidated (soft)	47	Dictyopteris Jamaicensis	1
Millepora spp.	45	Caulerpa	1
Favia Gravida	42	Agaricia Humilis	1
Cyanobacteria	37	Macroalgae: Articulated calcareous	1
Zoanthus spp.	32	Placospongia	1

Figure 5: Number of patches corresponding to each leaf node present in our custom benthic image annotation dataset.

Session 2

Prédiction des séquences culturales par apprentissage automatique : une comparaison entre grandes cultures et maraîchage diversifié

Baptiste Darnala[1], Clément Jonquet[2], Florence Amardeilh[1] et Konstantin Todorov[3]

[1] Elzeard, Villenave d'Ornon, France
[2] MISTEA, Université de Montpellier, INRAE, Institut Agro, Montpellier, France
[3] LIRMM, Université de Montpellier, CNRS, Montpellier, France

mél

Résumé

La rotation des cultures est une pratique consistant à alterner différentes cultures sur une même parcelle pour des bénéfices agronomiques. Cependant, cette pratique est difficile car elle est dépendante de règles agronomiques mais aussi d'une expérience du terrain permettant de connaître les rotations bénéfiques, neutres ou à éviter. Nous proposons une méthodologie d'implémentation de modèle d'apprentissage machine pour la prédiction de séquence de culture basé sur l'historique mais aussi sur le contexte de culture de l'agriculteur.

Mots-clés

Rotation des cultures, agroécologie, apprentissage machine, apprentissage profond, prediction de séquence

Abstract

Crop rotation is the practice of alternating different crops on the same plot for agronomic benefit. However, this practice is difficult because it depends on agronomic rules but also on experience in the field to know which rotations are beneficial, neutral or to be avoided. We propose a methodology for implementing a machine learning model for predicting crop sequences based on the farmer's cropping history and context.

Keywords

Crop rotation, agroecology, machine learning, deep learning, sequence prediction

1 Introduction

La rotation des cultures est une pratique agroécologique qui "consiste à organiser la successions culturale des espèces sur une parcelle". [1] Les objectifs de cette pratique sont en priorité d'améliorer la qualité des sols, les rendements et de prévenir certains aléas, comme par exemple, en coupant les cycles de développement de certaines maladies ou bioagresseurs. Pour construire leurs rotations, les agriculteurs utilisent des règles agronomiques ou métiers [14] comme par exemple alterner des cultures ou familles de cultures ou éviter d'enchaîner des cultures ayant les mêmes maladies ou bioagresseurs. Dans notre cas, nous nous intéressons au maraîchage qui est un domaine agricole tourné autour de la production de légumes et avec une forte diversité de culture. Si les agriculteurs se reposent sur leur connaissance pour formaliser leurs rotations, ils utilisent aussi leurs expériences du terrain pour construire leurs rotations culturales. Les agriculteurs analysent les historiques de cultures des années précédentes pour prendre des décisions sur la planification des cultures de l'année suivante. Dans ce processus décisionnel, les modèles d'apprentissage ont un avantage pour apprendre les pratiques culturales d'un ensemble de maraîchers. Pour la rotation de culture, ils peuvent analyser un ensemble de séquences de cultures, découvrir des relations latentes entre les cultures et proposer un ensemble de cultures ordonnées par préférence.

Nous présentons notre méthodologie de construction de modèles d'apprentissage pour répondre à la problématique de prédiction de séquences de cultures. Cette problématique consiste à proposer une culture aux maraîchers en fonction d'une séquence de taille n donnée en paramètre au modèle d'apprentissage.

1.1 Travaux antérieurs

Des méthodes automatiques pour aider à la planification des cultures existent. Ces approches peuvent être à bases de règles [9], soit en utilisant de la programmation linéaire [6][18][16], ou de la programmation par contraintes [1][3]. D'autres approches sont basés sur l'utilisation de modèles d'apprentissage et basés sur l'expérience acquises des agriculteurs.

[2] décrit une approche utilisant des chaînes de Markov pour modéliser les rotations culturales, avec des matrices de transition couplées à des règles expertes pour les grandes cultures. [15] utilise des "markov logic networks" pour intégrer des connaissances expertes et des données utilisateurs à l'échelle de la parcelle. [21] propose une méthodologie basée sur les réseaux de neurones profonds pour prédire les cultures futures, en analysant des cartes de cultures sur plusieurs années. [20] utilise des réseaux de neurones convolutionnels pour une approche similaire, adaptée aux grandes cultures. [7] présente une méthodologie basée sur

1. https://dicoagroecologie.fr/dictionnaire/rotation-des-cultures/

les chaînes de Markov, entraînée sur des données de fermes, permettant de prédire les cultures futures avec une bonne explicabilité. [8] utilise des réseaux de neurones profonds avec des neurones Long Short-Term Memory et l'algorithme BeamSearch pour améliorer la prédiction des séquences de cultures.

Ces méthodologies sont principalement appliquées aux grandes cultures comme les céréales et les protéagineux, mais leur application au maraîchage reste difficile en raison de la taille des parcelles et de la diversité des cultures. De plus l'accessibilité reste difficile pour la mise en place de méthode d'apprentissage pour la maraîchage.

2 Construction des jeux de données

Pour réaliser notre tâche de prédiction des cultures, nous avons cherché à comparer les résultats de plusieurs modèles d'apprentissage sur deux jeux de données.

- Le jeu de données pour représenter les cultures obtenu grâce à un recensement agricole est le Registre Parcellaire Graphique (RPG) [12]. Le RPG est "une base de données géographiques servant de référence à l'instruction des aides de la politique agricole commune (PAC)". Pour chaque zone de culture, la culture principale annuelle est enregistrée. Un jeu de données, que nous utilisons dans nos modèles, contenant les séquences de cultures a été développé et publié en open data [10]. Le jeu de données est présenté sous la forme d'un ensemble de fichiers organisés par département et nous avons décidé d'utiliser celui de l'Hérault. Ce choix arbitraire est réalisé pour réduire la quantité de calculs nécessaire aux développements des modèles.
- Pour les données du maraîchage, il n'existe pas à notre connaissance de base de données publiques présentant des séquences de cultures orientée pour le maraîchage permettant de construire des modèles imitant le pouvoir décisionnel des maraîchers pour les aider dans leur prise de décision. Pour cela, nous nous sommes basés sur les données utilisateurs issues de l'application Elzeard, une application d'aide à la planification et la gestion de fermes maraîchères, avec des données allant de 2021 à 2025. Ces données contiennent les planifications des cultures décrites par des maraîchers avec des périodes de culture, des éléments de contexte et des tâches planifiées pour les cultures.

Contrairement aux données du RPG, les cultures ne sont pas enregistrées de manière annuelle, mais ont des périodes sélectionnées par les maraîchers et représentant la durée planifiée des cultures sur un emplacement. La Table 1 compare les informations entre le jeu de données du RPG et celui d'Elzeard. Nous observons une différence majeure en termes de données avec environ 185 100 zones de culture enregistrées dans les données du RPG contre 759 pour les données d'Elzeard, les zones de cultures qui représentent chacune une séquence différente. Le nombre de cultures est assez similaire, mais c'est la taille maximale des séquences qui diffère avec une taille maximale de six pour le RPG si nous ne prenons que la base de données allant de 2015 à 2021 alors que la plus longue séquence pour les données d'Elzeard est de 16. Suite aux pré-traitements, nous analyserons la distribution des séquences selon leur taille.

TABLE 1 – Tableau comparatif des jeux de données du RPG et d'Elzeard.

	RPG	Elzeard
Date de création	2022-08-03	2024-05-29
Nombre de zones de culture	185 100	759
Nombre de cultures différentes	218	222
Taille de séquence maximale	6	16
Origine	Publique	Privée
Cible	Toutes cultures	Horticulture, Maraîchage
Méthodologie de construction	Recensement national	Application métier
Période	2015-2020	2021-2025

3 Traitements des données

Les deux bases de données, décrites dans la Table 1, que nous analysons et que nous traitons pour cette tâche d'apprentissage contiennent des informations sur des successions de cultures. Chaque jeu de données a ses spécificités et comme l'objectif est de construire des modèles d'apprentissage et de comparer les résultats obtenus sur les deux jeux de données, nous devons construire une structure de données similaires pouvant être introduite dans les modèles d'apprentissages. L'objectif de cette section est de présenter les traitements appliqués à chaque jeu de données pour obtenir une structure similaire.

Les données issues du RPG sont constituées de séquences décrivant une parcelle où chaque culture de la séquence représente une année précise. Par exemple, comme nous nous intéressons à la base de données allant de 2015 à 2021, une séquence comme "VRC-VRC-VRC-VRC-VRC-/" représente cinq années successives de vigne (VRC=vigne_cuve) et une année vide représentée par le "/". Dans les données d'Elzeard, les cultures ont des périodes uniques avec par exemple une culture de tomate allant de mai à juillet suivie d'une culture d'ail allant de juillet à octobre. Divers prétraitements ont été appliqués sur les deux jeux de données pour arriver à une structure commune (suppression des valeurs manquantes pour les données du RPG, suppression des chevauchement de cultures sur une même parcelle au même moment et unification de culture similaire et proche temporellement pour les données d'Elzeard, création de séquence de taille fixe).

Ces prétraitements permettent d'obtenir une structure commune permettant l'introduction des données dans les modèles et la comparaison des résultats entre les deux jeux de données.

4 Analyse des jeux de données

Suite aux prétraitements, nous pouvons analysé les deux jeux de données. Premièrement, nous analysons le nombre de séquences selon la taille des séquences dans les jeux de données à l'aide de la Figure 1. Nous observons dans un premier temps que le nombre de séquences du RPG est beaucoup plus grand que celui d'Elzeard et cela pour chaque taille de séquences. La quantité de données peut avoir un impact sur la capacité des modèles à prédire correctement, surtout si une distribution inégale des labels existe. La Figure 1 montre aussi l'écart proportionnel entre la quantité de séquences pour chaque taille de séquence et pour chaque jeu de données. Cela montre que le jeu de données d'Elzeard est majoritairement composé de petites séquences.

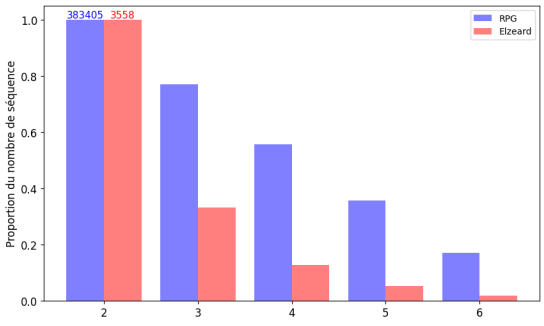


FIGURE 1 – Proportion du nombre de séquence par taille et par jeu de données par rapport au jeu le plus grand.

Ensuite, nous pouvons comparer grâce à la Figure 2 la distribution des cultures dans les séquences du RPG et d'Elzeard. Cette Figure montre le nombre d'instances prédictibles pour les 30 cultures les plus présentes dans les jeux de données. La distribution des jeux montre un problème d'équilibre dans les classes, explicable par le fait que les agriculteurs ont tendance à privilégier certaines cultures comme la tomate ou la laitue généralement pour des questions de rentabilité. Ce problème de distribution a des impacts sur les modèles qui vont favoriser les cultures sur lesquelles ils ont le plus été entraînés. Notre problématique ici est de construire des modèles favorisant la diversification des cultures. Avoir un jeu de données non équitable peut mener à construire un modèle avec de bons résultats, mais ne renvoyant qu'une sous-partie des sorties possibles. Pour contrer cela, nous analysons les résultats en tenant compte de ce problème en prenant en compte les k meilleures sorties.
Finalement pour comparer deux modes de production différents, nous avons analysé la diversité des séquences en elle-même. Pour cela, nous avons appliqué l'index de diversité de Shannon 1 [19] qui est une formule utilisée en écologie pour estimer la diversité des espèces dans une population donnée où H' désigne le score de diversité, R le nombre d'espèces et p_i la proportion d'individus dans la population. Plus le score est élevé, plus la population présente une diversité des espèces.

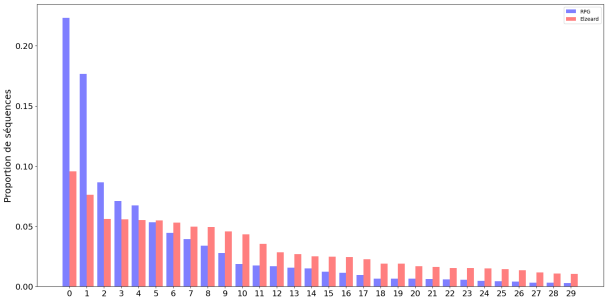


FIGURE 2 – Distribution ordonnée des 30 cultures les plus présentes dans chaque jeu de données.

$$H' = -\sum_{i=1}^{R} p_i ln p_i \quad (1)$$

La Figure 3 montre la proportion des séquences selon leurs scores de diversité par jeu de données et pour chaque taille de séquence. Ces graphiques montrent que les jeux de données ont des structures assez différentes avec une proportion supérieure de séquences avec une diversité assez faible pour les séquences issues du RPG comparé aux séquences issues d'Elzeard.
La diversité et le nombre de séquences peuvent rendre les apprentissages des modèles beaucoup plus complexes sur les données d'Elzeard. Des modèles ayant une aptitude supérieure pour gérer la diversité peuvent proposer des résultats supérieurs comme les modèles basés sur les réseaux de neurones. Cependant, ces algorithmes demandent généralement une quantité de données supérieure pour être efficaces tout en ne sur-apprenant pas.
Cette analyse nous permet de comprendre que les deux jeux de données ont des différences sur la structure des séquences et que ces différences peuvent mener à des différences significatives de résultats. La section suivante décrit la méthodologie mise en place pour entraîner les modèles pour pouvoir ensuite comparer les résultats.

5 Processus d'apprentissage

Pour pouvoir comparer l'efficacité de la prédiction de séquences de cultures sur les deux jeux de données, nous avons mis en place plusieurs architectures d'apprentissage. La première étape consiste à définir comment le modèle va lire la séquence pour prédire la culture suivante. Certains modèles, comme les modèles de réseaux de neurones implémentés et présentés ci-dessous, peuvent lire la séquence en analysant chaque culture indépendamment tout en gardant en mémoire les observations passées. D'autres modèles ne sont pas en capacité de lire les séquences, car chaque introduction de données est indépendante et le modèle ne peut pas analyser les liens entre les cultures qui pourrait être des choix pour les agriculteurs. Pour répondre à cette problématique, nous avons mis en place des fenêtres temporelles en concaténant les cultures des séquences pour

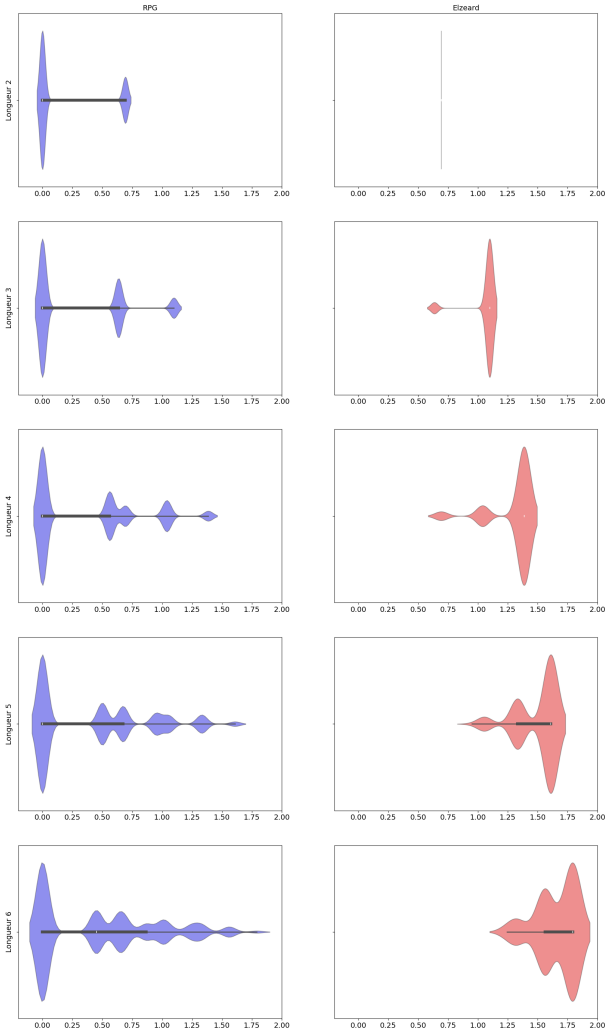


FIGURE 3 – Proportion des séquences par score de diversité pour chaque jeu de données et pour chaque longueur de séquence.

avoir une représentation unique de la séquence introductible dans les modèles.

Nous avons testé deux méthodes pour la représentation des données dans les séquences. Pour introduire des données dans un modèle d'apprentissage, nous devons les transformer numériquement. Pour cela, nous avons implémenté un encodage One-Hot et utilisé des Embeddings pour la représentation des plantes. L'encodage One-Hot consiste à construire pour chaque culture un vecteur de n bits égaux au nombre de cultures et de mettre le bit indiquant la plante à 1. L'utilisation d'Embeddings consiste à construire une représentation de plantes dans un espace latent et de faire évoluer leur représentation à travers les apprentissages pour que les plantes proches le soient aussi dans cet espace.

Pour la sélection des modèles, nous avons décidé d'implémenter des modèles d'apprentissage machine "classiques" et des modèles de réseaux de neurones. Par "classique", nous entendons les algorithmes d'apprentissage machine non-neuronaux. Nous faisons cette distinction, car ces modèles non-neuronaux ne sont pas normalement adaptés à la gestion de données séquentielles ; cependant, avec un aplanissement des données en un vecteur unique, nous pouvons procéder à leur introduction. L'intérêt des modèles non-neuronaux est d'être en partie adaptés à des jeux de données moins conséquents ce qui nous intéresse avec la faible volumétrie de données issues d'Elzeard. En outre certains des modèles ont un bon niveau d'explicabilité vis-à-vis des sorties proposées par les algorithmes. Concernant les modèles d'apprentissage machine "classiques", nous avons réalisé une sélection des modèles à expérimenter de manière empirique, car cette classe de modèle n'a pas été utilisée à notre connaissance pour la construction de modèles prédictifs pour les séquences de cultures. Nous avons testé un ensemble de modèles et sélectionné deux modèles selon différents critères que sont leur taux de précision et leur explicabilité : les arbres de décision et les forêts aléatoires. Concernant les modèles de réseaux de neurones, nous avons sélectionné deux architectures. Parmi les architectures de réseaux de neurones, certaines sont dédiées à l'encodage de la mémoire pour répondre aux problématiques de prédiction de séquences. Les choix méthodologiques concernant les réseaux de neurones ont été réalisés en suivant la méthodologie proposée par [8] pour avoir une référence pour notre expérience et les choix d'architectures. Nous avons décidé d'implémenter les neurones LSTM [11] qui permettent donc de représenter les mécanismes d'ordonnancement à l'intérieur des séquences. La problématique de ces modèles est qu'ils sont sensibles à la perte de mémoire sur de trop longues séquences. Une classe d'architecture de réseaux de neurones appelée Encoder-Decoder [4] permet de répondre en partie à cette problématique. Cette architecture fonctionne avec deux réseaux de neurones, un Encoder dans lequel une séquence d'entrée va être lue pour construire une représentation de la séquence d'entrée. Cette représentation est introduite dans un autre réseau de neurones, le Decoder, qui va construire la séquence de sortie. Ces deux couches sont construites à partir d'une ou de plusieurs couches de neurones LSTM ou équivalent. [8] ont utilisé cette architecture pour concevoir un système de prédiction de séquences de cultures.

Pour entraîner le modèle, nous avons séparé chaque jeu de données en un jeu d'entraînement, un jeu de validation et un jeu de test avec respectivement, 70%, 15% et 15% du jeu de données initial. Pour palier le problème d'équilibre entre les classes, nous avons introduit la même proportion de chaque classe dans chaque jeu de données. Nous avons implémenté et lancé les apprentissages des arbres de décisions et des forêts aléatoires avec la librairie Scikit Learn et les réseaux de neurones avec la librairie Pytorch. [2] [3] Pour chaque séquence du jeu de données, la dernière valeur de la séquence est retirée pour ensuite être comparée à la dernière sortie de la prédiction. Pour évaluer les prédictions

2. https://scikit-learn.org/stable/
3. https://pytorch.org/

du modèle, nous avons utilisé la fonction de perte *Cross Entropy* de Pytorch qui calcule la probabilité pour chaque culture d'être prédite à l'aide de la fonction Softmax 2 et qui calcule l'erreur entre la sortie prédite et la sortie attendue à l'aide de la fonction 3. Nous avons sélectionné cette fonction de perte, car elle permet d'orienter le modèle avec exactitude vers les sorties attendues et que cette fonction a été utilisée par les auteurs de [8] pour leur implémentation de l'architecture Encoder-Decoder pour la prédiction de séquences de cultures.

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{k} e^{z_k}} \text{ pour tout } j \in \{1, ..., K\} \quad (2)$$

$$H(p, q) = -\sum_i p_i log(q_i) \quad (3)$$

Les modèles ont été entraînés sur 100 époques, un nombre choisi pour permettre une analyse approfondie des données tout en évitant le sur-apprentissage grâce à la stratégie d'*Early Stopping*, qui arrête l'apprentissage dès que l'exactitude sur les données de validation n'évolue plus. L'algorithme Adam a été utilisé pour optimiser les poids du modèle en raison de sa capacité à réduire le nombre d'itérations nécessaires, bien qu'il puisse parfois rencontrer des difficultés de convergence. Un taux d'apprentissage de 0.0001 a été sélectionné pour éviter des modifications trop brusques des poids du modèle pendant l'apprentissage. Les modèles ont été entraînés avec des batchs de 32 séquences pour les données du RPG et de 8 séquences pour les données d'Elzeard, cette différence étant due à la quantité limitée de données d'Elzeard, qui nécessite une taille de batch plus petite pour permettre une analyse significative de l'évolution de l'apprentissage. Ces choix visent à optimiser l'apprentissage des modèles tout en tenant compte des contraintes liées à la quantité de données disponibles.

6 Résultats

Dans cette étude, l'objectif est de recommander plusieurs classes de cultures aux agriculteurs en utilisant la métrique Top-K pour évaluer les modèles. La métrique Top-K mesure l'exactitude des k premières recommandations du modèle. Les tests ont été effectués avec différentes valeurs de k, mais les résultats présentées dans cette section montre les performances des modèles pour un k=5, car cela permet de vérifier la présence de la valeur attendue malgré les problèmes de distribution.

Les résultats disponibles dans la table 2 pour les résultats issus des données du RPG montrent que les modèles peuvent prédire la culture suivante avec une précision de plus de 90%, indépendamment du modèle utilisé. Cependant, les prédictions pour des séquences de cultures diversifiées issues d'Elzeard disponibles dans la table 3, comme celles issues du maraîchage, sont moins efficaces. Les meilleurs scores pour des séquences d'entrée de taille 3 et 4 sont influencés par des jeux de données limités.

Les résultats obtenus avec les données du RPG sont meilleurs que ceux obtenus avec les données d'Elzeard, probablement en raison de la taille et de la diversité des jeux de données. Les modèles ont des difficultés à gérer la diversité, tendant à prédire les cultures les plus fréquentes dans le jeu de données. Cette tendance est observée dans les résultats des deux jeux de données, avec des difficultés à prédire les cultures moins représentées.

En conclusion, bien que les modèles montrent une bonne précision pour des prédictions simples, ils rencontrent des défis avec des séquences de cultures diversifiées, soulignant l'importance de la taille et de la diversité des jeux de données dans les prédictions agricoles.

TABLE 2 – Top-5 pour chaque modèle en fonction des tailles de séquence en entrée issue du RPG.

	Séquence longueur 1	Séquence longueur 2	Séquence longueur 3	Séquence longueur 4
Arbre de décision	**0.4507**	0.5273	0.5875	0.6296
Forêt aléatoire	0.4493	0.6227	**0.7125**	0.5926
LSTM One-Hot	0.4196	0.4738	0.5313	0.625
LSTM Embedding	0.415	**0.6262**	0.6094	**0.875**
Encoder-Decoder One-Hot	0.4177	0.5155	0.5156	0.625
Encoder-Decoder Embedding	0.4188	0.6083	0.5781	0.7083

7 Ajout des paramètres sur les données d'Elzeard

Pour cette tâche qui consiste à construire des modèles de prédiction de séquence sur des données du maraîchage, un des enjeux est d'apporter du contexte aux algorithmes pour améliorer les résultats grâce à une compréhension plus fine du contexte de culture. Pour cela, nous avons cherché à exploiter les données de contexte enregistrées par les maraîchers pour représenter leurs fermes et leurs cultures. Ces données sont enregistrées manuellement par l'utilisateur pendant le paramétrage de la ferme au sein de l'application Elzeard. De plus, nous avons aussi exploité la hiérarchie des plantes disponible grâce à nos alignements à TAXREF-LD [13], le référentiel taxonomique national français des organismes marins, terrestres et aquatiques présents en France métropolitaine et outre-mer, et FCU [17], un thésaurus des plantes cultivées organisé par usage en France, dans le graphe de connaissances d'Elzeard [5], pour tester les

TABLE 3 – Top-5 pour chaque modèle en fonction des tailles de séquence en entrée issue d'Elzeard.

	Séquence longueur 1	Séquence longueur 2	Séquence longueur 3	Séquence longueur 4
Arbre de décision	**0.9092**	0.9216	0.9164	0.9102
Forêt aléatoire	**0.9092**	**0.9224**	**0.9389**	**0.94**
LSTM One-Hot	0.9064	0.9189	0.9219	0.9174
LSTM Embedding	0.9062	0.9062	0.9254	0.9254
Encoder-Decoder One-Hot	0.905	0.92	0.9227	0.9235

prédictions des séquences de cultures sur les familles de cultures avec les familles d'usages et les familles botaniques. La Figure 4 présente les différents paramètres ajoutés aux modèles pour apporter du contexte à la prédiction et observer son impact sur les résultats obtenus précédemment. Les paramètres présentés, hors le mode de culture et les familles des plantes, ne sont pas des paramètres obligatoires et sont enregistrés par les maraîchers soit à l'échelle de la ferme, soit à l'échelle de la culture. Après analyse de la quantité de paramètres manquants, nous avons décidé de ne prendre que les paramètres indiqués seulement à l'échelle de la ferme, car l'enregistrement des paramètres à l'échelle de la culture est trop faible pour être pris en compte. La problématique est que le niveau de granularité est peut-être trop haut et peut se confondre entre les parcelles, diminuant la capacité des modèles d'apprentissage à prédire correctement. Les informations sur la hiérarchie des cultures et sur les paramètres externes ont été introduites indépendamment.

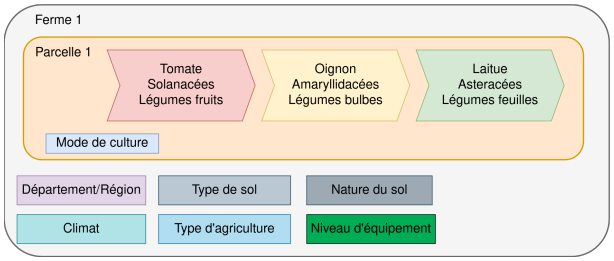


FIGURE 4 – Schéma d'une séquence de cultures et des paramètres de contexte.

Les résultats de l'expérience sur l'utilisation des paramètres

TABLE 4 – Top-5 pour les modèles entraînés sur les données d'Elzeard et enrichis avec la hiérarchie des cultures en fonction des tailles de séquence en entrée.

	Séquence longueur 1	Séquence longueur 2	Séquence longueur 3	Séquence longueur 4
Arbre de décision	0.4468 (-0.4%)	0.5242 (-0.1%)	0.575 (-1.6%)	0.625 (-0.1%)
Forêt aléatoire	**0.4536 (+1.4%)**	**0.6394 (+3.3%)**	**0.725 (+2.4%)**	**0.75 (+21.7%)**
LSTM One-Hot	0.4065 (-2.8%)	0.4310 (-9.5%)	0.4219 (-25.5%)	0.4167 (-49.6%)
LSTM Embedding	0.4265 (+3.1%)	0.481 (-29.7%)	0.4844 (-25.3%)	0.5 (-74.5%)
Encoder-Decoder One-Hot	0.2912 (-43.1%)	0.3262 (-57.7%)	0.4219 (-21.8%)	0.4583 (-35.9%)
Encoder-Decoder Embedding	0.3085 (-35.4%)	0.3286 (-84.8%)	0.4219 (-36.6%)	0.4167 (-69.6%)

de contexte, présentés dans la Table 5, montrent une amélioration générale des performances de prédictions sur la métrique évaulée par rapport aux résultats sans paramètres (Table 3). En moyenne, l'amélioration est de 16,7%, particulièrement notable pour les séquences d'entrée de taille un, deux et trois. Le modèle de forêts aléatoires se distingue avec de meilleurs résultats dans ces configurations. Cependant, les résultats sur l'introduction de la hiérarchie des plantes montrent une réduction des résultats visibles dans la table 4 avec une baisse en moyenne de 23.8% des performances.

8 Discussions

Durant le développement des modèles d'apprentissage, plusieurs défis et questionnements ont émergé. Les jeux de données du RPG et d'Elzeard présentent des différences structurelles significatives, remettant en question la pertinence de leur comparaison. Le RPG, basé sur un recensement annuel, diffère d'Elzeard, où les agriculteurs planifient leurs cultures via une application. Cette divergence soulève des problématiques :

— Les marqueurs temporels diffèrent, avec un recensement annuel pour le RPG, tandis qu'Elzeard peut inclure plusieurs cultures par an.
— Le RPG couvre divers modes d'agriculture, nécessitant un pré-traitement pour retirer les cultures pérennes comme la vigne.

TABLE 5 – Top-5 pour les modèles entraînés sur les données d'Elzeard et enrichis avec les paramètres de contexte en fonction des tailles de séquence en entrée.

	Séquence longueur 1	Séquence longueur 2	Séquence longueur 3	Séquence longueur 4
Arbre de décision	0.6464 (+30.9%)	0.5954 (+12%)	0.6 (+2.7%)	0.6667 (+6.2%)
Forêt aléatoire	**0.7072 (+37.2%)**	**0.7727 (+20.2%)**	**0.825 (+14.5%)**	0.7037 (+16.5%)
LSTM One-Hot	0.59 (+29.5%)	0.5607 (+16.1%)	0.5936 (+11.1%)	**0.8333 (+25.8%)**
LSTM Embedding	0.5765 (+28.6%)	0.656 (+5.2%)	0.6563 (+7.8%)	**0.8333 (-4.2%)**
Encoder-Decoder One-Hot	0.5704 (+27.3%)	0.5929 (+13.6%)	0.6719 (+23.9%)	0.7083 (+12.5%)
Encoder-Decoder Embedding	0.5819 (+28.6%)	0.6417 (+5.8%)	0.6563 (+12.6%)	**0.8333 (+15.8%)**

— La taille des jeux de données varie considérablement, avec Elzeard couvrant 80 fermes sur 2 ans, contre 200 000 parcelles pour le RPG sur 5 ans dans l'Hérault.

Ces disparités affectent la fiabilité des modèles d'Elzeard et la comparabilité des résultats. Pour améliorer la fiabilité, un recensement plus large des fermes maraîchères est nécessaire.

Le choix de séquences de taille fixe pour les modèles prédictifs, bien que cohérent avec une méthodologie établie, pose des questions quant à sa pertinence pour la planification des rotations culturales. La dimension temporelle, cruciale dans un contexte de changement climatique, est absente, rendant les séquences indifférenciées sur différentes périodes. Intégrer la temporalité et la saisonnalité permettrait des recommandations plus adaptées aux préférences des maraîchers.

L'intégration de la contextualisation, via la hiérarchie des cultures et des paramètres additionnels, montre des améliorations prometteuses. Cependant, des informations manquent pour les deux jeux de données. L'analyse de l'influence de chaque paramètre pourrait révéler ceux ayant le plus d'impact sur les décisions agricoles.

Enfin, les modèles actuels peinent à recommander des cultures peu représentées dans les données. Une approche alternative serait de prédire la viabilité d'une séquence de cultures plutôt que la culture suivante, facilitant ainsi l'apprentissage. Cependant, définir la viabilité d'une séquence reste un défi. L'entraînement des modèles pourrait être orienté vers des règles agroécologiques pour promouvoir des rotations plus durables ou alors l'analyse de facteurs de performance comme les rendements pour aider à dire si une séquence est viable pour des agriculteurs et aider à la prise de décision.

9 Conclusion

Une méthodologie a été développée pour comparer des modèles prédictifs sur les jeux de données du RPG, spécialisé dans les grandes cultures, et d'Elzeard, axé sur le maraîchage. Ces jeux de données diffèrent par la nature des cultures et leur temporalité, avec le RPG offrant une plus grande densité de données.

Les résultats montrent que les modèles sont plus performants sur le jeu de données du RPG, permettant de prédire efficacement les séquences de cultures pour les cultures les plus présentes. Cependant, les modèles peinent à prédire les cultures minoritaires. Sur le jeu de données d'Elzeard, les résultats sont moins favorables, en raison d'une plus faible quantité de données et d'une diversité accrue, rendant difficile la prédiction des séquences de cultures en maraîchage diversifié.

Les différences de performance peuvent s'expliquer par la taille des jeux de données et la diversité des séquences. Les modèles les plus performants diffèrent également entre les deux jeux de données : le modèle de forêt aléatoire pour le RPG et le modèle LSTM-Embedding pour Elzeard.

Pour améliorer les résultats sur Elzeard, l'introduction de la hiérarchie des cultures et des paramètres de contexte a été testée. Bien que la hiérarchie réduise les performances, les informations contextuelles améliorent les résultats.

Pour avancer, il est crucial d'augmenter la densité des données sur le maraîchage via des applications comme Elzeard ou des collectes de données expérimentales. Améliorer la description des données contextuelles pédo-climatiques et agricoles est également nécessaire. Enfin, développer des algorithmes pour des séquences de taille variable permettrait des recommandations plus flexibles pour les agriculteurs.

10 Biblio

Remerciements

Nous reconnaissons le soutien de BPI France et de l'ADEME dans le cadre du programme I-NOV Vague 8. Ce travail a également été partiellement réalisé avec le soutien du projet *Data to Knowledge in Agronomy and Biodiversity* (D2KAB - www.d2kab.org) qui a reçu un financement de l'Agence Nationale de la Recherche française (ANR-18-CE23-0017).

Références

[1] Mahuna Akplogan, Simon De Givry, Jean-Philippe Métivier, Gauthier Quesnel, Alexandre Joannon, and

Frédérick Garcia. Solving the Crop Allocation Problem using Hard and Soft Constraints. *RAIRO - Operations Research*, 47(2) :151–172, April 2013.

[2] M.S. Castellazzi, G.A. Wood, P.J. Burgess, J. Morris, K.F. Conrad, and J.N. Perry. A systematic representation of crop rotations. *Agricultural Systems*, 97(1-2) :26–33, April 2008.

[3] Margot Challand, Philippe Vismara, Dimitri Justeau-Allaire, and Stéphane De Tourdonnet. Supporting Sustainable Agroecological Initiatives for Small Farmers through Constraint Programming. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 5924–5931. International Joint Conferences on Artificial Intelligence Organization, August 2023.

[4] Kyunghyun Cho, Bo Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. *arXiv preprint arXiv :1406.1078*, 2014.

[5] Baptiste Darnala, Florence Amardeilh, Catherine Roussey, Konstantin Todorov, and Clément Jonquet. C3po : a crop planning and production process ontology and knowledge graph. *Frontiers in Artificial Intelligence*, 6 :1187090, 2023.

[6] S Dogliotti, W.A.H Rossing, and M.K Van Ittersum. rotat, a tool for systematically generating crop rotations. *European Journal of Agronomy*, 19(2) :239–250, May 2003.

[7] Ambre Dupuis, Camélia Dadouchi, and Bruno Agard. Predicting crop rotations using process mining techniques and Markov principals. *Computers and Electronics in Agriculture*, 194 :106686, March 2022.

[8] Ambre Dupuis, Camélia Dadouchi, and Bruno Agard. Methodology for multi-temporal prediction of crop rotations using recurrent neural networks. *Smart Agricultural Technology*, 4 :100152, August 2023.

[9] Jérôme Dury, Noémie Schaller, Frédérick Garcia, Arnaud Reynaud, and Jacques Eric Bergez. Models to support cropping plan and crop rotation decisions. A review. *Agronomy for Sustainable Development*, 32(2) :567–580, April 2012.

[10] Baptiste Girault, Philippe Martin, and Nicolas Piskiewicz. RPG Explorer Crop successions France 2007-2014, 2007-2020, 2015-2020, 2022.

[11] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9 :1735–1780, 1997.

[12] F. Levavasseur, P. Martin, C. Bouty, A. Barbottin, V. Bretagnolle, O. Thérond, O. Scheurer, and N. Piskiewicz. RPG Explorer : A new tool to ease the analysis of agricultural landscape dynamics with the Land Parcel Identification System. *Computers and Electronics in Agriculture*, 127 :541–552, September 2016.

[13] F. Michel, O. Gargominy, S. Tercerie, and C. Faron-Zucker. A model to represent nomenclatural and taxonomic information as linked data. application to the french taxonomic register, taxref. *Proceedings of ISWC 2017 Workshop on Semantics for Biodiversity (S4Biodiv 2017), Vienna, Austria*, pages 1–12, October 2017.

[14] Charles L. Mohler, Sue Ellen Johnson, and and Engineering Service Natural Resource, Agriculture, editors. *Crop rotation on organic farms : a planning manual*. Number 177 in NRAES. Natural Resource, Agriculture, and Engineering Service (NRAES) Cooperative Extension, 2009.

[15] Julien Osman, Jordi Inglada, and Jean-François Dejoux. Assessment of a Markov logic model of crop rotations for early crop mapping. *Computers and Electronics in Agriculture*, 113 :234–243, April 2015.

[16] C. Pahmeyer, T. Kuhn, and W. Britz. 'Fruchtfolge' : A crop rotation decision support system for optimizing cropping choices with big data and spatially explicit modeling. *Computers and Electronics in Agriculture*, 181 :105948, February 2021.

[17] Catherine Roussey. Frenchcropusage : Thésaurus sur les cultures françaises. le thésaurus décrivant les cultures françaises par leur utilisation au format skos. *FAIRsharing.org*, 2018.

[18] Martin Schönhart, Erwin Schmid, and title = CropRota – A crop rotation model to support integrated land use assessments Schneider, Uwe A. *European Journal of Agronomy*, 34(4) :263–277, May 2011.

[19] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27 :379–423, 1948.

[20] Raghu Yaramasu, Varaprasad Bandaru, and Koutilya Pnvr. Pre-season crop type mapping using deep neural networks. *Computers and Electronics in Agriculture*, 176 :105664, September 2020.

[21] Chen Zhang, Liping Di, Li Lin, and Liying Guo. Machine-learned prediction of annual crop planting in the U.S. Corn Belt based on historical crop planting maps. *Computers and Electronics in Agriculture*, 166 :104989, November 2019.

Vers une approche hybride combinant Intégration de données et IA pour la reconstitution d'un réseau d'adduction d'eau potable

Jacques Aimé RAZAFINDRAGY[1], Vincent ARMANT[2], Lalaina FANOMEZANTSOA[1], Jean-Christophe DESCONNETS[2], Eric DELAITRE[2], Aimé Richard HAJALALAINA[1], Hasina RAKOTONIRAINY[1], Thérèse LIBOUREL[2,3]

[1] Projet MADATLAS, Université de Fianarantsoa, Madagascar
[2] UMR 228, ESPACE-DEV, IRD, France
[3] Université de Montpellier, France

Résumé

Permettre le suivi de la distribution d'eau potable afin d'identifier les localisations de fuites, de branchements sauvages sur le réseau d'adduction est un enjeu majeur, spécialement dans la région Haute Matsiatra (Madagascar) où plusieurs mesures visant à la rationalisation de la distribution de l'eau sont décidées de façon saisonnière. Du fait de l'existence très ancienne et de la mise à jour continue du réseau d'adduction, la description de ce dernier est souvent incomplète et parfois erronée. Afin de permettre le suivi de la distribution d'eau (un des enjeux du projet MADATLAS co-porté par l'unité ESPACE-DEV), le premier objectif est de pouvoir reconstruire le graphe du réseau de distribution à partir des données parfois imprécises et incomplètes existantes. Dans ce contexte, nous envisageons différentes pistes qui nous permettront de vérifier la localisation du réseau réel.

Mots-clés

Gestion de la distribution d'eau potable, intégration de données hétérogènes, traitement des incertitudes.

Abstract

Monitoring the drinking water distribution network to anticipate unserved areas during periods of drought and potentially identify locations of leaks or unauthorized connections is a major challenge, especially in the Haute Matsiatra region (Madagascar), where a number of measures aimed at rationalising water distribution are decided on a seasonal basis. Because the water supply network has existed for a very long time and is constantly being updated, its description is often incomplete and sometimes erroneous. In order to monitor water distribution (one of the challenges of the MADATLAS project co-led by ESPACE-DEV), the first objective is to rebuild the graph of the distribution network from the existing erroneous and incomplete data. In this context, we are looking at various approaches that will enable us to verify the location of the actual network.

Keywords

Drinking water supply, spatial data heterogeneous, integration, treatment of uncertainties.

1 Introduction

L'eau est une ressource primordiale dans tout domaine d'activités. Assurer une gestion durable de cette ressource est l'un des objectifs du projet MADATLAS qui envisage la mise en place d'un observatoire de l'eau à Madagascar. Dans le territoire de la région de Haute Matsiatra à Madagascar, la gestion du réseau d'adduction d'eau potable est confiée à plusieurs structures. Chacune maintient son propre système d'information lui permettant de gérer la distribution de l'eau. Afin de permettre le suivi de la distribution d'eau potable à travers différentes limites administratives (quartiers, communes, régions...), il est essentiel de pouvoir intégrer et combiner les données en provenance des différents gestionnaires.

2 Contexte

Enjeu. La reconstruction spatiale du réseau d'adduction d'eau potable devrait permettre d'assurer un meilleur suivi de la distribution d'eau potable et d'alerter sur le risque de périodes de pénurie dans la région Haute Matsiatra à Madagascar. Le premier objectif répondant à ce défi est la reconstruction de la structure et de la localisation des ouvrages du réseau d'adduction d'eau potable. Répondre à ce challenge permettra de pouvoir exploiter efficacement un système d'information géographique harmonisé intégrant les données de différents acteurs.

Problématique. Le principal obstacle à la reconstruction spatiale du réseau d'adduction d'eau potable est l'hétérogénéité des données multi-sources et multi-formes. Dans la Figure 1 nous illustrons les différentes sources de données permettant de reconstruire le réseau d'adduction dans la région Haute Matsiatra. Dans la zone urbaine de Fianarantsoa, le gestionnaire du réseau d'eau est l'entreprise publique JIRAMA. La cartographie du réseau urbain présente de nombreuses imprécisions et informations manquantes, la localisation des conduits est approximative et de nombreux nœuds et canalisations ne sont pas connectés. De plus, cette carte n'est pas à jour et elle recense des canalisations anciennes ne servant plus à acheminer l'eau. Toujours dans la zone urbaine de Fianarantsoa, les éléments de mise

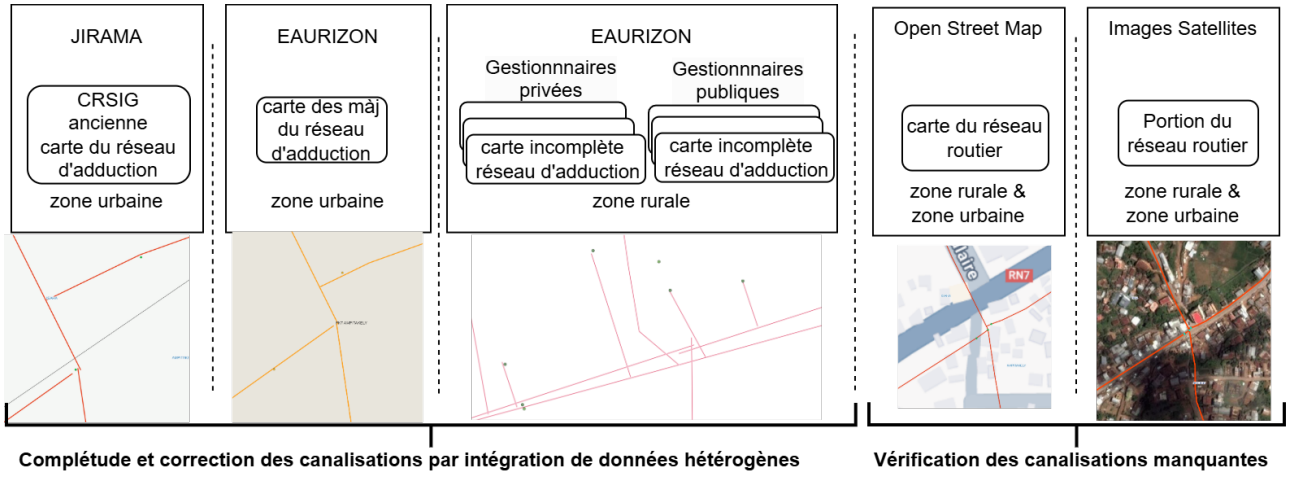


FIGURE 1 – Sources de données hétérogènes et approche envisagée pour la reconstruction du réseau d'adduction

à jour du réseau proviennent du programme EAURIZON, en charge de mettre à jour les infrastructures et par conséquent la carte du réseau d'adduction. En zone rurale, c'est le programme EAURIZON qui, par l'intermédiaire de gestionnaires publics ou privés, collecte les informations des réseaux dans les villages.

Challenges. Le premier challenge auquel s'adressera notre approche est celui de pouvoir intégrer les données hétérogènes et multi-sources mises à disposition par les différents gestionnaires en charge du réseau d'adduction. L'approche d'intégration de données seule ne suffira pas à compléter et corriger les données précises et incomplètes sur la structure spatiale des réseaux d'adduction. Le second challenge consistera à étendre l'approche d'intégration de données par une approche statistique. Cette approche statistique s'appuiera sur des informations indirectes telles que celles issues du réseau routier pour compléter la reconstruction du réseau d'adduction.

3 Travaux en cours

Les données collectées par les deux acteurs principaux (JIRAMA, EAURIZON) n'identifient pas de la même façon les ouvrages de traitement et les appareillages dont ils ont la charge. De même, ils ne localisent pas ouvrages et appareillages en question avec le même référentiel. L'entreprise JIRAMA utilise le système de référentiel géographique EPSG 29702 : Tananarive (Paris)/Laborde Grid approximation pour localiser les ouvrages et les associer à un toponyme (nom de quartier, ...). Le programme EAURIZON, de son côté, utilise le système de référencement géographique EPSG 4326 :WGS 84, mais ne localise pas ses ouvrages à l'aide de toponymes.

Un modèle spatial harmonisé intégrant des données multi-sources. Dans le contexte des systèmes d'information du domaine de l'eau potable, la commission de validation des données pour l'information spatialisée (COVADIS) [3] propose un standard de description des données pour les Réseaux d'Adduction d'Eau Potable et Assainissement (RAEPA). Ce standard présente la recommandation du cadre technique minimal pour l'échange de données descriptives du fonctionnement de ces réseaux. L'ontologie proposée par [1] et les recommandations COVADIS ont été conçues pour décrire les différents éléments et le fonctionnement des réseaux d'adduction et de traitement des eaux. Cependant, elles n'ont pas été conçues pour permettre le suivi de la distribution ou de la production d'eau potable. Afin d'harmoniser les données en provenance de différents gestionnaires, nous nous appuyons sur le modèle harmonisé présenté en Figure 2. Ce modèle s'inspire de la représentation issue du modèle COVADIS [3] mais diffère de ce dernier dans la représentation des concepts de réseaux et de canalisations. Dans le modèle présenté Figure 2, un réseau est une collection de conduits. Chaque conduit est délimité par deux extrémités représentant chacune un nœud. Un nœud est un élément du réseau ayant a minima une position géographique et se localise dans une limite administrative. Une canalisation, c.à.d., une suite de conduits connectés en ligne) est une spécialisation d'un réseau. Un ouvrage, comme par exemple un point de captage, une station de traitement, un réservoir ou encore un point de distribution, est un nœud localisé dans une limite administrative et produisant un type d'eau (brute, traitée, stockée, distribuée). Un appareillage comme par exemple une connexion en T entre des conduits est aussi un nœud localisé permettant d'interconnecter des canalisations.

4 Perspectives

L'intégration et l'harmonisation de la structure spatiale du réseau d'adduction en provenance des différents gestionnaires ne suffisent pas. L'hétérogénéité des données collectées par JIRAMA et EAURIZON entraîne des incohérences qui compromettent l'analyse du suivi de l'adduction d'eau potable. L'un des défis majeurs réside dans la gestion de l'incertitude des données. Cette incertitude est due

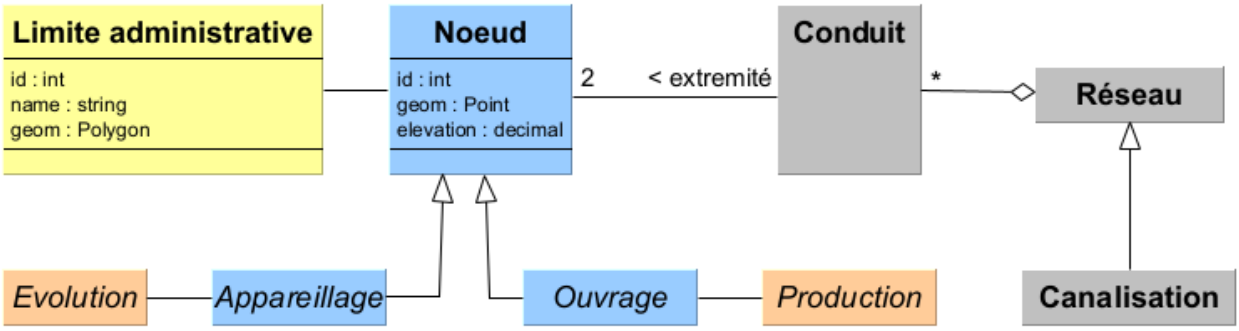


FIGURE 2 – Modèle du réseau d'adduction

à plusieurs facteurs : erreurs de mesure, hétérogénéité des sources de données, incohérences de saisie, lacunes spatiales et temporelles, et conditions environnementales fluctuantes. Dans le contexte de la reconstruction spatiale du réseau d'adduction, le défi est de corriger la localisation incomplète, et parfois erronée, des ouvrages afin de pouvoir reconnecter les conduits qui apparaissent anormalement déconnectés. Pour répondre à ce défi, nous partons de l'hypothèse que le réseau d'adduction est à proximité du réseau routier. Cette hypothèse repose sur le fait que, pour maintenir des canalisations d'eau, les équipes d'intervention doivent y accéder. Ainsi, nous comptons appuyer notre approche sur la localisation du réseau routier comme par exemple la description du réseau issue d'Open Street Map (OSM) ou encore les images satellites (c.f., Figure 1). Les données issues d'Open Street Map sont des données issues de collectes participatives. Ces données, bien que souvent représentatives du réseau de routes existantes, sont parfois incomplètes et non mises à jour.

Modélisation de l'incertitude spatiale et temporelle. Les données spatiales et temporelles relatives au réseau d'adduction d'eau sont souvent sujettes à des variations dues aux méthodes de collecte et aux écarts entre différents référentiels géographiques [4]. Pour quantifier ces incertitudes, nous investiguerons :
- Les réseaux bayésiens : qui modélisent les relations probabilistes entre différentes variables (ex. : localisation des infrastructures, volumes d'eau consommés).
- Les modèles de logique floue : permettant de représenter des informations incertaines ou incomplètes en assignant des degrés de confiance aux données spatiales et temporelles. Ces techniques nous permettront d'attribuer des marges d'incertitude aux valeurs collectées et d'améliorer la robustesse des analyses sur la distribution de l'eau potable.

Détection et correction des erreurs dans les bases de données. Nous comptons nous appuyer sur le tracé de routes disponible dans OSM et mettre en œuvre des modèles d'apprentissage automatique pour compléter les éventuelles routes manquantes à partir d'images satellites. Nous disposons d'images satellites Sentinel-2 à 10 mètres de résolution. La segmentation sémantique sera réalisée via un modèle U-Net (architecture CNN), spécialisé dans l'extraction des réseaux routiers. Les résultats seront ensuite affinés par un post-traitement incluant les opérations morphologiques pour corriger les artefacts et la vectorisation pour générer des tracés exploitables [5]. Pour la correction des erreurs, nous utiliserons des modèles bayésiens qui permettent d'évaluer l'incertitude des détections, complétés par des algorithmes d'imputation k plus proches voisins (k-NN) pour reconstruire les segments manquants et des règles topologiques pour garantir la cohérence du réseau [2]. Cette approche permet d'identifier et de corriger les discordances cartographiques, tout en préservant la structure du réseau.

Références

[1] Yassine Belghaddar, Abderrahmane Seriai, Ahlame Begdouri, Carole Delenne, Nanee Chahinian, Bachar Rima, and Mustapha Derras. Towards a generic fusion framework for underground networks involving model-driven engineering domain. *International Journal of Information Science and Technology*, 6(2) :8–19, 2022.

[2] BILAL BOUDJEHEM. Le traitement des données manquantes dans le «big data» médical. ., 2022.

[3] Covadis. Standard de données réseaux d'aep et d'assainissement, 2019.

[4] Aarthi Aishwarya Devendran and Gnanappazham Lakshmanan. A review on accuracy and uncertainty of spatial data and analyses with special reference to urban and hydrological modelling. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 09–12 December 2014, Hyderabad, India*, 2 :171–178, 2014.

[5] T Keijzer, Aafke Schipper, J Meijer, and W Nijland. Detecting roads from space : testing the potential of sentinel-1 sar imagery and deep learning for automated road mapping. *PBL Netherlands Environmental Assessment Agency*, 1(4417) :1–46, 2022.

MAEVa : Une approche hybride pour la mise en relation des variables expérimentales agroécologiques

Oussama Mechhour[1,2,3,4*], Sandrine Auzoux[1,2], Clément Jonquet[5,6], Mathieu Roche[3,4]

[1] CIRAD, UPR AIDA, F-34398 Montpellier, France
[2]AIDA, CIRAD, Université de Montpellier, Montpellier, France
[3] CIRAD, UMR TETIS, F-34398 Montpellier, France
[4] TETIS, Université de Montpellier, AgroParisTech, CIRAD, INRAE, Montpellier, France
[5] MISTEA, Univ. de Montpellier, INRAE, Institut Agro, Montpellier, France
[6] LIRMM, Univ. de Montpellier, CNRS, Montpellier, France
[*]Corresponding author: oussama.mechhour@cirad.fr

Résumé

Les variables sources ou les propriétés observables utilisées pour décrire les expérimentations agroécologiques sont hétérogènes, non standardisées et multilingues, rendant leur compréhension, explication et utilisation difficiles dans la modélisation des systèmes de culture et les évaluations multicritères de la performance des systèmes agroécologiques. L'annotation des données via un vocabulaire contrôlé, appelé variables candidates de Agroecological Global Information System (AEGIS), constitue une solution. Les mesures de similarité textuelle jouent un rôle clé dans la désambiguïsation du sens des mots, l'appariement de schémas dans les bases de données et l'annotation des données. Les approches courantes incluent (a) la similarité fondée sur les chaînes de caractères, (b) sur le corpus, (c) sur les connaissances et (d) les approches hybrides combinant deux ou plusieurs de ces méthodes. Ce travail propose une approche hybride, Matching Agroecological Experiment Variables (MAEVa), visant à apparier les variables sources et candidates selon (1) l'appariement des noms, (2) celui des descriptions, (3) une combinaison linéaire de (1) et (2), et (4) une méthode de sélection des résultats pour l'évaluation finale. Pour l'appariement des noms, nous étendons BERT-base avec une couche d'attention multi-têtes externe (BERTmha). Pour les descriptions, nous enrichissons celles existantes avec l'API GPT-3.5 Turbo et utilisons TF-IDF pour la représentation vectorielle. Nos résultats montrent que BERTmha améliore la précision de plus de 11% par rapport à BERT-base seul et que notre corpus améliore celle de TF-IDF de plus de 4%. Notre évaluation (étape 4) montre que MAEVa atteint une précision de plus de 66% de $P@1$ à $P@10$.

Mots-clés

Propriétés observables, Similarité fondée sur les chaînes de caractères, Similarité fondée sur les corpus, Similarité hybride, Modèles de langage pré-entraînés (PLMs), Grands modèles de langage (LLMs)

Abstract

Source variables or observable properties used to describe agroecological experiments are heterogeneous, non-standardized, and multilingual, making them challenging to understand, explain, and use in cropping system modeling and multicriteria evaluations of agroecological system performance. Data annotation via a controlled vocabulary, known as candidate variables from the Agroecological Global Information System (AEGIS), offers a solution. Text similarity measures play crucial roles in tasks such as word sense disambiguation, schema matching in databases, and data annotation. Commonly used measures include (a) string-based, (b) corpus-based, (c) knowledge-based, and (d) hybrid-based similarity, which combines two or more of these methods. This work introduces a hybrid approach called Matching Agroecological Experiment Variables (MAEVa), designed to match source and candidate variables based on (1) matching variable names, (2) matching variable descriptions, (3) a combination of (1) and (2) via a linear function, and (4) a method for selecting results for the final evaluation. For matching variable names, we propose a novel approach that extends BERT-base with an external multi-head attention layer (BERTmha). For matching variable descriptions, we augment existing descriptions using GPT-3.5 Turbo API to provide richer contextual information and employ TF-IDF to construct the vector space. Our experimental results demonstrate that BERTmha improves the precision of matching variable names by more than 11% compared to BERT-base alone, and that our constructed corpus enhances TF-IDF-based matching by more than 4%. Our evaluation (step 4) shows that MAEVa achieves a precision of over 66% from $P@1$ to $P@10$.

Keywords

Observable properties, String-based similarity, Corpus-based similarity, Hybrid-based similarity, Pre-trained language models (PLMs), Large language models (LLMs)

1 Introduction

Les expérimentations agroécologiques génèrent des bases de données pouvant être multi-échelles (plante, système de culture, exploitation, paysage, territoire), multi-espèces (cultures, plantes associées, adventices, plantes fourragères) et multidisciplinaires (agronomie, malherbologie, entomologie, sciences économiques et sociales, environnement, informatique). Les variables sources ou propriétés observables [1] utilisées pour décrire les expérimentations agroécologiques sont souvent nommées et décrites à l'aide d'homonymes, de synonymes, d'acronymes, de multiples langues et parfois de termes non standardisés, ce qui peut rendre leur interprétation et leur explication complexes. Elles sont également mesurées avec des unités hétérogènes et présentent une hétérogénéité d'un point de vue linguistique, structurel, sémantique, syntaxique et taxonomique. Cette complexité les rend difficiles à utiliser dans la modélisation des systèmes de culture et l'évaluation multicritère de la performance des systèmes agroécologiques. Pour relever ces défis, l'annotation des données via un vocabulaire contrôlé ou une ontologie commune constitue une solution. Afin d'harmoniser ces variables sources non standardisées et hétérogènes, le centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD) a développé *Agroecological Global Information System* (AEGIS) [1]. AEGIS intègre une chaîne harmonisée d'acquisition et de traitement des données utilisant un ensemble de variables candidates, combinant des termes sémantiques issus d'ontologies de référence (ontologie des plantes, des cultures, de l'environnement et de l'agronomie) et des connaissances expertes en agroécologie. L'approche adoptée dans AEGIS permet aux chercheurs d'utiliser librement leurs propres noms de variables sources, descriptions et unités de mesure, tout en fournissant une liste de variables candidates pour harmoniser et standardiser ces variables sources.

Dans ce travail, nous avons utilisé des variables sources hétérogènes collectées lors d'expérimentations sur la canne à sucre associée à des plantes de service, issues des projets CanecoH [2], Ecocanne [3], et AgriecoH, conduits par eRcane [4] en collaboration avec le CIRAD [5] à La Réunion.

La Table 1 présente des exemples de variables sources, et la Table 2 illustre des exemples de variables candidates. Ces tableaux mettent en évidence la complexité de ces variables. Par exemple, dans la Table 1, le premier nom de variable est multilingue : il contient *"yield"* en anglais et *"CAS"*, une abréviation française pour *"canne à sucre"*. Le second nom de variable est en français, et différentes manières d'écrire les unités de mesure existent. Par exemple, les deux notations suivantes représentent une division différemment : *"t.ha^{-1}"* et *"kg/m^2"*. Dans la Table 2, une caractéristique supplémentaire, *"méthode de calcul"*, est ajoutée, alors qu'elle est absente de la Table 1. Les lignes 2 et 3 de la Table 1 se réfèrent à la même variable que la ligne 3 de la Table 2, bien que leurs descriptions soient lexiquement différentes, elles reflètent néanmoins la même information.

Comme illustré dans les Tables 1 et 2, les variables sont principalement exprimées sous forme de descriptions textuelles, ce qui rend possible l'utilisation de mesures de similarité textuelle pour la mise en lien des variables sources et candidates. La mesure de similarité textuelle est un concept fondamental en théorie de l'information et est utilisée pour évaluer la proportion de contenu partagé entre deux textes [16]. Un score de similarité élevé indique une forte proximité entre les textes. La similarité textuelle consiste à évaluer la proximité entre deux textes en tenant compte à la fois des similarités lexicales et sémantiques. Cela signifie que même si deux textes ne partagent pas les mêmes mots, ils peuvent néanmoins véhiculer une sémantique proche. Ce concept est devenu central dans plusieurs domaines du traitement automatique des langues naturelles (TALN), notamment les systèmes de questions-réponses automatiques [11], la traduction automatique [30] et l'appariement de documents [23]. Les mesures de similarité textuelle courantes [6] peuvent être regroupées en quatre catégories : (1) les mesures fondées sur les chaînes de caractères (lexicales), (2) les mesures qui s'appuient sur les corpus, (3) les mesures fondées sur les connaissances et (4) les mesures hybrides.

Les mesures de similarité basées sur les chaînes de caractères évaluent la ressemblance entre deux textes en se basant uniquement sur la comparaison littérale des mots ou des caractères, sans prendre en compte les synonymes, le contexte ou les relations sémantiques entre les mots [9]. **Les mesures de similarité fondées sur les corpus** utilisent des informations dérivées d'un corpus pour calculer la similarité, en exploitant des caractéristiques textuelles ou des probabilités de cooccurrence. Par exemple, *TF–IDF* (*Term Frequency–Inverse Document Frequency*) [12] est une mesure qui se concentre principalement sur les aspects quantitatifs des mots dans un texte, sans prendre en compte l'ordre des mots, le contexte ou les relations sémantiques entre eux. Une autre approche est la représentation des mots sous forme de vecteurs denses, qui peuvent être statiques ou dynamiques. Les vecteurs statiques, tels que ceux produits par Word2Vec [18], FastText [13] et GloVe [21], sont indépendants du contexte. En revanche, les vecteurs dynamiques, tels que ceux produits par BERT [2], capturent les variations contextuelles des mots. **Les mesures de similarité basées sur les connaissances** reposent sur des réseaux structurés de concepts sémantiquement connectés pour évaluer la similarité entre les mots et peuvent être étendues à l'analyse au niveau des phrases. Ces réseaux sont souvent spécifiques à un domaine (par exemple, la biomédecine ou le droit) ou génériques, comme WordNet, qui est largement utilisé pour la mesure de similarité basée sur les connaissances [6]. **Les mesures de similarité hybrides** combinent deux ou plusieurs de ces approches de mesure de similarité

1. Une propriété observable est la description d'un élément observé ou dérivé.
2. https://ecophytopic.fr/dephy/concevoir-son-systeme/projet-canecoh
3. https://umr-pvbmt.cirad.fr/recherche/principaux-projets/ecocanne
4. https://www.ercane.re/en/home/
5. https://www.cirad.fr/

TABLE 1 – Exemples de variables sources. Chaque nom est une combinaison du nom de la variable et de son unité de mesure, avec l'unité apparaissant après le dernier underscore.

Noms	Descriptions	Unités de mesure
Yield_CAS_t.ha^{-1}	Cane yield (in fresh machinable stem)	t.ha^{-1}
Rec_globale_plein_%	Full weed and service plant coverage	%
Cov_end_CP1_%	Cover plant 1 coverage at the end de the trial	%
DM_end_weed_kg/m^2	Weed aerial dry mass at the end de the trial	kg/m^2
...	...	...

TABLE 2 – Exemples de variables candidates. Chaque nom est une combinaison du nom de la variable et de son échelle, avec l'échelle apparaissant après le dernier underscore.

Noms	Descriptions (Traits)	Échelles	Méthodes de calcul
leaf_plant_dm_kg	Measurement de foliar dry biomass at the individual level (plant scale)	kg	Common measurement method
abv_om_dm_content_%	Organic matter concentration de the WAB	%	Organic matter concentration defined based on dry matter concentration and mineral concentration
plant_ground_cover_%	Measurement de plant (or species) recovery by ceptometry	%	Common measurement method
fruit_plant_fm_kg	Measurement de fresh fruit biomass at the individual level (plant scale)	kg	Common measurement method
...	...	...	...

textuelle afin d'exploiter leurs forces respectives et d'améliorer la précision des mesures de similarité.

Cette étude introduit une approche hybride appelée "Matching Agroecological Experiment Variables" (MAEVa), conçue pour mettre en correspondance les variables sources et candidates selon (1) la mise en lien des noms de variables, (2) la mise en lien des descriptions de variables, (3) une combinaison de (1) et (2) via une fonction linéaire, et (4) une méthode de sélection des résultats pour l'évaluation finale.

Les principales contributions de cet article sont :

1. Une extension de BERT avec une couche d'attention multi-têtes externe (BERTmha) pour l'appariement des noms de variables.
2. Une augmentation des données via GPT-3.5 Turbo et l'utilisation de TF-IDF pour l'appariement des descriptions.
3. MAEVa, une approche hybride et générique adaptable à toute mesure de similarité textuelle, technique de construction de corpus et méthode de combinaison.

La suite de cet article est structurée comme suit. La Section 2 présente un état de l'art des mesures de similarité textuelle, en mettant en évidence leurs limites et ce que nous proposons dans nos travaux. La Section 3 introduit notre approche hybride MAEVa avec des expérimentations, résultats et une discussion des résultats détaillés en Section 4. Enfin, la Section 5 conclut l'article avec un résumé, les limites de notre travail et des perspectives.

2 État de l'art

Dans cette section, nous nous concentrerons sur les travaux connexes concernant les mesures de similarité fondées sur les corpus et les mesures de similarité hybrides.

2.1 Mesures de similarité basées sur les corpus

Les mesures de similarité basées sur les corpus ont été largement appliquées dans divers travaux. [20] a comparé TF–IDF (Baseline) [12], FastText [13], Doc2Vec [14], BERT [2] et ADA [8] dans des tâches de recommandation et l'analyse de rapports de bugs. [20] a démontré que BERT surpassait généralement les autres modèles en termes de rappel, suivi d'ADA, Doc2Vec, FastText et TF–IDF. Dans une autre étude, [7] a comparé TF–IDF, Doc2Vec, BERT et sentence-BERT (SBERT) [24] avec la similarité cosinus pour mesurer la similarité entre des thèses et mémoires académiques. Les résultats ont montré que TF–IDF surpassait les autres méthodes en termes de précision, rappel, score F1 et temps de traitement, suivi de Doc2Vec, BERT et SBERT.

2.2 Mesures de similarité hybrides

Deux ou plusieurs mesures de similarité textuelle, telles que les méthodes fondées sur les chaînes de caractères, les corpus et les connaissances, peuvent être combinées afin d'exploiter les avantages de chacune et d'améliorer la précision des mesures de similarité. Plusieurs études ont exploré ces approches. [4] a proposé une approche hybride appelée mesure de similarité sémantique agrégée (ASSM) pour comparer deux mots. La similarité sémantique entre deux mots (ou paires de mots) correspond à la similarité maximale obtenue à partir de deux méthodes : (1) la similarité cosinus entre les vecteurs obtenus avec Word2Vec préentraîné de Google, basé sur un corpus de 100 milliards de mots du Google News dataset, et (2) une fonction linéaire combinant la similarité cosinus de Word2Vec avec la similarité de Wu & Palmer [31]. [3] a comparé ELMo [22], les embeddings de bases de connaissances (KBs) via ComplEx [29] et la concaténation de ces méthodes pour les tâches de typage d'entités et de typage de relations, qui sont des tâches de classification multiclasse. Enfin, [27] a analysé l'efficacité de la similarité cosinus appliquée à plusieurs méthodes d'embedding de mots pour capturer la similarité lexicale entre des paires de mots de cinq ensembles de données différents. Les auteurs ont évalué Word2Vec [18], GloVe [21], FastText [13], LexVec [25] et ConceptNetNumberbatch, un modèle hybride combinant GloVe, Word2Vec et des réseaux sémantiques tels que ConceptNet [26] et PPDB [5].

2.3 Comparaison des approches

Nous avons résumé diverses méthodes et travaux existants pour mesurer la similarité textuelle. Chaque méthode présente ses propres avantages et inconvénients. Par exemple, les *approches fondées sur la fréquence ou les statistiques* traitent deux mots ayant une sémantique proche de manière indépendante et ne tiennent pas compte de l'ordre des mots. En raison de leur nature purement statistique, elles peuvent présenter certaines limites, telles que la *malédiction de la dimensionnalité*, où les vecteurs deviennent trop volumineux pour être gérés efficacement, ou encore des problèmes de *mots hors vocabulaire* (*out-of-vocabulary* - OOV) lorsque les mots dans de nouveaux textes ne sont pas présents dans le corpus d'origine. Les *mesures de similarité fondées sur une fenêtre contextuelle restreinte*, telles que Word2Vec, FastText et GloVe, qui utilisent une fenêtre de contexte limitée et attribuent un unique vecteur dense à chaque mot, présentent également des limites : quel que soit le contexte d'un mot, sa représentation vectorielle reste identique. À l'inverse, des méthodes comme BERT, RoBERTa et XLNet prennent en compte le contexte en générant plusieurs vecteurs pour un même mot en fonction de son environnement textuel immédiat, ce qui améliore la représentation sémantique du contexte. Cependant, ces méthodes sont moins efficaces lorsqu'un contexte étendu est nécessaire pour comprendre pleinement le sens du texte ou du mot. Pour exploiter les forces de chaque méthode tout en surmontant leurs limites, les *mesures de similarité hybrides* ont été introduites. Ces mesures hybrides combinent souvent des méthodes telles qu'elles sont, sans modification ni amélioration. Pour remédier à cette limite, nous proposons une approche hybride novatrice appelée *MAEVa*, qui combine les mesures de similarité fondées sur les corpus grâce à l'appariement des noms et des descriptions des variables. Pour l'appariement des noms, nous proposons une nouvelle approche où nous étendons BERT-base avec une couche d'attention multi-têtes externe (BERTmha). Pour les descriptions, nous enrichissons celles existantes avec l'API GPT-3.5 Turbo et utilisons TF-IDF pour la représentation vectorielle.

3 Approches proposées

MAEVa est un pipeline utilisé pour apparier les variables sources et candidates en procédant comme suit : (1) appariement des noms de variables, (2) appariement des descriptions de variables, (3) combinaison de (1) et (2) dans une fonction linéaire, et (4) une méthode de sélection des résultats pour l'évaluation finale. Avant de décrire chaque étape, nous commençons par présenter le prétraitement des données.

3.1 Prétraitement des données

Le prétraitement des données joue un rôle crucial en transformant le corpus, les noms et descriptions des variables dans une forme adaptée afin de faciliter leur appariement. Le processus utilisé dans cet article comprend plusieurs étapes appliquées dans un ordre précis. La fonction `clean_text()` nettoie les variables sources et candidates ainsi que le corpus en supprimant les chiffres, les parenthèses et leur contenu, et en convertissant les textes en minuscules, garantissant ainsi que l'appariement des variables est insensible à la casse. Ensuite, `remove_stopwords()` élimine les mots vides (*stopwords*), qui n'apportent pas de signification spécifique dans le contexte de l'agroécologie, mettant ainsi l'accent sur les termes clés pour une meilleure mise en correspondance des variables. La fonction `lemmatize()` applique la lemmatisation, qui réduit les mots à leur forme canonique (ex. *rédiges*, *rédigé* et *rédigeras* deviennent *rédiger*), améliorant ainsi la précision des calculs de similarité en comparant les lemmes plutôt que les différentes formes d'un mot. Par ailleurs, `remove_punctuation()` supprime la ponctuation des descriptions et du corpus pour favoriser l'appariement des variables. Enfin, `replace_synonyms()` facilite la mise en correspondance en remplaçant certains mots par leurs synonymes via WordNet, par exemple, *niveau* peut être remplacé par *degré*. Toutes ces fonctions sont appliquées dans leur ordre d'apparition, chacune prenant en entrée la sortie de la fonction précédente, au corpus ainsi qu'aux descriptions des variables sources et candidates. Étant donné que la majorité de nos noms de variables sont souvent des abréviations et ne contiennent pas de ponctuation, nous avons appliqué uniquement les trois premières fonctions pour leur prétraitement. Celles-ci peuvent, dans certains cas, modifier significativement le nom de la variable. Par exemple, la variable 'Number of leaves_n' est transformée en 'number leaf_n' après l'application des

trois fonctions. À l'inverse, certains noms restent inchangés après prétraitement, comme 'abv_om_dm_content_%'. Après cette étape de prétraitement des données, nous détaillons chaque étape du pipeline MAEVa dans les sous-sections suivantes.

3.2 Appariement des noms de variables

L'objectif de l'appariement des noms de variables est de comparer les variables à travers leurs noms (Étape 1). Dans cette étape, nous avons appliqué trois fonctions pour le prétraitement des noms de variables, comme décrit précédemment. Nous avons ensuite utilisé les sorties comme entrées de : (a) BERT (baseline) et (b) notre approche BERTmha pour apparier les noms de variables.

Nous avons utilisé BERT pour encoder chaque nom de variable, puis nous avons appliqué la similarité cosinus afin de mesurer la similarité entre chaque nom de variable source et tous les noms de variables candidates. Ensuite, les résultats ont été classés du candidat le plus similaire au moins similaire pour chaque nom de variable source. Concernant l'approche originale BERTmha que nous proposons, nous avons introduit les noms de variables sources prétraités dans BERT et avons extrait les derniers états cachés contenant l'encodage de chaque nom de variable source. Ensuite, les tenseurs des derniers états cachés du modèle BERT pour les noms de variables sources ont été utilisés comme entrées pour un mécanisme d'attention multi-têtes externe. Ces tenseurs ont été traités à travers des couches linéaires afin de générer les matrices de requête (Q), clé (K) et valeur (V) nécessaires au calcul de l'attention. Dans notre cas, les requêtes (Q) et les clés (K) ont été dérivées des états cachés de BERT pour les noms de variables sources, nous permettant ainsi de mieux comprendre les relations entre les mots composant chaque nom de variable source.

L'attention multi-têtes est définie par :

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (1)$$

où d_k est la dimension des vecteurs de clé. Le mécanisme multi-têtes applique ce calcul sur plusieurs sous-espaces en parallèle :

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_h)W^O \qquad (2)$$

où chaque tête est calculée comme suit :

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \qquad (3)$$

avec W_i^Q, W_i^K, et W_i^V étant les matrices de projection pour la $i^{\text{ème}}$ tête, et W^O la matrice de projection finale. Ce mécanisme permet de capturer les dépendances contextuelles sous plusieurs perspectives, enrichissant ainsi la représentation des noms de variables.

Le mécanisme d'attention multi-têtes produit ensuite un nouvel encodage pour chaque nom de variable source. Ce processus comprend le calcul des scores d'attention, l'application de la fonction *softmax* pour obtenir des probabilités, l'utilisation du *dropout* pour la régularisation, et enfin l'agrégation des vecteurs de contexte pour produire la sortie finale de l'attention. Le même processus de BERTmha a été appliqué aux noms de variables candidates.

Ensuite, nous avons calculé la similarité entre chaque nom de variable source et tous les noms de variables candidates à travers leurs nouvelles embeddings via la similarité cosinus. Les résultats ont ensuite été classés du plus au moins similaire pour chaque nom de variable source. Dans ce travail, pour initialiser les couches d'attention multi-têtes, nous avons adopté une technique de distribution uniforme, dans laquelle les poids ont été initialisés aléatoirement dans l'intervalle $[-1, 1]$.

3.3 Appariement des descriptions de variables

L'objectif de l'appariement des descriptions de variables est de comparer les variables sur la base de leurs descriptions en s'appuyant sur un corpus (Étape 2). Nous avons commencé par générer automatiquement un nouveau corpus contenant une nouvelle description pour chaque variable en utilisant le prompt suivant sur l'API GPT-3.5 Turbo : *You are a farmer. You have knowledge about sugarcane culture and other relevant information. Always answer with the goal of describing [variable's description] and use these descriptions as data for training TF–IDF. Provide the results in a complete paragraph, with a maximum of 500 words.* Nous nous référons au corpus généré sous le nom de *Corpus (GPT-prompt)*. Ce corpus contient toutes les nouvelles descriptions générées pour l'ensemble de nos variables. Ensuite, nous avons prétraité le corpus en utilisant les cinq fonctions décrites en Subsection 3.1, puis nous avons utilisé le *Corpus (GPT-prompt)* prétraité en appliquant une pondération TF-IDF pour construire l'espace vectoriel et extrait la représentation de chaque description de variable.

Comme pour l'appariement des noms de variables, nous avons calculé la similarité cosinus entre chaque description de variable source et toutes les descriptions de variables candidates, puis nous avons classé les résultats du candidat le plus similaire au moins similaire pour chaque description de variable source.

3.4 Méthode de combinaison

À l'étape 3, nous avons utilisé une fonction linéaire qui combine l'appariement des noms et des descriptions de variables afin d'augmenter la précision du processus de mise en correspondance, comme illustré dans l'Équation 4.

$$\text{combinaison} = \alpha \cdot \text{B} + (1 - \alpha) \cdot A \qquad (4)$$

Cela signifie que pour chaque variable source, nous avons combiné son appariement fondé sur le nom (B) et celui basé sur la description (A) avec chaque variable candidate. Les résultats ont ensuite été classés du plus au moins similaire pour chaque variable source. A représente la similarité cosinus entre un nom de variable source et un nom de variable candidate donné, et B la similarité cosinus entre la même description source et la description candidate corres-

pondante. Le paramètre $\alpha \in]0, 1[$ est un facteur de pondération ajustant l'importance de A et B (Étape 3). Il influence fortement le résultat : de faibles variations peuvent avoir un impact notable. Une expérimentation rigoureuse est donc nécessaire pour en fixer la valeur optimale.

3.5 Méthode de sélection des résultats

Nous proposons de rassembler et synthétiser les résultats de toutes nos méthodes. Nous proposons que si au moins une méthode identifie correctement la variable candidate réelle pour une variable source donnée, nous la considérons comme un appariement correct.

4 Expérimentations

Dans ce travail, nous avons utilisé des échantillons de variables sélectionnés par des experts en agroécologie afin d'illustrer différentes complexités et de mesurer l'efficacité de MAEVa. Nous disposons de 84 variables sources, de 170 variables candidates, ainsi que du corpus généré par GPT (*Corpus (GPT-prompt)*) [6].

4.1 Configuration expérimentale

Dans cette étude, nous avons utilisé le modèle BERT-base (*"bert-base-uncased"*) de Hugging Face [7] avec 2 couches cachées (2HLs). La méthode de combinaison utilise un seul paramètre, α, dont la valeur a été testée entre 0.01 et 0.99 avec un pas de 0.01. Les meilleurs résultats ont été obtenus avec $\alpha = 0.25$. Pour l'attention multi-têtes, nous avons utilisé une régularisation par dropout (DT = 0.1), une distribution uniforme des poids (UDW) et 256 têtes (Hs).

4.2 Résultats et discussion

4.2.1 Résultats

Pour chaque variable source, nous avons calculé sa similarité avec toutes les variables candidates et classé les résultats du plus similaire au moins similaire. Étant donné que chaque variable source possède exactement une variable candidate correcte, nous utilisons la précision au rang K ($P@K$) comme métrique d'évaluation.

Métrique d'évaluation La précision au rang K (P@K) évalue la capacité du modèle à classer la variable candidate correcte parmi les K premières positions de la liste ordonnée par similarité pour chaque variable source. Formellement, $P@K$ est définie dans l'Équation 5 :

$$P@K = \frac{N_{\text{correct@}K}}{N_{\text{total}}} \tag{5}$$

Où :

- $N_{\text{correct@}K}$: Nombre de variables sources où la variable candidate correcte est classée parmi les K premières.
- N_{total} : Nombre total de variables sources (84).

La Table 3 présente les résultats de l'appariement des noms de variables, tandis que la Table 4 montre les résultats de l'appariement des descriptions de variables. La Table 5 donne les résultats finaux obtenus en combinant les résultats de la mise en lien des noms de variables, des descriptions et de la méthode de combinaison, ce qui signifie que si au moins une méthode identifie correctement la variable candidate réelle pour une variable source donnée, nous la considérons comme un appariement correct (Subsection 3.5).

TABLE 3 – Résultats de l'appariement des noms de variables.

Méthode	P@1	P@3	P@5	P@10
BERTcos	11.90%	28.57%	36.90%	53.57%
BERTmha	**30.90%**	**40.33%**	**51.00%**	**64.69%**

BERTcos : BERT-base (2 HLs) + similarité cosinus, BERTmha : BERT-base (2 HLs) + Attention Multi-Têtes (256 Hs, DT=0.1 et UDW) + similarité cosinus, HLs : Couches cachées, Hs : Têtes, DT : Régularisation par dropout, et UDW : Distribution uniforme des poids.

TABLE 4 – Résultats de l'appariement des descriptions de variables.

Méthode	Corpus	P@1	P@3	P@5	P@10
TF-IDF	×	33.33%	42.86%	51.19%	60.71%
TF-IDF	Corpus (GPT-prompt)	**38.10%**	**50.00%**	**60.71%**	**66.67%**

TABLE 5 – Résultats finaux de l'appariement des variables (MAEVa pipeline).

Noms	Descriptions	Corpus	P@1	P@3	P@5	P@10
BERTcos	TF-IDF	×	59.52%	72.61%	77.38%	84.52%
BERTmha	TF-IDF	Corpus (GPT-prompt)	**66.66%**	**83.33%**	**85.71%**	**86.90%**

BERTcos : BERT-base (2 HLs) + similarité cosinus, BERTmha : BERT-base (2 HLs) + Attention Multi-Têtes (256 Hs, DT=0.1 et UDW) + similarité cosinus, HLs : Couches cachées, Hs : Têtes, DT : Régularisation par dropout, et UDW : Distribution uniforme des poids.

4.2.2 Discussion

La Table 3 montre que notre approche BERTmha a amélioré les résultats de plus de 11 points de $P@1$ à $P@10$. Cela démontre que l'attention multi-têtes ajoutée permet de mieux "comprendre" les noms de variables que l'auto-attention utilisée avec BERT-base. La Table 4 montre une amélioration des résultats de plus de 4 points de $P@1$ à $P@10$, ce qui signifie que les nouvelles descriptions générées par l'API GPT-3.5 Turbo sont efficaces et permettent une meilleure représentation des variables.
Notre pipeline MAEVa, qui combine plusieures techniques d'appariement, a permis d'améliorer les résultats de plus de 2 points par rapport à la baseline. De plus, les valeurs obtenues avec $P@3$ de MAEVa sont proches des résultats avec $P@10$ de la baseline, comme illustré dans la Table 5. L'amélioration des résultats obtenus avec MAEVa montre que nos méthodes sont complémentaires et que la meilleure approche consiste à utiliser l'ensemble du pipeline pour obtenir des performances optimales. Nos approches présentent certaines limites. Le modèle BERT-base, généraliste, n'est pas adapté au domaine agroécologique ; un af-

6. https://doi.org/10.18167/DVN1/9X3IVR
7. https://huggingface.co/google-bert/bert-base-uncased

finement ou une spécialisation sur ce domaine serait plus pertinent. De plus, les noms de variables, souvent abrégés et peu contextualisés, nuisent à l'explicabilité. À l'inverse, les unités de mesure textuelles (kilogramme) sont plus discriminantes et explicatives. Nous prévoyons de nous concentrer sur ces dernières, en les normalisant, pour l'appariement des noms de variables. Enfin, TF-IDF ne tenant pas compte du contexte, nous envisageons d'adapter des modèles généralistes comme BERT et AgriBERT au domaine agroécologique pour l'appariement des descriptions.

5 Conclusion et perspectives

Dans ces travaux, nous avons présenté une approche hybride appelée MAEVa, qui combine l'appariement des noms et des descriptions de variables pour mettre en lien les variables sources et candidates. Nous avons proposé une nouvelle approche appelée BERTmha pour l'appariement des noms de variables. BERTmha encode les noms de variables et nous extrayons ses couches d'états cachés pour les injecter dans une couche d'attention multi-têtes externe afin de mieux capturer une certaine sémantique des noms de variables. BERTmha a surpassé BERT-base de plus de 11 points de $P@1$ à $P@10$. Pour l'appariement des descriptions de variables, nous avons généré un nouveau corpus de descriptions en utilisant l'API GPT-3.5 Turbo avec un prompt spécifique, puis nous avons appliqué la pondération TF-IDF et la similarité cosinus pour l'appariement. Les résultats montrent que nos nouvelles descriptions comportent des informations plus adaptées que les descriptions originales, et l'appariement des descriptions de variables a été amélioré de plus de 4 points.

Malgré ces résultats encourageants, nous n'avons pas appliqué de méthodes de similarité sémantique ni utilisé d'informations sémantiques dans notre corpus. L'un des principaux défis rencontrés était l'hétérogénéité des variables, en particulier leurs unités de mesure, ce qui peut compliquer le processus de mise en correspondance. Pour y remédier, nous avons initié des travaux sur l'extraction d'informations sémantiques liées aux unités de mesure, y compris les labels et URIs, en utilisant des ontologies telles que QUDT (*Quantities, Units, Dimensions, and Types*), TO (*Plant Trait Ontology*), PO (*Plant Ontology*), UO (*Units de Measurement Ontology*) et OM (*Ontology de Units de Measure*). Notre objectif est de normaliser ces unités et d'utiliser les informations sémantiques extraites comme descripteurs pour faciliter le processus de mise en lien. Étant donné que MAEVa est une approche hybride et générique qui peut être facilement adaptée à différentes mesures de similarité textuelle, diverses techniques de construction de corpus et différentes méthodes de combinaison, nous prévoyons d'intégrer des informations sémantiques en affinant un modèle de langage pré-entraîné tel que RoBERTa [17], BERT [2] ou des grands modèles de langue tels que GPT-4 [19] ou LLaMA [28]. Pour la construction de corpus, il est envisageable d'explorer d'autres techniques, telles que l'extraction d'informations sémantiques à partir d'ontologies et de thésaurus. Nous prévoyons également d'utiliser des techniques de *Retrieval-Augmented Generation* (RAG) [15] afin de rechercher les documents ou textes pertinents pour une variable donnée et d'utiliser ce contexte pour améliorer la génération de GPT. Enfin, nous envisageons d'appliquer d'autres modèles génératifs, tels que GPT-4, LLaMa et Mistral, pour générer et contextualiser nos variables [10].

Remerciements. Ce travail est financé par l'Agence Nationale de la Recherche (ANR) au titre de France 2030 portant la référence ANR-16-CONV-0004 (#DigitAg) et par le programme de recherche et d'innovation Horizon Europe dans le cadre de la convention 101081973 - IntercropValuES. Ce travail a bénéficié du soutien du Conseil régional de La Réunion, du ministère français de l'Agriculture et de l'Alimentation, de l'Union européenne (programme Feader, subvention AG/974/DAAF/2016-00096 et programme Feder, subvention GURTDI 20151501-0000735).

Références

[1] Sandrine Auzoux, Mathias Christina, François-Régis Goebel, Alizé Mansuy, and Daniel Marion. A dictionary of variables to harmonize data from agro-ecological experiments on sugarcane. In *Proceedings of the 3rd ISSCT Agricultural Engineering, Agronomy and Extension Workshop*, page 22, Saint-Gilles, La Réunion, 2018. ISSCT.

[2] 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, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

[3] Léa Dieudonat, Kelvin Han, Phyllicia Leavitt, and Esteban Marquer. Exploring the combination of contextual word embeddings and knowledge graph embeddings. *arXiv preprint arXiv :2004.08371*, 2020.

[4] Aissa Fellah, Ahmed Zahaf, and Atilla Elçi. Semantic similarity measure using a combination of word2vec and wordnet models. *Indonesian Journal of Electrical Engineering and Informatics*, 12(2) :455–464, 2024.

[5] Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. Ppdb : The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 758–764, Atlanta, Georgia, 2013. Association for Computational Linguistics.

[6] Wael H. Gomaa and Aly A. Fahmy. A survey of text similarity approaches. *International Journal of Computer Applications*, 68(13) :13–18, 2013.

[7] Ramadan T. Hassan and Nawzat S. Ahmed. Evaluating the efficacy of semantic similarity methods for comparison of academic thesis and dissertation texts. *Science Journal of University of Zakho*, 11(3) :396–402, 2023.

[8] Aryan Jadon and Shashank Kumar. Leveraging generative ai models for synthetic data generation in healthcare : Balancing research and privacy. In *2023 International Conference on Smart Applications, Communications and Networking (SmartNets)*, pages 1–4. IEEE, 2023.

[9] Matthew A. Jaro. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, 84(406) :414–420, 1989.

[10] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv :2310.06825*, 2023.

[11] Nanjiang Jiang and Marie-Catherine de Marneffe. Do you know that florence is packed with visitors ? evaluating state-of-the-art models of speaker commitment. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4208–4213, Florence, Italy, 2019. Association for Computational Linguistics.

[12] Karen Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1) :11–21, 1972.

[13] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv :1607.01759*, 2016.

[14] Quoc V. Le and Tomas Mikolov. Distributed representations of sentences and documents. *arXiv preprint arXiv :1405.4053*, 2014.

[15] 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. *arXiv preprint arXiv :2005.11401*, 2020.

[16] Dekang Lin. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, pages 296–304, San Francisco, CA, USA, 1998. Morgan Kaufmann.

[17] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta : A robustly optimized bert pretraining approach. *CoRR*, abs/1907.11692, 2019.

[18] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv :1301.3781*, 2013.

[19] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv :2303.08774*, 2023.

[20] Avinash Patil, Kihwan Han, and Aryan Jadon. A comparative study of text embedding models for semantic text similarity in bug reports. *arXiv preprint arXiv :2308.09193*, 2023.

[21] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove : Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, 2014. Association for Computational Linguistics.

[22] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 2227–2237, New Orleans, Louisiana, 2018. Association for Computational Linguistics.

[23] Hieu Pham, Minh-Thang Luong, and Christopher D. Manning. Learning distributed representations for multilingual text sequences. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 88–94, Denver, Colorado, USA, 2015. Association for Computational Linguistics.

[24] Nils Reimers and Iryna Gurevych. Sentence-bert : Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, 2019. Association for Computational Linguistics.

[25] Alexandre Salle, Aline Villavicencio, and Marco Idiart. Matrix factorization using window sampling and negative sampling for improved word representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, pages 419–424, Berlin, Germany, 2016. Association for Computational Linguistics.

[26] Robyn Speer and Catherine Havasi. Representing general relational knowledge in conceptnet 5. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3679–3686, Istanbul, Turkey, 2012. European Language Resources Association (ELRA).

[27] Martina Toshevska, Frosina Stojanovska, and Jovan Kalajdjieski. The ability of word embeddings to capture word similarities. *International Journal on Natural Language Computing*, 9(3) :25–42, 2020.

[28] H. Touvron and et al. Llama : Open and efficient foundation language models. In *arXiv preprint arXiv :2302.13971*, 2023.

[29] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 2071–2080, New York, NY, USA, 2016. PMLR.

[30] Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F. Wong, and Lidia S. Chao. Learning deep transformer models for machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1810–1822, Florence, Italy, 2019. Association for Computational Linguistics.

[31] Zhibiao Wu and Martha Palmer. Verb semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 133–138, Las Cruces, New Mexico, USA, 1994. Association for Computational Linguistics.

Session 3

Approches hybrides basées sur l'IA symbolique et l'apprentissage automatique : Application au domaine automobile

Hamid Ahaggach
Syartec, Aix-en-Provence, France
hahaggach@syartec.com

Lylia Abrouk
LIB Laboratory, UBE, Dijon, France
lylia.abrouk@u-bourgogne.fr

Eric Lebon
Syartec, Aix-en-Provence, France
elebon@syartec.com

Résumé

Cet article explore l'apport de l'intelligence artificielle à la gestion des connaissances et l'aide à la décision dans le secteur automobile. En combinant l'IA symbolique et l'apprentissage automatique, nous proposons des solutions pour structurer les connaissances issues des interactions des acteurs et des processus industriels. Les défis liés à l'hétérogénéité des données, à l'absence de standardisation et à la variabilité des marchés sont abordés, avec des méthodologies hybrides pour améliorer la prise de décision et la précision des prédictions.

Mots-clés

Intelligence artificielle, ontologie, apprentissage automatique, gestion des connaissances, secteur automobile.

Abstract

This article explore the contribution of artificial intelligence to knowledge management and decision support in the automotive sector. By combining symbolic AI and machine learning, we propose solutions to structure knowledge derived from stakeholder interactions and industrial processes. Challenges related to data heterogeneity, lack of standardization, and market variability are addressed, with hybrid methodologies to improve decision-making and forecasting accuracy.

Keywords

Artificial intelligence, ontology, machine learning, knowledge management, automotive sector.

1 Introduction

L'industrie automobile subit une transformation majeure sous l'impulsion de la numérisation et de la croissance exponentielle des volumes de données [2]. Ces données, souvent hétérogènes et non structurées, posent des défis complexes en termes de gestion et d'exploitation. Dans ce contexte, cet article explore les défis et les méthodologies proposées pour surmonter ces obstacles, en particulier dans les domaines de la reconnaissance d'entités nommées (NER), de l'extraction de relations (RE) et de la prédiction des coûts de réparation des dommages.

1.1 Défis dans le secteur automobile

Les défis rencontrés dans le domaine automobile sont multiples et requièrent une approche intégrée pour surmonter les obstacles techniques et scientifiques. Parmi ces défis, on peut citer :

- **Hétérogénéité des données** : Les données automobiles se présentent sous des formes variées, structurées et non structurées, limitant leur exploitation pour des analyses globales.
- **Absence de standardisation** : Les rapports textuels produits par les experts manquent souvent de structures uniformes, rendant leur exploitation difficile.
- **Variabilité des marchés** : Les différences régionales influencent considérablement les ventes de véhicules, compliquant l'élaboration de modèles prédictifs universels.
- **Qualité et accessibilité des données** : La qualité des données peut être compromise par des erreurs ou des lacunes, tandis que l'accès aux données essentielles est souvent limité par des barrières techniques ou juridiques.

2 Solutions proposées

Pour répondre à ces enjeux, nous avons proposé :

- **Ontologie dédiée au domaine automobile [1]** : Une ontologie pour modéliser les dommages subis par les véhicules, intégrant des concepts spécifiques (types de dommages, gravité) et un processus de peuplement ontologique.
- **Apprentissage automatique** : Des techniques pour extraire des entités et des relations significatives à partir de documents textuels.
- **Approches hybrides** : La combinaison des ontologies et de l'apprentissage automatique pour renforcer la précision et la robustesse des modèles.

La Figure 1 illustre la méthodologie adoptée pour l'estimation des coûts de réparation des dommages automobiles.

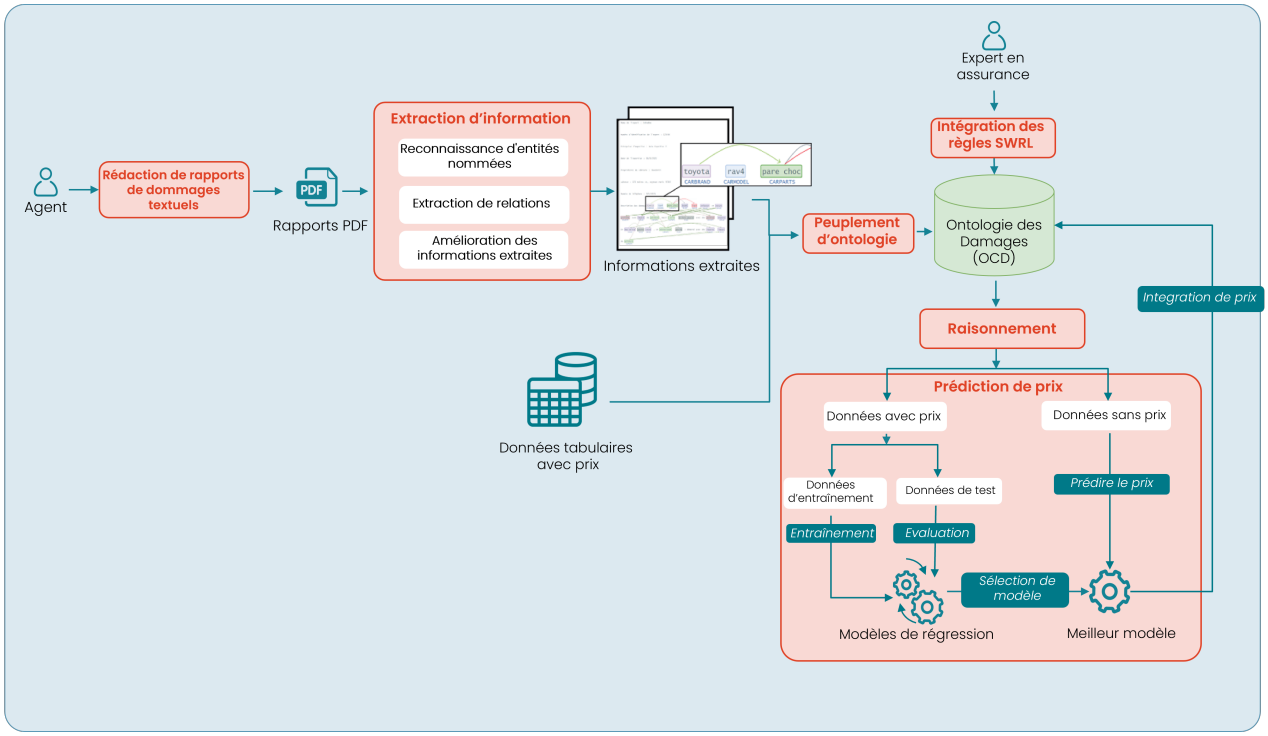


FIGURE 1 – Méthodologie pour la prédiction de prix de réparation des dommages

Cette approche se compose de trois étapes principales : l'extraction d'informations pour le peuplement de l'ontologie (*Ontology for Car Damage*), l'intégration des règles SWRL, et la prédiction des coûts de réparation.

2.1 Étape 1 : Extraction d'Informations et Peuplement d'Ontologie

Dans cette première étape, des informations clés sont extraites de rapports de dommages non structurés en utilisant des techniques de reconnaissance d'entités nommées et d'extraction de relations. Ces informations, sélectionnées en fonction de leur pertinence et de leur précision, sont ensuite intégrées dans l'ontologie *Ontology of Car Damages* (OCD). Par ailleurs, des données tabulaires issues des ateliers de réparation sont également exploitées pour enrichir cette ontologie.

2.2 Étape 2 : Intégration des règles SWRL

Cette étape consiste à intégrer des règles SWRL au sein de l'ontologie afin de permettre un raisonnement avancé sur les données collectées. Ces règles permettent d'identifier des liens logiques entre différentes entités et d'inférer de nouvelles connaissances qui ne sont pas explicitement mentionnées dans les rapports. L'intégration de ces règles améliore la qualité et la profondeur des informations disponibles pour l'étape suivante.

2.3 Étape 3 : Prédiction des coûts de réparation

Une fois les informations structurées et enrichies grâce au raisonnement ontologique, elles sont utilisées pour la prédiction des coûts de réparation. Cette phase repose sur des modèles de régression entraînés sur des données historiques, permettant d'estimer avec précision le coût des réparations en fonction des dommages observés.

3 Conclusion

Les travaux de recherche démontrent comment les ontologies et l'apprentissage automatique peuvent transformer la gestion des données dans l'industrie automobile. Les approches proposées surmontent les défis liés à l'hétérogénéité et au volume des données tout en ouvrant des perspectives pour des applications pratiques. Cependant, des défis persistent, notamment l'amélioration de la qualité des données et la structuration des données pour les exploiter ultérieurement dans le cadre de prédictions. Les méthodologies présentées offrent un cadre robuste pour exploiter des données complexes et soutenir la transformation numérique du secteur automobile.

Références

[1] Hamid Ahaggach, Lylia Abrouk, and Eric Lebon. Information extraction from automotive reports for ontology population. *Applied Ontology*, 19(2) :113–142, 2024.

[2] Meliha Koyuncuoğlu. The importance of digital transformation in the automotive industry. *Journal of Industrial Policy and Technology Management*, 7(1) :105–118, 2024.

Apprentissage profond séquentiel pour la prédiction/anticipation de l'hypotension artérielle non-invasive par le signal de pléthysmographie

Heni Lachouak[1],Youssef Mourchid[1], Yuehua Ding[2] ,Pierre-Grégoire Guinot[3]

[1] CESI LINEACT, Dijon, France
[2] CESI LINEACT, Strasbourg, France
[3] Department of Anesthesiology and Critical Care Medicine,
Dijon University Medical Centre, 21000 Dijon, France,
University of Burgundy Europe, LNC UMR1231, 21000 Dijon, France.

20 mai 2025

Résumé

L'hypotension peropératoire (une baisse anormale de la pression artérielle survenant pendant une intervention chirurgicale) représente un risque majeur pour les complications postopératoires, mais la surveillance actuelle reste invasive et coûteuse. La pléthysmographie numérique (méthode optique non invasive permettant de mesurer les variations du volume sanguin dans les tissus) offre une alternative non invasive, mais la prédiction en temps réel des épisodes hypotensifs demeure difficile en raison de la complexité des signaux physiologiques. Cet article propose une approche basée sur des réseaux Bidirectional Long Short-Term Memory (BiLSTM) pour capter les dépendances temporelles à court et long terme dans les signaux pléthysmographiques. Cette méthode assure une prédiction robuste de l'hypotension tout en maintenant une efficacité computationnelle. L'architecture bidirectionnelle permet une extraction complète des caractéristiques, anticipant les épisodes d'hypotension jusqu'à cinq minutes avant leur survenue. En améliorant la surveillance hémodynamique prédictive, cette approche permet une intervention proactive, réduisant les risques et optimisant les résultats postopératoires.

Mots-clés

Hypotension peropératoire, Pléthysmographie, BiLSTM, Surveillance hémodynamique non-invasive.

1 Introduction

L'hypotension artérielle peropératoire est un facteur majeur de complications graves telles que l'insuffisance rénale aiguë et les accidents vasculaires cérébraux, contribuant à la troisième cause de mortalité mondiale [8, 1]. Chaque année, plus de 300 millions de chirurgies sont réalisées, mais les outils de surveillance hémodynamiques restent insuffisants [4]. Les cathéters artériels invasifs, bien que précis, sont coûteux, complexes et peu utilisés, représentant moins de 10 % des interventions en raison de leur caractère invasif et de leur coût élevé [6]. En réponse, la pléthysmographie (PPG) numérique, corrélée aux variations hémodynamiques, s'avère une alternative prometteuse, bien que les méthodes actuelles manquent de précision et de prédiction anticipée [5]. Les premières tentatives avec l'apprentissage automatique ont montré des résultats prometteurs, mais leur précision reste insuffisante, notamment avec des signaux PPG, avec une précision prédictive inférieure à 70 % [3].D'autres approches basées sur l'apprentissage profond ont été développées, exploitant simultanément plusieurs signaux[2, 7].Les approches explorées dans ces travaux s'inscrivent dans différentes familles d'apprentissage profond supervisé, incluant des réseaux de neurones profonds multicanaux, des modèles séquentiels à base d'attention tels que les Transformers, ainsi que des architectures hybrides combinant analyse fréquentielle et réseaux profonds interprétables tous entraînés sur le benchmark ouvert VitalDB(Base de données des signaux cardiaque de patients ayant subi une chirurgie) .Cette étude propose une approche utilisant un modèle BiLSTM entraîné sur les données VitalDB pour prédire l'hypotension peropératoire à partir de signaux de pléthysmographie. .

2 Méthodologie

2.1 Préparation des données

Nous avons utilisé la base de données VitalDB. Les signaux PPG ont été extraits et synchronisés avec les enregistrements de pression artérielle à l'aide de corrélation croisée. Un protocole d'exclusion rigoureux a éliminé les enregistrements avec des fréquences cardiaques anormales (<40 bpm ou >150 bpm) et des valeurs de pression artérielle moyenne (PAM) non pertinentes (<40 mmHg ou >120 mmHg), garantissant ainsi la qualité et la fiabilité des données pour l'analyse.

2.2 Annotation des données

Pour définir l'hypotension peropératoire, un seuil de PAM < 65 mmHg pendant au moins 1 minute sera utilisé comme

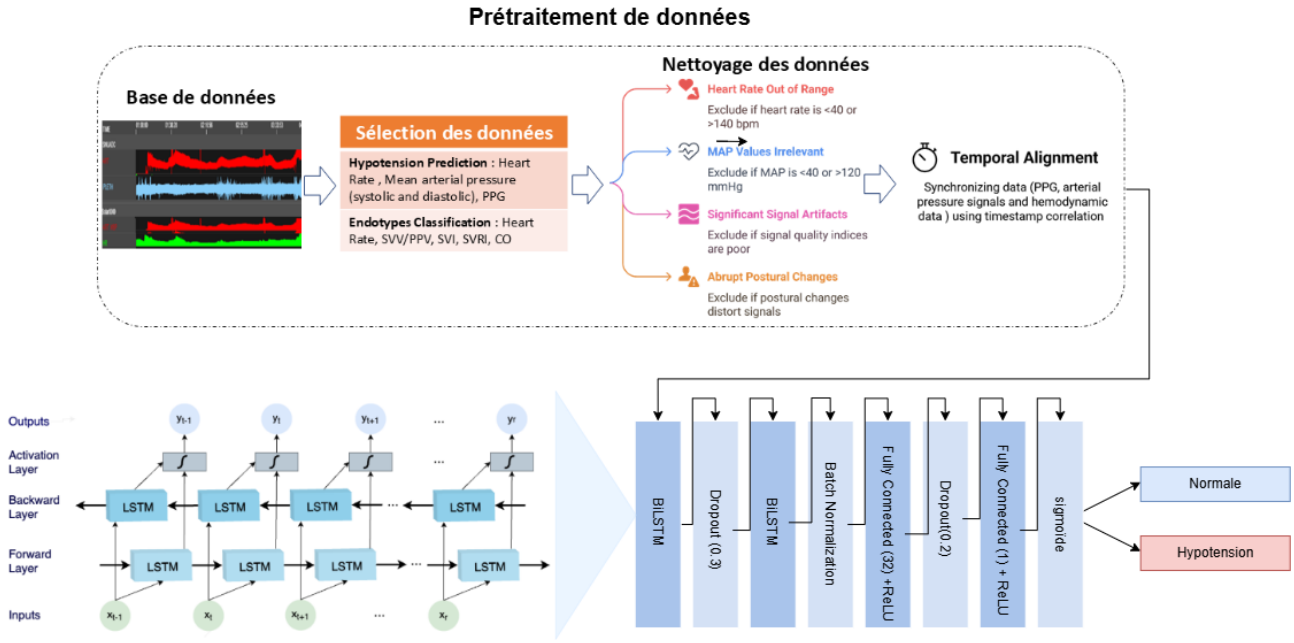


FIGURE 1 – Les étapes de notre approches proposée pour la prédiction des moments d'hypotension 5 min avant l'événement.

critère principal, basé sur la littérature hémodynamique. Les cas seront classés comme positifs si ce critère est vérifié et négatifs si la PAM reste au-dessus de 65 mmHg. Les données seront divisées en un intervalle de 10 minutes pour l'entrée du modèle et un intervalle de prévision de 5 minutes précédant un événement d'hypotension. Si aucun événement n'a lieu, le cas sera étiqueté négatif. Cette segmentation permettra au modèle de prédire l'hypotension peropératoire tout en étant adaptable à la surveillance en temps réel.

2.3 Architecture proposée

L'architecture proposée repose sur un modèle BiLSTM permettant la prédiction des épisodes d'hypotension jusqu'à cinq minutes avant leur survenue, en exploitant exclusivement le signal de pléthysmographie numérique, une méthode non invasive. Deux couches empilées de BiLSTM (128 unités chacune) capturent les dépendances temporelles du signal, avec une régularisation par dropout (30% et 20%) et une normalisation par lot (BatchNormalization) pour stabiliser l'apprentissage. Une couche dense de 32 neurones (ReLU) précède la sortie sigmoïde, assurant une classification binaire précise. Le modèle est optimisé avec Adam (learning rate = 5e-4, gradient clipping) et entraîné sur 150 époques, avec une taille de batch de 128 et une validation sur 30% des données. Des stratégies comme ReduceLROnPlateau, early stopping (20 époques) et ModelCheckpoint garantissent une convergence efficace. Cette approche améliore la prédiction précoce des épisodes hypotensifs en temps réel, offrant une alternative fiable aux techniques invasives actuelles.

3 Résultats

Le tableau 2 montre la supériorité de notre approche BiLSTM par rapport à l'état de l'art, avec une AUC-ROC (Area Under the Receiver Operating Characteristic Curve) de 97.3%. Cette métrique évalue la capacité du modèle à distinguer correctement les événements hypotensifs des non-événements, indépendamment du seuil de classification. De plus, notre modèle atteint une AUPRC (Area Under the Precision-Recall Curve) de 96.7%, une métrique particulièrement pertinente dans les cas de déséquilibre des classes, comme ici où les épisodes hypotensifs sont rares.

Notre approche présente également une précision positive (PPV) améliorée et une valeur prédictive négative (NPV) de 95.1%, réduisant significativement les faux positifs et les faux négatifs. Ces performances montrent l'efficacité du BiLSTM pour capturer les dynamiques complexes du signal de pléthysmographie (PPG) et permettre une prédiction fiable jusqu'à 5 minutes avant l'apparition de l'hypotension. Cela optimise ainsi le suivi hémodynamique non invasif et renforce le potentiel clinique de notre solution.

Métriques	Solam et al.[3]	Jeong et al.[2]	Jeong-Hyeon et al.[7]	Notre approche
AUC-ROC(%)	73.8	91.65	94.2	**97.3**
AUPRC(%)	–	–	–	**96.7**
PPV (%)	56.6	–	74.6	**89.9**
NPV (%)	77.5	–	92.18	**95.1**
Accuracy(%)	–	83.7	–	**92.3**

TABLE 1 – Comparaison des résultats entre l'état de l'art et notre approche pour une prédiction 5 minutes avant l'événement hypotensif, réalisée sur le benchmark public VitalDB

4 Conclusion et perspectives

Notre approche non invasive permet de prédire l'hypotension jusqu'à 5 minutes avant son apparition, évitant ainsi les risques des méthodes invasives. Grâce à un modèle d'apprentissage profond séquentiel (BILSTM), nous capturons les dépendances à long terme du signal de pléthysmographie et détectons des variations subtiles, souvent imperceptibles pour un observateur humain. Ce travail est le début d'un projet plus vaste, visant à développer un modèle plus précis (prédiction jusqu'à 10 minutes, différents seuils) et à l'intégrer dans un dispositif médical pour une validation clinique.

Références

[1] S. Bar, M. D. Moussa, and R. Descamps et al. Determinants of postoperative complications in high-risk noncardiac surgery patients optimized with hemodynamic treatment strategies. *J Clin Anesth*, 2024.

[2] Heejoon Jeong, Donghee Kim, Dong Won Kim, Seungho Baek, Hyung-Chul Lee, Yusung Kim, and Hyun Joo Ahn. Prediction of intraoperative hypotension using deep learning models based on non-invasive monitoring devices. *Journal of Clinical Monitoring and Computing*, pages 1–9, 2024.

[3] S. Lee, H. C. Lee, and Y. S. Chu et al. Deep learning models for the prediction of intraoperative hypotension. *Br J Anaesth*, 2021.

[4] F. Michard, M. S. Chew, and P. G. Guinot. Perioperative cardiac output monitoring : From yellow catheters to green algorithms. *ICU Management & Practice*, 2024.

[5] F. Michard and E. Futier. Predicting intraoperative hypotension : from hope to hype and back to reality. *Br J Anaesth*, 2023.

[6] F. Michard, E. Futier, and O. Desebbe et al. Pulse contour techniques for perioperative hemodynamic monitoring : a nationwide carbon footprint and cost estimation. *Anaesthesia Critical Care & Pain Medicine*, 2023.

[7] Jeong-Hyeon Moon, Garam Lee, Seung Mi Lee, Jiho Ryu, Dokyoon Kim, and Kyung-Ah Sohn. Frequency domain deep learning with non-invasive features for intraoperative hypotension prediction. *IEEE Journal of Biomedical and Health Informatics*, 28(10) :5718–5728, 2024.

[8] D. Nepogodiev, J. Martin, and B. Biccard et al. Global burden of postoperative death. *The Lancet*, 2019.

Génération de données de santé synthétiques : état de l'art en vue de la classification des mécanismes des traumatismes

Ariel Guerra-Adames[1,2,†], Océane Dorémus[1,†], Marta Avalos-Fernandez[2],
Vianney Jouhet[1,3], Emmanuel Lagarde[1], Cédric Gil-Jardiné[1,3]

[1] Équipe AHeaD, Université de Bordeaux, INSERM, BPH, U1219, F-33000 Bordeaux, France
[2] Équipe SISTM, Université de Bordeaux, INSERM, INRIA, BPH, U1219, F-33000 Bordeaux, France
[3] CHU de Bordeaux, INSERM, U1219, F-33000 Bordeaux, France

Résumé

Face aux impératifs de confidentialité et aux exigences de reproductibilité, la recherche sur l'IA en santé se tourne vers la conception de bases de données médicales synthétiques. Cet article présente un bref aperçu des méthodes d'IA pour générer des données tabulaires et textuelles synthétiques, en mettant l'accent sur leur application à la classification automatique des mécanismes des traumatismes.

Mots-clés

Données médicales synthétiques, GAN, VAE, DDPM, LLM

Abstract

Faced with the importance of confidentiality and reproducibility, research on healthcare AI is turning towards the conception of synthetic medical databases. This article presents a brief overview of AI methods for generating synthetic tabular and textual data, focusing their application to the automatic classification of trauma mechanisms.

Keywords

Synthetic medical data, GAN, VAE, DDPM, LLM

1 Introduction

La recherche en apprentissage automatique pour la santé fait face à un défi majeur de reproductibilité, concept qui renvoie à la capacité d'obtenir systématiquement les mêmes résultats que ceux rapportés initialement lorsqu'on dispose des données et du code d'origine. McDermott et al. [10] introduisent à cet égard la notion de « reproductibilité sans friction », définie comme la possibilité de reproduire aisément les résultats grâce à un partage transparent des données, à la disponibilité d'un code facilement ré-exécutable, ainsi qu'à une évaluation rigoureuse par des benchmarks standardisés. Pourtant, dans leur revue portant sur plus de 500 études, les auteurs constatent que de nombreuses recherches ne satisfont pas ces critères, notamment en raison de l'impossibilité de partager les données du fait de leur nature sensible.

Cette problématique est exacerbée par l'hétérogénéité des systèmes de collecte de données dans les établissements de santé, combinée à des biais introduits par des pratiques locales. Par exemple, les écarts dans la manière dont les dossiers médicaux sont structurés ou codés entre institutions peuvent entraîner des difficultés à reproduire les résultats obtenus dans une région ou un contexte spécifique [15]. Ce défi est particulièrement notable dans des tâches de classification appliquées à des affections médicales, où la sensibilité des données patient et la rareté de certains cas spécifiques peuvent limiter la constitution de jeux de données d'entraînement robustes et diversifiés.

Ces contraintes soulèvent une question fondamentale : La reproductibilité dite *sans friction* est-elle compatible avec l'apprentissage automatique en santé ? L'utilisation de données synthétiques pourrait permettre d'apporter une réponse à cette question. Ces données artificielles, imitant les propriétés et caractéristiques des données réelles, pourraient offrir une alternative pour surmonter les obstacles liés à la confidentialité et à l'accessibilité, tout en permettant le développement et l'évaluation d'algorithmes d'apprentissage reproductibles. Cependant, cette approche soulève à son tour deux interrogations principales auxquelles nous nous intéressons dans notre étude : *1) Peut-on générer des données de santé synthétiques de* ***qualité*** *qui respectent le principe de* ***privacité*** et *2) Comment peut-on* ***évaluer*** *ces deux aspects ?* Pour répondre à ces interrogations, nous envisageons de mettre en œuvre et d'évaluer des méthodes de génération de données synthétiques tabulaires et textuelles, dans le but d'entraîner des modèles performants pour la classification des mécanismes de traumatismes. En première étape, nous avons réalisé une revue de la littérature portant sur les méthodologies de génération de données synthétiques adaptées aux données médicales tabulaires et aux textes non structurés, ainsi que sur les approches d'évaluation de leur qualité et du respect de la confidentialité.

2 État de l'art

Nous avons consulté les bases bibliographiques Scopus, PubMed et Google Scholar pour des articles mentionnant la génération ou l'évaluation de données médicales synthétiques. Un total de 164 articles répondaient à nos critères. Nous présentons les principales méthodes de génération ainsi que d'évaluation.

†. Ces auteurs ont contribué équitablement à l'étude.

Méthodes de génération. Historiquement, en statistique, la génération de données simulées a été une étape clé pour évaluer et valider des modèles théoriques, en permettant de tester des hypothèses dans des conditions contrôlées. Cette approche repose sur des distributions de probabilité et des modèles paramétriques pour simuler des scénarios appropriés. En revanche, dans le domaine de l'apprentissage automatique, la simulation de données n'a pas été aussi courante. L'accent a plutôt été mis sur l'évaluation des algorithmes à partir de jeux de données réels, souvent bien documentés et normalisés, pour tester leur capacité à généraliser à des situations concrètes [2]. Récemment, la génération de données synthétiques a connu une évolution importante, passant de techniques purement statistiques à des approches d'apprentissage profond [3]. Parmi les approches statistiques, figurent l'échantillonnage aléatoire et la perturbation de données existantes, ou encore les arbres de décision [13] et les réseaux bayésiens [6]. Les autoencodeurs variationnels (VAE) [7] ont étendu l'approche probabiliste en combinant l'inférence bayésienne avec les réseaux de neurones profonds. Les réseaux antagonistes génératifs (GAN) [5] ont introduit, quant à eux, un principe d'opposition entre deux réseaux pour améliorer la qualité des données générées. Cependant, ces derniers sont connus pour leur instabilité [1]. Les modèles de diffusion, plus récents, fonctionnent en ajoutant progressivement du bruit aux données réelles, puis en apprenant à inverser ce processus [8]. Pour les données textuelles, les grands modèles de langage (LLM) offrent des perspectives intéressantes, bien que des approches plus complexes soient nécessaires pour les adapter correctement au domaine médical [9].

Évaluation des données générées. L'évaluation des données synthétiques demeure complexe en l'absence de standards établis ; toutefois, la littérature identifie plusieurs grandes catégories de métriques permettant d'en apprécier divers aspects. La qualité des données synthétiques s'articule principalement autour de deux axes : la fidélité et l'utilité [3]. La fidélité mesure à quel point les données générées ressemblent aux données originales : par des mesures telles que la distance de Wasserstein ou le test de Kolmogorov-Smirnov. L'utilité quant à elle évalue la pertinence des données synthétiques sur des applications spécifiques notamment en mesurant les performances sur des tâches concrètes, comme par exemple la classification. Au-delà de ces aspects, les métriques de privacité visent à garantir que les données synthétiques ne permettent pas de reconstruire ou d'identifier les individus présents dans le jeu de données original [16]. Enfin, la notion d'équité ou *fairness* gagne également en importance, cherchant à s'assurer que les données générées ne perpétuent pas ou n'amplifient pas les biais présents dans les données d'origine [14].

Équilibre entre privacité et utilité. L'équilibre entre confidentialité et utilité des données synthétiques constitue un défi majeur, leur renforcement mutuel étant souvent antagoniste [12] ; d'où la nécessité d'un encadrement rigoureux de leur génération et de leur usage afin d'en assurer la responsabilité et la sécurité.

3 Perspectives

Tenant compte des enjeux signalés, nous orientons à présent notre étude vers une application concrète : la classification des mécanismes de traumatismes physiques. Nous proposons une approche hybride séquentielle de génération de données synthétiques tabulaires et textuelles, en vue d'entraîner des modèles pour la classification des mécanismes de traumatisme des dossiers patients des urgences. L'objectif est de recréer fidèlement les distributions de nos données réelles tout en respectant la privacité, mais aussi à enrichir les jeux de données d'entraînement et d'évaluation, notamment pour les mécanismes traumatiques rares, afin d'améliorer la robustesse des futurs modèles de classification. Enfin, à partir de ces données tabulaires synthétiques, nous générerons ensuite des textes cliniques non structurés qui simuleront des anamnèses plausibles.
Cette méthodologie se déroulera en trois phases :

Phase 1 : Génération de données tabulaires. Nous entraînerons des modèles génératifs profonds intégrant des garanties de confidentialité différentielle [4], notamment des VAE et des modèles de diffusion. Ces modèles seront entraînés sur une base de données réelle issue du service d'urgences adultes du CHU de Bordeaux site Pellegrin, comprenant plus de 700.000 échantillons collectés entre 2013 et 2024. Chaque échantillon $x \in \mathbb{R}^d$ est constitué de $d = 25$ variables catégorielles et continues, incluant notamment les paramètres vitaux du patient (fréquence cardiaque, pression artérielle, etc), des informations contextuelles sur la visite aux urgences (heure d'arrivée, mode d'admission, etc), ainsi que des données démographiques (âge, sexe, etc). Formellement, nous cherchons à apprendre une distribution synthétique $p_\theta(x)$ approchant la distribution réelle inconnue $p_{\text{r}}(x)$, tout en respectant une contrainte stricte de confidentialité différentielle (Rényi differential privacy) [11]. Une fois entraîné, le modèle permettra d'échantillonner un nouveau jeu de données synthétiques : $X_{\text{synth}} = \{x_i^{\text{synth}}\}_{i=1}^{N}, \quad x_i^{\text{synth}} \sim p_\theta(x)$. Afin d'améliorer l'équilibre des classes, nous générerons volontairement un nombre plus important d'échantillons correspondant aux mécanismes traumatiques rares ou sous-représentés dans les données réelles initiales. Cette stratégie permettra d'obtenir un jeu de données synthétique équilibré facilitant l'apprentissage efficace des modèles prédictifs.

Phase 2 : Génération automatique de texte libre. Les données tabulaires synthétiques X_{synth} serviront d'entrée à un LLM dans une configuration *zero-shot* ou *few-shot*. Concrètement, pour chaque visite patient synthétique représentée par les variables tabulaires x_i^{synth}, le modèle générera une note clinique textuelle réaliste t_i à partir d'une consigne simple du type : « Étant donné les paramètres vitaux et informations suivants (x_i^{synth}), rédigez une anamnèse plausible ». L'ensemble final des données synthétiques ainsi obtenues sera donc composé à la fois des données tabulaires et textuelles : $D_{\text{synth}} = \{(x_i^{\text{synth}}, t_i)\}_{i=1}^{N}$. Cette phase est essentielle car les descriptions narratives des circonstances du traumatisme (anamnèses) contiennent des informations cruciales, souvent non capturées par les seules don-

nées tabulaires, pour une classification précise des mécanismes. L'utilisation de LLM pour générer ces textes à partir des données tabulaires synthétiques permet de créer des ensembles de données bimodales réalistes.

Phase 3 : Évaluation et Application à la classification des mécanismes traumatiques. Les données tabulaires synthétiques seront évaluées à l'aide des métriques présentées précédemment. De manière complémentaire, nous solliciterons également des professionnels de santé des urgences qui évalueront qualitativement ("naturel" ou "non naturel") un sous-échantillon aléatoire de (D_{synth}), composé conjointement des variables numériques et des notes cliniques associées. Cette phase finale permettra de justifier (ou pas) concrètement l'utilité des données synthétiques pour la tâche de classification envisagée. En entraînant des modèles de classification des mécanismes de traumatismes sur ces données, et en comparant leurs performances suivant le paradigme "*Train on Synthetic, Test on Real*" (TSTR) versus "*Train on Real, Test on Real*" (TRTR), nous évaluerons si l'approche proposée permet d'obtenir des modèles aussi performants, voire meilleurs, tout en surmontant les contraintes liées à l'utilisation de données réelles sensibles et parfois déséquilibrées.

Références

[1] Masoud Abedi, Lars Hempel, Sina Sadeghi, and Toralf Kirsten. GAN-based approaches for generating structured data in the medical domain. *Applied Sciences*, 12(14) :7075, 2022.

[2] Leo Breiman. Statistical modeling : The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16, 08 2001.

[3] Emmanuella Budu, Kobra Etminani, Amira Soliman, and Thorsteinn Rögnvaldsson. Evaluation of synthetic electronic health records : A systematic review and experimental assessment. *Neurocomputing*, page 128253, 2024.

[4] Cynthia Dwork. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer, 2006.

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11) :139–144, 2020.

[6] Dhamanpreet Kaur, Matthew Sobiesk, Shubham Patil, Jin Liu, Puran Bhagat, Amar Gupta, and Natasha Markuzon. Application of bayesian networks to generate synthetic health data. *Journal of the American Medical Informatics Association*, 28(4) :801–811, 2021.

[7] Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013.

[8] Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. TabDDPM : Modelling tabular data with diffusion models. In *International Conference on Machine Learning*, pages 17564–17579. PMLR, 2023.

[9] Gleb Kumichev, Pavel Blinov, Yulia Kuzkina, Vasily Goncharov, Galina Zubkova, Nikolai Zenovkin, Aleksei Goncharov, and Andrey Savchenko. Medsyn : LLM-based synthetic medical text generation framework. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 215–230. Springer, 2024.

[10] Matthew BA McDermott, Shirly Wang, Nikki Marinsek, Rajesh Ranganath, Luca Foschini, and Marzyeh Ghassemi. Reproducibility in machine learning for health research : Still a ways to go. *Science Translational Medicine*, 13(586) :eabb1655, 2021.

[11] Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pages 263–275. IEEE, 2017.

[12] Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. Synthetic data - A privacy mirage. *CoRR*, abs/2011.07018, 2020.

[13] Jaideep Vaidya, Basit Shafiq, Muazzam Asani, Nabil Adam, Xiaoqian Jiang, and Lucila Ohno-Machado. A scalable privacy-preserving data generation methodology for exploratory analysis. In *AMIA Annual Symposium Proceedings*, volume 2017, page 1695, 2018.

[14] Vibeke Binz Vallevik, Aleksandar Babic, Serena Elizabeth Marshall, Elvatun Severin, Helga MB Brøgger, Sharmini Alagaratnam, Bjørn Edwin, Narasimha Raghavan Veeraragavan, Anne Kjersti Befring, and Jan F Nygård. Can i trust my fake data–a comprehensive quality assessment framework for synthetic tabular data in healthcare. *International Journal of Medical Informatics*, page 105413, 2024.

[15] Jenna Wiens, John Guttag, and Eric Horvitz. A study in transfer learning : leveraging data from multiple hospitals to enhance hospital-specific predictions. *Journal of the American Medical Informatics Association*, 21(4) :699–706, 2014.

[16] Chao Yan, Yao Yan, Zhiyu Wan, Ziqi Zhang, Larsson Omberg, Justin Guinney, Sean D Mooney, and Bradley A Malin. A multifaceted benchmarking of synthetic electronic health record generation models. *Nature communications*, 13(1) :7609, 2022.

Graphe de connaissances pour la modélisation des projets de recherche équestre

Noama Adra[1], Lamine Bougueroua[1], Jean-Charles Huet[1], Faten Chakchouk[1]

[1] Université Paris-Panthéon-Assas, Efrei Research Lab, Paris, France

noama.adra@externe.efrei.fr, {lamine.bougueroua, jean-charles.huet,faten.chakchouk}@efrei.fr

Résumé

Nous avons développé un graphe de connaissances (KG) dédié au projet de recherche équestre [1]*. En effet, avec la massification des données, leur intégration et leur analyse deviennent un enjeu majeur, notamment en raison de leur diversité (capteurs, dossiers médicaux, vidéos, documents administratifs). Pour structurer ces données, nous utilisons ce graphe de connaissance. Nous avons interconnecté 96 entités clés comme chevaux, événements, cavaliers et compétitions. Ce KG facilite l'analyse avancée, les requêtes intelligentes et l'aide à la décision.*

Mots-clés

Graphe de connaissances, fusion de donnés, domaine équestre, outil d'aide à la décision.

Abstract

We developed a knowledge graph (KG) tailored to research equastrian projects.In fact, with the rise of big data, integrating and analyzing diverse information sources—such as sensors, medical records, videos, and administrative documents—has become a major challenge. To structure this information,we used the knowledge graph. We interconnected 96 key entities, including horses, events, riders, and competitions. Our KG enables advanced analysis, intelligent queries, and decision support.

Keywords

Knowledge Graph, Knowledge Fusion, Equestrian Sports, Decision-Making.

1 Introduction

L'intégration de sources de données diverses et complexes est un défi important dans de nombreux domaines, y compris le domaine équestre. À l'heure où les données sont au cœur de l'actualité, les organisations dépendent de plus en plus de la capacité à consolider et à analyser des ensembles de données disparates afin d'en tirer des informations exploitables. Cependant, la nature hétérogène des données, provenant de sources aussi variées que les relevés de capteurs, les dossiers médicaux, les séquences vidéo, les systèmes de surveillance de l'environnement et la documentation administrative, pose d'importants obstacles techniques et conceptuels. Il s'agit notamment de gérer les incohérences dans les formats de données, de gérer différents niveaux de granularité et de précision, et de garantir l'interopérabilité entre les systèmes. En outre, la nécessité d'harmoniser les données structurées (par exemple, les bases de données tabulaires) avec les formats non structurés (par exemple, le texte ou le multimédia) ajoute une nouvelle couche de complexité.

Dans le domaine équestre, où les données vont des paramètres physiologiques des chevaux et des cavaliers aux conditions environnementales et aux résultats des compétitions, la valeur potentielle de l'intégration est immense. La mise en relation de ces sources de données peut permettre le développement d'outils avancés pour le suivi des performances, la gestion de la santé, la prise de décision et la planification stratégique [4]. Cependant, cette intégration nécessite un processus innovant capable de combler ces silos de données et de transformer les informations brutes en connaissances significatives et exploitables.

Dans ce contexte, l'Institut Français du Cheval et de l'Équitation (IFCE) [2], créé en 2010, est un organisme public placé sous la double tutelle du ministère des Sports, de la Jeunesse et de la Ville et du ministère de l'Agriculture. Chargé de superviser le secteur équestre tant en France qu'à l'international, IFCE joue un rôle essentiel dans la promotion et la diffusion des connaissances du monde des sports équestres.

Actuellement, l'IFCE gère de larges volumes de données, mais l'un de ses principaux défis consiste à les structurer et à les exploiter efficacement.

Pour relever ce défi, cette étude vise à développer un graphe de connaissances (KG) pour la modélisation de projet de recherche équestre, en utilisant aussi des données d'autres projets pour généraliser au domaine équestre. Un KG offre une approche structurée pour organiser, relier et analyser des données hétérogènes, permettant ainsi une intégration et des capacités analytiques avancées [9].

La principale contribution de cette recherche est le développement d'un KG pour le domaine équestre. Ce KG constitue une base pour une meilleure gestion des données, une analyse approfondie et diverses applications dans le domaine des sports équestres.

1. https://noamaa.github.io/EquestrianDoamainKnowledgeGraph/

2. https://www.ifce.fr/, accessible au 07/05/2025

La suite de cet article est organisée comme suit : la section 2 présente un état de l'art des travaux connexes ; la section 3 décrit le modèle proposé ainsi que son processus de construction ; la section 4 expose les résultats obtenus. Enfin, la section 5 conclut en évoquant les perspectives futures.

2 Etat de l'art

Dans cette section, nous explorons les KGs , un outil puissant pour structurer et interconnecter des données complexes et spécifiques à un domaine. Nous mettrons en avant des exemples d'application des KGs dans le domaine du sport, illustrant leur potentiel à intégrer des ensembles de données variés.

2.1 Définition

Un KG [11] [8], également appelé réseau sémantique, est une manière structurée de représenter des entités du monde réel, telles que des objets, des événements, des situations ou des concepts, ainsi que les relations qui les lient [2] [5]. Ces informations sont généralement stockées dans une base de données orientée graphe et visualisées sous la forme d'une structure en graphe.
Un KG peut être défini comme une structure mathématique permettant de modéliser les relations binaires entre des entités. Il est généralement représenté comme suit :

$$G = (V, E)$$

où :

- V : un ensemble de noeuds qui répresente les entités ou les objets.
- E : L'ensemble des arêtes représente les relations ou interactions entre les nœuds du graphe.

2.2 Graphe de connaissance et ontologie

Une ontologie définit la structure générale et les règles d'un domaine [7]. Son but est de fournir un cadre conceptuel permettant de comprendre et de catégoriser les données. Elle traite des concepts plus larges et des relations générales qui existent au sein d'un domaine. Elle joue également un rôle essentiel en assurant une compréhension cohérente et une communication uniforme des informations. Ainsi, l'ontologie sert de base conceptuelle pour organiser la connaissance.
L'ontologie BBC Sport [1] a été conçue pour structurer les données liées aux événements sportifs (équipes, matchs, compétitions, lieux, etc.) en s'appuyant sur les principes du Web sémantique. Elle définit les entités et relations clés du domaine, facilitant l'organisation, la recherche et la publication de contenus. En étant alignée sur des standards ouverts, elle permet aussi de relier les données sportives de la BBC à d'autres sources sur le Web, posant ainsi les bases d'un KG dédié au sport.
La relation entre une ontologie et un KG est centrale dans la modélisation sémantique des données. L'ontologie joue le rôle de structure sémantique sous-jacente : elle définit les concepts du domaine (classes), les relations entre ces concepts (propriétés), ainsi que les contraintes logiques qui régissent leur usage. Cette structure constitue le schéma du graphe. Les ontologies définissent les types d'entités et les relations possibles dans un domaine, tandis que les graphes de connaissances appliquent ces définitions à des données concrètes pour représenter des faits spécifiques. En d'autres termes, une ontologie fournit le cadre conceptuel, et un graphe de connaissances est l'instanciation de ce cadre avec des données réelles[8].

2.3 Graphe de connaissances et sport

Les KG sont de plus en plus explorés dans divers domaines, y compris le sport, afin de relever des défis liés à l'intégration des données, aux recommandations personnalisées et aux requêtes intelligentes. À notre connaissance, la majorité des KGs existants dans le domaine du sport se concentrent principalement sur des applications liées aux systèmes de recommandation, en particulier dans le contexte du tourisme sportif [3].
Par exemple, Sun et al. [12] ont construit un KG pour interroger en masse des activités culturelles et sportives à partir de données multi-sources. En combinant ontologie et taxonomie et en utilisant un modèle basé sur BERT pour l'extraction d'entités, leur étude a permis d'améliorer la correspondance entre utilisateurs et activités grâce à des requêtes personnalisées, avec une validation effectuée sur des données réelles provenant de Pékin.
De même, Xu et al. [14] ont introduit un KG dédié au sport, intégrant des connaissances spécifiques au domaine sportif chinois et utilisant Neo4j comme base de données graphe. Leur travail met en évidence la manière dont les KGs peuvent améliorer le stockage des informations et permettre des requêtes intelligentes efficaces, contribuant ainsi à l'avancement des applications des KGs dans le sport.
Un domaine de recherche particulièrement prometteur concerne l'utilisation de la technologie des KGs pour améliorer la précision des outils de mesure de la condition physique [9]. En surmontant les limites des équipements traditionnels, les KGs permettent de cartographier les relations entre diverses variables de condition physique, offrant ainsi une vision complète et interconnectée de l'état physique. Ces avancées fournissent aux utilisateurs des recommandations personnalisées et exploitables, leur permettant d'améliorer leur condition physique sur la base d'évaluations précises et fondées sur les données.

3 Modèle proposé

Nous construisons un KG afin de représenter et d'organiser de manière structurée les connaissances relatives aux chevaux et à l'IFCE, facilitant ainsi l'intégration efficace des données et la découverte de relations. Ce graphe permet d'intégrer des données hétérogènes et de modéliser les liens entre différentes sources d'information.
Pour construire ce KG, nous avons utilisé Protégé [3], un éditeur d'ontologies open-source et un système de gestion des connaissances développé par l'Université de Stanford.

3. https://protege.stanford.edu/, accessible au 07/05/25

3.1 Données utilisées

Le KG a été construit en intégrant des données issues de plusieurs projets de recherche et de jeux de données open-source, permettant ainsi une représentation complète et interconnectée du domaine. La collecte initiale des données a impliqué l'agrégation d'informations structurées et non structurées provenant de diverses sources, telles que des bases de données publiques [4] (un benchmark pour l'estimation de pose comprenant 30 chevaux Pur-Sang, avec 22 parties du corps annotées sur 8114 images), des publications scientifiques [10] et les résultats de projets spécialisés. Le tableau 1 donne deux exemples des projets financés par le conseil scientifique de l'IFCE.

Chaque ensemble de données a été traité afin d'extraire les entités clés, telles que les objets, les événements et les relations pertinentes pour le domaine. Ces informations ont ensuite été harmonisées et projetés à un schéma commun, garantissant ainsi la cohérence et l'interopérabilité entre les différentes sources de données. Les données open-source, y compris les jeux de données accessibles publiquement provenant de sources reconnues, ont joué un rôle essentiel dans l'enrichissement du KG en apportant un contexte supplémentaire, tandis que les données issues de projets spécifiques ont permis d'intégrer des connaissances propres au domaine.

TABLE 1 – Exemples de données issues de projets scientifiques de l'IFCE

	Projet SoGen [5]	**Projet Able [10]**
Objectif	Caractériser les chevaux de saut d'obstacles à l'aide d'outils génomiques et morphométriques 3D.	Évaluer la locomotion et l'état physiologique des chevaux de dressage de haut niveau.
Taille de l'échantillon	2 000 jeunes chevaux.	10 chevaux de dressage de haut niveau en bonne santé.
Capteurs	Morphométrie 3D : 4 caméras synchronisées mesurant 28 points anatomiques.	9 IMUs, fréquence cardiaque (moniteur Polar®) et lactate (Akray® Lactate Pro 2).
Protocole	Filmage en marchant ; captures d'images pour des mesures précises.	Locomotion : Deux passages au trot/galop avant et après une séance personnalisée.
Les entités qu'on peut créer	Paramètres morphométriques (longueurs, angles, posture).	Fréquence cardiaque et niveaux de lactate , capteurs inertielles

Pour alimenter le KG, diverses sources de données ont été mobilisées, caractérisées par des formats variés et des contenus hétérogènes. Cette diversité a requis une procédure d'intégration adaptée à chaque type de support. Le Tableau 2 présente une synthèse des principaux formats traités, des sources correspondantes, des types d'informations extraites (manuellement), et d'exemples typiques issus des données collectées.

4. https://huggingface.co/datasets/mwmathis/Horse-30, accessible au 07/05/25
5. https://www.ifce.fr/ifce/sogen/-1200-chevaux-mesures/, accessible au 07/05/25

TABLE 2 – Extrait des divers données et des informations extraites

Format	**Source**	**Contenu**	**Exemple d'information extraites**
CSV	suivi santé cheval	fréquence cardiaque, age	cheval, fréquence cardiaque
JSON	système de capteurs	enregistrements IMU	CapteurIMUs, Vitesse , cheval
MP4	vidéos d'entraînement	métadonnées d'expérience (ID cheval, timestamps, étiquettes)	camera, ID cheval

3.2 Processus du construction du graphe de connaissances

Dans la construction d'un KG pour un domaine scientifique, les entités représentent les concepts clés, les objets ou les acteurs du domaine, tandis que les arêtes définissent les relations entre ces entités. Ces entités, telles que les événements, les objets et les acteurs, sont enrichies par des attributs tels que des données temporelles, des localisations et des détails contextuels, offrant ainsi une compréhension plus fine du domaine. Cependant, ce processus est itératif et s'affine continuellement avec l'ajout de nouveaux attributs afin d'améliorer la précision et l'utilité du KG.
Tamasauskaite et al. [13] ont réalisé une revue systématique pour définir le développement des graphes de connaissances, en proposant six étapes essentielles pour la construction du graphe (Figure 1) :

1. Identifier les données (Identify data)
2. Construire l'ontologie du KG (Construct the KG Ontology)
3. Extraire les connaissances (Extract Knowlegde)
4. Traiter les connaissances (Process Knowledge)
5. Construire le KG (Contruct the KG)
6. Maintenir les connaissances (Maintain the KG)

Dans notre travail actuel, nous avons complété les quatre premières étapes du processus de construction du KG. Étant donné que notre domaine ne disposait pas d'une ontologie préétablie à notre conaissance, nous avons accordé la priorité à l'établissement des éléments fondamentaux, notamment :

1. Collecte des données : Les sources et les détails des données sont décrits dans la section 3.1.
2. Construire l'ontologie du KG : Le domaine ne dispose pas d'une ontologie préétablie.
3. Extraction des connaissances : Cette étape a consisté à poser une série de questions ciblées pour identifier les éléments clés et comprendre la structure des données.
   (a) Identification des entités : Définition des principales entités du domaine.
   (b) Identification des relations : Détermination des relations entre ces entités.
   (c) Identification des attributs : Détermination des attributs des entités.

4. Traiter les connaissance : L'intégration de la connaissance et la construction d'une ontologie. Les premiers résultats sont présentés dans la section 4.

Il convient de souligner que ce travail a été réalisé manuellement. L'identification des entités et des relations s'est appuyée sur une analyse des sources, visant à repérer les concepts récurrents, à les structurer, puis à modéliser leurs interconnexions.

Par exemple, la figure 2 montre un exemple de processus de l'intégration des données issues des unités de mesure inertielles (IMU) suppose la modélisation de ces capteurs comme des nœuds dotés d'attributs (nom, identifiant), reliés à d'autres entités par des relations telles que IsAttachedTo (posé sur un cheval) ou IsUsedFor (utilisation dans une expérience). La première étape consiste à extraire l'information afin de définir les entités clés et leurs relations, comme détaillé dans la figure.

Ensuite, la modélisation du KG a été réalisée à l'aide de l'outil Protégé, un éditeur open source compatible avec le langage OWL (Web Ontology Language). Cet environnement permet de formaliser les entités (ou classes), de leur associer des propriétés descriptives (data properties), ainsi que des relations avec d'autres entités (object properties), conformément aux standards du Web sémantique.

La figure 3 présente une capture d'écran de l'interface de Protégé, illustrant la définition de l'entité Horse. Celle-ci est modélisée comme une classe OWL, associée à divers attributs (l'identifiant, age, race) et reliée à d'autres entités par des relations sémantiques associées tel que "isAttachedTo" reliant un cheval à un capteur. Et aussi, les propriétés de l'entité décrivant un cheval. Cet extrait illustre la manière dont les concepts métiers du domaine équestre ont été traduits en classes et propriétés RDF/OWL pour structurer les connaissances de manière exploitable et interopérable.

Cette approche structurée permet de capturer à la fois le contexte physique de la collecte des données et leur finalité scientifique, facilitant ainsi les requêtes avancées et le raisonnement sémantique au sein du KG. La même méthodologie est appliquée de manière itérative à d'autres entités du domaine, aboutissant à une représentation cohérente et enrichie des connaissances.

Dans cette première phase de développement, aucune ontologie existante n'a été directement réutilisée. Ce choix résulte de la volonté de concevoir une ontologie spécifiquement adaptée au domaine équestre, reflétant fidèlement les besoins métiers de l'IFCE et la structure hétérogène des données disponibles.

La construction d'une ontologie sur-mesure a permis de capturer les particularités des projets de recherche équestre.

4 Premier résultat

Le KG a été publiée en ligne [6] et documentée automatiquement à l'aide de l'outil Widoco [6]. Cette documentation comprend une visualisation des classes, des propriétés, ainsi qu'une description textuelle conforme aux standards du web sémantique.

Le KG contient actuellement 96 classes, le tableau 3 donne plus de détails sur la répartition des classes. Le tableau 4 présente les principales entités du graphe, parmi lesquelles "Cheval", "Entraînement", "Discipline équestre", "Humain", qui représentent les composantes essentielles du domaine. La figure 4 représente le KG avec quelque entités tel que "Horse" et "Sporting events" relié via la relation : "TrainsFor".

TABLE 3 – Caractéristiques du graphe

Nombre de classes	96
Nombre de propriétés utilisé	9
Nombre d'attributs	26
Nombre d'individus	3

TABLE 4 – Les entités principales représentées dans le graphe

Entité	**Description**
Horse	Représente le cheval
Experimental Devices	Différents équipements utilisés dans le projet
Training	Phases de l'entraînement
Event	Disciplines dans lesquelles le cheval peut concourir
Human	Représente les humains
Project	Représente les projets du domaine équestre

Les relations (ou arêtes) définissent la manière dont les entités interagissent ou sont liées entre elles dans le domaine. Dans Protégé, ces relations sont définies comme Object Properties. Voici quelques relations courantes :

- Associeted (Cheval - Cavalier) : Correspond aux différents couples formés entre un cheval et son cavalier.
- CompeteIn (Cheval - Compétitions) : Désigne les principaux événements auxquels le cheval participe.
- DependsOn (Cheval - Entraînement) : Permet d'analyser le niveau d'entraînement requis.
- hasParticipated (Cheval - Projet) : Associe un cheval aux projets dans lesquels il a été impliqué pour récupérer les données spécifiques par projet.
- isUsedFor (Capteurs - Projet) : Associe les capteurs aux projets dans lesquels ils sont utilisés.

Ces relations permettent de structurer le KG en définissant explicitement les interactions entre les entités, facilitant ainsi l'exploration et l'analyse des données du domaine équestre.

Nous avons réalisé une série de requêtes en utilisant SPARQL. Le Tableau 5 présente les différentes requêtes

6. https://noamaa.github.io/EquestrianDoamainKnowledgeGraph/

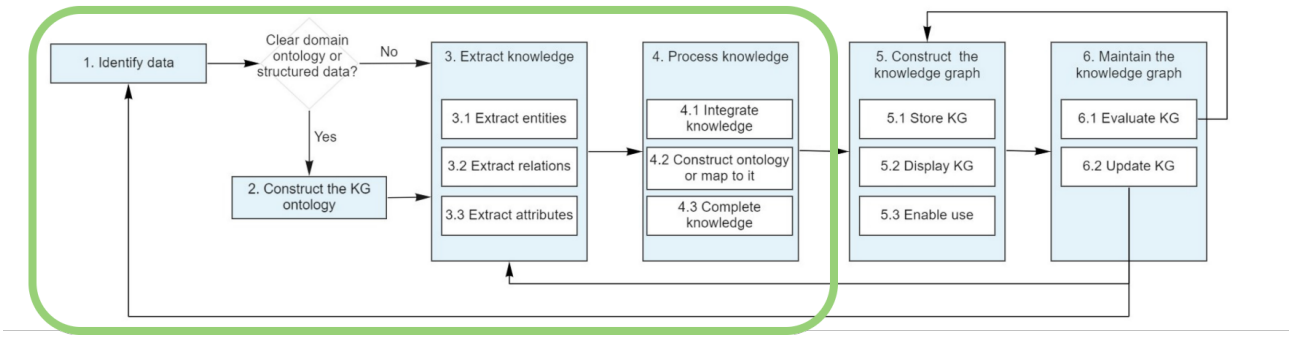


FIGURE 1 – Les étapes de construction du graphe [13]

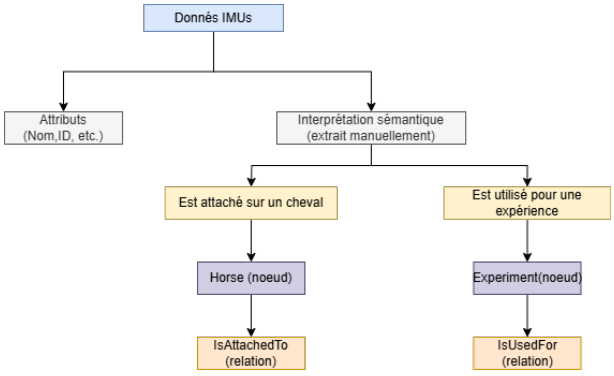


FIGURE 2 – Identifier les nœuds et leur relations

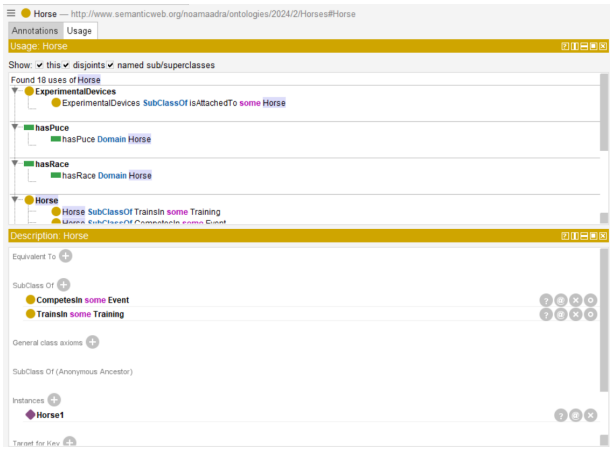


FIGURE 3 – Extrait de notre projet issu de Protégé qui affiche les informations sur l'entité cheval

effectuées ainsi que leurs résultats correspondants. Notre graphe est enrichi avec des étiquettes RDF (labels), ce qui améliore la clarté et l'interprétabilité des requêtes. Ces étiquettes facilitent une interrogation plus intuitive et renforcent la précision des résultats, rendant le graphe plus accessible pour l'exploration et l'analyse des données.

TABLE 5 – Exemples de requêtes

Questions	**Requêtes**	**Résultats**
Quels sont les dispositifs expérimentaux utilisés ?	SELECT ?device WHERE { ?device rdfs :label "Sensor" . }	Camera, IMU Sensor
Quels sont les différents projets ?	SELECT ?device WHERE { ?device rdfs :label "Project" . }	Toutes les entités des projets

5 Conclusion

Cette étude a présenté le développement d'un graphe de connaissances (KG) conçu pour représenter le domaine équestre et les données de l'Institut Français du Cheval et de l'Équitation (IFCE), offrant un moyen structuré et formalisé de capturer et d'analyser les connaissances liées à l'équitation. En tant que premier graphe de connaissances de ce type pour ce domaine à notre échelle de connaissance, il comble une lacune importante en fournissant une base pour mieux comprendre et améliorer les performances du cheval et cavalier grâce à une analyse avancée des données.

Malgré ces contributions, ce travail représente une première phase de développement. Le KG actuel reste limité en termes de périmètre et nécessite un enrichissement avec des données supplémentaires et des informations plus détaillées. Des tests approfondis sont également indispensables pour évaluer son efficacité dans diverses applications et affiner ses fonctionnalités.

Les travaux futurs se concentreront sur l'expansion du KG en intégrant des ensembles de données plus complets, en le déployant sur une plateforme centralisée pour une accessibilité accrue et en testant son utilité dans des scénarios réels. Ces évolutions garantiront sa scalabilité, son adaptabilité et son alignement avec les besoins en constante évolution de la communauté équestre.

À long terme, ce graphe de connaissances peut devenir un outil puissant d'aide à la prise de décision, en facilitant l'intégration, l'analyse et l'exploitation de données complexes. En structurant les connaissances, il permet d'optimiser la gestion des informations et d'orienter les choix stratégiques. Son application dans différents sports per-

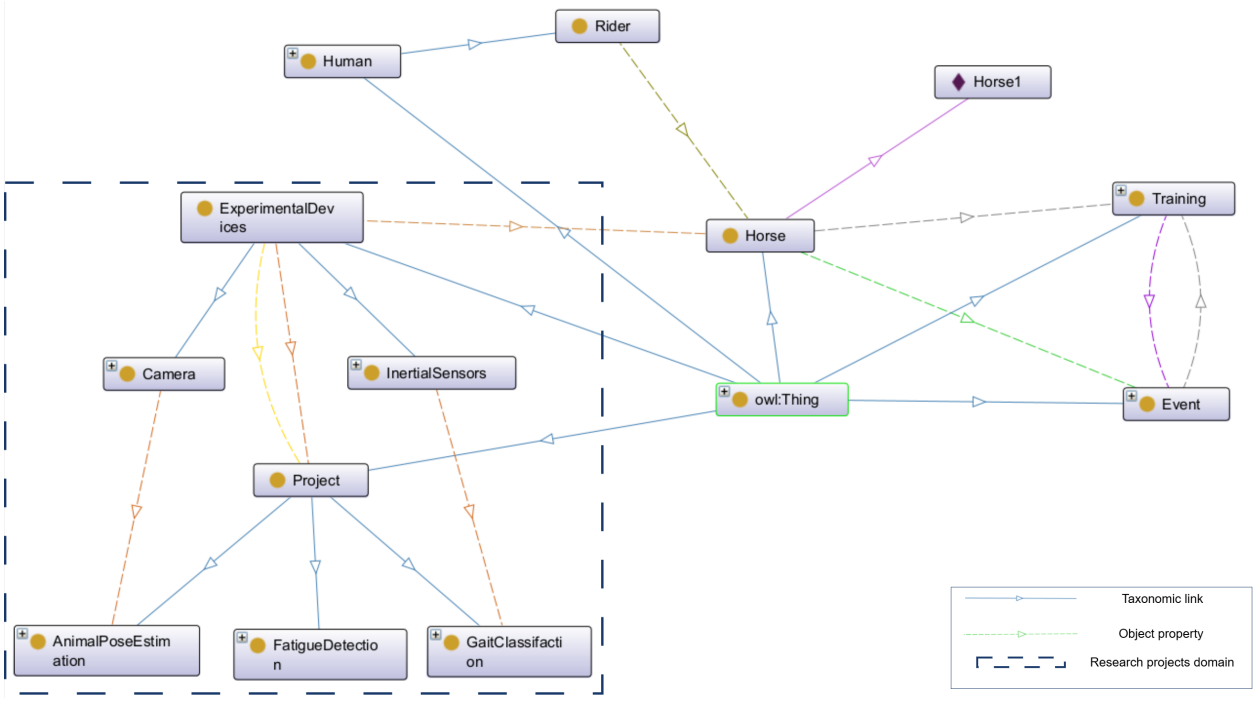


FIGURE 4 – Le KG pour les projets de recherche au sein du domaine équestre

mettra l'amélioration des performances et l'adoption d'approches basées sur les données. Il contribue ainsi à transformer un secteur en un écosystème intelligent et interconnecté, où la prise de décision est renforcée par des analyses précises et contextualisées.

Références

[1] BBC. Bbc sport ontology, 2024. Accessed : 2024-04-28.

[2] Antoine Bordes, Jason Weston, Ronan Collobert, and Yoshua Bengio. Learning Structured Embeddings of Knowledge Bases. *Proceedings of the AAAI Conference on Artificial Intelligence*, 25(1) :301–306, August 2011.

[3] Maoteng Cheng, Lu Zhang, and Baojian Zhang. A literature review of sports tourism research : knowledge graph analysis based on CiteSpace visualization. *Journal of Sport & Tourism*, 27(4) :293–313, 2023. Publisher : Taylor & Francis.

[4] Pierre Colle. What AI can do for horse-racing ?, July 2022.

[5] Michael Färber, Frederic Bartscherer, Carsten Menne, and Achim Rettinger. Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. *Semantic Web*, 9(1) :77–129, November 2017.

[6] Daniel Garijo. Widoco : a wizard for documenting ontologies. In *International Semantic Web Conference*, pages 94–102. Springer, Cham, 2017.

[7] Thomas R. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2) :199–220, 1993.

[8] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia D'amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. Knowledge graphs. *ACM Computing Surveys*, 54(4) :1–37, July 2021.

[9] Fanrong Kong, Guangpeng Li, and Hengchao Zhou. Design and Implementation of Sports Fitness Testing Equipment Based on Knowledge Graph Structure Optimization Algorithm. In *2023 International Conference on Data Science and Network Security (ICDSNS)*, pages 1–5. IEEE, 2023.

[10] Benoît Pasquiet, Sophie Biau, Quentin Trébot, Jean-François Debril, François Durand, and Laetitia Fradet. Detection of Horse Locomotion Modifications Due to Training with Inertial Measurement Units : A Proof-of-Concept. *Sensors*, 22(13) :4981, July 2022.

[11] Ciyuan Peng, Feng Xia, Mehdi Naseriparsa, and Francesco Osborne. Knowledge Graphs : Opportunities and Challenges. *Artificial Intelligence Review*, 56(11) :13071–13102, November 2023.

[12] Huanliang Sun, Yumeng Ma, and Junling Liu. Construction of Mass Cultural and Sports Activity Knowledge Graph based on the Fusion of Multi-Source Data. In *2022 International Conference on*

Knowledge Engineering and Communication Systems (ICKES), pages 1–6. IEEE, 2022.

[13] Gytė Tamašauskaitė and Paul Groth. Defining a Knowledge Graph Development Process Through a Systematic Review. *ACM Transactions on Software Engineering and Methodology*, 32(1) :1–40, January 2023.

[14] Zhangbo Xu, Tao Xu, and Fan Zhang. Construction of chinese sports knowledge graph based on neo4j. In *2020 IEEE 2nd International Conference on Civil Aviation Safety and Information Technology (ICCA-SIT*, pages 561–564. IEEE, 2020.

Session 4

Grands modèles de langage et exécution d'algorithmes. Application à une fonction arithmétique.

Farah Ben Slama, Frédéric Armetta

Université Claude Bernard Lyon 1, CNRS,
Ecole Centrale de Lyon, INSA Lyon,
Université Lumière Lyon 2, LIRIS, UMR5205,
69622 Villeurbanne, France

Mai 2025

Résumé

Les grands modèles de langage (LLM) ont récemment développé de nouvelles fonctionnalités avancées. Leur efficacité repose sur un apprentissage statistique et sur des capacités de généralisation. Ils connaissent cependant des limitations à s'imprégner des données qu'ils manipulent, et peinent par exemple à exécuter des algorithmes de manière autonome. Nous étudions dans cet article la possibilité d'étendre les capacités de ces modèles à l'exécution d'algorithmes à travers un entraînement spécifique supervisé dédié à la décomposition du raisonnement. Nous introduisons un modèle d'apprentissage appelé LLM-DAL (Large Language Model - Decompositional Algorithmic Learning), à travers lequel nous montrons que la capacité des LLM à réaliser des inférences complexes de nature algorithmique et à généraliser peut être améliorée significativement lorsque la méthode d'entraînement est adaptée et accompagne le modèle dans son apprentissage.

Mots-clés

Apprentissage algorithmique en langage naturel, Apprentissage supervisé par décomposition, Grand modèle de langage, Fine-tuning

Abstract

Large Language Models (LLMs) have recently developed new advanced functionalities. Their effectiveness relies on statistical learning and generalization capabilities. However, they face limitations in internalizing the data they process and struggle, for instance, to autonomously execute algorithms. In this paper, we investigate the possibility of extending these models' capabilities to algorithm execution through specialized supervised training focused on reasoning decomposition. We introduce a training model called LLM-DAL (Large Language Model - Decompositional Algorithmic Learning), through which we demonstrate that LLMs' ability to perform complex algorithmic inferences and generalize can be significantly improved when the training method is properly designed to guide the model in its learning process.

Keywords

Algorithmic learning in natural language, Supervised learning by decomposition, Large language model, Fine-tuning

1 Introduction

L'apprentissage profond a révolutionné le domaine de l'intelligence artificielle (IA), ouvrant de nouvelles perspectives dans des domaines variés, tels que la reconnaissance d'images, la reconnaissance vocale et le traitement du langage naturel. Grâce à sa capacité à modéliser des structures de données complexes, il s'impose comme un outil essentiel pour résoudre des problèmes de plus en plus diversifiés. Cependant, au-delà de ces applications bien établies, un défi reste encore largement inexploré : l'utilisation des réseaux de neurones pour l'apprentissage et pour l'exécution d'algorithmes.

Les réseaux de neurones sont reconnus pour leurs capacités d'apprentissage et de généralisation. Pourtant, leur application au domaine de l'apprentissage algorithmique présente des capacités spécifiques. En effet, l'objectif ne se limite pas à classer des données ou prédire des tendances, mais à permettre au réseau de produire, à sa sortie, le résultat d'un algorithme en fonction de paramètres fournis en entrée. Cette approche pourrait permettre d'automatiser l'exécution d'algorithmes complexes sans recourir à une implémentation explicite, ouvrant ainsi la voie à des modèles capables de raisonner de manière plus systématique et explicable mais aussi de généraliser au delà de leurs données d'entraînement.

Le problème soulevé par Searle, illustré par l'expérience de la chambre chinoise [8] où il imagine une personne ne comprenant pas le chinois mais suivant des instructions pour manipuler des symboles chinois de façon cohérente, met en évidence une limite fondamentale des systèmes d'intelligence artificielle : bien que l'échange semble signifiant pour un observateur extérieur, la personne à l'intérieur ne comprend rien au contenu. Cette expérience illustre le fait que simuler la compréhension ne correspond en réalité pas à une appropriation des concepts manipulés. De la même manière, les systèmes d'IA manipulent des symboles et ap-

pliquent des règles sans véritable compréhension ou appropriation des connaissances. Dans le cas des grands modèles de langage (LLM), cette limitation se traduit par un manque d'ancrage des connaissances, qui restent exogènes au système et ne résultent pas d'une construction subjective itérative basée sur l'expérience .

Ainsi, dans cet article, bien que nous ne proposions pas de dépasser ce problème fondamental d'ancrage des connaissances, nous proposons une méthode d'apprentissage permettant d'aider le système à s'approprier plus en profondeur la résolution de problèmes de nature algorithmique.

Un tel apprentissage impose au réseau de neurones de capturer de très longues dépendances au cours de son raisonnement algorithmique, de manipuler des variables dans ses représentations internes et de leur appliquer des transformations complexes pour produire le résultat attendu.

Une tentative pionnière pour relever ce défi a été introduite avec la Machine de Turing neuronale (*Neural Turing Machine*) [5], une architecture novatrice conçue pour simuler la logique algorithmique à travers des réseaux neuronaux. Bien que prometteuse, cette solution a mis en évidence des limites significatives, notamment des problèmes de convergence et une sensibilité extrême au paramétrage [4].

Ainsi, même si les LLM génèrent des réponses linguistiquement cohérentes, ils sont parfois qualifiés de "perroquets stochastiques" , c'est-à-dire qu'ils restituent des informations apprises sans comprendre leur logique, notamment lorsqu'elles sont algorithmiques, ce qui limite leurs capacités. Cette "mémorisation" des séquences peut conduire à des résultats satisfaisants dans des contextes préalablement rencontrés, mais ne garantit pas que ces modèles soient capables de résoudre des problèmes complexes ou de comprendre en profondeur les relations sous-jacentes entre les concepts.

C'est par exemple le cas dans l'article publié par *Razeghi et al.* [7], où il est démontré que la performance de résolution d'opérations arithmétiques est corrélée à la fréquence d'apparition des opérandes dans le corpus d'entraînement. Cela met en évidence que le réseau de neurones ne maîtrise pas la logique sous-jacente des opérations, mais se limite à reproduire des séquences apprises.

Cela soulève la question de savoir si ces modèles, fondés sur des réseaux de neurones, peuvent dépasser cette limitation de "perroquet stochastique" et accomplir des tâches nécessitant un raisonnement algorithmique plus complexe et une véritable compréhension logique.

Nous présentons dans un premier temps, un état de l'art sur les modèles de langage dans la section 2, puis nous décrivons notre modèle proposé dans la section 3. Nous détaillons ensuite le contexte expérimental et les résultats obtenus dans la section 4. Enfin, nous finissons par une conclusion du travail réalisé et des perspectives en section 5.

2 État de l'art

Les grands modèles de langage continuent de révolutionner le domaine de l'intelligence artificielle, en particulier pour les tâches complexes nécessitant du raisonnement mathématique, de la programmation ou encore des applications scientifiques. Ces dernières années, plusieurs modèles bien développés ont émergé, chacun apportant des innovations significatives en termes d'architecture ou de performance. L'engouement pour la diffusion des modèles de langage est en grande partie porté par les grands géants de l'IA. Ces entreprises disposent des ressources financières, des infrastructures de calcul massif (GPU, TPU), et des équipes de recherche nécessaires pour entraîner et affiner ces modèles sur d'énormes quantités de données.

D'un point de vue technique, les LLM s'appuient sur une architecture de réseaux de neurones profonds, héritant ainsi à la fois des capacités et des limitations propres à ces derniers. Leur fonctionnement consiste à prédire mot par mot par rapport au contexte formé par les mots précédents, et correspond à la mémorisation de *patterns* appris et une généralisation statistique, sans véritable compréhension ou raisonnement et donc leur capacité à "raisonner" reste limitée à la simple répétition probabiliste de séquences de mots, d'où l'hypothèse du "perroquet stochastique".

Ainsi, dans cet état de l'art, nous montrons dans la section 2.1 que des efforts significatifs ont été entrepris pour développer des modèles explicitement conçus autour du raisonnement logique, dans le but d'améliorer leurs capacités de résolution de tâches complexes. La section 2.2 s'intéresse plus spécifiquement au raisonnement mathématique, et en particulier à la multiplication, une opération représentative des difficultés rencontrées par les réseaux de neurones.

2.1 Avancées récentes en LLM orientés raisonnement algorithmique

Les récentes évolutions dans le domaine des grands modèles de langage témoignent d'un intérêt croissant pour leur capacité à résoudre des tâches relevant du raisonnement algorithmique, en particulier les tâches arithmétiques et logiques. De plus en plus, les modèles proposés par les grandes entreprises et laboratoires de recherche en intelligence artificielle intègrent explicitement ces dimensions, soulignant l'importance stratégique accordée à ce champ de compétences dans l'amélioration des performances générales des LLM.

Dans cette section, nous passons en revue quelques avancées dans le développement de modèles spécialisés dans ce type de raisonnement. La table 1 dresse un panorama de quelques modèles récents tels que LLEMMA, Mathstral 7B, QwQ-32B-Preview, ainsi que les versions o1 et o3 développées par OpenAI. Chaque modèle y est décrit selon ses spécificités, ses performances sur des benchmarks pertinents en relation avec son domaine d'expertise, et ses principales limites selon ses applications, identifiées à travers les retours de la communauté scientifique et nos propres expérimentations.

Modèle	Paramètres	Spécialisation	Performances	Limitations
LLEMMA[a]	7B / 34B	Raisonnement mathématique, preuve de théorèmes	Surpasse le modèle Minerva de Google[f] sur le benchmark MATH[g], base de données de test des capacités de raisonnement mathématique des modèles de langage.	Rencontre des difficultés en raisonnement algébrique avancé, notamment dans la preuve de théorèmes complexes nécessitant une compréhension approfondie et un raisonnement séquentiel.
Mathstral 7B[b]	7B	Raisonnement mathématique avancé	Sur le benchmark MATH[g], score de 56,6% en génération directe, 68,37% avec *vote majoritaire* (plusieurs réponses générées, choix de la plus fréquente), et 74,59% avec *reward model* (sélection de la meilleure réponse via un modèle de récompense). 63,47% de score sur le benchmark MMLU[h] évaluant la compréhension multitâche dans des domaines académiques variés.	Présente des difficultés dans le raisonnement avec des problèmes très complexes et dans les contextes nécessitant un raisonnement logique étendu, notamment en ce qui concerne les relations non-linéaires entre les éléments d'un problème.
QwQ-32B-Preview[c]	32,5B	Raisonnement mathématique et algorithmique	Sur les benchmarks MATH-500[i] et AIME[j], le modèle atteint respectivement 90,6% et 50% de réussite. Il obtient également 65,2% sur GPQA[k] et 50% sur LiveCodeBench[l].	Présente des limitations notables : mélange de langues, boucles de raisonnement récursif, compréhension limitée du sens commun et préoccupations éthiques et de sécurité[m].
OpenAI o1[d]	Non spécifié	Raisonnement complexe, RLHF(Apprentissage par renforcement à partir de retours humains)	Le modèle atteint un QI estimé à 120 selon des équivalences Mensa[a], se classe dans le top 500 du benchmark AIME[b], et atteint le 89e percentile sur Codeforces[c], plateforme de programmation compétitive. Il obtient également de bons résultats sur le benchmark GPQA[d], mesurant le raisonnement factuel et conceptuel.	Montre des limites en raisonnement long et multi-étapes, notamment dans des contextes ouverts ou nécessitant des inférences implicites[e].
OpenAI o3[e]	Non spécifié	Chaîne de pensée, raisonnement algorithmique	Sur le benchmark GPQA Diamond[f], o3 obtient un score de 87,7%. Il surpasse son prédécesseur o1 d'un facteur de 3 sur le test ARC-AGI[g], évaluant des capacités de généralisation analogiques. Il atteint également 71,7% SWE-bench Verified[h], mesurant l'aptitude à résoudre des problèmes de génie logiciel à partir de documents de bugs réels, et atteint un score de 2727 Elo sur Codeforces[i].	Peine encore à maintenir une cohérence parfaite sur des chaînes de raisonnement très longues ou incertaines, et peut échouer à synthétiser efficacement plusieurs sources d'information dans des scénarios non vus[j].

TABLE 1 – Quelques modèles récents spécialisés dans le raisonnement évalués de différentes manières selon leurs spécialisations

[a] arXiv – Llemma [b] Mistral – Mathstral [c] Reuters – QwQ [d] Wikipedia – OpenAI o1 [e] Wikipedia – OpenAI o3 [f] arXiv – Minerva [g] arXiv – MATH benchmark [h] arXiv – MMLU [i] Jurnals.net – QwQ benchmark [j] DeepLearning.ai – QwQ AIME [k] DataCamp – QwQ GPQA [l] Jurnals.net – LiveCodeBench [m] QwenLM blog – limitations QwQ [n] Mensa – QI estimé [o] AIME – benchmark math [p] Codeforces – compétition algorithmique [q] GPQA benchmark [r] OpenAI blog – limitations o1 [s] OpenAI Research – benchmarks o3 [t] SWE-bench – benchmark génie logiciel

La table 1 illustre clairement l'intérêt croissant de la communauté scientifique et industrielle pour le développement de modèles de langage capables de raisonner de manière algorithmique. Cette tendance se manifeste par la multiplication des travaux visant à doter les LLM de mécanismes cognitifs avancés leur permettant de résoudre des tâches complexes, telles que des problèmes mathématiques ou des processus logiques structurés. Ces efforts traduisent une volonté de combler les limites des approches actuelles fondées majoritairement sur la prédiction statistique, en intégrant des capacités plus proches du raisonnement humain, essentielles à la résolution de tâches nécessitant une exécution étape par étape ou une généralisation algorithmique.

2.2 Limites des réseaux de neurones dans l'apprentissage algorithmique : le cas de la multiplication

En plus des avancées récentes en matière de modèles de langage capables de traiter des tâches de raisonnement, des efforts notables ont ainsi été entrepris pour doter les architectures neuronales d'une forme de raisonnement procédural, en étendant les architectures usuelles ou encore en hybridant les modèles neuronaux avec des outils externes capables d'exécuter la tâche et transférer le résultat au réseau sans qu'il fasse le raisonnement lui même. Toutefois, malgré ces avancées, les réseaux de neurones souffrent encore de limitations structurelles qui compromettent leur capacité à apprendre et exécuter des algorithmes de manière fiable et généralisable.

Lors de notre étude nous avons pu tester une approche couramment utilisée qui consiste à déléguer l'exécution de l'algorithme de multiplication à des *API* externes capables d'effectuer des calculs mathématiques complexes et de renvoyer le résultat au modèle de langage. Dans ce cadre, le modèle dispose d'une liste d'*API* prédéfinies qu'il peut appeler pour exécuter la tâche algorithmique en dehors de son architecture, avant de récupérer le résultat et de l'intégrer à sa réponse.

Cette méthode s'avère efficace pour les tâches algorithmique simples avec un taux de succès de 100%, mais elle présente des inconvénients, notamment la difficulté d'exécution des appels imbriqués, la dépendance à des services externes, la latence des appels *API* et la confidentialité des données. En outre, elle ne permet pas au modèle de langage d'apprendre à exécuter l'algorithme par lui-même, ce qui limite sa capacité à généraliser en dehors de la liste de ressources qu'il a à sa disposition.

Le problème de multiplication sur lequel nous illustrons notre modèle d'apprentissage, d'apparence assez simple, est pourtant difficile lorsqu'il est résolu par un réseau de neurones ou un LLM.

Une étude menée par Hoshen et al. [6] montre que la multiplication est très difficile à apprendre de bout en bout (*end-to-end*) à l'aide d'un perceptron multi-couches.

Pour un grand modèle de langage, la performance de résolution d'opérations arithmétiques est par exemple corrélée à la fréquence d'apparition des opérandes dans le corpus d'entraînement [7]. Cela met en évidence que le réseau de neurones ne maîtrise pas la logique sous-jacente des opérations, mais se limite à reproduire des séquences apprises. Ce résultat met l'accent sur la légitimité de l'hypothèse du « perroquet stochastique », selon laquelle ces modèles ne font que répéter des patterns observés dans les données sans une réelle compréhension ou capacité à généraliser en dehors de ces exemples spécifiques.

L'inférence algorithmique exige une exécution étape par étape avec des mises à jour structurées des données, ce qui dépasse les capacités des LLM dans leur utilisation courante, qui fonctionnent principalement par association de motifs et génération probabiliste de texte. Contrairement aux architectures conçues pour l'exécution symbolique, les LLM ne disposent pas d'une mémoire de travail persistante ni d'un mécanisme interne permettant d'appliquer des règles de manière itérative avec un état contrôlé, ce qui limite leur aptitude à effectuer des raisonnements algorithmiques complexes.

Une démarche permettant de faciliter l'apprentissage de tâches complexes consiste à accompagner le modèle pour lui permettre d'apprendre des tâches de complexité croissante [3]. Cette décomposition a permis d'améliorer l'apprentissage de la multiplication de bout en bout pour une réseau de neurones récurent [2, 1]. Malgré les résultats prometteurs, ces travaux restent limités par la difficulté intrinsèque de l'apprentissage algorithmique, notamment en raison des dépendances à long terme et des problèmes d'aptitude à l'entraînement *(trainability problems)*. Les modèles de langage utilisés peinent à généraliser lorsque la variabilité des entrées est importante. De plus, les méthodes reposent encore sur une supervision complète, posant la question de la capacité du modèle à inférer des étapes manquantes.

Dans le cadre de cet article, nous souhaitons ainsi apprendre à un grand modèle de langage à réaliser des inférences complexes. Nous nous inspirons du raisonnement en chaîne de pensée (Chain of Thought, CoT) [9] qui permet d'améliorer les capacités des LLM à structurer leur raisonnement, en intégrant au raisonnement des tâches apprises de complexité croissante et un transfert algorithmique entre elles.

3 Modèle d'apprentissage proposé

Dans cette section, nous détaillons la méthodologie d'apprentissage employée pour entraîner un grand modèle de langage sur la tâche de multiplication et le corpus d'apprentissage que nous avons construit.

L'idée fondamentale de notre approche introduite sous le nom de LLM-DAL (Large Language Model - Decompositional Algorithmic Learning), repose sur la décomposition de la tâche étudiée en plusieurs sous-tâches simples, chacune représentant une étape clé du processus global. Cette stratégie est applicable à de nombreux problèmes algorithmiques, et nous l'adaptons ici spécifiquement à la multiplication. L'ordre d'apprentissage des sous-tâches est soigneusement structuré pour renforcer la capacité du modèle à généraliser ses connaissances et à apprendre de manière progressive.

L'intérêt de cette méthode est double : elle permet non seulement d'améliorer la capacité de raisonnement et de généralisation du modèle, mais aussi de limiter la complexité en scindant la tâche en étapes assimilables et facilement apprenables.

Ainsi, l'approche de *fine-tuning* utilisée pour ce modèle d'apprentissage présente une structure qui pourrait être appliquée à d'autres tâches algorithmiques, et elle démontre comment un apprentissage progressif, structuré en sous-tâches, peut améliorer la capacité d'un modèle de langage à raisonner pour résoudre des problèmes complexes.

3.1 *Chain of Thought* et décomposition de tâches

Le raisonnement en chaîne de pensée (Chain of Thought, CoT) [9] est une approche qui consiste à décomposer une tâche complexe en plusieurs sous-tâches plus simples, qui sont ensuite résolues séquentiellement pour arriver au résultat final. Cette approche permet de simplifier la tâche globale en la décomposant et la simplifiant par différentes étapes, tout en permettant au modèle de langage de raisonner de manière plus systématique et de généraliser plus efficacement.

Dans le cadre de l'apprentissage de la multiplication, cette approche permet de décomposer l'opération en sous-tâches élémentaires, ensuite combinées pour obtenir le résultat final.

Ces sous-tâches incluent :

— La multiplication chiffre par chiffre, $t1_{mult}$: qui

consiste à multiplier chaque chiffre du multiplicateur par chaque chiffre du multiplicande. Le modèle de langage doit pour cette tâche pouvoir s'appuyer sur une connaissance des tables de multiplication chiffre à chiffre.

— L'addition des résultats partiels, $t2_{add}$: qui consiste à additionner le résultat de la multiplication intermédiaire (chiffre à chiffre) avec la retenue qu'on a obtenu de la multiplication précédente. Les additions à apprendre sont donc des additions de deux chiffres au maximum du résultat de la multiplication intermédiaire chiffre à chiffre et de la retenue qui ne peut être qu'un seul chiffre.

— L'extraction du nième chiffre d'un nombre, $t3_{extract}$: qui consiste à extraire le chiffre à la position n d'un nombre donné.

— La concaténation à gauche d'un nombre à un autre, $t4_{concat}$: qui consiste à ajouter un chiffre à gauche d'un nombre donné.

Après un apprentissage supervisé progressif sur ces sous-tâches, le modèle entraîné sera capable de générer une description en langage naturel du raisonnement suivi pour effectuer une multiplication de deux nombres. Cette description permet d'évaluer la capacité du modèle à raisonner de manière logique et à s'adapter à des exemples complexes, même en dehors de son corpus d'entraînement.

Par exemple, lorsque interrogé sur le résultats de la multiplication de 5847 par 2, le modèle devrait être capable de générer une telle description du raisonnement lui permettant de trouver le résultat attendu :

```
multiplying 5847 by 2: 5847*2=
carry=0
digit 1 of 5847 is 7
multiply digit 1 of 5847 which is 7 by
2: temp_mult=7*2=14
Add the multiplication result to the
carry: temp_add=carry+temp_mult=0+14=14
compare the addition result to 10:
temp_add=14>=10
the first digit of temp_add=14 is
fd_result=4
first digit of temp_add which is
4 is concatenated to the left of
temp_result=
the result of the concatenation is 4
the second digit of temp_add is 1 which
will be the value of the carry: carry=1

carry=1
digit 2 of 5847 is 4
multiply digit 2 of 5847 which is 4 by
2: temp_mult=4*2=8
Add the multiplication result to the
carry: temp_add=carry+temp_mult=1+8=9
compare the addition result to 10:
temp_add=9<10
carry=0
temp_add is concatenated to the left of
temporary_result=4

...
final carry=1>0
the final carry which is 1 is
concatenated to the left of the final
result which is 1694
the result of the concatenation is
11694
the final_result is 11694
```

3.2 Corpus d'apprentissage

Dans le cadre de notre modèle d'apprentissage, des corpus d'entraînement ont été générés synthétiquement pour chacune des sous-tâches identifiées. Cette génération a été rendue nécessaire en l'absence de bases de données existantes répondant précisément à nos besoins expérimentaux et à la granularité requise pour l'apprentissage algorithmique supervisé. Des descriptions en langage naturel ont accompagné chaque entrée du jeu de données généré afin de permettre au modèle de langage de les maîtriser parfaitement.

Des extraits de ces données d'entraînement sont présentés sur les tables suivantes :

Pour la tâche $t1_{mult}$, un exemple d'entrée-sortie est présenté sur la table 2. Chaque exemple est composé de l'opération de multiplication (le multiplicateur, le multiplicande et les opérateurs) et le résultat de la multiplication en sortie.

Entrée	6 by 8 : 6*8=
Sortie	48

TABLE 2 – Tâche de multiplication chiffre à chiffre

Pour la tâche $t2_{add}$, un exemple d'entrée-sortie est présenté sur la table 3. Chaque exemple est composé de l'opération d'addition (les deux nombres et l'opérateur) et le résultat de l'addition en sortie.

Entrée	12 plus 6 : 12+6=
Sortie	18

TABLE 3 – Tâche d'addition d'un nombre à 2 chiffres à un seul chiffre

En ce qui concerne la tâche $t3_{concat}$, un exemple d'entrée-sortie est présenté sur la table 4. Chaque exemple est composé du nombre à ajouter à gauche et du nombre de destination, ainsi que le résultat de la concaténation en sortie.

Entrée	concatenating 16 to 375347 on the left gives
Sortie	16375347

TABLE 4 – Tâche de concaténation à gauche d'un nombre à un autre

Pour la tâche $t4_{extract}$, un exemple d'entrée-sortie est présenté sur la table 5. Chaque exemple est composé du nombre et de la position du chiffre à extraire, ainsi que le chiffre extrait en sortie.

Entrée	the 3rd digit from 393721 is
Sortie	7

TABLE 5 – Tâche d'extraction du nième chiffre

Un dernier corpus d'entraînement a été créé, combinant les quatre sous-tâches précédentes pour générer une description complète du raisonnement nécessaire à la multiplication de deux nombres, comme indiqué dans l'exemple à la fin de la section 3.1. Cette approche permet au modèle de fusionner les connaissances acquises à travers les sous-tâches individuelles et de produire une explication détaillée du processus de multiplication.
L'objectif de ces entraînements est de tirer parti des capacités de génération de texte des LLM pour leur apprendre à accomplir une tâche algorithmique complexe, en combinant des algorithmes exprimés en langage naturel avec des données d'entraînement synthétiques.

4 Expérimentations

4.1 Contexte expérimental

Le LLM que nous utilisons pour exécuter l'ensemble des étapes d'apprentissage est 3.2, un modèle de langage de 3 milliards de paramètres dans sa version Instruct [1].
Cette version a été affinée pour suivre des instructions. Contrairement aux modèles non adaptés à l'instruction, qui génèrent simplement une suite à l'invite, ces modèles sont conçus pour répondre de manière précise et contextualisée aux demandes de l'utilisateur.
Ce modèle a été choisi pour sa compatibilité avec les tâches à effectuer, les ressources de calcul à notre disposition mais aussi pour sa performance très honorable pour sa taille sur les différents *benchmarks* sur lequel il a été évalué.
Dans le cadre de notre approche algorithmique, la mémoire de travail nécessaire à la manipulation des variables intermédiaires est intégrée dans le contexte du LLM utilisé. Nous avons exploité la capacité étendue de la fenêtre de contexte de Llama 3.2, qui atteint jusqu'à 128 000 jetons [1]. Cette capacité permet au modèle de conserver et de traiter efficacement des séquences complexes et de longue durée, essentielles pour l'apprentissage de tâches arithmétiques multi-étapes.
Afin de pouvoir reproduire les résultats présentés, nous mettons à disposition le corpus d'apprentissage par tâche que nous avons constitué, accessible via les liens fournis en notes de bas de page.
Pour la tâche $t1_{mult}$, un corpus d'entraînement de toutes les combinaisons possibles de multiplication de deux chiffres de 0 à 9 , donnant un total de 100 exemples, a été généré. Chaque exemple est composé de l'opération de multiplication (le multiplicateur, le multiplicande et les opérateurs) et le résultat de la multiplication.
De la même façon, nous avons généré un corpus d'entraînement qui regroupe toutes les combinaisons relatives à la tâche $t2_{add}$ avec les chiffres allant de 0 à 9. Nous obtenons ainsi une base de données à 1000 exemples d'additions. Chaque exemple est composé de l'opération d'addition (les deux nombres et l'opérateur) et le résultat de l'addition.
Les deux bases de données précédentes relatives à $t1_{mult}$ et $t2_{add}$, ont été fusionnées pour former un seul corpus d'entraînement d'addition et de multiplication [2], d'une taille de 1100 exemples, qui sera ensuite utilisé pour réaliser l'étape initiale de l'entraînement du modèle.
Pour la tâche $t3_{extract}$, un corpus d'entraînement de 5000 exemples aléatoires d'extraction de nième chiffre à partir de nombres 4 à 6 chiffres a été généré [3]. Chaque exemple est composé du nombre et de la position du chiffre à extraire, ainsi que le chiffre extrait.
Pour la tâche $t4_{concat}$, un corpus d'entraînement de 5000 exemples aléatoires de concaténation de nombres de 1 à 2 chiffres à des nombres de 4 à 6 chiffres a été généré [4]. Chaque exemple est composé du nombre à ajouter à gauche et du nombre de destination, ainsi que du résultat de la concaténation.
L'ensemble des bases de données précédentes ont été réparties comme suit : 70%, soit 3500 exemples, pour la base d'entraînement et 30%, soit 1500 exemples, pour la base d'évaluation.
Enfin, une dernière base de données a été générée à l'aide d'un algorithme qui regroupe les 4 sous-tâches précédentes pour produire une description du raisonnement adopté pour effectuer la multiplication globale de deux nombres [5]. Cette base de données contient 2000 exemples aléatoires de multiplications de nombres de 4 à 6 chiffres à un seul chiffre. Chaque exemple est composé des deux nombres à multiplier, de la description du raisonnement et du résultat de la multiplication. La base de données est relativement petite par rapport aux autres pour des raisons de longueur des descriptions générées qui peuvent être assez longues et complexes à apprendre pour le modèle dans un temps raisonnable. La base de données a été répartie comme suit : 90%, soit 1800 exemples, pour la base d'entraînement et 10%, soit 200 exemples, pour la base d'évaluation. La répartition diffère des autres bases de données en raison de la taille relativement petite de la base de données et de la complexité des descriptions générées à apprendre pour le modèle.

4.2 Processus d'apprentissage incrémental

Le protocole expérimental a été conçu en plusieurs étapes intermédiaires d'entraînement, permettant au modèle d'acquérir progressivement les différentes sous-tâches nécessaires à la résolution complète d'une multiplication. Cette approche progressive, détaillée dans la table 6, vise à renforcer la compréhension du modèle en lui fournissant un apprentissage structuré et décomposé, avant d'aborder la tâche globale.
Une première étape a consisté à réaliser un *fine-tuning* du modèle de base Llama-3.2-3b-instruct sur les tâches $t1_{mult}$ et $t2_{add}$, aboutissant à un modèle M1. Étant donné l'im-

1. https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct
2. https://huggingface.co/datasets/farahbs/onetwo-digits-mult-add
3. https://huggingface.co/datasets/farahbs/extraction_db
4. https://huggingface.co/datasets/farahbs/left-concat
5. https://huggingface.co/datasets/farahbs/cot-multiplication-2k

Étape	Tâche	Type d'entraînement
1	$t1_{mult}$ + $t2_{add}$	sur-apprentissage
2	$t3_{extract}$ + $t4_{concat}$	*affinage*
3	$t_{globale}$	*affinage*

TABLE 6 – Configurations d'entraînement pour l'apprentissage algorithmique de la multiplication

portance capitale de ces deux sous-tâches par rapport aux autres, nous avons opté pour un sur-apprentissage du modèle spécifiquement sur ces deux tâches. En effet, bien que limitées en nombre, ces tâches sont fondamentales, car elles constituent les opérations de base sur lesquelles le modèle s'appuie pour effectuer un raisonnement complet et cohérent. La capacité de généralisation interviendra lors de l'exécution de l'opération complète.

La base de données combinée a été utilisée uniquement pour l'entraînement, avec une partie de validation mais sans évaluation sur des exemples non vus. Le *fine-tuning* s'est arrêté lorsque la fonction de perte utilisée a atteint un plateau, indiquant que le modèle avait appris les tâches d'addition et de multiplication de manière satisfaisante.

L'étape suivante a été de faire le *fine-tuning* du modèle M1 sur la tâche $t3_{extract}$, aboutissant à un modèle M2. Cette tâche est cruciale pour le processus global, car elle permet de récupérer les chiffres nécessaires pour effectuer les opérations de multiplication chiffre par chiffre.

L'étape suivante a été de faire le *fine-tuning* du modèle M2 sur la tâche $t4_{concat}$, aboutissant à un modèle M3. Cette tâche est également cruciale pour l'ensemble du raisonnement, car elle permet de combiner les chiffres extraits et les résultats intermédiaires pour obtenir le résultat final de la multiplication.

Les modèles M2 et M3 ont été entraînés sur leurs corpus d'apprentissage respectifs en appliquant la même approche de *fine-tuning* que pour M1. La progression de la fonction de perte a été surveillée tout au long de l'entraînement, aussi bien sur la base d'entraînement que sur la base d'entraînement, afin d'éviter tout phénomène de sur-apprentissage (où le modèle mémorise excessivement les données d'entraînement sans généraliser) ou de sous-apprentissage (où il ne parvient pas à extraire des tendances utiles des données). Une fois le *fine-tuning* terminé, les performances des modèles ont été évaluées sur la base d'évaluation pour mesurer leur capacité à généraliser sur des exemples inédits.

Enfin, une dernière étape a regroupé l'entraînement du modèle M3 sur la base de données de la tâche globale de multiplication, aboutissant ainsi au modèle final de la méthode LLM-DAL.

Après cet apprentissage supervisé sur la tâche globale de multiplication, les performances du modèle ont été évaluées sur la base d'évaluation. Toutefois, lors des premières évaluations, le modèle ne parvenait pas à produire une réponse complète incluant toutes les étapes intermédiaires nécessaires au raisonnement pour effectuer la multiplication demandée. C'est pour cela qu'un mécanisme de *prompting* récursif a été mis en place pour permettre au modèle de générer une réponse complète en plusieurs étapes. Ce mécanisme consiste à demander au modèle de générer une réponse partielle à la question posée, puis de lui demander de générer la suite de la réponse en fonction de sa réponse fournie précédemment. Un parseur vérifie à chaque fois si la phrase indiquant la fin de la réponse est générée, si ce n'est pas le cas, le modèle est relancé pour générer la suite de la réponse. Ce processus est répété jusqu'à ce que le modèle produise une réponse complète à la question posée ou qu'un nombre maximal d'itérations, fixé à 10 dans notre cas, soit atteint, afin de prévenir tout phénomène de divergence ou de répétition indéfinie.

4.3 Résultats

La table 7 ci-dessous illustre les résultats de l'évaluation du modèle entraîné sur la tâche globale de multiplication à l'aide de LLM-DAL. Cette approche a permis d'entraîner le modèle sur les différentes sous-tâches identifiées, d'explorer la tâche globale finale de multiplication et d'exploiter le *prompting* récursif pour générer l'intégralité des réponses à cette tâche. Les performances obtenues sont comparées à celles du modèle Vanilla, correspondant à la version initiale de *Llama-3.2-instruct*, entraînée et évaluée directement sur le corpus d'apprentissage de la tâche globale sans étapes intermédiaires afin d'évaluer le vrai impact de notre méthode d'entraînement.

Tâche	Modèle Vanilla	**LLM-DAL**
Taux de précision	13,5%	**42,1%**

TABLE 7 – Résultats d'évaluation sur la tâche globale de multiplication

Les résultats montrent que le modèle entraîné à l'aide de LLM-DAL présente une amélioration de performance de plus de 28% sur le raisonnement et la tâche globale de multiplication.

La table 8 présente les résultats de l'évaluation du modèle initial entraîné séparément sur chacune des sous-tâches. Les résultats montrent que le modèle a obtenu des performances presque à 100% sur chacune des étapes intermédiaires, ce qui nous a permis de passer à l'étape suivante de l'entraînement.

Tâche	Taux de précision
$t1_{mult}$ + $t2_{add}$	**100%**
$t3_{extract}$	**97,6%**
$t4_{concat}$	**99,2%**

TABLE 8 – Taux de précision par sous-tâche lors de l'entraînement

Pour la suite des expériences, le modèle M1, ayant été spécifiquement entraîné sur les tâches $t1_{mult}$ et $t2_{add}$, a été sélectionné comme point de départ pour les entraînements suivants.

L'étape suivante d'entraînement consistait en l'apprentissage supervisé du modèle M1 de la tâche $t3_{extract}$, abou-

tissant au modèle M2. Les résultats de précision obtenus sur la base de données d'entraînement relative à cette tâche sont présentés dans la table 9.

Tâche	Taux de précision
$t1_{mult}$ + $t2_{add}$	**100%**
$t3_{extract}$	**99,3%**

TABLE 9 – Taux de précision par sous-tâche lors de la deuxième étape d'entraînement

La dernière expérience a pris le modèle M2 comme point de départ, qui a été entraîné sur la tâche $t4_{concat}$ à son tour, aboutissant au **modèle M3**.

Les résultats de cette expérience, illustrés sur la table 10, ont montré que le modèle M3 a obtenu les précisions suivantes sur les différentes tâches :

Tâche	Taux de précision
$t1_{mult}$ + $t2_{add}$	**99,81%**
$t3_{extract}$	**99%**
$t4_{concat}$	**99,53%**

TABLE 10 – Taux de précision par sous-tâche lors de l'entraînement combiné

Ces résultats montrent que le réglage fin sur les différentes tâches a permis d'améliorer les performances du modèle sur ces tâches spécifiques. Le modèle a bien montré sa capacité à généraliser ses connaissances et à fournir des résultats précis et cohérents sur des tâches qu'il n'avait pas vues pendant l'entraînement ainsi que sa capacité à transférer des connaissances apprises sur des tâches spécifiques vers des tâches plus complexes. Ce processus de transfert de connaissances, où un modèle utilise les informations apprises sur une tâche ou un jeu de données pour améliorer ses performances sur une autre tâche ou jeu de données apparenté, joue un rôle clé dans l'apprentissage de la tâche globale de multiplication qui nécessite la maîtrise de toutes les sous-tâches identifiées.

De plus, notre méthode d'apprentissage introduite nous a permis de se situer dans un cadre d'IA frugale, où le modèle a pu acquérir une capacité à résoudre des opérations de manière logique et à s'adapter à des situations nouvelles, au-delà des données sur lesquelles il a été entraîné. Grâce à la taille réduite du modèle et des bases de données d'entraînement, le modèle a appris à effectuer une tâche algorithmique complexe en un temps raisonnable et avec une puissance de calcul limitée. Par exemple, il a pu combiner les quatre sous-tâches pour générer une description en langage naturel du raisonnement d'une multiplication de deux nombres en moins de 2 heures, sur une machine équipée d'un GPU NVIDIA A40 à 48 Go de mémoire. Cela se distingue des modèles plus grands et des méthodes d'entraînement alternatives, qui exigent des semaines de formation sur des infrastructures de calcul coûteuses.

5 Conclusion

L'apprentissage algorithmique est un enjeu central pour dépasser les limitations des LLM, permettre une meilleure appropriation des données manipulées et sortir du stéréotype du "perroquet stochastique", qui limite leur capacité à raisonner de manière autonome.

Bien que plusieurs travaux explorent cette direction, il s'agit toujours d'un problème ouvert et d'actualité. L'intégration d'un apprentissage algorithmique structuré pourrait permettre aux modèles de langage de mieux généraliser et de produire des résultats plus qualitatifs, basés plus sur du raisonnement logique et pas que sur de la simple mémorisation.

Dans ce travail, nous nous sommes intéressés à la multiplication, une tâche de nature algorithmique dont l'apprentissage s'est révélé particulièrement difficile pour les LLM. Nous avons mis en évidence les limites des modèles actuels face à cette tâche et proposé une approche permettant aux LLM d'acquérir une logique de résolution dans des temps d'apprentissage raisonnables. Il est cependant important de noter que ce problème est abordé comme un cas d'étude illustratif, à travers lequel nous explorons les limites actuelles de ces modèles, en nous concentrant sur leur capacité à généraliser et à raisonner de manière algorithmique.

Nous avons mesuré l'intérêt et l'apport potentiel de l'apprentissage algorithmique pour améliorer la capacité des LLM à exécuter des algorithmes. Pour cela, nous avons introduit LLM-DAL, un modèle d'apprentissage en plusieurs étapes, combinant un apprentissage supervisé par sous-tâches et un prompting structuré pour renforcer la cohérence du raisonnement. Les résultats obtenus montrent une amélioration significative de la précision du modèle ainsi qu'une meilleure capacité de généralisation sur des tâches similaires.

Enfin, ce travail ouvre plusieurs perspectives pour la suite. Une application à d'autres tâches algorithmique permettrait d'évaluer la robustesse et la transférabilité des approches étudiées. De plus, il serait intéressant d'explorer des méthodes pour extraire automatiquement les raisonnements de type Chain of Thought sans supervision, afin d'améliorer la capacité des modèles à structurer leur apprentissage de manière plus autonome et efficace.

Références

[1] Frédéric Armetta, Anthony Baccuet, and Mathieu Lefort. Algorithmic learning, a next step for ai. an application to arithmetic operations. In *Artificial Intelligence and Cognition (AIC)*, 2023.

[2] Frédéric Armetta, Anthony Baccuet, and Mathieu Lefort. L'apprentissage algorithmique, une nouvelle étape pour l'ia. une application aux opérations arithmétiques. In *CNIA 2023-Conférence Nationale en Intelligence Artificielle, PFIA*, pages 31–39, 2023.

[3] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Procee-*

dings of the 26th annual international conference on machine learning, pages 41–48. ACM, 2009.

[4] Mark Collier and Joeran Beel. Implementing neural turing machines. In *Artificial Neural Networks and Machine Learning–ICANN 2018 : 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27*, pages 94–104. Springer, 2018.

[5] Alex Graves. Neural turing machines. *arXiv preprint arXiv :1410.5401*, 2014.

[6] Yedid Hoshen and Shmuel Peleg. Visual learning of arithmetic operations. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 3733–3739. AAAI Press, 2016.

[7] Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. Impact of pretraining term frequencies on few-shot reasoning. *arXiv preprint arXiv :2202.07206*, 2022.

[8] John R. Searle. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(03) :417–424, 1980.

[9] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35 :24824–24837, 2022.

CoA-Text2OWL: A Chain-of-Agents Framework for Ontology Learning 2025

Hussam Ghanem[1], Samir Jabbar[1], Christophe Cruz[1]

[1] Université Bourgogne Europe, CNRS,
Laboratoire Interdisciplinaire Carnot de Bourgogne ICB UMR 6303, 21000 Dijon, France

hussam.ghanem@u-bourgogne.fr

Résumé

L'apprentissage d'ontologies à partir de texte non structuré reste un défi complexe, en particulier pour les sources textuelles volumineuses et complexes. Cet article présente (CoA-Text2OWL), un cadre multi-agent qui exploite les grands modèles de langage (LLMs) au sein d'une chaîne d'agents afin d'améliorer la génération d'ontologies. Contrairement aux approches traditionnelles basées sur un seul LLM, CoA-Text2OWL répartit la tâche entre plusieurs agents de travail, chacun traitant une partie du texte d'entrée, tandis qu'un agent gestionnaire synthétise leurs résultats en une ontologie cohérente. Nous évaluons notre approche par rapport à une méthode Text2OWL basée sur un seul LLM, et montrons des améliorations en termes d'extraction de propriétés d'objet et de complétude de l'ontologie. Toutefois, des défis persistent, notamment pour la préservation des structures hiérarchiques. Nos résultats mettent en évidence le potentiel de l'IA multi-agent pour l'apprentissage d'ontologies et suggèrent des améliorations futures, notamment des rôles spécialisés d'agents pour l'extraction de termes, la classification et la validation. Cette recherche contribue au domaine émergent des systèmes multi-agents alimentés par les LLMs et à leur application dans la représentation des connaissances.

Mots-clés

Apprentissage d'Ontologies, Agentic AI, Text2OWL, Grands Modèles de Langage.

Abstract

Ontology learning from unstructured text remains a complex challenge, particularly for large and intricate textual sources. This paper introduces CoA-Text2OWL, a multi-agent framework that leverages Large Language Models (LLMs) within a Chain-of-Agents to improve ontology generation. Unlike traditional single-LLM approaches, CoA-Text2OWL distributes the task across multiple worker agents, each processing a chunk of the input text, while a manager agent synthesizes their outputs into a coherent ontology. We evaluate our approach against a baseline single-LLM-based Text2OWL method, demonstrating improvements in object property extraction and ontology completeness. However, challenges remain in preserving hierarchical structures. Our results highlight the potential of multi-agent AI for ontology learning and suggest future enhancements, including specialized agent roles for term extraction, classification, and validation. This research contributes to the evolving field of LLM-powered multi-agent systems and their application in knowledge representation.

Keywords

Ontology Learning, Agentic AI, Text2OWL, Large Leanguage Models.

1 Introduction

Ontology learning, the task of automatically constructing and populating ontologies from unstructured text, has traditionally relied on a combination of linguistic, statistical, and logic-based methods [2, 7]. However, the recent emergence of large language models (LLMs) has opened new avenues for approaching this complex task. LLMs, with their ability to understand and generate human language, reason, and extract relevant information, offer the potential to streamline and automate aspects of ontology learning [7, 3].

Concurrently, the field of artificial intelligence has witnessed the rise of Agentic AI, a paradigm shift towards creating autonomous systems capable of pursuing complex goals with minimal human intervention [1]. Agentic AI systems demonstrate adaptability, advanced decision-making capabilities, and self-sufficiency, enabling them to operate dynamically in evolving environments [5]. This evolution in AI technology promises transformative applications across various domains, including ontology learning. However, the increasing agency of AI systems also raises important ethical considerations and challenges, particularly in terms of accountability and societal impact[5, 8].

One promising application of LLMs in the domain of ontology learning is direct ontology generation from text, often referred to as Text2OWL [7]. This approach aims to directly translate textual information into a formal ontology representation, typically using the Web Ontology Language (OWL). While this method holds significant promise, it can be challenging for longer and more complex texts, where a single LLM may struggle to maintain coherence and extract all relevant information.

To address these limitations, recent research has introduced

the Chain-of-Agents (CoA) framework, a multi-agent system architecture that has demonstrated significant improvements in performance on long-context tasks [10]. CoA distributes the processing of input text across multiple "worker" agents, which collaboratively analyze different segments of the text and communicate their findings. A "manager" agent then synthesizes these contributions to produce a final, coherent output.

In this paper, we propose a novel approach that combines the power of LLMs with the Chain-of-Agents framework for ontology learning. By applying CoA to the LLM Text2OWL approach, we aim to overcome the challenges of coherence and completeness in automatically generated ontologies, particularly when dealing with extensive and intricate textual sources. We hypothesize that by breaking down the ontology generation task into smaller, more manageable steps and leveraging multiple LLM agents to collaborate on the solution, we can achieve more accurate and complete ontologies, especially when dealing with complex or lengthy textual sources.

The contributions of this paper are as follows :

- We propose a novel CoA-based architecture for LLM Text2OWL ontology generation, leveraging the principles of Agentic AI.
- We evaluate the performance of our approach on [I have to specify datasets, likely drawing from [7]], comparing it to existing single-LLM ontology learning method.
- We analyze the advantages and disadvantages of using CoA in this context, considering both performance metrics and ethical implications.
- We provide insights into the design of effective agent communication strategies for ontology learning, contributing to the broader field of LLM-based multi-agent systems.

This paper is structured as follows : Section 2 presents a comprehensive review of related work in Agentic AI, Multi-Agent Systems, and both traditional and LLM-based Ontology Learning approaches. In Section 3, we detail the proposed CoA-Text2OWL methodology, outlining the system architecture, agent prompts, and overall workflow. Section 4 presents the experimental setup, datasets, evaluation metrics, and the results obtained. Section 5 then offers a discussion of the findings, including performance comparisons, an analysis of the generated ontologies, and an examination of the impact of chain length and agent configurations, as well as limitations and challenges. Finally, Section 6 concludes the paper with a summary and an outline of future research directions.

2 Related Work

This section provides an overview of related research in two key areas : Agentic AI and Multi-Agent Systems, and Ontology Learning (Traditional and LLM-based approaches).

2.1 Agentic AI and Multi-Agent Systems

The field of Agentic AI is rapidly evolving, with a growing emphasis on creating autonomous systems capable of pursuing long-term goals, making decisions, and executing complex workflows [5]. Agentic AI shares conceptual roots with intelligent agents and multi-agent systems [9], but is distinguished by its capacity to tackle open-ended tasks and proactively generate novel solutions [5]. The survey by [4] provides an in-depth discussion of essential aspects of LLM-based multi-agent systems, including simulated environments, agent profiling, communication strategies, and mechanisms for capacity growth. LLM-based Multi-Agent systems offer advanced capabilities compared to systems using a single LLM-powered agent, because the LLMs can be specialized into various distinct agents, each with different capabilities, and enabling interactions among these diverse agents to simulate complex real-world environments effectively [4].

The Chain of Agents (CoA) approach [10] aligns with the cooperative communication paradigm discussed by [4], where agents work together toward shared objectives by exchanging information. In CoA, worker agents, which they are parallel and independent, pass information and insights to a manager agent who synthesizes the final output.

The rise of agentic AI presents both opportunities and challenges [5] (Mukherjee & Chang, 2025). As AI systems take on roles with greater independence, it becomes crucial to address issues related to liability attribution, informed consent, and accountability [5]. Furthermore, the potential for emergent self-regulation within networks of agentic AI raises important questions about aligning these norms with societal values and ensuring transparency and accountability [5]. Our work builds upon the principles of agentic AI by leveraging multiple agents to collaboratively perform ontology learning tasks, aiming to improve both accuracy and interpretability. This research also aligns with efforts to specialize LLMs into agents with distinct capabilities and enable interactions to simulate complex environments, as noted by [4].

2.2 Ontology Learning (Traditional and LLM-based approaches

Ontology learning is the process of automatically constructing ontologies from data [2]. Traditional approaches to ontology learning have involved a combination of techniques from natural language processing, machine learning, and knowledge representation [2]. These techniques are often applied to sub-tasks such as term extraction, concept identification, relationship discovery, and axiom formulation [2].

With the advent of LLMs, new approaches to ontology learning have emerged [7]. LLMs have been explored for various ontology learning tasks, including term typing [3], relationship extraction [3], and axiom generation [7, 3]. One particularly relevant approach is direct ontology generation, where an LLM is prompted to directly generate an ontology from text [7]. This is also called Text2OWL, where textual information will be translated to formal OWL ontology [7]. [7] investigated the application of LLMs for unsupervised ontology learning, comparing different LLM-based pipelines to traditional methods. Their study highlighted the potential of LLMs for automating aspects of ontology lear-

ning, but also identified limitations related to coherence and completeness. They evaluated three ontology learning pipelines : LLM Text2OWL, OLAF LLM (using only LLM-based components), and OLAF no-LLM (using only non-LLM-based components) [7].
Our work builds upon these efforts by introducing a multi-agent approach to the LLM Text2OWL pipeline. By leveraging the Chain-of-Agents framework [10], we aim to overcome the limitations of single-LLM approaches and improve the accuracy and completeness of automatically generated ontologies. This aligns with current research to use LLM-based multi-agents to solve various tasks, such as software development, multi-robot systems, society simulation, policy simulation, and game simulation [4].

3 Methodology : CoA-Text2OWL

Our CoA-Text2OWL framework implements a novel approach to ontology generation from PDF documents (textual documents) using Chain-of-Agents (CoA) architecture [10]. This section details the system's components and workflow.

3.1 System Architecture

The CoA-Text2OWL system comprises three main components 1 :

1. **Input Processing Module** : Handles PDF parsing and text segmentation.
2. **Chain-of-Agents Core** : Consists of worker agents and a manager agent for distributed ontology construction.
3. **Output Generation Module** : Responsible for final ontology synthesis and serialization.

3.2 Agent Prompts

Our system utilizes carefully crafted prompts, drawing inspiration from the prompt format used by [7], to guide the behavior of both worker and manager agents. However, the content of the prompts was specifically designed and adapted to our Chain-of-Agents architecture.
For worker agents, the prompt emphasizes their role in ontology construction and extracting key ontological elements from a specific text chunk :

> You are a helpful assistant in building an ontology. You are fluent in the W3C Semantic Web stack and RDF, RDFS, and OWL languages. Use the given text to construct an OWL ontology in the Turtle format. Use this namespace : http ://example.org/example#. Return only the turtle file. Extract all the possible classes and subclasses.

This prompt is designed to focus workers on extracting classes and subclasses within the confines of their assigned chunk, while adhering to specific syntax guidelines for OWL ontologies in Turtle format.
The manager agent's prompt, on the other hand, emphasizes its role in integration and coherence, ensuring that the final ontology is a unified and consistent representation of the input text :

> You are an ontology learning expert and your goal is to integrate some sub ontologies into a coherent ontology while ensuring no redundancy. Your response should be structured, modular, and ready for further reasoning and inference.

These prompts are crucial in guiding the agents' behavior, ensuring that worker agents focus on extraction and initial structuring, while the manager agent concentrates on integration and overall coherence of the final ontology. While inspired by the format of [7], the content is uniquely tailored to leverage the strengths of our Chain-of-Agents approach.

3.3 Workflow

3.3.1 Input Processing

The system begins by ingesting a PDF document using PyMuPDF [1]. The extracted text undergoes a segmentation process, dividing it into chunks. We choose a size that balance context preservation with the optimal input capacity for the language model.

3.3.2 Chain-of-Agents Processing

The core of our system leverages a sequential Chain-of-Agents approach :

1. **Worker Agents** : Multiple instances of WorkerAgent, powered by the open-mistral-7b model [2], process individual text chunks. Each worker agent receives :
   — A chunk of text
   — The user's query (e.g., "Use the given text to construct an OWL ontology in the Turtle format")
   — The output from the previous agent (initially None)
   Workers generate partial ontologies in Turtle format based on their assigned chunk.
2. **Manager Agent** : A single ManagerAgent, also using the open-mistral-7b model, synthesizes the outputs from all worker agents into a coherent final ontology.

3.3.3 Agent Communication

The agents communicate through a context-preserving mechanism, allowing for the accumulation and propagation of knowledge across text chunks. Each worker agent can potentially build upon the context provided by previous agents, enhancing the coherence of the generated ontology.

3.3.4 Ontology Synthesis

The manager agent performs the crucial task of integrating partial ontologies from worker agents. This process involves :

1. PyMuPDF : https://pymupdf.readthedocs.io/en/latest/
2. Mistral-7b : https://mistral.ai/fr/news/announcing-mistral-7b

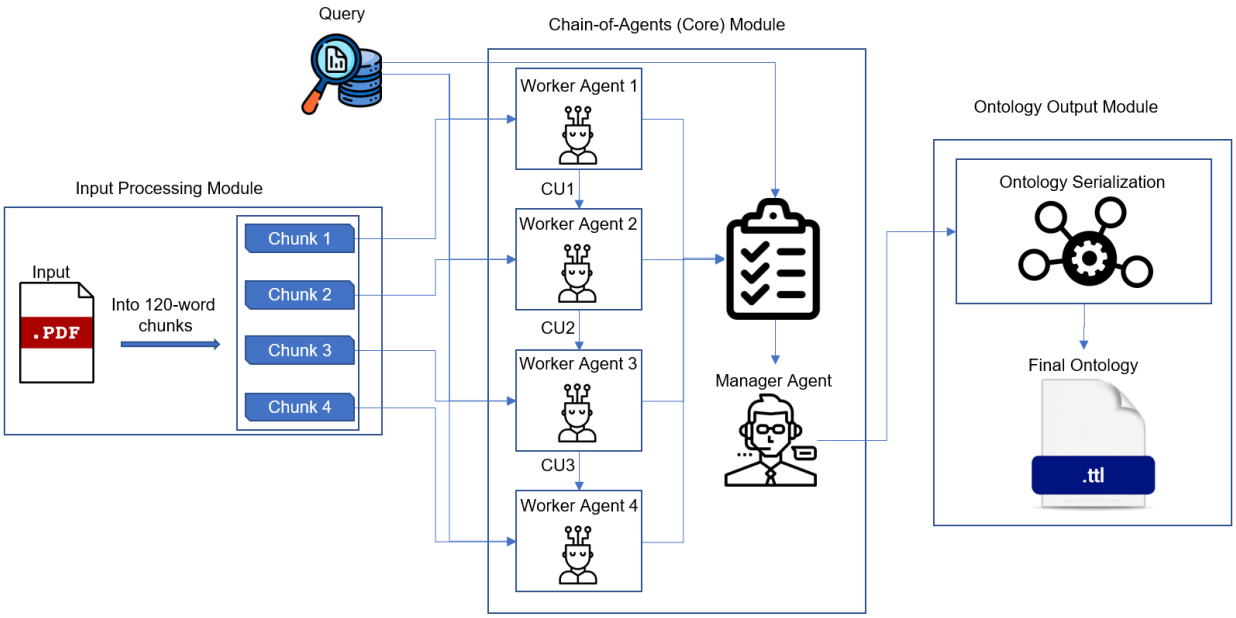


FIGURE 1 – CoA-Text2OWL Architecture

— Merging consistent concepts (classes) and relationships
— Resolving potential conflicts between partial ontologies
— Ensuring overall structural integrity of the final ontology

3.3.5 Output Generation

The final ontology is serialized into a Turtle (.ttl) file, adhering to standard RDF and OWL conventions.

This methodology leverages distributed processing of text chunks to generate a comprehensive ontology, with a focus on context preservation between chunks. The sequential processing of chunks allows for the construction of a coherent and potentially more comprehensive ontology than would be possible with isolated chunk processing.

4 Experiments and Results

4.1 Datasets

Our experimental evaluation leveraged pizza domain ontology dataset used in [7], where textual summaries of a pizza ontology were generated using LLMs. It comprises 10 structured textual descriptions covering diverse concepts related to pizza, including varieties, ingredient classifications, preparation methods, and compositional constraints. These descriptions inherently encode hierarchical relationships, taxonomic structures, and complex constraints, making them highly suitable for evaluating ontology extraction methods.

To adapt the dataset for ontology learning with our CoA-Text2OWL framework, we applied a segmentation process that divided texts into balanced chunks, ensuring compatibility with the agent-based architecture. No additional normalization or preprocessing (e.g., tokenization or linguistic modifications) was applied—only chunking was used to structure the input, maintaining a realistic and unaltered processing scenario.

The ontology used to generate the textual descriptions via LLMs is summarized the first colonne (Pizza Ontology) in Table 2 :

4.2 Baseline Methods

To evaluate the performance of our CoA-Text2OWL approach, we conducted a comparative analysis against a baseline follows the methodology presented by [7], where ontology extraction is performed using a single LLM instance without multi-agent coordination. This method, commonly referred to as direct Text2OWL, relies on a Mistral-7B model to generate OWL ontologies from text. The model is prompted to extract concepts, object properties, and hierarchical structures in a single pass. While this approach benefits from the expressive capabilities of large language models, it suffers from limitations in coherence, particularly when processing long texts. Without iterative refinement or structured communication between processing stages, the generated ontology may lack consistency and completeness.

By this comparison, we aim to demonstrate how our CoA-Text2OWL framework improves ontology learning by enabling multi-agent collaboration, allowing agents to iteratively refine and structure the extracted knowledge.

4.3 Evaluation Metrics

To assess the quality of the generated ontologies, we employ a set of well-established ontology evaluation metrics. These metrics allow us to measure how well the learned ontology aligns with a reference ontology by examining

its completeness, conciseness, and correctness. These three criteria are commonly used in ontology learning to ensure that the extracted knowledge is both meaningful and accurate [6].

Conciseness measures the extent to which the generated ontology avoids unnecessary or irrelevant elements. An ontology is considered concise if it does not contain extraneous concepts, relationships, or axioms that are not relevant to the domain. We compute conciseness as the proportion of elements in the generated ontology that are also present in the reference ontology, normalized by the total number of elements in the generated ontology.

$$Conciseness = \frac{|O_L \cap O_R|}{|O_L|}$$

where O_L represents the learned ontology and O_R represents the reference ontology.

Completeness evaluates whether the generated ontology sufficiently covers the domain knowledge present in the reference ontology. An ontology is considered complete if it includes all the key elements necessary for representing the target domain. Completeness is calculated as the proportion of elements in the reference ontology that are correctly captured in the learned ontology.

$$Completeness = \frac{|O_L \cap O_R|}{|O_R|}$$

A high completeness score indicates that the ontology learning approach has successfully captured most of the domain-relevant concepts and relationships.

Correctness is a composite metric that ensures that the generated ontology is both concise and complete. It is computed as the harmonic mean (F1-score) of conciseness and completeness.

$$Correctness = 2 \times \frac{Conciseness \times Completeness}{Conciseness + Completeness}$$

This metric provides a balanced assessment by penalizing cases where an ontology is either overly verbose (low conciseness) or missing key domain knowledge (low completeness).

These metrics allow us to quantify the effectiveness of our approach in generating ontologies that are both precise and comprehensive. By comparing the CoA-Text2OWL framework against baseline methods using these measures, we ensure a rigorous evaluation of the strengths and weaknesses of each approach in the ontology learning process.

4.4 Experimental Design

To evaluate the performance of our CoA-Text2OWL framework, we conducted controlled experiments with varying chunk sizes and agent chain lengths. These variations allowed us to analyze the impact of segmentation granularity and multi-agent collaboration on ontology learning.

We tested chain lengths from 2 to 5 worker agents, where each agent processed a text chunk before passing its output to the next. Chunk sizes ranged from 100 to 250 words, balancing granularity and context retention. The Mistral-7B model was used for all agent interactions with temperature = 0 to ensure deterministic outputs.

The best-performing configuration, maximizing accuracy and coherence in ontology extraction, is summarized in Table 1.

TABLE 1 – Best-performing experimental configuration for CoA-Text2OWL

Parameter	Value
Chunk Size	150
Chunks Number	3
Number of Worker Agents	3
Model	Mistral-7B
Temperature	0

4.5 Results

To evaluate the created ontologies, we begin with a quantitative analysis. Table 2 presents the counts of OWL named classes, object properties, named individuals, and `RDFS subClassOf` tuples for the reference Pizza Ontology, our CoA-Text2OWL approach, and the baseline Text2OWL method using Mistral-7B. This analysis provides an overview of the structure and richness of the generated ontologies.

Following this structural comparison, we assess the quality of the generated ontologies using the standard ontology evaluation metrics : conciseness, completeness, and correctness. These metrics are computed based on the alignment of extracted concepts, object properties, and hierarchical relations (subClassOf) with the reference ontology (Table 3).

Analysis of Results : The results in Table 3 demonstrate that CoA-Text2OWL significantly outperforms the baseline in object property extraction, achieving complete retrieval of relational structures, while the baseline fails in this aspect.

For class-level extraction, CoA-Text2OWL shows better precision by avoiding unnecessary elements, though its coverage of domain concepts remains only slightly higher than the baseline.

In hierarchical relations (subClassOf), the baseline method performs better, suggesting that CoA-Text2OWL struggles with dependency extraction in ontology structures. While it achieves improved correctness for class extraction, its performance is lower in maintaining subclass hierarchies.

Overall, CoA-Text2OWL excels in precision and relational structure identification but requires further refinement for capturing hierarchical relationships. These findings highlight the advantages of multi-agent ontology learning while pointing to areas for optimization.

TABLE 2 – Comparison of OWL axioms counts for the Pizza Ontology, CoA-Text2OWL, and baseline.

Count	Pizza Ontology	CoA-Text2OWL	Baseline
Named Classes	97	33	23
Object Properties	8	2	0
Named Individuals	5	0	0
Indiv. (ObjProps)	5	0	0
SubClassOf Tuples	141	32	23

TABLE 3 – Comparison of ontology learning performance between CoA-Text2OWL and the baseline method using Mistral-7B.

Metrics	**CoA-Text2OWL**	**Baseline [7]**
Classes conciseness	0.84	**1.00**
Classes completeness	**0.29**	0.25
Classes correctness	**0.43**	0.41
Object properties conciseness	**0.25**	0.00
Object properties completeness	**1.00**	0.00
Object properties correctness	**0.40**	0.00
SubClassOf pairs conciseness	0.12	**0.56**
SubClassOf pairs completeness	0.01	**0.05**
SubClassOf pairs correctness	0.02	**0.09**

5 Discussion

5.1 Performance Comparison

The evaluation of CoA-Text2OWL against the baseline Text2OWL using Mistral-7B reveals key insights into ontology extraction performance. The most significant improvement is observed in object property extraction, where CoA-Text2OWL achieves 100% completeness, whereas the baseline fails to extract any object properties. This suggests that the multi-agent processing framework is better suited for capturing relational structures in text.

For class-level metrics, CoA-Text2OWL exhibits higher conciseness, reducing unnecessary elements, while maintaining a completeness score slightly above the baseline. However, subclass relationships remain a challenge, with the baseline method achieving better `RDFS subClassOf` conciseness and completeness. This indicates that while CoA-Text2OWL enhances certain aspects of ontology learning, further refinements are needed to improve taxonomic structure extraction.

In summary, the primary advantage of CoA-Text2OWL lies in its ability to identify object properties, which the baseline approach fails to extract. However, both methods exhibit a significant reduction in hierarchical relationships, suggesting that ontology learning from text remains a challenging task, particularly for subclass extraction.

5.2 Limitations and Challenges

Despite its improvements over the baseline, CoA-Text2OWL faces several challenges :

1. Hierarchical Structure Loss : The reduced number of `RDFS subClassOf` relationships suggests that multi-agent processing does not fully preserve hierarchical dependencies, likely due to segmentation and inter-agent communication gaps.

2. Limited Concept Coverage : While CoA-Text2OWL outperforms the baseline in extracting object properties, it still captures significantly fewer named classes than the reference ontology. Expanding coverage without increasing noise remains an open problem.

3. LLM Stochasticity : Despite setting `temperature=0` for deterministic outputs, minor variations in generated ontologies were observed across repeated runs. This highlights the inherent non-deterministic nature of LLM-based ontology extraction and suggests that additional stabilization techniques, such as structured verification layers, could be beneficial.

4. Scalability Constraints : While CoA-Text2OWL demonstrates improvements over direct Text2OWL, its scalability to large, complex domains remains an open question. Future research should explore optimizations for processing efficiency and memory retention to enable broader applicability.

6 Conclusion and Future Work

In this work, we introduced CoA-Text2OWL, a multi-agent framework for ontology learning, and evaluated it against an LLM-based Text2OWL approach. Our results show that CoA-Text2OWL excels in object property extraction, achieving complete identification of relational structures, unlike the baseline.

At the class level, it produces a more concise ontology, effectively filtering irrelevant elements. However, its coverage of domain concepts remains only slightly higher than the baseline.

A key limitation is in subclass extraction, where the baseline performs better, suggesting challenges in maintaining hierarchical dependencies within the agent-based approach.

While CoA-Text2OWL demonstrates meaningful progress

in ontology learning, several challenges remain open for future research. An important direction involves refining the CoA framework to support ontology learning at each stage of the process. By involving it in term extraction, concept classification, relation identification, hierarchy construction, and axiom generation.

A promising enhancement is the introduction of task-specific workers within the CoA scheme. Rather than having a homogeneous sequence of agents, we propose an approach where each worker W_i is followed by a dedicated worker W_j that handles a specialized task such as extracting only object properties, identifying only hierarchical relationships, or validating only class-to-class mappings.

In summary, extending CoA-Text2OWL to a fully modular, multi-stage ontology learning framework with specialized agents for different tasks and dynamic agent coordination presents an exciting research direction. These advancements would enhance hierarchical extraction, relational accuracy, and scalability, ultimately improving the quality of automated ontology generation from text.

Acknowledgments

The authors thank the French company DAVI (Davi The Humanizers, Puteaux, France) for their support.

Références

[1] Deepak Bhaskar Acharya, Karthigeyan Kuppan, and B Divya. Agentic ai : Autonomous intelligence for complex goals–a comprehensive survey. *IEEE Access*, 2025.

[2] Pauline Armary, Cheikh Brahim El-Vaigh, Ouassila Labbani Narsis, and Christophe Nicolle. Ontology learning towards expressiveness : A survey. *Computer Science Review*, 56 :100693, 2025.

[3] Hamed Babaei Giglou, Jennifer D'Souza, Sameer Sadruddin, and Sören Auer. Llms4ol 2024 datasets : Toward ontology learning with large language models. In *Open Conference Proceedings*, volume 4, pages 17–30, 2024.

[4] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents : A survey of progress and challenges. *arXiv preprint arXiv :2402.01680*, 2024.

[5] Anirban Mukherjee and Hannah Hanwen Chang. Agentic ai : Expanding the algorithmic frontier of creative problem solving. *arXiv preprint arXiv :2502.00289*, 2025.

[6] Joe Raad and Christophe Cruz. A survey on ontology evaluation methods. In *International conference on knowledge engineering and ontology development*, volume 2, pages 179–186. SciTePress, 2015.

[7] Marion Schaeffer, Matthias Sesboüé, Léa Charbonnier, Nicolas Delestre, Jean-Philippe Kotowicz, and Cecilia Zanni-Merk. On the pertinence of llms for ontology learning. 2024.

[8] Yonadav Shavit, Sandhini Agarwal, Miles Brundage, Steven Adler, Cullen O'Keefe, Rosie Campbell, Teddy Lee, Pamela Mishkin, Tyna Eloundou, Alan Hickey, et al. Practices for governing agentic ai systems. *Research Paper, OpenAI*, 2023.

[9] Michael Wooldridge and Nicholas R Jennings. Intelligent agents : Theory and practice. *The knowledge engineering review*, 10(2) :115–152, 1995.

[10] Yusen Zhang, Ruoxi Sun, Yanfei Chen, Tomas Pfister, Rui Zhang, and Sercan Arik. Chain of agents : Large language models collaborating on long-context tasks. *Advances in Neural Information Processing Systems*, 37 :132208–132237, 2025.

Véhicules autonomes : simulation multi-agents pour explorer l'importance de la structuration de réseaux

Lorenzo Ferlito[1], Ophélie Bugnon[1], Thomas Matagne[1], Baudouin Dafflon[1,2]

[1] Ecov, ecov.fr, 44200 Nantes, France
[2]Disp-lab, UR4570, Université Lyon 1, France

prenom.nom@ecov.fr
prenom.nom@univ-lyon1.fr

Résumé

La mobilité domicile-travail constitue un enjeu majeur dans les zones périurbaines, où les alternatives aux transports individuels restent limitées. Les transports collectifs conventionnels (bus, cars, trains...) sont limités dans leur déploiement en raison d'une trop faible densité, empêchant un remplissage suffisant à leur exploitation comme en centre-ville. Parallèlement, avec les avancées récentes d'automatisation, des flottes de véhicules autonomes sont désormais exploitées commercialement, sous forme de taxis, à la fois aux États-Unis (Waymo) et en Chine (We-Ride, Pony.ai, Baidu Apollo...). Cependant, cette stratégie de déploiement du véhicule autonome repose sur le marché solvable, mais restreint, du taxi en zones denses, avec le déploiement de "robots taxis". Sur ce modèle, les véhicules autonomes n'apportent pas une alternative viable à la voiture individuelle qui reste la solution la plus pratique et économique, notamment en zone à faible densité de population où les alternatives sont très limitées. Pour répondre efficacement aux besoins de mobilité tout en garantissant une cohérence avec les dynamiques territoriales, une possibilité est de construire des réseaux de transports en s'appuyant sur les véhicules autonomes, avec une approche comparable à un transport collectif. Cette étude propose d'évaluer ces différentes approches à l'aide d'une simulation multi-agent, permettant de comparer leur efficacité sous l'angle de la performance énergétique. En modélisant les interactions entre usagers, véhicules et infrastructures, nous cherchons à identifier la solution la plus pertinente pour adapter un réseau de transport aux réalités des zones périurbaines.

Mots-clés

véhicules autonomes, covoiturage, simulation multi-agent

Abstract

Commuting to work is a major challenge in periurban areas, where alternatives to private transportation remain limited. Conventional public transportation systems (buses, trains...) are restricted in their deployment due to low population density, preventing a sufficient occupancy rate for efficient operation as seen in city centers. At the same time, with recent advances in automation, fleets of autonomous vehicles are now operating commercially, in the form of cabs, both in the USA (Waymo) and China (We-Ride, Pony.ai, Baidu Apollo...). However, this deployment strategy for autonomous vehicles is based on the solvent but limited cab market in densely populated areas, with the deployment of "robot cabs". In this model, autonomous vehicles do not provide a viable alternative to the private car, which remains the most practical and economical solution, especially in low-density areas where alternatives are highly limited. To effectively meet mobility needs while ensuring consistency with territorial dynamics, one possibility is to build transport networks based on autonomous vehicles, in a similar approach to public transportation. This study aims to assess these different approaches using a multi-agent simulation, allowing for a comparison of their efficiency from an energy performance perspective. By modeling interactions between users, vehicles and infrastructure, we seek to identify the most relevant solution for structuring a coherent and efficient transport network adapted to the realities of periurban areas.

Keywords

autonomous vehicles, carpooling, multi-agent simulations

1 Introduction

Au cours de la dernière décennie, la mobilité en zone périurbaine a connu une croissance significative. Cette évolution s'explique en partie par la pression foncière dans les centres-villes et la recherche d'un cadre de vie plus agréable, incitant de nombreux ménages à s'installer en périphérie. Toutefois, cette expansion résidentielle s'est accompagnée d'une augmentation des distances domicile-travail, complexifiant les déplacements quotidiens et accentuant les problèmes de congestion routière.

L'épidémie de COVID-19 a accéléré cette tendance d'exode urbain [19], poussant davantage de travailleurs à s'établir dans des zones éloignées des pôles d'emploi. Bien que le télétravail ait partiellement atténué la nécessité des trajets quotidiens, la reprise progressive de l'activité en présentiel remet en lumière les défis de la mobilité en péri-

phérie [3]. Aujourd'hui, les déplacements domicile-travail sont non seulement de plus en plus longs et coûteux pour les usagers, mais ils pèsent également sur l'environnement en raison de la forte dépendance à la voiture individuelle et de la consommation énergétique associée [8].

Face à ces enjeux, plusieurs solutions ont émergé pour optimiser les déplacements domicile-travail et réduire leur impact économique et environnemental. Parmi celles-ci, le covoiturage [2], le transport à la demande (TAD), l'autopartage [17] et le développement de réseaux de transport optimisés constituent des alternatives prometteuses. Cependant, l'efficacité réelle de ces solutions dépend de divers facteurs, notamment l'infrastructure disponible, le taux d'adhésion des usagers et l'organisation spatiale des zones concernées [15].

Dans ce contexte, la maturité des véhicules autonomes croissante apparaît comme un facteur clé dans la conception d'un système de transport périurbain. [11] propose une vision des flottes de véhicules autonomes partagés, en mettant en avant leur rôle transformateur dans les systèmes de transport.
Dans cette étude, nous proposons de simuler plusieurs scénarios de mobilité afin d'évaluer leur efficacité d'un point de vue énergétique. Trois configurations seront analysées :

1. Situation actuelle : déplacements en porte-à-porte en voiture individuelle.
2. Scénario de Mobilité Autonome à la Demande (MAD) : flotte de taxis autonomes répondant à la demande et inspirée de solutions comme Waymo.
3. Scénario de MAD Adaptatif (MADA) : flotte de vehicules autonomes intégrée et exploitée dans un réseau de lignes de transport collectif favorisant un taux de remplissage plus élevé des véhicules.

Ces simulations permettront d'identifier les leviers les plus efficaces pour améliorer la mobilité périurbaine, en tenant compte des aspects énergétiques et environnementaux. L'objectif est d'explorer des alternatives viables aux déplacements individuels afin de proposer des solutions plus durables et efficientes pour les trajets domicile-travail, et par extension la mobilité du quotidien.

Pour cette contribution, nous commencerons par une étude de la littérature et dresserons l'état des lieux de la mobilité aujourd'hui, puis nous proposerons un modèle de simulation basé sur plusieurs types d'agents en interaction et nous finirons par une évaluation des scénarios simulés.

2 État de l'art

L'optimisation des déplacements domicile-travail en zone périurbaine est un enjeu majeur en matière d'efficacité énergétique et de réduction des externalités négatives du transport. Plusieurs approches ont été explorées dans la littérature, notamment le covoiturage, l'autopartage et le transport à la demande, chacune présentant des avantages et des limites en fonction du contexte d'application. Pour chacun de ces modes, nous pouvons voir émerger une offre basée sur des véhicules autonomes en *free-floating* ou en gestion centralisée.

Le covoiturage planifié repose sur des plateformes numériques mettant en relation conducteurs et passagers en fonction d'horaires et d'itinéraires prédéfinis [12]. Cette approche, bien adaptée aux trajets longue distance (e.g., BlaBlaCar [1]), peine à s'imposer pour les déplacements quotidiens en raison de la rigidité des horaires, de la dispersion géographique des usagers et des contraintes d'engagement. L'absence de contrainte sur le trajet des conducteurs conduit à de fortes contraintes temporelles pour les passagers [10].

À l'inverse, le covoiturage spontané fonctionne sur des lignes structurées, analogues à un réseau de transport collectif, avec des points d'arrêt fixes et une organisation plus flexible dans le temps pour les usagers [6]. La mise en place d'une contrainte dans l'espace permet une liberté dans le temps : c'est la stratégie poursuivie par le monde des transports collectifs. En France, Ecov [2] est le principal opérateur de cette approche, intégrant le covoiturage dans une logique de service public. Il revient à l'opérateur de définir l'ensemble des facteurs qui peuvent permettre de construire une offre : structuration du réseau en fonction des flux du territoire, structuration du service, déploiement d'incitations financières et non financières, marketing et communication, opérations et exploitation (pilotage, monitoring, assistance en temps réel). Cette approche innovante de la voiture vue comme un réseau de transports collectifs a montré des résultats probants [16], mais reste encore émergente et fait face à des défis importants.

L'autopartage permet, quant à lui, un accès ponctuel à un véhicule sans nécessité de propriété, selon deux modèles principaux : en boucle (retour du véhicule au point de départ, e.g., Citiz) et en *free-floating* (libre dépôt dans une zone définie, e.g., ShareNow) [14]. En diminuant le nombre de véhicules en circulation et en favorisant leur usage collectif, l'autopartage contribue à une meilleure utilisation des ressources et à une réduction des émissions liées à la production automobile mais peine à contribuer aux efforts de décongestion. De plus, bien qu'efficace en milieu urbain dense, l'autopartage spontané rencontre des obstacles en périphérie : faible densité d'usagers, coût opérationnel élevé et logistique complexe pour la redistribution des véhicules [13].

Enfin, le transport à la demande permet d'optimiser les trajets en adaptant l'offre à la demande, offrant ainsi une flexibilité supérieure aux lignes de bus fixes [9]. Il se décline en plusieurs modèles :

— TAD zonal : desserte libre dans une aire définie (ex :

1. blablacar.com
2. ecov.fr

Flexigo).
— TAD en rabattement : connexion avec un réseau structurant (ex : Résa'TAD).
— TAD dynamique : ajustement en temps réel des itinéraires (ex : Kutsuplus).

Bien qu'il permette de réduire le nombre de véhicules en circulation, son impact environnemental dépend fortement de l'optimisation des itinéraires pour minimiser les trajets à vide et du taux de mutualisation des courses, en particulier dans les zones à faible densité de population [7]. De plus, le TAD reste limité par un coût d'exploitation élevé et une dépendance aux subventions publiques, rendant son déploiement dépendant des politiques locales de mobilité [3]. L'intégration de véhicules autonomes, en supprimant le coût du conducteur et en optimisant les déplacements d'une flotte favorisant les trajets partagés, est une piste étudiée pour améliorer la viabilité de ce modèle en périurbain [18].

Chaque approche présente des forces et des limites en fonction des contraintes périurbaines. Le covoiturage spontané et le TAD se rapprochent des modèles de transport collectif, tandis que l'autopartage repose sur une logique d'optimisation des ressources individuelles. Dans ce contexte très compétitif, nous proposons de modéliser et de simuler les déplacements domicile-travail autour d'un service de mobilité à la demande exploitant une flotte de véhicules autonomes pour évaluer les optimisations possibles dans une optique d'organisation de l'offre de transport favorisant la mutualisation des trajets.

3 Modèle proposé

Notre modèle repose sur une approche multi-agent où différents types d'agents interagissent au sein d'un environnement géographique simulé. Ce cadre permet d'étudier les dynamiques de mobilité en intégrant diverses formes de transport et d'organisation des trajets.

L'**environnement** constitue le cadre spatial de la simulation et joue plusieurs rôles essentiels. Le premier rôle est la représentation de l'espace sous forme d'une carte où se déroulent les déplacements des agents. L'environnement joue aussi le rôle de tableau blanc où sont calculées les informations perçues pour chaque agent, facilitant ainsi leur prise de décision. Enfin, il collecte les décisions individuelles des agents, les intègre de manière cohérente et garantit le bon déroulement de la simulation en assurant la compatibilité des actions.

Quatre catégories d'agents interagissent au sein de cet environnement : Conducteurs, Véhicules autonomes, Passagers et Orchestrateurs. Les interactions entre ces agents peuvent être résumées par la figure 1.

L'**agent conducteur** représente un utilisateur humain qui effectue un trajet domicile-travail en voiture personnelle. Cet agent possède un véhicule et est capable de transporter un ou plusieurs passagers. Le comportement de cet agent

3. https ://wiki.lafabriquedesmobilites.fr

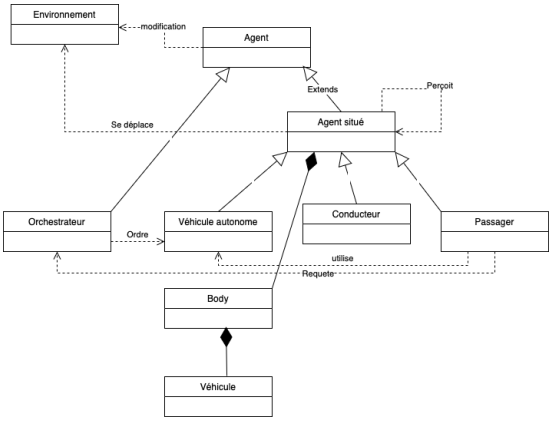


FIGURE 1 – Interactions entre agents

est simple : il effectue le trajet directement de porte à porte, sans aucune modification ou optimisation du parcours. Le taux d'occupation moyen de son véhicule est de 1,07 pour les trajets domicile-travail [1].

L'**agent passager** représente un individu souhaitant effectuer un trajet domicile-travail sans posséder de véhicule personnel. Cet agent utilise un service de transport, tel qu'un réseau de covoiturage ou un système de véhicules autonomes, pour se déplacer d'un point à un autre. Contrairement aux autres agents, l'agent passager ne possède pas de voiture et doit attendre l'affectation d'un véhicule pour réaliser son trajet. L'agent passager est défini par plusieurs caractéristiques essentielles, qui influencent son comportement dans le cadre de la simulation : temps d'attente, nombre de correspondances, demande de trajet. L'agent passager peut accepter ou refuser une offre de transport en fonction de ses contraintes personnelles. Si les conditions proposées (temps d'attente, correspondances, etc.) ne correspondent pas à ses attentes, il peut choisir de ne pas utiliser le service ou de renouveler sa demande jusqu'à ce qu'une offre satisfaisante soit trouvée. La présence d'un agent passager se répercute dans l'environnement à travers son poids défini par :

$$\text{Poids}(x, y) = \sum_{i \in \mathcal{B}(x,y,r)} P_{\text{agent},i}$$

Où $\mathcal{B}(x, y, r)$ représente l'ensemble des agents situés dans un rayon r autour de la position (x, y) et $P_{\text{agent},i}$ est défini par :

$$P_{\text{agent},i} = T_{a_i} + \text{Nb}_{\text{voisin},a_i}$$

Où :
— $T_{a_i} = exp(t)$ ou t est le temps d'attente pour $0 < t \leq 10$ (secondes)
— $\text{Nb}_{\text{voisin},a_i}$ est le nombre d'agents partageant la même étape voisins de i à la position (x, y) dans un rayon r.

Ce poids est utilisé par l'orchestrateur et les véhicules autonomes pour prendre leurs décisions.

L'**agent orchestrateur** de véhicules autonomes est responsable de la gestion et de la coordination de la flotte de véhicules autonomes. Son rôle principal est d'assurer que toutes les demandes de transport des passagers soient traitées efficacement, en optimisant l'affectation des courses aux véhicules disponibles. L'orchestrateur prend en compte des critères tels que la localisation des passagers, la disponibilité des véhicules et la capacité de transport pour minimiser les temps d'attente et les distances parcourues pour calculer un $Cout$ qui permettra de trier et seletionner les meilleurs couples "véhicule/passager". Ce processus de gestion est basé sur un algorithme de répartition en temps réel qui permet à l'orchestrateur de maximiser l'efficacité globale du système de covoiturage. La prise de décision de l'orchestrateur suit la logique suivante :

Algorithm 1 Affectation d'un véhicule a une demande

$D = \{D_1, D_2, \ldots, D_n\}$
$V = \{V_1, V_2, \ldots, V_n\}$
for chaque demande D_i dans D **do**
  for chaque véhicule V_i dans V **do**
    $c_i = \mathrm{C}(D_i, V_i)$
    $Cout \leftarrow (c_i, D_i, V_i)$
  end for
end for
$Cout$ = tri($Cout$)
retour $Cout[0]$

La fonction de coût est définie par :

$$C(D, V) = c + 1/(D_{v,p} + D_{v,d} + d_{p,demande})$$

où

- c est la capacité du véhicule
- $D_{v,p}$ est la distance au carrée de prise de passager si le véhicule est réservé
- $D_{v,d}$ est la distance au carrée de dépose de passager si le véhicule est occupé
- $d_{p,demande}$ est la distance au carrée pour prendre la nouvelle demande.

Cette fonction de coût est un moyen de trier les couple (passager / véhicule) afin de minimiser les critères de déplacement comme le temps d'attente, le nombre de kilometres parcourus, etc.

L'orchestrateur peut gérer un réseau de transport (scénario 3), qui est représenté par un graphe orienté. Ce graphe modélise les chemins et les connexions possibles entre les différents nœuds représentant les zones de transport (par exemple, des hub, des points de départ/arrivée ou des zones de passage). L'une des principales fonctions de l'orchestrateur est d'assigner un poids à chaque requête de passager.

L'**agent voiture autonome** modélise un véhicule autonome, tel un taxi autonome de type Waymo. Contrairement à l'agent conducteur classique, l'agent voiture autonome est déployé de manière dynamique dans l'environnement de la simulation au début du processus. Une fois positionné, cet agent reçoit des "courses" d'un orchestrateur central qui l'affecte aux passagers en fonction des demandes qui émergent au cours de la simulation. L'agent voiture autonome est équipé d'un module de décision autonome, lui permettant de prendre une course sans intervention humaine lorsque des passagers sont détectés dans son champ de vision et que le véhicule est libre. Ce module de décision est conçu pour maximiser l'efficacité du système en minimisant les trajets à vide et en optimisant les trajets partagés. Cependant, ce comportement autonome peut être ajusté ou interrompu par l'orchestrateur si nécessaire, afin de répondre à des situations spécifiques ou d'optimiser la gestion globale du trafic. Dans cet environnement, les véhicules autonomes sont en compétition pour répondre aux requêtes des passagers. Chaque véhicule tente de maximiser son efficacité en répondant aux demandes les plus « rentables » en termes de poids, ce qui reflète la priorité des passagers et la densité de la demande. Lorsqu'un véhicule autonome accepte une requête, il retire le poids associé au nœud correspondant dans le graphe du réseau. Ce retrait est une récompense pour le véhicule, qui est alors affecté à ce passager et commence son trajet vers la destination. Le véhicule peut ensuite se déplacer vers un autre nœud en fonction des demandes restantes, en fonction de son trajet et des zones à desservir.

Ce modèle permet ainsi d'analyser les interactions entre ces différentes entités et d'évaluer l'impact de divers scénarios de mobilité sur les trajets domicile-travail.

4 Expérimentations de simulation

4.1 Outil de simulation utilisé

Dans le cadre de cette simulation, nous utilisons une version dérivée du framework PAMELA [5], un outil open-source de simulation multi-agent (MAS). Ce framework est conçu pour faciliter la création de simulations distribuées, tout en étant accessible aux utilisateurs débutants et suffisamment flexible pour des scénarios plus complexes. Il a été utilisé par le passé dans des études sur la mobilité [4].

La version dérivée de PAMELA que nous utilisons intègre des bibliothèques graphiques supplémentaires pour améliorer la visualisation des résultats. Nous avons ainsi ajouté des outils comme DeckGL [4] et Lumagl [5], qui permettent de représenter graphiquement les trajectoires des véhicules autonomes, les flux de passagers, ainsi que les dynamiques du réseau de transport.

4.2 Scénarios de mobilité autonome à la demande

4.2.1 *Scénario 1 :* Porte à porte en véhicules privés

Le premier scénario repose sur l'extraction de la base de données MobPro [6] de l'INSEE pour le département du 69

4. deck.gl
5. luma.gl
6. Données de mobilité professionnelle des individus en France en 2020 en OpenData sur le site de l'INSEE.

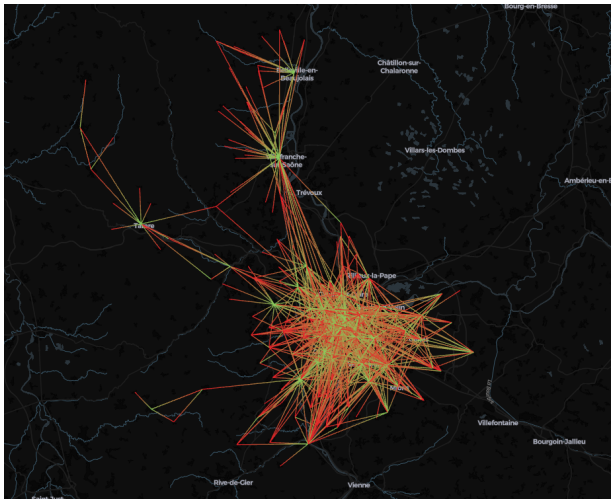


FIGURE 2 – Agrégation des trajets domicile-travail autour de Lyon.

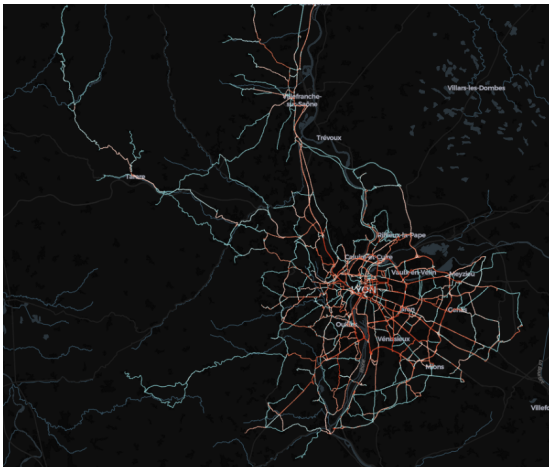


FIGURE 3 – Modélisation du réseau et des portions communes aux déplacements domicile-travail

(Rhône), afin de récupérer les flux de mobilité intercommunaux. Après avoir extrait les données brutes, un premier filtre est appliqué pour éliminer les petits flux, en nombre insuffisant pour générer un modèle significatif de transport, ainsi que les flux non motorisés. Ensuite, un autre filtre est appliqué pour exclure les origines-destinations (OD) trop longues, qui ne représenteraient pas des trajets réalisables dans le cadre d'un transport quotidien. Enfin, à partir de la matrice origine-destination filtrée, un agent est créé pour chaque déplacement individuel. Ces déplacements peuvent être représentés par la figure 2.

4.2.2 *Scénario 2 :* Mobilité Autonome à la Demande (type Waymo)

Le deuxième scénario repose sur les mêmes données de mobilité que dans le scénario 1, extraites de la base MobPro de l'INSEE pour le département du 69. Contrairement au scénario 1, ici, tous les agents sont des passagers qui utilisent un service de véhicules autonomes pour leurs trajets domicile-travail. Dans ce scénario, les véhicules autonomes fonctionnent comme un service de taxi/VTC. L'orchestrateur n'a pas de représentation du réseau de transport sous forme de graphe, ce qui signifie qu'il ne dispose pas d'informations préalables sur les itinéraires ou la structure d'un réseau de transport. Les véhicules autonomes sont dispatchés aléatoirement dans l'environnement au début de la simulation, et doivent naviguer de manière autonome pour récupérer les passagers et les transporter à leurs destinations. L'orchestrateur est responsable de la gestion des requêtes en temps réel, mais ne contrôle pas les trajets des véhicules de manière centralisée.

4.2.3 *Scénario 3 :* Système Adaptatif de Mobilité Autonome à la Demande exploitant un réseau

Le scénario 3 reprend les éléments du scénario 2, en fournissant à l'orchestrateur un graphe représentant un réseau de transport. Ce graphe permet à l'orchestrateur de modéliser les demandes en termes de chemins passant par les nœuds et de dispatcher les véhicules autonomes dans l'environnement de manière adaptative en tenant compte de ce réseau. Ainsi, l'orchestrateur obtient une gestion plus optimisée des déplacements des véhicules, permettant de densifier les trajets des agents passagers selon des axes prédéfinis. Si le chemin calculé par l'orchestrateur implique un trop grand détour pour le passager, celui-ci se verra affecté un véhicule autonome effectuant un service porte-à-porte type scénario 2.

Le réseau des portions communes aux déplacements domicile-travail autour de Lyon est présenté (figure 3). Le niveau de détail de ce réseau étant trop fin pour pouvoir organiser efficacement du covoiturage ou du partage de trajet, un prétraitement est effectué pour fournir à l'orchestrateur un réseau plus grossier et plus facile à exploiter. Seuls les axes avec un flux suffisamment important sont gardés dans la structure du réseau final. La figure 4 représente le réseau fourni à l'orchestrateur une fois simplifié.

4.2.4 Données d'entrées

Les scénarios étudiés sont résumés avec leurs variables d'entrée dans la première partie de la table 1. Les trois scénarios étudiés cherchent à remplir les besoins de déplacements d'un nombre de trajets équivalents (N_{trajet}) dont les couples origine-destination sont issus des flux de déplacements professionnels présentés lors de la description du scénario 1. Le nombre de passagers (N_{pass}) correspond au nombre d'agents conducteurs et au nombre d'agents passagers simulés, respectivement dans le scénario 1 et les scénarios 2 et 3, auquel s'ajoute un compagnon avec une probabilité de 0.3 permettant d'atteindre un taux d'occupation des véhicules de 1.07 tel qu'observé par les statistiques de l'INSEE pour les déplacements domiciles-travail. Il est précisé qu'un agent et son compagnon effectuent le même besoin de déplacement correspondant à un trajet commun.

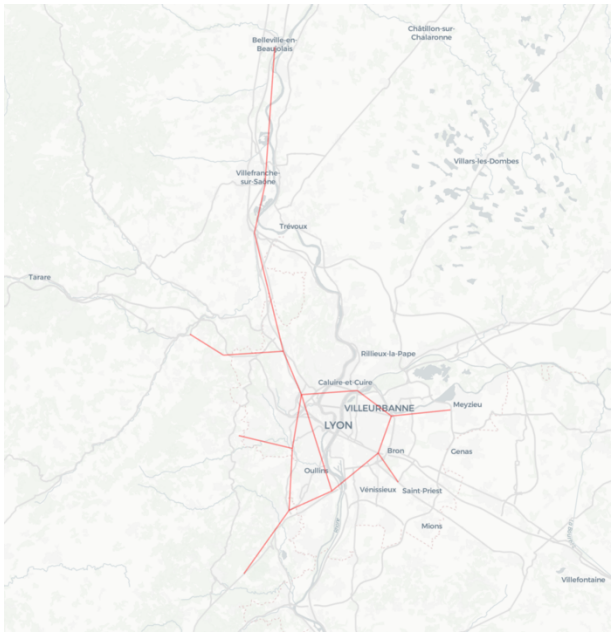


FIGURE 4 – Réseau simplifié

4.2.5 Métriques

Plusieurs métriques couramment utilisées dans la littérature ont été définies afin d'évaluer l'efficacité et la performance du système de transport simulé.

- N_{veh} : le nombre de véhicules utilisés pour effectuer les trajets
- τ_{succes} : le pourcentage de passagers arrivant à la destination voulue (équivalent à la fiabilité du service)
- τ_{occup} : le taux d'occupation moyen des véhicules pondéré par les distances parcourues
- $d_{\text{tot,veh}}$: le nombre de kilomètres parcourus par les véhicules
- $d_{\text{vide,veh}}$: le nombre de kilomètres parcourus par les véhicules à vide
- $\tau_{\text{d,vide}}$: le pourcentage de kilomètres parcourus par les véhicules à vide
- $\bar{t}_{\text{pass}}$: le temps de déplacement moyen des passagers
- $\bar{t}_{\text{att,pass}}$: le temps d'attente moyen des passagers

Ces métriques permettront de comparer les scénarios entre eux et de valider le bon fonctionnement du modèle proposé.

4.3 Résultats

Dans les trois scénarios simulés, l'analyse a principalement porté sur trois indicateurs clés : le taux d'occupation des véhicules, le taux de succès des trajets effectués et le temps moyen de déplacements des passagers. Ces métriques permettent d'évaluer l'efficacité du système en termes d'optimisation des trajets et de satisfaction des demandes de transport. En complément, nous avons également

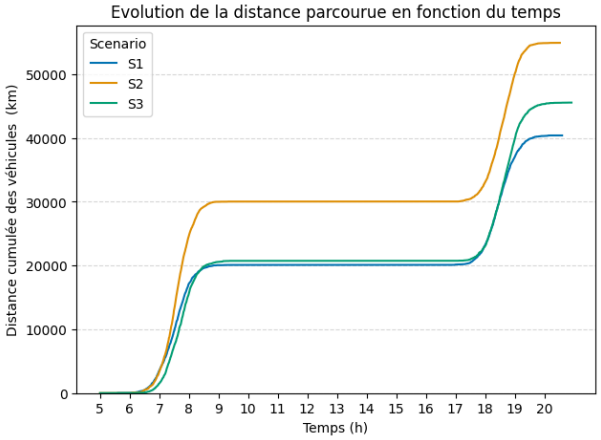


FIGURE 5 – Cumul des distances parcourues par les véhicules en fonction du temps par scénario

examiné le nombre total de kilomètres parcourus, en distinguant les trajets avec passagers à bord et les trajets à vide.

La figure 5 illustre l'évolution de la distance cumulée parcourue en fonction du temps pour les trois scénarios simulés. On observe deux phases principales d'augmentation de la distance : la première entre 6h et 9h, correspondant au pic des trajets domicile-travail du matin, et la seconde entre 17h et 20h, reflétant les déplacements du retour. Entre ces périodes, la distance totale parcourue reste relativement stable. Le scénario S2 (MAD type Waymo) affiche la plus grande distance totale, suivi de S3 (MAD adaptatif) et enfin S1 (véhicules privés). Cette tendance suggère que les véhicules autonomes parcourent davantage de kilomètres en raison de trajets à vide liés à la redistribution des véhicules dans l'environnement simulé.

Les graphiques de la figure 6 détaillent l'évolution de la distance cumulée parcourue en fonction du temps pour les trois scénarios simulés, en distinguant les kilomètres effectués avec passagers et ceux réalisés à vide. Dans le scénario S1 (véhicules privés), la totalité des trajets s'effectue avec des passagers, ce qui est cohérent avec un mode de transport individuel. En revanche, dans les scénarios S2 (MAD type Waymo) et S3 (MAD adaptatif), une proportion significative de kilomètres est parcourue à vide, notamment lors des phases de redistribution des véhicules. Le scénario S3, bien que générant également des trajets à vide, minimise ceux-ci grâce à la structuration d'un réseau regroupant les passagers à des points spécifiques.

Les résultats montrés dans la table 1 montrent que le scénario S3 (MAD Adaptatif) optimise le taux d'occupation des véhicules (1,75) tout en réduisant la distance parcourue à vide (12%), offrant ainsi une meilleure efficacité que le scénario S2 (MAD type Waymo), qui génère davantage de trajets à vide (27%). En termes de confort des passagers au

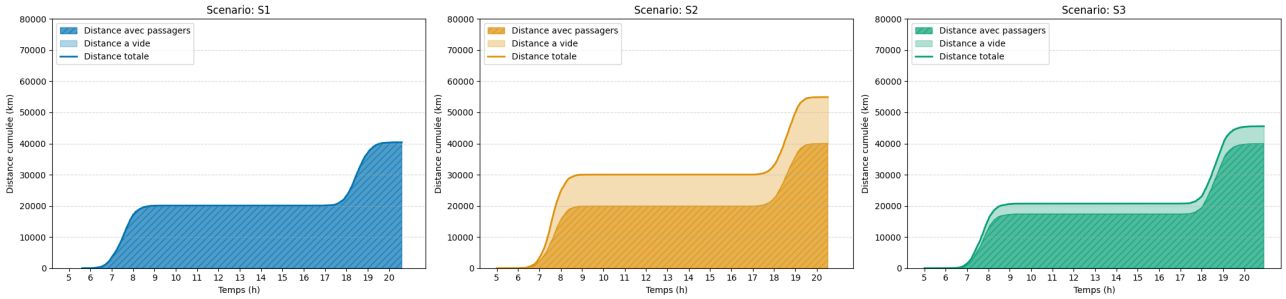


FIGURE 6 – Cumul des distance parcourues par les véhicules à vide et avec passagers en fonction du temps selon chaque scénario

	S1	S2	S3
Mode	Véhicule privé	MAD	MADA
N_{trajet}	3242	3242	3242
N_{pass}	3462	3491	3471
N_{veh}	1621	443	675
τ_{succes} (%)	100	99	99
τ_{occup}	1.07	0.78	1.75
$d_{\text{tot,veh}}$ (km)	40400	54922	45551
$d_{\text{vide,veh}}$ (km)	0	14900	5560
$\tau_{\text{dist,vide}}$ (%)	0	27	12
$\bar{t}_{\text{pass}}$ (min)	10	15	17
$\bar{t}_{\text{att,pass}}$ (min)	0	4	4

TABLE 1 – Données d'entrées et résultats des simulations multi-agents par scénario de mobilité

quotidien, le temps d'attente dû à l'utilisation des services de VA dans S2 et S3 induit une augmentation de 40% du temps de trajet par rapport à l'emploi de véhicule privé. Il est toutefois intéressant de constater que ce temps d'attente n'augmente pas dans le scénario S3 tout en maintenant un niveau de service optimal. Dans S3, le temps de trajet augmente par rapport à S2, traduisant les détours imposés par les itinéraires définis par le réseau. Par ailleurs, dans ce scénario, les premiers et derniers kilomètres sont effectués à pied si la distance au réseau le permet. Ces résultats soulignent l'intérêt d'une approche structurée pour améliorer l'efficacité énergétique et réduire les déplacements inutiles.

5 Conclusion

Dans cette étude, nous avons cherché à évaluer plusieurs politiques de déplacement, allant des solutions les plus flexibles aux plus efficaces. En utilisant un modèle basé sur des agents, nous avons simulé ces politiques dans un environnement performant, permettant d'obtenir des résultats concrets et exploitables. Les différentes politiques ont été comparées à l'aide de métriques précises, telles que le taux de remplissage des véhicules, reflétant le nombre de passagers transportés en moyenne par kilomètre parcouru par les véhicules, le nombre de kilomètres parcourus et l'efficacité du système global.

Nos résultats en simulation concernant les scénarios de mobilité autonome à la demande (MAD et MADA) montrent qu'une politique de déplacement s'appuyant sur un réseau structuré présente un avantage significatif en termes d'optimisation des trajets, de réduction des distances parcourues par les véhicules avec passagers, et se rapproche d'un service de transport collectif. De la même façon qu'un réseau urbain, la structuration d'un réseau de mobilité sur le système routier en zone périurbaine facilite le covoiturage grâce à un service de mobilité à la demande construit autour de grandes lignes de transport collectives appelé "Mobility as a Network" (MaaN).

Cependant, l'arrivée certaine des véhicules autonomes implique également une réflexion sur leur modèle d'exploitation et leur impact sur les performances et habitudes de mobilité. Utilisés comme un service de taxi/VTC, ils n'offrent pas une alternative viable à la voiture individuelle tant d'un point de vue économique pour l'usager que pour leur impact négatif sur le taux de remplissage des véhicules, en dessous de celui actuellement observé pour les trajets domicile-travail. Nos futurs travaux nécessiteront donc de construire un réseau de transport rivalisant, sur le plan économique et écologique, avec les avantages des trajets porte-à-porte pour répondre à la problématique d'acceptabilité par les usagers. Pour cela, il sera nécessaire de concentrer l'effort tant sur la conception d'un réseau adapté que sur l'optimisation de la gestion de la flotte de véhicules. Il sera également nécessaire d'orienter les travaux sur la réduction du temps d'attente des passagers, critère ayant un impact important sur l'adoption d'un système de transport publics. Enfin, il sera nécessaire de réfléchir à l'accompagnement au changement de comportement dans une population. Plusieurs outils comme les voies réservées au covoiturage (VR2+) et des systèmes d'incitation peuvent être identifiés comme leviers à ce changement. Il serait intéressant d'utiliser les données de l'INSEE autre que MobPro comme les Enquêtes ménages déplacement pour constuire

une population synthétique plus hétérogene.

Références

[1] Jimmy Armoogum, Élisabeth Bouffard-Savary, Yann Caenen, Christine Couderc, Jérémy Courel, François Delisle, Philippe Duprat, Laurent Fouin, Dominique François, Marie-Odile Gascon, Dominique Godineau, Richard Grimal, Jean-Paul Hubert, Yan Le Gal, Josiane Le Guennec, Thomas Le Jeannic, Zahia Longuar, Jean-Loup Madre, Jean-Pierre Nicolas, Dominique Pallez, Francis Papon, Christelle Paulo, Bernard Quételard, Amélie Ranty, Marina Robin, Sophie Roux, Sébastien Seguin, Thierry Siméon, Régis de Solère, Bruno Tisserand, and Damien Verry. *La mobilité des Français, panorama issu de l'enquête nationale transports et déplacements 2008*. « La Revue » du Service de l'Observation et des Statistiques (SOeS) du Commissariat Général au Développement Durable (CGDD), Ministère de l'écologie, Paris, 2010.

[2] Natalia Boichuk. Sustainability : carpooling as a component of smart mobility. *Education excellence and innovation management : A*, 2025.

[3] Clara Bosch-Checa, Edgar Lorenzo-Sáez, Maria Joaquina Porres de la Haza, Victoria Lerma-Arce, and Eloina Coll-Aliaga. Evaluation of the accessibility to urban mobility services with high spatial resolution—case study : Valencia (spain). *Applied Sciences*, 15(3) :1382, 2025.

[4] Baudouin Dafflon, Maxime Guériau, Yacine Ouzrout, and Ivana Dusparic. Emerging micro-communities for ride-sharing enabled mobility-on-demand systems. In *ATT@ ECAI*, pages 63–70, 2020.

[5] Baudouin Dafflon, Maxime Guériau, Yacine Ouzrout, and Sylvain Touchard. Pamela : a generic and light multi-agent platform. In *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 1422–1426. IEEE, 2021.

[6] Yuval Efron, Shyamal Patel, and Cliff Stein. A simple algorithm for dynamic carpooling with recourse. In *2025 Symposium on Simplicity in Algorithms (SOSA)*, pages 196–201. SIAM, 2025.

[7] France Mobilités. Étude sur le recensement et la conceptualisation du transport à la demande, 2023. Consulté le 18 mars 2025.

[8] Florence Goffette-Nagot, Carl Gaigné, Stéphane Riou, and Yves Schaeffer. Coûts environnementaux du développement périurbain : impact des configurations urbaines et des choix résidentiels. *Vivre en ville hors des villes, Aubertel, P.(dir.)*, pages p–51, 2014.

[9] Todd Hansen and Ipek Nese Sener. Strangers on this road we are on : a literature review of pooling in on-demand mobility services. *Transportation Research Record*, 2677(3) :1368–1381, 2023.

[10] Iftikhar Hussain, Luk Knapen, Stephane Galland, Tom Bellemans, Davy Janssens, Geert Wets, et al. Agent-based simulation model for long-term carpooling : Effect of activity planning constraints. *Procedia Computer Science*, 52 :412–419, 2015.

[11] Christos Karolemeas, Stefanos Tsigdinos, Evi Moschou, and Konstantinos Kepaptsoglou. Shared autonomous vehicles and agent based models : a review of methods and impacts. *European Transport Research Review*, 16(1) :25, 2024.

[12] Alix Le Goff, Guillaume Monchambert, and Martin Koning. Effects of numerical platforms on individual choices and social welfare : The case of short-distance carpooling. *Economics of Transportation*, 41 :100389, 2025.

[13] Cheng Li, David Parker, and Qi Hao. Optimal online dispatch for high-capacity shared autonomous mobility-on-demand systems. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 779–785. IEEE, 2021.

[14] Santhanakrishnan Narayanan and Constantinos Antoniou. Shared mobility services towards mobility as a service (maas) : What, who and when ? *Transportation research part A : policy and practice*, 168 :103581, 2023.

[15] Isac Olave-Cruz, Maïté Stéphan, Alexandre Volle, and Dianzhuo Zhu. Does carpooling reduce carbon emissions ? the effect of environmental policies in france. *Environmental and Resource Economics*, pages 1–34, 2025.

[16] Jean-Baptiste Ray. Planning a real-time ridesharing network : critical mass and role of transfers. In *5th Conference Transport Research Arena (TRA)*, 2014.

[17] Bo Sun, Shukai Chen, and Qiang Meng. Optimizing first-and-last-mile ridesharing services with a heterogeneous vehicle fleet and time-dependent travel times. *Transportation Research Part E : Logistics and Transportation Review*, 193 :103847, 2025.

[18] Frank Urbansky. New mobility-roboshuttles and the 15-minute city. *MTZ worldwide*, 86(2) :8–13, 2025.

[19] Stephan D Whitaker. Did the covid-19 pandemic cause an urban exodus ? *Cleveland Fed District Data Brief*, (20210205), 2021.